Oona VOICAN, PhDStudent
E-mail: oona.voican@yahoo.com
Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies

USING DATA MINING METHODS TO SOLVE CLASSIFICATION PROBLEMS IN FINANCIAL-BANKING INSTITUTIONS

Abstract*. Data Mining refers to the analysis of large amounts of data stored in computers. The Big Data era is already present, with current sources indicating that more data have been created over the last two years than they have been generated throughout the entire human history. Big Data involves data sets so large that traditional data analysis methods are no longer usable due to the huge amount of data. Lacking or ignoring the data structure is an extremely important aspect, even more important than size, in data analysis, transformation, innovation and value for business. It was said that there are important data about analysis rather than data themselves.*

Keywords*: classification, data mining, finance, customer segmentation, panel data.*

JEL Classification: C10, G20.

1. Introduction

Davenport (2014) considers that at present the following futures define Big Data: data too large to use a single computer server; data not structured enough to be organized in a rows and column database; data that change continuously cannot be entered into a statistic data repository. Data will become a genuine raw material for business, a vital economic resource and a prime source of value for any business.

Functioning of different economic systems (enterprise, banks, national economy), emails used in electronic communication, blogs appearing on the website, automation systems and business leadership lead to huge volumes of structured and semi-structured information within any organization. External data are also available in large volumes, many of which are provided free of charge via the Internet, but also from commercial vendors. There are also data that can be obtain from the social media (Facebook, Twitter, LinkedIn).

Data Mining is not limited to business. In elections in the US, but also in other countries of the world, the exploitation of potential voter data is practiced to predict their preferences in election(Bradley and Mangasarian, 1998).Data Mining

DOI: 10.24818/18423264/54.1.20.11

is increasingly being used in the medical field, from diagnosing patient suffering to identifying and disseminating best practices (Caiand Zhu, 2015).

The commercial use of Data Mining is becoming more and more dominant. Data Mining is increasingly used by banks to track credit cards in order to detect fraud. Data Mining is used to improve food safety as well as in tourism and in justice. Micromarketing targets small groups of customers with a high response to companies ‘offers. Consumer and lifestyle data are widely available, allowing personalized marketing campaigns to be customized. This is enabled by developing customer profiles, identifying those subsets of customers that are most likely attracted by the products offered, and thus lead to profitable acquisitions for the business, as well as targeting, determining the characteristic of the most profitable customers.

2. Data Mining Requirements

Data Mining requires identification of a problem, along with data collection that can lead to a better understanding, and computer models to provide statistical data or other means of analysis. Data access is also required. Current management systems include data repositories and data marts quite often. This data mart concept is used to efficiently manage large amounts of data. Analyses performed within them are performed on smaller sets of data, so they can be organized into online analytical processing systems. Thus, data volumes generated by cash registers, scanning of products sold or stored in warehouses and databases specific to each company are explored, analyzed, minimized and reused. Searches in data repositories and data marts are performed using different models and used to estimate sales, measure advertising responsiveness and profit achieved by any company.

Classical statistical approaches are fundamental to Data Mining. Automatic AI methods are also used. Data Mining tools must be versatile, scalable, capable of accurately predicting the links between actions and results, and capable of automatic deployment. Versatile refers to the ability of the tool to apply a wide variety of models. Scalable methods imply that if the tools operate on a small volume of data, they should also work on a larger set of data.

Automation is useful, but its application is relative. Some analytical functions are often automated, but human presence is required before procedures are applied. In fact, the analyst's judgment is extremely important for the successful implementation of data processing. Data Mining expands rapidly, with many business benefits. Two of the most profitable areas of application are customer segmentation by marketing organizations to identify those people with greater likelihood of positively responding to different forms of marketing media, namely the banking activity that uses Data Mining to determine more precisely the likelihood of people responding favorably to the offers of the various services offered. Many companies use this technology to identify their major customers so they can provide them with the services they need to keep them.

DOI: 10.24818/18423264/54.1.20.11

3. Data Mining in Business

Data Mining is very effective in many business areas. The key is to find ways or information that can be used in a concrete way to improve business profitability. Some of the first applications were in retail, especially in the form of market basket analysis. Table 1 lists some general areas of application that we will discuss later. The data in this table are more indicative than exhaustive.

Table 1. Data Mining Applications

Scope	Applications	Comments
Retail trade	Cross selling Developing and maintaining customer loyalty	Positioning by affinity Determining more customer products
Bank	Customer Relationship Management (CRM)	Identifying customer value Developing revenue maximization programs
Credit card management		Identify significant market segments Identify probable income per customer
Insurance	Fraud detection	
Telecommunications		
Telemarketing	Online information	
Human resources management		

3.1. Retail trade

Data Mining offers retailers, in general, and grocery stores, in particular, valuable predictive information extracted from large data volumes. Affinity positioning is based on identifying products that the same customer will want. For example, products intended for cold treatment could be placed together with warm fabrics. It would have a commercial sense to locate both items within the visual limits of the other. Cross selling is a related concept. Knowing the products that can be sold together can be used to market the complementary product. Food stores achieve this by positioning the products on the shelf. Advertising-based retail stores can send sales announcements on shirts and ties to those who have recently bought garments. These strategies have been used for a long time by wise

DOI: 10.24818/18423264/54.1.20.11

retailers.The recommendation systems are actually used by Amazon and other online retailers. Exploring the data provides the opportunity to identify the less-expected affinities between the products and the cross-selling opportunities. These actions develop and maintain customer loyalty. The grocery stores generate data masses from the cash registers which require automatic data analysis tools.

In the past, it was considered that the data in the cash registers are so many that they could not be analysed quickly. However, current technology allows commercial bank to analyse customers who have defected from store, the history of their purchases, and the features of other potential customers likely to defect.

3.2.Banking Industry

Banks were among the first Data Mining users. Using this technology, they are trying to find out what motivates customers and how they can retain them in order to maintain and expand their business (Customer Relationship Management - CRM). CRM involves customer monitoring, a much-improved feature by using data mining technology. Understanding the value provided to the bank by the customer allows for a rational assessment of additional costs to keep this client as long as possible. There are many opportunities to exploit data in the banking field. Data Mining gives banks the opportunity to identify models. This is true in assessing loan applications as well as in target marketing. Banks use Data Mining to estimate the profitability of their customers as well as to monitor the effectiveness of their marketing programs and their subsidiaries and agencies. Bank uses information on a large scale to promote their promotions. Once identification data recognizes customers by features, direct mail and telemarketing are used to present promotional programs. The mortgage market has experienced massive refinancing over several periods. Bank has quickly recognized that they need to keep their customers happy if they want to preserve their business. This has led to banks contacting current customers if these customers hold a mortgage at a significantly higher rate than the market rate. Although they can reduce their own lucrative financial packages, banks realize that if they do not offer better service to borrowers, a competitor will do it.

3.3.Credit card management

The credit card industry has proven to be very profitable both for card issuers and for banks. It has attracted many card issuers, and many customers now have four or five credit cards. It is quite common practice for a credit card user to pay an old debt with a new card. They are not considered good customers, and one of the frequent uses of Data Mining is to identify such customers.

The profitability of the industry has also attracted those who want to increase the credit risk, both from the customer's perspective and from the issuer's perspective. As a rule, cardholder response rates are very low, with 1 response per 1,000 requests. This very low rate, which can hardly be used to deduce certain trends, is significantly improved through Data Mining analysis. Credit score is among the most commonly used instruments in this area, intensively used by the

DOI: 10.24818/18423264/54.1.20.11

banks. The credit score is a rigorous quantitative analysis of the loan applicants in anticipation of their intention to repay the bank loan on time. A key condition is a consolidated data repository covering all banking products, including demand deposits, savings, loans, credit cards, insurance, annuities, retirement schemes, underwriting transferable securities and all other banking products. The credit score provides a unique value for each loan applicant bank customer. These credit scores can be used to accept or reject loan applications as well as to determine the size of a credit line. The credit score was used by bank credit officers who took into account several tested variables such as employment, income, age, assets, debt and loan history. The news wave of banking technology extends the use of applications to targeted databases and marketing strategies. Credit card management software is used by more advanced credit card issuers using data mining tools such as neural networks to recognize customer behavior patterns in their future relationships with the bank.

3.4. Insurance

The insurance industry uses Data Mining in marketing, just as retail organizations and commercial banks do. There are also specialized applications for insurance only. For example, the US Farmers Insurance Group has developed an underwriting system that generates millions of dollars of higher incomes and lower claims. The system allows companies that subscribe to an insurance company to better understand narrow niche markets and to predict losses for certain insurance lines (life, goods, etc.). One achievement of this system was that it was possible to decrease the insurance rates for sports cars, which significantly increased the market share for these cars. Unfortunately, unlawful commercial transactions, including insurance frauds, frequently occur in our complex society. Those who resort to these underground operations often use more people to mislead insurance companies, especially in the field of car insurance. Fraud Detection Software uses a search engine that analyses the information provided by companies and individuals claiming insurance payments.

3.5. Telecommunications

Deregulation of the communications industry has led to an unprecedented increase in competition between telephone service providers who struggle hard for customers. Once a client is obtained, it is attacked by competitors, and keeping customers is very difficult. The phenomenon of exchanging customers between different providers of telephony, cable and Internet services is called "porting", a fundamental concept in telemarketing, as well as in other areas. A marketing director for a telecommunications company believes that one third of the network leavings are due to the low quality of the services provided, and up to half is due to obsolete equipment. Companies typically use a performance monitor for phones that tracks low-performance phones. This system reduced the number of leavings

DOI: 10.24818/18423264/54.1.20.11

to an estimated 61%, accounting for about 3% of the company's global subscribers over the course of a year. Another method to reduce porting is to protect customers from abuse and fraudulent cloning. Cloning was estimated at millions of euro in the telecommunications industry. Many fraud prevention systems are marketed nowadays. These systems provide transparent verification for legitimate subscribers.

3.6. Human Resources Management

Business Intelligence is a way to truly understand markets, competitors and processes. Software technology, such as data repositories, data marts, online analytics processing (OLAP), and Data Mining, allow data analysis to identify trends and methods that can be used by the company to improve profitability. In the field of human resources, this analysis can lead to identifying people who can leave the company if no additional compensation or benefits are provided. Data Mining can be used to expand knowledge about things that are apparently already known. A company may know that 20% of its employees use 80% of the services they offer, but may not know which of the individuals falls under this 20%. Business intelligence provides a means of identifying segments so that programs can be designed to reduce costs and increase employee productivity. Data Mining can also be used to examine how an organization uses its employees. It might be a question of whether the most talented people are working for companies where they will have the greatest impact on profit. Good human resource management would identify the right people so that organizations can treat them as well as possible and therefore can keep them. This requires tracking key performance indicators and collecting talent data, company needs and competitors' requirements.

4.Data Mining Tools

One of the first business applications of Data Mining is customer segmentation. This application uses large databases (obtained from a variety of sources) to segment the market into categories, which are then researched using feature mining tools to anticipate customer response to specific advertising campaigns. Due to the probabilistic nature of the data mining process, this segmentation is not perfect. The idea is that it would be better to send catalogues or calls to a target group of customers with a 5% probability of buying than to lose these costly marketing resources to customers with a probability of purchase of less than 0.05%.

As we have already shown above, Data Mining plays a key role in managing customer relationships, managing credit cards, banking, insurance, telecoms and many other business areas.Expanding business data is possible by generating data masses from computer systems. Understanding this generation of information, the system and tools available that lead to analysis are fundamental for 21st century students. Support for Data Mining is necessary to understand the large volumes of business data generated by computer technology. Below we will describe some of the main applications of Data Mining. By doing so, there will also

DOI: 10.24818/18423264/54.1.20.11

be opportunities to demonstrate some of the various techniques that have proven useful in this area.

Table 2. Lists of the most important aspects of these applications

Application	Function	Statistical method	AI tool
Catalogue sales	Customer segmentation Optimization	Cluster analysis	K-means Neural networks
CRM	Consumer score Leaving analysis	Cluster analysis	Neural networks
Credit scoring	Customer score	Classification	K-means
Banking credit	Prediction of bankruptcy	Prediction Discriminant analysis	Decision trees
Investment risk	Risk prediction	Prediction	Neural networks
Insurance	Customer retention Evaluation of the price	Prediction Logic regression	Decision trees Neural networks

5.Customer segmentation - basic elements

Customer segmentation is the process of dividing customers into distinct, meaningful and homogeneous subgroups, based on their various attributes and features. Segmentation is commonly used as a differentiation tool in strategic marketing. It allows organizations to understand customers and develop differentiated strategies tailored to their characteristics. In general, customers may be segmented according to their value, socio-demographic information and life status, behavior, needs/attitude and loyalty characteristics. The type of segmentation to be used depends on the specific business objective.

The following types of segmentation are used more frequently (Tsiptsis and Chorianopoulos, 2009):

- **Segmentation based on customer value**: Customers are grouped according to their value. It is one of the most important types of

DOI: 10.24818/18423264/54.1.20.11

segmentation, as it can be used to identify the most valuable customers and to track changes in their value over time.

- **Segmentation based on customer behavior**: It is a highly efficient and useful segmentation type. Data required includes product ownership and usage data that are usually available in the organization's databases. Customers are divided according to their identified and shaped behavior. This type of segmentation is commonly used to develop personalized product offering strategies, launching new products, and designing loyalty schemes applied to customers.
- **Trend-based segmentation:** In this type of segmentation, customers are grouped according to their propensity scores, such as earnings scores, cross selling scores etc., which are estimated using the respective rating (propensity) models.
- **Loyalty-based segmentation**: Segmentation based on loyalty involves estimating customer loyalty and identifying segments based on loyalty, such as 'loyal' and 'migrant' segments. The retention actions can then be centered on high-value clients but with a non-loyal profile, while product offering is targeted at customers supposed to be loyal.
- **Socio-demographic segmentation**: Discloses different groups of customers based on socio-demographic and/or personal information such as age, address, income, or marital status. This type of segmentation is suitable for promoting life-based products as well as promoting marketing at different stages of customer life.
- **Segmentation by needs/attitudes**: This type of segmentation is typically based on market research data and identifies customer segments according to their needs, desires, attitudes, preferences and perceptions of the products and services offered by the company. It can be used to support the introduction and development of new products and to determine the brand image and key product characteristics that must be communicated to customers.

6.Application on panel data type

Panel data notice certain features of entities over several periods of time. Depending on the time horizon and the units observed, two types of panels are distinguished:

- *Micro panel* - characterized by a large number of individuals (N is in the order of thousands or hundreds) over a short period of time (T = 2 years, maximum 20 years);
- *Macro panel* - for an N number of countries (N is usually 7, 20 but also 100), annual variables are observed over a 20-60-year horizon.

With the help of panel structures, the heterogeneity of the units analyzed (countries, households, companies, individuals, banks) can be controlled because they (unlike time series or cross-sectional data) provide the possibility of including variables that capture specific aspects of each unit. In regression models performed on panel data, it is possible to include, besides the exogenous variables that vary

DOI: 10.24818/18423264/54.1.20.11

over time but also from one unit to another, variables that vary from one unit to another without varying in time (variables that express the characteristics of each unit). By omitting these aspects, we can obtain displaced estimators of coefficients of regression models.

6.1. Regression model on panel data

A regression model for panel data differs from the models built on cross sectional data or time series, first of all, in that its specification uses two indices (t = 1 ...T for time and i = 1 ...N for units):

$$\gamma_{it} = \alpha + \beta x_{it} + \mu_{it} \text{ , where i = 1 ...N, t-1 ...T}$$

Depending on the composition of the error termμ_{it} there are two types of regression models:

- Models (one-way error component) in which the error term breaks down according to the relationship:

$$\mu_{it} = \mu_i + \nu_{it}$$

- in a component that contains the unexplained part of the model that does not depend on time, but which refers to certain characteristics of the unit and not considered in the model (μ_i);
- and an error term vi_t(the common term in regression models) that varies from one unit to another but also by time.

- The models (two-way error component) where the error additionally contains a term that varies only by time (λ_t):

$$\mu_{it} = \mu_i + \lambda_t + \nu_{it}$$

The capturing of the individual effects is achieved by dividing the variable of discrepancy (deviation) μ_{it} into three components:

- variable$\mu_i$estimates the individual, unobservable specific event and is invariant over time: it estimates the effect of variables not included in the model on the endogenous, in unit i (specific individual effect);
- variable λ_t estimates the invariable, temporal specific effect in cross structures: estimates the effect of variables not included in the model on the endogen, at time t (fixed effect over time);
- error variable (residual discrepancy) ν_{it} is variable both between individuals and over time.

6.2.Fixed effects/Random effects models

In estimating regression models based on panel data, one of two options is chosen:

- models with fixed effects;
- models with random effects.

DOI: 10.24818/18423264/54.1.20.11

In the case of fixed effects models:$\mu_i$is considered a fixed parameter to be estimated and ν_{it} ~ iid(0, σ_v^2). The estimation assumes the insertion of (N-1) dummy variables (z):

$$\gamma = \alpha + \beta x + \mu z + \nu$$

Parameters α, β, μ are estimated by the OLS technique. But if the number of units is very large, then such a model is not feasible, introducing a N-1 dummy variables leading to the loss of a large number of degrees of freedom. Choosing a fixed effect model when a number of N units (usually countries) is being analyzed will assume that the statistical inference occurs only on this set of units.

In the case of models with random effects: $\mu_i$are representations of a random variable: μ_i~ iid (0, σv^2), ν_{it}~ iid (0, σ_v^2), x$i_t$is independent of ν_{it}∀i and ∀ t and μ_i and ν_{it} are independent.

It is advisable to use this model when randomly mining individuals from a larger population. It is obvious that in this case *N* (the number of individuals) being large, the use of fixed effects would lead to the loss of a large number of degrees of freedom. The method by which the parameters of such a model are estimated is the generalized method of least squares (OLS), and the inference is on the population from which the sample was mined.

6.2.1. Using dummy variables to estimate a fixed effects model

The dummy variable is a variable that can take two values: 0 or 1. Such a variable separates the sample of the observations into two categories. In a regression model, the coefficient associated with a dummy variable must be interpreted by reference to the base (the reference group is that for which the dummy variable is zero).

A model with fixed effects assumes that the slope of the regression model is the same regardless of country, but the free term differs from one country to another (but it is invariant over time). It is possible to construct such a model if dummy variables are used.

6.3. Analysis

In this application, we will use panel data that contains information about ***the value of the profit*** **and** ***investment in vocational training*** for 5 companies.
The time horizon is 4 years, i.e. the period 2008 - 2011.

A first step in conducting the analysis is to represent data in graphical form for the purposes of highlighting their heterogeneity over companies and years.

DOI: 10.24818/18423264/54.1.20.11

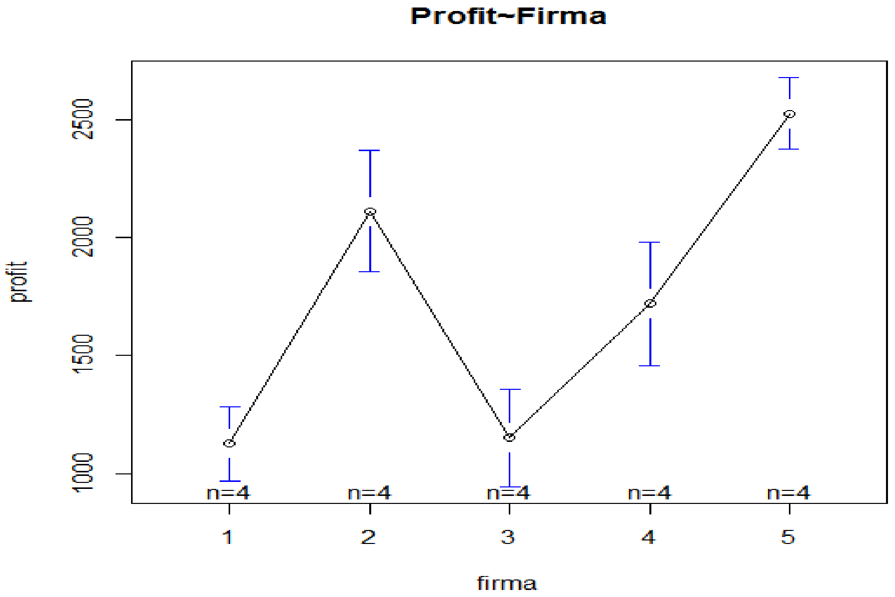


Figure 1.The average of the profit for each company for a confidence interval of 95%.

For each company we highlight the average return of 95% confidence interval. Company number 5 has the highest average return, amounting to 2,500 u.m. Company number 1 together with company number 3 have the lowest average return of 1,000 u.m.

Moreover, it can be observed that the range size for the first company is lower than that of other companies. This can happen because of variations that increase from year to year. Therefore, one can assert that companies are heterogeneous in terms of return.

In Figure 2 one can observe the company's average profits for eachyearfor a confidence interval of 95%.

DOI: 10.24818/18423264/54.1.20.11

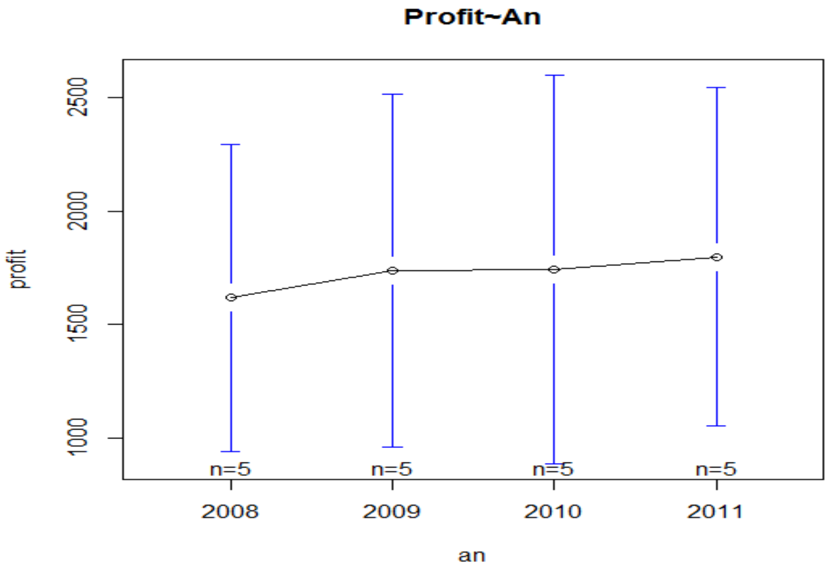


Figure 2.The average of the profit for each year for a confidence interval of 95%.

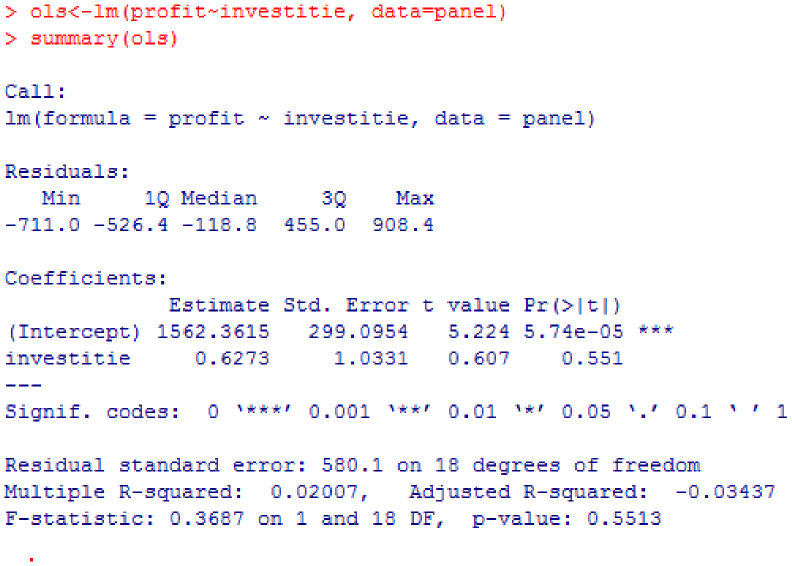
One can observe that the company's average profits are maintained over the years between 1,500 u.m. and 2,000 u.m. This average is increasing, and in terms of ranges, they are of different sizes, therefore one can assert there is also a heterogeneity per years, but less than that for companies.

The next step in this analysis is to estimate the results of the data set using the OLS(Ordinary Least Squares) technique, without taking into account the heterogeneity between companies or over time. The following output will result:

The results obtained using the OLS technique show that the estimated value of the investment is 0.6273, meaning that when the investment changes by 1 unit, the value of the profit changes by 0.6273 units. Thus, it is noticed that for *the effect of investment in vocational training on profit* there is a *p-value = 0.551*, which is higher than the 5% risk threshold, so it is not statistically significant, which means that **investment in vocational training does not explain the variation in profit.**

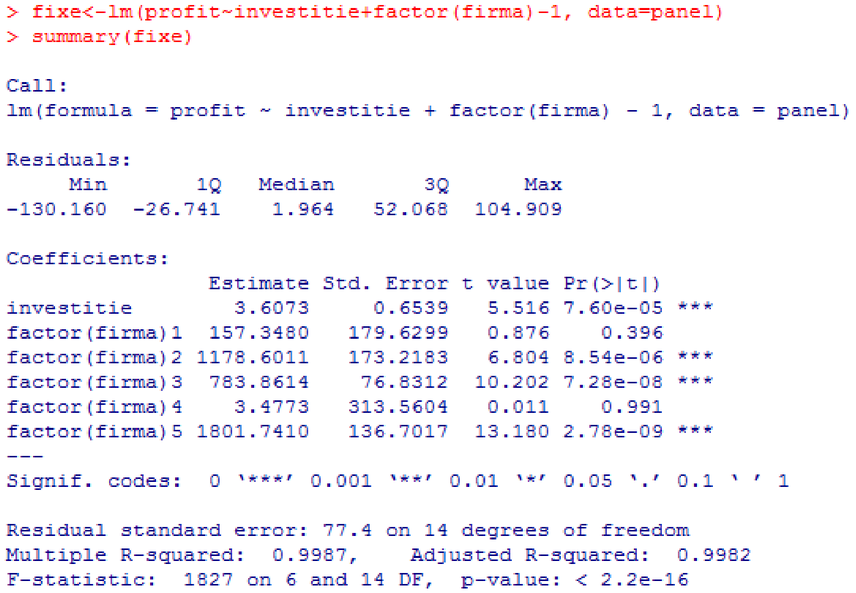
As one can observe in the output obtained, the model determined by this technique is not a very good one as the R-squared value is very small (0.02007).

DOI: 10.24818/18423264/54.1.20.11



```
> ols<-lm(profit~investitie, data=panel)
> summary(ols)

Call:
lm(formula = profit ~ investitie, data = panel)

Residuals:
   Min     1Q Median     3Q    Max 
-711.0 -526.4 -118.8  455.0  908.4 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1562.3615   299.0954   5.224 5.74e-05 ***
investitie     0.6273     1.0331   0.607    0.551    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 580.1 on 18 degrees of freedom
Multiple R-squared:  0.02007,	Adjusted R-squared:  -0.03437 
F-statistic: 0.3687 on 1 and 18 DF,  p-value: 0.5513
```

If we estimate the results of the data set using the OLS technique, taking into account the heterogeneity between companies and time we get the following result:



```
> fixe<-lm(profit~investitie+factor(firma)-1, data=panel)
> summary(fixe)

Call:
lm(formula = profit ~ investitie + factor(firma) - 1, data = panel)

Residuals:
     Min       1Q   Median       3Q      Max 
-130.160  -26.741    1.964   52.068  104.909 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
investitie        3.6073     0.6539   5.516 7.60e-05 ***
factor(firma)1  157.3480   179.6299   0.876    0.396    
factor(firma)2 1178.6011   173.2183   6.804 8.54e-06 ***
factor(firma)3  783.8614    76.8312  10.202 7.28e-08 ***
factor(firma)4    3.4773   313.5604   0.011    0.991    
factor(firma)5 1801.7410   136.7017  13.180 2.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 77.4 on 14 degrees of freedom
Multiple R-squared:  0.9987,	Adjusted R-squared:  0.9982 
F-statistic:  1827 on 6 and 14 DF,  p-value: < 2.2e-16
```

As a result of this analysis, the differences between the companies under assessment are highlighted. One can now notice at *the effect of investments on* *profits* that the value of p-value is very small (7.60e-05) which means it is significant and one can assert that **investment in vocational training explains the variation in profit.**

While companies 2, 3, and 5 are significant to the model, companies 1 and 4 are not significant to this model, but all companies and the investment account for 99.8% of the profit.

In the following figure we will represent the estimated regression lines for each company and the OLS regression line:

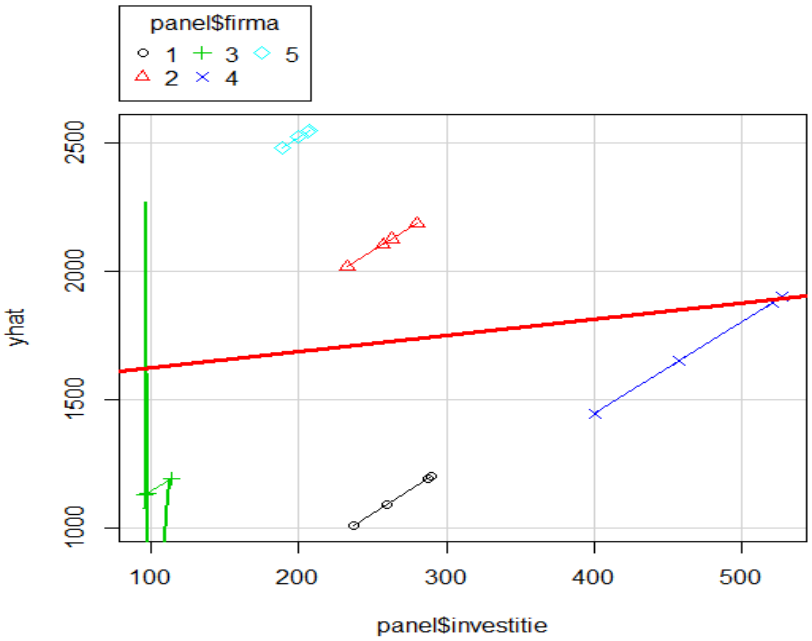


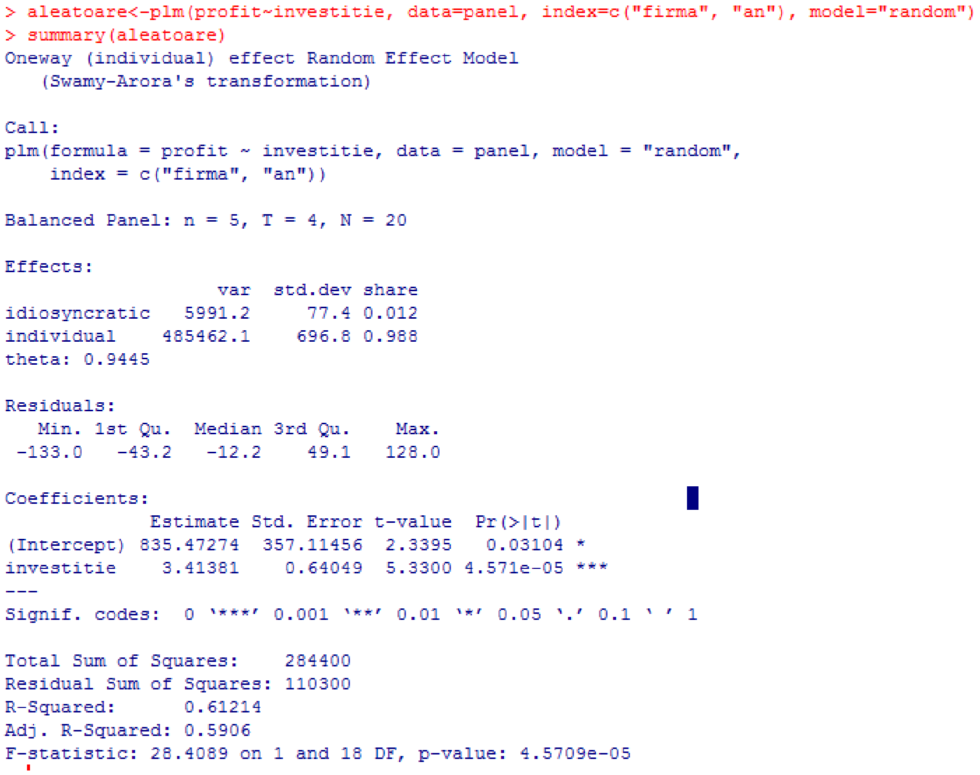
Figure 3.The estimated regression lines and the OLS regression line

From the resulting output and illustrated in Figure 3, one can state that the regression lines of the companies are parallel (so all companies have the same regression line). We note, however, that each regression line starts from different points. Furthermore, because of the dispersion of companies, a regression model that encompasses all five companies is not representative because it does not take into account the specificity of each company.

In conclusion, the OLS model, which takes into account the heterogeneity between companies, is better than the one in which heterogeneity between companies is not taken into account because it takes into account the specificity of each companies, more precisely.

DOI: 10.24818/18423264/54.1.20.11

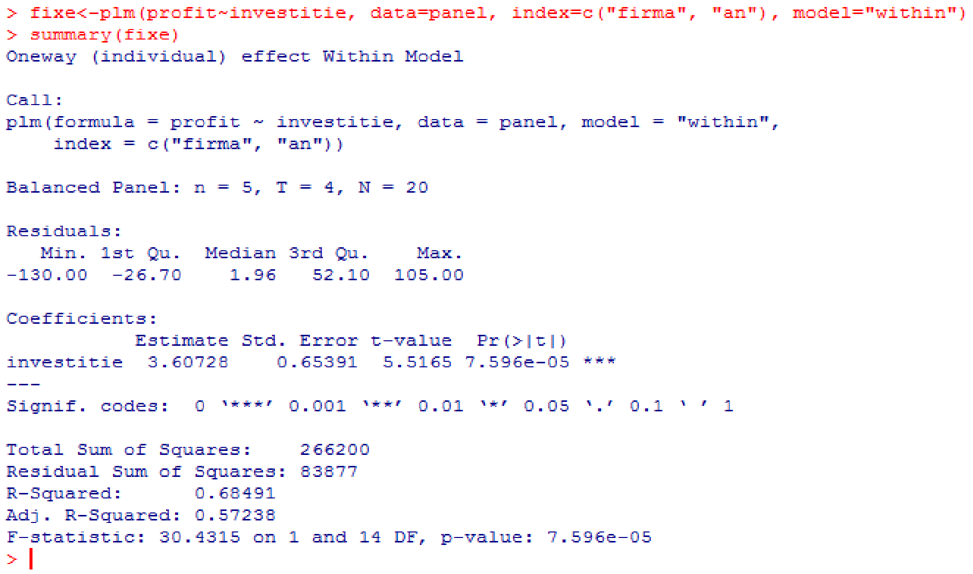
The random effects model has the next output:



```
> aleatoare<-plm(profit~investitie, data=panel, index=c("firma", "an"), model="random")
> summary(aleatoare)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = profit ~ investitie, data = panel, model = "random",
    index = c("firma", "an"))

Balanced Panel: n = 5, T = 4, N = 20

Effects:
                   var  std.dev share
idiosyncratic   5991.2     77.4 0.012
individual    485462.1    696.8 0.988
theta: 0.9445

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -133.0   -43.2   -12.2    49.1   128.0

Coefficients:
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept) 835.47274  357.11456  2.3395   0.03104 *
investitie    3.41381    0.64049  5.3300 4.571e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    284400
Residual Sum of Squares: 110300
R-Squared:      0.61214
Adj. R-Squared: 0.5906
F-statistic: 28.4089 on 1 and 18 DF, p-value: 4.5709e-05
```

The fixed effects model has the next output:

The output below shows that the investment dependent variable is statistically valid (the value of the test t is higher than the critical value and the probability is lower than the accepted 5% risk threshold), this being 3.60728. R-squared = 0.68491 shows that 64.491% of the profit variation is explained by the investment level, therefore the investment is a decisive factor included in the model.

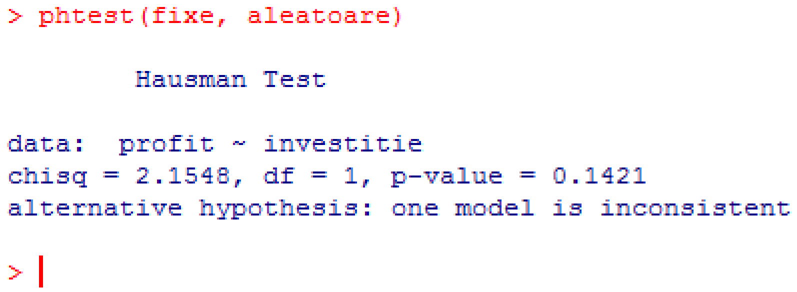


```
> fixe<-plm(profit~investitie, data=panel, index=c("firma", "an"), model="within")
> summary(fixe)
Oneway (individual) effect Within Model

Call:
plm(formula = profit ~ investitie, data = panel, model = "within",
    index = c("firma", "an"))

Balanced Panel: n = 5, T = 4, N = 20

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-130.00  -26.70    1.96   52.10  105.00

Coefficients:
           Estimate Std. Error t-value  Pr(>|t|)
investitie  3.60728    0.65391  5.5165 7.596e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    266200
Residual Sum of Squares: 83877
R-Squared:      0.68491
Adj. R-Squared: 0.57238
F-statistic: 30.4315 on 1 and 14 DF, p-value: 7.596e-05
> 
```

The Hausman test

The Hausman test is used to decide whether to choose the fixed effects model or the random effects model, so we formulate the assumptions:

- *H0: Random effects (errors are not correlated with regressions);*
- *H1: Fixed effects (errors are correlated with regressions).*



```
> phtest(fixe, aleatoare)

        Hausman Test

data:  profit ~ investitie
chisq = 2.1548, df = 1, p-value = 0.1421
alternative hypothesis: one model is inconsistent

> 
```

With the Hausman test, for a very small value (less than 5%) we will reject the null hypothesis and accept the random hypothesis, so errors are correlated with regressors. In our case we have a p-value = 0.1421, that is a probability of 14.21% that leads us to the indecision area. Therefore, we cannot assert that the effects are of a certain type, but that we cannot decide.

7. Conclusion

This example reviewed at the relationship between profit and the investment effect in vocational training for five companies and the period for which the analysis was conducted is 2008-2011. A panel model was used, the decision between the random and the fixed effects model was based on the Hausman test and the information criteria associated with the two models.

The explicit variable of investment in vocational training has not proven to be significant in the original OLS model. However, if we controlled the differences between the sections (the companies analyzed) through the dummy variables, the effect of the investments in the professional training on the profit became significant.

REFERENCES

[1] **Abdul-Muhmin, A., Umar, Y. A. (2007),*Credit Card Ownership and Usage Behavior in Saudi Arabia: The Impact of Demographics and Attitudes toward Debt;*** Journal of Financial Services Marketing, 12, pp. 219–234;
[2] **Allwein, E.L., Schapire, F.E., Singer,Y. (2000), *Reducing Multiclass to Binary Unifying Approach for Margin Classifier;*** Journal of Machine Learning Research, 1, pp.113-141;
[3]**Banasiewicz, A. (2013), *Marketing Database Analytics - Transforming Data for Competitive Advantage*,** Routledge;
[4]**Basnet, H. C., Donou-Adonsou, F. (2016) ,*Internet, Consumer Spending and Credit Card Balance: Evidence from US Consumers***; Review of Financial Economics, 30, pp.11–22;
[5]**Bradley, P.S., Mangesarian, O.L. (1999), *Feature Selection via Concave Minimization and Support Vector Machines;*** Machine Learning Proceedings of the Fifteenth International Conference on Machine Learning (ICML98), pp. 82-90;
[6] **Cai, L., Zhu, Y. (2015),*The Challenges of Data Quality and Data Quality Assessment in the Big Data Era;*** Data Science Journal, Vol.14;http://datascience.codata.org/articles/10.5334/djs-2015-002/print/;
[7] **Calcagno, R., Monticone, C. (2015), *Financial Literacy and the Demand for Financial Advice;*** Journal of Banking & Finance, 50, pp. 363–380;
[8]**Carlsson, H., Larsson, S., Svensson, L., Åström, F. (2017),*Consumer Credit Behavior in the Digital Context: A Bibliometric Analysis and Literature Review*;** Journal of Financial Counseling and Planning, 28(1), pp. 76–94;
[9] **Dahlbäck, O. (1991), *Saving and Risk Taking*;** Journal of Economic Psychology, 12(3), pp. 479–500;
[10] **Davenport, T., Dyche, J. (2013),*Big Data in Big Companies*;** International Institute for Analytics;
[11] **Davenport, T.H. (2014),*Big Data at Work;*** Boston, MA: Harvard Business Review Press;
[12]**Delamaire, L., Abdou, H., Pointon, J. (2009),*Credit Card Fraud and Detection Techniques: A Review, Banks and Bank Systems*;** Volume 4, Issue 2, 2009, pp. 57-68;

DOI: 10.24818/18423264/54.1.20.11

[13]**Dudyala A. K. and Ravi V. (2008),** ***Predicting Credit Card Customer Churn in Banks Using Data Mining*****.** International Journal of Data Analysis Techniques and Strategies, Vol. 1, No. 1, pp.4-28;

[14]**Dunis, Cr., Theofilatos, K., Middleton, P., Karathanasopolous, A., (2016),** ***Artificial Intelligence in Financial Markets. Cutting-Edge Applications for Risk Management, Portfolio Optimization and Economics*****;** *London:Palgrave, Macmillan.;*

[15] **Evans, M. (2015),** ***Healthcare Data Mining***; Modern Healthcare 45, no. 39, pp. 24;

[16] **Isson, J.P. (2018),*****Unstructured Data Analytics. How to Improve Customer Acquisition, Customer Retention and Fraud Detection and Prevention;*** *John Wiley & Sons,* Inc., Hoboken, New Jersey;

[17] **Mayer-Schonberger, V., Cukier K. (2013),** ***Big Data: A Revolution that Will Transform HowWe Live, Work, andThink.*** New York: *Houghton Mifflin Harcourt;*

[18]**Pardoe, I. (2012),*****Applied Regression Modeling*****;**2nd ed., *John Wiley&Sons,* Inc.;

[19] **Pascual, D. G. (2015),*****Artificial Intelligence Tools. Decision Support Systems in Condition Monitoring and Diagnosis***; Taylor & Francis Group, LLC.;

[20] **Perner, P. (2013),** ***Advances in Data Mining. Applications and Theoretical Aspects***; 13th Industrial Conference, ICDM 2013 New York, NY, USA;

[21]**RaghavendraPatidar, LokeshSharma (2011),*****Credit Card Fraud Detection Using Neural Network;***International Journal of Soft Computingand Engineering (IJSCE), Volume-1, pp. 32-38;

[22] **Sathi, A. (2012),*****Big Data Analytics: Disruptive Technologies for Changingthe Game;***Mc Press, ISBN: 1583473807;

[23]**Tsiptsis, K., Chorianopoulos, A. (2009),*****Data MiningTechniques in CRM. Inside Customer Segmentation;****John Wiley&Sons Ltd Publication*;

[24]**Westphal, C. (2008),*****Data Mining for Intelligence, Fraud & Criminal Detection. Advanced Analytics & Information Sharing Technologies*****;** CRC Press, Taylor & Francis Group, ISBN: 978-142006723.

DOI: 10.24818/18423264/54.1.20.11