Empirical data analysis in accounting and finance

Some examples on quantitative empirical research problems in accounting and finance

- Pattern recognition
- Financial classification models
  - Financial distress prediction
- Web Questionnaires
  - ANOVA, MANOVA and MRA modelling
- Causality models
  - Association between accounting data and financial market reactions
  - Causality patterns on international financial markets
- Time series modelling and prediction
- Optimization models, e.g.
  - Portfolio optimization
  - product mix optimization

A typical process for empirical data analysis

- Define the test problem
- Collect data
  - Data bases for financial data, e.g. market data, financial statements, interest rates, exchange rates
  - Surveys for opinion data
- Select the analysis method
- Control for the suitability of the data to the selected method
  - Different methods have different assumptions on the properties of the data, e.g. approximate normality

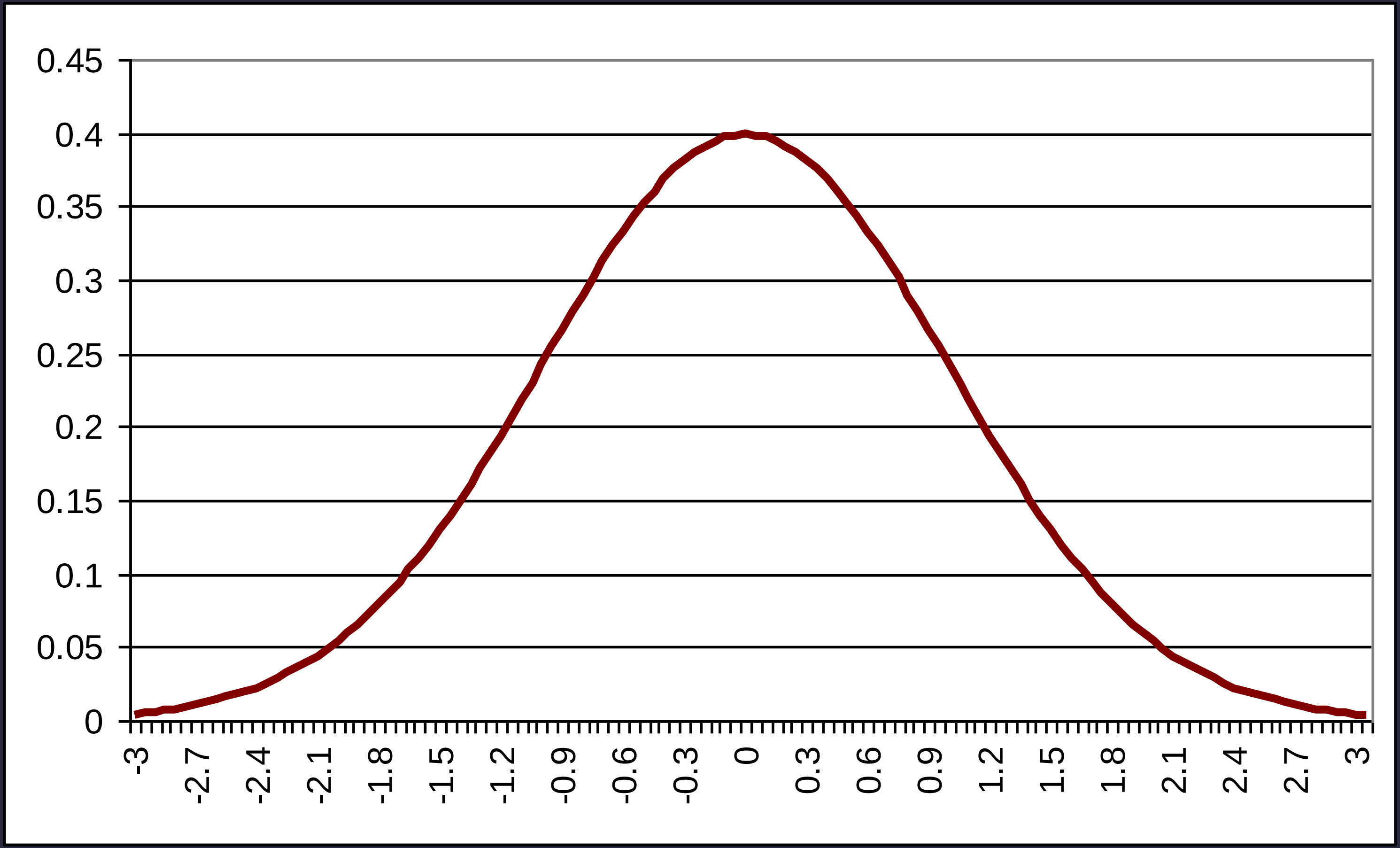
A typical process for empirical data analysis ...

- If necessary, improve the quality of the data
  - Different transformations
    - Taking logarithms of the data
    - Differencing (for time series data)
  - Removing the outliers
- Perform the data analysis by a suitable statistical program, e.g. SPSS, SAS
- Interpret the results critically

The quality of the data

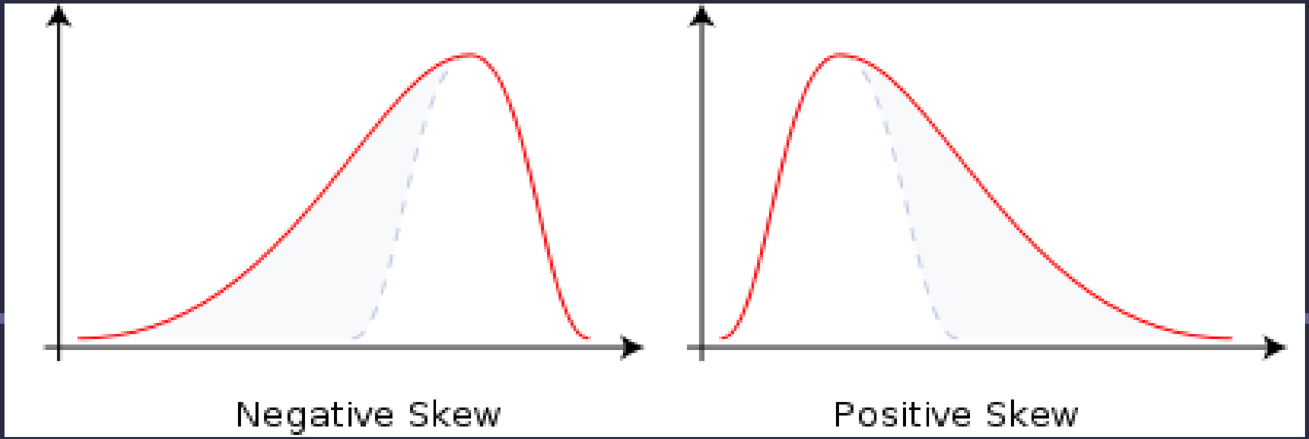
- Practically all statistical methods assume the data to follow some predefined distribution (probability density function or pdf), e.g.
  - Standard normal distribution
  - Normal distribution
- Most empirical data sets fail to satisfy the basic assumption on normal distribution
- Typical problems are
  - Skewness of the data
  - Leptokurtosis (thick tails)
  - Outliers

The standard normal (Gaussian) distribution ($\mu = 0$, $\sigma = 1$)



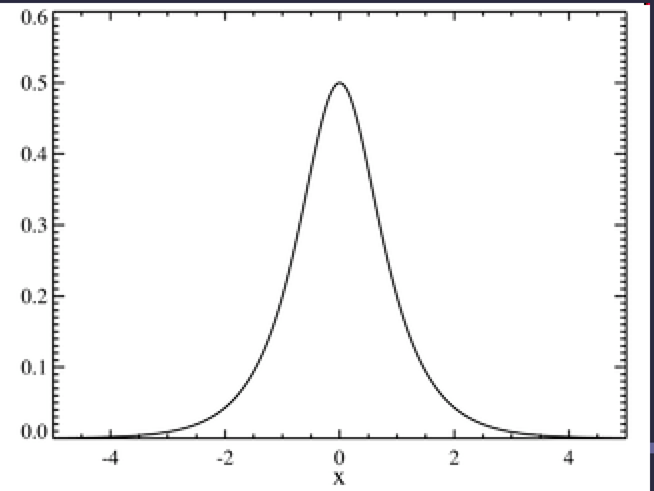
Skewness

- A measure of the **asymmetry** of the probability distribution
- Skewness = 0 for a symmetric distribution
- **Negative skewness (left skewed pdf)**: The left tail is longer; the mass of the distribution is concentrated on the right of the figure
- **Positive skewness (right-skewed pdf)**: The right tail is longer; the mass of the distribution is concentrated on the left of the figure

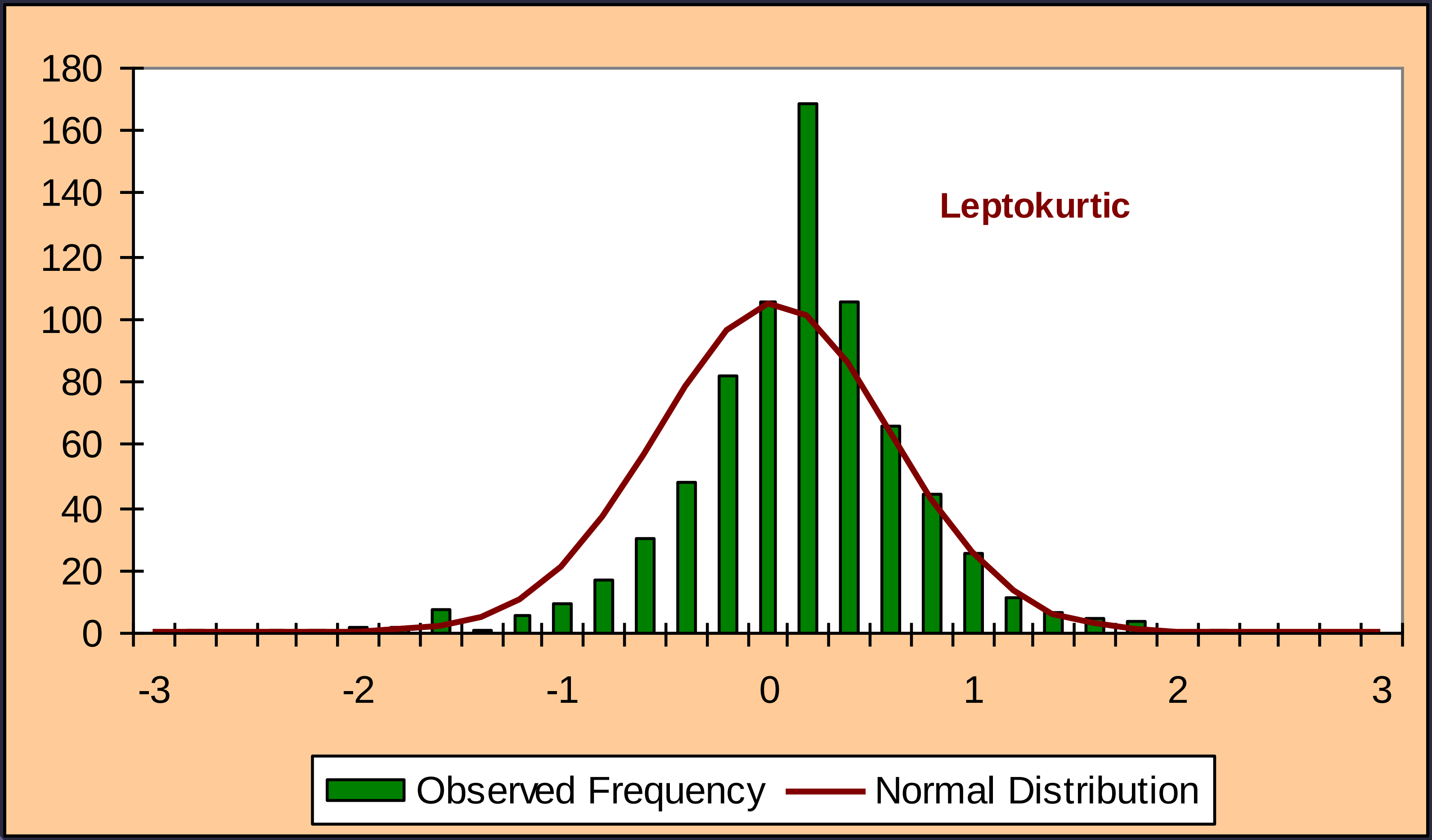


Kurtosis

- A measure of the **peakedness** of the probability distribution
- Many financial data series (for example, stock returns) have **leptokurtic** distributions
- A leptokurtic distribution has a more acute peak around the mean and fatter tails



Histogram of the Canadian stock market returns and a normal distribution with the observed mean and standard deviation



Testing for the normality of a data set

- There are several tests for measuring the normality of a data set, for example
  - Kolmogorov-Smirnov test (Massey, 1951)
  - Pearson's chi-square test (Pearson, 1900)
  - Jarque-Bera test (Jarque & Bera, 1980)
  - Shapiro-Wilk test (Shapiro & Wilk, 1965)

SPSS-output for the K-S test with the Canadian data

One-Sample Kolmogorov-Smirnov Test

		Can
N		752
Normal Parameters[a,,b]	Mean	,0340
	Std. Deviation	,57275
Most Extreme Differences	Absolute	,077
	Positive	,047
	Negative	-,077
Kolmogorov-Smirnov Z		2,103
Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

Data not normal

($\alpha < 0.01$)

The classification problem

- In a traditional classification problem the main purpose is to assign one of *k labels* (or classes) to each of *n objects*, in a way that is consistent with some observed data, i.e. to **determine the class of an observation based on** a set of variables known as **predictors** or **input variables**
- Typical classification problems in finance are for example
  - Financial failure/bankruptcy prediction
  - Credit risk rating

Classification models

- Discriminant analysis
- Logistic regression
- Recursive partitioning algorithm (RPA)
- Mathematical programming
  - Linear programming models
  - Quadratic programming models
- Neural network classifiers

Discriminant analysis

- Discriminant analysis is the most common technique for classifying a set of observations into predefined classes
- The model is built based on a set of observations for which the classes are known
- This set of observations is sometimes referred to as the **training set** or **estimation sample**

Discriminant analysis...

- Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that

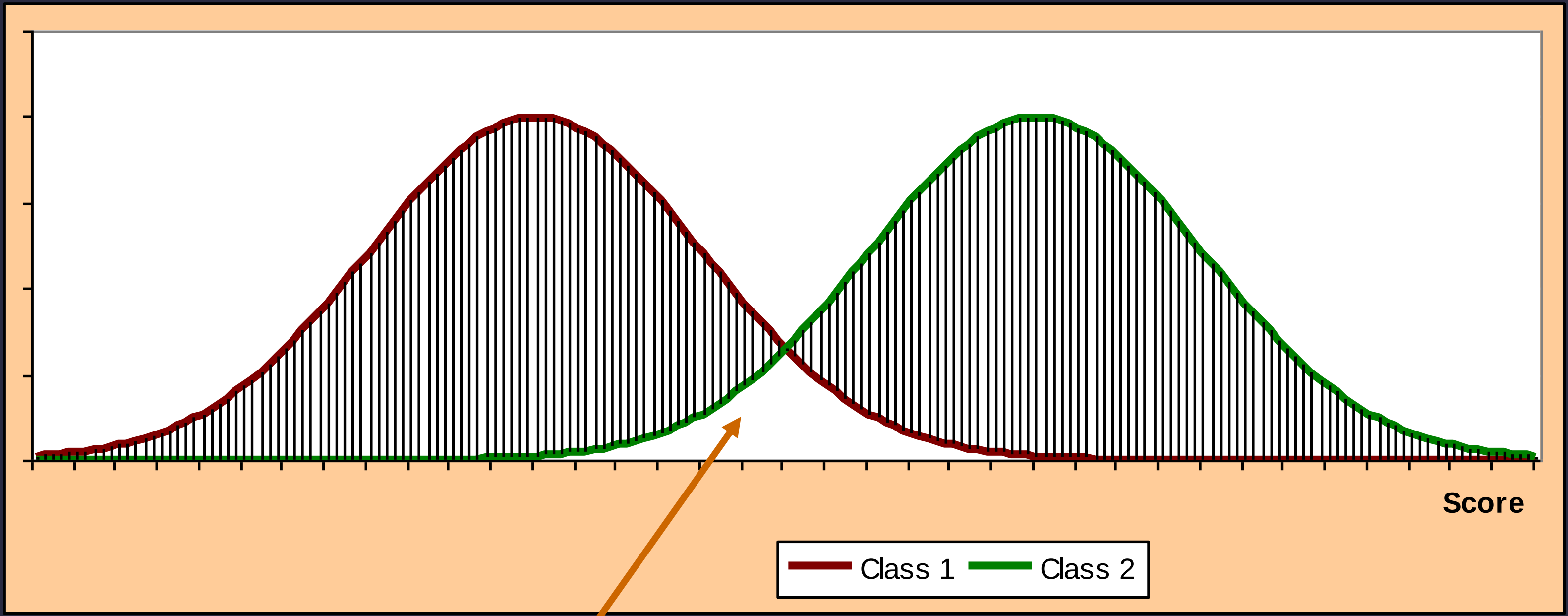
$$L = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + c,$$

where the β's are **discriminant coefficients**, the *x*'s are the input variables or predictors and *c* is a constant.

Discriminant functions

- The discriminant functions are optimized to provide a classification rule that minimizes the probability of misclassification
- In order to achieve optimal performance, some statistical assumptions about the data must be met
  - Each group must be a sample from a multivariate normal population
  - The population covariance matrices must all be equal
- In practice the discriminant has been shown to perform fairly well even though the assumptions on data are violated

Distributions of the discriminant scores for two classes



A discriminant function is optimized to minimize the common area for the distributions

Case: Bankruptcy prediction in the Spanish banking sector

- Reference: Olmeda, Ignacio and Fernández, Eugenio: "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction", *Computational Economics* **10**, 1997, 317-335.
- Sample: 66 Spanish banks
  - 37 survivors
  - 29 failed
- Sample was divided in two sub-samples
  - Estimation sample, 34 banks, for estimating the model parameters
  - Holdout sample, 32 banks, for validating the results

Case: Bankruptcy prediction in the Spanish banking sector

Input variables

- Current assets/Total assets
- (Current assets-Cash)/Total assets
- Current assets/Loans
- Reserves/Loans
- Net income/Total assets
- Net income/Total equity capital
- Net income/Loans
- Cost of sales/Sales
- Cash flow/Loans

Empirical results

- Analyzing the total set of 66 observations
  - Group statistics – comparing the group means
  - Testing for the equality of group means
  - Correlation matrix
- Classification with different methods
  - Estimating classification models using the estimation sample of 34 observations
  - Checking the validity of the models by classifying the holdout sample of 32 observations

Group statistics

	Class 0 N=37		Class 1 N=29		Total N=66	
	Mean	St.dev	Mean	St.dev	Mean	St.dev
CA/TA	,410	,114	,370	,108	,393	,112
(CA-Cash)/TA	,268	,089	,264	,092	,266	,089
CA/Loans	,423	,144	,390	,117	,409	,133
Reserves/Loans	,038	,054	,016	,012	,028	,043
NI/TA	,008	,005	-,003	,019	,003	,014
NI/TEC	,167	,082	-,032	,419	,079	,299
NI/Loans	,008	,005	-,003	,020	,003	,015
CofS/Sales	,828	,062	,957	,188	,885	,147
CF/Loans	,018	,029	,004	,012	,012	,024

Tests of equality of group means

Significant if close to zero

	Wilks' Lambda	F	df1	df2	Sig.
CA/TA	,969	2,072	1	64	,155
(CA-Cash)/TA	1,000	,027	1	64	,871
CA/Loans	,985	,981	1	64	,326
Reserves/Loans	,932	4,667	1	64	,034
NI/TA	,864	10,041	1	64	,002
NI/TEC	,889	8,011	1	64	,006
NI/Loans	,863	10,149	1	64	,002
CofS/Sales	,805	15,463	1	64	,000
CF/Loans	,918	5,713	1	64	,020

No significant difference in group means

Fisher's discriminant function coefficients

	Survived	Failed
Constant	-758.242	-758.800
CA/TA	48.588	34.572
CA_Cash/TA	9.800	23.506
CA/Loans	-18.031	-16.947
Res/Loans	351.432	342.204
NI/TA	-246 563.200	-236 546.700
NI/TEC	774.368	740.035
NI/Loans	23 681.300	21 4974.000
CofS/Sales	1 499.659	1 505.547
CF/Loans	14 625.844	14 245.368

Example on classifying an observation by discriminant functions

	Obs. 1	Survived	Score	Failed	Score
Constant		-758.24	-758.24	-758.800	-758.80
CA/TA	0.4611	48.59	22.40	34.572	15.94
CA_Cash/TA	0.3837	9.80	3.76	23.506	9.02
CA/Loans	0.4894	-18.03	-8.82	-16.947	-8.29
Res/Loans	0.0077	351.43	2.71	342.204	2.63
NI/TA	0.0057	-246563.2	-1405.41	-236546.7	-1348.32
NI/TEC	0.0996	774.37	77.13	740.035	73.71
NI/Loans	0.0061	23681.3	1364.46	214974.0	1311.34
CofS/Sales	0.8799	1499.66	1319.55	1505.547	1324.73
CF/Loans	0.0092	14625.84	134.56	14245.368	131.06
Total Score			752.08		753.02

Larger score $\Rightarrow$

Classification: Failed

Confusion matrix – Classification results for the holdout sample

		Predicted class	
		Survived	Failed
True class	Survived	15	3
		83.33 %	16.67 %
	Failed	4	10
		28.57 %	71.43 %

Summary of classifications (Estimation sample)

Method	Correct class	Errors		Total number	Percents		
		SW	NE		Correct	SW	NE
RPA	30	1	3	34	88.24 %	2.94 %	8.82 %
MDA	30	0	4	34	88.24 %	0.00 %	11.76 %
MDA-Q	31	0	3	34	91.18 %	0.00 %	8.82 %
MDA-W	31	0	3	34	91.18 %	0.00 %	8.82 %
LogR	33	0	1	34	97.06 %	0.00 %	2.94 %
LP	28	1	5	34	82.35 %	2.94 %	14.71 %
LP-Q	34	0	0	34	100.00 %	0.00 %	0.00 %
LPG	33	0	1	34	97.06 %	0.00 %	2.94 %
LPGQ	34	0	0	34	100.00 %	0.00 %	0.00 %
Kohonen	24	3	7	34	70.59 %	8.82 %	20.59 %

Summary of classifications (Holdout sample)

Method	Correct class	Errors		Total number	Percents		
		SW	NE		Correct	SW	NE
RPA	27	2	3	32	84.38 %	6.25 %	9.38 %
MDA	25	4	3	32	78.13 %	12.50 %	9.38 %
MDA-Q	20	7	5	32	62.50 %	21.88 %	15.63 %
MDA-W	25	5	2	32	78.13 %	15.63 %	6.25 %
LogR	28	3	1	32	87.50 %	9.38 %	3.13 %
LP	24	5	3	32	75.00 %	15.63 %	9.38 %
LP-Q	21	7	4	32	65.63 %	21.88 %	12.50 %
LPG	25	4	3	32	78.13 %	12.50 %	9.38 %
LPGQ	21	7	4	32	65.63 %	21.88 %	12.50 %
Kohonen	16	4	12	32	50.00 %	12.50 %	37.50 %

Factor analysis

- A statistical method used to describe variability among observed variables in terms of fewer unobserved variables called factors
- The observed variables are modeled as linear combinations of the factors plus error terms
- The information gained about the interdependencies can be used later to reduce the set of variables in a dataset

Factor analysis

Factor analysis – an example: Financial ratios

Variables
ΔSales
ΔAssets
EBIT-%
ROI
ROE
CF/Sales
Equity Ratio
QR
CR

→

Factors
Growth
Profitability
Solidity

Factor analysis – an example: Financial Ratios for Finnish listed companies

- 9 variables
  - ΔSales, ΔAssets, EBIT-%, ROI, ROE, Cash Flow(Operations)/Sales, Equity Ratio, Quick Ratio Current Ratio
- Fixed number of factors: 3
  - Predefined assumption on three factors: Growth, Profitability and Solidity
- Extraction method: Principal Components Analysis
- Rotation method: Varimax

Factor analysis: Varimax-rotated component matrix

	Component		
	1	2	3
DSales (%)	,132	-,055	,953
DAssets (%)	,100	-,048	,960
EBIT-%	,869	,344	,128
CF(Oper)/Sales	,671	,183	,248
ROI	,875	,177	,003
ROE	,834	,037	,031
Equity Ratio	,274	,795	-,086
Quick Ratio	,173	,911	,011
Current Ratio	,111	,911	-,042

Factor analysis – an example: Financial ratios for Finnish listed companies§

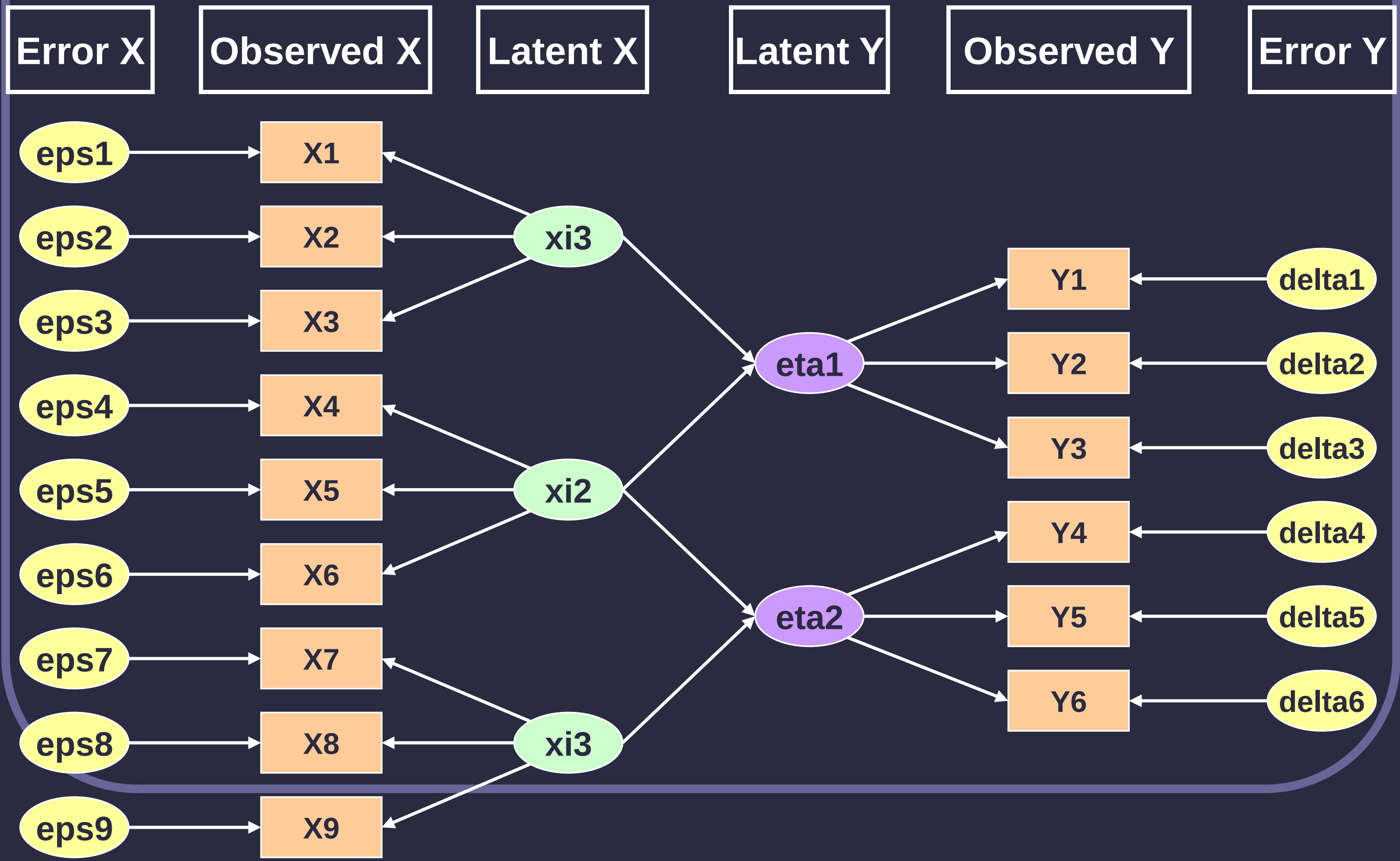
- The three pre-assumed factors – **Growth**, **Profitability** and **Solidity** – may be clearly identified in the rotated component matrix
- For example **Growth** is represented by component 3 combining the major part of ratios ΔSales and ΔAssets with minor influences from the other seven variables
- In the same manner **Profitability** is represented by component 1 and **Solidity** by component 2
- The component matrix may be further transformed into a Component score coefficient matrix to be used to create new ratios X§describing the factors

Linear Structural Relationships – Lisrel

- Lisrel-modeling combines underlying factor analyses with simultaneous estimation of structural relationship between the extracted latent factors
- The general form basic Lisrel model consists of
  - Observed explanatory variables (X)
  - Observed dependent variables (Y)
  - Latent explanatory factors (ξ)
  - Latent dependent factors (η)
  - Error (residual) terms (ε and δ) for each X- and Y-variable respectively

connected to each other as shown in the next page

Basic Lisrel model



In order to learn more about applying statistical methods...

- Participate in the course ”Advanced Financial Accounting (AFA) II”
- Lectures on statistical methods suitable for analyzing financial data and adapted to accounting terminology
- Practical assignments on each method, useful for your career

References

- Jarque, Carlos M. & Anil K. Bera (1980). “Efficient tests for normality, homoscedasticity and serial independence of regression residuals”. *Economics Letters* **6**(3): 255–259
- Massey, Frank J. Jr. (1951). “The Kolmogorov-Smirnov test for goodness of fit”, *Journal of the American Statistical Association* **46**(253): 68-78
- Pearson, Karl (1900). “On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”, *Philosophical Magazine* **50**: 157-175
- Shapiro, S. S. & M. B. Wilk (1965). “An analysis of variance test for normality”, *Biometrika,* **52**(3): 591-599