Multi-Stratification for Outlier Detection based on the Graphical Model： Evaluation by Chow Test and AIC

Dr. Kiyomi Shirakawa,
National Statistics Center, Tokyo, JAPAN

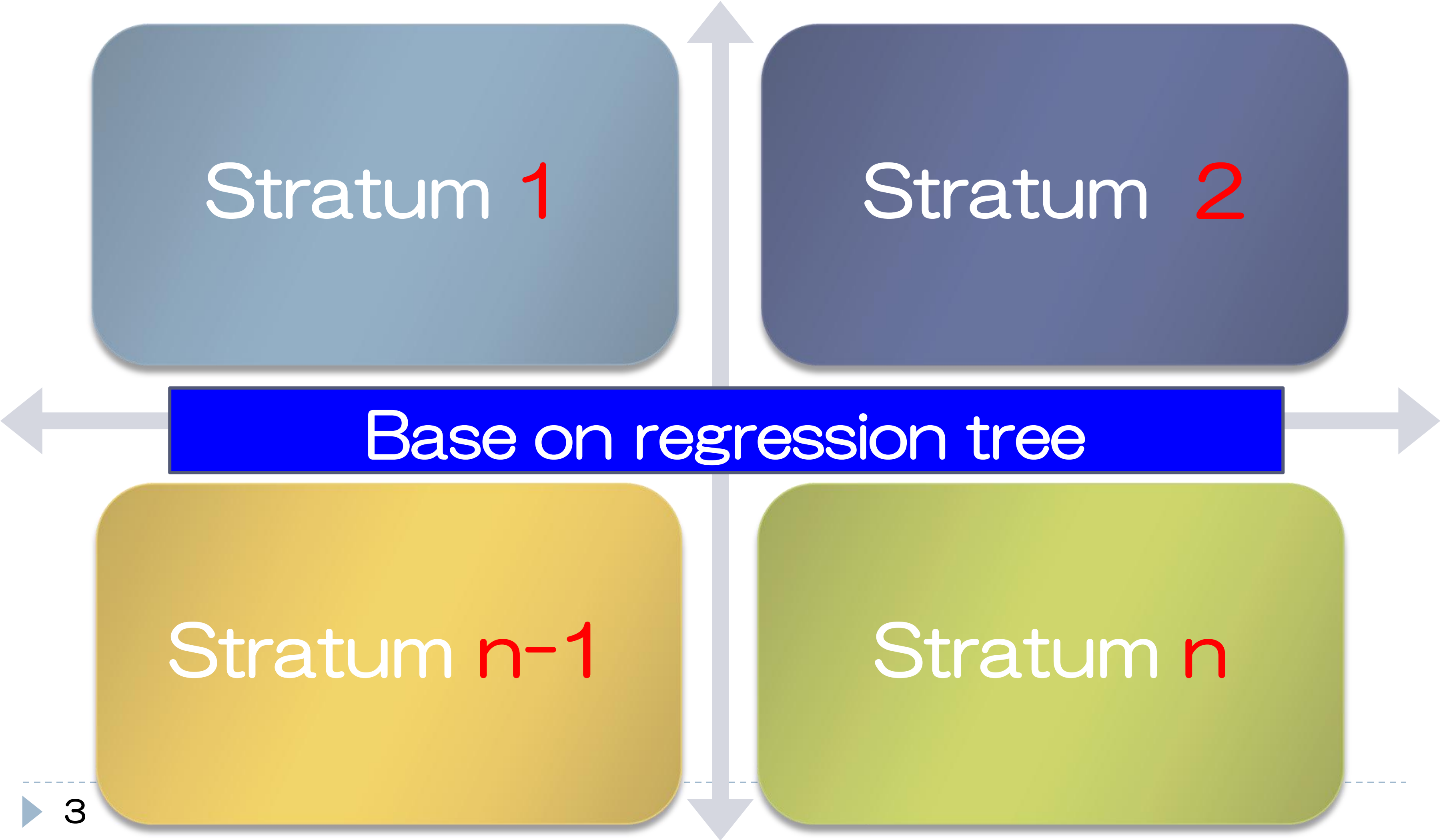
e-mail: kshirakawa@nstac.go.jp

Table of contents

1. Purpose
2. Background
- Verification Procedures
3. Data Analysis by Regression Tree
4. Evaluation of Boundary Value by Chow Test
5. Evaluation of Linear Regression Analysis for Chow Test by AIC
6. Conclusion

1. Purpose

Multi-Stratification for Outlier Detection



Stratum 1

Stratum 2

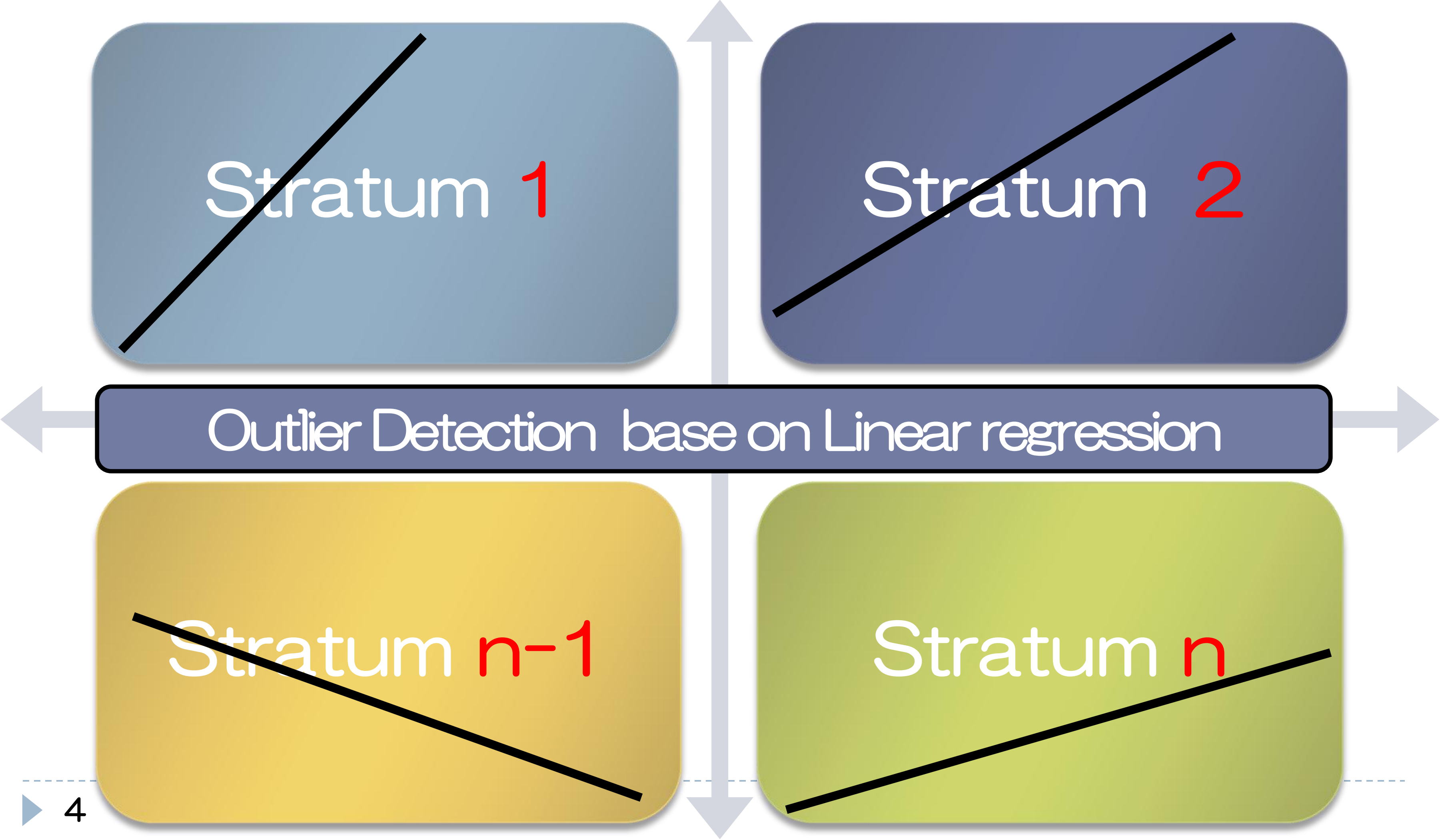
Base on regression tree

Stratum n-1

Stratum n

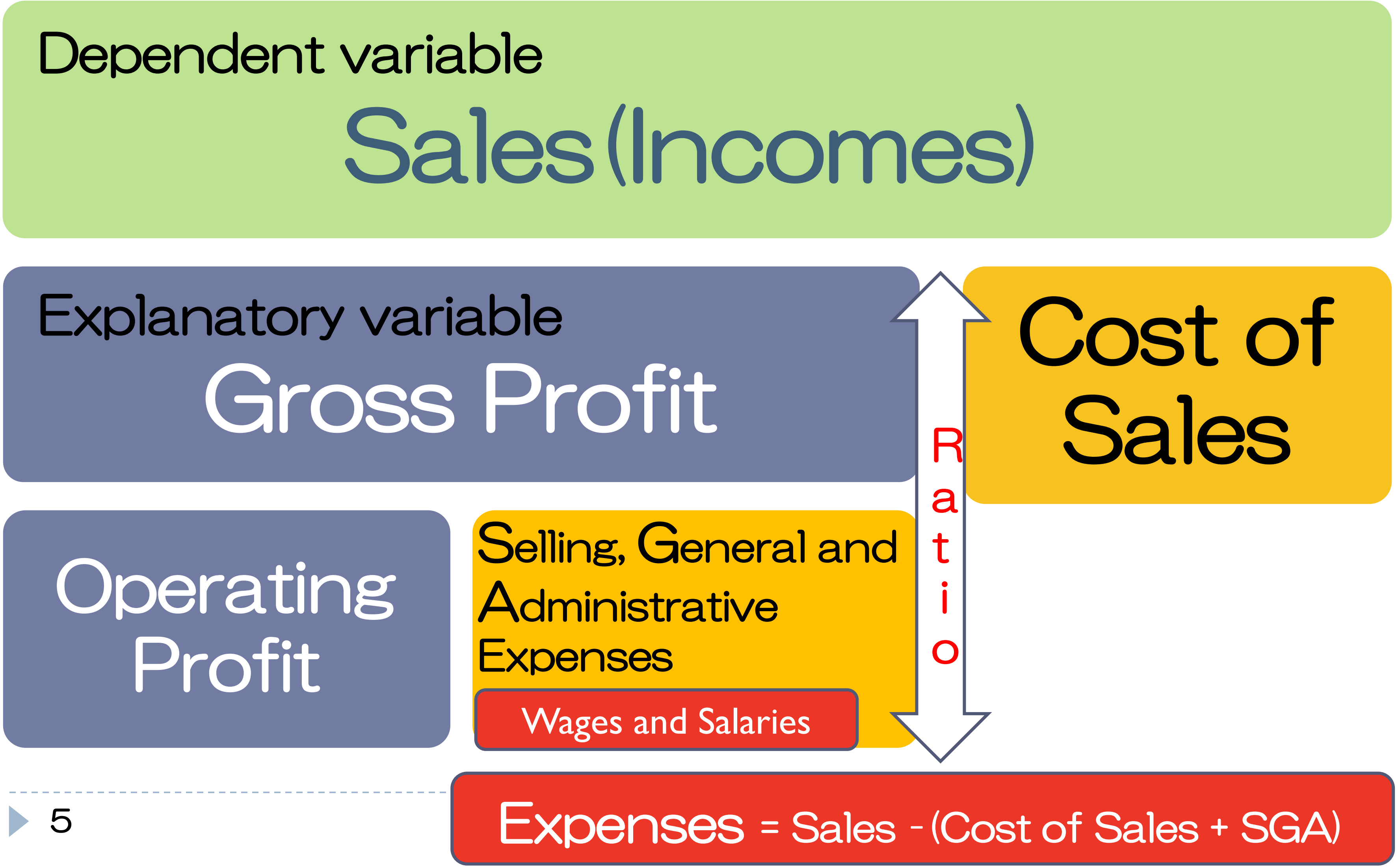
1. Purpose

Multi-Stratification for Outlier Detection



1.1 Relationship of each variable

Profit and Loss Statement



Dependent variable

Sales (Incomes)

Explanatory variable

Gross Profit

Cost of Sales

Ratio

Operating Profit

Selling, General and Administrative Expenses

Wages and Salaries

Expenses = Sales - (Cost of Sales + SGA)

1.2 Accounting items(Ratio), Tabulation of Enterprises

Item		Wholesale and Retail Trade	Manufacturing
Sales (Income)		100.0	100.0
Expenses	*2	97.2	96.1
Cost of sales		78.9	77.7
Gross profit	*3	21.1	22.3
SGA	*1	18.3	18.4
Operating profit	*4	2.8	3.9
Total wages and salaries		7.1	11.1

Data source: the 2012 Economic Census for Business Activity, Tabulation of Enterprises Table 8 in the preliminary summary, Statistics Bureau of Japan

*1 SGA: Selling and Generally Administrative expenses
*2 Expenses = Sales - (Cost of sales + SGA)
*3 Gross profit = Expense - Cost of sales
*4 Operating profit = Gross profit - SGA

1.3 Correlation coefficient between each accounting item

Correlation coefficient for the Sales is also as high as 0.9 or more.

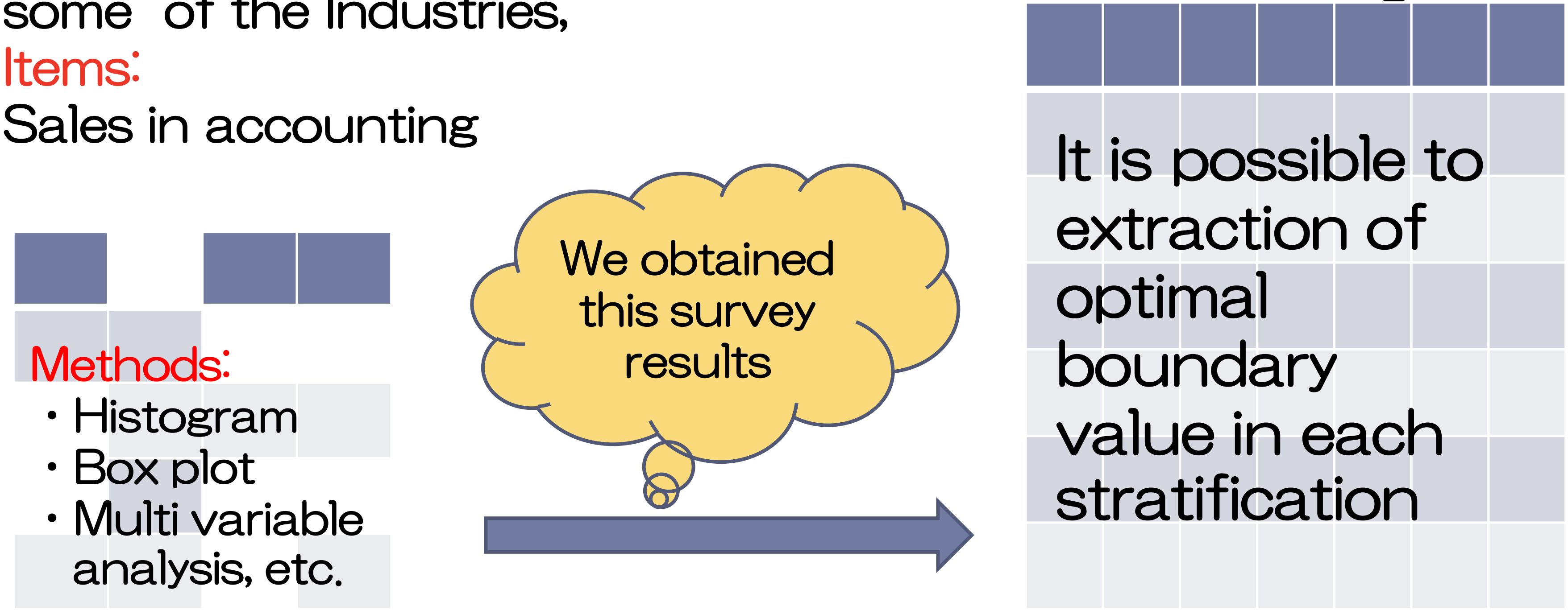
	Sales (Income)	Expenses	Cost of Sales	Gross profit	SGA	Operating profit	TWS
Sales (Income)	1.000						
Expenses	1.000	1.000					
Cost of sales	0.999	0.999	1.000				
Gross profit	0.988	0.987	0.981	1.000			
SGA	0.990	0.989	0.983	0.999	1.000		
Operating profit	0.953	0.950	0.943	0.979	0.970	1.000	
TWS	0.950	0.948	0.943	0.960	0.955	0.961	1.000

SGA: Selling and Generally Administrative expenses
TWS: Total Wages and Salaries

2 Background

The 2012 Economic Census for Business Activity was held in Japan.

Targets：
Establishments in some of the Industries,
Items:
Sales in accounting



All establishments,
Main accounting items

Methods:
- Histogram
- Box plot
- Multi variable analysis, etc.

We obtained this survey results

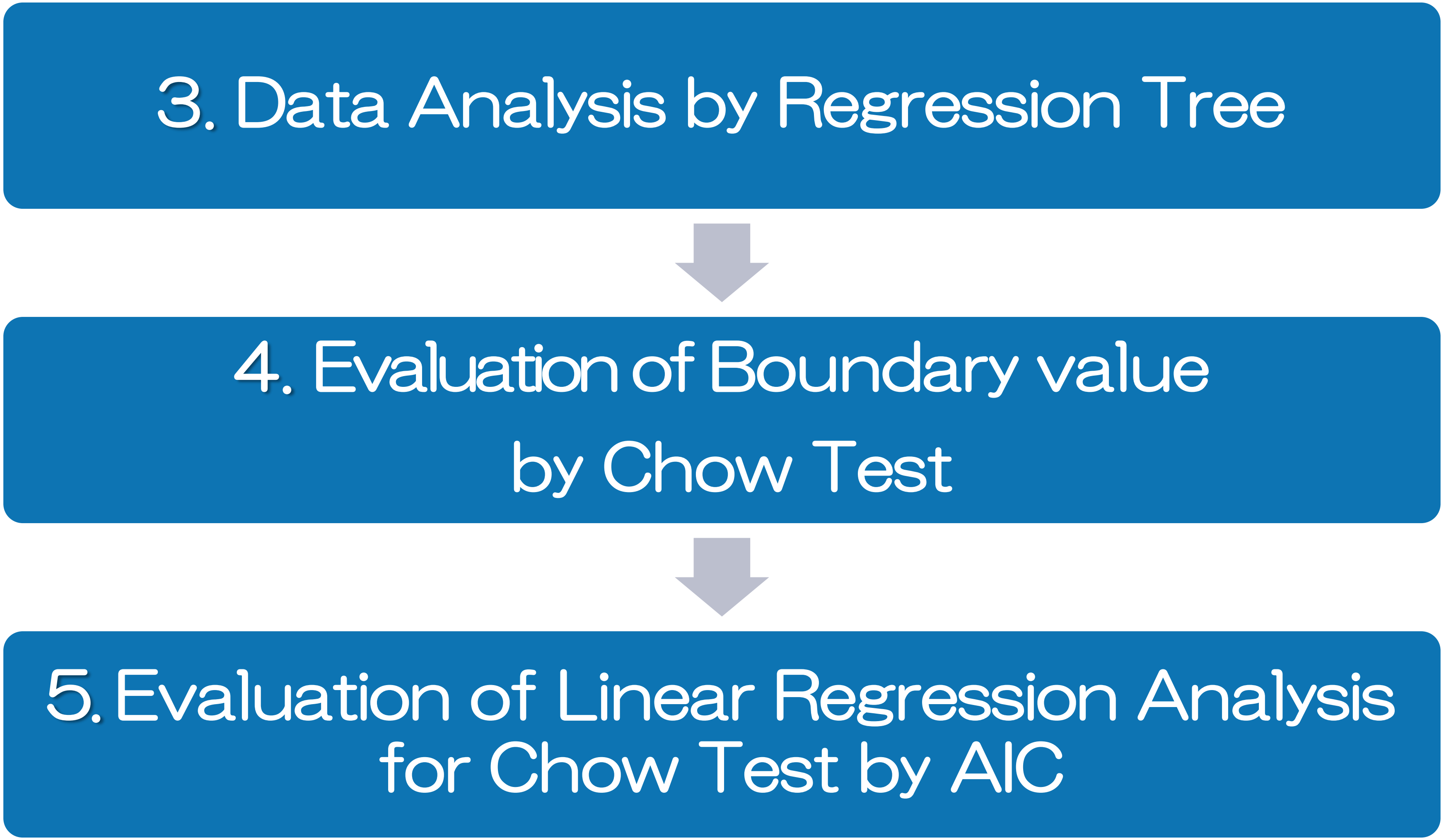
It is possible to extraction of optimal boundary value in each stratification

Kind of histogram analysis：
Evaluation for each method based on the AIC

Sample size	10	20	30	50	100	200	500	1000
Minimum	14.937	17.879	18.450	16.874	16.825	16.961	15.714	14.937
Maximum	24.699	23.699	25.359	23.770	26.153	27.659	26.347	27.383
Sample mean	20.657	21.021	21.296	20.227	21.217	21.024	21.034	20.980
USSD ∗1	3.273	1.590	1.676	1.628	1.929	2.025	1.927	2.021
IQR	3.430	2.299	1.755	2.330	2.719	3.087	2.865	2.793
(i) Sturges' formula								
Num. of bins	4	5	6	7	8	9	10	11
AIC	40.04	99.51	168.76	322.31	757.99	1,764.22	5,275.23	-
(ii) Scott's normal reference rule								
Num. of bins	2	3	4	5	7	9	13	18
AIC	34.85	95.02	164.66	318.86	761.87	1,774.81	5,303.48	11,943.10
(iii) Freedman-Diaconis' choice								
Num. of bins	4	4	7	6	8	11	15	23
AIC	42.92	99.55	177.62	323.63	767.23	1,782.76	5,315.59	11,972.54

∗1 USSD: Uncorrected sample standard deviation

Verification Procedures



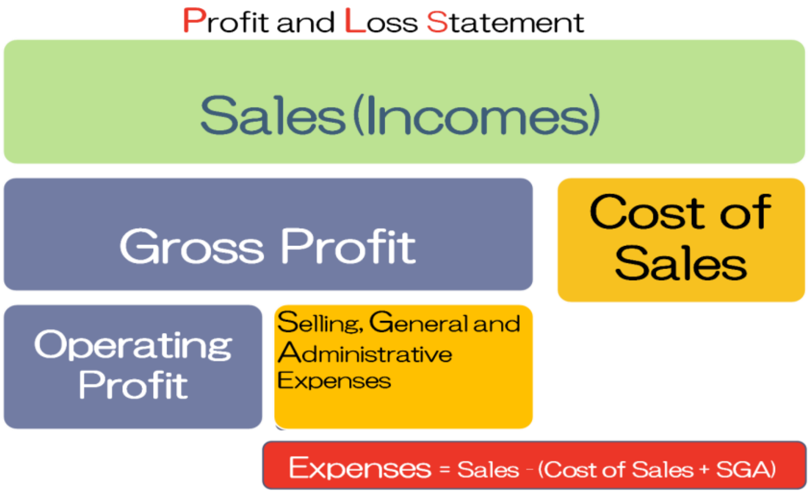
3. Data Analysis by Regression Tree

4. Evaluation of Boundary value by Chow Test

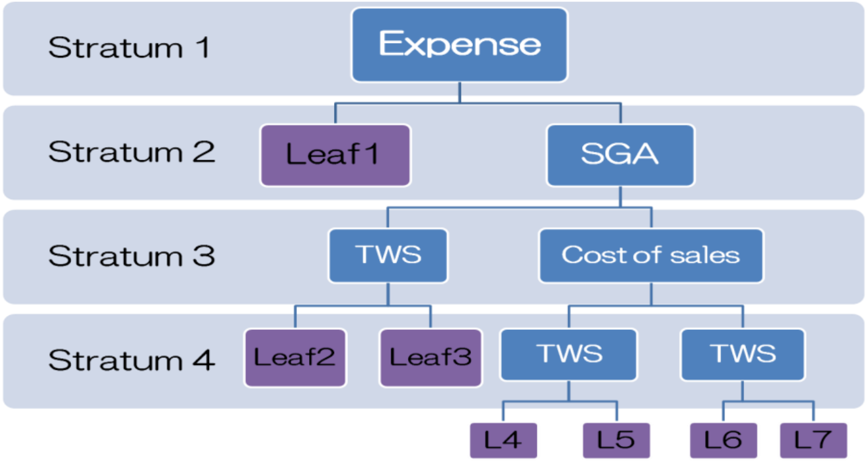
5. Evaluation of Linear Regression Analysis for Chow Test by AIC

Illustrate of Verification Procedures

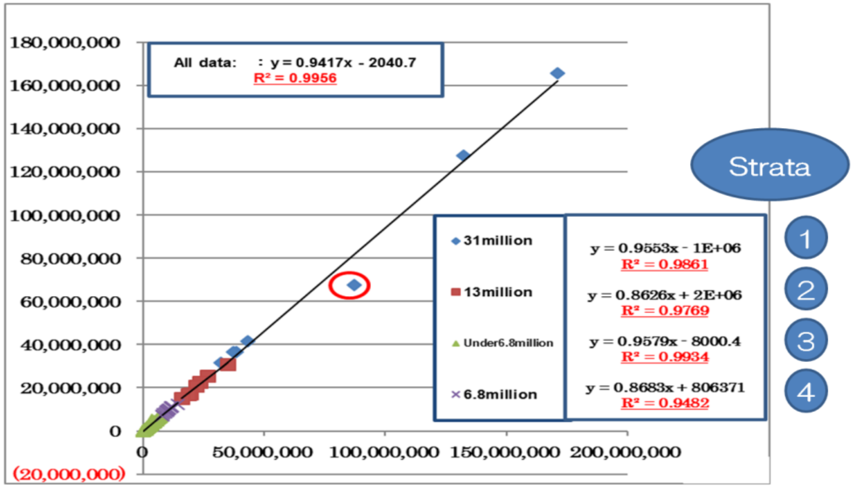
①Data Analysis


Profit and Loss Statement
Sales(Incomes)
Gross Profit
Cost of Sales
Operating Profit
Selling, General and Administrative Expenses
Expenses = Sales - (Cost of Sales + SGA)

② Stratified based on regression tree



③ Evaluation of Boundary value



④ Linear Regression Analysis

Coefficient	(Intercept)	Expense	Cost of sales	SGA	TWS	df	AIC
lm All	4605.5	1.021	-0.026	0.236	-0.060	6	18,917.4
lm 1	-1679.7		1.042	0.346	2.075	5	22,158.6
lm 2	-12023.5		1.083		2.457	4	22,167.0
lm 3	146400.0		1.023	1.263		4	22,293.6
lm 4	267900.0		1.296			3	22,489.0

Data Analysis

1. Data set
   - The 2012 Economic Census for Business Activity, Tabulation of Enterprises Table8 in the preliminary summary
   - Dependent variable：Sales (Income)
   - Explanatory variable：Expenses, so on
2. Method
   - The introduction of Regression Tree
   - R package of "mvpart"
3. Evaluation
   - Boundary value by Chow Test and AIC

List of calculation for histogram by Sturges' formula

No	Data section		(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Minimum	Maximum	Freq. ratio (Theoretical value)	Cumulative freq.	Freq. (n=721)	Ratio of (3)	Ratio of (2)	(3) × Ln(4)	Ln(3)!
1	89	17,130,280	0.95907	0.959	708	0.982	0.982	-12.9	—
2	171,30,280	34,260,471	0.04081	1	6	0.008	0.990	-28.7	6.579
3	34,260,471	51,390,662	0.00012	1	4	0.006	0.996	-20.8	3.178
4	51,390,662	685,20,853	1.1E-08	1	0	0	0.996	0	0
5	68,520,853	85,651,044	2.8E-14	1	0	0	0.996	0	0
6	85,651,044	102,781,235	0	1	0	0	0.996	0	0
7	102,781,235	119,911,426	0	1	0	0	0.996	0	0
8	119,911,426	137,041,617	0	1	1	0.001	0.997	-6.58	0
9	137,041,617	154,171,808	0	1	0	0	0.997	0	0
10	154,171,808	171,301,999	0	1	1	0.001	0.999	-6.58	0

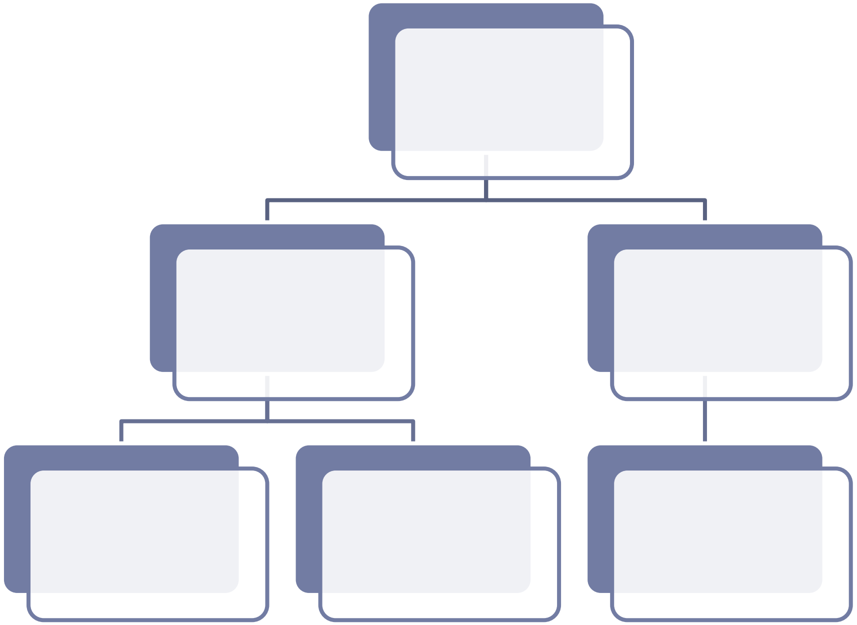
$$\mathrm{AIC} = (-2) \times (-6.58 - 0) + 2(10-1) = \underline{31.16}$$

3. Data Analysis by Regression Tree

Tree-based model has various main advantages:

(i) Simple to understand and interpret

$$GI = 1 - \sum_{i=1}^{m} [p(i|t)]^2$$ GI: Gini index



(ii) Able to handle both numerical and categorical data

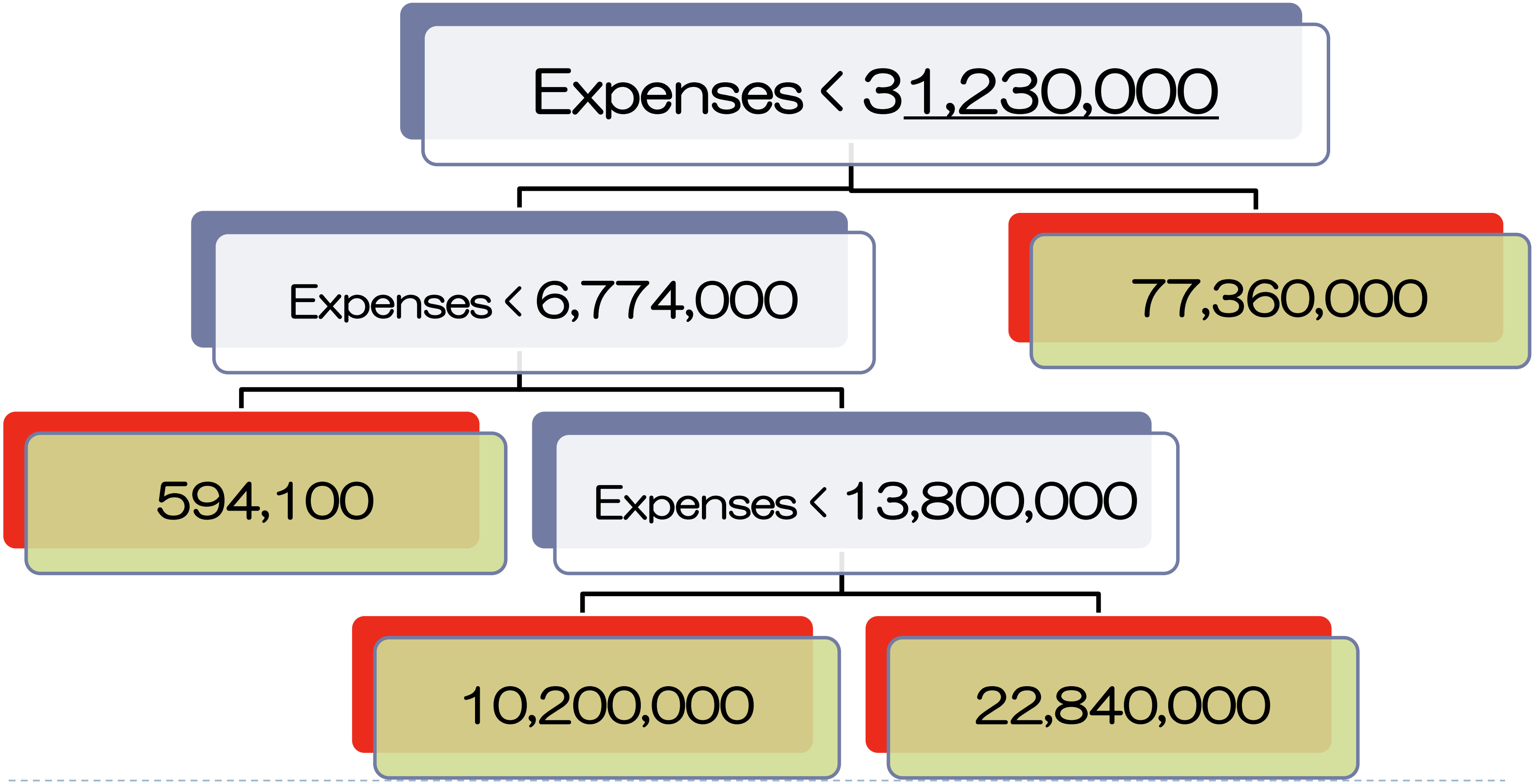
(iii) Uses a white box model and probabilistic graphical model

(iv) Performs well with large datasets

(v) Supervised learning, and prediction

3.1 Result of Analysis

- The Sales is computed by the Expenses in the explanatory variable.



Expenses < 31,230,000

Expenses < 6,774,000

77,360,000

594,100

Expenses < 13,800,000

10,200,000

22,840,000

3.2 Analysis Results by Other Variables

(i) When omitted the Expenses:

Node), Split	n	Deviance	Y value	
1) root	543	5.643E+16	2,125,002	
2) SGA < 4,577,904	536	4.84E+15	1,263,691	
4) wages and salaries< 784,186.5	510	5.39E+14	735,514	*
5) wages and salaries>=784,186.5	26	1.37E+15	11,624,090	*
3) SGA>=4,577,904	7	2.07E+16	68,076,790	*

The SGA and the wages and salaries are effective to split, the sales is divided by three classes.

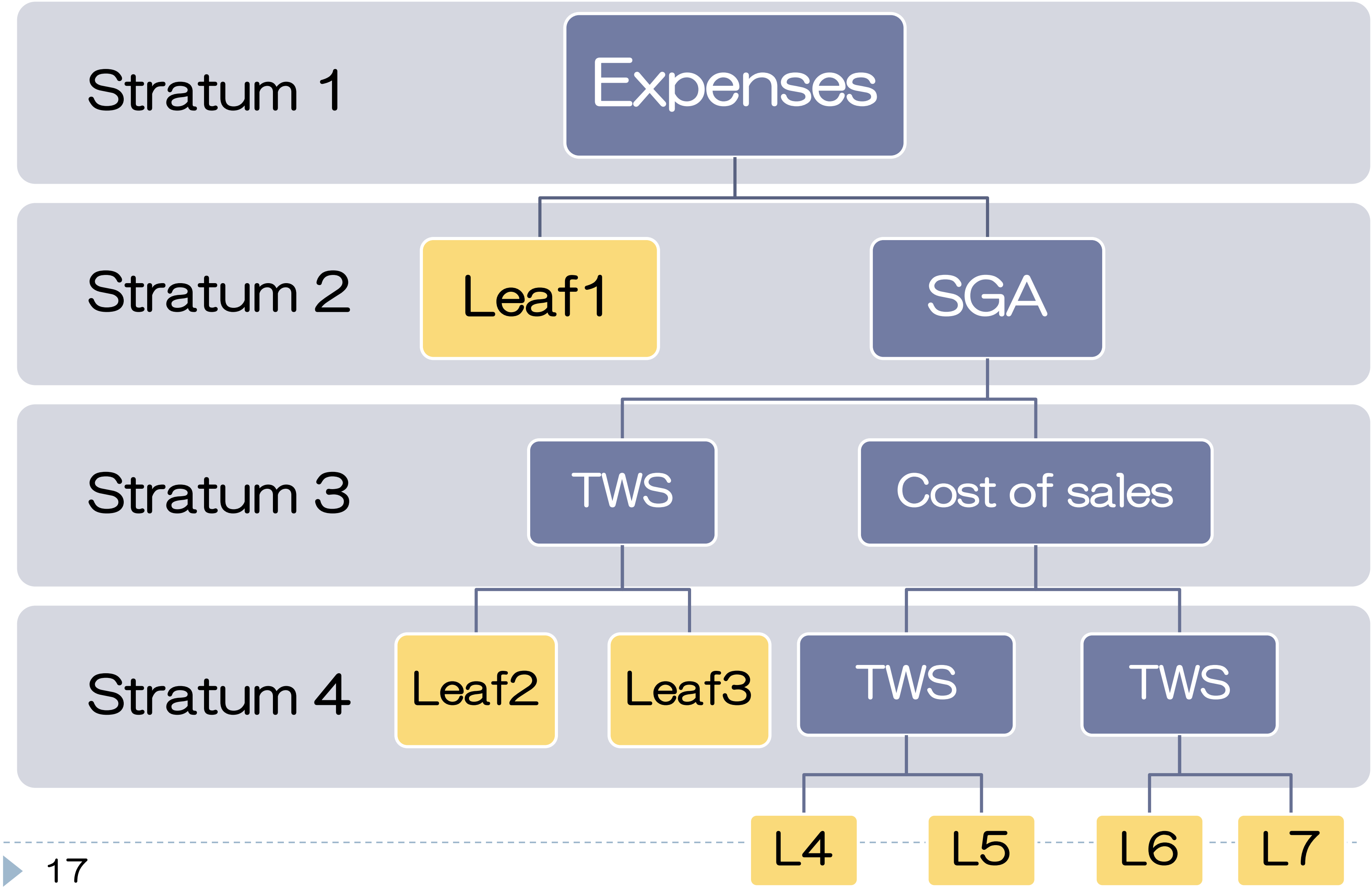
(ii) When omitted the Expenses and SGA:

The Sales is divided four classes.

(iii) When omitted the Expenses, SGA and Cost of sales:

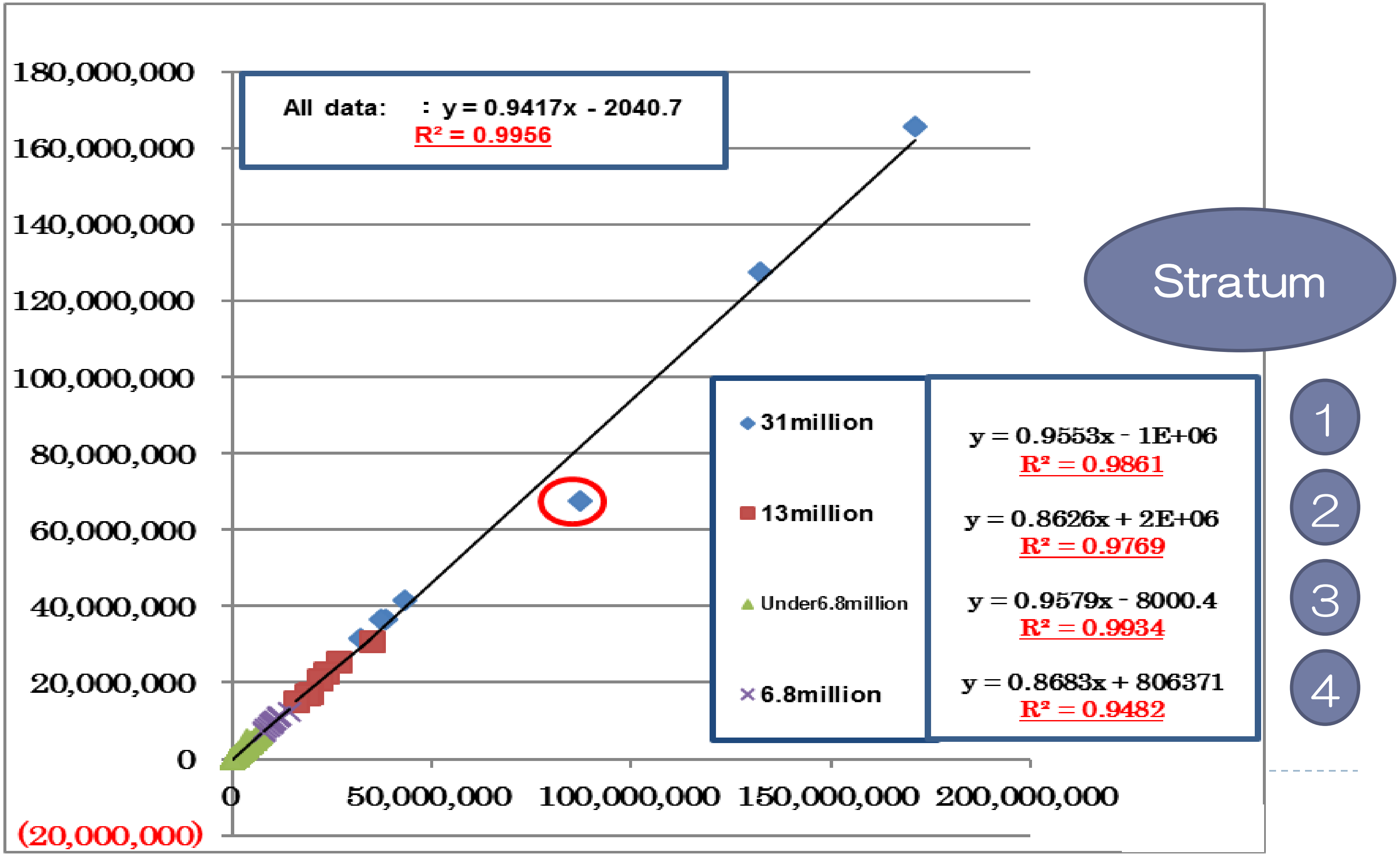
The Sales is divided four classes.

3.3 Integrated some analysis results

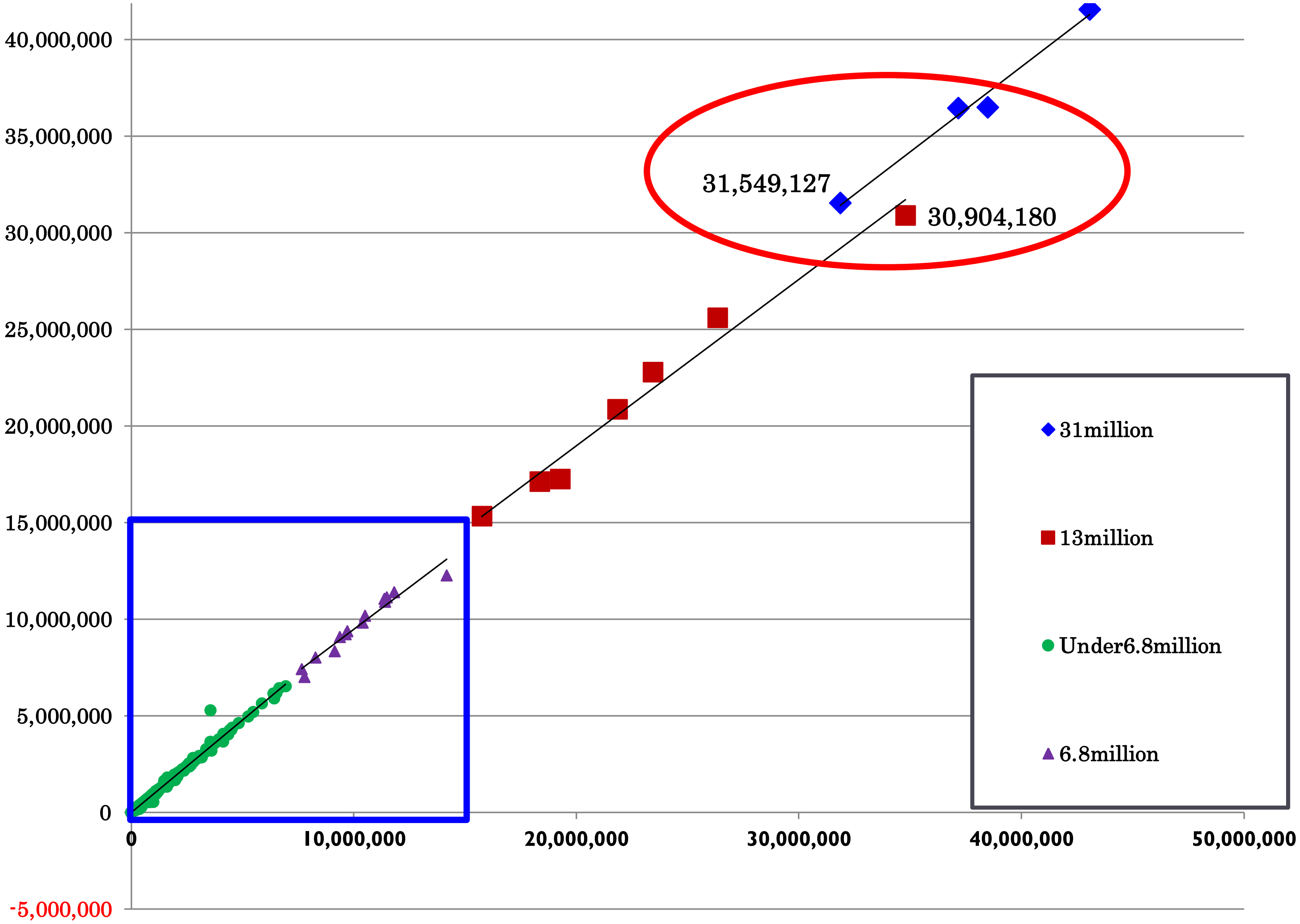


4. Evaluation of Boundary Value by Chow Test

Dependent variable is the sales, and explanatory variable is the Expenses.



4. Evaluation of Boundary Value by Chow Test



4.Evaluation of Boundary Value by Chow Test

The Expenses was divided boundary value of under 6.8 million and 6.8 million to 13 million yen by each stratification.

- Result of the Chow Test
  - F = 20.0103, df1 = 2, df2 = 781,
  - P-value = 3.35e-09
  - Evaluation of F value:
  - When $1 \leqq F \leqq F\alpha$, $P > 0.05$ is equal variables,
  - And $F > F\alpha$, $P < 0.05$ is unequal variables.

- P value is under 0.05, therefore, its boundary value is effective.

5. Evaluation of Linear Regression Analysis for Chow Test by AIC

Results of linear regression analysis

Coefficient	(Intercept)	Expenses	Cost of sales	SGA	TWS	df	AIC
lm All	4,605.5	1.021	-0.026	0.236	-0.060	6	18,917.4
lm 1	-1,679.7		1.042	0.346	2.075	5	22,158.6
lm 2	-12,023.5		1.083		2.457	4	22,167.0
lm 3	146,400.0		1.023	1.263		4	22,293.6
lm 4	267,900.0		1.296			3	22,489.0

SGA: Selling and Generally Administrative expenses
TWS: Total Wages and Salaries

6 Conclusion

- Achievement of the study
  1. Multi-stratification of the Sales based on the regression tree
     - Evaluation
  2. Boundary value by Chow Test
  3. Linear Regression Analysis for Chow Test by AIC
- Future research is an extension to other economic surveys based on the experience of authentic information in the aggregate the EC2012.

References

[1] Kiyomi Shirakawa, A post-aggregation error record extraction based on naive Bayes for statistics survey enumeration. 59th ISI world Statistics Congress (2013), Hong Kong, China.

http://www.statistics.gov.hk/wsc/CPS004-P4-S.pdf

[2] Sturges, H. A. “The choice of a class interval” . (1926). J. American Statistical Association: 65-66.

[3] Scott, David W. (1979). “On optimal and data-based histograms” . Biometrika 66 (3): 605-610. doi:10.1093/biomet/66.3.605

[4] Freedman, David; Diaconis, P. “On the histogram as a density estimator: L2 theory” . (1981).Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57 (4): 453-476. doi:10.1007/BF01025868

[5] Akaike, H., "Information theory and an extension of the maximum likelihood principle", Proceedings of the 2nd International Symposium on Information Theory, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: 267-281 (1973).

[6]Cristopher. M. Bishop, Pattern Recognition and Machine Learning, Springer, (2006)

[7] Kiyomi Shirakawa, Teisei deta ni motoduku kigyoukouzou no choukaichi bunseki, Japanese Joint Statistical Meeting, (2013), Osaka .(In Japanese)

[8]Takayuki Ito, Kiyomi Shirakawa, Keirikoumoku ni motoduku kigyou no kouzouka bunseki: Kouzou no kyoukaichi kentei, Japanese Joint Statistical Meeting, (2013), Osaka .(In Japanese)

[9] Kiyomi Shirakawa, Keizai sensasu kisochousa shukeikekka ni motoduku kigyou group ni kansuru kousatsu, Japanese Joint Statistical Meeting, (2012), Hokkaido. (In Japanese)

[10] Masato, Okamoto.　“Tahenryo Hazurechi kenshutsu no kenkyu douko oyobi Canada oroshiuri kourigyochosa ni okeru tahenryo hazurechi kenshutsuho “, Seihyo gijyutsu kenkyu report 1, National Statistics Center.（2004）（Non-disclosure)(In Japanese)

[11] Gordon S. Linoff, Michael J. A. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley,（2011）

[12] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Applied Statistics, Vol. 29, No.2.(1980), PP. 119-127.

[13] Breiman, L. Friedman, J. H. Friedman, and Stone Olshen. ″CJ, 1984. Classification and regression trees.″ Pacific Grove, Kalifornien

[14] Dawid, A. Philip. ″Conditional independence in statistical theory.″ Journal of the Royal Statistical Society. Series B (Methodological)（1979）: 1-31.

Thank you very much for your attention.

e-mail: kshirakawa@nstac.go.jp