



# Statistical-Modeling Approach for Limited Data Forecast —The Japanese Secret of Low-Maturity and High-Quality—

Shinobu Minamikawa and Yoshinobu Yamamura  
YARNE and Company



# Purpose

The Purpose of this presentation is:

based on our practices, we explain the method to analyze and predict performance of organizations and projects with limited and possibly poor quality data.



# Background 1

## Organizations

Management needs reasonable analysis and prediction to estimate their budget for the next year –without further cost to collect and analyze data.

## Projects

One of the goals of projects is to deliver products which work(at least).

The more they would get budget, the more they would use it for time and people – not for data.



## Background 2

With poor quality data, we need to make reasonable analysis and suggestion to the management.

Have you done this before?



## In Japan

For most Japanese organizations, CMMI is still one of the brand-new methods of process improvements. There are small numbers of CMMI high-maturity companies.

But “made in Japan” is the brand.

What is the key?



## 4 steps

- 1 Identify the problems
- 2 gather objective data
- 3 use statistical analysis
- 4 build a model using a forecasting target variable



## Step1: Identify the problems

What is the biggest concern for our organization?

What does the management need to know in the  
whitepaper this year?

...



## Step 2: Gather objective data

e.g. money, a number of staffs, lines of code, time in each phase, a number of errors found in each test, etc.

We carefully omit subjective data such as how difficult the projects are, how skillful the staffs are, etc.





## Step 3: Use statistical analysis

Try statistical analysis listed below:

- regression
- correlation
- analysis of variance



## Step 4: Build a model

Build a model using a forecasting target variable (e.g., number of failures, etc.) with usable data obtained above.

Example of Step 1:



How can we decrease defects?

*Yarn*

## Example of Step 2:

This is a sample data excerpted from a real performance data set of an organization. It is collected for “application” whitepaper.


Subjective!

ID	term	total size	distribute d system	host system	framew arok	% to framew ork	aaa	a+b	ratio
1	200803								
2	200603	732937	469481	64	263456	36			
3	200609	720852	454046	63	266806	37	3506	1248	35
4	200703	640863	371779	58	269084	42	2827	1035	36
5	200809	627579	361951	58	265628	42	2805	1025	36
6	afterd2	558199	353719	63	204480	37			41

*Yarn*

## Example of Step 2-(2):

Then make a subset excel data.



ID	term	total size	distribute d system	host system	framew arok	% to framew ork	num	num2	num3	oth
1	200803						551	491	54	
2	200603	732937	469481	64	263456	36	459	387	53	
3	200609	720852	454046	63	266806	37	340	267	57	
4	200703	640863	371779	58	269084	42	257	179	66	
5	200809	627579	361951	58	265628	42	231	160	61	
6	afterd2	558199	353719	63	204480	37				

Yam

## Example of Step 3:

Import this subset data to “minitab” and see if there is significant correlation between variables.

The screenshot shows the Minitab software interface. The top window is the 'セッション' (Session) window, which displays the date and time '2008/10/20 15:50:46' and a message: 'MINITABへようこそ、ヘルプを表示するにはF1を押してください。' (Welcome to MINITAB, press F1 to display help). The bottom window is the 'ワークシート 1 \*\*\*' (Worksheet 1) window, which contains a data table with 12 columns (C1 to C11) and 9 rows. The data is as follows:

	C1	C2-T	C3	C4	C5	C6	C7	C8	C9	C10	C11
	ID	term	total size	distributed system	host system	framework	% to framework	num	num2	num3	others
1	1	200803	*	*	*	*	*	551	491	54	6
2	2	200603	732937	469481	64	263456	36	459	387	53	19
3	3	200609	720852	454046	63	266806	37	340	267	57	16
4	4	200703	640863	371779	58	269084	42	257	179	66	12
5	5	200809	627579	361951	58	265628	42	231	160	61	10
6	6	afterd2	558199	353719	63	204480	37				
7											
8											
9											

Yam

# Example of Step3-(2):

Result and formula are as shown below:

MINITAB-20080218whitepaper.MPJ - [セッション]

ファイル(F) 編集(E) データ(A) 計算(C) 統計(S) グラフ(G) エディタ(D) ツール(T) ウィンドウ(W) ヘルプ(H)

回帰分析: [redacted]

回帰式  
トラブル件数 = - 150 + 0.140 [redacted]

3つのケースが使用されました。2つケースには欠損値が含まれています

予測変数	Coef	標準誤差Coef	T	p値
定数	-150.21	88.13	-1.70	0.338
[redacted]	0.13992	0.02877	4.86	0.129

S=18.2142 R二乗=95.9% R二乗 (調整済) 値=91.9%

分散分析

変動源	自由度	平方和	平均平方	F値	p値
回帰	1	6219.1	6219.1	23.88	0.129
残差誤差	1	262.9	262.9		
合計	2	6482.0			

見かけない観測値

観測値	トラブル件数	適合値	標準誤差適合値	残差	
3	350.6	340.00	340.37	18.21	-0.37

観測値 標準化残差  
3 -1.00 X

XIは、X値が大きな影響を与える観測値を示します。

回帰分析: トラブル件数対内部 + 外部

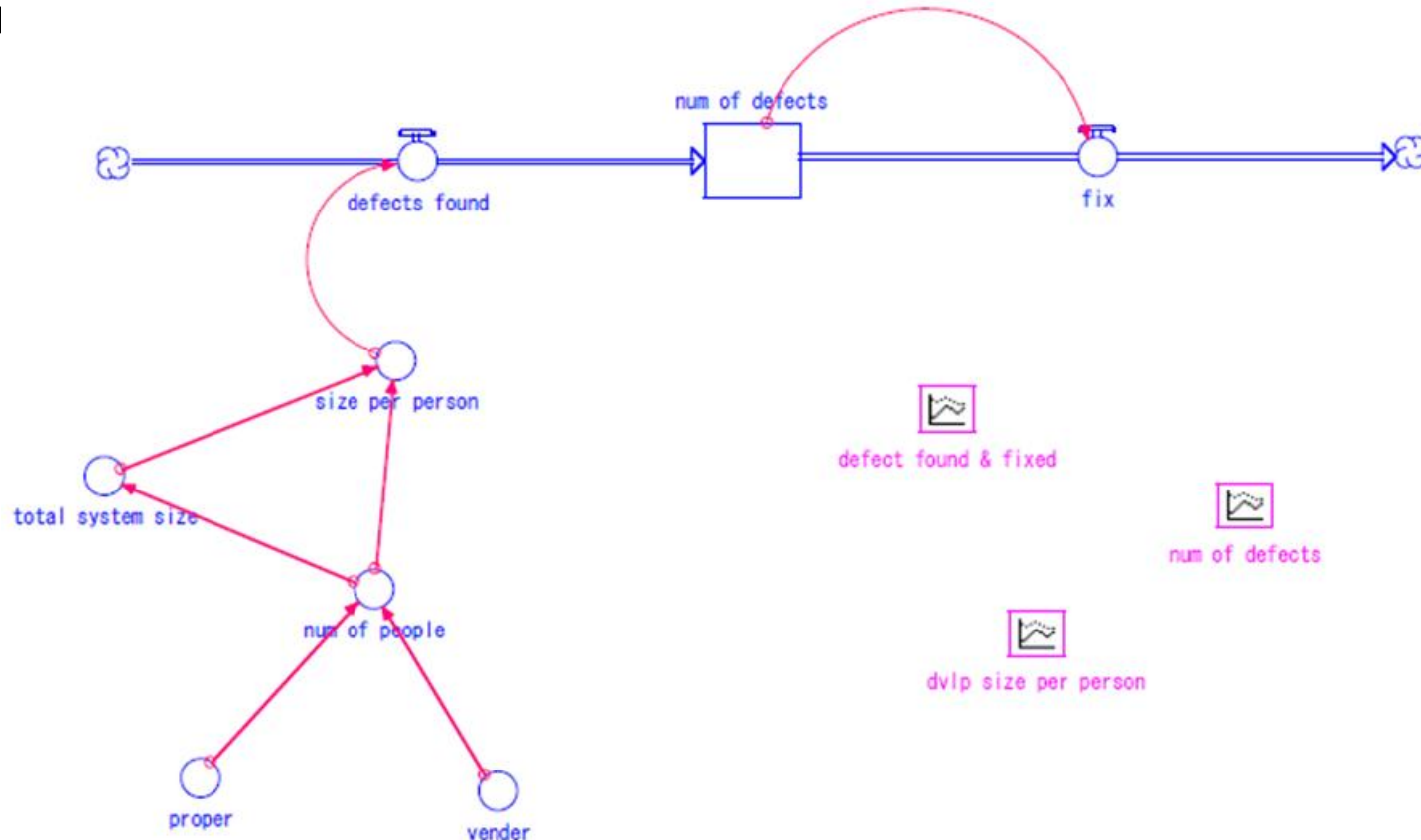
回帰式

Num. of defects  
= - 150 + 0.140 total system size / (proper + vender(FP))

*Yarn*

## Example of Step 4:

Input the identified variables and the formula to “ithink”, and make a model. An example can be as shown below

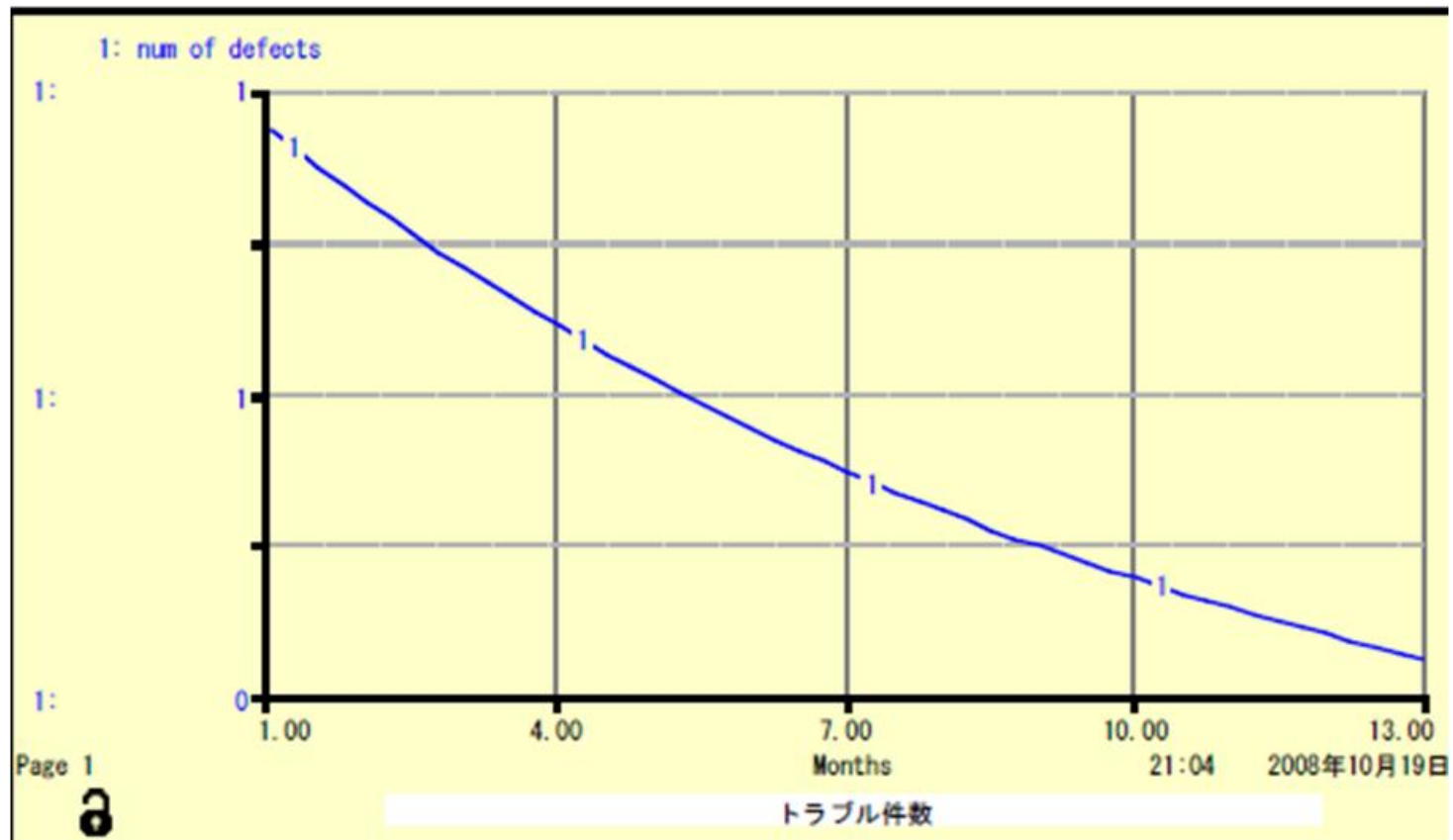




Yamada

# Example of Step 4-(2):

Run the model and see the output.





## Result 1

These metrics shows:

- Number of defects is decreasing at a certain rate and is on the ground in 14 months.

→assign appropriate amount of cost(people) to the organization for 14 months.

It would be better than taking a risk to upset the on-going project with causing a big change to decrease defects.



## Result 2

Management needs to know the method to decrease defects.

Model shows if measures listed below can be decreased, number of defects can be decreased.

- Total size of system
- Develop size per person



## Result 3

→ It is effective to educate engineers to know the way not to develop:

- design techniques should be skilled.
- general knowledge of system and the organization should be understood.

This is an investment plan, not a painkiller.



Another Example of Step 1:

How can we prevent project overrun?



## Example of Step 2:

This is a sample data excerpted from a real performance data set of an organization. It is collected for project whitepaper.

Subjective!

term	ID	PRMKey	PJterm	Pjname	system	dep	group	difficulty	critical	SI	ratio	mng	are
2003A	001	2003A001	15	a	w	1	s	B	L	TRUE	34		a
2003A	002	2003A002	15	b	e	2	g	B	M	FALSE	2		a
2003A	003	2003A003	15	c	g	3	c	C	H	FALSE	65		a
2003A	004	2003A004	15	d	q	4	j	A	L	TRUE	34		b
2003A	005	2003A005	15	e	v	5	s	A	M	TRUE	23		c
2003A	006	2003A006	15	a	w	6	b	A	H	TRUE	9		c
2003A	012	2003A012	15	b	e	7	s	AA	L	FALSE	18		b
2003B	001	2003B001	15	c	g	8	g	H	M	FALSE	34		b
2003B	002	2003B002	15	d	q	9	s	B	H	FALSE	2		a
2003B	003	2003B003	15	e	v	10	s	C	L	FALSE	65		c
2003B	004	2003B004	15	a	w	11	g	A	M	TRUE	34		b
2003B	005	2003B005	15	b	e	12	c	A	H	FALSE	23		c
2003B	006	2003B006	15	c	g	13	j	C	L	TRUE	9		c
2003B	007	2003B007	15	d	q	14	s	B	M	TRUE	18		b

*Yarn*

## Example of Step 2-(2):

Then make a subset excel data.

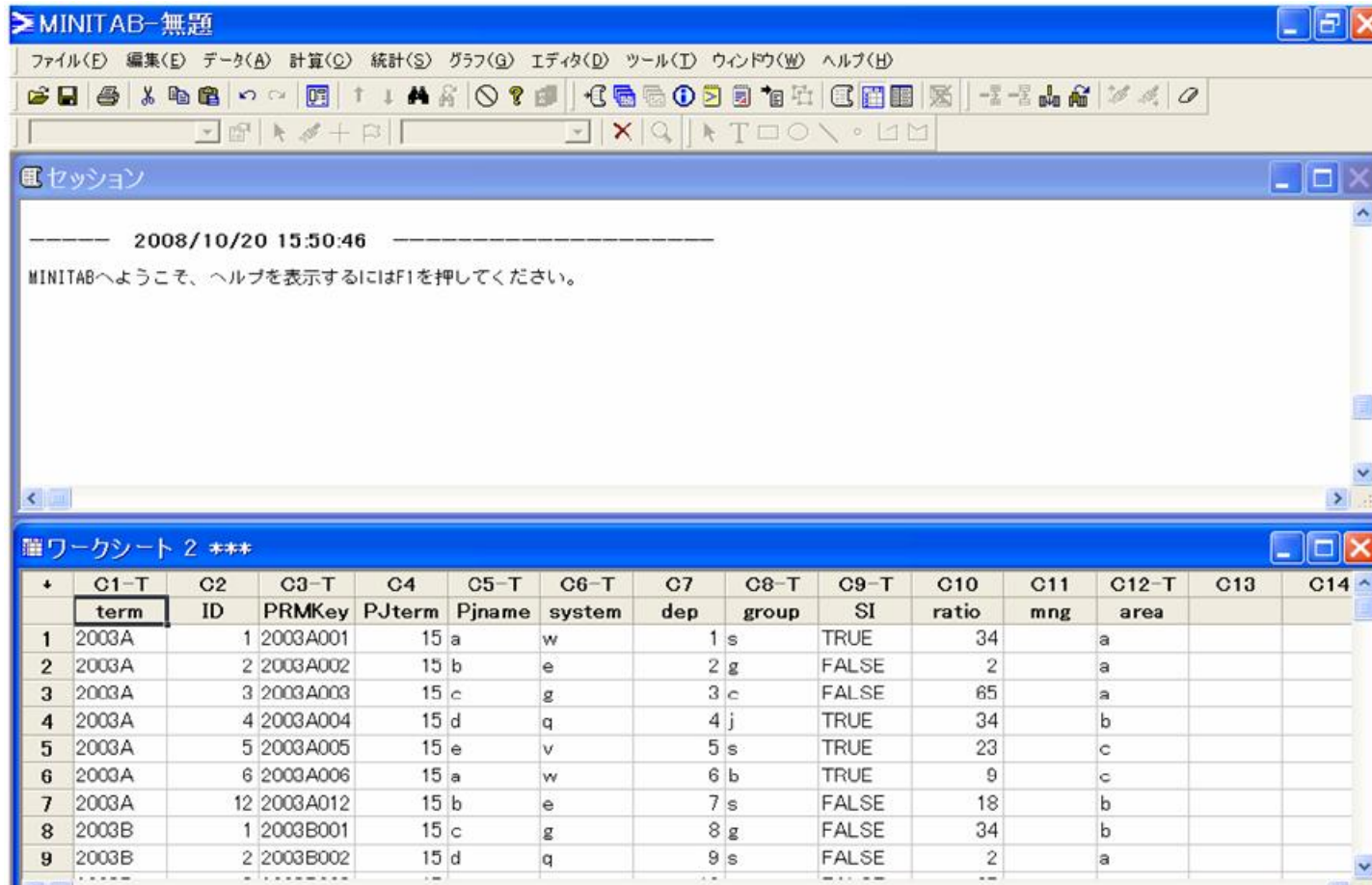


term	ID	PRMKey	PJterm	Pjname	system	dep	group	SI	ratio	mng	are:
2003A	001	2003A001	15	a	w	1	s	TRUE	34		a
2003A	002	2003A002	15	b	e	2	g	FALSE	2		a
2003A	003	2003A003	15	c	g	3	c	FALSE	65		a
2003A	004	2003A004	15	d	q	4	j	TRUE	34		b
2003A	005	2003A005	15	e	v	5	s	TRUE	23		c
2003A	006	2003A006	15	a	w	6	b	TRUE	9		c
2003A	012	2003A012	15	b	e	7	s	FALSE	18		b
2003B	001	2003B001	15	c	g	8	g	FALSE	34		b
2003B	002	2003B002	15	d	q	9	s	FALSE	2		a
2003B	003	2003B003	15	e	v	10	s	FALSE	65		c
2003B	004	2003B004	15	a	w	11	g	TRUE	34		b
2003B	005	2003B005	15	b	e	12	c	FALSE	23		c
2003B	006	2003B006	15	c	g	13	j	TRUE	9		c
2003B	007	2003B007	15	d	q	14	s	TRUE	18		b
2003B	008	2003B008	15	e	v	15	s	TRUE	23		c

Yam

# Example of Step 3:

Import this subset data to “minitab” and see if there is significant correlation between variables.



The screenshot shows the Minitab software interface. The main window displays a data table with 14 columns and 9 rows of data. The columns are labeled C1-T through C14, and the rows are numbered 1 through 9. The data includes terms, IDs, PRMKeys, PJterms, Pjnames, systems, departments, groups, SI values, ratios, mng values, and areas.

	C1-T	C2	C3-T	C4	C5-T	C6-T	C7	C8-T	C9-T	C10	C11	C12-T	C13	C14
	term	ID	PRMKey	PJterm	Pjname	system	dep	group	SI	ratio	mng	area		
1	2003A	1	2003A001	15	a	w	1	s	TRUE	34		a		
2	2003A	2	2003A002	15	b	e	2	g	FALSE	2		a		
3	2003A	3	2003A003	15	c	g	3	c	FALSE	65		a		
4	2003A	4	2003A004	15	d	q	4	j	TRUE	34		b		
5	2003A	5	2003A005	15	e	v	5	s	TRUE	23		c		
6	2003A	6	2003A006	15	a	w	6	b	TRUE	9		c		
7	2003A	12	2003A012	15	b	e	7	s	FALSE	18		b		
8	2003B	1	2003B001	15	c	g	8	g	FALSE	34		b		
9	2003B	2	2003B002	15	d	q	9	s	FALSE	2		a		





# Example of Step3-(2):

Result and formula are as shown below:

MINITAB-MUSK\_WPwork02.MPJ - [セッション]

ファイル(F) 編集(E) データ(A) 計算(C) 統計(S) グラフ(G) エディタ(D) ツール(T) ウィンドウ(W) ヘルプ(H)

回帰分析: 目的変数対初期トラブル件数

回帰式  
目的変数 = 0.119 + 0.0314 初期トラブル件数

67つのケースが使用されました、16つケースには欠損値が含まれています

予測変数	Coef	標準誤差Coef	T	p値
定数	0.11925	0.05300	2.25	0.028
初期トラブル件数	0.03135	0.01948	1.61	0.112

S=0.368816 R二乗=3.8% R二乗 (調整済) 値=2.4%

分散分析

変動源	自由度	平方和	平均平方	F値	p値
回帰	1	0.3524	0.3524	2.59	0.112
残差誤差	65	8.8416	0.1360		
合計	66	9.1940			

見かけない観測値 “outliners”

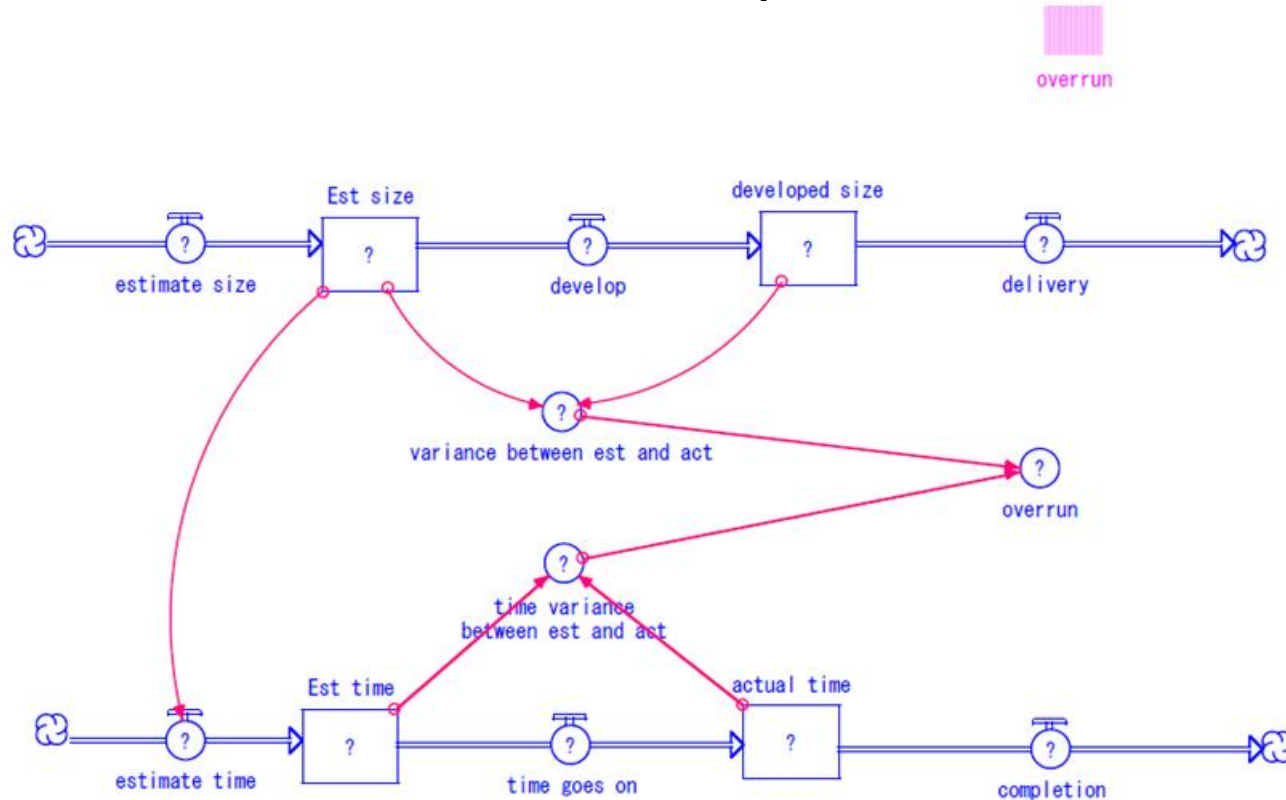
観測値	初期トラブル件数	目的変数	適合値	標準誤差適合値	残差	標準化残差
7	2.0	1.0000	0.1820	0.0464	0.8180	2.24R
8	2.0	1.0000	0.1820	0.0464	0.8180	2.24R
9	2.0	1.0000	0.1820	0.0464	0.8180	2.24R
15	13.0	0.0000	0.5269	0.2298	-0.5269	-1.83 X
18	5.0	1.0000	0.2760	0.0828	0.7240	2.01R
23	8.0	0.0000	0.3701	0.1356	-0.3701	-1.08 X
31	2.0	1.0000	0.1820	0.0464	0.8180	2.24R
33	1.0	1.0000	0.1506	0.0458	0.8494	2.32R
34	3.0	1.0000	0.2133	0.0544	0.7867	2.16R
38	2.0	1.0000	0.1820	0.0464	0.8180	2.24R

overrun (the size not completed)  
=0.119 + 0.034 \* num. of defects

Yarn

# Example of Step 4:

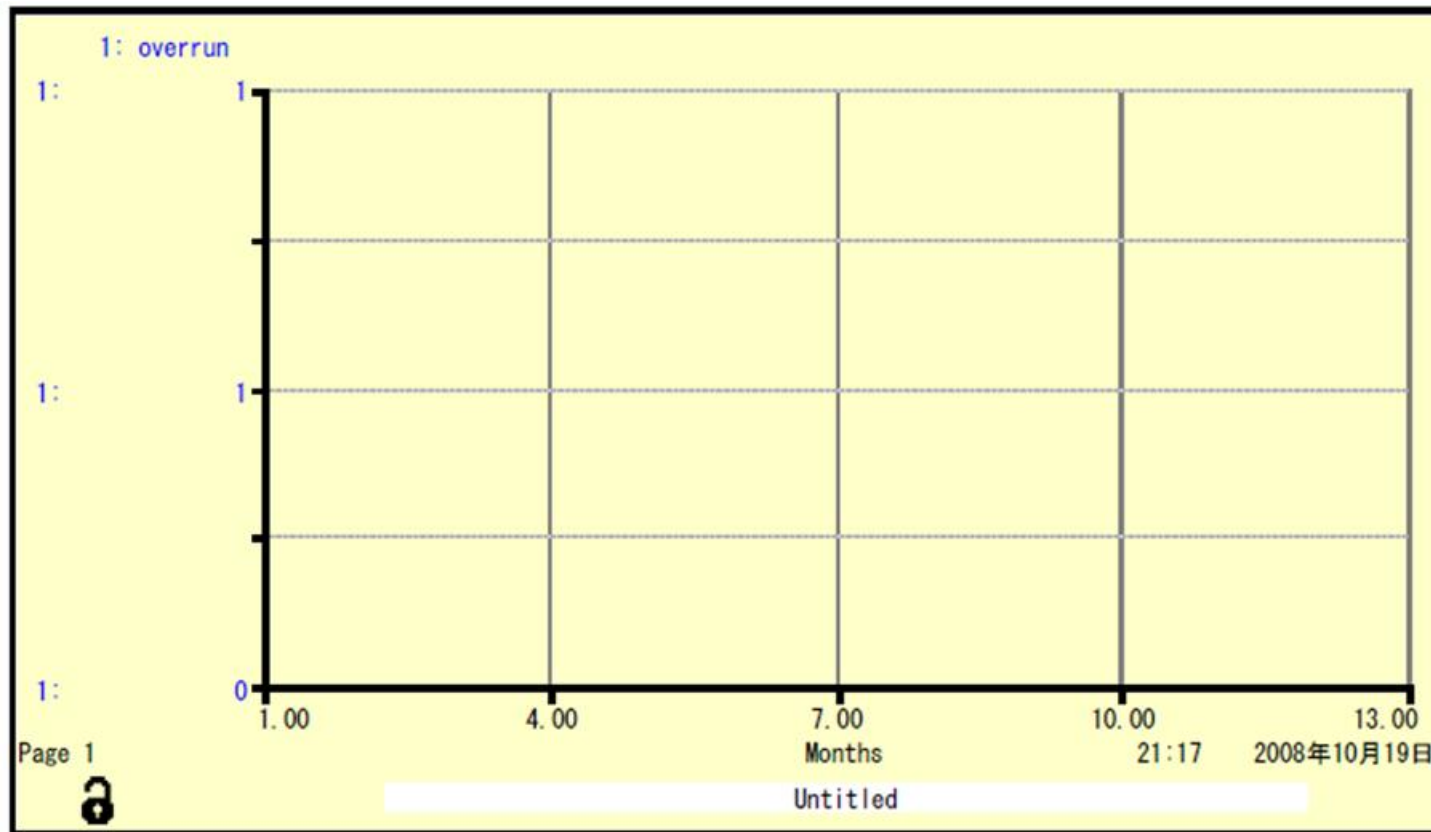
Input the identified variables and the formula to “ithink”, and make a model. An example can be as shown belc



Yann

## Example of Step 4-(2):

Run the model and see the output.





## Result 1

- There are few variables which are significantly related.
- Statistical relationship between time and cost is too weak.

This result shows that project members do their best effort (as a hidden effort) not to cause overrun enough to bend statistical analyzing.

In this case, no one can make accurate estimation.



## Result 2: Interpretation

But every project must make estimation and each project must go on based on the budget.



## Result 3: Recommendation

- Educate project engineers to improve their estimation accuracy.
- Educate project managers to improve their estimation review quality.

This is a plan to decrease “hidden effort” and increase the transparency of project performance.

# Benefit of this Stat-Modeling approach 1



As shown above, data, analysis, recommendations are:

- simple
- easy to understand intuitively
- focus on a part rather than entire organization

Simple and intuitive model works.

## Benefit of this Stat-Modeling approach 2



If you keep focus on the intuitive variables and keep making models straightforward,

- You don't waste too much time for getting involved in the model.
- Management, project members and SEPG members can communicate with the same language.

If you keep models simple, it works.

(Your due is to support projects and organization!)





# Traps for modelers

- You may want to make a big model, put every variable into the model and explain everything.
- The more variables used, the more the model is difficult.
- No one but you can understand and use the big complicated model.



# Traps for data guys 1

Data guy can be...

- Enthusiastic data collector
- Nervous to data quality
- If there is no significant relationships among variables, he will drop the data from his head.

Data doesn't lie!



## Traps for data guys 2

$P \text{ value} > 0.05 \rightarrow$  “reject the hypothesis”

This is just a statistical general rule.

In real environment, it can be a big message, specially when the result is against your guess.

There is a turning point: do you accept the hidden message and analyze closely, or just reject it?

Who judge it?  $\rightarrow$  We need skilled engineers.



## The essential element: skilled person

As we discussed above, this approach needs appropriate educated people. Expecting educations are as follows:

- (software) development experiences
- understanding dynamics of project (processes) (→PSP etc.)
- skills for statistical analysis (minitab etc.)
- skills for modeling (ithink etc.)
- on-the-job training to develop the engineers' intuition and hunch



## Prerequisites

It is necessary to invest a considerable amount of education in advance.

(You should convince the management.)

It is also necessary to campaign to management in order to build skills to understand this approach(as reviewers).



## Problems 1

On implementing this approach, problems occurred:

- If recommendations to the problems are not what management want to know, they could be rejected. (e.g. page 18, painkiller)
- Recommendations sounded “just too right”.



## Problem 2

e.g.: A patient complains of severe palpitation and a mild heart attack. Doctor uses freely statistical – Model approach and tell the patient:

*Overeating, inactivity, hypertension and hyperlipemia are suspected. Check your dietary habits, exercise more, and improve your physical constitution.*

→Right! But please stop this symptom now!

e.g. project manager complains uncontrollable overrun. They are in serious “death march”. Statistical-modeling guy uses this approach and tell the manager:

*Estimation seminar and design seminar can improve your project productivities.*

→Right!



## Conclusion

- Do not expect quality data in real environment. If it is, it's just a luck.
- Analysis and recommendations should be in simple and intuitive terms. Do not use "I know stat." terms.





## In Japan 2

- Keep it simple
- Focus on parts  
(not a whole view/not an entire organization)
- Educate management, project managers/members
- Educate statistical-modeling person(as an outliner)

Yarn

?



## Contact us

Shinobu Minamikawa (YARNE and Company)

[shinobu@yarneandcompany.com](mailto:shinobu@yarneandcompany.com)

+81-80-3094-0517

Yoshinobu Yamamura(YARNE and Company)

[yamamura@yarneandcompany.com](mailto:yamamura@yarneandcompany.com)



*End Of Presentation*

**YARNE and Company**