有数据才可以做统计分析，下面参考的案例，

芬兰某商业银行的软件维护数据统计分析

先看总体分析结论，再看从原始数据到分析结果的步骤。

总结报告

背景

银行管理层理解，如果不监控维护工作量，很可能会失控例如，有些软件系统因为在开发期间没有管理好（例如，引起很多缺陷），就会影响到后面维护工作量提升所以从 86 年开始，都一直统计软件维护相关数据。

- 从 1987 年到 1995 年，银行开发了 250 IBM 应用软件，其中部分是系统迁移（从原来的 BULL 主机迁移到 IBM 主机）
- 從中抽取了 67 有充分数据的应用软件，做统计分析。

希望解答以下问题:

1. 那些因素主要影响软件维护工作量（成本）
2. 有什么办法可以降低维护成本？
3. 利用预测模型，预估下一年度的维护成本

1. 影响维护工作量的主因

<< 表 5.4 里 >>

Dependent Variable	Significant Variables	Effect on Dependent Variable	Total Variance Explained
corrective effort	Size in function points	Positive	61.9%
	Batch processing integration	Positive	
	Internal business sector	Depends on type	
	Change management flexibility	Negative	
	Number of months maintained	Negative	

五个因素代表了 62% 的原因:

- 第一因素

是软件的规模大小占 27% 你可能疑问为什么没有看到缺陷密度？因为缺少密度与规模大小，息息相关相关，所以只需要考虑规模便可以

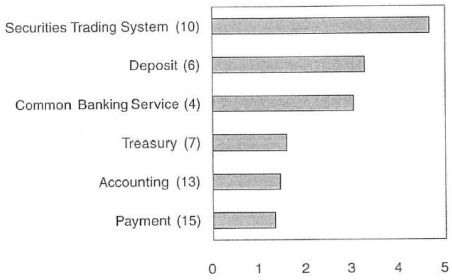
- 第二因素

批量处理的集成度

- 第三因素

是哪一类应用软件，看图 5.2

<< 图 5.2>>



股票买卖相关的维护工作量最高一般银行应用，例如捐款转钱最低原因：股票买卖的规则很复杂，嗯，大家大家都想你高竞争力竞争力，导致没有标准的模式

- 第四因素

是变更管理的灵活性如果应用很依赖其他系统更多的利益相关者，需要商户协商达成一致才可以变更维护成本就低了

- 第五因素

最后一个因素是应用软件投产后多少个月大概每年降低 17.4% 例如，如果第一年是 100 小时，第二年就会降到 82.6 小时，下一年会降到 68.2 小时

2. 什么办法可以降低维护成本

初步分析发现使用 Telon 语言的系统要维护工作量较少，只要 1.19 小时/FP，相对不使用 Telon 的软件系统需要是 2.47 小时/FP。

针对使用 Telon 语言的十一个系统统计分析：

- 发现语言的百分比与维护工作量正相关; 使用的比例越高，维护工作量越低
- 用功能点规模归一以后的维护工作量，也是强相关

<< 详见表 5.6>>

Dependent Variable	Significant Variable	Effect on Dependent Variable	Total Variance Explained
Corrective effort (11 observations)	Percentage of Telon code	Negative	72%
Corrective effort per function point (11 observations)	Percentage of Telon code	Negative	64%

3. 预估下一年度的维护成本

统计分析，除了帮我们洞察那些主要影响因素外，也可以帮我们做预测:

<< 表格 5.5>>

Predicted Using:	On 1994 Applications:	Num Obs	MMRE (%)	MedMRE (%)	Pred(.25) (%)
1993 actuals	Ongoing	17	76	42	35
	New	2	NA	NA	NA
1993 model	Ongoing	17	102	56	29
	New	2	48	48	50

但从表格 5.5 看到预测的准确度低预测 1994 年的工作量为例，用模型预测的准确度还不如直接使用 1993 的数字

在前面估算章节讨论过，很多因素影响软件开发的工作量难以准确预估

数据分析步骤

不用误以为只要有统计分析工具，把项目数据输入，便自动得出上面分析结果。以下对应每个步骤，用简单例子说明，如想自己动手，使用案例数据，完成整个数据分析过程，请参照 Maxwell 第 5 章 (详见 Ref)。

汇总数据，并明确每个变量的操作定义

- 本来数据分散在三个电子表单，每个表单有两百多个项目数据，但很多数据不齐全
- 最终汇总出 67 个项目数据，它们都在 1993 年有数据
- 明确每个变量的操作定义 (本来数据表都是用芬兰语言，也有很多银行专业术语)，确保大家的理解一致

使用描述性统计检查数据完整性和正确性

- 用数据总结功能，找出 28 个变量的均值 / 标准差 / 最大 / 最小:
- 发现某些变量最小值是零，不合理。例如：avetrans, disksp, cpu
- 有些变量数据不全，为空，请银行经理尽力填上。例如：r1-r10, dbms,tpms

<< Example 5.1 (p.209)>>

. summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
mid	67	34	19.48504	1	67
correff	67	515.0746	720.5377	22	3031
totfp	67	470.6119	514.5435	18	2328
pcobol	67	.3792537	.2943773	0	1
ptelon	67	.0704255	.1888824	0	.8691589
peasy	67	.0943003	.1332916	0	.5244565
pjcl	67	.4561443	.2620808	0	1
t	59	3.322034	.7529409	1	4
ageend	67	39.35821	20.5903	8	85
avetrans	67	14.10448	46.84149	0	345
disksp	67	1817	6019.997	0	39012
cpu	67	312.5672	535.797	0	2197
r1	56	3.392857	1.302844	1	5
r2	56	2.660714	1.391883	1	5
r3	56	2.375	1.272971	1	5
r4	56	2.375	.9256447	1	4
r5	56	2.160714	.968162	1	4
r6	56	2	.6875517	1	4
r7	56	4.053571	.9614316	1	5
r8	56	2.928571	1.616474	1	5
r9	56	3.660714	1.32496	1	5
r10	56	1.839286	.9298443	1	3
appdef	67	9.432836	22.3825	0	163
borg	67	3.567164	1.940196	1	6
morg	67	9.492537	6.013553	1	17
apptype	67	2.298507	.853068	1	4
dbms	59	1.033898	.1825208	1	2
tpms	66	2.772727	.7604746	1	5

Creation of New Variables 建立新变量

例：有些项目是在 1993 年一月份以后才开始维护，为了要与其他 1993 年一月份或者以前已开始的项目可以比较，创建新变量 (acorreff)，例如，1993 年只是维护了 10 个月，用了 30 小时，就要换成 36 小时 [=(30/10)*12]

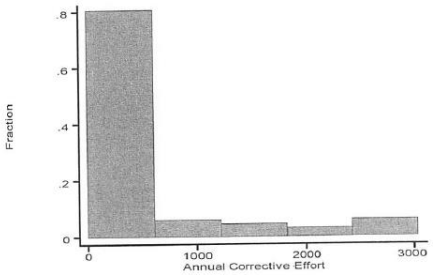
Data Modifications 数据修改

- Identify subsets of Categorical Variables 识别同一范畴的变量
- 模型选择 - Preliminary Analyses 初步分析:
- Q: 哪些因素影响到 “年度维护工作量”

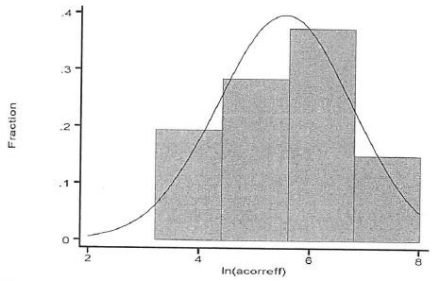
Histograms 直方图

- 画柱状图发现维护工作量 (acorreff)，规模大小 (totfp)，都是极度偏左，
- 使用自然对数使它变成较近似正态分布（因为变量不是正态分布，会影响回归分析不准确）

<< FIg 5.6 (p.217)>>



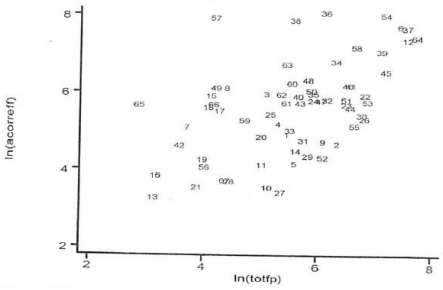
<< Fig 5.8 (p.218)>>



Scatter plot 散点图

- 从维护工作量 ln(acorreff) 与规模大小 ln(totfp) 的散点图，看到两者之间有线性关系

<< Fig 5.29 (p.228)>>



针对某个变量看各种分组的数量和维护工作量

例如，在内部业务部分组：

<< Example 5.14 (p.238)>>

```
. table morg, c(n acorreff mean acorreff)
```

Internal Business Unit	N(acorreff)	mean(acorreff)
Account	13	186
BUC	1	282
Common	4	283
CusInt	2	144
DecSup	1	1508
Deposit	6	1891
ITInfra	1	271
ITServ	1	39
ITSupp	1	112
IntlBank	1	795
LetCred	1	463
Loan	1	1122
Payment	15	328
Person	1	196
Resto	1	29
SecTrade	10	819
Treasury	7	255

- 存款业务 (Deposit) 的维护工作量最高
- 户口管理 (Account) 要维护工作量最低

相关 **(Correlation)** 分析

两变量的相关从-1 到 1，关于这系数的意义，可参考附件。因为如果两个变量是强相关，如果把这两个高度相关的变量放在一起做预测模型，就会会导致模型不稳定，所以要预先删除、处理。下表是相关系数大于 0.51 的汇总表：

<< Tab 5.9 (p.241)>>

Summary of Correlation Coefficients

Variables	Num Obs	Correlation
r2 and *r9*	56	0.53
r7 and *r9*	56	0.52
adefect and *totfp*	67	0.52
dsplev and *cpulev*	67	0.53
acorreff and *totfp*	67	0.52
r4 and *r5*	56	0.60
r2 and *r3*	56	0.74

例如，因为 r2 和 r3 强相关，它们也与其他变量相关，所以决定把 r2 和 r3 剔除

回归分析

- 利用以下连续变量做回归分析 [Y 是维护工作量 (lacorreff)]

(ltotfp, pcobol, pjcl, ageend, ladefect)

ln(acorreff) = 2.532 + 0.541*(ltotfp)

Adj. R-square = 0.27

- 最终得出以下 5 变量回归模型

ln(acorreff) = 3.768 + 0.555*ln(totfp) + r9_coef + submorg_coef + r3_coef
- 0.016*ageend
Adj. R-square = 0.619

- r10 因与 submorg 强相关，被剔除，剩下 5 变量
- r9, submorg, r3 因为是分组数据，系数会依据类型，按下表选对应系数：

<< Table 5.14 (p.257)>>

Categorical Variable Multipliers

r3 level	*r3_mult*	*r9* level	*r9_mult*	*morg*	*morg_mult*
1 (51)	**1**	1 (1)	0.1367*	*Account* (1)	1.2888
2 (2)	0.5440	2 (2)	0.4306	*Common* (3)	5.0544*
3 (3)	0.4338*	3 (3)	0.6446	*Deposit* (6)	9.8649*
4 (4)	0.2384*	4 (4)	0.6675	*Payment* (63)	**1**
5 (5)	0.2044*	5 (5)	**1**	*SecTrade* (16)	1.6233

使用回归模型预测

分析使用对维护工作量的作用

回顾与总结

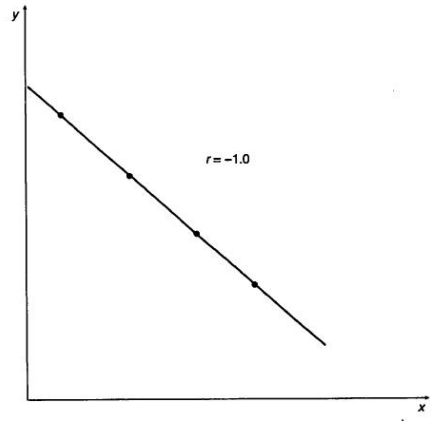
附件

项目变量列表

X 与 **Y** 相关性图

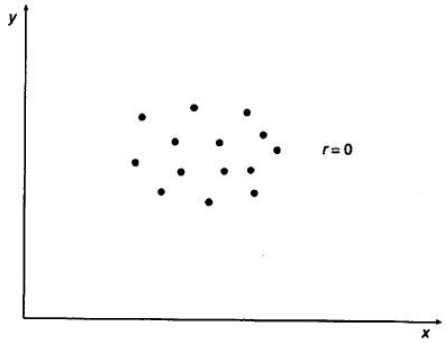
- 完美负相关 (-1):

<<Fig 6.7 ,pg291>>



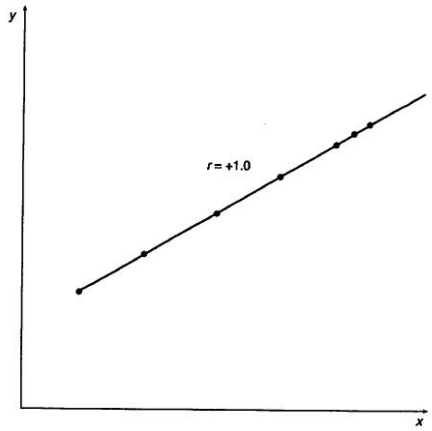
- 零相关 (0):

<<Fig 6.9 ,pg292>>



- 完美正相关 (+1):

<<Fig 6.8 ,pg292>>



Prunning by Wrapper

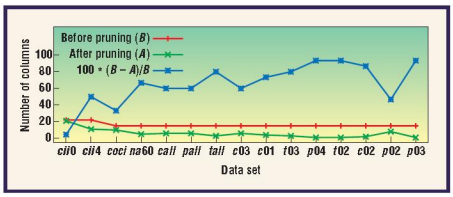
当我们收集到一定数量的项目数据，就可以尝试做组织级的数据分析。很多分析数据时，没有考虑项目之间的差异，假定项目的特性都类似，然后直接就用统计分析方法，求模型的方程式参数。
例如下图，6 个项目是 5 个迭代的系统测试密度数据，你觉得把 6 个项目的缺陷密度放在一起分析合适吗？很明显两个项目缺陷比较低，有些很散，所以如果没有考虑项目之间的差异，直接总体分析就会变成很宽，比如对 A 的缺陷比较低，没有什么参考意义。

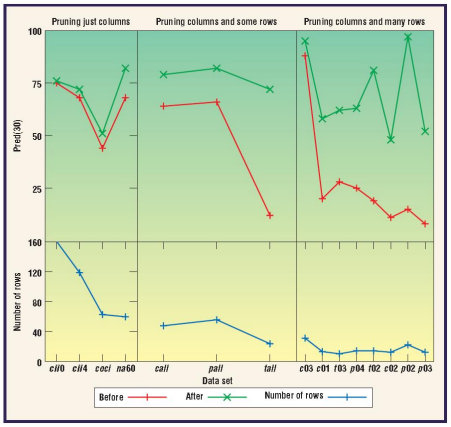
<<handDrawPicture.jpg>> to be provided

除了我们要细分项目就是要考虑保留那些项目参数（影响因素），很多时候我们看到一些公司级的数据分析，都是一大堆表，可能要有二三十个变量来分析，从模型出来的变量越多，其实不是好事：

1. 可能做了过度的调整，导致那个预测模型没有预测的意义，只适合用在这堆项目数据上；
2. 变量多也会导致要花很多精力去收集数据，不划算。
3. 使用模型的人也很难理解这个模型的意义，不知道如何去使用。

我们在前面 EST2 里的乐高模型数据案例，简单用了两个变量来做个预测模型，一是积木的数量，即它的规模大小；二是团队人数；可以比较好地估计工作量，这是比较理想的模型。
所以当开始时，收集到很多变量，我们就需要利用数据分析删除一些意义不大的变量。
怎么挑选呢？
比如一个方式是用 wrapper 机械性地去挑选，如果增加一个变量准确度更好就增加，如果不是就不增加，直到挑选出最佳的搭配。





References

1. Maxwell, Katrina: *Applied Statistics for Software Managers* Prentice-Hall 2002.

2. Chen, Zhihao: ”Finding the Right Data for Software Cost Modeling” IEEE Software Nov/Dec 2005

---===<<< END >>>===---