



R在市场研究中的应用

主讲：祝迎春

时间：20091206


联系方式：ereree@126.com




自我介绍



市场研究中的数据分析



R如何在市场研究中发挥作用



如何在市场研究中普及推广**R**



实际模块演示



自我介绍

- ◆ 2000年开始接触统计软件，9年以上的统计软件使用经验。
- ◆ 2002月开始进入市场研究行业，熟悉并开发各类调研和数据挖掘模型。
- ◆ 爱好研究各种统计软件，涉及：定性（质性）资料处理、市场研究、数据挖掘、抽样、缺失值处理、社会网络分析、可视化软件。
- ◆ 参与《**SPSS**统计分析高级教程》(张文彤主编)的编写工作，该书于2004年10月出版，高等教育出版社。
- ◆ 积极参与推动各种分析软件在社会科学领域的应用。



市场研究中的数据分析

市场研究数据分析的特点



市场研究没有自己独立的统计技术和方法论，但是有具体的商业业务（解决方案）、行业特点。

数据量大

- 需要对大量的数据进行分析

频率多

- 依赖重复的日常性分析

控制性

- 可视化的过程与结果

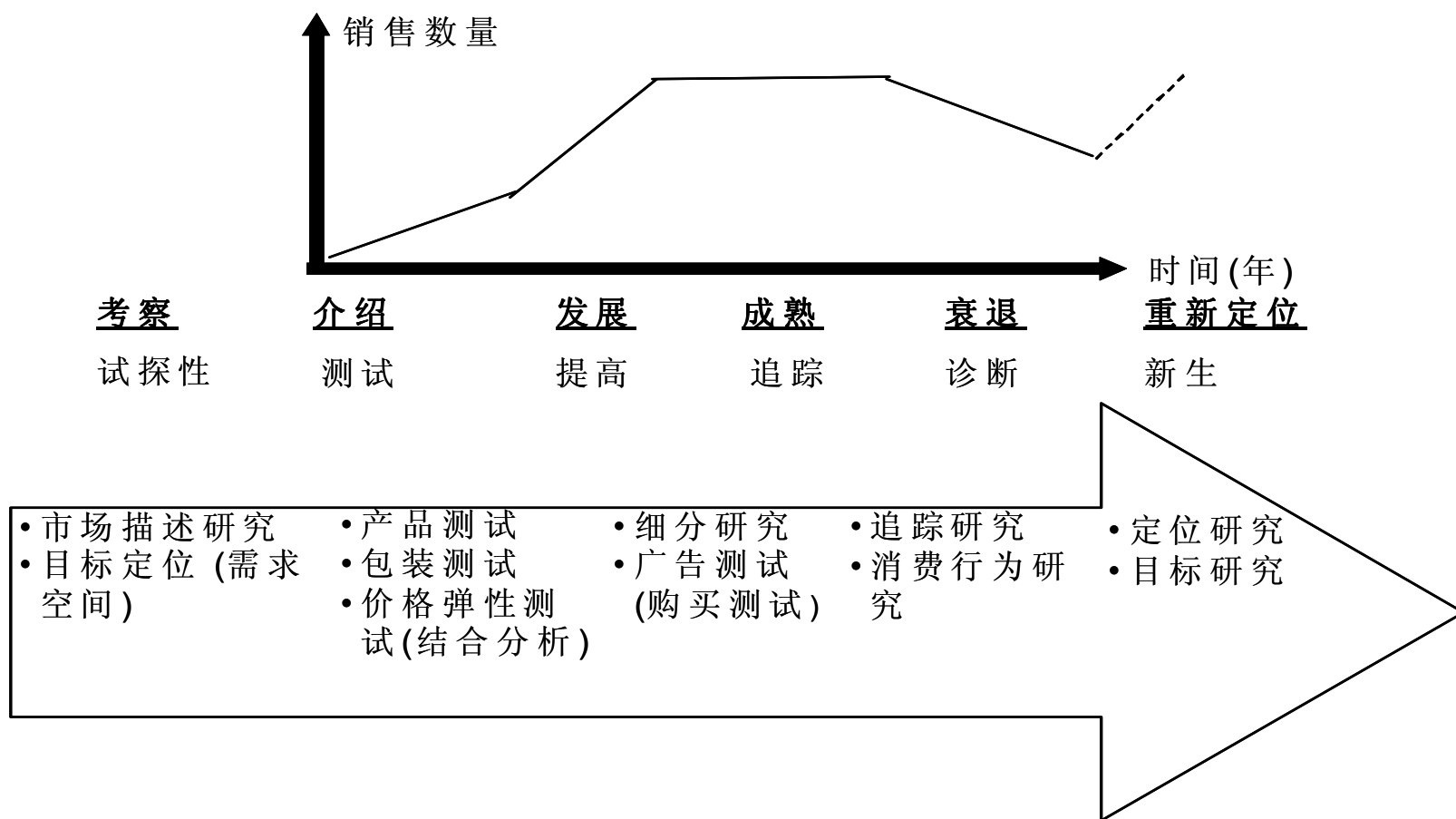
行业特点

- 数据格式、分析算法独特



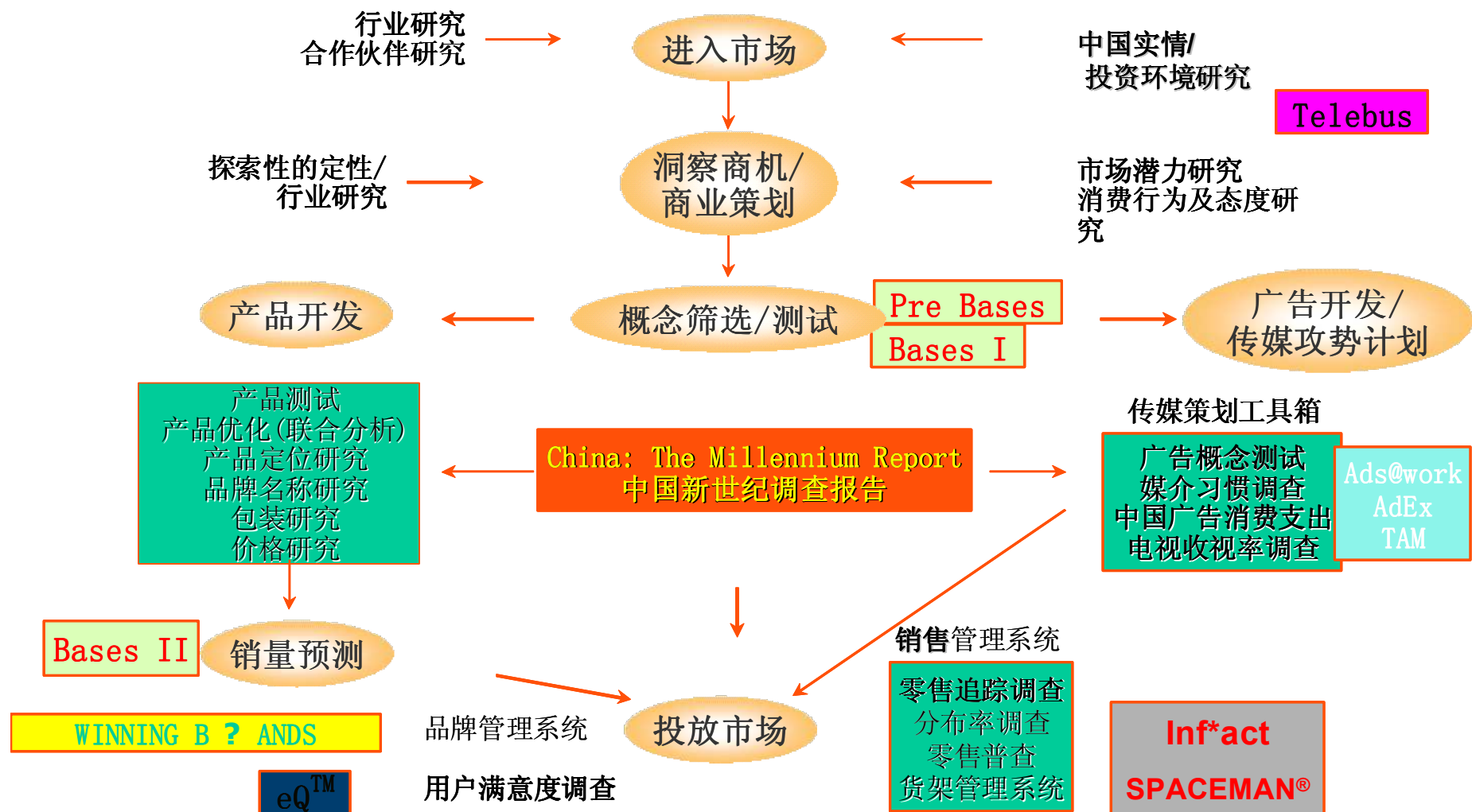
基于产品生命周期的调研思路

市场研究的一个比喻

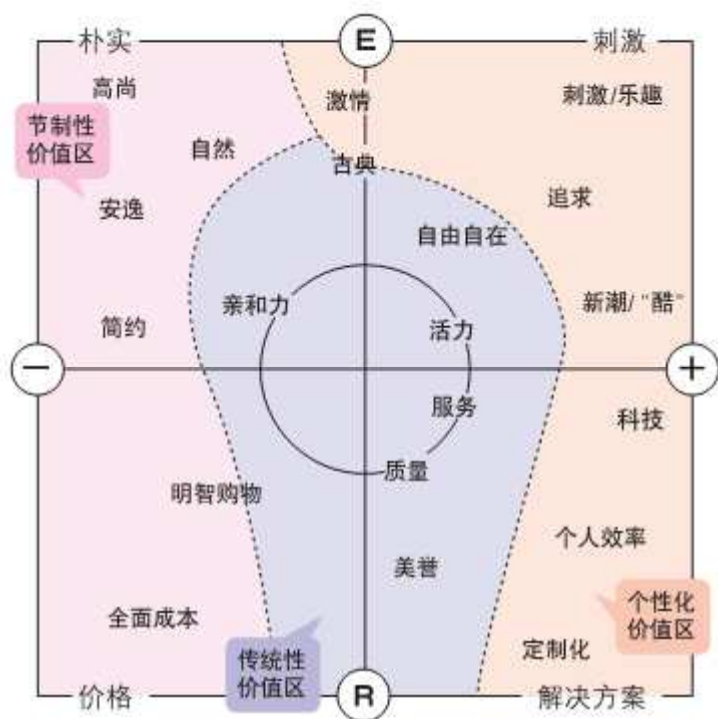




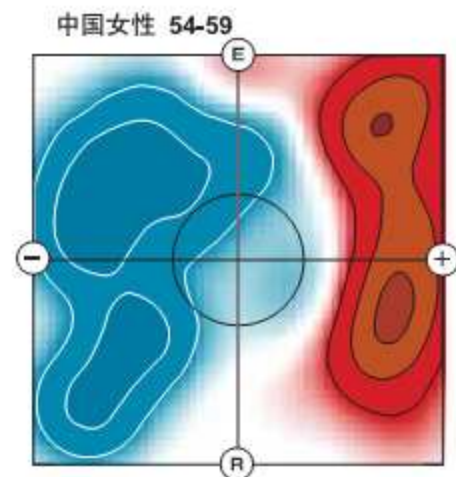
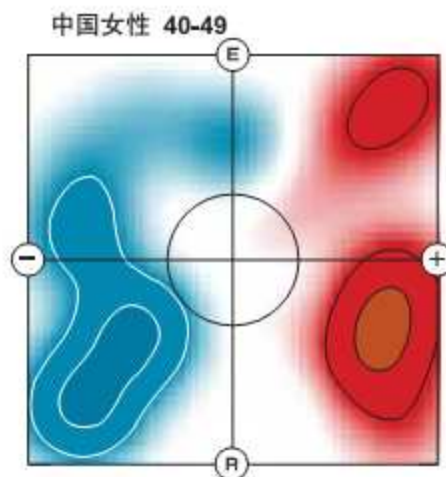
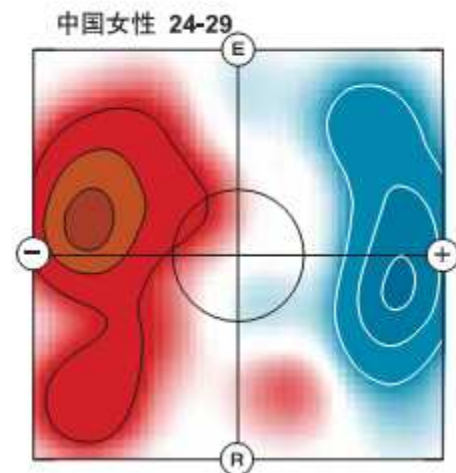
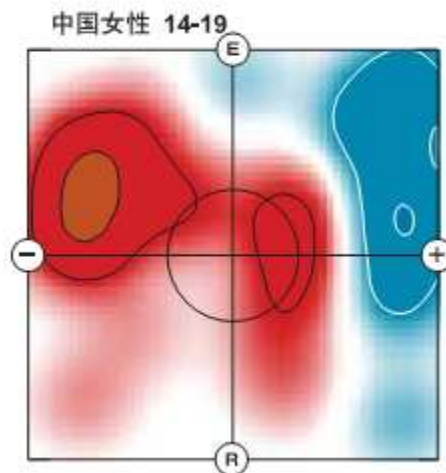
(举例) 尼尔森的市场研究及信息分析服务



（举例）罗兰贝格消费者研究模型



||



原理：数据降维技术和作图模型
现有方法论其实有很大技术改进的空间



市场研究中的研究工具

市场研究中的常用兵器库



Microsoft Excel



PASW Statistics

PASW Modeler

PASW Amos



Xlstat



SAS



STATISTICA



Xcelsius



Sawtooth



R/S-Plus



R如何进入市场研究中？

R如何吸引我



算法先进且更新快



小、免费且开放



作图专业类型多



算法先进且更新快

专门的包，对症下药。R可以引发一次调研技术的变革

Direct Marketing Analytics with R (直销研究模块): <https://r-forge.r-project.org/projects/dma/>



smart visualization 的介绍

假期上了个可视化的暑期班，听了很多新名词，其中有一个是smart visualization，IBM研究中心在弄这个。Smart vis和AI很像，想把可视化智能化，他们的野心很大，要求终端用户输入数据和目的，然后系统就会自己组合出若干套可视化方案供选择。这个要涉及很多东西，比如用户习惯、数据库匹配、自然语义分析处理、图形库的简历和分类等等。

个人感觉smart visualization很难做，但是很有做头，很有前途，现在越来越多的人进了这一行。感兴趣的不妨关注一下。

<http://cos.name/bbs/read.php?tid=16428&page=e#&a>





免费且开放

2个山寨版的免费SPSS

1

PSPP

基于JAVA的

<http://www.gnu.org/software/pspp/>

2

Rcmdr

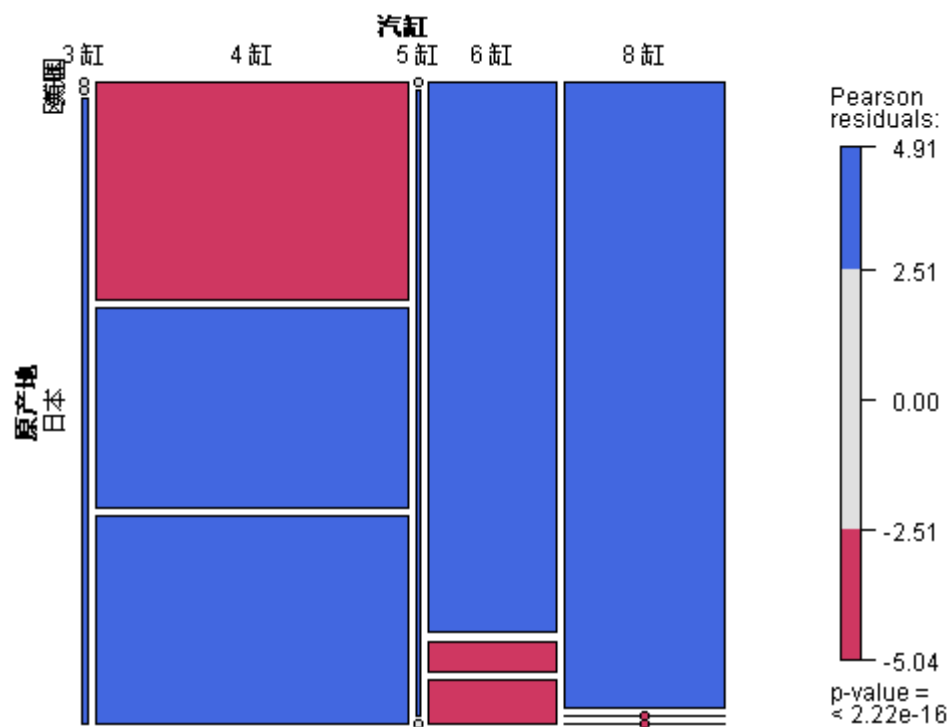
基于R的

<http://cran.r-project.org/web/packages/Rcmdr/index.html>



作图好看且专业

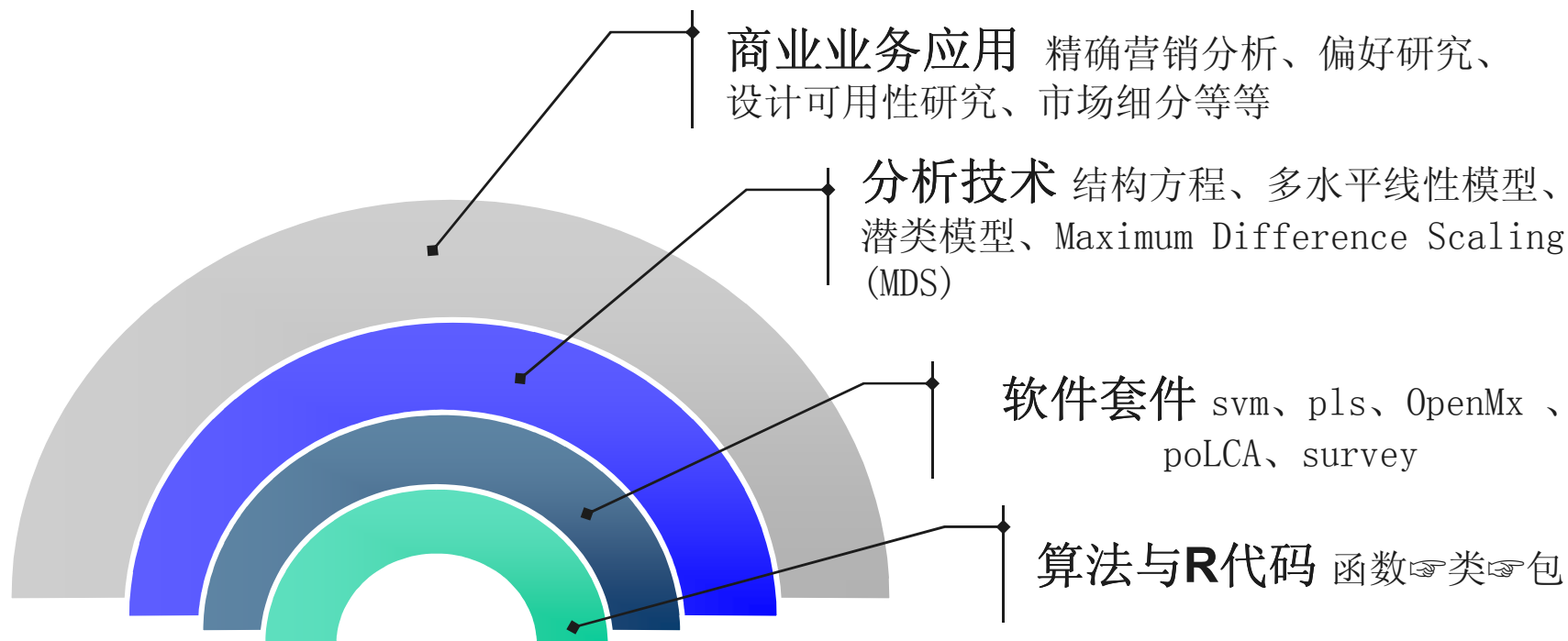
马赛克图





如何普及推广R

R在市场研究中的架构



变量选择 → 样本量 → 抽样方法 → 分析技术 → 输出形式



影响R在商业普及的因素

- 不支持多核 ➡ 应该是不能简单直接的支持
- 内存读取 ➡ 数据量和支持格式上的问题
- 输入输出 ➡ 查询和快速数据存取管理
- 流程控制 ➡ 即见即所得的分析与管理
- R是在国内教育还在起步阶段，但发展迅猛
- 不能自动化扩展商业统计软件的统计功能
- R是一种编程语言，比较高的可编程明确要求，在学习R技术上的投资，很难保证投资回报
- 认知与试用率低
- 其他替代软件众多



一个很好的样板：REvolution

他们为对开源数据分析软件R提供商业服务与支持。REvolution最近获得了North Bridge Venture Partners和Intel Capital总计900万美元的风险投资，并且任命Norman Nie为该公司的CEO。Norman Nie 是斯坦福大学的教授，更重要的是，他是SPSS公司的创立者和前CEO。

本文来自CSDN博客，转载请标明出处：<http://blog.csdn.net/C6H5NO2/archive/2009/10/26/4728135.aspx>

REvolution R Enterprise in Customer Analytics

Data Management

- Data extract and transformation
- Data quality analysis and data cleaning
- Data exploration

Marketing Planning

- Experimental design
- Sample size analysis
- Market research and survey planning

Analysis and Reporting

- Campaign analysis and reporting
- Graphical dashboards
- Customer trends and segment migration
- Churn analysis
- Mapping and spatial analysis
- Significance testing

Customer Acquisition

- Clone (look-alike) models
- Response models
- Prospect value models
- Channel and product affinity models
- Prospect segmentation

Customer Growth

- Cross-sell models
- Next-likely product models
- Market basket analysis
- Lifetime value models
- Campaign response models
- Customer segmentation

Customer Retention

- Churn (at-risk) models

Customer Winback

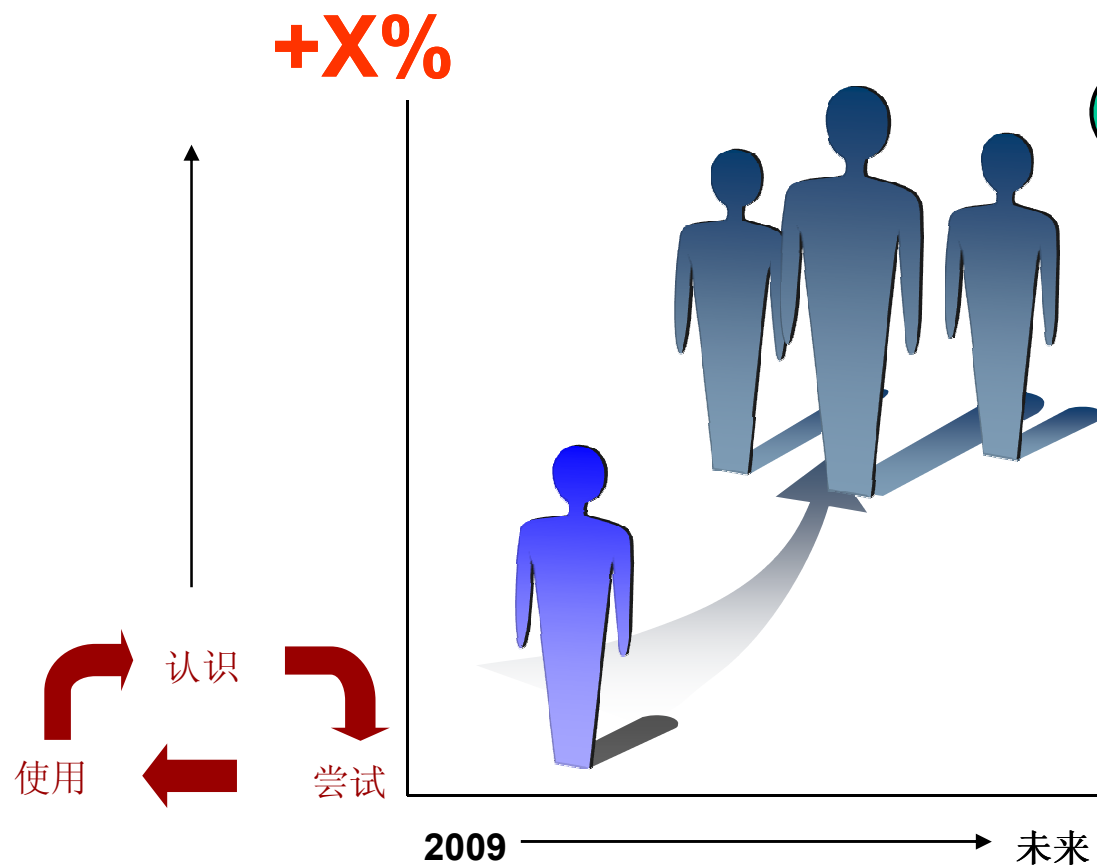
- Winback response models
- Winback offer models

Marketing Optimization

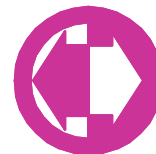
- Offer Optimization
- Channel Optimization
- Budget Optimization
- Campaign Optimization
- Price Optimization
- Media Optimization
- Digital Media Optimization



提高R 在调研行业的普及率



R的普及与发展预测



高端玩家为完成特定专业统计技术而使用的分析工具.



由于早期使用者的传播导致产生的工具免费而服务收费的商业模式.



用户形成规模，得到公认，分工明细，制度创新模式产生，形成产业链.

R可能给调研或者咨询行业带来革命性的影响：免费的商业模式.



R的插件化



第一，可以让你集中精力放在分析上，避免了繁杂的编程



第二，和已掌握的分析工具结合起来，减少转移成本

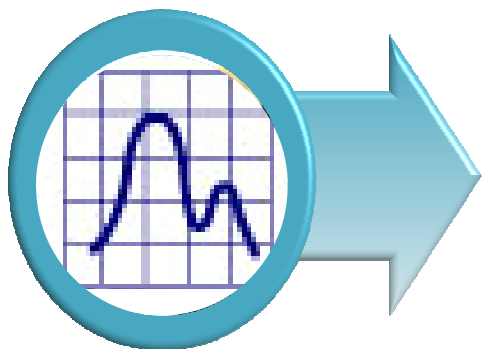


第三，优势互补，取长补短



第四，减少工具与工具之间的切换

2个支持R接口的商业统计分析软件



使用 **R (D)COM Server** 进行语言之间的衔接，不太好掌握，需要计算机专业知识。



使用**PASW Statistics-R Essentials** 插件作为接口，自动安装。



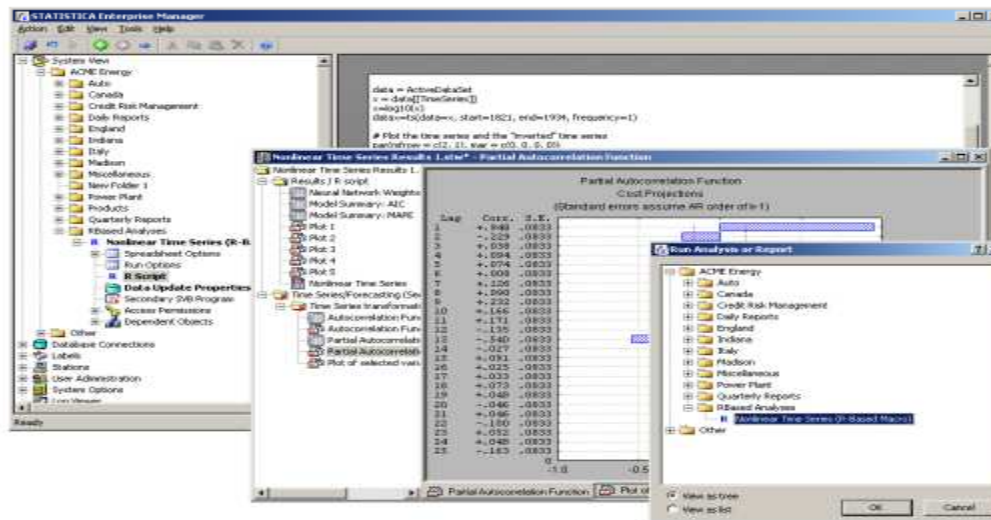
也支持R:

<http://sunsite.univie.ac.at/rcom/>
《R Through Excel》



- <http://www.statsoft.com/solutions/r-language-platform/>
- <http://cran.r-project.org/contrib/extra/dcom>

- Generate automated **reports** from reusable R-based analysis configurations, which deliver the power of R to users not familiar with R.



With **WebSTATISTICA Server**, you can:

安装PASW Statistics-R Integration Package

> SPSS Developer Central

SPSS Programmability Plugins

The SPSS Integration Plug-In is a complete, freeware example plug-in for integrating the open source Python programming language, the R statistical programming language or VB.NET with the SPSS Programmability Extension.

[Download the tools you need](#) for using R within PASW Statistics 18.

[Download the installation guide for the plugins here.](#)

[Download Python 2.4 \(SPSS 14 and 15\) or Python 2.5 \(SPSS 16 and SPSS Statistics 17\) from the Python website, if not on your CD.](#)

SPSS Product

PASW Statistics 18.0
PASW Statistics 17.0.2
PASW Statistics 17.0.1
PASW Statistics 17.0.1
PASW Statistics 17.0
PASW Statistics 17.0
SPSS 16.0.1
SPSS 15.0.1
SPSS 14.0.2

Operating System

Windows 32
Select OS
Windows 32
Windows 64 Server
Windows 32
Windows 64 Server
Windows
Windows
Windows

Plugin Download

Select Plugin
Select Plugin
Python
VB.NET
R
Python
Python | VB.NET | R
Python | VB.NET
Python | VB.NET
Python | R

> Downloads

[Download](#) code and tools from SPSS and the SPSS user community.



SPSS Developer Forums

Share code, solve problems, and be more productive - access SPSS Developer Forums now.

- » [DevCentral Forums](#)
- » [SPSS Programmability](#)
- » [Advanced Visualization](#)

Great Links

SPSS Blogs

- » [SPSS Developer Blog](#)
- » [Developer Video Blog](#)

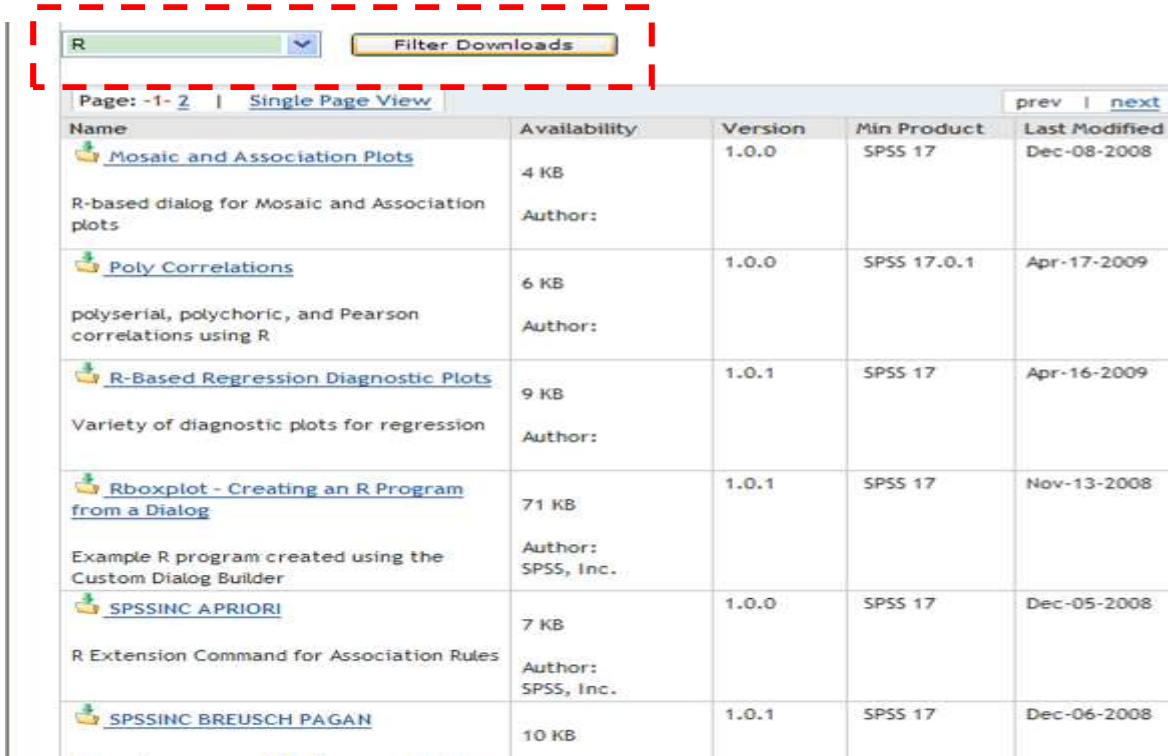
Articles







- » [Converting an SPSS Syntax Job to a Python Job](#)
- » [Directions Conference Presentation on Programmability](#)
- » [Installing the SPSS](#)



PASW Statistics-R小零件下载

- <http://www.spss.com/devcentral/index.cfm?pg=downloads>
- 需要注册下，即可免费下载
- 下载后直接双击安装即可



| Name | Availability | Version | Min Product | Last Modified |
|---|-----------------------------|---------|-------------|---------------|
|  Mosaic and Association Plots R-based dialog for Mosaic and Association plots | 4 KB Author: | 1.0.0 | SPSS 17 | Dec-08-2008 |
|  Poly Correlations polyserial, polychoric, and Pearson correlations using R | 6 KB Author: | 1.0.0 | SPSS 17.0.1 | Apr-17-2009 |
|  R-Based Regression Diagnostic Plots Variety of diagnostic plots for regression | 9 KB Author: | 1.0.1 | SPSS 17 | Apr-16-2009 |
|  Rboxplot - Creating an R Program from a Dialog Example R program created using the Custom Dialog Builder | 71 KB Author: SPSS, Inc. | 1.0.1 | SPSS 17 | Nov-13-2008 |
|  SPSSINC APRIORI R Extension Command for Association Rules | 7 KB Author: SPSS, Inc. | 1.0.0 | SPSS 17 | Dec-05-2008 |
|  SPSSINC BREUSCH PAGAN | 10 KB | 1.0.1 | SPSS 17 | Dec-06-2008 |



PASW Statistics-R编程

最简单的一条语句（在SPSS的sps文件中输入）：

```
BEGIN PROGRAM R.
```

```
  casedata <-
```

```
  spssdictionary.GetDictionaryFromSPSS()
```

```
  print(casedata)
```

```
END PROGRAM.
```

/*R程序开始

/*读取当前spss文件，并输出查看

/*R程序结束

详细请看：《Programming and Data Management for PASW[®] Statistics 18:
A Guide for PASW Statistics and SAS[®] User》

免费下载：www.spss.com/sites/dm-book



R插件的模块化

Cars.sav [数据集2] - PASW Statistics 数据编辑器

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 自定义(C) 实用程序(U) 窗口(W) 帮助

1: mpg 18

| | mpg | 引擎 | 马力 | 重量 | 加速 | 年份 | 原产地 | 汽缸 |
|----|-----|-----|-----|------|----|----|-----|----|
| 1 | 18 | 307 | 130 | 3504 | 12 | 70 | 1 | |
| 2 | 15 | 350 | 165 | 3693 | 12 | 70 | 1 | |
| 3 | 18 | 318 | 150 | 3436 | | | | |
| 4 | 16 | 304 | 150 | 3433 | | | | |
| 5 | 17 | 302 | 140 | 3449 | | | | |
| 6 | 15 | 429 | 198 | 4341 | | | | |
| 7 | 14 | 454 | 220 | 4354 | | | | |
| 8 | 14 | 440 | 215 | 4312 | | | | |
| 9 | 14 | 455 | 225 | 4425 | | | | |
| 10 | 15 | 390 | 190 | 3850 | | | | |
| 11 | | 133 | 115 | 3090 | | | | |
| 12 | | 350 | 165 | 4142 | | | | |
| 13 | | 351 | 153 | 4034 | | | | |
| 14 | | 383 | 175 | 4166 | | | | |
| 15 | | 360 | 175 | 3850 | | | | |
| 16 | 15 | 383 | 170 | 3563 | | | | |
| 17 | 14 | 340 | 160 | 3609 | | | | |
| 18 | | 302 | 140 | 3353 | 8 | 70 | 1 | |
| 19 | 15 | 400 | 150 | 3761 | 10 | 70 | 1 | |
| 20 | 14 | 455 | 225 | 3086 | 10 | 70 | 1 | |
| 21 | 24 | 113 | 95 | 2372 | 15 | 70 | 3 | |
| 22 | 22 | 198 | 95 | 2833 | 16 | 70 | 1 | |

已安装的扩展束

下列扩展束已安装:

| 名称 | 摘要 |
|-----------------------|------------------------------|
| SPSSINC TOBIT REGR | Estimate a regression |
| Rboxplot | Generate boxplots using |
| 偏最小二乘法 | 偏最小二乘法... |
| SPSSINC QUANTREG | Estimate one or more |
| SPSSINC ROBUST REGR | Estimate a linear regression |
| SPSSINC HETCOR | Calculate correlations |
| SPSSINC RASCH | Estimate a Rasch model |
| SPSSINC BREUSCH PAGAN | Estimate a linear model |

不显示未作为扩展束的组成部分进行安装的定制

关闭 帮助

扩展束详细信息

详细信息

摘要(S): 偏最小二乘法

描述(D): 由祝迎春在R 2.8.1和SPSS18的xp2环境下编写而成。

作者(A): 祝迎春 发布日期: 20091120

版本(V): 1.0.0 最低 PASW Statistics 版本: 18

链接(L): ereree@126.com

查看目录

组件

定制对话框:

| 名称 | 菜单位置 |
|--------|--------------------|
| 偏最小二乘法 | 分析(A)>回归(R)>偏最小二乘法 |

扩展命令:

相关性

需要 Python 插件: ☐ 否 需要 R 插件: ☒ 是

R 程序包: PLS

Python 模块:

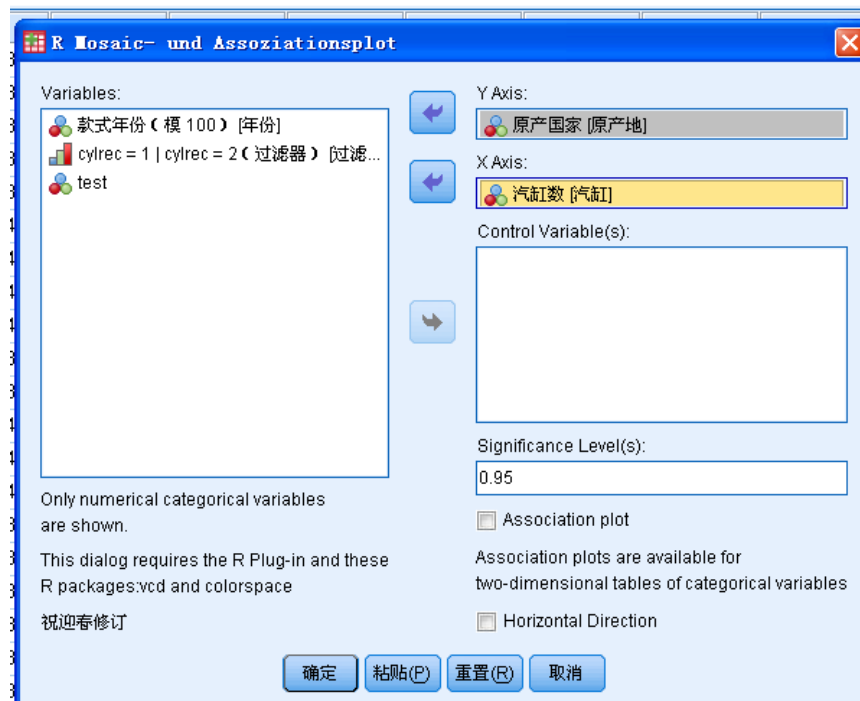
关闭 帮助



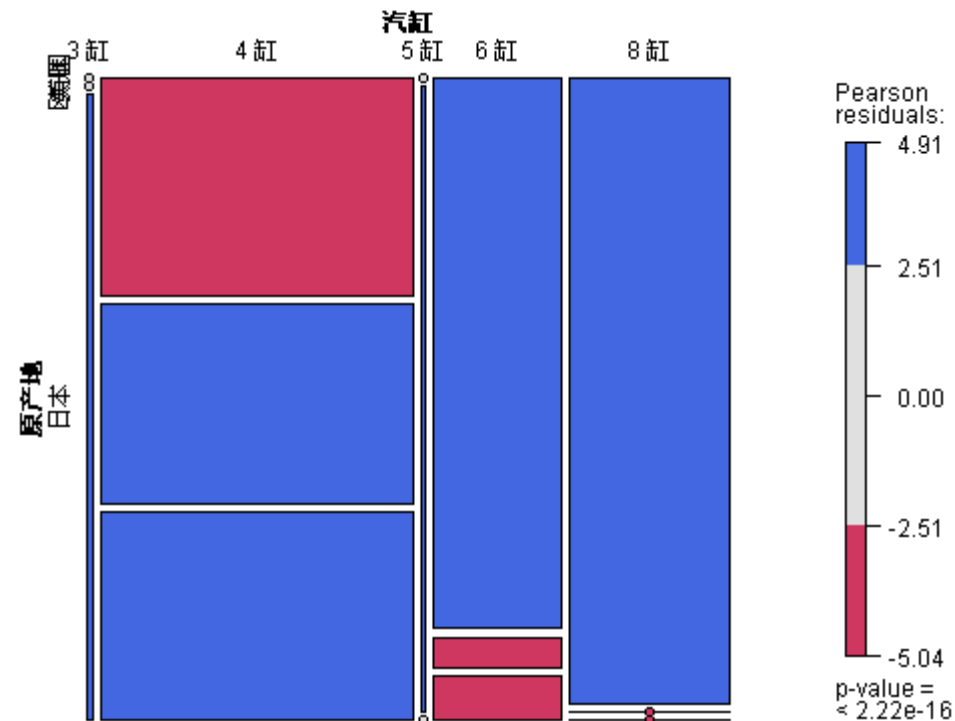
实际模块演示



(1) 作图



马赛克图





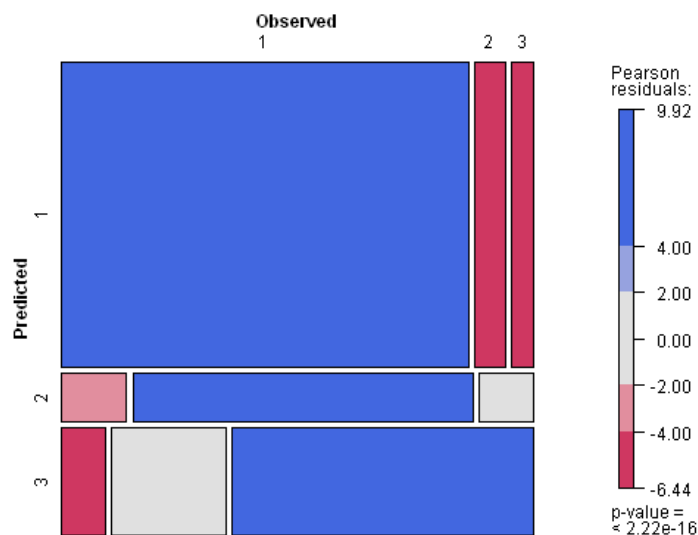
(2) 市场细分：支持向量机

Best Support Vector Machine

| | |
|----------------------|-----------------------|
| Type SVM | C-Classification |
| Kernel | Radial Basis Function |
| Best error rate | 0.220641025641026 |
| N?Classes | 3 |
| Cost | 64 |
| Gama | 4 |
| Total Suport Vectors | 193 |
| Sup.Vect. class 1 | 77 |
| Sup.Vect. class 0 | 58 |

Confusion table (n's and %s)

| Predicted | Observed | | | | | |
|-----------|----------|--------|--------|------|------|------|
| | 1 | 2 | 3 | 1(%) | 2(%) | 3(%) |
| 1 | 234.000 | 17.000 | 13.000 | .586 | .043 | .033 |
| 2 | 6.000 | 31.000 | 5.000 | .015 | .078 | .013 |
| 3 | 9.000 | 23.000 | 61.000 | .023 | .058 | .153 |



个人经验：一般条件下在没有大量缺失值的条件下，SVM的效果在是超过PASW Statistics中其他方法的。



(3) 满意度权重：偏最小二乘法1

偏最小二乘法

变量:

- 总体满意度
- 增值业务满意度
- 话费信息满意度
- 营业厅满意度
- 客服热线满意度
- 俱乐部会员
- 俱乐部满意度
- 客户经理满意度
- 区内话费

因变量:

自变量:

潜变量个数: 2

拟合系数指标

☐ R2

☒ RMSEP

☐ MSEP

验证

☒ L₀₀

☐ CV

☐ none

此插件需要R PLS(祝迎春)

得分数据保存在C盘下

确定 粘贴(P) 重置(R) 取消 帮助

偏最小二乘法作图

变量:

- 俱乐部会员
- 序号

因变量:

- 区内话费
- 漫游费
- 客户贡献
- 贡献价值

自变量:

- 总体满意度
- 增值业务满意度
- 话费信息满意度
- 营业厅满意度
- 客服热线满意度
- 俱乐部满意度

验证

☒ none

☐ CV

☐ L₀₀

此对话框需要R PLS(祝迎春)

潜变量个数: 2

确定 粘贴(P) 重置(R) 取消 帮助



(3) 满意度权重：偏最小二乘法2

回归系数标签

| | 总体满意度.1 comps | 总体满意度.2 comps |
|-------------|---------------|---------------|
| (Intercept) | 30.485 | 26.762 |
| 增值业务满意度 | .085 | .015 |
| 话费信息满意度 | .034 | -.143 |
| 营业厅满意度 | .179 | .285 |
| 客服热线满意度 | .158 | .202 |
| 俱乐部满意度 | .002 | -.046 |
| 客户经理满意度 | .115 | .275 |

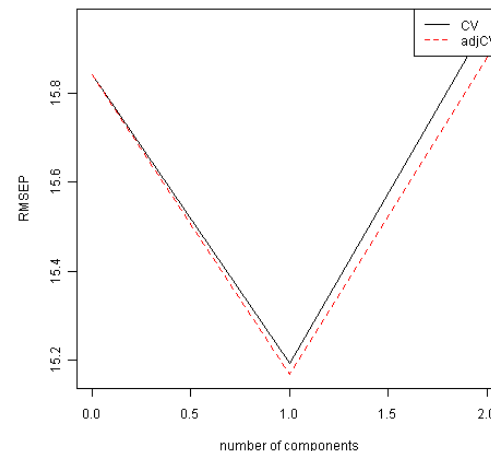
loadings系数标签

| | Comp 1 | Comp 2 |
|---------|--------|--------|
| 增值业务满意度 | .491 | -.755 |
| 话费信息满意度 | .466 | -.637 |
| 营业厅满意度 | .581 | .283 |
| 客服热线满意度 | .598 | -.002 |
| 俱乐部满意度 | .093 | -.157 |
| 客户经理满意度 | .209 | .314 |

Yloadings系数标签

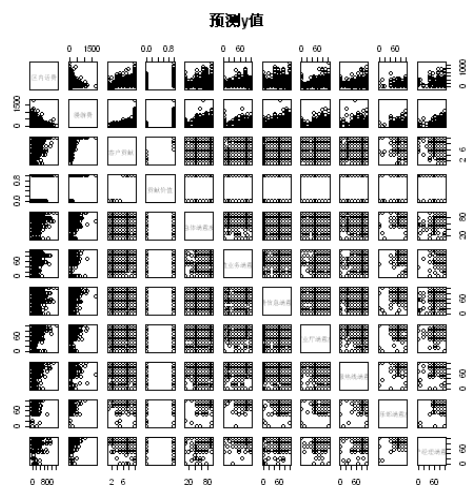
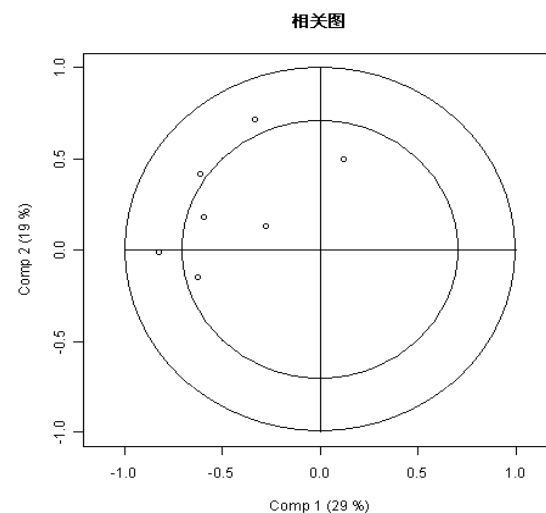
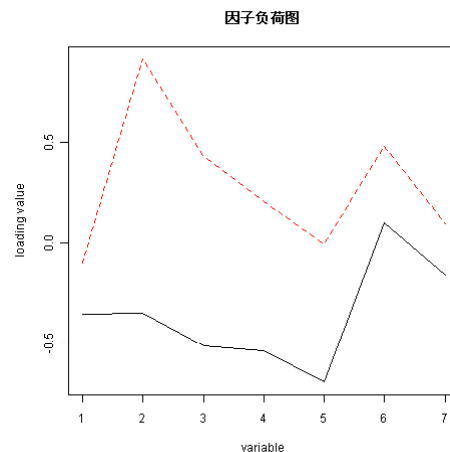
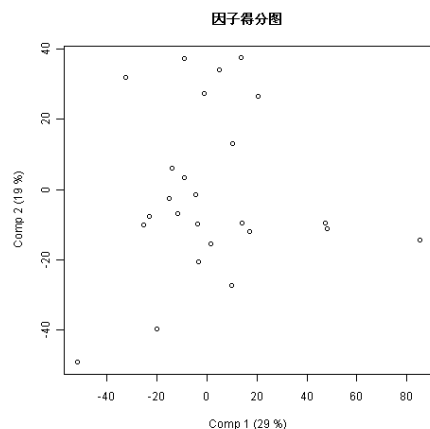
| | X |
|--------|------|
| Comp 1 | .280 |
| Comp 2 | .254 |

总体满意度

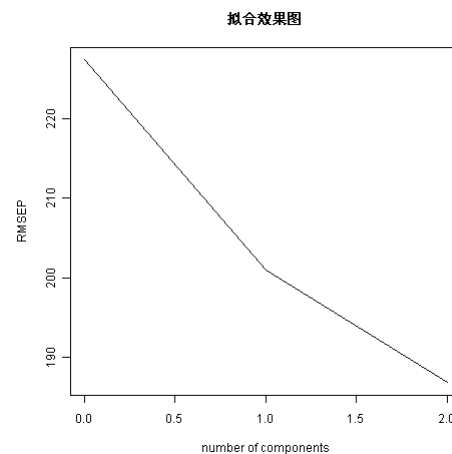
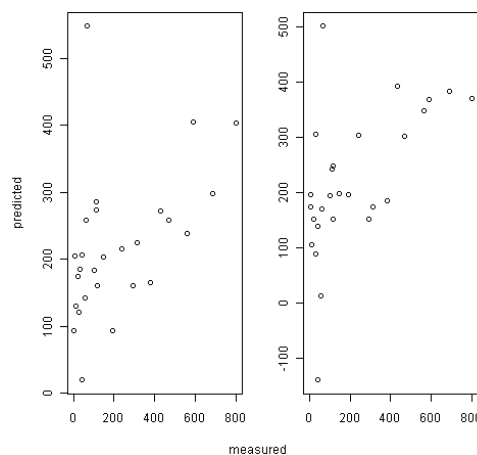




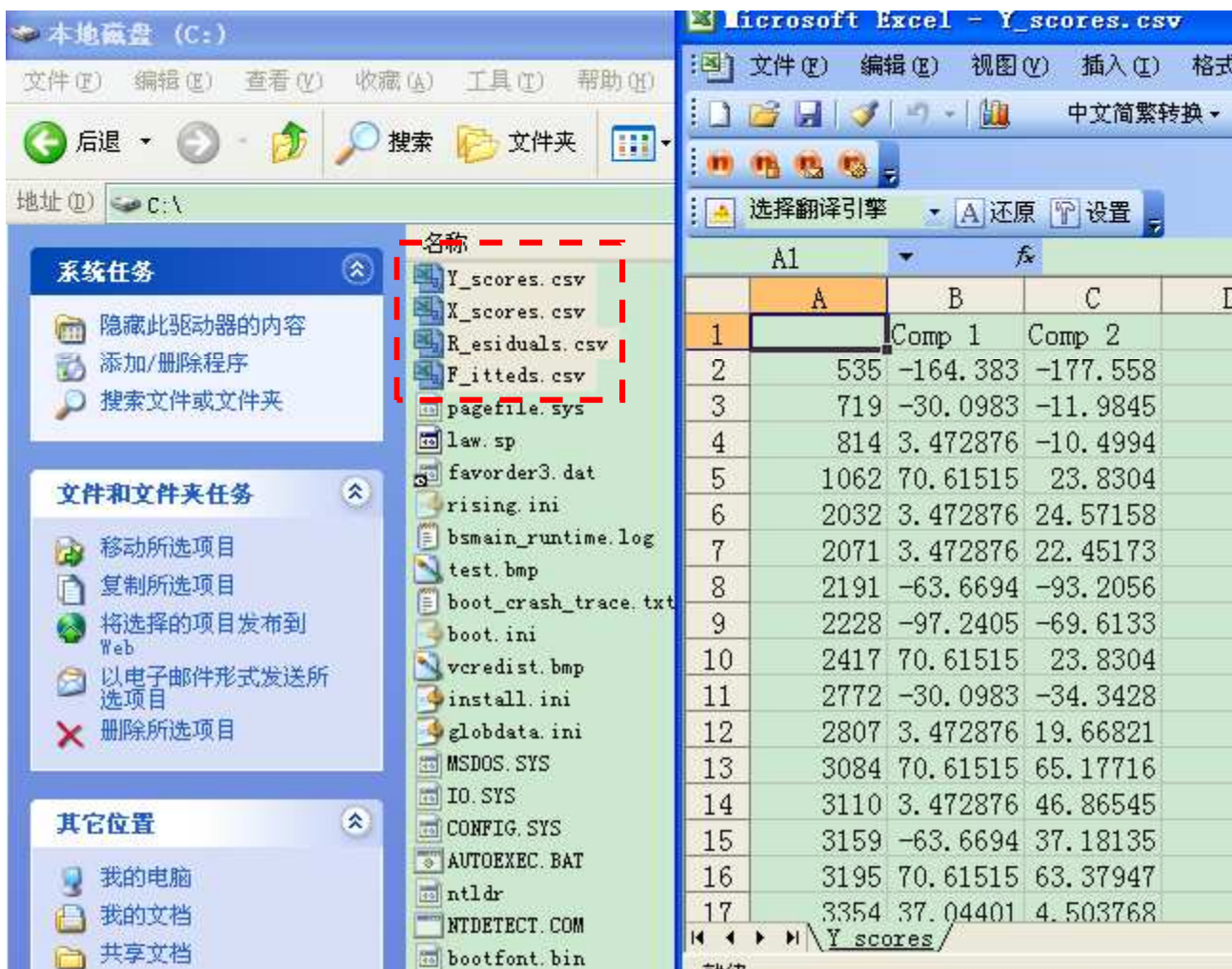
(3) 满意度权重：偏最小二乘法3



! + 漫游费 + 客户贡献 + 贡献价值; + 漫游费 + 客户贡献 + 贡献价值,



(3) 满意度权重：偏最小二乘法⁴

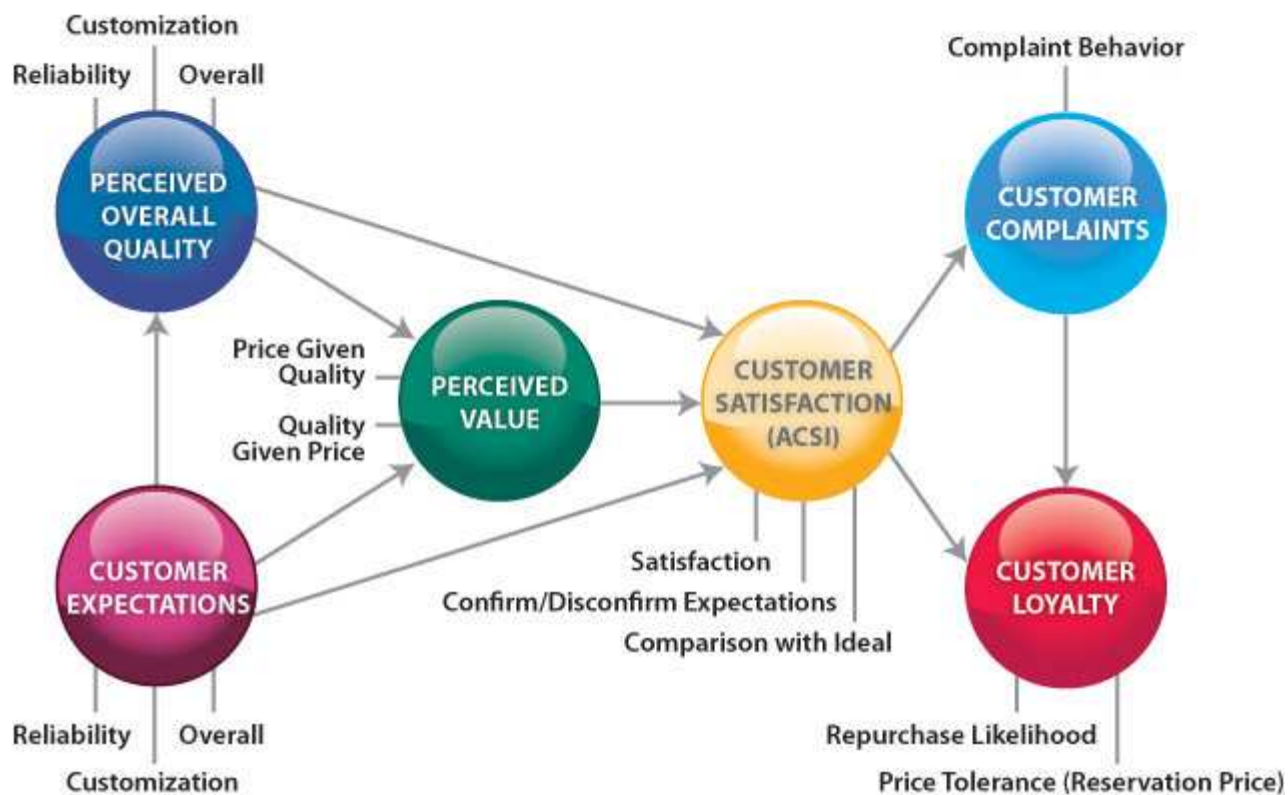


The screenshot shows a Windows Explorer window on the left displaying the contents of the C:\ drive. A red dashed box highlights a group of files: Y_scores.csv, X_scores.csv, R_residuals.csv, F_fitteds.csv, and pagefile.sys. To the right, a Microsoft Excel window titled 'Y_scores.csv' is open, showing a table of data. The table has columns A, B, and C, and rows numbered 1 through 17. The data in column A represents scores, while columns B and C represent component scores (Comp 1 and Comp 2).

| | A | B | C | D |
|----|------|----------|----------|---|
| 1 | | Comp 1 | Comp 2 | |
| 2 | 535 | -164.383 | -177.558 | |
| 3 | 719 | -30.0983 | -11.9845 | |
| 4 | 814 | 3.472876 | -10.4994 | |
| 5 | 1062 | 70.61515 | 23.8304 | |
| 6 | 2032 | 3.472876 | 24.57158 | |
| 7 | 2071 | 3.472876 | 22.45173 | |
| 8 | 2191 | -63.6694 | -93.2056 | |
| 9 | 2228 | -97.2405 | -69.6133 | |
| 10 | 2417 | 70.61515 | 23.8304 | |
| 11 | 2772 | -30.0983 | -34.3428 | |
| 12 | 2807 | 3.472876 | 19.66821 | |
| 13 | 3084 | 70.61515 | 65.17716 | |
| 14 | 3110 | 3.472876 | 46.86545 | |
| 15 | 3159 | -63.6694 | 37.18135 | |
| 16 | 3195 | 70.61515 | 63.37947 | |
| 17 | 3354 | 37.04401 | 4.503768 | |

美国满意度指数

(ACSI (American Customer Satisfaction Index))



基于偏最小二乘法的路径分析

基于PLS路径分析

变量: id, gender

测量变量: CC3, CC1, CC2, CC4, CC5, CE1

bootstrap: ☒ 否 ☐ 是

潜变量个数: 6

指标类型: ☐ 形成型 ☒ 反映型

此插件需要R plspm(祝迎春)

确定 粘贴(P) 重置(R) 取消

路径系数标签

| | CE | PQ | PV | CS | CC | CL |
|----|------|------|------|------|------|------|
| CE | .000 | .000 | .000 | .000 | .000 | .000 |
| PQ | .530 | .000 | .000 | .000 | .000 | .000 |
| PV | .130 | .798 | .000 | .000 | .000 | .000 |
| CS | .256 | .004 | .656 | .000 | .000 | .000 |
| CC | .000 | .000 | .000 | .813 | .000 | .000 |
| CL | .000 | .000 | .000 | .271 | .536 | .000 |

效应分解标签

| | relationships | dir.effects | ind.effects | tot.effects |
|----|---------------|-------------|-------------|-------------|
| 1 | CE->PQ | .530 | .000 | .530 |
| 2 | CE->PV | .130 | .423 | .553 |
| 3 | CE->CS | .256 | .364 | .620 |
| 4 | CE->CC | .000 | .504 | .504 |
| 5 | CE->CL | .000 | .438 | .438 |
| 6 | PQ->PV | .798 | .000 | .798 |
| 7 | PQ->CS | .004 | .523 | .527 |
| 8 | PQ->CC | .000 | .428 | .428 |
| 9 | PQ->CL | .000 | .372 | .372 |
| 10 | PV->CS | .656 | .000 | .656 |
| 11 | PV->CC | .000 | .533 | .533 |
| 12 | PV->CL | .000 | .464 | .464 |
| 13 | CS->CC | .813 | .000 | .813 |
| 14 | CS->CL | .271 | .436 | .707 |
| 15 | CC->CL | .536 | .000 | .536 |

goodnes-of-fit标签

| | GoF | value |
|---|-----------|-------|
| 1 | Absolute | .618 |
| 2 | Relative | .771 |
| 3 | Outer.mod | .998 |
| 4 | Inner.mod | .595 |



学习推荐:

- 《显微镜 如何深入发掘消费者需求》
- 《瞬间的真实》
- 《R语言与统计分析》
- 网站
 - 统计之都: <http://cos.name/>
 - R语言中文论坛 :
<http://rbbs.biosino.org/Rbbs/forums/list.page>
 - 人大经济论坛: www.pinggu.org/
 - 百岛潮论坛: <http://stat.baidao.net>



感谢你的耐心

Thank You !

