REXcel
数据挖掘发展趋势

关菁菁
中国人民大学统计学院06级本科
Email： cattyguan@gmail.com

2009-12-5

目录

1. REXcel
   - REXcel之创始
   - REXcel之启动
   - REXcel之应用
     - 数据导入
     - 数据分析
     - 结果保存

2. 数据挖掘
   - 数据挖掘之目标
   - 理想预测与实行
   - 寻找最fit模型
   - 模型回顾
   - 重点介绍
     - 集成运算（Emsemble Learning）

statconn之“幕后黑手”

（The masterminds behind statconn）

- **Thomas Baier**（1971-）
- 在不同环境中应用**R**
  - R/Scilab (D)COM Server
  - RExcel（1998)

- **Erich Neuwirth**（1948-）
- **RExcel** 的主要作者

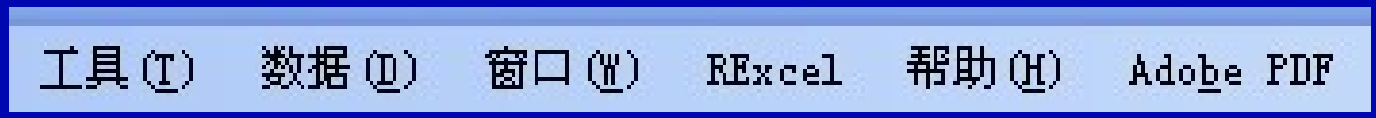
University of Vienna

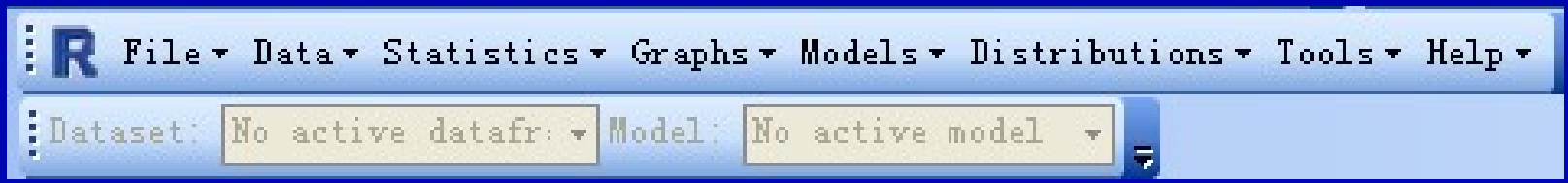
RExcel之创始

http://rcom.univie.ac.at/

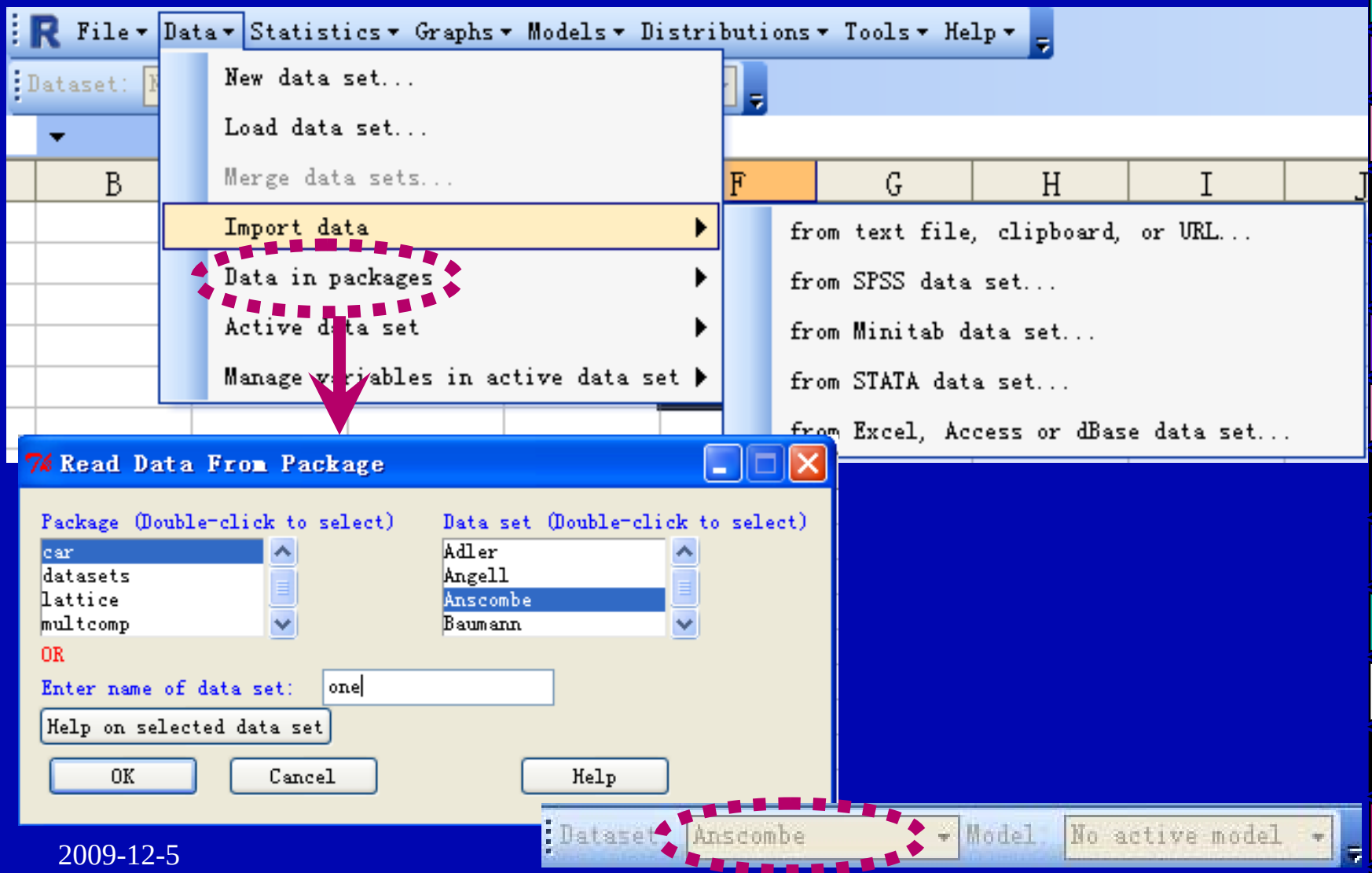
安装：http://rcom.univie.ac.at/

- 安装
  - 据网页提示手动逐步安装
  - 直接下载RAndFriends压缩包
- 安装须知
  - R的版本2.9.0以上
  - Excel的版本03、07均可



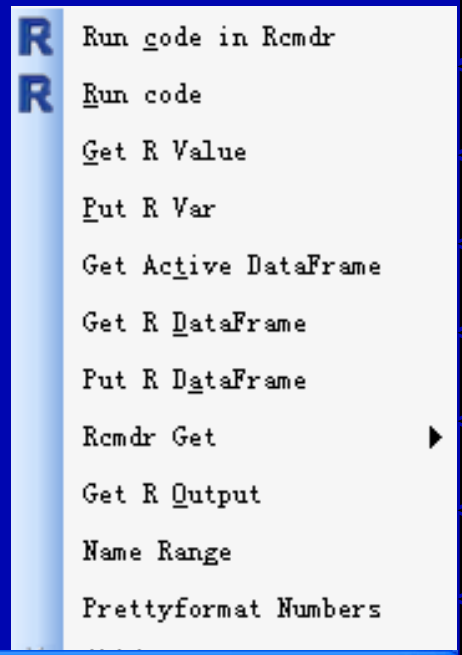


数据导入


File
Data
Statistics
Graphs
Models
Distributions
Tools
Help
Dataset:
New data set...
Load data set...
Merge data sets...
Import data
Data in packages
Active data set
Manage variables in active data set
from text file, clipboard, or URL...
from SPSS data set...
from Minitab data set...
from STATA data set...
from Excel, Access or dBase data set...
Read Data From Package
Package (Double-click to select)
car
datasets
lattice
multcomp
Data set (Double-click to select)
Adler
Angell
Anscombe
Baumann
OR
Enter name of data set:
one
Help on selected data set
OK
Cancel
Help
Dataset: Anscombe
Model: No active model

数据分析

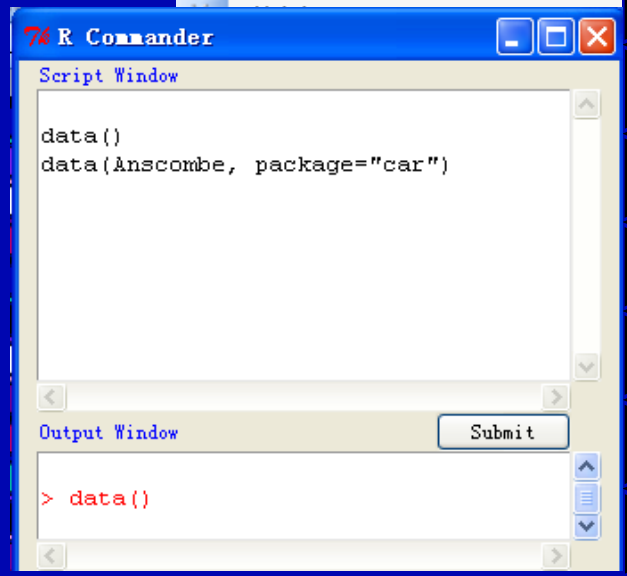
- 任何程序均可写在单元格、Commander、R console中
- 右图为右键功能



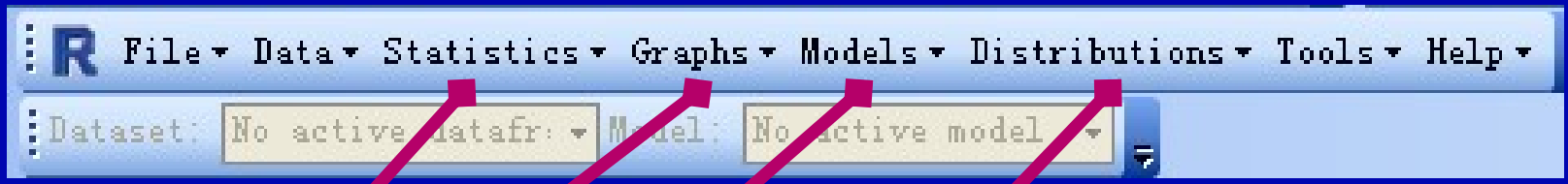
三个部分

RCommander

- Script Window
- Output Window
- Messages



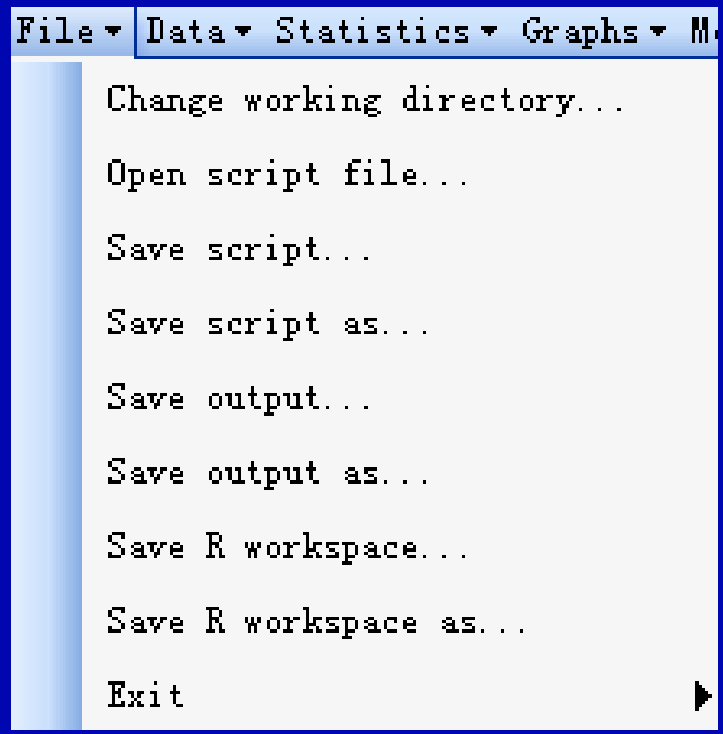
数据分析



- Statistics
  - 描述统计、简单参数和非参数检验、线性模型
- Graphs
  - 各种统计图表
- Models
  - 经典统计模型
- Distributions
  - 各种分布的分位数、分布图、抽样、尾概率等

结果保存

- 可直接储存在Excel中
- 其它储存方法如右图



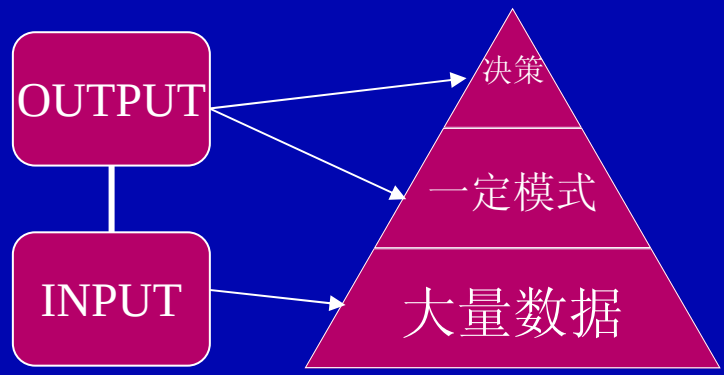
目录

1. REXcel
   - REXcel之创始
   - REXcel之启动
   - REXcel之应用
     - 数据导入
     - 数据分析
     - 结果保存

2. 数据挖掘
   - 数据挖掘之目标
   - 理想预测与实行
   - 寻找最fit模型
   - 模型回顾
   - 重点介绍
     - 集成运算（Emsemble Learning）

数据挖掘的目标

- 传统意义上，，面对大量数据我们要做的是“Extract the pattern of data and know what the data says”。



•但这远远不够。

有监督学习（Supervised Learning）

- Outcome Y
- Predictor X
  - 当Y是数值型变量时，我们可以用回归
  - 当Y取值与有限的无序集合时，可进行分类
- Training data训练集:$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$

目标

- 1、精确预测结果未知的测试集
- 2、理解哪些INPUT影响OUTPUT，怎样影响
- 3、评估我们预测和推断的质量

目录

1. REXcel
   - REXcel之创始
   - REXcel之启动
   - REXcel之应用
     - 数据导入
     - 数据分析
     - 结果保存
   - REXcel之优点

2. 数据挖掘
   - 数据挖掘之目标
   - 理想预测与实行
   - 寻找最fit模型
   - 模型回顾
   - 重点介绍
     - 集成运算（Emsemble Learning）

理想预测与实行

- Y是数值变量，以 $Ave\left(Y-f\left(x\right)\right)^2$ 作为误差测量，当INPUT 为X=x时，理想预测为 $f\left(x\right)=Ave\left(Y\mid X=x\right)$
- Y是定性变量，取值于{1,2,…,M}，$\Pr\left(Y=j\mid X=x\right)$ 取值最大则Y=j.

寻找最fit模型——模型评价

overfit	High variance
underfit	High bias

模型评价要做的事情：

1、Choose a value for a tuning parameter(*s*) for a technique.

2、Estimate the future prediction ability of the chosen model.

两个基本方法

- 最小二乘方法
- KNN（K近邻）方法
  - 投影追踪（Projection Pursuit）

分类器和回归树

- 好处
  - 处理大样本
  - 可以预测连续型变量和定性变量
  - 容易剔除多余变量
  - 规模小的树便于理解
- 不足
  - 规模大的树很难理解
  - 预测能力较差

模型回顾

Linear Models
Generalized Additive Models
Neural Networks
Trees, Random Forests and Boosted Tree Models 
Support Vector and Kernel Machines 

OUTLINE： Ensemble Learning

- 基展开与正则化
- 新观念
- 一般步骤和可解决的问题
- 一个简单例子
- 更好的解决方法

基展开与正则化
Basis Expansions & Regularization

步骤

1.用附加的变量（X的变换）增广或替换X

2.得到新导出的输入特征空间

3.在这个空间上使用线性模型

$$f(X)=\sum_{m=1}^{M}\beta_m h_m(X)$$

$h_m(X)$称为basis

基展开与正则化
Basis Expansions & Regularization

- 如何使用基函数（Basis）控制模型？

一：限制法——在处理前确定函数类限制

二：选择法——自适应地扫描Basis Function Space，选取对模型拟合有显著贡献的基函数，如Greedy Approache

三：正则化——使用整个词典但限制系数，如领回归，lasso

集成运算（Ensemble Learning）

- 新观念——Thomas G. Dietterich（2000）
- Rather than finding **one best hypothesis** to explain the data, **a set of hypotheses** are constructed and then have those hypotheses 'vote' in some fashion to predict the label of new data points

给定一系列的假设$\{h_1,\cdots,h_K\}$

选择一系列的权重$\{w_1,\ldots,w_N\}$

投票分类器为$\mathrm{H}(x)=w_1h_1(x)+\cdots+w_Kh_K(x)$

Ensemble Learning 两步走

- Step1: Developing a population of base learners from training data

  （Basis Function Space）

- Step2: Combining them to form the composite predictor

Ensemble Learning解决的问题

- Statistical ——

  只选择一个假设的风险
- Computational ——

  寻找最优假设的方法
- Representational ——

  不存在最优假设的情况

A Simple Example

考虑一个函数

$$f(x) = \alpha_0 + \sum_{T_k \in T} \alpha_k T_k(x),$$

- T is a dictionary of basis functions
- Friedman and Popescu (2003) proposed

  Step1: A finite dictionary T_L

  Step2: A family of functions $f_\lambda(x)$ is built by fitting a lasso path in this dictionary

$$T_L = \{T_1(x), T_2(x), ..., T_M(x)\},$$

$$\alpha(\lambda) = \arg\min_{\alpha} \sum_{i=1}^{N} L\left[y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x_i)\right] + \lambda \sum_{m=1}^{M} |\alpha_m|$$

*: λ 为正则化参数

A Better Ensemble Learner

目标：找到well-covered function space的basis functions

- Friedman & Popescu 借助了 numerical quadrature & importance sampling
- 未知函数定义为 $f(x)=\int \beta(\gamma)b(x;\gamma)d\gamma$
- $\gamma \in \Gamma$ 用于表示basis functions b(x; γ)
- 数值积分是为了找到M个估计点 γ_m 和相应的权重 α_m，使得在x的有效域内 $f_m(x)=\alpha_0+\sum_{m=1}^{M}\alpha_m b(x;\gamma_m)$能最接近$f(x)$

- Loss function – a measure of relevance

$$Q(\gamma) = \min_{c_0, c_1} \sum_{i=1}^{N} L\left(y_i, c_0 + c_1 b(x_i; \gamma)\right)$$

$$\gamma^* = \arg\min_{\gamma \in \Gamma} Q(\gamma)$$

- Width σ – a measure of SAMPLING SCHEME

$$\sigma = E_S\left[Q(\gamma) - Q(\gamma^*)\right]$$

Narrow	Too many b(x; γ) look alike and and similar to b(x; γ *)
Wide	A large spread in the b(x; γ m), possibly involved irrelevant cases

ISLE Ensemble Generation

1. $f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$

2. $For \quad m = 1 to\ M \qquad do$

$(a)\ \gamma_m = \arg\min_{\gamma} \sum_{i \in S_m(\eta)} L\left(y_i, f_{m-1}(x_i) + b(x_i;\gamma)\right)$

$(b)\ f_m(x) = f_{m-1}(x) + \nu b(x;\gamma_m)$

3. $T_{ISLE} = \{b(x;\gamma_1), b(x;\gamma_2), \ldots, b(x;\gamma_M)\}$

$\eta \in (0,1], \nu \in [0,1],$

$S_m(\eta)$ Refers to a sub sample of $N \bullet \eta$ of training data

$\eta \in (0,1], \nu \in [0,1],$

$S_m(\eta)$ Refers to a subsample of $N \bullet \eta$ of training data

- 目标：找到well-covered function space的basis functions

- η越小，N·η越小，$S_{M(\eta)}$ 的元素数目越少，但可能被抽取的subsample（子样本）个数越多，随机性因此越大，σ正好是这种随机性的衡量

$$\sigma = E_S\left[Q(\gamma) - Q(\gamma^*)\right]$$

η，ν

Ensemble Learning

- Bagging: $\eta = 1$，　$\nu = 0$
- Random Forest: $\eta < 0.5$ ≈reduce m
- Importance Sample Learning Ensemble：
  $\eta \leqslant 0.5 \cap \nu = 0.1$

- ISLE实际上还受到正则化参数 λ 的影响

- randomForest
- rattle
- nlme
- rpart
- TeachingDemos

部分参考文献

- Thomas G. Dietterich, Ensemble learning, The Handbook of Brain Theory and Neural Networks, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002
- Robert E. Schapire, The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, Nonlinear Estimation and Classification. Springer, 2003.
- Peter Bühlmann and Torsten Hothorn, Boosting algorithms: regularization, prediction and model fitting. Statistical Science, 22(4):477-505, 2007.
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second edition, Springer, 2009


谢谢
REXcel
数据挖掘发展趋势
关菁菁
中国人民大学统计学院06级本科
Email: cattyguan@gmail.com