



# 北京中科资环信息技术 研究院文件

中科资环发(2018)52号

## 基于 R 语言的回归分析与机器学习实践 技术应用研修班

### 各企事业单位:

统计学关心的是数据的收集、分析及其解释。随着计算机的普及和人们采集数据能力的提升,很多行业对充分认识数据、从数据中挖掘有用信息的需求,变得越来越为迫切。本课程将基于开源统计软件 R 语言,结合丰富的实际案例数据,详细地介绍如何使用 R 语言进行回归及其拓展统计建模分析。通过本课程的学习,学员们将会初步掌握基于 R 语言的回归分析及其拓展应用。

R 是功能强大、免费、开源的编程语言,具有强大的统计建模和画图功能,能够运行在 Linux、Windows、Macintosh 操作系统上。R 具有丰富和强大的包,几乎在各个领域都有可免费下载安装的 R 包供使用者参考,是统计领域乃至整个科学计算领域的主程序程序设计语言。

本课程将从 R 的基本安装操作开始,逐步介绍如何使用 R 语言拟合和解释线性回归模型和方差分析模型;将详细介绍 R 基础包 stats 中线性回归分析函数 lm 的具体应用。针对二项数据和多项数据,我们分别介绍二项回归模型和多项回归模型的基本原理及其 R 语言实现。针对计数数据,我们详细介绍泊松回归模型、拟泊松回归模型、负二项回归模型的统计原理及其 R 语言实现。对上述两类数据,我们重点介绍 R 基础包 stats 中广义线性模型拟合函数 glm 的具体应用。此外,我们介绍一般混合效应模型的建模原理及其 R 语言实现,并结合不同类型数据特点介绍常用的混合效应模型。我们重点介绍 lme4 包中的 lmer 函数的语法规则。最后,我们结合具体案例,简要介绍回归和分类树、随机森林、集成学习 (AdaBoost, Bagging)、神经网络的算法原理及其 R 语言实现。

本课程将从 R 的基本操作开始,一步步讲解,从数据可视化到常见数据分析方法的使用,并结合相关学科具体实例,让学员能从中借鉴学习。现将有关事宜通知如下:

- 一、**主办单位:** 北京中科资环信息技术研究院 **会议指定协办单位:** 北京中科硕博研计算技术中心  
二、**时间地点:** 2018年8月17日---8月21日 呼和浩特 (时间安排:第一天报到、授课四天)  
三、**培训目标:**

1. 掌握 R 基础操作。
2. 能通过 R 包和 R 语言编程解决实际问题。
3. 掌握 R 集成开发环境 Rstudio 的使用。
4. 掌握线性回归分析基本原理及其 R 语言实现。
5. 掌握二值和多分类回归模型原理及其 R 语言实现。
6. 掌握常用计数数据回归模型原理及其 R 语言实现。
7. 掌握随机效应模型基本原理及其 R 语言实现。
8. 了解回归树、分类树、随机森林、集成学习、神经网络模型及其 R 语言实现(简介)。

### 四、主讲专家:

主讲专家来自北京师范大学,长期从事统计学基础核心课程的教学工作,主要研究领域为复杂数据分析和统计建模、应用统计,有丰富的实际数据建模分析经验。

主办单位:北京中科资环信息技术研究院

协办单位:北京中科硕博研计算技术中心

二零一八年七月十日

二零一八年七月十日

101060538

1060652442



### 五、培训对象:

各省市、自治区对应用回归分析建模解决行业实际问题有需求的企事业单位技术骨干、科研院所研究人员和大专院校相关专业教学人员及在校研究生、硕士、博士等相关人员,以及广大 R 语言爱好者。

### 六、培训费用:

**报名费用: 3900 元** (报名费、培训费、资料费、午餐费) 住宿自理,可预订房间。

(**发票可开具: 培训费、会议费、资料费等,会场发放红头文件,用于参会人员报销使用**)

### 七、报名办法:

请各单位接此通知后,尽快确定参加培训人员,认真填写报名回执并传真 010-53853500 或 E-mail 至:[676741493@qq.com](mailto:676741493@qq.com),会务组收到回执后通知报到相关事项。

### 颁发证书:

参加会议的学员可以获得《R 编程技术》专业技术培训证书。此证书作为个人学习和知识更新、专业技能提升、单位人才聘用的参考依据。

**注:** 请学员准备身份证、学历证复印件各 1 份,2 寸蓝底证件照 1 张(背面注明姓名)。

### 七、联系方式:

**联系人: 宋鹏 17731236430**

**微信咨询: sq11544**

**电话/传真: 010-53853500**

**QQ 咨询: 676741493**

**精通 R 语言 qq 群: 512765924 (验证人填写: 宋鹏邀请)**

### 八、内训流程简介:

了解企事业单位的培训需求→电话访谈或面谈,诊断需求→提交/修改培训解决方案→签订培训合作协议→培训实施→培训效果评估→培训效果跟踪与回访。

如有相关单位对内训有需求或感兴趣的老师,您可以详细来电咨询!

## 报名回执表

单位名称							
发票抬头				发票类型	<input type="checkbox"/> 增值普票 <input type="checkbox"/> 增值专票		
纳税人识别号				发票内容	<input type="checkbox"/> 培训费 <input type="checkbox"/> 会议费		
通讯地址							
学员姓名	性别	部门	工作(研究)方向	职务	联系方式	邮箱	
希望通过会议解决哪些问题(请认真填写)		问题一: 问题二: 问题三:					
培训费		万 仟 佰 拾 元			电汇日期		
培训费用: <input type="checkbox"/> 电汇 <input type="checkbox"/> 现场办理费用 (请用“√”标注)							
汇款账户	账户: 北京中科硕博研计算技术中心			账户: 北京中科资环信息技术研究院			
	开户行: 中国工商银行股份有限公司北京学院路支行			开户行: 中国工商银行股份有限公司北京学院路支行			
账号	账号: 0200 0255 0920 0106 214			账号: 0200 0255 0920 0091 221			
	注: 请将银行汇款凭证传真至会务组。			注: 请将银行汇款凭证传真至会务组。			
是否需要住宿: <input type="checkbox"/> 是 <input type="checkbox"/> 否 入住日期_____, 住宿天数____天, 标准间_____间。				参会单位盖章: 参会学员签名: 会议负责人: 宋鹏			
二零一八年 月 日							



# 课程大纲

<p>R 基础</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ R 及 RStudio 下载安装</li> <li>■ R 基本功能介绍 (setwd、search、library 等函数)</li> <li>■ RStudio 介绍</li> <li>■ RMarkdown 的使用</li> <li>■ 数据框及其 R 生成函数 data.frame 语法规则</li> <li>■ 因子及其 R 生成函数 gl 语法规则</li> <li>■ 其他 R 数据结构 (列表、矩阵等)</li> <li>■ R 绘图 (高级绘图函数、低级绘图函数)</li> <li>■ 参数设置函数 options</li> <li>■ R 数据读入写出 (read.table, write.table 函数)</li> <li>■ 循环、判断语句</li> <li>■ 函数</li> </ul>
<p>线性回归分析 (stats 包 lm 函数、anova 函数、glmnet 包 glmnet 函数)</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 保险数据案例及探索性分析</li> <li>■ 总统选举案例及探索性分析</li> <li>■ 线性回归分析模型原理</li> <li>■ lm 函数的语法规则</li> <li>■ F 检验、t 检验</li> <li>■ 置信区间</li> <li>■ R2 及调整 R2</li> <li>■ qq 图</li> <li>■ 基于 AIC、BIC 及 LASSO 的变量选择</li> <li>■ 预报</li> <li>■ 方差分析</li> <li>■ anova 函数的语法规则</li> </ul>
<p>Logit, Binomial 及 MultiLogit 回归分析 (stats 包 glm 函数、nnet 包 multinom 函数)</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 心脏病案例数据及探索性分析</li> <li>■ 二项工程试验案例数据及探索性分析</li> <li>■ 党派属性类别数据及探索性分析</li> <li>■ Logit 回归模型介绍</li> <li>■ Odds (优势比)</li> <li>■ 残差</li> <li>■ Deviance (偏差: 线性模型残差平方和推广)</li> <li>■ 变量选择</li> <li>■ 拟合优度检验</li> <li>■ 灵敏度、特异度、及 ROC 曲线</li> <li>■ 偏差 (deviance) 分析表</li> <li>■ 广义 R2</li> <li>■ 二项回归模型</li> <li>■ Pearson 卡方统计量</li> <li>■ 多项回归模型</li> <li>■ multinom 函数的语法规则</li> </ul>
<p>计数数据回归分析</p>	<p><b>学习内容:</b></p>



<p>(stats 包 glm 函数、MASS 包 glm.net 函数、pscl 包 zeroinf、hurdle 函数)</p>	<ul style="list-style-type: none"> <li>■ Poisson 分布及负二项分布</li> <li>■ 年重大科学发现数量案例及探索性分析</li> <li>■ 岛屿物种案例数据及探索性分析</li> <li>■ 博士期间论文发表数案例数据及探索性分析</li> <li>■ Poisson 回归分析模型原理</li> <li>■ Poisson Rate Model</li> <li>■ Deviance 统计量</li> <li>■ Pearson 卡方统计量</li> <li>■ Overdispersion</li> <li>■ 拟 Poisson 回归</li> <li>■ 变量选择</li> <li>■ ZIP (Zero Inflated Poisson) 回归</li> <li>■ Hurdle 模型</li> <li>■ 负二项回归</li> </ul>
<p>混合效应模型 (lme4 包 lmer 函数, nlme 包 lme 函数)</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 分组数据、重复测量数据、层级数据</li> <li>■ 中学生各学年成绩数据案例及探索性分析</li> <li>■ 混合效应模型原理 (固定效应、随机效应)</li> <li>■ 参数估计和假设检验</li> <li>■ lmer 函数语法规则及案例应用</li> <li>■ lme 函数语法规则及案例应用</li> <li>■ 家庭收入追踪调查数据及纵向回归模型</li> <li>■ 视觉灵敏度数据及重复观测模型</li> <li>■ 学生多门成绩数据及多响应多水平模型</li> </ul>
<p>决策树、随机森林 ( rpart 包 rpart 函数、prune.rpart 函数 , randomForest 包 randomForest 函数)</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 回归树</li> <li>■ 树的剪枝</li> <li>■ rpart 函数</li> <li>■ prune.rpart 函数</li> <li>■ 随机森林</li> <li>■ 分类树</li> <li>■ 臭氧数据案例分析</li> <li>■ 车型数据案例分析</li> </ul>
<p>集成学习 ( adabag 包 bagging 函数、boosting 函数, )</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 集成学习基本思想</li> <li>■ Bagging 算法原理</li> <li>■ AdaBoost 算法原理</li> <li>■ bagging 函数</li> <li>■ boosting 函数</li> <li>■ 商业银行数据案例分析</li> </ul>
<p>神经网络 ( nnet 包中的 nnet 函数)</p>	<p><b>学习内容:</b></p> <ul style="list-style-type: none"> <li>■ 神经网络基本思想</li> <li>■ 前馈神经网络</li> <li>■ nnet 函数</li> <li>■ 臭氧数据的前馈神经网络分析</li> </ul>