机器学习实践手册

用R语言提供机器学习问题的解决方案

关于作者

Atul Tripathi在机器学习和量化投资领域有超过11年的从业经历，并拥有14年的软件开发和研究经验。他一直致力于先进的机器学习技术，如神经网络和马尔可夫模型的研究。在研究机器学习技术的同时，他利用其解决了图像处理、电信、语音识别和自然语言处理等领域相关的诸多问题。他还利用神经网络模型开发了文本挖掘工具。在量化投资领域，他利用蒙特卡洛仿真开发了价值风险，极值定理，期权定价和能源衍生品等模型 。

关于审稿人

Ryota Kamoshida是Python库MALSS (MAchine Learning Support System), (<https://github.com/canard0328/malss)>的开发者，目前在Hitachi, Ltd公司担任计算机科学领域的高级研究员。

前言

当今世界，数据已经成为新的“黑金”并以指数级的速度增长着。这种增长既包括现存数据的增长，也包括新数据的增长，这些新的数据以结构化和非结构化的形式展现，并来源于社交媒体，互联网，文档文献以及物联网等多种多样的数据源。数据流必须实时地被收集、处理、分析、并最终展现出来以确保数据的使用者能够在如今快速变化的环境中作出理性且明智的决定。机器学习技术将待解决问题的上下文信息应用于这些数据上，用统计学技术确保不断快速到达的复杂数据能够以科学的方式加以分析。并利用机器学习算法从数据中进行迭代学习，发现数据中的隐藏模式和规律。机器学习的这种迭代学习的模式是非常重要的，正因如此，当机器学习模型被暴露在新的数据中时，他们才能从新的数据集中独立地适应和学习以产出可靠的结论。

我们将首先对本书中包含的多种不同的机器学习主题进行简介，随后，我们将会基于现实世界的问题在不同的章节中对各个主题进行一一探讨，例如分类、聚类、模型选择和正则化、非线性问题、监督学习、无监督学习、强化学习、结构化预测、神经网络、深度学习，还有最后的实例研究。本书的机器学习算法以R语言作为编程语言。本书适用于R语言的初学者，但是熟悉R语言对理解和使用本书的代码肯定是会有所帮助的。

您将学习如何合理地决定使用哪类算法以及如何应用这些算法得到最佳的效果。如果您想要对图像、文字、语音、或者其他形式的数据都建立有意义的多功能的应用，这本书绝对会成为您的得力助手。

这本书包括什么

第1章，机器学习引言，涵盖了机器学习的各种概念。本章使读者初步了解本书涵盖的各个主题 。

第2章，分类，包括以下算法：判别分析、多元逻辑回归、Tobit回归，泊松回归。

第3章，聚类，包括以下主题和算法：层次聚类，二元聚类，k-均值聚类。

第4章，模型选择和正则化，包括以下主题和算法：收缩方法（shrinkage methods），降维方法和主成分分析。

第5章，非线性，包括以下主题和算法：广义加性模型，平滑样条，局部回归。

第6章，监督学习，包括以下主题和算法：决策树学习，朴素贝叶斯，随机森林，支持向量机，随机梯度下降。

第7章，无监督学习，包括以下主题和算法：自组织映射和矢量量化。

第8章，强化学习，包括以下主题和算法：马尔可夫链，蒙特卡洛模拟法。

第9章，结构化预测，包括以下主题和算法：隐马尔可夫模型。

第10章，神经网络，包括以下主题和算法：神经网络。

第11章，深度学习，包括以下主题和算法：递归神经网络。

第12章，案例研究-探索世界银行数据，包括世界银行数据分析。

第13章，案例研究-再保险合同定价，包括再保险合同定价。

第14章，案例研究-用电量预测，包括用电量预测。

这本书你需要什么？

本书的的重点是用R语言构建基于机器学习的应用。我们已经使用R语言构建过各种解决方案。我们的重点是利用R语言库和函数以最佳的方式来克服现实世界的挑战。我们尽量保持所有代码的友好和可读性。我们认为这将使我们的读者能够很容易地理解代码， 并在不同的场景中随时使用它。

这本书的目标读者是谁？

这本书是为想构建实用的基于机器学习的应用的，从事于统计，数据分析，机器学习，计算机科学，或其他专业的学生和专业人士准备的 。本书适用于R语言的初学者，但是熟悉R语言对理解和使用本书的代码肯定是会有所帮助的。对于那些希望在现有技术栈中探索机器学习技术的有经验的 R语言程序员来说，本书也将是非常有用的。

特殊段落

在本书中，您讲频繁看到如下标题（**准备工作**和**怎样做**）。

为了更加清晰的说明怎样完成一个流程指南，我们使用了如下特殊段落。

准备工作

这部分告诉你该流程指南需要哪些准备，并描述了如何设置该流程指南所要求的软件或其他先决条件。

怎样做……

该部分包含了流程指南的各个步骤。

公约

在本书中，您将发现许多用于区别不同信息类型的文本样式。下面是一些文本样式的例子和它们的含义解释。

文本中的代码、数据库表名、文件夹名称、文件名、文件扩展名、路径名、虚拟URL、用户输入和 Twitter 句柄用下面的格式表示：“我们将把数据保存到 fitbit\_details 帧中"

任何命令行输入和输入以如下格式表示：



**新名词**和**重要的词语**用加粗字体表示。你在屏幕上看到的文字，例如菜单或对话框中的文字，以这样的方式显示：“**Monte Carlo v/s Market n Zero Rates** ”

 警告或者重要的提示

 小提示和技巧

读者反馈

我们永远欢迎读者的反馈。让我们知道你对这本书的看法——你喜欢或者不喜欢的内容。读者反馈对我们来说很重要, 因为它能帮助我们了解最能让您受益的主题 。

您通过发送电子邮件到 [feedback@packtpub.com](mailto:feedback@packtpub.com)将意见反馈给我们并请在邮件中注明该书的标题 。

如果您擅长某个话题并有兴趣写作或贡献一本书，请见我们的作者指南[www.packtpub.com/authors](http://www.packtpub.com/authors)。

客户支持

现在您已经是Packt book的一名骄傲的拥有者，我们有许多帮助你从这次购买中获益更多的方式。

下载示例代码

您能从您在<http://www.packtpub.com>的账户中下载这本书的示例代码文件。如果您是从其他渠道购买的这本书，您可以访问<http://www.packtpub.com/surpport> 并注册，这些文件就会通过email的方式直接发送给你 。

您可以通过以下步骤下载代码文件：

1、使用您的电子邮件地址和密码登录或注册我们的网站 。

2、将鼠标悬停在顶部的**SUPPORT**选项卡上。

3、点击**Code Downloads & Errata**。

4、在搜索框中输入该书的名称。

5、选择您要下载代码文件的那本书。

6、从下拉菜单中选择您购买这本书的地方。

7、点击**Code Download**。

您也可以通过点击Packt出版网站中这本书页面的**Code Files**按钮来下载代码文件。通过在搜索框中输入这本书的书名可以访问到该页面。请注意您需要登录您的Packt帐号。

文件被下载后，请使用最新版本的解压软件解压或提取文件夹：

Windows: WinRAR / 7-Zip

Mac: Zipeg / iZip / UnRarX

Linux: 7-Zip / PeaZip

这本书的代码包也被托管在GitHub上，https://github.com/PacktPubl ishing/Practical-Machine-Learning-Cookbook。我们种类丰富的书和录像中的代码包同样被托管在GitHub上，地址是<https://github.com/PacktPublishing/> 。快去查看吧！

下载这本书的彩色图片

我们还为您提供了一个PDF文件，其中包含了本书的彩色截图和图表。这些彩色图片将帮助您更好的理解输出的变化。您能够从下面的地址下载该文件<https://www.packtpub.com/sites/default/files/down loads/PracticalMachineLearningCookbook_ColorImages.pdf>。

勘误表

虽然我们已经尽己所能来保证内容的准确性，但错误难免发生。如果您在我们的书籍中发现了错误（可能是文字中的错误或代码中的错误）并报告给我们的话，我们将不胜感激。通过这样做，您可以帮助其他读者免受错误的影响，并帮助我们改进本书的后续版本。如果您发现了任何错误，请通过如下方式报告给我们。访问<http://www.packtpub.com/submit-errata>，选择你的书名，点击**Errata Submission Form**链接，并输入错误的详细内容。一旦您的勘误被验证，您的提交将被接受，勘误将上传到我们的网站或者被添加到标题为勘误表的勘误章节下。

如果您想查看以前提交的勘误，请访问https://www.packtpub.com/books/content/support，并在搜索框中输入书名。所需的信息将出现勘误部分下。

盗版

互联网上的各种媒体形式的盗版版权材料是一个屡禁不止的问题。在Packt，我们采取的版权和许可证保护非常严格。如果您在互联网上遇到我们作品的任何形式的非法复制品，请立即向我们提供地址或网站名称，以便我们采取措施。

请通过copyright@packtpub.com联系我们，并附上疑似盗版材料的链接。

我们感谢您帮助保护我们的作者，以及我们为您带来有价值内容的能力。

问题

如果您有任何关于本书的问题，请通过 [questions@packtpub.com](mailto:questions@packtpub.com) 联系我们，我们将尽力处理您的问题。

1 机器学习引言

在本章中，我们将介绍机器学习及其涵盖的多个话题。您将了解以下内容：

•什么是机器学习？

•分类方法概述

•聚类方法概述

•模型的选择和正则化概述

•非线性方法概述

•监督学习概述

•无监督学习概述

•增强学习概述

•结构预测概述

•神经网络概述

•深度学习概述

什么是机器学习？

人类自出生起就被暴露在各种数据中 。眼睛、耳朵、鼻子、皮肤、舌头不断收集着各种形式的数据，然后大脑将其转换成视觉、听觉、嗅觉、触觉和味觉。大脑处理各种形式的感觉器官收到的原始数据，并将其转换成语音，进而用语音表达对于这些原始数据的观点。

当今世界，我们用与机器相连的传感器收集数据。从各式各样的网站和社交网络收集数据。之前的手写材料也在电子化、数字化后被加入数据集中。 这些形式丰富的从不同数据源中采集的数据需要经过处理才能得到更有洞察力的，更有意义的结果。

机器学习算法有助于从不同数据源收集数据，转换富数据集的形式，并基于分析结果，帮助我们采取明智的行动。机器学习算法被设计成高效、准确的方法，并提供通用的学习方法来解决以下问题：

•处理大规模问题

•做出准确的预测

•解决各种不同的学习问题

•学习哪些结果可以被得出，以及在何种条件下这些问题能够被学习

机器学习算法的一些应用领域如下：

•基于销售数据的价格预测

•预测药物的分子反应

•检测汽车保险欺诈

•分析股市回报

•识别高风险贷款

•风力发电厂产量预测

•跟踪和监测医疗保健设备的利用率和位置

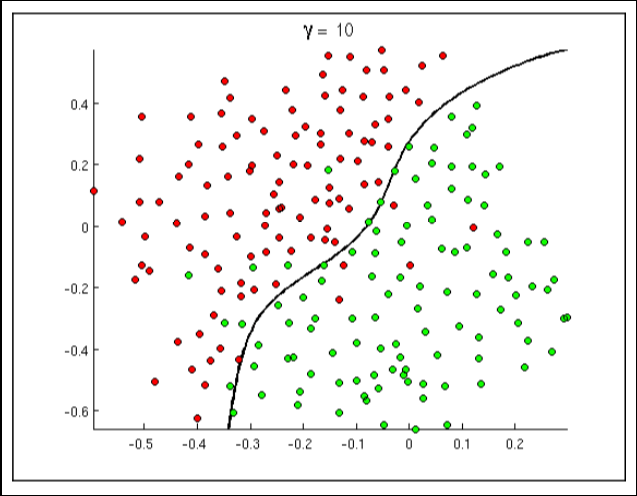
•计算能源有效利用率

•分析智能城市交通运输增长趋势

•采矿业矿产储量估算

分类方法概述

线性回归模型本质上是进行量化的响应，但是这样的响应本质上却是定性的。就像态度（强烈不同意，不同意，中立，同意和强烈同意）这样的响应，其本质上就是定性的。对于一个观察来说，预测一个定性的响应可以被视作对这个观察进行分类，因为这涉及到把这个观察分配给一个类别或种类。分类器对于今天的许多问题来说是非常有价值的工具，比如药物或基因组学预测，垃圾邮件检测，面部识别和财务问题。



聚类方法概述

聚类是将相似对象聚合在成一簇的过程。每一个簇由彼此之间相似并且与其他类的对象不相似的对象组成。聚类的目标是确定一组未标记数据的内在分组。 聚类可用于数据挖掘（DNA分析，营销研究，保险研究等），文本挖掘，信息检索，统计计算语言学和基于语料库的计算词典学等不同应用领域。 聚类算法必须满足的要求如下：

•可扩展性

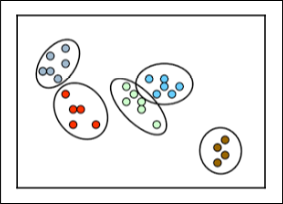
•处理各种类型的属性

•发现任意形状的簇

•处理噪音和异常值的能力

•可解释性和可用性

下图是聚类的一个示例：



监督学习概述

监督学习需要学习一组输入变量（通常为向量）和输出变量（也称为监控信号）之间的映射，并应用此映射来预测未知数据的输出。监督学习的方法尝试发现输入变量和目标变量之间的关系。发现的关系在被称为“模型”的结构中表示。通常隐藏在数据集中的模型描述和现象解释，在知道输入属性后，可以用于预测目标属性的值。

监督学习是从有监督的训练数据（训练样本集）推测出预测函数的机器学习任务。训练数据由一组训练样本组成。在监督学习中，每个例子是一个由输入对象和期望输出值组成的。监督学习算法分析训练数据并学习出预测函数。

为了解决监督学习问题，必须执行以下步骤：

1、确定训练样本的类型。

2、收集训练集。

3、确定预测函数的输入变量。

4、确定预测函数的结构和相应的学习算法。

5、完成设计

6、评估预测函数的准确性。

监督学习的方法可以被应用在各个领域，如市场营销，财务和制造业。

在监督学习中要考虑的一些问题如下：

•权衡有偏变量

•函数复杂性和训练数据量

•输入空间的维度

•输出数据的噪声

•数据的异构性

•数据冗余

•相互作用和非线性问题的存在

无监督学习概述

无监督学习针对全体输入样本，学习出一种特定的模型来表征输入样本整体的统计结构。无监督学习是重要的，因为它在大脑的学习过程中比监督学习更常见。例如，眼睛中光感受器的活动是随着视觉世界而不断变化的。他们持续提供着可用于显示世界上有什么对象，如何呈现，照明条件等等的所有信息。然而，基本上没有关于场景内容的信息在视觉学习期间是可用的。这就使得无监督的方法至关重要，并被用作适合神经突触的计算模型。

在无监督学习中，机器接收输入，但是既没有有监督的目标输出，也没有从环境中获得奖励或者反馈。想象一下机器在没有得到坏境的任何反馈时可能会学到什么，这似乎有些神秘。 然而，建立一个正式的无监督学习框架是可行的，因为无监督学习是基于这样的概念——机器学习的目标是建立一种用于决策制定，预测未来输入，高效传输输入到其他机器等目的的输入的表示。 某种意义上来说，无监督学习可被认为是在上述数据中发现模式和规律，并且考虑噪声的影响。

无监督学习的一些目标如下：

•在不需要目标输出的前提下，在大型数据集中发现有用的结构

•提高输入数据集的学习速度

•通过为每个可能的数据向量分配分数或概率来构建数据向量模型

增强学习概述

增强学习是训练一个会自我行动的“代理人”来最大化它从世界中获取奖励的问题。它是关于如何行动以及如何将事件情况映射到动作，以最大化量化的奖励信号的问题。像大多数机器学习方法一样，学习者刚开始不知道要采取哪些行动，而是要通过尝试来发现哪些行为能够产生最大的奖励。强化学习的两个最重要的区别特征是“尝试、错误，搜索”的过程和延迟奖励。 强化学习的一些例子如下：

•当一名棋手思考下一步棋时，他是通过计划下一步可能的反馈以及计数的反馈来做决策的。

•自适应控制器实时调整炼油厂的操作参数。控制器在指定边际成本的基础上权衡优化收益/成本/质量，而不是严格遵守工程师最初建议的设定参数。

•一只瞪羚在出生后几分钟就挣扎地站起来。半小时后，它已经能够每小时奔跑20英里。

•教一只狗一个新的技巧—— 不告诉它做什么，但是如果它做正确/错误的事情，就给它可以奖励/惩罚。它不得不弄清楚它是如何得到奖励/惩罚，这被称为信用分配问题。

增强学习就像试错学习一样。代理人应该从环境经验中发现良好的策略，以至在过程中不失去太多的奖励。“探索”是关于寻找有关环境的更多信息，而“利用”是利用已知信息来最大化奖励。例如：

•餐厅选择：利用：去你最喜欢的餐厅。探索：尝试一个新的餐厅。

•石油钻探：利用：在最有名的地点进行钻探。探索：在新的位置钻探。

增强学习的主要内容如下：

•策略：策略是代理人的行为函数。它确定从环境的感知状态到所采取的行为之间的映射。它对应于心理学所称的一组刺激反应规则或关联。

•价值函数：价值函数是对未来奖励的预测。某个状态的价值是从该状态开始代理人期望在未来积累的奖励总额。而奖励决定了环境状态的即时、内在的可取性，价值表示了状态长期的可取性，其中考虑了该状态之后的状态序列以及伴随这些状态的奖励。

•模型：模型预测下一步环境将会如何变化。它预测下一个状态和下一个状态的即时奖励。

结构化预测概述

结构化预测是机器学习问题的重要应用领域。考虑输入x和输出y，x和y可能是如下类型的数据：一组标注过的时间序列，一张图片的一组属性，句子的解析成分，或者从一张图片上分割出的一组对象，结构化预测问题是很有挑战性的，因为y’s是包含它的输出变量的数量的指数级别。由于预测需要搜索巨大的参数空间以及统计学因素的考虑，结构预测问题的计算上是具有挑战性的。因为从有限的数据中学习精确的模型需要推导不同结构化输出之间的共同点。结构化预测本质上是一个映射问题，其中映射必须捕捉x和y之间的区别性交互，并且还允许在y上进行有效的组合优化。

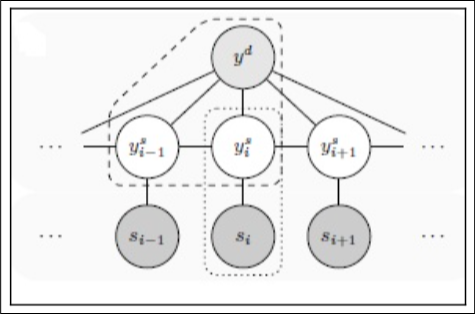
结构化预测是关于从输入数据中预测结构化的输出的问题，不同于只预测一个数值的分类或回归问题。例如：

•自然语言处理 - 自动翻译（输出：句子）或句子解析（输出：解析树）

•生物信息学 - 二级结构预测（输出：二分图）或酶功能预测（输出：树中的路径）

•语音处理 - 自动转录（输出：句子）或文本转语音（输出：音频信号）•

•机器人 - 规划（输出：动作顺序）



神经网络概述

神经网络代表了进行信息处理的仿大脑结构。这些模型受生物学的启发，而不是大脑实际功能结构的精确复制。神经网络能够从数据中学习的能力很强，已被证明是许多预测和分析应用场景中非常有应有价值的系统。

人工神经网络通过更新网络架构和连接权重来学习，使网络能够有效地执行任务。它可以从可用的训练模式中学习，或者从训练样本或输入-输出关系中自动学习。学习过程由以下之一设计：

•了解可用信息

•学习范例 - 从环境中获取模型

•学习规则 - 找出更新权重的过程

•学习算法 - 通过学习规则确定调整权重的过程

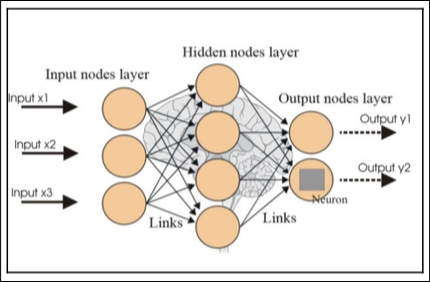
学习规则有四种基本类型：

•纠错规则

•玻尔兹曼

•赫布

•竞争学习



输入节点层 隐含节点层 输出节点层

输入x1 输出y1

输入x2 输入y2

输入x3 神经元

连接 连接

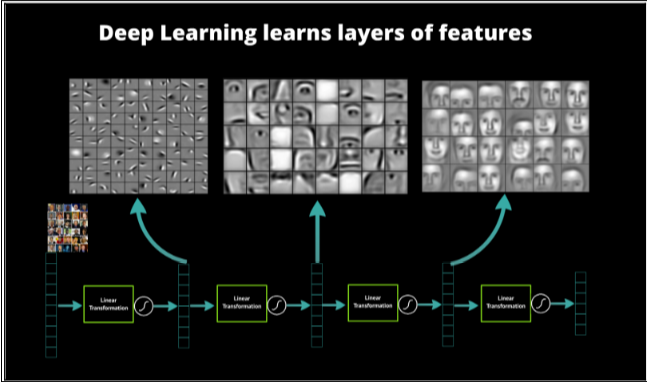
深度学习概述

深度学习是指一组相当广泛的机器学习技术和架构，他们的特点是使用了多层非线性信息处理结构化模型架构。 深度学习架构有三大类：

•无监督或生成式深度学习网络

•有监督深度学习网络

•混合深度学习网络



深度学习模型学习各特征层

2 分类

本章将涵盖如下内容：

•判别函数分析-地下卤水地质学分析

•多元逻辑回归-分析理解学生的课程计划选择

•Tobit回归－评估学生的学术能力

•泊松回归－加拉帕戈斯群岛物种分析

引言

判别分析（Discriminant analysis）用来对观察对象进行区分并分组，进而分配新的观察对象到先前定义好的组中。举例来说，如果进行一项研究来调查有哪些变量能够被用来区分（1）灵长类动物（2）鸟类 或（3）松鼠食用的水果，研究者可以收集多个被这三类动物食用的水果的特点。那么大多数水果会自然分到这三类中。判别分析可被用来确定哪些变量能够最好的预测一种水果是否会被鸟类、灵长类或者松鼠食用。此外，判别分析还常被用于生物物种分类，医学肿瘤分类，以及信用卡和保险行业的风险界定。判别分析的主要目标是判别和分类。关于判别分析的三个假设是各判别变量之间具有多元正态分布，变量间的低多重共线性，各组变量的协方差矩阵相等。

多元逻辑回归（Multinomial logistic regression）基于多个自变量，预测样本所在的分类或者分类在某因变量上的概率。它一般使用在因变量有两个以上名字型分类或无序分类的情况下，而且这种问题中自变量的虚拟编码相当普遍。自变量可以是二分的（二进制）或连续的（区间或比例）。多元逻辑回归使用最大似然估计来进行分类问题求解，而不是传统多项式回归中使用的最小二乘估计。样本概率分布的一般形式被事先假设，在参数估计的过程中，首先设定参数的初值，在设定的参数初值下，计算从事先假定分布进行样本抽样的似然值。该过程被迭代进行直到在某个参数估计下得到最大似然值。

Tobit回归（Tobit regression）是用来描述非负自变量和因变量之间关系的。它也被称为审查回归模型（censored regression model）。该模型被设计用来在因变量存在或左或右的删剪时估计变量之间的线性关系。因变量删剪发生在下面的情况下，当一个变量的值等于或者高于某个阈值时，该变量取值为该阈值，所以变量的真实值可能等于该阈值，也可能高于该阈值。Tobit模型被用在许多样本的因变量为零或特定值的应用中（如汽车支出，医疗支出，工作时间，工资等等）。该模型是为了度量受限的因变量，这些因变量 只在它高于或低于某个截断阈值时才能管观察到。例如：

•工资最低额被最低工资限制，所以工资额不会低于最低工资

•捐献给慈善机构的捐赠额度

•最高的编程收入

•个人的可用时间和休闲活动时长

泊松回归（Poisson regression）处理因变量是一个计数总和的情况。除了因变量（Y）是一个符合泊松分布的计数之外，泊松回归与传统的多项式回归很相似。因此，Y的可能值是非负整数：0，1，2，3等等。一般来说非常大的计数是很少见的。泊松分布与逻辑回归相似，因为逻辑回归也是有一个离散的响应变量。有所不同的是，泊松回归的响应不像逻辑回归那样局限于特殊的值。

判别分析－地下卤水地质学分析

假设有一项对从矿井收集的古代工艺品的研究需要进行。从矿井采集岩石样品并对采集的岩石样品进行地质化学测量。一个相似的地质化学测量已经在收集来的古代工艺品上进行过。为了将岩石样本进行分类，归类到他们被挖掘出来的矿井，DFA可以作为一个很好的工具。然后，建立好的DFA函数就可以用于判定这些古代工艺品来源于哪个矿井。

准备工作

为了应用判别函数分析，我们需要使用一组从矿井采集的数据集合。

第一步－收集并描述数据

我们将使用标题为BRINE的地质学数据分析数据集。该数据集可以从<http://www.kgs.ku.edu/Mathgeo/Books/Stat/ASCII/BRINE.TXT>获取。该数据集以标准格式存储，每行是一个样本，每列是一类变量。每个样本被指派给一个地质单元并列在最后一列。该数据集一共包含19个样本和8类变量。这8类数值型变量如下：

•No

•HCO3

•SO4

•CL

•CA

•MG

•NA

•Group

如何做……

让我们进入详细步骤。

第二步－探索数据

第一步是载入下面的包：



版本信息：本页的代码在R 3.2.3中测试（2015-12-10）

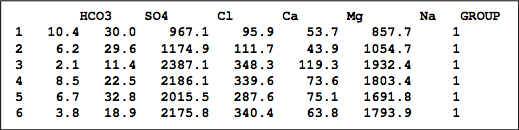
让我们探索一下这些数据并初步理解变量之间的关系。我们将从导入名为brine.txt的文本数据开始。我们将这些数据保存到brine数据帧中，具体操作如下：



接下来，我们将打印brine数据帧。head()函数返回brine数据帧的内容。brine数据帧被当作输入变量传入head()函数。具体命令如下：



结果如下：

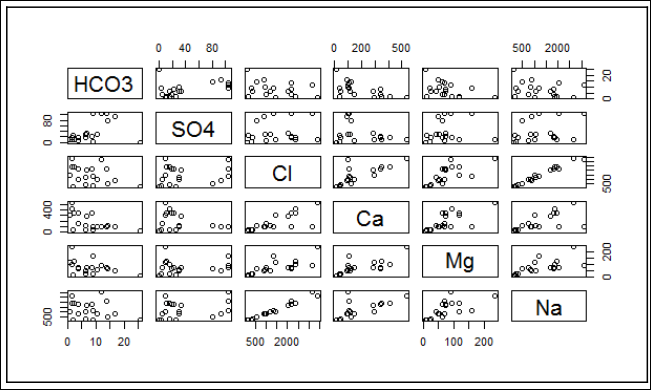


DFA假设了变量符合多元正态分布。因此在分析之前需要验证数据是否符合多元正态性。

为了验证数据集是否合适进行转换，我们首先画出这些数据。pairs()函数被用来画出数据，该函数可产生一个离散点矩阵。交叉作图法只交叉比较1－6列的变量。最后一个变量（第7列）是组名所以不列入比较。操作如下：

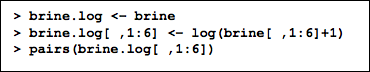


画出的图表如下面的截图：



第三步－转换数据

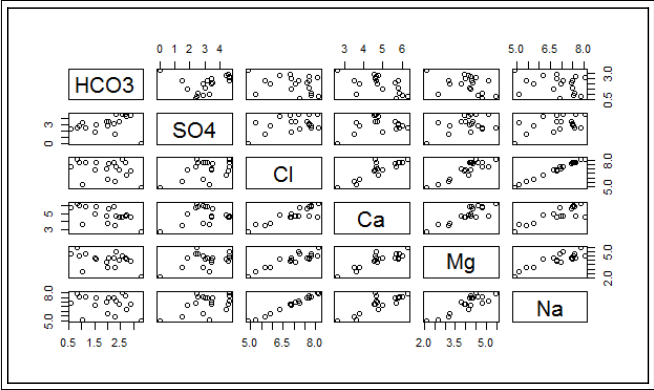
明显可见数据呈现彗星状分布。这表明需要对数据进行log变换，log变换对地质化学数据来说是常见的变换方法。首先对整个数据集进行复制是很好的实践经验，然后对地质化学指标进行log变换。因为数据中包含了0，所以我们对这份数据集采用log+1变换，而不是log变换。brine数据帧被拷贝到brine.log数据帧。对该数据帧进行log变换。代码如下：



在对数据进行变换之后，为了重新评估多元正态性。使用pairs()函数重新绘出brine.log。现在的分布看上去更加正态。相比之前的图偏斜减少了。



画出的图表如下面的截图：



第四步－训练模型

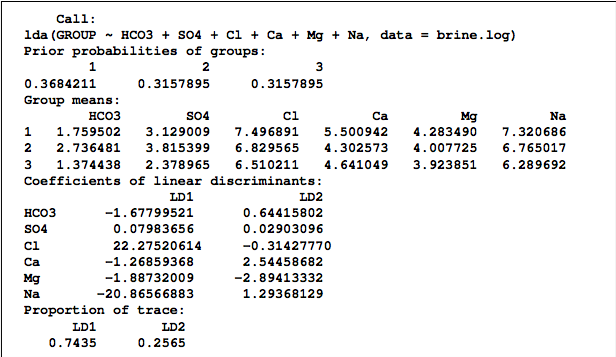
这一步是通过判别函数分析训练模型。调用lda()函数进行判别函数分析如下：



调用的形式很像线性回归或ANOVA，都需要传入一个定义好的方程。该问题中，GROUP变量被当作因变量，地质化学指标被当作自变量。这次试验中，变量间没有交叉，所以变量被用＋操作加起来而不是＊操作。因为没有调用attach()函数，所以数据帧的名字也必须传入。在运行DFA之后，通过如下命令查看结果：



结果如下所示：



•输出结果的第一部分显示拟合的公式。

•第二部分是各组的先验概率，它反映了数据集内各组的占比。换句话说，如果你没有度量变量，并且实验中各组的样本数量代表了实际各组的相对丰富程度，那么这个先验概率将可描述任何未知样本属于各组的概率。

•第三部分用各组各变量平均值表的形式显示了各组的均值。扫描此表可以帮助您发现某组是否在一个或多个变量中具有特殊性。

•第四部分显示了判别函数的系数（a组，b组和c组）。因为该实验一共有三个分组，所以有3-1个线性判别（如果只有两个分组，则只需要1个(2-1)线性判别）。每个变量对于每个线性判别（LD1和LD2） 有一个相应的系数。

•最后，第五部分显示了迹（trace）的比例，该比例给出了每个判别函数的方差解释。在这里，第一个判别解释了将近75%的方差，其余的由第二个判别解释。

第五步－分类数据

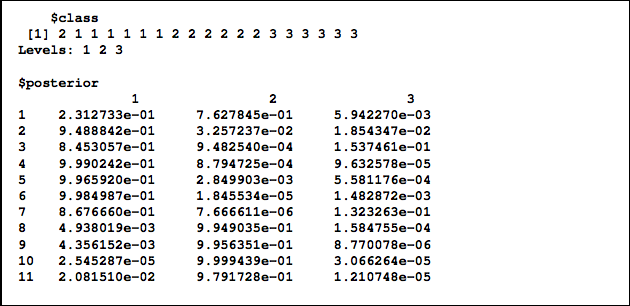
predict()函数也是MASS包中的一部分，该函数使用lda()的结果将样本分配给各组。换句话说，lda()得到了一个能够分组的线性函数，predict()可以在相同的数据中应用该函数，来检验分类函数的分类能力有多强。按照统计学惯例，x-hat是x的预测（hat加到对象名字上，以明确他们是预测）。代码如下：

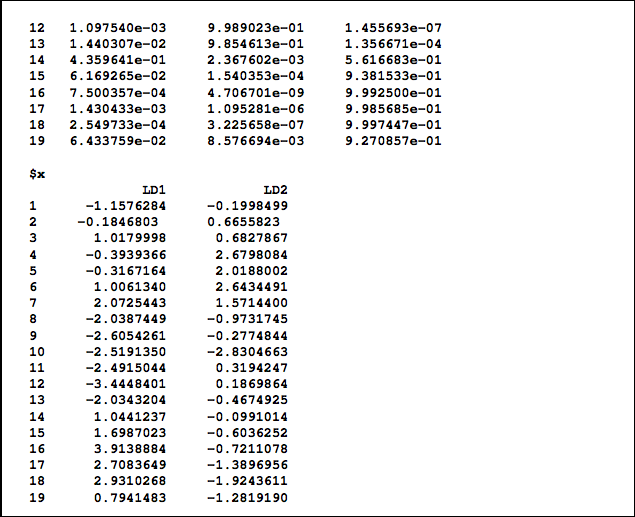


打印brine.log.hat如下：



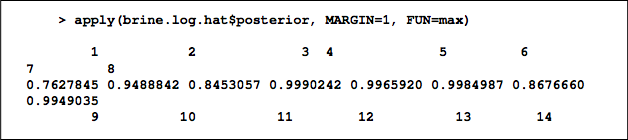
结果如下：

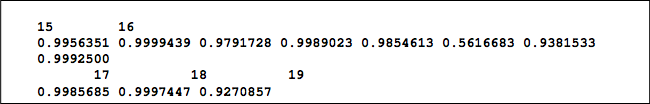




上面的输出以每个样本被指定的分类开始。接下来列出了每个样本属于每个分组的后验概率，每一行（即每个样本）的概率之和为1.0。这些后验概率衡量了样本属于每个分类的强度。如果一个样本属于某个组的概率比其他概率高出很多，那么该样本属于这个组的可能性就很高。如果两个或多个概率接近，那么样本的分组就比较不确定。

接下来的命令可以在有多个分组的情况下以一个快捷的方式找到每个样本的最大概率：





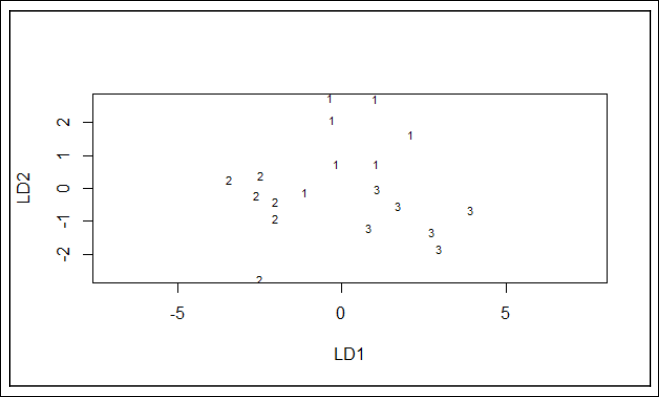
因为数据集中大多数概率值是较大的（>0.9），这表明大多数样本已经被成功分配给一个组。

如果大多数概率是大的，整体的分类就是成功的。predict()的最后一部分输出列出了每个样本对于每个判别函数轴的分数。就像主成分分析中的分数也可以被画出一样，这些分数被画出以展现分组在判别函数中是如何分布的。代码如下：

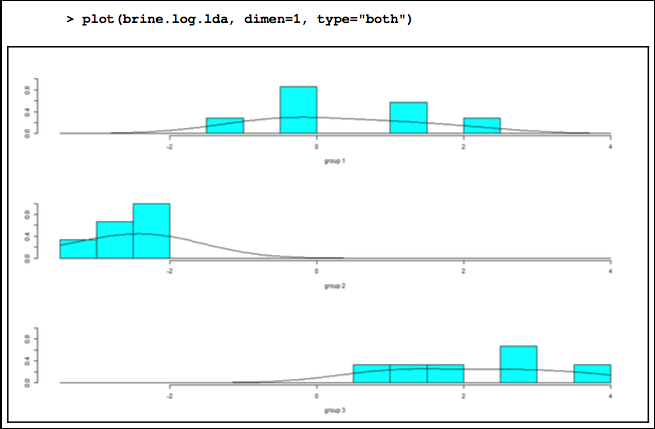


这三组占据了完全不同并且不重叠的区域。只有一个1组的样本靠近2组，所以我们可以明确地说这个判别是成功的。

画出的图表如下：



第二种图表展示一个特定的判别函数轴上的数据如下：



同样的，判别函数1上有着良好的组分割，特别是2组更加明显。

第六步－评估模型

DFA的分类效率需要被评估，可以通过比较predict()的分类结果和真实分组来进行评估。table()函数是一个很有用的用于评估的函数。按照约定，调用该函数时，需将实际分组当作第一个参数，预测的分组当作第二个参数，代码如下：



打印tab的值



结果如下：



（该图应该有错误，列坐标应该左移一格）

输出的每一行对应每个分组的原始数据，每一列对应于DFA做出的分组。在一个完美的分类中，大的数值会落在对角线上，对角线外的值应该是0，这就表明所有属于1组的样本被DFA判别为属于1组，其他组如是。该表的形式能够让你很好的洞察哪些分组被可信的判别了。它也能表明哪些分组可能是混乱的以及哪种错分类比其他常见。

下面的命令可以计算整体预测准确度，即落在对角线上的样本比例：



结果如下：



这里的预测准确率接近95%，非常成功。该方法衡量了所谓的再代入误差（resubstitution error），即当所有样本都被用来构建判别函数时的分类准确率。

第二个衡量DFA的手段是留一交叉验证法（也叫对折验证法），这种验证法在训练时排除一个样本。使用排除一个样本后余下的n－1个样本进行DFA训练。交叉验证技术自动地应用在数据集中的每一个样本上。为了完成这个验证，需添加CV=TRUE(即交叉验证)到lda()调用中，代码如下：



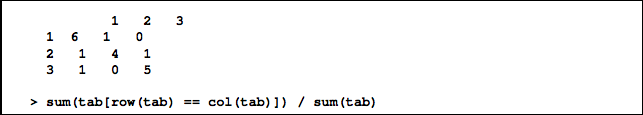
分组判别的准确性能够被上文类似的方法衡量，代码如下：



打印tab的值如下：



结果如下：



结果如下：



该数据集中，对折验证准确率较低（准确率只有79%），这反映了再代入误差往往造成对DFA性能的高估。这样的问题在类似这样的小数据集中非常常见，而判别函数分析往往在大数据集中更加成功。

多元逻辑回归—分析理解学生的课程计划选择

让我们假设高中学生在进行课程计划选择。学生们拥有选择课程计划的机会，课程计划有三种，分别是普通课程，职业课程和学术课程。每个学生的选择是基于自己的写作分数和社会经济情况。

准备工作

为了完成这个任务我们需要使用学生的数据集。第一步是收集数据。

第一步－收集数据

名为hsbdemo的学生数据集可被使用。该数据集可从http://voia.yolasite.com/resources/hsbdemo.csv 下载，格式为MS Excel。该数据集有201个数据行和13个变量。八个数值型变量如下：

•id

•read

•write

•math

•science

•socst

•awards

•cid

非数值型变量如下：

•gender

•ses

•schtyp

•prog

•honors

如何做……

让我们详细讲解细节

第二步－探索数据

第一步是载入类库。如果类库不存在，library()函数将返回错误。使用的命令如下：



版本信息：本页的代码在R 3.2.3版本中测试（2015-12-10）

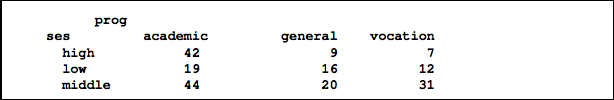
探索数据的过程可以发现这些数据之间关系的一些有益的结论。名为hsbdemo.csv的CSV文件需要被载入到R环境。被导入的数据保存在以ml为标题的数据帧中，代码如下：



我们可以使用with()函数来查探感兴趣变量的描述性统计信息，代码如下：



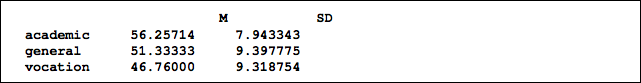
结果如下：



我们可以通过如下代码获取均值和标准差：



结果如下：



通过上面的结果可以发现学术课程的均值最高，通用课程的标准差最高。

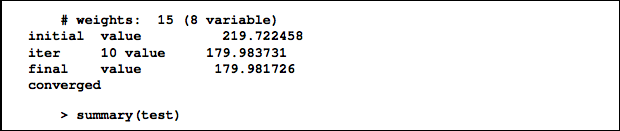
第三步－训练模型

我们使用multinom()函数估计多元逻辑回归。multinom()函数不需要对数据进行变形。

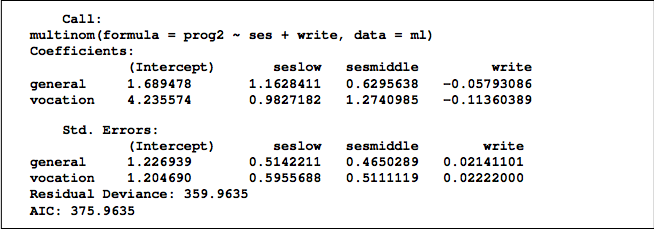
选择结果的对照组很重要。我们可以通过rellevel()函数选择我们期望作为基准的结果水平。然后，我们使用multinom()函数运行我们的模型。因为没有为回归系数进行p值计算，这里我们使用Wald检验（z检验）进行p值检验 。传入Multinom()的公式具有响应－预测器（response-predictors）的形式。数据帧ml是公式中出现的变量的解释，代码如下：



结果如下：



结果如下：



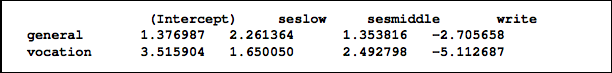
接下来，系数检验摘要除以标准误差检验摘要，如下：



显示z的值如下：



结果如下：



第四步－测试模型结果

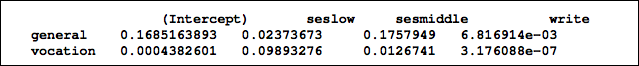
用如下代码进行双尾z检验：



展示p值如下：



结果如下：



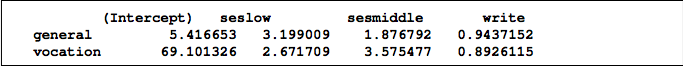
第五步－模型性能改进

相对风险率的定义是指定结果类别和基准类别的概率的比值。相对风险率是线性方程右边部分的指数。指数回归系数是预测变量中一个单元变化的相对风险比率。

从模型中提取出模型系数，然后计算系数的指数如下：



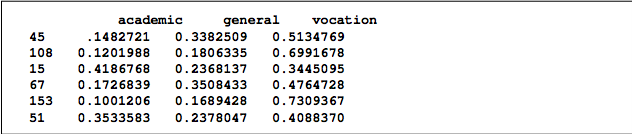
结果如下：



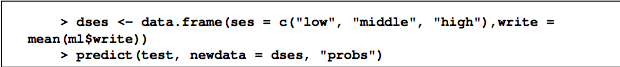
变量write的普通课程对比学术课程的相对风险比率单元增长是0.9437。普通课程对比学术课程的相对风险比率转换从ses=1到3是0.3126。使用已经被预测的概率来得到一个模型的洞察。fitted()函数被用来计算每个结果水平下的预测概率如下：



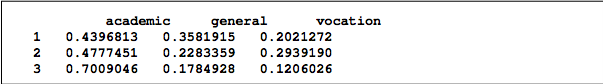
结果如下：



检查与ses和write两个变量之一相联系的概率的变化。在保持一个变量不变的同时，创建一个在另一个变量上变化的小数据。首先，保持write变量为它的均值，然后检查每个ses变量水平的预测概率如下：



结果如下：



看看对于拥有连续预测值的变量的不同值的平均预测概率，用预测概率如下：



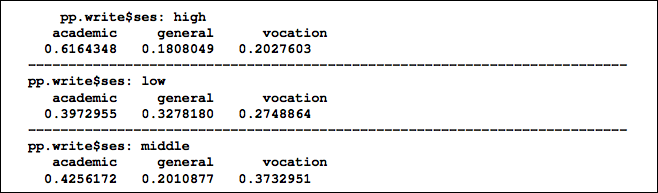
存储每个ses和write值的预测概率如下：



计算每个ses等级内的平均概率如下：



结果如下：



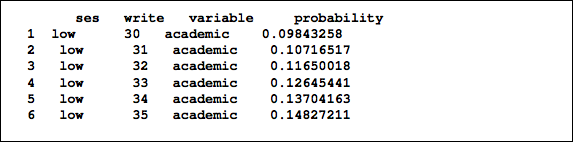
有时，一对图像的对比能够更好地传递信息。使用先前为pp.write对象生成的预测，我们能够画出不同ses等级下writing分数的预测概率变化。melt() 函数能够把数据转换成以宽表的格式的表示，并将一组列堆栈压进单独的数据列中。lpp数据帧被用来存储转换后的数据帧如下：



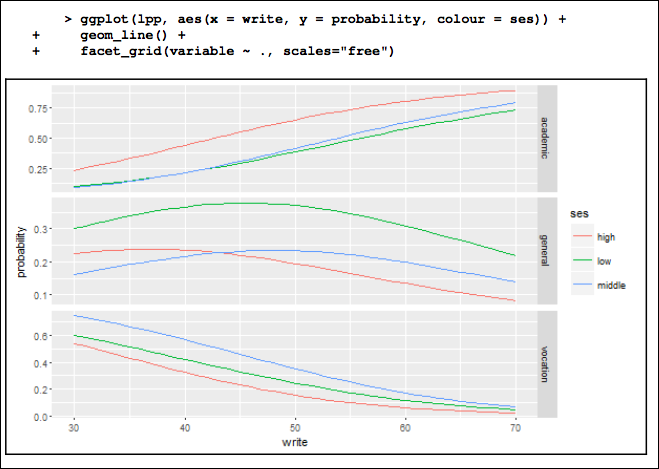
打印数值的头几行如下：



结果如下：



下面我们画出预测概率在write数值对于每个ses等级以课程类型分片的图表如下：



Tobit回归－评估学生的学术能力

让我们通过打分来评估一个学生的学术能力，分数范围是200-800分。该评估基于使用阅读和数学分数的模型。学生参与的课程计划的类型特点也被考虑进去。有三种课程计划类型：学术型，普通型和职业型。这种打分的评分方式存在一个问题，就是一些学生可能回答对了所有的学术能力测试问题并得到800分，但这些学生可能在学术能力上并不真正相同。同样的情况也会发生在那些全部答错所有问题并且得到200分的学生上。

准备工作

为了完成这个实验我们需要用到学生的数据集。第一步是收集数据。

第一步－收集数据

为了建立Tobit回归模型，我们需要使用名为tobit的学生数据集，通过http://www.ats.ucla.edu/stat/data/tobit.csv 可以获取到该数据集，格式为MS Excel。该数据集有201行数据和5个变量。4个数值型变量如下：

•id

•read

•math

•apt

非数值型变量如下：

•prog

如何做……

让我们详细讲解。

第二步－探索数据

第一步是载入如下库。require()函数被设计用来在其他函数的内部使用；如果包不存在它返回FALSE并给出警告（而不是像library()一样默认返回错误）。所用的代码如下：



版本信息：本页的代码在R 3.2.3版本中测试（2015-12-10）

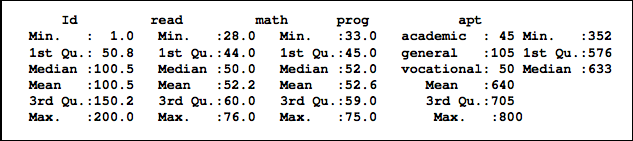
下面我们开始探索数据并理解不同变量间的关系。我们从导入名为gala.txt的CSV数据文件开始，将保存数据到dat数据帧中，代码如下：



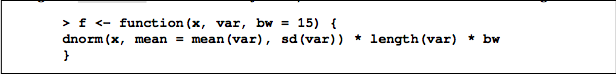
在该数据集中，apt的最低值是352。这表明没有学生得到最低分数200。即使下界删剪是可能的，也不需要在本数据集中考虑该问题。使用如下的命令：



结果如下：



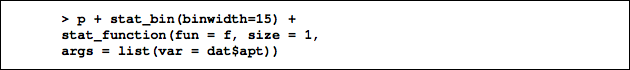
第三步－画出数据

Write是一个函数，它给出了给定均值和方差的正态分布的概率密度，该概率密度已经被拉伸到总数度量。我们使用如下代码生成以density\*sample size\*bin width为高度的柱状图：

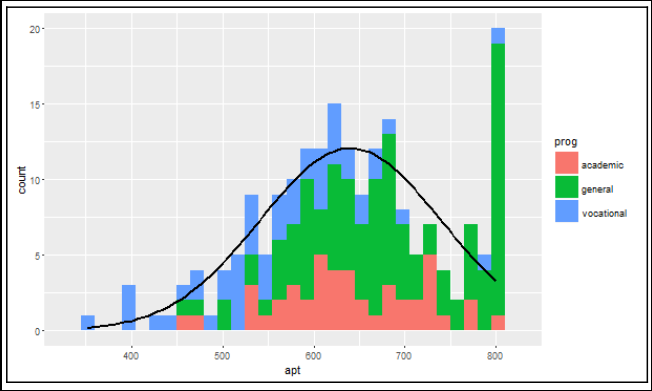
设置基准线如下：



现在我们准备绘制柱状图，以不同颜色表示不同课程计划的占比，同时绘出正态分布曲线覆盖其上作为对比，代码如下：



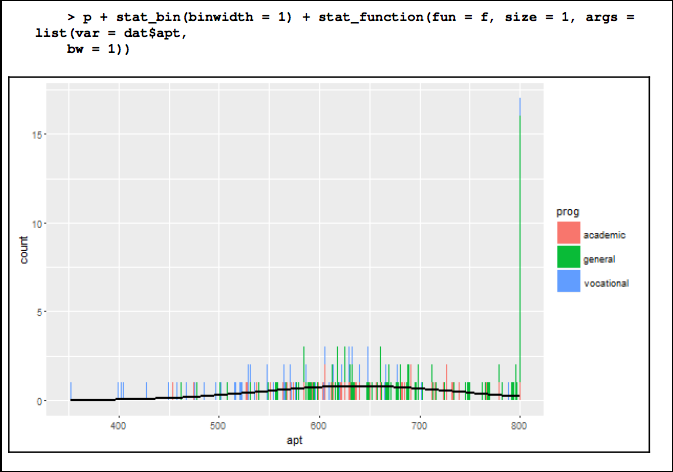
柱状图如下：



观察上面的柱状图，我们能够发现apt明显被截断了，在750-800的分段上，相比分布的其余部分，样本数量比期望数量多很多。

在下面的另一种形式的柱状图中，apt＝800的样本数量被着重标出 。在下面的柱状图中，通过设置breaks 选项使每一个apt独立值都有自己的柱子（设置breaks等于创建了一个包含从最小的apt到最大值apt的矢量）。虽然接近分布中央的位置，一些apt值有两到三个样本，但因为apt值是连续的，所以数据集中大多数apt值只有唯一的样本。

柱状图最右边的尖峰是apt＝800时的样本柱子，相对于其他柱子来说，这个柱子的高度清晰的显示了该值的样本数大大超出常规的数量。用如下命令生成该图：

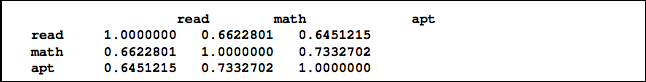


第四步－探索关系

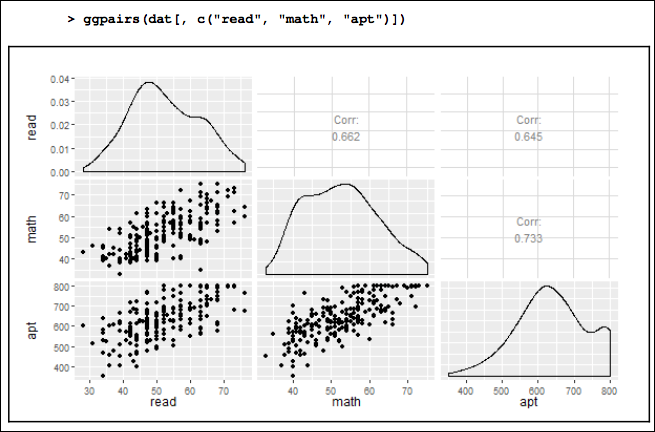
下面的命令用来探索数据集中的二元关系：



结果如下：



画出关系矩阵如下：



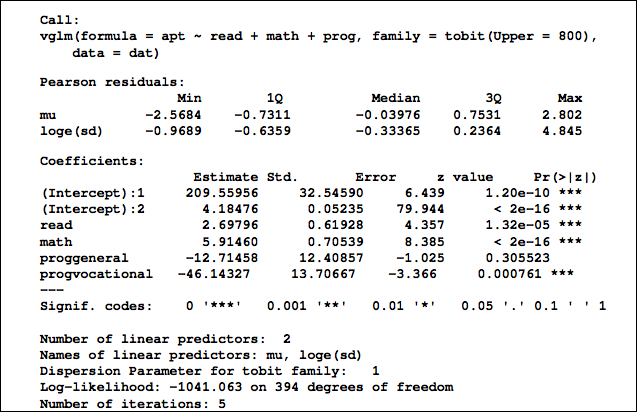
在散点图矩阵的最下面一行中，散点图显示了read和apt的关系。math和apt的关系也被建立并展现出来 。

第五步－训练模型

使用VGAM包中的vglm函数来运行Tobit模型，命令如下：



结果如下：



前面部分的输出告诉我们选项设置。

以coefficients为标签的表给出了标准误差和z统计的系数。摘要表中不包含p值。

Tobit回归系数的解释与OLS回归系数的解释相似。线性系数影响指的是在删剪隐藏变量上的影响：

•read分数增加1个单位，apt预测值增加2.6981（这里错了吧，应该是2.6980）分。

•math分数增加1个单位， apt预测值增加5.9146个单位 。

•prog的相关项有一个稍微不同的解释。选择职业课程计划的学生的apt预测值是－46.1419，即相比选择学术课程计划的学生低。

•标签为（Intercept）:1的系数是模型的截断或常数。

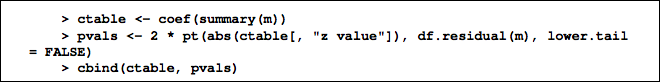
•标签为（Intercept）:2的系数是一个附属统计量。该值的指数，类似于OLS回归残余方差的平方根。65.6773的值相比学术能力的标准差99.21，是一个比较大的降低。

输出的最下面显示了最终的对数相似度，－1041.0629；该值可被用来做模型的交叉对比。

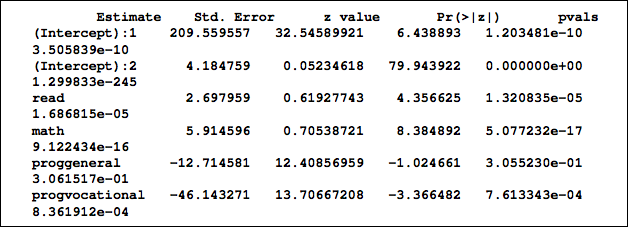
第六步－测试模型

使用z值计算模型中每个系数的p值。然后以扁平表的形式显示这些值。

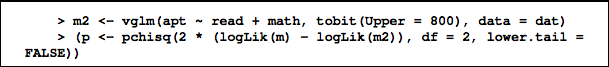
read，math，和prog＝3（职业课程计划）的系数统计上非常显著，命令如下：



结果如下：



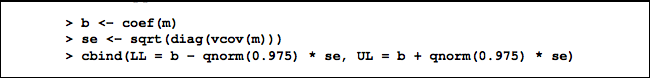
我们使用去除课程计划类型标签的数据整体上适配一个模型，然后使用似然比来检验课程计划类型的显著性，代码如下。



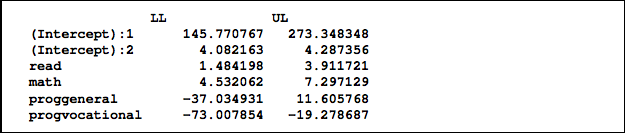
结果如下：



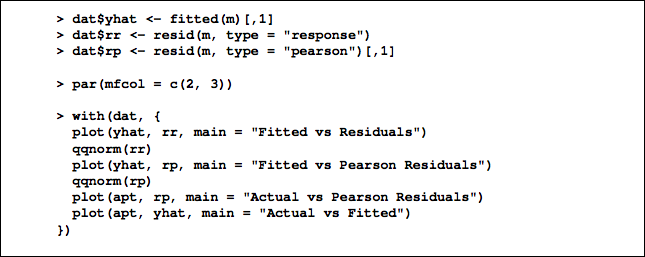
prog变量的p值等于0.0032，该值表示了prog变量的统计显著性。我们计算该系数的上限和下限分别为95%的置信区间如下：



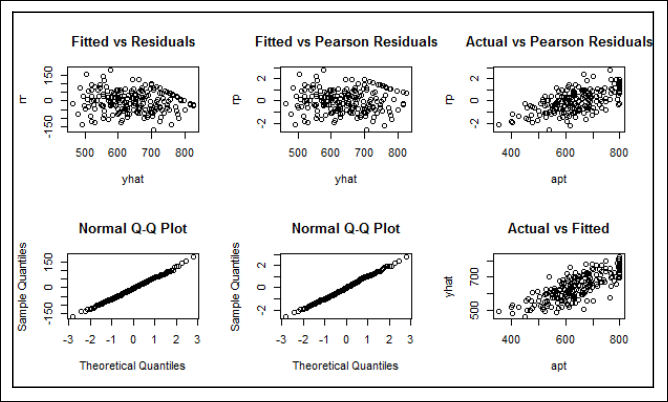
结果如下：



通过在同一个图中画出残差，我们可以方便的比较绝对的和相对的（Pearson）皮尔逊值以及正态性和同质变量的假设。这有利于我们检查模型和数据匹配情况，代码如下：



画出的图表如如下截图：



建立相关性如下：



结果如下：



方差计算如下：



结果如下：



apt预测值和实际观测值的相关性是0.7825。如果计算该值的平方，我们可以得到平方相关系数，这表示预测值与apt变量分享了61.23%的方差。

泊松回归－加拉帕戈斯群岛现存物种分析

加拉帕戈斯群岛坐落在距离厄瓜多尔海岸大约1000公里远的太平洋中。该群岛由13个岛屿组成，其中5个有人居住。该群岛的动植物种类十分丰富。科学家至今仍非常困惑为什么如此小且遥远的群岛能够具有如此繁盛多样的物种群。

准备工作

为了完成这个实验，我们需要利用物种数据集。第一步是收集数据。

第一步－收集并描述数据

我们利用名为gala的物种数量数据集，该数据可从<https://github.com/burakbayramli/kod/blob/master/books/Practical_Regression_Ano ve_Using_R_Faraway/gala.txt> 获得。该数据集包含30个样本和7个变量。这7个数值型的变量包括：

•Species

•Endemics

•Area

•Elevation

•Nearest

•Scruz

•Adjcacent

如何去做……

让我们进入详细步骤。

第二步－探索数据

探索数据将会帮助我们发现一些数据间关系的线索。我们从导入名为gala.txt的文本数据文件开始。我们将数据保存到gala数据帧中，命令如下：



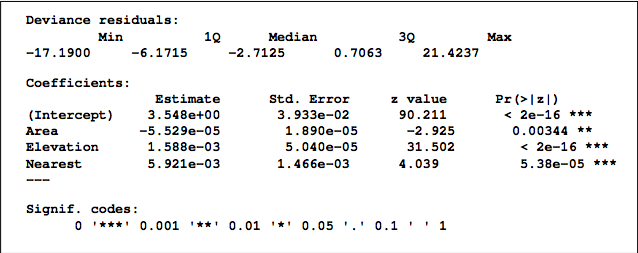
regpois()函数给出了生态学角度非常重要的几个变量的泊松回归，命令如下：

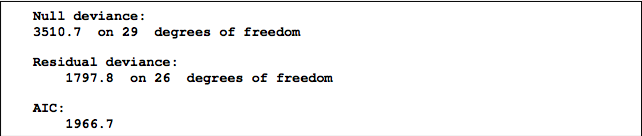


下面给出数据的摘要：



summary函数将提供偏差残差，系数，signif编码（signif codes），零偏差，剩余偏差(Null deviance)，AIC，和费雪分数迭代轮数。结果如下：

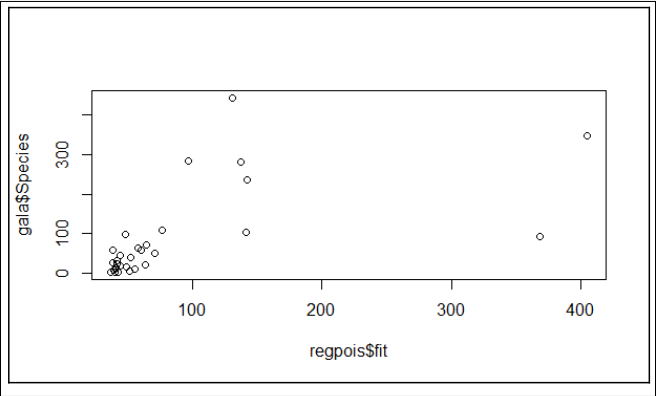




费雪分数迭代轮数：



绘图的结果如下截图：



第三步－绘出数据并测试经验数据

ppois()是泊松的分布函数，输入参数是lambda=regpois$fit，计算gala$Species变量的泊松分布如下：

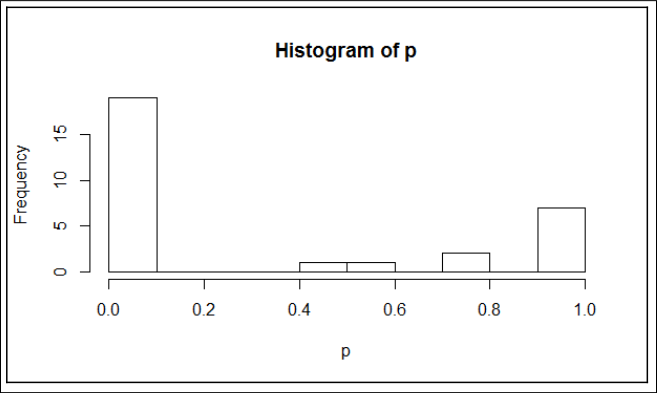


该值应该天然地接近平均分布。

通过画出p的分布来检查该值是否符合平均分布



绘图结果如下截图所示：



这个画图清楚的显示该值不是平均分布的 。

现在我们进行Kolmogorov-Smirnov检验，检验经验数据是否符合一个给定的分布。

Kolmogorov Smirnov检验是对拟合优度的检验，它通常用来检验一个从未知分布中随机抽取的样本是否符合一个已知的特定分布函数。我们也通常用Kolmogorov Smirnov检验来检验方差分析中的正态性假设。

Kolmogorov Smirnov检验是一个统计学假设检验。我们首先确定一个零假设，H0—我们检验的两个样本来自于同一个分布。然后，我们寻找这个假设被拒绝的证据，并用概率的方式表示这个证据。如果样本来自于不同分布的可能性超过了一个置信水平，我们就需要原假设被拒绝，而转而支持H1假设，H1假设是假设两个样本来源于不同的分布。

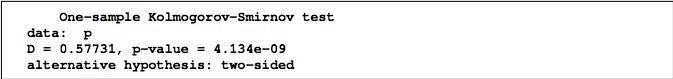
为此，我们设计了一个由样本计算的单一数值，即一个统计量。设计的诀窍是找到一个具有不依赖于我们未知事实（例如这种情况下的实际潜在分布）的取值范围的统计量。

Kolmogorov Smirnov检验中的统计量是非常简单的，它只是两个样本的经验累积分布函数之间的最大垂直距离。样本的经验累积分布是小于或等于给定值的样本值的比例。

一个样本的Kolmogorov Smirnov检验如下：



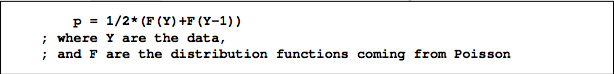
结果如下：



基于上面的结果我们能够较肯定地得出结论，这个模型是不适当的。

第四步－校正离散化泊松模型

因为泊松分布是离散的，所以我们需要做一个校正。改变如下：



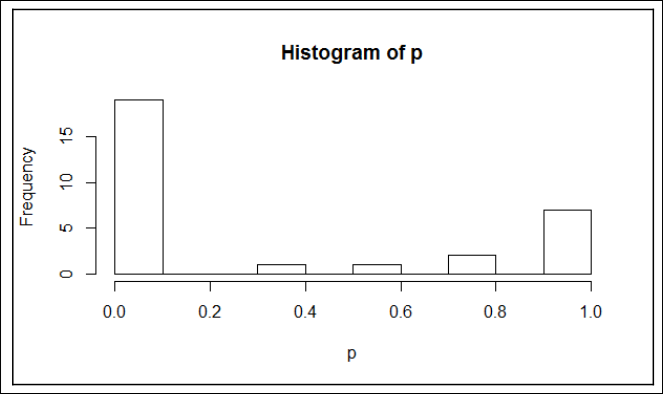
考虑离散分布，我们进行校正的步骤如下：



通过作图检查均运性如下：



作图结果如下图所示：

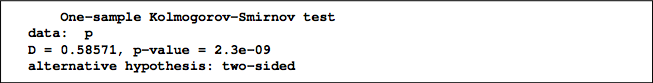


校正过程没有对结果产生不同的影响。作图结果清楚的显示p值分布不是平均的。

现在我们再次进行Kolmogorov-Smirnov检验来验证经验数据是否符合一个给定的分布，命令如下：



结果如下：



第五步－用链接方程训练和评估模型

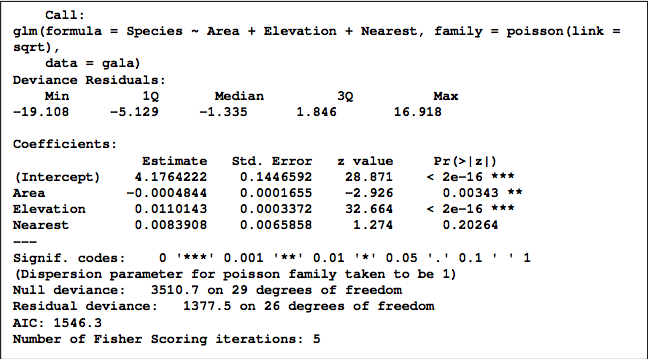
我们应该用glm()函数来观察广义线性模型拟合效果如下：



打印regpois2的命令如下：



结果如下：



第六步－用泊松模型再次评估

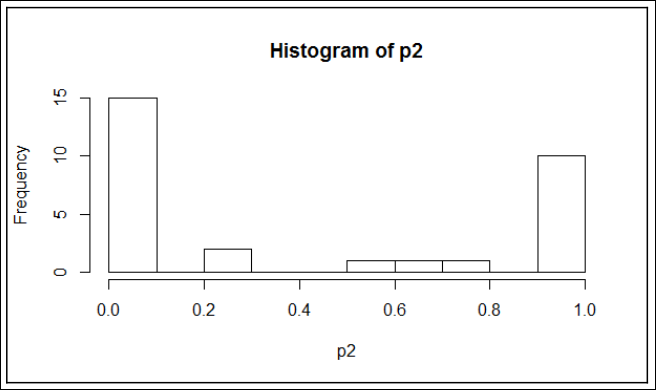
考虑离散分布的校正过程被实施如下：



通过作图检查均匀性如下：



作图结果如下图所示：



现在我们再次进行Kolmogorov-Smirnov检验来验证经验数据是否符合一个给定的分布如下：



一个Kolmogorov-Smirnov检测的样本如下：



结果仍然没有通过测试。

第七步－ 用线性模型重新评估

应用通用的线性模型：lm()函数用来拟合线性模型。该函数可用于回归，单层方差分析，和协方差分析（虽然aov可能提供一个对这些功能来说更加便捷的接口）。

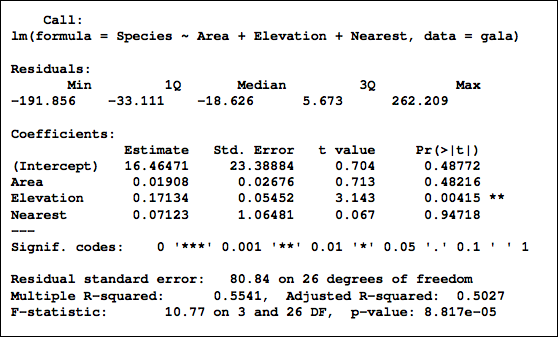
Reg数据帧用来存储lm()函数返回的结果，命令如下：



观察reg数据帧的结果，使用如下命令：



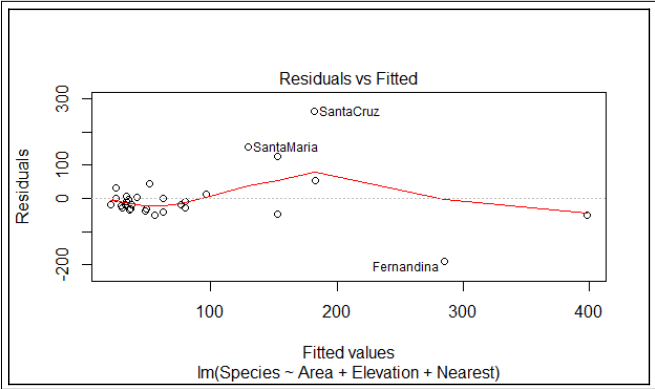
结果如下：



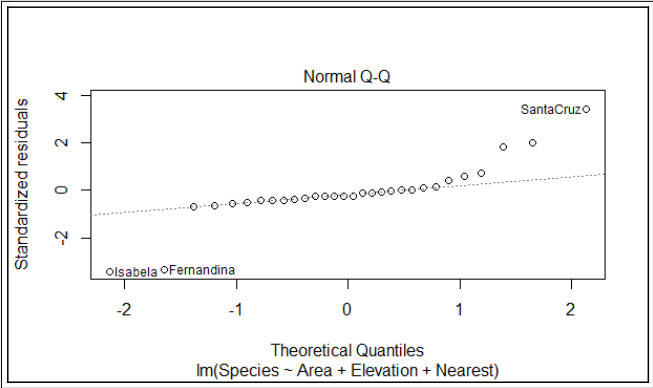
绘制reg数据帧如下：



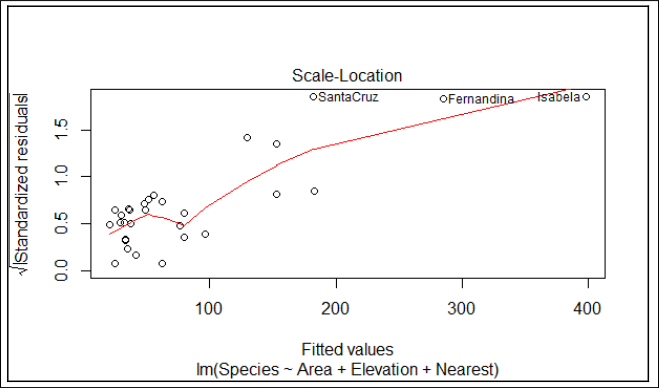
剩余残差VS拟合图如下：



正态Q－Q线性模型图如下：



线性模型位置尺度图如下所示：



现在我们使用下面的平方根函数进行一个变换。

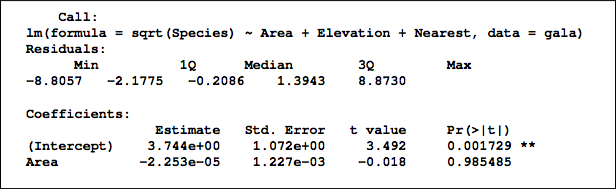
Reg2数据帧用来存储lm函数的结果：

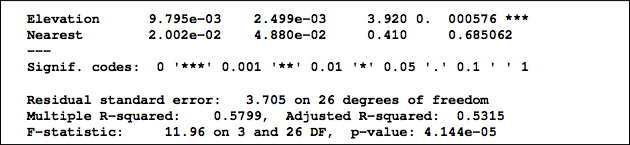


观察reg数据帧的结果如下：



结果如下：

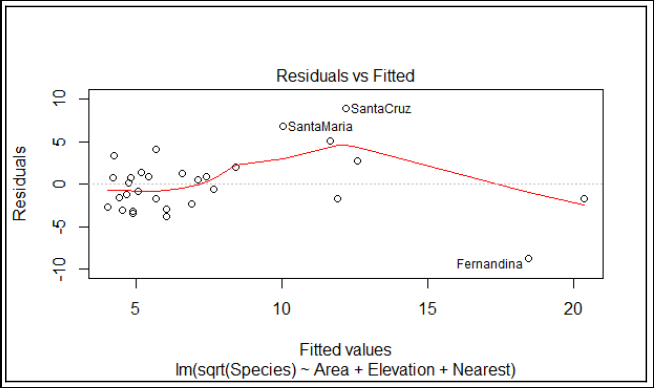




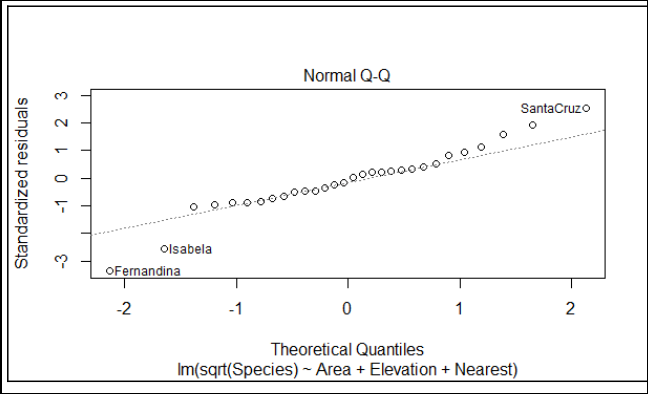
现在我们绘制reg2数据帧如下：



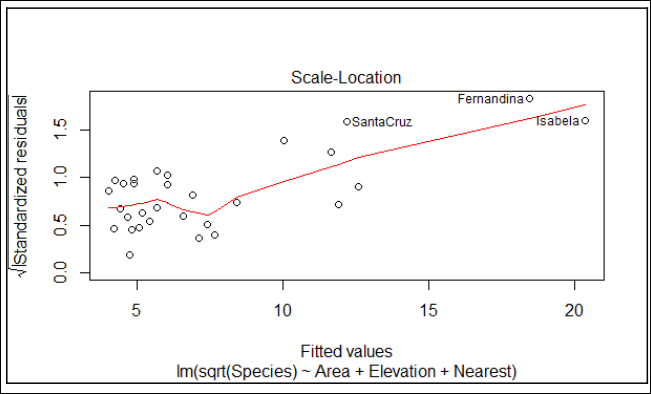
残差vs拟合图如下：



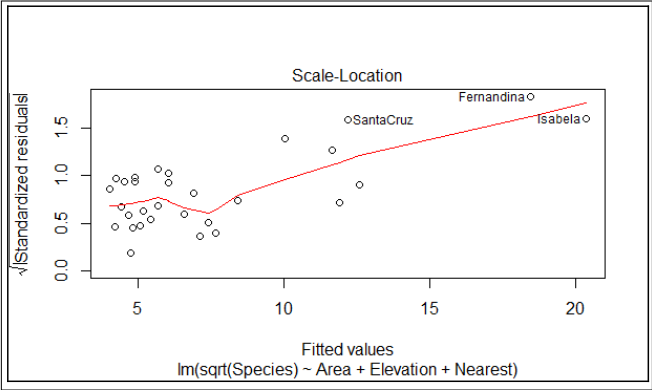
线性模型正态Q－Q图如下：



泊松回归位置尺度图如下：



线性模型位置尺度图如下：



让我们进行Shapiro检测。给定n个真实观测值的样本X1，……，Xn，Shapiro－Wilk检测（Shapiro 和Wilk, 1965）是一个针对数据是独立同分布和正态性的复合假设检验，即对对于未知的u和某个o>0的。命令如下：



结果如下：



现在我们使用log函数进行变换如下：

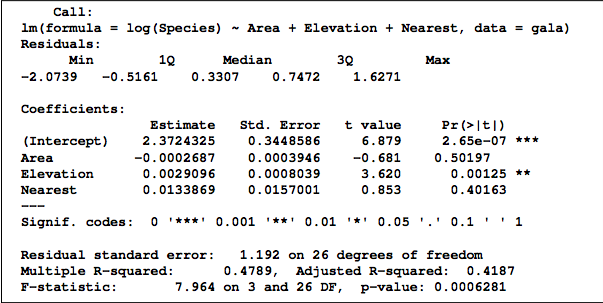
reg3数据帧被用来存储从lm()函数返回的结果如下：



观察reg3数据帧的结果如下：



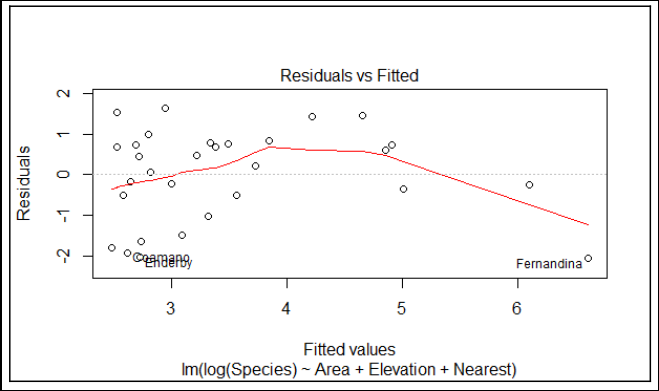
结果如下：



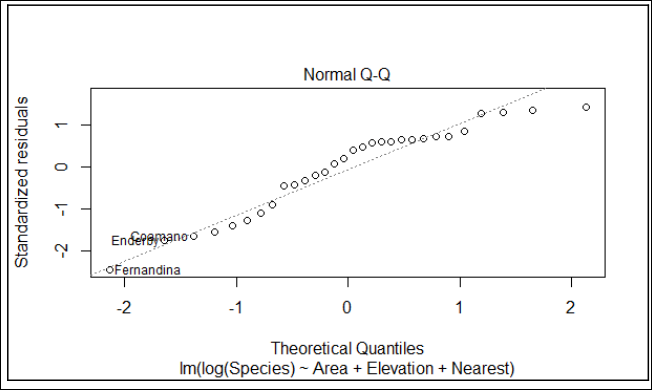
绘制reg3数据帧如下：



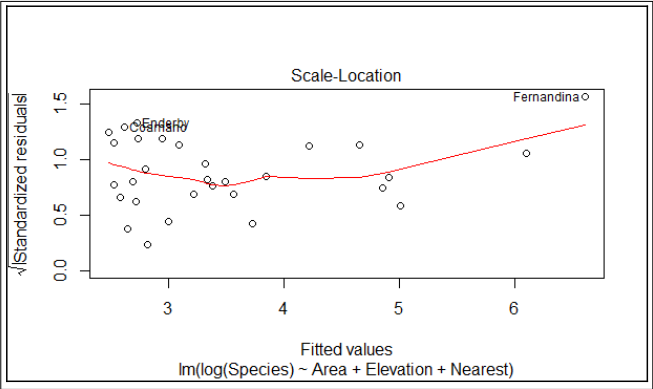
残差VS拟合图如下：



线性模型正态Q－Q图如下：



线性模型位置尺度图如下：



进行Shapiro检测如下：



结果如下：

