再看看相对重要性问题。模型所有的预测变量中，哪一个最重要，哪一个第二重要，哪一个最无关紧要？

正如你所看到的，我们会涵盖许多方面的内容。有效的回归分析本就是一个交互的、整体的、多步骤的过程，而不仅仅是一点技巧。为此，本书并不将它分散到多个章中进行讲解，而是用单独的一章来讨论。因此，这一章将成为本书最长最复杂的一章。只要坚持到最后，我保证你一定可以掌握所有的工具，自如地处理许多研究性问题！

## 8.1 回归的多面性

回归是一个令人困惑的词，因为它有许多特殊变种（见表8-1）。对于回归模型的拟合，R提供的强大而丰富的功能和选项也同样令人困惑。例如，2005年Vito Ricci创建的列表表明，R中做回归分析的函数已超过了205个（http://cran.r-project.org/doc/contrib/Ricci-refcardregression.pdf）。

表8-1 回归分析的各种变体

| 回归类型 | 用 途 |
|---|---|
| 简单线性 | 用一个量化的解释变量预测一个量化的响应变量 |
| 多项式 | 用一个量化的解释变量预测一个量化的响应变量，模型的关系是n阶多项式 |
| 多元线性 | 用两个或多个量化的解释变量预测一个量化的响应变量 |
| 多变量 | 用一个或多个解释变量预测多个响应变量 |
| Logistic | 用一个或多个解释变量预测一个类别型响应变量 |
| 泊松 | 用一个或多个解释变量预测一个代表频数的响应变量 |
| Cox比例风险 | 用一个或多个解释变量预测一个事件（死亡、失败或旧病复发）发生的时间 |
| 时间序列 | 对误差项相关的时间序列数据建模 |
| 非线性 | 用一个或多个量化的解释变量预测一个量化的响应变量，不过模型是非线性的 |
| 非参数 | 用一个或多个量化的解释变量预测一个量化的响应变量，模型的形式源自数据形式，不事先设定 |
| 稳健 | 用一个或多个量化的解释变量预测一个量化的响应变量，能抵御强影响点的干扰 |

在这一章中，我们的重点是普通最小二乘（OLS）回归法，包括简单线性回归、多项式回归和多元线性回归。OLS回归是现今最常见的统计分析方法，其他回归模型（Logistic回归和泊松回归）将在第13章介绍。

### 8.1.1 OLS 回归的适用情境

OLS回归是通过预测变量的加权和来预测量化的因变量，其中权重是通过数据估计而得的参数。现在让我们一起看一个改编自Fwa（2006）的具体示例（此处没有任何含沙射影之意）。

一个工程师想找出跟桥梁退化有关的最重要的因素，比如使用年限、交通流量、桥梁设计、建造材料和建造方法、建造质量以及天气情况，并确定它们之间的数学关系。他从一个有代表性的桥梁样本中收集了这些变量的相关数据，然后使用OLS回归对数据进行建模。

这种方法的交互性很强。他拟合了一系列模型，检验它们是否符合相应的统计假设，探索了所有异常的发现，最终从许多可能的模型中选择了“最佳”的模型。如果成功，那么结果将会帮助他完成以下任务。

- 在众多变量中判断哪些对预测桥梁退化是有用的，得到它们的相对重要性，从而关注重要的变量。
- 根据回归所得的等式预测新的桥梁的退化情况（预测变量的值已知，但是桥梁退化程度未知），找出那些可能会有麻烦的桥梁。
- 利用对异常桥梁的分析，获得一些意外的信息。比如他发现某些桥梁的退化速度比预测的更快或更慢，那么研究这些“离群点”可能会有重大的发现，能够帮助理解桥梁退化的机制。

可能桥梁的例子并不能引起你的兴趣。而我是从事临床心理学和统计的，对土木工程也是一无所知，但是这其中蕴含的一般性思想适用于物理、生物和社会科学的许多问题。以下问题都可以通过OLS方法进行处理。

- 铺路表面的面积与表面盐度有什么关系（Montogomery，2007）？
- 一个用户哪些方面的经历会导致他沉溺于大型多人在线角色扮演游戏（MMORPG；Hsu，Wen& Wu，2009）？
- 教育环境中的哪些因素与最能影响学生成绩得分？
- 血压、盐摄入量和年龄的关系是什么样的？对于男性和女性是相同的吗？
- 运动场馆和职业运动对大都市的发展有何影响（Baade & Dye，1990）？
- 哪些因素可以解释各州的啤酒价格差异（Culbertson & Bradford，1991）？（这个问题终于引起了你的注意！）

我们主要的困难有三个：发现有趣的问题， 设计一个有用的、可以测量的响应变量，以及收集合适的数据。

### 8.1.2　基础回顾

下面的几节，我将介绍如何用R函数拟合OLS回归模型、评价拟合优度、检验假设条件以及选择模型。此处假定读者已经在本科统计课程第二学期接触了最小二乘回归法，不过，我还是会尽量少用数学符号，关注实际运用而不是理论细节。有大量优秀书籍都介绍了本章提到的统计知识。我最喜欢的是John Fox的*Applied Regression Analysis and Generalized Linear Models*（偏重理论）和*An R and S-Plus Companion to Applied Regression*（偏重应用），它们为本章提供了主要的素材。另外，一份不错的非技术性综述可参考Licht（1995）。

## 8.2　OLS 回归

在本章大部分内容中，我们都是利用OLS法通过一系列的预测变量来预测响应变量（也可以说是在预测变量上“回归”响应变量——其名也因此而来）。OLS回归拟合模型的形式：