8.5　改进措施

我们已经花费了不少篇幅来学习回归诊断，你可能会问：“如果发现了问题，那么能做些什么呢？”有四种方法可以处理违背回归假设的问题：

- 删除观测点；
- 变量变换；
- 添加或删除变量；
- 使用其他回归方法。

下面让我们依次学习。

8.5.1　删除观测点

删除离群点通常可以提高数据集对于正态假设的拟合度，而强影响点会干扰结果，通常也会被删除。删除最大的离群点或者强影响点后，模型需要重新拟合。若离群点或强影响点仍然存在，重复以上过程直至获得比较满意的拟合。

不过，我对删除观测点持谨慎态度。若是因为数据记录错误，或是没有遵守规程，或是受试对象误解了指导说明，这种情况下的点可以判断为离群点，删除它们是十分合理的。

不过在其他情况下，所收集数据中的异常点可能是最有趣的东西。发掘为何该观测点不同于其他点，有助于你更深刻地理解研究的主题，或者发现其他你可能没有想过的问题。我们一些最伟大的进步正是源自于意外地发现了那些不符合我们先验认知的东西（抱歉，我说得夸张了）。

8.5.2　变量变换

当模型不符合正态性、线性或者同方差性假设时，一个或多个变量的变换通常可以改善或调整模型效果。变换多用Y^{λ}替代Y，λ的常见值和解释见表8-5。

表8-5　常见的变换

	−2	−1	−0.5	0	0.5	1	2
变换	$1/Y^2$	$1/Y$	$1/\sqrt{Y}$	$\log(Y)$	$\sqrt{Y}$	无	Y^2

若Y是比例数，通常使用logit变换[$\ln(Y/1-Y)$]。

当模型违反了正态假设时，通常可以对响应变量尝试某种变换。car包中的powerTransform()函数通过λ的最大似然估计来正态化变量X^{λ}。代码清单8-10是对数据states的应用。

代码清单8-10　Box-Cox正态变换

```
> library(car)
> summary(powerTransform(states$Murder))

bcPower Transformation to Normality
```

```
                Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
states$Murder         0.6     0.26             0.088               1.1

Likelihood ratio tests about transformation parameters
                       LRT df  pval
LR test, lambda=(0) 5.7   1 0.017
LR test, lambda=(1) 2.1   1 0.145
```

结果表明，你可以用Murder$^{0.6}$来正态化变量Murder。由于0.6很接近0.5，你可以尝试用平方根变换来提高模型正态性的符合程度。但在本例中，λ= 1的假设也无法拒绝（p=0.145），因此没有强有力的证据表明本例需要变量变换，这与图8-9的Q-Q图结果一致。

当违反了线性假设时，对预测变量进行变换常常会比较有用。car包中的boxTidwell()函数通过获得预测变量幂数的最大似然估计来改善线性关系。下面的例子为用州的人口和文盲率来预测谋杀率，对模型进行了Box-Tidwell变换：

```
> library(car)
> boxTidwell(Murder~Population+Illiteracy,data=states)

             Score Statistic p-value MLE of lambda
Population           -0.32    0.75          0.87
Illiteracy            0.62    0.54          1.36
```

结果显示，使用变换Population$^{0.87}$和Illiteracy$^{1.36}$能够大大改善线性关系。但是对Population（p=0.75）和Illiteracy（p=0.54）的计分检验又表明变量并不需要变换。这些结果与图8-11的成分残差图是一致的。

响应变量变换还能改善异方差性（误差方差非恒定）。在代码清单8-7中，你可以看到car包中spreadLevelPlot()函数提供的幂次变换应用，不过，states例子满足了方差不变性，不需要进行变量变换。

谨慎对待变量变换

统计学中流传着一个很老的笑话：如果你不能证明A，那就证明B，假装它就是A。（对于统计学家来说，这很滑稽好笑。）此处引申的意思是，如果你变换了变量，你的解释必须基于变换后的变量，而不是初始变量。如果变换得有意义，比如收入的对数变换、距离的逆变换，解释起来就会容易得多。但是若变换得没有意义，你就应该避免这样做。比如，你怎样解释自杀意念的频率与抑郁程度的立方根间的关系呢？

8.5.3 增删变量

改变模型的变量将会影响模型的拟合度。有时，添加一个重要变量可以解决我们已经讨论过的许多问题，删除一个冗余变量也能达到同样的效果。

删除变量在处理多重共线性时是一种非常重要的方法。如果你仅仅是做预测，那么多重共线性并不构成问题，但是如果还要对每个预测变量进行解释，那么就必须解决这个问题。最常见的方法就是删除某个存在多重共线性的变量（某个变量$\sqrt{\text{vif}} > 2$）。另外一个可用的方法便是岭回

归——多元回归的变体，专门用来处理多重共线性问题。

8.5.4　尝试其他方法

正如刚才提到的，处理多重共线性的一种方法是拟合一种不同类型的模型（本例中是岭回归）。其实，如果存在离群点和/或强影响点，可以使用稳健回归模型替代OLS回归。如果违背了正态性假设，可以使用非参数回归模型。如果存在显著的非线性，能尝试非线性回归模型。如果违背了误差独立性假设，还能用那些专门研究误差结构的模型，比如时间序列模型或者多层次回归模型。最后，你还能转向广泛应用的广义线性模型，它能适用于许多OLS回归假设不成立的情况。

在第13章中，我们将会介绍其中一些方法。至于什么时候需要提高OLS回归拟合度，什么时候需要换一种方法，这些判断是很复杂的，需要依靠你对主题知识的理解，判断出哪个模型提供最佳结果。

既然提到最佳结果，现在我们就先讨论一下回归模型中的预测变量选择问题。

8.6　选择“最佳”的回归模型

尝试获取一个回归方程时，实际上你就面对着从众多可能的模型中做选择的问题。是不是所有的变量都要包括？抑或去掉那个对预测贡献不显著的变量？还是需要添加多项式项和/或交互项来提高拟合度？最终回归模型的选择总是会涉及预测精度（模型尽可能地拟合数据）与模型简洁度（一个简单且能复制的模型）的调和问题。如果有两个几乎相同预测精度的模型，你肯定喜欢简单的那个。本节讨论的问题，就是如何在候选模型中进行筛选。注意，“最佳”是打了引号的，因为没有做评价的唯一标准，最终的决定需要调查者的评判。（把它看做工作保障吧。）

8.6.1　模型比较

用基础安装中的anova()函数可以比较两个嵌套模型的拟合优度。所谓嵌套模型，即它的一些项完全包含在另一个模型中。在states的多元回归模型中，我们发现Income和Frost的回归系数不显著，此时你可以检验不含这两个变量的模型与包含这两项的模型预测效果是否一样好（见代码清单8-11）。

代码清单8-11　用anova()函数比较

```
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost,
            data=states)
> fit2 <- lm(Murder ~ Population + Illiteracy, data=states)
> anova(fit2, fit1)

Analysis of Variance Table

Model 1: Murder ~ Population + Illiteracy
```