2013.4.24

第二十讲. 回归诊断 (I):
残差分析和Box-Cox变换

回归诊断 (regression diagnostics) 指的是使用模型分析数据之后，基于回归分析结果 (主要是残差)，判别模型拟合数据的好坏.

回归诊断主要包含两部分内容：残差分析和影响分析

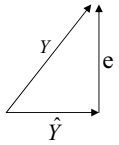
（1）残差分析主要对模型假设的合理性进行诊断；

（2）影响分析主要用于发现对回归分析结果影响较大的异常点。

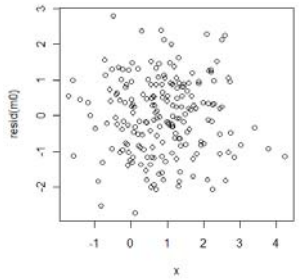
1. 回归诊断：残差分析

残差图：y-轴为残差，x-轴拟合值(或自变量)

通过残差图检查：
(i) 响应变量均值函数是否是线性
(ii) 误差方差是否为常数

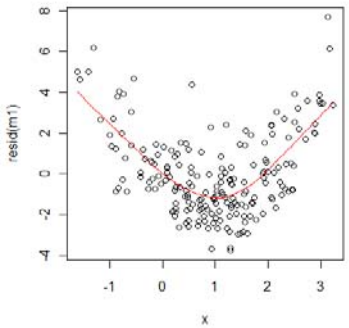


“好”的残差图:



无非线性趋势，
误差方差稳定

(i) 响应变量均值函数非线性



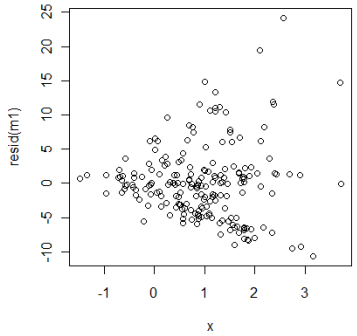
解决方案：

（1）增加x的高阶项、交互作用项

（2）变量变换，正态化。

（3）非线性回归或非参数回归

(ii) 误差方差不是常数



解决方案：

(1) 正态化变换（Box-Cox变换或方差稳定化变换）

(2) GLS（加权最小二乘）

(3) 广义线性回归模型(glm)

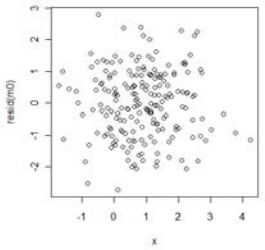
方差稳定化变换：若$\mu = E(Y), \sigma^2 = \mathrm{var}(Y)$,若$\sigma = \sigma(\mu)$,

则$Y \to \tilde{Y} = g(Y) = \int_0^Y \frac{1}{\sigma(\mu)} d\mu$的方差基本上为常数.

“好”的残差图并不代表模型假设完全正确，需要注意的是： 残差图无法检测出 “误差与自变量线性相关”

这是因为正则方程使用了线性不相关假设：

$$0 = \mathrm{X}'\varepsilon = X'(Y - X\beta)$$



“坏”的残差图表明模型假设不适用于当前数据，要么需要对模型做一些修正或使用其它模型，要么对数据做一些修正（最为常用的是Box-Cox 变换）

数据变换

- 数据变换总的原则是：变换后每个连续变量比较对称、均衡，换言之联合分布接近正态。常用变换包括
  - **Box-Cox变换**
  - 变量合并、加工
  - 连续变量离散化
  - 有次序的因子变量的连续化
  - 无次序因子变量的类别合并
  - **……**
- **log** 原则

  如果一个非负变量的取值不在一个尺度或量级**(magnitude)**上, 则取对数后分析可能是有益的. **log**变换是**Box-Cox**变换的一种主要形式.

2. Box-Cox变换

George E. P. Box

George Edward Pelham Box FRS (born 18 October 1919) is a statistician, who has made important contributions in the areas of quality control, time-series analysis, design of experiments, and Bayesian inference

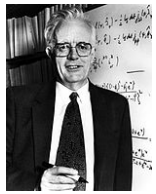
Contributions: ***Box–Cox transformations***, ***Box-Jenkins models, Box–Behnken designs***, ***robust statistics,*** etc.

Box famously wrote that "***essentially, all models are wrong, but some are useful***"

Son-in-law of Fisher.

Founder of Dept Stat, University of Wisconsin–Madison (1960).

David Cox



Sir **David Roxbee Cox** FRS, FBA (born 1924, Birmingham, England) is a prominent British statistician.

In 1966, he took up the Chair position in Statistics at Imperial College London where he later became head of the mathematics department. In 1988 he became Warden of Nuffield College and a member of the Department of Statistics at Oxford University

He has made pioneering and important contributions to numerous areas of statistics and applied probability, of which the best known is perhaps the ***proportional hazards model (Cox model)***, which is widely used in the analysis of survival data. The ***Cox process*** was named after him.

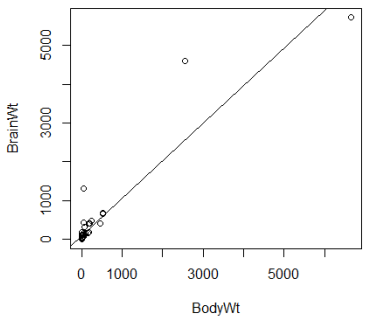
Box访问伦敦大学时，姓氏的相似使得他们觉得应该有所合作，选取的合作内容是变量变换，但两人兴趣的巨大差异使得该合作拖延了很久（～10年），直到1964年才得以发表

Box, George E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26 (2): 211–252.

该文所提出的变量变换方法称为Box-Cox变换。

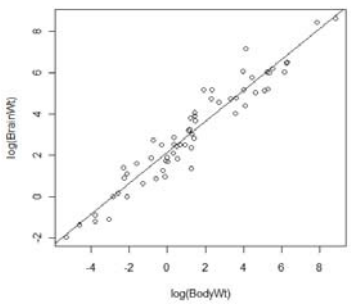
例1. 62种哺乳动物的脑重量与体重数据。

$$BrainWt = a + b \times BodyWt + \varepsilon$$



对两个变量取对数

$$\log(BrainWt) = a + b\log(BodyWt)$$



假设y为正的连续变量，Box - Cox变换为：

$$y \to y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0 \end{cases}$$

当响应变量和自变量联 合服从正态分布时，线 性模型假设成立. 基于此， Box - Cox变换的基本思想是变换 后的响应变量服从正态分 布：

$$y^{(\lambda)} \mid \mathbf{x} \sim \text{Normal}$$

对于自变量也可类似应用Box - Cox变换

$$x_k^{(\lambda)} \mid y, \mathbf{x}_{(-k)} \sim \text{Normal}$$

响应变量$\mathrm{Y} \to \mathrm{Y}^{(\lambda)} = (y_1^{(\lambda)},...,y_n^{(\lambda)})'$

假设存在某λ-变换，变换后满足正态模型：

$$\mathrm{Y}^{(\lambda)} = X\beta + \varepsilon,\ \varepsilon \sim N(0,\sigma^2 I_n) \Leftrightarrow \mathrm{Y}^{(\lambda)} \sim N(X\beta,\sigma^2 I_n)$$

$\mathrm{Y}^{(\lambda)} = (y_1^{(\lambda)},...,y_n^{(\lambda)})'$的联合密度函数为

$$f(\mathrm{Y}^{(\lambda)}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \| \mathrm{Y}^{(\lambda)} - X\beta \|^2\right)$$

所以数据$\mathrm{Y} = (y_1,...,y_n)'$的联合密度-似然函数为

$$\mathrm{L}(\lambda,\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \| \mathrm{Y}^{(\lambda)} - X\beta \|^2\right) \times \prod_{i=1}^{n} y_i^{\lambda-1}$$

log-似然函数$l(\lambda,\beta,\sigma^2) = \log \mathrm{L}(\lambda,\beta,\sigma^2)$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \| \mathrm{Y}^{(\lambda)} - X\beta \|^2 + (\lambda-1)\sum \log(y_i)$$

如何极大化$l(\lambda,\beta,\sigma^2)$，求解参数$\lambda,\beta,\sigma^2$的估计？

注意到：若λ给定，则β,σ^2的极大似然估计容易求得：

$$\hat{\beta}(\lambda) = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}^{(\lambda)}$$

$$\hat{\sigma}^2(\lambda) = \| \mathrm{Y}^{(\lambda)} - \mathrm{X}\hat{\beta}(\lambda) \|^2 / n = RSS(\lambda)/n$$

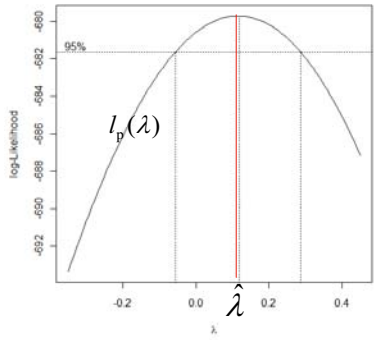
所以给定λ时的最大log-似然函数为

$$l_{\mathrm{p}}(\lambda) = l(\lambda,\hat{\beta}(\lambda),\hat{\sigma}^2(\lambda)) = \mathrm{C} - \frac{n}{2}\log RSS(\lambda) + (\lambda-1)\sum \log(y_i)$$

$l_{\mathrm{p}}(\lambda)$称为剖面似然函数，是单个参数λ的函数。

极大化剖面似然函数$l_{\mathrm{p}}(\lambda)$即可得到最优的λ的估计。

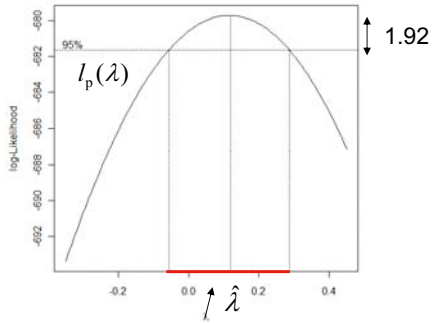
最简单的优化方法是逐点搜索，求得最优解：$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, l_{\mathrm{p}}(\lambda)$



但我们通常不一定取$\hat{\lambda}$，而是取其附近(95%置信区间内)的"容易解释"的一个λ值。比如$\hat{\lambda} = 1.819$，我们可能取$\lambda = 2$对应的平方变换.

具体地，λ的95%置信区间为：$\left\{\lambda : l_{\mathrm{p}}(\lambda) \geq l_{\mathrm{p}}(\hat{\lambda}) - 1.92\right\}$

从中选取“好”的λ。



λ的95%置信区间

这是因为近似地有 $2(l_{\mathrm{p}}(\hat{\lambda}) - l_{\mathrm{p}}(\lambda)) \sim \chi_1^2$

所以$0.95 = \mathrm{P}\left(2(l_{\mathrm{p}}(\hat{\lambda}) - l_{\mathrm{p}}(\lambda)) \leq 3.84\right) = \mathrm{P}\left(l_{\mathrm{p}}(\lambda) \geq l_{\mathrm{p}}(\hat{\lambda}) - 1.92\right)$

Box-Cox变换：应用于响应变量和连续型自变量

R： library(MASS)中的函数boxcox可确定最优的λ:

确定响应变量y的最优变换：$boxcox(y \sim x+z+...)$
确定自变量x的最优变换：$boxcox(x \sim y+z+...)$

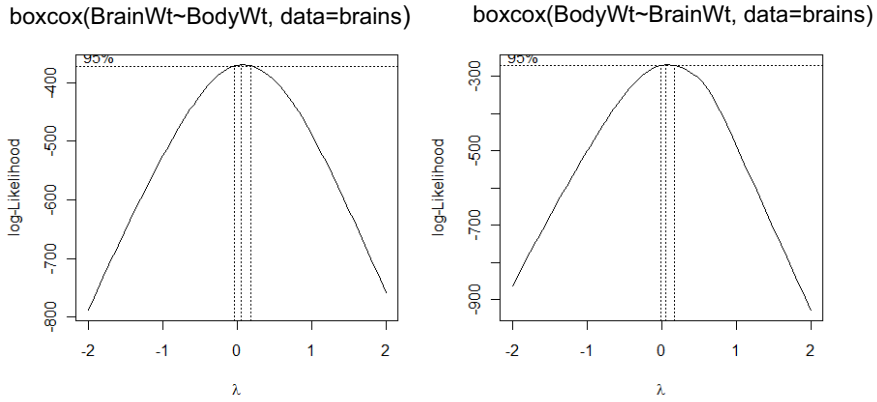
注1. 如果y取负值，对$y+c>0$做Box - Cox变换。

注2：下面的半参数模型容 许一般的未知变换 $f's$

Additive model： $y=\beta_0+f_1(x_1)+...+f_{p-1}(x_{p-1})+\varepsilon$

Single - index model： $y=\beta_0+f\ (\beta_1 x_1+...+\beta_{p-1}x_{p-1})+\varepsilon$

例1. 62种哺乳动物的脑重量与体重数据。



$\lambda \approx 0$, 对BrainWt做 log变换

$\lambda \approx 0$, 对BodyWt做 log变换

对两个变量取对数, LS拟合得回归直线:

$\log(BrainWt)=2.14+0.75\log(BodyWt) \Leftrightarrow$

$BrainWt=8.5\times BodyWt^{3/4}$