回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第24讲 残差分析：数据变换

2020.5.15

Box and Cox

含义: 轮流、互斥、小人物.

Two people who always miss each other and thus are never together.

来源：

The term comes from the comic play “Box and Cox - A Romance of Real Life in One Act”, by John Maddison Morton. This was first produced at the Royal Lyceum Theatre, London, in November 1847. Box and Cox were two lodges who shared their rooms - one occupying them by day and the other by night.

例句：

a Box and Cox arrangement.

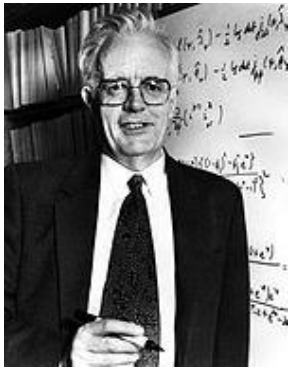
Since I've been on night work all week, Irene and I are like Box and Cox these days, constantly missing each other. I hope to actually spend time with her over the weekend!

In Statistics, Box and Cox are two prominent statisticians. They created “Box-Cox transform”

George E. P. 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.

- Son-in-law of Fisher.
- Founder of Dept Stat, University of Wisconsin–Madison(1960).
- Contributions: *Box–Cox transformations*, *Box-Jenkins models, Box–Behnken designs*, *robust statistics,* etc.

Sir David Roxbee Cox: FRS, FBA (born 1924, Birmingham, England) is a prominent British statistician. He worked in the Imperial College London, 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.

回归诊断

R^2度量了模型对数据的拟合程度，是一种整体上的拟合优度度量。拟合细节，主要是模型的合理性，还需通过分析残差进行诊断。

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，对模型拟合数据恰当与否进行分析，回归诊断主要包含两部分内容：

回归诊断：

- 残差分析：对模型假设的合理性进行诊断(线性、方差齐性)；
- 影响分析：发现对回归分析结果影响较大的点。

残差分析

模型假设误差项与自变量无关，故自变量-残差散点图上应没有任何明显的趋势关系。

因为自变量有多个, 故自变量-残差图有多个。

简化：以拟合值-替代所有自变量 (因为拟合值是自变量的最优组合)。

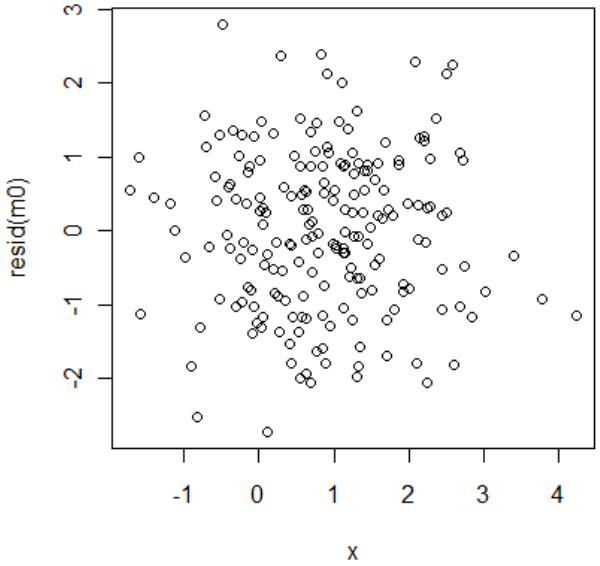
残差图：y-轴为残差，x-轴为拟合值。用于检查：

- 响应变量均值函数是否线性.
- 误差方差是否为常数.

当然，有时也需要研究每个自变量与残差的散点图

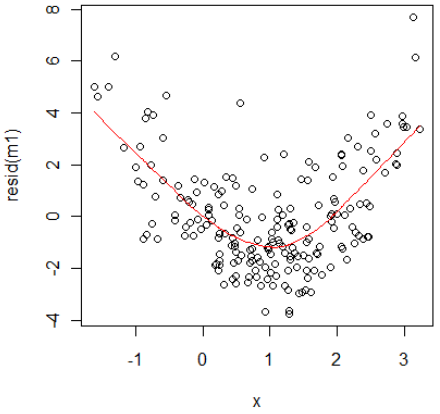
“好”的残差图:

无非线性趋势，误差方差稳定

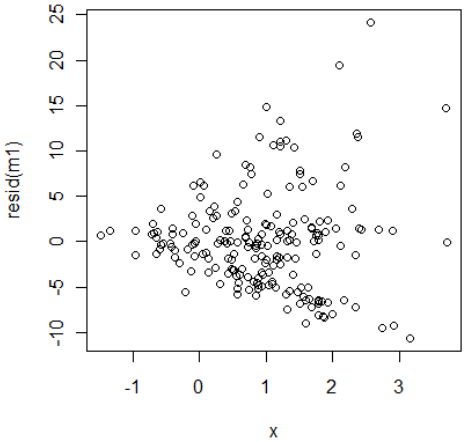


“不好”的残差图:

响应变量均值函数非线性



误差方差不是常数



问题解决方案：数据变换，或其它形式的回归

“非线性”问题的解决方案：

- 增加自变量的高阶项或非线性函数.
- 变换（Box-Cox变换或 方差稳定化变换）
- 其它回归（非线性回归或非参数回归）

“方差不齐”问题的解决方案：

- 变换（Box-Cox变换或 方差稳定化变换）
- GLS（加权最小二乘，但通常应用于方差结构已知的情况）
- 其它回归（广义线性回归模型，非线性回归）

数据变换

连续变量变换的一般原则：
变换之后的变量的分布对称、均衡，接近正态。

常用变换：

- 对数变换 (log-rule)：当正值变量取值不在一个量级/尺度上。
- Box-Cox变换：幂次变换，变换后服从正态分布。
- 方差稳定化变换：若方差是均值的函数，且函数形式已知。
- 离散变量的处理 (合并，计分…)

Box-Cox变换

事实：正态分布情形下线性模型假设成立。

基于此, Box - Cox变换的基本思想是：对于随机变量$y > 0$，寻找某种变换，使得变换之后的变量近似地服从正态分布。因为变换有无穷多，变换的范围限制在幂次变换。

Box - Cox变换：

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

选择λ，使得 $y^{(\lambda)}$ 的分布近似为 正态分布.

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

问题框架：

数据为$(y_i, x_i), i = 1,..., n,$对某个$\lambda \in R,$ 响应变量做Box - Cox：

$$\mathbf{y} = (y_1,..., y_n)^{\mathsf{T}} \to \mathbf{y}^{(\lambda)} = (y_1^{(\lambda)},..., y_n^{(\lambda)})^{\mathsf{T}}$$

假设变换后满足正态模型：

$$\mathbf{y}^{(\lambda)} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n) \Leftrightarrow \mathbf{y}^{(\lambda)} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)。$$

我们使用极大似然法（具体地，剖面似然法）求解最优的λ。

下面求解最优BC变换：对于给定的λ, $\mathbf{y}^{(\lambda)} = (y_1^{(\lambda)},..., y_n^{(\lambda)})^{\mathsf{T}}$的联合密度函数为

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

所以似然函数（即数据$\mathbf{Y} = (y_1,..., y_n)^{\mathsf{T}}$的联合密度）为：

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

$$l(\lambda, \beta, \sigma^2) = \log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \| \mathbf{y}^{(\lambda)} - X\boldsymbol{\beta} \|^2 + (\lambda - 1)\sum \log(y_i)$$

剖面似然法(profile likelihood)求解最优的λ-变换：

λ给定时,$\boldsymbol{\beta},\sigma^2$的极大似然估计有显式表达：

$$\hat{\boldsymbol{\beta}}(\lambda)=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}^{(\lambda)}$$

$$\hat{\sigma}^2(\lambda)=\|\mathbf{y}^{(\lambda)}-X\hat{\boldsymbol{\beta}}(\lambda)\|^2/n\hat{=}RSS(\lambda)/n$$

对于给定的λ，最大log-似然函数为

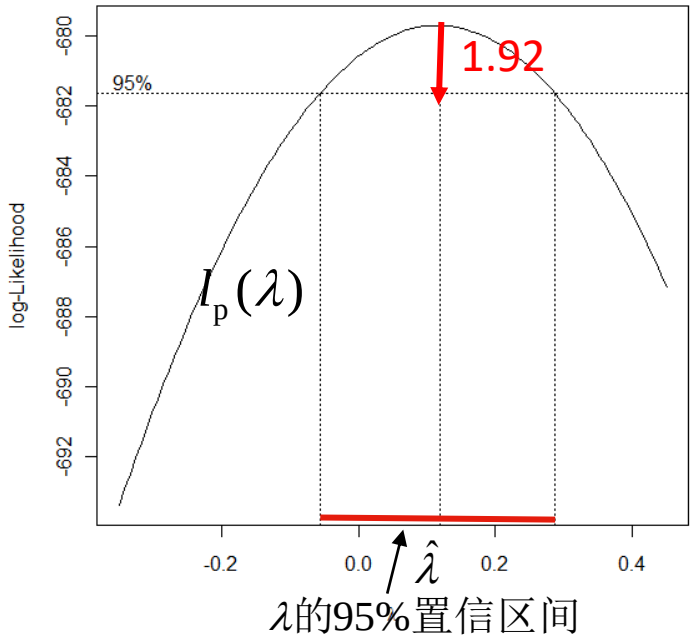
$$l_{\mathrm{p}}(\lambda)=l(\lambda,\hat{\beta}(\lambda),\hat{\sigma}^2(\lambda))=\mathrm{C}-\frac{n}{2}\log RSS(\lambda)+(\lambda-1)\sum\log(y_i)$$

$l_p(\lambda)$称为log-剖面似然函数，极大化$l_p(\lambda)$即可得到最优的λ的估计。

极大化$l_p(\lambda)$没有显式解，因为λ是实数，我们采用原始的逐点搜索极大化方法.

逐点计算$l_p(\lambda)$，$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, l_p(\lambda)$，如下图：

但我们通常不一定取极大点$\hat{\lambda}$，而是取其附近(95%置信区间内)的"容易解释"的一个λ值。比如$\hat{\lambda} = 0.61$，我们可以取$\lambda = 0.5$。



λ的95%置信区间

从最高点下拉1.92处的水平线与$l_p(\lambda)$的两个交点之间，构成λ的95%置信区间：

$$\left\{\lambda : l_p(\lambda) \geq l_p(\hat{\lambda}) - 1.92\right\}$$

这是因为似然比近似地服从卡方分布

$$2(l_p(\hat{\lambda}) - l_p(\lambda)) \sim \chi_1^2$$

$$\Rightarrow 0.95 = P\left(2(l_p(\hat{\lambda}) - l_p(\lambda)) \leq 3.84\right)$$

$$= P\left(l_p(\lambda) \geq l_p(\hat{\lambda}) - 1.92\right)$$

R： library(MASS)中的boxcox函数

自变量的Box-Cox变换

对于自变量也可类似应用Box - Cox变换

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

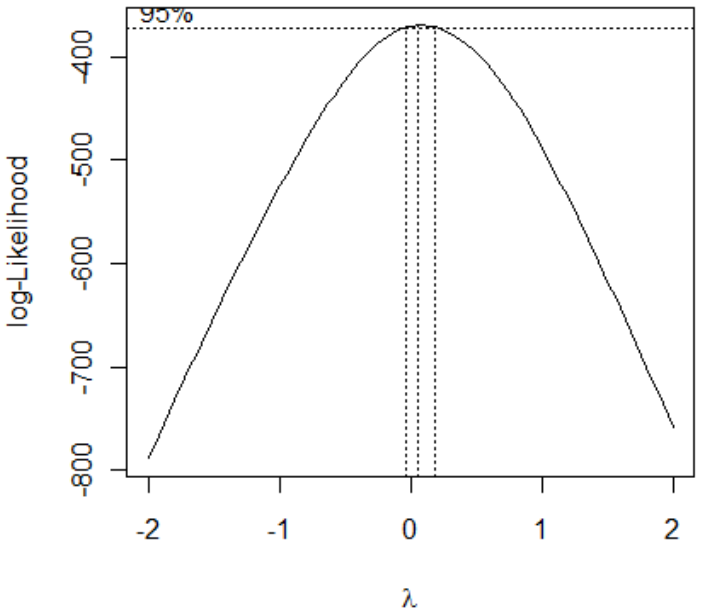
确定响应变量y的最优变换：$boxcox(y \sim x + z + ...)$

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

注. 如果y取负值，可对$y + c > 0$做Box - Cox变换。

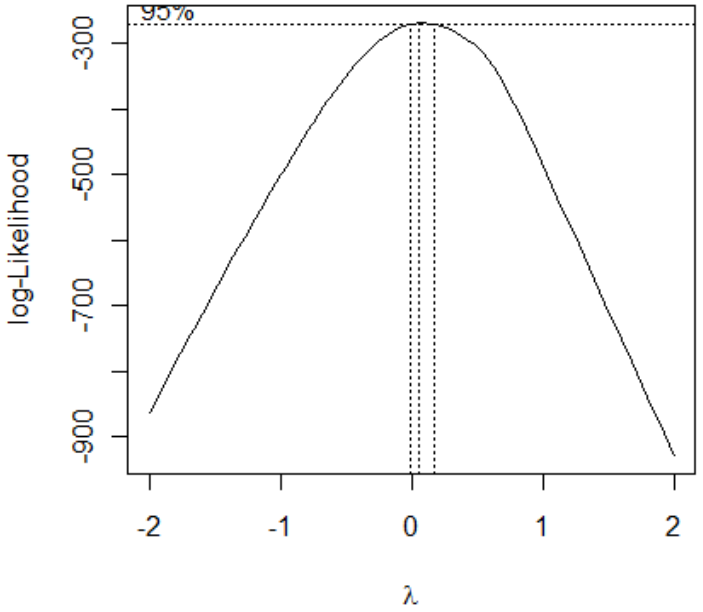
例1. 62种哺乳动物的脑重量与体重数据。

boxcox(BrainWt~BodyWt, data=brains)



$\lambda \approx 0$, 对BrainWt做 log变换

boxcox(BodyWt~BrainWt, data=brains)



$\lambda \approx 0$, 对BodyWt做 log变换

方差稳定化变换

正态分布$N(\mu,\sigma^2)$的方差μ与均值σ^2是两个无关的参数(样本均值$\bar{x}$和样本方差s^2独立!)。但对于非正态总体问题，中心极限定理(CLT)得到的渐近正态分布的渐近方差和渐近均值可能有关，比如

假设伯努利变量$x_1,...,x_n \ iid \sim B(1,p)$, 由CLT，$\frac{\sqrt{n}(\bar{x}-p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$, 近似地$(np>5$时$)$

比率/频率：$\bar{x} \overset{近似}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$,

计数：$X = n\bar{x} \sim B(n,p) \overset{近似}{\sim} N\left(np, np(1-p)\right)$

再如，泊松分布$X \sim Pois(\lambda)$, 则$\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1), \ \lambda \to \infty$

当$\lambda > 5$时，近似地$X \sim N(\lambda,\lambda)$.

伯努利变量$x_1,...,x_n \ iid \sim B(1,p)$, 若$p = p_n \to 0, np_n \to \lambda > 0$，则$X = \sum_{i=1}^{n} x_i \sim B(n,p) \to Pois(\lambda)$

如果响应变量的方差与均值存在某种已知关系, 为了满足线性模型方差齐性的要求，我们可以通过方差稳定化变换，将其变换为方差近似为常数（与均值无关）的变量：

方差稳定化变换：

设$\mu = E(Y)$，$\sigma^2 = \mathrm{var}(Y)$，若$\sigma = \sigma(\mu)$与$\mu$有关，则

$$\tilde{Y} \propto g(Y) = \int_0^Y \frac{1}{\sigma(\mu)} d\mu$$

称为方差稳定化变换（$\tilde{Y}$的方差近似地不依赖于其均值)

例2. (1) 比率数据(proportion data) 设$x \sim \mathrm{B}(n, p)$, 则

$$\mu = \mathrm{E}(x) = np，\ \sigma^2 = \mathrm{var}(x) = np(1-p) = \mu(1-\mu/n).$$

方差稳定化变换：

$$\mathrm{g}(x) \propto \int_0^x \frac{1}{\sigma(\mu)} d\mu = n\int_0^x \frac{1}{\sqrt{\mu(1-\mu/n)}} d\mu = 2\arcsin\sqrt{x/n}$$

略去常数2，可取x的方差稳定化变换为$\tilde{x} = \arcsin\sqrt{x/n}$。

结论：比率数据$\hat{p} = x/n$的方差稳定化变换为$\tilde{\mathrm{p}} = \arcsin\sqrt{\hat{p}}$

（2）计数数据。设$x \sim \mathrm{Poisson}(\lambda)$, 则$\sigma^2 = \mathrm{var}(x) = \lambda$与均值$E(x) = \lambda$有关。

变换：$\mathrm{g}(x) \propto \int_0^x \frac{1}{\sigma(\lambda)} d\lambda = \int_0^x \frac{1}{\sqrt{\lambda}} d\lambda = 2\sqrt{x}$

略去常数可取x的方差稳定化变换为$\tilde{x} = \sqrt{x}$

(3) 样本相关系数. 近似地样本相关系数 $r \sim N(\rho,(1-\rho^2)^2/n)$

则近似地有 $E(r)=\rho$， $\sigma^2=\mathrm{var}(x)=(1-\rho^2)^2/n$,

$$g(r)\propto\sqrt{n}\int_0^r\frac{1}{(1-\rho^2)}d\rho=\sqrt{n}\log\left(\frac{1+r}{1-r}\right),$$

第三讲命题2

样本：$(x_1,y_1),...,(x_n,y_n)$ iid, 设ρ为总体相关系数,r为样本相关系数，则

$\sqrt{n}\,(r-\rho)\xrightarrow{d}N\left(0,(1-\rho^2)^2\right)$, 当$n\to\infty$

样本相关系数r的方差稳定化变换为：

$$\tilde{r}=\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)=\mathrm{atanh}(r)$$

称为r的Fisher变换。

应用例子：

基于近似分布：$r \sim N(\rho,(1-\rho^2)^2/n)$

可以如下构造ρ的95%置信区间：$\left\{\rho: \left|\frac{\sqrt{n}(r-\rho)}{1-\rho^2}\right| \le 1.96\right\}$,

但该置信区间表现不是太好（因为r的方差与均值有关）。

基于*Fisher*变换，ρ的95%置信区间可以构造为：

$$\left\{\rho: \left|\sqrt{n}\left\{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)-\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)\right\} \le 1.96\right|\right\}$$

该区间比上面的区间具有更好的性质（覆盖率更精确地接近0.95,长度更短）。

Delta-方法与方差稳定化变换的推导

*Delta*方法：

若$\sqrt{n}(X_n-\theta)\xrightarrow{d}N(0,\sigma^2)$,则$\sqrt{n}(g(X_n)-g(\theta))\xrightarrow{d}N(0,[g'(\theta)]^2\sigma^2)$,
$n\to\infty$,其中假设$g'(\theta)$存在且非0.

证明：泰勒展开，近似地：$\sqrt{n}(g(X_n)-g(\theta))\approx\sqrt{n}g'(\theta)(X_n-\theta)$。
(*Delta*方法在统计中应用非常广泛!)。

推导方差稳定化变换：在Delta方法中,如果X_n的渐近方差$\sigma^2=\sigma^2(\theta)$
与均值θ有关，那么为了使$g(X_n)$的渐近方差与θ无关，只需：

$$[g'(\theta)]^2\sigma^2(\theta)=C\text{（常数）}$$

解方程得：$g(\theta)\propto\int\frac{1}{\sigma(\theta)}d\theta$,称为方差稳定化变换

其它变换

- 变量合并：多个变量合并加工为一个少数几个变量；

  比如：血压研究中height, weight合并为BMI?

  政策研究中 各区GDP，Pop(总人口)合并为人均GDP？

  因子变量相近的水平合并为一个水平？

- 连续变量离散/因子化：将含义不太具体明确的连续变量转化为离散变量甚或因子变量；

  比如：Day(日期)转化为月份或季节？

  血压值BP转化为高、中、低血压？

  百分制成绩转化为5分制？

- 因子变量连续化：有次序的因子变量通过打分转化为连续变量；

  比如：血压BP取值为高、中、低打分为4,2,1?

  职称Rank的三种职称(助理、副、正教授)打分为1,3,4?

  注意：打分最好由专业人士提供