

内容: Box-Cox 变换、回归诊断 (残差分析和影响分析)、WLS 和 IRLS、lowess 拟合
任务: 阅读例 1-2, 完成练习 1-4 (P6-8) 和/或练习 5 (P10)。

1 Box-Cox 变换

例 1. 程序包 `alr4` 数据集 `brains` 给出了 62 种哺乳动物的平均脑重 (g) 和平均体重 (kg),

- (a) 拟合线性模型 $\text{BrainWt} = \beta_0 + \beta_1 \text{BodyWt} + \epsilon$, 画出残差图 (R 命令: `plot(myfit, which=1)`, 其中 `myfit` 为 `lm` 的输出结果, `which=1` 指定画残差图)

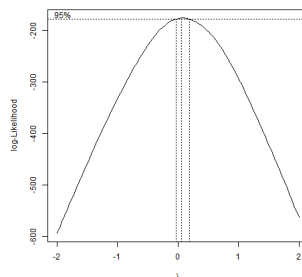
```
> myfit=lm(BrainWt~BodyWt, data=brains)
> plot(myfit,which=1) #residual plot
```

考察拟合效果 (残差是否接近正态分布, 误差方差是否为常数?)。

- (b) 考虑对响应变量 `BrainWt` 做 Box-Cox 变换, 应用 `library(MASS)` 中的函数 `boxcox` 求出变换:

```
library(MASS)
boxcox(BrainWt~BodyWt, data=brains)
```

输出曲线是对数剖面似然, 最大值点即是最优幂次 λ , 这里最优 $\lambda = 0$, 即对响应做对数变换。



- (c) `BrainWt` 变换之后重新拟合模型。结果可能仍不令人满意, 主要问题可能是自变量不均衡对称。所以考虑对自变量做 `boxcox` 变换:

```
boxcox(BodyWt~BrainWt, data=brains) # or log(BrainWt)
```

由此得到对自变量所应该进行的变换 (也是对数变换)。再次做残差分析观察拟合情况, 拟合效果令人满意 (参见第 14 讲例 1)。

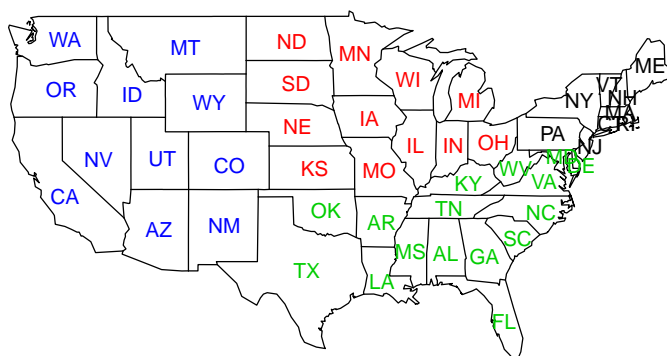
2 回归诊断：残差分析和影响分析

回归诊断通过残差分析判断模型假设的合理性，通过残差分析和影响分析发现高影响数据点。主要工具是残差图。发现问题后，解决问题的主要工具是 Box-Cox 变换。

例 2. 数据集<http://staff.ustc.edu.cn/~ynyang/2025/lab/edu.xls> 给出了 1975 年美国 50 个州的青少年教育花费数据, 变量解释如下

变量	描述
Expenditure	各州年度人均教育费用
Income	各州人均收入
Young	18 岁以下人口比例
Urban	城市人口比例
Region	地区, 1: 东北, 2: 中部和北部, 3: 南部, 4: 西部

```
> edu=read.table("http://staff.ustc.edu.cn/~ynyang/2025/lab/edu.xls",
+               head=T,row.names=1)
> install.packages("maps") #安装地图软件包
> library(maps)
> map("state")
> text(state.center, state.abb)
```



注意数据最后一列 Region 取值 1, 2, 3, 4 并不是实数或整数，他们分别代表东北，中北，南部和西部，所以首先需要将 Region 定义为因子 (factor):

```
> edu[,"Region"]=as.factor(edu[,"Region"])
> edu[,"Region"]
 [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 4
[39] 4 4 4 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
#
```

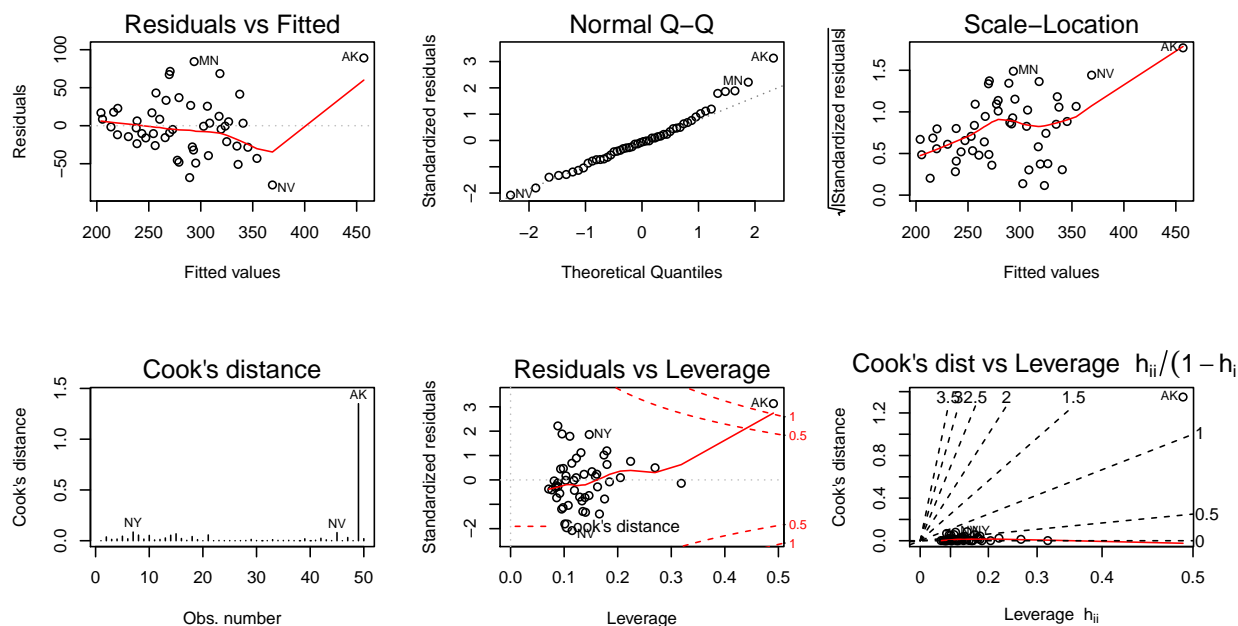
我们关心的是教育花费与其它变量的关系。假设回归模型

$$\text{Expenditure}_i = \beta_0 + \beta_1 \times \text{Income}_i + \beta_3 \times \text{Urban}_i + \sum_{k=2}^4 \alpha_k I(\text{Region}_i=k) + \epsilon_i, \epsilon_i, i = 1, \dots, 50 \text{ iid } \sim (0, \sigma^2)$$

回归诊断图： R 命令 `plot(lm.object, which=)` 绘出回归诊断图 (包括残差分析和影响分析, 共六个)。其中的选项 `which` 选择绘出哪几个图。缺省地, `which=c(1,2,3,5)`, 如果只需要第一个, 也即残差图, 可指定 `which=1`。

```
> fit1 = lm(Expenditure ~. , data=edu)
# 这里 '.' 代表除了Expenditure之外所有其它变量
> plot(fit1, which=1:6) #回归诊断图
# 注意这里实际上是调用的 plot.lm, 如果需要改变绘图设置, 需要查询帮助:
> ?plot.lm
# 特别地, plot.lm的帮助文档解释了scale-location散点图。
```

所有六个图如下:



各图含义分别解释如下:

图 1: 残差图, 横坐标为拟合值 $(\text{location})\hat{y}_i$, 纵坐标为残差 $e_i = y_i - \hat{y}_i$ 。图中的红色曲线是拟合值-残差的 lowess 拟合 (参见第 4 节), 如果该曲线在 0 点处平行于 x 轴, 则表明线性假设是合理的, 否则表明有一定的非线性趋势, 线性模型拟合欠佳。

从该图可以看到方差随拟合值增大而增大, 误差方差不是常数。AK 的残差为正数且异常 (即 AK 的响应变量 Expenditure 异常, 偏大)。

图 2: 残差的 qqnorm 图, 检查标准化残差

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}, i = 1, \dots, n$$

是否服从正态分布。

该图表明本数据误差基本服从正态分布（如果不呈一条直线，则表明误差非正态）。

图 3: scale-location 图，也称为 spread-location 图，注：一般情况下（未必限于回归问题），location 指的是均值、中位数等统计量，而 scale 或 spread 指的是与刻度、分散程度有关的统计量，比如标准差、极差（极大值与极小值的差）、IQR（inter-quantile, 75%, 25% 分位数之差）等。在残差分析的图 3 中，横坐标为拟合值 \hat{y}_i (location), 纵坐标为 $\sqrt{|r_i|}$ (scale), 主要用于检查方差 (scale) 齐性假设，即如果 Gauss-Markov 假设成立，则 $\sqrt{|r_i|}$ 应近似服从正态，且与自变量无关。

该图与图 1 反映出类似的问题，即方差不齐，AK 的残差异常（即响应变量异常），并有较明显的非线性趋势。

图 4: Cook 距离，横坐标为数据点编号 i (obs number), 纵坐标为 Cook 距离 D_i ,

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \times r_i^2 / p$$

该图表明 AK 的 Cook 距离很大，AK 是高影响点。

图 5: 残差-杠杆图，横坐标为杠杆 h_{ii} , 纵坐标为标准化残差 r_i , 两条红色虚线分别为 Cook 距离 $D = 0.5$ （影响较大）和 $D = 1$ 的等高线（影响很大）。

该图表明 AK 的 Cook 距离大于 1，其残差较大，leverage 也较大，即 AK 的响应变量和自变量都比较异常，是高影响点。

图 6: Cook 距离-杠杆图，横坐标为杠杆 h_{ii} , 纵坐标为 Cook 距离 D_i 。虚线为 $D_i/(h_{ii}/(1 - h_{ii})) = r_i^2/p$ 的等高线。

所有数据点都在等高线 $D_i/(h_{ii}/(1 - h_{ii})) = r_i^2/p = 1$ 下面，表明所有标准化残差 $|r_i| \leq \sqrt{p} = \sqrt{6}$ 。此外，AK 的 Cook 距离 D 和 leverage h_{ii} 都较大，高影响。

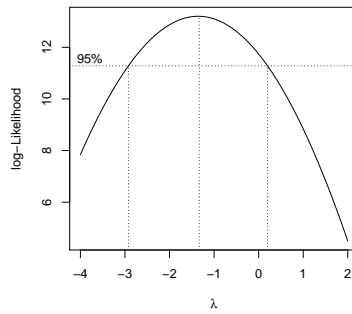
使用函数 `rstandard`, `hatvalues`, `cooks.distance`, `dffits`, `dfbetas` 可得到诸影响度量, `influence.measures` 给出所有度量。查看 AK（阿拉斯加），HI（夏威夷）的影响度量：

```
> influence.measures(fit1)
dfb.1_ dfb.Incm dfb.Yong dfb.Urbn dfb.Rgn2 dfb.Rgn3 dfb.Rgn4 dffit cov.r cook.d hat inf
CA 0.0179 2.3e-03 -0.0289 0.01505 9.6e-03 0.01376 0.0380 0.0611 1.40 5.5e-04 0.158
AK -2.2864 2.4e+00 2.0295 -1.74712 -7.2e-01 0.22849 0.0727 3.4571 0.38 1.3e+00 0.491 *
HI 0.0743 -8.0e-02 -0.0200 -0.07853 -1.2e-02 -0.06073 -0.1857 -0.3757 1.06 2.0e-02 0.098
```

AK 的影响比较大，HI 的影响不大。AK 和 HI 都在美国本土之外，但 AK 可能更特殊，尤其是它的自变量比较异常（杠杆值 $h_{ii} = 0.491$ 远远大于其它各州）。鉴于 AK 的特殊性（不在本土以及气候原因），我们可以考虑删除 AK。

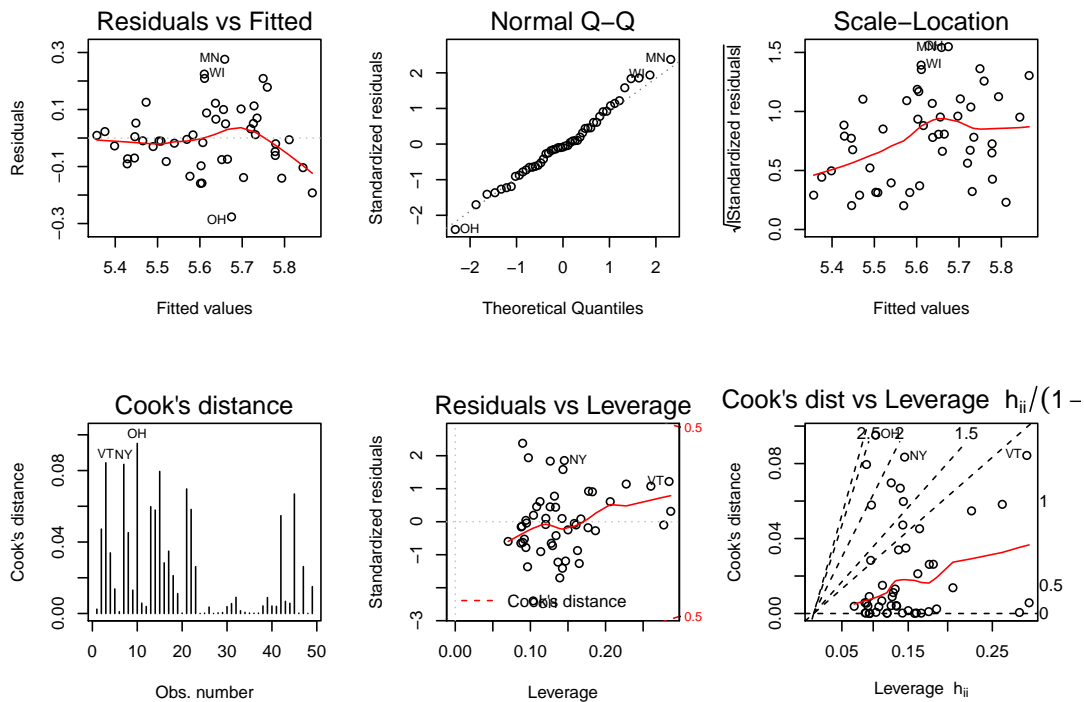
►Tip: 除非有特别的理由，不要轻易删除数据！

删除 AK 之后重新拟合，得到的回归诊断图与前面的类似，特别地，第一、三个图（残差图）都表明存在非线性和异方差现象。这说明高影响点 AK 并不是导致非线性或异方差的原因。下面试图做变量变换，观察是否能得到更好的拟合。对 Expenditure, Income 做 Box-Cox 变换：



Expenditure 的 BC 变换 λ 的置信区间为 $[-3, 0]$, 我们（暂时）选取对数变换, Income 也做对数变换。

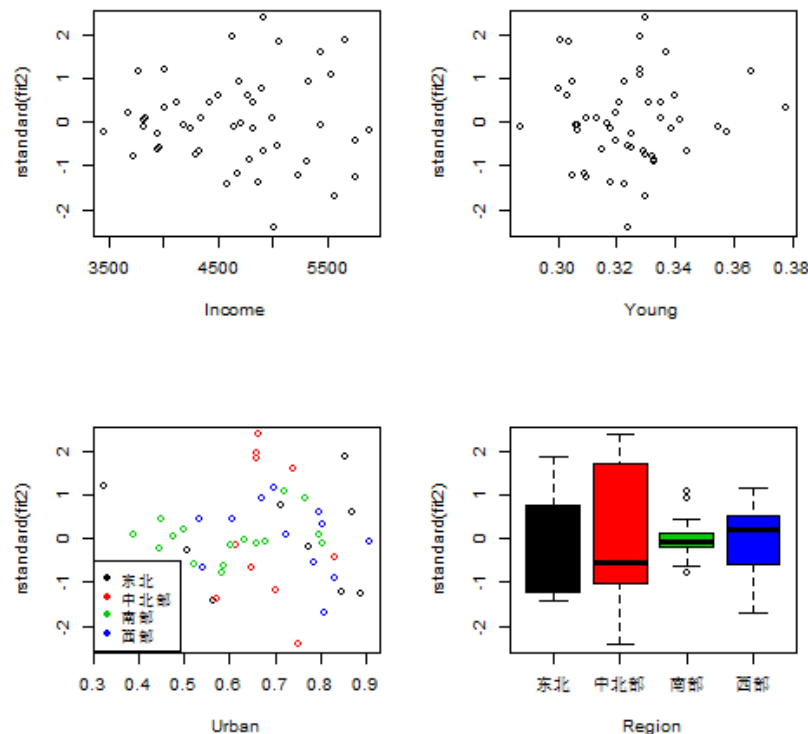
```
fit2=lm(log(Expenditure)~log(Income)+Young+Urban+Region,data=edu)
par(mfrow=c(2,3))
plot(fit2,1:6 )
```



通常, 响应变量和/或自变量的 Box-Cox 变换在很多问题中会消除或部分地消除残差的非线性和异方差现象, 从而完成数据分析过程。但本例比较特殊, 上图表明异方差现象和非线性现象在 BC 变换之后仍旧存在。进一步观察自变量-残差图, 研究残差与各个自变量的关系 (我们将发现异方差现象可能是因为四个地区的误差方差不同导致的)。

►Tip: 我们认为拟合值是所有自变量的最佳组合, 可以在某种程度上代替所有自变量, 所以 R 软件的残差图的横坐标是拟合值。为了更精细地考察残差, 有时, 我们也需要画出残差-自变量散点图, 考察残差与每个自变量是否存在非线性关系或方差不齐现象。

下面四个图是自变量为横轴，标准化残差为纵轴的散点图或盒型图。从图中可以看出，随 Income 增大残差方差有增大的趋势，但不太明显（第 1 图）。在 4 个地区 (region) 有较大的差异（第 4 图，盒型的高度代表方差），注意 Expenditure 是各州人均教育花费，其方差应该与各州总人口的倒数成正比，四个大区的人口差别较大，从而造成四个区的方差不同。针对这一现象，我们可以假设不同地区的误差方差不同，进而应用广义最小二乘法拟合（参见下节例 2(续)）。



►Tip: 盒型图 boxplot 的矩形高度称为 IQR(inter-quartile range), 其大小代表了数据的分散程度。如果数据是正态分布, 那么方差 $\sigma^2 = IQR^2/1.82 \propto IQR^2$

练习题

1. alr4 数据集 *fuel2001* 是美国 2001 年 51 个州的汽车汽油消耗量数据，变量如下

变量	描述
Drivers	持有驾照的人数
FuelC	汽车汽油销售总量（单位：1000 加仑）
Income	2000 年人均收入
Miles	该州内国有高速公路里程数（单位：英里）
MPC	人均驾驶里程数估计值（单位：英里/人）
Pop	16 岁以上人口数目
Tax	汽油州税（单位：加仑/美分）

本问题的目标是研究州税高的州是否汽油消耗较低。

提示：响应变量是什么？响应和自变量是否需要变换？检查有无高影响点或异常点，如果有高影响

的州，解释为什么（即高影响的州有什么特点，为什么是高影响的）？是否有足够的理由剔除高影响点？变量 Tax 显著吗？

2. **(Word Cities)** alr4 数据集 *BigMac2003* 给出了 2003 年世界上 70 个城市的数据 (变量如下表), 关心的是 BigMac (巨无霸汉堡包) 价格与其它因素的关系。

Variable	Description
BigMac	Minutes of labor to buy a Big Mac hamburger based on a typical wage averaged over 13 occupations
Bread	Minutes of labor to buy 1 kg bread
Rice	Minutes of labor to buy 1 kg of rice
Bus	Lowest cost of 10 km public transit
FoodIndex	Food price index, Zurich=100
TeachGI	Primary teacher's gross annual salary, thousands of US dollars
TeachNI	Primary teacher's net annual salary, thousands of US dollars
TaxRate	$100 \times (TeachGI - TeachNI)/TeackGI$. In some places, this is negative, suggesting a government subsidy rather than tax
TeachHours	Teacher's hours per week of work
Apt	Monthly rent in US dollars of a typical three-room apartment

- (a) 画出 *FoodIndex-BigMac* 散点图，找出恰当的变换使得两者关系接近线性（注意 BigMac 值特别大的两个城市可能会影响变换的选择，可考虑去掉这两个城市的数据）。
- (b) 发现合适的自变量的变换，并应用线性回归模型分析该数据。
3. 医疗保险计划 (health plan) 常包含多种方法来减少购买处方药的花费。其中两种常用的策略是 (1) Generic Substitution (GS)，要求医生开具过了专利保护期的传统药品而不是新上市的药（新药价格较高）；(2) Restrictiveness，限制医生开药的范围，例如对某症状如果有三种等效药物，那么要求医生只开其中一种（同一药物的购买量大则可以得到部分优惠），限制的严厉程度以 Restrictiveness Index (RI) 衡量，取值在 0-100 之间。alr4 数据集 *drugcost* 是美国 90 年代中期一个保险公司的 29 种医疗保险计划数据。变量描述如下

变量	描述
COST	响应变量，平均每天每次开处方医疗保险计划的支付，单位：美元
RXPM	医疗保险成员每年开药的平均次数
GS	保险计划中使用 GS 的百分比
RI	限制开同一种而不是多种等效药的限制强度 (0~100, 0: 没有, 100: 全部都要求)
COPAY	除了保险计划支付外的医疗保险成员所需的平均支付 (coypay, 比如挂号费)
AGE	平均年龄
F	女性比例
MM	所有会员的保险期限总和，体现保险计划的大小 (size)

主要关心的是 *GS* 和 *RI* 方法能否有效地减少保险计划的支付，但在分析 *COST* 与这两个变量的关系时，可能需要控制其它变量。使用线性模型分析该数据，诊断结果。

4. **(人工降雨, cloud seeding)** 为了研究人工降雨的有效性, 1975 年夏天在美国佛罗里达州 3000 平方英里的区域上空进行了试验。因为不是每天都适合人工降雨, 所以根据数学模型指标 *S* 是否大

于 1.5 来决定合适的日期, 共有 24 天 $S > 1.5$ 适合人工降雨。在这 24 天中, 每天通过抛均匀硬币的方式决定是否进行试验, 共有 12 天被选作试验日期, 通过飞机在云层中抛洒植入 (seeding) 碘化银的方式进行人工降雨, 其余 12 天不实施人工降雨。结果在 alr4 数据集 *cloud* 中, 变量描述如下:

Variable	Description
<i>A</i>	Action, 是否实施人工降雨 (1 = 实施人工降雨, 0 = 不实施)
<i>D</i>	Days, 第一次实施人工降雨 (1975 年 6 月 16 日, $D=0$) 之后的天数
<i>S</i>	Suitability for seeding, 度量是否适合进行人工降雨的数学模型指标
<i>C</i>	Cover, 试验区域云层覆盖率
<i>P</i>	Pre-wetness, 人工降雨之前 1 小时的降雨量 (单位: 10^7 立方米)
<i>E</i>	Echo motion category, 云层类型 (类别 1 或 2)
<i>Rain</i>	实施人工降雨之后的降雨量 (单位: 10^7 立方米)

本问题的目标是分析人工降雨的有效性 (即 *A* 与 *Rain* 是否存在显著的因果关系)。注意到这是一个随机化控制试验, 原则上只需要研究 *A* 与 *Rain* 的关系即可, 但因为只有 24 天的试验日期, $A = 1$ 的 12 天与 $A = 0$ 的 12 天之间在其它因素上可能还是有系统性差异的 (你可以考察 *A* 与其它变量是否相关), 为此可能需要在研究 *A* 与 *Rain* 的关系时控制其它因素, 这称为协方差分析 (即针对试验数据的、既有 treatment 因子变量也有连续控制变量的多变量回归分析)。试分析人工降雨是否有显著效果 (思考: 如何恰当处理变量 *D* 或许是一个关键, *D* 的具体数值是否有实数大小含义? 似应将它离散化为季节)。

3 广义/加权最小二乘和迭代加权最小二乘

例 2(续). 通过残差分析, 我们发现四个 Region 的误差方差差别较大, 这很可能是因为四个区的人口数量差别较大的原因。我们知道, 若分层抽样的每个层 (抽样单位, 这里是州) 的响应变量 y_i 是层内的平均值, 那么其方差与层内人口总数成反比, 当然也与层内面积、经济情况有关。

WLS

这里我们假设每个 Region 有不同的方差, 比例为 4 : 9 : 0.5 : 1

$$\log(\text{Expenditure})_i = \beta_0 + \beta_1 \times \log(\text{Income})_i + \beta_3 \times \text{Urban}_i + \sum_{k=2}^4 \alpha_k I_{(\text{Region}_i=k)} + \epsilon_i, \epsilon \sim (0, \sigma^2 G_0),$$

$$\text{var}(\epsilon) = G = \sigma^2 G_0 = \sigma^2 \text{diag}(4, \dots, 4, 9, \dots, 9, 0.5, \dots, 0.5, 1, \dots, 1).$$

下面我们应该广义最小二乘 GLS (实际上是加权最小二乘 WLS) 方法进行拟合:

```
G0=rep(0, 50)
G0[region==1]=4; G0[region==2]=9; G0[region==3]=0.5; G0[region==4]=1;
w=1/G0 #weights
fit3=lm(log(Expenditure)~log(Income)+Young+Urban+Region,data=edu,weights=w)
```

可以看到拟合效果有较大提升, 决定系数从 0.647 (fit2) 增加到 0.849 (fit3), 考虑到两个模型参数个数相同, 所以拟合效果的提升完全归因于考虑了异方差结构。需要注意的是, 基于 WLS 输出结果 fit3 的残差图 plot(fit3,which=1) 画出的拟合值-残差图仍然具有异方差特点。对于模型异方差模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim (0, \sigma^2 G_0)$$

为了考察 WLS 消除异方差的效果，我们需要考察

$$\mathbf{y}^* \triangleq G_0^{-1/2} \mathbf{y} = G_0^{-1/2} X \boldsymbol{\beta} + G_0^{-1/2} \boldsymbol{\epsilon} \triangleq X^* \boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* \sim (0, \sigma^2 I_n)$$

```
plot(fit3,which=1) #WLS fit3的残差图，仍然存在异方差
edu.standard =edu
edu.standard[,1:4]=edu.standard[,1:4]/sqrt(G0)
fit3.standard=lm(log(Expenditure)~log(Income)+Young+Urban+Region,data=edu.standard)
plot(fit3.standard,which=1,col=region)
```

IRLS

实际上，前述四个区的误差方差 G_0 并不是真实情况。假设

$$\text{var}(\epsilon_i) = \sigma_k^2, \text{ 若 } \text{Region}_i = k, k = 1, 2, 3, 4; i = 1, \dots, 50.$$

下面应用 IRLS 算法估计回归系数和误差方差 $\sigma_k^2, k = 1, 2, 3, 4$ 。IRLS 的基本思路是

- 给定 $\boldsymbol{\beta}$ 的估计，计算残差，并在每个区（Region）内计算误差方差 $\sigma_k^2, k = 1, 2, 3, 4$;
- 给定误差方差，应用 GLS 估计公式可求得 $\boldsymbol{\beta}$ 的估计;

上述两部迭代更新，直至收敛。

```
fit.ini = fit= lm(log(Expenditure)~., data=edu )

repeat{
  beta=coef(fit)
  res=resid(fit)
  sigmasq1 = sum(res[1:9]^2)/(9-1)
  sigmasq2 = sum(res[10:21]^2)/(12-1)
  sigmasq3 = sum(res[22:37]^2)/(16-1)
  sigmasq4 = sum(res[38:50]^2)/(13-1)

  sigma.sq=c(rep(sigmasq1,9),rep(sigmasq2,12),rep(sigmasq3,16),rep(sigmasq4,13))
  w=1/sigma.sq
  fit = lm(log(Expenditure)~log(Income)+Young+Urban+Region,
  data=edu , weight=w)
  beta.new=coef(fit)
  beta.new
  delta=sum(abs(beta.new-beta))
  print(delta)
  if (delta<1e-10) break
  beta=beta.new
}
fit3=fit # final fit
unique(sigma.sq)
```

练习题 (选做)

5. (最小一乘的 IRLS 解法) 对于线性模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim (0, \sigma^2 I_n), \quad y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), i = 1, \dots, n$$

最小一乘法极小化

$$\sum_{i=1}^n |y_i - \boldsymbol{\beta}^\top \mathbf{x}_i|$$

得到的估计称为最小一乘估计 (或 L_1 估计), 它对异常的响应值不太敏感, 因此是稳健估计。将最小一乘的目标函数改写为加权最小二乘的形式

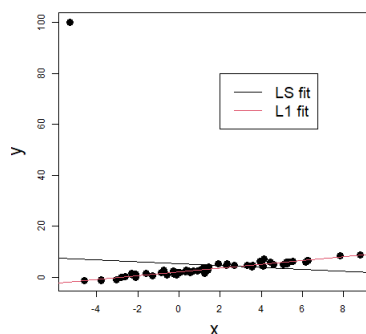
$$\sum_{i=1}^n w_i(\boldsymbol{\beta}) |y_i - \boldsymbol{\beta}^\top \mathbf{x}_i|^2$$

其中 $w_i(\boldsymbol{\beta}) = 1/|y_i - \boldsymbol{\beta}^\top \mathbf{x}_i|$, 为避免溢出, 可取 $w_i(\boldsymbol{\beta}) = 1/\max(0.0001, |y_i - \boldsymbol{\beta}^\top \mathbf{x}_i|)$ 。

任取 $\boldsymbol{\beta}$ 初始值, 比如 $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{OLS}$, IRLS 方法反复迭代如下两步:

$$\begin{aligned} \boldsymbol{\beta}_{new} &= \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n w_i(\boldsymbol{\beta}_0) |y_i - \boldsymbol{\beta}^\top \mathbf{x}_i|^2 = (X^\top W X)^{-1} X^\top W \mathbf{y}, W = \operatorname{diag}(w_1(\boldsymbol{\beta}_0), \dots, w_n(\boldsymbol{\beta}_0)) \\ \boldsymbol{\beta}_0 &\leftarrow \boldsymbol{\beta}_{new} \end{aligned}$$

- (a) 对于 alr4 数据集 *brains*, 考虑线性模型 $\log(\text{BrainWt}) = a + b \log(\text{BodyWt}) + \epsilon$, 求 a, b 的 LS 估计。
- (b) 假设由于某种错误第 14 行数据 y_{14} 被错误记录为 100, 求 a, b 的 LS 估计。
- (c) 写一个 IRLS 算法求解最小一乘估计的函数。对上述记录错误数据求 a, b 的最小一乘估计。与 (a) 中的结果比较, 错误记录对回归系数的最小一乘和最小二乘估计哪个影响更大? (提示: 你应该得到如下拟合结果)



4 附录: 探索非线性变换的两种方法 - lowess, IRP

```
#R函数lowess:
plot(x,y)
lowess(x,y,f=2/3)->lowess.fit ## f代表了曲线拟合的复杂度
lines(lowess.fit)
```

Box-Cox 是一种单调变换, 只能或有帮助解决残差中存在的单调的非线性现象。其它的非线性现象只能观察残差中的非线性趋势 (比如残差图中的红色拟合曲线, 它是由 lowess 方法拟合得到的), 猜测需要做的非线性变换。逆响应图 (IRP: inverse response plot) 与 lowess 类似。

LOWESS: 局部加权平滑方法 (Lowess, locally weighted scatterplot smoothing) 是一种一元非线性拟合方法, 是一种非参数方法。假设二元数据点 $(x_i, y_i), i = 1, \dots, n$ 满足非参数模型

$$y_i = f(x_i) + \epsilon_i, \epsilon_i \sim (0, \sigma^2)$$

其中 f 是未知的光滑函数。Lowess 方法在每个 $x_0 \in R$ 处最小化加权最小二乘

$$\min_{a,b} \sum_{i=1}^n w(x_i, x_0)(y_i - a - bx_i)^2,$$

其中 $w(u, v)$ 是权函数, 通常取高斯核函数 $w(u, v) = \phi((u - v)/h) = \frac{1}{2\pi} e^{-(u-v)^2/2h^2}$ 得到的解记为 $\hat{a} = \hat{a}(x_0), \hat{b} = \hat{b}(x_0)$, f 在 x_0 处的值估计为

$$\hat{f}(x_0) = \hat{a} + \hat{b}x_0$$

以工资-工龄数据为例

```
se=read.table("http://staff.ustc.edu.cn/~ynyang/2025/lab/salary-experience.txt",
             head=T, row.names=1)
se=se[,2:1]
plot(se)
lowess.fit=lowess(se, f=2/3)
lines(lowess.fit, col=2)
```

IRP (Inverse response plot) - lowess 的多自变量情形下的推广: 假设响应变量 y_i 与自变量 \mathbf{x}_i 满足模型:

$$\psi(y_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = 1, \dots, n \quad (1)$$

其中 ψ 是未知函数。

假设拟合线性模型 $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ 得拟合值 $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, 其中 $\hat{\boldsymbol{\beta}}$ 是 LS 估计。Inverse response plot 以响应变量为 x 轴, 拟合值为 y -轴, 画出二维散点图 $(y_i, \hat{y}_i), i = 1, \dots, n$, 观察并猜测两者之间的函数关系 $\hat{y}_i \approx \hat{\psi}(y_i)$, 此函数 $\hat{\psi}$ 可看作是模型 (1) 中的 ψ 函数的估计 (注意: 当只有一个自变量的时候, $\hat{y}_i = \hat{a} + \hat{b}x_i$, 此时 (y_i, \hat{y}_i) 散点图与 (y_i, x_i) 散点图等价)。如果从此图猜测 $\hat{y}_i = \psi(y_i)$, 则我们可将该函数用于变化响应变量 $y_i \rightarrow \psi(y_i)$ 。需要强调的是, 从散点图猜测变换通常并不容易, IRP 只能作为一个发现非线性变换的补充工具。

```
y=se[, "Salary"]
myfit=lm(Salary~Experience, data=se)
y.hat=fitted(myfit)
plot(y, y.hat) #y: response, y.hat: fitted response by LS
lowess(y, y.hat, f=2/3) -> lowess.fit
lines(lowess.fit)
```