Lab-5: 回归诊断 **2022 年 12 月 16-23 或 30 日**

内容: 回归诊断的残差分析和影响分析, Box-Cox 变换，lowess 非参数拟合
任务: 阅读例 1-2，完成练习题 1-3。

1 残差分析及 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)
```

变换之后重新拟合模型并做残差分析

(c) (b) 的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。再次做残差分析观察拟合情况。

2 简介：探索非线性变换的两种方法 - lowess, IRP

Box-Cox 是一种单调变换，只能或有希望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 $(x_i, y_i), i = 1, ..., n$ 满足非参数模型

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

其中 f 时未知的光滑函数。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

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

R 函数 lowess 使用方法如下

```
plot(x,y)
lowess(x,y,f=2/3)->lowess.fit
lines(lowess.fit)
```

以工资-工龄数据为例

```
se=read.table("http://staff.ustc.edu.cn/~ynyang/2022/lab/salary-experience.txt", head=T,row.names=1)
se=se[,2:1]
plot(se)
lowess(se,f=2/3)->lowess.fit
lines(lowess.fit)
```

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

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

其中 ψ 是未知函数。

假设拟合线性模型 $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ 得拟合值 $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$, 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 $(y_i, \hat{y}_i), i = 1, ..., n$，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 ψ 被看作是模型 (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 $y_i \to \psi(y_i)$。另外，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)
```

3 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/2022/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

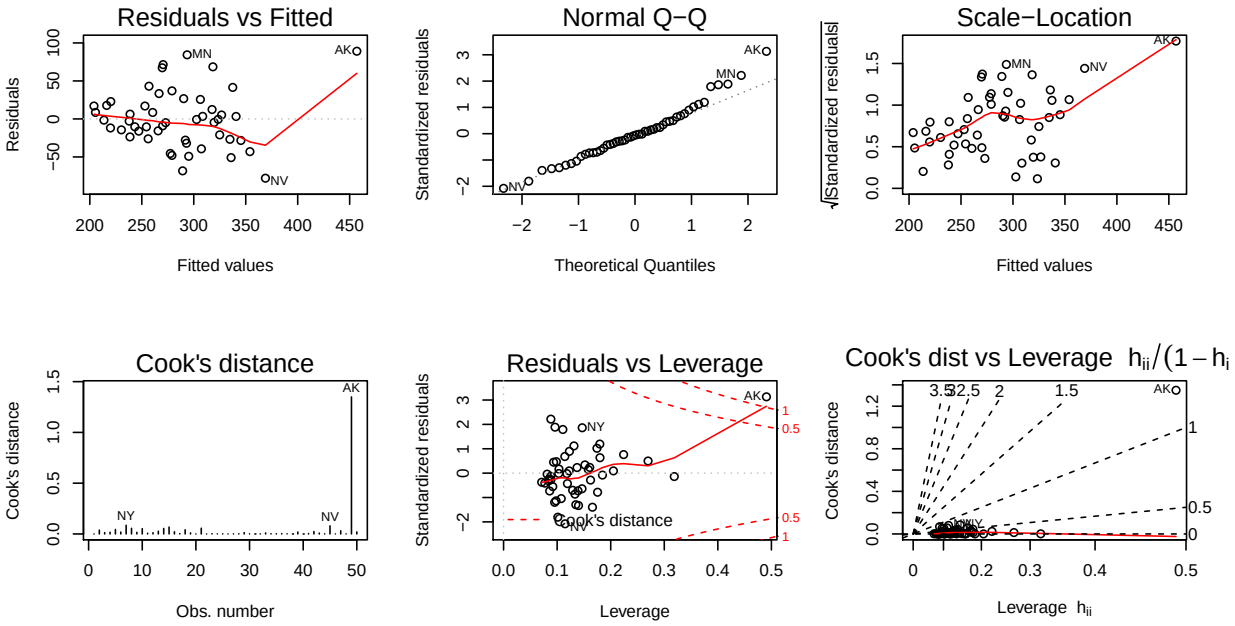
关心的问题是教育花费与其它变量的关系。假设回归模型

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

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。

```
> edu=read.table("http://staff.ustc.edu.cn/~ynyang/2022/lab/edu.xls", head=T,row.names=1)
> fit1 = lm(Expenditure ~. , data=edu)
> plot(fit1,which=1:6)
```

所有六个图如下：



各图分别是：

图 1: 残差图，横坐标为拟合值 (location)$\hat{y}_i$, 纵坐标为残差 $e_i = y_i - \hat{y}_i$;

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 的残差异常（即响应变量异常）。

图 2: 残差的 qqnorm 图，检查标准化残差

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

是否服从正态分布.

该图表明误差基本服从正态分布。

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）、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。在图 3 中 $\sqrt{|r_i|}$ 被当作 spread，其分布近似为正态。

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。

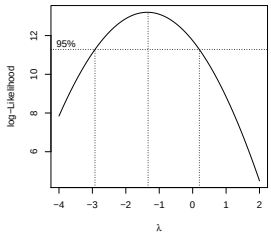
图 6: Cook 距离-杠杆图, 横坐标为杠杆 h_{ii}, 纵坐标为 Cook 距离 D_i。虚线为 $D_i/(h_{ii}/(1-h_{ii})) = r_i^2/p$ 的等高线。

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 leverageh_{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:

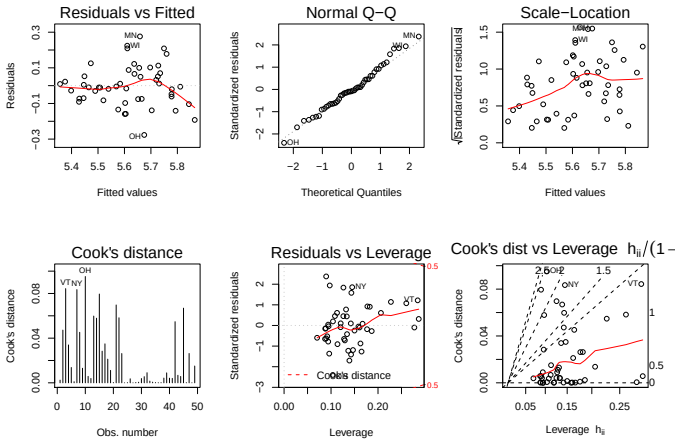
```
> influence.measures(fit1)
dfb.1_ dfb.Incm dfb.Yong dfb.Urbn dfb.Rgn2 dfb.Rgn3 dfb.Rgn4   dffit cov.r  cook.d   hat inf
CA  0.0179  2.3e-03  -0.0289  0.01505  9.6e-03  0.01376   0.0380  0.0611  1.40 5.5e-04 0.158
AK -2.2864  2.4e+00   2.0295 -1.74712 -7.2e-01  0.22849   0.0727  3.4571  0.38 1.3e+00 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 $h_{ii} = 0.491$ 远远大于其他各州）。删除 AK 之后，并对 Expenditure,Income 做 Box-Cox 变换。

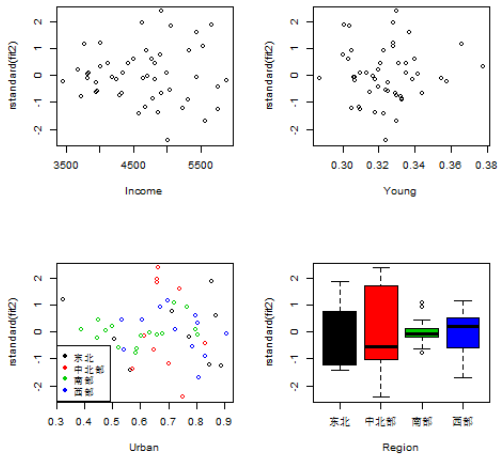


Expenditure 的 BC 变换 λ 的置信区间为 $[-3, 0]$, 我们（暂时）选取对数变换，Income 也做对数变换。

```
fit2=lm(log(Expenditure)~log(Income)+Young+Urban,data=edu[-49,])
par(mfrow=c(2,3))
plot(fit2,1:6 )
```



通常，响应变量和/或自变量的 Box-Cox 变换一般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：



我们发现，残差方差在 4 个地区 (region) 有较大的差异（第 4 图），对此我们可以假设地区之间的异方差模型，应用 IRLS 计算。另外第 3 图也有部分异方差现象，残差方差随 Urban 增大而增大，方差最大的中北部（红色）的城市人口比例 Urban 大多比较大，在右端，从而图 3 看起来是异方差的，这也说明异方差现象与 Region 有关。我们假设每个 Region 有不同的方差：

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

$$E(\epsilon_i) = 0, \quad var(\epsilon_i) = \sigma_k^2, \text{ 若 } \text{Region}_i = k, \ k = 1,2,3,4; i = 1,...,50.$$

下面是广义最小二乘估计的 IRLS 算法。

```
fit.ini = fit= lm(log(Expenditure)~., data=edu[-49,] )

repeat{
       beta=coef(fit)
       res=resid(fit)
       sigmasq1 = sum(res[1:9]^2)/(9)
       sigmasq2 = sum(res[10:21]^2)/(12)
       sigmasq3 = sum(res[22:37]^2)/(16)
       sigmasq4 = sum(res[38:49]^2)/(12)
       sigma2=c(rep(sigmasq1,9),
       rep(sigmasq2,12),rep(sigmasq3,16),rep(sigmasq4,12))
       w=1/sigma2
       fit = lm(log(Expenditure)~log(Income)+Young+Urban,
       data=edu[-49,], 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(sigma2)
```

练习题

1. 对于例 2 中的异方差模型（4 个 Region 的方差不同），试画出加权残差图 ($\sqrt{w_i}\hat{y}_i$, $\sqrt{w_i}e_i$)，是否还有方差不齐的现象？

2. 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？

3. **(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 前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
A	Action, 是否实施人工降雨 (1 = 实施人工降雨, 0 = 不实施)
D	Days, 第一次实施人工降雨 (1975 年 6 月 16 日，D=0) 之后的天数
S	Suitability for seeding, 度量是否适合进行人工降雨的数学模型指标
C	Cover, 试验区域云层覆盖率
P	Pre-wetness, 人工降雨之前 1 小时的降雨量 (单位：10^7 立方米)
E	Echo motion category, 云层类型 (类别 1 或 2)
$Rain$	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（即针对试验数据的多重回归分析）。试分析人工降雨是否有显著效果（提示：如何恰当处理变量 D 或许是一个关键）。