Lab-4: 多重线性回归和回归诊断 **2023** 年 **12** 月 **15-22** 日

内容：多重线性回归、F-检验、回归诊断的残差分析和影响分析、Box-Cox 变换
任务：阅读例 1-2，完成练习题 1-4。

1 多重线性回归模型

例 1. 我们下面以成年人身高-体重数据介绍多重回归模型

`http://staff.ustc.edu.cn/~ynyang/2023/lab/height-weight.txt`

读入 R. 该数据的三个变量为：sex (1：M, 0: F), weight (kg), height (m).

1.1 多重回归

显然，例子中性别 sex 既与体重 weight 有关，也与身高 height 有关，在群体中研究体重与身高的关系时，需在回归模型中对性别加以控制，这可以在简单回归模型 $\log(weight) \sim \log(height)$ 中添加 sex 一项：

$$\log(\text{weight}) = a + b \times \log(\text{height}) + c \times \text{sex} + \epsilon, \epsilon \sim (0, \sigma^2) \tag{1}$$

该模型蕴含了如下事实：log(height) 的回归系数 b 对于不同性别都是一样的 (但截距项有差别)：

$$\begin{aligned} \text{sex} = 0: &\quad \log(\text{weight}) = a + b \times \log(\text{height}) + \epsilon \\ \text{sex} = 1: &\quad \log(\text{weight}) = (a + c) + b \times \log(\text{height}) + \epsilon \end{aligned} \tag{2}$$

lm 拟合结果 fit 是个列表 (list), 它包含的内容可以以查看其各个分量名称的方式看到：

```
> fit = lm(log(weight)~log(height)+sex, data=hw )
> coef(fit) #fitted(fit), residuals(fit)分别给出拟合值和残差
      (Intercept) log(height)         sex
           3.0087      2.0572      0.1241
> names(fit)
      [1] "coefficients" "residuals" "effects" "rank"
      [5] "fitted.values" "assign" "qr" "df.residual"
      [9] "xlevels" "call" "terms" "model"
```

我们可用如下方式提取列表 fit 的各个分量：

```
fit$coeff, fit$res ...
```

对于系数估计、残差、拟合值也可以使用函数提取：

```
系数估计: coef(fit) or coefficients(fit)
残差: resid(fit) or residuals(fit)
拟合值: fitted(fit)
```

1.2 summary 函数

关于统计推断，以及更全面的结果可以使用 summary 函数得到. summary 包含如下内容 (主要是单个回归系数的检验、回归方程显著性检验以及拟合优度等):

```
> summary(fit)
      Coefficients:
      Estimate Std. Error t value Pr(>|t|)
      (Intercept) 3.00871 0.11560 26.028 < 2e-16
      log(height) 2.05716 0.23092 8.909 3.55e-16
      sex 0.12408 0.02427 5.113 7.51e-07
      ---
      Residual standard error: 0.1145 on 196 degrees of freedom
      Multiple R-squared: 0.6601, Adjusted R-squared: 0.6567
      F-statistic: 190.4 on 2 and 196 DF, p-value: < 2.2e-16
```

由该输出结果可以看出：

1. 各个回归系数都显著地非 0，比如 log(height) 的回归系数估计为 $\hat{b} = 2.05716$, 标准差为 $sd(\hat{b}) = \sqrt{\widehat{var}(\hat{b})} = 0.23092$, $t = \hat{b}/sd(\hat{b}) = 8.909$, p-值 $= P(|t_{n-p}| \geq 8.909) = 3.55\text{e-}16$.

2. 误差方差估计 $\hat{\sigma}^2 = 0.1145^2 = 0.0131$, 也可以由公式 $\hat{\sigma}^2 = RSS/(n-p)$ 求出：

```
e=resid(fit)
sum(e^2)/(199-3)
```

3. 复相关系数平方 $R^2 = 0.6601$, 验证 $R^2 = VAR(\hat{y})/VAR(y) = r^2_{\hat{y},y}$ (这里 $y = log(weight)$, VAR 代表样本方差)。

```
y.hat=fitted(fit)
y=log( hw[,"weight"])
( R2=cor(y.hat,y)^2 ) #=0.6601
( R2=var(y.hat)/var(y) ) #=0.6601
```

4. 回归方程显著性检验指的是同时检验所有自变量的回归系数是否为 0，这里检验 $H_0 : b = c = 0$, 检验统计量为 sumary 最后一行的 F-statistic $F = 190.4$, 自由度为 2 (检验的自变量的个数 k) 和 196 ($= n - p = 199 - 3$, pvalue$< 2.2e-16$。验证：$F = \frac{n-p}{k} \times \frac{R^2}{1-R^2}$.

```
R2=var(y.hat)/var(y)
n=nrow(hw)
p=ncol(hw)
k=2 # H0: b=c=0
F=(n-p)/k*R2/(1-R2)
F
```

summary 中所含的具体内容可以如下方式看到：

```
> a = summary(fit)
> names( a )
      [1] "call" "terms" "residuals" "coefficients"
      [5] "aliased" "sigma" "df" "r.squared"
      [9] "adj.r.squared" "fstatistic" "cov.unscaled"
```

如果希望提取 summary 中的某些信息，比如回归方程显著性检验统计量 F-statistic，则可

```
> summary(fit)$fstatistic
      value numdf dendf
      190.3614 2.0000 196.0000
```

给出了 F 值，分子自由度 numdf (df for numerator)，分母自由度 dendf (df for denumerator)。再如，提取 R^2:

```
> R.sq = summary(fit)$r.squared
```

1.3 一般线性假设的 F 检验

函数 *summary* 主要概括了其中的统计推断结果，包括 LS 估计及其 t 检验、回归方程的显著性 F 检验，但不直接提供其它的多个参数的同时 F 检验。为了检验一般的假设检验问题，可以调用 *anova* 函数。

anova(sub.model, full.model)

该函数计算 F 统计量

$$F = \frac{n-p}{q} \times \frac{RSS_0 - RSS}{RSS},$$

它比较子模型 sub.model 的残差平方和 RSS_0 与全模型 full.model 的残差平方和 RSS, 其中 n 为样本量，p 为回归系数的个数, q 为线性假设中待检验的参数个数或线性约束的个数。比如，检验模型 (1) 的显著性 $H_0 : \beta_1 = \beta_2 = 0$:

```
model0=lm(log(weight) ~ 1, data=hw)
              # ~1: intercept only (no covariates in the model)
fit2=lm(log(weight) ~ log(height) + sex , data=hw)
anova(model0, fit2)
```

这与 summary(fit2) 给出的结果是一样的。

再如，我们考虑检验模型 (1) 中 $H_0 : b = c$。该假设成立时的零模型为

$$\log(\text{weight}) = a + b \times [\log(\text{height}) + \text{sex}] + \epsilon, \epsilon \sim (0, \sigma^2) \tag{3}$$

因此我们需要先定义新变量 $z = \log(\text{height}) + sex$

```
model.full=lm(log(weight) ~ log(height) + sex , data=hw)
z=log(hw[,"height"])+hw[,"sex"]
model.null=lm(log(weight) ~ z, data=hw)
anova(model.null, model.full)
```

1.4 交互作用

如果认为第一部分 (简单回归) 中所得两条直线不平行，即 log(height) 的回归系数与性别有关，那么我们可以考虑在模型中加入交互作用项：

$$\log(\text{weight}) = \beta_0 + \beta_1 \log(\text{height}) + \beta_2 \text{sex} + \gamma \log(\text{height}) \times \text{sex} + \epsilon \tag{4}$$

你可以将该模型理解为有 3 个自变量 (log(height), sex, log(height)×sex) 的多重回归模型，它表明对于不同的性别，log(height) 的效应（回归系数）是不同的：

$$\begin{aligned} \text{sex} = 0: &\quad \log(\text{weight}) = \beta_0 + \beta_1 \log(\text{height}) + \epsilon \\ \text{sex} = 1: &\quad \log(\text{weight}) = (\beta_0 + \beta_2) + (\beta_1 + \gamma) \log(\text{height}) + \epsilon \end{aligned} \tag{5}$$

```
fit.interaction= lm(log(weight)~log(height)*sex, data=hw ) #or
fit.interaction= lm(log(weight)~log(height)+sex+log(height):sex,data=hw)
#a:b代表a,b的交互作用，也可以用a*b
```

为了检验模型 (1) 是否合理，我们可以检验模型 (4) 中回归系数 γ 是否为 0：

```
> summary(fit.interaction)
              Coefficients:
              Estimate Std. Error t value Pr(>|t|)
              (Intercept) 3.1203 0.1572 19.848 < 2e-16
              log(height) 1.8333 0.3147 5.826 2.32e-08
              sex  -0.1378 0.2513 -0.548 0.584
              log(height):sex 0.4848 0.4631 1.047 0.296
              ---
              Residual standard error: 0.1145 on 195 degrees of freedom
              Multiple R-squared: 0.662, Adjusted R-squared: 0.6568
              F-statistic: 127.3 on 3 and 195 DF, p-value: < 2.2e-16
```

所以 $H_0 : \gamma = 0$ 的检验 $t = 1.047$,p 值 =0.296, 不显著, 即没有理由认为两条直线是不平行的.

1.5 因子变量 (factor)

因子变量 (factor) 取值为类别, 其取值称为水平 (levels)。因子变量的取值一般使用类别的名称, 比如大学教师的职称（Rank）是因子变量，有 3 个水平：Assistant Professor, Associate Professor, Full Professor; 因子变量的取值有时也使用数字代表, 比如我们可以分别用 1, 2, 3 分别代表 Rank 的 3 个水平，但这里的 1,2,3 只是代表因子的水平而不具有实数含义, 所以使用任何三个不同的数字代表这三个水平都可以。在实际数据中如果某个因子变量的各个水平是用数字表示的，那么你首先应确认它是因子变量而不是实数 (numeric) 变量. R 中判断是否为因子的函数为 is.factor(), 而转化为因子的函数为 factor(), as.factor(). 举例如下：

```
> (x=c(1,4,1,2,3,3,2,2))
[1] 1 4 1 2 3 3 2 2
> is.numeric(x)
[1] TRUE
> x1=as.factor(x)
> x1
[1] 1 4 1 2 3 3 2 2
Levels: 1 2 3 4
> is.numeric(x1)
[1] FALSE
> is.factor(x1)
[1] TRUE
```

为了在数学上和计算机上能够处理因子变量，需要把因子变量表示成数值形式的变量，最常用的方法是哑变量/示性函数表示方法，统计中称为 treatment effect contrasts。通常，它是 R 软件缺省的因子变量表示方法。（但可以用改为其它表示方法）。

假设我们有如下数据 bpdata，为成年男性的血压 (BP), 体重 (Weight), 和人种 (Race).

```
BP Race Weight
112 White 71
122 White 82
133 Black 77
131 Yellow 68
127 Black 62
122 White 79
。。。
```

定义示性变量 RaceWhite $=1_{(Race=White)}$ 和 RaceYellow$=1_{(Race=Yellow)}$，Black 是基准 (baseline)，因此原数据实际上为：

```
BP RaceWhite RaceYellow Weight
112 1 0 71
122 1 0 82
133 0 0 77
131 0 1 68
127 0 0 62
122 1 0 79
```

为了研究血压与身高的关系, 我们在线性模型中控制 Race, R 命令为:

$$BP \sim Weight + Race$$

以数学公式表达如下：

$$BP = \alpha + \beta * Weight + \gamma_1 * RaceWhite + \gamma_2 * RaceYellow + \epsilon$$

共有 4 个回归系数，其中 γ_1,γ_2 分别是 RaceWhite,RaceYellow 的效应。上述模型是线性的，特别地，称为是可加的 (additive) 模型, Race 取值的改变并不改变 Weight 的效应 β, 反之，Weight 的变化不影响 RaceWhite, RaceYellow 的效应 γ_1, γ_2. 上述方程可以拆解为

$$\begin{aligned} Race = Black: &\quad BP = \alpha + \beta * Weight + \epsilon \\ Race = White: &\quad BP = \alpha + \gamma_1 + \beta * Weight + \epsilon \\ Race = Yellow: &\quad BP = \alpha + \gamma_2 + \beta * Weight + \epsilon \end{aligned}$$

无论 Race 为何，Weight 的效应都是 β。假设 LS 拟合得到如下结果

```
> lm(BP~Weight+Race, data=bpdata)
Coefficients:
(Intercept) Weight RaceWhite RaceYellow
52.2227 0.4589 -14.9278 -0.5649
```

所以三类人的拟合方程分别如下：

$$\begin{aligned} Race = Black: &\quad BP = 52.2227 + 0.4589 * Weight \\ Race = White: &\quad BP = 37.2949 + 0.4589 * Weight \\ Race = Yellow: &\quad BP = 51.6578 + 0.4589 * Weight \end{aligned}$$

如果希望改变 R 缺省的基准，比如上述问题中我们希望以 White 为基准，下面的命令将基准 (base) 改变为 Race 的第 2 个水平（即 White）。

```
> contrasts(bpdata[,"Race"]) = contr.treatment(3,base=2)
# base=2 (number of levels is 3, the 2nd level (White)
is set to be the base)
> bpdata[,"Race"]
[1] White White Black Yellow Black White Yellow
attr(,"contrasts")
1 3
Black 1 0
White 0 0
Yellow 0 1
Levels: Black White Yellow

> lm(BP~Weight+Race, data=bpdata)
Coefficients:
(Intercept) Weight RaceBlack RaceYellow
37.2950 0.4589 14.9278 14.3629
```

拟合得到的模型为

$$BP = 37.2950 + 0.459 \times Weight + 14.9278 \times RaceBlack + 14.3629 \times RaceYellow, \tag{6}$$

看起来，上述结果与 Black 做 baseline 的时候似乎不同，但实际上是完全相同的，比如当 Race 为 Black 时，拟合方程为

$$BP = 37.2950 + 0.459 \times Weight + 14.9278 = 52.2227 + 0.459 * Weight.$$

与前面得到的拟合方程相同，F 检验结果也是相同的。

t

注意：R 缺省地把名称的首字母次序最靠前 (或数字最小) 的水平作为基准。

1.6 方差分析 (anova) 与协方差分析 (ancova)

简单来说，对试验数据而言方差分析指的是线性回归模型中所有自变量皆为因子变量的情形，协方差分析是既有因子变量也有其它控制变量（连续或者因子）的情形。对于观察研究数据我们不妨也沿用这两种称呼。所以上述 1.6 种的例子可看作是协方差分析。当然，方差分析和协方差分析更为关注因子的检验及其依赖的平方和分解。R 中方差分析的专用函数是 aov (其调用方式与 lm 相似, 比如 aov (response ~ block+factor1+factor2+factor1:factor2)). 而协方差分析中检验因子变量可通过 anova 函数。例如

```
> summary(aov(BP~Race,data=bpdata))
Df Sum Sq Mean Sq F value Pr(>F)
Race 2 252.55 126.27 5.931 0.0636 .
Residuals 4 85.17 21.29
       ---

> summary(lm(BP~Race,data=bpdata))

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.000 3.263 39.843 2.37e-06 ***
RaceWhite -11.333 4.212 -2.691 0.0546 .
RaceYellow 1.500 4.614 0.325 0.7614
       ---
Residual standard error: 4.614 on 4 degrees of freedom
Multiple R-squared: 0.7478, Adjusted R-squared: 0.6217
F-statistic: 5.931 on 2 and 4 DF, p-value: 0.0636
```

aov 函数给出的结果是 lm 函数结果的一部分，因此所有 aov 分析结果都可在 lm 结果中找到，比如上面的 lm 结果中的回归方程显著性检验 $F = 5.931$ 与 aov 的 F 检验一致。

练习题

1. 上世纪 80 年代美国中西部一个大学女教师曾经起诉学校在工资待遇上歧视女性, 数据集 *salary* (在 alr4 程序包中) 是当时该校 52 个正式教工的工资数据, 变量描述如下:

变量	描述
Sex	1: 女，0: 男
Rank	职称. 1: Assistant Prof, 2: Associate Prof, 3:Full Prof
Year	拥有当前职称 (Rank) 的时间 (单位：年)
Degree	最高学位. 1: 博士, 0: 硕士
YSdeg	工龄: 获得最高学位至今的时间 (单位：年)
Salary	年薪 ($)

   我们需要研究数据是否说明了女性在工资待遇上确实受到了歧视。

   (a) 假设男女工资 (Salary) 各服从 等方差 的正态分布，检验男女教师工资是否相同。

   t.test(Salary ~ Sex, data=salary, var.equal=T)

   也可通过如下简单线性模型

   lm(Salary ~ Sex, data=salary)

   进行检验. 两个结果是否相同？根据模型拟合输出结果, 男女平均工资的差异等于多少？结果是否显著 (显著性水平 0.1)？该结论是否说明有歧视女性的现象？是否存在干扰因素？

   注：对于上述两样本 t-检验，如果认为两个总体方差不等，需要在 t.test 中设定 var.equal=F, 即所谓的 Welch two-sample t-test。

   t.test(Salary ~ Sex, data=salary, var.equal=F) # Welch's t-test

   (b) 一个可能的干扰因素是职称 (Rank), 试给出它与工资 (Salary) 以及与性别 (Sex) 相关的证据.

   (c) 我们在上述简单回归模型中增加 Rank 变量，用来控制（消除）Rank 的干扰：

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Rank} + \epsilon$$

该模型蕴含了如下事实：不论 Rank 属于哪个类，Salary 与 Sex 的关系（Sex 的回归系数 b）保持不变, 请验证这个事实是否近似成立。

lm(Salary ~ Sex , data=salary,subset= (Rank==1))

lm(Salary ~ Sex , data=salary,subset= (Rank==2))

lm(Salary ~ Sex , data=salary,subset= (Rank==3))

(d) 应用多重线性回归模型

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Rank} + \beta_3 \text{Year} + \beta_4 \text{Degree} + \beta_5 \text{YSdeg} + \epsilon \tag{7}$$

检验模型 (7) 中 $H_0 : \beta_1 = \beta_2 = 0$.

2. 假设如下三组数据分别来自于正态总体 $N(\mu_k, \sigma^2), k = 1, 2, 3$:

组 1: $-1.7, -1.5$;

组 2: $-0.4, -1.1, 1.3, -0.3$;

组 3: $2.0, 1.2, 0.6$;

检验: $H_0 : \mu_1 = \mu_2 = \mu_3$ (提示: 你需要首先定义一个因子变量表示分组).

2 回归诊断

2.1 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)
```

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)
```

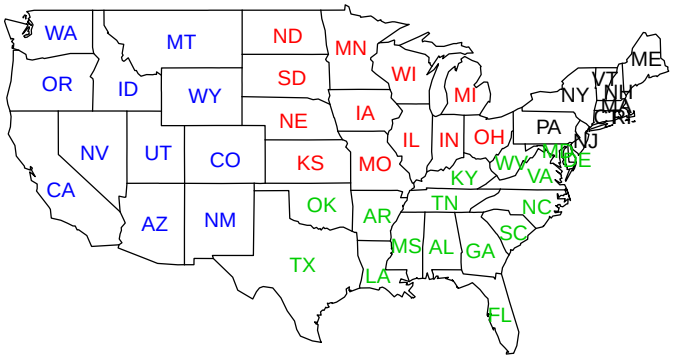
由此得到对自变量所应该进行的变换。再次做残差分析观察拟合情况。

2.2 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/2023/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/2023/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)
#如何在上图中以4种颜色标记四个地区 （Region）？
```

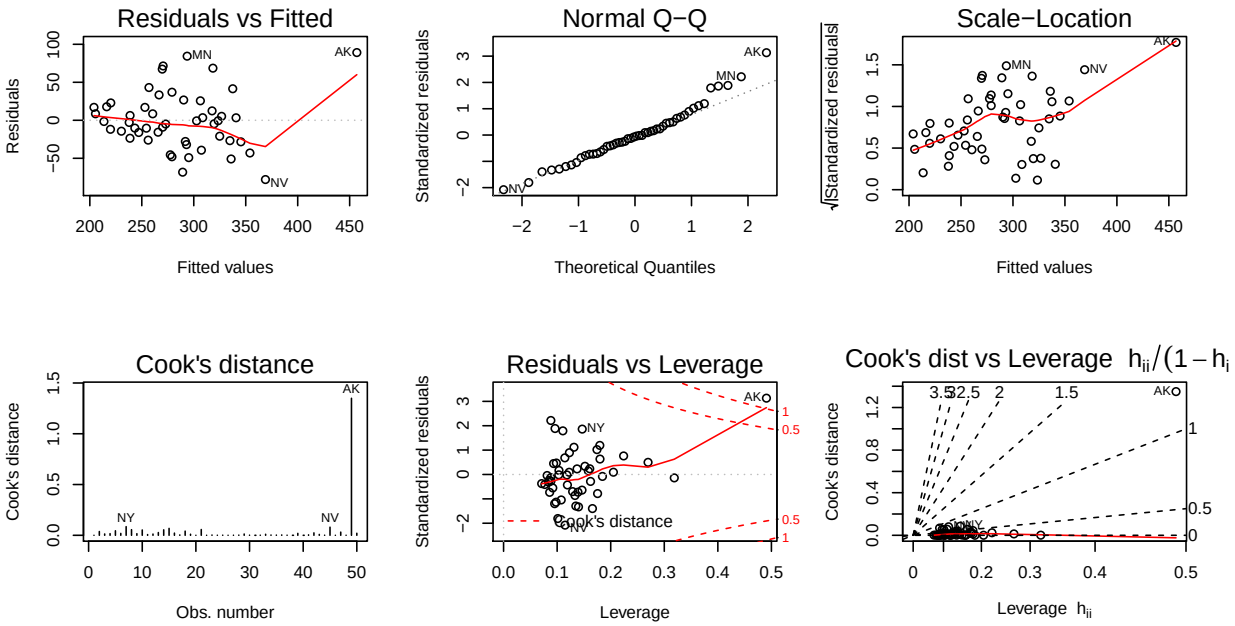
我们关心的问题是教育花费与其它变量的关系。假设回归模型

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

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。

```
> 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

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

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）、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。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。

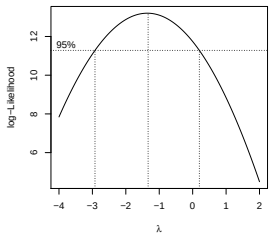
图 6: Cook 距离-杠杆图, 横坐标为杠杆 h_{ii}, 纵坐标为 Cook 距离 D_i。虚线为 $D_i/(h_{ii}/(1-h_{ii})) = r_i^2/p$ 的等高线。

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 $|r_i| \le \sqrt{p} = \sqrt{6}$。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:

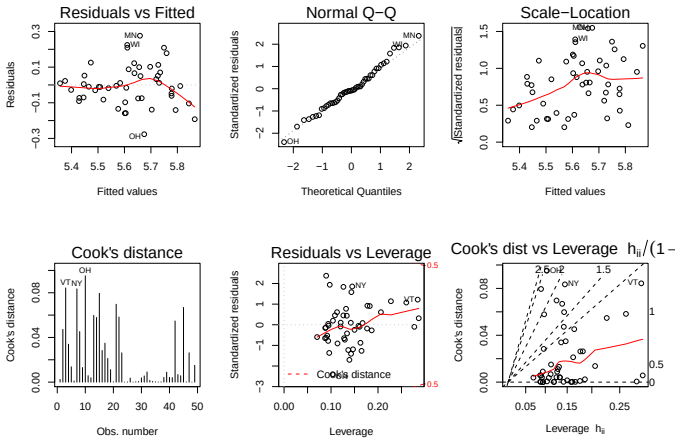
```
> influence.measures(fit1)
dfb.1_ dfb.Incm dfb.Yong dfb.Urbn dfb.Rgn2 dfb.Rgn3 dfb.Rgn4   dffit cov.r  cook.d   hat inf
CA  0.0179  2.3e-03  -0.0289  0.01505  9.6e-03  0.01376   0.0380  0.0611  1.40 5.5e-04 0.158
AK -2.2864  2.4e+00   2.0295 -1.74712 -7.2e-01  0.22849   0.0727  3.4571  0.38 1.3e+00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

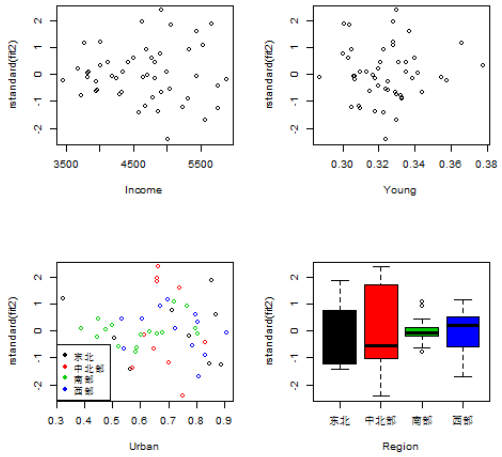


Expenditure 的 BC 变换 λ 的置信区间为 $[-3, 0]$, 我们（暂时）选取对数变换，Income 也做对数变换。

```
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），对此我们可以假设不同地区的误差方差不同，进而应用广义最小二乘法拟合（此处略）。

2.3 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 ## f代表了曲线拟合的复杂度
lines(lowess.fit)
```

以工资-工龄数据为例

```
se=read.table("http://staff.ustc.edu.cn/~ynyang/2023/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 与自变量 $\mathbf{x}_i$ 满足非线性模型：

$$\psi(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = 1, ..., n \tag{8}$$

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)
```

练习题

3. 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？

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 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„ D 代表了季节?）。