Lab-3: 多重线性回归 **2025** 年 **11** 月 **24** 日

内容: 多重线性回归，学习因子变量、交互作用
任务: 完成练习题 1-4。

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/2025/lab/height-weight.txt}`

该数据的三个变量为： `sex (1: M, 0: F), weight (kg), height (m).` 样本量`n=199.`

1 多变量线性回归

1.0 简单回归

我们首先做简单线性回归分析. 只对男性拟合模型：

$$\log(\text{weight}) = a_1 + b_1 \times \log(\text{height}) + \epsilon, \epsilon \sim (0, \sigma_1^2) \tag{1}$$

只对女性拟合模型：

$$\log(\text{weight}) = a_0 + b_0 \times \log(\text{height}) + \epsilon, \epsilon \sim (0, \sigma_0^2) \tag{2}$$

```
> hw=read.table("http://staff.ustc.edu.cn/~ynyang/2025
/lab/height-weight.txt",head=T)

> attach(hw)
> plot(log(height),log(weight), col=sex+1)

> fit.all=lm(log(weight)~log(height), data=hw)
> fit.all
Coefficients:
(Intercept) log(height)
2.599 2.930

> fit.female= lm(log(weight)~log(height), data=hw,subset=sex==0)
> fit.female
Coefficients:
(Intercept) log(height)
3.120 1.833

> fit.male= lm(log(weight)~log(height), data=hw,subset=sex==1)
> fit.male
Coefficients:
(Intercept) log(height)
2.982 2.318

> abline(fit.all,col=3)
```

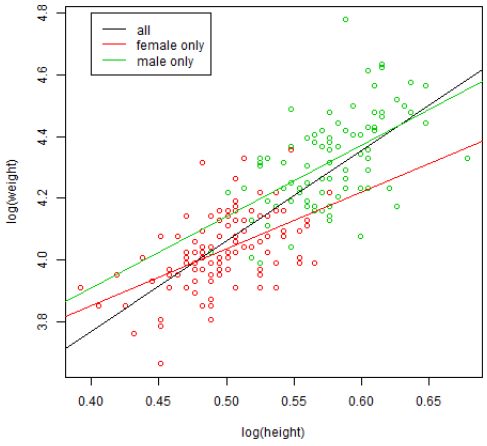


Figure 1: 简单线性回归

```
> abline(fit.female,col=1)
> abline(fit.male,col=2)
> legend(0.4,4.8,c("all","female","male"), col=c(3,1,2), lty=c(1,1,1))
> summary(fit.female) #more details
```

从输出结果和 Figure 1 看出，两条 sex=1，sex=0 两组数据的回归直线近似平行 (LS 估计分别为 $\hat{a}_0 = 3.12, \hat{b}_0 = 1.833$ 和 $\hat{a}_1 = 2.982, \hat{b}_1 = 2.318$), 另外，$\hat{\sigma}_0 = 0.103, \hat{\sigma}_1 = 0.127$。

1.1 多变量回归

显然，例 1 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：

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

该模型中样本量 $n = 199$, 回归系数个数 $p = 3$ (a, b, c), 该模型蕴含了如下事实：

$$\begin{aligned} \text{sex} = 0: &\quad \log(\text{weight}) = a + b \times \log(\text{height}) + \epsilon \\ \text{sex} = 1: &\quad \log(\text{weight}) = (a + c) + b \times \log(\text{height}) + \epsilon \end{aligned} \tag{4}$$

即两变量模型要求 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)，误差方差也相同。我们在 Figure 1 看到这些要求基本满足。

```
> myfit = lm(log(weight)~log(height)+sex, data=hw )
> names(myfit) #fit是个列表(list)，它包含的内容如下：
      [1] "coefficients" "residuals" "effects" "rank"
      [5] "fitted.values" "assign" "qr" "df.residual"
      [9] "xlevels" "call" "terms" "model"
> myfit$coeff #提取系数估计，这等价于 coef(myfit)
> coef(myfit)
> myfit$residuals # 等价于resid(myfit)
```

```
> resid(myfit)
> myfit$fitted.balues
> fitted(myfit)
##总结：既可以直接用"$"提取myfit的内容，也可以使用以下函数：
系数估计：coef(myfit) or coefficients(myfit)
残差：resid(myfit) or residuals(myfit)
拟合值：fitted(myfit)
```

1.2 LS 估计的分量

划分线性模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

其中 X_1 的第一列是 $\mathbb{1}$。我们知道 $\boldsymbol{\beta}_2$ 的 LS 估计

$$\hat{\boldsymbol{\beta}}_2 = (X_2^{\perp\top} X_2^{\perp})^{-1} X_2^{\perp\top} \mathbf{y}$$

其中 $X_2^{\perp} = X_2 - P_{X_1} X_2$ 消除了 X_2 与 X_1 的相关性，是 $lm(X_2 \sim X_1)$ 的残差。因此 $\hat{\boldsymbol{\beta}}_2$ 可由两步回归得到

- $lm(X_2 \sim X_1)$ 得到残差 $X_2^{\perp}$;
- $lm(y \sim X_2^{\perp})$ 得到的回归系数估计就是 $\hat{\boldsymbol{\beta}}_2$.

我们用例 1 的数据验证如下，求模型 (3) 中 log(weight) 的系数 b 的 LS 估计：

```
> fit1 = lm(log(height) ~ sex, data=hw )
> height.perp = resid(fit1)
# 消除了log(height)与sex的相关性， 即两个自变量 log(height), sex实现了正交化.
> fit2=lm(log(weight) ~ height.per,data=hw)
> coef(fit2) #与myfit中的log(height)的系数完全相同
```

1.3 summary 函数

关于统计推断，以及更全面的结果可以用 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(myfit)
      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 检验统计量 (t value):

$$t = \hat{b}/sd(\hat{b}) = 2.05716 \div 0.23092 = 8.909,$$

   p-值 $= P(|t_{n-p}| \geq 8.909) = 3.55\text{e-}16$.

2. 误差方差估计 $\hat{\sigma}^2 = 0.1145^2 = 0.0131$. R summary 输出结果“Residual standard error: 0.1145 on 196 degrees of freedom”中的“Residual standard error”就是 $\hat{\sigma}$, “degrees of freedom”指的是 $n - p = 199 - 3$。当然，误差方差也可以由公式 $\hat{\sigma}^2 = RSS/(n-p)$ 求出：

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

3. 决定系数（也称为复相关系数平方）$R^2 = 0.6601$。我们知道 R^2 是回归平方和在总平方和中的占比，即拟合值的样本方差与响应变量样方差之比。另外我们知道 R^2 是响应响应变量与拟合值之间样本相关系数的平方：$R^2 = r^2_{\hat{y},y}$ (这里 $y = log(weight)$)。下面验证这两种计算方式得到的结果一致：

```
y.hat=fitted(myfit)
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 $p - 1$) 和 196 ($n - p = 199 - 3$), pvalue= $P(F_{p-1,n-p} \geq 190.4) < 2.2e - 16$。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

#输出结果给出的回归方程显著性检验的p值$<2.2e-16$，具体多大？
pf(190.4, df1=2,df2=196, lower.tail=FALSE) #pvalue，右端尾概率
  [1] 1.149795e-46
pf(190.4, df1=2,df2=196, lower.tail=FALSE,log.p=TRUE) #log(pvalue)
  [1] -105.7793
```

summary 中所含的具体内容可以如下方式看到：

```
> a = summary(myfit)
> 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(myfit)$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(myfit)$r.squared
```

1.4 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)

这里 sub.model 指的是原假设成立时的模型拟合结果，full.model 指的是全模型的拟合结果。anova 函数计算 F 统计量将在第 11 讲给出：

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

它比较原假设成立时的子模型 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 (3) 的显著性 $H_0 : \beta_1 = \beta_2 = 0$:

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

这与 summary(model.full) 给出的回归方程显著性检验结果是一样的。

再如，我们考虑检验模型 (3) 中 $H_0 : b = c$。该假设成立时的零模型为

$$\log(\text{weight}) = a + b \times [\log(\text{height}) + \text{sex}] + \epsilon, \epsilon \sim (0, \sigma^2) \tag{5}$$

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)
```

1.5 交互作用

例 1 中如果我们认为 log(height) 的回归系数与性别有关（即简单回归中对男性和女性分别拟合所得的两条直线不平行），那么我们可以考虑在模型中加入交互作用项：

$$\log(\text{weight}) = \beta_0 + \beta_1 \log(\text{height}) + \beta_2 \text{sex} + \gamma \log(\text{height}) \times \text{sex} + \epsilon \tag{6}$$

我们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（即回归系数）是不同的，它们分别是 β_1 和 $\beta_1+\gamma$：

$$\begin{aligned} \text{sex}=0: &\quad \log(\text{weight})=\beta_0+\beta_1\log(\text{height})+\epsilon \\ \text{sex}=1: &\quad \log(\text{weight})=(\beta_0+\beta_2)+(\beta_1+\gamma)\log(\text{height})+\epsilon \end{aligned} \tag{7}$$

```
fit.interaction= lm(log(weight)~log(height)+sex+log(height):sex,data=hw) #or
fit.interaction= lm(log(weight)~log(height)*sex, data=hw )
#其中 a:b 代表a,b的交互作用，也可以用 a*b (后者既包含交互项也包含主项)
```

对应于交互模型 (6)，我们称模型 (3) 是可加的。为了检验可加模型 (3) 是否合理，我们检验交互模型 (6) 中回归系数 γ 是否为 0（当 $\gamma=0$ 时，模型 (6) 退化为模型 (3)）：

```
> 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
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 γ 的 LS 估计等于 0.4848, $H_0:\gamma=0$ 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, 从而否定交互模型 (6) 而接受可加模型 (3)。

2 因子变量简介

我们将在后续课程中学习方差分析方法，将碰到因子变量、交互作用等概念。

2.1 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。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

$$BP = 37.2950 + 0.459 \times Weight + 14.9278 \times RaceBlack + 14.3629 \times RaceYellow, \tag{8}$$

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。
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。

2.2 方差分析与协方差分析

简单来说，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。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))，当然也可使用 lm 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

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Rank} + \beta_3 \text{Year} + \beta_4 \text{Degree} + \beta_5 \text{YSdeg} + \epsilon \tag{9}$$

检验模型 (9) 中 $H_0 : \beta_1 = \beta_2 = 0$.

2. salary 数据中职称 (Rank) 变量取值为 1、2、3 为实数 (numeric), 上题中我们把它当成了实数变量，实际上 1，2，3 分别代表 Assistant Prof, Associate Prof 和 Full Prof，所以 Rank 其实是一个因子变量，需要用哑变量/示性变量将其表示为实数变量，应用如下模型分析数据

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \gamma_2 \text{Rank2} + \gamma_3 \text{Rank3} + \epsilon \tag{10}$$

其中 Rank2, Rank3 分别是 Associate Prof, Full Prof 的示性变量. 通常只需首先把 Rank 转化为因子变量：

$$lm(Salary \sim as.factor(Rank) + ..., data = salary),$$

但这里我们手工定义 Rank2, Rank3 以方便后续讨论：

```
Rank=salary[ , ''Rank'']
Rank2 <- Rank==2
Rank3<- Rank==3
```

(a) 可加性：模型 (10) 中以 Rank 与 Sex 线性可加的形式控制 Rank，这样做的前提是不同的职称等级中男女工资差异是相同的，试检验该前提是否合理。

(b) 第一题中将因子变量 Rank 当作数值 (numeric) 变量是否合理？

模型 (10) 中 Rank1(assistant prof), Rank2(associate prof), Rank3 (full prof) 的效应分别是 $\gamma_1 \equiv 0, \gamma_2, \gamma_3$, 如果它们呈等差数列，那么我们将 Assistan P, Associate P, Full P 赋值为实数 1, 2, 3 就是合理的。为此，我们需要在模型 (10) 中检验

$$H_0 : \gamma_3 - \gamma_2 = \gamma_2 - \gamma_1,$$

即 $H_0 : \gamma_3 = 2\gamma_2$, 注意原假设下模型 (3) 变成

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \gamma_2 (\text{Rank2} + 2 \times \text{Rank3}) + \epsilon \tag{11}$$

因此需定义新变量 Rank.null = Rank2+2∗Rank3 = 0, 1, 2 (分别为对应于 Assistan P, Associate P, Full P). 使用 anova 函数比较模型 (4) 与模型 (3) 的拟合差别就是检验假设 $H_0 : \gamma_3 = 2\gamma_2$,

结果是否显著？模型 (3) 中检验 $H_0 : \gamma_2 = \gamma_3 = 0$ 与 (4) 中检验 $H_0 : \gamma_2 = 0$ 都是检验职称的显著性，这两个结果相比，那个更显著？

注：一般地, 我们认为被检验的参数个数越少，检验功效通常会越大. (b) 的分析过程中，将因子变量 Rank 当作数值变量的好处是减少了参数的个数，因而我们认为这可能导致我们有更高的检验功效。当然，这样做的前提是 Rank 的三个级别工资是近似呈等差数列的，也即在模型 (3) 中 $\gamma_3 \approx 2\gamma_2$.

3. 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

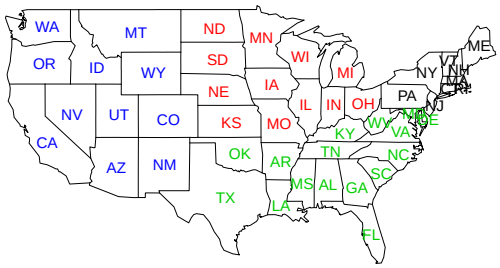
关心的问题是教育花费与其它变量的关系。

(a) 首先在地图上看看各个区的位置：美国地图可使用 R 包 maps 中的函数 map 以及 text 函数：

> map("state")

> text(state.center, state.abb, color=Region)

其中 state.center 是美国各州中心的坐标，state.abb 为州名缩写, 它们都在 R 自带的 package:datasets 中。而 Region 由本教育费用数据集提供。



(b) 使用回归模型

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

分析该数据，画出各个区的残差的盒型图 (boxplot), 检查误差方差齐性假设是否成立（提示：boxplot(resid(myfit) ~ Region) ）。

4. 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).