Lab-3: 多重线性回归模型 **2022** 年 **11** 月 **25** 日**-12** 月 **9** 日

内容: 主要内容分三节，分别是：
1. 多重线性回归模型与 F 检验（11 月 25 日-12 月 2 日）
2. 蒙特卡罗简介（选）
3. 广义最小二乘和迭代加权最小二乘（12 月 2 日-12 月 9 日）

任务: 重复例子中所提供的 R 代码命令（请手工输入命令!）, 完成 1、3 节中的练习 1-5.

1 多重线性回归模型与 F 检验

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/2022/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 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：

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

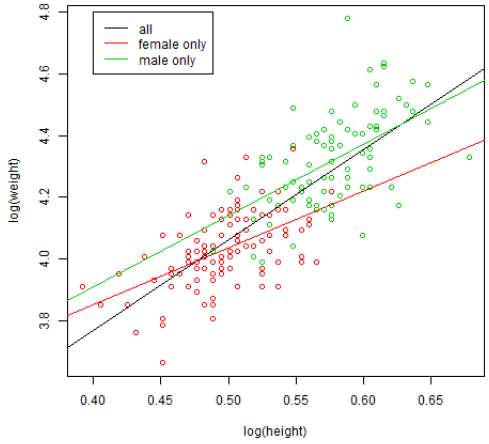
只对女性拟合模型：

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

```
> hw=read.table("http://staff.ustc.edu.cn/~ynyang/lm2020
/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)
> 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
```

从输出结果和下图看出，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，$\sigma_0 = 0.103 \sigma_1 = 0.127$。



1.2 多重回归

显然，上述例子中性别 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)：

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

在第一部分（简单回归）中我们已经发现 b_0 近似等于 b_1! 所以多重模型 (1) 基本可以说是一个合理的模型。所得到的拟合结果，在 log(height)-log(weight) 散点图上添加拟合回归直线:

```
> fit = lm(log(weight)~log(height)+sex, data=hw )
> coef(fit)
(Intercept) log(height) sex
3.0087 2.0572 0.1241

> abline(3.0087, 2.0572, col=2,lty=2,lwd=2) #for female, intercept=3.0087
> abline(3.0087 +0.1241, 2.0572, col=3, lty=2,lwd=2)
# for male, intercept=3.0087+0.1241
```

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：

```
> 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)
```

1.3 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
```

1.4 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 (3) 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 (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)
```

练习题

练习 1. 利用例 1 中两变量回归 log(weight)~ log(height)+sex 得到的结果，给出 BMI (weight/height^2) 的正常值区间 (20% -80%)。检验 $H_0 : b = 2$, 其中 b 是 log(height) 的回归系数。

提示：你可以通过在模型 (3) 中令 $b^* = b - 2$, 检验 $H_0 : b^* = 0$，也可以（等价地）使用 anova 函数比较全模型 (3) 和零模型

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

拟合该模型时，$2 \times \log(\text{height})$ 不是回归自变量（即不需要估计其回归系数），需要用 offset 指定：

```
fullmodel=lm(log(weight)~sex+log(height),data=hw)
nullmodel=lm(log(weight)~sex+offset(2*log(height)),data=hw)
anova(nullmodel,fullmodel)
```

练习 2. 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。

(1) 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

(2) 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.

(3) 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))

(4) 应用多重线性回归模型

$$\text{Salary} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Rank} + \beta_3 \text{Year} + \beta_4 \text{Degree} + \beta_5 \text{YSdeg} + \epsilon \tag{6}$$

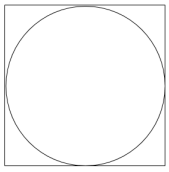
检验模型 (6) 中 $H_0 : \beta_1 = \beta_2 = 0$.

---

2 蒙特卡洛简介 (选)

蒙特卡洛方法 (Monte Carlo method，MC), 也称统计模拟方法, 通过计算机产生随机数模仿真实抽样过程，通过反复计算机抽样求解难以理论计算的问题（下面的 (a)-(d)），或对某一个数据分析方法进行评估（下面的 (e)）。MC 方法的理论依据是大数定律。

1. 假设我们不知道单位圆的面积公式。若随机变量 X 服从边长为 2 的正方形内的均匀分布，则它落在内切单位圆内的概率 p 为单位面积 (S) 与正方形面积 (4) 之比。如果我们知道该比率 p, 那么单位圆面积就是 $4p$。为了计算概率 p, 我们在计算机上产生大量 (比如 $n = 100000$ 个) 正方形内的均匀随机数，统计落在单位圆内的点的个数 m, 由大数定律知 $m/n \to p, n \to \infty$, 即当 n 足够大时我们可用比例 m/n 作为概率 p 的估计。



```
n=100000
x=runif(n,-1,1); y=runif(n,-1,1)
m=sum(x^2+y^2<=1) # 落在单位圆内的点的个数
p=m/n # 落在单位圆内的点数的比例
S=4*p # 单位圆面积 S=4*p
```

2. 假设 $X \sim N(0,1)$, $E\sqrt{|X|}$ =?$var(\sqrt{|X|})$ =?。 $\sqrt{|X|}$ 分布是否近似正态？

```
x=rnorm(n)
mean(sqrt(abs(x))) #0.822
var(sqrt(abs(x))) #0.122
hist(x) # 近似正态 N(0.822, 0.122)
```

3. $I = \int_{-\infty}^{\infty} \exp(-x^4)dx$ =?

   记 $\phi(x)$ 为标准正态分布的密度函数，改写上述积分为

$$I = \int_{-\infty}^{\infty} \Big\{\exp(-x^4)/\phi(x)\Big\}\phi(x)dx = \int_{-\infty}^{\infty} \Big\{\sqrt{2\pi}\exp(-x^4+x^2/2)\Big\}\phi(x)dx$$

   转化成了求解某个正态随机变量函数的期望的问题，这称为 Importance sampling (当然可以选择其它支撑在整个实数轴上的其它分布而不一定是正态分布)。所以我们只需对 $X \sim N(0,1)$，估算 $E\Big\{\sqrt{2\pi}\exp(-X^4+X^2/2)\Big\}$：

```
x=rnorm(n)
y=sqrt(2*pi)*exp(-x^4+x^2/2)
mean(y)  #I=1.813
```

4. 计算 $I = \frac{1}{\sqrt{2\pi}}\int_5^{\infty} x^2\exp(-x^2/2)dx$ =?

   我们可以产生标准正态随机数 $x_1,...,x_n$ iid $\sim N(0,1)$, 由大数定律

$$\sum_{i=1}^{n} x_i^2 I_{(x_i>5)}/n \to I, \quad \text{as } n \to \infty$$

   因为从 $N(0,1)$ 抽到大于 5 的随机数的可能性非常小，所以除非模拟次数 n 非常大，否则我们很少能得到大于 5 的随机数，从而上述逼近非常不准确。我们可以采用 Importance sampling 方法，用一个在区间 $[5,\infty]$ 上容易采到样本的分布，比如 $N(5,1)$ 分布：$f(x) = \frac{1}{\sqrt{2\pi}}\exp(-(x-5)^2/2)$,

$$I = \int_5^{\infty} x^2\exp(-x^2/2+(x-5)^2/2)f(x)dx \triangleq \int_5^{\infty} h(x)f(x)dx$$

   所以我们可产生 $x_1,...,x_n$ iid $\sim N(5,1)$, 并计算 $\sum_{i=1}^{n} h(x_i)I_{(x_i>5)}/n \approx I$。

```
> n=100000
> x=rnorm(n) #N(0,1)
> sum (x[x>5]^2)/n
[1] 0
> x=rnorm(n) +5 #N(5,1)
> h=x^2*exp(-x^2/2+(x-5)^2/2)
> sum(h[x>5])/n
[1] 7.766223e-06
```

5. 对于简单线性模型 $y_i = a + bx_i + \epsilon_i, \epsilon \sim (0, \sigma^2)$, $i = 1, ..., n$, b 的 LS 估计为

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{7}$$

其方差为

$$\text{var}(\hat{b}) = \sigma^2 / \sum (x_i - \bar{x})^2 \tag{8}$$

假设我们有一组常数 $z_1, ..., z_n$，比如 $z_i = i/n$, 构造 b 的另一个估计

$$\tilde{b} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})y_i}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}. \tag{9}$$

该估计是否是 b 的无偏估计？其方差比 $\hat{b}$ 的方差大多少？(假设我们不知道如何证明).

我们可以通过模拟，研究 $\tilde{b}$ 的性质。当然模拟研究不可能模拟所有情况，我们只需模拟一种或几种情况即可。比如，我们假设 $x_i = 1, i \le n/2$, $x_i = 0, i > n/2$，误差从正态分布产生。指定真正的参数 $a = 2, b = 1, \sigma = 1, n = 100$。下面我们产生 $N = 10000$ 次数据，每次计算估计 $\tilde{b}$, 共得到 $n = 10000$ 个估计（存放于 B.tilde），这些 B.tilde 可看作是 $\tilde{b}$ 的 N 个随机样本，计算它们的样本均值 m 和样本方差 v，如果 m（近似）等于真正的参数 $b = 1$，则说明至少在我们模拟的情景下 $\tilde{b}$ 是无偏的；如果 v 大于 (3) 式的方差，则说明 $\tilde{b}$ 不如 LS 估计精确（注意这是模拟实验，不是理论证明！但模拟实验确实能发现一些事实）。

```
a=2; b=1; sigma=1
n=100
x=c(rep(1,n/2), rep(0, n/2))
z=(1:n)/n

## 下面反复产生 N 次数据，每次都计算 b.hat,b.tilde,
N=10000
B.hat=B.tilde=NULL #all b.hat, all b.tilde

for (k in 1:N){
        epsilon=rnorm(n)*sigma # 产生误差项
        y=alpha+beta*x+epsilon # 产生响应 y
        # 下面基于产生的数据 (y,x) 计算 LS 估计 b.hat 和新的估计 b.tilde
        x.bar=mean(x); z.bar=mean(z)
        b.hat=sum((x-x.bar)*y)/sum((x-x.bar)^2)
        b.tilde=sum((z-z.bar)*y)/sum((z-z.bar)*(x-x.bar))
        B.hat=c(B.hat, b.hat)
        B.tilde=c(B.tilde, b.tilde)
} #end of simulation

mean(B.hat)
m=mean(B.tilde) # 是否等于真参数 b=1?

var(B.hat) # 是否等于 (2) 式的理论结果，即
v.true=sigma^2/sum((x-x.bar)^2)

v=var(B.tilde) # 方差比 b.hat 的方差大多少倍?
v/v.true
```

3 广义最小二乘、迭代加权最小二乘

加权最小二乘 (GLS/WLS)

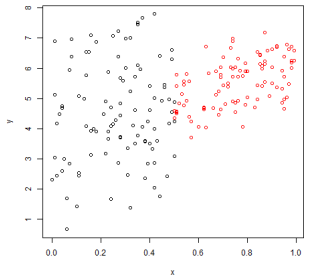
lm(y~x, weights=...) #WLS。weights：权重，缺省为等权

练习题

练习 3（WLS，加权最小二乘）数据集 cp.txt（`http://staff.ustc.edu.cn/~ynyang/2022/lab/cp.txt`），我们研究两个变量 x, y 之间的关系。x-y 散点图表明误差方差不是常数。假设线性模型

$$y_i = a + bx_i + \epsilon_i, var(\epsilon_i) = \sigma_i^2, i = 1, ..., 200$$

假设 $\sigma_1^2 = ... = \sigma_m^2 = \sigma^2$, $\sigma_{m+1}^2 = ... = \sigma_n^2 = \tau^2$.



假设我们已知 $m = 100, \tau^2 = \sigma^2/4$ 试求 a, b 的加权 LS 估计（lm 函数中指定 weights）。

练习 4 (转变点检测) 假设 cp 数据中转变点 $2 \le m \le n-2$ 未知，同时 a, b, σ^2, τ^2 也是未知的 (不假设 $\tau^2 = \sigma^2/4$)。假设误差服从正态分布，试设计一个算法，极小化目标函数

$$-2\log L = \sum_{i=1}^{m}(y_i - a - bx_i)^2/\sigma^2 + \sum_{i=m+1}^{n}(y_i - a - bx_i)^2/\tau^2 + m\log(\sigma^2) + (n-m)\log(\tau^2).$$

给出 $a, b, \sigma^2, \tau^2, m$ 的估计。

练习 5 (最小一乘的 IRLS 解法) 最小一乘方法比最小二乘法稳健，它对异常的响应值不太敏感。对于线性模型

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

最小一乘法极小化

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

改写为加权最小二乘的形式

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

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

任取 β 初始值, 比如 $\beta_0 = \hat{\beta}_{OLS}$, IRLS 方法反复迭代如下两步：

$$\beta_{new} = argmin_\beta \sum_{i=1}^{n} w_i(\beta_0)|y_i - \beta^\top \mathbf{x}_i|^2$$
$$\beta_0 \leftarrow \beta_{new}$$

(a) 考虑 alr4 中的 brains 数据集, y =log(BrainWt), x =log(BodyWt)，假设模型

$$y_i = a + bx_i + \epsilon_i, \epsilon \sim (0, \sigma^2)$$

求 a, b 最小二乘估计。画出残差图 (x-轴为 x =log(BodyWt), y-轴为残差)，该图有无非线性趋势？方差齐性假设是否合理？

(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。与 (a) 中结果相差大吗？

(c) 写一个基于 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)、(b) 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影响大吗？(提示：你应该得到如下拟合结果)