Lab-4: **2023** 年 **12** 月 **22** 日-**2024** 年 **1** 月 **5** 日

内容: 广义或加权最小二乘（GLS/WLS），迭代加权最小二乘 IRLS，变量选择，预测和交叉验证。
任务: 练习 1-4。

1 广义/加权最小二乘和迭代加权最小二乘

例 1 (**加权最小二乘** WLS). 数据集 cp.txt （`http://staff.ustc.edu.cn/~ynyang/2023/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 $m = \sigma^2/4$, 且 $\tau^2 = \sigma^2/4$, 此时我们可以应用加权最小二乘法 WLS 求解参数。WLS 估计的求解依然调用函数 lm, 但需要指定权重 weights:

```
cp=read.table("http://staff.ustc.edu.cn/~ynyang/2023/lab/cp.txt",head=T)
plot(cp)
n=nrow(cp)
m=100
w=c(rep(1, m),rep(4, n-m))
lm(y~x, weights=w,data=cp)
```

例 1（**续，迭代加权最小二乘** IRLS). 假设 $m = 100$，τ^2 未知，记 $\boldsymbol{\beta} = (a, b)^\top, \boldsymbol{\theta} = (\sigma^2, \tau^2)^\top$ 假设误差服从正态分布，极大似然法极小化目标函数

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

我们可直接极大上述目标函数，也可采用迭代加权最小二乘法 IRLS 求解 $\boldsymbol{\beta}$ 和 $\boldsymbol{\theta}$，即

(1) 对于给定的 $\boldsymbol{\theta}$，$\Sigma = diag(\sigma^2, ..., \sigma^2, \tau^2, ..., \tau^2)$ 应用 WLS 求得 WLS 估计

$$\hat{\boldsymbol{\beta}} = (\hat{a}, \hat{b})^\top = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \mathbf{y}$$

(2) 计算残差 $e_i = y_i - \hat{a} - \hat{b}x_i$，利用该残差估计 $\boldsymbol{\theta} = (\sigma^2, \tau^2)$:

$$\hat{\sigma}^2 = \sum_{i=1}^{m} e_i^2/m, \quad \hat{\tau}^2 = \sum_{i=m+1}^{n} e_i^2/(n-m)$$

例如:

```
y=cp[,2]
x=cp[,1]
X=cbind(1, x)
n=nrow(cp)
m=100
beta=beta.initial=c(1,1)
repeat {
       e=y-X%*%beta
       sigma2=sum(e[1:m]^2)/m
       tau2=sum(e[(m+1):n]^2)/(n-m)
       Sigma.inv=diag( c(1/rep(sigma2,m),1/rep(tau2,n-m)) )
       beta.update=solve(t(X)%*%Sigma.inv%*%X)%*%t(X)%*%Sigma.inv%*%y
       delta = sum(abs(beta.update-beta))
       print(delta)
       beta=beta.update
       if (delta<1e-9) break
}
print(beta)
```

练习 1.（可选）如果例 1 中 m 也是未知的，我们也需要估计转变点 m, 称为转变点 (change-point) 估计问题。例 1(续) 中的目标函数与未知量 m 有关，记作 $Q(m,\boldsymbol{\beta},\boldsymbol{\theta})$。类似于 Box-Cox 变换方法中的剖面似然法，对于任一固定的正整数 $2 \le m \le n-2$, 我们可以用例 1(续) 的方法得到估计 $\hat{\boldsymbol{\beta}}(m), \hat{\boldsymbol{\theta}}(m)$, 代入 Q 中

$$q(m) = Q(m, \hat{\boldsymbol{\beta}}(m), \hat{\boldsymbol{\theta}}(m))$$

然后极大化 $q(m)$，即比较 $q(2),...,q(n-2)$ 得到 m 的极大点 $\hat{m}$。

练习 2 (最小一乘的 IRLS 解法) 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

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

得到的估计称为最小一乘估计，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

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

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

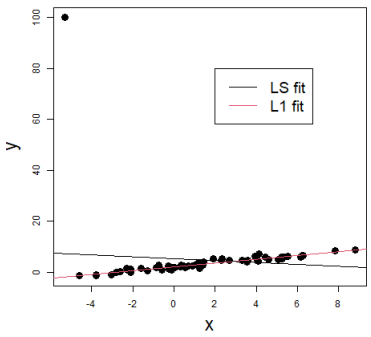
任取 β 初始值, 比如 $\beta_0 = \hat{\beta}_{OLS}$, IRLS 方法反复迭代如下两步：

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

(a) 对于 alr4 数据集 *brains*, 考虑线性模型 $\log(BrainWt) = a + b\log(BodyWt) + \epsilon$, 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）



2 变量选择与逐步回归介绍

回归分析中，如果有 $p-1$ 个自变量，每个自变量都可以选入回归模型也可以被排除在外，所以共有 2^{p-1} 个子模型。因为自变量越多不一定预测效果越好，所以我们有必要选择一个具有较少变量个数的子模型，使其预测精度最高。这个过程称为 “变量选择 (variable selection)”。通常的变量选择方法是选择使得某个变量选择准则达到最小的模型，常用的选择变量准则包括：Mallow C_p 准则,AIC 准则 (Akaike's information criterion) 或 BIC 准则 (Bayes information criterion)，这些准则都在某种意义下度量含 k 个自变量的模型的预测误差或者均方误差：

$$C_k = RSS_k/\hat{\sigma}^2 - n + 2k$$

$$AIC_k = n\log(RSS_k) + 2k, \quad BIC_k = n\log(RSS_k) + \log(n)k$$

其中 n 是样本量，RSS_k 是使用 k 个变量回归时的残差平方和。

在 p 较大的时候，比较所有 2^{p-1} 个子模型的计算复杂度是相当大的，普通计算机难以完成。逐步回归是一种快速的变量选择方法，它在特定准则（比如 AIC 准则）下，从全模型开始，逐步删除一个变量（或者从零模型开始，逐步增加变量），直至收敛，这种方法只需做 $p-1$ 次模型的比较。

函数 *step* 使用逐步回归方法搜索 C_p 或 AIC 或 BIC 最小 (或接近于最小) 的选变量模型。

```
fit.full = lm(response~., data=...)
step(fit.full, scale=sigma) # Mallow's Cp
step(fit.full, k=2) # AIC, default value of k is 2
step(fit.full, k=log(n)) # BIC, n: sample size
```

这里 *fit.full* 是包含所有自变量的全模型拟合结果，函数 *step* 中指定 *scale*=σ (全模型误差的标准差估计) 时，变量选择准则是 Mallow's C_p。*scale* 的缺省值为 0, 对应于 AIC ($k=2$) 或 BIC ($k=\log n$). *step* 输出结果给出的是最终选取的模型及其 LS 拟合结果。

例 2. alr4 程序包中的数据集 *Highway* 给出了汽车交通事故发生率与一些潜在可能有关的因素的资料。数据包含 1973 年明尼苏达州 $n = 39$ 个高速公路路段的有关信息。

变量	描述
rate	事故发生率（每百万里程）
len	路段长度 (单位：英里)
adt	平均每天通过车辆个数（average daily traffic, 单位：千辆）
trks	通过卡车的比例
slim	时速限制（Speed limit）
shld	行车路线边线之外的部分的宽度 (width of shoulder, 单位：英尺)
sigs	有信号灯的交叉口的个数（单位：个/英里）
lane	车道 (lane) 数目
lwid	车道宽度（单位：英尺）
Itg	每英里高速路交叉口 (interchanges) 个数
acpt	每英里入口 (access points) 个数
htype	路的类型或资助来源。"fai": 联邦高速 Federal interstate highways, "pa": 重要主干道 principal arterial highway, "ma": 主干道 major arterial highways, "mc": 其它 major collector,

本问题的主要目的是研究交通部门可以控制的变量 *Acpt, Slim, Sigs, Shld* 与事故率的关系, 其它变量不是交通部门能够掌控的，但可以作为协变量包括进模型。我们考虑变量选择以期望模型能够较好地预测事故发生率。

```
> fit =lm(rate~., data=Highway) #全模型
> vs = step(fit, k=2) #AIC选择变量，选出的模型如下：
> vs
 Call:
 lm(formula = rate ~ acpt + sigs + slim + len, data = Highway)

 Coefficients:
 (Intercept) acpt sigs slim len
 8.81443 0.08940 0.48538 -0.09599 -0.06856

 ## 在至少包含变量acpt+ slim+ sigs+ shld前提下选择变量：
> step(fit , scope =list(lower=rate~acpt+ slim+ sigs+ shld, upper=Rate~.) )
```

练习 3. 例 2 中使用 BIC 准则选择变量。

3 预测

岭估计是一种压缩估计, 对于线性模型 $Y = X\beta + \epsilon$, 其中 β 长度为 p, 岭估计定义为

$$\widehat{\beta}(\lambda) = (X'X + \lambda I_p)^{-1} X'Y,$$

其中 $\lambda \geq 0$ 为常数 ($\lambda = 0$ 时，即为 LS 估计)。恰当选取 λ 可使得岭估计具有较好的预测效果，所以 λ 称为可调节参数 (tuning parameter), 因为岭估计预测误差或者均方误差没有简单的表达，最优的 λ 没有显式解或者容易计算的恰当的估计，我们下面介绍的交叉验证方法 (cross-validation, CV) 可用来搜索最优的 λ 值使得预测效果最好。

交叉验证 (Cross Validation, CV)

理论上基于 C_p，AIC，BIC 的变量选择方法在全模型是正确的假设下，选出的模型具有最优或接近最优的预测准确度，同时这些方法也受限于模型假设的正确性。对于一般的预测问题，特别是对于线性模型假设未必正确的问题，变量选择方法不再是一个好的方法。

一般来讲，预测准确度应以预测方法在测试数据上的预测效果来判定（而不是由理论假设下得到的 AIC、BIC、Cp 等决定）。用于拟合模型的数据集称为训练数据集 (training sample)，用于测试预测效果的数据集称为测试数据集 (testing sample), 后者不应该用于训练拟合。因此，在构建预测方法时，我们应该预留部分数据用做测试。为了使得预测方法具有较强的泛化能力，我们希望每个样本点都应该有机会被用作测试数据。交叉验证 (CV) 方法就是一种数据点轮流用于训练和测试的方法，也是通用的选择模型和变量的方法。

Leave-one-out cross-validation

Leave-one-out 方法是最简单的 CV 方法，假设有 n 个数据点 $(y_1,x_1),...,(x_n,y_n)$, 对每个 $i=1,2,...,n$，我们将数据点 (x_i,y_i) 用于测试而 (x_i,y_i) 之外的其它 $n-1$ 个数据点作为训练集，训练得到的 y_i 的预测变量记为 $\hat{y}_i^{(-i)}=f^{(-i)}(x_i)$，则预测误差为 $\hat{y}_i^{(-i)}-y_i$, 预测误差平方和（predicted residual error sum of squares，PRESS）为

$$PRESS=\sum_{i=1}^{n}(\hat{y}_i^{(-i)}-y_i)^2.$$

因为所有的训练集基本相同，通常认为 PRESS 不是一个很好的度量.

K-fold cross-validation(K-*重交叉验证*)

K-重 CV 方法 (K-fold CV)，将数据集 $(y_i,\mathbf{x}_i), i=1,...,n$ 划分为大小近似的 K 个子集，通常 $K=5$ 或 10。与 Leave-one-out cross-validation 类似，依次将其中一个子集作为测试集，其余的用于训练/拟合，得到预测方法并对测试集的响应进行预测。K-CV 算法如下：

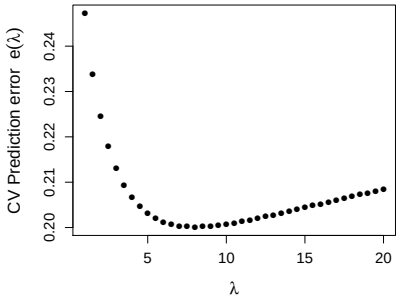
记所有响应变量组成的向量为 Y, 设计阵为 X, 假设线性模型 $Y=X\beta+\epsilon, \epsilon\sim(0,\sigma^2 I_n$.

1. 划分数据标号 $S=\{1,...,n\}$ 成 K 个子集，$\{S_1,...,S_K\}$, 每个子集的大小大约为 n/K.
2. 对每个 $k=1,...,K$,
   (a) 基于训练集 X_{-S_k},Y_{-S_k} 得到 β 的估计 $\widetilde{\beta}^{(-S_k)}$ (未必是 LS 估计)，计算测试集中响应变量 Y_{S_k} 的预测 $\widetilde{Y}_{S_k}=X_{S_k}\widetilde{\beta}^{(-S_k)}$.
   (b) 计算平均预测误差平方和 $e_k=\|Y_{S_k}-\widetilde{Y}_{S_k}\|^2/|S_k|$ 其中 $|S_k|$ 为集合 S_k 的元素个数。
3. 计算总的平均预测误差平方和：$e=\sum_{k=1}^{K}e_k/K$.

岭估计中参数 λ 的选取

现在我们使用 K-CV 方法选取使得平均预测误差平方和最小的调节参数 $\lambda_{optimal}$。对任意给定的 $\lambda>0$，我们可以得到平均预测误差平方和 $e(\lambda)$ (CV 最后一步得到的 error)，得到如下图所示的散点图，

并由此得到最优 $\lambda_{optimal}$



岭估计调节参数的 CV 选取算法:

记所有响应变量组成的向量为 Y, 设计阵为 X, 假设线性模型 $Y = X\beta + \epsilon, \epsilon \sim (0, \sigma^2 I_n)$. 划分数据标号 $S = \{1, ..., n\}$ 成 K 个子集: $\{S_1, ..., S_K\}$, 每个子集的大小大约为 n/K.

对 $\lambda = 0.1, 0.2, 0.3, ...$

对每个 $k = 1, ..., K$,

基于训练集 X_{-S_k}, Y_{-S_k} 得到 β 的岭估计 $\widetilde{\beta}^{(-S_k)} = (X'_{-S_k} X_{-S_k} + \lambda I_p)^{-1} Y_{-S_k}$，计算测试集中响应变量 Y_{S_k} 的预测 $\widetilde{Y}_{S_k} = X_{S_k} \widetilde{\beta}^{(-S_k)}$. 这里 $-S_k$ 表示除了 S_k 之外的其它数据标号。计算平均预测误差平方和 $e_k(\lambda) = \|Y_{S_k} - \widetilde{Y}_{S_k}\|^2 / |S_k|$. 其中 $|S_k|$ 为集合 S_k 的元素个数。

计算总的平均预测误差平方和：$e(\lambda) = \sum_{k=1}^{K} e_k / K$.

取 $\hat{\lambda} = \text{argmin } e(\lambda)$。

练习 4 下面使用数据集 houston (`http://staff.ustc.edu.cn/~ynyang/2023/lab/houston_train.txt`) 练习使用 CV 方法确定岭估计中的最优 λ. 该数据给出了 Houston 2006 年 1500 个小区房产价格数据。变量如下：

变量	描述
price	2006 年小区房价中位数, 基于售出的 *numbber* 套房子。单位：\$/foot^2
number	售出的房子数目
new	2006 年小区售出的房产中新房所占百分比
foreclosure	2006 年小区售出的房产数拍卖所占的百分比

(a) 试使用 10-fold 交叉验证，确定岭估计中的最优的 λ 值，其中，你可能需要考虑变量的变换、删除异常点、选择变量等等。对于你得到的最优值 λ，基于所有数据，计算岭估计 $\widetilde{\beta}_{ridge}$.

(b) 你的预测模型可在下述测试集上进行验证：(`http://staff.ustc.edu.cn/~ynyang/2023/lab/houston_test.txt`)，该数据是 422 个小区的自变量数据 (X_{test}, 422×3 矩阵)，不含价格数据。试用你的 $\widetilde{\beta}_{ridge}$ 值预测测试集的 422 个小区的价格: $\widetilde{\mathbf{y}}_{test} = X_{test} \widetilde{\beta}_{ridge}$。

(c) 下载 (b) 中 422 个小区的价格的真实价格 $\mathbf{y}_{true}$（这些真实价格不要在 (a) 中的训练中使用）：`http://staff.ustc.edu.cn/~ynyang/2023/lab/price_test.txt`.
提交你的预测误差值 $\|\widetilde{\mathbf{y}}_{test} - \mathbf{y}_{true}\|^2$。