回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第23讲 广义最小二乘

2020.5.13

$$\mathrm{var}\left(\frac{\sum y_i/\sigma_i^2}{\sum 1/\sigma_i^2}\right) \le \mathrm{var}\left(\frac{\sum y_i}{n}\right), \mathrm{var}(y_i)=\sigma_i^2$$

例1. 若 $y_1,...,y_n$ 独立, $y_i \sim (\mu, \sigma_i^2)$, μ未知, σ_i^2已知, 则普通的最小二乘法 (OLS):

$$\hat{\mu}_{OLS} = \bar{y} = \arg\min_{\mu} \sum (y_i - \mu)^2$$

因为诸y_i方差不等，μ的估计中方差大的数据应该给予较小权重，事实上下述加权最小二乘所定义的加权估计$\tilde{\mu}$具有更好的性质(*BLUE*):

$$\tilde{\mu} = \frac{\sum y_i / \sigma_i^2}{\sum 1/\sigma_i^2} = \arg\min_{\mu} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma_i^2}$$

$$E(\tilde{\mu}) = \mu, \quad \text{var}(\tilde{\mu}) = \frac{1}{\sum 1/\sigma_i^2} \le \frac{\sum \sigma_i^2}{n^2} = \text{var}(\hat{\mu}_{OLS}),$$

实际上，$\tilde{\mu}$是BLUE.

目标函数 $\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma_i^2}$ 中为什么以 $\frac{1}{\sigma_i^2}$ 作为权？

- 当 $y_i \sim N(\mu,\sigma_i^2)$ 时， 似然函数 $L(\mu)=-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma_i^2}$， $\max L(\mu) \Leftrightarrow \min\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma_i^2}$。

  而当方差齐性 $\sigma_i^2 \equiv \sigma^2$ 的时候, $\max L(\mu) \Leftrightarrow \min\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} \Leftrightarrow \min\sum_{i=1}^{n}(y_i-\mu)^2$

- 以线性模型表示：$y_i=\mu+\varepsilon_i, \varepsilon_i \sim (0,\sigma_i^2)$, 令 $\tilde{y}_i=\sigma_i^{-1}y_i, \tilde{x}_i=\sigma_i^{-1}, \tilde{\varepsilon}_i=\sigma_i^{-1}\varepsilon_i \sim (0,1)$, 上述模型变换为如下无截距项的方差齐性的简单回归模型：

$$\tilde{y}_i=\tilde{x}_i\mu+\tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim (0,1)$$

LS估计为：$\tilde{\mu}=\frac{\sum\tilde{x}_i\tilde{y}_i}{\sum\tilde{x}_i^2}=\frac{\sum\sigma_i^{-2}y_i}{\sum\sigma_i^{-2}}$

LS目标函数：$\sum(\tilde{y}_i-\tilde{x}_i\mu)^2=\sum(\sigma_i^{-1}y_i-\sigma_i^{-1}\mu)^2=\sum(y_i-\mu)^2/\sigma_i^2$

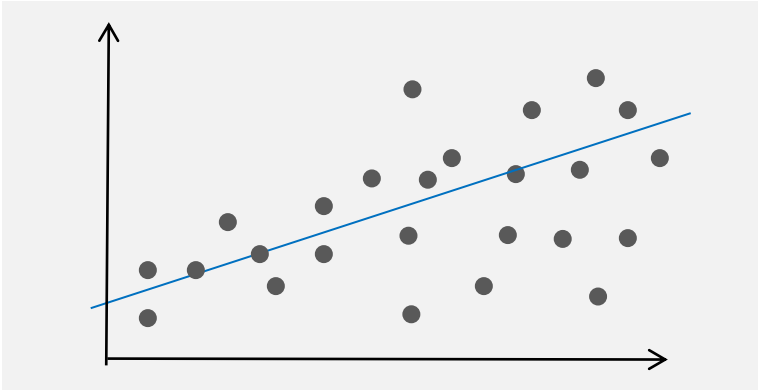
变换后的模型满足GM假设(方差齐性)，基于变换模型的 $\tilde{\mu}$ 是BLUE。

方差齐性与异方差

Gauss - Markov假设线性模型方差齐性：

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\text{与}X\text{独立},\ E(\boldsymbol{\varepsilon}) = \mathbf{0},\ \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

但有些情况下误差方差不是常数，甚至$\boldsymbol{\varepsilon}$的协方差矩阵非对角。



上图误差方差在某一个点(称为转变点,change - point)突然增大。误差方差随某些自变量变化，可能是由误差与自变量相依引起的，也可能是因变量的分布远非正态。

异方差（heteroscedasticity）模型：

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}$与$X$独立，$\boldsymbol{\varepsilon} \sim (0, G),\ G \neq \sigma^2 I_n$

回忆：LS方法与误差方差无关，且LS估计的无偏性不依赖于方差齐性假定.
如果在异方差模型中协方差部分或完全未知，我们仍可应用最小二乘法，
但LS估计未必是最优的 (BLUE依赖于方差齐性假设)。

换言之，如果在异方差模型拟合过程中考虑进方差不齐这个特点，我们
能得到比普通的LS估计更好的估计,即广义最小二乘估计 (GLS：
Generalized Least Squares),或加权LS估计 (WLS：Weighted LS).

下面我们首先考虑$G \neq \sigma^2 I_n$且G完全或几乎完全已知情形下的GLS，
然后考虑G仅含少数未知参数情形下的迭代加权最小二乘法(IRLS)。

广义最小二乘（GLS）

假设如下异方差模型：

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = G > 0 \ , \quad \boldsymbol{\varepsilon}\text{与}X\text{独立}$$

假设G已知.

广义最小二乘法 (GLS)：GLS极小化加权误差平方和

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} G^{-1} (\mathbf{y} - X\boldsymbol{\beta})$$

最优解称为GLS估计：$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (X^{\mathsf{T}} G^{-1} X)^{-1} X^{\mathsf{T}} G^{-1} \mathbf{y}.$

对加权误差平方和$(\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} G^{-1} (\mathbf{y} - X\boldsymbol{\beta})$求导，得正则方程

$$X^{\mathsf{T}} G^{-1} (\mathbf{y} - X\boldsymbol{\beta}) = 0,$$

$\Rightarrow$ GLS估计$\hat{\boldsymbol{\beta}}_{GLS} = (X^{\mathsf{T}} G^{-1} X)^{-1} X^{\mathsf{T}} G^{-1} \mathbf{y}.$

为什么GLS的目标函数具有形式$(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G^{-1}(\mathbf{y}-X\boldsymbol{\beta})$?

正态模型的$-\log L$

下面将异方差模型转化为方差相等的情形，我们将看到上述目标函数是转化后的等方差模型的误差平方和

方程 $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim(\mathbf{0},\ G)$, 两边左乘$G^{-1/2}$，得

$$G^{-1/2}\mathbf{y}=G^{-1/2}X\boldsymbol{\beta}+G^{-1/2}\boldsymbol{\varepsilon},$$

记$\tilde{\mathbf{y}}=G^{-1/2}\mathbf{y}$，$\tilde{X}=G^{-1/2}X$，$\tilde{\boldsymbol{\varepsilon}}=G^{-1/2}\boldsymbol{\varepsilon}$，得到如下方差齐性模型：

$$\tilde{\mathbf{y}}=\tilde{X}\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}},\quad \tilde{\boldsymbol{\varepsilon}}\sim(\mathbf{0},I_n), \qquad (*)$$

对模型(*)应用普通LS (OLS, Ordinary LS)，其误差平方和为

$$(\tilde{\mathbf{y}}-\tilde{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\mathbf{y}}-\tilde{X}\boldsymbol{\beta})=(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G^{-1}(\mathbf{y}-X\boldsymbol{\beta})$$

所以对模型(*)应用OLS方法（左端）$\Leftrightarrow$ 对原模型极小化右端函数。

模型(*)的LS解为$\hat{\boldsymbol{\beta}}_{GLS}=(\tilde{X}^{\top}\tilde{X})^{-1}\tilde{X}^{\top}\tilde{\mathbf{y}}$, 代入$\tilde{\mathbf{y}}=G^{-1/2}\mathbf{y}$，$\tilde{X}=G^{-1/2}X$，

$$\hat{\boldsymbol{\beta}}_{GLS}=(\tilde{X}^{\top}\tilde{X})^{-1}\tilde{X}^{\top}\tilde{\mathbf{y}}=(X^{\top}G^{-1}X)^{-1}X^{\top}G^{-1}\mathbf{y}$$

命题1. $E(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid X)=\boldsymbol{\beta}$，$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid X)=(X^{\top}G^{-1}X)^{-1}$,
$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$是$\boldsymbol{\beta}$的最优线性无偏估计(*BLUE*)。

证明：$E(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid X)=E((X^{\top}G^{-1}X)^{-1}X^{\top}G^{-1}\mathbf{y}\mid X)=(X^{\top}G^{-1}X)^{-1}X^{\top}G^{-1}X\boldsymbol{\beta}=\boldsymbol{\beta}.$

$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\mid X)=(X^{\top}G^{-1}X)^{-1}X^{\top}G^{-1}\,\mathrm{var}(\mathbf{y}\mid X)G^{-1}X(X^{\top}G^{-1}X)^{-1}=(X^{\top}G^{-1}X)^{-1}.$

对模型(*),有GM定理，$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$是BLUE，则对原异方差模型它也是BLUE。

GLS与OLS

线性回归模型的等方差假设不是本质的，即使不满足该假设，OLS估计也是无偏的，但OLS估计的方差偏大。

如果对原模型应用通常的最小二乘法(*OLS*)得到的解记为$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} = \operatorname{argmin} \| \mathbf{y} - X\boldsymbol{\beta} \|^2$$

容易验证：

$$E(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \mid X) = \boldsymbol{\beta}, \quad \operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \mid X) = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}GX(X^{\mathsf{T}}X)^{-1}$$

因为$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$是BLUE，则必有$\operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} \mid X) \geq \operatorname{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} \mid X)$，即

$$(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}GX(X^{\mathsf{T}}X)^{-1} \geq (X^{\mathsf{T}}G^{-1}X)^{-1}$$

GLS未必一定要求G完全已知

如果*GLS*中误差协方差矩阵除了一个常数倍数外已知，即

$$G=\sigma^2 G_0,\ \text{其中}G_0\text{已知},\sigma^2\text{未知},$$

那么GLS方法仍可用。此时 $(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G^{-1}(\mathbf{y}-X\boldsymbol{\beta})=(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G_0^{-1}(\mathbf{y}-X\boldsymbol{\beta})/\sigma^2,$

$$\min_{\boldsymbol{\beta}}(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G^{-1}(\mathbf{y}-X\boldsymbol{\beta})\Leftrightarrow\min_{\boldsymbol{\beta}}(\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}}G_0^{-1}(\mathbf{y}-X\boldsymbol{\beta})$$

$\Rightarrow GLS$ 估计为 $\hat{\boldsymbol{\beta}}_{\text{GLS}}=(X^{\mathsf{T}}G_0^{-1}X)^{-1}X^{\mathsf{T}}G_0^{-1}\mathbf{y}$，$G=\sigma^2G_0$中的未知参数$\sigma^2$不起作用。

- 最重要的例子是等方差情形：$G=\sigma^2 I_n$,即通常的GM假设。此时GLS等价于OLS。
- 另外一个重要的例子为$G=\text{var}(\boldsymbol{\varepsilon})=\sigma^2 diag(1/w_1,...,1/w_n)$,其中$w_1,...,w_n$已知.
  此时，GLS也称为加权最小二乘 (WLS, Weighted LS)

WLS与抽样调查

在抽样调查cluster survey中，$\mathbf{x}_i$为第i个*cluster*的特征/自变量，y_i是基于*cluster*内m_i个个体的数据汇总，比如平均、中位数、极值等，则其方差通常具有如下形式

$$\text{var}(y_i \mid \mathbf{x}_i) = \sigma^2 / m_i,$$

则误差的方差-协方差矩阵为 $G = \text{var}(\boldsymbol{\varepsilon}) = \sigma^2 diag\left(\frac{1}{m_1}, \frac{1}{m_2}, ..., \frac{1}{m_n}\right)$

加权最小二乘（WLS, Weighted LS)：

假设模型$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, E(\boldsymbol{\varepsilon}) = 0,\ G = \text{var}(\boldsymbol{\varepsilon}) = \sigma^2 diag(1/w_1, ..., 1/w_n)$, 其中权重$w_1, ..., w_n$已知.

记$W = diag(w_1, ..., w_n)$即G_0^{-1}，加权最小二乘目标函数形式如下：

$$(\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} W (\mathbf{y} - X\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2 = \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i}\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{WLS} = (X^{\mathsf{T}} W\ X)^{-1} X^{\mathsf{T}} W \mathbf{y} = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}\right)^{-1} \left(\sum_{i=1}^{n} w_i \mathbf{x}_i y_i\right),$$

加权残差：$e_i = \sqrt{w_i}(y_i - \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{WLS})$

R：lm函数中指定weights：

$\text{lm}(y \sim x, \text{weights} = (w_1, w_2, ..., w_n))$

例1(续). $y_1,...,y_n$独立, $y_i \sim (\mu, \sigma_i^2)$, μ未知, 假设$\sigma_i^2 = \sigma^2 / m_i$, m_i已知,σ^2未知，

则GLS/WLS: $\tilde{\mu} = \dfrac{\sum m_i y_i}{\sum m_i}$为加权平均.

如果y_i是第i个*cluster*的组内平均，该估计为所有$\sum m_i$个样本点的平均（虽然没有cluster内的个体值）。

如果y_i是第i个*cluster*的组内中位数,分位数或其它汇总统计量，可仍然以$m_1,...,m_n$加权(极大或极小值的方差可能与m_i成反比，也可能不是)。

例2(异方差简单线性回归). 假设$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 / w_i), w_i$已知. *WLS*极小化

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{\sum w_i (x_i - \bar{x}_w) y_i}{\sum w_i (x_i - \bar{x}_w)^2}, \quad \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}; \quad \hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w, \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum w_i (x_i - \bar{x}_w)^2}, \quad \mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2}{\sum w_i} + \frac{\bar{x}_w{}^2 \sigma^2}{\sum w_i (x_i - \bar{x}_w)^2}$$

注意：如果将加权最小二乘目标函数写成普通最小二乘的形式(变换y, x)：

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i} \beta_0 - \beta_1 \sqrt{w_i} x_i)^2$$

记$\tilde{y}_i = \sqrt{w_i} y_i, \tilde{x}_i = \sqrt{w_i} x_i, \tilde{z}_i = \sqrt{w_i}$,上述右端平方和为$\sum_{i=1}^{n} (\tilde{y}_i - \tilde{z}_i \beta_0 - \tilde{x}_i \beta_1)^2$

对应的方差齐性模型为：$\tilde{y}_i = \tilde{z}_i \beta_0 + \tilde{x}_i \beta_1 + \tilde{\varepsilon}_i, \ \tilde{\varepsilon}_i \sim (0, \sigma^2)$,它不含截距项。

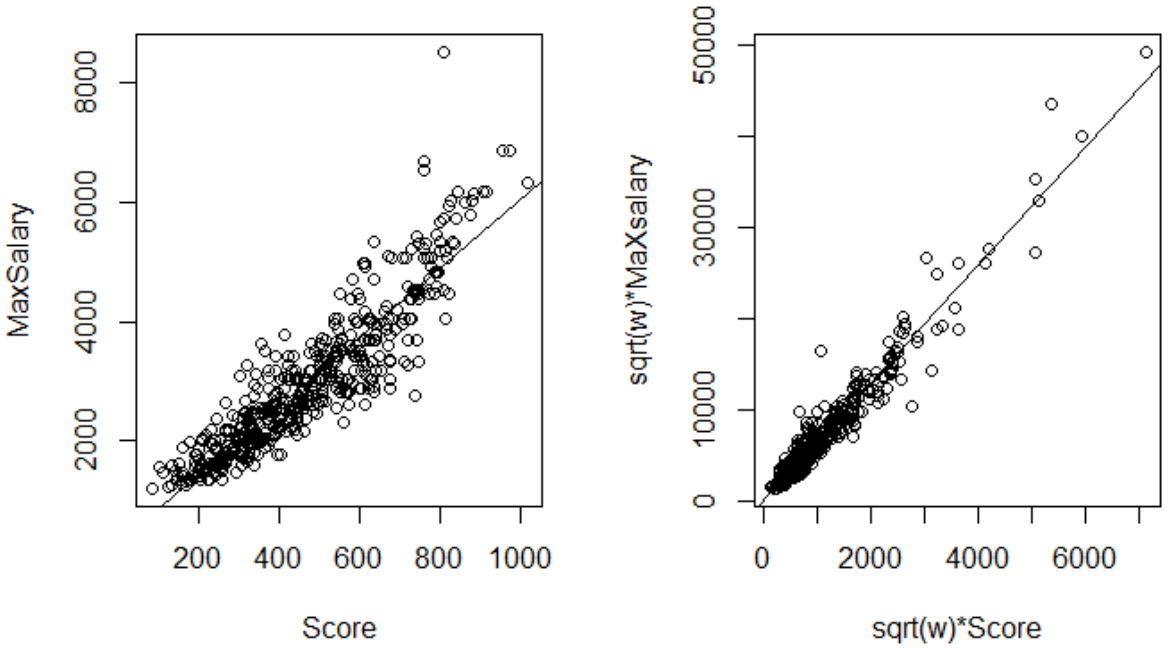
例3. 数据集 salarygov (alr3) 汇总了美国政府某部门495种职位的信息，把包括每种职位的最高工资、每种职位的人数、女性人数和难度系数(Score)。目的是研究工资与职位难度的关系。变量具体描述如下：

变量	描述
MaxSalary	职位最高工资
Score	职位难度系数（82-1017）
NE	该职位的雇员总数(number of employees)
NW	该职位的女性人数

JobClass	NW	NE	Score	MaxSalary
Account_clerk	52	68	258	1549
Account_clerk_Intermediate	26	29	269	1712
Account_clerk_Principal	10	13	321	2182
Account_clerk_Senior	16	24	273	1982
Accountant	1	12	352	2555
Accountant_Chief	0	5	709	4060
Accountant_Principal	0	4	505	3424
Accountant_Senior	2	18	404	3031
…				

$$WLS: \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2 = \sum_{i=1}^{n} (\sqrt{w_i} y_i - \sqrt{w_i} \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})^2$$

左图为(x_i, y_i)散点图，右图为$(\sqrt{w_i} x_i, \sqrt{w_i} y_i)$散点图. 显然右图有更好的线性性



迭代加权最小二乘方法(IRLS）

如果误差方差矩阵$G = \mathrm{var}(\boldsymbol{\varepsilon})$完全未知,则$G$不可估计；但如果$G = G(\boldsymbol{\theta})$只与少数未知参数$\boldsymbol{\theta}$有关，我们可以同时对$\boldsymbol{\beta}$和$\boldsymbol{\theta}$极小化目标函数：

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} (\mathbf{y} - X\boldsymbol{\beta})^{\mathsf{T}} G(\boldsymbol{\theta})^{-1} (\mathbf{y} - X\boldsymbol{\beta})$$

同时估计$\boldsymbol{\beta}$和$G = G(\boldsymbol{\theta})$一般不太容易。注意到如果$\boldsymbol{\theta}$已知,则$G$已知，应用GLS方法容易估计$\boldsymbol{\beta}$；而如果$\boldsymbol{\beta}$已知，则可利用残差估计误差G中的参数$\boldsymbol{\theta}$。优化问题可以分为两步反复迭代：

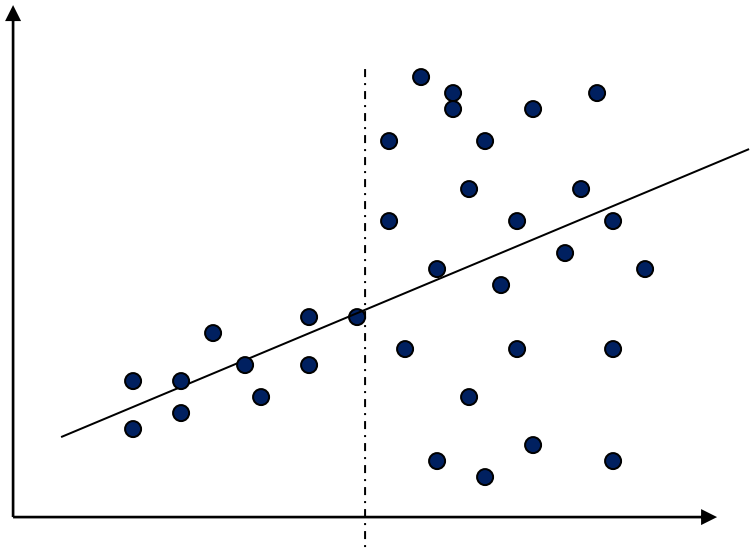
- 给定$\boldsymbol{\theta}$，即给定G，以GLS方法估计$\boldsymbol{\beta}$；
- 给定$\boldsymbol{\beta}$，使用残差更新G中的未知参数估计。

该方法称为迭代加权LS (IRLS, Iteratively Reweighted LS)。

例4. 对于下图所示的独立样本情形，异方差模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, G), G = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$$

假设$\sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2 = \sigma^2$(未知)，$\sigma_{m+1}^2 = ... = \sigma_n^2 = \tau^2$（未知），$m$已知。



$$GLS: \min_{\beta,\sigma^2,\tau^2} (\mathbf{y}-X\boldsymbol{\beta})^{\mathsf{T}} G^{-1} (\mathbf{y}-X\boldsymbol{\beta}) = \min_{\beta,\sigma^2,\tau^2} \left(\sum_{i=1}^{m} \frac{(y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{\sigma^2} + \sum_{i=m+1}^{n} \frac{(y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{\tau^2} \right)$$

IRLS解法:

Step0. $k=0$, 初始化$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$,

Step1. $k=k+1$, 计算残差$\mathbf{e}=\mathbf{y}-X\hat{\boldsymbol{\beta}}^{(k-1)}$, 并估计 σ^2 , τ^2 :

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

$$\Rightarrow G^{(k)} = diag(\hat{\sigma}^{(k)2}, ..., \hat{\sigma}^{(k)2}, \hat{\tau}^{(k)2} ..., \hat{\tau}^{(k)2})$$

Step2. 计算*GLS*估计$\hat{\boldsymbol{\beta}}^{(k)} = (X^{\mathsf{T}} G^{(k)^{-1}} X)^{-1} X^{\mathsf{T}} G^{(k)^{-1}} \mathbf{y}$,

goto Step1, 重复至收敛。

例4(续,转变点检测). 如果转变点m是未知的, 我们需要估计m，但例4中的加权平方和对任何m都是常数，故不能作为目标函数。为此假设误差正态，

$$-2\log L(\beta,\sigma^2,\tau^2,m)=\sum_{i=1}^{m}\frac{(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta})^2}{\sigma^2}+\sum_{i=m+1}^{n}\frac{(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta})^2}{\tau^2}+m\log(\sigma^2)+(n-m)\log(\tau^2)$$

它是在例4加权平方和基础上增加了惩罚项$m\log(\sigma^2)+(n-m)\log(\tau^2)$.

如果只是对$\boldsymbol{\beta}$优化，惩罚项$m\log(\sigma^2)+(n-m)\log(\tau^2)$与$\boldsymbol{\beta}$无关，有关的项即是GLS的目标函数。

对于任何给定的$1\le m\le n-1$,极小化$-2\log L(\beta,\sigma^2,\tau^2,m)$与例4实际上相同，可以应用IRLS方法求得最优解：

$$(\hat{\beta}_m,\hat{\sigma}_m^{\,2},\hat{\tau}_m^{\,2})=\arg\min_{\beta,\sigma^2,\tau^2}\left(\sum_{i=1}^{m}\frac{(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta})^2}{\sigma^2}+\sum_{i=m+1}^{n}\frac{(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta})^2}{\tau^2}+m\log(\sigma^2)+(n-m)\log(\tau^2)\right)$$

记上述最小值为$SS(m)=SS(m,\hat{\beta}_m,\hat{\sigma}_m^{\,2},\hat{\tau}_m^{\,2})$,则$m$的估计为

$$\hat{m}=\arg\min SS(m),$$

β,σ^2,τ^2的估计分别为$\hat{\beta}=\hat{\beta}_{\hat{m}},\hat{\sigma}^2=\hat{\sigma}_{\hat{m}}^{\,2},\hat{\tau}^2=\hat{\tau}_{\hat{m}}^{\,2}$.

注：关于目标函数

假设模型 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2)$

为了求$\boldsymbol{\beta}$的估计，最小二乘法极小化 $\| \mathbf{y} - X\boldsymbol{\beta} \|^2$,得到$\hat{\boldsymbol{\beta}}$之后定义$\hat{\sigma}^2 = \| \mathbf{y} - X\hat{\boldsymbol{\beta}} \|^2 / (n-p)$

事实上，我们极小化下述目标函数可同时得到$\boldsymbol{\beta}$和σ^2的估计

$$f(\boldsymbol{\beta}, \sigma^2) = \| \mathbf{y} - X\boldsymbol{\beta} \|^2 / \sigma^2 + n\log(\sigma^2)$$

误差正态时, $f = -2\log L$, 其中似然函数 $L = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \| \mathbf{y} - X\boldsymbol{\beta} \|^2 \right)$

$$\partial f / \partial \boldsymbol{\beta} = 2X^T(\mathbf{y} - X\boldsymbol{\beta}) / \sigma^2 = 0 \Leftrightarrow X^T(\mathbf{y} - X\boldsymbol{\beta}) \Rightarrow \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

$$\partial f / \partial \sigma^2 = -\| \mathbf{y} - X\boldsymbol{\beta} \|^2 / \sigma^4 + n / \sigma^2 = 0 \Rightarrow \hat{\sigma}^2 = \| \mathbf{y} - X\hat{\boldsymbol{\beta}} \|^2 / n.$$

因为 f 中第二项$n\log(\sigma^2)$与$\boldsymbol{\beta}$无关，如果重点是求解$\boldsymbol{\beta}$,我们可忽略$n\log(\sigma^2)$

以及f中第一项中的$1/\sigma^2$,只需极小化 $\| \mathbf{y} - X\boldsymbol{\beta} \|^2$,此即LS方法。

但例4（续）不能丢弃与log-方差有关的惩罚项。

例5 假设模型$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim(0,G)$,其中$G=\sigma^2\left(\rho^{|i-j|}\right)_{1\le i,j\le n}$

误差满足1阶自回归AR(1)模型：$\varepsilon_{i+1}=\rho\varepsilon_i+\delta_i,\delta_i\sim(0,\sigma^2)$与$\varepsilon_i$独立

$\boldsymbol{\beta}$的初始估计可取为OLS估计$\tilde{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$，

IRLS反复迭代下面两步：

- 对于给定的$\tilde{\boldsymbol{\beta}}$，$\mathbf{e}=\mathbf{y}-X\tilde{\boldsymbol{\beta}}$，

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

$$\hat{G}=\hat{\sigma}^2\left(\hat{\rho}^{|i-j|}\right)_{1\le i,j\le n}$$

- 给定$\hat{G}$，GLS方法重新估计$\boldsymbol{\beta}$：$\tilde{\boldsymbol{\beta}}=(X^{\mathsf{T}}\hat{G}^{-1}X)^{-1}X^{\mathsf{T}}\hat{G}^{-1}\mathbf{y}$

WLS用于优化非线性回归

许多统计问题的优化问题可转化为如下加权最小二乘问题（比如稳健回归、广义线性模型）：

样本为$(y_i,\ \mathbf{x}_i)$，$i=1,...,n$

$$\min \sum w_i(\boldsymbol{\beta},\theta)\left(y_i-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2$$

其中一般θ代表与方差有关的参数。

$$IRLS:\boldsymbol{\beta}^{(k)}=\arg\min_{\boldsymbol{\beta}} \sum w_i(\boldsymbol{\beta}^{(k-1)},\theta^{(k-1)})\left(y_i-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2$$

例6 (线性模型的M估计方法 - 稳健回归). 模型 : $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$,

$$\min \sum \rho(y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})$$

$\rho(\cdot) \geq 0$关于0对称，$\rho(0) = 0.$

特别，当$\rho(t) = |t|^p$ 时 :

$$\min \| \mathbf{y} - X\boldsymbol{\beta} \|_p = \min \sum | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^p$$

称为最小L_p方法。$p = 1$时称为最小一乘法(LAD : least absolute deviation).

最小一乘法比较稳健，它不太受异常大或异常小的响应变量值的影响。

例如，对于模型 $\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$, 最小一乘估计为中位数

$$m = \text{median}(y_1, .., y_n) = \min \sum | y_i - \mu |$$

中位数m比LS估计$\bar{y}$稳健。

对于最小L_p方法, 改写目标函数 :

$$\sum | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^p = \sum w_i(\boldsymbol{\beta}) | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^2, \text{ 其中} w_i(\boldsymbol{\beta}) = | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^{p-2} .$$

$$IRLS : \ w_i(\boldsymbol{\beta}) = | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^{p-2}, \ \boldsymbol{\beta}^{(k+1)} = \arg\min_{\beta} \sum w_i(\boldsymbol{\beta}^{(k)}) | y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} |^2.$$

例7（logistic回归模型）响应变量y_i取值0,1，自变量$\mathbf{x}_i, i=1,..,n$。logistic回归模型是一个研究响应变量与自变量关系的模型，它假设回归函数$p_i = E(y_i \mid \mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i)$具有如下形式：

$$p_i = p_i(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})},$$

成败概率之比$\frac{p_i}{1-p_i}$称为odds，上述假设等价地为

$$\text{logit}(p_i) \triangleq \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^{\top}\boldsymbol{\beta}$$

似然函数

$$L(\boldsymbol{\beta}) = \prod p_i^{y_i}(1-p_i)^{1-y_i} = \prod\left(\frac{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})}\right)^{y_i}\left(1-\frac{\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})}\right)^{1-y_i}$$

log - 似然函数：

$$\log L = \sum [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

*score*函数：

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum \mathbf{x}_i (y_i - p_i) = X^{\mathsf{T}} (\mathbf{y} - \mathbf{p}),$$

其中记$\mathbf{y} = (y_1, ..., y_n)^{\mathsf{T}}$, $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^{\mathsf{T}}$设计阵, $\mathbf{p} = (p_1, .., p_n)^{\mathsf{T}}$。

极大似然估计是似然方程$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = 0$的解。

Newton - Raphson算法：

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left(\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} \right)^{-1} \frac{\partial \log L}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{\text{old}}}$$

Newton - Raphson算法快速高效，对于logistic回归，它实际上是一种IRLS算法：

注意到Hessian matrix

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} = -\sum \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} p_i (1 - p_i) = -X^{\mathsf{T}} W X$$

其中$W = \mathrm{diag}(p_1(1-p_1), \ldots, p_n(1-p_n))$

所以Newton - Raphson迭代方程可写为

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} + \left(X^{\mathsf{T}} W X\right)^{-1} X^{\mathsf{T}} (\mathbf{y} - \mathbf{p})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\text{old}}} = \left(X^{\mathsf{T}} W X\right)^{-1} X^{\mathsf{T}} W \tilde{\mathbf{y}},$$

其中$\tilde{\mathbf{y}} = X\boldsymbol{\beta}^{\text{old}} + W^{-1}(\mathbf{y} - \mathbf{p})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\text{old}}}$

给定初值$k = 0, \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$,

(1)$k = k + 1$, 计算W矩阵，$\tilde{\mathbf{y}} = X\boldsymbol{\beta}^{(k-1)} + W^{-1}(\mathbf{y} - \mathbf{p})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k-1)}}$,

(2)更新$\boldsymbol{\beta}: \boldsymbol{\beta}^{(k)} = \left(X^{\mathsf{T}} W X\right)^{-1} X^{\mathsf{T}} W \tilde{\mathbf{y}}.$

go to (1)

更一般地，广义线性模型(Generalized Linear Model, GLM)的MLE求解方法一般都采用IRLS方法。

GLM是一类广泛的指数族分布的回归模型，其重要情形有：
(记$\mu = \mu(\mathbf{x}) = \mathrm{E}(y \mid \mathbf{x})$为回归函数)

- 正态线性回归模型（正态响应数据：$\mu = \alpha + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$）；
- logistic回归模型（0-1响应数据：$\mathrm{logit}(\mu) = \alpha + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$）；
  Probit回归模型（0-1响应数据：$\Phi^{-1}(\mu) = \alpha + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$）；
- Poisson回归模型（计数响应数据：$\log(\mu) = \alpha + \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$）；

Poisson回归模型$(y_i, \mathbf{x}_i)$, y_i为计数响应, $\mathbf{x}_i$为自变量, 通常假设

$$y_i \sim Pois(\mu_i), \quad \mu_i = \exp(\alpha + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})$$

即μ_i与$\mathbf{x}_i$有关. 例如$\mathbf{x}_i = 1,0$, 则为两样本问题

$$\mathbf{x}_i = 0 : y_i \sim Pois(e^{\alpha}); \quad \mathbf{x}_i = 1 : y_i \sim Pois(e^{\alpha+\beta})$$