《线性回归》
— 线性模型的诊断

杨　　瑛

清华大学　数学科学系
Email: yyang@math.tsinghua.edu.cn
Tel: 62796887

2019.05.07

Outline

Outline

回归曲面和偏残差

- 设$\mathbf{E}[\mathbf{Y}|\mathbf{x}]$表示给定$\mathbf{x}$之后响应变量$\mathbf{Y}$的条件期望。
  这里有两个模型：
  - ✓ 真的回归函数$\mathbf{E}[\mathbf{Y}|\mathbf{x}] = \mu(\mathbf{x})$
  - ✓ 线性模型$\mathbf{E}[\mathbf{Y}|\mathbf{x}] = \mathbf{x}^T\beta$

  为了评判线性模型是否合适，需要能够可视化真的回归曲面！以便我们决定$\mu(\mathbf{x})$是否可以充分的用$\mathbf{x}$线性的表示.

回归曲面和偏残差

可视化回归曲面：一维情形

- 当$\mathbf{x}$是一维的协变量时，可以利用以前学习过最近邻平均估计，(R:) losse或者其它方法得到$\mu(\mathbf{x})$的光滑估计. 然后画出数据点和光滑的曲线, 见图1.
- 如果关系看上去是线性的，我们可以拟合线性模型.
- 如果关系不是线性的，则需要对数据做合适变换，例如，幂变换，对数变换或者Box-Cox变换等.
- 也可以拟合多项式模型【黑板：简要说明】.

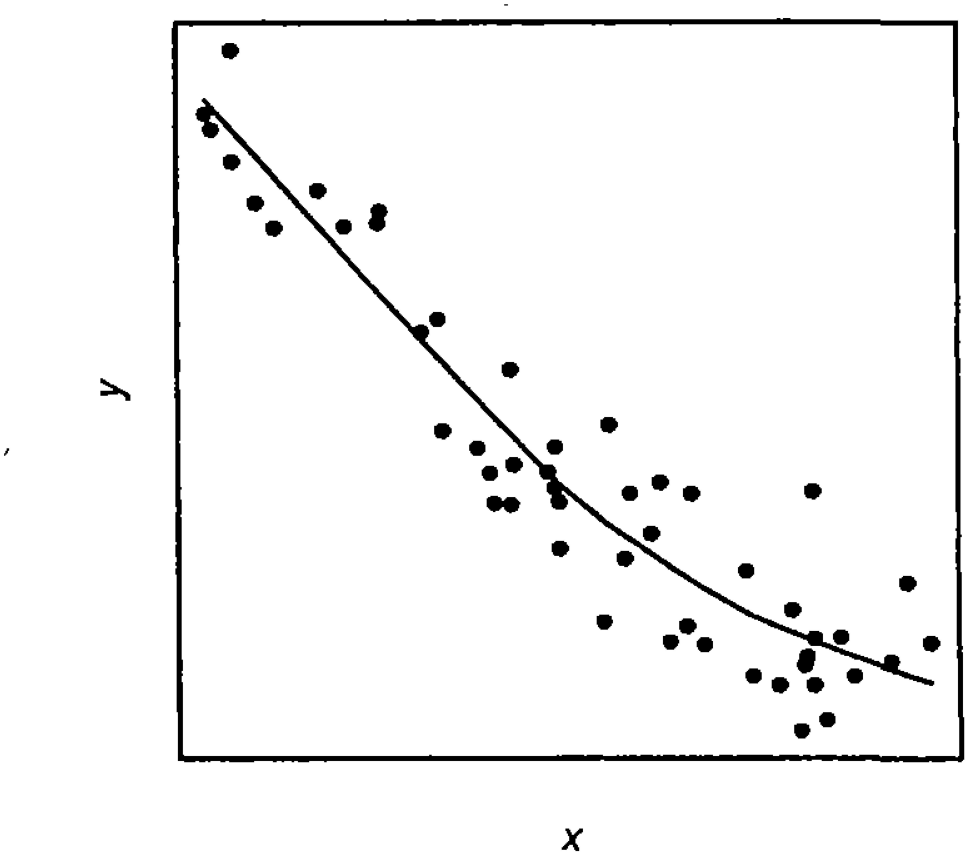

Figure: Smoothing a plot of $\mathbf{Y}$ versus $\mathbf{x}$. [来源：Seber and Lee (2003), Figure 10.1]

可视化回归曲面：二维情形

- 当$\mathbf{x} = (x_1, x_2)^T$是二维的协变量时，可以用R的package：plot3D或者rgl画出$\mathbf{Y}$和$\mathbf{x} = (x_1, x_2)^T$的三维图形。这类图形通常是可以旋转的，以便从不同的角度观察，进而确定是平面还是曲面.
- 如果关系看上去是平面，我们可以拟合线性模型.
- 如果关系看上去是曲面，则需要对数据做合适变换
- 可以拟合多项式模型【想一想：是什么样的多项式模型？】.

可视化回归曲面：三维以上

- 当$\mathbf{x} = (x_1, \cdots, x_p)^T$是$p$维的协变量时($p \geq 3$)，则很难将曲面可视化.
- 但是，可以画出残差和拟合值的残差图，如果模型正确，残差将没有确定的模式，其分布在一个水平的带状区域内。

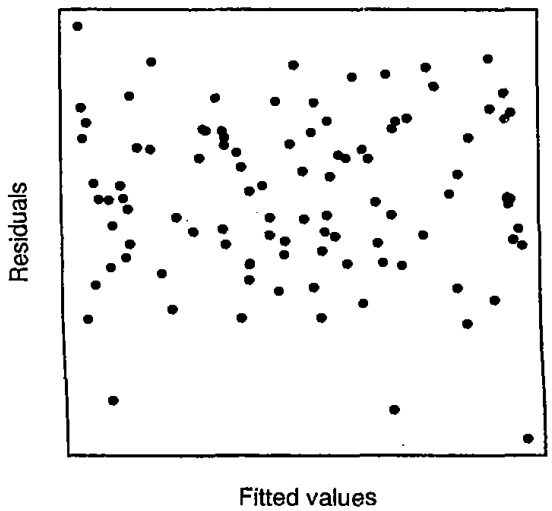

Fig. 10.2 Plotting residuals versus fitted values.

Figure: [来源：Seber and Lee (2003), Figure 10.2]

偏残差图(partial residual plot)

- 为了揭示曲面的性质，我们可以使用各种类型的偏残差图。这些图是指适当修正的残差与协变量的图形。
- 假定真的回归模型是：

$$E[\mathbf{Y}|\mathbf{x}] = \beta_0 + \beta_1 g(x_1) + \beta_2^T \mathbf{x}_2, \tag{1}$$

其中$\mathbf{x} = (x_1, x_2)^T$, $g(\cdot)$是未知函数. [模型(1)是一种特殊的半参数模型, 若(1)中$\beta_0 = 0$且$\beta_1 = 1$,则称为部分线性模型].

- 如果我们能够发现函数g的性质，则我们做变换$x_1^* = g(x_1)$, 代替x_1, 然后来拟合模型：

$$E[\mathbf{Y}|\mathbf{x}] = \beta_0 + \beta_1 x_1^* + \beta_2' \mathbf{x}_2. \tag{2}$$

这个模型是正确的。偏残差图用来发现g的形式。
这种类型的图可以追溯到Ezekiel (1924), Ezekiel and Fox (1959), Larsen and McCleary (1972).

偏残差图

- 偏残差图的直观想法是这样的：
  当(1)是真的模型，假定我们拟合的线性模型是：

$$E[\mathbf{Y}|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2^T \mathbf{x}_2. \tag{3}$$

即

$$\begin{aligned} \mathbf{Y} &= \beta_0 + \beta_1 x_1 + \beta_2' \mathbf{x}_2 + [\beta_1 g(x_1) - \beta_1 x_1 + \varepsilon] \\ &\approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2' \mathbf{x}_2 + e \end{aligned}$$

从这个线性拟合可以看出，残差e可以近似地表示为：$\beta_1 [g(x_1) - x_1] \approx \hat{\beta}_1 [g(x_1) - x_1]$.

- 以$e_i^\dagger = e_i + \hat{\beta}_1 x_{i1} \left[\approx \hat{\beta}_1 g(x_{i1})\right]$为纵坐标，$x_{i1}$为横坐标偏残差图，就可以揭示$g$的形状。
- 修正的残差$e_i^\dagger$称之为偏残差。

偏残差图的特性：

- 对‘数据’$(e_i^{\dagger}, x_{i1})$，$i = 1, \cdots, n$, 拟合过原点的直线，其系数的最小二乘估计是：

$$\frac{\sum_i x_{i1} e_i^{\dagger}}{\sum_i x_{i1}^2} = \frac{\sum_i x_{i1} \left(\hat{\beta}_1 x_{i1} + e_i \right)}{\sum_i x_{i1}^2} = \hat{\beta}_1 + \frac{\sum_i x_{i1} e_i}{\sum_i x_{i1}^2} = \hat{\beta}_1, \quad (4)$$

  即，斜率是$\widehat{\beta}$. (这里用到了：$\mathbf{Xe} = \mathbf{X}(\mathbf{I}_p - \mathbf{H})\mathbf{Y} = \mathbf{0}$.)

- 偏残差图也有两个缺陷：
  - ✓ 这个偏残差图会过分的去掉解释变量x_1的重要性，因为图中的非常靠近拟合直线. 后面会给出一个替代的图形.
  - ✓ 各位摘要的是，如果g是高度的非线性的，则系数β_2的LSE不是很好的估计.【问题：如何度量一个函数的线性程度?】

方差齐次性的检验

- 考虑通常的线性模型：

$$Y_i = \mathbf{x}_i'\beta + \varepsilon_i \tag{5}$$

其中ε_i是独立正态均值为0的随机误差。其标准的假设是：

$$\text{var}\left[\varepsilon_i\right] = \sigma^2 \quad (i = 1, \ldots, n),$$

这里，我们假定：

$$\text{var}\left[\varepsilon_i\right] = \sigma_i^2, \tag{6}$$

其中方差σ_i^2可以即依赖于均值$E\left[Y_i\right] = \mathbf{x}_i'\beta$, 又依赖于其它的参数，或者依赖于解释变量$\mathbf{z}_i$.

- 如果随机误差的方差不全相等，则系数β的LSE不是有效的. 我们需要检查方差是否相等. 如果必要，使用更为有效的估计方法.

方差齐次性的检验(续)

线性模型中的方差齐次性的检验和异方差的判断是一个非常重要的课题.

- 在实验时，如果安排了重复试验，则可以利用Bartlett检验，levene检验. 详见Draper and Smith (1998) Section 2.2中的内容
- 如果没有安排重复试验，需要对方差的结构方式做出假设，然后在检验方差的齐次性.
- 最好的办法就是在实验时考虑安排重复试验。否则，退而求其次，对误差方差的结构进行研究和假设.
- 以下的内容集中在方差有结构的情形.

方差齐次性的检验(续)

- 假定

$$\sigma_i^2 = w\left(\mathbf{z}_i, \boldsymbol{\lambda}\right), \tag{7}$$

其中$\mathbf{z}_i$是第i个观测的已知的协变量向量，w是方差函数，对于某个$\boldsymbol{\lambda}_0$, $w\left(\mathbf{z}, \boldsymbol{\lambda}_0\right)$不依赖于$\mathbf{z}$.

- 在(7)中，w的形式是已知的，但是λ是未知的. 例如，w可以取为

$$w(\mathbf{z}, \boldsymbol{\lambda}) = \exp\left(\mathbf{z}^T \boldsymbol{\lambda}\right) \tag{8}$$

其中$\mathbf{z} = (1, z_1, \cdots, z_k)^T$. 在这种情形下，$\lambda = (\lambda_0, 0, \cdots, 0)^T$,

$$\operatorname{var}\left[Y_i\right] = w\left(\mathbb{Z}, \lambda_0\right) = e^{\lambda_0} = \sigma^2. \tag{9}$$

需要说明的是(7)中的方差形式不包括方差是均值函数的情形. 这种情形将在后面讨论.

- 上面的假设完全有响应的分布确定，我们可以检验假设

$$H_0 : \lambda = \lambda_0$$

来检验方差的齐次性.

- 有最小二乘拟合得到的残差**e**含有方差的信息. 如果$\Sigma = \text{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$, 则

$$\begin{aligned} \text{Var}[\mathbf{e}] &= \text{Var}\left[\left(\mathbf{I}_n - \mathbf{H}\right)\varepsilon\right] \\ &= \left(\mathbf{T}_n - \mathbf{H}\right)\Sigma\left(\mathbf{I}_n - \mathbf{H}\right), \end{aligned}$$

因而,

$$\text{var}\left[e_i\right] = \left(1 - h_i\right)^2 \sigma_i^2 + \sum_{k:k\neq i} h_{ik}^2 \sigma_k^2. \tag{10}$$

通常，$h_{ik} \ll h_i, k \neq i$, 这样大的方差意味着大的残差，这种情况可能不一定意味着是高的杠杆点.

- 注意到$E[e_i]=0$, 故有$\text{var}[e_i]=E[e_i^2]$, 下面的量

$$b_i=\frac{e_i^2}{1-h_i} \tag{11}$$

在做各种图的时有用. 当所有方差都相等时，

$$E[b_i]=(1-h_i)\sigma^2+\sum_{k:k\neq i}\frac{h_{ik}^2}{1-h_i}\sigma^2=\sigma^2$$

上式成立是因为幂等矩阵$\mathbf{I}_n-\mathbf{H}$蕴含着

$$(1-h_i)^2+\sum_{k:k\neq i}h_{ik}^2=1-h_i. \tag{12}$$

由此，当方差是常数时，$\mathbf{b}_i$的期望是常数.

- 当随机误差的方差不全相等时，拟合值$\hat{Y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$的期望仍然是$E[Y_i]$. 但是，均值大的观测常常有大的方差.
- 因此，如果方差随着均值增加的话，b_i与拟合值$\widehat{Y}_i$的图形呈现V型(wedge-shaped). 见图10.3(a).
- 另一种图形是：b_i和其它的协变量作图, 解释相同. 可以揭示方差和均值之间的关系.
- 通常的原始的残差和拟合值残差图. 例如, 扇形图, 也说明方差随着均值增加, 见图10.3(b)

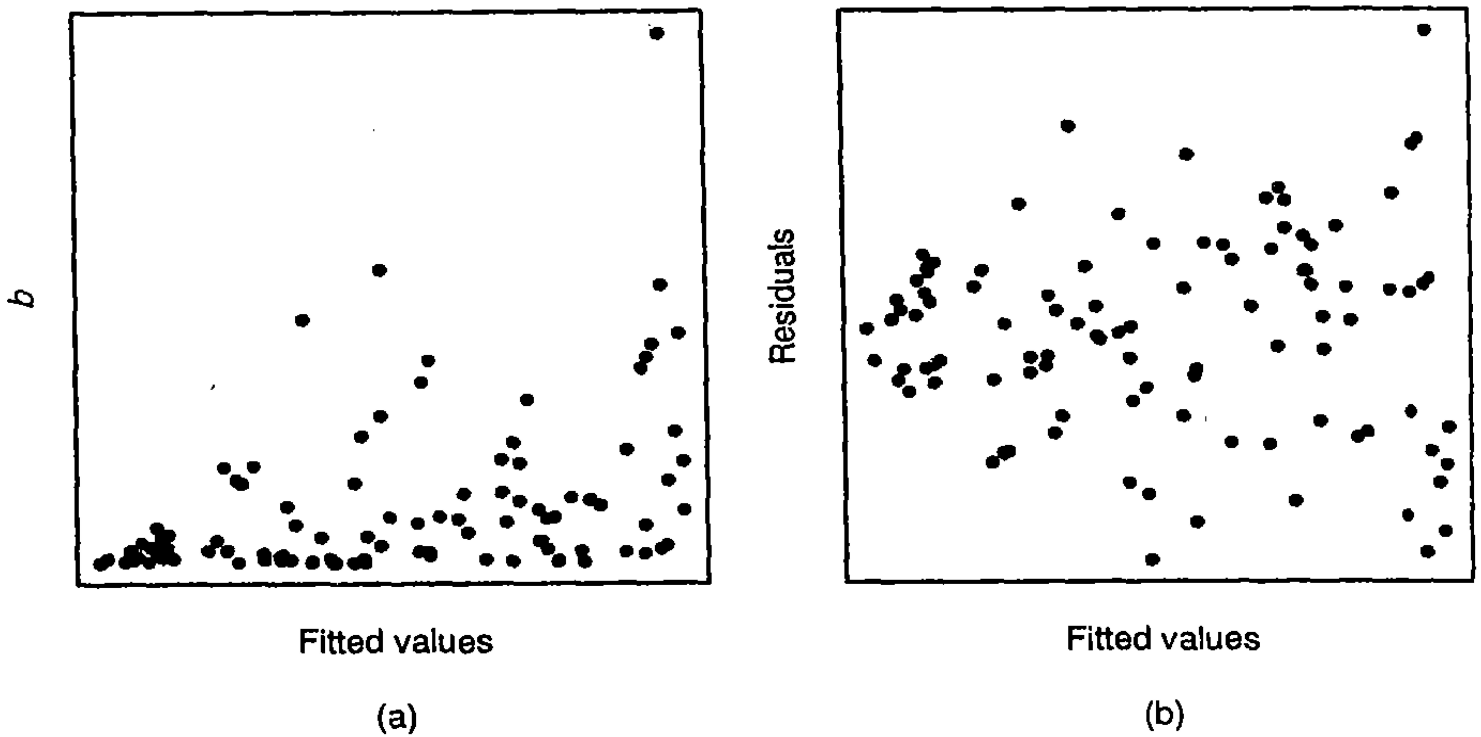

Fig. 10.3 Patterns resulting from the variances being a function of the means.

Figure: 来源于：Seber and Lee (2003), Fig. 10.3

其它的图形：

- 假定w是光滑的，做Taylor展开, 有

$$w(\mathbf{z}_i, \boldsymbol{\lambda}) \approx w(\mathbf{z}_i, \boldsymbol{\lambda}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)' \dot{w}(\mathbf{z}_i, \boldsymbol{\lambda}_0)$$

其中,$\dot{w} = \partial w / \partial \lambda$. 则利用(10)和(12)，得到

$$\begin{aligned} E[b_i] &= (1 - h_i) w(\mathbf{z}_i, \lambda) + \sum_{k:k \neq i} \frac{h_{ik}^2 w(\mathbf{z}_i, \lambda)}{1 - h_i} \\ &\approx w(\mathbf{z}_i, \lambda_0) + (\lambda - \lambda_0)' \left\{ (1 - h_i) \dot{w}(\mathbf{z}_i, \lambda) \right. \\ &\quad \left. + \sum_{k:k \neq i} \frac{h_{ik}^2 \dot{w}(\mathbf{z}_i, \lambda)}{1 - h_i} \right\}. \end{aligned}$$

- Cook and Weisberg (1983)建议画出b_i和上式中$\{\}$中量的图形. 非齐次方差将在图形中显示出线性趋势.
- 这些图形可以得到假设$\lambda = \lambda_0$的基于标准渐近检验的支持.
- 在正态分布的假定之下，对数似然函数$\ell(\beta, \lambda)$是

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= c - \frac{1}{2}\left\{\log\det(\Sigma) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \\ &= c - \frac{1}{2}\left\{\sum_{i=1}^{n}\log w_i + \sum_{i=1}^{n}\frac{\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2}{w_i}\right\} \end{aligned} \tag{13}$$

其中, $w_i = w(\mathbf{z}_i, \lambda)$, c是一个常数.

- $\ell(\beta, \lambda)$关于β和λ求偏导数得：

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \Sigma^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{X}^T \Sigma^{-1} \boldsymbol{\varepsilon}, \tag{14}$$

$$\frac{\partial l}{\partial \boldsymbol{\lambda}} = -\frac{1}{2}\left[\sum_{i=1}^{n}\left\{\frac{1}{w_i} - \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{w_i^2}\right\}\frac{\partial w_i}{\partial \boldsymbol{\lambda}}\right] \tag{15}$$

求解β和λ的算法

1. 令$\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$.
2. 计算$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{Y}$ [这是加权LSE]
3. 关于λ 求解方程：$\partial \ell / \partial\, \lambda\big|_{\beta=\hat{\beta}} = \mathbf{0}$
4. 重复步骤2和3直至收敛.

Example

假定$w(\mathbf{z},\boldsymbol{\lambda}) = \exp\left(\mathbf{z}^T\boldsymbol{\lambda}\right)$, 其中$\mathbf{z} = (1, z_1, \cdots, z_k)$, 第三步可以用LSE实现。以下是计算$\lambda$的细节.
对于w,$\partial w_i/\partial\lambda = w_i\mathbf{z}i$, 故

$$\frac{\partial\ell}{\partial\lambda} = -\frac{1}{2}\sum_{i=1}^{n}\left(1 - \frac{\varepsilon_i^2}{w_i}\right)\mathbf{z}_i \tag{16}$$

记$d_i = \varepsilon_i^2/w_i, \mathbf{d} = (d_1, \cdots, d_n)^T$, $\mathbf{Z}$是由$\mathbf{z}_i$为行构成的矩阵.
则(13)可以写为:

$$\frac{\partial\ell}{\partial\boldsymbol{\lambda}} = \frac{1}{2}\mathbf{Z}^T\left(\mathbf{d} - \mathbf{1}_n\right). \tag{17}$$

以下计算信息矩阵:

$$\mathrm{Var}\left[\frac{\partial\ell}{\partial\boldsymbol{\beta}}\right] = \mathrm{Var}\left[\Sigma^T\Sigma^{-1}\boldsymbol{\varepsilon}\right] = \mathbf{X}^T\Sigma^{-1}\,\mathrm{Var}[\varepsilon]\Sigma^{-1}\mathbf{X} = \mathbf{X}^T\Sigma^{-1}\mathbf{X}$$

Example (续)

$$\mathrm{Var}\left[\frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right] = \mathrm{Var}\left[\frac{1}{2}\mathbf{Z}^T(\mathbf{d} - \mathbf{1}_n)\right] = \frac{1}{4}\mathbf{Z}^T \mathrm{Var}[\mathbf{d}]\mathbf{Z}$$

因为ε的方差为w_i, 故$\mathbf{d}$的元素是iid χ^2_1, 方差为2, 故有$\mathrm{Var}[\mathbf{d}] = 2\mathbf{I}_n$, $\mathrm{Var}\{\partial\ell/\partial\lambda\} = \frac{1}{2}\mathbf{Z}^T\mathbf{Z}$.

进一步，由于$E\left[\varepsilon_i\varepsilon_j^2\right] = 0 (i \neq j)$, $E\left[\varepsilon_j^3\right] = 0$, 故有$\mathrm{Cov}[\varepsilon, d] = 0$,

$$\mathrm{Cov}\left[\frac{\partial \ell}{\partial \boldsymbol{\beta}}, \frac{\partial \ell}{\partial \boldsymbol{\lambda}}\right] = \frac{1}{2}\mathbf{X}^T\Sigma^{-1}\mathrm{Cov}[\varepsilon, \mathbf{d}]\mathbf{Z} = \mathbf{0}$$

故由(3.19)得到【练习】,

$$\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \begin{pmatrix} \mathbf{X}^T\Sigma^{-1}\mathbf{X} & \mathbf{0} \\ 0 & \frac{1}{2}\mathbf{Z}^T\mathbf{Z} \end{pmatrix}.$$

Example (续)

利用Fisher Scoring方法[黑板]可以求解似然方程. 迭代方程为：

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(m+1)} &= \left(\mathbf{X}^T\Sigma_{(m)}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\Sigma_{(m)}^{-1}\mathbf{Y}, \\
\boldsymbol{\lambda}_{(m+1)} &= \boldsymbol{\lambda}_{(m)} + \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\left(\mathbf{d}-\mathbf{1}_n\right), \\
\Sigma_{(m+1)} &= \operatorname{diag}\left[w\left(\mathbf{z}_1,\boldsymbol{\lambda}_{(m)}\right),\ldots,w\left(\mathbf{z}_n,\boldsymbol{\lambda}_{(m)}\right)\right], \\
& \quad m=1,2,\cdots.
\end{aligned} \tag{18}$$

(18)可以改写为各位紧凑的形式：

$$\mathbf{Z}^T\mathbf{Z}\lambda_{(m+1)} = \mathbf{Z}^T\left(\mathbf{d}-\mathbf{1}_n+\mathbf{Z}\lambda_{(m)}\right) \tag{19}$$

上面的式子恰好可以解释为$\mathbf{d}-\mathbf{1}_n+\mathbf{Z}\boldsymbol{\lambda}_{(m)}$(形式的响应变量)对$\mathbf{Z}$做形式的线性回归，因而可以利用R中的lm来实施.

在统计中，有时建立‘伪回归’以方便计算!

检验：$\lambda = \lambda_0$ (方差齐次性的检验)

- 前面介绍的估计方法其中的一个重要作用是做假设检验:

$$H_0 : \lambda = \lambda_0.$$

在这个假设之下, $w_i(\mathbf{z}, \lambda_0) = \sigma^2$.

- 可以利用似然比(LRT)和得分检验(score test). 其实还有Wald test [黑板].

关于这三个检验的内容，可以参考Bickel and Doksum (2001), 2nd, 398–400

- 很容易得到上面假设检验的LRT为:

$$\mathrm{LR} = -2\left[\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \hat{\sigma}^2\right) - l(\hat{\beta}, \hat{\boldsymbol{\lambda}})\right],$$

其中，$\hat{\boldsymbol{\beta}}$ 和OLS, $\hat{\sigma}^2$是H_0成立时β和σ^2的ML估计，而$\hat{\boldsymbol{\beta}}$和$\hat{\boldsymbol{\lambda}}$是上面的算法给出的无约束的$\beta$和$\sigma^2$的MLE.

- 在H_0之下，LR $\sim \chi^2_k$.
- Score test【细节从略】
- Wald test 【细节从略】
  - ✓ .