—— 回归分析 ——

线性回归的数学、统计观点

回归是一种有监督学习

好像想说些什么，好像又什么也说不出口。

——*Arthur*

目录

高斯-马尔可夫定理：

完整来说有五个假定条件：

- 假定 SLR.1: 线性于参数，即要求回归参数 β_i 皆为常数以保证模型的线性.
- 假定 SLR.2: 随机抽样，即样本必须是以某种方式随机抽样得到的.
- 假定 SLR.3: 解释变量的样本有波动，即没有自变量是常数，没有自变量之间具有完全共线性.
- 假定 SLR.4: 零条件均值，即误差的条件期望应为零且不受自变量的影响.
- 假定 SLR.5: 同方差性.

定理内容为，没有多重共线性情况下，若误差项满足零均值、同方差、不相关，这时 OLS 就是 BLUE，而不需要误差独立同分布于正态分布，这个条件太强，这个的证明思路为设 $\boldsymbol{a}^T\boldsymbol{y}$ 是 $\boldsymbol{c}^T\boldsymbol{\beta}$ 的任意一个线性无偏估计，有：

$$\begin{aligned}\|\boldsymbol{a}\|_2 &= \|\boldsymbol{a}-\boldsymbol{c}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}+\boldsymbol{c}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\|_2 \\ &= \|\boldsymbol{a}-\boldsymbol{c}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\|_2+\|\boldsymbol{c}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\|_2+2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{a}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}) \\ &= ①+②+③ = ①+\frac{Var(\boldsymbol{c}^T\boldsymbol{\beta})}{\sigma^2}+0\end{aligned}$$

$\Rightarrow \forall \boldsymbol{a}^T\boldsymbol{y}\ s.t.\mathbb{E}(\boldsymbol{a}^T\boldsymbol{y})=\boldsymbol{c}^T\boldsymbol{\beta}$，都有 $Var(\boldsymbol{a}^T\boldsymbol{y})\geqslant Var(\boldsymbol{c}^T\boldsymbol{\beta})$，
当且仅当 $\boldsymbol{a}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}$ 时等号成立.

当误差服从正态分布时，截距项与回归系数向量的 OLS 与 MLE 等价，而且在 Cramér–Rao 下界的意义下是渐进有效的；当误差独立同分布时，截距项与回归系数向量的 OLS 与广义矩法估计等价，他们都为 $\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，但随机误差的方差的 OLS 与 MLE 不同，这时 MLE 有偏，而以自由度 $n-p$ 为分母的 OLS 才是无偏的.

$$\hat{\sigma}^2_{MLE}=\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2=\frac{RSS}{n}$$

$$\hat{\sigma}^2_{OLS}=\frac{1}{n-p}\sum_{i=1}^{n}\hat{e}_i^2=\frac{RSS}{n-p}$$

$RSS=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=\boldsymbol{y}^T\boldsymbol{y}-\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$，其中 $\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$ 为回归平方和

事实上只要满足前四条假设，截距、回归系数的 OLS 就一定是无偏的；除了无偏性，若误差项还满足第五条假设即同方差，则进一步地有：

$$Cov(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})=\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

而如若还有 $\boldsymbol{\varepsilon}\sim N_n(\boldsymbol{0},\sigma^2\boldsymbol{I})$ 即误差独立同分布于正态分布，那么：

i $\widehat{\boldsymbol{\beta}}\sim N_n(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

ii $\frac{RSS}{\sigma^2}\sim\chi^2(n-p)$

iii $\widehat{\boldsymbol{\beta}}$ 与 RSS 相互独立

线性回归中一般考虑 OLS 而非 MLE（有些情况下他们是等价的），这是由问题的假设条件决定的，这样的情况下 MLE 准则下的损失函数反而不是凸函数，使用优化策略、优化算法时可能陷入局部最优解，而 OLE 在高斯-马尔可夫定理的条件下对于线性回归是一个凸优化问题，再加上 OLS 在我们的假定下具有上述诸多良好性质，我们考虑 OLS 而非 MLE；但在逻辑斯蒂回归中我们却应用的是 MLE、以交叉熵作为损失函数，这个问题我在分类数据分析的总结归纳中讨论过了.

OLS、正则方程与帽子矩阵：普通最小二乘估计

模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 的偏差为 $\boldsymbol{\varepsilon}=\boldsymbol{y}-\boldsymbol{X\beta}$，将偏差向量模的的平方

$$\mathcal{L}(\boldsymbol{\beta})=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2=(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})$$

作为损失函数，令 $\mathcal{L}(\boldsymbol{\beta})$ 的矩阵偏导为 0 得到方程组 $\boldsymbol{X}^T\boldsymbol{X\beta}=\boldsymbol{X}^T\boldsymbol{y}$，称为正则方程，正则方程有唯一解析解的充要条件为 $\boldsymbol{X}^T\boldsymbol{X}$ 满秩（$\boldsymbol{X}$列满秩、没有复共线性），他的解就是 OLS，$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，任意 $\boldsymbol{\beta}$ 有 $\mathcal{L}(\boldsymbol{\beta})\geqslant\mathcal{L}(\widehat{\boldsymbol{\beta}})$，这是由于：

$$\mathcal{L}(\boldsymbol{\beta})=\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2^2=\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+2(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

当且仅当 $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\boldsymbol{0}$ 取等.

$$\widehat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\dots)=argmin_{\hat{\beta}_0,\hat{\beta}_1,\dots}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

称 $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ 为帽子矩阵，易见 $\hat{\boldsymbol{y}}=\boldsymbol{X}\widehat{\boldsymbol{\beta}}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\Rightarrow\hat{\boldsymbol{y}}=\boldsymbol{H}\boldsymbol{y}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
帽子矩阵是一个幂等矩阵；残差与帽子矩阵有非同寻常的关系，$\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}=(\boldsymbol{I}-(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$，进而有 $RSS=\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}^T\boldsymbol{y}-\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$；① $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ 是在由 $\boldsymbol{X}$ 列向量张成的线性空间 $\mu(\boldsymbol{X})$ 中的正交投影阵，② $\hat{\boldsymbol{y}}=\boldsymbol{H}\boldsymbol{y}$ 是 $\boldsymbol{y}$ 在 $\mu(\boldsymbol{X})$ 上的投影，③ $\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$ 是 $\boldsymbol{y}$ 在$\mu(\boldsymbol{X})^{\perp}$ 上的投影.

容易看出，损失函数 $\mathcal{L}=$ RSS，考虑到$RSS=n\times MSE$，也可以说损失函数为 MSE. 最小二乘法核心思想即最小化 RSS，为了方便求导后的表示也常常写作$\frac{1}{2}$RSS，效果是一致的；虽然我们有计算公式 $\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，但由于维数灾难的存在使得数据量充分大时求解困难、内存成本过高，通常用一些最优化算法迭代求数值解，这时 MSE 准则下的求解问题为凸优化问题的优势就很明显了. 另外，MSE 准则作为综合了无偏性与有效性（综合考虑偏差——一阶矩与方差——二阶矩）的一种评估办法本身也是可行的，再加上 MSE 准则下得到的最小二乘估计在一定条件下的诸多优良性质，进一步说明了为什么我们考虑最小二乘估计. 最小二乘法有非常多的证明方式，例如这里我们直接求矩阵导数，还可以用更“代数”的方法：用欧氏空间的理论加以证明.

一般 OLS 问题用梯度下降法、*Nnewton-Raphson 法*、BFGS 算法等算法，这属于最优化理论，这里不做详细说明了.

高斯-马尔可夫假设下最小二乘问题是凸优化问题：

$$\mathcal{L}(\boldsymbol{\beta})=\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2=(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})$$

$$=\boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta}-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{y}^T\boldsymbol{y}$$

由于 $\boldsymbol{X}^T\boldsymbol{X}$ (半)正定，显然 $\mathcal{L}(\boldsymbol{\beta})$ 是关于 $\boldsymbol{\beta}$ 的凸二次函数.

可证明 OLS 的诸多优良性质：相合性，无偏性，有效性（方差最小的无偏估计），渐近正态性；OLS 的多种变种估计也具有这些性质，后文会逐一提到，不再赘述.

```
> fit <- lm(hwy ~ displ + class, data = mpg)
> summary(fit)

Call:
lm(formula = hwy ~ displ + class, data = mpg)

Residuals:
   Min     1Q Median     3Q    Max
-5.572 -1.569 -0.245  1.355 14.724

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    38.9533     1.7976  21.669  < 2e-16 ***
displ          -2.2976     0.2132 -10.778  < 2e-16 ***
classcompact   -5.3122     1.5283  -3.476 0.000610 ***
classmidsize   -4.9471     1.4722  -3.360 0.000914 ***
classminivan   -8.7986     1.5939  -5.520 9.26e-08 ***
classpickup   -11.9232     1.3687  -8.711 6.46e-16 ***
classsubcompact  -4.6988     1.5097  -3.112 0.002095 **
classsuv      -10.5851     1.3268  -7.978 7.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.745 on 226 degrees of freedom
Multiple R-squared:  0.7939,    Adjusted R-squared:  0.7875
F-statistic: 124.3 on 7 and 226 DF,  p-value: < 2.2e-16
```

特别地，一元线性回归的 OLS：

模型 $y_i = \alpha + \beta x_i + \varepsilon_i$，两个参数的 OLS 用初等数学表达出来为

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \dfrac{\sum x_i y_i - \sum x_i \sum y_i}{\sum x_i^2 - n\bar{x}^2} \end{cases}$$

标准化（中心化）后的线性回归：

以标准化为例，标准化操作可以消除量纲的影响、便于比较自变量之间一些数字特征差异，如果仅对线性回归而言这不是必需的操作，但若涉及 PCA 和正则化的岭回归、Lasso 回归则标准化是有必要的，因为可以从理论上证明 PCA 的操作中越前的主成分越倾向于赋予方差更大数据更多的权重，岭回归、Lasso 回归如果不进行标准化，正则化的操作会带来更大的偏差.

中心化令 $y_i \leftarrow y_i - \bar{y}$、$x_{ii} \leftarrow x_{ii} - \bar{x_i}$，最终只改变截距项 $\hat{\beta}_0$ 的值，相当于将模型平移至原点处并得到了$\hat{\beta'}_0 = 0$，而标准化中截距项、回归系数与残差都会改变，但对于线性回归模型而言没有实质的变化.

线性回归：$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ 改写为 $\boldsymbol{y'} = \boldsymbol{X}_{std}\boldsymbol{\beta'} + \boldsymbol{\varepsilon'}$，其中

$$\boldsymbol{y}' = \begin{pmatrix} \frac{y_1 - \bar{y}}{\hat{\sigma}_y^2} \\ \frac{y_2 - \bar{y}}{\hat{\sigma}_y^2} \\ \vdots \\ \frac{y_n - \bar{y}}{\hat{\sigma}_y^2} \end{pmatrix}, \quad \boldsymbol{X}_{std} = \begin{pmatrix} 1 & \frac{x_{11} - \bar{x}_1}{\hat{\sigma}_1^2} & \frac{x_{12} - \bar{x}_2}{\hat{\sigma}_2^2} & \cdots & \frac{x_{1,p-1} - \bar{x}_{p-1}}{\hat{\sigma}_{p-1}^2} \\ 1 & \frac{x_{21} - \bar{x}_1}{\hat{\sigma}_1^2} & \frac{x_{22} - \bar{x}_2}{\hat{\sigma}_2^2} & \cdots & \frac{x_{2,p-1} - \bar{x}_{p-1}}{\hat{\sigma}_{p-1}^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x_{n1} - \bar{x}_1}{\hat{\sigma}_1^2} & \frac{x_{n2} - \bar{x}_2}{\hat{\sigma}_2^2} & \cdots & \frac{x_{n,p-1} - \bar{x}_{p-1}}{\hat{\sigma}_{p-1}^2} \end{pmatrix}$$

记 $\boldsymbol{X}_{std}$ 去掉第一列后为 $\boldsymbol{X}'$，最终有

$$\begin{cases} \hat{\beta}'_0 = \bar{y} \\ (\hat{\beta}'_1, \hat{\beta}'_2, \ldots, \hat{\beta}'_{p-1}) = (\boldsymbol{X}'^T \boldsymbol{X}')^{-1} \boldsymbol{X}'^T \boldsymbol{y}' \end{cases}$$

$\boldsymbol{X}'$ 也是比较重要的一个矩阵，$\boldsymbol{X}'^T \boldsymbol{X}'$ 和一些多重共线性检验有密切联系，在后文会提到.

拟合优度 R^2 与修正后的拟合优度：

$$\text{总平方和 } SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^2 = SSR + SSE$$

一般认为模型中能解释的部分为回归平方和 SSR，而残差平方和 RSS 也即 SSE 是不能解释的部分，因此定义取值范围在 0 至 1 的 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 代表模型中 $\boldsymbol{y}$ 变化能被解释的比例，通常认为 R^2 越大模型拟合越佳，但由于自变量个数单纯地越多 R^2 也会越大（直观理解来是“参数越多模型拟合越好”，但注意参数过多会造成过拟合问题）因此只用 R^2 判断模型优劣是片面的，容易损失自由度，也可能会造成过拟合、重共线性性的问题（过拟合可以通过一些准则方法来减少参数解决，也可以增大样本量，在今天大数据时代，大型深度学习模型常常有上百亿个参数）.

因此提出新的指标：修正后的拟合优度，$R^2_{adjusted} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$，但值得注意的是修正后的拟合优度可能是大于 1 的.

CLS：约束最小二乘估计

k 个线性约束 $\boldsymbol{\alpha}_k^T \boldsymbol{\beta} = \boldsymbol{b}_k$ 记为 $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b}$，在这一组约束下约束最小二乘估计为

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T (\boldsymbol{A} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T)^{-1} \left(\boldsymbol{A} \widehat{\boldsymbol{\beta}} - \boldsymbol{b} \right)$$

$$= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T (\boldsymbol{A} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{A}^T)^{-1} (\boldsymbol{A} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{b})$$

考虑 *Lagrange* 乘子法证明.

WLS：加权最小二乘估计

WLS、GLS 都是为了处理异方差问题而提出的.

WLS 用以处理误差不相关但不同分布的问题；GLS 用以处理误差相关的问题：可以讲 OLS 是 WLS 的特例，WLS 是 GLS 的特例. 对于高斯-马尔可夫定理的条件下的问题，这些方法都应该得到相同的结果.

换句话说，对于误差的协方差阵，非主对角线元素皆为 0 且主对角线元素全部相等，即 $Cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$ 时是 OLS 能够处理的问题；非主对角线元素皆为 0 但主对角线元素不全相等，即异方差但互相不相关，这时可以考虑 WLS；非主对角线元素不全为 0，这时是 GLS 处理的场景.

如果模型设定是正确的，那么 OLS 和 WLS 方法都是一致的，但 WLS 或许更有效一些，得到的估计的标准误 SE 更小一些；然而如果模型设定错误，OLS 和 WLS 就可能呈现截然不同的结果，并且两者的估计结果都是错误的.

损失函数：加权后的 RSS，或者讲加权后的 MSE（这时因为 OLS 无偏性）.

$$\mathcal{L} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

容易知道 $\mathcal{L}$ 也是凸函数，假设误差的协方差阵为

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

通常令 $w_i = \frac{1}{\sigma_i^2}$，即权重矩阵 $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$，

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

这时正则方程为 $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$（取单位权重即 $\boldsymbol{W} = \boldsymbol{I}$，WLS 完全等价于 OLS）.

Alexander Aitken 证明了在损失函数——残差平方和加权最小（即残差平方的每一项乘权重再求和最小）的情况下，如果权重取方差估值的倒数（即取权重矩阵为 $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$），则估计为 BLUE.

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots) = argmin_{\hat{\beta}_0, \hat{\beta}_1, \dots} \left\| \boldsymbol{W}^{\frac{1}{2}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\|_2 = argmin_{\hat{\beta}_0, \hat{\beta}_1, \dots} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

$$= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}$$

估计参数的协方差满足 $Cov(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \widehat{\boldsymbol{\Omega}} \boldsymbol{W}^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$，在取 $\boldsymbol{W} = \widehat{\boldsymbol{\Omega}}^{-1}$ 的时候进一步有 $Cov(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$.

关于置信限 Parameter confidence limits

It is often *assumed*, for want of any concrete evidence but often appealing to the central limit theorem—see Normal distribution#Occurrence and applications—that the error on each observation belongs to a normal distribution with a mean of zero and standard deviation σ. Under that assumption the following probabilities can be derived for a single scalar parameter estimate in terms of its estimated standard error SE_β (given here):

68% that the interval $\widehat{\boldsymbol{\beta}} \pm SE_\beta$ encompasses the true coefficient value

95% that the interval $\widehat{\boldsymbol{\beta}} \pm 2\,SE_\beta$ encompasses the true coefficient value

99% that the interval $\widehat{\boldsymbol{\beta}} \pm 2.5\,SE_\beta$ encompasses the true coefficient value

> The assumption is not unreasonable when *n* >> *m*. If the experimental errors are normally distributed the parameters will belong to a Student's t-distribution with *n* − *m* degrees of freedom. When *n* ≫ *m* Student's t-distribution approximates a normal distribution. Note, however, that these confidence limits cannot take systematic error into account. Also, parameter errors should be quoted to one significant figure only, as they are subject to sampling error.
>
> When the number of observations is relatively small, Chebychev's inequality can be used for an upper bound on probabilities, regardless of any assumptions about the distribution of experimental errors: the maximum probabilities that a parameter will be more than 1, 2, or 3 standard deviations away from its expectation value are 100%, 25% and 11% respectively.

FWLS：可行加权最小二乘估计

很多情况下 WLS 不可行，因为残差的协方差阵常常未知，并且对矩阵做有效估计是困难甚至不可完成的任务，这时可以考虑在合理的假设下进行 FWLS.

假设误差的方差不相等且未知但观测值之间并不相关，FWLS 需要做的是取合适的权重，即针对未知的误差方差取合适的估计；FWLS 是迭代的过程，算法步骤为：

（1）先计算 OLS $\widehat{\boldsymbol{\beta}}_0$，得到残差 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0$；

（2）用一种方法构建第 i 个协方差阵的估计 $\widehat{\boldsymbol{\Omega}}_i$，记帽子矩阵 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$，通常可取的办法有：

（a）$\widehat{\boldsymbol{\Omega}}_i = diag(\hat{\varepsilon}_0^2, \hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_{n-1}^2)$

（b）$\widehat{\boldsymbol{\Omega}}_i = diag\left(\frac{\hat{\varepsilon}_0^2}{1-\boldsymbol{H}_{0,0}}, \frac{\hat{\varepsilon}_1^2}{1-\boldsymbol{H}_{1,1}}, \dots, \frac{\hat{\varepsilon}_{n-1}^2}{1-\boldsymbol{H}_{n-1,n-1}}\right)$

（c）$\widehat{\boldsymbol{\Omega}}_i = diag\left(\frac{\hat{\varepsilon}_0^2}{(1-\boldsymbol{H}_{0,0})^2}, \frac{\hat{\varepsilon}_1^2}{(1-\boldsymbol{H}_{1,1})^2}, \dots, \frac{\hat{\varepsilon}_{n-1}^2}{(1-\boldsymbol{H}_{n-1,n-1})^2}\right)$

（d）$\widehat{\boldsymbol{\Omega}}_i = diag(log\,\hat{\varepsilon}_0^2, log\,\hat{\varepsilon}_1^2, \dots, log\,\hat{\varepsilon}_{n-1}^2)$（这条暂时没有找到支持的文献）

（e）$\widehat{\boldsymbol{\Omega}}_i = \boldsymbol{I}$，这样的选择会使下一步得到的 WLS 等价于 OLS

（3）以 $\widehat{\boldsymbol{\Omega}}_i$ 作为权重矩阵计算 WLS $\widehat{\boldsymbol{\beta}}_{i+1}$，$\widehat{\boldsymbol{\beta}}_{i+1}=\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_i^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_i^{-1}\boldsymbol{y}$；

（4）误差达到终止准则则输出结果 $\widehat{\boldsymbol{\beta}}_{i+1}$，否则更新残差，返回步骤 2，$i\leftarrow i+1$.

对于中小样本量，可能 FWLS 没有 OLS 有效，不过大样本的异方差问题下 FWLS 优于 OLS，因此样本量不大的异方差问题可以稍加处理后或直接应用 OLS，不过此时 OLS 得到的回归系数协方差阵的估计不应该再被考虑.

称 $\sqrt{Var\left(\widehat{\boldsymbol{\beta}}_j^*\right)}$ 为 *White* 稳健标准误，在 FGLS 可以推广至 HAC 稳健标准误.

高斯-马尔可夫条件下一般对单个参数进行检验时考虑 t 检验，系数除其标准误应服从 t 分布，但异方差情况下这样检验效果可能有很大偏差，这时“异方差、不相关”的问题应该应用 *White* 稳健标准误，“异方差、部分相关”的问题应该应用 HAC 稳健标准误，如果存在分组应使用聚类稳健标准误：对每组的系数分别进行可行的最小二乘估计，根号下方差即聚类稳健标准误.

GLS：广义最小二乘估计

“WLS、GLS 都是为了处理异方差问题而提出的.

WLS 用以处理误差不相关但不同分布的问题；GLS 用以处理误差相关的问题：可以讲 OLS 是 WLS 的特例，WLS 是 GLS 的特例. 对于高斯-马尔可夫定理的条件下的问题，这些方法都应该得到相同的结果.”

称 $\boldsymbol{\Sigma}^{-1}$ 为精度矩阵，作为权重矩阵 $\boldsymbol{W}$ 的推广；设异方差的形式为 $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{0}$、$Cov(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\boldsymbol{\Sigma}$，$\boldsymbol{\Sigma}$ 正定且已知，记 $\boldsymbol{Z}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{y}$、$\boldsymbol{U}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$、$\boldsymbol{E}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$，则有模型：

$$\boldsymbol{Z}=\boldsymbol{U\beta}+\boldsymbol{E}$$

且有

$$\mathbb{E}(\boldsymbol{E}|\boldsymbol{X})=\boldsymbol{0},\ Cov(\boldsymbol{E}|\boldsymbol{X})=\sigma^2\boldsymbol{I}$$

损失函数：残差向量的马氏距离

$$\mathcal{L}=\boldsymbol{\varepsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}\overset{\min}{=}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

这时再对 $\boldsymbol{Z}=\boldsymbol{U\beta}+\boldsymbol{E}$ 应用 OLS，得到原模型的 GLS：

$$\widehat{\boldsymbol{\beta}}^*=argmin_{\widehat{\beta}_0,\widehat{\beta}_1,\dots}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)=argmin_{\widehat{\beta}_0,\widehat{\beta}_1,\dots}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}-2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

$$=(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z}=(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

可以看出，正则方程为 $\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}=\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$.

GLS 有些情况下可能不太实用，因为大多数问题中做到预知 $\boldsymbol{\Sigma}$ 是比较困难的.

在 $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{0}$、$Cov(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\boldsymbol{\Sigma}$ 成立时，$\widehat{\boldsymbol{\beta}}^*$ 仍具有相合性、无偏性（$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}^*\middle|\boldsymbol{X}\right)=\boldsymbol{\beta}$）、有效性及渐进正态性，且 $Cov\left(\widehat{\boldsymbol{\beta}}^*\middle|\boldsymbol{X}\right)=\sigma^2(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$.

对比 $\boldsymbol{\beta}$ 的 OLS 与 GLS，二者在 $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\boldsymbol{0}$、$Cov(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\boldsymbol{\Sigma}$ 条件下均为无偏估计，恒

有 $Var(\hat{\beta}^*_i) \leqslant Var(\hat{\beta}_i)$，高斯-马尔可夫定理的条件成立时二者等价，存在异方差情况下 GLS 比 OLS 更优.

FGLS：可行广义最小二乘估计

Feasible generalized least squares [edit]

If the covariance of the errors Ω is unknown, one can get a consistent estimate of Ω, say $\hat{\Omega}$,[3] using an implementable version of GLS known as the **feasible generalized least squares** (**FGLS**) estimator. In FGLS, modeling proceeds in two stages: (1) the model is estimated by OLS or another consistent (but inefficient) estimator, and the residuals are used to build a consistent estimator of the errors covariance matrix (to do so, one often needs to examine the model adding additional constraints, for example if the errors follow a time series process, a statistician generally needs some theoretical assumptions on this process to ensure that a consistent estimator is available); and (2) using the consistent estimator of the covariance matrix of the errors, one can implement GLS ideas.

Whereas GLS is more efficient than OLS under heteroscedasticity (also spelled heteroskedasticity) or autocorrelation, this is not true for FGLS. The feasible estimator is, provided the errors covariance matrix is consistently estimated, *asymptotically* more efficient, but for a small or medium size sample, it can be actually less efficient than OLS. This is why some authors prefer to use OLS, and reformulate their inferences by simply considering an alternative estimator for the variance of the estimator robust to heteroscedasticity or serial autocorrelation. But for large samples FGLS is preferred over OLS under heteroskedasticity or serial correlation.[9][4] A cautionary note is that the FGLS estimator is not always consistent. One case in which FGLS might be inconsistent is if there are individual specific fixed effects.[5]

In general this estimator has different properties than GLS. For large samples (i.e., asymptotically) all properties are (under appropriate conditions) common with respect to GLS, but for finite samples the properties of FGLS estimators are unknown: they vary dramatically with each particular model, and as a general rule their exact distributions cannot be derived analytically. For finite samples, FGLS may be even less efficient than OLS in some cases. Thus, while GLS can be made feasible, it is not always wise to apply this method when the sample is small. A method sometimes used to improve the accuracy of the estimators in finite samples is to iterate, i.e. taking the residuals from FGLS to update the errors covariance estimator, and then updating the FGLS estimation, applying the same idea iteratively until the estimators vary less than some tolerance. But this method does not necessarily improve the efficiency of the estimator very much if the original sample was small. A reasonable option when samples are not too large is to apply OLS, but throwing away the classical variance estimator

$$\sigma^2 * (X'X)^{-1}$$

(which is inconsistent in this framework) and using a HAC (Heteroskedasticity and Autocorrelation Consistent) estimator. For example, in autocorrelation context we can use the Bartlett estimator (often known as Newey–West estimator estimator since these authors popularized the use of this estimator among econometricians in their 1987 *Econometrica* article), and in heteroskedastic context we can use the Eicker–White estimator. This approach is much safer, and it is the appropriate path to take unless the sample is large, and "large" is sometimes a slippery issue (e.g. if the errors distribution is asymmetric the required sample would be much larger).

The ordinary least squares (OLS) estimator is calculated as usual by

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$$

and estimates of the residuals $\hat{u}_j = (Y - X\hat{\beta}_{\text{OLS}})_j$ are constructed.

For simplicity consider the model for heteroscedastic and not autocorrelated errors. Assume that the variance-covariance matrix Ω of the error vector is diagonal, or equivalently that errors from distinct observations are uncorrelated. Then each diagonal entry may be estimated by the fitted residuals $\hat{u}_j$ so $\hat{\Omega}_{OLS}$ may be constructed by

$$\hat{\Omega}_{\text{OLS}} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_n^2).$$

It is important to notice that the squared residuals cannot be used in the previous expression; we need an estimator of the errors variances. To do so, we can use a parametric heteroskedasticity model, or a nonparametric estimator. Once this step is fulfilled, we can proceed:

Estimate β_{FGLS1} using $\hat{\Omega}_{OLS}$ using[4] weighted least squares

$$\hat{\beta}_{FGLS1} = (X'\hat{\Omega}_{\text{OLS}}^{-1}X)^{-1}X'\hat{\Omega}_{\text{OLS}}^{-1}y$$

The procedure can be iterated. The first iteration is given by

$$\hat{u}_{FGLS1} = Y - X\hat{\beta}_{FGLS1}$$

$$\hat{\Omega}_{FGLS1} = \text{diag}(\hat{\sigma}^2_{FGLS1,1}, \hat{\sigma}^2_{FGLS1,2}, \ldots, \hat{\sigma}^2_{FGLS1,n})$$

$$\hat{\beta}_{FGLS2} = (X'\hat{\Omega}_{FGLS1}^{-1}X)^{-1}X'\hat{\Omega}_{FGLS1}^{-1}y$$

This estimation of $\hat{\Omega}$ can be iterated to convergence.

Under regularity conditions any of the FGLS estimator (or that of any of its iterations, if we iterate a finite number of times) is asymptotically distributed as

$$\sqrt{n}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{d} \mathcal{N}(0, V),$$

where n is the sample size and

$$V = \text{p-lim}(X'\Omega^{-1}X/T)$$

here p-lim means limit in probability

参考：https://en.wikipedia.org/wiki/Generalized_least_squares

Feasible Generalized Least Squares

Feasible generalized least squares (FGLS) estimates the coefficients of a multiple linear regression model and their covariance matrix in the presence of nonspherical innovations with an unknown covariance matrix.

Let $y_t = X_t\beta + \varepsilon_t$ be a multiple linear regression model, where the innovations process ε_t is Gaussian with mean 0, but with true, nonspherical covariance matrix Ω (for example, the innovations are heteroscedastic or autocorrelated). Also, suppose that the sample size is T and there are p predictors (including an intercept). Then, the FGLS estimator of β is

$$\hat{\beta}_{FGLS} = \left(X^{\mathsf{T}}\hat{\Omega}^{-1}X\right)^{-1}X^{\mathsf{T}}\hat{\Omega}^{-1}y,$$

where $\hat{\Omega}$ is an innovations covariance estimate based on a model (e.g., innovations process forms an AR(1) model). The estimated coefficient covariance matrix is

$$\hat{\Sigma}_{FGLS} = \hat{\sigma}^2_{FGLS}\left(X^{\mathsf{T}}\hat{\Omega}^{-1}X\right)^{-1},$$

where

$$\hat{\sigma}^2_{FGLS} = \frac{y^{\mathsf{T}}\left[\hat{\Omega}^{-1} - \hat{\Omega}^{-1}X\left(X^{\mathsf{T}}\hat{\Omega}^{-1}X\right)^{-1}X^{\mathsf{T}}\hat{\Omega}^{-1}\right]y}{T-p}.$$

FGLS estimates are computed as follows:

1. OLS is applied to the data, and then residuals $(\hat{\varepsilon}_t)$ are computed.
2. $\hat{\Omega}$ is estimated based on a model for the innovations covariance.
3. $\hat{\beta}_{FGLS}$ is estimated, along with its covariance matrix $\hat{\Sigma}_{FGLS}$.
4. Optional: This process can be iterated by performing the following steps until $\hat{\beta}_{FGLS}$ converges.
   a. Compute the residuals of the fitted model using the FGLS estimates.
   b. Apply steps 2–3.

If $\hat{\Omega}$ is a consistent estimator of Ω and the predictors that comprise X are exogenous, then FGLS estimators are consistent and efficient.

Asymptotic distributions of FGLS estimators are unchanged by repeated iteration. However, iterations might change finite sample distributions.

参考：https://www.mathworks.com/help/econ/fgls.html

Box-Cox 变换：

对解释变量进行 $Box-Cox$ 变换也是一种处理异方差问题的手段，让数据看起来更符合正态分布.

注意 $Box-Cox$ 是针对因变量的处理，例如原本的模型是 $Y=\beta_0+\beta_1X+\varepsilon$，为使方差近似服从同方差的正态分布，变换为 $Y^{(0)}=log(Y)=\beta_0+\beta_1X+\varepsilon$ 再进行回归.

$$y_i^{(\lambda_1,\lambda_2=0)}=\begin{cases}\dfrac{(y_i+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1\neq 0, y_i>-\lambda_2\\ log(y_i+\lambda_2), & \lambda_1=0, y_i>-\lambda_2\end{cases}$$

此即双参数的 $Box-Cox$ 变换，令 $\lambda_2=0$ 即为单参数的 $Box-Cox$ 变换.

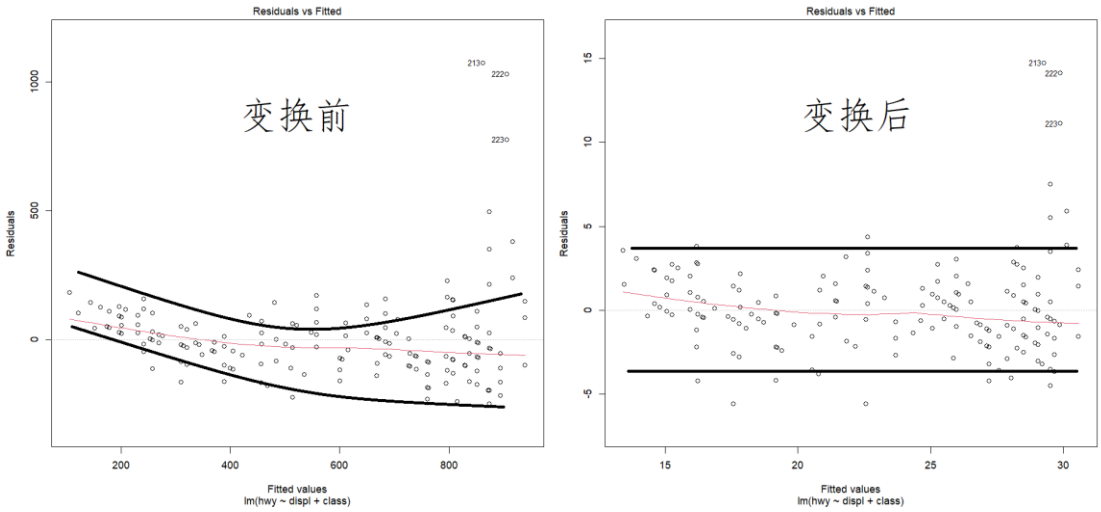
如何确定参数的取值？配合拟合优度检验，选择使得似然函数 $L(\boldsymbol{\beta},\sigma^2,\lambda)$ 最大的 λ，由于解析解难以得到，逐个枚举是一个可以一试的办法.

常用的 $Box-Cox$ 变换有 $log(y_i)$、$\sqrt{y_i}$ 和 $\frac{1}{y_i}$.

幂变换：$y_i^{(\lambda)}=\begin{cases}\dfrac{y_i^{\lambda}-1}{\lambda(y_1y_2\ldots y_n)^{\frac{\lambda-1}{n}}}, & \lambda\neq 0\\ \sqrt[n]{(y_1y_2\ldots y_n)}\,log(y_i), & \lambda=0\end{cases}$，$(y_1y_2\ldots y_n)^{\frac{1}{n}}$ 为几何平均值.

$Box-Cox$ 变换的雅可比矩阵正是 $(y_1y_2\ldots y_n)^{\lambda-1}$.

某数据在对因变量进行 $Box-Cox$ 变换前后残差的分布：



多重共线性（复共线性）：

多重共线性即自变量间的线性关系，若自变量间存在严格的线性关系则称这些自变量间具有完全共线性，但完全共线性通常并不多见，更多的表现为近似共线性.

多重共线性在设计矩阵 $\boldsymbol{X}$ 上的直接体现是 $\boldsymbol{X}$ 的列相关性较强，这可能导致 $\hat{\boldsymbol{\beta}}$ 的一些估计值异常地大（这将在稍后证明）；此外直观角度理解起来，如果存在多重共线性，不妨假设有模型 $Z=4.58+0.4X-2Y+\varepsilon$、$Y=2X$，那么回归方程 $\hat{Z}=4.58+2.4\hat{X}-3\hat{Y}$ 与 $\hat{Z}=4.58-1.6\hat{X}-\hat{Y}$ 是完全等价的，回归系数可能不止有一种可行的估计，也有可能使得回归系数的符号与实际问题中其现实意义完全背离，如同我们举的例子中出现的问

题一样. 以上种种情况都会使我们对回归结果做出违背真实情况的错误解读.

首先考虑两个变量间的线性关系，我们有无量纲的相关系数来衡量两变量间线性关系的强弱，它可以衡量两个变量间的相关关系，如果两个变量存在$Y=\pm X+\mu+\varepsilon$的关系则$\rho(X,Y)$应近乎等于 1 或 -1，我们认为两个变量相关性非常强（一般相关系数大于 0.8 就认为存在强烈的线性关系），那么如何判断多个变量间可能存在的相关关系呢？比如$Z=10+2X-7Y+\varepsilon$，这时三个变量间存在强烈的复共线性，简单的观察相关系数的办法不再可行，前文已经提到多重共线性会对回归产生很多糟糕的影响，我们迫切需要一些可用的办法来判断数据的多重共线性. 判断多重共线性有多种方式，从理论出发最直观的方法是特征根法，最常见的方法是 VIF（方差膨胀因子）检验.

特征根判别法：

设计矩阵$\boldsymbol{X}$列满秩（如果都不列满秩说明列向量间存在异常严重的线性相关），记去掉第一列的设计矩阵标准化后为$\boldsymbol{X}'$，考虑$\boldsymbol{X}'^T\boldsymbol{X}'$：如果存在很小的特征值说明$\boldsymbol{X}'^T\boldsymbol{X}'$线性相关性强烈，设$\boldsymbol{X}'^T\boldsymbol{X}'$的特征值为$\lambda_1,\lambda_2,\dots,\lambda_n>0$，某特征值$\lambda_k$对应的标准正交特征向量记为$\xi_k$，若存在一个明显近似于 0 的特征值，不妨设$\exists\lambda_i\approx 0$，则有

$$\lambda_i\approx 0\Leftrightarrow \boldsymbol{X}'^T\boldsymbol{X}'\xi_k=\lambda_i\xi_i\approx\boldsymbol{0}\Rightarrow\|\boldsymbol{X}'^T\boldsymbol{X}'\xi_k\|=\lambda_i\approx 0\Leftrightarrow\left\|\frac{1}{\|\xi_k^T\|}\boldsymbol{X}'^T\xi_k^T\boldsymbol{X}'\xi_k\right\|=\lambda_i\approx 0$$

$$\Leftrightarrow\|\boldsymbol{X}'\xi_k\|^2=\lambda_i\approx 0\Leftrightarrow\boldsymbol{X}'\xi_k\approx\boldsymbol{0}\Leftrightarrow\sum_{j=1}^{p-1}\xi_k^{(j)}\boldsymbol{x}_j\approx\boldsymbol{0}$$

这说明$\boldsymbol{X}'$，或者讲$\boldsymbol{X}$的列线性相关性很强，因此可以通过观察特征值判断多重共线性，如果$\boldsymbol{X}'^T\boldsymbol{X}'$有一个较小的特征值则说明多重共线性可能很强（在稍后证明多重共线性导致$\widehat{\boldsymbol{\beta}}$的一些估计值异常大的过程中，会从另一个方面体现这一点的原因）；不过因为没有标准所以这样判断很不准确，经验而谈，设

$$\omega=\frac{\lambda_{max}}{\lambda_{min}}$$

如果$\omega<100$认为复共线性弱，$100\leqslant\omega\leqslant 1000$认为可能存在一定的共线性，$\omega>1000$则认为有很强的复共线性，但这样判断是十分粗糙的.

特征根法的原理对后文阐述复共线性使得估计向量长度异常增大原因有重要作用.

VIF 检验：

VIF 检验的思想是对每个X_i做新因变量、剩余原自变量做新的自变量进行建模、做线性回归，即

$$X_i=\zeta_0+\zeta_1X_1+\cdots+\zeta_{i-1}X_{i-1}+\zeta_{i+1}X_{i+1}+\cdots+\zeta_{p-1}X_{p-1}+\epsilon_i,\ \ i=1,2,\dots,p-1$$

这个想法很非常直观，每个X_i做因变量模型进行回归得到的拟合优度记为R_i^2，VIF 检验的想法也较为直观，如果原模型的自变量间完全无关，那么新模型的拟合效果应该非常差（因为他们之间根本没有线性关系），RSS 将会较大，则$\forall i,\ R_i^2$应该比较小，但若反之原自变量间有很强的多重共线性，那么$\exists i_0$使得$R_{i_0}^2$较大，令

$$VIF_i = \frac{1}{1-R_i^2} = \frac{\hat{\beta}_i \hat{\sigma}_i^2}{RSS_i}，\ 容忍度 = \frac{1}{VIF_i} = 1 - R_i^2$$

原变量间多重共线性若较弱，则 R_i^2 都较小，相应的 VIF_i 都较小；原变量间共线性较强，则某个 R_i^2 偏大，相应的 VIF_i 较大，综上所述，我们希望得到的 VIF_i 比较小，这样就不需要后续关于多重共线性的数据处理了.

VIF_i 取值范围为 $(1, +\infty)$，经验上看，模型自变量较少则在 $VIF_i > 5$ 时认为有较明显的复共线性，自变量较多则在 $VIF_i > 10$ 时认为有较明显的共线性，假使 $VIF_i > 100$ 则表明已经存在异常严重的多重共线性了，在实际应用中也可以对比观察各 VIF_i 数值，着重关注显著偏大者.

判断出多重共线性的存在后该如何处理？最“简单粗暴”、最“不漂亮”的办法是直接删掉“冗余”的具有线性关系的变量再进行回归，但一般不会这样做，因为没有充分的理由选择具体删掉某几个变量而保留某几个变量，通常 PCA 考虑主成分或对模型添加约束是可取的，即进行正则化处理，比如进行岭回归、Lasso 回归，这将在后文提到.

最后我们用 MSE 评估一下多重共线性对估计的影响，MSE 综合考虑了无偏性与有效性，是行之有效的评估准则.

$$MSE(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2 = \mathbb{E}\left(\left\|\widehat{\boldsymbol{\beta}}\right\|_2^2\right) - \boldsymbol{\beta}^T\boldsymbol{\beta} = tr\left(Cov(\widehat{\boldsymbol{\beta}})\right) + \left\|\mathbb{E}\,\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2$$

可以看出，估计 $\widehat{\boldsymbol{\beta}}$ 的长度的期望可以分解为一个包含 $MSE(\widehat{\boldsymbol{\beta}})$ 的式子，这是我们为什么提到 MSE 的原因之一；为了消除量纲的影响这里考虑 $\boldsymbol{X}$ 标准化后去掉第一列得到的矩阵 $\boldsymbol{X}'$ 与标准化后新模型的回归系数 $\boldsymbol{\beta}'$ 及其估计 $\widehat{\boldsymbol{\beta}}'$（去掉截距项，新的截距项等于 $\bar{y}$），注意到 $\boldsymbol{X}'^T\boldsymbol{X}'$ 作为对称的(半)正定矩阵，存在可以使 $\boldsymbol{X}'^T\boldsymbol{X}'$ 对角化的 p 维正交阵 $\boldsymbol{\Phi}$，不妨设 $\boldsymbol{\Phi}$ 的第 i 列为 $\boldsymbol{X}'^T\boldsymbol{X}'$ 的第 i 个标准正交的特征向量，对应特征值为 λ_i：

$$\mathbb{E}\left(\left\|\widehat{\boldsymbol{\beta}}'\right\|_2^2\right) = \|\boldsymbol{\beta}'\|_2^2 + MSE(\widehat{\boldsymbol{\beta}}')$$

$$= \|\boldsymbol{\beta}'\|_2^2 + tr\left(Cov(\widehat{\boldsymbol{\beta}}')\right) + \left\|\mathbb{E}\,\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}'\right\|_2^2$$

$$= \|\boldsymbol{\beta}'\|_2^2 + tr(\sigma^2(\boldsymbol{X}'^T\boldsymbol{X}')^{-1}) + 0$$

$$= \|\boldsymbol{\beta}'\|_2^2 + \sigma^2\, tr\left(\boldsymbol{\Phi}\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{pmatrix}\boldsymbol{\Phi}^T\right)^{-1}$$

$$= \|\boldsymbol{\beta}'\|_2^2 + \sigma^2\, tr\left(\boldsymbol{\Phi}\begin{pmatrix}\frac{1}{\lambda_1} & & & \\ & \frac{1}{\lambda_2} & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_p}\end{pmatrix}\boldsymbol{\Phi}^T\right)$$ 注意到: $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$

$$= \|\boldsymbol{\beta}'\|_2^2 + \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

在前文特征根法的部分中我们提到，如果 $\boldsymbol{X'}^T\boldsymbol{X'}$ 存在一个较小特征值说明很可能存在多重共线性，可以看出如果自变量间具有多重共线性，会存在一个较小的特征值使得估计的长度期望 $\mathbb{E}\left(\left\|\widehat{\boldsymbol{\beta}}'\right\|_2^2\right)$ 偏离真实回归系数 $\|\boldsymbol{\beta}'\|_2^2$，对单个系数而言某个回归系数的绝对值会异常地大，距离真实值有很大偏差，且较小特征值越小偏离便越大.

系数的假设检验：特殊的方差分析

一般线性假设的检验（高斯-马尔可夫假设）

H_0：线性约束 $\boldsymbol{A\beta} = \boldsymbol{b}$ 成立，$\boldsymbol{A}$ 是秩为 m 的 $m \times p$ 矩阵

H_1：$\boldsymbol{A\beta} \neq \boldsymbol{b}$

无约束模型 OLS 的 RSS 为 $RSS_{\text{OLS}} = \hat{\boldsymbol{\varepsilon}}_{\text{OLS}}^T\hat{\boldsymbol{\varepsilon}}_{\text{OLS}} = \boldsymbol{y}^T\left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}_{\text{OLS}}\right) = \boldsymbol{y}^T(\boldsymbol{I} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}$，在 H_0 成立条件下约束模型 CLS 的 RSS 为 $RSS_{\text{CLS}} = \boldsymbol{y}^T\left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}_{\text{CLS}}\right)$，因此

$$\begin{cases} RSS_{\text{OLS}} = \boldsymbol{y}^T(\boldsymbol{I} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} \\ RSS_{\text{CLS}} = \boldsymbol{y}^T\left(\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T(\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{b})\right) \end{cases}$$

有检验统计量：

$$F = \frac{\frac{RSS_{\text{CLS}} - RSS_{\text{OLS}}}{m}}{\frac{RSS_{\text{OLS}}}{n-p}} \overset{L}{\sim} F(m, n-p)$$

拒绝域：$F > F_\alpha(m, n-p)$ 或者 p-value $\leqslant \alpha$，p-value $= P(F(m, n-p) \geqslant F)$

F 检验：回归方程的显著性检验 由给定特殊约束一般线性假设检验导出（高斯-马尔可夫假设）

H_0：$\beta_1 = \beta_2 = \ldots = \beta_p = 0$（$\left(\boldsymbol{0}, \boldsymbol{I}_{p-1}\right)\boldsymbol{\beta} = \boldsymbol{0}$）

H_1：β_1、β_2、…、β_p不全为 0（$\left(\boldsymbol{0}, \boldsymbol{I}_{p-1}\right)\boldsymbol{\beta} \neq \boldsymbol{0}$）

在 H_0 成立条件下，检验统计量：

$$F = \frac{\frac{TSS - RSS}{p-1}}{\frac{RSS}{n-p}} = \frac{\frac{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} - \bar{y}\boldsymbol{1}^T\boldsymbol{y}}{p-1}}{\frac{RSS}{n-p}} \overset{L}{\sim} F(p-1, n-p)$$

拒绝域：$F > F_\alpha(p-1, n-p)$ 或者 p-value $\leqslant \alpha$，p-value $= P(F(p-1, n-p) \geqslant F)$

t 检验：回归系数的显著性检验 由给定特殊约束一般线性假设检验导出（高斯-马尔可夫假设）

H_0：$\beta_i = 0$

H_1：$\beta_i \neq 0$

在 H_0 成立条件下，检验统计量：

$$\frac{\hat{\beta}_i}{\sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}} \sim N(0,1),\ \frac{RSS}{\sigma^2} \sim \chi^2(n-p) \text{ 相互独立}$$

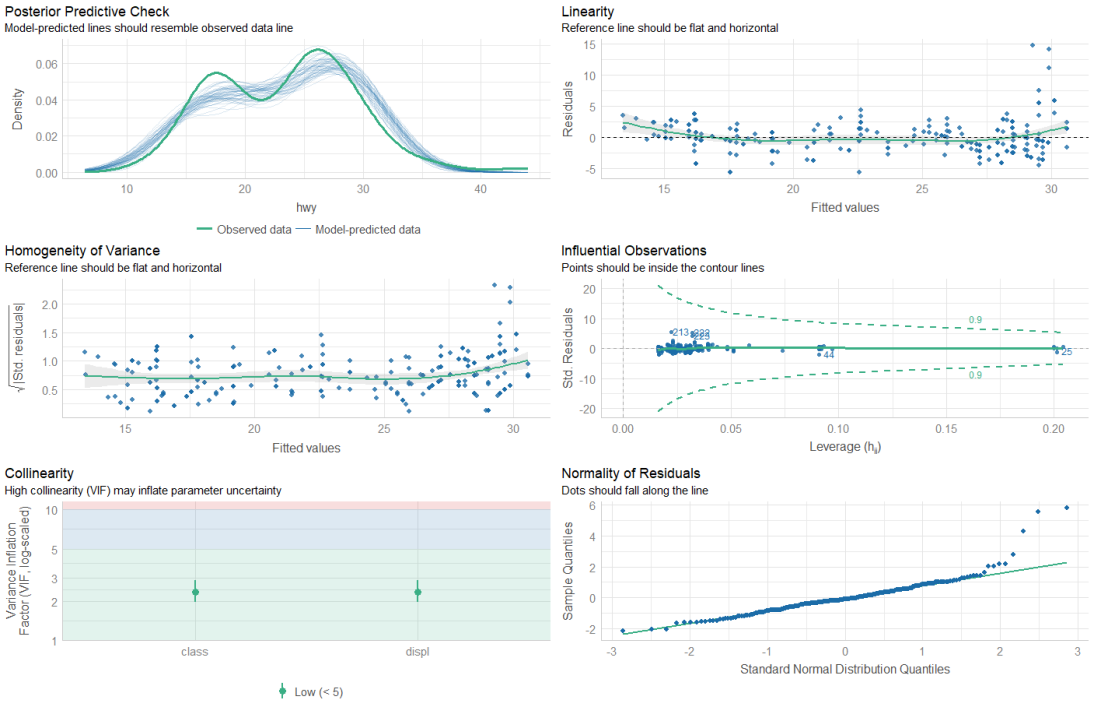
$$\Rightarrow\ t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}} = \frac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}} \sim t(n-p)$$

拒绝域：$|t| > t_{\frac{\alpha}{2}}(n-p)$ 或者 p-value $\leqslant \frac{\alpha}{2}$，p-value $= P(\,t(n-p) \geqslant |t|\,)$

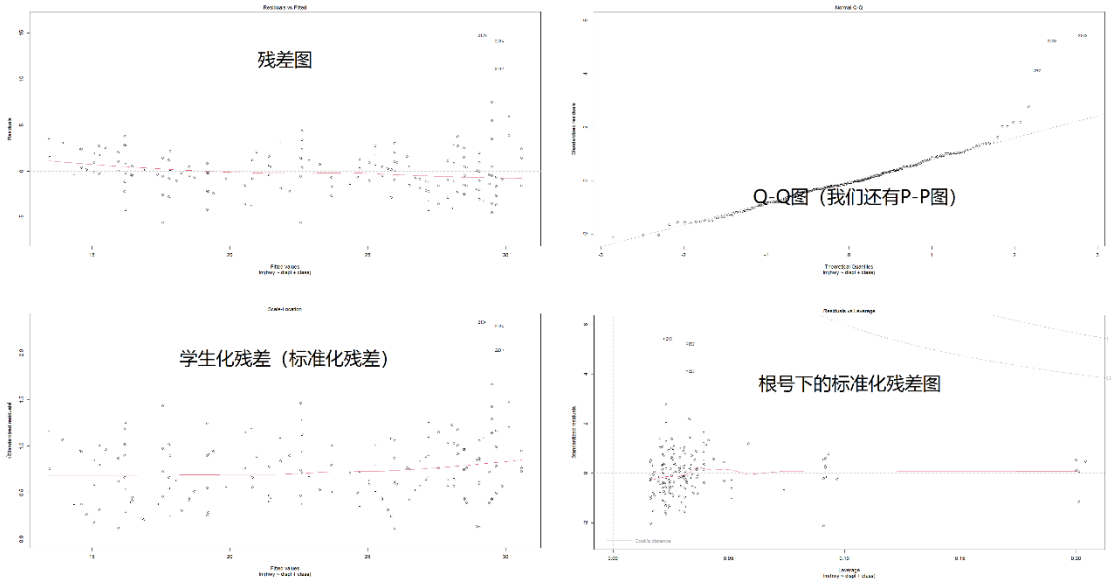
回归诊断：

常见的诊断图有学生化残差图、Q-Q 图等，不再这篇文档中做详细说明，原理并不复杂. 值得一提的是，R 的“performance”包我感觉挺不错的.

```
library(tidyverse)
library(tidymodels)
library(performance)

# use data "MPG"

fit <- linear_reg() %>%

set_engine("lm") %>%

fit(hwy ~ displ + class, data = mpg)

check_model(fit)
```



例如：



学生化残差图、Q-Q 图等等图像的原理没有复杂到难以理解，比较清晰明了，可用于初步判断数据正态性、峰度、偏度与趋势等等，需要说明的是 Q-Q 图斜率为标准差、截距为算术平均值，如果直线在靠近左下角处明显向下偏移说明数据分布左厚尾，

正态性检验不是本文重点，不在赘述.

岭回归与岭估计：

注意!!! 岭估计与 Lasso 估计都是有偏估计，是带特定约束的最小二乘估计——作为 OLS 的改良，相较于 OLS，岭估计与 Lasso 估计用无偏性换取了有效性，损失了部分信息，但能减小回归系数估计的方差，对病态数据的回归效果均远远强于 OLS. 一定程度上减轻了多重共线性带来的影响.

此外，岭回归一定程度上能缓解多重共线性的问题，提升模型泛化能力.

岭估计于 1970 年由 *Hoerl* 和 *Kennard* 首次提出，在 OLS 基础上对回归系数添加了 L^2 范数约束（这类操作称为 $\boldsymbol{L^2}$ **正则化**），为什么做约束后能减轻多重共线性的影响呢？在前文关于多重共线性的部分有证明：

$$\mathbb{E}\left(\left\|\widehat{\boldsymbol{\beta}}'\right\|_2^2\right)=\|\boldsymbol{\beta}'\|_2^2+MSE\left(\widehat{\boldsymbol{\beta}}\right)=\|\boldsymbol{\beta}'\|_2^2+\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

这表明多重共线性会使得某些回归系数绝对值偏大，如果我们限制回归系数的长度 $\left\|\widehat{\boldsymbol{\beta}}\right\|$ 再进行回归则能减轻多重共线性的影响，这样便能减少 $MSE\left(\widehat{\boldsymbol{\beta}}\right)$.

定理：$\exists k>0,\ s.t.\ MSE\left(\widehat{\boldsymbol{\beta}}(k)\right)<MSE\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)$，这个定理的证明需要用到引理：岭估计或 OLS 的 MSE 与他的典则回归系数的岭估计或 OLS 的 MSE 相等，定理证明的思路是从 $\widehat{\boldsymbol{\beta}}(k)$ 的典则回归系数的岭估计 $\widehat{\boldsymbol{\alpha}}(k)$ 出发，利用 $MSE\left(\widehat{\boldsymbol{\alpha}}(k)\right)tr\left(Cov\left(\widehat{\boldsymbol{\alpha}}(k)\right)\right)+\|\mathbb{E}\,\widehat{\boldsymbol{\alpha}}(k)-\widehat{\boldsymbol{\alpha}}(k)\|$，这里略去证明过程. 这个定理说明了岭估计的价值，在 MSE 的意义下选取合适的 k 能使得其优与 OLS.

岭估计是一种压缩估计：$\forall k>0,\ \left\|\widehat{\boldsymbol{\beta}}(k)\right\|<\left\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\|$.

正则化优点是能防止过拟合，但相应地可能导致欠拟合，这包括岭估计的 L^2 正则化.

岭回归模型：

$$\begin{cases}Y=\beta_0+\beta_1X_1+\ldots+\beta_{p-1}X_{p-1}+\varepsilon\\ \left\|\boldsymbol{\beta}_{1,2,\ldots,p-1}\right\|_2\propto\beta_1^2+\beta_2^2+\ldots+\beta_{p-1}^2\le k\end{cases}$$

损失函数：带有 L^2 范数约束长度的 MSE，显然这也是凸函数.

$$\begin{aligned}\mathcal{L}&=MSE+nk\|\boldsymbol{\beta}\|_2^2\\&=\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+nk\|\boldsymbol{\beta}\|_2^2\\&=\boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X}+nk\boldsymbol{I})\boldsymbol{\beta}-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{y}^T\boldsymbol{y}\\&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+nk\sum_{i=1}^{n}{\beta_i}^2\\&\propto\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+k\|\boldsymbol{\beta}\|_2^2\\&=SSE+k\|\boldsymbol{\beta}\|_2^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+k\sum_{i=1}^{n}{\beta_i}^2\end{aligned}$$

考虑 *Lagrange* 乘子法求解约束下的 MSE 来证明，并不复杂，这里略去步骤，最终得到

$$\widehat{\boldsymbol{\beta}}(k) = argmin_{\widehat{\beta}_0,\widehat{\beta}_1,\dots}(\|\boldsymbol{y}-\boldsymbol{X\beta}\| + k\|\boldsymbol{\beta}\|)$$
$$= (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$Cov\left(\widehat{\boldsymbol{\beta}}(k)\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}$$

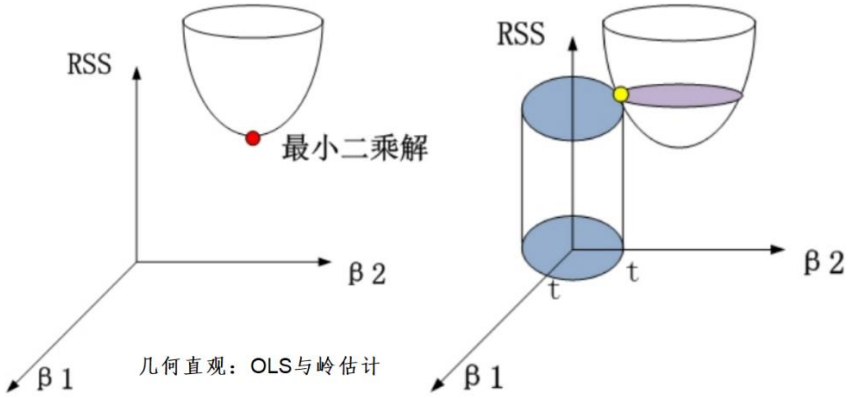
有偏性 $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}(k)\right) = (\boldsymbol{X}^T\boldsymbol{X} + nk\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta} \neq \boldsymbol{\beta},\ as\ k \neq 0, \boldsymbol{\beta} \neq \boldsymbol{0}$

本文只对岭回归的损失函数确为 $MSE + nk\|\boldsymbol{\beta}\|_2^2$ 证明：

$$\mathcal{L} = MSE + k\|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}$$
$$grad(\mathcal{L}) = 2(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y} \overset{order}{=} \boldsymbol{0}$$
$$\Rightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$，这与应用 $Lagrange$ 乘子法求解结果一致

k 实质上不过是 *Lagrange* 乘子.

$k = 0$ 岭估计等同于 OLS，k 越大与 OLS 相差越大；岭估计相较于 OLS 减小了 MSE（这不难观察出，增大了偏差但更多地减小了方差），但这增大了 RSS（增大 RSS 的原因是 OLS 以 RSS 为损失函数的，已经使得 RSS 最小）.



图源：徐老师教学 PPT

但该如何确定一个合适的 k，使得岭估计相较于 OLS 不会有不可接受的精度丢失（表现为 RSS 过大）、回归系数出现不合理符号并出现过大的偏差、MSE 较大？事实上这些就是我们希望得到的 k 应当具备的性质，但直到今天都没有一个泛用、最佳的办法，这里介绍岭迹法和方差膨胀因子法，均为经验方法.

事实上，在证明参数为某个 k 的岭估计 MSE 小于 OLS 的 MSE 过程中可以得到，最优的 k，即使得岭估计 MSE 最小都应满足一个四次多项式，但由于式子包含了典则回归系数与误差的方差，这个四次多项式是不可解的.

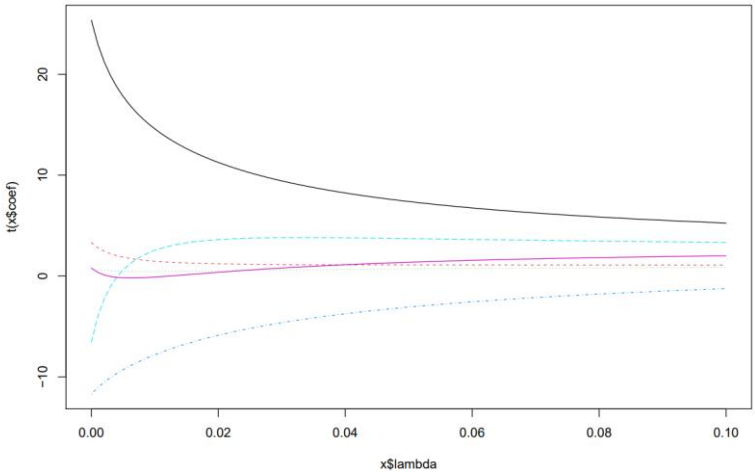
***Hoerl-Kennard* 公式：**

1970 年由 *Hoerl-Kennard* 提出.

$\hat{k} = \frac{\hat{\sigma}^2}{\hat{\boldsymbol{\alpha}}^T\hat{\boldsymbol{\alpha}}}$ 或 $\frac{\hat{\sigma}^2}{\max_i \hat{\alpha}_i^2}$，$\hat{\alpha}_i$为点则回归系数的估计(即标准化模型的 OLS)

岭迹法：

观察岭迹图——一张自变量为 k、因变量为 $\hat{\beta}_i(k)$ 的图像，通常我们倾向于选择尽可能小但使得岭估计相对 k 大体稳定、不再随着 k 增大而显著变化的一个 k 值.



如果岭迹在一定的 k 范围内上下波动，说明这时 OLS 可能就不是一个良好的估计，因为岭迹并不平稳，有明显的多重共线性，这时需要选取一个合理的 k 使得岭迹在此以后接近平稳；也可能岭回归系数估值近乎不受 k 影响，以极其缓慢趋势变化，近乎水平直线，这时说明多重共线性较弱，OLS 即是一个理想的估计（OLS 可视为 $k = 0$ 的特殊岭估计）.

注：一时间没找到合适的数据能绘一个趋势明显的岭迹图……不得已再次借用徐老师 PPT 上的图片

方差膨胀因子法：

由 $Cov\left(\hat{\boldsymbol{\beta}}(k)\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1} = \sigma^2 VIF$ 得出，VIF_i 与 k 呈反比例关系，选择合适的 k 使得对所有的 $VIF_i < j$，j 为一个常数，常用的经验数值譬如 10.

双 h 公式：

这个方法的原理相对麻烦……可以在一些文献中参考，综合而言是不错的方法.

岭估计可以通过假定零均值的正态分布的先验分布推导而来，最大后验概率的估计即岭估计，包括 Lasso 估计也可以这样从贝叶斯后验概率角度解释，这可能不是那么容易理解. 有一篇文献描述得非常细致详细，我会把他放在附录, 英文文献参考：https://pdf.sciencedirectassets.com/271708/1-s2.0-S0167947320X00028/1-s2.0-S0167947320300086/am.pdf?X-Amz-Security-Token=IQoJb3JpZ2l

---

Lasso 回归与 Lasso 估计：

注意，在岭估计的部分就有提到的：岭估计与 Lasso 估计都是有偏估计，是带特定约束的最小二乘估计——作为 OLS 的改良，相较于 OLS，岭估计与 Lasso 估计用无偏性换取了有效性，损失了部分信息，但能减小回归系数估计的方差，对病态数据的回归效果均远远强于 OLS. 一定程度上减轻了多重共线性带来的影响，这值得强调两次.

Lasso 估计于 1986 年被首次提出，1996 年 *Robert Tibshirani* 独立发现并推广了 Lasso 回归，Lasso 估计在 OLS 基础上对回归系数添加了 L^1 范数约束（这类操作称为 $\boldsymbol{L^1}$ 正则化），为什么做约束后能减轻多重共线性的影响在前文已经讨论过了.

Lasso 回归不仅能应对多重共线性情况，同时还具有变量筛选的功能：Lasso 回归通过强制回归系数的绝对值之和小于 λ，能够强制某些系数为 0，达到了变量筛选的目的，而这岭估计并不能做到. 因此相较于岭回归，Lasso 回归的结果更具备可解释性..

Lasso 回归模型：

$$\begin{cases} Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon \\ |\beta_1| + |\beta_2| + \ldots + |\beta_{p-1}| \le \lambda \end{cases}$$

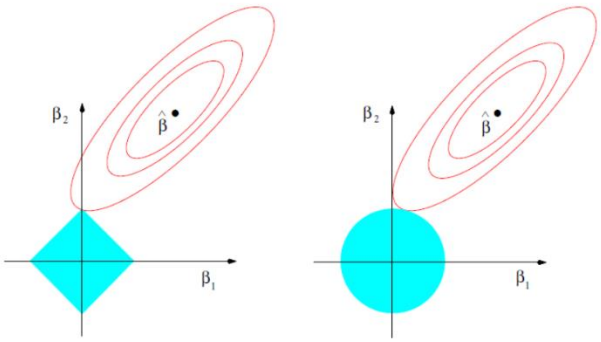
损失函数：带有 L^1 范数约束长度的 MSE.

$$\begin{aligned} \mathcal{L} &= MSE + n\lambda \|\boldsymbol{\beta}\|_1 \\ &= \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + n\lambda \|\boldsymbol{\beta}\|_1 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + n\lambda \sum_{i=1}^{n} |\beta_i| \\ &\propto \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \end{aligned}$$

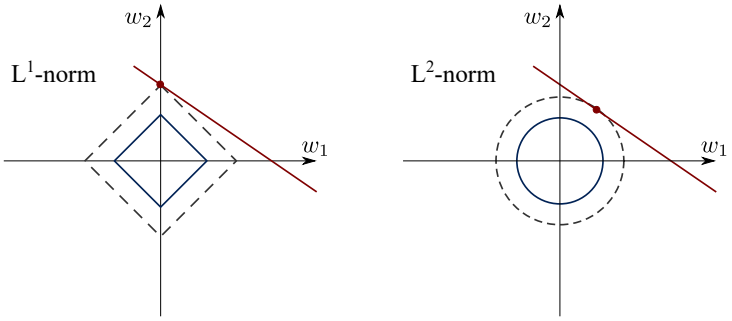
$$= SSE + \lambda\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{n}|\beta_i|$$

LASSO 估计与岭估计

LASSO 估计与岭估计不同之处是在最小化 $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ 时，对 $\boldsymbol{\beta}$ 施加不同的约束条件. 以二元线性回归模型为例，岭估计（右图）约束条件是 $\hat{\beta}_1^2 + \hat{\beta}_2^2 \leq t$，而 LASSO 估计的约束条件是 $|\hat{\beta}_1| + |\hat{\beta}_2| \leq t$. LASSO 估计不仅是压缩估计，还有变量选择的功能.



图源：徐老师教学 PPT



以上是普通 Lasso 估计，此外我们还有 Group Lasso、Fused Lasso、Adaptive Lasso 和 Prior Lasso 等等多种回归，方便起见这里不展开细说了.

Elastic Net 回归：

2005 年由 *Zou* 和 *Hastie* 提出，Elastic Net 回归综合了 L^1、L^2 正则化，混合了两种约束条件，可以通过控制混合参数决定何种正则化占何种地位.

当样本量 n 小于自变量数量 p 时，Lasso 回归只能选取至多 n 个自变量且倾向于从任何一组高度相关的自变量中；即使在 $n > p$ 时，面对强相关的自变量岭回归通常效果比 Lasso 好很多.

主成分估计与主成分回归：

主成分回归（PCR）是基于主成分分析理论 PCA 的，是将 PCA 技术应用于回归系数估计的手段.

主成分

记线性回归模型为 $\boldsymbol{y}=\alpha_0\mathbf{1}+\boldsymbol{X\beta}+\boldsymbol{e}$，设计矩阵 $\boldsymbol{X}$ 已经中心化，即 $\mathbf{1}^T\boldsymbol{X}=\mathbf{0}$. 记 $\boldsymbol{X}^T\boldsymbol{X}$ 的 $p-1$ 个特征值 $\lambda_1\geq\cdots\geq\lambda_{p-1}$，对应的标准正交化的特征向量为 $\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_{p-1}$，记 $\boldsymbol{\Phi}=(\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_{p-1})$，于是

$$\boldsymbol{y}=\alpha_0\mathbf{1}+\boldsymbol{X\beta}+\boldsymbol{e}=\alpha_0\mathbf{1}+\boldsymbol{X\Phi\Phi}^T\boldsymbol{\beta}+\boldsymbol{e}.$$

记 $\boldsymbol{Z}=\boldsymbol{X\Phi}$（正交变换），$\boldsymbol{\alpha}=\boldsymbol{\Phi}^T\boldsymbol{\beta}$，可得线性回归模型的**典则形式**

$$\boldsymbol{y}=\alpha_0\mathbf{1}+\boldsymbol{Z\alpha}+\boldsymbol{e}$$

定义：典则设计矩阵 $\boldsymbol{Z}\triangleq(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{p-1})=(\boldsymbol{X\varphi}_1,\cdots,\boldsymbol{X\varphi}_{p-1})$，有

$$\boldsymbol{z}_j=\boldsymbol{X\varphi}_j\triangleq(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{p-1})\begin{pmatrix}c_{1,j}\\ \vdots\\ c_{p-1,j}\end{pmatrix}=c_{1,j}\boldsymbol{x}_1+\cdots+c_{p-1,j}\boldsymbol{x}_{p-1}$$

显然，$\boldsymbol{z}_j$ 是原自变量 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{p-1}$ 的线性组合，称为**主成分**. 最大特征值 λ_1 对应的 $\boldsymbol{z}_1$ 称为**第一主成分**，以此类推. 显然，主成分之间线性无关.

主成分与特征值

假设 $\boldsymbol{X}$ 已经中心化，λ_j 为 $\boldsymbol{X}^T\boldsymbol{X}$ 的第 j 个特征值，$\boldsymbol{\varphi}_j$ 为对应的标准正交化的特征向量，$\boldsymbol{z}_j=\boldsymbol{X\varphi}_j$ 为第 j 主成分. 记 $\boldsymbol{\Phi}=(\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_{p-1})$，$\boldsymbol{Z}=(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{p-1})$，其中 $\boldsymbol{z}_j=(z_{1j},\cdots,z_{nj})^T$.

因 $\mathbf{1}^T\boldsymbol{Z}=\mathbf{1}^T\boldsymbol{X\Phi}=\mathbf{0}$，故 $\boldsymbol{Z}$ 已中心化，即 $\bar{z}_j=\frac{1}{n}\sum\limits_{i=1}^{n}z_{ij}=0.$

又因为 $\boldsymbol{X}^T\boldsymbol{X\varphi}_j=\lambda_j\boldsymbol{\varphi}_j$，有

$$\begin{aligned}\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)^2=\sum_{i=1}^{n}z_{ij}^2=\boldsymbol{z}_j^T\boldsymbol{z}_j=(\boldsymbol{X\varphi}_j)^T\boldsymbol{X\varphi}_j=\boldsymbol{\varphi}_j^T\boldsymbol{X}^T\boldsymbol{X\varphi}_j\\ =\boldsymbol{\varphi}_j^T\lambda_j\boldsymbol{\varphi}_j=\lambda_j\end{aligned}$$

这说明特征值 λ_j 可度量主成分 $\boldsymbol{z}_j$ 的“变化”. 因此，各主成分变化程度从大到小排序为：第一主成分 $>$ 第二主成分 $>\cdots>$ 第 $p-1$ 主成分.

主成分估计

实施步骤：

1. 通过正交变换 $\boldsymbol{Z}=\boldsymbol{X\Phi}$，获得主成分 $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{p-1}$，即 $\boldsymbol{Z}$ 的所有列（不存在复共线性）；
2. 剔除对应特征值较小的主成分，剩下前 r 个主成分 $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_r$，其中 r 值取决于**贡献率** $\sum_{j=1}^{r}\lambda_j/\sum_{j=1}^{p-1}\lambda_j$ 即“前 r 个主成分对应特征值的和”与“所有特征值的和”之比；
3. Y 对前 r 个主成分最小二乘回归 $\hat{Y}=\hat{\alpha}_0+\hat{\alpha}_1Z_1+\cdots+\hat{\alpha}_rZ_r$，再将 $Z_j=c_{1,j}X_1+\cdots+c_{p-1,j}X_{p-1}$ 代入上式，获得主成分估计

$$\widetilde{Y}=\tilde{\beta}_0+\tilde{\beta}_1X_1+\cdots+\tilde{\beta}_{p-1}X_{p-1}.$$

性质：主成分估计 $\tilde{\boldsymbol{\beta}}$ 有偏. 若适当选择 r，有 $MSE(\tilde{\boldsymbol{\beta}})<MSE(\hat{\boldsymbol{\beta}})$，即就 MSE 而言，存在优于最小二乘估计的主成分估计.

LOWESS：局部加权最小二乘估计

***

PLS：偏最小二乘估计

***

附录：收敛性定义

1. 函数列的点态收敛 $\lim\limits_{n\to\infty} f_n(x)=f(x)$

对给定的一点x_0，$\forall\varepsilon>0$，$\exists N\in\mathbb{N}^+$，$s.t.\ \forall n>N$，$|f_n(x_0)-f(x_0)|<\varepsilon$

2. 一致收敛 $f_n(x)\rightrightarrows f(x)\ (n\to\infty)$

$\forall\varepsilon>0$，$\exists N\in\mathbb{N}^+$，$s.t.\ \forall n>N$、$\forall x\in E$，$|f_n(x)-f(x)|<\varepsilon$

3. 近一致收敛

$\forall\delta>0$，$\exists E$ 的可测子集 E_δ，$s.t.$在 $E\backslash E_\delta$ 上 $\{f_n\}$ 一致收敛于f，且 $m(E\backslash E_\delta)<\delta$

4. 几乎处处收敛 $f_n(x)\to f(x)\ (n\to\infty)$ $a.e.\ x_n\in E$，概率测度下记为 $a.s.$

$$m\left(\left\{x:\lim_{n\to\infty} f_n(x)\neq f(x)\right\}\right)=0$$

即$f_n(x)\to f(x)$几乎对所有$x\in E$成立，不成立点的集合为零测集.

5. 依测度收敛

$\forall\varepsilon>0$，$\lim\limits_{n\to\infty} m(x:|f_n(x)-f(x)|\geq\varepsilon)=0$
依测度收敛可能无处收敛，处处收敛也可能不依测度收敛.

• 但若 $m(E)<\infty$（例如在概率空间的意义下）且 $\{f_n\}$ 几乎处处有限，若 $f_n\xrightarrow{a.e.}f(x)$，则 $\{f_n\}$ 依测度收敛到 f.

• 同上，若 $m(E)<\infty$，则 f_n 依测度收敛到 f 充要条件是 $\{f_n\}$ 的任何子列 $\{f_{n_k}\}$，都存在处处收敛的二级子列 $\{f_{n_{k_j}}\}$.

• Riesz 定理：若 f_n 依测度收敛到 f，则存在几乎处处收敛的子列 $\{f_{n_k}\}$.

• Ерóров 定理：设f、$\{f_n\}$分别是定义在可测集 E 上的可测函数和可测函数列且几乎处处有限，并且$m(E)<\infty$，如果 $f_n\to f(x),n\to\infty\ a.e.x\in E$，则 $\{f_n\}$ 近一致收敛于 f.

• Лузин 定理和多维推广：设 $f:[a,b]\to\mathbb{C}$上的 Lebesgue 可测函数（记μ是$\mathbb{R}^n$上的正则$Borel$测度，$f:\mathbb{R}^n\to\mathbb{R}^m$上的$\mu$可测函数，$X$是$\mathbb{R}^n$中$\mu$可测集且$\mu(X)\mathrm{K}<\infty$），则$\forall\varepsilon>0$，存在紧集$E$（存在$X$中的紧急$K$），$s.t.m([a,b]\backslash E)<\varepsilon(\mu(X\backslash K)<\varepsilon)$，且 f 在其上连续.

• Лузин 定理推论：任意可测函数存在一列几乎处处收敛于其的连续函数.

6. 依 Lp 意义收敛 (p 次幂平均收敛)

$$\lim_{n\to\infty}\int_E|f_n(x)-f(x)|^p\,dx=0$$

7. 函数列的依范数收敛

$$\lim_{n\to\infty}\|f_n(x)-f(x)\|=0$$

8. 弱收敛 $x_n \xrightarrow{W} x$、称 w-$\lim\limits_{n\to\infty} x_n = x$ 为弱极限

X 是赋范线性空间, X 的对偶空间为 X'，对于 X 中的点列 $\{x_n\}$，若 $\exists x \in \mathrm{X}\ s.t.\ \forall f \in \mathrm{X}'$ 都有 $f_n(x)$ 依范数收敛到 $f(x)$，则称 $\{x_n\}$ 弱收敛到 x.

9. 弱*收敛 $f_n \xrightarrow{W^*} f$、称 w*-$\lim\limits_{n\to\infty} f_n(x) = f(x)$ 为弱*极限

对于 X'中的点列 $\{f_n\}$，若 $\exists x \in \mathrm{X}\ s.t.\ \forall f \in \mathrm{X}'$ 都有 $f_n(x)$ 依范数收敛到 $f(x)$，则称 $\{f_n\}$ 弱*收敛到 x.

10. 强收敛

即点态的依测度收敛.

11. 依分布收敛 $(\Leftrightarrow$ 分布函数的弱 $*$ 收敛$)$ $F_n \xrightarrow{W} F$或$X_n \xrightarrow{L} X$

随机变量 $X, X_1, X_2, \ldots$的分布函数分别为 $F(x), F_1(x), F_2(x), \ldots$，若 $F(x)$ 的任意连续点 x，有 $\lim\limits_{n\to\infty} F_n(x) = F(x)$，则称随机变量序列 $\{X_n\}$ 依分布收敛于 X.

12. 以概率 1 收敛 $(\Leftrightarrow$ 概率意义下的几乎处处收敛，应用有强大数率、强相合性$)$ $X_n \xrightarrow{a.s.} X$

$$P\left(\left\{X: \lim_{n\to\infty} X_n = X\right\}\right) = 1$$

13. 依概率收敛 $(\Leftrightarrow$ 依概率测度收敛，应用有弱大数率、弱相合性$)$ $X_n \xrightarrow{P} X$

$\forall \varepsilon > 0$，$\lim\limits_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$

- 以概率 1 收敛(几乎处处收敛)一定依概率收敛.
- 依概率收敛一定依分布收敛.

14. p 阶收敛 $($特别地，$p = 2$时称均方收敛$)$

$\lim\limits_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$

- 均方收敛一定依概率收敛，这可以由 Чебышёв (切比雪夫)不等式证明.

三个重要的收敛定理/极限与积分交换次序定理

1. Beppo Levi 非负渐升列积分定理

设置定义在 E 上非负可测函数渐升列 $\{f_n(x)\}$ 满足 $f_1(x) \leq f_2(x) \leq \ldots \leq f_k(x) \leq \ldots$，且有

$$\lim_{n\to\infty} f_n(x) = f(x), a.e.\ x \in E，则 \lim_{n\to\infty} \int_E f_n(x)\,dx$$

$$= \int_E f(x)\,dx.$$

- 引理：非负可积函数是几乎处处有限的.

·非负可积渐降列积分定理

对于 E 上的非负渐降列 $\{f_n(x)\}$，如在 E 上还是可积的，且有 $\lim\limits_{n\to\infty} f_n(x) =$

$f(x), a.e.\ x \in E$，则 $\lim\limits_{n\to\infty} \int_E f_n(x)\,dx = \int_E f(x)\,dx.$

2. Fatou 引理

若 $\{f_n(x)\}$ 是 E 上的非负可测函数列，则

$$\int_E \varliminf_{n\to\infty} f_n(x)\,dx \le \varliminf_{n\to\infty} \int_E f_n(x)\,dx$$

·当存在 E 上的可积非负函数 $g(x)$，使得对任意的 n 有 $f_n(x) \ge g(x)\ a.e.$，则可以将 Fatou 引理适用范围推广至值域为扩展实数轴的 $\{f_n(x)\}$.

·假设函数列几乎处处收敛到 f 或依测度收敛到 f，命题仍然成立，有

$$\int_E f(x)\,dx \le \varliminf_{n\to\infty} \int_E f_n(x)\,dx$$

·反向的 Fatou 引理

设 $\{f_n(x)\}$ 是 E 上的值域为扩展实数域的可测函数列，若存在 E 上的可积非负函数 $g(x)$，使得对任意的 n 有 $f_n(x) \le g(x)\ a.e.$，则

$$\int_E \varlimsup_{n\to\infty} f_n(x)\,dx \ge \varlimsup_{n\to\infty} \int_E f_n(x)\,dx$$

·备注：上下极限可以等价改写

$$\varliminf_{n\to\infty} f_n(x) = \lim_{n\to\infty} \inf f_n(x),\ \varlimsup_{n\to\infty} f_n(x) = \lim_{n\to\infty} \sup f_n(x)$$

3. Lebesgue 控制收敛定理

设 $f_n \in \mathrm{L}(E),\ n = 1,2,\ldots$(即任意 n，f_n 是 E 上的可积函数)，且

$$\lim_{n\to\infty} f_n(x) = f(x),\quad a.e.\ x \in E$$

若存在 E 上的可积函数 $F(x)$ 使得

$$|f_n(x)| \le F(x),\quad a.e.\ x \in E\ \ (n = 1,2,\ldots)$$

则 $f_n(x)$ 在 E 上依 L^1 意义收敛于 f，于是得到一个更常用的结论

$$\lim_{n\to\infty} \int_E f_n(x)\,dx = \int_E f(x)\,dx$$

·依测度收敛的控制收敛定理

设 $f_n \in \mathrm{L}(\mathbb{R}^k),\ n = 1,2,\ldots$(即任意 n，f_n 是 $\mathbb{R}^k$ 上的可积函数)，且 f_n 在 $\mathbb{R}^k$ 上依测度收敛到 f，若存在 $\mathbb{R}^k$ 上的可积函数 $F(x)$ 使得

$$|f_n(x)| \le F(x),\quad a.e.\ x \in \mathbb{R}^k\ \ (n = 1,2,\ldots)$$

则 $f_n(x)$ 在 $\mathbb{R}^k$ 上依 L^1 意义收敛于 f，同样地，有

$$\lim_{n\to\infty} \int_E f_n(x)\,dx = \int_E f(x)\,dx$$

附录：逻辑斯蒂回归

Logistic 回归总体而言是一种广义线性模型，用以解决单个因变量的分类问题，对应分类变量的问题；区别于对数线性模型，对应关于列联表的变量的问题.

1. Logistic 变换及 Logistic 线性回归模型

Logistic 回归是一种对数几率模型 (Logit model)，为了因变量为离散变量时进行回归而提出，即对分类问题进行回归，例如利用 Logistic 回归，可以通过样本的回归判断任意指定的一个拥有某 BMI 指数的人是否大概率患有心血管病（二分类、二值回归）、仅由某人的一些面部特征数据判断他是成年女士、成年男士还是未成年儿童等等，这些都可以通过 Logistic 回归实现.

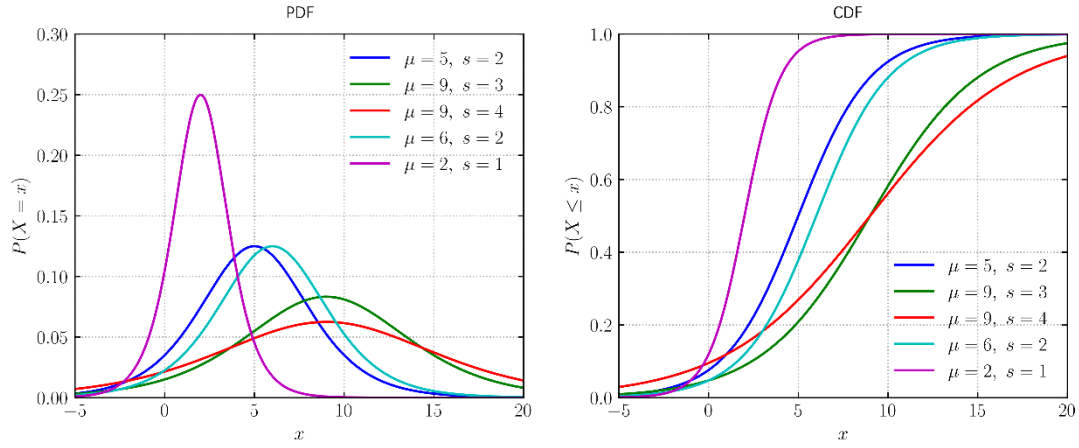
Logistic 分布：一种指数族分布，设 X 为连续型随机变量 CDF 与 PDF 分别为：

$$F(x)=\frac{1}{1+e^{-\frac{x-\mu}{\gamma}}}=\frac{1}{2}+\frac{1}{2}\tanh\left(\frac{x-\mu}{2\gamma}\right),\ x\in\mathbb{R}$$

$$f(x)=\frac{e^{-\frac{x-\mu}{\gamma}}}{\gamma(1+e^{-\frac{x-\mu}{\gamma}})^2}=\frac{1}{4\gamma}\operatorname{sech}^2\left(\frac{x-\mu}{2\gamma}\right),\ x\in\mathbb{R}$$

其中称 μ 为位置参数，$\gamma>0$ 为形状参数；$F(x)$ 是一条 S 形曲线，$f(x)$ 关于 $x=\mu$ 对称；γ 越大越厚尾，而在 $x=\mu$ 附近增长越慢.

特别地，对于二值 Logistic 回归，常常令$(Y=1\mid x)=\frac{e^{\boldsymbol{\beta X}}}{1+e^{\boldsymbol{\beta X}}}$，$P(Y=0\mid x)=\frac{1}{1+e^{\boldsymbol{\beta X}}}$即服从 Logistic 分布.



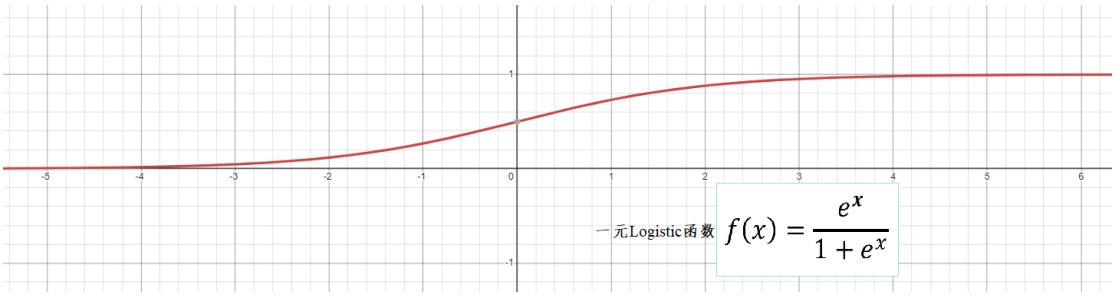
由于理想的分类函数是符号函数 sign( ) 并不可微，(对二值 Logistic 回归)提出对 Y 进行 Logistic 变换，等价于“优势比的对数等于 $\boldsymbol{\beta X}$”. Logistic 变换是针对因变量 Y 为离散变量（属性数据）且回归函数为 Logistic 函数的变换，通过这种方式使原函数线性化、成为关于 Z 的线性函数；假设 X 为自变量而 Y 为因变量，记 $\boldsymbol{\beta}$ 为系数向量、$\boldsymbol{X}$ 为设计矩阵，其中 $\boldsymbol{\beta}$ 常取极大似然估计 MLE，定义 Logistic 函数：

$$P(Y\mid X)=\frac{e^{\boldsymbol{\beta X}}}{1+e^{\boldsymbol{\beta X}}}\ (\text{Logistic 函数})$$

Logistic 函数是一种定义域为 $\mathbb{R}$、值域为 $(-1,1)$ 的非线性函数，但通过 Logistic 变换可以线性化：

$$Z=\log\left(\frac{Y}{1-Y}\right)$$

这就是 Logistic 变换.



由于概率 $y=P(Y=1\mid X)$ 的值取值在 0 到 1 之间，如若把 y 假设为多项式函数等取值在 0 到 1 之间的函数是不合适的，需要做一个映射处理，这个目的常用一些变换得到，例如：

① Logistic 变换：$f(y)=\log\frac{y}{1-y}$，特别地在 Logistic 线性模型中 $f(y)=\boldsymbol{\beta X}$

② Probit 变换：$f(y)=\Phi^{-1}(y)$

③ 双对数变换：$f(y)=\log(-\log(1-y))$

其中就包括 Logistic 变换，实质是 Logistic 函数的逆. 这样变换以后再假设 $f(y)=\boldsymbol{\beta X}$ 服从某回归模型，譬如在线性回归模型中 $f(y)=\boldsymbol{\beta X}$，从而拟合 p. 容易看出，相合性检验比独立性检验更深入，进一步地 Logistic 回归比相合性检验更深入，他直接给出了一个可能的关系式.

二值 Logistic 回归模型：假设响应变量即因变量 Y 仅有两个状态，我们分别用 0 和 1 表示，现研究 $y=P(Y=1)$，若一共有 k 个因素 $x_1,x_2,\ldots,x_k$ 影响Y 的取值，则称

$$\log\frac{y}{1-y}=g(x_1,x_2,\ldots,x_k)$$

为二值 Logistic 回归模型，简称 Logistic 回归模型.

当 $g(x_1,x_2,\ldots,x_k)$ 为一个线性函数时称为 Logistic 线性回归模型，即：

$$\log\frac{y}{1-y}=\boldsymbol{\beta X}=\beta_0+\sum_{i=1}^{k}\beta_i x_i$$

容易知道 y 服从 Logistic 分布，即 $y=P(Y=1|X)=\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}$，$1-y=P(Y=0|X)=\frac{1}{1+\exp(\boldsymbol{\beta X})}$，这样设置可以让 $P(Y=1|X)$ 的回归系数使得 $\boldsymbol{\beta X}=0$；$\boldsymbol{\beta}$ 取其 MLE.

当协变量向量一共有 t 种而第 i 种有 n_i 个，其中有 r_i 个响应变量值为 1 而有 n_i-r_i 个取值为 0，参数 $\boldsymbol{\beta}$ 的似然函数为

$$\prod_{i=1}^{t}\left(P(Y=1|X)\right)^{r_i}\left(P(Y=0|X)\right)^{n_i-r_i}=\prod_{i=1}^{t}\left(\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n}}\right)^{r_i}\left(\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n}}\right)^{n_i-r_i}$$

若将上式记为 $\sup L$，则有

$$-2\log\Lambda=-2\log\frac{\left(\frac{n_{Y=1}}{n}\right)^{n_{Y=1}}\left(\frac{n_{Y=0}}{n}\right)^{n_{Y=0}}}{\sup L}\overset{L}{\sim}\chi^2(1)$$

MLE 准则下 Logistic 回归损失函数：交叉熵 记 $P(Y=1\mid X)=p(x)$

$$\textbf{交叉熵}=\textbf{KL散度}+\textbf{信息熵}:\ -\sum_{i=1}^{n}p(x_i)\log\big(1-p(x_i)\big)=D_{KL}(p||1-p)-\sum_{i=1}^{n}p(x_i)\log p(x_i)$$

∗ 为什么 Logistic 回归中 $\boldsymbol{\beta}$ 取其 MLE 而非于线性回归中更常用的、一定条件下具有优良性质的普通最小二乘估计 OLS？换句话说，在均方误差 MSE 意义下得到的最佳估计便是 OLS，那为什么损失函数不再选用 MSE（而考虑交叉熵），进而用 MLE 代替 OLS？

答：

① 本质原因是分类问题中属性的分布是**多项分布**（二值 Logistic 回归中是二项分布），并没有做残差正态分布的假设；在普通线性回归中我们有正态残差假设.

注意：此处值得指出的是 OLS 是非参数方法，MLE 是参数统计方法，这正好适用于 Logistic 回归问题；最小二乘方法是一个凸优化的问题，MSE 综合了无偏性与 SSE 的值，MSE $=n\cdot$ SSE，然而这个问题下 MSE 准则得到的损失函数不是凸优化问题，极大似然的方法不一定是凸优化问题，不过在此处成立——凸性拥有极好的优化性质；OLS 可以视作使得残差在 L^2 范数意义下最小的最优解，即使得 SSE 最小，MLE 是使得似然函数最大的最优解.

正态残差假设是普通线性回归中 OLS 具有很好的一些性质所必要的前提（参考 Gauss–Markov theorem 的条件），这时 OLS 是一致最小方差无偏估计，是最佳线无偏估计；甚至在一些 MLE 如残差是有偏的，因此 Gauss–Markov 假设、正态残差假设成立情况下，在线性回归中一般考虑 OLS 而非 MLE.

但对于分类问题，在属性服从多项分布的假定下 OLS 并没有 Gauss–Markov 假设下普通线性回归的那些优良性质，而参数的 MLE 具有此时具有无偏性、渐进正态性等性质，所以考虑 MLE，而 MLE 准则下的损失函数就是交叉熵，当然这只是考虑 MLE 一方面的原因. 也可以说，他们的不同可以讲是源自提出的假设不同，这是问题实际情况决定的.

② 最小二乘损失函数，或者讲 MSE 在这种情况下是非凸的，Hessian 矩阵非正定，难以数值迭代最优化，算法很有可能收敛到某局部最优解；但似然函数是可导凸函数，拥有唯一的最优解，局部最优解等价于全局最优解，总能收敛到最优点.

③ MSE 损失函数在靠近 0 和 1 时存在梯度消失的现象，同时似然损失函数的梯度只与参数有关，与 Logistic 函数的梯度无关：一元 Logistic 函数导数最大值点在四分之一处，可能会导致其他损失函数参数更近变慢.

④ MSE 损失函数的学习比交叉熵慢很多，对错误分类惩罚不够重，当损失函数为交叉熵（负对数似然）时样本分类错误越严重惩罚会越严重，呈，也会在导数最大点附近急剧减小，而最小二乘损失函数相对而言变化平缓，并不“陡峭”，响应不够敏感.

$proof$：

① $w=\boldsymbol{\beta X}$，二项分布假设，记 $P(Y=1\mid X)=p(x)$，$P(Y=0\mid X)=1-p(x)$

$$l(w)=\log\big(L(w)\big)=\log\left(\prod_{i=1}^{n}\big(p(x_i)\big)^{y_i}\big(1-p(x_i)\big)^{1-y_i}\right)$$

$$=\sum_{i=1}^{n}\left(y_i\log\left(\frac{p(x_i)}{1-p(x_i)}\right)+\log\big(1-p(x_i)\big)\right)$$

$$= \sum_{i=1}^{n} [y_i w x_i - \log(1 + e^{w x_i})] = -n\,J(w) \quad \Leftrightarrow \quad \text{交叉熵损失函数梯度为 0}$$

② 利用下文③的结果，

$$\frac{\partial^2 MSE}{\partial w^2} = \sum_{i=1}^{n} \hat{y}_i(1-\hat{y}_i)x_i^2\left(-y_i + 2(1+y_i)\hat{y}_i - 3\hat{y}_i^2\right)$$

由于 y_i 只取 0 或 1，可以导出 MSE 二阶导数不一定总大于 0.

③ 假设损失函数为 MSE

$$\frac{\partial MSE}{\partial w} = \frac{\partial \sum(\hat{y}_i - y_i)^2}{\partial w} = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\frac{\partial(\hat{y}_i - y_i)}{\partial w} = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\frac{\partial\left(\frac{1}{1+e^{-wx_i}}\right)}{\partial w}$$

$$= \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{1}{1+e^{-wx_i}}\right)^2 e^{-wx_i}x_i = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)[\hat{y}_i(1-\hat{y}_i)x_i]$$

可以看出在 $\hat{y}_i$ 靠近 0 或 1 时梯度趋近于 0，不能有效迭代.

Logistic 回归模型的解释

$$\log\frac{y}{1-y} = \boldsymbol{\beta X} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

我们认为，若 x_i 增加 1，则优势比 $\frac{y}{1-y} = \frac{P(Y=1\mid X)}{P(Y=0\mid X)}$ 增长至原来的 e^{β_i} 倍.

对于自变量都是定量数据的问题， $\log\frac{y}{1-y}$ 一般用 $\log\frac{\hat{y}}{1-\hat{y}} = \log\frac{r_k}{n_i - r_k}$ 进行估计，其中 r_k 是 k 组观测数据的组合中取 1 的自变量的个数、n_k 为观测值个数，但当 $r_k = 0\ or\ 1$ 时情况有些麻烦，通常在 $r_k = 0$ 时用 $\log\frac{0.5}{n_k+0.5}$ 代替 $\log\frac{r_k}{n-r_k}$，在 $r_k = n_i$ 时用 $\log\frac{n_k+0.5}{0.5}$ 代替 $\log\frac{r_k}{n-r_k}$，在 $0 < r_k < n_k$ 时用 $\log\frac{r_k+0.5}{n_i-r_k+0.5}$ 代替 $\log\frac{r_k}{n-r_k}$.

$y = P(Y=1\mid X) \approx 0$ 或 $1\left(y \approx 0\right.$ 即 $\left.1 - y = P(Y=0\mid X) \approx 1\right)$时我们认为自变量对因变量 Y 影响很小而且非常有可能实际的因变量取值就是 0 或 1，而当 $y = P(Y=1\mid X) = 0.5$ 时认为自变量对因变量取值的影响很大，毕竟直观理解起来是因变量取两个可能值的概率都是 0.5.

得到模型后，做预测时再将 $y = P(Y=1\mid X)$ 作映射，根据实际问题情况规定大于某值时取因变量为 1，反之取为 0，比方可以取这个分界点为 0.5，这样便得到了一个分类.

2. **含有名义数据的二分类 Logistic 线性回归模型**

例如要研究年龄、血型与死亡率之间的关系，年龄为定量数据而血型为名义数据. 通常血型大致可以分为四种：A 型、B 型、AB 型和 O 型，不能在 Logistic 回归中简单令离散随机变量 Q，让 A 型$\Leftrightarrow Q = 1$，B 型$\Leftrightarrow Q = 2$，… 因为这样使四种血型间有了顺序关系，事实上他们只作为名义数据，相互之间并没有大小之分，更没有“3 倍的 A 型血型等于 AB 型血型”的荒谬说法；应设三个随机变量 Q_1、Q_2、Q_3，选取一个名义数据作为基线，例如O型血：A型$\Leftrightarrow Q_1 = 1 \Leftrightarrow Q = (1,0,0)$，B型$\Leftrightarrow Q_2 = 2 \Leftrightarrow Q = (0,1,0)$，AB 型$\Leftrightarrow Q_3 = 1 \Leftrightarrow Q = (0,0,1)$，O 型$\Leftrightarrow Q = (-1,-1,-1)$，Logistic 回归模型为：

$$\log\frac{y}{1-y} = \mu + \beta_1 \cdot age + \gamma_1 Q_1 + \gamma_2 Q_2 + \gamma_3 Q_3$$

此外，由上式还可以分别得到四种血型情况下的年龄与死亡率的关系的回归方程.

因此对于含有名义数据的二分类 Logistic 线性回归模型，设有 n 维定量数据、j 维名义数据，$\boldsymbol{\beta}$ 表示定量数据的系数向量、$\boldsymbol{\gamma}$ 表示名义数据的系数向量，$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$ 的估计取其 MLE，**每种定性数据只需要引进一个变量，而 j 维名义数据需要引进 $j-1$ 个变量**，这种办法叫基线法，第 j 个变量既可以取 $j-1$ 个变量都取零也可以都取负一，根据问题要求而定，有

Logistic 线性回归模型：

$$\log\frac{y}{1-y}=\mu+\beta_1x_1+\cdots+\beta_nx_n+\gamma_1\lambda_1+\ldots+\gamma_j\lambda_{j-1}$$

Logistic 线性回归方程：

$$\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\hat{\beta}_1x_1+\cdots+\hat{\beta}_nx_n+\hat{\gamma}_1\lambda_1+\ldots+\hat{\gamma}_j\lambda_{j-1}$$

配合四格表独立性检验，当不独立性显著时进行 Logistic 回归更有说服力.

3. **含有有序数据的二分类** Logistic **线性回归模型**

例如文化程度，小学以下、小学、初中、高中、大学及以上分别可以用 0、1、2、3、4 表示，这是一组有序数据，可以认为他们之间是有大小的，**每种有序数据也仅引进一个变量**.

方法一致.

Logistic 线性回归模型：

$$\log\frac{y}{1-y}=\mu+\boldsymbol{\beta}A+\boldsymbol{\gamma}B+\boldsymbol{\nu}C$$

Logistic 线性回归方程：

$$\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\widehat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$$

C 即有序数据.

4. Logistic **判别分析**

对于 Logistic 回归方程，假设 $\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\widehat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$，令 $u(A,B,C)=\hat{\mu}+\widehat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$，当 u 较大认为 $Y=1$ 概率较大，若取分界点为 0.5，则当 $u>0$ 时可以简单认为 $Y=1$ 会发生；具体的判别方式与分界点的选取视实际问题的情况与需求而定.

5. **多项** Logistic **回归**

多项 Logistic 回归中，假设因变量有多个可能的取值 $(0,1,\ldots,N)$，基线法要求选取一个（常常是最后一个）使得其 $w=\boldsymbol{\beta X}$ 值为 0，其中 $\boldsymbol{X}$ 为设计矩阵. 一般模型：

$$\log\frac{P(Y=0\mid X)}{P(Y=N\mid X)}=\mu^{(0)}+\beta_1^{(0)}x_1+\cdots+\beta_n^{(0)}x_n$$

$$\log\frac{P(Y=1\mid X)}{P(Y=N\mid X)}=\mu^{(1)}+\beta_1^{(1)}x_1+\cdots+\beta_n^{(1)}x_n$$

…

$$\log\frac{P(Y=k\mid X)}{P(Y=N\mid X)}=\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n$$

$$0=\log\frac{P(Y=N\mid X)}{P(Y=N\mid X)}=\mu^{(N)}+\beta_1^{(N)}x_1+\cdots+\beta_n^{(N)}x_n$$

有约束$\sum_{i=1}^{N}P(Y=i\mid X)=1$

$\boldsymbol{\beta}^{(k)}$ 用极大似然估计 MLE 代替，

$$P(Y=k\mid X)=\frac{\exp\left(\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n\right)}{\sum_{i=1}^{N}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}=\frac{\exp\left(\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n\right)}{1+\sum_{i=1}^{N-1}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}$$

特别地$P(Y=N\mid X)=\frac{1}{\sum_{i=1}^{N}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}=\frac{1}{1+\sum_{i=1}^{N-1}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}$

6. **如何利用计算机做** Logistic **回归?**

R 语言：线性回归

有必要先介绍一下线性回归如何操作，这就要先了解 lm() 函数.

lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)

重要参数的选择：

· formula：模型的关系式，如 formula = Z~X+Y 表示拟合 $Z=X+Y+intercept$ 的线性模型，也可以写作 formula = Z~X+Y+1，注意此时模型中有截距项；而 formula = Z~X+Y-1 表示拟合 $Z=X+Y$ 的线性模型，此时不含截距项，事实上 formula 遵循 R 表达式的语法：

“~”为变量类型的分隔，左边是响应变量，右边是解释变量；

“+”用以隔开两个解释变量；

“:”即连接两个变量表示他们的交互项，比如 formula = Z~X+Y+X:Y；

“*”表示两个(或多个)变量自己与他们之间所有可能的交互项，例如 formula = Z~A*B*C，等价于 formula = Z~A+B+C+A:B+ A:C+ B:C+ A:B:C；

“^”表示交互项次数，例如 formula = Z~(A+B+C)^2，等价于 formula = Z~ Z~A+B+C+A:B+ A:C+ B:C；

“.”表示除因变量外所有变量，可以极大简化写法，比如 formula = Z~.；

“-”移除某变量，例如 formula = Z~(A+B+C)^2- A:C，特别地“-1”表示删除截距项；

“I()”将某变量进行算术平方，比如有 formula = Z~A+I((B+C)^2)，此时等价于 formula = Z~A+D，D 是一个新变量，值为 B 与 C 的和的平方；

“$function$”可以在模型中使用一些函数，例如 formula = log(Z)~X+Y.

· data：数据集，要求是数据框类型.

· weights：应该被赋值“NULL”或一个数值向量，当赋值为“NULL”使用普通最小二乘法 OLS，当赋值为数值向量则把他作为权重使用加权最小二乘估计 WLS（即最小化 sum(weights * e^2)）.

· na.action：当数据包含“NA”时应该如何处理.

a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.

· method：目前只有 method = qr 可用，也可以选择 method = "model.frame"，会返回模型数据库且不进行

拟合，和 model = TRUE 效果相同.

- data：数据集，要求是数据框类型.

- data：数据集，要求是数据框类型.

- 剩下的一些参数不是很重要，比如 subset 可以选择一个用来观察的子集.

lm() 除了进行线性回归还可以做单因素方差分析、协方差分析，不过后者有封装好的函数 aov()，这个函数实质也是在调用 lm().

lm() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm

aov() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov

---

关于模型的函数：

- summary(fit)：提供一个详细的结果，包括多种参数与指标，含部分以下函数列出的参数.

```
Call:
glm(formula = CAD ~ college * TC * SBP, family = binomial(link = "logit"),
    data = dataSet)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.73574  -0.48116  -0.31985   0.00008   2.81342

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.938e+00  5.012e-01  -7.858  3.9e-15 ***
college         2.350e+01  1.484e+03   0.016   0.9874
TC              4.956e-01  2.049e-01   2.419   0.0156 *
SBP             4.187e-01  2.028e-01   2.064   0.0390 *
college:TC     -4.956e-01  9.577e+02  -0.001   0.9996
college:SBP    -4.187e-01  9.160e+02   0.000   0.9996
TC:SBP          3.012e-03  8.425e-02   0.036   0.9715
college:TC:SBP -3.012e-03  7.834e+02   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3034.5  on 2396  degrees of freedom
Residual deviance: 1112.0  on 2389  degrees of freedom
AIC: 1128

Number of Fisher Scoring iterations: 18
```

- coefficients(fit)：列出拟合的参数，包括截距(如果有).

- cofint(fit)：列出拟合参数的置信区间.

- fitted(fit)：列出预测值.

- residuals(fit)：列出残差.

- vcov(fit)：列出拟合参数的协方差阵.

- AIC(fit)、BIC(fit)：列出 AIC、BIC 统计量的值.

- plot(fit)：绘制一系列用以评价模型的回归诊断图，包含 Q-Q 图.

- predict(fit, data)：预测.

R 语言线性回归的例子：

```r
dataSet <- read.csv("data.csv")

fit <- lm(Effect ~ Gender + Age * Method, data = dataSet)
summary(fit)
```

R 语言：Logistic 回归

前文提到了线性回归函数 lm()，这里我们需要广义线性回归函数 lm() 函数.

glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(…), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, …)

glm.fit(x, y, weights = rep(1, nobs), start = NULL, etastart = NULL, mustart = NULL, offset = rep(0, nobs), family = gaussian(), control = list(), intercept = TRUE, singular.ok = TRUE)

重要参数的选择：

・family：选择相应变量分布的假设与连接函数，比如 family = binomal(link = 'logit') 表示响应变量分布假设为二项分布，用 Logistic 函数作为连接函数，这时做的便是逻辑回归；family = binomal(link = 'probit') 表示响应变量分布假设为二项分布，用 Probit 函数作为连接函数；family = poisson(link = 'identity') 表示响应变量分布假设为泊松分布，用 $f(x) = x$ 作为连接函数.

・control：控制误差与最大迭代次数，例如 control = list(epsilon=1e-8, maxit=25) 中 epsilon 为终止准则的误差，maxit 为最大迭代次数.

glm() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

---

R 语言 Logistic 回归的例子：

```
dataSet <- read.csv("data.csv")

fit <- glm(Effect ~ Gender + Age + Method, family = binomial(link = "logit"), data = dataSet,
control = list(maxit = 200))
summary(fit)
```

Python：Logistic 回归（sklearn 库）

```
1.  import pandas as pd
2.  import sklearn as sl
3.
4.
5.  data = pd.read_csv('data.csv')
6.
7.  X = data.loc[:, ['Gender']]
8.  Y = data.loc[:, ['Effect']]
9.
10.
11.  model = sl.linear_model.LogisticRegression()  # Logistic Regression (逻辑回归)
12.  model.fit(X, Y)
13.  pre = model.predict(X)
14.  print("Line regression predict result: ", pre)
15.
16.  epsilon = pd.sqrt(sl.metrics.mean_squared_error(Y, pre))
17.  print("Line regression mean squared error: ", epsilon)
```

7. 参数的选取：信息准则

AIC

AIC 建立在信息论上，又称赤池信息量准则，是 KL 散度的估计量；一般假设模型误差服从相互独立的正态分布，记 k 为参数的数量、RSS 为残差平方和、C 是依赖于数据的常量，

$$AIC = 2k - 2\log\hat{L} = 2k + n\log RSS - (n\log n + 2C)$$

由于 $(n\log n + 2C)$ 不影响相同一批数据不同模型 AIC 值的差别，所以可以只考虑 AIC 的一部分用以对比

$$AIC \doteq 2k + n\log\left(\frac{RSS}{n}\right) \text{或 } 2k + n\log RSS$$

倾向于选择 AIC 较小的模型，当 k 增大通常能“更多地拟合”，这会让似然函数最大值 $\hat{L}$ 的值也增大而使得 AIC 的值减小，但 k 太大的时候会大大影响 AIC 的值使之也变得庞大，与其同时出现的问题是过拟合，这是应该避免的情况，因此倾向于选择 AIC 小的模型.

当样本量较小的时候一般更正 AIC 为 AICc，即更正后的赤池信息量准则，而 AICc 在样本量增加的时候又会收敛到 AIC，可以证明在任何大小的样本量下都可以使用 AICc，同时还有另一种指标 AICu.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \doteq \log\left(\frac{RSS}{n}\right) + \frac{n+k}{n-k-2}$$

$$AICu = \log\left(\frac{RSS}{n-k}\right) + \frac{n+k}{n-k-2}$$

BIC

BIC 又称贝叶斯信息量准则，BIC 对过量参数的惩罚比 AIC 更重，AIC 的惩罚是 $2k$ 而 BIC 的惩罚是 $k\log n$，因此应用 BIC 更容易选出一个参数更少的模型.

$$BIC = k\log n + n\log\left(\frac{RSS}{n}\right)$$

其他信息准则

我们还有 FIC、HQC 等准则与多种散度用来解决模型选择问题.

AIC: https://en.wikipedia.org/wiki/Akaike_information_criterion

BIC: https://en.wikipedia.org/wiki/Bayesian_information_criterion

FIC: https://en.wikipedia.org/wiki/Focused_information_criterion

HQC: https://en.wikipedia.org/wiki/Hannan%E2%80%93Quinn_information_criterion

KL 散度: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

JS 散度: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

WAIC: https://en.wikipedia.org/wiki/Watanabe%E2%80%93Akaike_information_criterion

8. 统计检验

一般用 F 检验、Wald 检验、似然比检验和拉格朗日乘子检验来假设检验回归系数与 0 是否存在显著差异（实质是对若干约束假设进行的检验），其中 Wald 检验在某些情况下会给出错误的结论，如标准误 SE 较大时（比方数据比较极端，$P(Y=1\mid X)$ 在某点激增）.

这些内容过于复杂了，暂不打算深入讨论，仅简单说明.

F 检验：原假设为 k 个回归系数都为 0，备择假设为回归系数不全为 0，则

$$\frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} \overset{L}{\sim} F(k, n-k-1)$$

F 检验针对线性约束且需要满足随机扰动项满足误差正态分布，对误差分布没有假设时 F 检验失效，

这时 Wald 检验会是个可能的不错选择. Wald 统计量是一个标准化后的二次型，Wald 检验是一致参数统计方法，直接检验了参数的 MLE 与原假设的差异，这与似然比检验、得分检验底层逻辑上不同；更一般的 Wald 检验不仅能用来检验若干回归系数是否显著为 0，设原假设为 $\left(\beta_{i_1},\beta_{i_2},\ldots,\beta_{i_k}\right)=\mathbf{0}$，即约束 $R\beta-q=0$ 成立，$W=\left(R\hat{\beta}-q\right)^T(R\hat{\sigma}^2(X^TX)R^T)^{-1}\left(R\hat{\beta}-q\right)\overset{L}{\sim}\chi^2(k)$，$k$ 为约束个数，$\hat{\sigma}^2$ 为残差方差的估计；特别地，对于检验某一个回归系数是否显著为 0，Wald 检验的渐进分布为

$$\frac{\hat{\beta}_j}{SE_{\hat{\beta}_j}}\overset{L}{\sim}N(0,1)$$

剩下的许多检验，不再赘述.

附录：方差分析

***

附录：Bayesian empirical likelihood for ridge and lasso regressions

Bayesian
empirical likelihoc