多重共线性 Multicollinearity

假设 Designed Matrix X "不再" 满秩：

$$rank(X) \approx p \qquad X \in \mathbb{R}^{n\times p}$$

Assumption：① 假设 Model 均满足 Gauss-Markov 条件

② 假设已中心化，即

$\{y, x_1, \cdots, x_p\}$ 中的向量的分量均值为 0，方差为 1

$$\frac{1}{n}\sum_{j=1}^{n} y_j = 0 \qquad \frac{1}{n}\sum_{j=1}^{n}(y_j - 0)^2 = 1$$

$$\frac{1}{n}\sum_{j=1}^{n} x_{ij} = 0 \qquad \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - 0)^2 = 1$$

③ β 不含截距项 $\beta \in \mathbb{R}^p$

一、定义 Definition

1. 完全多重共线性

若存在非 0 向量 $v \in \mathbb{R}^p$，使设计矩阵 X 满足 $Xv = 0$，即 X 列向量线性相关，称为 $(x_1, \cdots, x_p)$ 完全多重共线性 Perfect Multicollinearity

2. 多重共线性

若存在非 0 向量 $v \in \mathbb{R}^p$，使设计矩阵 X 满足 $Xv \approx 0$，即 X 列向量线性相关，称为 $(x_1, \cdots, x_p)$ 多重共线性 Multicollinearity

二、影响 Effect

1. 完全多重共线性

当自变量 $\{x_{(1)}, x_{(2)}, \cdots, x_{(p)}\}$ 存在完全多重共线性时，设计矩阵 X 不满秩，$rank(X) < p$ 则 $rank(X^TX) = rank(X) < p$，即有 X^TX 不可逆，之后的 OLSE 均不可进行

2. 多重共线性

当自变量 $\{x_{(1)}, x_{(2)}, \cdots, x_{(p)}\}$ 存在多重共线性时，质量差，因为方差 $var(v^T\hat{\beta})$ 的上界趋向 $+\infty$ （OLSE：无偏）

$\exists\ \lambda_{min}\{(X^TX)\} \longmapsto 0$

Proof ①

$$cov(\hat{\beta}) = cov((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^T \cdot cov(y) \cdot X(X^TX)^{-1} = \sigma^2 \cdot (X^TX)^{-1}$$

$$\begin{aligned}
var(v^T\hat{\beta}) &= cov(v^T\hat{\beta}, v^T\hat{\beta}) \\
&= v^T cov(\hat{\beta})\, v \\
&= v^T \cdot \sigma^2 (X^TX)^{-1} \cdot v \\
&\leq \sigma^2 \cdot \|v\|^2 \cdot \lambda_{max}\{(X^TX)^{-1}\} \\
&= \sigma^2 \cdot \|v\|^2 \cdot \left[\lambda_{min}\{(X^TX)\}\right]^{-1}
\end{aligned}$$

X^TX 存在接近 0 的特征根，故 $\dfrac{1}{\lambda_{min}\{(X^TX)\}} \longmapsto \infty$

Proof ②

$\hat{\beta}$ 各分量方差加和 $tr\{\sigma^2(X^TX)^{-1}\} = \sigma^2 \cdot \sum_{j=1}^{p} \lambda_j^{-1} \longrightarrow \infty$

三、诊断 Diagnoses

1. 方差扩大因子法 Variance inflation factor (VIF)

因为已经标准化，故 $corr(\hat{\beta}) = \sigma^2 \cdot (X^TX)^{-1} = (X^TX)^{-1}$，记

$$C = (X^TX)^{-1}$$

称 C 的第 j 个对角元素为：

$$VIF_j = C_{[j,j]}$$

称 VIF_j 为自变量 $x_{(j)}$ 的方差扩大因子，由于 $cov(\hat{\beta}) = \sigma^2 C$，故

$$var(\hat{\beta}_{[j]}) = \sigma^2 \cdot VIF_j$$

衡量标准：

① VIF_j 衡量 $\hat{\beta}_j$ 方差因多重共线性而扩大的程度

② 经验判断： $VIF_j \geq 10$，认为 $x_{(j)}$ 与其他有明显多重共线性

③ 检验整体： $\frac{1}{p}\sum_{j=1}^{p} VIF_j >> 1$，认为整体有明显多重共线性

2. 特征根判定法： Eigenvalue - based

$\exists$ 多重共线性 $\Longleftrightarrow$ $\lambda_{min}\{(X^TX)\}$ 接近 0

经验判断：

考虑 X^TX 的特征根 λ：

① $0 < \frac{\lambda_{max}}{\lambda_{min}} < 100$ 不存在多重共线性

② $100 < \frac{\lambda_{max}}{\lambda_{min}} < 10000$ 存在较强多重共线性

③ $\frac{\lambda_{max}}{\lambda_{min}} > 10000$ 存在严重多重共线性

3. 直观判断 Intuitive judgement

- 增加或剔除某个自变量，或者改变某个观测值时，回归系数的估计发生较大变化;
- 定性分析认定的一些重要自变量在拟合结果中没有通过显著性检验;
- 某些自变量回归系数估计值的正负号与定性分析结果不符;
- 自变量间的相关系数较大;
- 某些重要自变量回归系数估计值的标准误差较大。