第 6 章 多重共线性的情形及其处理

李 杰

数据科学学院，浙江财经大学

2019 年 12 月 4 日

内容提要

- 多重共线性产生的背景和原因
- 多重共线性对回归模型的影响
- 多重共线性的诊断
- 消除多重共线性的方法
- 本章小结与评注

6.1 多重共线性产生的背景和原因

☞ 共线性

- **多元线性回归模型的重要假设:** $\text{rank}(\mathbf{X}) = p+1$, 即如果 $\mathbf{X}$ 按列分块, 得到列向量组 $(x_0, x_1, \cdots, x_p)$, 则该向量组线性无关.
- **完全共线性:** 如果存在一组不全为 0 的实数 $(c_0, c_1, \cdots, c_p)$ 使得

$$c_0x_0 + c_1x_1 + \cdots + c_px_p = \mathbf{0}$$

  则称自变量 $x_1, \cdots, x_p$ 之间存在**完全共线性** (perfect collinearity).
- **多重共线性:** 如果存在一组不全为 0 的实数 $(c_0, c_1, \cdots, c_p)$ 使得

$$c_0x_0 + c_1x_1 + \cdots + c_px_p \approx \mathbf{0}$$

  则称自变量 $x_1, \cdots, x_p$ 之间存在**多重共线性** (multi-collinearity).

☞ 共线性产生的背景和原因:

- 解释变量之间的多重共线性几乎不可避免, 只是严重程度的问题.
- 所研究的问题涉及时间序列数据时, 由于很多经济变量存在共同的变化趋势, 这些变量间容易出现严重的多重共线性.
- 在研究经济、社会问题时, 由于问题的复杂性, 涉及的因素往往很多, 在建立模型时, 由于研究者认识水平的局限性, 很难在众多因素中找到一组互不相关同时对因变量 y 有显著影响的变量.

6.2 多重共线性对回归模型的影响

① 如果数据存在完全共线性，即 $\text{rank}(\mathbf{X}) < p+1$，此时 $(\mathbf{X}'\mathbf{X})^{-1}$ 不存在，此时 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 不存在，而正规方程组 $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ 的解有无穷多个.

② 当数据存在较严重的多重共线性时，估计量是相合的、无偏的，但估计量的方差很大，估计值不稳定，数据的微小变化将导致估计量的巨大波动，甚至导致估计量符号的变化.

【说明】 假设向量组 $x_1, \cdots, x_p$ 存在较严重的多重共线性，假设变量 x_j 可近似表示成变量 $x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p$ 的线性组合. 令 $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ 表示变量 x_j 的总波动，R_j^2 表示辅助回归

$$x_j = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{j-1}x_{j-1} + \alpha_{j+1}x_{j+1} + \cdots + \alpha_p x_p + \mu$$

的决定系数，则对于回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{j-1}x_{j-1} + \beta_j x_j + \beta_{j+1}x_{j+1} + \cdots + \beta_p x_p + \varepsilon$$

估计量 $\hat{\beta}_j$ 的方差 $\text{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1 - R_j^2)}$. 由此可见，$x_j$ 与 $x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_p$ 的共线性程度越高，R_j^2 越大，$\hat{\beta}_j$ 的方差也就越大.

☞ **方差扩大因子法** (variance inflation factor, VIF)

- 因 $\mathrm{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{SST_j(1-R_j^2)}$, 定义 x_j 的**方差扩大因子法**为 $c_{jj} = \frac{1}{1-R_j^2} \triangleq VIF_j$.
- $VIF_j \geqslant 1$.
- $VIF_j \geqslant 10$ 则认为存在较严重的多重共线性.
- 若 $\bar{VIF} = \dfrac{1}{p}\sum_{i=1}^{p} VIF_j >> 1$, 也可认为存在较严重的多重共线性.
- **R 代码**

```
library(car)

states <- as.data.frame(state.x77[, c("Murder", "Population",
                              "Illiteracy", "Income", "Frost")])
murder_fit <- lm(Murder ~ Population + Illiteracy + Income +
                                       Frost, data = states)
vif(murder_fit)
```

☞ 特征根判别法

- **特征根分析:** 如果 $|\mathbf{X}'\mathbf{X}| \approx 0$ ($\mathbf{X}$ 是标准化处理后的资料矩阵), 则 $\mathbf{X}'\mathbf{X}$ 至少有一个特征根接近 0. 反之, $\mathbf{X}'\mathbf{X}$ 有多少个特征根接近 0, 这 $\mathbf{X}$ 中就存在多少个多重共线性关系.
- **条件数**
  - 记 $\mathbf{X}'\mathbf{X}$ 的特征值为 $\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_p$, $\lambda_{\max} = \max\{\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_p\}$, $\lambda_{\min} = \min\{\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_p\}$, 则 $\mathbf{X}'\mathbf{X}$ 的条件数定义为

$$k(\mathbf{X}'\mathbf{X}) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

  - 条件数度量了矩阵特征根的散布程度，可用来判断多重共线性是否存在以及多重共线性的严重程度。
  - $k < 100$ 时可认为多重共线性关系很弱; $100 \leqslant k_i \leqslant 1000$ 时认为存在较强的多重共线性; 而 $k_i \geqslant 1000$ 时, 存在较严重的多重共线性.
  - R 代码

```
library(car)
states <- as.data.frame(state.x77[, c("Murder", "Population",
                              "Illiteracy", "Income", "Frost")])
S_states <- cor(states)
kappa(S_states,exact=T)
eigen(S_states)
```

直观判别法

① 如果增加或删除一个自变量或观测值, 回归系数的估计值发生很大的改变, 则认为回归方程存在严重的多重共线性.

② 从定性的角度看, 当一些重要的自变量在回归方程中没有通过显著性检验时, 可初步判断存在较严重的多重共线性.

③ 当有些自变量的回归系数所带正负号与定性分析结果不一致时, 认为存在多重共线性.

④ 自变量的相关矩阵中, 当自变量间的相关系数较大时, 认为可能存在多重共线性.

⑤ 当一些重要的自变量的回归系数的标准误较大时, 认为存在多重共线性.

☞ 剔除不重要的解释变量

- 找到 *VIF* 最大值对应的解释变量，删除后重新估计回归模型，再进行 *VIF* 检验，再删除......

☞ 增大样本容量

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

☞ 回归系数的有偏估计

```
library(foreign)
library(QuantPstc)

avi <- read.csv("E:\\Documents\\ZUFE\\数科学院\\教学课件\\应用回归分析\\数据\\civil.csv")
avi_lm <- lm(guests~.,data=avi)

# 方差扩大因子法
vif(avi_lm)
avi_lm_1 <- lm(guests~consume+realguests+civilmiles+travors,data=avi)
vif(avi_lm_1)
avi_lm_2 <- lm(guests~realguests+civilmiles+travors,data=avi)
vif(avi_lm_2)

lm.beta(avi_lm_2)

# 特征根法

aviT <- cor(as.matrix(avi))
kappa(aviT,exact=T)
eigen(aviT)
```