统计方法与机器学习

第四章：多重共线性 - 1

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

目录

目录

多重共线性的定义与原因

动机

- 在线性模型中，最小二乘估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

- 在自变量 $x_1, x_2, \cdots, x_p$ 中，如果某些自变量可以由另外某些自变量线性表示，那么，我们有

$$|\boldsymbol{X}'\boldsymbol{X}| = 0.$$

这样我们无法得到最小二乘估计 $\hat{\boldsymbol{\beta}}$。

多重共线性的定义与原因

动机

- 在线性模型中，最小二乘估计

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

- 在自变量 $x_1, x_2, \cdots, x_p$ 中，如果某些自变量并不是由另外某些自变量线性表示，那么，我们仍可以得到最小二乘估计 $\hat{\boldsymbol{\beta}}$。但是，当

$$|\boldsymbol{X}'\boldsymbol{X}| \approx 0$$

时，所得到的最小二乘估计的方差会很大，导致估计精度低（不稳定）。

- 这种情形就是我们所讨论的**多重共线性**。

多重共线性的定义与原因

例子：特征相关性与参数估计稳定性的关系

- 考虑对因变量 y 和两个自变量 x_1 和 x_2 建立线性回归.
- 假定 y 与 x_1, x_2 都已经中心化.
- 回归方程为
$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$
- 将 x_1 与 x_2 之间的相关系数记为
$$r_{12} = \frac{l_{12}}{\sqrt{l_{11} l_{22}}}.$$

其中
$$l_{11} = \sum_{i=1}^{n} x_{i1}^2, \quad l_{22} = \sum_{i=1}^{n} x_{i2}^2, \quad l_{12} = \sum_{i=1}^{n} x_{i1} x_{i2}.$$

多重共线性的定义与原因

例子（续）：特征相关性与参数估计稳定性的关系

- 最小二乘估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的方差-协方差矩阵为

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \sigma^2 \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix}^{-1} \\
&= \sigma^2 \begin{pmatrix} l_{11} & l_{12} \\ l_{12} & l_{22} \end{pmatrix}^{-1}
\end{aligned}
$$

多重共线性的定义与原因

线性代数知识（补充）

- 如何 n 阶方阵 A 的逆矩阵？采用**伴随矩阵**的方法.
- 第一步，计算 A 的行列式 $|A|$.
- 第二步，计算行列式 $|A|$ 的各个元素的代数余子式 A_{ij}.
- 第三步，构造方阵 A 的伴随矩阵，即

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{12} & \cdots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix}$$

- 第四步，A 的逆矩阵为 $A^{-1} = \frac{A^*}{|A|}$.

多重共线性的定义与原因

例子（续）：特征相关性与参数估计稳定性的关系

- 最小二乘估计 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的方差-协方差矩阵为

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \begin{pmatrix} l_{11} & l_{12} \\ l_{12} & l_{22} \end{pmatrix}^{-1} \\
&= \sigma^2 \frac{1}{|\boldsymbol{X}'\boldsymbol{X}|} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix} \\
&= \sigma^2 \frac{1}{l_{11}l_{22} - l_{12}^2} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix} \\
&= \sigma^2 \frac{1}{l_{11}l_{22}(1 - r_{12}^2)} \begin{pmatrix} l_{22} & -l_{12} \\ -l_{12} & l_{11} \end{pmatrix}
\end{aligned}
$$

多重共线性的定义与原因

例子（续）：特征相关性与参数估计稳定性的关系

- 因此，我们得到

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(1-r_{12}^2)l_{11}} \tag{1}$$

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)l_{22}} \tag{2}$$

- 结论：
  - 随着自变量 x_1 和 x_2 的相关性**增加**，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差均会**增大**.
  - 特别地，当 $r_{12}^2 \approx 1$ 时，那么回归参数的估计值的方差将**趋于无穷**。此时，x_1 与 x_2 接近于完全正（负）相关。

多重共线性的定义与原因

例子（续）：特征相关性与参数估计稳定性的关系

- 因此，我们得到

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{(1-r_{12}^2)l_{11}} \quad (1)\\
\mathrm{Var}(\hat{\beta}_2) &= \frac{\sigma^2}{(1-r_{12}^2)l_{22}} \quad (2)
\end{aligned}
$$

- 结论：
  - 随着自变量 x_1 和 x_2 的相关性**增加**，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差均会**增大**.
  - 特别地，当 $r_{12}^2 \approx 1$ 时，那么回归参数的估计值的方差将**趋于无穷**。此时，x_1 与 x_2 接近于完全正（负）相关。

多重共线性的定义与原因

说明

- 自变量之间完全不相关的情形非常少见。尤其，当自变量个数较多时，我们很难找到一组自变量，它们不但互相不相关，而且对因变量有显著影响。

- 我们在考虑多个自变量时，
  - 当自变量之间的相关性**较弱**时，我们一般会采用最小二乘估计；
  - 当自变量之间的相关性**较强**时，我们需要优化最小二乘估计。

多重共线性的定义与原因

常见场景

- 对于分类变量，设置过多的虚拟变量。
  - 以**性别**为例。
  - 通常采用的虚拟变量

    $$x_m = I(\text{性别为男性}) \quad \text{和} \quad x_f = I(\text{性别为女性})$$

  - 问题：在建模时，不会将这两个虚拟变量同时纳入模型. 这是为什么?
  - 原因：$x_f = 1 - x_m$，即 x_f 可由 x_m 完全线性表示。
  - 解决方案: 通常有 J 个分类的变量, 至多可以设置 $J-1$ 个虚拟变量。

多重共线性的定义与原因

常见场景

- 对于分类变量，设置过多的虚拟变量。
  - 以**性别**为例。
  - 通常采用的虚拟变量

$$x_m = I(\text{性别为男性}) \quad \text{和} \quad x_f = I(\text{性别为女性})$$

  - 问题：在建模时，不会将这两个虚拟变量同时纳入模型. 这是为什么?
  - 原因：$x_f = 1 - x_m$，即 x_f 可由 x_m 完全线性表示。
  - 解决方案: 通常有 J 个分类的变量, 至多可以设置 $J-1$ 个虚拟变量。

多重共线性的定义与原因

常见场景

- 某一个变量是由其他变量计算而成的。
  - 例如，在研究体型越大的鸟类更容易找到配偶的问题中，这类鸟有一种特别形态的尾部，研究者想探索鸟的整体大小和尾部大小是否有助于其找到配偶。
  - 研究者考虑了三个自变量：鸟的体长、尾长以及整体的长度。
  - 注意到，整体长度是体长和尾长之和。
  - 问题：是否可以将这三个自变量同时纳入模型?
  - 不能!
  - 解决方案：挑选合适的自变量。

多重共线性的定义与原因

常见场景

- 模型中选用同样的或相似的自变量。
    - 同一概念但采用不同的测量方法，由此构造自变量纳入模型。
    - 例如，在研究收入与压力水平的关系时，度量收入的自变量有很多，如：个人收入、家庭收入等。由于这些自变量都可以度量“收入”这一概念，往往具有很高的相关性。
    - 解决方案：找到一个最为合适代表“收入”的特征，放入模型。

目录

方差扩大因子法

定义

- 第 j 个自变量 $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \cdots, x_{nj})'$。
- 自变量 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$ 的样本相关矩阵定义为

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

其中,

$$\rho_{jj'} = \rho_{j'j} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \cdot \sum_{i=1}^{n}(x_{ij'} - \bar{x}_{j'})^2}}$$

方差扩大因子法

定义

- 令
$$x_{ij}^{**}=\frac{x_{ij}-\bar{x}_j}{\sqrt{l_{jj}}}, l_{jj}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2.$$
为**标准化**后第 j 个的自变量。

- 令
$$\boldsymbol{X}_s=(\boldsymbol{x}_1^{**},\boldsymbol{x}_2^{**},\cdots,\boldsymbol{x}_p^{**}),$$
其中，$\boldsymbol{x}_j^{**}=(x_{1j}^{**},x_{2j}^{**},\cdots,x_{nj}^{**})$。

- 于是，
$$\boldsymbol{R}=\boldsymbol{X}_s'\boldsymbol{X}_s.$$

方差扩大因子法

定义

- 这个样本相关矩阵的逆矩阵记为

$$(\boldsymbol{X}_s{}'\boldsymbol{X}_s)^{-1} = \boldsymbol{C} = (c_{ij})$$

- 我们称方阵 $\boldsymbol{C}$ 主对角线元素

$$\mathrm{VIF}_j = c_{jj}$$

为第 j 个自变量的**方差扩大因子**（Variance Inflation Factor, VIF）。

方差扩大因子法

命名的由来

- 在线性回归模型中，不考虑回归常数，并比较原始的自变量以及标准化后的自变量所对应的回归系数，两类最小二乘估计之间的关系为

$$\hat{\beta}_{s,j} = \sqrt{l_{jj}}\hat{\beta}_j$$

其中，$l_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$。

- 于是，

$$\hat{\beta}_j = \frac{1}{\sqrt{l_{jj}}}\hat{\beta}_{s,j}.$$

方差扩大因子法

命名的由来

- 可以证明

$$\mathrm{Var}(\boldsymbol{\beta}_s) = \sigma^2(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}$$

- 因此,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{c_{jj}}{l_{jj}}\sigma^2, \quad j = 1, 2, \cdots, p$$

- 说明：由于 c_{jj} 越大，自变量 x_j 所对应回归系数 β_j 的最小二乘估计 $\hat{\beta}_j$ 的方差也越大。

问题：**为什么方差扩张因子能够用于诊断自变量间存在多重共线性呢?**

方差扩大因子法

另一个角度来看 VIF

- 考虑以下回归模型

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

- 令 R_j^2 为该模型的决定系数。
- 可以证明（留作习题），

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}.$$

- 说明：
  - R_j^2 度量了自变量 $\boldsymbol{x}_j$ 与其余自变量的线性相关程度。
  - 如果 R_j^2 越大，VIF_j 也越大。

方差扩大因子法

如何利用 VIF 判断?

- 基本思想：VIF_j 越大，自变量 x_j 与其他自变量之间的多重共线性程度更严重。
- 判断标准：
  - 当 $\mathrm{VIF}_j < c_{\mathrm{VIF}}$ 时，自变量 $\boldsymbol{x}_j$ 与其余自变量之间不存在多重共线性；
  - 当 $\mathrm{VIF}_j \geq c_{\mathrm{VIF}}$ 时，自变量 $\boldsymbol{x}_j$ 与其余自变量之间存在多重共线性；
- 临界值 c_{VIF} 常见的取值有 $5, 10, 100$。

方差扩大因子法

如何利用 VIF 判断?

- 如何度量整个设计矩阵的多重共线性?

- 用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性，即

$$\overline{\mathrm{VIF}} = \frac{1}{p}\sum_{j=1}^{p} \mathrm{VIF}_j.$$

- 判断准则：当 $\overline{\mathrm{VIF}}$ 特别大时，表示存在严重的多重共线性问题。

- 值得注意的是，当样本量比较小时，R^2 较容易接近 1。因此 $\overline{\mathrm{VIF}}$ 的讨论需要基于样本量而讨论。

特征值判定法

线性代数知识（补充）

- 假定一个 n 阶方阵 $\boldsymbol{A}$ 是一个实对称矩阵。根据特征值分解，

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$$

  注意到，

  - $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$，其中特征值分别为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.
  - $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$，其中 $\boldsymbol{v}_i$ 是特征值 λ_i 所对应的特征向量，$i = 1, 2, \cdots, n$。

- A 的行列式等于其特征值的乘积，即

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$$

特征值判定法

概述

- 这里仅仅考虑自变量是经过标准化后的，记为 $\boldsymbol{X}_s$。
- 根据线性代数的知识可知，行列式 $|\boldsymbol{X}_s'\boldsymbol{X}_s| \approx 0$ 时，矩阵 $\boldsymbol{X}_s'\boldsymbol{X}_s$ 至少存在一个特征值近似为零。
- 反之，当矩阵 $\boldsymbol{X}_s'\boldsymbol{X}_s$ 至少存在一个特征值近似为零时，$\boldsymbol{X}_s$ 的列向量间必然存在多重共线性。

特征值判定法

具体来说

- 假定 λ 是矩阵 $\boldsymbol{X}_s'\boldsymbol{X}_s$ 的一个近似为零的特征值，即
$$\lambda \approx 0$$
- 而 $\boldsymbol{v} = (v_1, \cdots, v_p)'$ 是特征值 λ 所对应的单位特征向量，则
$$\boldsymbol{X}_s'\boldsymbol{X}_s\boldsymbol{v} = \lambda\boldsymbol{v} \approx 0$$
- 于是，在上式中等式两端都左乘 $\boldsymbol{v}'$，可得
$$\boldsymbol{v}'\boldsymbol{X}_s'\boldsymbol{X}_s\boldsymbol{v} \approx 0 \Rightarrow (\boldsymbol{X}_s\boldsymbol{v})'\boldsymbol{X}_s\boldsymbol{v} \approx 0 \Rightarrow \boldsymbol{v}'\boldsymbol{X}_s \approx 0$$
- 这与多重共线性的定义是一致的。

特征值判定法

判定方法

- 假设 $\boldsymbol{X}_s'\boldsymbol{X}_s$ 的特征值分别为 $\lambda_1 \geq \cdots \geq \lambda_p$。
- 称
$$\kappa_j = \sqrt{\frac{\lambda_1}{\lambda_j}}, j = 1, 2, \cdots, p$$
为特征值 λ_j 的条件数（condition index）
- 基本想法：
  - 如果设计矩阵 $\boldsymbol{X}_s$ 没有多重共线性，即最小特征值 λ_p 不会接近零，那么条件数 κ_p 不会特别大；
  - 设计矩阵 $\boldsymbol{X}_s$ 存在多重共线性，即最小特征值 λ_p 接近零，那么条件数 κ_p 会特别大。

特征值判定法

判定方法

- 常用的判断标准
  - $0 < \kappa_p < c_\kappa$ 时，设计矩阵 $\boldsymbol{X}_s$ 没有多重共性性；
  - $\kappa_p \geq c_\kappa$ 时，设计矩阵 $\boldsymbol{X}_s$ 存在多重共线性；
- 临界值 c_κ 的常见取值有 $10, 100, 1000$。

直观判定法

总结

- 量化标准
  - 方差扩大因子
  - 条件数
- 这种量化标准并不是识别多重共线性的绝对标准，还应该结合一些直观方法综合识别多重共线性。

直观判定法

判定方法

- **增加或剔除**一个自变量，其他自变量的回归系数的估计值或显著性发生**较大**变化；
- 从定性分析或者背景知识的角度，一些**重要**的自变量在回归方程中从**没有通过显著性检验**；
- 有些自变量的回归系数的数值大小与预期相差数**很大**，甚至正负号与定性分析结果数**相反**；
- 计算自变量的相关矩阵，自变量间的相关系数数**较大**。

消除多重共线性的方法

三种常见方法

- 选择合适的自变量；
- 增加数据（样本量）；
- 改进最小二乘估计（岭回归、主成分回归）。

假定

- 在介绍岭回归和主成分回归时，这里假定了设计矩阵 $\boldsymbol{X}$ 是经过标准化，而因变量 y 未经过标准化。