六、多重共线性

目录

多重共线性及其产生的原因

- 多重共线性及其产生的原因
  - 完全共线性 (代数形式)：存在不全为 0 的常数 $\lambda_i (i = 1, 2, \cdots, k)$ 有 $\sum \lambda_i x_i = 0$
  - 完全共线性 (矩阵形式)：令 $\boldsymbol{X} = (\mathbf{1}_{n\times 1}, x_{1j}, x_{2j}, \cdots, x_{kj})$，则 $|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}| = 0$ 或 $\mathrm{rank}(\boldsymbol{X}) < k+1$
  - 产生原因：

    经济变量之间在时间上往往存在同方向的变化趋势；

    经济变量之间往往存在着密切的关联度；

    在模型中引入滞后变量；

    解释变量选择不当

多重共线性的影响

- 二元线性回归模型下多重共线性的影响 (OLS)
  - $\mathrm{Var}(\hat{b}_i) = \dfrac{\sigma^2}{\sum(x_{it} - \bar{x}_i)^2} \cdot \mathrm{VIF}_i$，完全共线性下 $\hat{b}_i$ 为不定式，方差为无穷大；方差膨胀因子 $\mathrm{VIF}_i = \dfrac{1}{1 - R_i^2}$
  - 近似共线性下：

    $\hat{b}_1, \hat{b}_2$ 的方差随 x_1, x_2 共线性的增加而增加；

    $\hat{b}_1, \hat{b}_2$ 的置信区间随 x_1, x_2 共线性的增加而变大；

    x_1, x_2 存在严重共线性时，t 检验失效，预测精度降低；

    $\hat{b}_1, \hat{b}_2$ 的经济含义不合理，回归模型缺乏稳定性

    保持线性、无偏性、一致性，不具备有效性

    引入与解释变量无关的变量，估计量导致不再具有最小方差性，精度下降

多重共线性的检验

- 多重共线性的检验
  - 相关系数检验法：

    计算 x_i, x_j 之间的相关系数矩阵 $\boldsymbol{A} = [r_{ij}]_{k\times k}$；

    $r_{ij}^2 > R^2$，认为 x_i, x_j 之间的共线性较为严重
  - 法勒-格劳伯检验 (F-G 检验)：

    第一，χ^2 检验：$H_0 : x_i$ 之间是正交的 (不存在多重共线性)，$H_1 : x_i$ 之间不是正交的 (存在多重共线性)；

    检验统计量 $\chi^2 = -\left[n-1-\frac{1}{6}(2k+5)\right]\cdot|\boldsymbol{A}| \sim \chi^2\left[\frac{k(k-1)}{2}\right]$；

    $\chi^2 > \chi_\alpha^2$，拒绝 H_0，χ^2 越大，多重共线性越大；$\chi^2 < \chi_\alpha^2$，接受 H_0；

多重共线性的检验

- 多重共线性的检验
  - 法勒-格劳伯检验 (F-G 检验):

    第二，F 检验：将 x_i 对其余 x_j 进行回归，分别求得

    $R_i^2, F_i (i = 1, 2, \cdots, k)$;

    若其中最大的 R_i^2 接近 1，F_i 显著大于临界值，则 x_i 与其余 x_j 之间存在多重共线性;

    第三，t 检验：计算 x_i 与其余每一个 x_j 之间的偏相关系数 r_{ij}，构建统计量 $t = \dfrac{r_{ij}\sqrt{n-k}}{\sqrt{1-r_{ij}}} \sim t(n-k)$;

    $|t| > t_{\alpha/2}(n-k)$，认为 x_i, x_j 的偏相关系数是显著的，x_i, x_j 是引起多重共线性的原因；$|t| < t_{\alpha/2}(n-k)$，认为 x_i, x_j 的偏相关系数不是显著的，x_i, x_j 不是引起多重共线性的原因

多重共线性的检验

- 多重共线性的检验

  - 方差膨胀因子检验：

    $\mathrm{VIF}_i = \dfrac{1}{1-R_i^2}$，$R_i^2 \to 1, \mathrm{VIF}_i \to \infty$；

    $\mathrm{VIF}_i > 5$ 或 $\mathrm{VIF}_i > 10$ 时，认为模型存在较严重的多重共线性

  - 特征值检验：

    令 $\boldsymbol{X} = (\mathbf{1}_{n\times 1}, x_{1j}, x_{2j}, \cdots, x_{kj})$，$|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}| = \lambda_1 \cdot \lambda_2 \cdots \lambda_{k+1} \approx 0$；

    特征值 λ_i 中至少有一个近似地等于 0；

    条件数 $\mathrm{CN} = \dfrac{\lambda_{\max}}{\lambda_{\min}}$；条件指数 $\mathrm{CI} = \sqrt{\mathrm{CN}}$；

    CN 和 CI 数值越大，多重共线性越严重；$\mathrm{CI} > 10$ 即认为存在多重共线性，大于 30 认为存在严重的多重共线性

多重共线性的解决方法

- 多重共线性的解决方法
  - 保留重要的解释变量，去掉可替代的解释变量：x_i, x_j 之间的 $r_{ij}^2 > R^2$，保留其中之一
  - 用先验信息改变参数的约束形式
  - 变换模型的形式：

    一阶差分法，令 $\Delta y_t = y_t - y_{t-1}, \Delta x_{it} = x_{it} - x_{i,t-1}, \Delta u_t = u_t - u_{t-1}$，则

    $$\Delta y_t = b_1 \Delta x_{1t} + b_2 \Delta x_{2t} + \cdots + b_k \Delta x_{kt} + \Delta u_t$$
  - 使用时序数据与截面数据：用不同类型的数据分别估计模型中一部分的参数

多重共线性的解决方法

- 多重共线性的解决方法
  - 逐步回归法 (Frisch 综合分析法):

    利用相关系数从所有解释变量中选与被解释变量相关性最强的变量建立一元回归模型;

    引入第二个变量，建立 $k-1$ 个二元回归模型，选其中较优的，(影响显著，参数符号正确，$\bar{R}^2$ 有所提高);

    重复步骤，直至无法引入
  - 增加样本容量

多重共线性的解决方法

- 多重共线性的解决方法
  - 主成分分析：

    对原始数据进行标准化，计算相关系数矩阵 $\boldsymbol{R}$；

    计算 $\boldsymbol{R}$ 的特征值及对应的标准化特征向量，检验多重共线性；

    主成分为 $z_i = u_{i1}x_1 + u_{i2}x_2 + \cdots + u_{ik}x_k, i = 1, 2, \cdots, k$，取其中不近似为 0 的 z_i 进行回归，得到 $\hat{y} = \hat{a}_1 z_1 + \hat{a}_2 z_2 + \cdots + \hat{a}_m z_m$；

    将主成分代回 y 与 z 的方程，得到原回归参数

    $\hat{b}_i = \frac{s_y}{s_i}\hat{\beta}_i, \hat{b}_0 = \bar{y} - \sum \hat{b}_i \bar{x}_i$

代码输出结果分析

▶ 回归分析结果

同第二章：

常数和解释变量	参数估计值	参数标准误差	t统计量	双侧概率
C(b_0)	331.5264	57.16954	5.799003	0.0000
PI(b_1)	0.692812	0.006279	110.3337	0.0000
决定系数	0.997297	被解释变量均值		4662.514
调整的决定系数	0.997215	被解释变量标准差		4659.100
回归标准误差	245.8925	赤池信息准则		13.90311
残差平方和	1995283.	施瓦兹信息准则		13.99199
对数似然函数	−241.3044	汉南准则		13.93379
F统计量	12173.53	DW统计量		0.180221
F统计量的概率	0.000000			

代码输出结果分析

- 各种检验的输出结果分析

同第四章：

英文	含义	英文	含义
Heterpskedasticity Test	检验方法	F-statistic	回归模型的 F 统计量
Obs*R-squared	F 检验统计量	Prob. Chi-Square(2)	F 统计量对应的 p 值
Prob. F(a, b)	自由度为 a,b 的 F 分布临界值	Scaled explained SS	LM 统计量