三、多元线性回归模型

目录

多元线性回归模型的估计

- 总体与样本回归模型与方程与矩阵表示
  - 总体回归模型：$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_k x_{kt} + u_t$，$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U}$
  - 总体回归方程：$E(y_t) = b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_k x_{kt}$，$E(\boldsymbol{Y}) = \boldsymbol{XB}$
  - 样本回归模型：$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_{1t} + \hat{b}_2 x_{2t} + \cdots + \hat{b}_k x_{kt} + e_t$，$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{B}} + \boldsymbol{e}$
  - 样本回归方程：$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_{1t} + \hat{b}_2 x_{2t} + \cdots + \hat{b}_k x_{kt}$，$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}}$
  - $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$，$\boldsymbol{X} = (\boldsymbol{1}_{n\times 1}, x_{1j}, x_{2j}, \cdots, x_{kj})$，$\boldsymbol{B} = (b_0, b_1, \cdots, b_k)^{\mathrm{T}}$，$\boldsymbol{U} = (u_1, u_2, \cdots, u_n)^{\mathrm{T}}$，$\boldsymbol{e} = (e_1, e_2, \cdots, e_n)^{\mathrm{T}}$
  - 能够得出参数估计值要求 $n > k+1$

多元线性回归模型的估计

- 多元线性回归模型的基本假定
  - $E(u_t) = 0$，$E(\boldsymbol{U}) = 0$
  - $\mathrm{Cov}(u_t, u_s) = 0$
  - $\mathrm{Var}(u_t) = \sigma^2$，$E(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}) = \sigma^2 I_n$
  - $\mathrm{Cov}(x_{jt}, u_t) = 0$，$E(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}) = \boldsymbol{0}$
  - $u_t \sim N(0, \sigma^2)$，$\boldsymbol{U} \sim N(\boldsymbol{0}, \sigma^2 I_n)$
  - 解释变量之间不存在多重共线性，$\mathrm{rank}(\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) = k+1$

多元线性回归模型的估计

- 多元线性回归模型的估计
  - 回归参数的最小二乘估计：$\hat{\boldsymbol{B}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$
  - 最小二乘估计量的性质：

    线性：$\hat{\boldsymbol{B}} = \boldsymbol{B} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}$；

    无偏性：$E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$；

    最小方差性：$\mathrm{Var}(\hat{\boldsymbol{B}}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$；

    $\hat{\boldsymbol{B}} \sim N(\boldsymbol{B}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$，$\hat{b}_j \sim N(b_j, \sigma^2 c_{jj})$ (c_{jj} 为 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 的第 j 个主对角元素)
  - 随机误差项的方差：$\hat{\sigma}^2 = \dfrac{\sum e_t^2}{n-k-1}$，$E(\hat{\sigma}^2) = \sigma^2$，$s(\hat{b}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}$

多元线性回归模型的估计

- 多元线性回归模型的估计
  - 极大似然估计法 (ML):

    极大似然函数：$L(\theta)=\prod_{i=1}^{n} f(y_i,\theta)$;

    $\hat{\theta}_{ML}$ 使得 $\max L(\theta)$，则 $p(\lim \hat{\theta}_{ML})=\theta_0$，$\hat{\theta}_{ML}\sim N(\theta_0,V(\theta_0))$
  - ML 法得到的 $\hat{\sigma}^2$ 是 σ^2 的有偏、一致估计量
  - 对于线性回归模型，用极大似然估计法得到的系数估计值与用最小二乘估计法得到的结果完全相同
  - 参数置信区间:

    b_j 的置信区间：$[\hat{b}_j-t_{\alpha/2}(n-k-1)s(\hat{b}_j),\hat{b}_j+t_{\alpha/2}(n-k-1)s(\hat{b}_j)]$

多元线性回归模型的检验

▶ 多元线性回归模型的检验

- 拟合优度检验：

  TSS 的自由度为 $n-1$，RSS 的自由度为 $n-k-1$，ESS 的自由度为 k；

  决定系数：$R^2 = \dfrac{\text{ESS}}{\text{TSS}} = 1 - \dfrac{\text{RSS}}{\text{TSS}}$；

  修正的决定系数：$\bar{R}^2 = \dfrac{\text{ESS}/k}{\text{TSS}/(n-1)} = 1 - \dfrac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$；

  $\bar{R}^2 = 1 - \dfrac{n-1}{n-k-1}(1-R^2) = R^2 - \dfrac{k}{n-k-1}(1-R^2)$，$\bar{R}^2 < R^2$

- 赤池信息准则：$\text{AIC} = \ln \dfrac{\sum e_t^2}{n} + \dfrac{2(k+1)}{n}$，值越小，拟合优度越好
- 施瓦兹准则：$\text{SC} = \ln \dfrac{\sum e_t^2}{n} + \dfrac{k}{n} \ln n$，值越小，拟合优度越好

多元线性回归模型的检验

▶ 多元线性回归模型的检验

- 回归方程的 F 检验：$H_0: b_1 = b_2 = \cdots = b_k = 0, H_1: b_j$ 不全为 0

  $\mathrm{ESS} \sim \chi^2(k), \mathrm{RSS} \sim \chi^2(n-k-1)$；$F = \dfrac{\mathrm{ESS}/k}{\mathrm{RSS}/(n-k-1)}$；

  $F > F_\alpha(k, n-k-1)$，拒绝 H_0，回归方程显著；$F < F_\alpha(k, n-k-1)$，

  接受 H_0，回归方程不显著；

  $R^2 = \dfrac{kF}{(n-k-1)+kF}, \bar{R}^2 = 1 - \dfrac{n-1}{(n-k-1)+kF}$

- 回归参数的 t 检验：$H_0: b_j = 0, H_1: b_j \neq 0$

  $t = \dfrac{\hat{b}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k-1)$；

  $|t| \geq t_{\alpha/2}(n-k-1)$，拒绝 H_0，x_j 对 y 的影响是显著的；

  $|t| < t_{\alpha/2}(n-k-1)$，接受 H_0，x_j 对 y 的影响是不显著的；

  一元情况下，$F = t^2$

多元线性回归模型的预测

- 多元线性回归模型的预测
  - 点预测：给定 $\boldsymbol{X}_f$，代入样本回归方程，求得 $\hat{y}_f$
  - 区间预测：

    总体均值：$E(y_f) = \hat{y}_f \pm t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{X}_f(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_f^{\mathrm{T}}}$；

    样本预测值：$y_f = \hat{y}_f \pm t_{\alpha/2}s(\hat{y}_f)$，其中 $s(\hat{y}_f) = \hat{\sigma}\sqrt{1+\boldsymbol{X}_f(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_f^{\mathrm{T}}}$

多元线性回归模型的预测

- 多元线性回归模型的预测
  - 预测评价：

    平均绝对误差：$\mathrm{MAE} = \frac{1}{n}\sum |\hat{y}_t - y_t|$；

    均方根误差：$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (\hat{y}_t - y_t)^2}$；

    平均相对误差：$\mathrm{MPE} = \frac{1}{n}\sum \left|\frac{\hat{y}_t - y_t}{y_t}\right|$，值低于 10，预测精度较高；

    Theil 不等系数：$\text{Theil IC} = \frac{\sqrt{\frac{1}{n}\sum (\hat{y}_t - y_t)^2}}{\sqrt{\frac{1}{n}\sum \hat{y}_t^2} + \sqrt{\frac{1}{n}\sum y_t^2}}$，$\text{Theil IC} \in [0,1]$，值越小，预测精度越高

非线性回归模型

▶ 可线性化模型

- 对数模型：$\ln y = b_0 + b_1 \ln x + u$，令 $y^* = \ln y, x^* = \ln x$，则 $y^* = b_0 + b_1 x^* + u$，b_1 是 y 关于 x_1 的弹性 (xy'/y)；
- 半对数模型：$y = b_0 + b_1 \ln x + u$ 或 $\ln y = b_0 + b_1 x + u$，令 $y^* = \ln y$ 或 $x^* = \ln x$ 即可，b_1 是 x 的相对 (绝对) 变化引起 y 的期望值绝对 (相对) 变化；
- 倒数模型：$y = b_0 + b_1 \dfrac{1}{x} + u$ 或 $\dfrac{1}{y} = b_0 + b_1 x + u$，令 $y^* = \dfrac{1}{y}$ 或 $x^* = \dfrac{1}{x}$ 即可；
- 多项式模型：$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k + u$，设 $x_t = x^t$ 即可；
- 逻辑成长曲线模型：$y_t = \dfrac{K}{1 + b_0 \mathrm{e}^{-b_1}}$，两边取倒数，再取 ln 即可；
- 龚珀兹成长曲线：$y_t = \mathrm{e}^{K + b_0 b_1^t}$，两边取两次 ln 即可

代码输出结果分析

▶ 代码输出结果分析

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			