二、一元线性回归模型

目录

一元线性回归模型的基本假定

- 一元线性回归模型的概念
  - 变量之间的两种关系：函数关系、相关关系
  - 模型：$y_t = b_0 + b_1 x_t + u_t$
  - x_t 是解释变量，y_t 是被解释变量，b_0, b_1 是回归参数，b_0 是常数项，b_1 是回归系数，u_t 是随机误差项
  - 线性指：

$$\frac{\partial y_t}{\partial x_t} = b_1, \frac{\partial^2 y_t}{\partial x_t^2} = 0, \frac{\partial y_t}{\partial b_0} = 1, \frac{\partial y_t}{\partial b_1} = x, \frac{\partial^2 y_t}{\partial b_0^2} = \frac{\partial^2 y_t}{\partial b_1^2} = 0$$

一元线性回归模型的基本假定

- 一元线性回归模型的基本假定
  - $E(u_t) = 0$
  - $\mathrm{Var}(y_t) = \mathrm{Var}(u_t) = \sigma^2$
  - $\mathrm{Cov}(y_t, y_s) = \mathrm{Cov}(u_t, u_s) = 0$
  - $\mathrm{Cov}(x_t, u_t) = 0$
  - $u_t \sim N(0, \sigma^2)$
  - 总结：x_t 非随机变量，u_t 为独立的随机变量，$u_t \sim N(0, \sigma^2)$ 或 $y_t \sim N(b_0 + b_1 x_t, \sigma^2)$

一元线性回归模型的参数估计

- 总体与样本回归模型与方程
  - 总体回归模型：$y_t = b_0 + b_1 x_t + u_t$
  - 总体回归方程：$E(y_t) = b_0 + b_1 x_t$
  - 样本回归模型：$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_t + e_t$
  - 样本回归方程：$\hat{y}_t = \hat{b}_0 + \hat{b}_1 x_t$
  - x 与 y 的真实线性关系：$y_t = b_0 + b_1 x_t + u_t$

一元线性回归模型的参数估计

- 普通最小二乘法 (OLS)
  - 最小二乘准则：$\min\sum e_t^2 = \min\sum(y_t - \hat{b}_0 - \hat{b}_1 x_t)^2$
  - 满足：$\sum(y_t - \hat{y}_t) = 0$
  - 解为：

$$\hat{b}_1 = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2} = \frac{n\sum x_t y_t - \sum x_t \sum y_t}{n\sum x_t^2 - (\sum x_t)^2}, \quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

  - 估计量与最大似然估计的结果完全一致 (无论大小样本)
  - $\bar{\hat{y}} = \bar{y}$，样本方程经过 $(\bar{x}, \bar{y})$，$\sum y_t = \sum \hat{y}_t$，$\sum e_t = 0$

一元线性回归模型的参数估计

- 最小二乘估计量的性质
  - 线性：$\hat{b}_0, \hat{b}_1$ 分别是 y_t, u_t 的线性函数或线性组合
  - 无偏性：$E(\hat{b}_0) = b_0, E(\hat{b}_1) = b_1$
  - 有效性 (方差最小)：$\mathrm{Var}(\hat{b}_1) = \dfrac{\sigma^2}{\sum(x_t - \bar{x})^2}, \mathrm{Var}(\hat{b}_0) = \dfrac{\sigma^2 \sum x_t^2}{n\sum(x_t - \bar{x})^2}$
- 回归参数的区间估计
  - 回归参数的分布：$\hat{b}_0 \sim N(b_0, \mathrm{Var}(\hat{b}_0)), \hat{b}_1 \sim N(b_1, \mathrm{Var}(\hat{b}_1))$
  - 总体 (随机误差项) 的方差：
    $\hat{\sigma}^2 = \dfrac{\sum e_t^2}{n-2} = \dfrac{\sum(y_t - \bar{y})^2 - \hat{b}_1 \sum(x_t - \bar{x})(y_t - \bar{y})}{n-2}$，$E(\hat{\sigma}^2) = \sigma^2$
    对于 $\hat{b}_0, \hat{b}_1$ 的方差可以用 $\hat{\sigma}^2$ 来估计 σ^2

一元线性回归模型的参数估计

- 回归参数的区间估计
  - 总体 (随机误差项) 的方差：

    大样本时：$\hat{b}_0 \sim N(b_0, \mathrm{Var}(\hat{b}_0)), \hat{b}_1 \sim N(b_1, \mathrm{Var}(\hat{b}_1))$

    小样本时：$\dfrac{\hat{b}_1 - b_1}{s(\hat{b}_1)} \sim t(n-2)$ (自由度为 $n-2$ 的 t 分布)
  - 回归系数的区间估计：

    σ^2 已知，区间为 $[\hat{b}_1 - 1.96\sigma(\hat{b}_1), \hat{b}_1 + 1.96\sigma(\hat{b}_1)]$

    σ^2 未知，样本容量较大，区间为 $[\hat{b}_1 - 1.96s(\hat{b}_1), \hat{b}_1 + 1.96s(\hat{b}_1)]$

    σ^2 未知，样本容量较小，区间为

    $[\hat{b}_1 - t_{\alpha/2}(n-2)s(\hat{b}_1), \hat{b}_1 + t_{\alpha/2}(n-2)s(\hat{b}_1)]$
  - 缩小置信区间的方法：增大样本容量 n，提高模型的拟合优度 R^2

一元线性回归模型的检验

- 一元线性回归模型的检验
  - 经济意义检验：检验回归参数的正负是否符合经济意义
  - 回归参数的显著性检验：$H_0: b_1 = 0$(不显著), $H_1: b_1 \neq 0$(显著)

    t 检验：

    大样本：计算 $Z = \dfrac{\hat{b}_1}{\sigma(\hat{b}_1)}$ 与 $Z_{\alpha/2}$；

    $|Z| \leq Z_{\alpha/2}$，接受 H_0；$|Z| > Z_{\alpha/2}$，拒绝 H_0

    小样本：计算 $t = \dfrac{\hat{b}_1}{s(\hat{b}_1)}$ 与 $t_{\alpha/2}(n-2)$；

    $|t| \leq t_{\alpha/2}(n-2)$，接受 H_0；$|t| > t_{\alpha/2}(n-2)$，拒绝 H_0；

    p 值检验法：$p < \alpha$，拒绝 H_0；$p \geq \alpha$，接受 H_0

一元线性回归模型的检验

▶ 一元线性回归模型的检验

- 拟合优度检验：$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$；

  总离差平方和 $\mathrm{TSS} = \sum (y_t - \bar{y})^2$；

  回归平方和 $\mathrm{ESS} = \sum (\hat{y}_t - \bar{y})^2$，反映模型中解释变量所解释的那部分离差；

  剩余 (残差) 平方和 $\mathrm{RSS} = \sum e_t^2 = \sum (y_t - \hat{y}_t)^2$，反映模型中解释变量未解释的那部分离差；

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

  $R^2 \in [0,1]$，R^2 约接近 1，拟合优度越好

- 相关系数检验：$\rho_{xy} = \dfrac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} = \dfrac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum (x_t - \bar{x})^2 \sum (y_t - \bar{y})^2}}$；

  $R^2 = r_{xy}^2$，$r_{xy} \in [-1,1]$

一元线性回归模型的检验

▶ 一元线性回归模型的检验

- 相关系数检验：$H_0:\rho=0, H_1:\rho\neq 0$

  相关系数检验：$|r|\leq r_\alpha(n-2)$，接受 H_0，不存在显著的线性相关关系；$|r|>r_\alpha(n-2)$，拒绝 H_0，存在显著的线性相关关系；

  t 检验：$t=\dfrac{r-\rho}{s(r)}=\dfrac{r\sqrt{n-2}}{1-r^2}\sim t(n-2)$；

  $|t|\leq t_{\alpha/2}(n-2)$，接受 H_0；$|t|>t_{\alpha/2}(n-2)$，拒绝 H_0

- 正态性检验：偏度系数 $S=\dfrac{\sum(x_t-\bar{x})^3}{n\sigma_x^3}$；峰度系数 $K=\dfrac{\sum(x_t-\bar{x})^4}{n\sigma_x^4}$；

  JB (雅克-贝拉) 统计量：$\mathrm{JB}=\dfrac{n}{6}\left[S^2+\dfrac{(K-3)^2}{4}\right]\sim\chi^2(2)$；

  $\mathrm{JB}\leq\chi^2_\alpha$，接受 H_0；$\mathrm{JB}>\chi^2_\alpha$，拒绝 H_0

一元线性回归模型的预测

▶ 一元线性回归模型的预测

- 点预测：给定 x_f，代入样本回归方程，求得 y_f
- 区间预测：

  总体均值：$E(y_f) = \hat{y}_f \pm t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_t - \bar{x})^2}}$；

  样本预测值：$y_f = \hat{y}_f \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_t - \bar{x})^2}}$
- 影响预测区间的大小的因素：

  $\hat{\sigma}^2$ 越小，预测精度越高；

  n 越大，预测精度越高；

  $\sum(x_t - \bar{x})^2$ 越大，预测精度越高；

  $(x_f - \bar{x})^2$ 越小，预测精度越高

代码输出结果分析

- 代码输出结果分析

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			