Chapter5

小样本 OLS

二元线性回归

$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \epsilon_i$

```
reg y x1 x2

predict lny1 // 计算拟合值，命名为 lny1
predict e, re // 计算残差，命名为 e，re 可选参数指计算残差，没有则为计算拟合值

list lny lny1 e // 列出真实值和上面算出来的拟合值
```

古典线性回归模型的假定

- **线性假定**：每个解释变量对y_i的边际效应为常数。如果边际效应可变（解释变量对因变量的影响可能不是恒定的，而是随变量值的变化而变化），可加入平方项（x_{i2}^2）或交叉项（$x_{i2}x_{i3}$）。
- **严格外生性（零条件均值）**：误差项（残差）与所有的解释变量（自变量）在任何时刻都是不相关的。
- **不存在严格多重共线性**：不存在某个解释变量是另一个解释变量的倍数，或可由其他解释变量线性表出的情形。
- **球型扰动项**：扰动项条件同方差且无自相关。
- **随机抽样**：在重复抽样中，样本数据是从总体中随机抽取的，保证观测值之间相互独立。

高斯-马尔可夫定理

扰动项理想情况下必须满足四个条件，这些条件被称为**高斯-马尔可夫条件（Gauss-Markov Conditions）**：

1. **零条件均值** $E(\epsilon_i|X_i) = 0$
2. **同方差性** $E(\epsilon|X_i) = \sigma^2$ 对于所有 i均成立
3. **非自相关性** $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$
4. **自变量非随机性** $\text{Cov}(X_i, u_i) = 0$

如果所有高斯-马尔可夫条件都满足，OLS是所有估计量中最优的，因为在数学上它被证明是**最佳线性无偏估计量**（Best Linear Unbiased Estimator, BLUE）。这个结论被称为**高斯-马尔可夫定理**。

OLS 的小样本性质

- 线性性：OLS 估计量$\beta = (X'X)^{-1}X'y$为 y 的线性组合。
- 无偏性：$\mathrm{E}(\boldsymbol{b} \mid \boldsymbol{X}) = \boldsymbol{\beta}$,即$\boldsymbol{b}$不会系统地高估或低估$\boldsymbol{\beta}$。
- 估计量$\boldsymbol{b}$的方差为 $\mathrm{Var}(\boldsymbol{b} \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- 高斯-马尔可夫定理证明 OLS 得到的是 BLUE
- 方差的无偏估计：$E\left(s^2|X\right) = \sigma^2$

假设检验

单系数 t 检验

检验某个特定自变量（或解释变量）对因变量是否具有显著的线性影响。

小样本理论（有限样本理论）：不要求样本容量$n \to \infty$

假定：$\epsilon \mid X \sim N\left(0, \sigma^2 I_n\right)$（数据来自正态分布的总体）

> **正态分布的特点**
>
> 密度函数完全由均值和协方差矩阵决定
>
> 两个随机变量不相关就意味着相互独立
>
> 正态分布变量的线性函数仍然是正态分布

原假设H_0：$\beta_k = 0$（假定自变量 (X_k) 对因变量 (Y) 没有线性影响）。检测在原假设成立的前提下，是否导致不太可能发生的小概率事件在一次抽样的样本中实现。如果小概率事件在一次抽样的样本被观测到，那么假设不可信，拒绝原假设，接受替代假设H_1：$\beta_k \neq 0$

步骤

1. 计算 t 统计量：$t_k \equiv \frac{\hat{\beta}_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} \sim t\,(n)$，其中$\mathrm{SE}\left(\hat{\beta}_k\right) = \frac{s}{\sqrt{n}}$
2. 计算显着性水平为α的临界值$t_{\alpha/2}\,(n-K)$，通常$\alpha = 5\%, \alpha/2 = 2.5\%$
3. 如果$|t_k| \geqslant t_{\alpha/2}\,(n-K)$，则 t 落入拒绝域，拒绝原假设。

计算 $\text{p-value} \equiv P\left(|T| > |t_k|\right),\ T \sim t\,(n-K)$

p值是一个概率值，表示在零假设为真的情况下，得到一个像 (t_k) 或更极端的样本统计量（即偏离原假设的程度更大）的概率。

F 检验

$$H_0:\ \beta_2 = \ldots = \beta_k = 0$$

计算临界值 $F_\alpha\,(m,\ n-K)$，如果大于临界值，则落入拒绝域。

Stata 实现

$$\ln w = \beta_1 + \beta_2 s + \beta_3 \mathrm{expr} + \beta_4 \mathrm{tenure} + \beta_5 \mathrm{smsa} + \beta_6 \mathrm{rns} + \epsilon$$

```
reg lnw s expr tenure smsa rns //... 先做线性回归
```

Source	SS	df	MS	Number of obs	=	758
				F(5, 752)	=	81.75
Model	49.0478814	5	9.80957628	Prob > F	=	0.0000
Residual	90.2382684	752	.119997697	R-squared	=	0.3521
				Adj R-squared	=	0.3478
Total	139.28615	757	.183997556	Root MSE	=	.34641

F 落在拒绝域

p-value 均小于 0.05

lnw	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
s	.102643	.0058488	17.55	0.000	.0911611	.114125
expr	.0381189	.0063268	6.02	0.000	.0256986	.0505392
tenure	.0356146	.0077424	4.60	0.000	.0204153	.0508138
smsa	.1396666	.0280821	4.97	0.000	.0845379	.1947954
rns	-.0840797	.0287973	-2.92	0.004	-.1406124	-.0275471
_cons	4.103675	.085097	48.22	0.000	3.936619	4.270731

```
vce // variance convariance matrix estimated 协方差矩阵
```

Covariance matrix of coefficients of **regress** model

e(V)	s	expr	tenure	smsa	rns	_cons
s	.00003421					
expr	8.660e-06	.00004003				
tenure	-3.997e-08	-.00001107	.00005994			
smsa	-.0000144	3.261e-06	-7.819e-06	.00078861		
rns	8.524e-06	7.334e-07	7.259e-06	.00012486	.00082928	
_cons	-.00046567	-.00016778	-.00008646	-.00038746	-.00043997	.0072415

```
reg lnw s expr tenure smsa if rns // 删掉 rns 为 0 的行进行回归
reg lnw s expr tenure smsa if !rns // 删掉 rns 不为 0 的行进行回归
reg lnw s expr tenure smsa if s>12 // 删掉 s 小于 12 的行进行回归

quietly reg lnw s expr tenure smsa rns // 不输出回归结果，以便后面使用 predict
```

进行假设检验，原假设为 s 对应的 $\beta_2 = 0.1$

```
test s=0.1
```

返回 F（1, 752）和 p-value

又检验 expr-tenure = 0

```
test expr = tenure
```