多元统计分析—R 与 Python 的实现

回归

吴喜之

June 28, 2019

经典线性回归的数学假定:

1. 模型对于参数的线性形式为:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \ \text{ 或 } \ y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i.\ i = 1, 2, \ldots, n;$$

2. $\boldsymbol{y}$ 的所有值都是互相独立的, 如果把 $\boldsymbol{X}$ 看成是固定的, 则等价于 $\boldsymbol{\epsilon}$ 的所有值都是互相独立的, 这意味着没有自相关;
3. 对于每个 $\boldsymbol{X}$ 的值, $\boldsymbol{y}$ 的分布是正态分布的, 如果把 $\boldsymbol{X}$ 看成是固定的, 则等价于 $\boldsymbol{\epsilon}$ 的分布是正态分布;
4. 条件期望 $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X\beta}$ 或 $E(\boldsymbol{\epsilon}|\boldsymbol{X}) = \boldsymbol{0}$, 协方差阵 $\mathrm{Cov}(\boldsymbol{y}|\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$, 这意味着所有观测值独立同方差 (homoscedasticity), 对于正态分布意味着所有观测值都独立同分布 (当然误差项也独立同分布);
5. 自变量和 $\boldsymbol{\epsilon}$ 独立;

注意: 如果假定自变量是弱外生的 (weak exogeneity), 可以近似地把 $\boldsymbol{X}$ 看成是固定的, 上面的概率就不是条件概率了, 自变量和 $\boldsymbol{\epsilon}$ 独立的条件就用不着了.

当自变量被看成是固定的, 人们所说的最小二乘线性回归的基本假定就是一句话: 观测值独立同正态分布. 或者用公式表示为: $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, $\quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

注意

1. 所有上述假定都无法核对.
2. 把因变量和自变量的关系假定为线性关系的主要原因是当年人们的能力只能勉强应对线性模型所带来的数学及计算等一些问题.
3. 把样本假定为独立同正态分布，这大多不合乎实际，但基于与上一款同样的理由，这种假定在数学上是比较方便的. 这种分布假定和最小二乘回归方法本身无关，仅仅和与回归相关的各种推断有关，比如 t 检验及 F 检验统计量的分布、各种估计量的分布及性质等等.

没有假定则无法做任何经典统计推断 (诸如假设检验等)，只有基于计算机算法的机器学习出现之后，才有可能摆脱主观假定的束缚.