社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年10月22日

CONTENTS

CONTENTS

Multiple Linear regression 多元线性回归模型可以表示为：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

1. Y_i表示第i名个体在因变量Y上的取值
2. X_i表示第i名个体在自变量X上的取值
3. X_{ik}中的第二个下标k表示第k个自变量（k=1,...K）
4. β_0、β_1、β_2...β_k是模型的待估参数，称为回归系数（regression coefficients）
5. ϵ_i是随机误差项

回归系数直接反映了自变量对因变量的影响。

线性回归系数的基本含义是，**在控制其他自变量不变的条件下**，某个自变量每变化一个单位，导致因变量变化的平均值。

1 Y=a+bx，b具有边际变化率的意义

2 lnY=a+b lnx，b具有弹性的意义

1. 自变量为连续型变量

   reg lninc educ_y

   解释：教育年限每增加一年，将导致对数收入增加β_1，也就是收入增加$e^{\beta_1}-1$。

2. 自变量为二分类变量
3. 自变量为多分类变量

   xi: reg lninc i.edu

1. 自变量为连续型：在控制其他因素下，自变量每变化一个单位，因变量平均变化多少。
2. 自变量为二分类：自变量取分类“1”时，因变量的值平均比自变量分类取“0”时高多少。
3. 自变量为多分类：自变量取该分类时，因变量的值平均比**基准组**高多少。

We start by converting the religious denomination variable into a set of four dichotomous variables,one for each religious group, with each variable scored 1 for persons with that religion and scored 0 otherwise. （教材第116页）

1. R_1 if the respondent is Protestant, and =0 otherwise
2. R_2 if the respondent is Catholic, and =0 otherwise
3. R_3 if the respondent is Jewish, and =0 otherwise
4. R_4 if the respondent has another religion, no religion, or failed to respond, and =0 otherwise

如果变量为n类的分类变量，那么需要将其拆分为n-1个哑变量，并将其中一组视为参照组。

1 性别（女性=1，男性=2），重新编码为性别（男性=1，其余=0），以female为参照组。recode sex 1=0 2=1

2 宗教（Protestant=1，Catholic=2，Jewish=3，no religion=4），拆分为：

(1) Protestant（Protestant=1，其余=0）
(2) Catholic（Catholic=1，其余=0）
(3) Jewish（Jewish=1，其余=0）
(4) 以no religion为参照组。

The coefficients of the dummy variables included in the OLS equation are interpreted as deviations from the value for the *omitted*, or *reference*, category.

对于“定性数据”(qualitative data)或“分类数据”(categorical data)，需引入“哑变量”，即取值为0或1的变量。

比如，性别分男女，可定义 $D=\begin{cases}1, & \text{男}\\ 0, & \text{女}\end{cases}$。

对于全球的五大洲，则需要四个哑变量，即

$$D_1=\begin{cases}1, & \text{Asia}\\ 0, & \text{other}\end{cases},\quad D_2=\begin{cases}1, & \text{America}\\ 0, & \text{other}\end{cases},\quad D_3=\begin{cases}1, & \text{Europe}\\ 0, & \text{other}\end{cases}$$

$$D_4=\begin{cases}1, & \text{Africa}\\ 0, & \text{other}\end{cases}$$

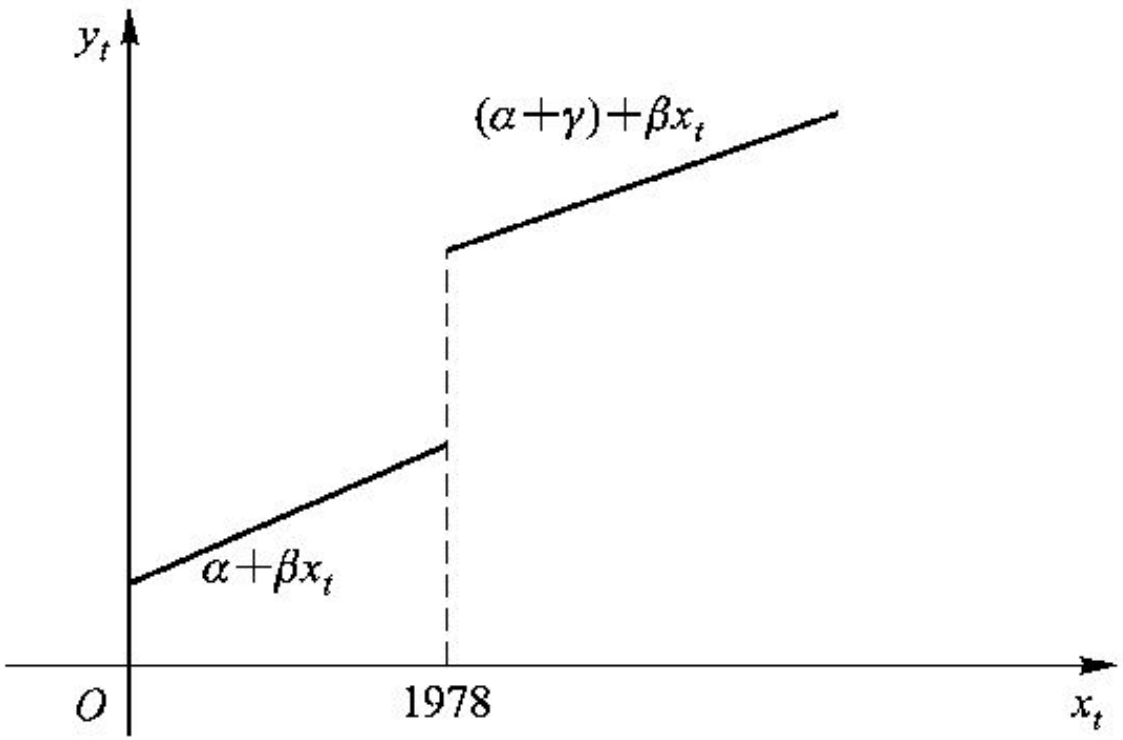
如果D_1=D_2=D_3=D_4=0，则表明为大洋洲。

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t$$

引入哑变量D_t，该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若} t < 1978, D_t=0 \\ (\alpha + \gamma) + \beta x_t + \varepsilon_t, & \text{若 } t \geq 1978, D_t=1 \end{cases}$$

仅引入哑变量，相当于在不同时期使用不同的截距项。



仅引入哑变量的效果

OLS估计的统计性质

1. 残差平方和最小。
2. 扰动项ϵ条件独立于自变量X，意味着ϵ与自变量X不相关。
   - $E(\epsilon|x) = 0 \Rightarrow Cov(x_i, \epsilon) = 0$
3. 扰动项ϵ条件期望值为0，意味着ϵ的无条件期望值也为0。
   - $E(\epsilon|x) = 0 \Rightarrow E(\epsilon) = 0$
4. $\bar{\hat{y}_i} = \bar{y}$
5. $Cov(\hat{y}, \epsilon) = 0$
6. 直线通过$(\bar{y}, \bar{x})$点。

Linear regression assumptions:

1. 零条件均值
2. 独立同分布
3. 正态分布假定
4. 不存在多重共线性

LINEAR REGRESSION ASSUMPTIONS（零条件均值）

- A key assumptions that yields the identification of the unknown parameters is the **independence between ϵ_i and x variables.**

$$Cov(x_i, \epsilon) = 0$$

- **(Assumption)** Zero Conditional Mean: The error ϵ has an expected value of **zero** given any value of X.（零条件均值）

$$E(\epsilon|x) = 0$$

Zero Conditional Mean

残差项均值为0，意味着：

1. 模型没有遗漏任何重要自变量，也没有模型识别错误（specification error）（教材第102页）。
2. 若不满足，会导致内生性（endogeneity）问题，引起OLS的有偏估计（教材366-367页）。
   - 当自变量与残差项为正相关时，为高估。
   - 当自变量与残差项为负相关时，为低估。

LINEAR REGRESSION ASSUMPTIONS（独立同分布）

Other assumptions: ϵ_i is independent of one another and identically distributed (i.i.d). （独立同分布假定）

1. The independence assumption implies that the correlation in ϵ between a pair of observations is zero.

$$Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$$

2. The identical distribution assumption assures a common variance of σ_ϵ^2 (i.e. homoscedasticity).

$$var(\epsilon|x) = \sigma^2$$

LINEAR REGRESSION ASSUMPTIONS（正态分布假定）

- 正态分布假定：残差项（扰动项）服从正态分布。

$$\epsilon_i \sim N(0, \sigma^2)$$

- 因变量背后的决定机制：
  扰动项正态分布 $\longrightarrow$ 因变量正态分布

Multiple Linear regression assumptions

No perfect collinearity

- There are no exact linear relationships among the independent variables.
- 共线性就是同语反复。

Example 4.1

小学生的年龄与上学年数

用样本所计算的统计量作为总体参数的点估计时，通常有三个评价标准：

1. 无偏性(unbiased)：估计值的平均数恰好等于总体的参数值。
2. 有效性（efficient）：抽样分布方差最小的那个估计就是对总体参数最有效率的估计。
3. 一致性（consistent）：随着样本规模扩大，估计的偏差随之减小，或者说估计值以更大的概率趋近总体参数值。

1. $\hat{\beta}$ ARE UNBIASED AND CONSISTENT（若满足零均值假定）

Zero Conditional Mean: The error ϵ has an expected value of **zero** given any value of X. （自变量与扰动项无关）

$$E(\epsilon|x) = 0$$

$$E(\epsilon|x_1, x_2, \ldots, x_k) = 0$$

- Under this assumption, $\hat{\beta}_0$ is unbiased for β_0, $\hat{\beta}_1$ is unbiased for β_1.

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1$$

无偏的(unbiased)：估计值的平均数恰好等于总体的参数值。

1. $\hat{\beta}$ ARE UNBIASED AND CONSISTENT（若满足零均值假定）

Zero Conditional Mean: The error ϵ has an expected value of **zero** given any value of X. （自变量与扰动项无关）

$$E(\epsilon|x) = 0$$

$$E(\epsilon|x_1, x_2, \ldots, x_k) = 0$$

- Under this assumption, $\hat{\beta}_0$ is unbiased for β_0, $\hat{\beta}_1$ is unbiased for β_1.

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1$$

无偏的(unbiased)：估计值的平均数恰好等于总体的参数值。

2. $\hat{\beta}$ ARE EFFICIENT（若满足同分布假定）

(Homoskedasticity Assumption) The error ϵ has the same variance given any value of X.

$$var(\epsilon|x) = \sigma^2$$

$$var(\epsilon|x_1, x_2, \ldots, x_k) = \sigma^2$$

- 有效性（efficient）：抽样分布方差最小的那个估计就是对总体参数最有效率的估计。
- V($\hat{\beta}$) is the lowest possible among all other estimators, thus has the most **precision** among all estimators.

BLUE（Gauss-Markov Theorem）

1. Best（efficient）: the sampling distributions of b_0 and b_1 having the smallest variance.
2. Linear: $\hat{\beta}$ can be expressed as a linear function of the data on the dependent variable. 估计值可以表示为因变量的函数。
3. Unbiased: $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$
4. Estimator: the estimated b_0 and b_1 tend to β_0 and β_1 as n tends to infinite.

参考文献

1 郭志刚，2015，《社会统计分析方法——SPSS软件应用（第二版）》，北京: 中国人民大学出版社。
2 邱嘉平，2020，《因果推断实用计量方法》，上海：上海财经大学出版社。
3 谢宇，2013，《回归分析》（修订版），北京：社会科学文献出版社。