计量经济学

多元回归模型

张晨峰

2016年5月11日

华东理工大学商学院

3. 多元回归模型

主要内容

- 多元回归分析：估计
- 多元回归分析：推断
- 多元回归分析：OLS的渐近性
- 多元回归分析：虚拟变量
- 多元回归分析：异方差

3.1 多元回归分析：估计

多元线性回归模型

- 使用多元回归的动因
- 多元线性回归模型形式：$y = X\beta + \mu$
- 偏回归系数的含义

多元回归模型的最小二乘法估计量

$$Min_{\hat{\beta}} S(\hat{\beta}) = \hat{\mu}'\hat{\mu} = (y - X\hat{\beta})'(y - X\hat{\beta})$$

化简得到

$$S(\hat{\beta}) = y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

最小化的必要条件是

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

于是，得到

$$\hat{\beta} = (X'X)^{-1}X'y$$

多元回归模型的最小二乘法估计量的另一种表达

考虑如下回归模型：$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$，一种表示$\hat{\beta}_1$的方式为：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r_{i1}} y_i}{\sum_{i=1}^{n} \hat{r_{i1}^2}}$$

其中$\hat{r_{i1}}$是利用现有样本将x_1对x_2进行简单回归而得到的OLS残差。

遗漏变量偏误

模型一：

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

模型二：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

如果模型二是正确的模型，则

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

工资方程

一个人的工资水平与他的可测教育水平及其他非观测因素的关系为

$$wage = \beta_0 + \beta_1 educ + \mu$$

如果遗漏了重要变量——能力*(ability)*，且能力对收入有正影响，与教育水平正相关，那么偏误为正还是为负呢？

3.1 多元回归分析：估计

对函数形式的进一步讨论

- 对数形式
- 含二次项的模型
- 含交互作用项的模型

对拟合优度和回归元选择的讨论

- 调整的拟合优度
- 回归元的选择

3.1 多元回归分析：估计

多元线性回归（MLR）的假定

- MLR.1 线性于参数
- MLR.2 随机抽样
- MLR.3 不存在完全共线性
- MLR.4 零条件均值 $E(\mu|x_1, x_2, ...x_k) = 0$
- MLR.5 同方差性 $Var(\mu|x_1, x_2, ...x_k) = \sigma^2$

OLS的无偏性

在假定MLR.1至MLR.4下，下式对总体参数对β_j的任意值都成立，

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, ..., k$$

即OLS估计量是总体参数的无偏估计量。

OLS斜率估计量的抽样方差

在假定MLR.1至MLR.5下，以自变量的样本值为条件，对所有的$j = 1, 2, ..., k$，都有

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

其中，$SST_j = \sum_{i=1}^{n}(x_i - \bar{x})^2$是$x_j$的总样本变异，而$R_j^2$则是将$x_j$对所有其他自变量（并包括一个截距项）进行回归得到的R^2。

σ^2的无偏估计

在高斯-马尔可夫假定MLR.1至MLR.5下，有

$$E(\hat{\sigma^2}) = \sigma^2$$

高斯-马尔可夫定理

在假定MLR.1至MLR.5下，$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$分别是$\beta_0, \beta_1, ..., \beta_k$的最优线性无偏估计量。

3.2 多元回归分析：推断

多元线性回归（MLR）的假定

- MLR.6 正态性：总体误差μ独立于解释变量，且服从均值为零和方差为σ^2 的正态分布，即$\mu \sim Normal(0, \sigma^2)$

正态抽样分布

在CLM假定MLR.1至MLR.6下，以自变量的样本值为条件，有

$$\hat{\beta}_j \sim Normal[\beta_j, Var(\hat{\beta}_j)]$$

因此，

$$\frac{(\hat{\beta}_j - \beta_j)}{sd(\hat{\beta}_j)} \sim Normal(0, 1)$$

标准化估计量的t分布

在CLM假定MLR.1至MLR.6下，

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

其中，$k+1$是总体模型中未知参数的个数。

小时工资方程

使用 WAGE1. RAW 中的数据得到如下估计方程

$$\widehat{\log(wage)} = 0.284 + 0.092educ + 0.004\,1exper + 0.022tenure$$
$$(0.104)\quad (0.007)\quad (0.001\,7)\quad (0.003)$$
$$n = 526,\ R^2 = 0.316$$

t统计量的另一个例子

一般的t统计量可以写成(估计值-假设值)/标准误。

住房价格和空气污染

对于一个由波士顿地区506个社区组成的样本，我们估计了一个联系社区中平均住房价格（*price*）与各种社区特征的模型：*nox*表示空气中氧化亚氮的含量，以每社区的百万分子数度量；*dist*表示该社区相距五个商业中心的加权距离，以英里为单位；*rooms*表示该社区平均每套住房的房间数；而*stratio*则表示该社区学校的平均学生—教师比。总体模型是

$$\log(price)=\beta_0+\beta_1\log(nox)+\beta_2\log(dist)+\beta_3 rooms+\beta_4 stratio+u$$

其中，β_1是*price*对*nox*的弹性。我们希望针对对立假设H_1：$\beta_1\neq-1$来检验H_0：$\beta_1=-1$。做这个检验的t统计量是$t=(\hat{\beta}_1+1)/\mathrm{se}(\hat{\beta}_1)$。

利用HPRICE2.RAW中的数据，估计模型是

$$\widehat{\log(price)}=11.08-0.954\log(nox)-0.134\log(dist)+0.255rooms-0.052stratio$$
$$(0.32)\quad(0.117)\qquad(0.043)\qquad(0.019)\qquad(0.006)$$

$n=506$，$R^2=0.581$

检验关于参数的一个线性组合假设

可以采用的方式是估计一个能直接给出我们所需标准误的不同模型。

$$\log(wage)=\beta_0+\beta_1 jc+\beta_2 univ+\beta_3 exper+u \tag{4.17}$$

其中，jc 是就读两年制大学的年数，而 $univ$ 是就读四年制大学的年数，$exper$ 是参加工作的月数。注意，大专和大学的任意组合都是允许的，包括 $jc=0$ 和 $univ=0$。

对多个线性约束的检验：F检验

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)}$$

举一个例子，考虑如下方程：

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u \quad (4.47)$$

其中，*price* 为住房价格，*assess* 为住房的评估价值（在房屋售出以前），*lotsize* 为以英尺为单位的占地面积，*sqrft* 为平方英尺数，而 *bdrms* 则为卧室数。现在，假设我

3.3 多元回归分析：OLS的渐近性

多元线性回归（MLR）的假定

- MLR.4′ 零均值和零相关：对所有的$j = 0, 1, ..., k$，都有$E(u) = 0$和$Cov(x_j, u) = 0$。

OLS的一致性

在假定MLR.1至MLR.4下，对所有的$j = 0, 1, ..., k$，OLS估计量$\hat{\beta}_j$都是β_j的一致估计。

OLS估计量的渐近正态性

在大样本容量的情况下，OLS估计量是近似正态分布的。

3.4 多元回归分析：虚拟变量

对定性信息的描述

在计量经济学中，二值变量最常见的称呼是虚拟变量。

$$\widehat{wage} = -1.57 - 1.81\,female + 0.572\,educ + 0.025\,exper + 0.141\,tenure$$
$$(0.72)\quad (0.26)\qquad (0.049)\qquad (0.012)\qquad (0.021)$$
$$n = 526,\ R^2 = 0.364$$

$$\widehat{wage} = 7.10 - 2.51\,female$$
$$(0.21)\,(0.30)$$
$$n = 526,\ R^2 = 0.116$$

3.5 多元回归分析：异方差

异方差性对OLS所造成的影响

异方差性并不会导致OLS估计量出现偏误或产生不一致性。但在出现异方差性的情况下，我们在高斯-马尔可夫假定下用来检验假设的统计量都不再成立。

OLS估计后的异方差-稳健推断

在一般多元回归模型中

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \mu$$

那么$Var(\hat{\beta}_j)$的一个有效估计量是

$$Var(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{\mu}_i^2}{SSR_j^2}$$

稳健标准误和稳健t统计量只有在样本容量越来越大时才能使用。

异方差性的怀特检验

同方差假定$Var(\mu_1|x_1, x_2, ..., x_k) = \sigma^2$可由如下较弱的假定所取代，即误差平方$\mu^2$与所有自变量，所有自变量的平方和所有自变量的交叉乘积都不相关。这促使怀特提出对异方差性的一种检验方法。

当模型包含$k = 3$个自变量时，怀特检验则基于如下估计：

$$\hat{\mu}^2 = \delta_0+\delta_1 x_1+\delta_2 x_2+\delta_3 x_3+\delta_4 x_1^2+\delta_5 x_2^2+\delta_6 x_3^2+\delta_7 x_1 x_2+\delta_8 x_1 x_3+\delta_9 x_2 x_3+e$$