第一部分 横截面数据的回归分析

第二章 线性回归模型的基本概念

2．1．线性回归模型

回归分析是一种重要金融计量工具

•可看作一个动态变化的系统，有些基本的要素在驱动这个系统，需要分析的变量与其他变量之间存在某种内在联系的系统

•寻找潜在的未知函数关系

•大部分的情况下只需要考虑线性关系

•一般的回归模型可以表示为方程

$$y = \beta_o + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

•通常称为线性回归模型。y称为因变量或被解释变量、响应变量、被预测变量或回归变量；$x_1, \cdots, x_k$ 称为自变量、解释变量、控制变量、预测变量或回归因子。

•变量ε 称为误差或干扰项。

•在上面的线性回归模型中，本质上是把对y有影响的所有因素分解为两个部分

- $x_1, \cdots, x_k$ 是可以被控制的，被观测的，或感兴趣的部分
- 把对y有影响的除了之外的因素都看做是不可以被观测到的。因此人们通常也把ε 看做是不可观测的部分。
- 把包含在ε 中的所有其它因素看做是固定的
- 则称 x_1 对y有线性影响，记为 $\Delta y = \beta_1 \Delta x$

- β_j 可以看做是其它因素不变时，x_j 对y的影响
- 模型设定是否合适，看这一分解或分割是否合适

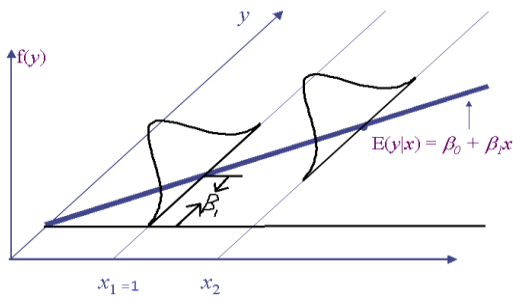
$$y = \underbrace{\beta_o + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{可观测部分}} + \underbrace{\varepsilon}_{\text{不可观测的干扰}}$$

- 需要一个假设限制ε 与自变量的关系
- 度量ε 和任意一个自变量之间关系的最自然的度量就是它们之间的相关系数
- 相关系数只能限制它们之间的线性关系
- 比较合适的假设需要条件期望

$$E(\varepsilon \mid x_{1,\cdots,} x_k) = E(\varepsilon) = 0$$

- 学习成绩与工作起薪之间关系的例子中，这一假设表示各位同学在入学时的起点是一样的，各位的学习能力是相同的
- 由于不能观测到人们的学习能力，也就无法知道平均的学习能力是否与学习结果有关
- 只有一个自变量时，这一假设条件给出β_1 的直观表示 $E(y \mid x_1) = \beta_0 + \beta_1 x_1$
- 对任一个给定的x_1值，y是以$E(y \mid x_1)$为中心分布
- 把y分解为两个部分：
  - $\beta_0 + \beta_1 x_1$ 称为系统部分
  - ε 不可以解释的部分

对给定的 x，y分布在其中心 E(*y*|*x*)的附近 E(*y*|*x*) 作为 x的线性函数，



2．2．线性回归模型的最小二乘估计

- 如何用样本得到多元线性模型的最小二乘估计
- 记 $\{(x_{i1}, x_{i2}, \cdots x_{ik}, y_i), i = 1, \cdots, n\}$
- 为收集到的n组观测值
- 对每一个i应该满足

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- 已知各解释变量和干扰项的关系

$$E(\varepsilon) = 0 \qquad Cov(x_j, \varepsilon) = E(x_j \cdot \varepsilon) = 0, i = 1, \cdots, k$$

- 得到了因变量y和 自变量x的观测值代入得

$$\begin{cases} \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})=0 \\ \sum_{i=1}^{n}x_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})=0 \\ \cdots \\ \sum_{i=1}^{n}x_{ik}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})=0 \end{cases}$$

- 方程为一阶条件
- 求解上面的方程组就得到

$$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$$

7

- 利用随机变量均值的基本性质 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$
- 当k=1时，即为 $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_1$
- 代入另一个方程为

$$\sum_{i=1}^{n}x_{i1}(y_i-(\bar{y}-\hat{\beta}_1\bar{x}_1)-\hat{\beta}_1x_{i1})=0$$

$$\sum_{i=1}^{n}x_{ii}(y_i-\bar{y})=\hat{\beta}_1\sum_{i=1}^{n}x_{i1}(x_{i1}-\bar{x}_1)=\hat{\beta}_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i1}-\bar{x}_1)$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

8

- 除了前面的假设之外，得到这一估计只需要解释变量的方差大于零
- 代入模型参数的估计值而给出残差项的估计

$$\hat{\varepsilon}_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}$$

- 当k>0时，采用向量和矩阵的记号

$$\vec{y}=(y_1,\cdots,y_n)' \qquad \vec{\beta}=(\beta_0,\beta_1,\cdots,\beta_k)'$$

$$\vec{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)'$$

$$X=\begin{pmatrix}1 & x_1 & \cdots & x_k\end{pmatrix}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1k}\\ \cdots & \cdots & \cdots & \cdots\\ 1 & x_{n1} & \cdots & x_{nk}\end{pmatrix}$$

9

- 向量形式的回归模型为 $\vec{y}=X\hat{\vec{\beta}}+\vec{\varepsilon}$
- 一阶条件方程可以表示为

$$X'(\vec{y}-X\hat{\vec{\beta}})=0$$

- 假定矩阵X'X是非退化的，它存在逆矩阵 $(X'X)^{-1}$ 则有OLS系数估计的一般表示式

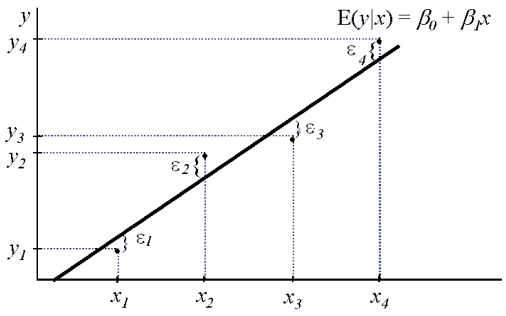
$$\hat{\vec{\beta}}=(X'X)^{-1}X'\vec{y}$$

- 残差序列的估计为

$$\hat{\varepsilon}_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik},i=1,\cdots,n$$

10

模型拟合的回归线和残差



- 如果要求所得到的估计的残差平方和尽可能小

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})^2$$

对各未知参数求一阶导数并令其等于零，可以得到前面的方程组，称为OLS估计的一阶条件

最小二乘估计的名称就来源于上面的估计方法

也可以考虑最小化残差的其它函数形式

最小化绝对值和

其它的函数形式

12

- 模型的拟合回归线为
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$
- 真实的回归方程为
$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
- 真实的回归方程是固定的，未知的。
- 估计的回归方程是由样本决定。
- 一组新的样本可能会得到另外的一个回归方程
- 当k=1时，我们很容易得到 $\hat{\beta}_1 = \Delta\hat{y} / \Delta x_1$
- 当k>1时，$\hat{\beta}_j$称为有偏效应（partial effect）或固定其它条件相同时的影响

13

- y的改变量表示为 $\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \cdots + \hat{\beta}_k \Delta x_k$
- 只有 x_j 改变，其它条件不变（其他变量不改变）
$$\Delta x_1 = \Delta x_{j-1} = \Delta x_{j+1} = \cdots = \Delta x_k = 0$$
$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j$$
- 回归模型最重要的作用就是在考虑一个具体自变量对因变量的影响时，包含了其它变量在模型中
- 只有包含了其它变量在模型中，我们才能得到所关注系数的正确估计

14

- 例1 CEO收入和公司股本收益
- y为以万元为单位的年收入，x表示过去一年该CEO所在公司的年平均股本收益（ROE）
$$salary = \beta_0 + \beta_1 roe + \varepsilon$$
- 采用中国1055家上市公司的2007年相关数据使用OLS估计得到估计的模型为
$$\hat{salary} = 48.765(\Delta roe)$$
- 股本收益每提高一个百分点，CEO的年收入预期增加487,650元

15

- 例2 小时工资和教育程度
- 使用526个美国工人的数据，educ表示受教育的年限，exper表示工龄，tenure表示受聘于目前公司的年数，wage表示每小时工资。代入数据估计得到
$$\widehat{\log(wage)} = 0.284 + 0.092 educ + 0.0041 \exp er + 0.022 tenne$$
- 如果只考虑受教育的年限，样本中的平均小时工资为5.9美元，估计的方程为
$$\widehat{\log(wage)} = 0.584 + 0.083 educ$$

回归模型中，教育年限的系数0.092是在控制了工龄和目前工作年数后多接受一年教育所带来的工资增加

16

2．3．最小二乘估计的结构

- 2．3．1 拟合值和残差
- 得到了参数的估计值后，对每一组自变量观测值可以得到一个拟合值 $\hat{y}_i$在模型的OLS回归线上
- $\hat{\varepsilon}_i$表示的是实际观测值与拟合值之间的差异
- OLS的拟合值与残差之间具有一些重要特征
  - 残差的平均值为0
  - 每一个自变量与残差的样本协方差为0，OLS拟合值与残差的样本协方差也为0
  - 点 $(\bar{x}_1, \cdots, \bar{x}_k, \bar{y})$ 总是在回归线上

17

- 把每一个y_i表示为拟合值与残差之和 $y_i = \hat{y}_i + \hat{\varepsilon}_i$
- 由第一个特征，有 $\bar{\hat{y}} = \bar{y}$
- 定义因变量的总平方和（SST）、可解释部分平方和（SSE）和残差平方和（SSR）分别为
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$SSR = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$$
- y的总变化也可以分解为两个部分，有
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSR + 2\sum_{i=1}^{n} \hat{\varepsilon}_i (\hat{y}_i - \bar{y}) + SSE$$
- 分解式成立，只需要 $\sum_{i=1}^{n} \hat{\varepsilon}_i (\hat{y}_i - \bar{y}) = 0$

18

2.3.2 模型的拟合度

- 如何度量线性回归模型中自变量解释能力
- 总平方和分解式除以SST得1=SSE/SST+SSR/SST
- 线性回归模型的决定系数定义为

$$R^2 = SSE/SST = 1 - SSR/SST$$

- 表示了y的样本方差中可以被x解释的部分
- 也可表示为观测值与拟合值间相关系数的平方

$$R^2 = [\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2 / [\sum_{i=1}^n (y_i - \bar{y})^2 \cdot \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2]$$

- R^2这一说法的来源

$$\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})((\hat{y}_i - y_i + y_i - \bar{\hat{y}})$$

$$= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(\hat{y}_i - y_i) + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{\hat{y}})$$

$$= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})\hat{\varepsilon}_i + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{\hat{y}})$$

$$= \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})$$

- 根据定义有 $R^2 = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 / \sum_{i=1}^n (y_i - \bar{y})^2$

$$= \frac{[\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} = \frac{[\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}$$

- 实证研究中回归的R方很小是常见的
- R方小并不是说OLS回归方程没有意义
- CEO的年薪与ROE之间的关系是否合适并不完全依赖于R方
- 随着自变量数目的增加，回归的R方也增加
- 要确定一个变量是否应该加入模型要看解释变量对y的偏效应是否为0
- 在CEO年薪与股本收益的例子中，R^2=0.083

2.4 “控制其它因素不变”的意义

- 多元回归模型可以对不是在同等条件下收集到的数据给出其它条件相同时应有的分析结果
- 2．4．1 控制其他因素不变
- 控制其他因素不变实际上是一种“剥离”方法
- 需要直接给出 $\hat{\beta}_j$ 的公式表达式
- 考虑k=2的情形，估计模型为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- 假若我们只对 $\hat{\beta}_1$ 感兴趣，可以把它表示为

$$\hat{\beta}_1 = (\sum_{i=1}^n \hat{r}_{i1} y_i) / (\sum_{i=1}^n \hat{r}_{i1}^2)$$

$\hat{r}_{i1}$ 为把第一个自变量作为因变量用第二个自变量来进行回归而得到的残差

- 把 x_1 表示为拟合值与回归残差之和 $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$
- 代入一阶条件(2．6)的第二个方程得

$$\sum_{i=1}^n (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})$$

$$= \sum_{i=1}^n \hat{x}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) +$$

$$\sum_{i=1}^n \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

- 第一项满足OLS估计的条件,为零，所以

$$\sum_{i=1}^n \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

- $\hat{r}_{i1}$是回归模型 $x_1 = \hat{\delta}_0 + \hat{\delta}_1 x_2 + \hat{r}_1$ 的残差

$$\sum_{i=1}^n \hat{r}_{i1}\hat{\beta}_0 = 0 \quad \sum_{i=1}^n \hat{r}_{i1}\hat{\beta}_2 x_{i2} = 0$$

- 上式等价于 $\sum_{i=1}^n \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$
- 再代入 $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ 得到

$$\sum_{i=1}^n \hat{r}_{i1}[y_i - \hat{\beta}_1(\hat{x}_{i1} + \hat{r}_{i1})] = 0$$

- 拟合值与残差的协方差为零

$$\sum_{i=1}^n \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0$$

$$\hat{\beta}_1 = (\sum_{i=1}^n \hat{r}_{i1} y_i) / (\sum_{i=1}^n \hat{r}_{i1}^2)$$

- $\hat{\beta}_1$估计表达方式就相当于用得到的残差 $\hat{r}_1$ 对y进行回归而得到的系数估计，偏效应的一种表达
- $\hat{r}_{i1}$是 x_{i1} 中通过“剥离”了 x_{i2} 影响之后剩余的效果
- k>2的一般情形，$\hat{\beta}_1$ 仍然可以写成上面的形式，只是残差项 $\hat{r}_1$ 为使用 $x_2,\cdots,x_k$ 对 x_1 进行回归后得到的残差
- $\hat{\beta}_1$ 刻划了在“剥离” $x_2,\cdots,x_k$ 之后 x_1 对y的影响

25

- 2．4．2 一元回归与多元回归估计的比较
- 什么情况下，给出相同的估计？
- 模型分别为 $\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1 x_1$ $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2$
- 可得到它们之间存在如下关系 $\tilde{\beta}_1=\hat{\beta}_1+\hat{\beta}_2\tilde{\delta}_1$
- $\tilde{\delta}_1$ 为用 x_1 对 x_2 进行回归的系数
- 只有在两种情况下才能相等
  - x_2 对y的偏效应为0，即 $\hat{\beta}_2=0$
  - x_1 与 x_2 是无关的，即 $\tilde{\delta}_1=0$
- 多个自变量时也有类似的结果

26

2．5．度量单位和函数形式

- 自变量或因变量的度量单位改变对OLS估计的影响
- CEO的例子中，采用万元为单位来表示年薪，而ROE的单位采用百分比，采用元为单位来表示年薪

$$\widehat{salardol}=763{,}490+487{,}650\,roe$$

- 改变因变量，模型估计系数都被放大了10000倍
- 一般地，我们可以得到，当因变量被乘以一个常数c时，则意味着OLS估计的系数都要乘以c
- 自变量被除以（或乘以）一个非零常数c，则它对应的系数将分别要乘以（或除以）常数c ，但其它变量对应的系数不变

27

- 通过对变量使用函数变换，可在线性回归模型中处理变量之间的非线性关系
- 教育程度与收入水平的例子，假定再增加一年教育工资的增加是按百分比增加，百分比增加与最初的收入水平有关
- 满足具有相同的增长比率

$$\log(wage)=\beta_0+\beta_1 educ+\varepsilon$$

$$\Delta wage\approx\beta_1\Delta educ$$

- 把模型表示为 $wage=\exp(\beta_0+\beta_1 educ+\varepsilon)$

28

- 例如，log工资与教育，代入数据估计得到

$$\widehat{\log(wage)}=0.584+0.083\,educ,\quad n=526,R^2=0.186$$

- 例 CEO年薪和公司的销售收入，把CEO的年薪认为与公司的销售收入间具有固定的弹性

$$\log(sal\,\text{arg})=\beta_0+\beta_1\log(sales)+\varepsilon$$

- 采用OLS估计得到

$$\widehat{\log(sal\,\text{arg})}=11.442+0.0923\log(sales)$$

$$n=1055,R^2=0.082$$

- 公司的销售收入每增加一个百分点，CEO的年薪将大约增加0.0923个百分点

29

- 因变量取对数后是近似比率的改变，所以对斜率不再有影响
- 对每个 y_i 乘以c，原方程为 $\log(y_i)=\beta_0+\beta_1 x_i+\varepsilon_i$
- 两边都加log(c)得

$$\log(c)+\log(y_i)=[\log(c)+\beta_0]+\beta_1 x_i+\varepsilon_i=\log(cy_i)$$

- 因此斜率系数没有改变 ,截距项变为 $\log(c)+\beta_0$
- 类似地如果在取对数的自变量log(x)改变单位，其斜率系数仍然保持不变，只会改变截距项
- 变量进入模型是直接取值时，称为是水平变量
- 因变量y和自变量x都直接取值的模型被称为水平-水平模型

30

- 线性回归模型中因变量和自变量的形式

模型	因变量	自变量	β_1的表示
水平-水平	y	x	$\Delta y = \beta_1 \Delta x$
水平-对数	y	log(x)	$\Delta y = (\beta_1/100)(\Delta x\%)$
对数-对数	log(y)	log(x)	$\Delta y\% = \beta_1(\Delta x\%)$
对数-水平	log(y)	x	$\Delta y\% = (100\beta_1)\Delta x$

- x和y表示变量最初的形式，称为水平-水平模型
- 回归模型中线性的意思是指回归方程中的估计参数 $\beta_0, \beta_1, \cdots, \beta_k$ 是线性关系，而对y和 $x_1, \cdots, x_k$ 的取值并没有作限制

31

2．6．最小二乘估计的无偏性

- 2．6．1 OLS估计的无偏性(unbiasedness)
- 考虑OLS估计的性质需要对模型所属于的总体有一些假设条件
- 假设1：（参数是线性的）因变量与自变量之间的关系是线性的，与干扰项也是线性的。即模型为 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$
- 假设2：（随机抽样）我们所得到的样本 $\{(x_{i1}, x_{i2}, \cdots, x_{ik}, y_i), i = 1,2,\cdots,n\}$ 是从真实模型中随机抽取的

32

- 把回归模型表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, i = 1,2,\cdots,n$$

- 对CEO年薪的例子

$$\log(sal\arg_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + \varepsilon_i$$

- 假设3（条件期望为0）给定任何一组自变量的取值后，干扰项ε 的期望值为0。也即

$$E(\varepsilon \mid x_1, x_2, \cdots, x_k) = 0$$

- 对随机抽取的样本，这一假设表明对所有

$$i = 1,2,\cdots,n, E(\varepsilon_i \mid x_{i1}, x_{i2}, \cdots, x_{ik}) = 0$$

33

- 假设3不成立的一种情况是模型设定不合适
- 因变量与自变量之间的关系不完整
- 函数形式不当，在应该使用变量的对数时直接使用了变量的水平，模型给出的估计就会有偏
- 由于数据的局限而使某些与某个或几个自变量相关但没有包含在模型中也会使假设3不成立
- 还有其它可使ε 与解释变量存在相关性的问题
- 当假设3成立时，我们通常称为得到的解释变量都是外生的。如果因为某种原因而与ε 有关，我们就其称为内生的解释变量

34

- 假设4（不完全共线性）在样本中没有一个解释变量是常数，自变量之间没有完全的线性关系。任何一个自变量不能被其它的自变量线性表出
- 在单变量的情况，假设4等价于 $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$
- 利用 $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$
- 模型参数的估计表示为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{SST_x}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{SST_x}$$

$$= \frac{[\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i]}{SST_x}$$

- 所以有 $\hat{\beta}_1 = \beta_1 + (\frac{1}{SST_x}) \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i)$
- 由此可见，OLS的估计实际上是真实参数再加上干扰项的一个线性组合
- 在观测值已知的条件下，$\hat{\beta}_1$ 的不确定性完全由样本的干扰项决定

36

- **定理1（OLS的无偏性）在假设条件1-4之下，** $E(\hat{\beta}_j)=\beta_j, j=0,1,\cdots,k$ **对任何一个总体模型的参数，OLS估计是真实参数的无偏估计。**
- **当k=1时，由上面的表示式很容易验证**

$$E(\hat{\beta}_1)=\beta_1+E\left[\frac{1}{SST_x}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i)\right]$$

$$=\beta_1+\frac{1}{SST_x}\sum_{i=1}^{n}(x_i-\bar{x})E(\varepsilon_i)=\beta_1$$

$$\beta_0=\bar{y}-\hat{\beta}_1\bar{x}=\beta_0+\beta_1\bar{x}+\bar{\varepsilon}-\hat{\beta}_1\bar{x}=\beta_0+(\beta_1-\hat{\beta}_1)\bar{x}+\bar{\varepsilon}$$

$$E(\hat{\beta}_0)=\beta_0+E[(\beta_1-\hat{\beta}_1)\bar{x}]+E(\bar{\varepsilon})=\beta_0$$

37

- **无偏性只是把** $\hat{\beta}_j$ **看成一个随机变量，它所具有的样本分布性质**
- **无偏性不成立是4个假设中有至少一个不成立**
- **应用中我们需要对每一个假设进行考虑**
- **假设1要求自变量和因变量之间具有线性关系和一个可加的干扰项**
- **假设2不成立，横截面数据有时可能并不能真实地代表总体，可能对某一部分取样太多**
- **假设3成立的时候，OLS估计是无偏的，而当它不成立时，OLS估计通常是有偏的**

38

- **假设4并不是不允许自变量之间存在相关性，而只是限制不能完全相关**
- **自变量的不同非线性函数同时包含在一个回归模型中，就可能出现这一问题**
- **样本量比较小而选择的参数又偏多时，也很容易出现共线性**
- **2．6．2变量过多的影响**
- **回归模型中包含了过多的变量，意味着至少有一个自变量它对y的偏效应为0**
- **在模型中包含了一个系数为0的自变量之后会带来什么影响？**

39

- **假设模型满足假设1~4，设为**

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$$

- **在控制了** x_1 **和** x_2 **之后，** x_3 **对y没有偏效应**
- **进行估计得到了回归线**

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3$$

- **从无偏性来看，它没有什么影响**

$$E(\hat{\beta}_0)=\beta_0, E(\hat{\beta}_1)=\beta_1, E(\hat{\beta}_2)=\beta_2, E(\hat{\beta}_3)=0$$

- **是否意味着加入过多的无关变量对多元回归模型没有什么损失呢？答案是否定的**
- **加入无关变量将会影响OLS估计的方差**

40

2．6．3 缺失变量的偏差

- **确定缺失重要变量的偏离原因是模型设定不当分析的一个部分**
- **假定真实模型为** $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$
- **主要关注的是** x_1 **对y的偏效应** β_1
- **由于疏忽或数据没法得到 使用单变量得到**

$$\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1x_1$$

- **从单变量回归模型中得到的估计**

$$\tilde{\beta}_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$$

41

- **记** $SST_1=\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2$
- **上式的分子**

$$\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i=\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i)$$

$$=\beta_1\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2+\beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2}+\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)\varepsilon_i$$

$$=\beta_1\cdot SST_1+\beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2}+\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)\varepsilon_i$$

- **我们有** $E(\tilde{\beta}_1)=\beta_1+\beta_2\cdot\left[\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i2}}{SST_1}\right]$

方括号中是用 x_1 **对** x_2 **进行回归而得到的斜率**

$$\tilde{\tilde{x}}_2=\tilde{\delta}_0+\tilde{\delta}_1x_1$$

42

- 因为 $\tilde{\delta}_1$ 是一个已知量
- 估计的期望可以表示为 $E(\tilde{\beta}_1)=\beta_1+\beta_2\tilde{\delta}_1$
- 通常称偏差 $\beta_2\tilde{\delta}_1$ 为缺失变量的偏差
- 只有当 x_1 和 x_2 是样本无关的，$\tilde{\beta}_1$ 才是无偏的
- 估计偏差的方向

	Corr(x_1 ,x_2)>0	Corr(x_1,x_2)<0
β_2 >0	正偏差	负偏差
β_2 <0	负偏差	正偏差

43

- 偏差通常根据相关的理论和经验来判断
- 先天能力比较强的人有高的工资 β_2 >0
- 一元回归 $wage=\beta_0+\beta_1 educ+u$ 中OLS估计的系数平均而言是偏大
- 只是一组样本我们不能确定0.083一定大于 β_1

当 $E(\tilde{\beta}_1)>\beta_1$ 时，常称向上偏离（upward bias）

当 $E(\tilde{\beta}_1)<\beta_1$ 时，常称向下偏离（downward bias）

有时也称偏向零（biased towards zero）

44

- 真实的模型为 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$
- 满足假设1~4，去掉 x_3 而估计模型

$$\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1x_1+\tilde{\beta}_2x_2$$

若 x_2 与 x_3 不相关，但 x_1 与 x_3 相关；可能会根据前面的结论而认为 $\tilde{\beta}_1$ 是有偏的，而 $\tilde{\beta}_2$ 是无偏的

一般来说，$\tilde{\beta}_1$ 和 $\tilde{\beta}_2$ 都可能是有偏的

唯一的例外是当 x_2 与 x_1 也无关时

此时很难确定偏离的方向

实际中比较有效的处理方式是采用近似

45

- 假定 x_1 与 x_2 是不相关的，则可把 x_2 看成是在真实和简化模型中都不存在，来考察 $\tilde{\beta}_1$ 的偏差
- 可以得到

$$E(\tilde{\beta}_1)=\beta_1+\beta_3\cdot\left[\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)x_{i3}}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\right]$$

- 类似前面的方法可以讨论的 $\tilde{\beta}_1$ 和 $\tilde{\beta}_2$ 偏差
- 处理比较复杂的回归模型时，也通常是按照上面所示的方式进行

46

2．7．最小二乘估计的方差

- 除知道 $\hat{\beta}_1$ 的样本分布的中心 β_1（无偏估计）
- 还想知道，$\hat{\beta}_1$ 偏离的程度将会有多大
- 在假设1~4之下，可以得到OLS估计的方差
- 2.7.1 估计参数的方差
- 要使所得到的方差是合理估计还需要另一个假设
- 假设5 （同方差，Homoskedasticity），对任意一组样本都有

$$Var(\varepsilon_i\mid x_{i1},\cdots,x_{ik})=\sigma^2$$

47

- 假设5是为了使OLS估计的方差具有一定范围内的有效性（efficiency）
- 如果直接假定干扰项与自变量独立，则干扰项的分布与自变量无关

$$E(\varepsilon\mid x_1,\cdots,x_k)=E(\varepsilon)=0$$

$$Var(\varepsilon\mid x_1,\cdots,x_k)=Var(\varepsilon)=\sigma^2$$

- 因为 $Var(\varepsilon\mid x_1,\cdots,x_k)=Var(y\mid x_1,\cdots,x_k)$
- 因此当 $Var(y\mid x_1,\cdots,x_k)$ 是 $x_1,\cdots,x_k$ 的函数时，一定存在异方差

48

假设1~5也称为高斯-马尔可夫(Gauss-Markov)假设

- 定理2 （OLS估计的样本方差）在假设1~5之下，已知自变量的取值时，OLS估计的样本方差为：

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)} \qquad j = 1,2,\cdots,k$$

其中， $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x})^2$

R_j^2为用其它自变量对 x_j 进行回归而得到的R方

一元回归模型中，因为 $R_1^2 = 0$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \qquad Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

49

- 因为 $\hat{\beta}_1 = \beta_1 + \frac{1}{SST_x}\sum_{i=1}^{n}(x_i - \bar{x})^2 \varepsilon_i$
- 所以有

$$Var(\hat{\beta}_1) = (\frac{1}{SST_x})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 Var(\varepsilon_i) = \frac{\sigma^2}{SST_x}$$

- OLS估计的方差依赖于干扰项的方差
- 不可观测部分对因变量y的影响越大越难给出参数准确的估计
- 增加样本量也能增加自变量的总变差 SST_j
- 其它条件相同，通常倾向于选择变化幅度比较大或方差较大的作为自变量

50

- 如果 SST_j 太小就可能会使假设4不成立
- 影响OLS估计方差的另一个重要因素是 R_j^2
- 考虑k=2的情形 $Var(\hat{\beta}_j) = \frac{\sigma^2}{[SST_1(1-R_1^2)]}$

R_1^2 实际上是自变量 x_1 和 x_2 之间相关系数的平方

一般情况下，R_j^2 平方正好是自变量 x_j 的方差中可以被其它自变量解释的部分

当一个或几个自变量之间具有比较高的时候，称为多重共线性（multi-collinearity）

没有一个明确的意义和界限

51

- 例如 R_j^2 =0.9表明的样本方差有90%可以被回归模型中的其它自变量表示，但它是否会使太大而变得没有意义，还取决于 σ^2和 SST_j
- 收集更多的数据来降低无偏估计的方差
- 通过去掉一些自变量来克服多重共线性问题
- 如果去掉了真实模型中应该有的变量，又可能带来估计偏差
- 值得特别注意的是某个变量与其它变量之间存在比较高的相关性，可能并不影响我们需要估计的其它变量

52

- 例如，$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- 其中 X_3 和 X_2 高度相关，
- $Var(\hat{\beta}_2)$ 和 $Var(\hat{\beta}_3)$ 可能非常大，
- 但可能并不直接影响 $Var(\hat{\beta}_1)$
- 这一观点在实证分析中是非常重要的，
- 计量学家通常会引入许多控制变量来隔离一些特定变量对因果分析的影响
- 但这些变量之间的高度相关并不会影响我们确定分类

53

- 2．7．2 模型设定不当时的估计方差
- 自变量的添加涉及到估计的偏差和方差之间的权衡和选择
- 满足Gauss-Markov假设的真实模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- 缺少变量 x_2 的情况下，得到估计 $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

在 $\beta_2 \neq 0$和 x_1与 x_2有关时，估计 $\tilde{\beta}_1$是有偏的

估计的方差 $Var(\tilde{\beta}_1) = \sigma^2 / SST_1$

$$Var(\hat{\beta}_1) = \sigma^2 / [SST_1(1-R_i^2)]$$

54

假定x_1与 x_2是相关的，则我们可得到结论

当 $\beta_2 \neq 0$ 时， $\tilde{\beta}_1$是有偏的，$\hat{\beta}_1$ 是无偏的

$Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$ ？

当 $\beta_2 = 0$ 时， $\tilde{\beta}_1$ $\hat{\beta}_1$都是无偏的

$Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$

当 $\beta_2 \neq 0$ 时，需要权衡是带来的偏差还是增加的方差对模型影响更大

在样本量比较大时，我们倾向于 $\hat{\beta}_1$

x_2 没有包含在模型中时，干扰项的方差还包含 x_2 的部分，所以要更大

2．8．干扰项方差的估计

- 干扰项与残差的区别

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{\varepsilon}_i$$

- 误差是永远无法观测到的，而残差是可以用样本数据计算出来

$$\hat{\varepsilon}_i = \varepsilon_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_{i1} - \cdots - (\hat{\beta}_k - \beta_k)x_{ik}$$

首先 $\sigma^2 = E(\varepsilon^2)$ 它的一个无偏估计是 $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$

所以 $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = \bar{\varepsilon}_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x + \cdots + (\hat{\beta}_k - \beta_k)x_k$

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}[\hat{\varepsilon}_i + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_{i1} - \cdots - (\hat{\beta}_k - \beta_k)x_{ik}]^2$$

OLS估计的残差和为0

$$\hat{\varepsilon}_i = (\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_{i1} - \bar{x}_1) - \cdots - (\hat{\beta}_k - \beta_k)(x_{ik} - \bar{x}_k)$$

当k=1时， $\hat{\varepsilon}_i = (\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$

$$\hat{\varepsilon}_i^2 = (\varepsilon_i - \bar{\varepsilon})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(\varepsilon_i - \bar{\varepsilon})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$$

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})$$

- 第一项的期望值为 $(n-1)\sigma^2$

$$E[(\hat{\beta}_1 - \beta_1)^2] = Var(\hat{\beta}_1) = \sigma^2 / SST_x^2$$

$$\sum_{i=1}^{n}\hat{\varepsilon}_i(x_i - \bar{x}) = \sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\varepsilon})(x_i - \bar{x}) = 0$$

- 第三项可以表示为 $2(\hat{\beta}_1 - \beta_1)^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$
- 其期望为 $2\sigma^2$
- 因此有 $E(\sum_{i=1}^{n}\hat{\varepsilon}_i^2) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$
- 当k>1时，采用矩阵形式的记号

$$\hat{\Psi} = (\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)' \quad \Psi = (\varepsilon_1, \cdots, \varepsilon_n)'$$

$$M = I_n - X(X'X)^{-1}X'$$

$$Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = MY = M\Psi = \hat{\Psi}$$

$$\hat{\Psi}'\hat{\Psi} = \Psi'M'M\Psi = \Psi'M\Psi$$

- 因为 $\Psi'M\Psi$是一个数等于它的迹

$$E(\Psi'M\Psi \mid X) = E[tr(\Psi'M\Psi \mid X)] = E[tr(M\Psi\Psi' \mid X)]$$

$$= tr[E(M\Psi\Psi' \mid X)] = tr[ME(\Psi\Psi' \mid X)]$$

$$= tr(M\sigma^2 I_n) = \sigma^2 tr(M) = \sigma^2(n-k-1)$$

$$tr(M) = tr(I_n) - tr[X(X'X)^{-1}X']$$

$$= n - tr[(X'X)^{-1}X'X] = n - tr(I_{k+1}) = n - k - 1$$

- 定理3（ σ^2的无偏估计） 在Gauss-Markov假设1~5之下，干扰项方差σ^2的无偏估计是

$$E(\hat{\sigma}^2) = E[\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2] = E[\frac{SSR}{n-k-1}] = \sigma^2$$

- 标准差 $sd(\hat{\beta}_j) = \frac{\sigma}{[SST_j(1-R_j^2)]^{1/2}}$
- 标准误差 $se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[SST_j(1-R_j^2)]^{1/2}}$
- 当假设5不成立时，上面给出的OLS估计的标准差估计也就不是的一个合理的估计

2．9．OLS估计的有效性Gauss-Markov定理

- 一个无偏估计就是估计的期望值等于参数的真值。线性意味着一个估计当且仅当它可以表示为因变量的一个线性函数，即 $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$
- 定理4 （GM定理）在假设1~5之下，OLS估计 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ 是最优线性无偏估计（blue）
  - 定理告诉我们，当5个假设都成立的时候，不用再考虑其它线性无偏估计，OLS就是最好的
  - 当假设条件不成立时，OLS估计就不一定好，需要对它进行必要的调整

61

2．10． OLS估计的样本分布

- 严格统计推断希望知道 $\hat{\beta}_j$ 的样本分布
- OLS估计的样本分布完全依赖于干扰项的分布
- 假设6（正态性）模型的干扰项与 $x_1, x_2, \cdots, x_k$ 自变量独立且服从均值为0，方差为 σ^2 的正态分布：$\varepsilon \sim N(0, \sigma^2)$
- 在假设6之下，ε 与 x_j 独立，
- 自然要求假设3和5 $E(\varepsilon \mid x_1, \cdots, x_k) = E(\varepsilon) = 0$

$$Var(\varepsilon \mid x_1, \cdots, x_k) = Var(\varepsilon) = \sigma^2$$

62

- 假设1~6称为经典线性模型假设（classical linear model,CLM）
- CLM假设下，OLS有比GM假设更强的有效性
- 它是所有无偏估计中方差最小的估计
- 定理5（最优无偏估计（Best unbiased estimator ））在CLM假设下，OLS估计 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ 是所有无偏估计中方差最小的估计，称为最优无偏估计。
- 在中心极限定理起作用时是假定所有影响y的不可观测因素是可分的和可加的
- 当不可观测的因素是以比较复杂的函数形式作用于干扰项时，可能就有问题了

63

- 定理6（正态样本分布）在CLM假设1~6之下，在给定自变量取值的条件下，OLS的样本分布也是正态分布，$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$,
- 进一步有：$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim N(0,1)$

$$\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij} \varepsilon_i \qquad w_{ij} = \hat{r}_{ij} / SSR_j$$

- 在假设6条件下，干扰项是独立同分布的正态随机变量。因为独立正态分布随机变量的线性组合仍然是正态分布，因此 $\hat{\beta}_j$ 服从正态分布

64

总结

- 介绍了最常用的线性模型参数最小二乘（OLS）估计方法
- 讨论了最小二乘估计的结构，通过偏效应的剥离方式给出了多元回归模型参数的直观表示
- 介绍了模型的拟合度及改变自变量或因变量的度量单位对模型参数的影响，变量通过对数函数变换后所表达的意义
- 讨论了线性回归模型参数估计满足无偏性所需要假设条件：参数是线性的，随机抽样，条件期望为零和不完全共线性。
- 讨论了模型设定不当时，参数估计的偏差程度和方向
- 给出了最小二乘估计方差的结构和干扰项方差的估计方法
- 给出了OLS估计是最优线性无偏估计的GM条件
- OLS估计是最优无偏估计的经典线性模型假设条件
- 给出对模型OLS估计参数进行统计推断所需要的样本分布

65