大数据统计与计量分析II

西安交通大学管理学院
信息系统与电子商务系
刘笑笑

第三章 大数据统计与计量分析II
——横截面数据

西安交通大学管理学院
信息系统与电子商务系
刘笑笑

第三章 横截面数据

- §1.1 回归分析
- §1.2 回归分析：深入专题与探讨

§1.1 回归分析

- 一、回归分析基本理解
- 二、普通最小二乘法
- 三、OLS估计量的期望值
- 四、OLS估计量的方差
- 五、OLS的有效性：高斯-马尔科夫定理
- 六、小结

一、回归分析基本理解

- 简单回归模型

$$y = \beta_0 + \beta_1 x + u$$

➢有n个观察值的随机样本

(x_1, y_1) ← 第一个观察值

(x_2, y_2) ← 第二个观察值

(x_3, y_3) ← 第三个观察值

⋮

(x_n, y_n) ← 第n个观察值

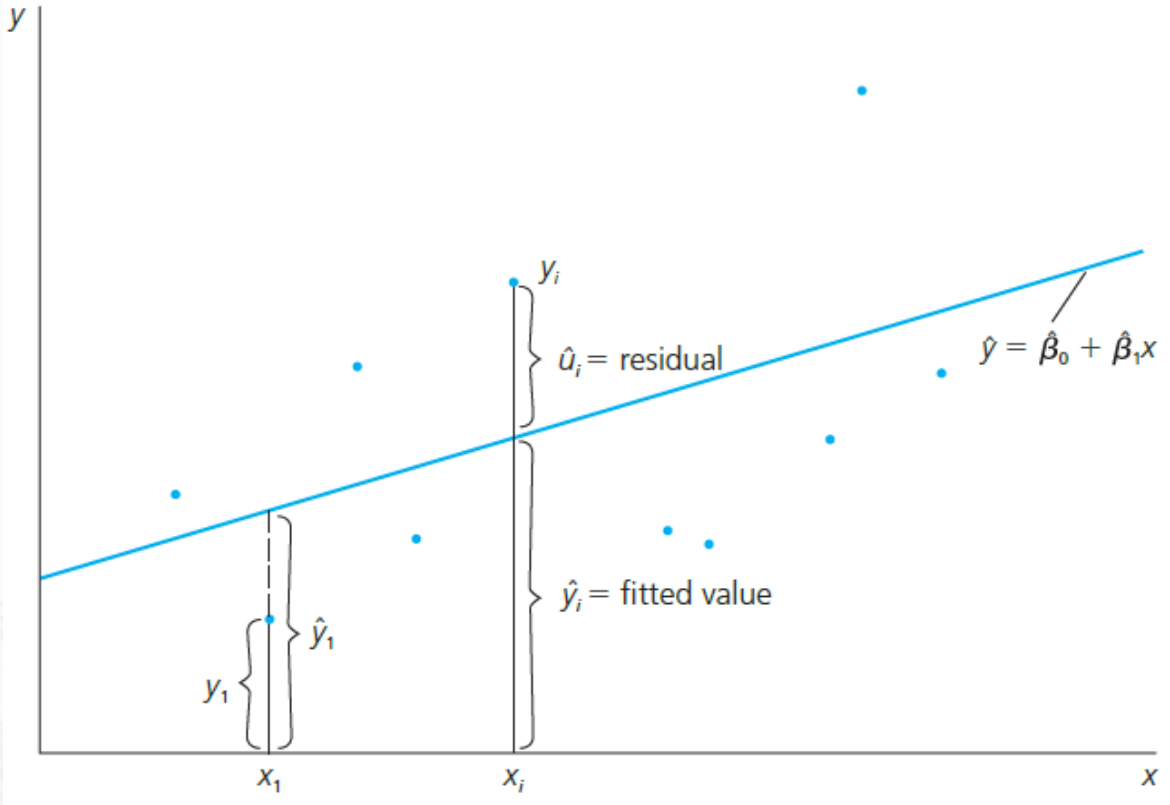
$$\{(x_i, y_i) : i = 1, \ldots, n\}$$

- 普通最小二乘法Ordinary Least Squares (OLS)

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$


y
y_i
$\hat{u}_i$ = residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y}_i$ = fitted value
$\hat{y}_1$
y_1
x_1
x_i
x

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

证明：$\min \sum_{i=1}^{n} \hat{u}_i^2 \rightarrow \hat{\beta}_0, \hat{\beta}_1$

$$min \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \Rightarrow \begin{cases} \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \bar{y} - \hat{\beta}_0 - \hat{\beta}_1\bar{x} = 0 \Rightarrow \begin{cases} \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i) = 0$$

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

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

若$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$，有

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

一、回归分析基本理解

线性回归的含义？

注意：变量的取值范围

- 多元线性回归模型

截距　　斜率参数

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

因变量
被解释变量
响应变量
被预测变量
回归子

自变量
解释变量
控制变量
预测元变量
回归元

误差项
干扰项

- 对多元回归模型的解释

$$\beta_j = \frac{\Delta y}{\Delta x_j}$$

如果第j个自变量增加一个单位，并且保持所有其他自变量和误差项恒定，则因变量的变化量是多少

- 使其他解释变量的值保持固定，即使它们与所考虑的解释变量可能相关联
- 偏效应，其他条件不变
- 必须假设，如果解释变量发生变化，则未观察到的因素不会发生变化

- 总体回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

➢关键假定：$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$

$$E(y \mid x_1, x_2, \ldots, x_k) = E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mu | x_1, x_2, \ldots, x_k)$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + E(\mu | x_1, x_2, \ldots, x_k)$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

讨论：用医生在线回复速度(Responsiveness)和医生职称(Title)来解释咨询患者数(Patients)的一个简单模型是

$$Patients=\beta_0+\beta_1 Title+\beta_2 Responsiveness+\mu$$

μ中包含了一些什么因素？你认为上述关键假定有可能成立吗？

- 相对于简单回归，使用多元回归的动机
  - 可以将更多的解释变量放入模型中
  - 明确的控制其他影响因变量的因素（在u中的）
  - 可以用于引入一般化的函数关系

例子：工资方程

在其他条件不变的情况下（明确工作经历不变）受教育程度的影响

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

u：所有的其他因素

$wage$：小时工资　$educ$：受教育程度　$exper$：工作经历

在其他条件不变的情况下（明确工作经历不变）受教育程度的影响

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

小时工资　受教育程度　工作经历　所有的其他因素

比较下：$wage = \beta_0 + \beta_1 educ + u$

- 受教育程度与工作经历相关
- 遗漏工作经历将会导致受教育程度对小时工资影响估计的偏差
- 在一个简单回归中，受教育程度的影响将部分包括工作经历对小时工资的影响，为什么？

➢可以用于引入一般化的函数关系

例子：家庭消费(cons)与家庭收入(inc)

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

线性回归的含义

例子：CEO的薪水(salary)，企业销售量(sales)，CEO的任期(ceoten)

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

- 方程是其诸参数的一个线性函数

§1.1 回归分析

- 一、回归分析基本理解
- **二、普通最小二乘法**
- 三、OLS估计量的期望值
- 四、OLS估计量的方差
- 五、OLS的有效性：高斯-马尔科夫定理
- 六、小结

二、普通最小二乘法

最小二乘：根据被解释变量的所有观测值与估计值之差的平方和最小的原则求得参数估计量

- **OLS (Ordinary Least Squares)**

➢OLS回归线或样本回归函数

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

➢随机样本(Random sample)

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \ i = 1, \ldots, n\}$$

➢回归残差(Regression residuals)

$$\hat{\mu}_i = y_i - \hat{y}_i \qquad \hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}$$

➢残差平方和最小(Sum of squared residuals, SSR)

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

➢残差平方和最小(Sum of squared residuals, SSR)

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

➢OLS一阶条件

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

假定有唯一解

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\ldots$$

$$\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

- 正规方程组的矩阵形式

$$\begin{pmatrix} n & \sum X_{1i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{ki} \\ \cdots & \cdots & \cdots & \cdots \\ \sum X_{ki} & \sum X_{ki}X_{1i} & \cdots & \sum X_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

条件？

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

- **对OLS回归方程的解释**

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

用变化量表示

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k$$

x_1的系数度量的是，在所有其他条件不变的情况下，因提高一个单位的x_1而导致的$\hat{y}$变化。即在 $x_2, x_3, \ldots, x_k$ 保持不变的情况下，

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

- 例子：大学GPA的决定因素（仅系数）

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$

大学平均成绩　　高中平均成绩　　大学能力测验分数

在MLR.1-MLR.3假设下

- **OLS对任一样本数据的性质(*)** $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{\mu}_i$

➢拟合值和残差(Fitted values and residuals)

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$ $\quad \hat{u}_i = y_i - \hat{y}_i$

OLS回归的代数性质

➢残差的样本均值为零，即 $\sum_{i=1}^{n} \hat{u}_i = 0$ 或 $\bar{y} = \bar{\hat{y}}$

➢每个自变量和OLS残差之间的样本协方差为零，OLS拟合值与OLS残差之间的样本协方差也为零

➢点 $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ 总是位于OLS回归线上 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_k \bar{x}_k$

回顾：求和算子与几个重要的性质

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n$$

$$\bar{x} = (1/n)\sum_{i=1}^{n} x_i$$

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

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

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

$$= \sum_{i=1}^{n} (x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - n(\bar{x}\cdot\bar{y})$$

残差的样本均值为零，即 $\sum_{i=1}^{n}\hat{u}_i=0$ 或 $\bar{y}=\bar{\hat{y}}$

➢证明：

$$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\ldots-\hat{\beta}_kx_{ik})=0$$

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

$$\sum_{i=1}^{n}x_{i2}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\ldots-\hat{\beta}_kx_{ik})=0$$

$$\cdots$$

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

$$\sum_{i=1}^{n}\hat{u}_i=0$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=0 \Rightarrow \bar{y}=\bar{\hat{y}}$$

每个自变量和OLS残差之间的样本协方差为零，OLS拟合值与OLS残差之间的样本协方差也为零

$$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0 \qquad \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(\hat{\mu}_i-\bar{\hat{\mu}}) = 0$$

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})(\hat{\mu}_i-\bar{\hat{\mu}}) = 0$$

➢证明：

$$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\ldots-\hat{\beta}_kx_{ik})=0$$

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

$$\sum_{i=1}^{n}x_{i2}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\ldots-\hat{\beta}_kx_{ik})=0$$

$$\ldots$$

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

$$\Rightarrow \sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$$

每个自变量和OLS残差之间的样本协方差为零，OLS拟合值与OLS残差之间的样本协方差也为零

$$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0 \qquad \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(\hat{\mu}_i-\bar{\hat{\mu}}) = 0$$

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})(\hat{\mu}_i-\bar{\hat{\mu}}) = 0$$

➢证明：

$$\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(\hat{\mu}_i-\bar{\hat{\mu}}) = \sum_{i=1}^{n}(x_{ij}-\bar{x}_j)\hat{\mu}_i$$

$$=0$$

因为$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$

所以$\hat{y}_i$与$\hat{\mu}_i$之间的样本协方差也为零

点$(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$总是位于OLS回归线上 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_k\bar{x}_k$

➢证明

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

…

$$\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\Downarrow$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_k\bar{x}_k$$

- 讨论

讨论：用高中GPA和ACT分数来解释大学GPA的OLS拟合线为：

$$colGPA = 1.29 + 0.453hsGPA + 0.0094ACT$$

若样本中的平均高中GPA约为3.4，而平均ACT分数约为24.2，那么大学的平均GPA是多少呢？你利用了什么性质？

（弗里施-沃定理，Frisch-Waugh theorem）

- **对多元回归“排除其他变量”的解释(*)**

多元回归中解释变量的估计系数可以通过两个步骤来获得：

- 1) 将解释变量对所有其他解释变量进行回归，并得到其OLS残差
- 2) 将y对上一步得到的OLS残差进行回归，得到该解释变量的估计系数

工作原理

- 第一次回归的残差是解释变量中与其他解释变量不相关的部分
- 第二回归的斜率系数表示解释变量对因变量的孤立影响

说明

解释变量的估计系数代表了**该解释变量剔除了与其他解释变量相关的部分对y的影响**

（弗里施-沃定理，Frisch-Waugh theorem）

设多元回归模型$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mu$

将x_1对自变量x_2回归，保留OLS残差$\hat{\gamma}_1$；做y对$\hat{\gamma}_1$的简单回归，$\hat{\gamma}_1$的系数就是估计原方程时x_1的系数$\hat{\beta}_1$。

➢证明：将x_{i1}对x_{i2}回归，得：

$$\hat{x}_{i1} = \hat{\alpha}_0 + \hat{\alpha}_2 x_{i2}$$
$$x_{i1} = (\hat{\alpha}_0 + \hat{\alpha}_2 x_{i2}) + \hat{\gamma}_{i1} = \hat{x}_{i1} + \hat{\gamma}_{i1}$$

代入多元回归模型的OLS一阶条件的第二个方程 $\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$

得：$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{\gamma}_{i1})\hat{\mu}_i = 0 \Rightarrow \sum_{i=1}^{n} \hat{x}_{i1}\hat{\mu}_i + \sum_{i=1}^{n} \hat{\gamma}_{i1}\hat{\mu}_i = 0$

（弗里施-沃定理，Frisch-Waugh theorem）

由于 $$\sum_{i=1}^{n}\hat{x}_{i1}\hat{\mu}_i=\sum_{i=1}^{n}(\hat{\alpha}_0+\hat{\alpha}_2x_{i2})\hat{\mu}_i=0$$

则 $$\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{\mu}_i=0 \Rightarrow \sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})=0$$

针对$x_{i1}=\hat{\alpha}_0+\hat{\alpha}_2x_{i2}+\hat{\gamma}_{i1}$，

根据OLS代数性质，有$\sum_{i=1}^{n}\hat{\gamma}_{i1}=0,\sum_{i=1}^{n}x_{i2}\hat{\gamma}_{i1}=0$

因此，$$\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2})=\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_1x_{i1})=0$$

（弗里施-沃定理，Frisch-Waugh theorem）

$$\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_1x_{i1})=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}x_{i1}=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}(\hat{x}_{i1}+\hat{\gamma}_{i1})$$

$$=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{x}_{i1}-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2=0$$

对于$\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{x}_{i1}=\sum_{i=1}^{n}\hat{\gamma}_{i1}(\hat{\alpha}_0+\hat{\alpha}_2x_{i2})=0$

因此，

$$\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2=0 \Rightarrow \hat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2}$$

（弗里施-沃定理，Frisch-Waugh theorem）

做y对$\hat{\gamma}_{i1}$的简单回归，$y = \eta_0 + \eta_1\hat{\gamma}_1 + \varepsilon, \hat{\gamma}_{i1}$的系数估计值是

$$\hat{\eta}_1 = \frac{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)^2} = \frac{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}^{\,2} - \sum_{i=1}^{n}2\hat{\gamma}_{i1}\bar{\hat{\gamma}}_1 + \sum_{i=1}^{n}\bar{\hat{\gamma}}_1^{\,2}}$$

由$\sum_{i=1}^{n}\hat{\gamma}_{i1} = 0$得

$$\hat{\eta}_1 = \frac{\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}^{\,2}} = \hat{\beta}_1$$

（弗里施-沃定理，Frisch-Waugh theorem）

以估计x_1的系数$\hat{\beta}_1$为例

设含有k的解释变量的多元回归模型$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$

将x_1对其余所有自变量x_2、…、x_k回归，保留OLS残差$\hat{\gamma}_1$；做y对$\hat{\gamma}_1$的简单回归，$\hat{\gamma}_1$的系数就是估计原方程时x_1的系数$\hat{\beta}_1$。

➢证明：将x_{i1}对x_{i2}、…、x_{ik}回归，得：

$$\hat{x}_{i1} = \hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + \cdots + \hat{\alpha}_k x_{ik}$$

$$x_{i1} = (\hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + \cdots + \hat{\alpha}_k x_{ik}) + \hat{\gamma}_{i1} = \hat{x}_{i1} + \hat{\gamma}_{i1}$$

代入多元回归模型的OLS一阶条件的第二个方程 $\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$

得：$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{\gamma}_{i1})\hat{\mu}_i = 0 \Rightarrow \sum_{i=1}^{n} \hat{x}_{i1}\hat{\mu}_i + \sum_{i=1}^{n} \hat{\gamma}_{i1}\hat{\mu}_i = 0$

（弗里施-沃定理，Frisch-Waugh theorem）

由于 $$\sum_{i=1}^{n}\hat{x}_{i1}\hat{\mu}_i=\sum_{i=1}^{n}(\hat{\alpha}_0+\hat{\alpha}_2x_{i2}+\cdots+\hat{\alpha}_kx_{ik})\hat{\mu}_i=0$$

则 $$\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{\mu}_i=0 \Rightarrow \sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})=0$$

针对$x_{i1}=\hat{\alpha}_0+\hat{\alpha}_2x_{i2}+\cdots+\hat{\alpha}_kx_{ik}+\hat{\gamma}_{i1}$，

根据OLS代数性质，有$\sum_{i=1}^{n}\hat{\gamma}_{i1}=0,\sum_{i=1}^{n}x_{ij}\hat{\gamma}_{i1}=0\ (\forall j=2,3,\ldots,k)$

因此，$$\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_kx_{ik})=\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_1x_{i1})=0$$

（弗里施-沃定理，Frisch-Waugh theorem）

$$\sum_{i=1}^{n}\hat{\gamma}_{i1}(y_i-\hat{\beta}_1x_{i1})=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}x_{i1}=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}(\hat{x}_{i1}+\hat{\gamma}_{i1})$$

$$=\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{x}_{i1}-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2=0$$

对于$\sum_{i=1}^{n}\hat{\gamma}_{i1}\hat{x}_{i1}=\sum_{i=1}^{n}\hat{\gamma}_{i1}(\hat{\alpha}_0+\hat{\alpha}_2x_{i2}+\cdots+\hat{\alpha}_kx_{ik})=0$

因此，

$$\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i-\hat{\beta}_1\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2=0 \Rightarrow \hat{\beta}_1=\frac{\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}{}^2}$$

（弗里施-沃定理，Frisch-Waugh theorem）

做y对$\hat{\gamma}_{i1}$的简单回归，$y = \eta_0 + \eta_1\hat{\gamma}_1 + \varepsilon, \hat{\gamma}_{i1}$的系数估计值是

$$\hat{\eta}_1 = \frac{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)^2} = \frac{\sum_{i=1}^{n}(\hat{\gamma}_{i1} - \bar{\hat{\gamma}}_1)y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}^{\,2} - \sum_{i=1}^{n}2\hat{\gamma}_{i1}\bar{\hat{\gamma}}_1 + \sum_{i=1}^{n}\bar{\hat{\gamma}}_1^{\,2}}$$

由$\sum_{i=1}^{n}\hat{\gamma}_{i1} = 0$得

$$\hat{\eta}_1 = \frac{\sum_{i=1}^{n}\hat{\gamma}_{i1}y_i}{\sum_{i=1}^{n}\hat{\gamma}_{i1}^{\,2}} = \hat{\beta}_1$$

总结：弗里施-沃定理，Frisch-Waugh theorem

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

$$x_{i1} = (\hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + \cdots + \hat{\alpha}_k x_{ik}) + \hat{\gamma}_{i1}$$

$$y = \eta_0 + \eta_1 \hat{\gamma}_1 + \varepsilon$$

弗里施-沃定理表明：可以把x_1分为两部分，一部分与$x_2, \ldots, x_k$相关，一部分与$x_2, \ldots, x_k$不相关

x_1的估计系数代表了

x_1剔除了与其他解释变量相关的部分对y的影响

练习题3-1

- 计算机练习题

利用WAGE1数据集中526个工人的观测数据，估计*educ*（受教育年数）、*exper*（在劳动力市场上的工作经历）和*tenure*（任现职的任期）对收入（*wage*）的影响。

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \mu$$

现进行“排除其他影响”的练习，证实对OLS估计值做“排除其他影响”的解释。要求：

- 将*educ*对*exper*和*tenure*进行回归，并保留残差$\hat{\gamma}_1$
- 将$\log(wage)$对$\hat{\gamma}_1$进行回归
- 将$\hat{\gamma}_1$的系数与在$\log(wage)$对*educ*、*exper*和*tenure*的回归中*educ*的系数相比较

- 拟合优度(Goodness-of-Fit)

“解释变量对因变量的解释程度如何”

➢波动的测量(Measures of Variation)

$$SST \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SSR \equiv \sum_{i=1}^{n}\hat{u}_i{}^2$$

总平方和(Total sum of squares)
表示因变量的总波动

解释平方和(Explained sum of squares)
表示通过回归解释的波动

残差平方和(Residual sum of squares)
表示不能通过回归解释的波动

➢总波动的分解(Decomposition of total variation)

$$SST = SSE + SSR$$

总波动　　解释的部分　　不能解释的部分

➢拟合优度度量(Goodness-of-fit measure) (R-squared)

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \qquad \bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

➢讨论：

1. 在回归中增加解释变量，R-squared 绝对不会减少，通常会增大（没有缺失数据时）

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

2. 系数估计是否可靠不取决于R-squared的大小

即使R-squared很小，OLS估计值仍有可能是每个解释变量在其它条件不变的情况下对y影响的可靠估计

3. 较低的R-squared一般表明很难准确地预测某个观测的y值

- R-squared的其它表达式

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

因变量的实际值与其拟合值的相关系数的平方

例子：对拘捕记录的解释

1986年被拘捕的次数

1986年前被捕导致定罪的比例

1986年在监狱中度过的月数

1986年被雇佣的季度数

$$\widehat{narr86} = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$

$$n = 2,725, \quad R^2 = .0413$$

$$\widehat{narr86} = .707 - .151\, pcnv + .0074\, avgsen - .037\, ptime86 - .103\, qemp86$$

此前定罪平均判刑时间

$$n = 2,725, \quad R^2 = .0422$$

§1.1 回归分析

- 一、回归分析基本理解
- 二、普通最小二乘法
- **三、OLS估计量的期望值**
- 四、OLS估计量的方差
- 五、OLS的有效性：高斯-马尔科夫定理
- 六、小结

三、OLS估计量的期望值

- 多元回归模型(MLR)的标准假设

➢Assumption MLR.1 线性于参数(Linear in parameters)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

➢Assumption MLR.2 随机抽样 (Random sampling)

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): \ i = 1, \ldots, n\}$$

数据是从整体中随机抽取的样本

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

每个数据点都遵从总体方程

➢Assumption MLR.3 不存在完全共线性 (No perfect collinearity)

在样本中（因而在总体中），没有一个自变量是常数，
自变量之间也不存在严格的线性关系

✓关于MLR.3的讨论

- 该假设仅排除了解释变量之间的完美共线性/相关性，允许不完全相关
- 如果一个解释变量是其他解释变量的完美线性组合，则该解释变量是多余的，可以消除
- 常数变量也被排除（与截距共线）
- 小样本可能会导致完全共线

例子：

expendA+*expendB*=*totexpend*

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u$$

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u$$

shareA+*shareB*=1

讨论

如果使用*expendA*、*expendB*、*shareA* (100**expendA*/*totexpend*)作为解释变量，是否违背MLR.3?

练习题3-2

- 以下方程表示，在由某国各个县构成的总体中，各种税收比例对随后就业增长方面的影响：

$$growth = \beta_0 + \beta_1 share_P + \beta_2 share_I + \beta_3 share_S + \beta_4 share_F + Control + \mu$$

其中，*growth*是就业从1980年到1990年的变化百分比，$share_P$是总税收收益中财产税的比例，*shareI*是所得税税收收益的比例，*shareS*是销售税税收收益的比例，而*shareF*包括收费和杂项税收。所有这些变量都是以1980年的货币度量。根据定义，这四个比例之和为1。其他因素将包括对教育、基础设施等的支出（均以1980年货币度量）。

（1）在模型中，保持*shareI*、*shareS*、*shareF*不变而改变$share_P$是否有意义？

（2）解释为什么这个模型违背了MLR.3？

（3）应如何重新构建这个模型，才能使得它的参数具有一个有用的解释，而又不违背假定MLR.3？

（4）对重新构建的模型中的β_1给一个仔细的解释。

➢Assumption MLR.4 零条件均值 (Zero conditional mean)

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

✓关于零条件均值的讨论

– 与误差项相关的解释变量称为**内生变量**。内生性违反了假设MLR.4

– 与误差项不相关的解释变量称为**外生变量**。如果所有解释变量都是外生的，则MLR.4成立

– 外生性是回归中因果解释的关键性假设，也是OLS估计量无偏性的关键假设

- **定理1 OLS的无偏性 (Theorem 1, Unbiasedness of OLS) (*)**

$$MLR.1-MLR.4 \quad \Rightarrow \quad E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$