

回归：预测与拟合

司继春

2025年9月



线性回归：两种不同的视角

线性回归是计量经济学中最常用的工具，然而在实践中，根据使用目的不同，线性回归这一工具在建模时的具体使用方法是不同的。

- 作为条件期望的预测工具：目标是预测的准确
 - 在因果推断中，核心问题其实就是预测反事实
- 作为因果推断的工具：目标是得到处理效应的某种平均
 - 要求外生性

这一节我们首先讲解作为预测工具的线性回归，接下来更重要的：外生性条件下的线性回归。

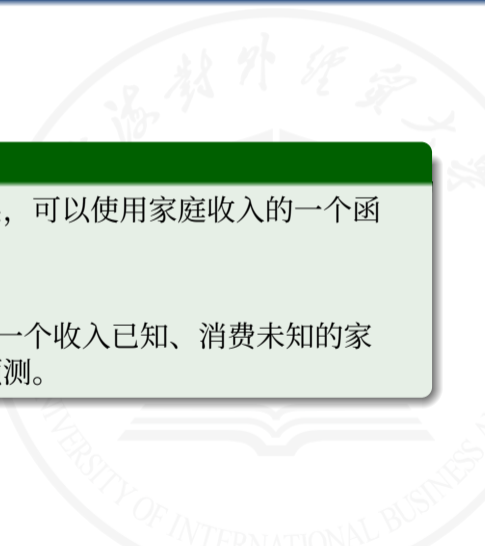
一元线性回归

一元线性回归的例子：收入与消费

一个非常经典的问题是家庭收入与消费之间的关系，可以使用家庭收入的一个函数：

$$f(I) = \alpha + \beta \cdot I$$

对家庭消费进行预测。只要我们知道了 α, β ，对于一个收入已知、消费未知的家庭，我们就可以使用以上函数对未知的消费进行预测。

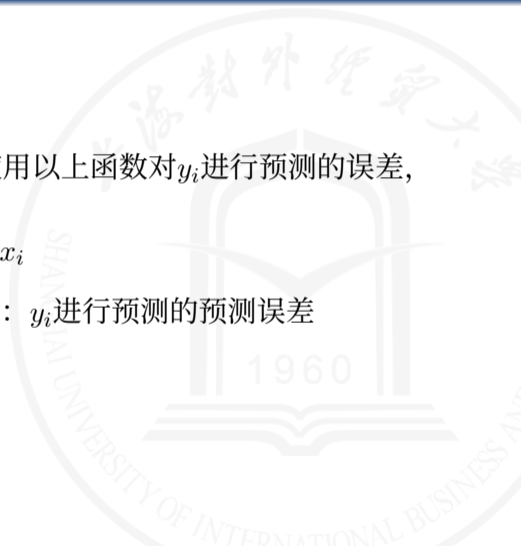


预测误差

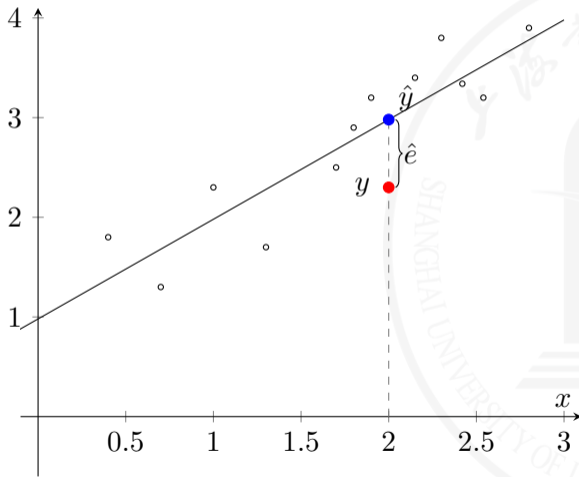
如果给定一个 α 和 β 的值 $(\tilde{\alpha}, \tilde{\beta})$ ，我们可以计算使用以上函数对 y_i 进行预测的误差，即残差 (residuals)：

$$\tilde{e}_i = y_i - \tilde{\alpha} - \tilde{\beta}x_i$$

- 残差即使用函数 $f(x) = \tilde{\alpha} + \tilde{\beta} \cdot x$ 对个体 i 的 y ： y_i 进行预测的预测误差
- 残差应该越“小”越好。



预测误差



最小二乘法

为了使得预测误差（残差）更小，一个最常见的办法是最小化均方误差（mean squared error）：

- 首先将残差计算平方，从而当预测误差为0（完美预测）时，残差的平方为0，否则不管高估还是低估，都是残差平方越小越好
- 其次对所有样本的残差平方求平均：

$$\frac{1}{N} \sum_{i=1}^N e_i^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2$$

进而，我们只要选择一个 α, β 使得以上的残差平方和最小化即可：

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \sum_{i=1}^N e_i^2 = \arg \min_{\alpha, \beta} \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2$$

最小二乘法

- 解上述最小化问题，得到：

$$\begin{cases} \frac{\partial \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2}{\partial \alpha} = -2 \sum_{i=1}^N (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2}{\partial \beta} = -2 \sum_{i=1}^N (y_i - \alpha - \beta x_i) x_i = 0 \end{cases}$$

- 化简上述问题，得到：

$$\begin{cases} \alpha = \bar{y} - \beta \bar{x} \\ \alpha \bar{x} = \frac{1}{N} \left(\sum_{i=1}^N x_i y_i - \beta \sum_{i=1}^N x_i^2 \right) \end{cases}$$

最小二乘法

- 最终得到:

$$\begin{cases} \hat{\beta} = \frac{\frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{x} \bar{y}}{\frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \\ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \end{cases}$$

即最小二乘估计量 (least-squares estimator)。

- 在得到 $\hat{\alpha}$ 、 $\hat{\beta}$ 以后, 给定任意一个 x , 可以计算其对应的对 y 的预测值:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

而残差 \hat{e} 是对于已知的 x_i, y_i , 使用 \hat{y} 对 y_i 进行预测的误差:

$$\hat{e} = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

一元线性回归示例

收入与消费

在下面的程序中，我们使用2017年CHFS的数据，使用家庭总消费对总收入做简单的一元线性回归：

代码 1: 一元线性回归示例

```
1 // file: reg_one_variate.do
2 use datasets/chfs2017_hh.dta, clear
3 drop if total_income < 0
4 drop if max(censor_total_consump, censor_total_income)
5 reg total_consump total_income
6 outreg2 using reg_one_variate.tex, replace
7 predict pred_consump
8 label variable pred_consump "预测的消费"
9 sort total_income
10 twoway (scatter total_consump total_income) (line pred_consump
        total_income)
11 graph export reg_one_variate.pdf, replace
```


一元线性回归的三个性质

- 如果我们将 x_i 的平均值 \bar{x} 带入到拟合公式中，可以得到：

$$\hat{\alpha} + \hat{\beta}\bar{x} = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}\bar{x} = \bar{y}$$

因而使用最小二乘法进行预测时，在 x_i 的平均值 \bar{x} 处的预测即 \bar{y} 。

- 残差的和：

$$\sum_{i=1}^N \hat{e}_i = \sum_{i=1}^N (y_i - \hat{\alpha} - \hat{\beta}x_i) = N\bar{y} - N\hat{\alpha} - N\hat{\beta}\bar{x} = 0$$

- 残差和 x 之间不相关：

$$\sum_{i=1}^N x_i \hat{e}_i = 0$$

从而残差和 x 之间的样本相关系数为0。

0/1型解释变量的一元线性回归

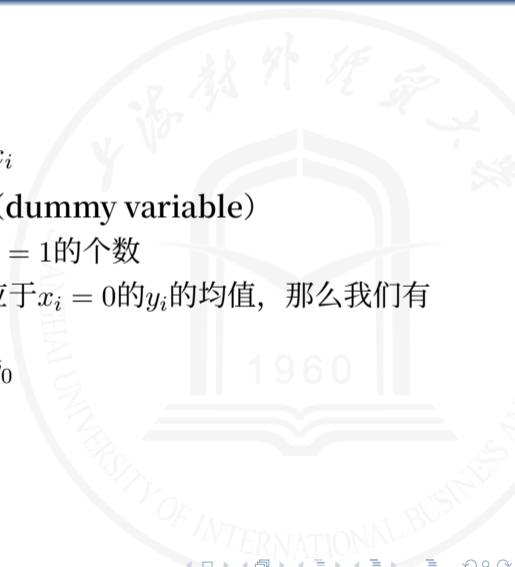
对于回归:

$$\hat{y}_i = \alpha + \beta \times x_i$$

x_i 只能取0/1两个值，此时我们称 x_i 为虚拟变量（dummy variable）

- 令 N_0 为样本中 $x_i = 0$ 的个数， N_1 为样本中 $x_i = 1$ 的个数
- 记 \bar{y}_1 为对应于 $x_i = 1$ 的 y_i 的均值，记 \bar{y}_0 为对应于 $x_i = 0$ 的 y_i 的均值，那么我们有 (how?) :

$$\begin{cases} \hat{\beta} = \bar{y}_1 - \bar{y}_0 \\ \hat{\alpha} = \bar{y}_0 \end{cases}$$



0/1型解释变量的一元线性回归

将以上结论带入到预测方程中，可以得到：

- 当 $x_i = 0$ 时，有：

$$\hat{y} = \hat{\alpha} + \hat{\beta}x_i = \hat{\alpha} = \bar{y}_0$$

- 当 $x_i = 1$ 时，有：

$$\hat{y} = \hat{\alpha} + \hat{\beta}x_i = \hat{\alpha} + \hat{\beta} = \bar{y}_1 - \bar{y}_0 + \bar{y}_0 = \bar{y}_1$$

即，当 x_i 只能取0/1两个值时，对 y 的预测即分组的均值：**分组均值即最优预测。**

虚拟变量回归与均值比较

不同性别收入比较

我们使用2017年CHFS数据比较不同性别个人收入的不同。我们使用以下程序分别使用描述性统计和回归的方法进行比较。

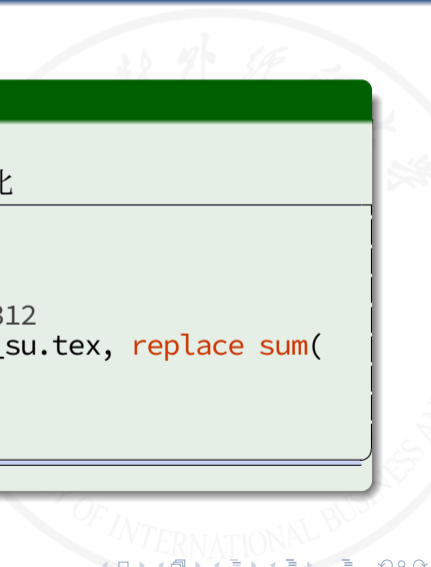
- 接着使用outreg2命令根据性别将描述性统计（只导出了观测数和收入的均值）导出
- 女性平均收入为38811.26元，而男性平均收入为48108.3元，男性收入比女性多了9297.04元。
- 接下来，我们使用回归的方法对不同性别的收入进行了比较。
- 由于gender=0代表为女性，因而截距项实际上度量了女性的平均收入，为38811.26元。而回归中的斜率项代表了gender=1（男性）与gender=0（女性）之间的收入差异，为9297.04元

虚拟变量回归与均值比较

不同性别收入比较

代码 2: 不同性别的收入对比

```
1 // file: reg_with_dummy.do
2 use datasets/chfs2017_ind.dta, clear
3 gen p_income = a3109*12
4 gen gender = 2-a2003 // 定义为男性, 为女性a200312
5 bysort gender: outreg2 using reg_with_dummy_su.tex, replace sum(
   log) eqkeep(N mean) keep(p_income)
6 reg p_income gender
7 outreg2 using reg_with_dummy.tex, replace
```



虚拟变量回归与均值比较

不同性别收入比较

VARIABLES	(1) p_income
gender	9,297*** (464.0)
Constant	38,811*** (356.8)
Observations	33,481
R-squared	0.012
Standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

多元线性回归

以上讨论了一元线性回归，即使用一个解释变量 x 对 y 进行预测。我们还可以继续推广，即使用多个 x 对 y 进行预测，即使用函数：

$$f(x_i|\beta) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

其中

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

一般而言，我们通常会保留常数项，不失一般性，我们令 $x_{i1} = 1$ 。



多元线性回归

在这里假设函数 $f(x_i, |\beta)$ 是参数线性的，即不存在 β_k 之间的非线性关系。比如：

- 我们排除了如下的函数形式：

$$\hat{y}_i = f(x_i|\beta) = \beta_1 x_{1i} + \beta_1^2 x_{2i}$$

由于存在 β_1 的非线性函数，从而以上设定不是线性回归设定。

- 包含 x_i 的非线性函数是可以的，比如：

$$\hat{y}_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{1i}^2$$

是完全允许的（此时不妨记 $x_{2i} = x_{1i}^2$ ）。

最小二乘法 (OLS)

如果我们记:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, X = [x_1, x_2, \dots, x_N]' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

那么残差向量为:

$$\hat{e} = Y - X\beta = \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_N \end{bmatrix}_{N \times 1}$$

而最小二乘法最小化:

$$\min_b \sum_{i=1}^N \hat{e}_i^2 = \min_b \hat{e}'\hat{e} = \min_b (Y - Xb)'(Y - Xb)$$

最小二乘法 (OLS)

对以上目标函数求导数并令其等于0，可以得到一阶条件：

$$\frac{\partial (Y - Xb)'(Y - Xb)}{\partial b} = \frac{\partial (Y'Y - Y'Xb - b'X'Y + b'X'Xb)}{\partial b} = -X'Y - X'Y + 2X'Xb = 0$$

解以上方程可以得到：

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

以上最大化问题的二阶导为：

$$\frac{\partial^2 (y - X\beta)'(y - X\beta)}{\partial \beta \partial \beta'} = 2X'X$$

为一个正定矩阵，因而以上根据一阶条件求得的解：

$$\hat{\beta} = (X'X)^{-1} X'Y = \left(\sum_{i=1}^N x_i x_i' \right)^{-1} \left(\sum_{i=1}^N x_i y_i \right)$$

识别条件

注意以上我们使用了矩阵 $X'X$ 的逆矩阵，由于 $\text{rank}(X'X) = \text{rank}(X)$ ，因而 $X'X$ 可逆性要求 $\text{rank}(X) = K$ ，即要求矩阵 X 是列满秩的（同时样本量 $N \geq K$ ）。为此我们必须引入如下假设：

识别条件

矩阵 X 为列满秩矩阵，即 $\text{rank}(X) = K$ 。

矩阵 X 是列满秩的意味着：

- X 的列数小于行数，即 $K < N$ 。
- X 的任何一列不能被其他列线性表示出来——无完全共线性（**perfect colinearity**）。

识别条件

收入与消费、储蓄

家庭收入 (I) 等于家庭的消费 C 加储蓄 S , $I = C + S$, 那么 I, C, S 不能同时出现在 X 里面, 否则 X 列不满秩。但是由于 $\ln(I) \neq \ln(C) + \ln(S)$, 因而 X 中同时包含 $\ln(I), \ln(C), \ln(S)$ 理论上仍然是可以的。

识别条件

- 如果以上假设不满足，且 $N > K$ ，那么最小化最小二乘目标函数的解不唯一
- 如果识别条件不满足，将会有无穷多个解使得最小化最小二乘目标函数
 - 比如，如果我们使用两个变量： x_{i2}, x_{i3} 和常数项1共同预测 y ，且 $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ 最小化最小二乘目标函数
 - 如果矩阵 X 不满秩，比如， $x_{i2} + x_{i3} = 1$ ，令 $\hat{\beta}_2^* = \hat{\beta}_2 + c$ ， c 为任意常数，那么：

$$\begin{aligned}
 \hat{y}_i &= \hat{\beta}_1 + (\hat{\beta}_2^* - c) x_{i2} + \hat{\beta}_3 x_{i3} \\
 &= \hat{\beta}_1 + \hat{\beta}_2^* x_{i2} - c x_{i2} + \hat{\beta}_3 x_{i3} \\
 &= \hat{\beta}_1 + \hat{\beta}_2^* x_{i2} - c(1 - x_{i3}) + \hat{\beta}_3 x_{i3} \\
 &= (\hat{\beta}_1 - c) + \hat{\beta}_2^* x_{i2} + (\hat{\beta}_3 + c) x_{i3} \\
 &\triangleq \hat{\beta}_1^* + \hat{\beta}_2^* x_{i2} + \hat{\beta}_3^* x_{i3}
 \end{aligned}$$

因而 $(\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*)$ 这组参数与 $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ 这组参数得到了完全一模一样的预测，因而 $(\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*)$ 也是最小二乘问题的解。

虚拟变量

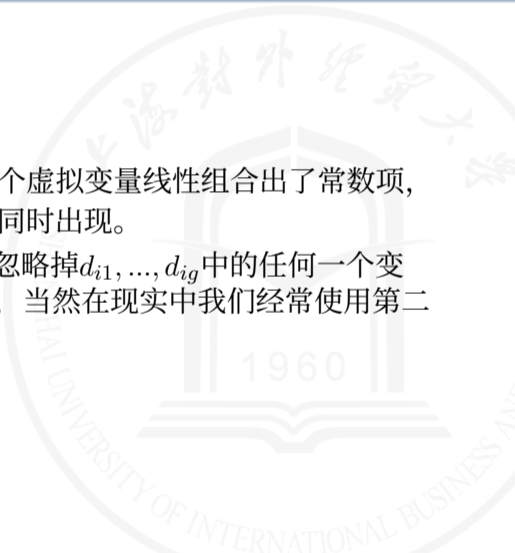
虚拟变量

对于「文化程度」这个分类变量 (G_i)，可能有7种不同的取值，比如 $G_i = 0$ 代表文盲， $G_i = 6$ 代表研究生等等，那么虚拟变量可以如下定义：

G	d_0	d_1	d_2	d_3	d_4	d_5	d_6
0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0
2	0	0	1	0	0	0	0
3	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0
5	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1

虚拟变量

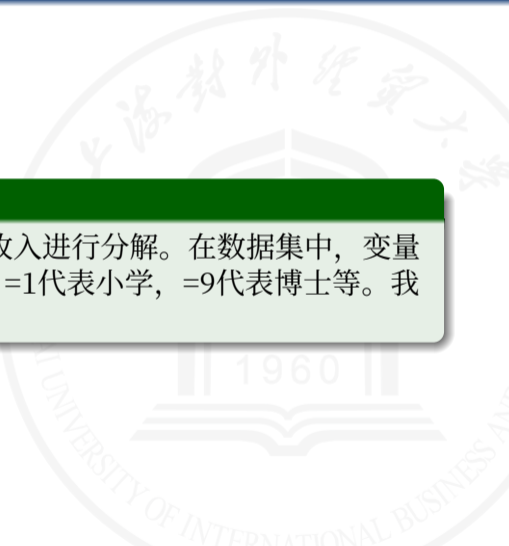
- 由于在虚拟变量定义中， $\sum_{j=0}^g d_{ij} = 1$ ，即7个虚拟变量线性组合出了常数项，所以在包含常数项的回归中， d_{i1}, \dots, d_{ig} 不能同时出现。
- 解决以上问题的方法是忽略掉常数项，或者忽略掉 d_{i1}, \dots, d_{ig} 中的任何一个变量，以上两种方法都可以使得矩阵 $X'X$ 可逆，当然在现实中我们经常使用第二种方法，即抛弃其中的一个分组虚拟变量。



虚拟变量回归

不同教育程度的收入

同样使用2017年CHFS数据，对不同教育程度的收入进行分解。在数据集中，变量 a_{2012} 代表教育程度，比如 $a_{2012}=0$ 时表示文盲， $=1$ 代表小学， $=9$ 代表博士等。我们使用如下程序计算分组差异或者分组平均：



虚拟变量回归

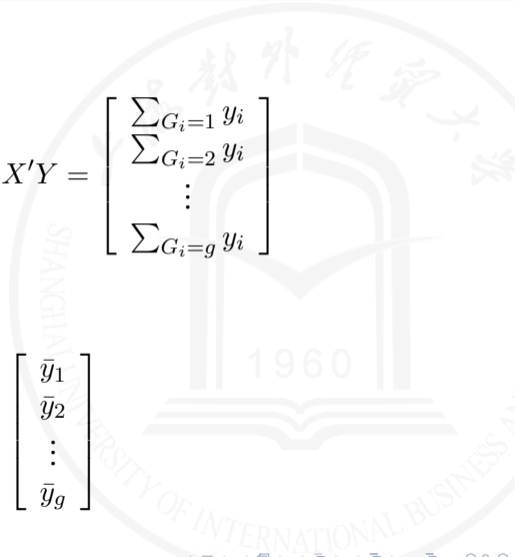
使用分块矩阵的乘法：

$$X'X = \begin{bmatrix} N_1 & 0 & \dots & 0 \\ 0 & N_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N_g \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum_{G_i=1} y_i \\ \sum_{G_i=2} y_i \\ \vdots \\ \sum_{G_i=g} y_i \end{bmatrix}$$

从而，最小二乘估计量：

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_g \end{bmatrix}$$



虚拟变量回归

如果我们包含常数项，而忽略了 d_1 ，只保留 d_2, \dots, d_g ，即：

$$\tilde{X} = \begin{bmatrix} \iota_{N_1} & 0 & \cdots & 0 \\ \iota_{N_2} & \iota_{N_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \iota_{N_g} & 0 & \cdots & \iota_{N_g} \end{bmatrix} = X \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{bmatrix} \triangleq X \cdot Q$$

其中 Q 为 $K \times K$ 的矩阵，将以上定义的 X 矩阵转换为第一列变成常数项的矩阵 \tilde{X} ，因而最小二乘估计：

$$\begin{aligned} \tilde{\beta} &= (\tilde{X}'\tilde{X})^{-1} \tilde{X}'Y = (Q'X'XQ)^{-1} Q'X'Y \\ &= Q^{-1} (X'X)^{-1} Q^{-1} Q'X'Y = Q^{-1} \hat{\beta} \end{aligned}$$

条件期望与回归

重要假设：

线性函数假设

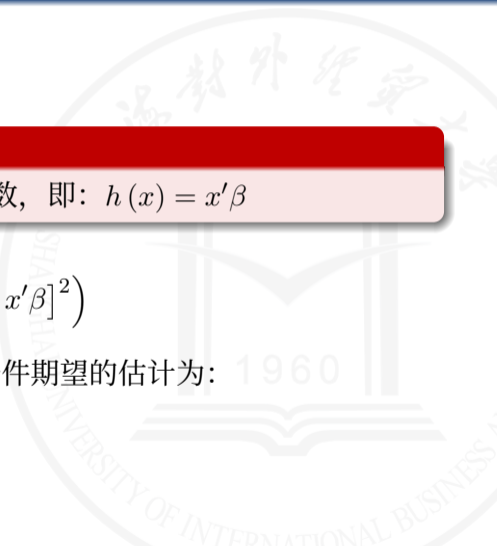
假设随机变量y给定x的条件期望 $\mathbb{E}(y|x)$ 为线性函数，即： $h(x) = x'\beta$

那么：

$$\beta_0 = \arg \min_{\beta} \mathbb{E} \left([y - x'\beta]^2 \right)$$

即为真实参数，OLS估计量 $\hat{\beta}$ 为 β_0 的估计量，而条件期望的估计为：

$$\widehat{\mathbb{E}(y|x)} = x'\hat{\beta}$$



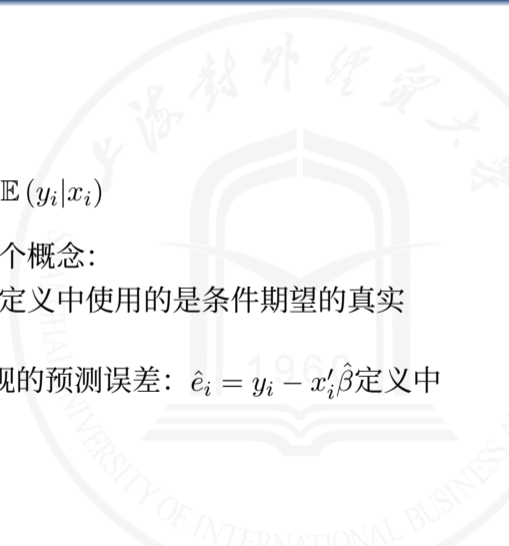
误差项

定义误差项 (error term) :

$$e_i = y_i - x_i' \beta_0 = y_i - \mathbb{E}(y_i | x_i)$$

为总体的误差。注意这里误差项和残差并不是一个概念：

- 误差项是一个总体的不可观测的随机变量，定义中使用的是条件期望的真实值 β_0 ；
- 而残差是在得到 β_0 的估计值 $\hat{\beta}$ 之后得到的实现的预测误差： $\hat{e}_i = y_i - x_i' \hat{\beta}$ 定义中使用的是估计值 $\hat{\beta}$ 。



误差项

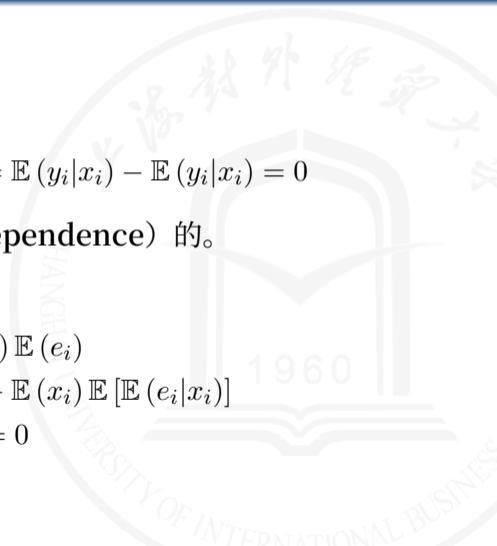
- 根据 e_i 的定义，有：

$$\mathbb{E}(e_i|x_i) = \mathbb{E}(y_i - \mathbb{E}(y_i|x_i) |x_i) = \mathbb{E}(y_i|x_i) - \mathbb{E}(y_i|x_i) = 0$$

我们称误差项与 x_i 是均值独立（mean independence）的。

- 注意均值独立意味着不相关：

$$\begin{aligned} C(x_i, e_i) &= \mathbb{E}(x_i e_i) - \mathbb{E}(x_i) \mathbb{E}(e_i) \\ &= \mathbb{E}[\mathbb{E}(x_i e_i|x_i)] - \mathbb{E}(x_i) \mathbb{E}[\mathbb{E}(e_i|x_i)] \\ &= \mathbb{E}[x_i \mathbb{E}(e_i|x_i)] = 0 \end{aligned}$$



总体回归方程

在定义了误差项之后，我们就可以将 y_i 分解为均值独立的两部分：

$$y_i = \mathbb{E}(y_i | x_i) + e_i = x_i' \beta_0 + e_i$$

- 以上方程我们通常称为**总体回归方程 (population regression equation)**，接下来将经常使用以上方程代表我们的模型。
- 注意在这里由于我们是以拟合和预测作为目的，误差项 e_i 是根据条件期望定义出来的。这与下一章中我们需要假设均值独立是有区别的。

总体回归方程

- e_i 与 x_i 虽然是均值独立的, $\mathbb{E}(e_i|x_i) = 0$, 但是并没有对 $\mathbb{E}(e_i^2|x_i)$ 做任何假设, 因而 $\mathbb{E}(e_i^2|x_i)$ 或者条件方差 $\mathbb{V}(e_i|x_i)$ 可以是 x_i 的任意函数。
- 如果 $\mathbb{V}(e_i|x_i) = \mathbb{E}(e_i^2|x_i)$ 不为常数, 那么我们称 e_i 具有异方差 (heteroscedasticity) ;
- 如果 $\mathbb{V}(e_i|x_i) = \mathbb{E}(e_i^2|x_i) = \mathbb{E}(e_i^2)$, 那么我们称 e_i 具有同方差 (homoscedasticity) 性质。
- 均值独立意味着 x_i 对 e_i 没有预测能力, 而异方差的存在意味着 x_i 对 e_i 的方差仍然具有预测能力, 两者并不矛盾。
- 注意, 由于:

$$\mathbb{V}(y_i|x_i) = \mathbb{V}(x_i'\beta_0 + e_i|x_i) = \mathbb{V}(e_i|x_i)$$

即 $y_i|x_i$ 的条件方差也就是 $e_i|x_i$ 的条件方差。

最小二乘的统计性质

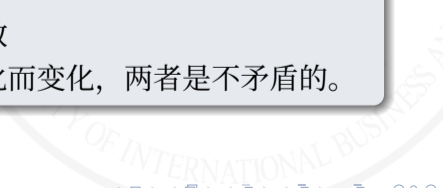
我们现在将最小二乘估计 $\hat{\beta}$ 看成是总体回归方程中真值 β_0 的一个估计。在此基础上，我们继续讨论最小二乘估计 $\hat{\beta}$ 的统计性质，包括 $\hat{\beta}$ 的无偏性、一致性。为此我们引入如下假设：

独立同分布假设

设样本 $[x_i', y_i]'$, $i = 1, 2, \dots, N$ 独立同分布。

注意独立同分布假设与异方差并不矛盾：

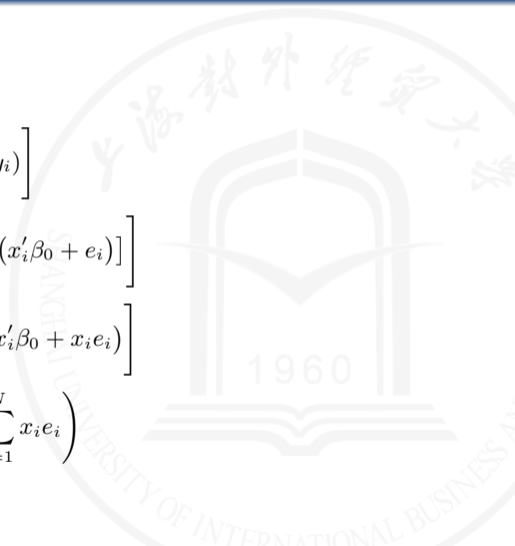
- 异方差指的是条件方差 $V(y_i|x_i) = \sigma^2(x_i)$ 不为常数
- 然而同分布则意味着无条件方差 $V(y_i)$ 不随 i 的变化而变化，两者是不矛盾的。



OLS的统计性质

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'Y \\ &= \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \left[\sum_{i=1}^N (x_i y_i) \right] \\ &= \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \left[\sum_{i=1}^N [x_i (x_i' \beta_0 + e_i)] \right] \\ &= \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \left[\sum_{i=1}^N (x_i x_i' \beta_0 + x_i e_i) \right] \\ &= \beta_0 + \left[\sum_{i=1}^N (x_i x_i') \right]^{-1} \left(\sum_{i=1}^N x_i e_i \right) \\ &= \beta_0 + (X'X)^{-1} X'e\end{aligned}$$

其中 $e = [e_1, \dots, e_N]'$ 。



无偏性

进一步，有：

$$\begin{aligned}
 \mathbb{E}(\hat{\beta}) &= \mathbb{E}\left[\mathbb{E}(\hat{\beta}|X)\right] \\
 &= \mathbb{E}\left[\mathbb{E}\left(\beta_0 + \left[\sum_{i=1}^N (x_i x_i')\right]^{-1} \left(\sum_{i=1}^N x_i e_i\right) \mid X\right)\right] \\
 &= \beta_0 + \mathbb{E}\left[\left[\sum_{i=1}^N (x_i x_i')\right]^{-1} \mathbb{E}\left(\sum_{i=1}^N x_i e_i \mid X\right)\right] \\
 &= \beta_0 + \mathbb{E}\left[\left[\sum_{i=1}^N (x_i x_i')\right]^{-1} \sum_{i=1}^N (x_i \mathbb{E}(e_i|X))\right] \\
 &= \beta_0
 \end{aligned}$$

一致性

如果 $[x_i', y_i]'$ 是独立同分布的, 且 $\mathbb{E}(x_{ik}^2) < \infty$ 以及 $\mathbb{E}|x_{ik}e_i| < \infty, k = 1, \dots, K$, 根据大数定律, 有:

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N (x_i x_i') \xrightarrow{p} \mathbb{E}(x_i x_i') \\ \frac{1}{N} \sum_{i=1}^N (x_i e_i) \xrightarrow{p} \mathbb{E}(x_i e_i) = 0 \end{cases}$$

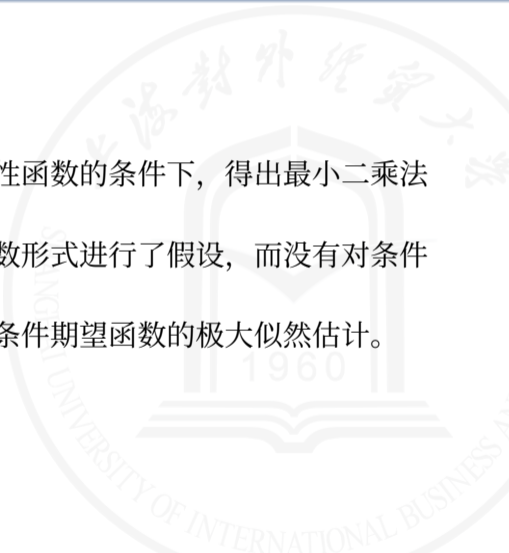
由于矩阵求逆为连续映射, 因而:

$$\hat{\beta} - \beta_0 = \left[\frac{1}{N} \sum_{i=1}^N (x_i x_i') \right]^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i e_i \right) \xrightarrow{p} \mathbb{E}(x_i x_i')^{-1} \mathbb{E}(x_i e_i) = 0$$

正态性假设与条件极大似然估计

在前面的讨论中，我们在假设条件期望函数为线性函数的条件下，得出最小二乘法可以看做是条件期望的估计。

- 注意在以上过程中，我们只对条件期望的函数形式进行了假设，而没有对条件分布做任何假设。
- 接下来，我们将加强条件分布的假设，讨论条件期望函数的极大似然估计。



正态性假设

条件正态性假设

设样本 $[y_i, x_i']'$, $i = 1, \dots, N$ 独立同分布，且 y_i 给定 x_i 的条件分布为同方差的正态分布，其条件期望为线性函数，即：

$$y_i|x_i \sim \mathcal{N}(x_i'\beta_0, \sigma^2)$$

或者等价的：

$$Y|X \sim \mathcal{N}(X\beta_0, \sigma^2 I)$$

以上假设等价于假设误差项 $e_i|x_i \sim \mathcal{N}(0, \sigma^2)$ ，或者 $e|X \sim \mathcal{N}(0, \sigma^2 I)$ 。

正态性假设

这里需要注意的是，即使我们假设了 $y_i|x_i$ 服从正态分布，这也不意味着 y_i 也服从正态分布：

条件正态与正态

假设数据生成过程：

$$y_i = 85 - 5 \times d_i + e_i$$

其中 $d_i = 0/1$ 为虚拟变量，且假设 $P(d_i = 1) = 0.3$ ， $e_i|x_i \sim \mathcal{N}(0, 4)$ ，那么 y_i 本身为一个混合正态分布（作业）。

条件极大似然估计

根据以上假设，条件密度函数为：

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - x'_i\beta_0)^2}{2\sigma^2}\right\}$$

因而条件似然函数为：

$$L(\beta, \sigma|y, x) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \sigma^2 - \sum_{i=1}^N \frac{(y_i - x'_i\beta)^2}{2\sigma^2}$$

最大化以上函数，得到：

$$\begin{cases} \hat{\beta} = \left(\sum_{i=1}^N x_i x'_i\right)^{-1} \left(\sum_{i=1}^N x_i y_i\right) = (X'X)^{-1} X'Y \\ \hat{\sigma}^2 = \frac{\sum_{i=1}^N (y_i - x'_i\hat{\beta})^2}{N} = \frac{\sum_{i=1}^N \hat{e}_i^2}{N} \end{cases}$$

其中 $\hat{e}_i = y_i - x'_i\hat{\beta}$ 为残差。再次，我们得到了最小二乘估计量。

最小二乘与条件期望

条件期望即我们使用自变量 x 对因变量 y 的最优预测。然而从条件期望得到OLS的过程中，我们假设了条件期望的线性函数形式：

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_{K-1} x_{K-1} + e$$

然而这一条件未必满足：

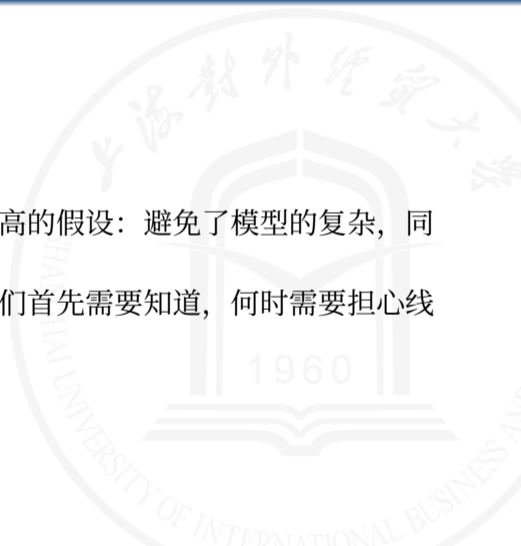
- 如果条件期望函数的确为线性函数，OLS是对条件期望函数 $\mathbb{E}(y|x)$ 的最优预测；
- 如果条件期望函数不是线性函数，则OLS是对条件期望函数的最优线性逼近：

$$\beta_0 = \arg \min_{\beta} [x' \beta - \mathbb{E}(y|x)]^2$$

最小二乘与条件期望

然而，很多时候条件期望函数并非线性函数。

- 大多数情况下线性函数是一个“性价比”极高的假设：避免了模型的复杂，同时也足够精准
- 然而有的时候，线性函数会有很大问题。我们首先需要知道，何时需要担心线性假设的问题。



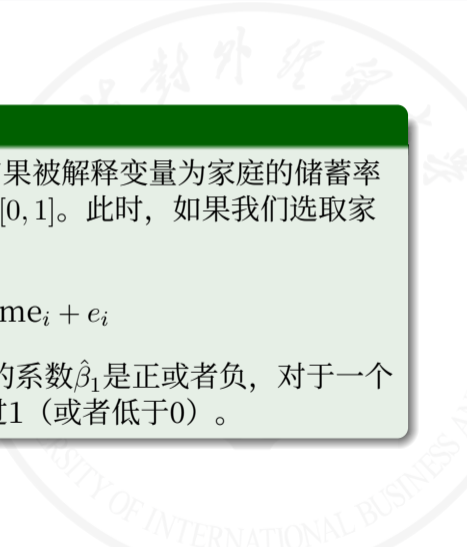
条件期望函数形式问题

支撑集问题

支撑集 (support) 即一个随机变量的取值范围。如果被解释变量为家庭的储蓄率 (saving_rate)，我们知道 $\text{supp}(\text{savin_rate}_i) = [0, 1]$ 。此时，如果我们选取家庭资产规模 (wealth) 作为解释变量：

$$\text{savin_rate}_i = \beta_0 + \beta_1 \cdot \text{income}_i + e_i$$

由于 $\text{supp}(\text{income}_i) = [0, \infty)$ ，因而不管回归得到的系数 $\hat{\beta}_1$ 是正或者负，对于一个资产规模足够大的家庭，总会使得预测的储蓄率超过1（或者低于0）。



条件期望函数形式问题

经济增长

如果令 y_t 为时期 t 时国家的GDP，根据索洛模型（Acemoglu, 2009, Chaper 3）， y_t 满足如下关系式：

$$g_t = \beta_0 + \beta_1 \ln y_{t-1} + e_t$$

其中 $g_t = \ln y_t - \ln y_{t-1}$ 为GDP的对数增长率。根据上式，得到：

$$y_t = \exp \{ \beta_0 + (1 + \beta_1) \ln y_{t-1} + e_t \} = e^{\beta_0} y_{t-1}^{1+\beta_1} e^{e_t}$$

从而条件期望函数：

$$\mathbb{E}(y_t | y_{t-1}) = e^{\beta_0} y_{t-1}^{1+\beta_1} \mathbb{E}(e^{e_t} | y_{t-1})$$

因而条件期望函数为一个指数函数形式，而非线性函数。

条件期望函数形式问题

引力模型

在国际贸易理论中 (Head and Mayer, 2014) , 双边贸易与两个国家的GDP之间存在着被称为“引力模型”的关系, 即:

$$X_{ni} = GY_i^a Y_n^b \phi_{ni}$$

其中下标*i*代表国家, 而*n*代表出口目的地国, X_{ni} 为两国之间的贸易额, G 为常数, Y 为国家的GDP, ϕ_{ni} 则是两国之间贸易成本的函数。双边贸易额与GDP之间的关系并非简单的线性关系。

条件期望函数形式问题

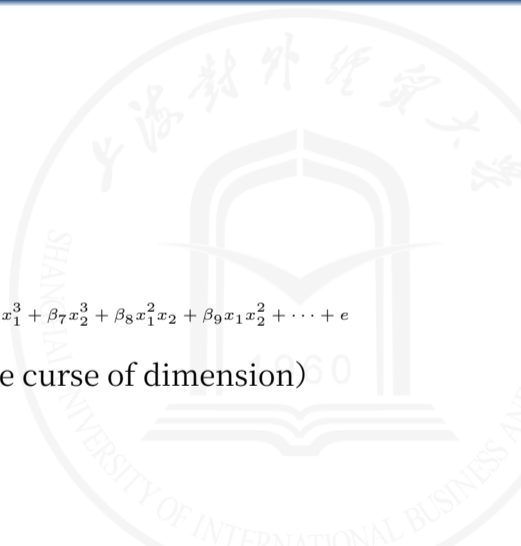
解决方案:

- 使用非参数回归 (kernel, sieve)
- 机器学习方法
- 引入多项式 (平方项、交叉项) :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 x_1^3 + \beta_7 x_2^3 + \beta_8 x_1^2 x_2 + \beta_9 x_1 x_2^2 + \dots + e$$

- 以上方案可能有其缺点, 如维数的诅咒 (the curse of dimension)

更常用的解决方案——对数据进行变换:



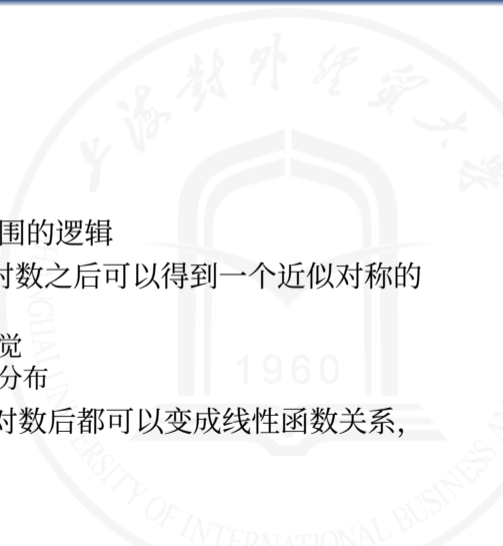
对数变换

对数变换的最常用的变换：

$$x \rightarrow \ln(x)$$

进行对数变换的理由：

- 将 $[0, +\infty)$ 变换到 $(-\infty, \infty)$ 上，更符合取值范围的逻辑
- 有偏的分布，如收入、财富等右偏分布，取对数之后可以得到一个近似对称的分布
 - 解释变量和被解释变量都是对称的更符合直觉
 - 一些右偏的分布比较难以线性组合出对称的分布
- 理论预期。如上例中的GDP和出口，变量取对数后都可以变成线性函数关系，这是经济学理论预期的。



对数变换

- 具有弹性 (elasticity) 解释, 经过取对数后, 其变化可以解释为百分比变化:

$$d \ln y = \frac{dy}{y}$$

人口、GDP等具有比较平稳的增长率, 取对数更容易与其他变量之间满足线性关系。

- 比如对于GDP: $y_t \propto (1 + \beta_1)^t y_0$, 取对数后:

$$\ln y_t = C + t \log(1 + \beta_1) + \log y_0$$

更容易与其他变量形成线性关系

- 如果GDP是指数增长, 那么:

$$X_{nit} \propto (1 + \beta_{i1})^{ta} y_{i0}^a (1 + \beta_{n1})^{tb} y_{n0}^b$$

从而出口也类似, 取对数后容易与其他变量形成线性关系

对数变换

- 实际上对于一些“比例”型的数据，取对数有时也会有比较好的解释。
- 比如储蓄率例子中，储蓄率 $\text{saving_rate} = \frac{\text{saving}}{\text{income}}$ ，如果我们将其取对数：

$$\ln \text{saving_rate} = \ln \text{saving} - \ln \text{income}$$

从而如果将之前回归的被解释变量和解释变量取对数，即：

$$\ln \text{saving_rate}_i = \beta_0 + \beta_1 \cdot \ln \text{income}_i + e_i$$

等价于：

$$\ln \text{saving}_i - \ln \text{income}_i = \beta_0 + \beta_1 \cdot \ln \text{income}_i + e_i$$

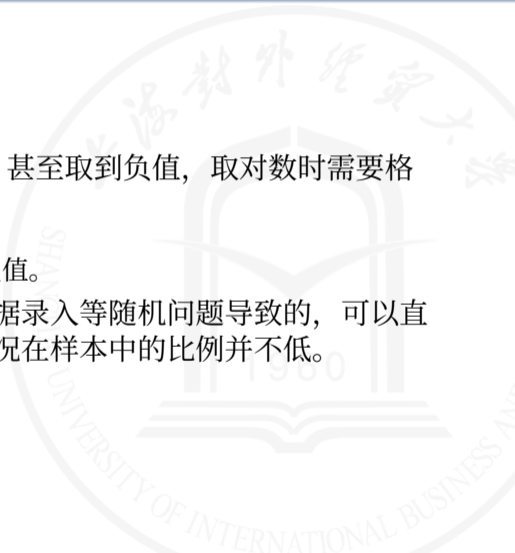
- 实际上我们可以证明（练习1.11），以上回归与以下回归是等价的：

$$\ln \text{saving}_i = \delta_0 + \delta_1 \cdot \ln \text{income}_i + e_i$$

且OLS估计量 $\hat{\beta}_0 = \hat{\delta}_0$, $\hat{\beta}_1 = \hat{\delta}_1 - 1$ 。

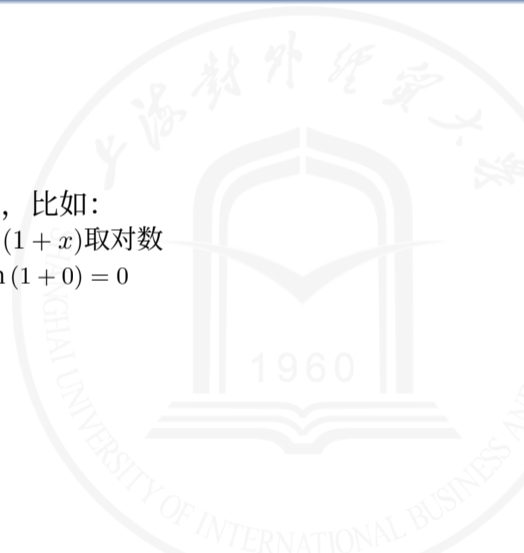
对数变换

- 最后需要注意的是，有些变量可能取到0值，甚至取到负值，取对数时需要格外小心。
 - 需要对收入取对数，然而很多人的收入为0；
 - 需要对净出口取对数，然而净出口可能为负值。
- 如果这些情况样本量比较少，可能是由于数据录入等随机问题导致的，可以直接忽略这些样本，然而更多的情况是这些情况在样本中的比例并不低。



对数变换

- 一些不太严谨的方法可以大概处理这些问题，比如：
 - 对于数据可能为0的问题，我们可以使用 $\ln(1+x)$ 取对数
 - 如此哪些 $x=0$ 的样本取“对数”之后， $\ln(1+0)=0$
 - 且这个变换是一个单调变换



对数变换

一些方法也许可以帮助解决这一问题

- 比如如果需要对被解释变量 y 取对数，我们发现很多的 $y = 0$:
 - 使用Tobit一类回归，比如第I类Tobit等。
 - 在国际贸易中，Eaton和Tamura (1994) 在处理引力模型时，提出可以使用 $\ln(a + X_{ni})$ 作为被解释变量，而将 a 作为一个待估参数，即ET Tobit (Head和Mayer, 2014) 。
 - Caprettini和Voth (2023) 推荐使用泊松拟极大似然回归 (PPML)
- 如果需要取对数的变量为解释变量，一个简单的处理方法是定义两个新的变量：

$$\ln x = \begin{cases} \ln x & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases}$$

$$d = \mathbf{1}\{x = 0\}$$

从而使用回归：

$$y_i = \beta_0 + \beta_1 \ln x_i + \beta_2 d_i + u_i$$

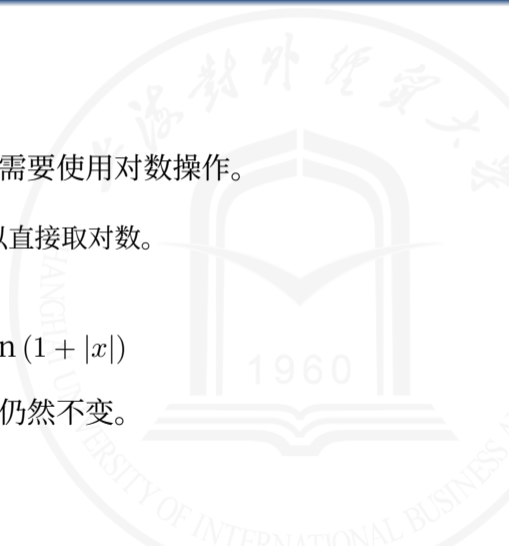
负数的对数变换

一些方法也许可以帮助解决这一问题

- 此外，还有些可以取到负值的变量仍然可能需要使用对数操作。
 - 比如，净出口额、人口净流入等变量
 - 取对数是一个合理的操作，然而负数不可以直接取对数。
- 对于可能为负的变量，一种方法是使用：

$$g(x) = \text{Sign}(x) \cdot \ln(1 + |x|)$$

该变换同样也是单调变换，经过变换后符号仍然不变。



负数的对数变换

或者，也可以使用反双曲正弦函数（如Caprettini和Voth, 2023）

- 双曲正弦函数的定义为：

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

- 以上函数的定义域和值域都是 \mathbb{R}
- 当 $x \rightarrow \infty (-\infty)$ 时，以上函数趋向于 $\frac{e^x}{2} \left(-\frac{e^{-x}}{2}\right)$ ，从而当 $|x|$ 足够大时，以上函数近似于 $\frac{e^x}{2} \left(-\frac{e^{-x}}{2}\right)$
- 其反函数为：

$$\operatorname{arsinh}(x) = \ln\left(x + \sqrt{x^2 + 1}\right)$$

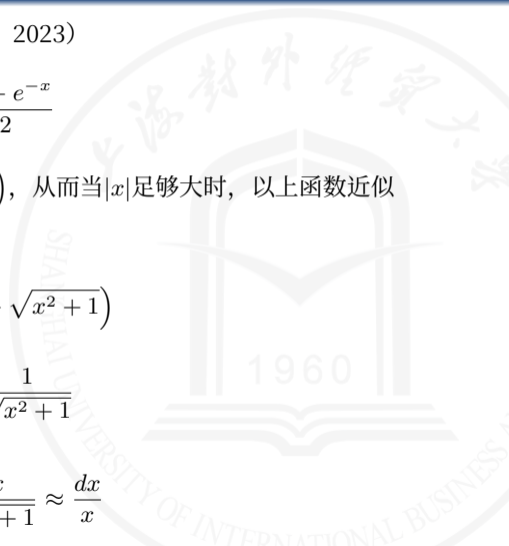
其导函数为：

$$\frac{d\operatorname{arsinh}(x)}{dx} = \frac{1}{\sqrt{x^2 + 1}}$$

当 $|x|$ 足够大时，以上导函数与 $\frac{1}{x}$ 近似，从而：

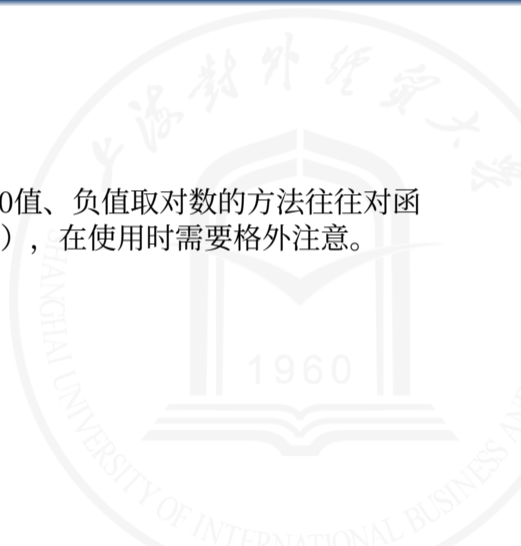
$$d\operatorname{arsinh}(x) = \frac{dx}{\sqrt{x^2 + 1}} \approx \frac{dx}{x}$$

也可以近似解释为百分比变动。



Log with Zeros

- 最后，仍然需要再次提示的是，以上的解决0值、负值取对数的方法往往对函数形式有很强的假设（Chen和Roth, 2023），在使用时需要格外注意。
- 其他方法：
 - 拟泊松回归



其他变换

其他变换:

- Box-Cox变换 (不推荐):

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

- Logistic逆变换: 使用

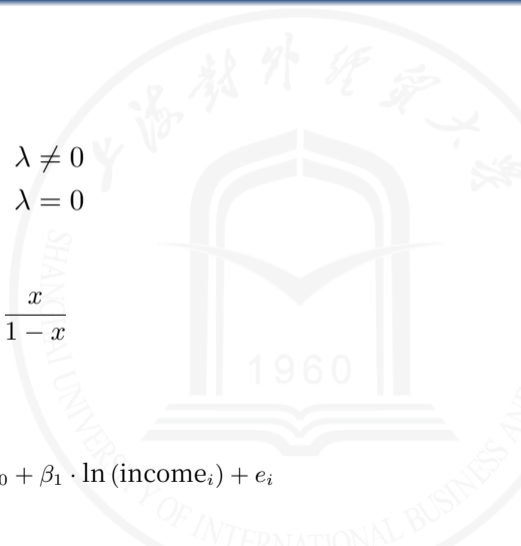
$$f(x) = \ln \frac{x}{1-x}$$

将(0, 1)区间上的实数映射到 $(-\infty, \infty)$ 上

- 比如对于储蓄率, 我们使用:

$$\ln \frac{\text{saving_rate}_i}{1 - \text{saving_rate}_i} = \beta_0 + \beta_1 \cdot \ln(\text{income}_i) + e_i$$

从而左边和右边取值范围都是 \mathbb{R}



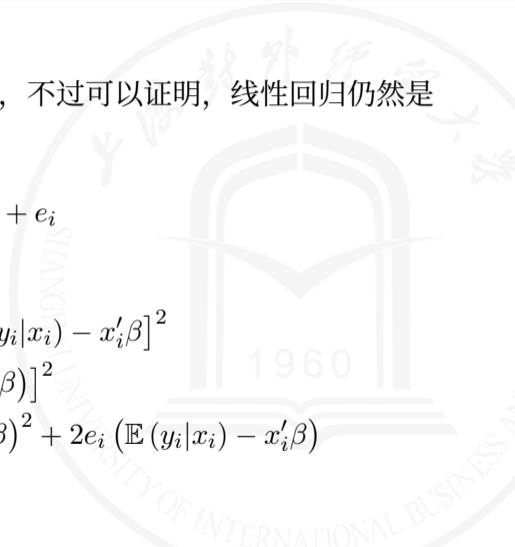
条件期望的最优逼近

- 真实的条件期望函数我们是永远无法知道的，不过可以证明，线性回归仍然是条件期望函数的最优线性近似。
- 根据定义：

$$y_i = \mathbb{E}(y_i|x_i) + e_i$$

而最小二乘法的目标函数可以写为：

$$\begin{aligned} (y_i - x_i'\beta)^2 &= [y_i - \mathbb{E}(y_i|x_i) + \mathbb{E}(y_i|x_i) - x_i'\beta]^2 \\ &= [e_i + (\mathbb{E}(y_i|x_i) - x_i'\beta)]^2 \\ &= e_i^2 + (\mathbb{E}(y_i|x_i) - x_i'\beta)^2 + 2e_i(\mathbb{E}(y_i|x_i) - x_i'\beta) \end{aligned}$$



条件期望的线性逼近

由于

$$\mathbb{E} [e_i (\mathbb{E} (y_i | x_i) - x_i' \beta)] = \mathbb{E} (\mathbb{E} [e_i (\mathbb{E} (y_i | x_i) - x_i' \beta)] | x_i) = 0$$

从而:

$$\mathbb{E} [(y_i - x_i' \beta)^2] = \mathbb{E} (e_i^2) + (\mathbb{E} (y_i | x_i) - x_i' \beta)^2$$

其中第一项跟 β 无关, 因而最小化 $\mathbb{E} [(y_i - x_i' \beta)^2]$ 等价于最小化 $(\mathbb{E} (y_i | x_i) - x_i' \beta)^2$, 即 $x_i' \beta_0$ 是条件期望函数 $\mathbb{E} (y_i | x_i)$ 在均方误差标准下的最优线性逼近。

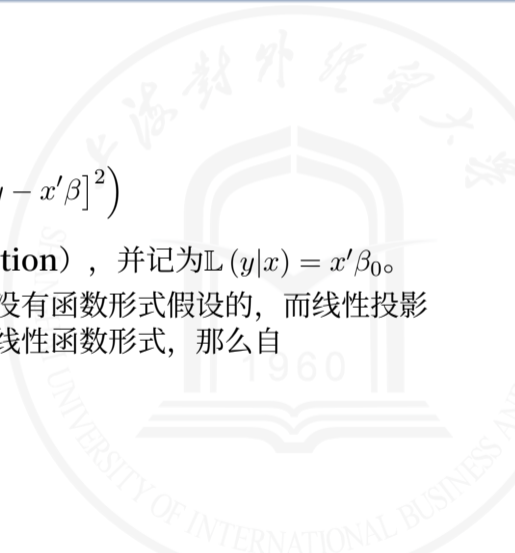
线性投影

- 对于最优化问题:

$$\beta_0 = \arg \min_{\beta} \mathbb{E} \left([y - x'\beta]^2 \right)$$

我们称 $x'_i\beta_0$ 其称为**线性投影 (linear projection)**，并记为 $\mathbb{L}(y|x) = x'\beta_0$ 。

- 线性投影区别于条件期望，因为条件期望是没有函数形式假设的，而线性投影有函数形式假设，如果真实的条件期望就是线性函数形式，那么自然 $\mathbb{E}(y|x) = \mathbb{L}(y|x)$ 。



线性投影的性质

- ① 令 $e = y - \mathbb{L}(y|x)$, 那么 $\mathbb{E}(ex) = 0$, 且 $\mathbb{L}(e|x) = 0$
- ② $\mathbb{L}(a_1y_1 + a_2y_2|x) = a_1\mathbb{L}(y_1|x) + a_2\mathbb{L}(y_2|x)$
- ③ $\mathbb{L}(y|x) = \mathbb{L}[\mathbb{L}(y|x, w)|x]$
- ④ $\mathbb{L}(y|x) = \mathbb{L}[\mathbb{E}(y|x, w)|x]$



变换后的预测

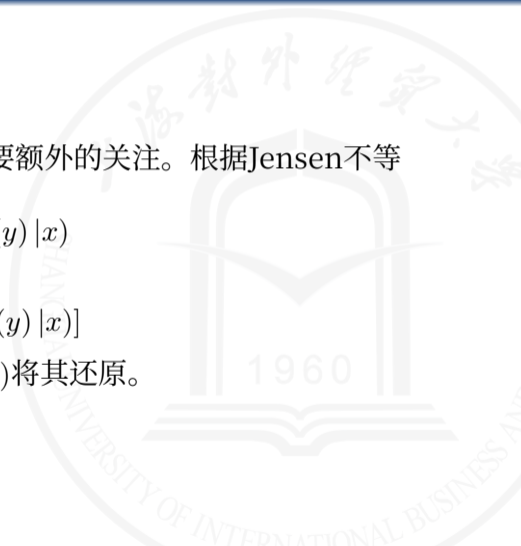
当我们使用 y_i 的非线性变换时，对于 y_i 的预测需要额外的关注。根据Jensen不等式，由于：

$$f(\mathbb{E}(y|x)) \neq \mathbb{E}(f(y)|x)$$

因而

$$\mathbb{E}(y|x) \neq f^{-1}[\mathbb{E}(f(y)|x)]$$

为了预测 y 的值，不能先预测 $f(y)$ ，再使用 $f^{-1}(\cdot)$ 将其还原。



对数的预测

- 比如，如果我们设定如下方程：

$$\ln y = x'\beta + e$$

使用以上方程，我们得到的实际上是对于条件期望函数： $\mathbb{E}(\ln y|x)$ 的最优线性估计

- 然而，根据Jensen不等式：

$$\mathbb{E}(\ln y|x) \leq \ln [\mathbb{E}(y|x)]$$

从而：

$$\mathbb{E}(y|x) \geq \exp \{ \mathbb{E}(\ln y|x) \}$$

因而如果我们使用 $\exp(x'\hat{\beta})$ 对 y 进行预测，会低估 y 的条件期望。

对数的预测

- 注意到: $y = \exp(x'\beta) \cdot \exp(e)$, 从而

$$\mathbb{E}(y|x) = \exp(x'\beta) \mathbb{E}(\exp(e)|x)$$

- 如果假设 e 和 x 独立且 $e \sim \mathcal{N}(0, \sigma^2)$, 那么

$$\mathbb{E}(\exp(e)|x) = \mathbb{E}(\exp(e)) = e^{\sigma^2/2}$$

从而:

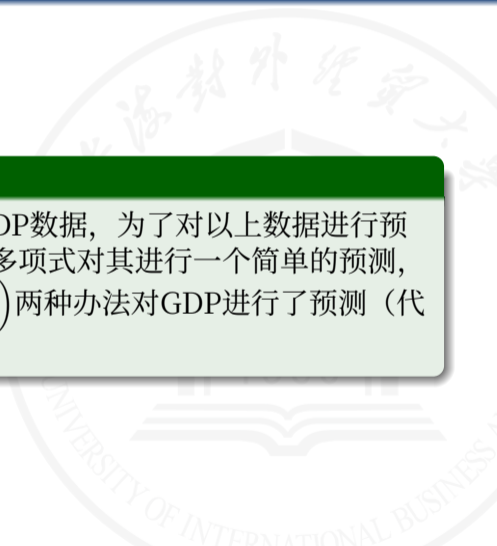
$$\mathbb{E}(y|x) = \exp\left(x'\beta + \frac{\sigma^2}{2}\right)$$

将 β 和 σ^2 使用极大似然回归结果, 替代即可得到 y 的条件期望的预测值。

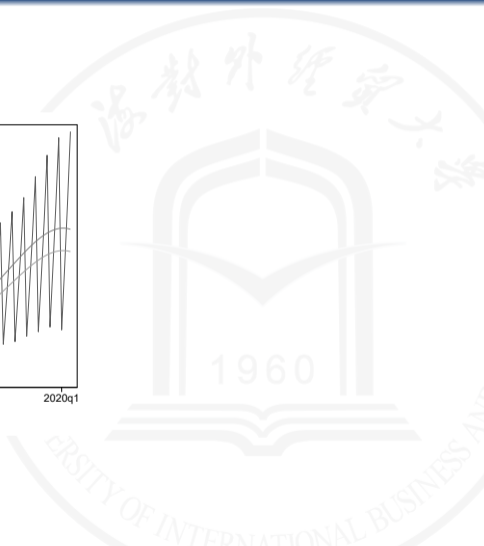
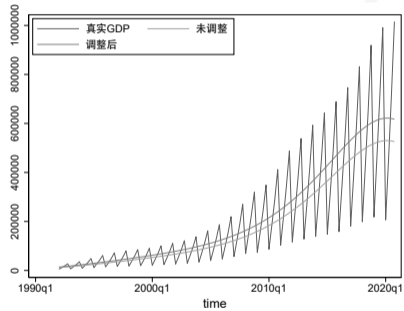
预测GDP

季度GDP的预测

数据集quarterlyGDP.dta中包含了我国的季度GDP数据，为了对以上数据进行预测，我们可以将GDP先取对数，然后使用时间的多项式对其进行一个简单的预测，接下来，我们分别使用 $\exp(x'\beta)$ 和 $\exp\left(x'\beta + \frac{\sigma^2}{2}\right)$ 两种办法对GDP进行了预测（代码：predict_log.do）



预测GDP



分步回归

- 对于总体回归:

$$y_i = x_i' \beta + e_i$$

如果我们将解释变量 x_i 分为两部分: $x_i = [x_{i1}', x_{i2}']'$, 那么总体回归方程可以写为:

$$y_i = x_{i1}' \beta_1 + x_{i2}' \beta_2 + e_i$$

- 如果我们将以上方程两边同时对 x_{i2} 求条件期望, 得到:

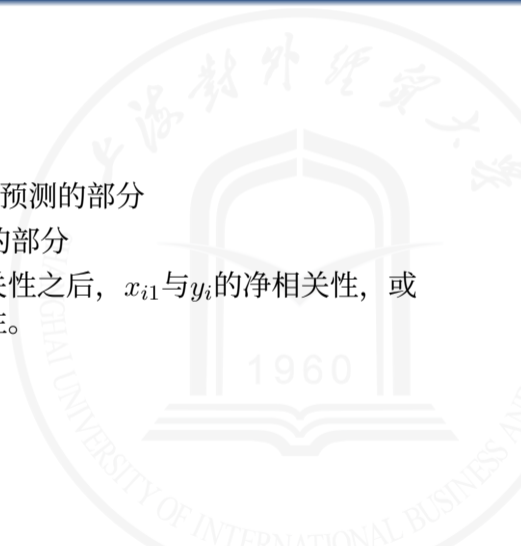
$$\begin{aligned} \mathbb{E}(y_i | x_{i2}) &= \mathbb{E}(x_{i1} | x_{i2})' \beta_1 + x_{i2}' \beta_2 + \mathbb{E}(e_i | x_{i2}) \\ &= \mathbb{E}(x_{i1} | x_{i2})' \beta_1 + x_{i2}' \beta_2 \end{aligned}$$

在总体回归方程两边同时减去上式, 得到:

$$\begin{aligned} y_i - \mathbb{E}(y_i | x_{i2}) &= x_i' \beta + e_i - \mathbb{E}(x_{i1} | x_{i2})' \beta_1 - x_{i2}' \beta_2 \\ &= [x_{i1} - \mathbb{E}(x_{i1} | x_{i2})]' \beta_1 + e_i \end{aligned}$$

分步回归

- 在上式中, $y_i - \mathbb{E}(y_i|x_{i2})$ 代表 y_i 中 x_{i2} 所不能预测的部分
- 而 $x_{i1} - \mathbb{E}(x_{i1}|x_{i2})$ 代表 x_{i1} 中 x_{i2} 所不能预测的部分
- 因而系数 β_1 代表的是排除 x_{i2} 与 x_{i1} 和 y_i 的相关性之后, x_{i1} 与 y_i 的净相关性, 或者在保持 x_{i2} 不变的条件下, x_{i1} 与 y_i 的相关性。



分步回归

- 类似的结果对于线性投影也成立:

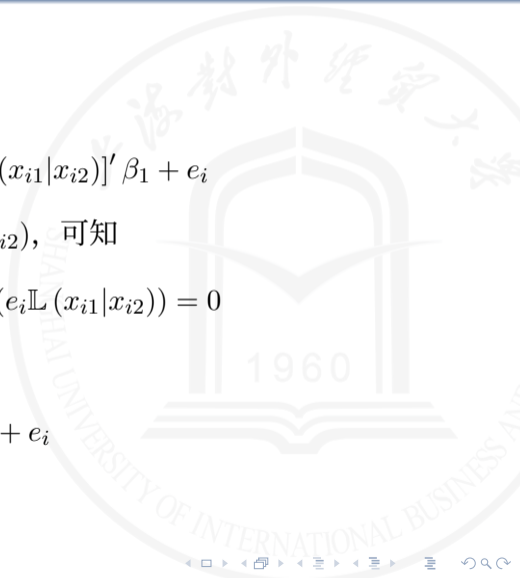
$$y_i - \mathbb{L}(y_i|x_{i2}) = [x_{i1} - \mathbb{L}(x_{i1}|x_{i2})]' \beta_1 + e_i$$

- 令 $e_{iy} = y_i - \mathbb{L}(y_i|x_{i2})$, $e_{i,x1} = x_{i1} - \mathbb{L}(x_{i1}|x_{i2})$, 可知

$$\mathbb{E}(e_i e_{i,x1}) = \mathbb{E}(e_i x_{i1}) - \mathbb{E}(e_i \mathbb{L}(x_{i1}|x_{i2})) = 0$$

- 从而相减后得到的式子可以写为:

$$e_{iy} = e'_{i,x1} \beta_1 + e_i$$

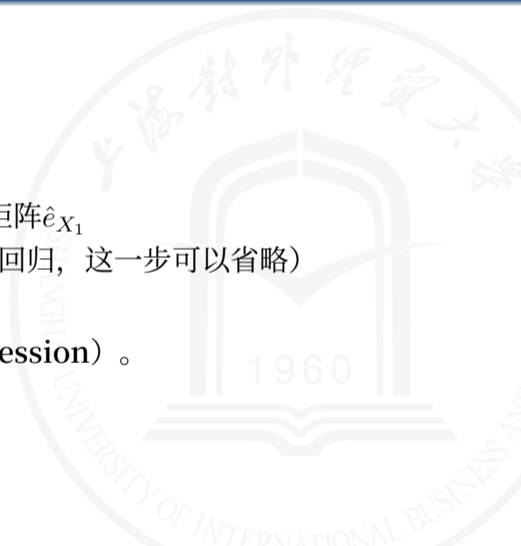


分步回归

为了计算 β_1 ，我们可以通过如下步骤进行计算：

- ① 使用对 X_1 的每一列对 X_2 做回归，得到残差矩阵 \hat{e}_{X_1}
- ② 使用 Y 对 X_2 做回归，得到残差 \hat{e}_y （对于线性回归，这一步可以省略）
- ③ 使用 \hat{e}_y 对 \hat{e}_{X_1} 做回归，得到系数 $\hat{\beta}_1$

如上的步骤被称为分步回归（partitioned regression）。



分步回归

Frisch-Waugh-Lovell定理

使用以上分步回归得到的 $\hat{\beta}_1$ 与式：

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + e_i$$

的最小二乘回归得到的 $\hat{\beta}_1$ 是完全等价的。

证明见讲义。

分步回归示例

消费与收入

partitioned_regression.do使用对数收入和对数资产、是否农村预测了对数消费。

- 如果直接进行回归，得到的对数收入的系数为0.2088
- 而使用分步回归得到的系数同样为0.2088，两者严格相等。
- 此外，代码中还计算了 \hat{e}_y 和 \hat{e}_{x_1} 的相关系数，即偏相关系数，这度量了资产规模、是否农村等变量之后对数收入与对数消费之间的相关性，约为0.3288，而不排除这些的相关系数为0.5224。

