

回归：预测与拟合

司继春

2025年8月



什么是计量经济学?

应用场景分类:

- ① 微观计量
- ② 宏观计量
- ③ 金融计量

工具方法分类:

- ① 简约式 (reduced form)
- ② 结构式 (structural form)



应用计量：如何学习？

我们的目标是：

- ① 看得懂：经典应用文献在说什么，语言的问题
- ② 做得出：能够复制出经典文献的结果
- ③ 会解释：知道软件中的结果该怎么解释
- ④ 想得到：能够将方法与技术完美的结合，找到合适的选题
- ⑤ 写得妙：如何motivate你的选题、模型？细节决定成败！

所以我们需要：

- ① 掌握计量经济学的语言体系
- ② 理解识别（identification），掌握基本的识别策略
- ③ 该如何估计
- ④ 软件怎么操作

微观计量经济学的任务

Heckman(2008)提出，因果推断的三个任务：

- ① 定义问题、假说或者反事实——逻辑、理论与想象
- ② 从总体数据中识别参数——如果没有抽样偏差，能否还原反事实？
- ③ 从实际数据中识别参数——统计理论，估计与检验

「Effect of Causes」 versus 「Causes of Effects」

实证论文的结构

- ① 导言
- ② 制度背景
- ③ 数据与描述性统计
- ④ 实证模型与结果
- ⑤ 稳健性检验
- ⑥ 结论



预测问题

- 拟合 (fitting) 以及预测是最经典的统计问题之一，而回归 (regression) 是解决这类问题最常用的手段。
- 如果我们观察到一系列数据 $(y_i, x_i), i = 1, \dots, N$ ，我们希望使用 x_i 的线性函数： $f(x_i) = \alpha + \beta x_i$ 对 y_i 进行预测，那么只要确定了其中的参数 α 和 β 就确定了这个预测的函数。
- 我们称：
 - x_i 为自变量 (independent variable) 或者解释变量 (explanatory variable)、回归元 (regressor)
 - 而 y_i 为因变量 (dependent variable) 或者被解释变量 (explained variable)、结果变量 (outcome variable)。

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。

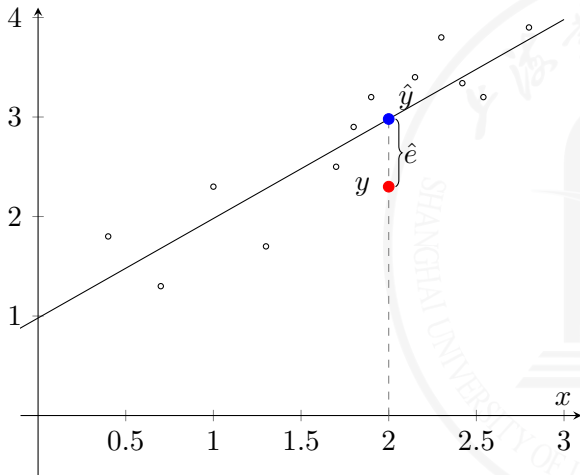
预测误差

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：

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

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
```

一元线性回归示例

收入与消费

在以上程序中：

- 首先剔除了收入和消费的异常值（收入小于0的值以及收入和消费存在删失问题数据）
- 接着使用reg命令计算了消费（total_consump）对收入（total_income）的回归，回归结果为：

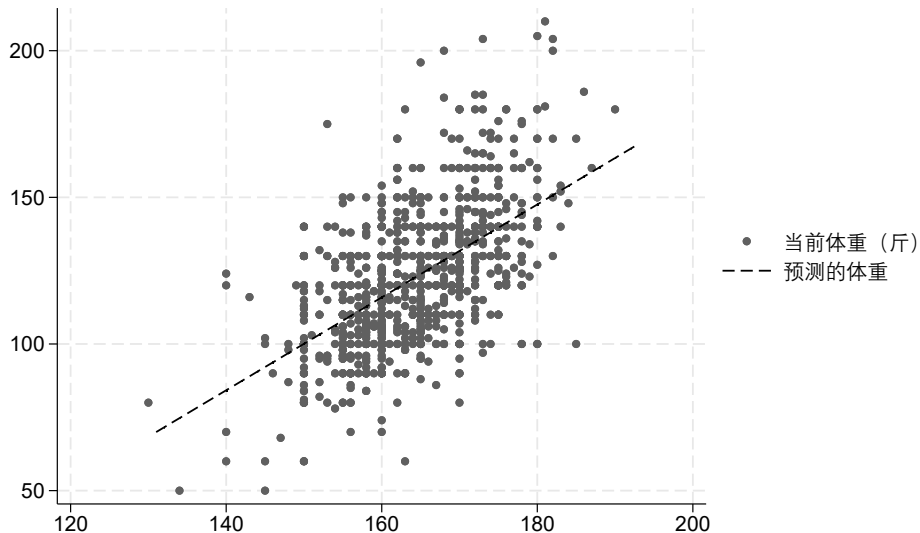
$$\hat{y} = 45696.74 + 0.15 \times x$$

意味着收入每增加1元，消费平均会增加0.15元。

- 此外，如果我们知道某个家庭收入为30万元，那么对其消费的最优预测为

$$\hat{y}_{175} = 45696.74 + 0.15 \times 300000 = 90696.74$$

一元线性回归示例



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

	(1)	(2)	(3)	(4)	(5)	(6)
	gender 0		gender 1		gender .	
VARIABLES	N	mean	N	mean	N	mean
p_income	13,688	38,811	19,793	48,108	2	36,000

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：

$$\hat{y}_i = f(x_i|\beta) = \beta_1 x_{i1} + \beta_1^2 x_{2i}$$

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 $x_{2i} = x_{1i}^2$ ）。

多元线性回归

为了方便起见，我们一般用向量表述上述方程：

$$f(x_i) = x_i' \beta$$

其中：

$$x_i = \begin{pmatrix} 1 \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

为两个 K 维列向量。给定一个 β ，我们可以使用 $x_i' \beta$ 对 y_i 进行预测的预测值：

$$\hat{y}_i = x_i' \beta$$

以及预测的误差，即残差：

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

最小二乘法 (OLS)

如果我们记：

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

那么残差向量为：

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

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。

虚拟变量

分类变量加入回归中需要使用虚拟变量（dummy variables），即如果一个变量的取值范围为 $G_i = 1, 2, \dots, g$ ，我们可以相应的定义 g 个虚拟变量：

$$d_{ij} = \mathbb{1} \{G_i = j\} = \begin{cases} 1 & \text{if } G_i = j \\ 0 & \text{otherwise} \end{cases}$$

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 α_{2012} 代表教育程度，比如 $\alpha_{2012}=0$ 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

不同教育程度的收入

代码 3: 不同性别的收入对比

```
1 // file: reg_with_dummies.do
2 use datasets/chfs2017_ind.dta, clear
3 gen p_income = a3109*12
4 tab a2012, gen(edu)
5 reg p_income edu*
6 outreg2 using reg_with_dummies.tex, replace
7 reg p_income edu*, noconstant
8 outreg2 using reg_with_dummies.tex, append
9 reg p_income edu2-edu9
10 outreg2 using reg_with_dummies.tex, append
```

虚拟变量回归

不同教育程度的收入

VARIABLES	(1) p_income	(2) p_income	(3) p_income
edu1	-100,601*** (4,080)	26,116*** (1,493)	
edu2	-96,302*** (3,845)	30,415*** (605.8)	4,299*** (1,611)
edu3	-91,942*** (3,816)	34,775*** (380.2)	8,659*** (1,540)
edu4	-87,936*** (3,839)	38,781*** (564.9)	12,665*** (1,596)
edu5	-85,926*** (3,870)	40,791*** (749.1)	14,675*** (1,670)
edu6	-77,183*** (3,838)	49,534*** (561.0)	23,418*** (1,595)
edu7	-58,928*** (3,833)	67,789*** (524.4)	41,673*** (1,582)
edu8	-13,441*** (4,107)	113,276*** (1,567)	87,160*** (2,164)
o.edu9	-		
edu9		126,717*** (3,797)	100,601*** (4,080)
Constant	126,717*** (3,797)		26,116*** (1,493)
Observations	33 434	33 434	33 434

虚拟变量回归

不同教育程度的收入

- 如果一定要加入edu9这个虚拟变量，那么可以在reg命令后面加入noconstant选项，该选项即防止线性回归中包含常数项，从而我们可以包含edu9这个变量。
- 实际上，如果包含edu9而不包含常数项，那么估计的系数就是每个分组的收入的平均值，比如，edu1的系数为26115.86，意味着文化程度为文盲的平均收入为26115.86元。
- 而如果包含常数项而把edu9忽略掉，那么edu1-edu8估计的系数即每个组的收入与edu9这个组（基准组）的差异，比如edu1的系数为-100600.7，那么意味着文化程度为文盲的平均收入比文化程度为博士的平均收入低100600.7元。
- 第3列同理，如果把edu1去掉，那么edu2-edu9的系数都是与edu1组（基准组）相比的收入差异。

参数回归：条件期望的估计

回忆条件期望的定义：

$$\mathbb{E}(y|x) = \arg \min_{h \in L^2} \left\{ \mathbb{E} \left[(y - h(x))^2 \right] \right\}$$

如果假设函数 $h(x) = x'\beta$ ，即 x 的一个线性函数，那么上式变为：

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

从而 $\mathbb{E}(y|x) = x'\beta_0$ ，使用样本平均代 $(\frac{1}{N} \sum)$ 替总体期望 $(\mathbb{E})$ ，得到：

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

即普通最小二乘估计量 (ordinary least squares)。

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：

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

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。

是并没有对 $\mathbb{E}(e_i^2|x_i)$ 做任何假设，
的任意函数。
们称 e_i 具有异方
们称 e_i 具有同方
方差的存在意味着 x_i 对 e_i 的方差仍然

$$\mathbb{V}(e_i|x_i) = \mathbb{V}(e_i|x_i)$$

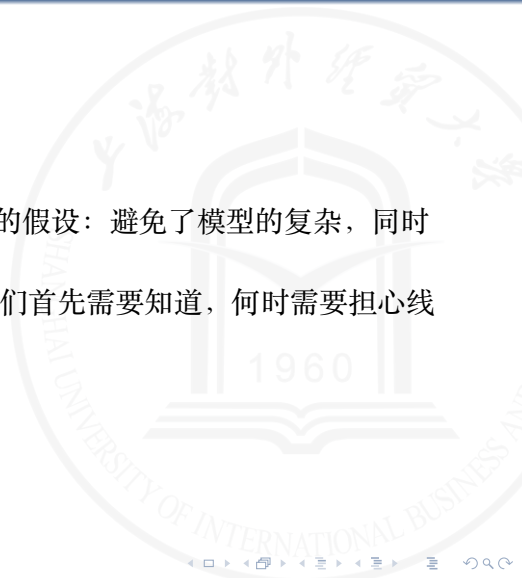
- 即 $y_i|x_i$ 的条件方差也就是 $e_i|x_i$ 的条件方差。

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

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

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

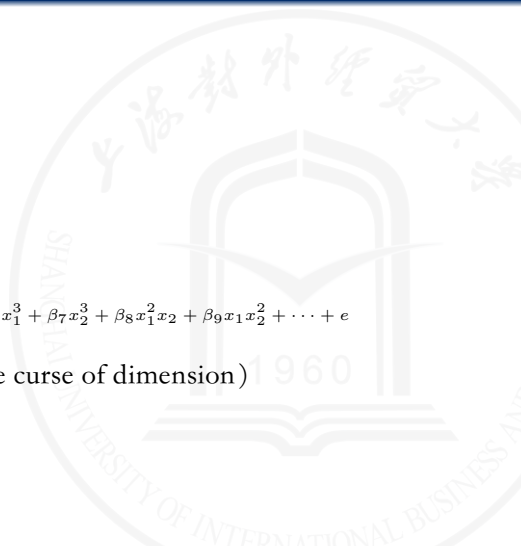
解决方案：

- 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)：

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

- 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

- 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

- 虽然上述方法被广泛应用，然而这些方法是不严谨的。
 - 原本对数变换的一个优良性质的可以“去量纲”，即不同量纲取对数只差一个常数
 - 例如收入 x 如果以“万元”为单位，那么 $10000x$ 就是以“元”为单位，取对数后：

$$\ln(10000x) = \ln 10000 + \ln x$$

两者只差一个常数。

- 然而如果使用以上 $\ln(1+x)$ 的方法，该“对数”不再有此性质：

$$\ln(1+10000x) - \ln(1+x) = \ln \frac{1+10000x}{1+x} \neq C$$

- 为何是 $\ln(1+x)$ 而不是 $\ln(0.1+x)$ 或者 $\ln(10000+x)$?

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：

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

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

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。

