回归：预测与拟合

司继春

2023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。

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$(y_i, x_i), i = 1, ..., N$，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$\alpha$和$\beta$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：身高与体重

一个非常传统的问题是个人身高与体重的关系，一般而言我们会用身高对体重进行预测，即使用身高的一个函数：

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

对体重进行预测。只要我们知道了α, β，对于一个身高已知、体重未知的人，我们就可以使用以上函数对未知的体重进行预测。

回归与插值

我们需要区分函数关系与相关关系：

- 如果给定一个输入，有确定的输出，那么两个变量是函数关系，比如：

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

- 如果给定一个输入，可能有不同的值与之相对应，那么是相关关系，比如身高都为170的人，体重完全可能不同

在此基础上：

- 回归需要解决的是相关关系，我们使用一个具有确定性的函数输出对不确定的y进行预测
- 如果一个函数是未知的，为了确定这个函数，解决方案为插值（interpolation）而非回归。

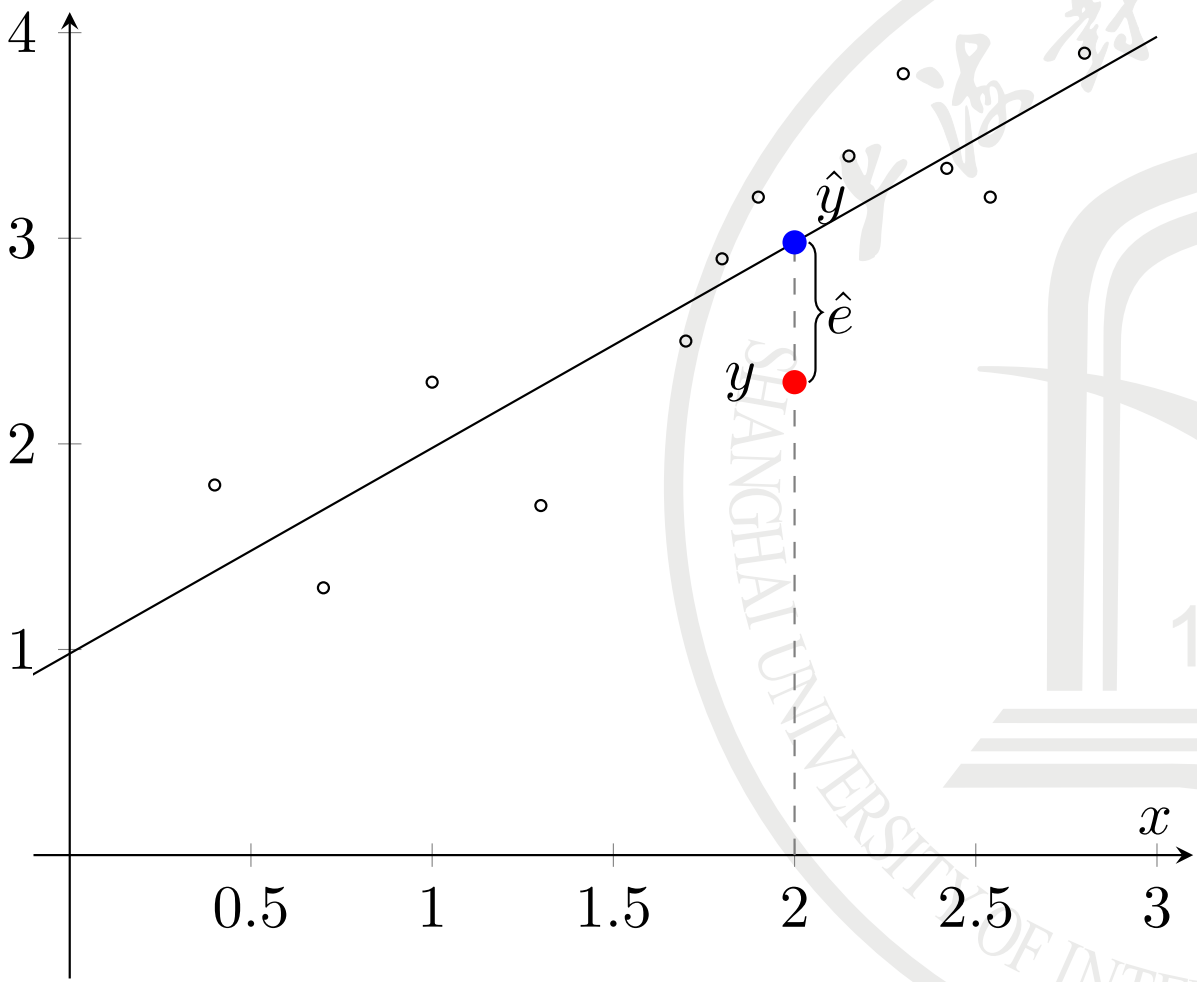
预测误差

如果给定一个α和β的值$\left(\tilde{\alpha}, \tilde{\beta}\right)$，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：

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

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

身高与体重

在下面的程序中，我们使用2014年CFPS的数据，使用体重对身高做简单的一元线性回归，回归结果为：

$$\hat{y} = -128.2 + 1.528 \times x$$

由于身高的单位为厘米，体重的单位为斤，所以该回归结果意味着，身高每增加1cm，平均而言体重会增加大约1.528斤。此外，如果我们知道某个人身高为175cm，而不知道其具体身高，那么对其身高的最优预测为：

$$\hat{y}_{175} = 1.528 \times 175 - 128.2 = 139$$

即身高175cm的人平均身高为139斤。

一元线性回归示例

身高与体重

```
1  // file: reg_one_variate.do
2  use datasets/cfps_adult, clear
3  drop if qp102<0
4  drop if qp101<130
5  reg qp102 qp101
6  outreg2 using reg_one_variate.tex, replace
7  predict p_weight
8  label variable p_weight "预测的体重"
9  sort qp101
10 twoway (scatter qp102 qp101 if mod(_n,30)==20) (
       line p_weight qp101 if qp101>130 & qp101<200),
       xscale(range(130 200))
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：

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

- 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2014年CFPS的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。

- 我们首先使用剔除了收入的异常值（即个人收入<0的观测）
- 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
- 可以看到，女性平均收入为5751元，而男性平均收入为12287元，男性收入比女性多了6536元。

虚拟变量回归与均值比较

不同性别收入比较

代码 1: 不同性别的收入对比

```
1 // file: reg_with_dummy.do
2 use datasets/cfps_adult, clear
3 keep cfps_gender p_income
4 drop if p_income<0
5 bysort cfps_gender: outreg2 using reg_with_dummy_su
      .tex, replace sum(log) eqkeep(N mean) keep(
      p_income)
6 reg p_income cfps_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)
	cfps_gender 0		cfps_gender 1	
VARIABLES	N	mean	N	mean
p_income	18,308	5,751	18,398	12,287

虚拟变量回归与均值比较

不同性别收入比较

	(1)
VARIABLES	p_income
cfps_gender	6,536***
	(194.3)
Constant	5,751***
	(137.5)
Observations	36,706
R-squared	0.030

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：

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

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$f\left(x_i,|\beta\right)$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：

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

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：

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

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 coliearity）

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$\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$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：

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

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

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$G_i = 1, 2, ..., g$，我们可以相应的定义g个虚拟变量：

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

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，$d_{i1}, ..., d_{ig}$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$d_{i1}, ..., d_{ig}$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

我们同样使用2014年CFPS数据，对不同教育程度的收入进行分解。在数据集中，变量te4代表教育程度，比如te4=0时表示文盲，te4=1代表小学等等，te4总共有7个可能的取值（文化程度）。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 2: 不同性别的收入对比

```
// file: reg_with_dummies.do
use datasets/cfps_adult, clear
drop if p_income<0
drop if te4<0
tab te4, gen(edu)
reg p_income edu*
outreg2 using reg_with_dummies.tex, replace
reg p_income edu*, noconstant
outreg2 using reg_with_dummies.tex, append
reg p_income edu2-edu7
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	-33,868***	8,211***	
	(6,705)	(1,092)	
edu2	-28,200***	13,879***	5,669***
	(6,661)	(781.4)	(1,343)
edu3	-27,527***	14,551***	6,341***
	(6,638)	(554.9)	(1,225)
edu4	-27,441***	14,638***	6,428***
	(6,670)	(850.1)	(1,384)
edu5	-18,867***	23,212***	15,002***
	(6,743)	(1,306)	(1,702)
edu6	-17,647***	24,432***	16,221***
	(6,773)	(1,451)	(1,817)
o.edu7	-		
edu7		42,079***	33,868***
		(6,615)	(6,705)
Constant	42,079***		8,211***
	(6,615)		(1,092)
Observations	3,226	3,226	3,226
R-squared	0.042	0.383	0.042

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

虚拟变量回归

- 如果一定要加入edu7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edu7这个变量。
- 如果包含edu7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，
- 如果包含常数项而把edu7忽略掉，那么edu1-edu7估计的系数即每个组的收入与edu7这个组（基准组）的差异
  - 比如edu1的系数为-33868，那么意味着文化程度为文盲的平均收入比文化程度为硕士的平均收入低33868元。
  - 第3列同理，如果把edu1去掉，那么edu2-edu7的系数都是与edu1组（基准组）相比的收入差异。

虚拟变量回归

以上的结果并非偶然。如果在回归中不加入常数项而是加入所有的分组虚拟变量，不失一般性，我们将所有的观测按照G_i进行排序，那么X应该是一个分块对角矩阵：

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

其中N_j为$G_i = j$组的观测个数。

虚拟变量回归

使用分块矩阵的乘法：

$$X'X = \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N_g \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum_{G_i=1} y_i \\ \sum_{G_i=2} y_i \\ \vdots \\ \sum_{G_i=g} y_i \end{bmatrix}$$

从而，最小二乘估计量：

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

虚拟变量回归

如果我们包含常数项，而忽略了d_1，只保留$d_2, ..., d_g$，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：

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

虚拟变量回归

从而：

$$\begin{cases} \bar{y}_1 &= \tilde{\beta}_1 \\ \bar{y}_2 &= \tilde{\beta}_1 + \tilde{\beta}_2 \\ \vdots & \\ \bar{y}_g &= \tilde{\beta}_1 + \tilde{\beta}_g \end{cases}$$

或者等价的：

$$\begin{cases} \tilde{\beta}_1 &= \bar{y}_1 \\ \tilde{\beta}_2 &= \bar{y}_2 - \bar{y}_1 \\ \vdots & \\ \tilde{\beta}_g &= \bar{y}_g - \bar{y}_1 \end{cases}$$

参数回归：条件期望的估计

回忆条件期望的定义：

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

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：

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

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）：

$$u_i = y_i - x_i'\beta_0 = y_i - \mathbb{E}\left(y_i|x_i\right)$$

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：$\hat{u}_i = y_i - x_i'\hat{\beta}$定义中使用的是估计值$\hat{\beta}$。

误差项

- 根据u_i的定义，有：

$$\mathbb{E}\left(u_i|x_i\right)=\mathbb{E}\left(y_i-\mathbb{E}\left(y_i|x_i\right)|x_i\right)=\mathbb{E}\left(y_i|x_i\right)-\mathbb{E}\left(y_i|x_i\right)=0$$

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：

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(x_i,u_i\right)&=\mathbb{E}\left(x_iu_i\right)-\mathbb{E}\left(x_i\right)\mathbb{E}\left(u_i\right)\\&=\mathbb{E}\left[\mathbb{E}\left(x_iu_i|x_i\right)\right]-\mathbb{E}\left(x_i\right)\mathbb{E}\left[\mathbb{E}\left(u_i|x_i\right)\right]\\&=\mathbb{E}\left[x_i\mathbb{E}\left(u_i|x_i\right)\right]=0\end{aligned}$$

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：

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

- 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u_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

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

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

即$y_i|x_i$的条件方差也就是$u_i|x_i$的条件方差。

异方差

异方差示例

记y_i为学生i的分数，$d_i = 0/1$代表性别的虚拟变量，假设：

$$\mathbb{E}\left(y_i|d_i\right) = 85 + 0 \times d_i = 85$$

即男生和女生的平均成绩相同，都是85分，然而如果记误差项：$u_i = y_i - 85$，那么：

$$\mathbb{V}\left(u_i|d_i\right) = 40 + 10d_i$$

是完全有可能出现的。如果：

$$\begin{cases} \mathbb{E}\left(y_i|d_i\right) = 85 - 5 \times d_i \\ \mathbb{V}\left(y_i|d_i\right) = 40 + 10d_i \end{cases}$$

即男生平均分比女生低5分，但是同时异方差仍然存在，这当然也是可能的情况，虽然此时$u_i = y_i - 85 - 5d_i$与d_i之间是均值独立的。

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$(x_i', y_i)', i = 1, 2, ..., N$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$\mathbb{V}(y_i|x_i) = \sigma^2(x_i)$不为常数
- 然而同分布则意味着无条件方差$\mathbb{V}(y_i)$不随i的变化而变化，两者是不矛盾的。

OLS的统计性质

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

其中$u = [u_1, ..., u_N]'$。

无偏性

进一步，有：

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

一致性

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

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

由于矩阵求逆为连续映射，因而：

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

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$(y_i, x_i')', i = 1, ..., N$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：

$$y_i|x_i \sim N\left(x_i'\beta_0, \sigma^2\right)$$

或者等价地：

$$Y|X \sim N\left(X\beta_0, \sigma^2 I\right)$$

以上假设等价于假设误差项$u_i|x_i \sim N\left(0, \sigma^2\right)$，或者$u|X \sim N\left(0, \sigma^2 I\right)$。

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：

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

其中$d_i = 0/1$为虚拟变量，且假设$P(d_i = 1) = 0.3$，$u_i|x_i \sim N(0, 4)$，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：

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

因而条件似然函数为：

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

最大化以上函数，得到：

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

其中$\hat{u}_i=y_i-x_i'\hat{\beta}$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：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{K-1} x_{K-1} + u$$

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：

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

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），我们知道储蓄率的取值范围应该在0到1之间。此时，如果我们选取家庭资产规模（wealth）作为解释变量：

$$saving\_rate_i = \beta_0 + \beta_1 \cdot wealth_i + u_i$$

由于家庭资产规模的取值范围应为$wealth_i \in [0, \infty)$，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：

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

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：

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

从而条件期望函数：

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

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）：

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

- 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：

$$x \to \ln(x)$$

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$saving_rate = \frac{saving}{income}$，如果我们将其取对数：

$$\ln saving\_rate = \ln saving - \ln income$$

从而如果将之前回归的被解释变量和解释变量取对数，即：

$$\ln saving\_rate_i = \beta_0 + \beta_1 \cdot \ln income_i + u_i$$

等价于：

$$\ln saving_i - \ln income_i = \beta_0 + \beta_1 \cdot \ln income_i + u_i$$

- 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：

$$\ln saving_i = \delta_0 + \delta_1 \cdot \ln income_i + u_i$$

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

- 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）。
- 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:

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

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

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$x \to \infty\,(-\infty)$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。

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：

$$\ln \frac{saving\_rate_i}{1 - saving\_rate_i} = \beta_0 + \beta_1 \cdot \ln(wealth_i) + u_i$$

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：

$$y_i = \mathbb{E}\left(y_i|x_i\right) + u_i$$

而最小二乘法的目标函数可以写为：

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

条件期望的线性逼近

由于

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

从而：

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

其中第一项跟β无关，因而最小化$\mathbb{E}\left[\left(y_i-x_i'\beta\right)^2\right]$等价于最小化$\left(\mathbb{E}\left(y_i|x_i\right)-x_i'\beta\right)^2$，即$x_i'\beta_0$是条件期望函数$\mathbb{E}\left(y_i|x_i\right)$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：

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

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$u = y - \mathbb{L}(y|x)$，那么$\mathbb{E}(ux) = 0$，且$\mathbb{L}(u|x) = 0$
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：

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

因而

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

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

对数的预测

- 比如，如果我们设定如下方程：

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

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$\exp\left(x'\hat{\beta}\right)$对$y$进行预测，会低估$y$的条件期望。

对数的预测

- 注意到：$y = e^{x'\beta}e^{u}$从而：$\mathbb{E}(y|x) = e^{x'\beta}\mathbb{E}(e^{u}|x)$
- 如果假设u和x独立且$u \sim N(0, \sigma^2)$，那么

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

从而：

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

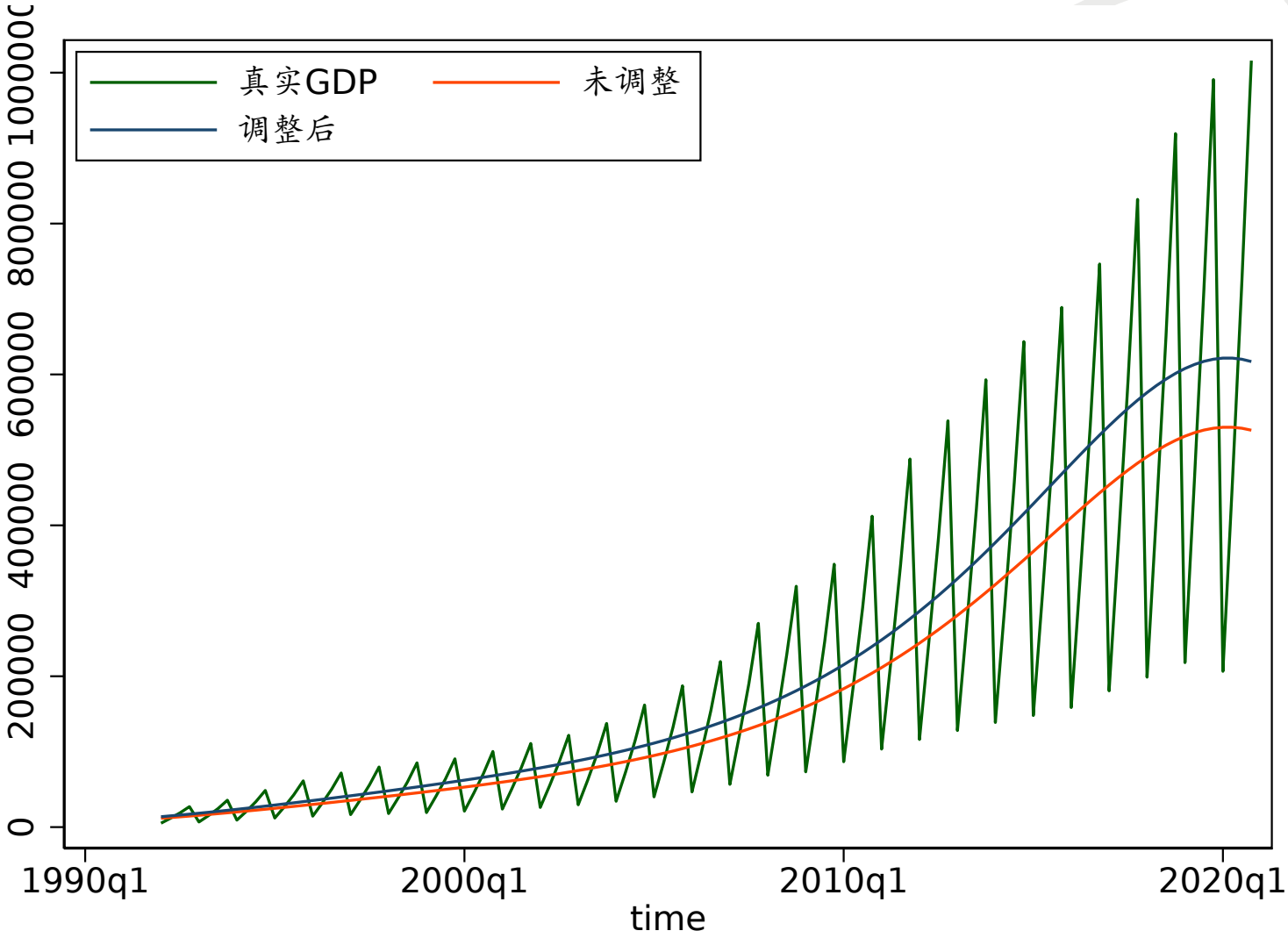
将β和σ^2使用极大似然回归结果，替代即可得到y的条件期望的预测值。

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$e^{x'\beta}$和$e^{x'\beta+\frac{\sigma^2}{2}}$两种办法对GDP进行了预测（代码：predict_log.do）

预测GDP


真实GDP
未调整
调整后
0
200000
400000
600000
800000
1000000
1990q1
2000q1
2010q1
2020q1
time

分步回归

- 对于总体回归：

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

如果我们将解释变量x_i分为两部分：$x_i = [x_{1i}', x_{2i}']'$，那么总体回归方程可以写为：

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

- 如果我们将以上方程两边同时对x_{i2}求条件期望，得到：

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

在总体回归方程两边同时减去上式，得到：

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

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：

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

- 令$e_{iy} = y_i - \mathbb{L}\left(y_i|x_{i2}\right), e_{i,x_1} = x_{i1} - \mathbb{L}\left(x_{i1}|x_{i2}\right)$，可知

$$\mathbb{E}\left(u_i e_{i,x_1}\right) = \mathbb{E}\left(u_i x_{i1}\right) - \mathbb{E}\left(u_i \mathbb{L}\left(x_{i1}|x_{i2}\right)\right) = 0$$

- 从而相减后得到的式子可以写为：

$$e_{iy} = e'_{i,x_1}\beta_1 + u_i$$

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

定理

使用以上分步回归得到的$\hat{\beta}_1$与式：

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

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

克强指数

为了展示分步回归的步骤，我们使用代码partioned_regression.do复制了“克强指数”，即使用工业用电量新增、铁路货运量新增和银行贷款新增拟合GDP增长，如果直接进行回归，得到的用电量新增的系数约为0.1637，而使用分步回归得到的用电量新增的系数同样为0.1637，两者严格相等。此外，我们还计算了$\hat{e}_y$和$\hat{e}_{x_1}$的相关系数，即偏相关系数，这度量了排除了铁路货运量新增、银行贷款新增的影响后，工业用电量新增与GDP增长的“净相关性”，大约为0.3861，表现了很强的相关性。

拟合优度

- 拟合优度（goodness of fit）即使用x对y进行拟合的效果，即被解释变量y有多少可以被解释变量x所解释。
- 在回归分析中，最常用的拟合优度的度量为可决系数（coefficient of determination），或称R^2（R-quared）。

总平方和的分解

- 问题：y的方差中有多少是可以被x解释的。
- 记y的方差的分子为总平方和（total sum of squares）：

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 = Y' M_0 Y$$

其中$M_0 = I - \frac{1}{N}\iota\iota'$。

- 分解（过程见讲义）：

$$TSS = ESS + RSS$$

其中：

$$ESS = \hat{Y}' M_0 \hat{Y} = \sum_{i=1}^{N} \left(\hat{y}_i - \bar{\hat{y}}\right)^2 = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

$$RSS = \hat{e}'\hat{e} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

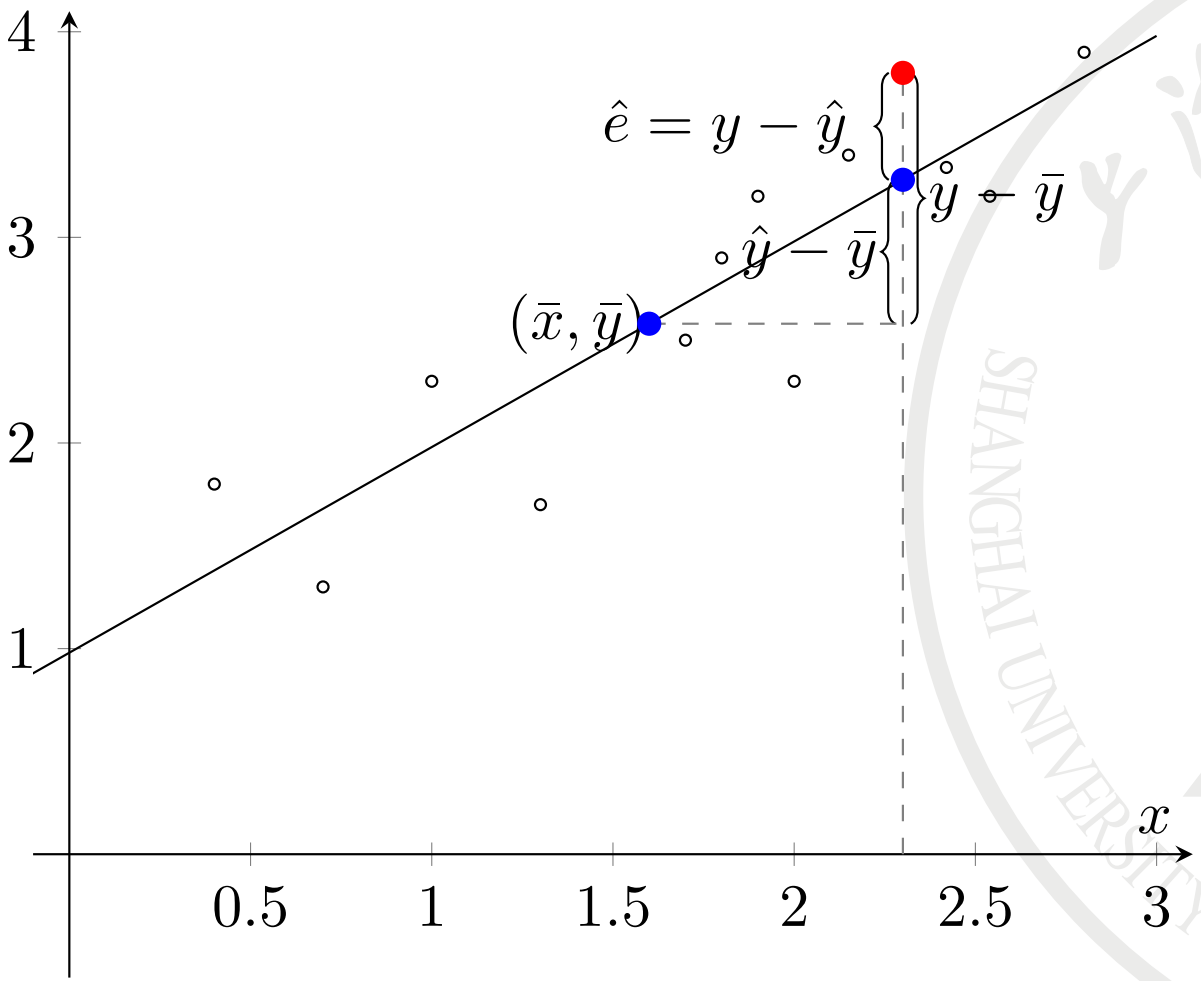
总平方和的分解

$$ESS = \hat{Y}' M_0 \hat{Y} = \sum_{i=1}^{N} \left(\hat{y}_i - \bar{\hat{y}}\right)^2 = \sum_{i=1}^{N} \left(\hat{y}_i - \bar{y}\right)^2$$

$$RSS = \hat{e}'\hat{e} = \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2$$

- 回归平方和（explained sum of squares）：x可以解释的部分
- 残差平方和（residual sum of squares）：x不能解释的部分

总平方和的分解



R^2的定义

定义：

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{Y}' M_0 \hat{Y}}{Y' M_0 Y} = 1 - \frac{\hat{e}'\hat{e}}{Y' M_0 Y} = 1 - \frac{RSS}{TSS}$$

R^2度量了使用x对y进行预测时，x可以解释多少部分的y的方差。R^2一定是在$[0,1]$之间的。

- 当$R^2 = 0$时，对应于$RSS = TSS$，此时$\hat{y}_i = \bar{y}$，也就是说不管有没有x，对y的最优预测都是$\bar{y}$，意味着x对y完全没有任何解释能力。
- 而如果$R^2 = 1$，此时$ESS = TSS$，即$\hat{y}_i = y_i$，达到了完美拟合。

R^2的取值范围

- TSS完全可以看作是只包含常数项的RSS
- 如果回归中包含常数项，那么预测结果不可能比只有常数项更差，从而必然会有$RSS < TSS$。
- 但是，如果回归中不包含常数项，得到的预测可能效果会非常差（比只用常数项预测的结果还差），从而可能会出现$TSS < RSS$的情况，从而得到$R^2 < 0$。
- 所以，$R^2 \in [0, 1]$的前提条件是回归中包含常数项！

变量增加时的R^2

- 如果我们在回归中添加新的解释变量，由于出现了更多的信息，因而R^2值不会降低。
- 们假设x_1是一个$N \times 1$维列向量，而X_2为$N \times K$维其他解释变量，线性回归:

$$Y = \beta_1 x_1 + X_2 \beta_2 + u$$

的最小二乘估计分别为$\hat{\beta}_1, \hat{\beta}_2$。

- 令残差：$\hat{u} = Y - \hat{\beta}_1 x_1 - X_2 \hat{\beta}_2$，那么拟合优度

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{Y' M_0 Y}$$

变量增加时的R^2

- 如果只对X_2做回归，即：

$$Y = X_2\gamma + e$$

记其最小二乘估计为$\hat{\gamma}$，残差为$\hat{e} = Y - X_2\hat{\gamma}$，那么其$R^2$为：

$$R_2^2 = 1 - \frac{\hat{e}'\hat{e}}{Y'M_0Y}$$

变量增加时的R^2

定理

在上述回归模型中，$R_2^2 \leq R^2$。

目标函数为：

$$\min_{\beta_1,\beta_2} (Y - \beta_1 x_1 + X_2\beta_2)' (Y - \beta_1 x_1 + X_2\beta_2)$$

而如果只对X_2做回归，目标函数为：

$$\min_{\gamma} (Y - X_2\gamma)' (Y - X_2\gamma)$$

相当于限制$\beta_2 = \gamma, \beta_1 = 0$，从而必然有：$\hat{u}'\hat{u} \leq \hat{e}'\hat{e}$

变量增加时的R^2

定理

R^2和R_2^2满足：

$$R^2 = R_2^2 + \left[\mathbb{C}\text{orr}\left(\hat{e}, \hat{\epsilon}\right)\right]^2 \left(1 - R_2^2\right)$$

其中$\hat{\epsilon}$为回归$x_1 = X_2\delta + \epsilon$的残差，即$\hat{\epsilon} = x_1 - X_2\hat{\delta}$。

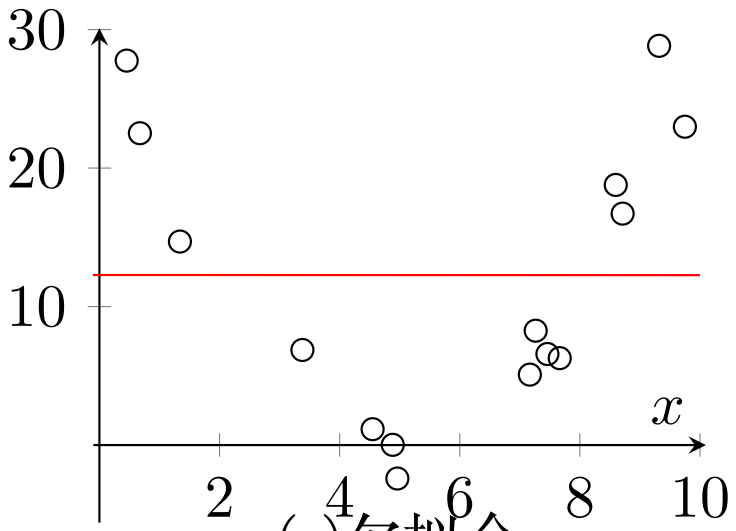
证明见讲义。其中$\mathbb{C}\text{orr}\left(\hat{e}, \hat{\epsilon}\right)$度量了$Y$和$x_1$同时不能被$X_2$解释的部分相关系数，即排除$X_2$影响之后的相关系数，因而也被称为$Y$和$x_1$的偏相关系数（partial correlation）。

模型选择

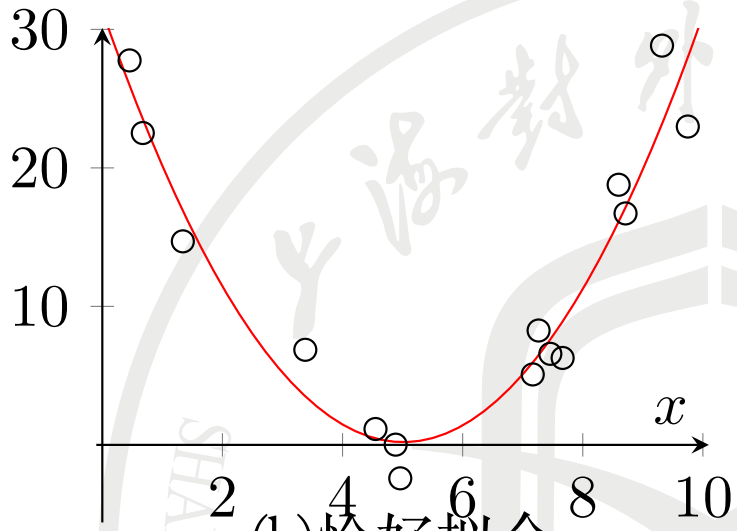
从预测的角度，我们不仅仅需要对样本内进行预测并达到比较好的预测效果，更希望当有新的样本进来时，也达到非常好的预测效果。通常有三种拟合情况：

- 欠拟合（under-fitting）：没有发现数据中本来有的pattern
  - 通常起源于太『小』的模型
  - 比如本该是二次方关系，错误设定为线性关系
  - 即使在现有样本内都表现不好，更不要说样本外预测
- 过拟合（over-fitting）：发现了本来没根本有的pattern
  - 通常起源于太『大』的模型
  - 比如本来应该是二次方关系，但是使用了10阶多项式拟合
  - 样本内预测非常好，但是样本外预测非常差
- 恰好拟合：具有非常好的样本内以及样本外预测效果

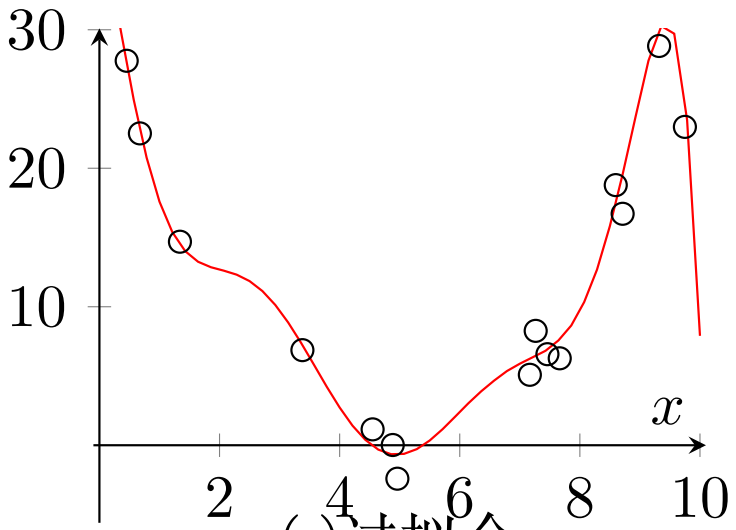
欠拟合、过拟合与恰好拟合



(a)欠拟合



(b)恰好拟合



(c)过拟合

过拟合问题

- 通常我们以R^2等指标为基础，更容易挑出R^2更大的模型来
- 但是R^2大并不代表样本外预测效果好，容易出现过拟合问题
- 过拟合会导致样本外预测效果很差
  - 我发现每次看国足都会都输球
  - 是否可以使用「我有没有看球」这个指标预测国足输赢?

模型选择

模型选择方法：

- 逐步回归法，准则包括：
  - $\bar{R}^2$
  - AIC
  - BIC
- 交叉验证（cross-validation）
- LASSO

问题：

1. 是否可以剔除不显著的变量？
2. 逐步回归方法是否正确？

调整后的R^2

- 拟合优度，$R^2 = 1 - \frac{\hat{u}'\hat{u}}{Y'M_0Y}$：
  - 度量了y中可预测部分（$\hat{y}$的方差）占总的方差（y的方差）的百分比
  - R^2越大，误差部分的方差越小，拟合效果越好
  - R^2大不代表得到了因果关系
- 调整后的R^2：$\bar{R}^2 = 1 - \frac{\hat{u}'\hat{u}/(N-K)}{Y'M_0Y/(N-1)}$
  - $\bar{R}^2$与R^2相比，分别使用了残差$\hat{u}$和y的方差的无偏估计
  - R^2随着解释变量的增加而单调变大，而$\bar{R}^2$不会

信息准则

- 信息准则
  - **赤池信息准则**（Akaike information creterion,**AIC**）：Akaike（1974）通过Kullback – Leibler距离，定义了如下信息准则：

$$AIC = -2Log\_Likelihood + 2K$$

  - **贝叶斯信息准则**（Bayesian information creterion,**BIC**）：Schwarz（1978）通过贝叶斯法则提出了如下信息准则：

$$BIC = -2Log\_Likelihood + \ln(N)\, K$$

线性回归中的信息准则

- 对于线性回归模型，如果使用极大似然估计，得到：

$$AIC = N \ln\left(\frac{\sum_{i=1}^{N} \hat{u}_i^2}{N}\right) + N + 2K$$

$$BIC = N \ln\left(\frac{\sum_{i=1}^{N} \hat{u}_i^2}{N}\right) + N + \ln(N) K$$

- 此外，小样本条件下可以使用调整的AIC，AICc（Hurvich和Tsai，1989）：

$$AIC_c = AIC + \frac{2K(K+1)}{N-K-1}$$

交叉验证

- 不管是欠拟合还是过拟合，都会导致样本外预测的误差变大，因而我们可以只使用一部分样本进行估计，而在另外一部分样本中检验模型的预测能力。
- 在交叉验证法中，我们将样本分为两部分：训练集（training set）用于估计模型、验证集（validation set）用于模型选择。交叉验证常见的用法如：
  1. S折交叉验证（S-fold cross validation）：将N个样本随机的分为大小相同的S组，然后利用S-1组的数据对数据进行拟合，并使用该模型对剩下的一组计算目标函数值（在这里即预测误差的度量，如误差平方）。将这一过程对S种组合重复进行，最终得到了N个目标函数值的加总（如均方误差）。对于不同的模型，选择使得验证集目标函数最优的那个，或者预测误差最小的那个模型。
  2. 留一验证（leave-one-out cross validation）：即S=N的S折交叉验证的特殊情形，每一次都用N-1个样本训练模型，对剩下的一个样本进行预测。

模型选择

多项式阶数选择

在以下程序中，我们首先产生了一个伪数据集：

$$y = e^x + u$$

其中$x \sim U(0,3)$，$u \sim N(0,4)$。接着，我们从x的一次方开始，逐渐向回归中添加$x^2, x^3, ..., x^{10}$对模型进行拟合，并计算R^2、$\bar{R}^2$、AIC、BIC以及留一验证的结果。
（model_selection.do cross_validation_reg.do）

模型选择

多项式阶数	(1) R^2	(2) $\bar{R}^2$	(3) AIC	(4) BIC	(5) AICC	(6) MSE(CV)
1	0.8230	0.8194	242.83	245.83	242.26	7.471
2	0.9244	0.9212	201.49	207.23	202.01	3.268
3	0.9296	0.9240	200.60	208.25	201.49	3.187
4	0.9300	0.9238	201.60	211.16	202.97	3.295
5	0.9302	0.9223	203.47	214.95	205.43	3.445
6	0.9303	0.9205	205.46	218.84	208.12	3.656
7	0.9317	0.9202	206.43	221.73	209.94	3.628
8	0.9319	0.9205	208.27	225.48	212.77	3.586
9	0.9319	0.9186	210.24	229.36	215.88	3.819
10	0.9319	0.9186	212.24	233.28	219.19	3.854

作业

- 1.6、1.7、1.8、1.11、1.14、1.15