线性回归：拟合与预测

司继春

上海对外经贸大学统计与信息学院

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。在这一节中，我们将讨论回归分析，特别是线性回归在拟合以及预测中的应用，以及回归与条件期望之间的关系。

1 一元线性回归

经典的一元线性回归即使用一个变量对另外一个变量进行拟合和预测。比如，我们可能希望使用一个人的身高对体重进行预测、使用一个人的中考成绩对考研成绩进行预测，或者使用性别对收入进行预测等等。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 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**）。

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**）：

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

如图 (1) 所示，红色点 y 为实际观察到的数据，我们使用图中直线进行拟合，那么当 $x = 2$ 时，预测的值为图中蓝色的点所对应，两者的差距即残差。

为了进行拟合，我们通常希望残差 $\tilde{e}_i$ 与 0 的距离越近越好。尽管有多种度量残差 $\tilde{e}_i$ 和 0 的距离的方法，然而最常用的方法是使用其平方：$\tilde{e}_i^2$。在这种距离的定义下，只有对于所有的 i 误差均为 0，即 $\tilde{e}_i = y_i - \tilde{\alpha} - \tilde{\beta} x_i = 0$ 时，对 y_i 的预测 $f(x_i) = \tilde{\alpha} + \tilde{\beta} x_i = y_i$，此时我们得到了完美的拟合。

然而现实中，完美拟合是非常罕见的，我们能够做的仅仅是使得平均误差最小化。如果我们最小化所有样本的残差 $\tilde{e}_i$ 的平方和，就得到了所谓的「**最小**

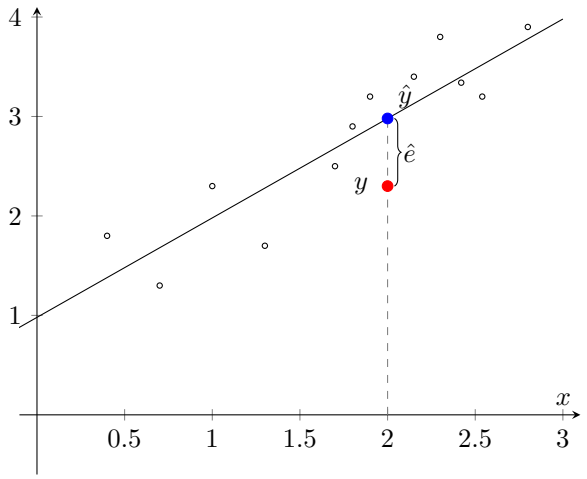


图 1: 预测值与残差

二乘法（Least squares）」:

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

求解上述最小化问题，可以对上述目标函数求导，并令导数等于 0，得到一阶条件为:

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

化简上述问题，得到:

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

继续化简，得到:

$$\bar{x}\bar{y}-\beta\bar{x}^2=\frac{1}{N}\sum_{i=1}^{N}x_iy_i-\beta\frac{1}{N}\sum_{i=1}^{N}x_i^2$$

因而解得:

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

至此，我们就得到了使用解释变量 x_i 对被解释变量 y_i 进行预测所需要的参数，α 和 β 的估计: $\hat{\alpha}$、$\hat{\beta}$, 我们称之为**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：

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

此外，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 $\bar{y}$。

例 1. 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：

代码 1: 一元线性回归示例

```
// file: reg_one_variate.do
use datasets/cfps_adult, clear
drop if qp102<0
drop if qp101<130
reg qp102 qp101
outreg2 using reg_one_variate.tex, replace
predict p_weight
label variable p_weight "预测的体重"
sort qp101
twoway (scatter qp102 qp101 if mod(_n,30)==20)/*
     */(line p_weight qp101 if qp101>130 & qp101<200),/*
     */ xscale(range(130 200))
graph export reg_one_variate.png, replace
```

在以上程序中，首先剔除了身高和体重的异常值（小于 0 的值），接着使用 reg 命令计算了体重 (qp102) 对身高 (qp101) 的回归，回归结果如表 (1) 所示。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 1.5 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。接下来我们使用 predict 命令计算了最小二乘的预测值 ($\hat{y}$)，并在同一张图上画出了数据的散点图和预测直线，如图 (2) 所示。可见身高和体重呈现了明显的正相关关系。

表 1: 身高与体重的关系

VARIABLES	(1) qp102
qp101	1.528***
	(0.0125)
Constant	-128.2***
	(2.055)
Observations	32,536
R-squared	0.315

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

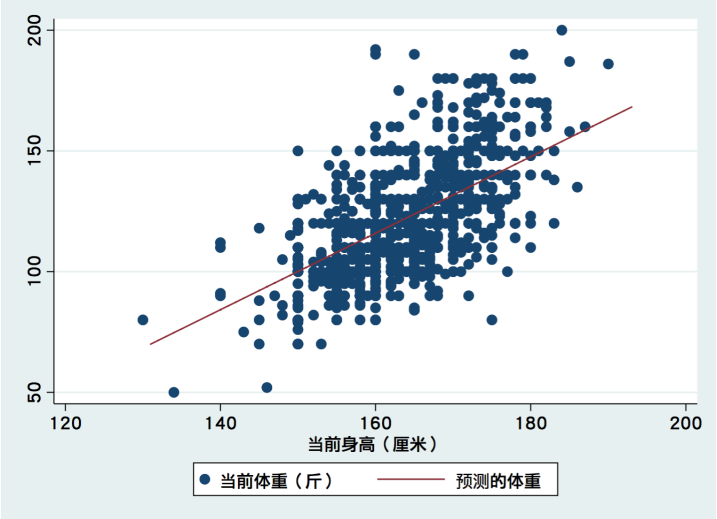


图 2: 身高与体重的关系

以上介绍了作为拟合的一元最小二乘法，实际上回归系数可以看成是简单的均值比较。如果以上回归方程中，x_i 只能取 0/1 两个值，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，同时记 $\bar{y}_1$ 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：

$$\begin{aligned}\hat{\beta} &= \frac{\frac{N_1}{N}\bar{y}_1 - \frac{N_1}{N}\bar{y}}{\frac{N_1}{N} - \left(\frac{N_1}{N}\right)^2} \\ &= \frac{\bar{y}_1 - \bar{y}}{1 - \frac{N_1}{N}} \\ &= \frac{\bar{y}_1 - \left(\frac{N_1}{N}\bar{y}_1 + \frac{N-N_1}{N}\bar{y}_0\right)}{1 - \frac{N_1}{N}} \\ &= \frac{\frac{N-N_1}{N}\bar{y}_1 + \frac{N-N_1}{N}\bar{y}_0}{1 - \frac{N_1}{N}} \\ &= \bar{y}_1 - \bar{y}_0\end{aligned}$$

因而实际上，如果 x_i 只能取 0/1 的值，那么使用 y 对 x 的回归实际上就是两组均值的比较。而同时：

$$\begin{aligned}\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ &= \frac{N_1}{N}\bar{y}_1 + \frac{N-N_1}{N}\bar{y}_0 - (\bar{y}_1 - \bar{y}_0)\frac{N_1}{N} \\ &= \frac{N-N_1}{N}\bar{y}_0 + \frac{N_1}{N}\bar{y}_0 \\ &= \bar{y}_0\end{aligned}$$

因而 $\hat{\alpha}$ 实际就是第 0 组的均值。当 $x_i = 0$ 时，有：

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

而当 $x_i = 1$ 时，有：

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

因而对于特定的 $x_i = 0/1$，其预测值就等于分组的平均值。

例 2. 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：

代码 2: 不同性别的收入对比

```
// file: reg_with_dummy.do
use datasets/cfps_adult, clear
keep cfps_gender p_income
drop if p_income<0
bysort cfps_gender: outreg2 using reg_with_dummy_su.tex,/*
```

表 2: 收入的描述性统计

	(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

表 3: 收入对性别的回归

	(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

```
6   */replace sum(log) eqkeep(N mean) keep(p_income)
7   reg p_income cfps_gender
8   outreg2 using reg_with_dummy.tex, replace
```

在以上代码中，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，结果如表 (2) 所示。从表中可以看到，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。

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。表 (3) 汇报了收入对性别回归的结果。根据以上的推测，在该回归中，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 5751 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 6536 元，这与我们的描述性统计的结果是相符的。

如果解释变量是连续的，同样也可以将回归系数看做均值比较。如图 (3) 所示，当自变量 x 从 1 增加到 1.5 时，或者 $\Delta x_0 = 0.5$，线性回归的系数 β 可以解释为随着 x 的增加，相应的 y 的增加为 $\Delta y_0 = \beta \cdot \Delta x_0$，或者平均而言，当 $x = 1.5$ 时 y 的均值比 $x = 1$ 时 y 的均值多 $\beta \cdot \Delta x_0$。

而线性回归中“线性”的含义是，当 x 的增量相等时，相应的 y 的增量也相

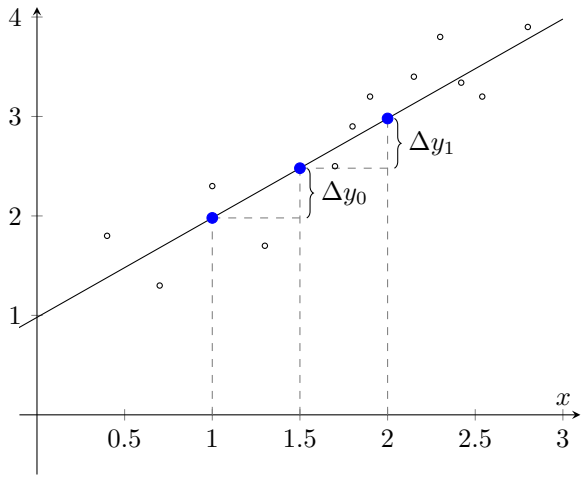


图 3: 预测值与残差

等。即当 $\Delta x_0 = \Delta x_1 = \Delta x$ 时，$\Delta y_1 = \Delta y_0 = \beta \cdot \Delta x$，或者：

$$\frac{\Delta y_1}{\Delta x_1} = \frac{\Delta y_0}{\Delta x_0} = \beta$$

2 多元线性回归

2.1 最小二乘

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：

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

其中 $x_i = (x_{i1}, ..., x_{iK})'$，$\beta = (\beta_1, ..., \beta_K)'$。一般而言，我们通常会保留常数项，不失一般性，我们一般令 $x_{i1} = 1$。我们在这里假设函数 $f\left(x_i, |\beta\right)$ 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：

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

同样的，我们称 y_i 为因变量或者被解释变量，而 x_{ik} 为自变量或者解释变量。为了方便起见，我们一般用向量表述上述方程：

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

其中：

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

为两个 K 维向量。与一元线性回归一样，给定一个 β，我们可以得到使用 $f(x_i) = x_i'\beta$ 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：$\hat{e}_i = y_i - \hat{y}_i = y_i - x_i'\beta$。

为了计算方便，我们记：

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}, X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix}_{N\times K} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1,K} \\ 1 & x_{22} & \cdots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \cdots & x_{N,K} \end{pmatrix}$$

因而残差向量为：

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

与一元线性回归一样，我们可以通过最小化残差的平方和 $\sum_{i=1}^{n} \hat{e}^2$，得到：

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

对以上目标函数求导数并令其等于 0，可以得到一阶条件：

$$\begin{aligned} \frac{\partial (Y - Xb)'(Y - Xb)}{\partial b} &= \frac{\partial (Y'Y - Y'Xb - b'X'Y + b'X'Xb)}{\partial b} \\ &= -X'Y - X'Y + 2X'Xb = 0 \end{aligned} \tag{3}$$

解以上方程可以得到：

$$X'Xb = X'Y \Rightarrow \hat{\beta} = (X'X)^{-1} X'Y \tag{4}$$

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：

$$\hat{\beta} = (X'X)^{-1} X'Y = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i y_i\right) \tag{5}$$

即为原最小化问题的解。我们称以上回归为**普通最小二乘回归**（**ordinary least squares**，**OLS**）。

注意以上我们使用了矩阵 $X'X$ 的逆矩阵，这就要求矩阵 $X'X$ 可逆。更进一步，由于 $\text{rank}(X'X) = \text{rank}(X)$，而 $X'X$ 为 $K \times K$ 维的矩阵，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 $N \geq K$）。因而为了使得式 (5) 成立，我们必须引入如下假设：

假设 1. （识别条件）矩阵 X 为列满秩矩阵，即 $rank(X) = K$。

矩阵 X 是列满秩的意味着：

1. X 的列数小于行数，即 $K < N$。
2. X 的任何一列不能被其他列线性表示出来。

以上第一个要求即要求解释变量的个数要小于样本量。实际上，如果 $K = N$，我们就得到了完美拟合，而如果 $K > N$，则会有不止一组参数达到完美拟合。因而在接下来的讨论中，默认条件下我们都假设 $K < N$。

以上那个的第二个要求意味着 X 的列之间不能存在着完美的线性关系，即不存在**完全共线性**（**perfect colinearity**）。例如，我们知道家庭收入（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，虽然这样做会带来解释上的困难。

以上假设被称为识别条件，是由于如果以上假设不满足，且 $N > K$，那么最小化目标函数 (2) 的解不唯一。实际上，在这里，如果识别条件不满足，将会有无穷多个解使得最小二乘目标函数最小化。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 $\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$ 是方程 (3) 的解，那么我们可以使用：

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$

对 y_i 进行预测。然而如果矩阵 X 不满秩，比如，$x_{i2} + x_{i3} = 1$，那么 $\left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$

不是方程 (3) 的唯一解。比如，我们令 $\hat{\beta}_2^* = \hat{\beta}_2 + c$，$c$ 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 $\left(\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*\right)$ 也是最优化问题 (2) 的解。为了保证最优化问题 (2) 有唯一解，必须保证矩阵 X 是满秩的。

另外一个更常见的例子是**虚拟变量**（**dummy variables**）的使用。在回归分析中，我们经常加入分类变量的虚拟变量，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：

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

比如，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

由于在以上的变量定义中，$\sum_{j=0}^{6} d_{ij} = 1$，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，$d_1, ..., d_6$ 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 $d_1, ..., d_6$ 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。

例 3. 在例2中，我们计算了不同性别的收入差异，即当分组变量 $G_i = 0, 1$ 时的回归。接下来我们同样使用 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：

表 4: 不同教育程度收入比较

VARIABLES	(1) p_income	(2) p_income
edu1	-33,868***	8,211***
	(6,705)	(1,092)
edu2	-28,200***	13,879***
	(6,661)	(781.4)
edu3	-27,527***	14,551***
	(6,638)	(554.9)
edu4	-27,441***	14,638***
	(6,670)	(850.1)
edu5	-18,867***	23,212***
	(6,743)	(1,306)
edu6	-17,647***	24,432***
	(6,773)	(1,451)
o.edu7	-	
edu7		42,079***
		(6,615)
Constant	42,079***	
	(6,615)	
Observations	3,226	3,226
R-squared	0.042	0.383

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

代码 3: 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
```

在以上程序中，我们使用 tab 命令产生了 te4 代表的不同教育程度的虚拟变量[1]，并使用个人收入对这些虚拟变量进行回归，回归结果如表 (4) 第一列所示。可以看到，为了保证矩阵可逆，Stata 自动忽略了 edu7 这个虚拟变量。

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 no-

[1]实际上也可以不用手动产生虚拟变量，而是在回归中直接使用 i.te4。

constant 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。实际上，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，比如，edu1 的系数为 8211，意味着文化程度为文盲的平均收入为 8211 元。而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。

实际上以上的结果并不是偶然。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。如此，使用分块矩阵的乘法：

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

$$X'Y = \begin{bmatrix} \iota' y_1 \\ \iota' y_2 \\ \vdots \\ \iota' y_g \end{bmatrix}$$

其中 y_j 为 $G_i = j$ 组的 y_i 的和。从而，最小二乘估计量：

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

即最小二乘估计为每个组的均值。

现在，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：

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

因而 $\hat{\beta} = Q\tilde{\beta}$，即：

$$\begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_g \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \vdots \\ \tilde{\beta}_g \end{bmatrix}$$

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：

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

因而在忽略虚拟变量 d_1 的情况下，常数项所估计的是 $G_i = 1$ 组的均值，而 $\tilde{\beta}_j$

估计的则是 $G_i = j$ 组的均值与 $G_i = 1$ 组的均值之差。

2.2 最小二乘的几何性质

如果我们需要获得 Y 的预测值，那么可以使用：

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

如果我们记 $P = X\left(X'X\right)^{-1}X'$，则 $\hat{Y} = PY$，即 P 矩阵将任意一个 N 维空间向量 Y 映射到其最小二乘的预测向量 $\hat{Y}$。注意由于：

$$\begin{aligned} P^2 &= X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' \\ &= X\left(X'X\right)^{-1}X' \\ &= P \end{aligned}$$

因而矩阵 P 为实对称投影矩阵。注意如果我们取出 X 矩阵的某一列 $X_{(j)} = XI_{(j)}$，其中

$$I_{(j)} = \begin{bmatrix} 0 & \\ 0 & \\ \vdots & \\ 1 & (j) \\ \vdots & \\ 0 & \end{bmatrix}$$

并将其使用 P 矩阵进行投影，那么：

$$\begin{aligned} PX_{(j)} &= PXI_{(j)} \\ &= X\left(X'X\right)^{-1}X'XI_{(j)} \\ &= XI_{(j)} \\ &= X_{(j)} \end{aligned}$$

即如果把 X 的某一列 $X_{(j)}$ 使用 P 进行投影，那么得到的投影仍然是 $X_{(j)}$ 本身。或者说，如果使用 X 预测 X 的某一列 $X_{(j)}$，那么预测值即 $X_{(j)}$ 本身。更进一步，对于任意的 X 的列向量的线性组合 $X\delta$，对其使用 P 进行投影，得到的都是 $X\delta$ 本身：

$$PX\delta = X\left(X'X\right)^{-1}X'X\delta = X\delta$$

特别的，由于我们假设回归中包含常数项，因而 ι 必然为 X 中的一列，因而必然有 $P\iota = \iota$。

同时，我们可以记残差为：

$$\hat{e} = Y - \hat{Y} = (I - P) Y$$

如果我们记 $M = I - P$，那么 M 矩阵将任意一个 N 维空间向量 y 映射到其最小二乘的残差向量 $\hat{e}$。我们可以计算残差的和：

$$\sum_{i=1}^{N} \hat{e}_i = \hat{e}'\iota = Y'M\iota = Y'(I - P)\iota = Y'(\iota - P\iota) = 0$$

因而残差之和必然为 0。由于 $y = \hat{y} + \hat{e}$，因而：

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} \hat{e}_i = \sum_{i=1}^{N} \hat{y}_i$$

上式意味着 y 的平均值等于 $\hat{y}$ 的平均值，即 $\bar{y} = \bar{\hat{y}}$。

注意 M 矩阵也为幂等矩阵：

$$M^2 = (I - P)(I - P) = I - P - P + P^2 = I - P = M$$

对于任意的 X 的列向量的线性组合 $X\delta$，对其使用 M 进行投影，得到的都是 0 向量：

$$MX\delta = (I - P) X\delta = X\delta - PX\delta = X\delta - X\delta = 0$$

最后，注意 $MP = (I - P) P = P - P^2 = 0$，同理 $PM = 0$，因而对于任意一个 N 维空间向量 y，有：

$$\hat{Y}'\hat{e} = (PY)'(MY) = Y'PMY = 0$$

即最小二乘得到的预测值向量与残差向量都是正交的。

因而我们可以把向量 y 分解为正交的两部分：

$$Y = PY + MY$$

且其长度满足「勾股定理」：

$$Y'Y = Y'PY + Y'MY = \hat{Y}'\hat{Y} + \hat{e}'\hat{e}$$

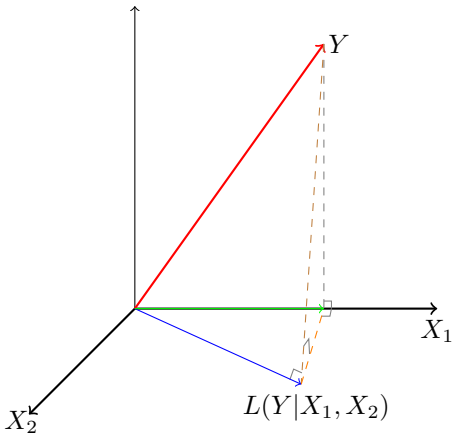


图 4: 最小二乘与投影

3 回归与条件期望

3.1 总体回归

线性回归与条件期望的概念密不可分。回忆一下，条件期望的定义即，对于随机变量 y 和随机向量 $x = (1, x_2, ..., x_K)'$，条件期望即使用 x 对 y 的最优预测：

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

假设 2. 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：

$$h(x) = x'\beta$$

在以上假设条件下，条件期望定义就变成了寻找 β 使得预测误差平方的期望（均方误差）最小化的过程：

$$\beta_0 = \arg\min_\beta \mathbb{E}\left([y - x'\beta]^2 \right) \tag{6}$$

因而条件期望 $\mathbb{E}(y|x) = x'\beta_0$，其中 β_0 为以上最小化问题的最优解，即条件期望的真实参数。

如果我们观察到一组样本：$(y_i, x_i')', i = 1, ..., N$，那么式 (6) 的样本等价形式为：

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

如此我们就得到了最小二乘的定义，因而 $\hat{\beta}$ 可以看成是 β_0 的估计，而 $x'\hat{\beta}$ 可

以看做条件期望 $\mathbb{E}(y|x)$ 的估计，即：

$$\widehat{\mathbb{E}(y_i|x_i)} = x_i'\hat{\beta}$$

现在我们定义**误差项**（**error term**）：

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

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：

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

定义中使用的是估计值 $\hat{\beta}$。根据 u_i 的定义，有：

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

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：

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

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：$\mathbb{E}(y_i|x_i) = x_i'\beta_0$ 和 u_i，且两者为相加的形式：

$$y_i = \mathbb{E}(y_i|x_i) + u_i = x_i'\beta_0 + u_i \tag{7}$$

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。

注意根据以上的定义，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 $\mathrm{Var}(u_i|x_i)$ 可以是 x_i 的任意函数。如果 $\mathrm{Var}(u_i|x_i) = \mathbb{E}(u_i^2|x_i)$ 不为常数，那么我们称 u_i 具有**异方差**（**heteroscedasticity**）；如果 $\mathrm{Var}(u_i|x_i) = \mathbb{E}(u_i^2|x_i) = \mathbb{E}(u_i^2)$，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。

3.2 最小二乘的统计性质

在上一小节中，我们介绍了总体回归方程，并将最小二乘估计 $\hat{\beta}$ 看成是总体回归方程中真值 β_0 的一个估计。在这一小节中，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。可以证明，最小二乘估计 $\hat{\beta}$ 是 β_0 的无偏且一致估计。

首先，将总体定义式 (7) 带入最小二乘估计量，有：

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

其中 $u = [u_1, ..., u_N]'$。

进一步，对于无偏性，我们有：

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

其中最后一个等号是由于：$\mathbb{E}(u_i|X) = 0$。在条件期望 $\mathbb{E}(y_i|x_i) = x_i'\beta_0$ 的线性假设条件下，最小二乘估计量 $\hat{\beta}$ 是 β_0 的无偏估计量。

类似地，我们也可以证明其为一致估计量，如果 (x_i, y_i) 是独立同分布的，

且 $\mathbb{E}\left(x_{ik}^2\right)<\infty$ 以及 $\mathbb{E}|x_{ik}u_i|<\infty, k=1,...,K$[2]，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：

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

从而 $\hat{\beta}\xrightarrow{p}\beta_0$，即 $\hat{\beta}$ 是 β_0 的一致估计量。

3.3 条件极大似然估计

在前面的讨论中，我们在假设2，即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，我们只对条件期望进行了假设，而没有对条件分布做任何假设。在这一节中，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。

假设 3. 设样本 $\left(y_i,x_i'\right)', i=1,...,N$ 为独立同分布的，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：

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

或者等价地：

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

实际上，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 $u|X\sim N\left(0,\sigma^2 I\right)$。因而相较于假设2，假设3额外假设了误差项 u_i 服从正态分布且同方差。

在以上条件分布的假设条件下，我们可以使用条件极大似然估计对总体参数 β_0 进行估计。在假设3的条件下，条件密度函数为：

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

因而条件似然函数为：

$$L\left(\beta,\sigma|y,x\right)=-\frac{N}{2}\ln\left(2\pi\right)-\frac{N}{2}\ln\sigma^2-\sum_{i=1}^{N}\frac{\left(y_i-x_i'\beta\right)}{2\sigma^2}\tag{8}$$

[2]在独立不同分布的条件下，需要使用相应的大数定律的假设。

最大化以上函数，得到：

$$\begin{cases} \hat{\beta} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i y_i\right) = (X'X)^{-1} X'Y \\ \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \left(y_i - x_i'\hat{\beta}\right)^2}{N} = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N} \end{cases} \tag{9}$$

其中 $\hat{u}_i$ 为残差。再次，我们得到了最小二乘估计量。

3.4 条件期望函数形式的讨论

在以上的讨论中，无论是假设2还是假设3，都假设了条件期望函数 $\mathbb{E}(y_i|x_i) = h(x_i)$ 是一个线性函数的形式。然而在对条件期望的定义中，条件期望函数 $h(x_i)$ 可以是任意的函数形式，并不受线性函数的限制。实际上，现实世界中变量之间的关系通常是非线性的，因而假设2往往很难保证成立。

例 4. （支撑集问题）线性函数假设最明显的问题是所谓的支撑集（support）问题。在这里，**支撑集**（**support**）即一个随机变量的取值范围[3]。比如，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：

$$saving\_rate_i = \beta_0 + \beta_1 \cdot wealth_i + u_i \tag{10}$$

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）。因而，使用家庭资产规模预测储蓄率时，真实的条件期望函数不可能是线性函数。

例 5. （经济增长）在经济增长理论中，索洛模型（Solow growth model）是标准的理论框架。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[4]。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：

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

[3]对于一个随机变量 $X : \Omega \to \mathbb{R}$，其支撑集 D 被定义为使得 $P(X \in D) = 1$ 成立的最小闭集。对于一个连续的随机变量，其支撑集即密度函数大于 0 的集合。

[4]由于当 $|r|$ 比较小时，$\ln(1+r) \approx r$，因而 $g_t = \ln y_t - \ln y_{t-1} = \ln\left(\frac{y_t}{y_{t-1}}\right) = \ln\left(1 + \frac{y_t - y_{t-1}}{y_{t-1}}\right) \approx \frac{y_t - y_{t-1}}{y_{t-1}}$。

假设 u_t 独立于 y_{t-1}，则上式化简为：

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

因而条件期望函数为一个指数函数形式，而非线性函数。

例 6. （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：

$$X_{ni}=GY_i^aY_n^b\phi_{ni}$$

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。再次，双边贸易额与 GDP 之间的关系并非简单的线性关系。

可见，现实中的变量，特别是经济变量之间的关系通常都是非线性的，因而线性回归可能不足以对条件期望函数进行准确的估计。如果条件期望函数的确是线性函数，那么我们称总体回归模型 (7) 是正确设定（correctly specified）的。

尽管我们永远无法知道条件期望函数的准确形式，然而我们通常可以对变量进行一些变换，使得我们的线性回归能够尽量逼近正确的条件期望函数。一些变换可以帮助我们对条件期望函数进行更好的建模，比如：

1. 对数变换。实际上，例5已经为我们展示了如何使用对数函数将一个非线性的关系变换成线性关系，类似的，例6也可以通过两边取对数转换成一个线性模型：
$$\ln X_{ni}=\ln G+a\ln Y_i+b\ln Y_n+\ln\phi_{ni}$$
在实际的数据建模中，对一些变量取对数是比较常用的操作。取对数的另外一个优点是其系数具有**弹性**（**elasticity**）的解释，我们将在后面详细介绍。

2. Box-Cox 变换。Box 和 Cox(1964) 提出了以下变换：
$$y\left(\lambda\right)=\begin{cases}\frac{y^\lambda-1}{\lambda} & \lambda\neq0\\ \ln y & \lambda=0\end{cases}$$
进而在回归中可以使用：
$$y_i\left(\lambda\right)=x_i'\beta+u_i$$
以上回归中，当 $\lambda=1$ 时即线性回归，当 $\lambda=0$ 时即 y 的对数的回归。以上变换经常要求 y 不能为负 Bickel 和 Doksum(1981) 提出了以下以下的 Box-Cox 变换：
$$y\left(\lambda\right)=\frac{|y|^\lambda\operatorname{Sign}\left(y\right)-1}{\lambda},if\ \lambda>0$$

其中：

$$\text{Sign}(y) = \begin{cases} 1 & if\ y \geq 0 \\ -1 & if\ y < 0 \end{cases}$$

然而在这个回归中，λ 为未知参数需要进行估计，因而不能用简单的线性回归估计 λ，可以使用极大似然估计等对 (λ, β) 进行联合估计。

3. Logistic 变换。使用函数：

$$f(x) = \frac{e^x}{1+e^x}$$

可以将 $(-\infty, \infty)$ 的实数映射到 $(0,1)$ 区间上，因而对于类似于式 (10) 的比率问题，我们可以使用：

$$saving\_rate_i = \frac{\exp\{\beta_0 + \beta_1 \cdot wealth_i + u_i\}}{1+\exp\{\beta_0 + \beta_1 \cdot wealth_i + u_i\}}$$

进行建模，或者等价的：

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

因而我们可以使用

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

将 $(0,1)$ 区间上的实数映射到 $(-\infty, \infty)$ 上，从而解决支撑集的问题。

当然，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。实际上，根据定义：

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

而最小二乘法的目标函数可以写为：

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

由于

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

从而：

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

其中第一项跟 β 无关，因而最小化 $\mathbb{E}\left[(y_i - x_i'\beta)^2\right]$ 等价于最小化 $(\mathbb{E}(y_i|x_i) - x_i'\beta)^2$，

因而根据定义式 (6)，回归参数即：

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

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。

当然，除简单的线性函数之外，我们还可以使用 x_i 的多项式对条件期望函数进行逼近。比如，如果有两个自变量 x_1 和 x_2，我们可以使用二次函数：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + u_i$$

对条件期望函数进行逼近等等。

最后，需要提示的是，当我们使用非线性变换时，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。

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：$\mathbb{E}\left(\ln y|x\right)$ 的最优线性估计，然而，根据 Jensen 不等式：

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

从而：

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

因而如果我们首先使用 $x'\hat{\beta}$ 对 $\ln y$ 进行预测，再使用 $\exp\left(x'\hat{\beta}\right)$ 对 y 进行预测，会低估 y 的条件期望。

对于以上问题，注意到：

$$y = e^{x'\beta}e^{u}$$

从而：

$$\mathbb{E}\left(y|x\right) = e^{x'\beta}\mathbb{E}\left(e^{u}|x\right)$$

如果假设 u 和 x 独立且 $u \sim N\left(0, \sigma^2\right)$，那么 $\mathbb{E}\left(e^{u}|x\right) = \mathbb{E}\left(e^{u}\right) = e^{\sigma^2/2}$，从而：

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

将 β 和 σ^2 使用极大似然回归结果，即式 (9) 替代即可得到 y 的条件期望的预测值。

4 分步回归

4.1 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：

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

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。

以上解释针对的是总体回归方程，而对于样本回归，即普通最小二乘回归，也有类似的结果。对于回归模型：

$$Y = X\beta + u$$

如果我们把 X 分为两部分变量：$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$，那么：

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

其中 $\begin{bmatrix} \beta_1' & \beta_2' \end{bmatrix}' = \beta$。如果对以上方程求解最小二乘，式 (4) 的一阶条件可以写为：

$$(X'X)\beta = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix} = X'Y$$

解以上方程，即：

$$\begin{cases} X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y \\ X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y \end{cases} \tag{11}$$

由第一个式子可以得到：

$$\hat{\beta}_1 = (X_1'X_1)^{-1}\left(X_1'Y - X_1'X_2\hat{\beta}_2\right) = (X_1'X_1)^{-1}X_1'\left(Y - X_2'\hat{\beta}_2\right)$$

即如果我们已经有了 β_2 的最小二乘估计值 $\hat{\beta}_2$，那么 $\hat{\beta}_1$ 的估计值等价于使用 $Y - X_2\hat{\beta}_2$ 整体对 X_1 做回归。将上式带入带入第二个式子：

$$\begin{aligned} X_2'X_2\hat{\beta}_2 &= X_2'Y - X_2'X_1\hat{\beta}_1 \\ &= X_2'Y - X_2'X_1(X_1'X_1)^{-1}\left(X_1'Y - X_1'X_2\hat{\beta}_2\right) \\ &= X_2'Y - X_2'X_1(X_1'X_1)^{-1}X_1'Y + X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 \end{aligned}$$

记 $P_1 = X_1(X_1'X_1)^{-1}X_1'$，则上式可以简记为：

$$X_2'X_2\hat{\beta}_2 - X_2'P_1X_2\hat{\beta}_2 = X_2'Y - X_2'P_1Y$$

整理得：

$$X_2'(I - P_1)X_2\hat{\beta}_2 = X_2'(I - P_1)Y$$

记 $M_1 = I - P_1$，那么：

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y$$

同理：

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$$

注意实际上 M_2X_1 即使用 X_1 的每一个列向量对 X_2 做回归，得到的残差所组成的矩阵，因而如果记：

$$\begin{cases} \hat{e}_{X_1} = M_2X_1 \\ \hat{e}_y = M_2Y \end{cases}$$

那么：

$$\begin{aligned} \hat{\beta}_1 &= (X_1'M_2X_1)^{-1}X_1'M_2Y \\ &= \left[(M_2X_1)'(M_2X_1)\right]^{-1}\left[(M_2X_1)(M_2Y)\right] \\ &= \left(\hat{e}_{X_1}'\hat{e}_{X_1}\right)^{-1}\hat{e}_{X_1}\hat{e}_y \end{aligned}$$

即如果我们对解释变量进行分组，$X=(X_1,X_2)$，那么 X_1 的系数 β_1 的最小二乘估计 $\hat{\beta}_1$ 等价于以下回归步骤得到的回归系数：

1. 使用对 X_1 对 X_2 做回归，得到残差 $\hat{e}_{X_1}$
2. 使用 Y 对 X_2 做回归，得到残差 $\hat{e}_y$
3. 使用 $\hat{e}_y$ 对 $\hat{e}_{X_1}$ 做回归，得到系数 $\hat{\beta}_1$

以上步骤与直接进行最小二乘估计是等价的。注意由于 M_2 为幂等矩阵，因而 $\hat{\beta}_1$ 也可以写为：

$$\begin{aligned}\hat{\beta}_1 &= \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 Y \\ &= \left(\hat{e}_{X_1}' \hat{e}_{X_1}\right)^{-1} \hat{e}_{X_1} Y\end{aligned}$$

即上述第 2 步是可以省略的，第 3 步直接用 Y 对 $\hat{e}_{X_1}$ 做回归即可。以上过程称为**分步回归**（**partioned regression**）。

4.2 分步回归的含义

分步回归意味着，我们对 X_1 的最小二乘系数的估计，实际上是在排除了 X_2 的相关性之后，X_1 对 Y 的净相关性。作为示例，我们考虑存在一个 0-1 型变量：$d_i=0/1$，一个解释变量 w_i 以及一个因变量 y_i。分步回归告诉我们，如果我们使用 y_i 对 w_i 和 d_i 做回归，即估计方程：

$$y_i = \alpha + \beta \cdot w_i + \gamma \cdot d_i + u_i \tag{12}$$

那么 β 的最小二乘估计 $\hat{\beta}$ 等价于：

1. 首先使用 w_i 对 d_i 做回归：

   $$w_i = \eta + \delta \cdot d_i + \epsilon_i$$

   得到残差 $\hat{\epsilon}_i$。由于 d_i 为虚拟变量，因而 $\hat{\eta}=\bar{w}_0$ 为 $d_i=0$ 组的 w_i 的均值，而 $\hat{\delta}=\bar{w}_1-\bar{w}_0$，为 $d_i=1$ 组与 $d_i=0$ 组的 w_i 的均值只差，因而其残差：

   $$\hat{\epsilon}_i = w_i - \bar{w}_0 - (\bar{w}_1 - \bar{w}_0)\cdot d_i = d_i\left(w_i - \bar{w}_1\right) + (1-d_i)\left(w_i - \bar{w}_0\right)$$

2. 使用 y_i 对 $\hat{\epsilon}_i$ 做回归，得到的系数为：

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^N \hat{\epsilon}_i y_i}{\sum_{i=1}^N \hat{\epsilon}_i^2} \\ &= \frac{\sum_{i=1}^N \left[d_i\left(w_i-\bar{w}_1\right)+\left(1-d_i\right)\left(w_i-\bar{w}_0\right)\right] y_i}{\sum_{i=1}^N \left[d_i\left(w_i-\bar{w}_1\right)+\left(1-d_i\right)\left(w_i-\bar{w}_0\right)\right]^2} \\ &= \frac{\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right) y_i+\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right) y_i}{\sum_{i=1}^N\left[d_i\left(w_i-\bar{w}_1\right)^2+\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]}\end{aligned}$$

其中第 3 个等号由于 $d_i^2=d_i, d_i\left(1-d_i\right)=0$。

注意到，如果我们只挑选出 $d_i=1$ 的样本，使用 y_i 对 w_i 做回归，那么得到的系数为：

$$\hat{\beta}_1=\frac{\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right) y_i}{\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right)^2}$$

同理，如果只选取 $d_i=0$ 的样本，那么得到的系数为：

$$\hat{\beta}_0=\frac{\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right) y_i}{\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2}$$

从而得到：

$$\begin{cases}\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right) y_i & =\hat{\beta}_1\left[\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right)^2\right] \\ \sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right) y_i & =\hat{\beta}_0\left[\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]\end{cases}$$

最终我们得到：

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right) y_i+\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right) y_i}{\sum_{i=1}^N\left[d_i\left(w_i-\bar{w}_1\right)^2+\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]} \\ &= \frac{\hat{\beta}_1\left[\sum_{i=1}^N d_i\left(w_i-\bar{w}_1\right)^2\right]+\hat{\beta}_0\left[\sum_{i=1}^N\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]}{\sum_{i=1}^N\left[d_i\left(w_i-\bar{w}_1\right)^2\right]+\sum_{i=1}^N\left[\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]} \\ &\triangleq \hat{\beta}_1 \omega_1+\hat{\beta}_0 \omega_0\end{aligned}$$

其中 $\omega_1+\omega_0=1$ 为一个权重，即方程 (12) 的最小二乘估计 $\hat{\beta}$ 是在 $d_i=0/1$ 的不同组内进行最小二乘估计的一个加权平均，而不涉及不同组之间的比较，因而 $\hat{\beta}$ 是一个组内（within-group）估计。

例 7. （Simpson 悖论）该悖论是指，当我们对某个感兴趣的变量 y 进行比较时，在每个组内比较的结果与忽略分组进行比较的结果可能是不相同的。比如我们考虑如下的思想实验。我们已知男性的平均寿命比女性的平均寿命要短，但是同时男性可能更愿意锻炼身体，而锻炼身体对寿命有正向的促进作用。以下代

码通过生成一些伪数据模拟了一个符合这个故事的数据：

代码 4: Simpson 悖论的模拟

```
// file: simpson_paradox.do
clear
set obs 1000
gen gender=runiform()<0.5
gen      exer=runiform()<0.8 if gender==1
replace exer=runiform()<0.3 if gender==0
gen y=80-10*gender+3*exer+rnormal()
reg y exer
outreg2 using simpson_paradox.tex, replace
reg y exer if gender==1
outreg2 using simpson_paradox.tex, append
reg y exer if gender==0
outreg2 using simpson_paradox.tex, append
reg y exer gender
outreg2 using simpson_paradox.tex, append
```

我们首先产生了一个 0-1 型的变量，性别 (gender)，接着分别根据性别产生了锻炼与否 (exer) 这个变量。在产生锻炼与否这个变量时，我们假设男性 (gender=1) 有 80% 的人会从事锻炼，而女性 (gender=0) 只有 30%。最后，我们生成了一个寿命的变量 (y)，在这里我们假设不锻炼的女性平均寿命为 80 岁，男性平均低 10 岁，而如果锻炼身体，平均可以延长三年的寿命。接下来，我们分别单独比较了：

1. 锻炼的群体与不锻炼的群体（不考虑性别）
2. 只比较锻炼的男性与不锻炼的男性之间的差别
3. 只比较锻炼的女性与不锻炼的女性之间的差别
4. 以性别作为控制变量，比较锻炼与不锻炼的差别

结果如表 (5) 所示。我们发现，不管是在男性内部（第 2 列）、女性内部（第 3 列），锻炼都能提高寿命，但是如果我们不考虑性别，单纯比较锻炼与否，却会得到锻炼有损寿命的结论（第 1 列）。图 (5) 展示了产生这一现象的原因，最关键的原因是性别（gender）和是否锻炼（exer）并不是独立的，而是呈现了相关性。在回归分析中，解决这一问题的方法之一是使用性别（gender）变量作为控制，根据上述推理，在第 4 列中，我们得到的 exer 的系数实际上是第 2 列和第 3 列的系数的一个加权平均，即在（性别）组内的估计量的加权平均。

在上述的例子中，我们通过控制性别得到了锻炼对寿命的真实影响。实际上，以上通过虚拟变量控制分组比较的思想可以扩展到任意的分类变量，我们

表 5: Simpson 悖论

VARIABLES	(1) y	(2) y	(3) y	(4) y
exer	-1.856***	2.912***	3.187***	3.059***
	(0.285)	(0.105)	(0.0990)	(0.0720)
gender				-10.05***
				(0.0720)
Constant	77.57***	70.02***	79.92***	79.96***
	(0.205)	(0.0915)	(0.0521)	(0.0483)
Observations	1,000	491	509	1,000
R-squared	0.041	0.613	0.671	0.953

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

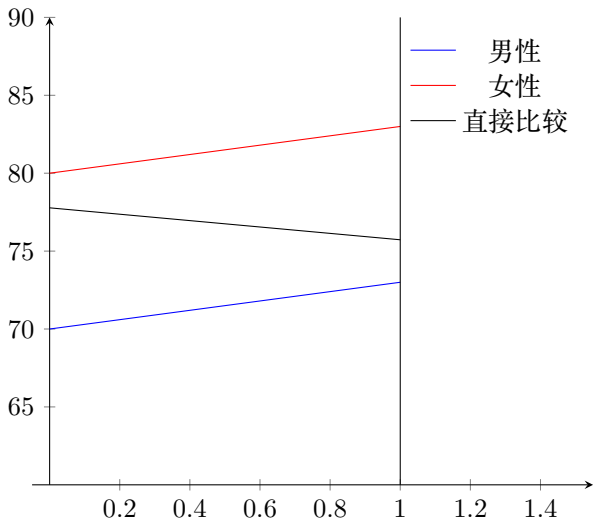


图 5: Simpson 悖论图示

经常把这些虚拟变量称之为**固定效应**（**fixed effects**），因为通过这些虚拟变量，我们将比较限制在分组内部，相当于控制了分组内部那些固定不变的不可观测因素。比如，当我们控制了地区的固定效应时，即限制了所有的回归和比较都来自于地区内部，不存在不同地区之间的比较，相当于控制了地区的特征，或者地区的固定效应。

此外，如果我们有两个分类变量（比如地区地区和学历），那么我们可以同时加入两个分类变量的虚拟变量进行控制：

$$y_i = x_i'\beta + \sum_{r=1}^{R-1}\gamma_{ir} + \sum_{m=1}^{M-1}\gamma_{im} + u_i$$

其中 γ_{ir} 为第一个分类变量的固定效应，代表个体 i 是否属于地区 r，γ_{im} 为第二个分类变量的固定效应，代表个体 i 是否属于学历 m，x_i 为其他解释变量。以上回归方程经常被简写为：

$$y_{irm} = x_i'\beta + \gamma_r + \gamma_m + u_{irm} \tag{13}$$

注意我们在加入这些固定效应时，都删掉了一个虚拟变量以保证矩阵可逆。

或者，我们可以使用两个分类变量定义一个新的分类变量。比如当两个分类变量为学历和地区时，可以定义新的虚拟变量为每个地区的每个学历。因而，如果我们有 30 个地区、7 个学历水平，那么我们可以定义出 30×7 个虚拟变量，并将其作为固定效应在回归中加以控制：

$$y_{irm} = x_i'\beta + \gamma_{rm} + u_{irm} \tag{14}$$

这是一种比同时加入两组虚拟变量更为严格的控制方式，相当于将比较限制在同一地区的同一学历水平进行比较。式 (13) 相当于在式 (14) 的基础上，假设了：

$$\gamma_{rm} = \gamma_r + \gamma_m$$

显然这一假设并不一定成立。

例 8. 如果我们希望比较是否读书（CFPS 数据中的 qq1101）的人收入是否有差异，我们可以通过回归的方法，使用收入对 qq1101 进行回归即可。然而，考虑到是否读书可能是教育程度带来的，而教育程度也会影响收入，因而我们可以通过加入教育程度的虚拟变量来将比较限制在教育程度相同的人群中。此外，不同教育程度的个体可能会出现排序（sorting）效应，即教育程度高的人更多的去往东部沿海城市，因而是否读书的收入差异也有可能是这种排序效应导致的，如果我们希望将比较限制在同一地区，也可以加入地区固定效应。我们可以使用如下程序进行比较：

代码 5: 读书与收入

表 6: 是否读书与收入

VARIABLES	(1) p_income	(2) p_income	(3) p_income	(4) p_income
qq1101	8,879***	3,712**	3,331**	2,471
	(1,528)	(1,593)	(1,593)	(1,668)
Constant	7,771***	2,462		
	(864.8)	(2,092)		
Observations	846	846	846	813
教育固定效应	No	Yes	Yes	No
地区固定效应	No	No	Yes	No
教育 × 地区固定效应	No	No	No	Yes
R-squared	0.038	0.136	0.195	0.258

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

```
// file: reg_readings.do
use datasets/cfps_adult, clear
drop if p_income<0
drop if te4<0
drop if qq1101<0
reg p_income qq1101
outreg2 using reg_readings.tex, replace
reg p_income qq1101 i.te4
outreg2 using reg_readings.tex, append keep(qq1101)
reghdfe p_income qq1101, absorb(i.te4 i.provcd14)
outreg2 using reg_readings.tex, append
reghdfe p_income qq1101, absorb(i.te4#i.provcd14)
outreg2 using reg_readings.tex, append
```

表 (6) 汇报了不同设计下的回归结果。其中第 1 列是单纯的比较是否读书群体的收入，而第 2 列则加入了教育程度的固定效应，我们发现如果控制了教育程度的影响，是否读书的收入差异更小了。第 3 列继续加入地区的固定效应，差距更小。最后一列加入了教育程度和地区相乘的固定效应，控制更为严格，系数也更小。注意以上我们使用了 reghdfe 命令，此命令专门用于在存在非常多虚拟变量时的回归分析，选项 absorb 括号里面是要加入的固定效应的分类变量，使用 # 号代表加入两个分类变量相乘的固定效应。

5 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-squared）。在本节中，我们将讨论拟合优度的度量和检验的问题。

5.1 拟合优度的度量

在拟合或者预测的应用中，我们经常会关注 x 对 y 的解释能力问题。特别的，我们关注 y 的方差中有多少是可以被 x 解释的。我们记 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 $M_0 = I - \frac{1}{N}\iota\iota'$。注意由于 M_0 也是幂等矩阵，因而 $TSS = Y'M_0Y = (M_0Y)' M_0Y$。我们可以通过分析 M_0Y 来将其分解为可被 x 解释的部分和不能被 x 解释的部分：

$$M_0Y = M_0PY + M_0MY$$

注意如果回归方程中包含常数项，那么 ι 为 X 矩阵的第一列，因而：

$$M_0M = \left(I - \frac{1}{N}\iota\iota'\right) M = M - \frac{1}{N}\iota\,(M\iota)' = M$$

因而上式可以化简为：

$$M_0Y = M_0PY + MY$$

而对于 M_0P，有：

$$M_0P = \left(I - \frac{1}{N}\iota\iota'\right) P = P - \frac{1}{N}\iota\iota'$$

注意以上矩阵仍然为实对称的幂等矩阵：

$$\begin{aligned} M_0PM_0P &= \left(P - \frac{1}{N}\iota\iota'\right)\left(P - \frac{1}{N}\iota\iota'\right) \\ &= P - \frac{1}{N}\iota\iota' - \frac{1}{N}\iota\iota' + \frac{1}{N^2}\iota\iota'\iota\iota' \\ &= P - \frac{1}{N}\iota\iota' - \frac{1}{N}\iota\iota' + \frac{1}{N}\iota\iota' \\ &= P - \frac{1}{N}\iota\iota' \\ &= M_0P \end{aligned}$$

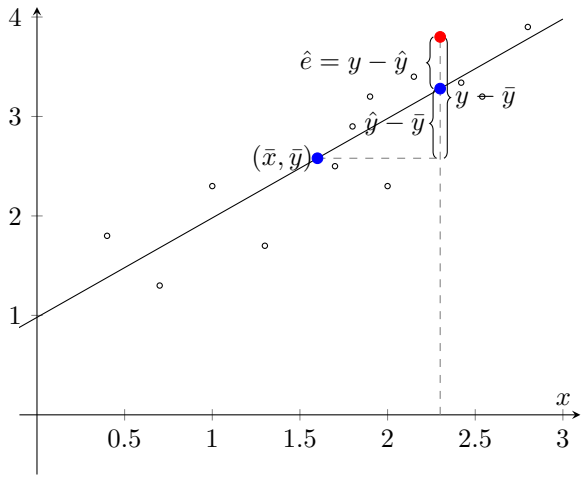


图 6: 三种平方和示意

此外，$MM_0P = M\left(P - \frac{1}{N}\iota\iota'\right) = MP - \frac{1}{N}M\iota\iota' = 0$，因而我们可以得到：

$$\begin{aligned} Y'M_0Y &= (M_0PY + MY)'(M_0PY + MY) \\ &= (M_0PY)'(M_0PY) + Y'MY + Y'MM_0PY + Y'PM_0MY \\ &= Y'M_0PY + Y'MY \\ &= \hat{Y}'M_0\hat{Y} + \hat{e}'\hat{e} \\ &\triangleq ESS + RSS \end{aligned}$$

其中

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

为**回归平方和**（**explained sum of squares**），而 $RSS = \hat{e}'\hat{e}$ 为**残差平方和**（**residual sum of squares**）。因而我们可以定义：

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{Y}'M_0\hat{Y}}{Y'M_0Y} = 1 - \frac{\hat{e}'\hat{e}}{Y'M_0Y}$$

R^2 度量了所谓的「**拟合优度**（**goodness of fit**）」，即使用 x 对 y 进行预测时，x 可以解释多少部分的 y 的方差。

5.2 变量增加时的拟合优度

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。为了证明这个结论，现在我们假设 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

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

而如果我们只对 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：

$$R_2^2 = 1 - \frac{\hat{e}'\hat{e}}{Y' M_0 Y}$$

接下来我们将比较两个 R^2 的大小，或者，当我们向回归方程中添加一个变量 x_1 时，R^2 的变化。

根据分步回归，我们知道 β_1 的最小二乘估计可以写成：

$$\hat{\beta}_1 = \frac{\hat{e}'\hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}}$$

其中 $\hat{\epsilon}$ 为回归 $x_1 = X_2\delta + \epsilon$ 的残差，即 $\hat{\epsilon} = x_1 - X_2\hat{\delta}$。我们将 $\hat{e} = y - X_2\hat{\gamma}$ 带入到 $\hat{u} = y - \hat{\beta}_1 x_1 - X_2\hat{\beta}_2$ 中得到：

$$\begin{aligned}
\hat{e} &= \hat{u} + \hat{\beta}_1 x_1 + X_2\hat{\beta}_2 - X_2\hat{\gamma} \\
&= \hat{u} + \hat{\beta}_1 \left(X_2\hat{\delta} + \hat{\epsilon} \right) + X_2\hat{\beta}_2 - X_2\hat{\gamma} \\
&= \hat{u} + \hat{\beta}_1\hat{\epsilon} + X_2 \left(\hat{\beta}_1\hat{\delta} + \hat{\beta}_2 - \hat{\gamma} \right)
\end{aligned}$$

根据式 (11) 可得：

$$\begin{aligned}
X_2\hat{\beta}_2 &= X_2 \left(X_2' X_2 \right)^{-1} X_2 \left(Y - x_1\hat{\beta}_1 \right) \\
&= P_2 y - P_2 x_1 \hat{\beta}_1 \\
&= X_2\hat{\gamma} - X_2\hat{\delta}\hat{\beta}_1
\end{aligned}$$

从而 $X_2 \left(\hat{\beta}_1\hat{\delta} + \hat{\beta}_2 - \hat{\gamma} \right) = 0$。因而残差平方和：

$$\hat{e}'\hat{e} = \hat{u}'\hat{u} + \hat{\beta}_1^2 \hat{\epsilon}'\hat{\epsilon} + 2\hat{\beta}_1 \hat{u}'\hat{\epsilon}$$

其中 $\hat{u} = MY, \hat{\epsilon} = M_2 x_1$，因而 $\hat{u}'\hat{\epsilon} = Y'MM_2x_1$。根据定义，有：

$$\begin{aligned} MM_2 &= (I - P)(I - P_2) \\ &= I - P - P_2 + PP_2 \\ &= I - P - P_2 + PX_2 (X_2'X_2)^{-1} X_2 \\ &= I - P - P_2 + X_2 (X_2'X_2)^{-1} X_2 \\ &= I - P = M \end{aligned}$$

而 $Mx_1 = 0$，因而 $\hat{u}'\hat{e} = 0$，从而

$$\hat{e}'\hat{e} = \hat{u}'\hat{u} + \hat{\beta}_1^2 \hat{\epsilon}'\hat{\epsilon} = \hat{u}'\hat{u} + \left(\frac{\hat{e}'\hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}}\right)^2 \hat{\epsilon}'\hat{\epsilon} = \hat{u}'\hat{u} + \frac{(\hat{e}'\hat{\epsilon})^2}{\hat{\epsilon}'\hat{\epsilon}}$$

带入到 R^2 和 R_2^2 的定义中，有：

$$\begin{aligned} R_2^2 &= 1 - \frac{\hat{e}'\hat{e}}{Y'M_0Y} \\ &= 1 - \frac{\hat{u}'\hat{u} + \frac{(\hat{e}'\hat{\epsilon})^2}{\hat{\epsilon}'\hat{\epsilon}}}{Y'M_0Y} \\ &= 1 - \frac{\hat{u}'\hat{u}}{Y'M_0Y} - \frac{(\hat{e}'\hat{\epsilon})^2}{\hat{\epsilon}'\hat{\epsilon}} \frac{1}{Y'M_0Y} \\ &= R^2 - \frac{(\hat{e}'\hat{\epsilon})^2}{\hat{\epsilon}'\hat{\epsilon}\hat{e}'\hat{e}} \frac{\hat{e}'\hat{e}}{Y'M_0Y} \\ &= R^2 - [\mathrm{Corr}(\hat{e}, \hat{\epsilon})]^2 \left(1 - R_2^2\right) \end{aligned}$$

从而：

$$R^2 = R_2^2 + [\mathrm{Corr}(\hat{e}, \hat{\epsilon})]^2 \left(1 - R_2^2\right) \tag{15}$$

上式中，R_2^2 是在 Y 对 X_2 的回归中可以被 X_2 解释的部分，而 $\left(1 - R_2^2\right)$ 可以看作在 Y 对 X_2 的回归中不能被 X_2 解释的部分。此外，$\hat{\epsilon}$ 是 x_1 中不能被 X_2 解释的部分，因而添加了变量 x_1 后，模型的确提高了拟合程度，提高的部分来自于 Y 和 x_1 同时不能被 X_2 解释的部分。其中 $\mathrm{Corr}(\hat{e}, \hat{\epsilon})$ 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**）。

5.3 拟合优度的检验

R^2 为我们提供了拟合优度的一个度量，然而更进一步的，我们有时还希望对模型的拟合优度进行检验。最常见的检验是，当使用 x 对 y 进行预测时，其预测效果是否比仅仅使用平均值 $\bar{y}$ 对 y 进行预测究竟更好，或者 x 是否为对 y 的预测带来了更多的信息。在线性回归的情景下，该检验相当于检验是否所有

的 β_j 都同时等于 0（除常数项）。如果设截距项为 β_1，那么原假设为：

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0 \tag{16}$$

对于以上假设检验的问题，在假设3的条件下，我们可以使用统计量：

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

对原假设进行检验。注意在这里我们使用了误差项 u_i 服从正态分布的假设。在原假设条件下，常数项 $\hat{\beta}_1 = \bar{y}$，因而无论 x 取任何值，对于 y 的最优预测都是 $\bar{y}$，从而 $R^2 = 0$，因而只需要检验 $F = 0$ 即可。

由于：

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

因而

$$F = \frac{\hat{Y}' M_0 \hat{Y}/(K-1)}{\hat{e}'\hat{e}/(N-K)}$$

其中分母上是残差平方和 $\hat{e}'\hat{e}$ 除以 $N-K$，我们称之为**均方误差**（**mean squared error**, **MSE**），即对于每一个观测，预测误差平方的平均值；而**均方根误差**（**root mean squared error**, **RMSE**）即均方误差的开平方：

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\hat{e}'\hat{e}}{N-K}}$$

接下来我们讨论 F 的抽样分布。注意到：

$$M_0 \hat{Y} = M_0 P Y = M_0 (X\beta + Pu)$$

在原假设条件下，$X\beta = \beta_1 \iota$，而 $M_0 \iota = \left(I - \frac{1}{N}\iota\iota'\right)\iota = \iota - \iota = 0$，因而在原假设条件下，$M_0 \hat{Y} = M_0 P u$。而由于：

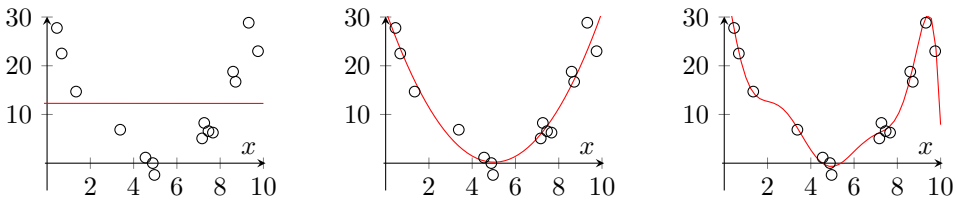
$$M_0 P = \left(I - \frac{1}{N}\iota\iota'\right) P = P - \frac{1}{N}\iota\iota'$$

为实对称幂等矩阵，且 $\operatorname{tr}(M_0 P) = \operatorname{tr}(P) - \operatorname{tr}\left(\frac{1}{N}\iota\iota'\right) = K - 1$，如果 $u \sim N\left(0, \sigma^2 I\right)$，那么：

$$\frac{1}{\sigma^2}\hat{Y}' M_0 \hat{Y} = \frac{1}{\sigma^2} u' M_0 P u \sim \chi^2(K-1)$$

同理 $\hat{e} = Mu$，$\operatorname{tr}(M) = \operatorname{tr}(I - P) = N - K$，因而：

$$\frac{1}{\sigma^2}\hat{e}'\hat{e} = \frac{1}{\sigma^2} u' M u \sim \chi^2(N-K)$$



(a) 欠拟合 (b) 恰好拟合 (c) 过拟合

图 7: 欠拟合、恰好拟合和过拟合

最后，注意到 $M_0 P \cdot M = M_0 (PM) = 0$，因而 $\frac{1}{\sigma^2}\hat{Y}' M_0 \hat{Y}$ 与 $\frac{1}{\sigma^2}\hat{e}'\hat{e}$ 为独立的 χ^2 分布，从而

$$F = \frac{\hat{Y}' M_0 \hat{Y} / (K-1)}{\hat{e}'\hat{e} / (N-K)} \sim F(K-1, N-K)$$

因而对于原假设 (16) 可以使用 F 统计量，使用 $F(K-1, N-K)$ 分布进行右侧检验，如果可以拒绝原假设，说明 x 的加入对于 y 的预测或者拟合有促进作用。

6 模型选择

为了拟合或者预测的目的，我们可以对模型做不同的设定，比如哪些变量可以作为自变量、自变量的函数形式等等。当存在多个不同的模型时，就要面临**模型选择**（**model selection**）的问题。一方面，我们选择的模型需要有足够的预测和拟合能力，对现有数据，模型应该有尽可能高的拟合和预测能力，如果模型的拟合能力不足，我们称之为**欠拟合**（**underfitting**），如图 (7.a) 所示；而另一方面，如果一味提高拟合和预测能力，那么不可避免的会选择复杂度更高的模型，错误地将一些数据中的噪音当做信号进行拟合，我们称之为**过拟合**（**overfitting**），如图 (7.c) 所示。

无论是欠拟合还是过拟合，对于预测都是非常不利的：欠拟合意味着没有充分的发现数据中的规律，预测效果差；过拟合意味着把数据中的一些噪音当做信号，进行样本外预测时，会出现重大偏差。比如在图 (7) 中，虽然真实的关系是二次方的关系，欠拟合没有发现这种关系，因而当 x 很大或者很小时都会预测失败；而过拟合过分的对数据进行拟合，当 x 很大时，甚至预测的方向都是相反的。为了避免上述问题，我们这一节将介绍一些常见的模型选择标准和方法。

6.1 模型选择标准

为了对不同的模型进行选择，一个最简单的方法是根据某些判定准则选择使用哪个模型。下面我们首先介绍一些常用的模型选择准则。

首先，一个自然的想法是使用拟合优度 R^2 进行模型选择。然而以上结论意味着，如果我们向回归中添加变量，那么 R^2 必然不会减少，因而如果使用 R^2，则必然会选择出过拟合的模型，因而 R^2 并不适合作为模型选择的标准。此时我们需要对 R^2 进行调整，即**调整后的** R^2（**adjusted** R^2）：

$$\bar{R}^2 = 1 - \frac{\hat{e}'\hat{e}/(N-K)}{Y'M_0Y/(N-1)} = 1 - \frac{N-1}{N-K}\left(1-R^2\right)$$

即分别使用残差平方和以及总平方和的自由度进行调整。实际上，由于：

$$R^2 = 1 - \frac{\hat{e}'\hat{e}}{Y'M_0Y} = 1 - \frac{\hat{e}'\hat{e}/N}{Y'M_0Y/N}$$

由于最后一项中，分子分母都不是 $\mathrm{Var}(\hat{e})$ 以及 $\mathrm{Var}(y)$ 的无偏估计，而 $\bar{R}^2$ 仅仅是将其替换为两者的无偏估计。

由于调整后的 $\bar{R}^2$ 是变量个数 K 的单调递减函数，相当于为更多的变量添加了惩罚项，因而 $\bar{R}^2$ 不再随着变量个数 K 的增加而单调递增。实际上，$\bar{R}^2$ 甚至可以为负。当回归方程中添加新的变量时，一方面残差平方和会变小，另一方面自由度对变量个数 K 进行了惩罚，因而只有当新添加的变量有足够的解释能力时 $\bar{R}^2$ 才会提高。

$\bar{R}^2$ 经常被用在诸如“逐步回归”的模型选择过程中，然而逐步回归的方法一方面缺乏逻辑基础，另一方面并不能保证能够选择出最优的模型，而是倾向于选出不相关的解释变量，因而现实中应该避免使用该方法（Leamer, 1983）。

即使使用 $\bar{R}^2$ 仍然不能保证足够的惩罚，此时一些信息准则如**赤池信息准则**（**Akaike information creterion**, **AIC**）和**贝叶斯信息准则**（**Bayesian information creterion**, **BIC**，又称**施瓦茨信息准则**，**Schwarz information creterion**, **SIC**）可以作为模型选择的标准。

AIC 和 BIC 都是基于极大似然估计而提出的，其中，Akaike（1974）通过 Kullback–Leibler 距离，定义了如下信息准则：

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

而 Schwarz（1978）通过贝叶斯法则提出了如下信息准则：

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

比如，对于线性回归模型，如果使用极大似然估计，将式 (9) 带入式 (8)，并带入 AIC、BIC 的定义，得到：

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

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

两种信息准则都是通过在对数似然函数中加入惩罚项实现模型选择的，不同的是惩罚项的函数形式不同。应用中，对于多个不同的模型，可以选择使得 AIC 或者 BIC 最小的那个模型。实际上，BIC 选出的模型几乎总是比 AIC 选出的模型更小。

此外，在使用 AIC 和 BIC 时需要注意的是，两者都只能对同一数据集的不同模型进行模型选择，而不适用于不同数据集下的模型选择问题。

例 9. （多项式阶数选择）我们接下来展示如何使用使用以上介绍的标准进行模型选择。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：，

代码 6: 基于 $\bar{R}^2$、AIC、BIC 的模型选择

```
// file: model_selection.do
clear
set obs 50
gen x=runiform()*3
gen y=exp(x)+rnormal()*2
local control ""
scalar K=0
forvalues i=1/10{
    gen x`i'=x^`i'
    local control "`control' x`i'"
    scalar K=K+1
    quietly: reg y `control'
    scalar r2=e(r2)
    scalar r2a=e(r2_a)
    scalar aic=-2*e(ll)+2*K
    scalar bic=-2*e(ll)+log(e(N))*K
    display `i' _skip r2 _skip r2a _skip aic _skip bic
}
```

在以上程序中，我们使用了一个循环产生 x 的高阶项，并通过 local 更新控制变量，进行回归，并计算 R^2、$\bar{R}^2$、AIC、BIC 等，得到的结果如表 (7) 所示。可以发现：

1. 随着高阶项不断地加入到回归中，R^2 逐渐升高；

表 7: 模型选择标准

多项式阶数	(1) R^2	(2) $\bar{R}^2$	(3) AIC	(4) BIC	(5) MSE(CV)
1	0.7119	0.7059	243.77	245.69	7.6158
2	0.8513	0.8450	212.70	216.52	3.4729
3	0.8590	0.8498	212.04	217.78	2.9516
4	0.8594	0.8470	213.90	221.55	3.0424
5	0.8633	0.8477	214.51	224.07	2.9887
6	0.8708	0.8528	213.66	225.13	3.0477
7	0.8719	0.8506	215.25	228.63	3.1342
8	0.8722	0.8509	217.14	232.43	3.1620
9	0.8725	0.8476	219.03	236.24	3.3287
10	0.8725	0.8438	221.00	240.12	3.3543

2. 随着高阶项不断地加入到回归中，$\bar{R}^2$ 先逐渐升高再逐渐下降，当加入 6 阶多项式时，$\bar{R}^2$ 达到了最高值；

3. AIC 和 BIC 在 3 阶、2 阶多项式时达到了最小值，使用 BIC 选择出的模型小于使用 AIC 选择出的模型。

6.2 交叉验证

以上介绍的方法基于回归分析（$\bar{R}^2$）或者基于极大似然，而接下来的方法则可以在任何目标函数下进行模型选择，即**交叉验证**（**cross validation**）法。

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。

可见，交叉验证的主要缺点是运算时间：对于 S 折交叉验证，必须重复估计模型 S 次，而对于留一验证，必须对模型估计 N 次。

实际上，Stone（1977）证明了在 OLS 中，交叉验证和 AIC 是渐进等价的，因而对于简单的普通最小二乘而言，如果样本量足够大，交叉验证法可以直接

使用 AIC 代替。交叉验证法的主要优点在其是一个比较一般的模型选择方法，不像 $\bar{R}^2$ 严重依赖于模型，或者 AIC、BIC 需要极大似然估计的情形下才能使用。

例 10. （留一验证）接下来我们使用留一验证法对例9中的多项式回归模型进行阶数的选择。在以下程序中，我们生成了与例9相同的数据，接下来每一次向模型中添加更高阶次方的 x，都使用留一法计算验证集的预测误差，共循环 N 次，计算产生了 N 个预测误差，最终汇报在该模型下，所有 N 个预测误差的平方和：

代码 7: 交叉验证示例

```
// file: cross_validation_reg.do
clear
set obs 50
gen x=runiform()*3
gen y=exp(x)+rnormal()*2
local control ""
scalar K=0
forvalues i=1/10{
  gen x`i'=x^`i'
  local control "`control' x`i'"
  local N= _N
  quietly{
    gen error=.
    forvalues j=1/`N'{
      reg y `control' if _n~=`j'
      predict resid, residual
      replace error=resid if _n==`j'
      drop resid
    }
  }
  gen error2=error^2
  quietly: su error2
  scalar mse=r(mean)
  display `i' _skip mse
  drop error error2
}
```

表 (7) 第 (5) 列给出了计算结果。我们发现使用 3 阶多项式时，验证集预测误差的平方和是最低的。使用留一验证法选择出了与 AIC 同样的模型。

练习题

练习 1. 重复例2中的程序，并画出散点图、预测直线，观察截距项和斜率项。

练习 2. 观察例3中产生的虚拟变量的形式，并验证例3中第二个回归结果计算的即分组的平均值。

练习 3. 重复例7中的数据生成过程，并进行以下回归：

1. 使用 exer 对 gender 做回归，保存残差为 resid_exer
2. 使用 y 对 resid_excer 做不带常数项的回归

观察上述 resid_excer 的估计系数是否与使用 y 对 exer 和 gender 做回归的回归系数相等。

练习 4. 自己设计一个数据生成过程，验证公式 (15)。

练习 5. 对于总体回归方程：

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

令 $y_i^* = y_i - x_i'\delta$，试证明在以下回归中：

$$y_i^* = x_i'\gamma + e_i$$

中，最小二乘估计：$\hat{\gamma} = \hat{\beta} - \hat{\delta}$。

练习 6. 使用 2014 年 CFPS 数据，进行以下回归：

1. 使用“家庭总支出”的对数($\ln(expense)$)对“全部家庭纯收入”的对数($\ln(income1)$)做回归
2. 使用家庭总支出占家庭总收入的比例（$\ln\left(\frac{expense}{income1}\right)$）对“全部家庭纯收入”的对数（$\ln(income1)$）做回归

观察以上两个回归的系数，是否与上题中的结论相一致。此外，观察两个回归中的 R^2，两者是否相等？出现这种情况的原因是什么？

练习 7. 现有两个回归模型：

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i1} \cdot x_{i2} + u_i$$

以及：

$$y_i = \delta_0 + \delta_1 (x_{i1} - \mu_1) + \delta_2 (x_{i2} - \mu_2) + \delta_3 (x_{i1} - \mu_1)(x_{i2} - \mu_2) + v_i$$

试证明其最小二乘估计量：$\hat{\delta}_3 = \hat{\beta}_3$。自己设计数据生成过程验证以上结论。

练习 8. 现有一个 0-1 变量 d_i、一个解释变量 x_i 以及一个被解释变量 y_i。我们可以对 $d_i = 0/1$ 分别做回归：

$$y_i = \alpha_0 + x_i\beta_0 + u_{0i} | d_i = 0$$
$$y_i = \alpha_1 + x_i\beta_1 + u_{1i} | d_i = 1$$

我们也可以使用如下回归模型：

$$y_i = \delta_0 + \delta_1 d_i + \delta_2 x_i + \delta_3 d_i \cdot x_i + u_i$$

试证明：

$$\begin{cases} \hat{\alpha}_0 = \hat{\delta}_0 \\ \hat{\beta}_0 = \hat{\delta}_2 \\ \hat{\alpha}_1 = \hat{\delta}_0 + \hat{\delta}_1 \\ \hat{\beta}_1 = \hat{\delta}_2 + \hat{\delta}_3 \end{cases}$$

并自己设计数据生成过程验证以上结论。

练习 9. 试证明：对于一元线性回归：

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

其最小二乘估计得到的 $R^2 = \mathrm{Corr}\,(x, y)$。

练习 10. 对于模型：

$$y = \beta_1 x_1 + x'\beta + u$$

以及模型：

$$y - x_1 = \delta_1 x_1 + x'\delta + e$$

试证明：

1. 最小二乘估计量 $\hat{\beta}_1 = \hat{\delta}_1 + 1$，$\hat{\beta} = \hat{\delta}$。
2. 两个模型的残差满足：$\hat{u} = \hat{e}$。
3. 讨论何时第二个回归的 R^2 大于第一个回归的 R^2？
4. 根据第二问的内容，请问何时两个回归的 R^2 可以进行比较？

参考文献

[1] Acemoglu, D. (2009). Introduction to Modern Economic Growth. New Jersey: Princeton University Press.

[2] Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

[3] Angrist, J. D., & Pischke, J.-S. (2008). Mostly Harmless Ecnometrics: An Empiricist's Companion. Princeton and Oxford: Princeton University Press.

[4] Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 211–252.

[5] Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. Journal of the American Statistical Association, 76(374), 296–311.

[6] Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: Methods and Applications (Vol. 100). Cambridge University Press.

[7] Greene, W. H. (2008). Econometric Analysis (6th ed.). New Jersey: Pearson Education.

[8] Head, K., & Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In Handbook of international economics (Vol. 4, pp. 131–195). Elsevier.

[9] Leamer, E.E., 1983. Chapter 5 Model choice and specification analysis, in: Handbook of Econometrics. Elsevier, pp. 285–330.

[10] Schwarz, G.E., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6.

[11] Stone, M., 1977. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical Society. Series B (Methodological) 39, 44–47.

[12] Wooldridge, J. M. (2010). Econometric Analysis of Cross Sectional and Panel Data (2nd ed.). Cambridge: The MIT Press.

[13] 李航. (2012). 机器学习方法. 北京: 清华大学出版社.

[14] 周志华. (2016). 机器学习. 北京: 清华大学出版社.