作为拟合的回归

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两个统计问题，而回归（regression）是解决这类问题的最常用手段。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。

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 的线性函数形式对 y_i 进行预测，即使用 x_i 的一个线性函数：

$$f(x_i) = \alpha + \beta x_i$$

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 为自变量或者解释变量，而 y_i 为因变量或者被解释变量。

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**）：

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

为了进行拟合，我们通常希望残差 $\tilde{e}_i$ 离 0 越近越好。实际上，有多种度量残差 $\tilde{e}_i$ 和 0 的距离的度量方法，然而最常用的方法是使用其平方：$\tilde{e}_i^2$。只有当误差为 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，即得到了所谓的「**最小二乘法**（**Least squares**）」：

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

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：

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

化简上述问题，得到：

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

继续化简，得到：

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

因而解得：

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

至此，我们就得到了使用 x_i 对 y_i 进行预测所需要的参数，或者 α 和 β 的估计，我们称之为最小二乘估计量。

在得到了 α 和 β 的估计以后，我们可以得到给定 x 对 y 的预测值：

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

以及预测的残差：

$$\hat{e} = y - \hat{y}$$

对于一个给定的 x，如果其对应的 y 未知，我们可以使用 $\hat{y}$ 对 y 进行预测，而残差 $\hat{e}$ 就是对于已知的 x_i, y_i，我们使用 $\hat{y}$ 对 y_i 进行预测的误差。

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. 身高和体重在历史上是线性回归最早研究的问题之一。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<0
```

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

表 1: 身高与体重的关系

```
reg qp102 qp101
outreg2 using reg_one_variate.tex, replace
predict p_weight
sort qp101
twoway (scatter qp102 qp101)/*
     */(line p_weight qp101)
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 (1) 所示。可见身高和体重呈现了明显的正相关关系。

以上介绍了作为拟合的一元最小二乘法，实际上回归系数还可以看成是简单的比较。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，

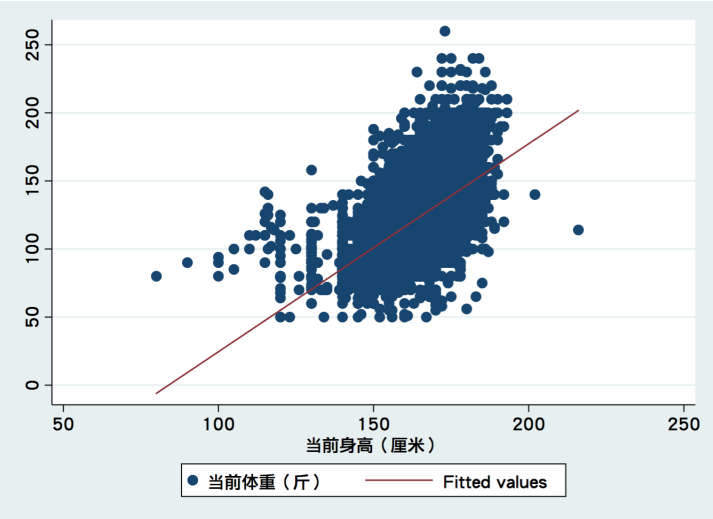

图 1: 身高与体重的关系

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：

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

VARIABLES	(1) cfps_gender 0 N	(2) cfps_gender 0 mean	(3) cfps_gender 1 N	(4) cfps_gender 1 mean
p_income	18,308	5,751	18,398	12,287

表 2: 收入的描述性统计

而当 $x_i = 1$ 时，有:

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

因而实际上，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,/*
  */replace sum(log) eqkeep(N mean) keep(p_income)
reg p_income cfps_gender
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。

VARIABLES	(1) 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

表 3: 收入对性别的回归

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：

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

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 $f\left(x_i\right) = x_i'\beta$ 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：

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

对以上目标函数求导数并令其等于 0，可以得到一阶条件：

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

解以上方程可以得到：

$$X'Xb = X'y \Rightarrow \hat{\beta} = (X'X)^{-1} X'y \tag{2}$$

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：

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

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**）。

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 $N > K$）。即矩阵 X 是列满秩的意味着 X 的任何一列不能被其他列线性表示出来。这就

排除了例如以下情况：

1. 完全相同或者成比例的 X；

2. 如果存在常数项，那么加其他几个或者某几个变量之和不能是常数。

比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。

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除了常数项，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：

代码 3: 不同性别的收入对比

```
1 // file: reg_with_dummies.do
```

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

表 4: 不同教育程度收入比较

```
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 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。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

[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。

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 回归与条件期望

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：

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

如果我们假设函数 $h(\cdot)$ 只能取线性函数的形式，即:

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

那么以上条件期望定义就变成了寻找 β 使得目标函数最小化的过程:

$$\beta_0 = \arg\min_{\beta} \mathbb{E}\left([y - x'\beta]^2 \right) \tag{3}$$

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 (3) 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。

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扰动，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，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:

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

在定义了误差项只有，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:

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

因而我们经常也会直接写出上述带误差项的模型。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。

实际上，对于模型：

$$y_i = x_i'\beta_0 + u_i$$

如果我们希望估计 β_0，在 $\mathbb{E}(u_i|x_i) = 0$ 满足的情况下，我们也可以用矩估计方法。在这里，矩条件为：

$$\mathbb{E}(x_i u_i) = \mathbb{E}[x_i(y_i - x_i'\beta_0)] = \mathbb{E}(x_i y_i) - \mathbb{E}(x_i x_i')\beta_0 = 0$$

因而如果我们使用样本平均代替期望，就得到了：

$$\frac{1}{N}\sum_{i=1}^{N} x_i\left(y_i - x_i'\hat{\beta}\right) = 0$$

解得：

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

同样得到了最小二乘估计。

2.3 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：

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

如果我们记 $P = X(X'X)^{-1}X'$，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：

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

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：

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

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 $\bar{y} = \bar{\hat{y}}$。

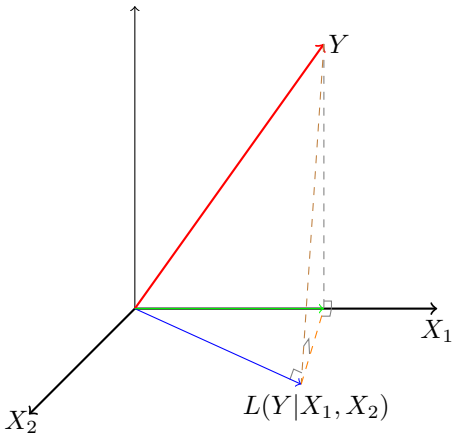

图 2: 最小二乘与投影

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」：

$$y'y = y'Py + y'My = \hat{y}'\hat{y} + \hat{e}'\hat{e}$$

2.4 分步回归

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 (2) 的一阶条件可以写为：

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

解以上方程，即：

$$\begin{cases} X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'y \\ X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'y \end{cases} \tag{4}$$

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：

$$\begin{cases} \hat{e}_{X_1} = M_2 X_1 \\ \hat{e}_y = M_2 y \end{cases}$$

那么：

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

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，直接用 y 对 $\hat{e}_{X_1}$ 做回归即可。

分布回归意味着，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 的影响之后，X_1 对 y 的净影响。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：

$$y_i = \alpha + \beta \cdot w_i + \gamma \cdot d_i + u_i \tag{5}$$

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：

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

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 (5) 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 4. （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。图 (3) 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

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

表 5: Simpson 悖论

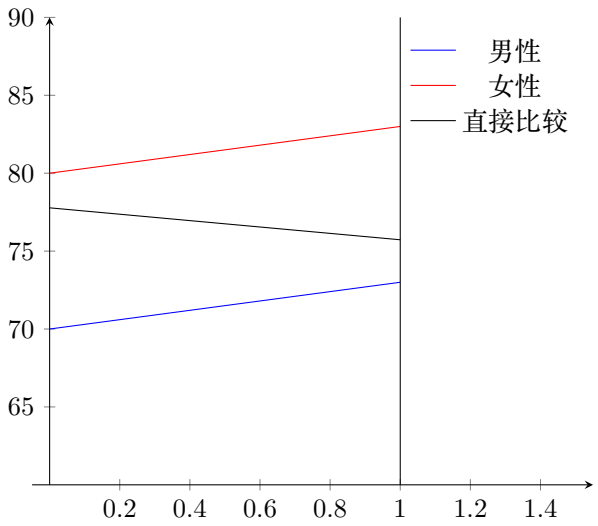

图 3: Simpson 悖论图示

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，或者说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：

$$y_i = x_i'\beta + \sum_{r=1}^{R-1}\gamma_r + \sum_{m=1}^{M-1}\gamma_m + u_i$$

其中 γ_r 为第一个分类变量的固定效应，γ_m 为第二个分类变量的固定效应，x_i 为其他解释变量。以上回归方程经常被写为：

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

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，比如如果我们有 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：

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

实际上这是一种比同时加入两组虚拟变量更为自由和严格的控制方式，实际上如果同时加入两组虚拟变量，相当于假设了：

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

当然这一假设并不一定成立。

例 5. 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

```
1 // file: reg_readings.do
```

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
R-squared	0.038	0.136	0.195	0.258

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

表 6: 是否读书与收入

```
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。

2.5 拟合优度

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 的总变分（total variation）中有多少是可以被 x 解释的，其中 y 的总变分为:

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

其中 $M_0 = I - \frac{1}{N}\iota\iota'$。注意由于 M_0 也是幂等矩阵，因而 $y'M_0y = (M_0y)'M_0y$，因而我们可以通过分析 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：

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

现在，我们可以得到：

$$\begin{aligned}
y'M_0y &= (M_0Py)'(M_0Py) + y'My + y'MM_0Py + y'PM_0My \\
&= y'M_0Py + y'My \\
&= \hat{y}'M_0\hat{y} + \hat{e}'\hat{e} \\
&= SSR + SSE
\end{aligned}$$

其中

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

为回归平方和，而 $SSE = \hat{e}'\hat{e}$ 为残差平方和。因而我们可以定义：

$$R^2 = \frac{SSR}{SST} = \frac{\hat{y}'M_0\hat{y}}{y'M_0y} = 1 - \frac{\hat{e}'\hat{e}}{y'M_0y}$$

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 的总变分。

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 (4) 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 $\hat{u}=My, \hat{\epsilon}=M_2x_1$，因而 $\hat{u}'\hat{\epsilon}=y'MM_2x_1$。根据定义，有：

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

而 $Mx_1=0$，因而 $\hat{u}'\hat{e}=0$，从而

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

带入到 R^2 和 R_2^2 的定义中，有：

$$\begin{aligned}R_2^2 &= 1-\frac{\hat{e}'\hat{e}}{y'M_0y}\\ &= 1-\frac{\hat{u}'\hat{u}+\frac{\left(\hat{e}'\hat{\epsilon}\right)^2}{\hat{\epsilon}'\hat{\epsilon}}}{y'M_0y}\\ &= 1-\frac{\hat{u}'\hat{u}}{y'M_0y}-\frac{\left(\hat{e}'\hat{\epsilon}\right)^2}{\hat{\epsilon}'\hat{\epsilon}}\frac{1}{y'M_0y}\\ &= R^2-\frac{\left(\hat{e}'\hat{\epsilon}\right)^2}{\hat{\epsilon}'\hat{\epsilon}\hat{e}'\hat{e}}\frac{\hat{e}'\hat{e}}{y'M_0y}\\ &= R^2-\left[\mathrm{Corr}\left(\hat{e},\hat{\epsilon}\right)\right]^2\left(1-R_2^2\right)\end{aligned}$$

从而：

$$R^2=R_2^2+\left[\mathrm{Corr}\left(\hat{e},\hat{\epsilon}\right)\right]^2\left(1-R_2^2\right) \tag{6}$$

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。

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。然而在实际应用中，如果以拟合为目的，为了防止过拟合（overfitting），经常需要做变量选择，即选择哪一些解释变量保留在模型中，哪些剔除，这时 R^2 就不再是一个好的指标，因为只要增加变量就可以增大 R^2。此时我们需要对 R^2 进行调整，即调整后的 R^2：

$$\overline{R^2}=1-\frac{\hat{e}'\hat{e}/\left(N-K\right)}{y'M_0y/\left(N-1\right)}=1-\frac{N-1}{N-K}\left(1-R^2\right)$$

由于调整后的 $\overline{R^2}$ 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 $\overline{R^2}$ 不再随着变量个数 K 的增加而单调递增。实际上，即使使用 $\overline{R^2}$ 仍然不能保证足够的惩罚，一些信息准则如 AIC（Akaike Information Creterion）和 BIC（Bayesian Information Creterion）可以作为模型选择的标准。

3 * 分位数回归初步

4 * 非参数与半参数回归初步

以上线性回归可以使用 x 对 y 进行拟合，然而使用了非常强的假设，即 x 和 y 之间存在着线性关系，然而这一假设并不一定满足。很多时候我们希望在没有函数形式假定的情况下使用 x 对 y 进行拟合，这就诞生了非参数回归。为了介绍非参数回归，我们先从密度函数的估计入手。

4.1 核密度估计

首先我们考虑对于随机变量 x 的密度函数的估计。为了便于叙述，我们首先考虑一元随机变量 x 的密度估计。

考虑一下密度函数的概念，密度函数就是分布函数的一阶导数。一般情况下，我们可以使用经验分布函数（empirical distribution function）对分布函数进行估计：

$$\hat{F}(t)=\frac{1}{N}\sum_{i=1}^{N}1\{x_i\le t\}$$

然而以上估计出的分布函数不可导，所以我们不能使用其对密度函数进行估计。

考虑导数的定义，如果假设分布函数连续可微，那么：

$$f(t)=F'(t)=\lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t-\Delta t)}{2\Delta t}$$

如果我们使用经验分布函数 $\hat{F}(t)$ 代替上式中的分布函数 $F(t)$，同时给定一个固定的 $\Delta t=h$，有：

$$\hat{f}(t)=\frac{\hat{F}(t+h)-\hat{F}(t-h)}{2h}$$

而根据经验分布函数 $\hat{F}$ 的定义，以上估计等价于：

$$\begin{aligned}\hat{f}(t)&=\frac{\#\{x_i\in(t-h,t+h)\}}{2hN}\\&=\frac{1}{Nh}\sum_{i=1}^{N}\frac{1\{t-h\le x_i\le t+h\}}{2}\\&=\frac{1}{Nh}\sum_{i=1}^{N}\frac{1}{2}1\left\{\left|\frac{x_i-t}{h}\right|\le 1\right\}\end{aligned}$$

即，给定一个 t，选取一个 h，样本落在在邻域 $(t-h, t+h)$ 中的比例即可以当做是密度函数的一个近似估计。

注意在以上的替代过程中，导数的定义要求 $h \to 0$，而我们在实际操作过程中我们不可能让 $h=0$，所以必须选取一个正的 h。然而 h 选取的太大，则会违背导数的定义，导致估计的偏差很大；如果太小，那么在一个邻域内样本量可能会非常小，甚至没有观测，导致估计的方差很大。这也就是非参数估计里面的 bias-variance tradeoff。实际使用中，理论上存在着一个能够平衡偏差和方差的最优的 h。我们通常把 h 成为窗宽（bandwidth）。

注意以上的密度函数是不光滑的。观察以上式子，如果记 $K_0(u) = \frac{1}{2}1\{|u| \leq 1\}$，那么：

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}(t)\,dt &= \int_{\mathbb{R}} \frac{1}{Nh}\sum_{i=1}^{N} K_0\left(\frac{x_i - t}{h}\right) dt \\ &= \frac{1}{Nh}\sum_{i=1}^{N}\int_{\mathbb{R}} K_0\left(\frac{x_i - t}{h}\right) dt \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}} K_0(u)\,du\end{aligned}$$

因而如果 $\int_{\mathbb{R}} K_0(u)\,du = 1$，那么估计出的密度函数等于 1。

因而我们经常会替换其中的 $K_0(u) = \frac{1}{2}1\{|u| \leq 1\}$ 为常用的连续随机变量的密度函数 $K(\cdot)$（比如正态分布密度函数），从而得到密度函数的一个光滑的估计。我们称 $K(\cdot)$ 为核函数（kernel function）。

如果我们用正态分布的密度函数对 x 的密度进行估计，那么：

$$\hat{f}(t) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - t}{h}\right)$$

其中 $K(\cdot)$ 取为正态分布的密度函数。

以上讨论的是一元随机变量 x 的核密度估计，以上方法还可以进行进一步推广。如果 $x_i = (x_{i1}, ..., x_{ik})'$ 为 k 维的随机样本，$i = 1, 2, ..., N$，那么核密度估计为：

$$\hat{f}(t) = \frac{1}{N \cdot h_1 \cdot \cdots \cdot h_k}\sum_{i=1}^{N} K_1\left(\frac{x_{i1} - t_1}{h_1}\right) \cdot \cdots \cdot K_k\left(\frac{x_{ik} - t_k}{h_k}\right)$$

4.2 非参数回归

如果我们有数据 $(y_i, x_i')'$，我们希望使用 x_i 拟合 y_i，如果我们有理由认为 y_i 与 x_i 之间存在线性关系，那么自然可以使用线性回归：

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

对以上函数进行拟合。然而如果我们并不知道函数形式，那么更一般的方法是对 x 与 y 之间的函数关系不多任何假设：

$$y_i = g(x_i) + u_i$$

其中 $u_i = y_i - \mathbb{E}(y_i|x_i)$。

然而如果对函数形式不做任何假设，以上估计过程就变得十分困难。在此我们需要一些平滑性的假设，假设 $g(x_i)$ 为一个足够平滑的函数。在此基础上，观察到：

$$\begin{aligned}\mathbb{E}(y_i|x_i) &= \int_{\mathbb{R}} yf(y|x)\,dy \\ &= \int_{\mathbb{R}} y\frac{f(x,y)}{f_X(x)}dy \\ &= \frac{\int_{\mathbb{R}} yf(x,y)\,dy}{f_X(x)}\end{aligned}$$

我们可以通过使用核密度估计替代以上方程中的两个密度函数，对 $\mathbb{E}(y_i|x_i)$ 进行估计。

由于：

$$\hat{f}(x,y) = \frac{1}{Nh_y h}\sum_{i=1}^{N} K\left(\frac{y-y_i}{h_y}\right)K\left(\frac{x-x_i}{h}\right)$$

而

$$f_X(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)$$

因而：

$$\begin{aligned}\frac{\int_{\mathbb{R}} yf(x,y)\,dy}{f_X(x)} &= \frac{\int_{\mathbb{R}} y\frac{1}{Nh_y h}\sum_{i=1}^{N} K\left(\frac{y-y_i}{h_y}\right)K\left(\frac{x-x_i}{h}\right)dy}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\frac{1}{h_y}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)\int_{\mathbb{R}} yK\left(\frac{y-y_i}{h_y}\right)dy}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\frac{1}{h_y}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)\int_{\mathbb{R}} (h_y u + y_i)K(u)h_y du}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\frac{1}{h_y}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)\left[h_y^2\int_{\mathbb{R}} uK(u)\,du + h_y y_i\int_{\mathbb{R}} K(u)\,du\right]}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)y_i\int_{\mathbb{R}} K(u)\,du}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)} \\ &= \frac{\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)y_i}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)}\end{aligned}$$

其中我们假设了使用了对称的核函数，因而 $\int_{\mathbb{R}} uK(u)\,du = 0$。以上即是非参数回归的表达式，即：

$$\hat{\mathbb{E}}(y_i|x_i = x) = \frac{\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{N} K\left(\frac{x_i-x}{h}\right)}$$

实际上，以上非参数回归可以看成是使用 $K\left(\frac{x-x_i}{h}\right)$ 作为权重的滑动平均，给予 x 距离近的点以更多的权重，而距离 x 远的点以更小的权重，如此进行加权平均，即得到了非参数回归。

注意非参数回归也有其应用局限。首先，x 的维数不能太大，实际上非参数回归仅仅适合维数比较小的情况下使用。对于任何一个点 $x_i = x$ 处，周围可以用以滑动平均的点随着维数变大迅速减少，因而其大样本性质随着维数增大也会逐渐变差。

其次，非参数回归不能做外延预测，即不能做超过数据集范围的预测。实际上，即使没有超过数据集 x_i 的取值范围，在 x_i 的边界处，预测的效果也会大打折扣。

由于非参数回归的这些缺点，我们可以将参数回归和非参数回归结合，得到半参数回归。即，如果我们有两部分自变量 x_i 和 w_i，我们可以对 x_i 进行参数假设，而对 w_i 不做任何参数假设，即设定模型：

$$y_i = x_i'\beta + g(w_i) + u_i$$

注意到，对上市两边对 w_i 求条件期望，由于：

$$\mathbb{E}(y_i|w_i) = \mathbb{E}(x_i|w_i)'\beta + g(w_i)$$

因而：

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

因而我们可以使用 y_i 和 x_i 分别对 w_i 做非参数回归，得到残差后使用得到的残差做线性回归，即可得到 β 的估计。

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 (4) 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 (6)。

练习 5. 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 6. 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：

$$\begin{cases} \alpha_0 = \delta_0 \\ \beta_0 = \delta_2 \\ \alpha_1 = \delta_0 + \delta_1 \\ \beta_1 = \delta_2 + \delta_3 \end{cases}$$

并自己设计数据生成过程验证以上结论。