线性回归

司继春

上海对外经贸大学统计与信息学院

1 一元线性回归

在应用中，我们经常碰到所谓的「拟合（fitting）」问题，即如果我们有一列数据 (y_i, x_i)，我们希望使用 x_i 的线性函数形式对 y_i 进行预测，即：

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

其中 u_i 为预测的误差，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。为了度量 $\tilde{e}_i$ 与 0 的距离，我们通常会使用 $\tilde{e}_i^2$，如果我们最小化所有样本的 $\tilde{e}_i$ 的平方和，即得到了所谓的「最小二乘法（Least squares）」：

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

求解上述最小化问题，即对上述目标函数求导，并令导数等于 0，得到：

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

化简上述问题，得到：

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

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：

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

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
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，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。

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。

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

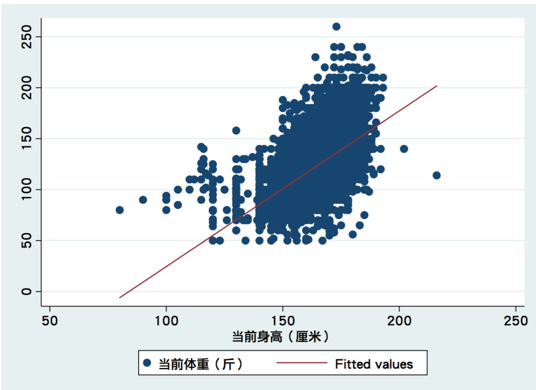

图 1: 身高与体重的关系

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，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,/*
```

	(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 2: 收入的描述性统计

	(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

表 3: 收入对性别的回归

```
  */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。

2 作为拟合的回归

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：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_{K-1} x_{i,K-1} + u_i$$

同样的，其中 u_i 为误差项，y_i 为因变量或者被解释变量，而 x_{ik} 为解释变量。为了方便起见，我们一般用向量表述上述方程：

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

其中：

$$x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{i,K-1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$$

为两个 K 维向量。为了计算方便，我们记：

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}, X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix}_{N\times K} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,K-1} \\ 1 & x_{21} & \cdots & x_{2,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{N,K-1} \end{pmatrix}$$

因而误差项向量为：

$$e = y - X\beta = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}_{N\times 1}$$

与一元线性回归一样，给定一个 b，我们可以得到残差：

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

我们希望最小化残差的平方和 $\sum_{i=1}^n \hat{e}^2$，因而我们可以最小化：

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

对以上目标函数求导数并令其等于 0，可以得到：

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

解以上方程可以得到：

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

以上最大化问题的二阶导为：

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

为一个正定矩阵，因而以上根据一阶条件求得的解：

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

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。

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，$X'X$ 可逆要求矩阵 X 是列满秩的（样本量 $N>K$），即矩阵 X 的任何一列不能被其他列表示出来。这就排除了例如一下情况：

1. 完全相同或者成比例的 X
2. 某一个解释变量 x_{ik} 可以被其他的几个解释变量线性表示出
3. 如果存在常数项，那么加虚拟变量的时候其和不能为常数项

比如我们知道

例如在回归分析中，我们经常加入分组的虚拟变量，即如果一个变量的取值范围为 $G_i=1,2,...,g$，我们经常设定如下回归：

$$y_i=\beta_0+\tilde{x}'\tilde{\beta}+\sum_{j=1}^{g}\delta_j 1\left\{G_i=j\right\}+u_i$$

然而以上设定违背了矩阵 X 是列满秩的要求，必须剔除一个虚拟变量，比如：

$$y_i=\beta_0+\tilde{x}'\tilde{\beta}+\sum_{j=1}^{g-1}\delta_j 1\left\{G_i=j\right\}+u_i$$

或者，我们可以剔除常数项：

$$y_i=\tilde{x}'\tilde{\beta}+\sum_{j=1}^{g}\delta_j 1\left\{G_i=j\right\}+u_i$$

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一种方法，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

```
// 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 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个分组的收入的平均值，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。

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：

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

[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。

	(1)	(2)
VARIABLES	p_income	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

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。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：

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

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 $\hat{e}$。注意 M 矩阵也为幂等矩阵：

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

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 M 进行投影，得

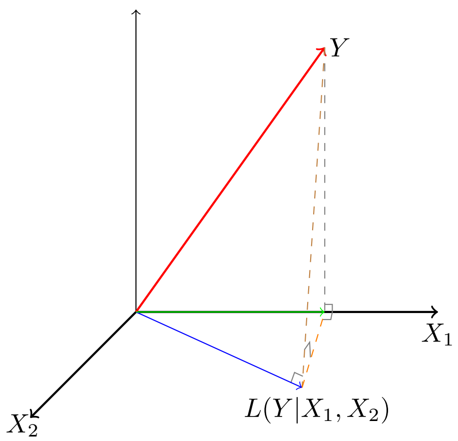

图 2: 最小二乘与投影

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」：

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

2.3 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：

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2 = y'M_0y$$

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

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

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。实际上，R^2 与方差分析有着密不可分的联系。

在实际应用中，当在回归方程中添加变量时，R^2 总是会提高的。因而为了防止过拟合，需要对 R^2 进行调整，即调整后的 R^2：

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

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_1, X_2，那么：

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

如果对以上方程求解最小二乘，我们可以得到如下的一阶条件：

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

解以上方程，即：

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

由第一个式子可以得到：

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

带入第二个式子：

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

记 $P_1 = X_1\left(X_1'X_1\right)^{-1}X_1'$，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：

$$X_2'\left(I-P_1\right)X_2\hat{\beta}_2 = X_2'\left(I-P_1\right)y$$

记 $M_1 = I - P_1$，那么：

$$\hat{\beta}_2 = \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 y$$

同理：

$$\hat{\beta}_1 = \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 y$$

注意实际上 $M_2 X_1$ 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 的系数等价于：

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。

3 其他拟合方法：非参数与半参数回归

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。

3.1 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 $K_0(u) = \frac{1}{2} 1\{|u| \le 1\}$，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 $K_0(u) = \frac{1}{2} 1\{|u| \le 1\}$ 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：

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

3.2 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：

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

而

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

因而：

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

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：

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

因而：

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

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。