第 15 章　倾向得分匹配

15.1　潜在结果框架

我们常希望评估某项目或政策实施后的效应，比如政府推出的就业培训项目(job training program)。

此类研究被称为**项目效应评估**(program evaluation)，而项目效应也常称为**处理效应**(treatment effect)。

项目参与者的全体构成**处理组**(treatment group)，而未参与项目者则构成**控制组**(control group)。

以就业培训的效应评估为例。

如果直接对比处理组与控制组的未来收入，常会发现参加就业培训者的未来收入比未参加者更低。

参加培训一般是参加者“自我选择”(self-selection)的结果。

岗位好收入高的人群并不需要参加就业培训，而就业培训的参加者多为失业、低收入者。

由于处理组与控制组成员的初始条件不完全相同，故存在“选择偏差”(selection bias)。

我们真正感兴趣的问题是，处理组的未来收入是否会比这些人若未参加培训的(假想)未来收入更高。

Rubin (1974) 提出**潜在结果框架**(potential outcomes framework)，也称为**反事实框架**(counterfactual framework)或**鲁宾因果模型**(Rubin causal model)。

鲁宾因果模型的三大要素为潜在结果(potential outcomes)、稳定性假定(the stability assumption)与分配机制(assignment mechanism)。

以虚拟变量$D_i = \{0,1\}$表示个体i是否参与项目，即 1 为参与，而 0 为未参与。

通常称D_i为“处理变量”(treatment variable)或“干预变量”(intervention variable)，反映个体i是否受到“处理”(treatment)或“干预”(intervention)。

记个体i的未来收入或其他感兴趣的结果(outcome of interest)为y_i。

我们想知道D_i是否对y_i有因果作用。

对于个体i，其未来收入y_i可能有两种状态，称为**潜在结果**(potential outcomes)，取决于是否参加此项目：

$$y_i = y_i(D_i) = \begin{cases} y_{0i} & \text{若 } D_i = 0 \\ y_{1i} & \text{若 } D_i = 1 \end{cases} \qquad (15.1)$$

潜在结果y_{0i}表示个体i未参加项目的未来收入。

潜在结果y_{1i}表示个体i参加项目的未来收入。

个体i参加项目的处理效应(treatment effect)，即所谓“个体层面的因果效应”(unit-level causal effects)，可定义为

$$\tau_i \equiv y_{1i} - y_{0i} \qquad (15.2)$$

如果个体i参加项目，则y_{1i}可观测，称为“观测结果”(observed outcome)，故$y_i = y_{1i}$；但看不到潜在结果y_{0i}，此时y_{0i}即为“反事实结果”(counterfactual outcome)。

反之，如果个体i未参加项目，则y_{0i}为观测结果(可观测)，故$y_i = y_{0i}$；却看不到y_{1i}，此时y_{1i}即为反事实结果。

由于个体只能处于一种状态(要么参加项目，要么不参加)，故只能观测到y_{0i}或y_{1i}，而无法同时观测到y_{0i}与y_{1i}。

这是一种“数据缺失”(missing data)问题，正是“因果推断的基本问题”(fundamental problem of causal inference)(Holland, 1986)。

因果推断的任务就是在一定的合理假设下，得到对反事实结果的无偏估计。

若个体i参与项目，则y_{1i}可观测；如果$\hat{y}_{0i}$为y_{0i}的估计值，则因果效应的估计值为$(y_{1i}-\hat{y}_{0i})$。

不同因果推断方法的区别主要在于如何估计$\hat{y}_{0i}$。

表达式(15.1)将y_i写为分段函数。更简洁地，可将y_i写为

$$y_i=(1-D_i)y_{0i}+D_iy_{1i}=y_{0i}+\underbrace{(y_{1i}-y_{0i})}_{=\tau_i}D_i=y_{0i}+\tau_iD_i \quad (15.3)$$

一般假定每位个体的处理效应定义良好(well defined)，且不依赖于其他个体是否受到处理。

此假定称为**个体处理值稳定假定**(Stable Unit Treatment Value Assumption，简记 SUTVA)，也简称为“稳定性假定”(the stability assumption)。

假定 15.1（稳定性假定，SUTVA）。任何个体的潜在结果不依赖于其他个体的处理状态，而且此个体的每个处理状态均对应于唯一的潜在结果。

假定 15.1 的前半部分意味着，不存在溢出效应，即个体i的潜在结果，与任何其他个体的处理状态无关。

这就排除了个体间的社会互动(social interactions)或一般均衡效应(general equilibrium effects)。

假定15.1的后半部分则意味着，给定个体i与处理水平(treatment level)$t \in \{0,1\}$，对应着唯一的潜在结果y_{ti}，故潜在结果是定义良好的。

表达式(15.1)已隐含地假定了SUTVA，因为它意味着$y_i = y_i(D_i)$只是D_i的函数，故定义良好，且不受D_j $(\forall\, j \neq i)$的影响。

由于处理效应$(y_{1i}-y_{0i})$为随机变量，可定义其期望值为**平均处理效应**(Average Treatment Effect，简记 ATE)，记为

$$\tau_{\text{ATE}} \equiv \text{E}(y_{1i}-y_{0i}) \qquad (15.4)$$

τ_{ATE}表示从总体中随机抽取某个体的期望处理效应，无论该个体是否参与项目。

另一常用概念仅考虑项目参与者的平均处理效应，称为**处理组平均处理效应**(Average Treatment Effect on the Treated，简记 ATT 或 ATET)，记为

$$\tau_{\text{ATT}} \equiv \text{E}(y_{1i}-y_{0i} \mid D_i=1) \qquad (15.5)$$

也可定义**控制组平均处理效应**(Average Treatment Effect on the Untreated，简记 ATU)，记为

$$\tau_{\text{ATU}} \equiv \text{E}(y_{1i} - y_{0i} \mid D_i = 0) \qquad (15.6)$$

这些被估计的对象(estimand)，统称为“因果效应参数”(causal estimands)。

我们主要关注τ_{ATE}与τ_{ATT}。

由于不能同时观测y_{0i}与y_{1i}，应如何估计 ATE 或 ATT？

例 假设总体包含 4 位病人，将其分为两组。

控制组($D=0$)采用保守的药物治疗，而处理组($D=1$)进行激进的手术治疗。

结果变量为治疗后的病人存活年限(y)。

表 1.1 列出了每位病人的潜在结果(y_{0i}与y_{1i})、处理效应(τ_i)、处理状态(D_i)与观测结果(y_i)。

表 1.1 因果效应参数的简单例子

个体	y_{0i}	y_{1i}	$\tau_i = y_{1i} - y_{0i}$	D_i	y_i
病人 1	1	7	6	1	7
病人 2	1	5	4	1	5
病人 3	6	5	-1	0	6
病人 4	8	7	-1	0	8

注：此例来自 Imbens and Rubin(2015, p. 14-15)。

平均处理效应为$\tau_{\text{ATE}} = (6+4-1-1)/4 = 2$

处理组平均处理效应为$\tau_{\text{ATT}} = (6+4)/2 = 5$

控制组平均处理效应为$\tau_{\text{ATU}} = (-1-1)/2 = -1$

或许由于医生知道哪种治疗方案对于不同病人最有利，故处理组中每位病人的处理效应均为正，而控制组中每位病人的处理效应都为负。

从观测结果来看，处理组的平均结果为$\bar{y}_{\text{treat}} = (7+5)/2 = 6$，反而低于控制组的平均结果$\bar{y}_{\text{control}} = (6+8)/2 = 7$。

若简单地以处理组与控制组的平均差异$(\bar{y}_{\text{treat}} - \bar{y}_{\text{control}})$来估计平均处理效应，则通常会导致选择偏差(selection bias)。

更一般地，可将处理组与控制组的平均差异分解为

$$
\begin{aligned}
&\mathrm{E}(y_i \mid D_i = 1) - \mathrm{E}(y_i \mid D_i = 0) \\
&= \mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0) \qquad \text{(根据(15.1)式)} \\
&= \underbrace{\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 1)}_{\tau_{\mathrm{ATT}}} + \underbrace{\mathrm{E}(y_{0i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)}_{\text{选择偏差}}
\end{aligned}
\tag{15.7}
$$

上式将处理组与控制组的平均差异分解为两部分，其中第一项为τ_{ATT}，而第二项为处理组的平均y_{0i}与控制组的平均y_{0i}之差(即这两类人若均未参与项目的平均差距)，即选择偏差。

回到就业培训的例子。

由于低收入者通常更倾向于选择参加培训项目，故选择偏差一般为负，导致处理组与控制组的平均差异(即$\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)$)可能低估处理组平均处理效应($\tau_{\mathrm{ATT}}$)。

如果选择偏差的绝对值足够大，甚至可导致$\left[\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)\right] < 0$，即处理组的平均收入反而低于控制组的情形。

由于个体通常根据其参加项目的期望收益$\mathrm{E}(y_{1i} - y_{0i})$而自我选择是否参加项目，导致对平均处理效应的估计带来困难，这称为**选择难题**(the selection problem)。

如何解决此选择难题，取决于究竟个体如何被分配到处理组与控制组，即所谓**分配机制**(assignment mechanism)。

如果分配机制为随机实验(randomized experiment)，选择难题可迎刃而解。

在随机分组(random assignment)的情况下，个体i的D_i(即是否参加项目)通过抛硬币或电脑随机数而决定，故D_i独立于(y_{0i}, y_{1i})，记为

$$D_i \perp (y_{0i}, y_{1i}) \qquad (15.8)$$

其中，“$\perp$”表示相互独立。

此时，处理组($D_i = 1$)与控制组($D_i = 0$)的潜在结果y_{0i}并无系统差异，故(15.7)式中的选择偏差为 0：

$$\mathrm{E}(y_{0i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0) = \mathrm{E}(y_{0i}) - \mathrm{E}(y_{0i}) = 0 \quad (15.9)$$

其中，由于y_{0i}独立于D_i，故y_{0i}均值独立于D_i，因此条件期望等于无条件期望。

在随机实验的情况下，可用处理组与控制组的平均差异来估计τ_{ATT}。

在随机分组的情况下，三种平均处理效应均相等，即$\tau_{\mathrm{ATE}} = \tau_{\mathrm{ATT}} = \tau_{\mathrm{ATU}}$(参见习题)。

随机实验并非在所有情况下都可行，可能成本太高，或道德上不可行。

如果只有观测数据，则D_i的决定一般将受到(y_{0i}, y_{1i})的影响，故存在选择偏差。

个体是否参与项目(D_i)的决定通常受到潜在结果(y_{0i}, y_{1i})的影响，这有时称为**罗伊模型**(Roy model)。

如果个体是否参与项目的分配机制为仅依据可观测变量进行选择，则可使用匹配的方法进行因果推断。

15.2 依可测变量选择

除了(y_i, D_i)之外，通常还可观测到个体i的一些特征，比如年龄、性别、培训前收入，记为向量$\mathbf{x}_i$，也称为“协变量”(covariates)或“处理前变量”(pretreatment variables)。

一般假定协变量$\mathbf{x}_i$为前定变量，并不受处理变量D_i的影响。

如果个体i对D_i的选择完全取决于可观测的$\mathbf{x}_i$，则称为**依可测变量选择**(selection on observables)，就可找到估计处理效应的合适方法(即使没有工具变量)。

如果个体对D_i的选择完全取决于$\mathbf{x}_i$，则在给定$\mathbf{x}_i$的情况下，D_i将独立于潜在结果(y_{0i}, y_{1i})，这就是 Rosenbaum and Rubin (1983)所引入的**非混杂性假定**(unconfoundedness assumption)。

假定 15.2（非混杂性假定，Unconfoundedness） 给定协变量$\mathbf{x}_i$，则D_i独立于(y_{0i}, y_{1i})，记为$D_i \perp (y_{0i}, y_{1i}) | \mathbf{x}_i$，其中“⊥”表示相互独立。

根据假定 15.2，给定$\mathbf{x}_i$，则个体i进入处理组$(D_i = 1)$或控制组$(D_i = 0)$，与潜在结果(y_{0i}, y_{1i})无关。

给定$\mathbf{x}_i$后，将“不存在混杂变量”(no confounder)，故名“非混杂性”(unconfoundedness)。

非混杂性的含义是，只要把$\mathbf{x}_i$包括在回归方程中，就能完全解决遗漏变量偏差(即使有遗漏变量，也与解释变量不相关)，从而避免变量间作用关系的混杂(confounding)，使得D_i不再有内生性。

给定$\mathbf{x}_i$后，(y_{0i}, y_{1i})对于D_i的影响可以忽略(因为二者相互独立)，故此假定也称为**可忽略性**(ignorability)或“依可测变量选择”(selection on observables)。

假定 15.2 意味着，在给定$\mathbf{x}_i$的条件下，D_i条件独立于(y_{0i}, y_{1i})，故也称为**条件独立假定**(Conditional Independence Assumption，简记 CIA)。

假定 15.2 意味着，在给定$\mathbf{x}_i$的条件下，D_i的取值可视为随机分配(as good as randomly assigned, conditional on $\mathbf{x}_i$)，也称为“基于协变量的随机实验”(randomization based on covariates)。

条件独立假定相当于一种**条件随机实验**(conditionally randomized trial)，也称为**分层随机实验**(stratified randomized experiment)或**分层随机抽样**(stratified random sampling)。

此时，对于$\mathbf{x}_i$完全相同的个体，其进入处理组的概率必然相等。

反之，对于$\mathbf{x}_i$不同的个体，其进入处理组的概率则可能不同，参见图 15.1。

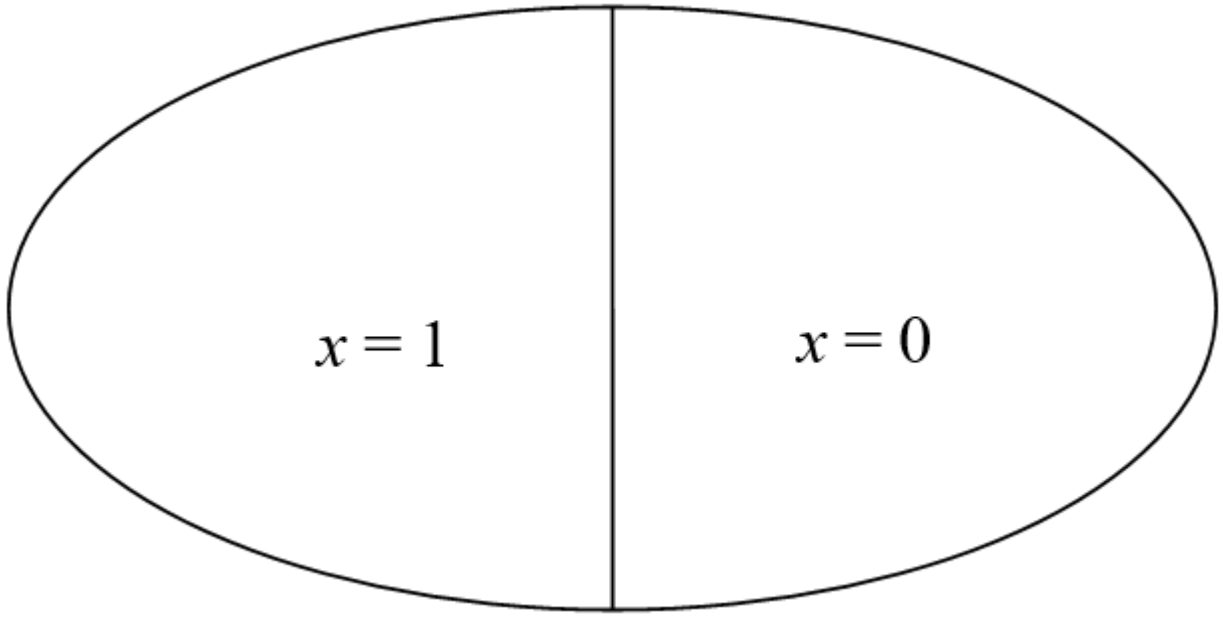

图 15.1 分层随机实验的示意图

在图 15.1 中，假设协变量$\mathbf{x}_i$仅含一个变量，若$x=1$则为男性，而$x=0$为女性。

假定样本包括 20 位个体，其中男女各 10 人，现欲将样本随机分为两组，即处理组与控制组。考虑以下几种随机实验的方法。

例 伯努利实验(Bernoulli trial)。使用一枚公平的硬币(a fair coin)，采用抛硬币的方法决定每位个体进入处理组或控制组。

每位个体进入处理组的概率均为 0.5，且相互独立。

伯努利实验的缺点：无法保证处理组与控制组的个体数目相等。

在极端情况下，甚至可能出现 20 位个体均进入处理组，或都分到控制组的情形；尽管此概率非常小(但依然为正数)。

例 完全随机实验(completely randomized experiment)。

完全随机实验将固定数目的个体随机地分配到处理组。

例如，可将 20 位个体的编号标签放入箱中，然后从中随机抽出 10 位个体。

完全随机实验并不能保证关键协变量的平衡。

在极端情况下，有可能抽中的 10 位个体均为男性，或都是女性。

例 分层随机实验(stratified randomized experiment)。

分层随机实验先将样本分为若干“层”(strata)或“块”(blocks)，使得在每一层内，个体的协变量相同或近似。

然后，在针对每一层，进行完全随机实验。

例如，将样本分为$x=1$的“男性层”(包含 10 位男性)与$x=0$的“女性层”(包含 10 位女性)；再通过完全随机实验从男性层与女性层分别抽取 5 位个体。

假定 15.2 的非混杂性假定相当于分层随机实验，这意味着给定协变量$\mathbf{x}$，则D_i的取值可视为随机分配。

15.3 回归 vs 匹配

与“潜在结果框架”(potential outcomes framework)相对应，若仅使用可观测结果y_i进行分析，则称为**观测结果框架**(observed outcome framework)。

作为对比，下面我们回到观测结果框架，以y_i作为被解释变量进行回归分析。首先，考虑最简单的一元回归：

$$y_i = \tau D_i + \varepsilon_i \qquad (15.10)$$

扰动项ε_i包括了不少遗漏变量，比如协变量$\mathbf{x}_i$与不可观测的潜在结果(y_{0i}或y_{1i})。

如果样本数据来自完全随机实验，则 OLS 估计量$\hat{\tau}$可一致地估计平均处理效应τ_{ATE}。

如果样本数据来自分层随机实验，则D_i一般并不独立于(y_{0i}, y_{1i})，故存在选择偏差，导致一元回归的估计量不一致。

由于非混杂性成立，故原则上可将$\mathbf{x}_i$作为控制变量引入回归方程，以解决遗漏变量问题：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \tau D_i + \varepsilon_i \qquad (15.11)$$

但我们通常并不清楚$\mathbf{x}_i$是否应以线性形式进入上述方程。

若遗漏了非线性项，依然可能存在遗漏变量偏差。

解决方法之一正是本章要介绍的匹配估计量。

在使用匹配估计量时，将分别估计每位个体的处理效应，若找不到合适的匹配则去掉该个体。

另一方面，若进行回归，则将所有个体都混在一起作回归(限制所有个体的处理效应均为τ)，无视个体间是否具有可比性。

非混杂性假定(unconfoundedness)是一个很强的假定。

它意味着回归方程已包含所有必要的控制变量，故不存在任何与解释变量相关的遗漏变量或混杂因素。

非混杂性假定排除了“依不可测变量选择”(selection on unobservables)的情形，比如根据不可观测的个体能力进行选择。

非混杂性假定本质上不可检验。

如果$\mathbf{x}_i$已包含较丰富的协变量(a rich set of covariates)，则或可认为 CIA 假定基本得到满足，遗漏变量偏差较小。

15.4 匹配估计量的思想

假设个体i属于处理组，匹配估计量的基本思路是，找到属于控制组的某个体j，使得个体j与个体i的可测变量取值尽可能相似(匹配)，即$\mathbf{x}_i \approx \mathbf{x}_j$。

基于非混杂性假定，则个体i与个体j进入处理组的概率相近(类似于随机分组)，具有可比性；故可将y_j作为y_{0i}的估计量，即$\hat{y}_{0i} = y_j$。

可将$(y_i - \hat{y}_{0i}) = y_i - y_j$作为对个体$i$处理效应的估计。

对处理组的每位个体都如此进行匹配。

对控制组每位个体也进行匹配，然后对每位个体的处理效应进行平均，即可得到**匹配估计量**(matching estimator)。

由于匹配的具体方法不同，故存在不同的匹配估计量。

首先，是否有放回。

如果**无放回**(no replacement)，则每次都将匹配成功的个体(i, j)从样本中去掉，不再参与其余匹配。

如果**有放回**(with replacement)，则将匹配成功的个体留在样本中，参与其余匹配(这导致一位个体可能与多位不同组个体匹配)。

一般地，无放回匹配的效率低于有放回匹配。

无放回匹配的另一缺点是，其匹配结果一般与样本中个体排序有关，故需先将个体随机排序。

其次，是否允许**并列**(ties)，比如控制组个体 j 与 k 的可测变量都与处理组个体 i 一样接近。

如果允许并列，则将 y_j 与 y_k 的平均值作为 y_{0i} 的估计量，即 $\hat{y}_{0i} = (y_j + y_k)/2$。

如果不允许并列，则算法将根据数据排序选择个体 j 或 k；此时，匹配结果可能与数据排序有关，故建议先将样本随机排序，再进行匹配。

允许并列的做法更有效率。

对于有放回且允许并列的匹配，其匹配结果与个体排序无关，故无须将观测值随机排序。

以上为**一对一**(one-to-one)匹配，也可进行**一对多**匹配。

比如，一对四匹配，即针对每位个体寻找四位不同组的最近个体进行匹配。

一般地，匹配估计量存在偏差(bias)，除非在**精确匹配**(exact matching)的情况下，即对于所有匹配都有$\mathbf{x}_i = \mathbf{x}_j$。

更常见的则是**非精确匹配**(inexact matching)，即只能保证$\mathbf{x}_i \approx \mathbf{x}_j$。

在非精确匹配的情况下，若进行一对一匹配，则偏差较小，但方差较大(因为仅使用一位最近的邻居)

进行一对多匹配可降低方差(因为使用了更多邻居的信息)，但代价是偏差增大(因为使用了更远邻居的信息)。

Abadie, Drukker, Herr and Imbens(2004)建议进行一对四匹配，在一般情况下可最小化均方误差(MSE)。

尽管如此，一对多匹配可能使得匹配结果更难满足数据平衡的要求，故实践中一对一匹配依然流行。

15.5 倾向得分匹配

为满足非混杂性假定，$\mathbf{x}_i$通常需包括较多变量，比如$\mathbf{x}_i$为K维向量。

若直接以$\mathbf{x}_i$进行匹配，则需在高维空间进行匹配，可能遇到数据稀疏的问题，很难找到与$\mathbf{x}_i$相近的$\mathbf{x}_j$与之匹配。

这是“维度诅咒”(curse of dimensionality)的一种表现。

一种解决方法是，使用某函数$f(\mathbf{x}_i)$，将K维向量$\mathbf{x}_i$的信息压缩到一维，再根据$f(\mathbf{x}_i)$进行匹配。

Rosenbaum and Rubin (1983)提出使用“倾向得分”(propensity score，简记 p-score)，将K维向量$\mathbf{x}_i$的信息压缩到一维的$[0,1]$区间。

定义(倾向得分) 个体i的倾向得分为，在给定$\mathbf{x}_i$的情况下，个体i进入处理组的条件概率，即$p(\mathbf{x}_i) \equiv \mathrm{P}(D_i = 1 \mid \mathbf{x} = \mathbf{x}_i)$，简记$p(\mathbf{x})$(省略下标$i$)。

在以样本数据估计$p(\mathbf{x})$时，可使用 probit 或 logit，而最流行的方法为 logit。

使用倾向得分度量个体间距离的好处在于，它不仅是一维变量，而且取值介于$[0,1]$之间。

比如，即使$\mathbf{x}_i$与$\mathbf{x}_j$距离较远，但仍可能$p(\mathbf{x}_i) \approx p(\mathbf{x}_j)$。

使用倾向得分进行匹配，称为**倾向得分匹配**(Propensity Score Matching，简记 PSM)。

PSM 的理论依据在于，如果非混杂性假定成立，则只须在给定 $p(\mathbf{x})$的情况下，D_i就独立于(y_{0i}, y_{1i})。

命题（倾向得分定理） 如果非混杂性假定成立，即 $D \perp (y_0, y_1) | \mathbf{x}$，则$D \perp (y_0, y_1) | p(\mathbf{x})$。

证明：使用迭代期望定律可证，参见附录。

倾向得分定理是 PSM 的理论基础。

它意味着倾向得分相同的个体i与个体j具有可比性(而不必强求协变量完全相同)，可视为随机分组，并据此进行反事实的因果推断。

为了能够进行匹配，需要在$\mathbf{x}$的每个可能取值上都同时存在处理组与控制组的个体。

这就是“重叠假定”(overlap assumption)或“匹配假定”(matching assumption)。

假定 15.3（重叠假定）对于$\mathbf{x}$的任何可能取值，都有$0 < p(\mathbf{x}) < 1$。

此假定意味着处理组与控制组这两个子样本存在重叠，故名“重叠假定”。

它是进行匹配的前提，故也称“匹配假定”。

它保证了处理组与控制组的倾向得分有共同的取值范围，即所谓**共同支撑**(common support)，参见图 15.2。

如果假定 15.3 不成立，则意味着可能存在某些$\mathbf{x}$，使得$p(\mathbf{x})=1$，即这些个体都属于处理组，无法找到与之匹配的控制组个体；

另一方面，也可能存在某些$\mathbf{x}$，使得$p(\mathbf{x})=0$，即这些个体都属于控制组，无法找到与其匹配的处理组个体。

在进行匹配时，为了提高匹配质量，有时仅保留倾向得分重叠部分的个体；尽管这样做可能损失样本容量。

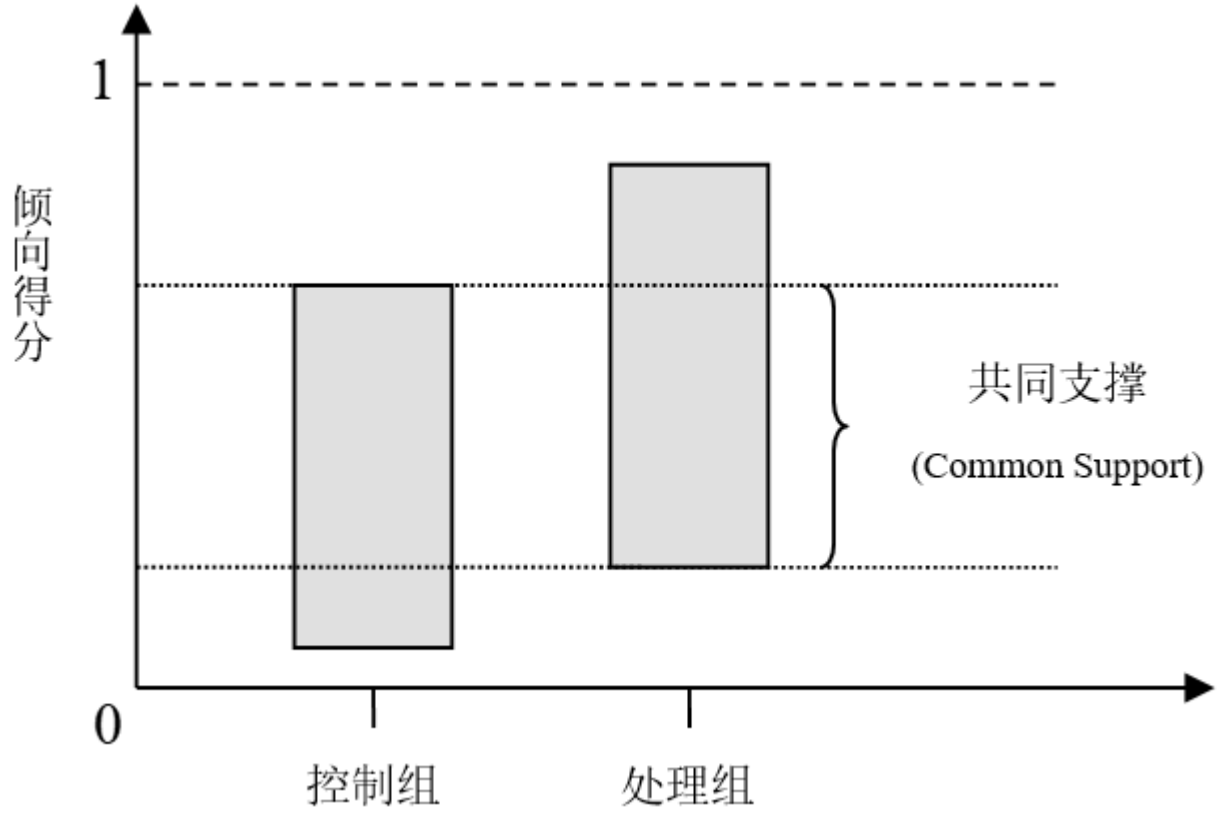

图 15.2 倾向得分的共同取值范围

通过 PSM 估计平均处理效应的一般步骤如下。

(1) 选择协变量$\mathbf{x}_i$。尽量将可能影响(y_{0i}, y_{1i})与D_i的相关变量包括进来，以保证非混杂性假定得到满足。如果协变量$\mathbf{x}_i$太少或选择不当，则可能导致非混杂性假定不满足，从而引起偏差。

(2) 估计倾向得分，一般使用 Logit 回归。Rosenbaum and Rubin (1985)建议使用形式灵活的 logit 模型，比如包含$\mathbf{x}_i$的高次项与交互项。

(3) 进行倾向得分匹配，具体方法参见下文。

(4) 检验匹配样本是否满足数据平衡的要求(参见下文)。若未满足，则回到第(2)步甚至第(1)步，重新估计倾向得分，直至通过数据平衡的检验。

(5) 根据匹配后样本(matched sample)计算平均处理效应。处理组平均处理效应(ATT)估计量的一般表达式为

$$\hat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{D_i=1} (y_i - \hat{y}_{0i}) \qquad (15.12)$$

$N_1 = \sum_{i=1}^{N} D_i$ 为处理组个体数，$\sum_{D_i=1}$ 表示仅对处理组个体加总。

控制组平均处理效应(ATU)估计量的一般表达式为

$$\hat{\tau}_{\text{ATU}} = \frac{1}{N_0} \sum_{D_j=0} (\hat{y}_{1j} - y_j) \qquad (15.13)$$

$N_0 = \sum_{j=1}^{N} (1 - D_j)$ 为控制组个体数，$\sum_{D_j=0}$ 表示仅对控制组个体加总

整个样本的平均处理效应(ATE)估计量的一般表达式为

$$\tau_{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{1i} - \hat{y}_{0i}) \qquad (15.14)$$

其中，$N = N_0 + N_1$为样本容量；如果$D_i = 1$，则$\hat{y}_{1i} = y_i$；如果$D_i = 0$，则$\hat{y}_{0i} = y_i$。

在进行倾向得分匹配时，有不同的具体方法。

方法之一为k近邻匹配(k-nearest neighbor matching)，即寻找倾向得分最近的k个不同组个体。

如果$k = 1$，则为“一对一匹配”(one-to-one matching)。

即使“最近邻居”也可能相去甚远，从而失去可比性。

方法之二限制倾向得分的绝对距离$\left|p_i - p_j\right| \le \varepsilon$，一般建议$\varepsilon \le 0.25\hat{\sigma}_{pscore}$，其中$\hat{\sigma}_{pscore}$为倾向得分的样本标准差；这被称为**卡尺匹配**(caliper matching)或**半径匹配**(radius matching)。

方法之三为**卡尺内k近邻匹配**(k-nearest neighbor matching within caliper)，即在给定的卡尺ε范围内，寻找最近的k个邻居进行匹配。

在实际进行匹配时，究竟应使用以上哪种具体方法或参数(比如，k近邻匹配的k取值，是否放回，是否允许并列)，目前文献中尚无明确指南。

一般认为，不存在适用于一切情形的绝对好方法，只能根据具体数据来选择匹配方法。

比如，若样本容量不大，则应进行有放回的匹配。

又比如，若存在较多具有可比性的控制组个体，则可考虑一对多匹配，以提高匹配效率。

在实践中，一般建议尝试不同的匹配方法，然后比较其结果(类似于稳健性检验)。

如果不同方法的结果相似，则说明结果稳健，不依赖于具体方法；反之，如果存在较大差异，则应考察造成此差异的原因。

15.6 倾向得分的平衡性质

倾向得分还具有**平衡性质**(balancing property)，即对于倾向得分相同的个体而言，其协变量在处理组与控制组的分布相同。

这是因为，给定$p(\mathbf{x})$之后，处理变量D与协变量$\mathbf{x}$相互独立，即$D \perp \mathbf{x} \mid p(\mathbf{x})$。

命题(**倾向得分的平衡性质**) $D \perp \mathbf{x} \mid p(\mathbf{x})$。

证明：由于D为虚拟变量，故只需证明，给定$p(\mathbf{x})$后，“$D=1$”的条件概率不再依赖于$\mathbf{x}$：

$$
\begin{aligned}
&\mathrm{P}\left(D=1 \mid \mathbf{x}, p(\mathbf{x})\right) \\
&=\mathrm{P}\left(D=1 \mid \mathbf{x}\right) \qquad \text{(给定}\mathbf{x}\text{，则}p(\mathbf{x})\text{也给定)} \qquad (15.15) \\
&=p(\mathbf{x}) \qquad \text{(倾向得分的定义)}
\end{aligned}
$$

由于 $\mathrm{P}\left(D=1 \mid \mathbf{x}, p(\mathbf{x})\right)=p(\mathbf{x})$，故给定 $p(\mathbf{x})$ 后，$\mathrm{P}\left(D=1 \mid \mathbf{x}, p(\mathbf{x})\right)$不再依赖于$\mathbf{x}$，因此$D \perp \mathbf{x} \mid p(\mathbf{x})$。

平衡性质是倾向得分本身的性质。

在此命题的证明中，并未使用非混杂性假定，故不依赖于非混杂性假定。

倾向得分的平衡性质意味着，对于倾向得分相同或接近的“匹配样本”(matched sample)而言，无论$D=1$(处理组)，还是$D=0$(控制组)，均不影响协变量$\mathbf{x}$的分布(或影响很小)。

如果倾向得分估计得较准确，则应使得$\mathbf{x}_i$在匹配后的处理组与控制组之间分布较均匀。

例如，匹配后的处理组均值$\bar{\mathbf{x}}_{treat}$与控制组均值$\bar{\mathbf{x}}_{control}$较为接近；这称为**数据平衡**(data balancing)。

但$\bar{\mathbf{x}}_{treat}$与$\bar{\mathbf{x}}_{control}$的差距显然与计量单位有关，故一般针对$\mathbf{x}$的每个分量$x$考察如下**标准化差距**(standardized differences)或**标准化偏差**(standardized bias)：

$$\frac{\left|\overline{x}_{treat}-\overline{x}_{contrl}\right|}{\sqrt{(s_{x,treat}^2+s_{x,control}^2)/2}} \qquad (15.16)$$

其中，$s_{x,treat}^2$与$s_{x,control}^2$分别为处理组与控制组变量x的样本方差。

一般要求此标准化差距不超过 10%。

如果标准化偏差超过 10%，则应回到 PSM 估计的第(2)步，甚至第(1)步，重新估计倾向得分；或者改变具体的匹配方法。

除了考察匹配后处理组与控制组的一阶矩差别(即$\overline{\mathbf{x}}_{treat}$与$\overline{\mathbf{x}}_{control}$的差别)，还可考察其二阶矩的差别。

比如，对于协变量x，可计算匹配后处理组与控制组的“方差比”(variance ratio)，即$s^2_{x,treat} \big/ s^2_{x,control}$。

若此方差比接近于 1，则说明匹配后处理组与控制组的方差相近。

通过标准化差距与方差比考察匹配后的数据平衡情况，称为**数据平衡检验**(data balancing test)。

15.7 倾向得分匹配的 Stata 案例

倾向得分匹配可通过 Stata 官方命令 `teffects psmatch` 来实现；其中，`teffects` 表示 treatment effects。

该命令的一般格式为

```
. teffects psmatch (y) (D x1 x2 x3, probit), atet
nn(#) caliper(#) osample(newvar) pstolerance(#)
```

“y”为结果变量(outcome variable)，“D”为处理变量(treatment variable)，而“x1 x2 x3”为协变量。

选择项“probit”表示使用 probit 估计倾向得分，默认使用 logit。

选择项“atet”表示估计 ATE on the treated (即 ATT)；默认估计 ATE。

选择项“nn(#)”表示进行 1 对#匹配，比如“nn(3)”为 1 对 3 匹配；默认为“nn(1)”。

选择项“caliper(#)”表示进行卡尺内匹配。

选择项“osample(newvar)”表示生成新变量，作为“overlap violation indicator”，记录违背重叠假定的观测值。

Stata 检查以下两种违背的情形。第一，指定进行卡尺内近邻匹配，但在卡尺内找不到足够的邻居。第二，个体的倾向得分过于靠近 0 或 1；默认标准为倾向得分小于 0.00005 或大于 0.99995；可使用选择项“pstolerance(#)”修改此默认标准。

Stata 官方命令 teffects psmatch 只能进行有放回且允许并列的匹配，故无须先将观测值随机排序。

若需进行无放回或不允许并列的匹配，可使用非官方命令 psmatch2，但 psmatch2 所提供的标准误不正确(未考虑倾向得分为估计所得而非已知)。

psmatch2 所提供的匹配方法更为丰富，详见陈强(2014，第 28 章)。

随机实验常被视为因果推断的“黄金标准”(gold standard)，而观测数据的推断结果一般可信度更低。

LaLonde (1986)使用实验数据的结果作为参照系，评估计量方法用于观测数据所得结果的可靠性。

为了支持弱势工人(disadvantaged workers)，美国在 1975 年 3 月至 1977 年 7 月推出一个就业培训的实验项目，称为“国家工作支持示范项目”(National Supported Work Demonstration Program，简记 NSW)。

参加者包括 AFDC 女性[①]、有吸毒前科者(ex-drug addicts)，有犯罪前科者(ex-criminal offenders)，以及高中辍学者(high school dropouts)。

项目参加者被随机地分为实验组与控制组。

[①] AFDC 指“Aid to Family with Dependent Children”，是美国政府的福利项目。参加 NSW 项目的 AFDC 女性须满足以下条件：(1) 当前处于失业状态；(2) 在过去 6 个月中，至少工作满 3 个月；(3) 没有 6 岁以下的小孩；(4) 在过去 36 个月中，至少有 30 个月从 AFDC 得到收入；参见 LaLonde (1986, p.605)。

实验组成员被安排真实的工作(比如加油站、复印店或建筑工地)

控制组成员则未获任何帮助。

实验组成员的工资低于正常工作的市场工资，但若业绩良好可获加薪。

LaLonde (1986) 首先使用实验数据，获得平均处理效应的无偏估计。

然后，LaLonde (1986)保留实验组的数据，但将控制组替换为观测数据。

观测数据来自“动态收入面板研究”(Panel Study of Income Dynamics，简记 PSID)或“当前人口调查”(Current Population Survey，简记 CPS)。

结果发现，无论使用 OLS，双重差分法或样本选择模型，所得估计结果均与实验结果相去甚远，甚至得到负的平均处理效应。

一时之间，观测数据及相应计量方法的可信度广受质疑。

但 Dehejia and Wahba (1999)发现，将 PSM 用于 LaLonde (1986)构造的混合数据(处理组为实验数据，而控制组为观测数据)，仍可得到与实验结果相近的估计结果。

从此倾向得分匹配变得流行起来。

我们以 Dehejia and Wahba (1999)所提供的 NSW 实验数据集 nsw_dw.dta 与 CPS 观测数据集 cps_controls.dta 为例进行 PSM 的演示。

数据集包括：结果变量 *re78* (1978 年实际收入)，处理变量 *treat*(是否参加就业培训)，协变量 *age*(年龄)，*education*(教育年限)，*black*(是否黑人)，*hisp*(是否拉丁裔)，*married*(是否已婚)，*nodegree* (无高中学历=1，反之为 0)，*re74*(1974 年实际收入)，*re75*(1975 年实际收入)。

首先，载入实验数据，并考察协变量在实验组与控制组之间是否平衡。

```
. use nsw_dw.dta, clear
. bysort treat: sum
```

其中，前缀“bysort treat”表示根据变量 *treat* 的取值划分子样本，并分别运行之后的命令 sum。

-> treat = 0

Variable	Obs	Mean	Std. dev.	Min	Max
data_id	0				
treat	260	0	0	0	0
age	260	25.05385	7.057745	17	55
education	260	10.08846	1.614325	3	14
black	260	.8269231	.3790434	0	1
hispanic	260	.1076923	.3105893	0	1
married	260	.1538462	.3614971	0	1
nodegree	260	.8346154	.3722439	0	1
re74	260	2107.027	5687.906	0	39570.68
re75	260	1266.909	3102.982	0	23031.98
re78	260	4554.801	5483.836	0	39483.53

-> treat = 1

Variable	Obs	Mean	Std. dev.	Min	Max
data_id	0				
treat	185	1	0	1	1
age	185	25.81622	7.155019	17	48
education	185	10.34595	2.01065	4	16
black	185	.8432432	.3645579	0	1
hispanic	185	.0594595	.2371244	0	1
married	185	.1891892	.3927217	0	1
nodegree	185	.7081081	.4558666	0	1
re74	185	2095.574	4886.62	0	35040.07
re75	185	1532.055	3219.251	0	25142.24
re78	185	6349.144	7867.402	0	60307.93

控制组($treat = 0$)有 260 个观测值，而实验组($treat = 1$)有 185 个观测值，除了结果变量 *re78* 外，所有协变量的均值在两组之间都较为接近，协变量基本平衡。

由于协变量较多，可用命令 `global` 定义一个名为“*cov*”(表示 covariates)的“全局宏”(global macro)，以指代所有协变量：

```
. global cov age education black hispanic married nodegree re74 re75
```

定义好全局宏 *cov* 后，只需使用`$cov`(在之前加上美元符号)，即可调用这些控制变量。

例如，聚焦于比较两组的协变量均值，可输入命令：

```
. tabstat $cov,by(treat)
```

命令 `tabstat` 表示“table of summary statistics”，即将统计指标列表。

选择项“`by(treat)`”表示根据 *treat* 的取值划分子样本分别计算。

```
Summary statistics: Mean
Group variable: treat

   treat |       age  educat~n     black  hispanic   married  nodegree      re74      re75
---------+------------------------------------------------------------------------------
       0 |  25.05385  10.08846  .8269231  .1076923  .1538462  .8346154  2107.027  1266.909
       1 |  25.81622  10.34595  .8432432  .0594595  .1891892  .7081081  2095.574  1532.055
---------+------------------------------------------------------------------------------
   Total |  25.37079  10.19551  .8337079  .0876404  .1685393  .7820225  2102.265  1377.138
------------------------------------------------------------------------------------------
```

若想汇报更多统计指标(比如标准差)，可输入命令：

```
. tabstat re78 $cov,by(treat) stat(mean sd) nototal
```

其中，选择项“stat(mean sd)”指定计算均值与标准差，默认为“stat(mean)”；选择项“nototal”表示不汇报整个样本的统计结果。

```
Summary statistics: Mean, SD
Group variable: treat

   treat |       age  educat~n     black  hispanic   married  nodegree      re74      re75
---------+------------------------------------------------------------------------------
       0 |  25.05385  10.08846  .8269231  .1076923  .1538462  .8346154  2107.027  1266.909
         |  7.057745  1.614325  .3790434  .3105893  .3614971  .3722439  5687.906  3102.982
---------+------------------------------------------------------------------------------
       1 |  25.81622  10.34595  .8432432  .0594595  .1891892  .7081081  2095.574  1532.055
         |  7.155019   2.01065  .3645579  .2371244  .3927217  .4558666   4886.62  3219.251
---------------------------------------------------------------------------------------
```

可使用回归的方法考察两组之间的协变量平衡。例如，将变量 *treat* 对所有协变量进行 logit 回归。

```
. logit treat $cov,r nolog
```

```
Logistic regression                                     Number of obs =    445
                                                        Wald chi2(8)  =  16.18
                                                        Prob > chi2   = 0.0399
Log pseudolikelihood = -293.60822                       Pseudo R2     = 0.0281
```

treat	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
age	.0046982	.0138994	0.34	0.735	-.0225441	.0319404
education	-.071239	.0705089	-1.01	0.312	-.209434	.0669559
black	-.2247005	.3671884	-0.61	0.541	-.9443765	.4949755
hispanic	-.8527817	.5095363	-1.67	0.094	-1.851454	.1458911
married	.1636176	.282491	0.58	0.562	-.3900545	.7172898
nodegree	-.9035053	.3157919	-2.86	0.004	-1.522446	-.2845646
re74	-.0000316	.0000245	-1.29	0.197	-.0000797	.0000164
re75	.0000616	.0000415	1.49	0.137	-.0000197	.0001429
_cons	1.177674	1.034817	1.14	0.255	-.8505304	3.205878

几乎所有协变量都不显著，对于分组变量 *treat* 无解释力；正如我们的预期。

但变量 *nodegree* 却在 1%的水平上显著为负。实验组无高中学历占比为 70.81%，而控制组则高达 83.46%，二者有明显差异；这或许为随机抽样的偶然性所致。

对于此实验数据，使用一元回归即可得到一致估计。

```
. reg re78 treat,r
```

```
Linear regression                               Number of obs     =        445
                                                F(1, 443)         =       7.15
                                                Prob > F          =     0.0078
                                                R-squared         =     0.0178
                                                Root MSE          =     6579.5

------------------------------------------------------------------------------
             |               Robust
        re78 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       treat |   1794.342   670.8245     2.67   0.008     475.9486    3112.736
       _cons |   4554.801   340.2038    13.39   0.000     3886.187    5223.415
------------------------------------------------------------------------------
```

平均处理效应为1794.34，即参加就业培训平均能使1978年实际收入提高1,794.34美元，且在1%的水平上显著。

此回归的R^2很低，仅为0.0178(即是否参加就业培训仅能解释1978年实际收入1.78%的变动)。

由于随机分组的偶然性，实验组与控制组的协变量也并非完全平衡。为此，下面引入协变量，进行更可信的多元回归。

```
. reg re78 treat $cov, r
```

Linear regression

					Number of obs	=	445
					F(9, 435)	=	2.66
					Prob > F	=	0.0052
					R-squared	=	0.0548
					Root MSE	=	6513.5

re78	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
treat	1676.343	676.7337	2.48	0.014	346.2683	3006.417
age	55.31668	41.2858	1.34	0.181	-25.82778	136.4611
education	395.7343	197.3434	2.01	0.046	7.86916	783.5995
black	-2159.522	1011.814	-2.13	0.033	-4148.174	-170.8708
hispanic	164.0327	1366.909	0.12	0.905	-2522.534	2850.6
married	-138.7253	873.4508	-0.16	0.874	-1855.434	1577.983
nodegree	-70.68064	1026.725	-0.07	0.945	-2088.638	1947.277
re74	.0821412	.1068727	0.77	0.443	-.1279099	.2921924
re75	.0527641	.1241808	0.42	0.671	-.1913049	.2968331
_cons	785.0614	3297.291	0.24	0.812	-5695.542	7265.665

在控制协变量后，平均处理效应降为 1676.34(变化不大)，且显著性水平接近 1%(*p* 值为 1.4%)。

在协变量中，除了 *education*(教育年限)与 *black*(是否黑人)在 5%水平上显著为，其余协变量均不显著。

综合以上一元与多元回归的结果，可认为真实的处理组平均处理效应(ATT)大约为 1700。

遵照 LaLonde (1986)与 Dehejia and Wahba (1999)的做法，去掉控制组(*treat* = 0)的观测值，然后加入 CPS 的数据作为控制组。

```
. drop if treat == 0
. append using cps_controls.dta
```

此命令将数据集 cps_controls.dta 直接附加(append)到内存中原有数据集之后。再次考察此混合数据集的协变量平衡性。

```
. tabstat $cov,by(treat) nototal
```

```
Summary statistics: Mean
Group variable: treat

   treat |       age  educat~n     black  hispanic   married  nodegree      re74      re75
---------+------------------------------------------------------------------------------
       0 |  33.22524  12.02751  .0735368   .072036  .7117309  .2958354   14016.8   13650.8
       1 |  25.81622  10.34595  .8432432  .0594595  .1891892  .7081081  2095.574  1532.055
---------------------------------------------------------------------------------------
```

来自 CPS 的控制组与实验组的协变量均值有很大差异。与实验组相比，控制组的平均年龄高 7 岁多，教育年限长一年多，黑人占比从 84.32%大幅降为 7.35%，拉丁裔占比从 5.95%略升为 7.20%，已婚占比从 18.92%跃升为 71.17%，无高中学历占比从 70.81%减少为 29.58%，而 1974 年与 1975 年的实际收入则分别高出近 7 倍与近 9 倍。

来自 CPS 的个体可视为 1970 年代美国人的典型抽样，而来自 NSW 的实验组则由底层民众所构成。

使用此样本进行 OLS 回归，将存在严重的选择偏差。

首先，进行一元回归。

```
. reg re78 treat,r
```

```
Linear regression                               Number of obs     =     16,177
                                                F(1, 16175)       =     213.24
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0087
                                                Root MSE          =       9629

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -8497.516   581.9158   -14.60   0.000    -9638.135   -7356.897
       _cons |   14846.66   76.29073   194.61   0.000     14697.12    14996.2
------------------------------------------------------------------------------
```

一元回归所估计的平均处理效应竟为-8497.52，且在 1%的水平上显著。处理组与控制组存在巨大的初始差异，而 NSW 项目的正效应并不足够抵消此负向的选择偏差。

其次，将协变量引入多元回归，增加处理组与控制组的可比性。

```
. reg re78 treat $cov, r
```

Linear regression

					Number of obs	=	16,177
					F(9, 16167)	=	1903.96
					Prob > F	=	0.0000
					R-squared	=	0.4758
					Root MSE	=	7003.7

re78	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
treat	699.1317	616.6538	1.13	0.257	-509.5781	1907.841
age	-101.8551	5.714843	-17.82	0.000	-113.0568	-90.65335
education	160.1864	28.74707	5.57	0.000	103.8389	216.5338
black	-836.9473	196.8362	-4.25	0.000	-1222.768	-451.1266
hispanic	-218.3184	218.0536	-1.00	0.317	-645.7277	209.0909
married	73.07576	144.4096	0.51	0.613	-209.983	356.1345
nodegree	372.2376	174.8348	2.13	0.033	29.54208	714.9331
re74	.2894924	.0150379	19.25	0.000	.2600164	.3189684
re75	.4707289	.01531	30.75	0.000	.4407196	.5007383
_cons	5735.731	443.758	12.93	0.000	4865.916	6605.546

在控制协变量后，平均处理效应的估计值上升为 699.13，但与上文实验结果相比，依然大幅低估，而且不显著(p 值为 0.257)。

下面使用倾向得分匹配。首先，使用默认的一对一匹配，估计处理组平均处理效应(ATT；Stata 记为 ATET)。

```
. teffects psmatch (re78) (treat $cov), atet
```

```
Treatment-effects estimation                    Number of obs      =     16,177
Estimator      : propensity-score matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Treatment model: logit                                      max =         43
------------------------------------------------------------------------------
             |               AI robust
        re78 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
       treat |
   (1 vs 0)  |   1793.966   751.8995     2.39   0.017     320.2699    3267.662
------------------------------------------------------------------------------
```

ATT 估计值为 1793.97，与上文实验数据的一元回归结果很接近。

`AI robust std. err.`为根据 Abadie and Imbens (2016)计算的稳健标准误(等于 751.90)。

此稳健标准误已考虑了倾向得分为估计所得而导致的不确定性。

根据 AI 稳健标准误计算的t值为 2.39，相应的 p 值为 0.017，故在 5%的水平上显著。

一对一匹配的方差可能较大，未必最有效率。尝试进行一对四匹配：

```
. teffects psmatch (re78) (treat $cov),atet nn(4) gen(match)
```

其中，选择项 gen(match)表示生成变量 *match1*、*match2* 等，表示第 1 个匹配结果、第 2 个匹配结果，以此类推。

```
Treatment-effects estimation                    Number of obs      =     16,177
Estimator      : propensity-score matching      Matches: requested =          4
Outcome model  : matching                                   min =          4
Treatment model: logit                                      max =         43
------------------------------------------------------------------------------
             |               AI robust
        re78 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
       treat |
  (1 vs 0)   |   1686.293   663.3691     2.54   0.011     386.1139    2986.473
------------------------------------------------------------------------------
```

ATT 估计值为 1686.29，很接近于上文实验数据的多元回归结果。AI 稳健标准误降低为 663.37，而 *p* 值也相应地减少为 0.011，几乎在 1%的水平上显著。

顺便考察前两位处理组个体的匹配结果：

```
. list match* in 1/2
```

1.	match1 8254	match2 9820	match3 1477	match4 12110	match5 12363	match6 .	match7 .	match8 .
	match9 .	match10 .	match11 .	match12 .	match13 .	match14 .	match15 .	
	match16 .	match17 .	match18 .	match19 .	match20 .	match21 .	match22 .	
	match23 .	match24 .	match25 .	match26 .	match27 .	match28 .	match29 .	
	match30 .	match31 .	match32 .	match33 .	match34 .	match35 .	match36 .	
	match37 .	match38 .	match39 .	match40 .	match41 .	match42 .	match43 .	

2.

match1 8247	match2 1982	match3 5501	match4 9411	match5 .	match6 .	match7 .	match8 .
match9 .	match10 .	match11 .	match12 .	match13 .	match14 .	match15 .	
match16 .	match17 .	match18 .	match19 .	match20 .	match21 .	match22 .	
match23 .	match24 .	match25 .	match26 .	match27 .	match28 .	match29 .	
match30 .	match31 .	match32 .	match33 .	match34 .	match35 .	match36 .	
match37 .	match38 .	match39 .	match40 .	match41 .	match42 .	match43 .	

(处理组的)第 1 位个体分别与(控制组的)第 8254、9820、1477、12110 与 12363 位个体相匹配；以此类推。由于存在并列的倾向得分，故第 1 位个体与五位控制组个体相匹配。

进行卡尺内一对四匹配。为此，首先计算倾向得分。

```
. predict pscore,ps tlevel(1)
```

其中，选择项“ps”表示 propensity score。

Stata 的倾向得分既可表示进入处理组的概率，也可表示进入控制组的概率；而默认为进入处理水平(treatment level)最小的那个组的概率(允许存在多个处理水平)，在此例对应于选择项“tlevel(0)”(即进入控制组的概率)。

此命令的选择项“tlevel(1)”指定计算进入处理组的倾向得分，并将所得结果记为变量 *pscore*。

然后，计算倾向得分的标准差，再乘以 0.25。

```
. sum pscore
```

Variable	Obs	Mean	Std. dev.	Min	Max
pscore	16,177	.011436	.0540848	3.77e-06	.4883897

```
. dis 0.25*r(sd)
.0135212
```

结果显示，$0.25\,\hat{\sigma}_{pscore} \approx 0.0135$，故下面使用选择项“caliper(0.0135)”将卡尺范围定为 0.0135，这意味着对倾向得分相差不超过 0.0135 的观测值进行一对四匹配。

```
. teffects psmatch (re78) (treat $cov),atet nn(4) caliper(0.0135) osample(outside)
```

```
  179 observations have fewer than 4
propensity-score matches within
  caliper .0135; they are identified in the
osample() variable
  r(459);
```

选择项“osample(outside)”表示生成变量 *outside*，以记录违反重叠假定的个体，比如该个体无法在卡尺内找到四个邻居。

共有 179 个观测值无法在 0.0135 的卡尺内找到四个邻居；故此命令报错，并退出运行。

仅使用 *outside* 取值为 0(即匹配结果均在卡尺内)的样本进行 PSM 估计。

```
. teffects psmatch (re78) (treat $cov) if outside==0,atet nn(4)
```

```
Treatment-effects estimation                   Number of obs      =     15,998
Estimator      : propensity-score matching     Matches: requested =          4
Outcome model  : matching                                   min =          4
Treatment model: logit                                      max =          9
```

		AI robust				
re78	Coefficient	std. err.	z	P>|z|	[95% conf.	interval]
ATET						
treat (1 vs 0)	1488.198	702.4305	2.12	0.034	111.4593	2864.936

ATT 的点估计下降为 1488.20，更加远离实验数据的结果，但依然在 5%的水平上显著。

这或许是由于样本容量降为 15,998(原为 16,177)，损失了 179 个观测值，故更不易找到高质量的匹配。

仍回到一对四匹配，不再限制在卡尺内匹配。然后，使用命令 teoverlap 检验重叠假设；其中“te”表示 treatment effects。

```
. qui teffects psmatch (re78) (treat $cov), atet nn(4)
. teoverlap,ptlevel(1) name(overlap,replace) xtitle(propensity score)
```

选择项“ptlevel(1)”指定倾向得分为进入处理组(即处理水平为 1)的概率；默认为“ptlevel(0)”(将倾向得分定义为进入控制组的概率)。

选择项“xtitle(propensity score)”指定横轴的标题。

选择项“name(overlap,replace)”将画图结果命名为overlap，并覆盖内存中可能的同名文件。

将画图结果命名后，可在 Stata 中同时显示多张图，不会覆盖之前的画图结果。

结果参见图 15.3。

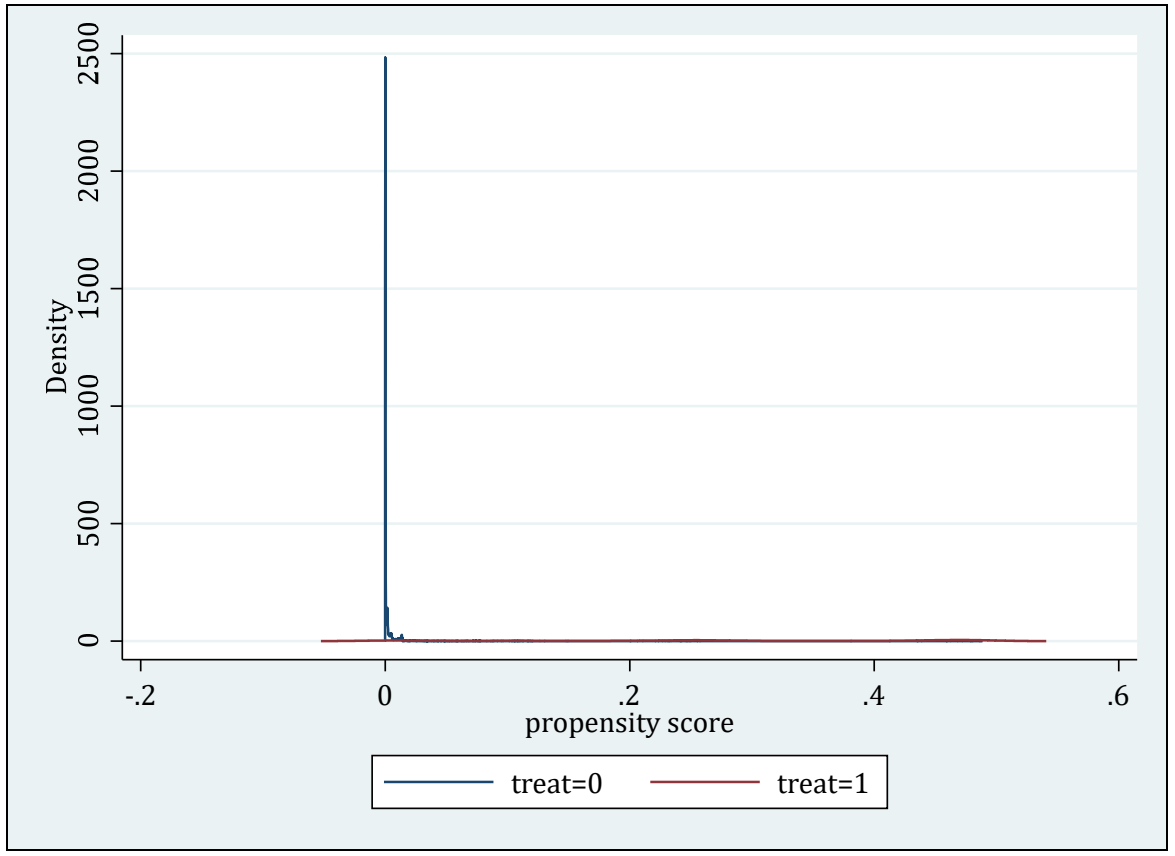

图 15.3 重叠假设的检验

从图 15.3 可见，倾向得分的分布(概率密度)在处理组与控制组之间似乎有很大差异。

为了进一步考察，分别画控制组与处理组的倾向得分直方图。

```
. hist pscore if treat == 0,fraction name(pscore_control,replace) xtitle(propensity score for the control group)

. hist pscore if treat == 1,fraction name(pscore_treat,replace) xtitle(propensity score for the treatment group)
```

选择项“fraction”指定纵轴为频率，结果参见图 15.4 与图 15.5。

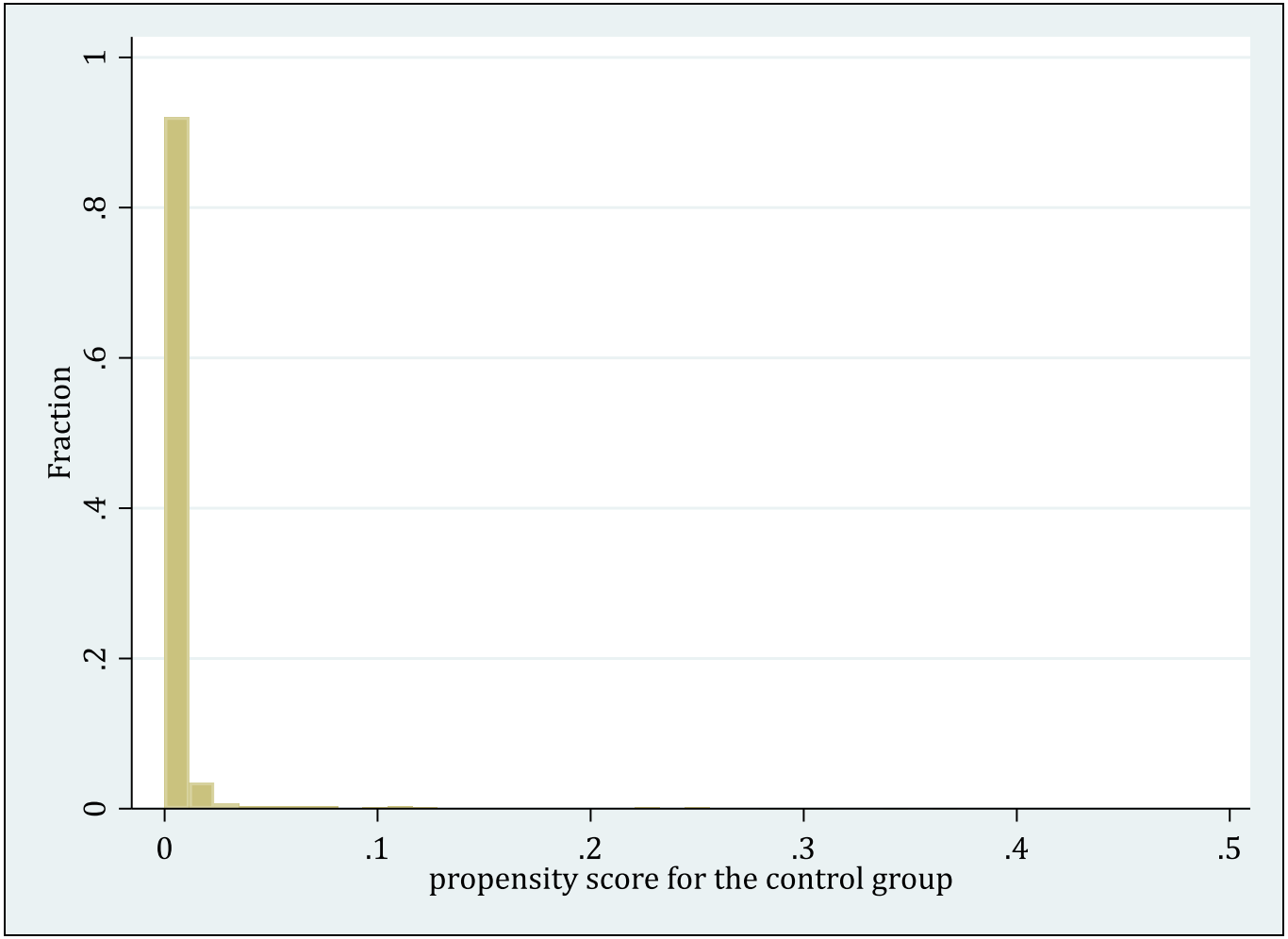

图 15.4 控制组倾向得分的直方图

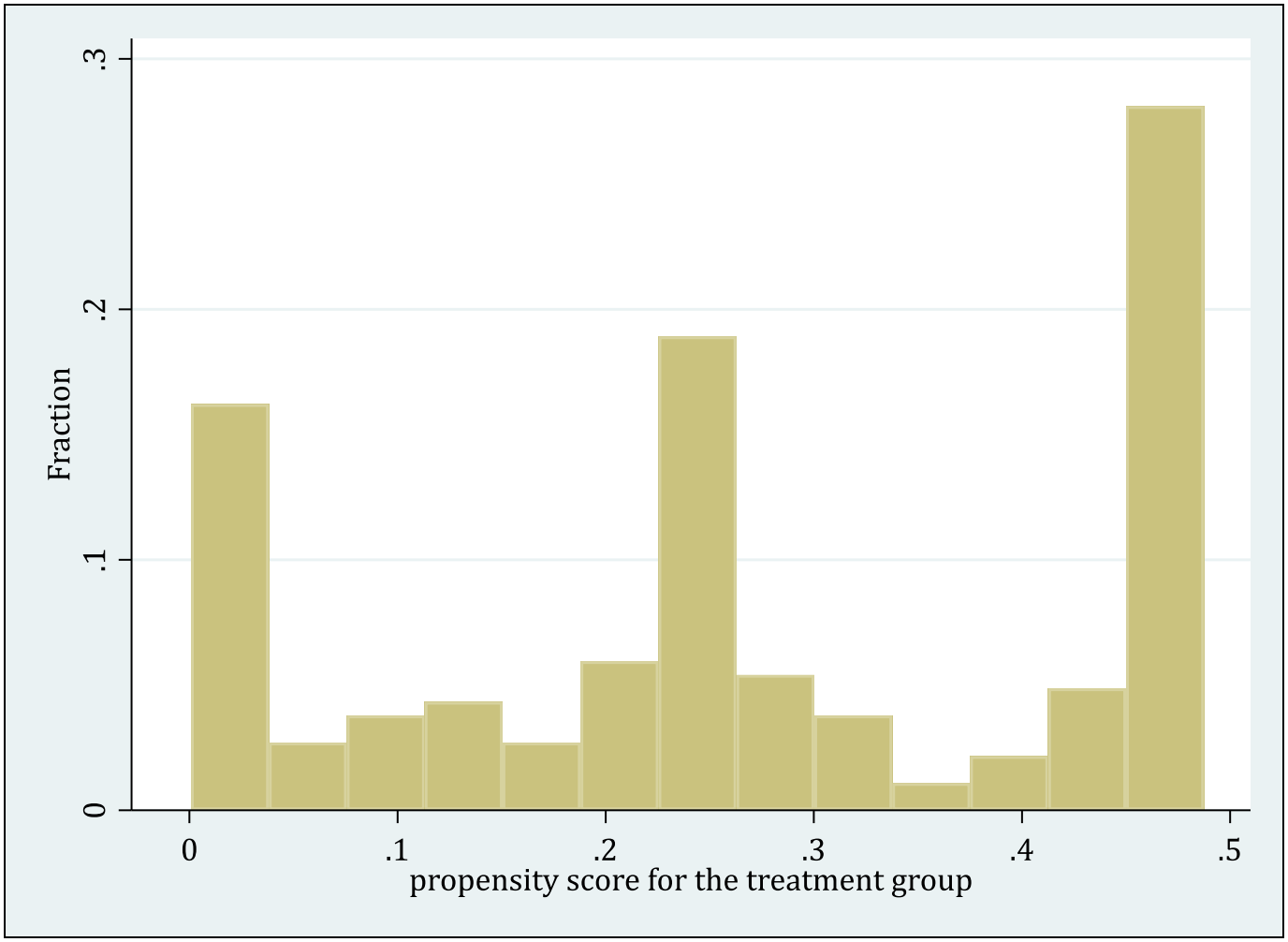

图 15.5 处理组倾向得分的直方图

控制组的倾向得分大量集中在 0 附近。

处理组的倾向得分则主要分布在 0 与 0.5 之间。

但处理组与控制组依然存在一定的共同支撑(common support)。

对于倾向得分接近于 0 的处理组个体而言，仍可匹配到倾向得分接近的控制组个体。

下面进行协变量的平衡性检验。重新运行命令(但不汇报结果)，然后使用 tebalance 的系列命令。

```
. qui teffects psmatch (re78) (treat $cov),atet nn(4)
```

```
. tebalance summarize
```

Covariate balance summary

	Raw	Matched
Number of obs =	16,177	370
Treated obs =	185	185
Control obs =	15,992	185

	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
age	-.7961833	.1872456	.4196365	.5026332
education	-.6785021	.0190581	.4905163	.6212854
black	2.427747	.0110627	1.950622	.979499
hispanic	-.0506973	.093457	.8410938	1.485232
married	-1.232648	.0738029	.7516725	1.135145
nodegree	.9038111	.054493	.9975255	.9546008
re74	-1.56899	.0361495	.2607425	1.893972
re75	-1.746428	.0682046	.1205903	1.603915

与匹配前(Raw)的结果相比，匹配后(Matched)大多数变量的标准化偏差均大幅缩小。

除变量 *age* 外，其余协变量的匹配后标准化偏差均小于 10%。

可考虑将变量 *age* 的高次项引入估计倾向得分的 logit 回归中(参见习题)。

从方差比(variance ratio)来看，匹配后大多数变量的方差比更接近于 1，但有些变量依然有一定差距。

命令 tebalance 还可通过箱形图(box plot)与密度图(density plot)来对比倾向得分以及协变量在处理组与控制组的分布差异。

所谓**箱形图**，也称为**箱线图**(box and whisker plot)，是一种以简洁方式呈现变量分布特征的画图方法，参见图 15.6。

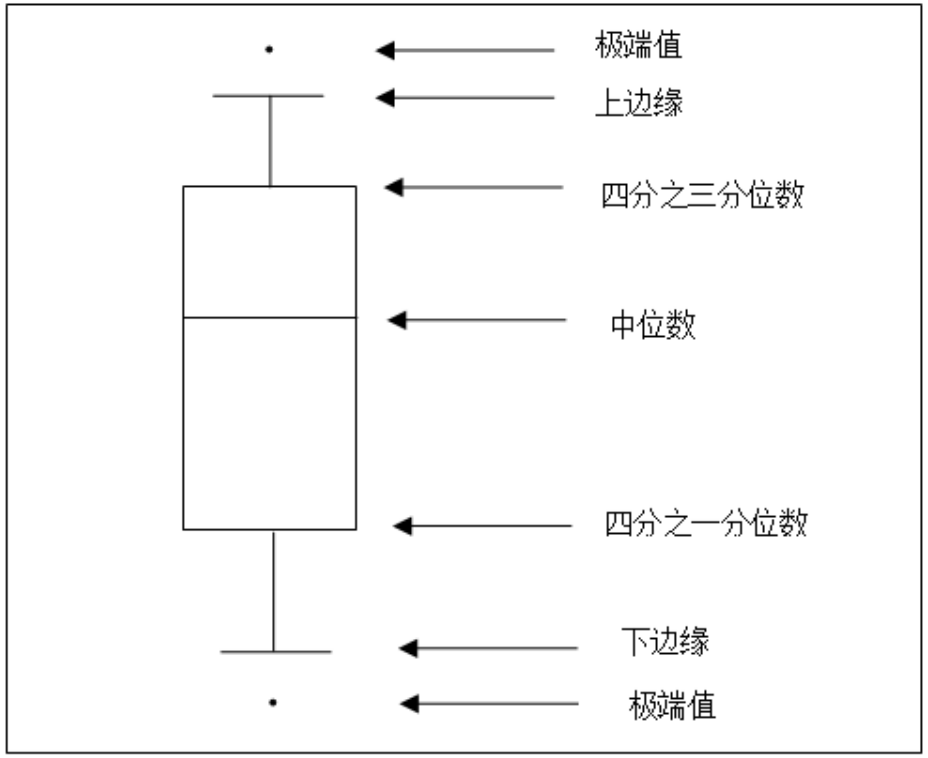

图 15.6 箱形图的示意图

箱体上端(upper hinge)为四分之三分位数(记为Q_3)，箱体下端(lower hinge)为四分之一分位数(记为Q_1)，而箱体中的横线表示中位数。

大于“上边缘” (upper adjacent value)与小于“下边缘”的观测值均为“极端值”或“异常值”(outlier)。

记$U = Q_3 + \frac{3}{2}(Q_3 - Q_1)$，其中$(Q_3 - Q_1)$称为“四分位矩”(interquartile range)；故U为四分之三分位数(Q_3)加上四分位矩$(Q_3 - Q_1)$的 1.5 倍。

假设画样本数据$\{x_1, x_2, \cdots, x_n\}$的箱形图，其中$x_1, x_2, \cdots, x_n$已按从小到大排列。

上边缘定义为观测值x_i，满足$x_i \le U$，而$x_{i+1} > U$。

下边缘也可以类似地定义。记$L = Q_1 - \frac{3}{2}(Q_3 - Q_1)$，则下边缘定义为观测值$x_j$，满足$x_j \ge L$，而$x_{j-1} < L$。

```
. tebalance box,name(box,replace)
```

由于此命令未指定任何协变量，故将画倾向得分的箱形图，结果参见图 15.7。

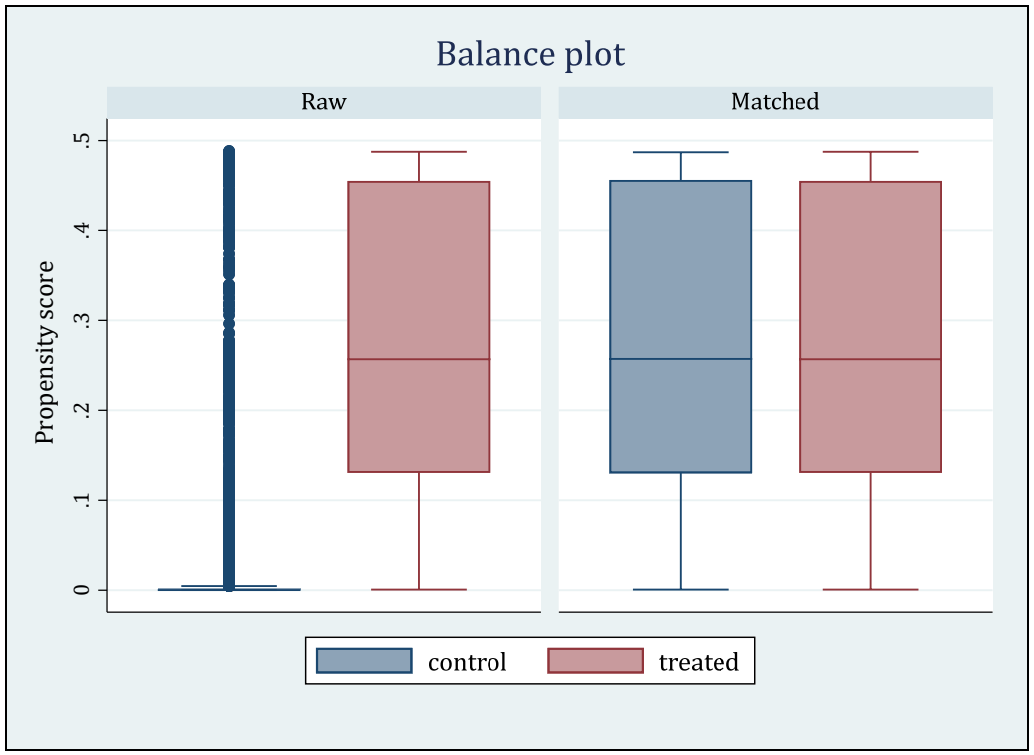

图 15.7 倾向得分的箱型图

在原始样本中，控制组的箱体很窄且靠近 0，且箱体之上则有不少极端值；而处理组的箱形图则比较“正常”。在匹配样本中，二者的箱形图则很相似，表明二者的分布接近。

下面画倾向得分的密度图，结果参见图 15.8。

```
. tebalance density,name(density,replace)
```

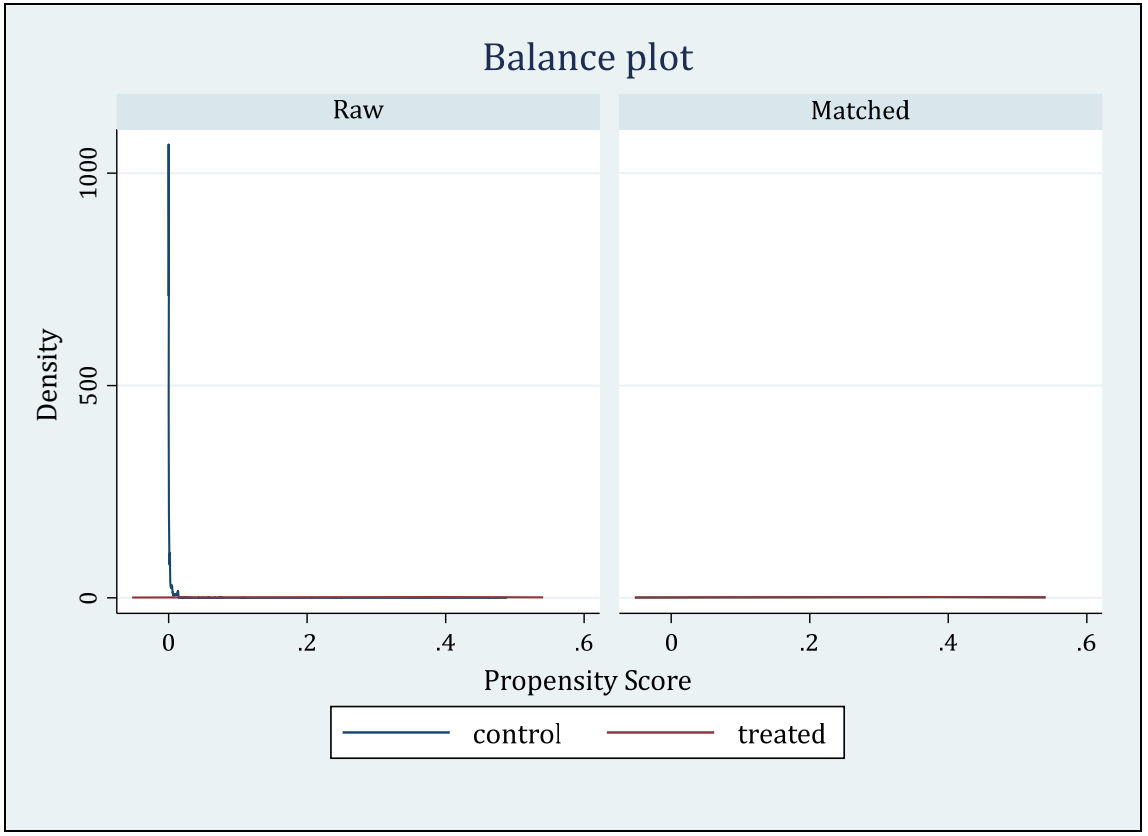

图 15.8 倾向得分的密度图

由于左图(原始样本)的比例尺缘故，不易看出右图(匹配样本)倾向得分的分布差异。

为此，我们分别手工画图 15.8 的左图与右图。

在画原始样本的左图时，由于控制组的倾向得分大量集中于 0 附近，导致此处的概率密度很大(超过 1000)，故使用两个纵轴以便看清控制组与处理组的分布。

```
. twoway kdensity pscore if treat == 0, lp(dash) xtitle(Propensity Score) yaxis(1) || kdensity pscore if treat == 1, yaxis(2) ytitle(Density) title(Raw) legend(label(1 Control) label(2 Treated))
```

选择项“yaxis(1)”与“yaxis(2)”分别指定第一与第二个纵轴。选择项“legend”用于设定图例，结果参见图 15.9。

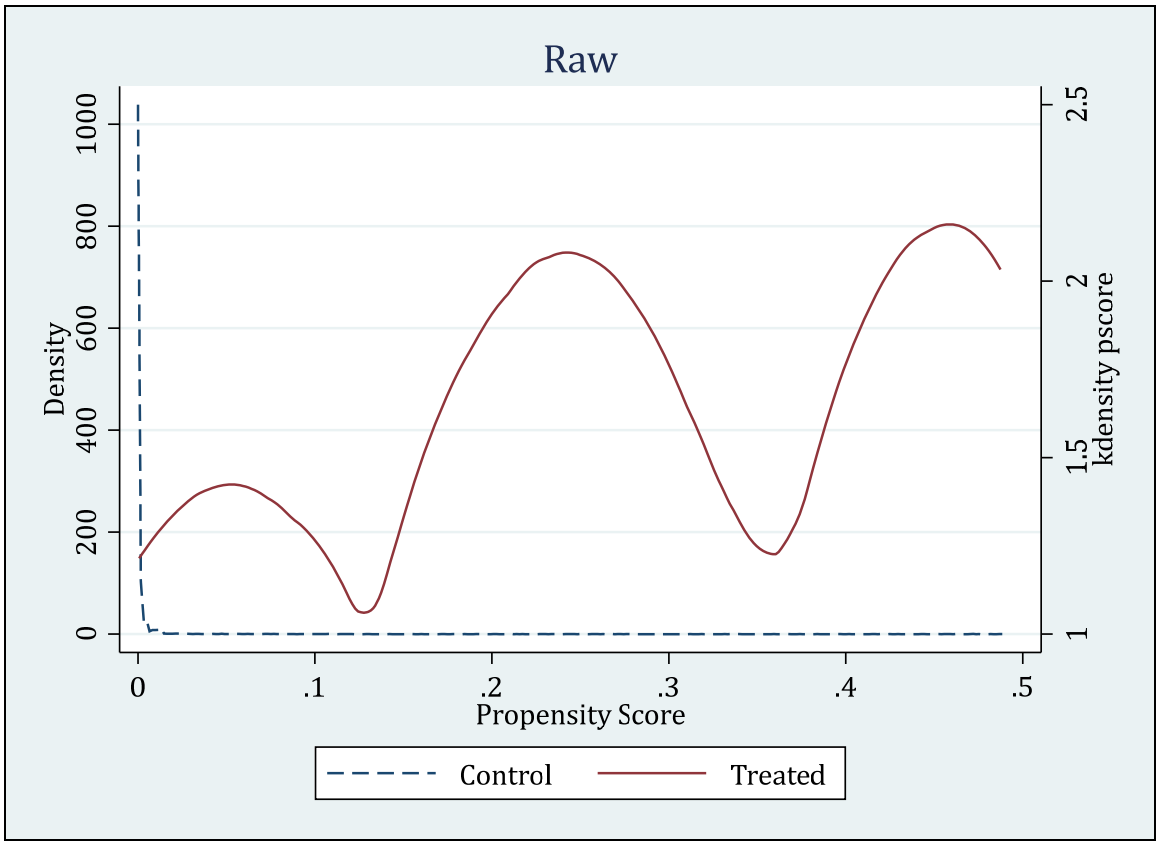

图 15.9 原始样本的倾向得分密度图

在原始样本中，控制组与处理组倾向得分的分布差别很大。

为了画匹配样本的密度图，首先根据最近邻的匹配结果*match1*，计算最近邻的倾向得分。

```
. gen pscore_match = pscore[match1]
```

其中，“pscore[match1]”表示变量 *pscore* 的第 *match1* 个观测值。例如，若 *match1* = 470，则“pscore[470]”表示变量 *pscore* 的第 470 个观测值。

此命令生成变量 *pscore_match*，用于记录最近邻的倾向得分。更具体地，考察处理组前 5 位个体的情况：

```
. list pscore pscore_match match1 in 1/5
```

	pscore	pscore~h	match1
1.	.4797528	.4797528	8254
2.	.2539605	.2538374	8247
3.	.2575564	.2575564	2780
4.	.2560519	.2561141	6231
5.	.3999529	.3996576	470

第 5 位个体的最近邻为第 470 位个体(因为 *match1* = 470)，而此最近邻的倾向得分为 0.3996576。验证第 470 位个体的倾向得分：

```
. list pscore in 470
```

	pscore
470.	.3996576

第 470 位个体的倾向得分正是 0.3996576。

下面画处理组倾向得分的密度图，以及与处理组匹配的控制组个体的倾向得分密度图，结果参见图 15.10。

```
. twoway kdensity pscore_match if treat==1,lp(dash) xtitle(Propensity Score) || kdensity pscore if treat==1 , ytitle(Density) title(Matched) legend(label(1 Control) label(2 Treated))
```

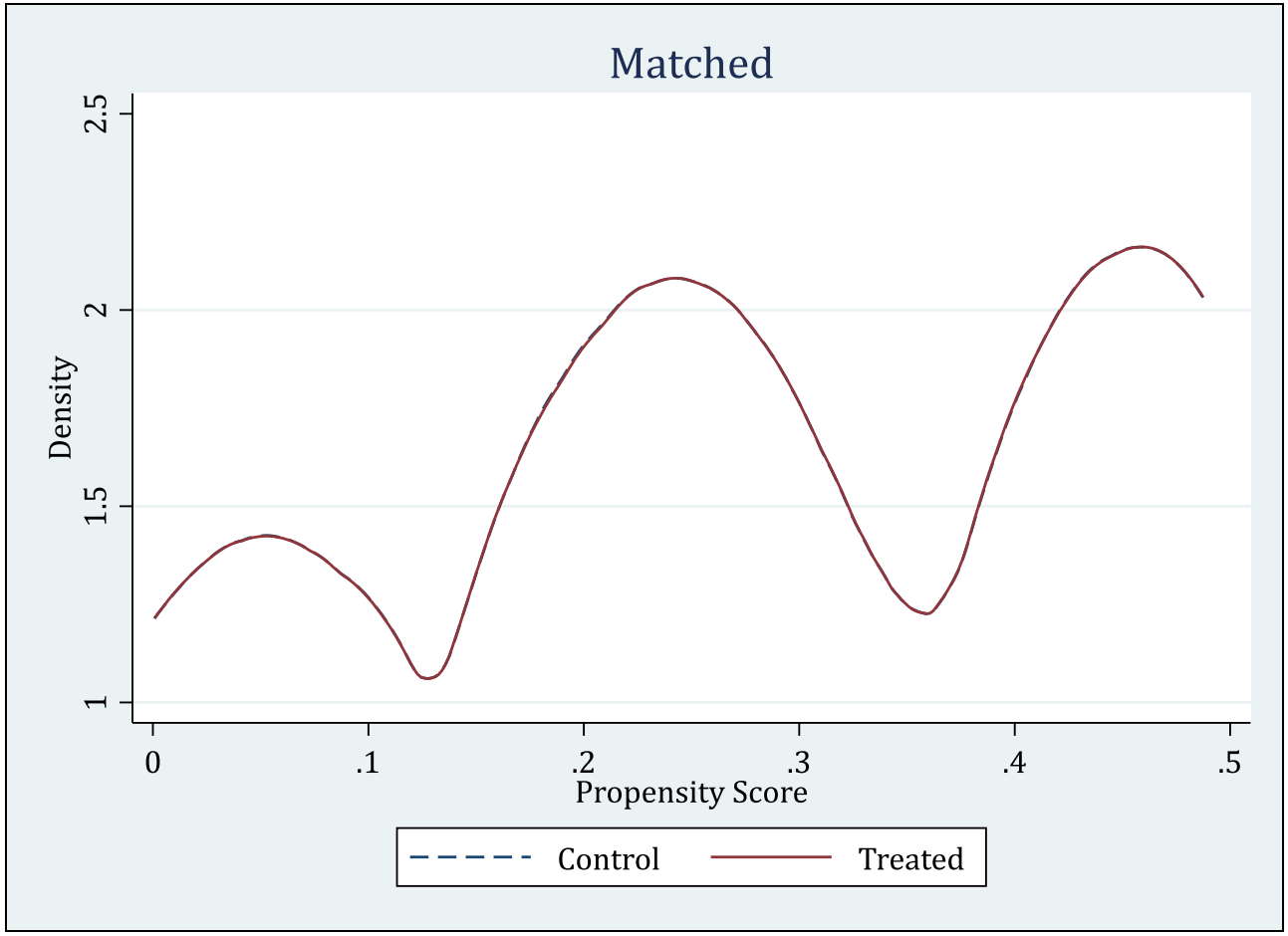

图 15.10 匹配样本的倾向得分密度图

匹配后，控制组与处理组倾向得分的分布非常接近，几乎重合。

直接考察二者的统计特征，也可得到进一步验证：

```
. sum pscore pscore_match if treat==1
```

Variable	Obs	Mean	Std. dev.	Min	Max
pscore	185	.2705936	.1681871	.0007016	.4874457
pscore_match	185	.2705289	.1681619	.0007012	.4874457

在匹配样本中，控制组与处理组倾向得分的均值、标准差、最小值与最大值均很接近。

以协变量*education*为例，画该变量的箱形图，结果参见图15.11。

```
. tebalance box education,
name(box_educ,replace)
```

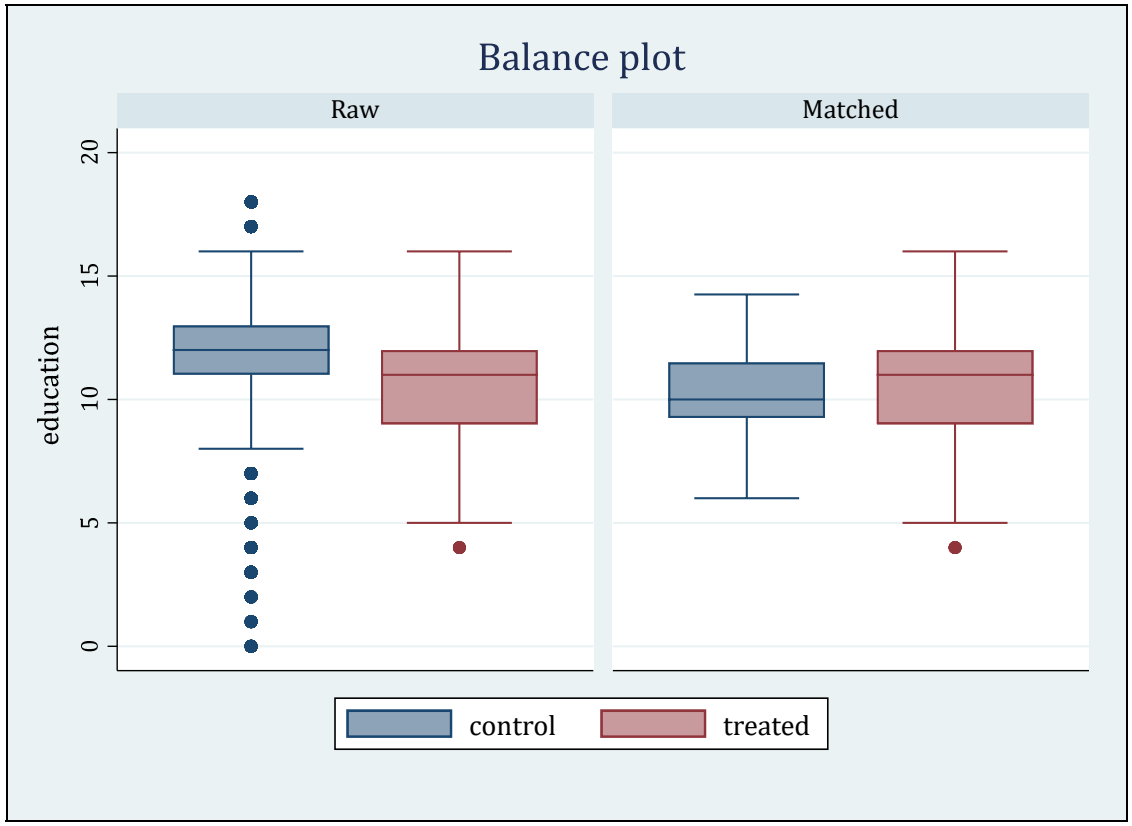

图 15.11 变量 *education* 的箱型图

匹配之后变量 *education* 的分布特征似乎在控制组与处理组之间更为接近了。

画协变量 *education* 的密度图，结果参见图 15.12。

```
. tebalance density education, name(density_educ,replace)
```

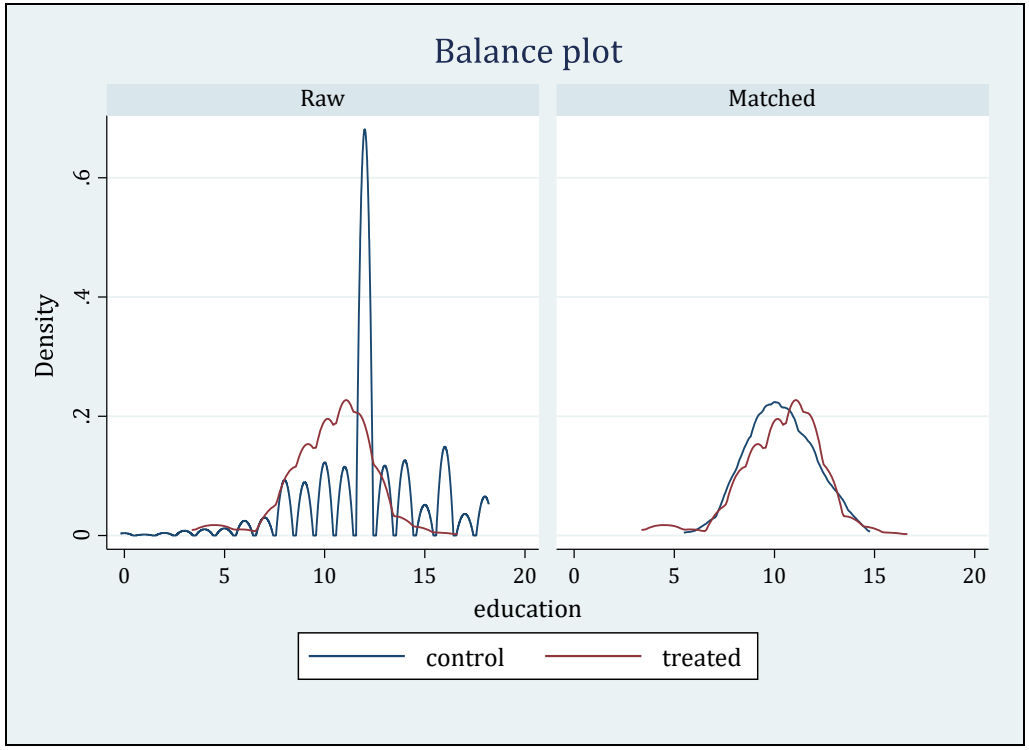

图 15.12 变量 *education* 的密度图

在原始数据中(左图)，变量 *education* 在控制组的分布有些不规则，与其在处理组的分布差别很大。

在匹配样本中(右图)，二者的分布已经比较接近。

在进行倾向得分匹配之后，是否应使用匹配样本(matched sample)再进行线性回归？

一般不建议这样做。

首先，PSM 相当于非参数方法(nonparametric method)，不依赖于函数形式，更为稳健；若使用匹配样本进行线性回归，则又回到线性假设，可谓“画蛇添足”。

其次，先匹配再回归的做法相当于二阶段估计(two-stage estimation)，一般无法直接使用第二阶段回归的标准误(因为未考虑第一阶段估计的误差)。

在进行 PSM 之后，一般不宜再进行回归。

当然，也有例外。

比如，适用于面板数据的**倾向得分匹配双重差分法**(Propensity Score Matching Difference in Difference，简记 PSM-DID)，就是先做 PSM，再使用双重差分法。

15.8 倾向得分匹配的局限性

通过使用反事实的因果推断框架，PSM 可能在相当程度上减少观测数据的偏差，但它本身仍有一定的局限性。

(1) PSM 通常要求比较大的样本容量以得到高质量的匹配。

(2) PSM 要求处理组与控制组的倾向得分有较大的共同支撑(common support)；否则，将丢失较多观测值，导致剩下的样本不具有代表性。

(3) 在估计倾向得分时，一般仍假设具体的函数形式(比如logit)，所得结果可能依赖于此函数形式，存在“模型依赖性”(model dependence)。

(4) 倾向得分相近的个体，其协变量未必接近。这可能导致男性与女性匹配，中学毕业生与大学毕业生匹配，从而导致偏差。

(5) PSM 所依赖的非混杂性假定较强。由于 PSM 只控制了可测变量的影响，如果存在依不可测变量选择(selection on unobservables)，则仍会带来“隐性偏差”(hidden bias)。