王岳龙：如何正确理解和运用matching方法

刘西川阅读写作课 4天前

本文讨论的论文来源：
Guido.W.Imbens，“Matching method in Practice Three Examles”，The Journal of Human Resource 2015

一、 引言

目前DID、RD、IV等因果识别方法在国内外应用十分广泛，我本人也推送了大量上述方法的应用，但是对Matching却少有介绍，该方法在前几年在国内公司金融、财务会计领域研究中用的特别多，但是在实际应用中却往往一知半解，存在了诸如matching可以解决内生性问题的误解，以及选择匹配的协变量和函数形式过于随意的问题，因此本文结合微观计量大神Imbens 2015年在JHR发表的一个论文，详细介绍该方法的原理和应用。

二、 Matching的基本原理

1. Matching的本质

国内很多文章都说matching能解决内生性问题，一直没有查到该说法的最原始出处，但是被国人以讹传讹流传到现在，所以在这里必须强调下matching和OLS一样都属于selection on observables的方法，识别假设都是是unconfoudness或者说是conditional independent assumption，（y0,y1），此外matching还多了个overlap（共同区间假设）。所谓CIA假设，就是控制住所有可观测因素x后，未观测变量不会对两组观测结果y产生系统性差异，也就是内生性问题。因为内生性问题意味着未观测因素是可能的confounding factor，既影响个体选择D，又影响结果变量y，即便控制可观测因素，选择性偏差仍然存在，因此我们说matching本质不能解决内生性问题。

2. Matching和OLS的比较

$$\tau_{\text{OLS}} = \sum_x \left(\frac{p(x)(1-p(x))\Pr[X_i = x]}{\sum_x [p(x)(1-p(x))]\Pr[X_i = x]}\right)\tau_x$$

$$\tau_{\text{ATT}} = \sum_x \left(\frac{p(x)\Pr[X_i = x]}{\sum_x p(x)\Pr[X_i = x]}\right)\tau_x$$

其中 p(x)=Pr(D=1|x),即倾向得分值

两个估计量形式十分类似，因此OLS是一种特殊的匹配方法。OLS是以条件方差p(x)(1-p(x))为权重进行加权，当p(x)=0.5时权重最大，因此它对层内两组个体数目相同的层赋予更大的权重。匹配估计是以倾向得分指数p(x)进行加权，对层内处理组个体更多的层赋予更大的权重。两个估计参数一般是不同的，只有当p(x)和τx是常数时，两者才相等。

3. OLS的不足和Matching的优点

既然matching和OLS本质是一样的，但是为什么还要在有了最简单实用的OLS基础上，再弄出一个matching方法，这两个方法各自有什么优缺点，这是一个值得关注的问题。由上式可知，OLS估计时对于倾向得分指数为0或者1的层，将赋予0权重，回归时自动丢弃那些仅有参考组或者处理组一组个体的层，因此得到的估计系数不能解释为总体的ATE。协变量匹配的好处在于匹配时，由于需要检验共同区间的要求是否满足，从而很清楚哪些样本进入匹配，从而知道是对哪部分子样本的估计，这是匹配方法相对于回归方法的优势之一。其次为了避免extrapolation和misspecification，使得OLS估计因果效应具有一致性，要求条件期望函数必须是线性，同时处理组和控制组的控制变量具有相同的分布。因此当控制变量组间差异非常大时，且数据对函数形式设定也非常敏感，OLS往往不能得到稳健的估计结果，而这些问题对matching并不存在。imbens在论文中引入一个参加NSW项目培训对收入影响的例子，通过线性和对数线性2个不同模型设定形式，显示了差异比较明显的ATT结果。

4. Matching的分类

Matching按照其寻找反事实的方法，可以分为协变量匹配和倾向得分匹配（PSM）两大类。协变量匹配是基于各变量空间距离（马氏、欧式距离），但是当样本小或者变量多时，通常难以找到与之匹配的个体。这时可以考虑采用倾向得分匹配，通过估计处理变量的影响因素方程，把多维控制变量降成一维的倾向得分值。King and Nielsen（2016）指出，倾向得分匹配将许多协变量综合成倾向得分值，由此可能导致倾向得分值接近而协变量差异更大，建议采用协变量匹配。由于协变量匹配是以分组随机实验为基础，由于考虑了协变量差异，可以保证匹配样本的平衡性。此外还有逆概率加权匹配、回归调整、核匹配方法，stata13新出的teffects命令可以很方便实

现上述方法。

三、 Matching的规范操作步骤

1. 设计阶段

首先计算normalized differenceΔ

$$(12)\quad \Delta_{X,k}=\frac{\bar{X}_{t,k}-\bar{X}_{c,k}}{\sqrt{(S^2_{X,t,k}+S^2_{X,t,k})/2}}$$

where

$$\bar{X}_{c,k}=\frac{1}{N_c}\sum_{i:W_i=0}X_{i,k},\ \bar{X}_{t,k}=\frac{1}{N_t}\sum_{i:W_i=1}X_{i,k},$$

$$S^2_{X,c,k}=\frac{1}{N_c-1}\sum_{i:W_i=0}(X_{i,k}-\bar{X}_{c,k})^2,\ \text{and}\ S^2_{X,t,k}=\frac{1}{N_t-1}\sum_{i:W_i=1}(X_{i,k}-\bar{X}_{t,k})^2.$$

Δ越接近0，样本越平衡，属于样本内特征，和样本大小无关。

$$t_{X,k}=\frac{\bar{X}_{t,k}-\bar{X}_{c,k}}{\sqrt{S^2_{X,t,k}/N_t+S^2_{X,c,k}/N_c}}.$$

Δ越接近0，样本越平衡，属于样本内特征，和样本大小无关。

Δ和样本均值差异t检验计算式很像，由于上式t检验与样本容量有关，当样本很大时，即使协变量平衡也可能出现显著差异结果，所以imbens不推荐采用。

如果normalized differenceΔ差异很大，需要考虑删减样本，以更好满足overlap假设。根据Crump etal(2009),保留α<e(x)<1-α的样本，其中

$$\frac{1}{\alpha\cdot(1-\alpha)}=2\cdot\mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))}\middle|\frac{1}{e(X)\cdot(1-e(X))}\le\frac{1}{\alpha\cdot(1-\alpha)}\right].$$

通常 α=0.1

2. 补充分析：评价CIA假设

$$H_0:\ E\{E[y_i^p\mid X_i,D_i=1]-E[y_i^p\mid X_i,D_i=0]\}=0$$

由于现实中只能观测到一个潜在结果，因此CIA假设本身无法直接被检验。Imbens(2004)和Imbens and Rubin(2015)提出了两种间接检验方法，类似DID里面的falsification test，并不是充分条件，不拒绝原假设也不能保证CIA一定成立，只是提供了暂时没有发现CIA不成立的证据，使我们有很大信心相信估计结果。

利用一个事实上没有受到干预影响的变量作为伪结果，用它作为潜在结果y0的代理变量，一般选择滞后的结果变量作为伪结果变量yp。如果控制X，影响潜在结果的所有混淆因素均被控制，未观测因素将不会对两组结果产生系统性影响，伪结果估计的ATE将为0。反之如果ATE显著不为0，说明仍然存在着未观测的混淆因素没有被控制，从而CIA假设不成立。

3. **分析阶段：确定合理的函数形式**

以PSM为例，选择哪些变量和变量形式进入logit模型估计是个十分重要的问题。变量和变量形式不同，算出来的倾向得分值自然就不同，倾向得分值不同匹配的个体就不同，从而得到不同的ATT，所以确定logit模型的形式很关键，但是这个问题在国内以往研究就忽略掉了，选择模型过于主观随意，没有经过严格检验，为此imbens提出了下列stepwise回归检验方法供参考：

（1）先根据相关理论和经验确定必须要加入的preslect变量（如果没有相关先验信息，则只放常数项）

（2）逐步加入其他变量一次项，对新增变量进行联合显著性检验的likelihood ratio test，此时临界值是Clin=1

（3）逐步加入二次项（平方项和交互项），同样对新增变量进行联合显著性检验的likelihood ratio test，此时临界值是Cqua=2.71

四、一个简单的例子

Imbens(2015，jhr)的论文里面原本提供了三个例子，这里选取第一个是否中彩票对日后劳动收入影响的例子进行重点讲解。

Table 1
Summary Statistics Lottery Data

	Losers (N=259)		Winners (N=237)			
Covariate	Mean	(Standard Deviation)	Mean	(Standard Deviation)	t-stat	nor-diff
Year won	6.38	(1.04)	6.06	(1.29)	−3.0	−0.27
Number of tickets	2.19	(1.77)	4.57	(3.28)	9.9	0.90
Age	53.2	(12.9)	47.0	(13.8)	−5.2	−0.47
Male	0.67	(0.47)	0.58	(0.49)	−2.1	−0.19
Education	14.4	(2.0)	13.0	(2.2)	−7.8	−0.70
Working then	0.77	(0.42)	0.80	(0.40)	0.9	0.08
Earn year−6	15.6	(14.5)	12.0	(11.8)	−3.1	−0.27
Earn year−5	16.0	(15.0)	12.1	(12.0)	−3.2	−0.28
Earn year−4	16.2	(15.4)	12.0	(12.1)	−3.4	−0.30
Earn year−3	16.6	(16.3)	12.8	(12.7)	−2.9	−0.26
Earn year−2	17.6	(16.9)	13.5	(13.0)	−3.1	−0.27
Earn year−1	18.0	(17.2)	14.5	(13.6)	−2.5	−0.23
Pos earn year−6	0.69	(0.46)	0.70	(0.46)	0.3	0.03
Pos earn year−5	0.68	(0.47)	0.74	(0.44)	1.6	0.14
Pos earn year−4	0.69	(0.46)	0.73	(0.44)	1.1	0.10
Pos earn year−3	0.68	(0.47)	0.73	(0.44)	1.4	0.13
Pos earn year−2	0.68	(0.47)	0.74	(0.44)	1.6	0.15
Pos earn year−1	0.69	(0.46)	0.74	(0.44)	1.2	0.10

首先对中彩票和没有中彩票的两组人进行了分组的统计描述，虽然理论上我们认为中彩票似乎是个随机事件，但是从统计分析结果来看，情况不完全是这么回事情，从均值差异检验来看，中彩票的人平均看来每周买彩票的次数更多、年纪更小、更多是女性、教育程度更低、之前6年的收入也越低。此外Δ也比较大，说明如果直接用原始数据匹配，overlap不大容易满足，匹配效果可能不会太好。为了比较，作者还是先直接用原始数据进行了匹配。

Table 2
Estimated Parameters of Propensity Score for the Lottery Data

Variable	Estimated	(Standard Error)
Intercept	30.24	(0.13)
Preselected linear terms		
Tickets bought	0.56	(0.38)
Education	0.87	(0.62)
Working then	1.71	(0.55)
Earnings year–1	–0.37	(0.09)
Additional linear terms		
Age	–0.27	(0.08)
Year won	–6.93	(1.41)
Pos earnings year–5	0.83	(0.36)
Male	–4.01	(1.71)
Second-order terms		
Year won × year won	0.50	(0.11)
Earnings year–1 × male	0.06	(0.02)
Tickets bought × tickets bought	–0.05	(0.02)
Tickets bought × working then	–0.33	(0.13)
Education × education	–0.07	(0.02)
Education × earnings year–1	0.01	(0.00)
Tickets bought × education	0.05	(0.02)
Earnings year–1 × age	0.002	(0.001)
Age × age	0.002	(0.001)
Year won × male	0.44	(0.25)

作者在估计倾向得分方程时，先将过去一周买票次数、教育程度、是否有工作、去年一年收入作为preselected变量放入模型，然后经过上述逐步回归检验，确定了上述变量进入了最终baseline模型。

Table 3
Alternative Specifications of the Propensity Score

	Baseline	No Preselected	$C_{lin} = 0, C_{qua} = -\infty$	Lasso
Degrees of freedom	18	18	18	12
Log likelihood function	–201.5	–201.5	–231.7	–229.1
Correlations of log odds Ratios				
Baseline	1.00	1.00	0.86	0.86
.No preselected	1.00	1.00	0.86	0.86
$C_{lin} = 0, C_{qua} = -\infty$	0.86	0.86	1.00	0.98
Lasso	0.86	0.86	0.98	1.00

在上述模型基础上，imbens又进行了其他几个备选模型检验，表三第二列是没有任何preselected变量，完全依靠逐步回归检验情形。第三列是只放一次项情形，最后一列是根据lasso回归情形。

Table 4
Sample Sizes for Subsamples with the Propensity Score between α and 1 − α (α = 0.0891) by Treatment Status

	Low $e(x) < \alpha$	Middle $\alpha \le e(X) \le 1-\alpha$	High $1-\alpha < e(X)$	All
Losers	82	172	5	259
Winners	4	151	82	237
All	86	323	87	496

通过将α设置为0.089，将参考组中倾向得分值低于0.089和高于0.911的分别82个和5个样本，作为离群值删去，同理处理组中向得分值低于0.089和高于0.911的分别4个和82个样本也作为离群值删去，最后得到包括151个处理组和172个参考组在内的323个样本。

Table 5
Summary Statistics Trimmed Lottery Data

	Losers (*N*=172)		Winners (*N*=151)			
Covariate	Mean	(Standard Deviation)	Mean	(Standard Deviation)	*t*-stat	nor-dif
Year won	6.40	(1.12)	6.32	(1.18)	−0.6	−0.06
Number of tickets	2.40	(1.88)	3.67	(2.95)	4.6	0.51
Age	51.5	(13.4)	50.4	(13.1)	−0.7	−0.08
Male	0.65	(0.48)	0.60	(0.49)	−1.0	−0.11
Education	14.0	(1.9)	13.0	(2.2)	−4.2	−0.47
Work then	0.79	(0.41)	0.80	(0.40)	0.2	0.03
Earn year−6	15.5	(14.0)	13.0	(12.4)	−1.7	−0.19
Earn year−5	16.0	(14.4)	13.3	(12.7)	−1.8	−0.20
Earn year−4	16.4	(14.9)	13.4	(12.7)	−2.0	−0.22
Earn year−3	16.8	(15.6)	14.3	(13.3)	−1.6	−0.18
Earn year−2	17.8	(16.4)	14.7	(13.8)	−1.8	−0.20
Earn year−1	18.4	(16.6)	15.4	(14.4)	−1.7	−0.19
Pos earn year−6	0.71	(0.46)	0.71	(0.46)	−0.0	−0.00
Pos earn year−5	0.70	(0.46)	0.74	(0.44)	0.9	0.10
Pos earn year−4	0.71	(0.46)	0.74	(0.44)	0.5	0.06
Pos earn year−3	0.70	(0.46)	0.72	(0.45)	0.2	0.03
Pos earn year−2	0.70	(0.46)	0.72	(0.45)	0.5	0.05

对经过trim处理后的323个样本又进行了统计描述，如上表5所示，各变量无论是组间均值差异t检验还是normalized difference都比之前小了许多，说明此时样本十分平衡，类似于通过了rct的变量平衡性测试，得到了一个类似随机实验环境，可以进行匹配。

Table 6
Estimated Parameters of Propensity Score for the Trimmed Lottery Data

Variable	Estimated	(Standard Error)
Intercept	21.77	(0.13)
Preselected linear terms		
Tickets bought	−0.08	(0.46)
Years of schooling	−0.45	(0.08)
Working then	3.32	(1.95)
Earnings year−1	−0.02	(0.01)
Additional linear terms		
Age	−0.05	(0.01)
Pos earnings year−5	1.27	(0.42)
Year won	−4.84	(1.53)
Earnings year−5	−0.04	(0.02)
Second-order terms		
Year won × year won	0.37	(0.12)
Tickets bought × year won	0.14	(0.06)
Tickets bought × tickets bought	−0.04	(0.02)
Working then × year won	−0.49	(0.30)

在该trim样本基础上，又重新估计了倾向得分方程，我们发现此时进入最终baseline模型的变量与之前有明显不同，主要是少了许多二次项。

Table 7
Assessing Unconfoundedness for the Lottery Data: Estimates of Average Treatment Effects for Pseudo Outcomes

	Blocking		Matching	
Pseudo Outcome	Estimated	(Standard Error)	Estimated	(Standard Error)
Y_{-1}	−0.53	(0.78)	−0.10	(0.95)
$(Y_{-1} + Y_{-2}) / 2$	−1.16	(0.83)	−0.88	(0.94)
$(Y_{-1} + Y_{-2} + Y_{-3}) / 3$	−0.39	(0.95)	−0.81	(0.98)

对trim后样本，接着又分别利用过去三年内的平均收入作为伪结果，对CIA假设进行了检验，结果发现无论是分组（blocking）匹配还是直接匹配，估计结果都不显著，从而暂时没有发现存在不可观测因素干扰。

Table 8

Lottery Data: Estimates of Average Treatment Effects (standard errors in parentheses)

	Full Sample		Trimmed			
Covariate	1 Block	Match	1 Block	2 Blocks	5 Blocks	Match
Number	−6.16	−4.03	−6.64	−6.05	−5.66	−4.53
	(1.34)	(1.32)	(1.66)	(1.87)	(1.99)	(1.36)
Few	−2.85	−4.29	−3.99	−5.57	−5.07	−4.19
	(0.99)	(1.31)	(1.16)	(1.30)	(1.46)	(1.36)
All	−5.08	−5.77	−5.34	−6.35	−5.74	−5.00
	(0.93)	(1.31)	(1.10)	(1.29)	(1.40)	(1.36)

Table 9

Lottery Data: Sensitivity of Estimates of Average Treatment Effects to Propensity Score Specification (standard errors in parentheses)

	Preferred Specification $C_L = 1, C_Q = 2.71$		Linear Specification $C_L = 0, C_Q = \infty$		Lasso	
Covariate	Blocked	Match	Blocked	Match	Blocked	Match
Number	−5.66	−4.53	−4.61	−3.62	−3.34	−3.50
	(1.99)	(1.36)	(2.08)	(1.56)	1.95	1.24
Few	−5.07	−4.19	−4.52	−2.39	−3.91	−2.97
	(1.46)	(1.36)	(1.46)	(1.56)	1.35	1.24
All	−5.74	−5.00	−5.21	−3.71	−3.64	−3.78
	(1.40)	(1.36)	1.60	1.56	1.31	1.24

上述两表估计结果表明，通过变换各种估计和匹配方法，都得到了显著为负的稳健性结果。上面这个例子，展现了一个标准的matching所需要的全部规范流程，对比国内之前研究，会发现我们之前研究是多么的不足。

五、小结

在过去的三年中，我陆续推送了包括IV、DID、RD在内许多推文，再加上今天这个matching的推文，基本上覆盖了目前微观应用计量经济学各种主要方法，下面我再对上述方法做个简单小结：

共同点：通过识别策略的设计，使得观测研究近似于随机实验。

区别：

OLS（匹配）：识别条件是CIA假设（条件均值独立假设），以可观测的控制变量为条件（selection>
IV:非依从的随机化实验，干预的随机化分配是工具变量，个体实际接受的干预状态是内生处理变量，利用干预分配的随机性可以识别出受工具变量影响的个体的因果效应（LATE）。识别条件是IV的排他性假设。

DID（FE）：增量上的分层随机实验，通过差分或者去均值的方法消除不随时间变化的未观测的混淆因素影响，识别条件是平行趋势假设。

RD：断点附近的随机实验，识别假设是局部随机化假设，即个体没有精确控制断点的能力，使得断点附近个体具有高度相似性。

RCT：完全随机化实验，个体是否接受干预完全随机，从而可观测因素和不可观测因素在组间是均值无差异的，识别假设是独立性假设（变量平衡性测试）

本文转载自微信公众号"计量经济学"。