Potential outcome model and regression

选择性偏误

为了说明选择性偏误问题，首先考虑一个经典的例子：去医院能使人变得更健康吗？为了让这个问题更加贴近实际，假设我们正在研究医院急诊室就诊的一群老年人，其中一部分人就诊后进一步被医院接收[1]。这种医疗方式浪费了医疗资源，并且人满为患的医院也可能使得相应的治疗不太有效（Grumbach, Keane and Bindman, 1993）。事实上，接触其他重病患者确实可能对这些老年人的健康而言有负面影响（如接触传染病患者）。

然而就诊后进一步被医院接收治疗的那些老年人也能够得到医生的专业服务，所以对医院是否能够让人变得更健康这一问题的回答似乎又应该是肯定的。但是数据层面是否支持这个说法？对于一个倾向于进行经验研究的人而言，自然而然地会考虑比较去过医院和没去医院的人在健康状况上的差异。全国健康采访调研（National Health Interview Survey，简记为NHIS）包含了相关的信息可以进行研究。具体而言，这个调研里包含这样一个问题，“在过去的12个月中，被访者是否曾因病在医院过夜？”，我们可以用这个问题来识别最近去过医院的人。全国健康采访调研还有一个问题是，“总体而言，你觉得你的健康水平是极好、非常好、好、一般还是差？”下面的表格给出了最近去过医院和没有去过医院的人的平均健康状况（对健康状况最差的人赋值1，对健康状况最好的人赋值5，数据来自2005年的NHIS）。

组别	样本大小	平均健康水平	标准差
去过医院	7774	3.21	0.014
没有去过医院	90049	3.93	0.003

两者之间的平均差距是0.72，没有去过医院的人健康状况更好，两者之差大且显著，其t统计值是58.9。
从表面上看，这个结果意味着去医院会使人的健康状况变差。由于医院往往充满了可能会使我们受到感染的各类重病患者，危险的医疗仪器和化学药剂也可能伤害到我们，所以去医院会使人健康状况变差未必不是正确答案。但是另一方面，我们也很容易解释为什么这个结果不能只从表面上看：去医院的人可能本身健康水平就比较差。更进一步讲，平均而言，即使在医院接受过治疗，那些到医院寻求治疗的人的健康水平可能还是不如没有去医院的人，也就是说，对于去医院的那些人而言，不去医院只能让他们的健康状况变得更差，但是去医院也未必能让这些人的健康水平赶上不去医院的人。
为了更精确地描述这个问题，我们将接受医院治疗描述为一个二值随机变量，$D_i = \{0, 1\}$。我们所考虑的研究对象的结果——对健康水平的度量，记为Y_i。我们的问题就是：Y_i是否受医疗的影响。为了回答这个问题，我们想象去了医院的人如果没有去医院将会发生什么，没有去医院的人如果去了医院将会发生什么。因此，对于任何个体而言，他们的健康状况都有两种潜在结果（Potential outcome）：

$$Potential\ outcome = \begin{cases} Y_{1i} & if\ D_i = 1 \\ Y_{0i} & if\ D_i = 0 \end{cases} \tag{1.1}$$

也就是说，假设一个人没有去医院，他的健康状况将是Y_{0i}（指没有去医院情况下的潜在结果），而不论他事实上有没有去；假设一个人去医院接受了治疗，他的健康状况将是Y_{1i}（指去了医院情况下的潜在结果），同样不论他事实上有没有去。我们想知道的是Y_{1i}和Y_{0i}之间的差距，这个差距就可以解释为第i个人在医院接受的治疗对其健康状况产生的影响。也就是我们一直希望研究的因果效应，这里的“因”是是否去医院接受治疗，“果”是两种选择下不同的健康状况，“因果效应”指的是两种健康状况之间的差别[2]。观察到的结果Y，可以用潜在结果的线性组合表示：

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & if\ D_i = 1 \\ Y_{0i} & if\ D_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \end{aligned} \tag{1.2}$$

可见在这个表达式中$Y_{1i} - Y_{0i}$，就是个体去医院接受治疗对其健康状况的影响。一般来说，Y_{1i}和Y_{0i}在总体中都有相应的分布，因此对于不同的人，去医院接受治疗的因果效应是不一样的。但是由于我们不可能同时看到某个人的两种潜在的健康状况，所以我们必须比较同一类人去医院治疗和不去医院治疗对其健康状况的影响。
尽管在是否去医院接受治疗所带来的不同结果间进行简单比较并非我们想要的，但是这种肤浅的比较还是能告诉我们一些关于潜在结果的有益信息。下面这个公式就将去医院接受治疗与否带来的对平均健康水平的差异与我们感兴趣的平均意义上的因果效应（average casual effect）联系了起来：

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]\ \text{（处理的平均因果效应）} \\ + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\ \text{（选择性偏误）} \end{aligned} \tag{1.3}$$

其中$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$是处理组的平均因果效应，$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$是选择性偏误。
而：

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 1] \tag{1.4}$$

就是那些接受医院治疗的人因为在医院得到治疗而获得的平均因果效应。这里$E[Y_{1i}|D_i = 1]$是接受住院治疗的人的平均健康水平，$E[Y_{0i}|D_i = 1]$是如果接受住院治疗的人本来没有得到治疗，他们的平均健康水平。我们能够观察到的健康状况的差异实际上由两部分组成，在我们关心的因果关系之外，剩下的那部分叫做选择性偏误（selection bias），它是去医院接受治疗与不去医院接受治疗的人如果没有被治疗时健康状况的平均差别。由于患病者比健康人更加倾向于寻求治疗，所以那些接受住院治疗的人的初始健康水平Y，本身就比较低，从而使得选择性偏误是负的。在这个例子中，选择性偏误的绝对值可能会很大，当它大过我们想要寻找的因果效应时，就足以掩盖我们所要寻找的因果关系的符号，使得观察到的情况和真实情况相反。因此，经济学中大部分经验研究的目的就是剔除这种选择性偏误，从而阐释某个变量带来的效果，比如这里的变量D_i。

用随机分配解决选择性偏误

对D_i进行随机分配可以解决上文提到的选择性偏误，因为随机分配使得D_i独立于潜在的结果。为了理解这一点，让我们考虑：

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\ + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \end{aligned}$$

其中，Y_{0i}和D_i之间的独立性使得我们可以知道$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$，从而可以将之前等式中的第二行选择性偏误消去。事实上，给定随机分配下D_i的独立性，我们还可以对因果效应继续简化：

$$\begin{aligned} E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

也就是说对接受医院治疗的人考虑因果效应等同于随机分配患者进行治疗得到的因果效应。主要的发现就是，随机分配D_i消去了选择性偏误。这并不意味着随机分配本身不存在问题，但是总的来说它解决了在经验研究中遇到的最重要的问题。

对实验的回归分析

无论使用的数据来自实验与否，回归都是研究因果关系的有用工具。假设因果效应对所有人都一样，也就是$Y_{1i}-Y_{0i}=\rho$，是个常数。如果因果效应被假设为常数，那么我们可以将等式：

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & if\ D_i=1 \\ Y_{0i} & if\ D_i=0 \end{cases} \\ &= Y_{0i}+(Y_{1i}-Y_{0i})D_i \end{aligned} \tag{1}$$

写为：

$$Y_i=\alpha+\rho D_i+\eta_i \tag{3.1}$$

其中，α为$E[Y_{0i}]$，ρ为$Y_{1i}-Y_{0i}$，η是Y_{0i}的随机部分$Y_{0i}-E(Y_{0i})$。根据处理状态（treatment status，指对被试者进行了何种处理）的有无，对上面这个等式求条件期望可得：

$$\begin{aligned} E[Y_i|D_i=1]&=\alpha+\rho+E[\eta_i|D_i=1] \\ E[Y_i|D_i=0]&=\alpha+E[\eta_i|D_i=0] \end{aligned} \tag{3.2}$$

于是：

$$E[Y_i|D_i=1]-E[Y_i|D_i=0]=\rho+E[\eta_i|D_i=1]-E[\eta_i|D_i=0] \tag{3.3}$$

其中ρ为处理效应，$E[\eta_i|D_i=1]-E[\eta_i|D_i=0]$为选择性偏误。因此选择性偏误意味着回归残差项$\eta_i$和回归元$D_i$之间的相关性。由于：

$$E[\eta_i|D_i=1]-E[\eta_i|D_i=0]=E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0] \tag{3.4}$$

上面等式指的是得到处理和没有得到处理的人的潜在结果的差别。在医院治疗的故事里，得到医院治疗的人的健康状况要差于没有得到医院治疗的人。

有个著名的教育学研究“田纳西州的师生比例改进计划”（Tennessee Student Teacher Achievement Ratio, STAR），用以评估小学小班教学的效果，开创性地使用了随机化研究方法（Krueger，1999）。在STAR实验中，D_i是随机分配的，所以选择性偏误项就消失了，对Y_i关于D_i的回归就估计出我们感兴趣的因果效应ρ。STAR使用不同回归模型时估计出了不同参数，主要是包含了变量D_i之外的一些控制变量。在用回归模型分析实验数据时使用控制变量有两个用处。首先，在STAR实验中使用了条件随机分配方法。具体而言，在同一学校内，将学生分配至不同的班级是随机的，但是在学校间，这种分配不是随机的（有些学生必然会在某个学校）。在不同类型学校（比如城市里的学校和农村的学校）接受教育可能会影响学生被分配进入小规模班级的可能性。在忽略了学校类型时，估计出的参数可能会被不同类型学校对学生成绩的影响而干扰。为了进行调整，Krueger在一些回归方程中包含了学校固定效应，也就是对每个进入STAR数据的学校估计一个截距项。但事实上，对学校固定效应的调整并没有对结果带来大的改变，但是如果我们不这样做，就不会知道这个事实。

在Krueger模型中的其他控制变量描述了学生的个体特征，这些特点包括诸如种族、性别、是否参与免费午餐等。在之前我们就已知道这些个体特征在不同班级类型之间已经得到平衡，也就是说这些特点已经系统性地与将学生分配至哪种类型的班级无关了。记这些控制变量为X_i，它们与D_i不相关，因此也就不会影响对ρ的估计。换句话说，在长的回归方程：

$$Y_i=\alpha+\rho D_i+X_i^{'}\gamma+\eta_i \tag{2}$$

里估计出的ρ与短回归方程$Y_i=\alpha+\rho D_i+\eta_i$中估计出来的$\rho$将会很接近。

参考文献

[1] Grumbach, K., Keane, D., Bindman, A., 1993. Primary care and public emergency department overcrowding. Am J Public Health 83, 372–378. https://doi.org/10.2105/AJPH.83.3.372
[2] Holland, P.W., 1986. Statistics and Causal Inference. Journal of the American Statistical Association 81, 945–960. https://doi.org/10.1080/01621459.1986.10478354
[3] Krueger, A.B., 1999. Experimental Estimates of Education Production Functions. The Quarterly Journal of Economics 114, 497–532. https://doi.org/10.1162/003355399556052
[4] Rubin, D.B., 1977. Assignment to Treatment Group on the Basis of a Covariate. Journal of Educational Statistics 2, 1–26. https://doi.org/10.3102/10769986002001001
[5] Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701. https://doi.org/10.1037/h0037350

---

1. 在最近几十年，由于社区医疗机构和医疗器械的缺乏，美国人将原本用于对重伤或者有生命危险的患者进行治疗的急诊室用于初级护理。而且以这种方式进行初级护理的人往往是那些低收入、年老的人群，他们雇不起家庭医生，因此到医院进行初级护理。 ↩

2. 这里使用的潜在结果的观点是目前对因果关系进行研究的基石。在发展这个概念过程中出现的重要参考文献包括Rubin（1974，1977）以及Holland（1986），其中Holland（1986）将潜在结果中蕴含的因果框架称为Rubin的因果模型。 ↩