统计因果推断及应用—2023 年春季期末作业

截止日期：2023 年 04 月 18 日晚 22：00

1 Presentation (必选题)

1. 请用课堂所学的方法分析实际数据，提供数据来源、数据整理过程、分析过程、及最终结论。整理成小组报告形式提交，请组长提交，第九周课堂展示。
2. 因果推断方面的论文读书报告。整理成小组读书报告形式提交，此部分内容可以个人 or 小组，第九周课堂展示。
3. 科研交流展示。提交展示 ppt 即可，此部分内容可以个人 or 小组，第九周课堂展示。

注: 前两个选项也建议做展示 ppt，每个小组控制在 15 分钟以内。

以下作业截止日期：2023 年 04 月 30 日晚 22：00

2 个人作业 (以下作业必选 5 题，11 大题换算为 2 题，12 大题换算为 3 题，13 题换算为 4 题，多选酌情加分)

1. 分别用一句话叙述 Simpson's paradox 和 Surrogate paradox 的内容。列举你的生活和研究中可能潜在的以上悖论的例子。
2. 叙述可识别性的定义。
3. 叙述潜在结果模型中潜在结果的定义，和平均因果作用的定义。
4. 叙述可忽略性的定义和含义。
5. 叙述倾向得分的定义和性质。
6. 叙述倾向得分匹配的原理和估计思想。
7. 叙述工具变量在单调性假设下的非参识别性。
8. 叙述工具变量两阶段最小二乘法的有效性。
9. 叙述无效工具变量下的一些识别方法。
10. 叙述双重差分和合成控制法的异同。

11. 给定 STUVA，一致性和可忽略性，利用倾向得分估计、回归估计、双稳健估计、和一些机器学习的方法完成平均因果作用的估计，包括数据生成和估计。

12. 中介分析，如图1所示，我们给定以下假设，

    (a) $Y_{x,m} \perp\!\!\!\perp X \mid C$

    (b) $Y_{x,m} \perp\!\!\!\perp M \mid (I, X)$

    (c) $M_x \perp\!\!\!\perp X \mid W$

    (d) $Y_{x,m} \perp\!\!\!\perp M_{x^*} \mid I$

    请思考：

    (1) 证明 $\mathbb{E}(Y_{0,M_1})$ 可识别。

    (2) 设计一个模拟试验完成多稳健估计，参考文献 Tchetgen Tchetgen & Shpitser (2012) 。

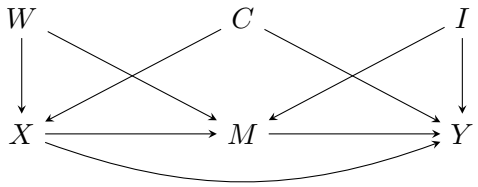

图 1: 中介分析，其中 $V = (W, C, I)$

13. 我们关心某个二值处理变量 X 对结果变量 Y 的因果作用。给定 STUVA，一致性和可忽略性，及完全观察的协变量 V，一个参数倾向得分

$$\pi(V;\alpha) = \text{pr}(X = 1 \mid V;\alpha)$$

和一个参数化结果回归

$$m(V;\beta) = \mathbb{E}(Y \mid X = 1, V;\beta)$$

这里 $\hat{E}$ 表示经验均值算子。我们可以用以下三种方法估计潜在结果均值 $\mu = \mathbb{E}(Y_1)$：

- IPW:

$$\hat{E}\left[\left\{\frac{X}{\pi(V;\hat{\alpha})} - 1\right\} \cdot g(V)\right] = 0$$
$$\hat{E}\left\{\frac{XY}{\pi(V;\hat{\alpha})} - \hat{\mu}_{\text{ipw}}\right\} = 0,$$

- REG:

$$\hat{E}\left[X\{Y - m(V;\hat{\beta})\} \cdot h(V)\right] = 0$$
$$\hat{E}\left\{m(V;\hat{\beta}) - \hat{\mu}_{\text{reg}}\right\} = 0,$$

- DR:

$$\hat{E}\left[\left\{\frac{X}{\pi(V;\hat{\alpha})} - 1\right\} \cdot g(V)\right] = 0$$
$$\hat{E}[X\{Y - m(V;\hat{\beta})\} \cdot h(V)] = 0$$
$$\hat{E}\left[\frac{XY}{\pi(V;\hat{\alpha})} + \left\{1 - \frac{X}{\pi(V;\hat{\alpha})}\right\} m(V;\hat{\beta}) - \hat{\mu}_{\text{dr}}\right] = 0.$$

作业如下：

(a) 设计一个模拟试验完成上面三种估计方法，包括数据生成和估计。(参考广义矩估计思想)

(b) 证明如果工作模型正确，根据一定的正则条件，所得到的估计量是一致且渐进正态的。(参考广义矩估计思想)

(c) 计算它们的渐进方差，并解释它们如何受到 g 和 h 的选择影响。如果一个或两个工作模型不正确会发生什么？

3 课程建议（必选题，建议不超过 500 字）

- 关于课程内容、课程难易程度、任何你对因果推断的认识、以及课程评价。
- 欢迎批评指正:-) 。

参考文献

1. Tchetgen E J T, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis[J]. Annals of statistics, 2012, 40(3): 1816.