SISCER Module 12
Lecture 3: Sensitivity analysis

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 2023

Plan

- Motivating example: Smoking and lung cancer. (Qingyuan)
- Sensitivity analysis for matching. (Qingyuan)
- Alternative methods to estimate the average treatment effect. (Ting)
- Sensitivity analysis for weighting. (Qingyuan)

Smoking and lung cancer: A brief review of the history

- A seminal case-control study by Doll and Hill (1950) showed strong correlation between cigarette smoking and lung cancer.
- This was followed up by many prospective studies that match on many covariates, which all pointed to the same causal relationship (Doll and Hill, 1954; Hammond and Horn, 1954).
- 1957 statement by the UK Medical Research Council and 1964 report by the U.S. Surgeon General concluded that smoking is the principal cause of lung cancer.

Smoking and lung cancer: A brief review of the history

- But this was challenged by several statisticians and epidemiologists. For example, Berkson (1958) questioned the usage of risk ratio (instead of risk difference) in the studies and the lack of "specificity".

Table: Standardized death rates (per 1,000 men) in relation to smoking status, reproduced from Table V in Doll and Hill (1956) and Table 29 in Berkson (1958). The last two columns compare the death rates of heavy smokers (>25 g.) versus non-smokers in two different measures.

Cause of death	Smoking a daily average of				Heavy vs. Non- smokers	
	0 g.	1-14 g.	15-24 g.	>25 g.	Ratio	Difference
Lung cancer	0.07	0.47	0.86	1.66	23.71	1.59
Other cancer	2.04	2.01	1.56	2.63	1.29	0.59
Other respiratory diseases	0.81	1.00	1.11	1.41	1.74	0.60
Coronary thrombosis	4.22	4.64	4.60	5.99	1.42	1.77
Other causes	6.11	6.82	6.38	7.19	1.18	1.08

Smoking and lung cancer: A brief review of the history

- More relevant to us is the criticism by Fisher (1958) and response by Cornfield et al. (1959).
- Fisher was also a geneticist and questioned whether the association between smoking and lung cancer can be explained by confounding genotypes. He offered some preliminary twin data suggesting smoking is genetically heritable.
- This prompted the first sensitivity analysis that established a mathematical inequality which amounts to the following in this example.

  If cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect.

Matched observational studies

- By matching units in an observational studies with very similar covariates, the hope is that we reconstruct a block randomized experiment.
- Consider the Neyman-Rubin causal model. Suppose treated observation $i = 1, \ldots, n$ is matched to control observation $i + n$. Define

$$\mathcal{M} = \{\boldsymbol{a} \in \{0,1\}^{2n} \mid a_i + a_{i+n} = 1, i = 1, \ldots, n\}$$

- Randomization analysis of matched observational studies assumes

$$\mathbb{P}\Big(\boldsymbol{A} = \boldsymbol{a} \,\Big|\, \boldsymbol{X}, \boldsymbol{A} \in \mathcal{M}\Big) = \begin{cases} 2^{-n_1}, & \text{if } \boldsymbol{a} \in M, \\ 0, & \text{otherwise.} \end{cases}$$

- This would be satisfied if $(\boldsymbol{X}_i, A_i)$ are drawn i.i.d. (independent and identically distributed) from a population and the matching is exact.
- By further assuming no unmeasured confounders $A_i \perp\!\!\!\perp Y_i(a) \mid \boldsymbol{X}_i$ for all a, randomization tests can be constructed as in randomized experiments.

No unmeasured confounders

Let U be the unmeasured confounder (e.g. $U_i = (Y_i(0), Y_i(1))$). The key assumption above is that

$$\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big) = \frac{1}{2}.$$

- No unmeasured confounders allows us to discard U_i, U_{i+n}.
- Let $\pi(\boldsymbol{x}) = \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x})$ be the **propensity score**. Matching by $\boldsymbol{X}$ (exactly) then establishes the equality, because

$$\begin{aligned}&\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}\Big)\\ =&\frac{\pi(\boldsymbol{X}_i)(1-\pi(\boldsymbol{X}_{i+n}))}{\pi(\boldsymbol{X}_i)(1-\pi(\boldsymbol{X}_{i+n})) + (1-\pi(\boldsymbol{X}_i))\pi(\boldsymbol{X}_{i+n})}.\end{aligned}$$

Rosenbaum's sensitivity model

- Inspired by Cornfield's sensitivity analysis, we would like to use a model that bounds the magnitude of unmeasured confounding.
- One option is the following model proposed by Rosenbaum (1987):

$$1/\Gamma \leq \mathrm{OR}(\mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u), \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u')) \leq \Gamma,$$

  where $OR(p, q) = \{p/(1-p)\}/\{q/(1-q)\}$ is the odds ratio and $\Gamma \geq 1$.
- This is equivalent to assume the following logistic model

$$\log \frac{\mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u)}{\mathbb{P}(A_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u)} = g(\boldsymbol{x}) + \gamma u, \ \ 0 \leq \gamma \leq \log \Gamma, 0 \leq U \leq 1.$$

Rosenbaum's sensitivity analysis (Rosenbaum, 2002)

Let $\pi_i = \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i, U_i)$, $i = 1, \ldots, 2n$. A consequence of Rosenbaum's sensitivity model is that

$$\frac{1}{1+\Gamma} \leq \mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big) = \frac{\pi_i(1-\pi_{i+n})}{\pi_i(1-\pi_{i+n}) + (1-\pi_i)\pi_{i+n}} \leq \frac{\Gamma}{1+\Gamma}.$$

- So within each pair, a fair coin toss is replace by a biased coin toss.
- We then seek the *least favorable* randomization distribution that is allowed by Rosenbaum's sensitivity model. This is usually given by the following (if we are trying to explain away a apparently positive treatment effect):

$$\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big) = \begin{cases} \frac{1}{1+\Gamma}, & \text{if } Y_i \geq Y_{i+n}, \\ \frac{\Gamma}{1+\Gamma}, & \text{if } Y_i < Y_{i+n}. \end{cases}$$

Sensitivity table and value

- A typical table of results of Rosenbaum's sensitivity analysis looks like the following.

Γ	1.0	2	4	8	**9**	10
Worst-case p-value	0.0001	0.0005	0.001	0.005	**0.01**	0.02

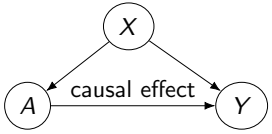
- The value of Γ where the worst-case p-value crosses the significance threshold (e.g. 0.01) is called the **sensitivity value** of the study. This is equal to 9 in Cornfield's example.
- The sensitivity value bears some similarity with the p-value. Both are random quantities determined by the data and indicate the strength of evidence.
- One may consider the problem of how to design an observational studies, not to minimize the p-value, but to maximize the sentivity value (Rosenbaum, 2010; Zhao, 2018).

What we learn from a sensitivity anlaysis

- A sensitivity analysis replaces qualitative claims about whether unmeasured biases are present with an **objective quantitative statement** about the magnitude of bias that would need to be present to change the conclusions.
- In this sense, a sensitivity analysis speaks to the assertion "it might be bias" in much the same way that a P-value speaks to the assertion "it might be bad luck".
- Because a genotype that had as large an effect on smoking and lung cancer might be considered unlikely in light of knowledge of the genotype's effect on other common diseases, the sensitivity analysis **strengthens the evidence** that smoking causes lung cancer although it does not prove that smoking causes lung cancer.

Recall: Observational studies under no unmeasured confounding

1. (No unmeasured confounders assumption, NUCA) Treatment among individuals with a particular value of X is essentially at random: $A \perp (Y(0), Y(1)) \mid X$.



2. (Positivity or Overlap assumption) $0 < P(A = 1 \mid X) < 1$ for all X

In previous lectures, we talked about

- Using **matching** to resemble randomized experiments under NUCA.
- **Rosenbaum's sensitivity analysis** to gauge the robustness of the study conclusion that presumed NUCA to deviation from it.

A toy observational study: all binary variables

- Y = death
- A = surgery
- X = severe injury
- $E(Y \mid A = 1) = 0.4$
- $E(Y \mid A = 0) = 0.6$
- What is the causal effect of surgery on mortality?
- Formally: what is ATE $= E(Y(1)) - E(Y(0))$?
- Note: prob. of receiving surgery are
  - 31% for non-severe injury
  - 86% for severe injury
- We need to break the correlation between A and X

$X = 0$		Outcome Y 0	1	Total
Treatment A	0	4	5	9
	1	3	1	4
	Total	7	6	13

$X = 1$		Outcome Y 0	1	Total
Treatment A	0	0	1	1
	1	3	3	6
	Total	3	4	7

Method 1: outcome regression (g-computation, standardization)

NUCA can be viewed as running a "randomized experiment" at each value of X.

$$\text{ATE} = \int \text{ATE}(x)P(X = x)dx = E[\mu_1(X) - \mu_0(X)]$$

where $\text{ATE}(x) = E[Y(1)|X = x] - E[Y(0)|X = x]$ and $\mu_a(x) = E(Y|X = x, A = a)$.

- Implementation:
  1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
  2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
  3. The estimator is $\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i)$.
- Advantages: more efficient; usually computationally stable
- Disadvantages: sensitive to model misspecification; potential danger of extrapolation

Outcome regression in our toy example

- $\mu_1(x=0) = 1/4$, $\mu_0(x=0) = 5/9$
- $\mu_1(x=1) = 3/6$, $\mu_0(x=1) = 1$

$$
\begin{aligned}
\text{ATE} &= \text{ATE}(0)P(X=0) + \text{ATE}(1)P(X=1) \\
&= \left(E[Y(1)|X=0] - E[Y(0)|X=0]\right)P(X=0) \\
&\quad + \left(E[Y(1)|X=1] - E[Y(0)|X=1]\right)P(X=1) \\
&= \left(\frac{1}{4} - \frac{5}{9}\right)\frac{13}{20} + \left(\frac{3}{6} - \frac{1}{1}\right)\frac{7}{20} \\
&= (-0.31) \times 0.65 + (-0.5) \times 0.35 \\
&= -0.37
\end{aligned}
$$

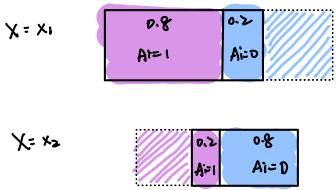
$X=0$		Outcome Y 0	1	Total
Treatment A	0	4	5	9
	1	3	1	4
	Total	7	6	13

$X=1$		Outcome Y 0	1	Total
Treatment A	0	0	1	1
	1	3	3	6
	Total	3	4	7

Method 2: Inverse probability weighting (IPW)

IPW assigns every unit different weight to create a pseudo-population in which A no longer depends on X.

$$\text{ATE} = E\left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)}\right]$$



where $\pi(X) = P(A = 1|X)$ is the propensity score.

- Implementation:
  1. Fit a propensity score model $\hat{\pi}(x)$
  2. The estimator is $\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}(X_i)}$.
- Advantages: sometimes the propensity score may be easier to model
- Disadvantages: sensitive to model misspecification; less efficient; can be unstable if some $\hat{\pi}(X_i)$ are close to zero (stabilized IPW helps a little (Glynn and Quinn, 2010))

IPW in our toy example

- $P(A = 1 \mid X = 0) = 4/13$
- $P(A = 1 \mid X = 1) = 6/7$

$$
\begin{aligned}
\text{ATE} &= \frac{1}{n}\sum_{i:X_i=0}\frac{A_iY_i}{4/13} - \frac{1}{n}\sum_{i:X_i=0}\frac{(1-A_i)Y_i}{9/13} \\
&\quad + \frac{1}{n}\sum_{i:X_i=1}\frac{A_iY_i}{6/7} - \frac{1}{n}\sum_{i:X_i=1}\frac{(1-A_i)Y_i}{1/7} \\
&= \frac{1}{20}\left(\frac{13}{4}\times 1 - \frac{13}{9}\times 5 + \frac{7}{6}\times 3 - 7\times 1\right) \\
&= -0.37
\end{aligned}
$$

$X = 0$		Outcome Y 0	1	Total
Treatment A	0	4	5	9
	1	3	1	4
	Total	7	6	13

$X = 1$		Outcome Y 0	1	Total
Treatment A	0	0	1	1
	1	3	3	6
	Total	3	4	7

Method 3: Augmented inverse probability weighting (AIPW)

Why not use both the outcome model and the propensity score model?

$$\text{ATE} = E\left[\frac{A\{Y-\mu_1(X)\}}{\pi(X)} - \frac{(1-A)\{Y-\mu_0(X)\}}{1-\pi(X)} + \mu_1(X) - \mu_0(X)\right]$$

- Implementation:
  1. Fit an outcome model $\hat{\mu}_1(x)$ using data from the treated group
  2. Fit another outcome model $\hat{\mu}_0(x)$ using data from the control group
  3. Fit a propensity score model $\hat{\pi}(x)$
  4. The estimator is $\frac{1}{n}\sum_{i=1}^{n}\frac{A_i(Y_i-\hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)(Y_i-\hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i)$.
- Advantages: doubly robust; more efficient and stable compared to IPW; can be used in combination with machine learning algorithms
- Disadvantages: still biased if all models are wrong; can be unstable if some $\hat{\pi}(X_i)$ are close to zero

AIPW in our toy example[1]

- $\mu_1(x=0) = 1/4$, $\mu_0(x=0) = 5/9$
- $\mu_1(x=1) = 3/6$, $\mu_0(x=1) = 1$
- $P(A=1|X=0) = 4/13$
- $P(A=1|X=1) = 6/7$

$$\begin{aligned}\text{ATE} = &\frac{1}{n}\sum_{i:X_i=0}\frac{A_i(Y_i - 1/4)}{4/13} - \frac{1}{n}\sum_{i:X_i=0}\frac{(1-A_i)(Y_i-5/9)}{9/13} \\ &+ \frac{1}{n}\sum_{i:X_i=1}\frac{A_i(Y_i - 3/6)}{6/7} - \frac{1}{n}\sum_{i:X_i=1}\frac{(1-A_i)(Y_i-1)}{1/7} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i) \\ &= 0 + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_0(X_i) = -0.37\end{aligned}$$

$X=0$		Outcome Y 0	1	Total
Treatment A	0	4	5	9
	1	3	1	4
	Total	7	6	13

$X=1$		Outcome Y 0	1	Total
Treatment A	0	0	1	1
	1	3	3	6
	Total	3	4	7

[1] When X is discrete and $\hat{\pi}(x) \in (0,1)$ for all x, all three estimators are equal.

Statistical inference

When $\mu_0(x), \mu_1(x), \pi(x)$ are estimated using parametric models

- We can use nonparametric (i.e., assumption-lite) bootstrapping for estimating standard errors, computing confidence intervals, and hypothesis testing
  - The lazy statistician's method
  - Sample with replacement to create a new sample of the same size as the study sample, estimate the effect estimate in that sample, repeat many (e.g., 1000) times, find 2.5 and 97.5 percentiles of the 1000 estimates as the 95% confidence interval
- Sandwich variance estimator (implemented in the CausalGAM package in R)

When $\mu_0(x), \mu_1(x), \pi(x)$ are estimated using machine learning algorithms

- Key: use AIPW + cross fitting (implemented in the AIPW package in R)

More recent ideas: "covariate-balancing weights"

- The purpose of IPW is to create a pseudo-population in which the observed covariates are balanced between the treated and control. However, the conventional approach of estimating $\pi(x)$ and use $1/\hat{\pi}(X_i)$ as the weight for unit i is sensitive to model misspecification and can be unstable
- Idea: directly find weights w_i's that balance the observed covariates

Entropy balancing (Hainmueller, 2012)

Let $\{c_j(x)\}$ be a set of covariate moments (e.g., expectation). Find weights w_i^{EB}'s using

$$\text{maximize}_w \; -\sum_{A_i=0} w_i \log w_i$$

$$\text{subject to } \sum_{A_i=0} w_i c_j(X_i) = \bar{c}_j(1) = \frac{1}{n_1} \sum_{A_i=1} c_j(X_i), \; j = 1 \ldots, p$$

$$\sum_{A_i=0} w_i = 1, \; w_i > 0, i = 1, \ldots, n$$

- The Entropy balancing estimator of average treatment effect for treated is $\bar{Y}_1 - \sum_{A_i=0} w_i^{\text{EB}} Y_i$.
- This can be understood as an IPW estimator in which $\hat{\pi}(X_i)$ is fitted using a special loss function. The entropy balancing estimator enjoys certain double robustness property (Zhao and Percival, 2016).

Sensitivity analysis for IPW estimator

- Recall the inverse-probability weighted estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}(X_i)}.$$

- This based on the assumption that $A \perp\!\!\!\perp (Y(0), Y(1)) \mid X$ and the identity

$$\text{ATE} = E\left[\frac{AY}{\pi(X)} - \frac{(1-A)Y}{1-\pi(X)}\right].$$

- Suppose there is an unmeasured confounder U such that $A \perp\!\!\!\perp (Y(0), Y(1)) \mid X, U$. Then

$$\text{ATE} = E\left[\frac{AY}{\pi(X,U)} - \frac{(1-A)Y}{1-\pi(X,U)}\right].$$

- Of course $\pi(X, U)$ cannot be estimated, but what if we assume it is not too different from $\pi(X)$?

A different sensitivity model

Marginal sensitivity model

Let $\pi(X) = \mathbb{P}(A = 1 \mid X)$ and $\pi(X, U) = \mathbb{P}(A = 1 \mid X, U)$. This model assumes

$$1/\Gamma \leq \mathrm{OR}(\pi(X), \pi(X, U)) \leq \Gamma, \text{ for all } X, U.$$

- This different from **Rosenbaum's sensitivity model**

$$1/\Gamma \leq \mathrm{OR}(\pi(X, U), \pi(X, U')) \leq \Gamma, \text{ for all } X, U, U'.$$

- But they are related. Let $\mathcal{M}(\Gamma)$ and $\mathcal{R}(\Gamma)$ be all the $\pi(X, U)$ that satisfies the marginal and Rosenbaum sensitivity models, respectively. Then

$$\mathcal{M}(\sqrt{\Gamma}) \subseteq \mathcal{R}(\Gamma) \subseteq \mathcal{M}(\Gamma).$$

Computation

- Similar to Rosenbaum's sensitivity analysis, the main idea here is to find the worst-case $\pi(X, U)$.
- The marginal sensitivity model puts a linear constraint on $\pi(X, U)$:

$$1/\Gamma \leq \mathrm{OR}(\pi(X), \pi(X, U)) \leq \Gamma \iff 1/\Gamma \leq \frac{1/\pi(X) - 1}{1/\pi(X, U) - 1} \leq \Gamma.$$

- To estimate $\beta = \mathbb{E}[Y(1)]$, we can minimize & maximize the stabilized IPW estimator

$$\hat{\beta} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(X_i, U_i)}\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\pi}(X_i, U_i)}\right],$$

  under the above constraint. This can be similarly done for $\mathbb{E}[Y(0)]$.
- The optimization problem is a **linear fractional program** and can be solved very efficiently.

Inference via bootstrap

- To obtain $(1-\alpha)$ confidence intervals, we can compute the minmum and maximum over bootstrap resamples, and then take the $\alpha/2$ quantile of the bootstrap minima and $(1-\alpha)/2$ quantile of the bootstrap maxima.
- Validity is guaranteed by the generalized minimax inequality:

$$Q_{\frac{\alpha}{2}}\left(\inf_{\eta}\hat{\hat{\beta}}(\eta)\right) \leq \inf_{\eta} Q_{\frac{\alpha}{2}}\left(\hat{\hat{\beta}}(\eta)\right) \leq \sup_{\eta} Q_{1-\frac{\alpha}{2}}\left(\hat{\hat{\beta}}(\eta)\right) \leq Q_{1-\frac{\alpha}{2}}\left(\sup_{\eta}\hat{\hat{\beta}}(\eta)\right).$$

Percentile bootstrap sensitivity interval

Union sensitivity interval

- More details: Zhao et al. (2019).
- R package: `https://github.com/qingyuanzhao/bootsens`.
- An improvement of the original method: Dorn and Guo (2021).

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *Journal of the American Statistical Association*, 53(281):28–38.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203. **Note:** This classic paper was reprinted in the *International Journal of Epidemiology*, 38(5), 2009 with commentaries.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682):739–748.

Doll, R. and Hill, A. B. (1954). The mortality of doctors in relation to their smoking habits. *BMJ*, 1(4877):1451–1455.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *BMJ*, 2(5001):1071–1081.

Dorn, J. and Guo, K. (2021). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing.

Fisher, R. A. (1958). Cancer and smoking. *Nature*, 182(4635):596–596.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46.

Hammond, E. C. and Horn, D. (1954). The relationship between human smoking habits and death rates. *Journal of the American Medical Association*, 155(15):1316–1328.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer Series in Statistics. Springer, New York.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer, New York.

Zhao, Q. (2018). On sensitivity value of pair-matched observational studies. *Journal of the American Statistical Association*, 114(526):713–722.

Zhao, Q. and Percival, D. (2016). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761.