Bayesian causal inference: day 1 practical

Chris Holmes and Andrew Yiu

September 16, 2024

1 Randomization inference

The first two examples in this practical uses a dataset studied by Dehejia and Wahba (1991), which is a subset of the data from LaLonde (1986). The dataset contains information on disadvantaged male workers in the mid-1970's, some of whom were randomly selected to join the National Supported Work Demonstration (NSW) program to help them enter the labour market. It is recommended to skim read the first two sections of Imbens and Xu (2024) to obtain some background knowledge of the study.

(i) The skeleton code for this section is found in the script `day1_practical1.R`, which starts by loading the Lalonde data:

```
load("data/lalonde.RData")
```

The experimental dataset that we study today is accessed by `ldw` (standing for LaLonde, Dehejia and Wahba). The outcome of interest Y is the calendar 1978 earnings in dollars, stored in the `re78` variable. The treatment variable T (given by `treat`) is a binary variable that takes the value 1 if the individual was selected to join the NSW program. Compute the difference-in-means estimator

$$S = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1 - T_i) Y_i}{\sum_{i=1}^{n} (1 - T_i)}$$

and interpret the result.

(ii) Recall that the potential outcome variables are linked to the observed outcomes via the SUTVA/consistency assumption:

$$Y_i = T_i Y_i^1 + (1 - T_i) Y_i^0.$$

The *Fisher sharp null hypothesis* is $H_0 : Y_i^1 = Y_i^0$ for all $i = 1, \ldots, n$. Interpret this hypothesis in the context of the study.

(iii) We will test the sharp null by treating all the potential outcomes $\{(Y_i^1, Y_i^0) : i = 1, \ldots, n\}$ as fixed, so that the only random variability arises from the treatment assignments $\{T_i\}$. In particular, we assume that $T_1, \ldots, T_n$ i.i.d.$\sim$ Ber(0.5) (called *Poisson sampling*). Our test statistic is S; write an algorithm to generate and store 10000 i.i.d. samples from the sampling distribution of S under the sharp null. Plot a histogram/density of the samples and indicate the value of the actual observed statistic of S (e.g. by a vertical line).

(iv) Use your samples to construct a Monte Carlo estimate of a one-sided p-value for the observed statistic of S and comment on your findings.

(v) We can generalize to the class of null hypotheses

$$H_0 : Y_i^1 - Y_i^0 = \beta, \quad i = 1, \dots, n$$

for a fixed value of β. Discuss how these hypotheses can be interpreted. For any value of β, we can similarly estimate a one-sided p-value for the observed statistic of S. By "inverting" these p-values, construct an estimate of a one-sided 95% confidence interval for β. *Hint: a set of β values for which the corresponding null hypothesis is* ***not*** *rejected will form a 95% confidence interval. This comes straight from the definition of a confidence interval (e.g. Theorem 9.2.2. of Casella and Berger, 2002)*

2 Bayesian model-based inference and predictive resampling

Instead of treating the potential outcomes as fixed (like the randomization inference approach in §1), we now view the potential outcomes as random variables drawn i.i.d. from a hypothetical "super-population". For instance, we could be interested in **all** disadvantaged male workers in the US in 1974 and then view the data as being randomly sampled from this superset of individuals. Unlike the randomization inference in §1, we now need a model for how this superpopulation is distributed. The stronger modelling assumptions required for superpopulation inference increases the risk of model misspecification bias, but we gain flexibility in our choice of estimand and statistical analysis.

Imbens and Rubin (2015) analysed the Dehejia and Wahba (1991) dataset using Bayesian model-based inference. To match their notation and reduce clutter, we will rescale Y, Y^1, Y^0 so that the units are now in \$1000. They posited a normal model[1]

$$\begin{pmatrix} Y_i^0 \\ Y_i^1 \end{pmatrix} \Big| \mu_c, \mu_t \sim \mathcal{N}\left(\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix}, \begin{pmatrix} 25 & 0 \\ 0 & 64 \end{pmatrix} \right), \tag{1}$$

where (μ_c, μ_t) are unknown parameters. Under the randomization assumption $(Y^0, Y^1) \perp\!\!\!\perp T$, the conditional likelihood given $T_{1:n}$ is

$$\mathcal{L}(\mu_c, \mu_t) = \prod_{i=1}^{n} f_{\mu_c}(Y_i)^{1-T_i} f_{\mu_t}(Y_i)^{T_i},$$

where f_{μ_c} and f_{μ_t} are the densities for $\mathcal{N}(\mu_c, 25)$ and $\mathcal{N}(\mu_t, 64)$ respectively.

(i) For the time being, we are interested in the *average treatment effect (ATE)* $\theta = \mu_t - \mu_c$. Comment on the difference in interpretation between θ and the β parameter from §1 (aside from the scaling).

[1] We have changed the variance of Y_i^0 from 100 to 25 to bring the value closer to the empirically observed sample variance.

(ii) What are the maximum likelihood estimates $\hat{\mu}_c$ and $\hat{\mu}_t$? Therefore, what is the plug-in maximum likelihood estimate $\hat{\theta} = \hat{\mu}_t - \hat{\mu}_c$?

Imbens and Rubin (2015) specified a diffuse prior

$$\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10000 & 0 \\ 0 & 10000 \end{pmatrix} \right)$$

with the intention of allowing the data to dominate the posterior behaviour. Due to the conjugate model structure, we can easily find the exact form for the posterior for (μ_c, μ_t). Let $N_c = \sum_{i=1}^{n}(1 - T_i) = 260$ and $N_t = \sum_{i=1}^{n} T_i = 185$. Then

$$\begin{pmatrix} \mu_c \\ \mu_t \end{pmatrix} \Big| (Y_{1:n}, T_{1:n}) \sim \mathcal{N}\left(\begin{pmatrix} \hat{\mu}_c \cdot \frac{N_c \cdot 10000}{N_c \cdot 10000 + 25} \\ \hat{\mu}_t \cdot \frac{N_t \cdot 10000}{N_t \cdot 10000 + 64} \end{pmatrix}, \begin{pmatrix} (N_c/25 + 1/10000)^{-1} & 0 \\ 0 & (N_c/64 + 1/10000)^{-1} \end{pmatrix} \right). \tag{2}$$

(iii) Construct a central 95% posterior credible interval for θ (i.e. use the 2.5% and 97.5% posterior quantiles for the lower and upper bounds respectively) and discuss your findings.

Using the same model, we now undertake a parallel analysis with *predictive resampling*. The method proceeds by repeatedly forward-simulating $(Y_{n+1}, T_{n+1}), (Y_{n+2}, T_{n+2}), \ldots$ based on the intuitive notion that an infinite dataset would allow us to recover our estimand exactly, i.e. $\theta = \theta(Y_{n+1:\infty}, T_{n+1:\infty})$. Across multiple independent iterations of this forward-simulation, we obtain a random sample of our estimand from its *martingale posterior* (Fong et al., 2021), which enables us to quantify uncertainty.

In practice, of course, we are unable to simulate an infinite amount of data. Fortunately, it is sufficient to truncate the procedure at $n+M$ for a large value of M, as the theory tells us that the finite approximation error disappears as $M \to \infty$. Pseudo-code can be found in Algorithm 1.

Algorithm 1: Predictive resampling for the average treatment effect

```
Input: $B$ (no. of posterior samples) and $M$ (no. of forward simulation steps) are large integers;
for $j \leftarrow 1$ to $B$ do
    for $m \leftarrow 1$ to $M$ do
        Sample $(\mu_c, \mu_t)$ from its posterior given $(Y_{1:n+m-1}, T_{1:n+m-1})$;
        Sample $T_{n+m}$ from Ber(0.5) (Assign a randomized treatment);
        Sample $Y_{n+m}$ from $f_{\mu_c}(y)$ if $T_{n+m} = 0$ or $f_{\mu_t}(y)$ if $T_{n+m} = 1$;
    end
    Evaluate $\theta_M^{(j)} = \frac{\sum_{m=1}^{M} T_{n+m} Y_{n+m}}{\sum_{m=1}^{M} T_{n+m}} - \frac{\sum_{m=1}^{M} (1-T_{n+m}) Y_{n+m}}{\sum_{m=1}^{M} (1-T_{n+m})}$;
end
Return $\{\theta_M^{(1)}, \ldots, \theta_M^{(B)}\}$.
```

(iv) In the `day1_practical2.R` script, there is an implementation of a single iteration of predictive resampling. Read the code and then execute it in your `R` session. The script also creates a line plot (Figure 1) that tracks the value of θ_m as m increases from 21 to $M = 3000$ (we start at 21 rather than 1 to ensure that the difference-in-means estimator is defined, i.e. we have forward simulated at least

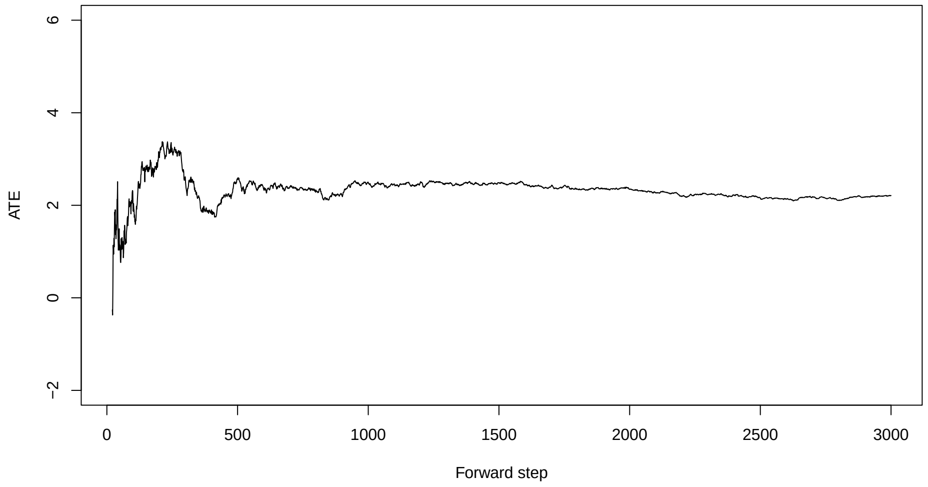

Figure 1: Single iteration of predictive resampling for the average treatment effect.

one treatment and one control). We can see that the value of θ stabilizes as M increases. You may like to experiment with the value of M. Can you devise a stopping criterion to automate the length of the procedure?

(v) Loop the procedure to obtain $B = 500$ independent samples of θ from its martingale posterior. Compare this posterior distribution with the one obtained previously.

(vi) Try changing the predictive distribution of T from Ber(0.5) to Ber(0.3). What do you notice? Can you explain why?

(vii) We now consider a slightly different estimand called the *sample average treatment effect (SATE)*

$$\theta_{SATE} = \frac{1}{n}\sum_{i=1}^{n}[Y_i^1 - Y_i^0].$$

This focuses on the missing potential outcomes in the observed sample, rather than the missing data in the superpopulation as operationalized in predictive resampling. Use (1) and (2) to draw posterior samples of θ_{SATE} and compare with the previous posteriors for θ. *Hint: Conditional on* (μ_c, μ_t)*, all the potential outcomes are jointly independent, i.e.*

$$Y_1^1 \perp\!\!\!\perp Y_1^0 \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_n^1 \perp\!\!\!\perp Y_n^0 \mid (\mu_c, \mu_t).$$

3 Observational data and target trial emulation

So far, we have been analysing data from a randomized experiment. We now proceed to study an observational dataset obtained from the NHANES I Epidemiologic Follow-up Study. More details can be found in Davis et al. (1994).

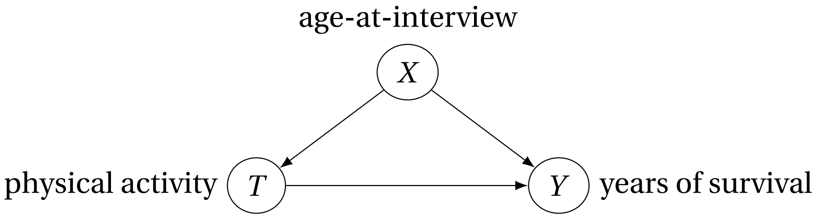

Figure 2: Causal diagram representing the assumed causal relationships for the NHANES dataset.

The NHANES I sample was interviewed in 1971 and followed for survival until 1992. Our interest is in the causal effect of being physically active (the treatment T) on the number of years of survival up to 1992 (the outcome Y). We will include the age at interview (pre-treatment covariate X) to adjust for confounding. As described in the lectures, we will introduce a regime indicator F_T that indexes the observational and hypothetical experimental regimes. In particular, we use $F_T = \mathscr{O}$ for the observational setting, from which the NHANES data was obtained. Our objective is to predict a hypothetical target trial to answer our causal question, which is indexed by $F_T = \mathscr{E}$.

(i) Let $P(Y, T, X \mid \mathscr{O})$ and $P(Y, T, X \mid \mathscr{E})$ denote the distribution of the variables in the observational and target populations respectively. In order to learn about the latter using our data, we require some transportability assumptions. First, we assume that

$$P(X \mid \mathscr{O}) = P(X \mid \mathscr{E}), \tag{3}$$

which states the distribution of the ages in our target population matches that of the participants in NHANES. Next, we require stability in the conditional outcome models:

$$P(Y \mid T, X, \mathscr{O}) = P(Y \mid T, X, \mathscr{E}). \tag{4}$$

What does this assumption mean in our context? (This may not be realistic but it suffices for the sake of simplicity for now; we consider modifying this assumption later).

The observed data from NHANES is denoted by $Z_{1:n} = (Y_{1:n}, T_{1:n}, X_{1:n})$. Recall from the lectures that we require a **joint predictive model** $p(z_{n+1:M} \mid z_{1:n}, F_T = \mathscr{E})$ to repeatedly impute the "missing" target trial data for a large population size M. We start by specifying a predictive called the *Bayesian bootstrap* for X on its own. This works very simply; given $X_{1:k}$, we predict X_{k+1} by sampling uniformly at random from the observed values $\{x_1, \ldots, x_k\}$. One could visualize this in terms of a **Pólya urn** that contains different coloured balls; at each step, we pick out one of the balls at random and put it back in the urn along with a new ball of the same colour. More formally, we are drawing X_{k+1} from the **empirical distribution** formed from $X_{1:k}$:

$$X_{k+1} \mid X_{1:k} \sim \mathbb{P}_k = \frac{1}{k} \sum_{i=1}^{k} \delta_{x_i}.$$

This forward simulation from the observational data is justified by our transportability assumption in (3). Pseudo-code for this procedure is given in Algorithm 2.

Algorithm 2: The Bayesian bootstrap

```
Input: $\mathbb{P}_n$ (empirical distribution of $X$ formed from the data); $B$ (no. of posterior samples) and
  $M$ (no. of forward simulation steps) are large integers;
for $j \leftarrow 1$ to $B$ do
    for $m \leftarrow 1$ to $M$ do
        Sample $X^{(b)}_{n+m}$ from $\mathbb{P}_{n+m-1}$;
        Update $\mathbb{P}_{n+m-1} \mapsto \mathbb{P}_{n+m}$ ;
    end
end
Return $\{X^{(1)}_{n+1:n+M}, \dots, X^{(B)}_{n+1:n+M}\}$.
```

(ii) Implement the Bayesian bootstrap for a single iteration ($B = 1$) for $M = 3000$ forward simulation steps. To verify convergence, track the proportion p_{n+m} of predicted X values equal to 65 up to $n+m$, i.e.

$$p_{n+m} = \frac{1}{m} \sum_{i=n+1}^{n+m} 1(X_i = 65),$$

and plot a line plot similar to the one in Figure 1.

Since treatment allocation for the target trial is fully randomized by design, we will predict each new T_k by carrying out an independent Ber(0.5) trial, i.e. flip a fair coin to decide treatment or control. For the conditional outcome prediction model, we will use a generalized t-distribution[2]

$$Y_{k+1} \mid x_{k+1}, t_{k+1}, z_{1:k} \sim t_{k-3}(w_{k+1}\hat{\beta}_k, s_k^2(1 + w_{k+1} V_k^{-1} w_{k+1}^{\mathrm{T}})),$$

where

$$\begin{aligned}
w_i &= (1, x_i, t_i)^{\mathrm{T}} \\
V_k &= (w_{1:k}^{\mathrm{T}} w_{1:k})^{-1} \\
\hat{\beta}_k &= V_k w_{1:k}^{\mathrm{T}} y_{1:k} \\
s_k^2 &= \frac{1}{k-3}(y_{1:k} - w_{1:k}\hat{\beta}_k)^{\mathrm{T}}(y_{1:k} - w_{1:k}\hat{\beta}_k).
\end{aligned}$$

This is based on a noninformative conjugate Bayesian linear regression analysis of Y on T and X with an intercept. The $k-3$ degrees of freedom arises from the fact that w_i has dimension 3. See Section 9 of the included document “BayesianLinearModel.pdf” for more details.

(iii) The full predictive resampling procedure for this model can be found in `day1_practical3.R`. Note: the Bayesian bootstrap is implemented here using a “shortcut” method instead of the Pólya urn scheme above; see the Appendix for more information. Read the code and execute; it should take around 2-3 minutes to run. You may wish to parallelize the code and increase the number of posterior samples and forward steps if such resources are available to you.

[2] A variable $X \sim t_\nu(\mu, s^2)$ is equivalent to $X \sim \mu + s \cdot T$, where T is distributed according to a standard Student t-distribution with ν degrees of freedom.

(iv) The code returns a vector of posterior samples for the average treatment effect, as well as a histogram and density estimator for the posterior distribution. Comment on the results.

(v) We made the crucial assumption in (4) that allowed us to transport a predictive model for the outcome from the observational data to the experimental setting. For simplicity, we only selected one covariate (age at interview) to form X. Looking at the data, can you think of other covariates that should be included to make the assumption more justifiable? Modify the code to augment the predictive model and compare your results.

4 Bonus theoretical exercise

Recall from the lecture material on potential outcomes that we made three assumptions for *identification*:

- (SUTVA/consistency) $Y_i = T_i Y_i^1 + (1 - T_i) Y_i^0$
- (Unconfoundedness/ignorability) $T_i \perp\!\!\!\perp (Y_i^1, Y_i^0) \mid X_i$
- (Overlap/positivity) $0 < \mathbb{P}(T_i = 1 \mid X_i) < 1$ with probability 1

Use these assumptions to derive an identification formula for the conditional variance $\mathrm{var}(Y_i^1 \mid X_i)$. (Recall that this means we want to write $\mathrm{var}(Y_i^1 \mid X_i)$ as as a functional of the observational distribution, i.e., the distribution of (Y_i, T_i, X_i)).

Appendix: An alternative implemention of the Bayesian bootstrap

In §3, we outlined a "Bayesian bootstrap" predictive model that predicted each new covariate by (re)sampling uniformly from the already observed covariate values. There is an alternative, more computationally efficient method that uses a two-stage sampling scheme. In the first stage, we sample a vector of uniform Dirichlet-distributed weights $(w_1, \dots, w_n)$. An equivalent way of sampling from the uniform Dirichlet distribution $\mathrm{Dir}(1, \dots, 1)$ is to draw a set of i.i.d. $\mathrm{Exp}(1)$ variables $(q_1, \dots, q_n)$ and normalize by their sum:

$$w_i = \frac{q_i}{\sum_{j=1}^n q_j}.$$

Since the Dirichlet weights sum to 1, the following:

$$F_n^w = \sum_{i=1}^n w_i \delta_{X_i}$$

is a probability distribution. In words, F_n^w is a discrete distribution supported only on the observed covariate values $\{x_1, \dots, x_n\}$ and takes the value x_i with probability w_i. In the second stage, we sample $X_{n+1:n+M}$ i.i.d. from F_n^w. It can be shown that this is equivalent to the Pólya urn scheme for the Bayesian bootstrap. Pseudo-code for this procedure can be found in Algorithm 3.

Algorithm 3: The Bayesian bootstrap (Dirichlet method)

```
Input: $X_{1:n}$ (covariate values observed in the data); $B$ (no. of posterior samples) and $M$ (no. of forward simulation steps) are large integers;
for $j \leftarrow 1$ to $B$ do
    Sample $(w_1, \ldots, w_n) \sim \mathrm{Dir}(1, \ldots, 1)$;
    Sample $X^{(b)}_{n+1:n+M} \sim \text{i.i.d.} \sum_{i=1}^{n} w_i \delta_{X_i}$;
end
Return $\{X^{(1)}_{n+1:n+M}, \ldots, X^{(B)}_{n+1:n+M}\}$.
```

References

G. Casella and R. Berger. *Statistical Inference*. Duxbury, 2002.

M. Davis et al. Health Behaviors and Survival among Middle-Aged and Older Men and Women in the NHANES I Epidemiologic Follow-Up Study. *Preventine Medicine*, 23:369–376, 1994.

R. Dehejia and S. Wahba. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1991.

E. Fong, C. Holmes, and S. G. Walker. Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*, 2021.

G. Imbens and D. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

G. Imbens and Y. Xu. Lalonde (1986) after nearly four decades: lessons learned. *arXiv preprint arXiv:2406.00827*, 2024.

R. LaLonde. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76:604–620, 1986.