Bayesian causal inference: day 1 practical solutions

Chris Holmes and Andrew Yiu

September 16, 2024

1 Randomization inference

The first two examples in this practical use 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.

Answer: On average, the participants who were selected to join the NSW program earned approximately \$1794 more than the controls in 1978.

(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 $i = 1, \dots, n$. Interpret this hypothesis in the context of the study.

Answer: Joining the NSW program had no causal effect on changing the calendar 1978 earnings for all participants. Note that this is much stronger than the hypothesis that there is no causal effect on average, which would allow some participants to benefit while others are hindered.

(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.∼ 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.

Answer: We can estimate a one-sided p-value by finding the proportion of samples that exceed the observed statistic, which is approximately 0.002. Thus, we would reject the sharp null at both the 5% and 1% levels.

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

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

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)*

Answer: Like the sharp null, these hypotheses posit a constant causal effect across all individuals. In the context of the study, the hypothesis states that every individual earns β dollars more by joining the NSW program. We can construct a one-sided 95% confidence interval by searching for a β value such that the observed statistic lies at the rejection boundary (i.e. has a p-value ≈0.05). This could be achieved through an automated iterative search or by trial and error. We obtain an approximate confidence interval of $\{\beta > 760\}$.

2 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

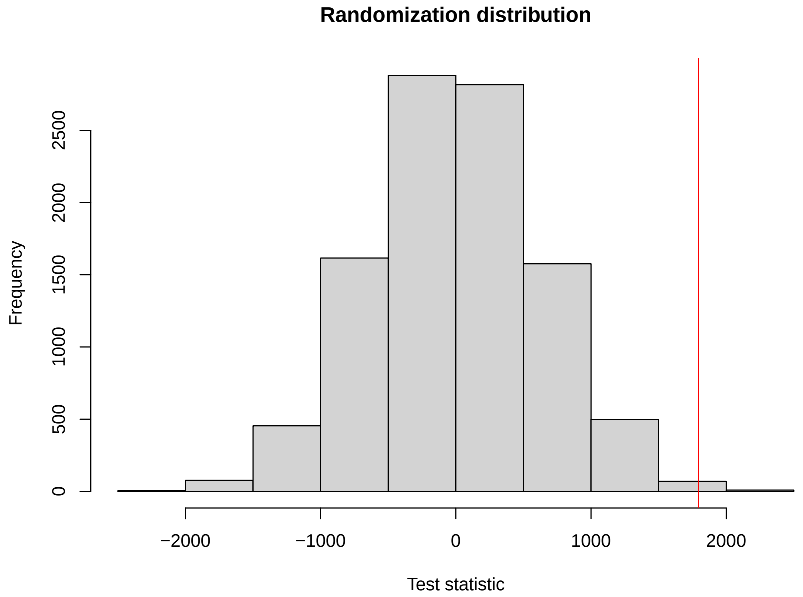

Figure 1: Histogram of the sampling distribution of S under the sharp null with the observed statistic indicated by the vertical red line.

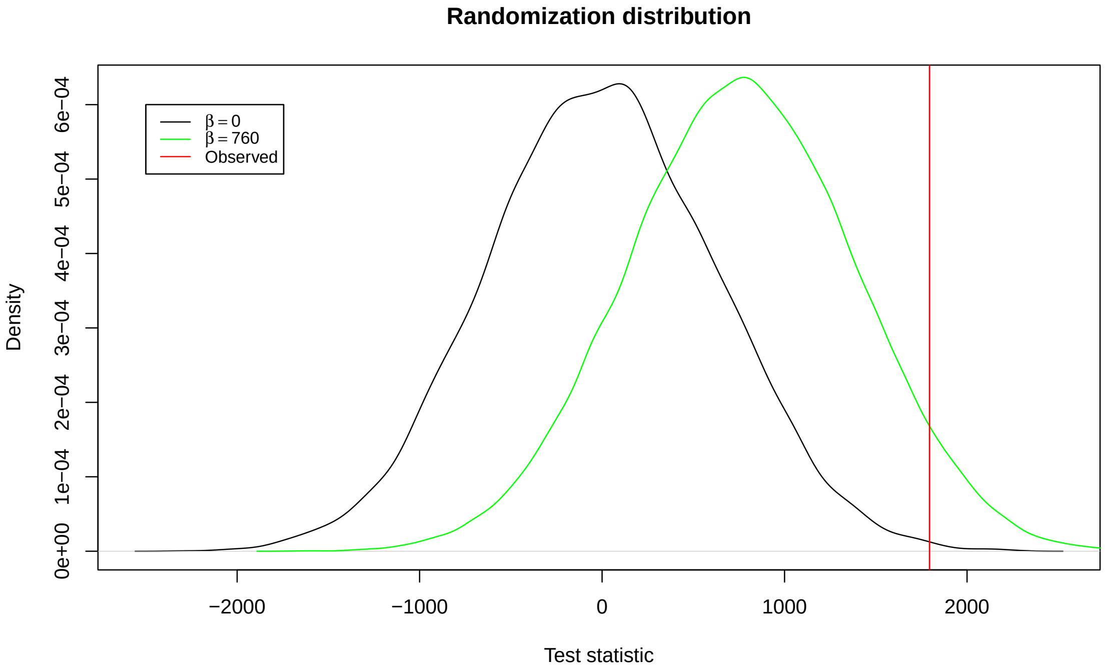

Figure 2: Comparison of sampling distribution densities under the hypotheses $\beta = 0$ and $\beta = 760$ with the observed statistic indicated by the vertical red line.

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

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

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).

*Answer: The β parameter from §1 is an **individual-level causal effect**, which is assumed to be constant across all individuals. In contrast, θ corresponds to an average causal effect for the superpopulation from which the sample was drawn. For instance, if $\theta = 0$, it is possible that some individuals benefit while others are hindered, as long as the effects cancel out on average. But if $\beta = 0$, then all individuals are completely unaffected.*

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

Answer: since we are working with a normal model, we know from standard MLE theory that $\hat{\mu}_c$ and $\hat{\mu}_t$ are the group-specific outcome means for the control and treatment groups respectively. Thus, $\hat{\theta}$ is (again) the difference-in-means estimator.

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.

Answer: We obtain a central 95% posterior credible interval of (\$491, \$3097). Similar to before, we obtain evidence that the NSW program has a positive effect on increasing earnings.

[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.

(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 3) 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 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.

Answer: See Figure 4. The sampling distribution from predictive resampling approximates the posterior obtained analytically from before.

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

*Answer: The predictive resampling procedure still approximates the same posterior. Intuitively, this is because our estimand is **conditional** on T, so the sampling distribution is unaffected by changes in the marginal distribution of T. This is only true to an extent, however. If T is modified to take the value 0 with probability 1, then the conditional distribution for $Y \mid T = 1$ is no longer defined. Hence, we require the marginal probability of T to lie within $(0,1)$ (this is the overlap assumption).*

(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.*

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

Answer: Due to the independence structure, we can sample θ_{SATE} as follows: first, draw θ from its posterior (2); next, draw all of the missing potential outcomes independently from (1); finally, combine all the observed and imputed potential outcomes to compute θ_{SATE}. The posterior densities for the ATE and SATE are compared in Figure 5. We can see that the SATE posterior is centred at the same mean but has much lower variance. This makes intuitive sense because there is much less uncertainty in the finite, observed population $(i = 1, \ldots, n)$ than there is in the superpopulation $(i = n+1, n+2, \ldots)$.

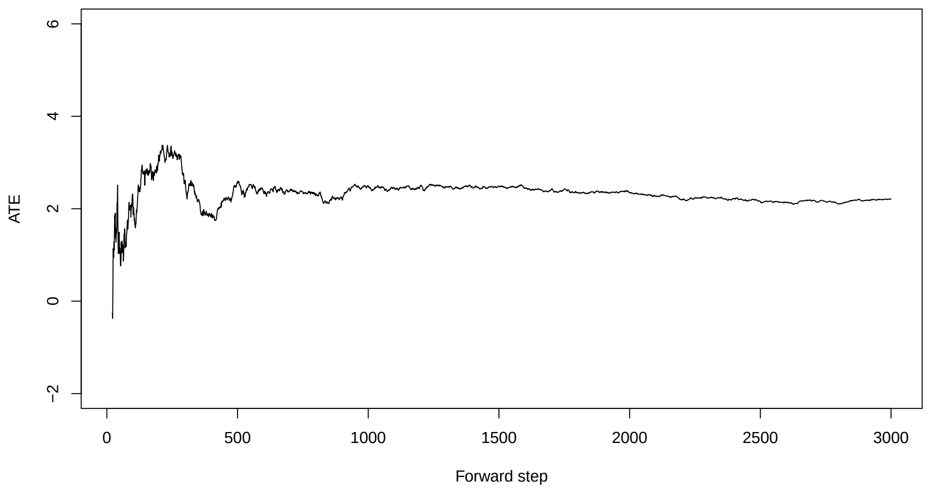

Figure 3: Single iteration of predictive resampling for the average treatment effect.

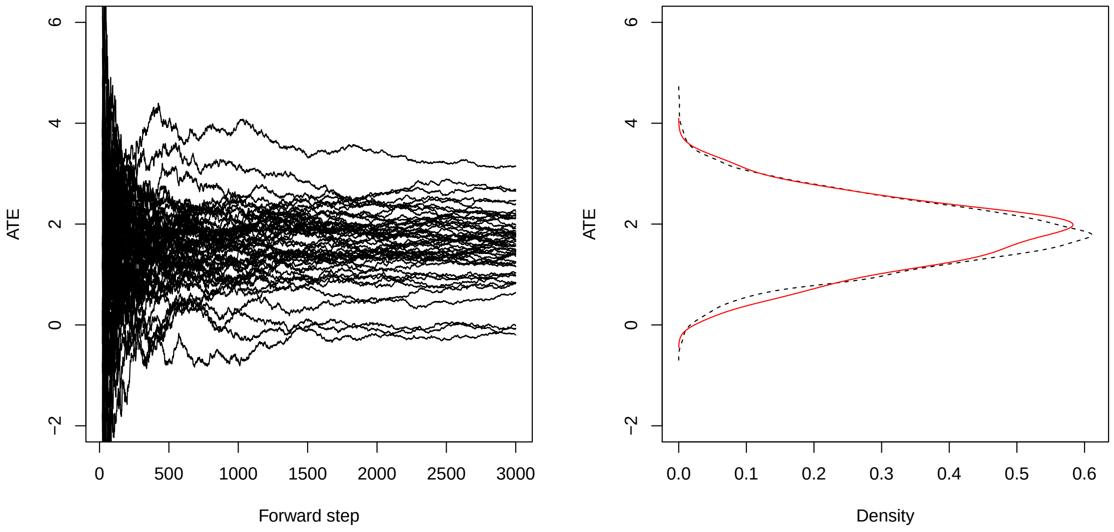

Figure 4: (Left) Multiple iterations of predictive resampling for the Imbens-Rubin model. (Right) Comparison of densities: (red) predictive resampling; (black dashed) exact Bayesian posterior.

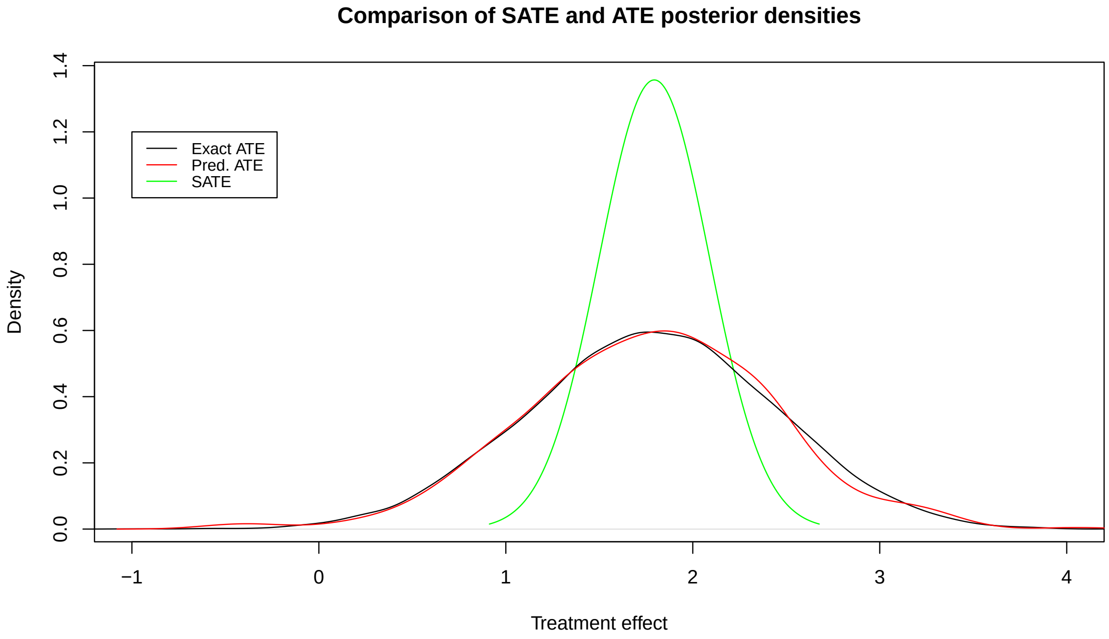

Figure 5: Comparison of posterior densities for the ATE and SATE.

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).

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 the NHANES dataset. 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).

Answer: Conditional on a participant's physical activity and age, the distribution of their remaining lifespan is the same regardless of whether their physical activity was observed or intervened on.

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 1.

Algorithm 1: 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 $\{X^{(1)}_{n+1:n+M}, \ldots, X^{(B)}_{n+1:n+M}\}$.
```

(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 3.

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:

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

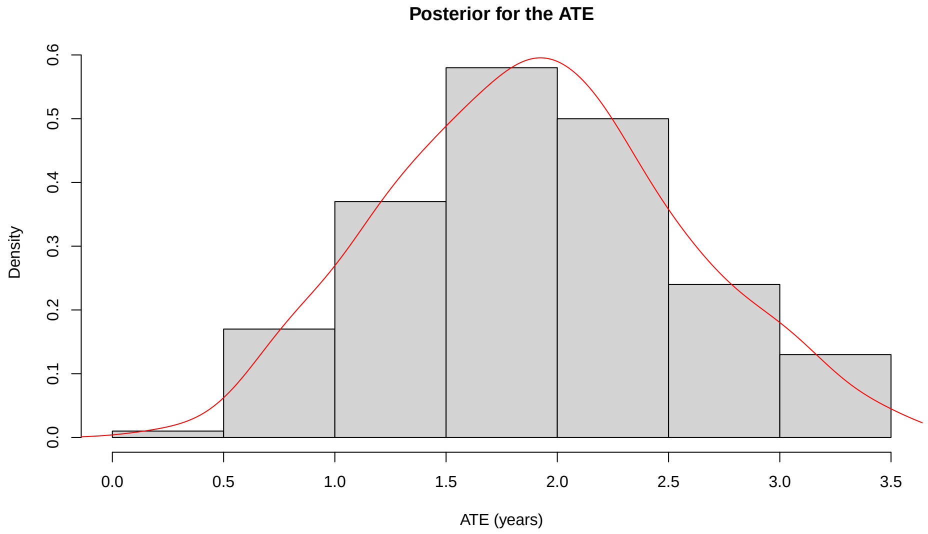

Figure 6: Histogram and density estimator for the average treatment effect for the NHANES dataset.

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.

(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.

Answer: See Figure 6. The analysis suggests (unsurprisingly!) that being physically active has a positive effect of increasing lifespan. The average causal effect is estimated to be around 2 (extra) years.

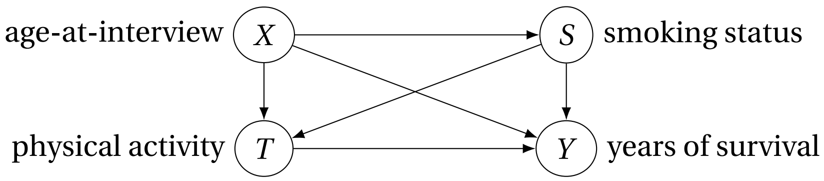

Figure 7: Causal diagram representing the assumed causal relationships for the NHANES dataset if we include smoking status.

(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.

Answers: The assumption may not be realistic as we might expect there to be other factors that both cause a participant to be more (or less) physically active and affect their lifespan. For example, those who do not smoke or have quit smoking may be more likely to be conscious of keeping fit through exercise, and smoking less is likely to have a direct causal impact on increasing life expectancy. Without conditioning on smoking status, we might then introduce spurious correlations that bias our conclusions. The new causal diagram is given by Figure 7. If we rerun the analysis and compare with before, we get the plot in Figure 8. Reassuringly, the densities are quite similar, but the density for the augmented model possibly suggests a slightly smaller magnitude for the ATE; perhaps the inclusion of smoking status in our model has removed some spurious correlations.

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)).

Answer:

$$\begin{aligned} \mathit{var}(Y_i^1 \mid X_i) &= \mathbb{E}[\{Y_i^1 - \mathbb{E}[Y_i^1 \mid X_i]\}^2 \mid X_i] \\ &= \mathbb{E}[\{Y_i^1 - \mathbb{E}[Y_i^1 \mid X_i]\}^2 \mid T_i = 1, X_i] \quad \textit{(unconfoundedness and overlap)} \\ &= \mathbb{E}[\{Y_i - \mathbb{E}[Y_i^1 \mid X_i]\}^2 \mid T_i = 1, X_i] \quad \textit{(consistency)} \end{aligned}$$

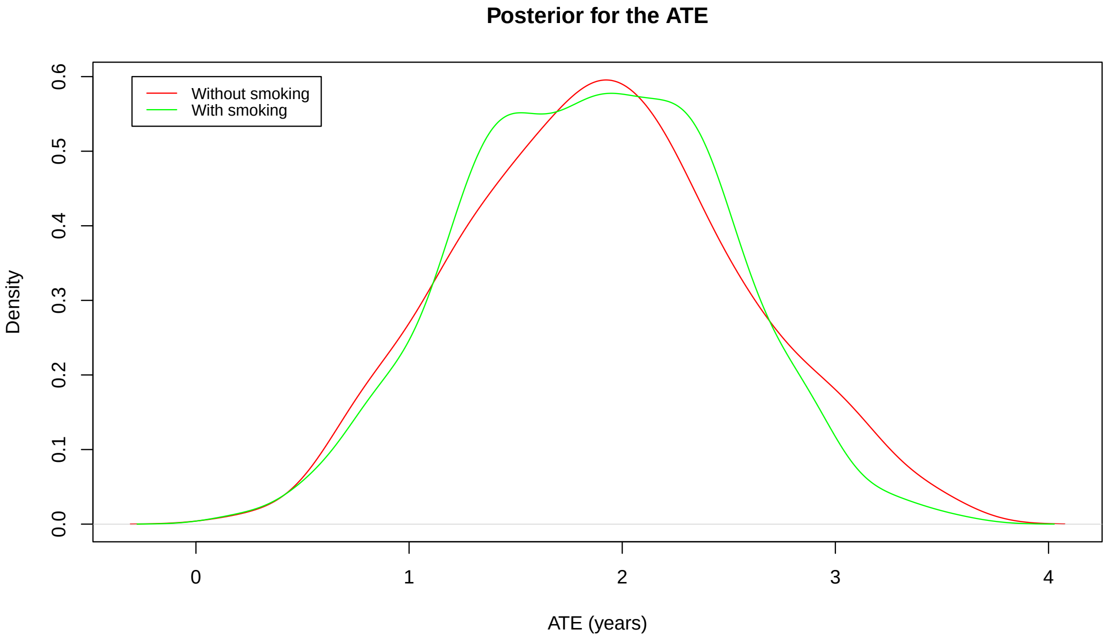

Figure 8: Comparison of densities for the ATE with and without smoking status as a confounder.

We already showed how to identify $\mathbb{E}[Y_i^1 \mid X_i]$ *in the lectures:*

$$\begin{aligned}\mathbb{E}[Y_i^1 \mid X_i] &= \mathbb{E}[Y_i^1 \mid T_i = 1, X_i] \quad \textit{(unconfoundedness and overlap)} \\ &= \mathbb{E}[Y_i \mid T_i = 1, X_i] \quad \textit{(consistency)}\end{aligned}$$

Putting every together, we have

$$var(Y_i^1 \mid X_i) = var(Y_i \mid T_i = 1, X_i).$$

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 $(w_1, \ldots, w_n)$. An equivalent way of sampling from the uniform Dirichlet distribution $\mathrm{Dir}(1, \ldots, 1)$ is to draw a set of i.i.d. $\mathrm{Exp}(1)$ variables $(q_1, \ldots, q_n)$ 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 $\{x_1, \ldots, x_n\}$ 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 2.

Algorithm 2: 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 $X_{n+1:n+M}^{(b)} \sim$ i.i.d. $\sum_{i=1}^{n} w_i \delta_{X_i}$;
end
Return $\{X_{n+1:n+M}^{(1)}, \ldots, X_{n+1:n+M}^{(B)}\}$.
```

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.

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.