Estimating the Effect of Maternal Smoking on Newborn Baby's Birth Weight

Máté Kormos

Supervised Machine Learning
Final Assignment, 1 March 2020

1. Introduction

WHO strongly recommends that pregnant women do not smoke (WHO et al., 2019) as tobacco smoking is known to have adverse health effects. In addition to direct effects on the smoking mother, such as increased risk of coronary heart disease or lung cancer, smoking leads to low birth weight of the newborn baby (WHO, 2019). In this report, I quantify this effect by answering the following research question. *How large is, on average, the effect of maternal smoking on the newborn baby's birth weight?*

2. Data

Hypothetically, the golden standard of answering the research question is based on experimental data, randomising mothers to smoking and non-smoking conditions. Needless to say, such an experiment is unethical and illegal. Therefore, I rely on observational data collected by the North Carolina State Center Health Services (NCSCHS) on first-time Caucasian mothers. I use a randomly chosen subset, including $n = 60,000$ observations, of the data set in Abrevaya et al. (2015), which was generously provided by Robert P. Lieli.

The data set contains information on the newborn baby's weight and whether the mother smokes or not. Covariates indicating the socioeconomic and health status of the mother, and neighbourhood-level economic measures are also recorded. The variables are described in Table 1 (source: Fan et al. (2019)), and summary statistics, grouped by mother's smoking, are presented in Table 2.

Table 2 suggests that the birth weight of smoking mothers' babies tends to fall short that of the nonsmoker mothers', by 240 grams on average. To formally test independence between mother's smoking and baby's birth weight, I perform a permutation test with $H_0 : \textit{bweight} \perp\!\!\!\perp \textit{smoke}$ against $H_1 : \textit{bweight} \not\perp\!\!\!\perp \textit{smoke}$. The test statistics used is the Pearson correlation coefficient between *bweight* and *smoke*, with 2000 random permutations to simulate its distribution under H_0. The test results in a test statistics of 1.0000, with corresponding p-value 0.0000 , hence we can reject independence at every conventional size.

However, the composition of smoker and nonsmoker mothers differs – a clear sign of the lack of randomisation and thus potential confounding. Smoking mothers appear to be younger, less

well-educated, non-married, they have more previously terminated pregnancies, and come from poorer, rural/small town neighbourhoods (as it could be inferred from *popdens*). Therefore, in answering the research question, I will use methods which adjust for confounding as much as possible given this data set.

Table 1: Description of Variables in the NCSCHS Data Set

Name	Support	Description
bweight	$\mathbb{R}_+$	birthweight of newborn baby
smoke	$\{0,1\}$	does the mother smoke?
mage	$\mathbb{Z}_+$	mother's age in years
meduc	$\mathbb{Z}_+$	mother's education in years
married	$\{0,1\}$	is the mother married?
terms	$\{0,1\}$	does mother have previous terminated pregnancies?
fagemiss	$\{0,1\}$	is father's age missing?
feducmiss	$\{0,1\}$	is father's education missing?
anemia	$\{0,1\}$	did mother suffer from anemia?
hyperpr	$\{0,1\}$	did mother suffer from hyperextension?
medinc	$\mathbb{Z}_+$	median income in mother's zip code in $
pcinc	$\mathbb{Z}_+$	per capita income in mother's zip code in $
long	$\mathbb{Z}_+$	longitude of mother's zip code
lat	$\mathbb{Z}_+$	latitude of mother's zip code
popdens	$\mathbb{R}_+$	population density in mother's zip code (units/km^2)

3. Method

3.1. Identification

To estimate the *causal* effect of mother's smoking on the baby's birth weight, I rely on the Rubin Causal Model (Rubin, 1974), the standard framework for causal inference. The estimand of interest is the average treatment effect (ATE), defined as

$$\theta := \mathbb{E}\left[Y(1) - Y(0)\right]$$

where the counterfactual outcomes $Y(0), Y(1)$ describe the baby's weight in two scenarios: $Y(1) \in \mathbb{R}_+$ is the baby's birth weight if his/her mother smokes, and $Y(0) \in \mathbb{R}_+$ is the baby's weight when the mother does not smoke. Clearly, only one scenario is observable. Let $D \in \{0,1\}$ denote the treatment: $D = 1$ if and only if the mother actually smokes. Then the observable birth weight is $Y = Y(0) + (Y(1) - Y(0))D$.

Identification of ATE from observational data requires the conditional independence assumption (CIA; also called unconfoundedness) to hold. Under CIA, if we condition on a set of

Table 2: Summary Statistics by Mother's Smoking

Variable	Count	Minimum	Median	Maximum	Mean	Standard Deviation
bweight nonsmokers	49669.0	255.1	3430.3	5669.9	3396.3	576.8
bweight smokers	10331.0	198.4	3203.5	5216.3	3156.5	574.5
mage nonsmokers	49669.0	13.0	25.0	46.0	25.4	5.6
mage smokers	10331.0	13.0	21.0	49.0	22.7	5.3
meduc nonsmokers	49669.0	0.0	13.0	17.0	13.6	2.3
meduc smokers	10331.0	0.0	12.0	17.0	11.6	1.9
married nonsmokers	49669.0	0.0	1.0	1.0	0.8	0.4
married smokers	10331.0	0.0	1.0	1.0	0.5	0.5
terms nonsmokers	49669.0	0.0	0.0	1.0	0.2	0.4
terms smokers	10331.0	0.0	0.0	1.0	0.3	0.4
fagemiss nonsmokers	49669.0	0.0	0.0	1.0	0.1	0.3
fagemiss smokers	10331.0	0.0	0.0	1.0	0.2	0.4
feducmiss nonsmokers	49669.0	0.0	0.0	1.0	0.1	0.3
feducmiss smokers	10331.0	0.0	0.0	1.0	0.2	0.4
anemia nonsmokers	49669.0	0.0	0.0	1.0	0.0	0.1
anemia smokers	10331.0	0.0	0.0	1.0	0.0	0.1
hyperpr nonsmokers	49669.0	0.0	0.0	1.0	0.1	0.3
hyperpr smokers	10331.0	0.0	0.0	1.0	0.1	0.2
medinc nonsmokers	49669.0	13750.0	38575.0	106022.0	41351.1	11881.8
medinc smokers	10331.0	13750.0	36427.0	106022.0	37749.2	8583.4
pcinc nonsmokers	49669.0	7981.0	18632.0	71301.0	20773.1	6558.2
pcinc smokers	10331.0	7981.0	17781.0	67484.0	18836.3	4616.4
long nonsmokers	49669.0	75.6	79.8	84.1	79.7	1.6
long smokers	10331.0	75.6	80.0	84.1	79.9	1.6
lat nonsmokers	49669.0	33.9	35.6	36.5	35.5	0.5
lat smokers	10331.0	33.9	35.6	36.5	35.5	0.5
popdens nonsmokers	49669.0	4.1	340.3	5543.4	698.6	840.8
popdens smokers	10331.0	4.1	253.0	5543.4	526.9	707.6

observable pre-treatment covariates ($\boldsymbol{x} \in \mathbb{R}^p$), the treatment ($D$) is as good as random, that is $(Y(1), Y0)) \perp\!\!\!\perp D \mid \boldsymbol{x}$. In our example, CIA means that if a smoker and nonsmoker mother have the same covariates, then there is nothing systematically different between them that affects the birth weight. CIA implicitly introduces the second required identifying assumption, the overlap (OL), which asserts that the treated group can be compared to the control group based on $\boldsymbol{x}$.

OL states that $\text{supp}(\boldsymbol{x} \mid D = 1) = \text{supp}(\boldsymbol{x} \mid D = 0)$, i.e. the covariates have the same support in both groups. As opposed to CIA, this is a testable assumption but I present no tests in this report.

I take all the variables in Table 1 (except *bweight* and *smoke*) as covariates, and, in addition, include their powers and interactions up to third degree, leading to $p = 403$ covariates.
CIA and OL provides the identification strategy:

$$\begin{aligned}\theta &= \mathbb{E}\left[Y(1) - Y(0)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y(1) \mid \boldsymbol{x}\right] - \mathbb{E}\left[Y(0) \mid \boldsymbol{x}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y(1) \mid \boldsymbol{x}, D\right] - \mathbb{E}\left[Y(0) \mid \boldsymbol{x}, D\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y \mid \boldsymbol{x}, D = 1\right] - \mathbb{E}\left[Y \mid \boldsymbol{x}, D = 0\right]\right].\end{aligned}$$

Let $g(D, \boldsymbol{x}) \coloneqq \mathbb{E}\left[Y \mid \boldsymbol{x}, D\right]$, so that $\theta = \mathbb{E}\left[g(1, \boldsymbol{x}) - g(0, \boldsymbol{x})\right]$. Let $m(\boldsymbol{x}) \coloneqq \mathbb{E}\left[D \mid \boldsymbol{x}\right]$ denote the propensity score.

3.2. Estimation

I estimate the ATE based on the procedure in Chernozhukov et al. (2016) (see Algorithm 1). Specifically, the nuisance parameters $\eta \coloneqq (m, g)$ are estimated with machine learning (ML) methods. The advantage of ML methods is that they are capable of learning complex relationships from the data. However, they are, in general, asymptotically biased because of the regularisation. Chernozhukov et al. (2016) addresses this issue by proposing the use of Neyman orthogonality scores and cross-fitting to ensure asymptotically unbiased ATE estimators for a large class of ML methods including ridge, lasso, regression and classification trees.

Algorithm 1 DML1 Estimation Strategy (Chernozhukov et al. (2016), p. 23, 35)

input: *i.i.d.* sample $\boldsymbol{w}_i = (Y_i, D_i, \boldsymbol{x}_i')'$ for $i = 1, 2, \dots N$

output: estimated ATE

1: $(I_k)_{k=1}^K \leftarrow$ random partition of $\{1, 2, \dots, N\}$ s.t. $n \coloneqq |I_k| = N/K \ \forall k \in \{1, 2, \dots, K\}$
2: $(I_k^c)_{k=1}^K \leftarrow \{1, 2, \dots, N\} \setminus I_k \ \forall k \in \{1, 2, \dots, K\}$
3: $\psi(\boldsymbol{w}; \theta, \eta) \leftarrow g(1, \boldsymbol{x}) - g(0, \boldsymbol{x}) + \frac{D(Y - g(1, \boldsymbol{x}))}{m(\boldsymbol{x})} + \frac{(1-D)(Y - g(0, \boldsymbol{x}))}{1 - m(\boldsymbol{x})}$ ▷ Neyman orthogonal score for ATE
4: **for** $k \in \{1, 2, \dots, K\}$ **do**
5: $\quad \hat{\eta}_k = (\hat{m}_k, \hat{g}_k) \leftarrow \hat{\eta}((\boldsymbol{w}_i)_{i \in I_k^c})$ ▷ cross-fitting ML estimators
6: $\quad \psi_k(\boldsymbol{w}) \leftarrow \hat{g}_k(1, \boldsymbol{x}) - \hat{g}_k(0, \boldsymbol{x}) + \frac{D(Y - \hat{g}_k(1, \boldsymbol{x}))}{\hat{m}_k(\boldsymbol{x})} + \frac{(1-D)(Y - \hat{g}_k(0, \boldsymbol{x}))}{1 - \hat{m}_k(\boldsymbol{x})}$
7: $\quad$ compute $\hat{\theta}_k$ s.t. $\mathbb{E}_{n,k}\left[\psi(\boldsymbol{w}; \hat{\theta}_k, \hat{\eta}_k)\right] = n^{-1} \sum_{i \in I_k} \psi_k(\boldsymbol{w}_i) = 0$
8: $\hat{\theta} \leftarrow K^{-1} \sum_{k=1}^K \hat{\theta}_k$
9: **return** $\hat{\theta}$

Estimation of m_k. The estimate is obtained by training a lasso logit on the data $(\boldsymbol{w}_i)_{i \in I_k^c}$. I use lasso to shrink the parameters, thereby selecting the covariates that are good predictors of the

treatment. As D is binary, I choose the logit based loss function to ensure that the predicted propensity score is in $[0, 1]$. Let $\boldsymbol{x}_*$ denote the standardised covariate vector[1], then

$$\begin{aligned}
\hat{m}_k(\boldsymbol{x}) &\coloneqq S(\tilde{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_k) \\
\tilde{\boldsymbol{x}} &\coloneqq (1, \boldsymbol{x}_*')' \\
S(z) &\coloneqq (1 + \exp(-z))^{-1} \\
\hat{\boldsymbol{\beta}}_k &\coloneqq \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} -n^{-1}\sum_{i\in I_k^c}[D_i \log(S(\tilde{\boldsymbol{x}}_i'\boldsymbol{\beta})) + (1 - D_i)\log(1 - S(\tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}))] + \lambda_k^* \left\|\boldsymbol{\beta}^{(-1)}\right\|_1
\end{aligned}$$

for $k = 1, 2 \ldots, K$, where $\boldsymbol{\beta}^{(-1)}$ is the coefficient vector excluding the coordinate of $\boldsymbol{\beta}$ corresponding to the constant. λ_k^* is the misclassification rate-optimal penalty selected by 10-fold cross-validation on I_k^c from the set $\{10^j\}_{j=-1,0,\ldots,7,8}$.

Estimation of g_k. The estimate is obtained by bagging a regression tree, capturing potentially complex dependencies between the outcome variable Y and $(D, \boldsymbol{x}')'$. Bagging reduces variance by resampling the data and aggregating the predictions across the samples. Formally,

$$\begin{aligned}
\hat{g}_k(D, \boldsymbol{x}) &\coloneqq B^{-1}\sum_{b=1}^{B} \hat{T}_k^b(D, \boldsymbol{x}) \\
\hat{T}_k^b(D, \boldsymbol{x}) &\coloneqq \sum_{l\in\mathcal{L}_k^b} \mathbb{1}_{(D,\boldsymbol{x}')'\in l_k}(D, \boldsymbol{x}) \sum_{i\in I_k^c:(D_i,\boldsymbol{x}_i')'\in l} \frac{Y_i}{|\{j \in I_k^c : (D_j, \boldsymbol{x}_j')' \in l\}|}, \quad b = 1, 2 \ldots, B
\end{aligned}$$

for $k = 1, 2 \ldots, K$ where $\mathbb{1}_{\cdot}(.)$ is the indicator function, taking on value 1 if $(D, \boldsymbol{x}')' \in l$, 0 otherwise, $\mathcal{L}_k^b$ is the set of terminal nodes in the bth bootstrap sample.[2] That is, we take the sample average of observations in I_k^c that are in the same node as $(D, \boldsymbol{x}')'$. The regression trees are all fitted with the parameters *maximum depth* set to 15, *minimum number of samples in terminal nodes* set to 500. The parameters are chosen to avoid overfitting and to ensure that there is a large number of observations in the terminal nodes so that the variance of the average in the node is reduced. I set the number of bootstrap samples to $B = 50$. Following the recommendation of Chernozhukov et al. (2016), K is set to 5.

Inference Chernozhukov et al. (2016) establish asymptotic normality of $\hat{\theta}$, so that the confidence interval

$$\begin{aligned}
\mathrm{CI}_{1-\alpha} &\coloneqq \left[\hat{\theta} \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\hat{\sigma}^2/N}\right] \\
\hat{\sigma}^2 &\coloneqq K^{-1}\sum_{k=1}^{K} \hat{\sigma}_k^2 \\
\hat{\sigma}_k^2 &\coloneqq \mathbb{E}_{n,k}\left[\psi(\boldsymbol{w}, \hat{\theta}, \hat{\eta}_k)^2\right]
\end{aligned}$$

has a uniform coverage of $100(1 - \alpha)\%$.

[1] Note that to standardise $(\boldsymbol{x})_{i\in I_k^c}$, only data in I_k^c are used.

[2] In each bootstrap sample, I randomly select $|I_k^c|$ observations from I_k^c and fit the tree T_k^b using the selected observations for the training.

4. Results

Applying Algorithm 1 results in the point estimates $(\hat{\theta}_k)_{k=1}^K$ = (-249.55, -214.07, -238.84, -231.04, -222.90). Hence, the final point estimate is $\hat{\theta} = -231.2812$. Note that this is a smaller effect than what was found by simply comparing birth weight sample averages for smoker and nonsmoker mothers (≈-240 grams). However, the difference between the effect estimates is not large, which becomes even more apparent after quantifying estimation uncertainty in $\hat{\theta}$.

Estimation of the variance of $\hat{\theta}$ results in the point estimates $(\hat{\sigma}_k^2)_{k=1}^K$ = (6.74e+09, 5.72e+09, 6.10e+09, 6.13e+09, 6.33e+09). The variance estimate is then $\hat{\sigma^2} = 6.2037e + 09$. Hence the 95% confidence interval is [-861.5218, 398.9594] grams. That is, at size 5% we do not reject the two-sided hypothesis of zero, or even positive, average treatment effect.

The main limitation of the report is the validity of CIA. It may well be the case that we do not observe enough covariates for CIA to hold. Unconfoundedness would be more credible if we observed more detailed data on smoking and health status.

5. Conclusion

I estimated the causal effect of the mother's smoking on her newborn baby's birth weight with machine learning methods proposed by Chernozhukov et al. (2016), using an observational data set comprising 60,000 first-time Caucasian mothers in North Carolina, USA. The point estimate of the average treatment effect is −231.2812 grams indicating an adverse health effect for the baby. However, the point estimate has a large variance: the 95% confidence interval for the average treatment effect is [-861.5218, 398.9594] grams. Thus, in conclusion, we cannot reject the null hypothesis that maternal smoking has zero average treatment effect on the birth weight.

6. Codes

Codes are available at `https://github.com/kmmate/SML_assignment`

References

Jason Abrevaya, Yu-Chin Hsu, and Robert Lieli. Estimating Conditional Average Treatment Effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015. doi: 10.1080/07350015.2014.975555.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters, 2016.

Qingliang Fan, Yu-Chin Hsu, Robert P. Lieli, and Yichong Zhang. Estimation of Conditional Average Treatment Effects with High-Dimensional Data, 2019.

Donald B Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688, 1974.

WHO. Fact sheets - Tobacco, 2019. URL `https://www.who.int/news-room/fact-sheets/detail/tobacco`.

WHO, United Nations Population Fund, and UNICEF. Pregnancy, Childbirth, Postpartum and Newborn Care. A Guide for Essential Practice, 2019. URL `https://www.who.int/maternal_child_adolescent/documents/imca-essential-practice-guide/en/`.