Double Robustness and Sequential Causal Inference

Robin J. Evans and Vanessa Didelez

University of Oxford and BIPS, University of Bremen

APTS Week 4, Glasgow
August 2023

Outline

1. Doubly-Robust Approaches

2. Machine Learning Methods
   - Post Double Selection Inference
   - Double Machine Learning

3. Graphical Sequential Treatment
   - Multiple Regression
   - Multiple Treatments

Doubly Robust Approaches

Note we've seen that if we specify

- the **outcome model** (i.e. $Y \mid A, \boldsymbol{X}$) correctly, we can obtain a consistent estimate of the ACE by averaging over the empirical $\boldsymbol{X}$ values;
- the **propensity score model** (i.e. $A \mid \boldsymbol{X}$) correctly, we can use the Horvitz-Thompson estimator which is also consistent.

Is there an estimator that uses both of these models, but only requires one of them to be correct?

Yes!

We can use the following approach: suppose we believe that

$$\mathbb{E}[Y \mid a, \boldsymbol{x}] = Q_a(\boldsymbol{x}; \beta, \gamma) \qquad \text{and} \qquad \pi(\boldsymbol{x}) = \pi(\boldsymbol{x}; \eta)$$

for **parametric** models Q_0, Q_1, and π.

These are sometimes called **working models**.

Doubly Robust Methods

Notice that the following function has expectation $Y(1)$ if **either** Q_1 or π is specified correctly:

$$\mu_1^{dr}(O) = Q_1(\boldsymbol{X}) + \frac{A}{\pi(\boldsymbol{X})}\{Y - Q_1(\boldsymbol{X})\}$$
$$= \frac{AY}{\pi(\boldsymbol{X})} + \left\{1 - \frac{A}{\pi(\boldsymbol{X})}\right\} Q_1(\boldsymbol{X}).$$

So fit 'nuisance' models Q and π to the data (e.g. by maximum likelihood). This gives parameter estimates $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\eta}$.

Then consider the following estimator of $\mathbb{E}Y(1)$:

$$\hat{\mu}_1^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i\{Y_i - Q_{A_i}(\boldsymbol{X}_i;\hat{\beta},\hat{\gamma})\}}{\pi(\boldsymbol{X}_i;\hat{\eta})} + Q_1(\boldsymbol{X}_i;\hat{\beta},\hat{\gamma})\right\}.$$

If **either** model is correctly specified, then by the above we can see that the estimate will be consistent.

This property is called **double robustness**.

Doubly Robust Methods

We can do something similar for $\hat{\mu}_0^{dr}$, and then

$$\hat{\beta}^{dr} := \hat{\mu}_1^{dr} - \hat{\mu}_0^{dr}. \qquad (*)$$

We call this the **augmented** inverse probability weighted estimator (AIPW).

In addition, each $\hat{\mu}_a^{dr}$ is **semi-parametric efficient** if both parametric models are correct, so it achieves the same rate (asymptotically) as maximum likelihood estimation.

If Q_a is wrong then MLEs will be difficult to interpret.

In practice, even under moderate misspecifications of both models, the doubly robust estimator mostly performs well in practice.

Demonstrations

The `R` package `causl`[1] allows one to simulate data from a parametrically specified causal model.

Suppose we want to have:

$$Z \sim \text{Exponential}(\lambda)$$
$$A \mid Z = z \sim \text{Bernoulli}\left(\text{logit}(\alpha_0 + \alpha_1 z)\right)$$
$$Y \mid do(A = a) \sim N(\beta a,\ \sigma^2)$$

with $\lambda = 2$, $\alpha_0 = 0$, $\alpha_1 = 1$ and $\beta = 1/2$.

```
library(causl)
forms <- list(Z ~ 1,
              A ~ Z,
              Y ~ A,
              ~ 1)  ## for the copula
```

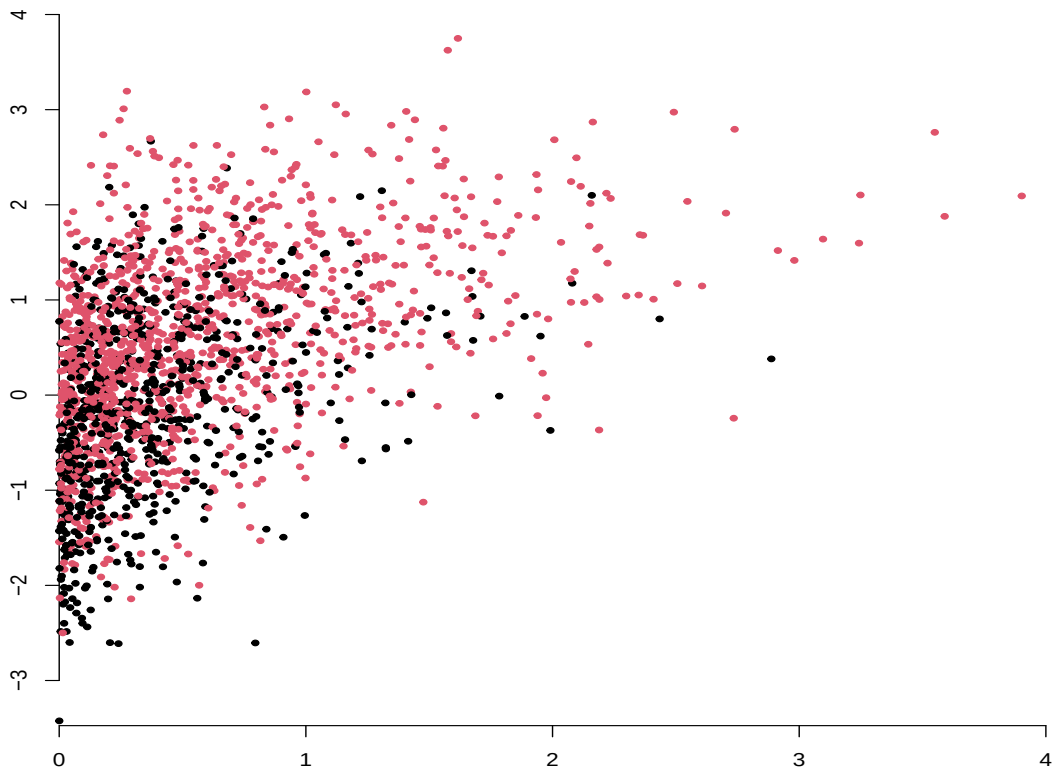
```
pars <- list(Z = list(beta = -log(2), phi=1),  ## we use log-link
             A = list(beta = c(0,1)),
             Y = list(beta = c(0,0.5), phi = 1),
             cop = list(beta = 1))
fam <- list(3,5,1,1)  # distributions: 1=normal, 3=Gamma, 5=binomial
```

[1] https://github.com/rje42/causl

Demonstrations

We can then use the `causalSamp` function to simulate our data:

```
set.seed(123)
dat <- causalSamp(1e4, formulas=forms, pars=pars, family=fam)
```



The plot shows the first 2000 data points.

Doubly Robust Methods

Let us suppose that $\mathbb{E}Y$ is linear in A and $\boldsymbol{X}$ separately, so

$$\mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = \beta_A a + \beta_{\boldsymbol{X}} \boldsymbol{x}.$$

```
# get propensity score
ps <- fitted(glm(A ~ Z, data=dat, family="binomial"))
dat <- dplyr::mutate(dat, ps = ps)  # add est. propensity score

# outcome model
modY <- lm(Y ~ A + Z, data=dat)

dat0 <- dat1 <- dat    ## set 0 and 1 in mock datasets
dat0$A <- 0; dat1$A <- 1

## compute mu_x for x = {0,1}
mu1 <- mean(dat$A*(dat$Y - predict(modY))/dat$ps
            + predict(modY, dat1))

mu0 <- mean((1-dat$A)*(dat$Y - predict(modY))/(1-dat$ps)
            + predict(modY, dat0))

mu1 - mu0


## [1] 0.5153306
```

References

Bang, H. and Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), pp.962-973, 2005.

Moore and van der Laan. Covariate adjustment in randomized trials with binary outcomes, *Statistics in Medicine*, 2008.

Robins, J.M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122-129, 1995.

Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55, 1983.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448), pp.1096-1120. 1999.

Outline

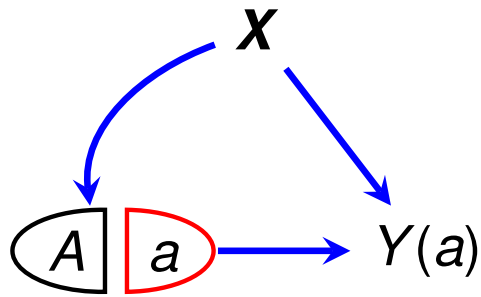
1. Doubly-Robust Approaches

2. Machine Learning Methods
   - Post Double Selection Inference
   - Double Machine Learning

3. Graphical Sequential Treatment
   - Multiple Regression
   - Multiple Treatments

Post 'Double Selection' Inference

Suppose we have the following set up, where $\boldsymbol{X}$, is high-dimensional (say $|\boldsymbol{X}| = p$).



It is clear that we can **identify** the causal effect of A on Y, since assuming independent observations and the model implied by the SWIG:

$$\mathbb{E}\,Y(a) \;=\; \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot \mathbb{E}[\,Y \mid a, \boldsymbol{x}] \;=\; \mathbb{E}\left[\frac{Y\,\mathbb{1}_{\{A=a\}}}{P(A = a \mid \boldsymbol{X})}\right];$$

however, statistically we may still have difficulties.

- We do not know what form the expressions for $\mathbb{E}[\,Y \mid a, \boldsymbol{x}]$, $P(\boldsymbol{x})$, or $P(a \mid \boldsymbol{x})$ should take.
- Even if we knew the families, actually estimating the parameters may be infeasible with a finite dataset of reasonable size.

Frisch-Waugh-Lovell Theorem

Suppose we have n i.i.d. observations $(\boldsymbol{X}_i, A_i, Y_i)$ such that

$$A_i = \alpha^T \boldsymbol{X}_i + \delta_i \qquad\qquad Y_i = \beta A_i + \gamma^T \boldsymbol{X}_i + \varepsilon_i,$$

where $\boldsymbol{X}_i$ has fewer than $n - 1$ entries.

Consider two different ways of obtaining an estimate of β:

1. regress Y on $\boldsymbol{X}$ and A using OLS, and look at $\hat{\beta}$;
2. regress Y on $\boldsymbol{X}$ to obtain residual r_Y; and then A on $\boldsymbol{X}$ to obtain r_A; then regress r_Y on r_A, and take the linear coefficient $\tilde{\beta}$.

Theorem (Frisch and Waugh (1933), Lovell (1963))

The estimates for β from methods 1 and 2 are the same.

Intuition

Why does this result hold?

Proof.

Note that $r_A = A - \hat{\alpha}^T \boldsymbol{X}$, so $r_A \perp\!\!\!\perp \boldsymbol{X}$.
Then

$$\begin{aligned}\mathbb{E}[Y \mid \boldsymbol{X}, A] &= \beta A + \gamma^T \boldsymbol{X} \\ &= \beta(r_A + \alpha^T \boldsymbol{X}) + \gamma^T \boldsymbol{X} \\ &= \beta r_A + (\alpha + \gamma)^T \boldsymbol{X}.\end{aligned}$$

Then, since $\boldsymbol{X} \perp\!\!\!\perp r_A$, we must have that regressing Y on $\boldsymbol{X}$ gives an estimate of $\alpha + \gamma$.
Hence

$$\mathbb{E} r_Y = \beta \mathbb{E} r_A,$$

giving the result. □

Sparsity

Suppose that we have

$$\mathbb{E}[A \mid \boldsymbol{X} = \boldsymbol{x}] = \alpha^T \boldsymbol{x}$$
$$\mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = \beta a + \gamma^T \boldsymbol{x}.$$

Assume also that $\log p = o(n^{1/3})$ and there exist subsets $\boldsymbol{B}$ and $\boldsymbol{D}$ of size at most $s_n \ll n$ such that:

$$\mathbb{E}[A \mid \boldsymbol{x}] = \alpha_{\boldsymbol{B}}^T \boldsymbol{x} + r_n$$
$$\mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = \beta a + \gamma_{\boldsymbol{D}}^T \boldsymbol{x} + t_n,$$

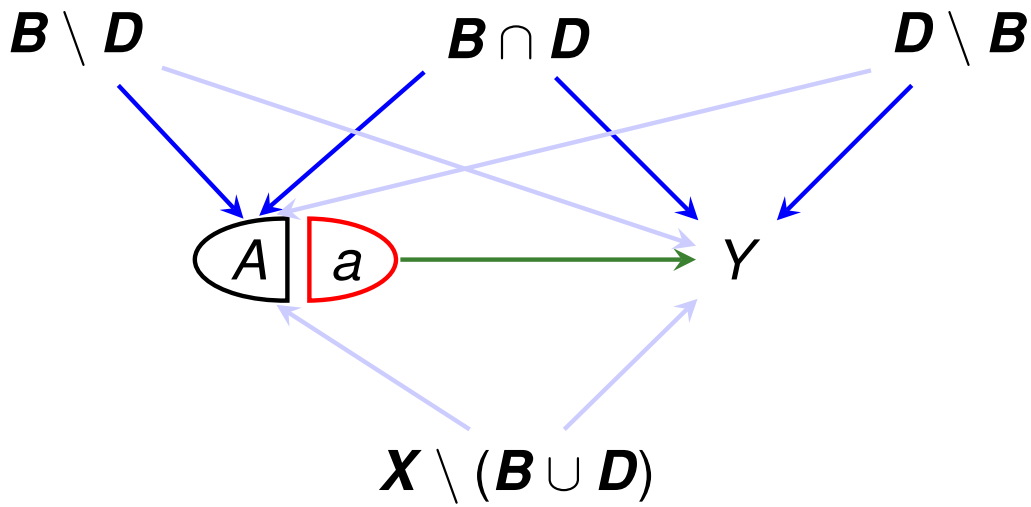
where the approximation error is stochastically smaller than the estimation error: i.e.

$$\mathbb{E}\|r_n\|_2 \lesssim \sqrt{\frac{s_n}{n}} \qquad \text{and} \qquad \mathbb{E}\|t_n\|_2 \lesssim \sqrt{\frac{s_n}{n}}.$$

In other words, a much smaller subset of covariates is sufficient to **approximately** make A and Y unconfounded.

Post 'Double Selection' Inference

Graphical representation:



The idea is that if we account for variables in **both $\boldsymbol{B}$ and $\boldsymbol{D}$**, then we will be guaranteed to have good control of the bias in estimating β.

In principle we can use any consistent selection method to choose $\boldsymbol{B}$ and $\boldsymbol{D}$. In practice, Belloni et al. recommend a version of the lasso.

Post 'Double Selection' Inference

Here we perform a simulated example. Suppose that

$$A_i = \alpha \sum_{i=1}^{7} X_i + \delta_i$$

$$Y_i = \beta A_i + \gamma \sum_{i=4}^{10} X_i + \varepsilon_i$$

where $\delta_i, \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ (independently), and we are given 1000 covariates in $\boldsymbol{X}$, where each $X_{ij} \sim N(0,1)$ independently.

Set $\beta = \gamma = 2$ and $\alpha = 1$, and pick $n = 100$.

Post 'Double Selection' Inference

```
alpha <- 1
gamma <- beta <- 2
n <- 100; p <- 1000

## simulate data
set.seed(123)
Z <- matrix(rnorm(n*p), n, p)
X <- Z %*% c(rep(alpha, 7), rep(0,p-7)) + rnorm(n)
Y <- Z %*% c(rep(0,3), rep(gamma, 7), rep(0,p-10)) + beta*X + rnorm(n)
dat <- data.frame(Y=Y, X=X, Z)
names(dat) <- c("Y","X",paste0("Z",seq_len(p)))

head(dat[,1:9])
```

```
##          Y      X      Z1     Z2     Z3     Z4      Z5      Z6
## 1  -1.932  0.876 -0.5605 -0.710  2.199 -0.715 -0.0736 -0.6019  1.07
## 2 -11.460  0.227 -0.2302  0.257  1.312 -0.753 -1.1687 -0.9937 -0.02
## 3   0.821  0.408  1.5587 -0.247 -0.265 -0.939 -0.6347  1.0268 -0.03
## 4  -0.752 -1.633  0.0705 -0.348  0.543 -1.053 -0.0288  0.7511 -1.51
## 5  -4.478 -1.284  0.1293 -0.952 -0.414 -0.437  0.6707 -1.5092  0.79
## 6  -2.355  0.906  1.7151 -0.045 -0.476  0.331 -1.6505 -0.0951 -0.21
```

Post 'Double Selection' Inference

We can try a naïve model, and obtain the wrong answer.

```
sum_lm <- summary(lm(Y ~ X, data=dat))
sum_lm$coef


##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.244      0.492   0.496 6.21e-01
## X              3.067      0.184  16.649 2.52e-30


coef <- sum_lm$coef
```

Notice that the estimate $\hat{\beta}$ = 3.07 is not within 2 s.e.s (0.37) of β = 2.

Post 'Double Selection' Inference

Then we can try using the R package `hdm`, which implements double selection.

```
library(hdm) ## library for implementation
lasso_out = rlassoEffect(y=dat[,"Y",drop=FALSE],
                         d=dat[,"X",drop=FALSE],
                         x=Z, method="double selection")

sum_out <- summary(lasso_out)
sum_out


## [1] "Estimates and significance testing of the effect of target var
##   Estimate. Std. Error t value Pr(>|t|)
## X     2.018      0.119    16.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note this solution $\tilde{\beta} = 2.02$, is (well) within two s.e.s (0.24) of $\beta = 2$.

Post 'Double Selection' Inference: Application

Let us try applying double selection to a wage dataset.

```
X <- model.matrix(~ -1 + female + (widowed +divorced + separated +
                  nevermarried + hsd08 + hsd911 + hsg + cg + ad + mw +
                  we + exp1 + exp2 + exp3)^2, data = cps2012)
X <- X[, apply(X, 2, var) != 0] # exclude all constant variables
y <- cps2012$lnw
effects_female <- rlassoEffects(x = X, y = y, index = "female")
summary(effects_female)


## [1] "Estimates and significance testing of the effect of target var
##        Estimate. Std. Error t value Pr(>|t|)
## female  -0.28067    0.00692   -40.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post 'Double Selection' Inference: Application

Now let's try fitting the other covariates too (note some are causally subsequent to sex).

```
data(cps2012)
X <- model.matrix(~ -1 + female + female:(widowed + divorced + separat
                  nevermarried +hsd08 + hsd911 + hsg + cg + ad + mw +
                  we + exp1 + exp2 + exp3) + (widowed +divorced + sepa
                  nevermarried + hsd08 + hsd911 + hsg + cg + ad + mw +
                  we + exp1 + exp2 + exp3)^2, data = cps2012)
X <- X[, apply(X, 2, var) != 0] # exclude all constant variables
index.gender <- grep("female", colnames(X))
y <- cps2012$lnw
```

Post 'Double Selection' Inference: Application

```
effects_female <- rlassoEffects(x = X, y = y, index = index.gender)
summary(effects_female)


## [1] "Estimates and significance testing of the effect of target var
##                     Estimate. Std. Error t value Pr(>|t|)
## female               -0.15492    0.05016   -3.09  0.00201 **
## female:widowed        0.13610    0.09066    1.50  0.13332
## female:divorced       0.13694    0.02218    6.17  6.7e-10 ***
## female:separated      0.02330    0.05321    0.44  0.66144
## female:nevermarried   0.18685    0.01994    9.37  < 2e-16 ***
## female:hsd08          0.02781    0.12091    0.23  0.81809
## female:hsd911        -0.11934    0.05188   -2.30  0.02144 *
## female:hsg           -0.01289    0.01922   -0.67  0.50252
## female:cg             0.01014    0.01833    0.55  0.58011
## female:ad            -0.03046    0.02181   -1.40  0.16241
## female:mw            -0.00106    0.01919   -0.06  0.95581
## female:so            -0.00818    0.01936   -0.42  0.67247
## female:we            -0.00423    0.02117   -0.20  0.84176
## female:exp1           0.00494    0.00780    0.63  0.52714
## female:exp2          -0.15952    0.04530   -3.52  0.00043 ***
## female:exp3           0.03845    0.00786    4.89  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.

Frisch, R. and F.V. Waugh (1933). Partial time regression as compared with individual trends. *Econometrica* 1 (October): 387–401.

Lovell, M.C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *JASA* 58 (December): 993–1010.

Double Machine Learning

Double (or **debiased**) **machine learning** is an increasingly common approach to estimating causal effects. See, e.g. Chernozhukov et al. (2018).

The basic idea is the same as the approach of Belloni et al. (2014).

We estimate separate **high-dimensional models** for the treatment and outcome.

The methods make extensive use of **cross-fitting**, i.e. splitting the data into separate components and using each to predict the other.

This allows for estimation while preventing **over-fitting**.

Mathematically speaking, much more **complicated models** can be used but still give an unbiased estimator of a (low-dimensional) causal effect.

Conditions for Double ML

A crucial condition for double ML to work is **Neyman orthogonality**, which says that the derivative of the estimating equation (at the true parameters) with respect to any nuisance parameters should be zero.

Suppose our score function is $\psi(W; \theta, \eta)$, with parameters of interest θ and nuisance parameters η. Then we need:

$$\frac{\partial}{\partial \eta} \mathbb{E}\psi(W; \theta_0, \eta)\Big|_{\eta=\eta_0} = 0,$$

where (θ_0, η_0) are the true parameters.

If we are given a score function that is **not** Neyman orthogonal, we can often change it to become so.

Conditions for Double ML

Consider the linear model example, where the usual score is

$$\tilde{\psi}_\beta(W;\beta,\gamma) = (Y - \beta A - \gamma^T \boldsymbol{X}) \cdot A$$
$$\tilde{\psi}_\gamma(W;\beta,\gamma) = (Y - \beta A - \gamma^T \boldsymbol{X}) \cdot \boldsymbol{X}.$$

Suppose we consider a directional derivative $\delta \cdot h$ with $h \in \mathbb{R}^{|\boldsymbol{X}|}$, then we have

$$\begin{aligned}
&\frac{\partial}{\partial \gamma} \tilde{\psi}_\beta(W;\beta,\gamma_0 + \delta h)\Big|_{\delta \to 0} \\
&= \lim_{\delta \to 0} \frac{(Y - \beta A - (\gamma_0 + \delta h)^T \boldsymbol{X}) \cdot A - (Y - \beta A - \gamma_0^T \boldsymbol{X}) \cdot A}{\delta} \\
&= -h^T \boldsymbol{X}.
\end{aligned}$$

In particular, this is **not** zero!

Conditions for Double ML

Now, we can reparametrize the nuisance parameter γ as $\eta = (\gamma, \mu)$, where we choose μ so that the new score for β is

$$\begin{aligned}\psi_\beta(W; \beta, \eta) &= \tilde{\psi}_\beta(W; \beta, \gamma) - \mu^T \tilde{\psi}_\gamma(W; \beta, \gamma) \\ &= (Y - \beta A - \gamma^T \boldsymbol{X})(A - \mu^T \boldsymbol{X}).\end{aligned}$$

If we pick $\mu = \alpha$, then note that the expectation of second factor is 0!

Hence, **small** errors in the estimation of γ and α will **not** affect the estimate of β.

In particular:

$$\frac{\partial}{\partial \gamma}\psi_\beta(W; \beta, \gamma, \alpha) = -\boldsymbol{X}(A - \alpha^T \boldsymbol{X})$$

and $$\frac{\partial}{\partial \alpha}\psi_\beta(W; \beta, \gamma, \alpha) = -\boldsymbol{X}(Y - \beta A - \gamma^T \boldsymbol{X}),$$

and these both have expectation 0.

Moral: Neyman orthogonality is very helpful for robustness to misspecification.

401(k) Example

Chernozhukov et al. (2018) analyse data on 401(k) savings plans, and whether eligibility to enroll leads to an increase in net assets.

They consider a dataset of 9,915 individuals, measuring:

age	age in years;
inc	income;
educ	years of education;
fsize	family size;
marr	indicator of being married;
twoearn	two earners in household;
db	member of defined benefit pension scheme;
pira	eligible for Individual Retirement Allowance;
hown	homeowner.

DML for 401(k) Example

```
library(DoubleML)
library(mlr3)
library(data.table)
library(dplyr)

## note that the DoubleML package uses data.table objects
dat <- fetch_401k(return_type = "data.table", instrument = TRUE)

# Initialize DoubleMLData (data-backend of DoubleML)
dml = DoubleMLData$new(dat,
                       y_col = "net_tfa",
                       d_cols = "e401",
                       x_cols = c("age", "inc", "educ", "fsize",
                        "marr", "twoearn", "db", "pira", "hown"))
mod <- DoubleMLIRM$new(dml,
               ml_m = lrn("classif.cv_glmnet", s = "lambda.min"),
               ml_g = lrn("regr.cv_glmnet",s = "lambda.min"),
               n_folds = 10, n_rep = 10)
mod$fit()    ## fit the model
```

```
c(beta=mod$coef, se=mod$se)


## beta.e401   se.e401
##      1648      3739
```

DML for 401(k) Example

We can also try using a more flexible set of covariates.

```
## add quadratic terms to age, income, education and family size
formula_flex = formula(" ~ -1 + poly(age, 2, raw=TRUE) +
  poly(inc, 2, raw=TRUE) + poly(educ, 2, raw=TRUE) +
  poly(fsize, 2, raw=TRUE) + marr + twoearn + db + pira + hown")
features_flex = data.frame(model.matrix(formula_flex, dat))
model_data = data.table("net_tfa" = dat[, net_tfa],
                        "e401" = dat[, e401], features_flex)

## initialize and fit model
dml_f <- DoubleMLData$new(model_data, y_col = "net_tfa",
                          d_cols = "e401")
mod_f <- DoubleMLIRM$new(dml_f,
              ml_m = lrn("classif.cv_glmnet", s = "lambda.min"),
              ml_g = lrn("regr.cv_glmnet",s = "lambda.min"),
              n_folds = 10, n_rep = 5)
mod_f$fit()
```

We obtain a much smaller standard error.

```
c(beta=mod_f$coef, se=mod_f$se)


## beta.e401   se.e401
##      8641      1261
```

References

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J.M. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1) C1–C68.

Outline

1. Doubly-Robust Approaches

2. Machine Learning Methods
   - Post Double Selection Inference
   - Double Machine Learning

3. Graphical Sequential Treatment
   - Multiple Regression
   - Multiple Treatments

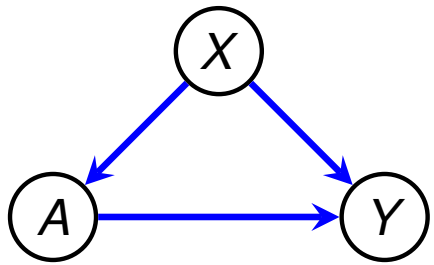
Regression

Question

Suppose we perform a regression of an outcome Y on **several** other variables. Under what circumstances are the coefficients estimating a **causal** quantity?

If they are, then what quantity is it, exactly?

If we compute
$\mathbb{E}[Y \mid A, X] = \beta_{AY \cdot X} A + \beta_{XY \cdot A} X$
under this graph, will we have:



$$p(y \mid do(a)) = \beta_{AY \cdot X} a \, ?$$

$$p(y \mid do(x)) = \beta_{XY \cdot A} x \, ?$$

$\beta_{XY \cdot A}$ **is** a causal effect, but it is the effect of X on Y when keeping A constant; this is called the **controlled direct effect**.

$\beta_{AY \cdot \boldsymbol{C}}$ denotes the regression coefficient for A on Y in the model that includes $\boldsymbol{C}$.

Table 2 Fallacy

JOURNAL ARTICLE

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich ✉, Sander Greenland Author Notes

American Journal of Epidemiology, Volume 177, Issue 4, 15 February 2013, Pages 292–298,
https://doi.org/10.1093/aje/kws412
Published: 30 January 2013 **Article history** ▾

Abstract

It is common to present multiple adjusted effect estimates from a single model in a single table. For example, a table might show odds ratios for one or more exposures and also for several confounders from a single logistic regression. This can lead to mistaken interpretations of these estimates. We use causal diagrams to display the sources of the problems. Presentation of exposure and confounder effect estimates from a single model may lead to several interpretative difficulties, inviting confusion of direct-effect estimates with total-effect estimates for covariates in the model. These effect estimates may also be confounded even though the effect estimate for the main exposure is not confounded. Interpretation of these effect estimates is further complicated by heterogeneity (variation, modification) of the exposure effect measure across covariate levels. We offer suggestions to limit potential misunderstandings when multiple effect estimates are presented, including precise distinction between total and direct effect measures from a single model, and use of multiple models tailored to yield total-effect estimates for covariates.

Multiple Exposures

Often in applications not a clear distinction between specific exposure and covariates used for adjustment.

There may well be multiple questions of interest: e.g. what is effect of smoking, diet, alcohol intake all **together**.

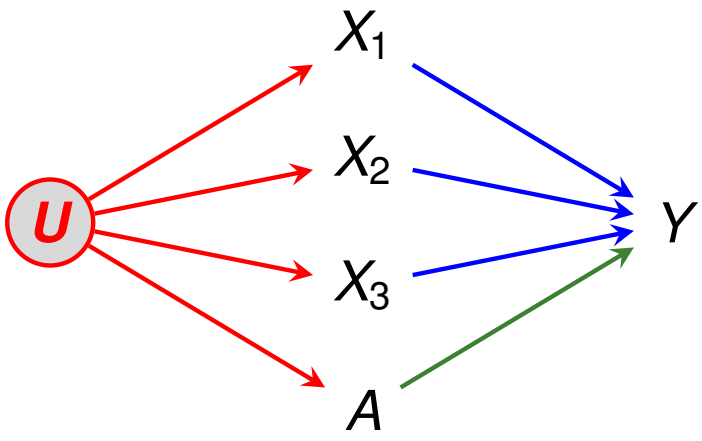
Be clear about causal question relating to multiple exposures: what would be your **ideal target trial**?

Many possible causal effects to define with multiple exposures:

- separate total effects;
- joint intervention effects;
- controlled direct effects;
- strategy for dynamic treatment effects;
- separable (or natural) direct and indirect effects.

Confounding Due to Common History

Suppose the causal structure is more like below—what should we adjust for?



Answer

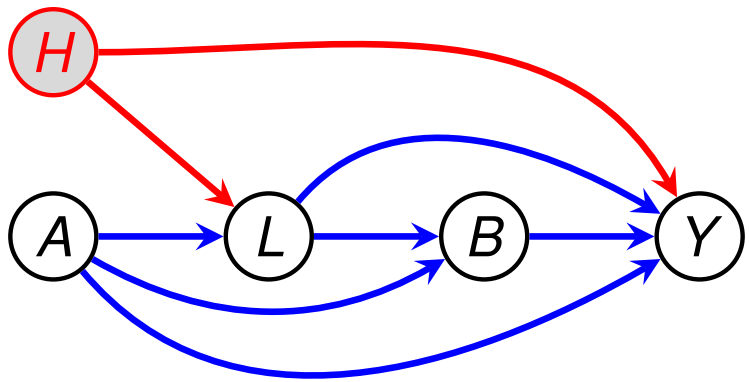
Condition on X_1, X_2, **and** X_3!

This is the only way to block all back-door paths from A to Y.

Joint Interventions

A **joint intervention** considers what would happen if we intervene on **multiple variables** at the same time.

For example, suppose we have the graph below:

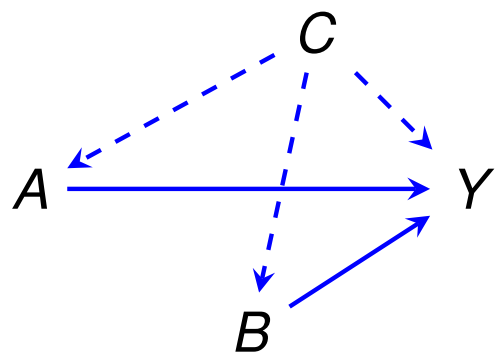


How can we identify $P(Y \mid do(A, B))$?

Examples I

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$

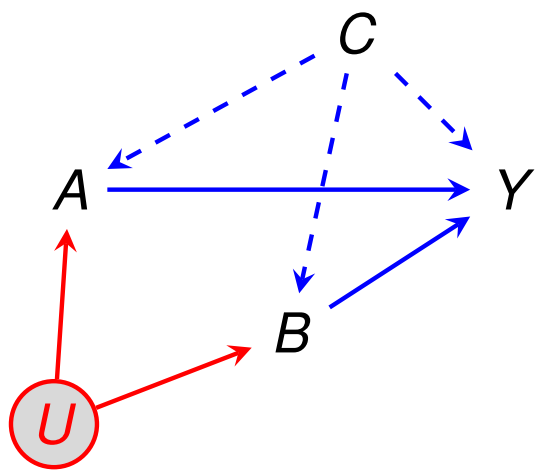


Note that $Y(a, b) \perp_d A, B \mid C$, so indeed (β_A, β_B) **are** the joint causal effects of A, B on Y.

Examples II

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$



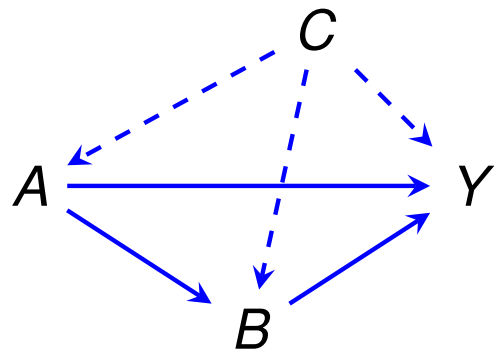
Is $Y(a, b) \perp_d A, B \mid C$?

Yes!

Examples III

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$



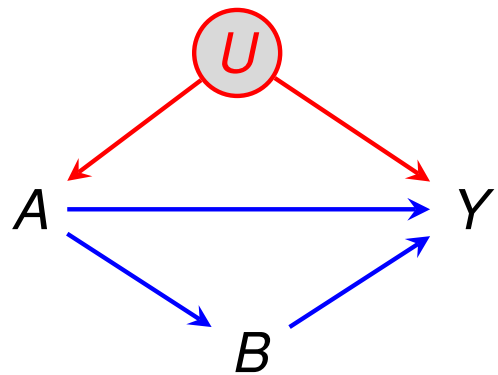
Is $Y(a, b) \perp_d A, B \mid C$?

Yes! **But:** β_A is a **controlled direct effect**, not a total effect.

Examples IV

What is the interpretation of β_A, β_B in the regression:

$$\mathbb{E}[Y \mid A = a, B = b] = \beta_A a + \beta_B b \quad ?$$



Is $Y(a, b) \perp_d A, B$?

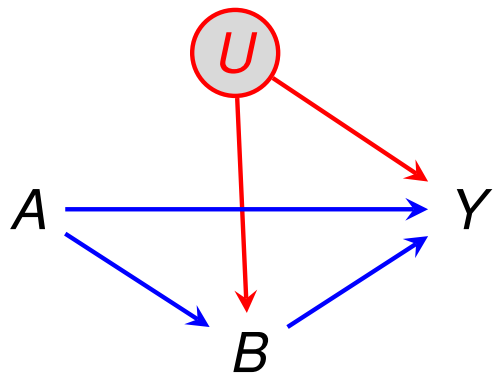
No! β_A has no causal interpretation.

However, $Y(b) \perp_d B \mid A$, so β_B **is** total effect of B on Y.

Examples V

What is the interpretation of β_A, β_B in the regression:

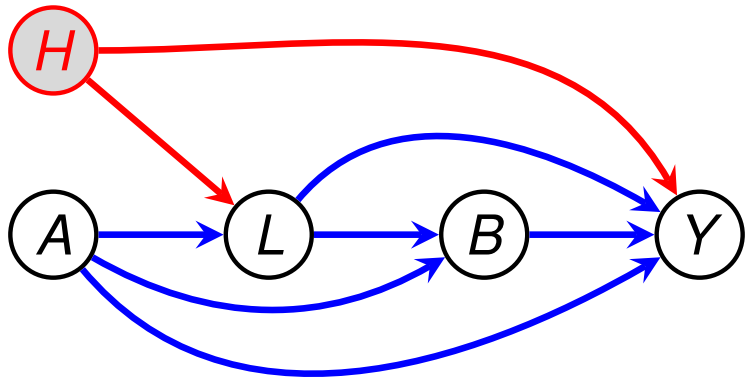
$$\mathbb{E}[Y \mid A = a, B = b] = \beta_A a + \beta_B b \quad ?$$



Is $Y(a, b) \perp_d A, B$?

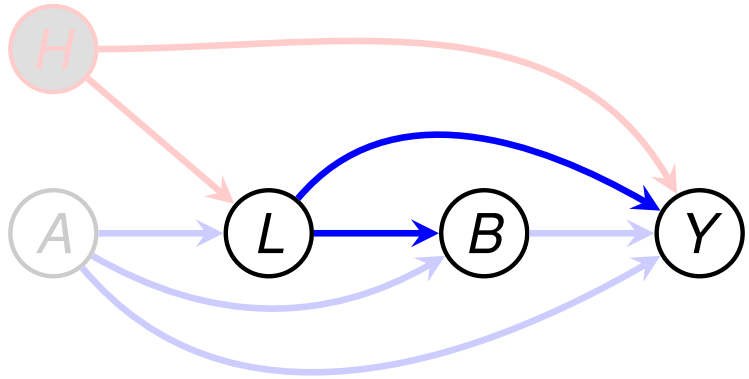
No! Neither coefficient has a causal interpretation.

Sequentially randomized experiment



- A and B are treatments;
- H is unobserved;
- L is a time varying confounder;
- Y is the final response;
- Treatment B is assigned randomly conditional on the observed history, A and L;
- Want to know $P(Y(\tilde{a}, \tilde{b}))$.

Time-Varying Confounding



Should we adjust for L?

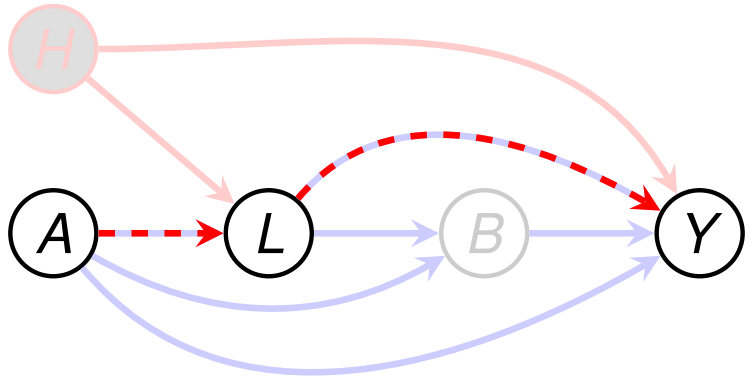
From B's perspective, L needs to be adjusted for to control for confounding.

From A's perspective, L is on the causal path and conditioning would open a spurious path.

Neither regression makes sense!

Need to **break** B's dependence on L in order to estimate the effect.

Time-Varying Confounding



Should we adjust for L?

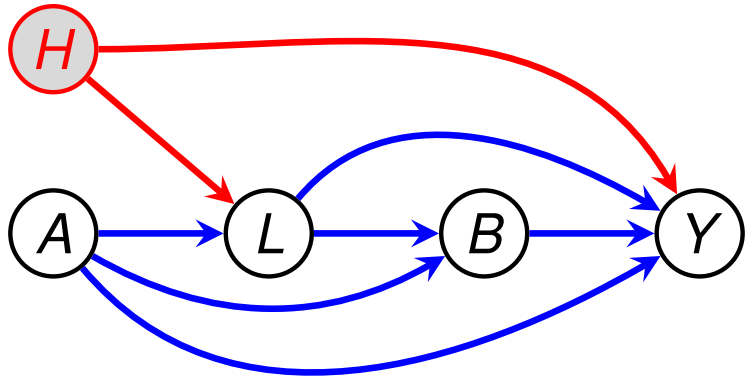
From B's perspective, L needs to be adjusted for to control for confounding.

From A's perspective, L is on the causal path and conditioning would open a spurious path.

Neither regression makes sense!

Need to **break** B's dependence on L in order to estimate the effect.

Sequentially randomized experiment



If the following holds:

$$A \perp\!\!\!\perp Y(a,b)$$
$$B(a) \perp\!\!\!\perp Y(a,b) \mid L(a), A$$

General result of Robins (1986) then implies:

$$P(Y(a,b)=y) = \sum_l P(L=l \mid A=a) \cdot P(Y=y \mid A=a, L=l, B=b).$$

Exercise: can you show this?

Do the independences hold?

Marginal Structural Models

Models of the quantity

$$P(Y \mid do(A, B)) = \sum_{l} P(Y \mid A, L = l, B) \cdot P(L = l \mid A)$$

are called **marginal structural models** (Robins et al., 2000).

They have various nice properties:

- can be modelled semi-parametrically (so no need to fully specify rest of the distribution)
- either the propensity scores ($P(A)$ and $P(B \mid A, L)$) **or** the outcome models $P(Y \mid A, L, B)$ and $P(L \mid A)$ can be used to identify parameters.
- there is also a doubly robust version!

Hard to simulate from, though Evans and Didelez (2023) provide a solution!

IPW Identification

The model suffers from time-dependent confounding, so weights must reflect this.

Idea of IPW is to obtain the g-formula by 'removing' pieces not in it. So:

$$P(A, L, B, Y) = P(A) \cdot P(L \mid A) \cdot P(B \mid A, L) \cdot P(Y \mid A, L, B)$$

So create a **pseudo-population** by fitting models for $P(A)$ and $P(B \mid A, L)$ and then reweight ith observation by

$$w_i = \frac{1}{\widehat{P}(A = a_i)} \cdot \frac{1}{\widehat{P}(B = b_i \mid A = a_i, L = l_i)}$$

Common to use logistic regression if A, B are binary.

IPW Remarks

Conditions are similar to single treatment case; see Robins et al. (2000) for details.

General form of weights with treatments $A_0, \ldots, A_K$ and covariates $L_0, \ldots, L_K$ is

$$\widehat{w}_i = \prod_{k=0}^{K} \frac{1}{\widehat{P}(A_k \mid \overline{A}_{k-1}, \overline{L}_k)},$$

where $\overline{A}_j = (A_0, \ldots, A_j)$.

- Models for $A_k | \overline{A}_{k-1}, \overline{L}_k$ must be correctly specified;
- in practice people use stabilised weights;
- can also estimate optimal dynamic treatments using *Q-learning* (Chakraborty and Moodie, 2013).

References

Chakraborty, B. and Moodie, E.M. Statistical Methods for Dynamic Treatment Regimes. *Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer, 2013.

Evans, R.J. and Didelez, V. Parameterizing and simulating from causal models (with discussion). *JRSS-B*, 2023.

Robins, J.M., Hernan, M.A. and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp.550-560, 2000.