Statistical analysis of causal parameters in epidemiology: the DAIFI study example

Antoine Chambaz[1]

[1] Laboratoire MAP5, Université Paris Descartes

Atelier INSERM "Avancées statistiques récentes en analyse causale", 7-8 juin 2011

The DAIFI study

- Collaboration in progress with J. Bouyer (Université Paris-Sud 11, INSERM, INED), E. de la Rochebrochard (INED, INSERM), S. Gruber (UC Berkeley), S. Rose (UC Berkeley) and M. J. van der Laan (UC Berkeley)
- DAIFI? From the DAIFI study website:

  > L'enquête DAIFI est une enquête scientifique menée par l'INSERM et l'INED sur le devenir des femmes et des couples après un traitement par FIV.
  >
  > The DAIFI study is a scientific investigation carried out by INSERM and INED on the lives of women and couples who underwent an IVF program.

- Thanks to:
  - Sophie Ancelet-Enjalric (INRA) for providing the dataset
  - Hôpital Cochin et CHU Clermont-Ferrand for allowing us to exploit the dataset

Describing the problem of interest and the statistical protocol

- Question of interest: estimate the

  probability that a woman who undergoes an IVF program with up to four cycles eventually gives birth.
- Statistical protocol (universal):
  1. *describe* as accurately as possible the observed data structure $O \sim P$ and its law $P \in \mathcal{M}$;
  2. *express* the parameter of interest under the form $\Psi(P)$;
  3. *study* the functional $\Psi : \mathcal{M} \rightarrow \mathbb{R}$;
  4. *derive* from this study *how to estimate* $\Psi(P)$;
  5. *carry out* the estimation.
- This 5-step protocol is typical of semi-parametric statistics.
- In step 4, we actually follow the Targeted Maximum Likelihood Estimation (TMLE) methodology.

  Original article by van der Laan et Rubin (2006), many other since then, and forthcoming large-audience book by Rose and van der Laan (June 2011) — a chapter is devoted to the DAIFI study in the Examples section.

Statistical protocol (step 1)

"1. *describe* as accurately as possible the observed data structure $O \sim P$ and its law $P \in \mathcal{M}$"

- Observed data structure: $O = (L_{0:3}, A_{0:2}) \sim P \in \mathcal{M}$ with
  - baseline covariates L_0:
    - $L_{0,1} \in \{0,1\}$, *IVF center* (Cochin or Clermont);
    - $L_{0,2} \in \mathbb{R}$, *age of woman* at first IVF cycle;
    - $L_{0,3} \in \mathbb{N}$, *number of embryos harvested* at first IVF cycle;
    - $L_{0,4} \in \{0,1\}$, *indicator of birth at first IVF cycle*;
  - for $j = 1, 2, 3$,
    - $A_{j-1} \in \{0,1\}$, *censoring* indicator after $(j-1)$-th cycle;
    - $L_j \in \{0,1\}$, *indicator of birth* at j-th IVF cycle.
- $\mathcal{M}$ is the (non-parametric) set of all laws P for O which are *compatible with the following constraints*:

$$\forall\, 0 \le j \le 2: \quad L_j = 1 \Rightarrow \begin{cases} \forall\, j \le j' \le 3, & L_{j'} = 1 \\ \forall\, j \le j' \le 2, & A_{j'} = 1 \end{cases}, \qquad A_j = 0 \Rightarrow \begin{cases} \forall\, j \le j' \le 2, & A_{j'} = 0 \\ \forall\, j < j' \le 3, & L_{j'} = 0 \end{cases}.$$

- Data: $n = 3000$ women followed during their IVF program

cycle	j	0	1	2	3
proportion of women still followed	$\frac{1}{n}\sum_{i\le n} \mathbf{1}\{A_j = 1\}$	75%	59%	49%	-
proportion of success so far	$\frac{1}{n}\sum_{i\le n} \mathbf{1}\{L_j = 1\}$	22%	32%	35%	37%

- Implicitly: we assume (*strong!*) that the number of embryos harvested at first IVF cycle is a reliable summary of the numbers of embryos possibly harvested later.

Statistical protocol (step 2)

"2. *express* the parameter of interest under the form $\Psi(P)$"

- Reminder: the question of interest is to estimate the

  probability that a woman who undergoes an IVF program with up to four cycles eventually gives birth.
- *Statistically speaking*, we are interested in $\Psi(P)$, where

$$\forall\, P \in \mathcal{M},$$
$$\Psi(P) = \sum_{\ell_{0:2}\in\{0,1\}^3} P(L_3 = 1 | L_{0:2} = \ell_{0:2}, A_{0:2} = (1,1,1)) \times P(L_2 = \ell_2 | L_{0:1} = \ell_{0:1}, A_{0:1} = (1,1)) \times P(L_1 = \ell_1 | L_0 = \ell_0, A_0 = 1) P(L_0 = \ell_0).$$

- Justification?...

  Fundamental: whether or not the assumptions presented in the *next* slide are met, $\Psi(P)$ is always a *well-defined statistical parameter* worth estimating.

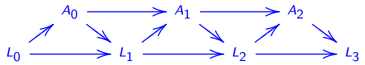
Justification

- Non-parametric modeling of the random phenomenon of interest:

NP-SEM (non-parametric system of structural equations)
$\exists\ (f_1, \ldots, f_6)$ *deterministic*,
$\exists\ (U_1, \ldots, U_6)$ sources of randomness s.t.,
once L_0 is drawn,

$$A_0 = f_1(L_0, U_1), \qquad L_1 = f_2(L_0, A_0, U_2),$$
$$A_1 = f_3(L_{0:1}, A_0, U_3), \qquad L_2 = f_4(L_{0:1}, A_{0:1}, U_4),$$
$$A_2 = f_5(L_{0:2}, A_{0:1}, U_5), \qquad L_3 = f_6(L_{0:2}, A_{0:2}, U_6).$$

Causal diagram



either *as NP-SEM* (with many missing arrows from L_0) or *constrained NP-SEM* (the influence of the past is conveyed by two nodes)

- Assumptions on the sources of randomness:

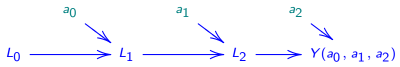
$U_1 \perp U_2 | L_0$; $U_3 \perp U_4 | (L_0, U_{1:2})$; $U_5 \perp U_6 | (L_0, U_{1:4})$

implicitly: $(L_0, U_1, \ldots, U_6)$ mutually independent

- The notion of *intervention*...

once L_0 is drawn,

$$A_0 = a_0, \qquad L_1 = f_2(L_0, A_0, U_2),$$
$$A_1 = a_1, \qquad L_2 = f_4(L_{0:1}, A_{0:1}, U_4),$$
$$A_2 = a_2, \qquad Y(a_0, a_1, a_2) = f_6(L_{0:2}, A_{0:2}, U_6).$$



- ... gives rise to the counterfactual variables $\{Y(a_0, a_1, a_2) : (a_0, a_1, a_2) \in \mathcal{A}\}$ s.t.
  - $Y(a_0, a_1, a_2)$ is *the outcome of the IVF program when one imposes* $(A_0, A_1, A_2) = (a_0, a_1, a_2)$;
  - *consistency*: in particular, $L_3 = Y(A_0, A_1, A_2)$;
  - *sequential randomization*: conditionally on the past, censoring is independent of two counterfactual outcomes.

- Then

$$\Psi(P) = \mathbb{E}_P[Y(1, 1, 1)].$$

Statistical protocol (step 3)

" 3. *study* the functional $\Psi : \mathcal{M} \to \mathbb{R}$"

- Statistical parameter Ψ is *differentiable* at any $P \in \mathcal{M}$:
  - if $P_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} P$ from *direction* S, i.e.

$$P_0 = P, \quad \frac{\partial}{\partial\varepsilon} \log P_\varepsilon(O)|_{\varepsilon=0} = S(O),$$

  - then

$$\frac{\partial}{\partial\varepsilon} \Psi(P_\varepsilon)|_{\varepsilon=0} = E_P\{S(O) \times D^\star_\Psi(P)(O)\}$$

  for some "efficient influence curve" (derivative) $D^\star_\Psi(P) \in L^2_0(P)$.
- The efficient influence curve $D^\star_\Psi(P)$ is *known explicitly here* (otherwise, we would have derived it *recursively*).
- The efficient influence curve $D^\star_\Psi(P)$ teaches us what is the *relevant information for the purpose of estimating* $\Psi(P)$.
- Furthermore, the asymptotic variance of *any* regular estimator of $\Psi(P)$ is lower-bounded by the variance $\mathrm{Var}_P D^\star_\Psi(P)(O) = E_P\{D^\star_\Psi(O)^2\}$ of the efficient influence curve at P.

Statistical protocol (step 4)

"4. *derive* from this study *how to estimate* $\Psi(P)$"

The TMLE methodology is a 4-step estimating procedure:

(A) build an *initial estimator* P_n^0 of P

recommended: aggregation of several estimators into one single better estimator (*e.g.*, by relying on multi-fold cross-validation); see the *super-learning* machine-learning methodology, and *remarkable* `R`-package `SuperLearner` by E. Polley

remark: heuristically, its *bias-variance trade-off* is optimized for the sake of estimating the whole law P

(B) build a *fluctuation* $P_n^0(\varepsilon)$ of P_n^0 from direction $D_\Psi^\star(P_n^0)$

remark: since all variables (except L_0) are binary, this mainly involves a series of *logistic regressions*! (see next slide)

(C) estimate by *maximum likelihood* the best model $P_n^* = P_n^0(\varepsilon_n)$ within the fluctuation

remark: heuristically, its *bias-variance trade-off* is optimized for the sake of estimating what we really care for *i.e.*, $\Psi(P)$!

(D) estimate $\Psi(P)$ by the TMLE $\Psi(P_n^*)$ (a substitution estimator)

On the fluctuation

Let's simply build a fluctuation for the *conditional distribution of* L_3 *given its past* (*i.e.*, given $(L_{0:2}, A_{0:2})$).

- Relevant feature of initial estimator P_n^0 is the conditional probability $P_n^0(L_3 = 1|L_{0:2}, A_{0:2})$.
- Define $P_n^0(\varepsilon)$ (first fluctuation of P_n^0) in such a way that
  - the past of L_3 has the same distribution under P_n^0 as under $P_n^0(\varepsilon)$
  - under $P_n^0(\varepsilon)$,

$$\text{logit}\, P_n^0(\varepsilon)(L_3 = 1|L_{0:2}, A_{0:2}) = \text{logit}\, P_n^0(L_3 = 1|L_{0:2}, A_{0:2}) + \varepsilon \times \frac{\mathbf{1}\{A_{0:2} = (1,1,1)\}}{g(P_n^0)(1,1,1)}, \quad (1)$$

  where $g(P_n^0)(1,1,1) = P_n^0(A_0 = 1|L_0) \times P_n^0(A_1 = 1|L_{0:1}, A_0) \times P_n^0(A_2 = 1|L_{0:2}, A_{0:1})$.

  (the $\mathbf{1}/g$-factor *targets* the relevant component of $D^\star_\Psi(P_n^0)$)
- Maximizing the likelihood under $P_n^0(\varepsilon)$ (wrt $\varepsilon \in \mathbb{R}$) amounts to fitting (1) (standard logistic regression)!
  Yields MLE ε_n^0.
- First update of P_n^0 is $P_n^1 = P_n^0(\varepsilon_n^0)$.

We're done with the conditional distribution of L_3 given its past, and go now for the update of the conditional distribution of L_2 given its past, and so on...

Here, the TMLE procedure *converges in one single updating step*.

Asymptotic properties of the TMLE (1/2)

from previous slide: estimate $\Psi(P)$ by the TMLE $\Psi(P_n^*)$ (a substitution estimator)

- TMLE is a substitution estimator: consequently, it automatically satisfies all the constraints on the parameter of interest (namely here, that $\Psi(P) \in [0,1]$)

  remark: by solving an estimating equation for $\Psi(P)$, one may end up with an estimator outside the range $[0,1]$!

- TMLE involves a maximization step

  remark: *maximizing* is much easier than *solving* an equation (in particular, one has seldom to cope with multiple solutions)!

- by construction, TMLE satisfies $P_n D_\Psi^\star(P_n^1) = 0$.

Asymptotic properties of the TMLE (2/2)

from previous slide: TMLE satisfies $P_n D^\star_\Psi(P_n^1) = 0$

- If P_n^1 converges in such a way that
  - we get the conditional distributions of A_0, A_1, A_2 given their past right,

  or- we get the conditional distribution of L_1, L_2, L_3 given their past right,

  then (under mild additional assumptions) TMLE is *consistent*!

  Example of the so-called *double-robustness* property.
- If the TMLE $\Psi(P_n^1)$ consistently estimates $\Psi(P)$ then (under mild additional assumptions) $\sqrt{n}(\Psi(P_n^1) - \Psi(P))$ is asymptotically Gaussian, centered with variance denoted by σ^2.

  Moreover:
  - if we get the conditional distributions of A_0, A_1, A_2 and L_1, L_2, L_3 given their past right, then $\sigma^2 = \mathrm{Var}_P D^\star_\Psi(P)(O)$ (the smallest possible value);
  - if we estimate the conditional distributions of A_0, A_1, A_2 given their past by maximum-likelihood on a well-specified model, then σ^2 is *conservatively* estimated by

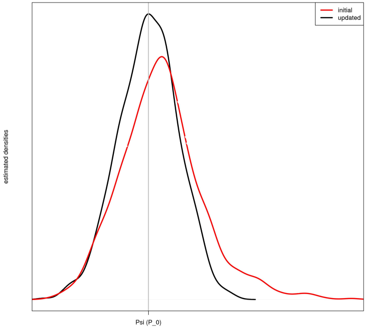
$$\frac{1}{n}\sum_{i=1}^{n} D^\star_\Psi(P_n^1)(O^{(i)}).$$

  - remark: one can always rely on the bootstrap to estimate σ^2.

Simulation study

- The simulation scheme *mimicks* the DAIFI dataset.
- True value of parameter: $\Psi(P) \approx 0.798$.
- We simulate $B = 1000$ datasets with $n = 3000$ observations.
- Summarized results:

  - $\frac{1}{B}\sum_{b\leq B}\Psi(P_n^{1,b}) \approx 0.798$
  - $\frac{1}{B}\sum_{b\leq B}\mathbf{1}\{\Psi(P_0) \in [\Psi(P_n^{1,b}) \pm 1.96\frac{\hat{\sigma}}{n}]\} \approx 0.926$ (wished level equal to 95%)



- Consistant estimator!
- Empirical cover slightly deficient.

- The update corrects the poor initial estimations!

Statistical protocol (step 5)

"5. carry out the estimation"

- Pointwise estimation:

$$\Psi(P_n^1) = 0.50$$

  95%-confidence interval:

$$[0.48; 0.53]$$

- Conclusion:

  The probability that a woman who undergoes an IVF program with up to four cycles will eventually give birth equals approximately $\frac{1}{2}$.

- A little bit disappointing in the sense that this is not significantly different from what one gets by adopting a standard survival analysis approach...
- Next step (work in progress!): do not assume anymore that the number of embryos harvested at first IVF cycle is a reliable summary of the numbers of embryos possibly harvested later!
  - this introduces time-dependent confounders...
  - standard survival analysis approach not possible anymore,
  - however TMLE methodology presented here can be slightly modified in order to cope with them!

to be continued...

[*sneak preview*:

- estimated probability equal to 0.39, with a 95% confidence interval equal to [0.34; 0.44] — remember that crude probability of success equals 0.37!
- *surprinsingly* suggests... something! (we need more time!)]

References

- *Causality*, Judea Pearl (2000)
- *Statistics for Epidemiology*, Nicholas Jewell (2004)
- `SuperLearner` `R`-package by Eric Polley (2009-2011)
- *Targeted maximum likelihood learning*, Mark van der Laan et Daniel Rubin, International Journal of Biostatistics (2006)
- *Targeted Learning*, Sherri Rose and Mark van der Laan (June 2011)
- *TMLE of the probability of success of an IVF program and the DAIFI study*, chapter in *Targeted learning*, AC (June 2011)