Estimation de l'effet dune exposition cumulée avec les modèles structuraux marginaux

Denis Talbot

Département de médecine sociale et préventive, Université Laval
Unité santé des populations et pratiques optimales en santé, CHU de Québec - Université Laval

Novembre 2016

Plan

- Contexte
- Le *Honolulu Heart Program* (HHP)
- Implantation des modèles structuraux marginaux (MSMs) avec SAS
- Conclusion

Contexte :

En statistique, les techniques d'**inférence causale** sont construites pour **prédire l'effet qu'aurait une intervention** potentielle, telle qu'un traitement, une campagne de santé publique ou une politique, soit à l'aide d'expériences randomisées ou de données observationnelles.

Contexte :

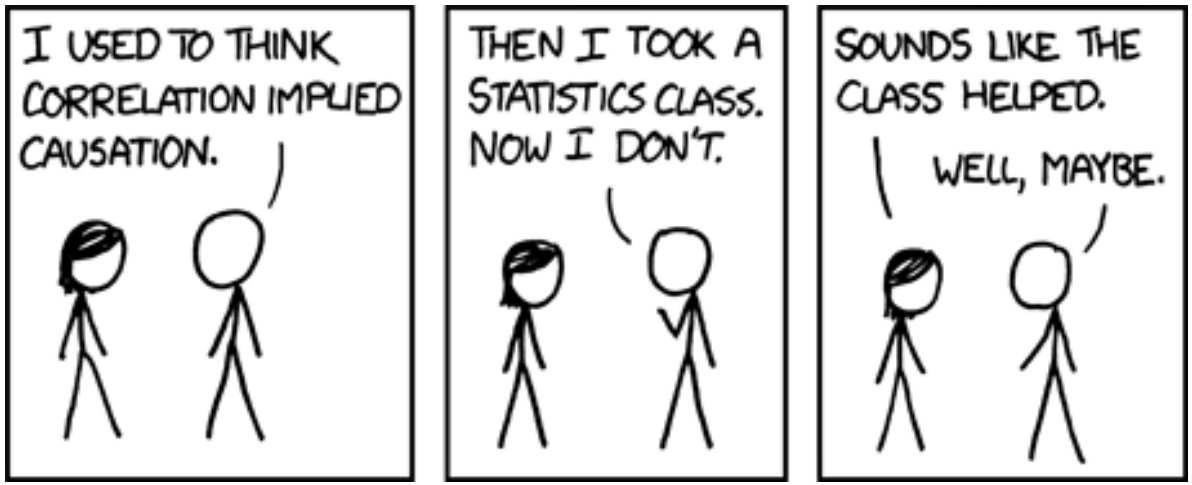
En statistique, les techniques d'**inférence causale** sont construites pour **prédire l'effet qu'aurait une intervention** potentielle, telle qu'un traitement, une campagne de santé publique ou une politique, soit à l'aide d'expériences randomisées ou de données observationnelles.

Plus facile à réaliser à l'aide d'études randomisées...

mais parfois nécessaire d'utiliser des études d'observation.

Contexte :

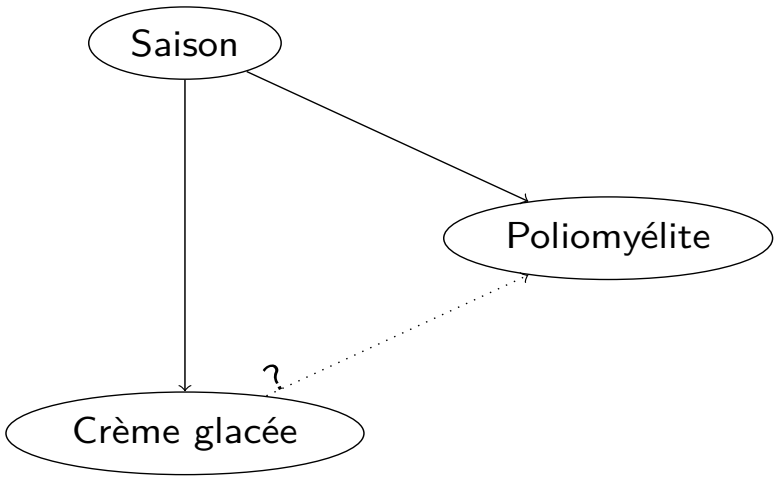
En absence de randomisation (idéale), l'association statistique $\neq$ causalité.



source : `https://xkcd.com/552/`

Contexte :

En absence de randomisation (idéale), différentes variables peuvent influencer à la fois l'exposition et l'issue et engendrer des associations non causales.

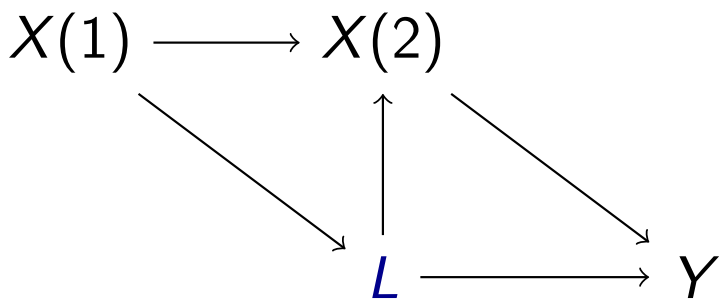


Contexte :

Estimer l'effet d'un **régime d'exposition** sur une issue ne peut généralement pas être fait à l'aide d'un modèle de régression avec ajustement pour les variables confondantes en raison de la **confusion dépendante du temps**.

Contexte :

Par exemple, si on est intéressé à l'effet sur Y d'une intervention jointe sur $\{X(1), X(2)\}$:



Alors **ni** le modèle $E[Y] = \beta_0 + \beta_1 X(1) + \beta_2 X(2) + \beta_3 L$ **ni** le modèle $E[Y] = \beta_0 + \beta_1 X(1) + \beta_2 X(2)$ **ne sont appropriés**.

Contexte :

Les modèles structuraux marginaux (MSMs) sont une solution à ce problème.

L'implantation la plus commune des MSMs consiste à ajuster une régression **pondérée** de l'issue en fonction de l'historique d'exposition.

L'objectif de ces poids est de **créer une pseudo-population** où l'historique d'exposition n'est plus associé aux variables confondantes, **répliquant une étude randomisée séquentielle**.

Le HHP :

- 8006 participants, Japonnais-Américains vivant sur l'île de Oahu, Hawaï et nés entre 1900 et 1919
- Recrutés entre 1965 et 1968 à l'aide d'une liste d'inscrits dans la réserve de l'armée (âgés entre 45 et 68 ans)
- Trois examens où des mesures d'activité physique et de pression artérielle (PA) ont été prises de façon similaire (**Examen 1, 1965-1968 ; Examen 2, 1968-1971** ; Examen 4, 1991-1993)

Les variables d'intérêt principal :

- Activité physique (actif/inactif) - `act_phy_T1` ($X(1)$) et `act_phy_T2` ($X(2)$)
- PA systolique (PAS - en mmHg) - `PAS_T2` (Y)

On s'intéresse à l'effet sur `PAS_T2` d'une intervention jointe sur $\{$`act_phy_T1`, `act_phy_T2`$\}$

Les variables suivantes supplémentaires ont été sélectionnées :

- Âge (en années) - `age_T1` et `age_T2`
- Statut d'emploi (oui/non) - `emploi_T1` et `emploi_T2`
- IMC (en kg/m^2) - `IMC_T1`
- Statut de fumeur (actuel, dans le passé, jamais) - `fumeur_actuel_T1`, `fumeur_passe_T1`, `fumeur_actuel_T2` et `fumeur_passe_T2`
- Utilisation d'antihypertenseur (oui/non) - `Anti_hyp_T1`

Parce qu'elles étaient mesurées de façon similaire aux examens et soit cliniquement importantes ou possiblement confondantes.

Variables potentiellement confondantes

Relation act_phy_T1 $\rightarrow$ PAS_T2 ($L(1)$) : age_T1, emploi_T1, fumeur_actuel_T1 et fumeur_passe_T1

Relation act_phy_T2 $\rightarrow$ PAS_T2 ($L(2)$) : age_T2, emploi_T2, fumeur_actuel_T2, fumeur_passe_T2, IMC_T1*, Anti_hyp_T1*, PAS_T1* et act_phy_T1.

* : variable potentiellement intermédiaire de la relation act_phy_T1 $\rightarrow$ PAS_T2.

Statistiques descriptives :

	Inactif à T1	Actif à T1
Âge à T1, moy (É-T)	54.3 (5.5)	54.4 (5.6)
Emploi à T1, n (%)	1316 (91%)	5970 (92%)
Fumeur passé à T1, n (%)	368 (25%)	1715 (26%)
Fumeur actuel à T1, n (%)	636 (44%)	2832 (44%)

Statistiques descriptives :

	Inactif à T2	Actif à T2
Âge à T2, moy (É-T)	56.7 (5.8)	56.9 (5.5)
Emploi à T2, n (%)	721 (81%)	5892 (89%)
Fumeur passé à T2, n (%)	301 (34%)	2003 (30%)
Fumeur actuel à T2, n (%)	341 (38%)	2683 (41%)
IMC à T1, moy (É-T)	24.4 (3.1)	23.8 (3.0)
PAS à T1, moy (É-T)	136.7 (22.8)	133.4 (20.4)
Actif à T1, n (%)	532 (60%)	5553 (85%)

Implantation des MSMs avec SAS

On veut estimer causalement les paramètres du modèle :

$$E[PAS\_T2] = \beta_0 + \beta_1 act\_phy\_T1 + \beta_2 act\_phy\_T2 + \beta_3 act\_phy\_T1 \times act\_phy\_T2$$

En raison de la confusion dépendante du temps, il est impossible de simplement ajuster pour les variables confondantes en les ajoutant comme des covariables.

On utilisera une pondération qui créera une pseudo-population où l'exposition à chaque période est indépendante des variables confondantes antérieures.

Implantation des MSMs avec SAS

Étape 1) Les données doivent être dans un **format large**

ID	act_phy_T1	PAS_T1	fumeur_actuel_T1	fumeur_passe_T1	IMC_T1	Anti_hyp_T1	emploi_T1	act_phy_T2	PAS_T2	fumeur_actuel_T2	fumeur_passe_T2	emploi_T2	age_T1	age_T2
1	0	133	0	0	23	0	0	1	100	1	0	1	64	66
2	1	163	0	0	23	0	0	0	148	0	0	0	53	55
3	1	107	1	0	17	0	1	1	123	0	1	1	54	56

Supposons pour l'instant qu'il n'y pas d'attrition.

Les poids stabilisés pour notre application sont :

$$W_i = \frac{P(X(1) = x_i(1))}{P(X(1) = x_i(k)|\boldsymbol{L}(1) = \boldsymbol{l}_i(1))} \times \frac{P(X(2) = x_i(2)|X(1) = x_i(1))}{P(X(2) = x_i(2)|X(1) = x_i(1), \boldsymbol{L}(2) = \boldsymbol{l}_i(2))}$$

De façon plus générale, les poids stabilisés sont :

$$W_i = \prod_{t=1}^{T} \frac{P(X(t) = x_i(t)|\bar{X}(t))}{P(X(t) = x_i(t)|\bar{X}_t, \bar{\boldsymbol{L}}(t))}$$

Étape 2) Calculer les poids.

Étape 2.1) Obtenir les valeurs prédites de modèles de régression logistique.

```
/*P(X(1) = 1) :*/
PROC LOGISTIC DATA = donnees DESCENDING;
    MODEL act_phy_T1 =  / LINK = logit;
    OUTPUT OUT = sortie1 P = num1;
RUN;
```

```
/*P(X(1) = 1|L(1)) :*/
PROC LOGISTIC DATA = donnees DESCENDING;
    MODEL act_phy_T1 = age_T1 emploi_T1 fumeur_passe_T1 fumeur_actuel_T1 / LINK = logit;
    OUTPUT OUT = sortie2 P = dnom1;
RUN;
```

```
/*P(X(2) = 1| X(1)) :*/
PROC LOGISTIC DATA = donnees DESCENDING;
    MODEL act_phy_T2 = act_phy_T1 / LINK = logit;
    OUTPUT OUT = sortie3 P = num2;
RUN;
```

```
/*P(X(2) = 1| X(1), L(1))*/
PROC LOGISTIC DATA = donnees DESCENDING;
    MODEL act_phy_T2 = age_T2 fumeur_passe_T2 fumeur_actuel_T2 emploi_T2
                       IMC_T1 Anti_hyp_T1 act_phy_T1 PAS_T1 / LINK = logit;
    OUTPUT OUT = sortie4 P = dnom2;
RUN;
```

Étape 2.2) Calculer les poids à l'aide des sorties des modèles :

```
DATA poids;
     MERGE sortie1-sortie4;
     BY ID;
     IF act_phy_T1 = 0 THEN DO;
          num1 = 1 - num1;
          dnom1 = 1 - dnom1;
     END;
     IF act_phy_T2 = 0 THEN DO;
          num2 = 1 - num2;
          dnom2 = 1 - dnom2;
     END;

     poids_stabilises = num1/dnom1*num2/dnom2;
     bidon = 1;
RUN;
```

Les pertes au suivi peuvent engendrer un biais de sélection.

Pour atténuer ce biais, on peut utiliser une pondération.

`Censure_T2` $= 1$ si le sujet n'a pas participé au temps 2,
`Censure_T2` $= 0$ sinon.

Les poids stabilisés de censure pour notre application sont :

$$W_i^{\dagger} = \frac{P(Censure\_T2 = 0|X(1))}{P(Censure\_T2 = 0|\boldsymbol{Z}(1))},$$

où $\boldsymbol{Z}(1)$ représente les variables disponibles à T1 qui pourraient prédire le risque censure à T2.

Les poids totaux pour notre application sont :

$$W^T = W_i \times W_i^{\dagger}$$

Ces poids cherchent à répliquer une étude randomisée séquentielle sans attrition.

Les poids stabilisés de censure généraux sont :

$$W_i^{\dagger} = \prod_{t=2}^{T} \frac{P(C(t) = 0 | C(t-1) = 0, \bar{X}(t-1))}{P(C(t) = 0 | C(t-1) = 0, \bar{X}(t-1), \bar{\boldsymbol{Z}}(t-1))}$$

et les poids totaux sont toujours $W^T = W_i \times W_i^{\dagger}$.

Étape 2) Calculer les poids.

Étape 2.1) Obtenir les valeurs prédites de modèles de régression logistique.

```
/*P(C(2) = 0|X(1))*/
PROC LOGISTIC DATA = donnees;
     MODEL Censure_T2 = act_phy_T1 / LINK = logit;
     OUTPUT OUT = sortie5 P = num2c;
RUN;
```

```
/*P(C(2) = 0|X(1), Z(1))*/
PROC LOGISTIC DATA = donnees;
     MODEL Censure_T2 = act_phy_T1 age_T1 emploi_T1 PAS_T1
                 Anti_hyp_T1 IMC_T1 fumeur_actuel_T1 fumeur_passe_T1 / LINK = logit;
     OUTPUT OUT = sortie6 P = dnom2c;
RUN;
```

Étape 2.2) Calculer les poids à l'aide des sorties des modèles :

```
DATA poids;
     MERGE sortie1-sortie6;
     BY ID;
     IF act_phy_T1 = 0 THEN DO;
          num1 = 1 - num1;
          dnom1 = 1 - dnom1;
     END;
     IF act_phy_T2 = 0 THEN DO;
          num2 = 1 - num2;
          dnom2 = 1 - dnom2;
     END;

     poids_stabilises = num1/dnom1*num2/dnom2*num2c/dnom2c;
     bidon = 1;
RUN;
```

Lorsque les modèles sont correctement spécifiés, la moyenne des poids devrait être proche de 1.

```
PROC MEANS DATA = poids MEAN;
     VAR poids_stabilises;
     OUTPUT OUT = poids2
     MEAN = poids_moy;
RUN;
```

The MEANS Procedure

Analysis Variable : poids_stabilises
Mean
1.0001879

Étape 3) normaliser les poids (facultatif, mais peut améliorer la performance) :

$$NW^T = \frac{W^T}{\sum W^T/n}$$

```
DATA poids2;
     SET poids2;
     bidon = 1;
RUN;
```

```
DATA poids3;
     MERGE poids poids2;
     BY bidon;
     poids_normalises_stabilises = poids_stabilises/poids_moy;
RUN;
```

Étape 4) tronquer les poids élevés (facultatif, mais peut améliorer la performance) :

```
PROC MEANS DATA = poids3 MIN Q1 MEDIAN MEAN Q3 MAX N;
    VAR poids_normalises_stabilises;
RUN;
```

The MEANS Procedure

Analysis Variable : poids_normalises_stabilises						
Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	N
0.2711439	0.9607172	0.9810791	1.0000000	1.0108855	2.5996987	7409

```
PROC UNIVARIATE DATA = poids3 NOPRINT;
    VAR poids_normalises_stabilises;
    OUTPUT OUT = pctls P99 = p99;
RUN;
```

```
DATA pctls;
    SET pctls;
    bidon = 1;
RUN;
```

```
DATA poids4;
    MERGE poids3 pctls;
    BY bidon;
    IF poids_normalises_stabilises > P99 THEN poids_normalises_stabilises = P99;
RUN;
```

Sur les données pondérées, l'exposition à chacun des temps devrait être indépendante des variables potentiellement confondantes antérieures :

	Inactif à T1	Actif à T1
Âge à T1, moy (É-T)	54.3 (5.5)	54.4 (5.6)
Emploi à T1, n (%)	1247 (92%)	5542 (92%)
Fumeur passé à T1, n (%)	356 (26%)	1583 (26%)
Fumeur actuel à T1, n (%)	593 (44%)	2638 (44%)

	Inactif à T2	Actif à T2
Âge à T2, moy (É-T)	57.0 (5.6)	57.0 (5.6)
Emploi à T2, n (%)	765 (88%)	5712 (88%)
Fumeur passé à T2, n (%)	269 (31%)	2005 (31%)
Fumeur actuel à T2, n (%)	358 (41%)	2662 (41%)
IMC à T1, moy (É-T)	24.0 (3.1)	23.8 (3.0)
PAS à T1, moy (É-T)	134.6 (22.1)	134.0 (20.7)
Actif à T1, n (%)	517 (59%)	5528 (85%)

Il est normal qu'il reste une association entre les expositions avec les poids stabilisés !

S'il reste des associations (différences de moyennes, de fréquences relatives ou d'écart-type), on pourrait modifier les modèles de régression logistique en ajoutant des termes d'ordre supérieurs (e.g., quadratiques) ou des termes d'interaction statistique.

Étape 5) ajuster le modèle $E[PAS\_T2] = \beta_0 + \beta_1 act\_phy\_T1 + \beta_2 act\_phy\_T2 + \beta_3 act\_phy\_T1 \times act\_phy\_T2$ sur les données pondérées.

Utiliser un estimateur robuste des variances permet d'obtenir des intervalles de confiance et des tests statistiques qui tiennent adéquatement compte de l'estimation des poids.

```
ODS SELECT GEEEmpPEst;
PROC GENMOD DATA = poids4;
    CLASS act_phy_T1 act_phy_T2 ID / DESCENDING PARAM = REF;
    MODEL PAS_T2 = act_phy_T1|act_phy_T2 / LINK = ID DIST = NORMAL;
    WEIGHT poids_normalises_stabilises;
    REPEATED SUBJECT = ID / TYPE = IND;
RUN;
```

The GENMOD Procedure

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter			Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept			136.3024	1.3141	133.7268	138.8780	103.72	<.0001
act_phy_T1	1		-1.3566	1.6064	-4.5050	1.7918	-0.84	0.3984
act_phy_T2	1		-2.4375	1.4779	-5.3342	0.4592	-1.65	0.0991
act_phy_T*act_phy_T2	1	1	1.4415	1.7647	-2.0172	4.9002	0.82	0.4140

On aurait pu modéliser différemment l'effet de notre exposition et utiliser les mêmes poids :

```
DATA poids5;
     SET poids4;
     act_phy_cum = act_phy_T1 + act_phy_T2;
RUN;

ODS SELECT GEEEmpPEst;
PROC GENMOD DATA = poids5;
     CLASS act_phy_cum ID / DESCENDING PARAM = REF;
     MODEL PAS_T2 = act_phy_cum / LINK = ID DIST = NORMAL;
     WEIGHT poids_normalises_stabilises;
     REPEATED SUBJECT = ID / TYPE = IND;
RUN;
```

The GENMOD Procedure

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

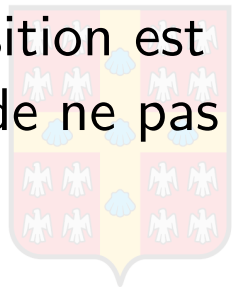
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		136.3005	1.3142	133.7246	138.8763	103.71	<.0001
act_phy_cum	2	-2.3506	1.3430	-4.9828	0.2815	-1.75	0.0801
act_phy_cum	1	-2.0643	1.4229	-4.8531	0.7245	-1.45	0.1468

Conclusion :

- L'implantation des MSMs avec SAS requière plusieurs étapes, mais chacune est relativement simple.
- On dispose de certains outils pour vérifier de l'ajustement de nos modèles

Conclusion :

- L'implantation des MSMs avec SAS requière plusieurs étapes, mais chacune est relativement simple.
- On dispose de certains outils pour vérifier de l'ajustement de nos modèles
- ... mais rien ne peut assurer qu'il ne reste pas de biais de confusion !
- La sélection des variables confondantes demeure un défi.
- Lorsque le modèle reliant l'issue à l'historique d'exposition est incorrectement spécifié, il est peut-être plus prudent de ne pas utiliser les poids stabilisés présentés.



Références :

- Robins, J.M., Hernán, M.A., Brumback, B. Marginal structural models and causal inference in epidemiology. Epidemiology, 2000 ;11(5) :550-560. (Référence générale sur les MSMs)
- Moodie, E.E.M., Delaney, J.A.C., Lefebvre, G., Platt, R.W. Missing confounding data in marginal structural models : a comparison of inverse probability weighting and multiple imputation. Int J Biostat, 2008 ;4(1) :1-23. (Traitement de l'attrition)
- Xiao, Y., Abrahamowicz, M., Moodie, E.E.M. Accuracy of conventional and marginal structural Cox model estimators : a simulation study. Int J Biostat, 2010 ; 6(2). (Normalisation des poids)

Références :

- Talbot, D., Atherton, J., Rossi, A.M., Bacon, S.L., Lefebvre, G. A cautionary note on the use of stabilized weights in marginal structural models. Stat Med, 2015 ; 34 (5) : 812-823. (Stabilisation des poids)
- Talbot, D., Rossi, A.M., Bacon, S.L., Atherton J., Lefebvre, G. (2016) A graphical perspective of marginal structural models : an application for estimating the effect of physical activity on blood pressure, soumis à SMMR. (Sélection des variables confondantes)