Section 9

Données manquantes II : pondération par probabilité inverse et GEE pondérées

Séance du 19 mars 2027 • Lectures : Wu (2010), § 3.3.2 (GEE pondérées) ; Robins, Rotnitzky et Zhao (1995) ; Fitzmaurice et coll. (2011), ch. 17–18 ; Tsiatis (2006), ch. 6–7 et 10 (doctorat)

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

À la fin de cette séance, vous devriez être en mesure de :

1. expliquer l'idée de Horvitz-Thompson et démontrer que l'estimateur par pondération inverse (IPW) d'une moyenne est sans biais sous MAR lorsque la probabilité d'observation est connue, et convergent lorsqu'elle est bien estimée ;
2. expliquer pourquoi les GEE ordinaires exigent MCAR et comment la pondération rétablit la validité sous MAR ;
3. construire des poids par probabilité inverse de censure (IPCW) dans une étude longitudinale avec attrition : modèles logistiques séquentiels, produit cumulatif, poids stabilisés et normalisés ;
4. justifier la restriction à la structure de travail d'indépendance lorsque les poids varient dans le temps ;
5. discuter l'estimation de la variance, les poids extrêmes et l'estimateur augmenté doublement robuste (AIPW) ;
6. comparer les hypothèses de la pondération, de l'imputation multiple et de la vraisemblance.

1 Mise en contexte

À la séance 8, nous avons traité les données manquantes en modélisant *les données* : loi conjointe pour la vraisemblance, lois conditionnelles pour l'imputation. Ces approches exigent un modèle correct pour la loi des valeurs manquantes sachant les valeurs observées. Dans les études longitudinales analysées par GEE, on veut justement éviter de spécifier la loi conjointe des mesures répétées : le modèle marginal ne précise que la moyenne $\mathrm{E}(Y_{ij} \mid X_i)$ (séance 6). Il serait incohérent de devoir ensuite postuler la loi conjointe pour traiter les abandons.

L'approche de cette séance modélise plutôt *le mécanisme de non-réponse* : on estime pour chaque unité observée sa probabilité d'avoir été observée, et on la pondère par l'inverse de cette probabilité. L'échantillon pondéré reconstitue alors, en moyenne, la population de départ.

Exemple 1.1 (Attrition dans une étude longitudinale)**.** Dans un essai clinique sur l'hypertension, on mesure la pression artérielle à quatre visites. Les patients dont la pression reste élevée à une visite sont plus susceptibles d'abandonner l'étude avant la suivante (inefficacité perçue, effets indésirables, changement de traitement). À la dernière visite, les patients encore présents ont donc, en moyenne, mieux répondu au traitement que l'ensemble des patients randomisés : la moyenne observée surestime l'efficacité. Si la probabilité d'abandon ne dépend que de l'historique observé (pression aux visites précédentes, groupe, covariables de base), le mécanisme est MAR et la pondération peut corriger ce biais.

2 L'idée de Horvitz-Thompson

L'idée vient de l'échantillonnage. Soit une population finie $\{y_1, \dots, y_N\}$ dont on veut estimer le total $t = \sum_{k=1}^{N} y_k$. On tire un échantillon selon un plan où l'unité k est incluse avec une probabilité connue $\pi_k > 0$; notons R_k l'indicatrice d'inclusion, de sorte que $\mathrm{E}(R_k) = \pi_k$ (ici, l'aléa vient du plan, les y_k sont fixes).

Proposition 2.1 (Horvitz et Thompson, 1952)**.** *L'estimateur* $\widehat{t}_{\mathrm{HT}} = \sum_{k=1}^{N} R_k\, y_k/\pi_k$ *est sans biais pour* t*.*

Démonstration. $\mathrm{E}(\widehat{t}_{\mathrm{HT}}) = \sum_k \mathrm{E}(R_k)\, y_k/\pi_k = \sum_k \pi_k y_k/\pi_k = t.$ □

Chaque unité échantillonnée « représente » $1/\pi_k$ unités de la population : une unité échantillonnée avec probabilité $1/4$ compte pour quatre. La moyenne $\widehat{t}_{\mathrm{HT}}/N$ est sans biais pour t/N ; l'estimateur de Hájek $\sum_k R_k y_k/\pi_k / \sum_k R_k/\pi_k$, qui remplace N par son estimation sans biais $\sum_k R_k/\pi_k$, est légèrement biaisé mais souvent plus stable.

Pour les données manquantes, on interprète la non-réponse comme une seconde phase d'échantillonnage dont les probabilités d'inclusion sont inconnues, mais modélisables sous MAR. Lorsque ces probabilités sont *connues par construction*, comme dans l'échantillonnage à deux phases (séance 10), on est exactement dans le cadre de Horvitz-Thompson.

3 Estimateur IPW d'une moyenne sous MAR

3.1 Probabilité d'observation connue

Reprenons le cadre de la séance 8 (§ 3.2) : (Y_i, V_i, R_i), $i = 1, \dots, n$, i.i.d., où V est toujours observée et Y observée si $R = 1$. On suppose

(H1) **MAR** : $\mathbb{P}(R = 1 \mid Y, V) = \pi(V)$;

(H2) **positivité** : $\pi(v) \geq \varepsilon > 0$ pour tout v dans le support de V.

Proposition 3.1. *Sous (H1)–(H2), avec π connue, l'estimateur*

$$\widehat{\mu}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi(V_i)}$$

est sans biais pour $\mu = \mathrm{E}(Y)$, et $\sqrt{n}(\widehat{\mu}_{\mathrm{IPW}} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,\ \mathrm{E}\{Y^2/\pi(V)\} - \mu^2\big)$ si $\mathrm{E}(Y^2) < \infty$.

Démonstration. On remarque d'abord que $R_i Y_i$ est toujours calculable : lorsque Y_i manque, $R_i = 0$. Par l'espérance itérée et MAR,

$$\mathrm{E}\Big\{\frac{RY}{\pi(V)}\Big\} = \mathrm{E}\Big\{\frac{Y}{\pi(V)}\ \mathrm{E}(R \mid Y, V)\Big\} = \mathrm{E}\Big\{\frac{Y\,\pi(V)}{\pi(V)}\Big\} = \mathrm{E}(Y) = \mu.$$

La positivité garantit que le quotient est bien défini et que $\mathrm{E}\{R^2Y^2/\pi(V)^2\} = \mathrm{E}\{Y^2/\pi(V)\} \leq \mathrm{E}(Y^2)/\varepsilon < \infty$. Le théorème central limite donne la normalité asymptotique avec la variance $\mathrm{Var}\{RY/\pi(V)\} = \mathrm{E}\{Y^2/\pi(V)\} - \mu^2$. □

La variance $\mathrm{E}\{Y^2/\pi(V)\} - \mu^2$ dépasse $\mathrm{Var}(Y)$: la pondération a un prix, d'autant plus élevé que certaines probabilités $\pi(V)$ sont petites. On voit aussi que la variance de l'estimateur IPW dépend de $\mathrm{E}(Y^2)$ et non de $\mathrm{Var}(Y)$: elle n'est pas invariante par translation de Y, ce qui explique la meilleure stabilité de la version de Hájek $\sum_i R_i Y_i/\pi(V_i) / \sum_i R_i/\pi(V_i)$, invariante par translation.

3.2 Probabilité d'observation estimée

En pratique, π est inconnue. On la modélise, typiquement par une régression logistique $\pi(V;\alpha) = \mathrm{expit}(\alpha^\top \tilde{V})$, où $\tilde{V}$ contient une constante et des fonctions de V, et on l'estime par maximum de vraisemblance avec *toutes* les unités (la réponse R et les prédicteurs V sont toujours observés).

Proposition 3.2. *Si le modèle $\pi(V;\alpha)$ est correctement spécifié, que $\widehat{\alpha}$ est l'EMV et que la positivité tient, alors $\widehat{\mu}_{\mathrm{IPW}}(\widehat{\alpha}) = n^{-1}\sum_i R_i Y_i/\pi(V_i;\widehat{\alpha})$ est convergent et asymptotiquement normal pour μ.*

Démonstration. On empile les équations d'estimation (séance 1) :

$$\sum_{i=1}^{n} \begin{pmatrix} S_i(\alpha) \\ R_i Y_i/\pi(V_i;\alpha) - \mu \end{pmatrix} = 0, \qquad S_i(\alpha) = \{R_i - \pi(V_i;\alpha)\}\tilde{V}_i,$$

où S_i est le score de la régression logistique. Aux vraies valeurs (α_0, μ), les deux composantes sont d'espérance nulle : la première parce que $\mathrm{E}(R \mid V) = \pi(V;\alpha_0)$, la seconde par la proposition 3.1. Sous les conditions de régularité usuelles des estimateurs M (en particulier la positivité, qui borne les dérivées), le système possède une racine convergente et asymptotiquement normale. □

Si le modèle de π est mal spécifié, $\pi(V;\widehat{\alpha})$ converge vers une fonction $\pi^*(V) \neq \pi(V)$ et la limite de l'estimateur, $\mathrm{E}\{Y\pi(V)/\pi^*(V)\}$, diffère en général de μ : **la validité de l'IPW repose entièrement sur le modèle de non-réponse**, comme celle de l'imputation repose sur le modèle des données.

Remarque 3.3 (Poids estimés plus efficaces que poids connus). Résultat contre-intuitif : même lorsque π est connue, l'estimer par un modèle correctement spécifié réduit (ou ne change pas) la variance asymptotique de l'estimateur IPW (Robins, Rotnitzky et Zhao, 1994). Nous le démontrons à la section 7. Une conséquence pratique : l'erreur type qui ignore l'estimation des poids est *conservatrice*.

4 GEE pondérées

4.1 Rappel et problème

Pour des données longitudinales $Y_i = (Y_{i1}, \ldots, Y_{iT})^\top$ avec le modèle marginal $\mathrm{E}(Y_{ij} \mid X_i) = \mu_{ij}(\beta) = g^{-1}(X_{ij}^\top \beta)$, les GEE (séance 6) résolvent

$$U(\beta) = \sum_{i=1}^{n} D_i^\top V_i^{-1}\{Y_i - \mu_i(\beta)\} = 0, \qquad D_i = \partial \mu_i / \partial \beta^\top, \quad V_i = A_i^{1/2} R_i(\varrho) A_i^{1/2}.$$

Leur convergence repose seulement sur $\mathrm{E}\{Y_i - \mu_i(\beta_0) \mid X_i\} = 0$. Avec des données manquantes, les logiciels utilisent les composantes observées : si $\mathcal{O}_i$ désigne l'ensemble des visites observées,

$$U_{\text{obs}}(\beta) = \sum_i D_{i,\mathcal{O}_i}^\top \left(V_{i,\mathcal{O}_i}\right)^{-1} \{Y_{i,\mathcal{O}_i} - \mu_{i,\mathcal{O}_i}(\beta)\}. \tag{1}$$

Prenons la structure d'indépendance pour simplifier : $U_{\text{obs}}(\beta) = \sum_i \sum_j R_{ij}\, d_{ij}\, v_{ij}^{-1}(Y_{ij} - \mu_{ij})$ où d_{ij} et v_{ij} ne dépendent que de X_i. Son espérance fait intervenir

$$\mathrm{E}\{R_{ij}(Y_{ij} - \mu_{ij}) \mid X_i\} = \mathbb{P}(R_{ij} = 1 \mid X_i)\ \mathrm{E}(Y_{ij} - \mu_{ij} \mid R_{ij} = 1, X_i).$$

— Sous **MCAR** (ou, plus généralement, si la non-réponse ne dépend que des covariables X_i du modèle), $\mathrm{E}(Y_{ij} \mid R_{ij} = 1, X_i) = \mathrm{E}(Y_{ij} \mid X_i) = \mu_{ij}$ et l'équation reste sans biais.

— Sous **MAR**, si l'abandon dépend des réponses passées $Y_{i1}, \ldots, Y_{i,j-1}$, corrélées à Y_{ij}, alors $\mathrm{E}(Y_{ij} \mid R_{ij} = 1, X_i) \neq \mu_{ij}$: les sujets encore présents à la visite j forment un sous-groupe sélectionné. Les GEE ordinaires sont biaisées.

C'est la différence fondamentale avec la vraisemblance : les GEE ne sont pas fondées sur une vraisemblance complète, le théorème d'ignorabilité ne s'applique pas, et **les GEE ordinaires exigent MCAR** (Wu, § 3.3.2 ; Fitzmaurice et coll., 2011, ch. 17).

4.2 Les GEE pondérées de Robins, Rotnitzky et Zhao (1995)

On suppose l'attrition monotone et le mécanisme MAR *séquentiel* : la probabilité d'être encore observé à la visite j ne dépend que de l'historique observé,

$$\pi_{ij} = \mathbb{P}(R_{ij} = 1 \mid Y_i, X_i) = \mathbb{P}(R_{ij} = 1 \mid \bar{H}_{i,j-1}), \qquad \bar{H}_{i,j-1} = (X_i, Y_{i1}, \ldots, Y_{i,j-1}),$$

avec $\pi_{ij} \geq \varepsilon > 0$. Les GEE pondérées résolvent

$$U_W(\beta) = \sum_{i=1}^{n} D_i^\top V_i^{-1} W_i \{Y_i - \mu_i(\beta)\} = 0, \qquad W_i = \mathrm{diag}\Big(\frac{R_{i1}}{\pi_{i1}}, \ldots, \frac{R_{iT}}{\pi_{iT}}\Big), \tag{2}$$

où D_i et V_i sont les matrices *complètes* de dimension T, fonctions de X_i seulement. L'équation est calculable : les composantes manquantes de Y_i sont multipliées par $R_{ij} = 0$.

Théorème 4.1 (Robins, Rotnitzky et Zhao, 1995)**.** *Si les π_{ij} sont connues ou estimées de façon convergente par un modèle correct, et si le modèle marginal pour la moyenne est correct, alors $\mathrm{E}\{U_W(\beta_0)\} = 0$ et la solution de (2) est convergente et asymptotiquement normale pour β_0, quelle que soit la matrice de travail V_i (pourvu qu'elle ne dépende que de X_i).*

Démonstration. Conditionnellement aux données complètes (Y_i, X_i), seul W_i est aléatoire, et par MAR $\mathrm{E}(R_{ij}/\pi_{ij} \mid Y_i, X_i) = \pi_{ij}/\pi_{ij} = 1$. Donc $\mathrm{E}(W_i \mid Y_i, X_i) = I_T$ et

$$\mathrm{E}\{U_W(\beta_0)\} = \mathrm{E}\big[D_i^\top V_i^{-1}\ \mathrm{E}(W_i \mid Y_i, X_i)\{Y_i - \mu_i(\beta_0)\}\big] = \mathrm{E}\big[D_i^\top V_i^{-1}\{Y_i - \mu_i(\beta_0)\}\big] = 0,$$

la dernière égalité étant la condition de validité des GEE en données complètes. On conclut par la théorie des estimateurs M ; si les poids sont estimés, on empile les scores du modèle de non-réponse (section 7). □

La pondération rend donc l'équation d'estimation sans biais « en moyenne sur le mécanisme de non-réponse », exactement comme l'estimateur de Horvitz-Thompson. Une variante pondère au niveau du sujet : seuls les sujets ayant complété l'étude ($R_{iT} = 1$) contribuent, avec le poids $1/\mathbb{P}(R_{iT} = 1 \mid Y_i, X_i)$. Elle est sans biais pour toute structure de travail, mais gaspille l'information des sujets partiellement observés ; la pondération par observation (2) est plus efficace.

5 Attrition et pondération par probabilité inverse de censure

5.1 L'idée sur un exemple

En raison des pertes au suivi, les sujets restants au temps t ne sont plus représentatifs de la cohorte initiale. On veut que les sujets surreprésentés reçoivent un poids inférieur à celui des sujets sous-représentés, de sorte que l'échantillon pondéré des sujets restants « ressemble » à la cohorte de départ.

Exemple 5.1 (100 femmes et 100 hommes)**.** Au début de l'étude, on compte 100 femmes et 100 hommes ; à la deuxième visite, il reste 75 femmes et 25 hommes. On attribue à chaque femme restante le poids $1/0{,}75 = 1{,}33$ et à chaque homme restant le poids $1/0{,}25 = 4$. Le poids total des femmes vaut $75 \times 1{,}33 = 100$ et celui des hommes $25 \times 4 = 100$: l'échantillon pondéré retrouve la composition initiale. Si l'issue diffère selon le sexe, la moyenne pondérée corrige le biais que produirait la moyenne non pondérée, dominée par les femmes. Ici, la probabilité de rester est estimée dans chaque strate par une proportion ; avec plusieurs variables, on la modélise par régression logistique.

5.2 Notation et poids

Soit C_{it} l'indicatrice de censure : $C_{it} = 1$ si le sujet i n'est plus dans l'étude au temps t, $C_{it} = 0$ sinon. On suppose qu'un sujet perdu au suivi ne revient pas (patron monotone) et que tous les sujets sont présents à la première visite ($C_{i1} = 0$). Soit $\bar{Z}_t$ l'ensemble des variables mesurées jusqu'à la visite t inclusivement qui servent à expliquer la censure : covariables X et réponses $Y_1, \ldots, Y_t$ (et d'éventuelles variables auxiliaires).

Visite 2. On estime $\mathbb{P}(C_{i2} = 0 \mid \bar{Z}_1 = \bar{z}_{i1})$ par une régression logistique dont la réponse vaut 1 si le sujet est encore présent à la visite 2 et 0 sinon, avec comme prédicteurs des variables mesurées à la visite 1.

Visite 3. Par la loi des probabilités totales,

$$\begin{aligned}\mathbb{P}(C_{i3}=0 \mid \bar{Z}_2) = \mathbb{P}(C_{i3}=0 \mid C_{i2}=0, \bar{Z}_2)\,\mathbb{P}(C_{i2}=0 \mid \bar{Z}_1) \\ + \mathbb{P}(C_{i3}=0 \mid C_{i2}=1, \bar{Z}_2)\,\mathbb{P}(C_{i2}=1 \mid \bar{Z}_1),\end{aligned}$$

et le second terme est nul puisqu'un sujet perdu ne revient pas. Le facteur $\mathbb{P}(C_{i3}=0 \mid C_{i2}=0, \bar{Z}_2)$ s'estime par une régression logistique ajustée sur les seuls sujets présents aux deux premières visites (l'*ensemble à risque*), avec des prédicteurs mesurés aux visites 1 et 2.

Cas général. Par récurrence, la probabilité d'être encore présent à la visite t est un produit de probabilités conditionnelles, et le poids IPCW est

$$\mathrm{IPCW}_{it} = \prod_{k=2}^{t} \frac{1}{\mathbb{P}(C_{ik}=0 \mid C_{i,k-1}=0,\ \bar{Z}_{i,k-1})}, \qquad t=2,\dots,T, \tag{3}$$

avec $\mathrm{IPCW}_{i1}=1$. On peut ajuster un modèle logistique distinct pour chaque visite ou un seul modèle « groupé » sur l'ensemble à risque en format long (une ligne par sujet et par visite à risque), avec des indicatrices de visite. Il faut, pour chaque ligne, disposer des variables de la visite précédente : on travaille en format large ou on ajoute en format long la réponse précédente comme variable (décalage).

Poids stabilisés. Les produits (3) peuvent devenir très grands, ce qui rend les estimations instables. On les stabilise par un numérateur qui ne dépend pas de l'historique de la réponse :

$$\mathrm{SIPCW}_{it} = \prod_{k=2}^{t} \frac{\mathbb{P}(C_{ik}=0 \mid C_{i,k-1}=0)}{\mathbb{P}(C_{ik}=0 \mid C_{i,k-1}=0,\ \bar{Z}_{i,k-1})}.$$

Le numérateur peut aussi dépendre des covariables de base X *qui figurent dans le modèle marginal* : multiplier les poids par une fonction de X et de t seulement ne biaise pas l'équation (2), car cela revient à changer la matrice de travail par une matrice diagonale fonction de X_i. Le numérateur ne doit pas dépendre des réponses passées.

Poids normalisés. Pour ne pas gonfler artificiellement la taille d'échantillon, on peut normaliser les poids à chaque visite :

$$\mathrm{NIPCW}_{it} = \frac{\mathrm{SIPCW}_{it}}{\sum_{i=1}^{n_t} \mathrm{SIPCW}_{it}/n_t},$$

où la somme porte sur les n_t sujets présents à la visite t. La moyenne des poids normalisés vaut 1 à chaque visite. Avec la structure d'indépendance et un modèle saturé en visite (paramètres propres à chaque visite), une constante multiplicative par visite ne change pas les estimations ponctuelles, mais elle peut changer les estimations de la variance fondées sur le modèle et la stabilité numérique ; dans les autres cas (p. ex. une tendance linéaire en t), elle modifie aussi les estimations ponctuelles, sans les biaiser puisqu'elle ne dépend que de t.

5.3 Pourquoi seule la structure d'indépendance est valide en pratique

Le document du H26 met en garde : avec des poids variant dans le temps, la seule structure de corrélation de travail qui donne des résultats valides est l'indépendance. Le théorème 4.1 semble

dire le contraire ; la clé est la façon dont les logiciels (`geeglm`, `gee`, PROC GENMOD avec `WEIGHT`) utilisent les poids. Ils résolvent

$$\sum_i D_{i,\mathcal{O}_i}^\top \left(V_{i,\mathcal{O}_i}^w\right)^{-1} \{Y_{i,\mathcal{O}_i} - \mu_{i,\mathcal{O}_i}\},$$

où $V_{i,\mathcal{O}_i}^w$ est la matrice de travail *restreinte aux visites observées* et dans laquelle les poids interviennent comme des poids de précision (en divisant les variances, d'où des facteurs $w_{ij}^{1/2} w_{ik}^{1/2}$ hors diagonale). Deux problèmes surgissent dès que V n'est pas diagonale.

1. L'inverse d'une sous-matrice n'est pas la sous-matrice de l'inverse : le coefficient qui multiplie le résidu $Y_{ij} - \mu_{ij}$ dépend de l'ensemble $\mathcal{O}_i$ des visites observées, donc des indicatrices *futures* R_{ik}, $k > j$. Or celles-ci dépendent de Y_{ij} sous MAR séquentiel (un Y_{ij} élevé provoque l'abandon). L'argument $\mathrm{E}(R_{ij}/\pi_{ij} \mid Y_i, X_i) = 1$, qui portait sur un facteur à la fois, ne s'applique plus au produit d'un poids et d'un coefficient aléatoire corrélé au résidu.

2. Les termes croisés $w_{ij}^{1/2} w_{ik}^{1/2} (Y_{ik} - \mu_{ik})$ font intervenir le poids de la visite $j > k$, qui dépend de Y_{ik} par le modèle de censure ; leur espérance n'est pas nulle.

Avec la structure d'indépendance, V_i est diagonale : le coefficient de chaque résidu ne dépend que de X_i, l'équation devient $\sum_i \sum_j w_{ij} R_{ij}\, d_{ij} v_{ij}^{-1} (Y_{ij} - \mu_{ij})$, et chaque terme est sans biais par l'argument du théorème 4.1. Une implantation fidèle de (2) avec une matrice de travail non diagonale est possible (on calcule V_i^{-1} sur les T visites, puis on multiplie par W_i), mais elle n'est pas offerte par les fonctions usuelles. La perte d'efficacité de l'indépendance est le prix de la simplicité ; l'estimateur augmenté (section 9) permet de la récupérer.

6 Exemple en R : attrition MAR

On simule $n = 400$ sujets mesurés aux visites $t = 0{,}1{,}2{,}3$, répartis au hasard en deux groupes ($X \in \{0{,}1\}$). Le modèle marginal est

$$\mathrm{E}(Y_{it} \mid X_i) = 10 + t - 0{,}5\, X_i + 0{,}5\, X_i t,$$

avec un effet aléatoire de sujet et une erreur en marche aléatoire (corrélation non échangeable). À chaque visite $t \geq 1$, un sujet encore présent abandonne avec probabilité $\mathrm{expit}\{-1{,}5 + 0{,}6\,(Y_{i,t-1} - 12) - 0{,}5\, X_i\}$: les sujets dont la réponse précédente est élevée abandonnent davantage. Le mécanisme est MAR (séquentiel) et monotone. On compare les GEE sur les données complètes (inaccessibles en pratique), les GEE non pondérées sur les données observées et les GEE pondérées.

```r
library(geepack)
d <- simuler(n = 400)                          # format long : id, t, X, Y, Yprec,
    present
tapply(d$present, d$t, sum)
## ensemble à risque : visites t >= 1 des sujets présents à la visite précédente
d$risque <- d$t > 0 & ave(d$present, d$id, FUN = function(r) c(0, r[-4])) == 1
ar <- d[d$risque, ]
m.den <- glm(present ~ factor(t) + Yprec + X, family = binomial, data = ar)
m.num <- glm(present ~ factor(t) + X, family = binomial, data = ar)
d$p.den <- d$p.num <- 1
d$p.den[d$risque] <- predict(m.den, type = "response")
d$p.num[d$risque] <- predict(m.num, type = "response")
```

```r
d$w  <- ave(1 / d$p.den, d$id, FUN = cumprod)              # IPCW, produit cumulatif
d$sw <- ave(d$p.num / d$p.den, d$id, FUN = cumprod)        # poids stabilisés
obs <- d[d$present == 1, ]
```

```
  0   1   2   3                     # sujets présents par visite
400 373 341 280
# modèle de P(C_t = 0 | C_{t-1} = 0, Y_{t-1}, X)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    8.419      0.789  10.667    0.000
factor(t)2     0.350      0.291   1.202    0.229
factor(t)3     0.121      0.293   0.414    0.679
Yprec         -0.575      0.070  -8.247    0.000
X              0.402      0.210   1.911    0.056
      0    1    2    3               # résumé des poids par visite
w     1 1.07 1.17 1.45               #   moyenne IPCW
sw    1 1.00 1.00 1.01               #   moyenne des poids stabilisés
w.max 1 1.76 3.05 6.05               #   poids IPCW maximal
Somme des poids w par visite (cible n = 400) :
400.0 399.9 399.3 405.3
```

Le modèle de censure retrouve la dépendance à la réponse précédente. Deux diagnostics sont rassurants : la somme des poids IPCW à chaque visite est proche de l'effectif initial (400), et la moyenne des poids stabilisés est proche de 1. Le poids maximal à la dernière visite (6,05) indique un sujet qui en représente six ; rien d'alarmant ici.

```r
f.ind  <- geeglm(Y ~ t * X, id = id, data = obs, corstr = "independence")
f.ech  <- geeglm(Y ~ t * X, id = id, data = obs, corstr = "exchangeable")
f.w    <- geeglm(Y ~ t * X, id = id, data = obs, weights = w,  corstr = "independence")
f.sw   <- geeglm(Y ~ t * X, id = id, data = obs, weights = sw, corstr = "independence")
f.wech <- geeglm(Y ~ t * X, id = id, data = obs, weights = w,  corstr = "exchangeable")
summary(f.w)$coefficients
```

```
             Estimate    Std.err        Wald     Pr(>|W|)
(Intercept)  9.8855637 0.12161070 6607.837112 0.000000e+00
t            1.0866107 0.05935062  335.194896 0.000000e+00
X           -0.4780209 0.16983593    7.921997 4.883733e-03
t:X          0.3975916 0.07363598   29.153733 6.685681e-08
```

TABLE 1 – Estimations (erreurs types sandwich) des coefficients du modèle marginal sur le jeu simulé. Vraies valeurs : 10 ; 1 ; −0,5 ; 0,5.

Méthode (structure de travail)	Constante	t	X	$t \times X$
Données complètes (ind.)	9,909 (0,120)	1,025 (0,028)	−0,499 (0,166)	0,507 (0,041)
Non pondérée (ind.)	9,939 (0,121)	0,872 (0,041)	−0,506 (0,168)	0,515 (0,059)
Non pondérée (éch.)	9,908 (0,120)	1,018 (0,032)	−0,487 (0,167)	0,493 (0,045)
IPCW (ind.)	9,886 (0,122)	1,087 (0,059)	−0,478 (0,170)	0,398 (0,074)
IPCW stabilisés (ind.)	9,892 (0,121)	1,081 (0,057)	−0,485 (0,169)	0,404 (0,072)
IPCW (éch.), à éviter	10,455 (0,182)	1,091 (0,034)	−0,812 (0,218)	0,457 (0,047)

Le tableau 1 et la figure 1 montrent que :

— la GEE non pondérée avec indépendance sous-estime nettement la pente (0,87 au lieu de 1) : les sujets dont la réponse augmente le plus quittent l'étude ;

— les GEE pondérées avec indépendance éliminent ce biais systématique (sur ce jeu, l'écart à la vraie pente, 0,087, est de l'ordre de 1,5 erreur type ; la simulation ci-dessous confirme l'absence de biais), au prix d'une erreur type plus grande (0,059 contre 0,041) ; les poids stabilisés réduisent un peu la variance ;

— la GEE pondérée avec structure échangeable donne des estimations aberrantes pour la constante et l'effet de groupe, conformément à la section 5.3 ;

— la GEE *non pondérée* avec structure échangeable est ici presque sans biais. Ce n'est pas une contradiction : avec une structure de travail non diagonale, les résidus observés aux visites précédentes « prédisent » les visites manquantes, comme dans une vraisemblance. Si la matrice de travail et la loi des données étaient correctes (modèle normal à corrélation bien spécifiée), cette GEE coïnciderait avec l'EMV, valide sous MAR. Mais rien ne garantit ce comportement lorsque la structure est mal spécifiée ou la réponse non gaussienne : on ne peut pas s'y fier en général (voir la simulation ci-dessous).

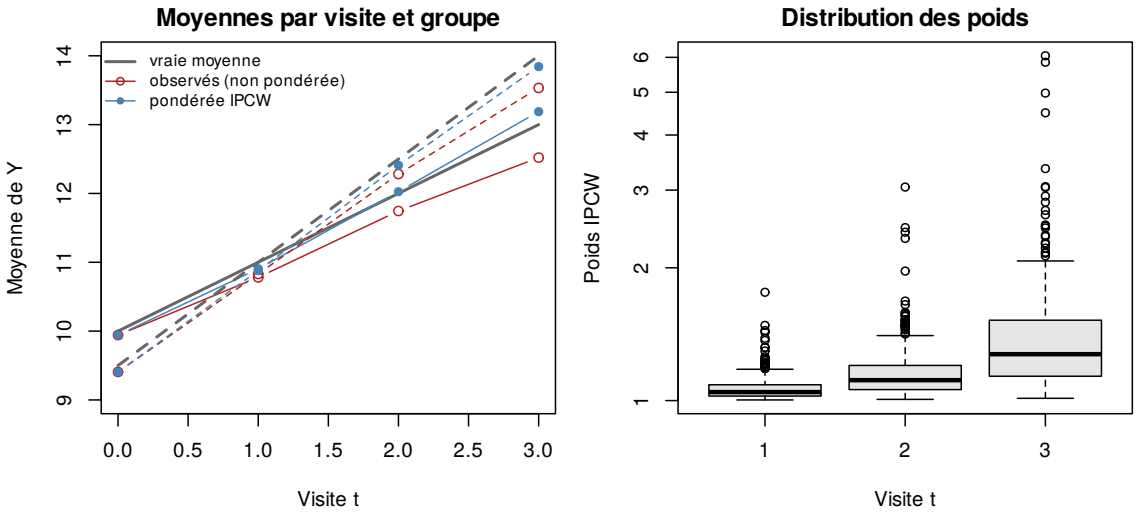


FIGURE 1 – À gauche : moyennes de Y par visite et par groupe (trait plein : $X = 0$; pointillé : $X = 1$), vraies (gris), observées non pondérées (rouge) et pondérées par IPCW (bleu). À droite : distribution des poids IPCW aux visites 1 à 3 (échelle logarithmique).

Simulation. Sur 200 jeux de données simulés selon le même modèle, on obtient pour la pente t (vraie valeur 1) les résultats du tableau 2.

La GEE non pondérée à structure échangeable a un biais modeste mais réel (couverture de 80 %) ; les GEE pondérées avec indépendance sont sans biais avec une couverture proche de 95 % ; la GEE pondérée échangeable est biaisée et son erreur type sandwich sous-estime fortement sa variabilité réelle.

Table 2 – Simulation (200 réplications, $n = 400$) : biais, écart type empirique, moyenne des erreurs types sandwich et couverture de l'intervalle à 95 % pour la pente t ; biais pour l'interaction $t \times X$.

Méthode (structure)	Biais (t)	ET emp.	ET moy.	Couverture	Biais ($t \times X$)
Données complètes (ind.)	0,000	0,028	0,029	0,960	−0,001
Non pondérée (ind.)	−0,222	0,046	0,046	0,000	0,046
Non pondérée (éch.)	−0,038	0,033	0,033	0,795	0,008
IPCW (ind.)	−0,001	0,079	0,076	0,935	−0,012
IPCW stabilisés (ind.)	0,000	0,074	0,073	0,945	−0,012
IPCW (éch.)	0,041	0,096	0,050	0,920	−0,008

7 Estimation de la variance

7.1 Sandwich en ignorant l'estimation des poids

La pratique la plus courante consiste à traiter les poids estimés comme fixes et à utiliser l'estimateur sandwich de la GEE pondérée, $\widehat{A}^{-1}\widehat{B}\widehat{A}^{-\top}$ avec $\widehat{A} = -n^{-1}\sum_i \partial U_{W,i}/\partial\beta^\top$ et $\widehat{B} = n^{-1}\sum_i U_{W,i}U_{W,i}^\top$. C'est ce que fait `geeglm`. Cette erreur type est *conservatrice* lorsque le modèle de non-réponse est correctement spécifié et estimé par maximum de vraisemblance.

Proposition 7.1 (Robins, Rotnitzky et Zhao, 1994)**.** *Soit $\widehat{\beta}$ la solution de $\sum_i U_{W,i}(\beta, \widehat{\alpha}) = 0$, où $\widehat{\alpha}$ est l'EMV d'un modèle correct de non-réponse de score $S_i(\alpha)$. Alors la variance asymptotique de $\sqrt{n}(\widehat{\beta} - \beta_0)$ est*

$$A^{-1}\operatorname{Var}\{U_W - \Pi(U_W \mid S)\}A^{-\top} \preceq A^{-1}\operatorname{Var}(U_W)A^{-\top},$$

où $\Pi(U_W \mid S) = \mathrm{E}(U_W S^\top)\{\mathrm{E}(SS^\top)\}^{-1}S$ est la projection linéaire de U_W sur le score S.

Démonstration. On empile $\sum_i S_i(\alpha) = 0$ et $\sum_i U_{W,i}(\beta, \alpha) = 0$. Par le développement habituel des estimateurs M (séance 1), $\sqrt{n}(\widehat{\alpha} - \alpha_0) = I_\alpha^{-1} n^{-1/2}\sum_i S_i + o_p(1)$ avec $I_\alpha = \mathrm{E}(SS^\top)$, et

$$\begin{aligned}\sqrt{n}(\widehat{\beta} - \beta_0) &= A^{-1}\Big\{\frac{1}{\sqrt{n}}\sum_i U_{W,i} + \mathrm{E}\Big(\frac{\partial U_W}{\partial \alpha^\top}\Big)\sqrt{n}(\widehat{\alpha} - \alpha_0)\Big\} + o_p(1)\\ &= A^{-1}\frac{1}{\sqrt{n}}\sum_i\Big\{U_{W,i} + \mathrm{E}\Big(\frac{\partial U_W}{\partial \alpha^\top}\Big)I_\alpha^{-1}S_i\Big\} + o_p(1).\end{aligned}$$

Il reste à montrer que $\mathrm{E}(\partial U_W/\partial\alpha^\top) = -\mathrm{E}(U_W S^\top)$. Pour toute valeur de α, l'équation pondérée par $\pi(\cdot;\alpha)$ est sans biais sous la loi où le mécanisme de non-réponse est $\pi(\cdot;\alpha)$ (la loi des données complètes ne dépend pas de α) : $\int U_W(\beta_0, \alpha)\, p(\text{données}; \alpha) = 0$. En dérivant par rapport à α sous le signe intégral, $\mathrm{E}(\partial U_W/\partial\alpha^\top) + \mathrm{E}(U_W S^\top) = 0$. Ainsi la fonction d'influence de $\widehat{\beta}$ est $A^{-1}\{U_W - \mathrm{E}(U_W S^\top)I_\alpha^{-1}S\} = A^{-1}\{U_W - \Pi(U_W \mid S)\}$. Comme un résidu de projection est non corrélé avec la projection, $\operatorname{Var}(U_W) = \operatorname{Var}\{U_W - \Pi(U_W \mid S)\} + \operatorname{Var}\{\Pi(U_W \mid S)\}$, d'où l'inégalité. □

L'interprétation est remarquable : estimer α « retire » de l'équation la partie de son bruit qui est expliquée par le mécanisme de non-réponse. Plus le modèle de non-réponse est riche (plus il contient de prédicteurs liés à Y), plus le gain est grand. Ignorer cette correction donne des intervalles un peu trop larges, ce qui est rarement grave.

7.2 Correction

Deux options permettent d'obtenir une erreur type correcte : (i) programmer l'estimateur sandwich du système empilé (score logistique et GEE pondérée), dont la fonction d'influence est donnée dans la preuve ; (ii) utiliser le *bootstrap par sujets*, en réestimant les poids à chaque réplique. Sur notre exemple, avec 300 répliques :

```
          (Intercept)      t      X    t:X
bootstrap      0.1218 0.0489 0.1691 0.0700
sandwich       0.1216 0.0594 0.1698 0.0736
```

Comme prévu par la proposition 7.1, l'erreur type sandwich qui traite les poids comme fixes est plus grande que l'erreur type bootstrap, surtout pour la pente t (0,059 contre 0,049), le coefficient le plus touché par l'attrition. Pour la constante et l'effet de groupe à la visite initiale, où les poids valent 1, les deux coïncident.

8 Poids extrêmes et troncature

Les poids extrêmes surviennent lorsque certaines probabilités π_{it} sont proches de 0 : quasi-violation de la positivité (certains profils ne restent presque jamais dans l'étude), modèle de non-réponse trop flexible, ou accumulation du produit (3) sur de nombreuses visites. Quelques unités dominent alors l'estimation, dont la variance explose et dont la loi s'éloigne de la normale. Pistes de solution :

- **diagnostiquer** : examiner la distribution des poids par visite (moyenne des poids stabilisés proche de 1, maximum, taille d'échantillon effective $(\sum w)^2 / \sum w^2$) ;
- **stabiliser** les poids (section 5) ;
- **tronquer** (ou « écrêter ») les poids à un centile élevé, par exemple le 99$^{\text{e}}$, ou à une valeur fixe ;
- revoir le modèle de non-réponse (forme fonctionnelle, interactions) ou restreindre la population cible lorsque la positivité est violée structurellement ;
- utiliser un estimateur augmenté (section 9), moins sensible aux petites probabilités.

La troncature réalise un compromis biais–variance : elle réduit la variance, mais réintroduit une partie du biais de sélection, puisque les poids tronqués ne corrigent plus complètement la sous-représentation. Sur l'exemple :

```
Quantiles 95 % et 99 % des poids : 1.58 2.47
      aucune 99e centile 95e centile
t      1.087       1.021       0.953
se.t   0.059       0.046       0.041
tX     0.398       0.463       0.506
se.tX  0.074       0.063       0.059
```

Plus on tronque, plus l'erreur type diminue et plus l'estimation de la pente se rapproche de l'estimation non pondérée biaisée (0,87). Il est recommandé de présenter une analyse de sensibilité à plusieurs seuils de troncature.

9 Estimateurs augmentés et double robustesse

L'IPW n'utilise les variables V des unités incomplètes que pour estimer π, ce qui est inefficace. L'estimateur *augmenté* (AIPW ; Robins, Rotnitzky et Zhao, 1994) ajoute un terme fondé sur un modèle de régression $m(V) = \mathrm{E}(Y \mid V)$ (ou $\mathrm{E}(Y \mid V, R = 1)$, égal sous MAR) :

$$\widehat{\mu}_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{R_i Y_i}{\widehat{\pi}(V_i)} - \frac{R_i - \widehat{\pi}(V_i)}{\widehat{\pi}(V_i)}\,\widehat{m}(V_i)\Big\} = \frac{1}{n}\sum_{i=1}^{n}\Big\{\widehat{m}(V_i) + \frac{R_i}{\widehat{\pi}(V_i)}\big(Y_i - \widehat{m}(V_i)\big)\Big\}. \tag{4}$$

La seconde écriture le montre : on part de l'estimateur par régression (imputation par $\widehat{m}$) et on corrige son biais éventuel par une moyenne pondérée des résidus observés.

Proposition 9.1 (Double robustesse)**.** *Sous MAR et positivité, si $\widehat{\pi} \xrightarrow{\mathbb{P}} \pi^*$ et $\widehat{m} \xrightarrow{\mathbb{P}} m^*$, la limite de $\widehat{\mu}_{\mathrm{AIPW}}$ est μ dès que $\pi^* = \pi$ ou $m^* = m$.*

Démonstration. La limite est $\mathrm{E}\{m^*(V) + R\,(Y - m^*(V))/\pi^*(V)\}$. Par l'espérance itérée sur (Y, V) et MAR, elle vaut

$$\mu + \mathrm{E}\Big[\{m(V) - m^*(V)\}\Big\{\frac{\pi(V)}{\pi^*(V)} - 1\Big\}\Big],$$

car $\mathrm{E}\{R(Y - m^*)/\pi^* \mid V\} = \pi(V)\{m(V) - m^*(V)\}/\pi^*(V)$ et $\mathrm{E}\{m^*(V)\} = \mu - \mathrm{E}\{m(V) - m^*(V)\}$. Le terme de biais est un produit de deux erreurs : il s'annule si $m^* = m$ ou si $\pi^* = \pi$. □

La structure « produit d'erreurs » a une autre conséquence : si les deux modèles sont estimés avec des vitesses n^{-a} et n^{-b} avec $a + b > 1/2$ (par exemple deux vitesses $o(n^{-1/4})$ obtenues par apprentissage automatique), l'AIPW reste $\sqrt{n}$-convergent. C'est le fondement de l'estimation ciblée et de l'apprentissage automatique « double » (séance 7). Lorsque les deux modèles sont corrects, l'AIPW atteint la borne d'efficacité semi-paramétrique du modèle où seul le mécanisme MAR est spécifié. Pour les données longitudinales, il existe des versions augmentées des GEE pondérées, qui récupèrent l'efficacité perdue par la structure d'indépendance.

```
ph <- predict(glm(R ~ V1 + V2, family = binomial, data = d), type = "response")
mh <- predict(lm(Y ~ V1 + I(V1^2) + V2, data = d[d$R == 1, ]), newdata = d)
y0 <- ifelse(d$R == 1, d$Y, 0)
c(IPW = mean(d$R * y0 / ph), Hajek = sum(d$R * y0 / ph) / sum(d$R / ph),
  Reg = mean(mh), AIPW = mean(d$R * y0 / ph - (d$R - ph) / ph * mh))
```

On illustre la double robustesse par 1000 simulations ($n = 500$) où $V = (V_1, V_2)$ est toujours observée, $Y = 2 + V_1 + 0{,}8V_1^2 + 0{,}5V_2 + \varepsilon$ (donc $\mu = 2{,}8$), et $\mathrm{logit}\,\pi(V) = 0{,}3 + 1{,}2V_1 - 0{,}4V_2$. Le modèle « mal » spécifié de π omet V_1 ; celui de m ne contient que V_2.

TABLE 3 – Biais (écart type empirique) de quatre estimateurs de $\mu = 2{,}8$ selon la spécification des modèles de π et de m ; 1000 réplications, $n = 500$. L'estimateur des cas complets a un biais de 0,313.

Scénario	IPW	Hájek	Régression	AIPW
π correct, m correct	−0,007 (0,176)	−0,005 (0,126)	0,000 (0,099)	0,000 (0,103)
π incorrect, m correct	0,395 (0,125)	0,395 (0,125)	0,000 (0,099)	0,000 (0,099)
π correct, m incorrect	−0,007 (0,176)	−0,005 (0,126)	0,394 (0,124)	−0,002 (0,121)
π incorrect, m incorrect	0,395 (0,125)	0,395 (0,125)	0,394 (0,124)	0,394 (0,125)

Le tableau 3 illustre la proposition 9.1 : l'AIPW est sans biais dès qu'un des deux modèles est correct, et il échoue seulement si les deux sont faux. Lorsque les deux modèles sont corrects, il est presque aussi précis que l'estimateur par régression (efficace dans ce cas) et beaucoup plus précis que l'IPW ; l'estimateur de Hájek est nettement plus stable que l'IPW non normalisé.

Niveau doctorat : la théorie de Tsiatis pour les données manquantes

Soit $\varphi^F(Z)$ la fonction d'influence d'un estimateur régulier et asymptotiquement linéaire (RAL) de β en données complètes (séance 6), et considérons un mécanisme MAR à deux niveaux ($R \in \{0,1\}$, V toujours observée, Z observée si $R = 1$) avec $\pi(V) \geq \varepsilon > 0$. Tsiatis (2006, ch. 7 et 10) montre que, lorsque π est connue, les fonctions d'influence des estimateurs RAL en données observées sont exactement

$$\frac{R\,\varphi^F(Z)}{\pi(V)} + \frac{R - \pi(V)}{\pi(V)}\,h(V), \qquad h \text{ arbitraire (de carré intégrable)},$$

où $\{(R - \pi(V))h(V)/\pi(V)\}$ est l'*espace d'augmentation* Λ_2. Lorsque π est modélisée par $\pi(V;\alpha)$, il faut en plus soustraire la projection sur l'espace tangent $\Lambda_\psi \subset \Lambda_2$ du modèle de non-réponse (c'est la proposition 7.1). Pour φ^F fixée, l'élément de variance minimale s'obtient en projetant $R\varphi^F/\pi$ sur Λ_2 : $h_{\text{opt}}(V) = -\,\mathrm{E}\{\varphi^F(Z) \mid V\}$, ce qui donne précisément l'estimateur AIPW (4) lorsque $\varphi^F(Z) = Y - \mu$. L'estimateur efficace s'obtient en optimisant ensuite sur φ^F ; avec un patron monotone à plusieurs visites, Λ_2 s'écrit comme une somme sur les temps de censure de termes $\{dN_C(t) - \lambda_C(t \mid \bar{H}_t)Y(t)\,\mathrm{d}t\}\,h_t(\bar{H}_t)/K(t \mid \bar{H}_t)$, et on retrouve les GEE pondérées augmentées. Les conditions de régularité comprennent la positivité stricte, la dérivabilité de $\pi(V;\alpha)$, et, si m et π sont estimés de façon non paramétrique, des conditions de type Donsker ou un ajustement croisé (*cross-fitting*).

10 Comparaison des approches

Aucune méthode ne traite MNAR sans hypothèse supplémentaire non vérifiable ; toutes peuvent servir de base à une analyse de sensibilité (séance 8, § 4.3). Le choix dépend surtout de ce qu'on juge le plus facile à modéliser correctement : la loi des données (vraisemblance, imputation) ou le mécanisme de non-réponse (pondération). Avec un patron monotone dû à l'attrition et un modèle marginal estimé par GEE, la pondération est naturelle ; avec des valeurs manquantes éparses dans plusieurs covariables, l'imputation multiple est plus pratique.

Lien avec l'échantillonnage à deux phases (séance 10). Dans un plan à deux phases, on mesure des variables peu coûteuses chez tous les sujets (phase 1) et une variable coûteuse (biomarqueur, exposition mesurée en laboratoire) chez un sous-échantillon (phase 2) tiré selon des probabilités qui dépendent des variables de phase 1 (par exemple, du statut cas–témoin et d'une strate). Les données de phase 2 sont manquantes *par construction* chez les autres sujets ; le mécanisme est MAR, et les probabilités de sélection sont connues. L'estimateur de vraisemblance pondérée de Breslow et Chatterjee (1999) est l'analogue exact de Horvitz-Thompson et des GEE pondérées de cette séance ; il reste convergent sans modèle pour la loi de la variable coûteuse. Les méthodes de pseudo-vraisemblance et de vraisemblance semi-paramétrique comparées dans cet article sont plus efficaces, au prix d'hypothèses supplémentaires, et l'estimation des poids (même connus) par calibration ou post-stratification améliore l'efficacité pour la raison donnée à la proposition 7.1.

TABLE 4 – Hypothèses et propriétés des principales méthodes de traitement des données manquantes.

	Cas complets	Vraisemblance (EMV, modèles mixtes)	Imputation multiple	IPW / GEE pondérées
Mécanisme requis	MCAR (ou non-réponse dépendant des seules covariables)	MAR + paramètres distincts	MAR	MAR
Modèle à spécifier correctement	modèle d'analyse	loi conjointe des données (y compris des covariables manquantes)	modèle d'imputation (lois conditionnelles ou conjointe), congénère à l'analyse	modèle de non-réponse π + modèle marginal de la moyenne
Variables auxiliaires	non	difficile	oui, facilement	oui (dans π)
Efficacité	faible	maximale si modèle correct	proche de l'EMV	faible à moyenne (indépendance)
Inférence	habituelle	information observée	règles de Rubin	sandwich (conservateur) ou bootstrap
Points faibles	biais hors MCAR	mauvaise spécification de la loi conjointe	compatibilité, congénialité	poids extrêmes, positivité
Double robustesse	–	–	–	oui avec AIPW

11 Synthèse

Grille méthodologique

1. **Objet d'intérêt** : paramètres β d'un modèle marginal (moyenne de population) ; ou une moyenne μ.
2. **Nuisance, données manquantes, biais** : réponses manquantes par attrition ; le mécanisme MAR qui sélectionne les sujets restants biaise les GEE ordinaires. Le mécanisme π est un paramètre de nuisance qu'on modélise ; la loi conjointe des réponses reste non spécifiée.
3. **Approche** : repondérer les observations par l'inverse de leur probabilité d'être observées (Horvitz-Thompson) ; éventuellement augmenter par un modèle de régression (AIPW).
4. **Critère** : équations d'estimation pondérées $\sum_i D_i^\top V_i^{-1} W_i (Y_i - \mu_i) = 0$, structure d'indépendance en pratique ; scores logistiques pour les poids.
5. **Calcul** : régressions logistiques séquentielles sur les ensembles à risque, produit cumulatif ; GEE pondérée (Fisher scoring) avec `geeglm(..., weights = w, corstr = "independence")`.
6. **Inférence** : sandwich en ignorant l'estimation des poids (conservateur), sandwich empilé ou bootstrap par sujets avec réestimation des poids.

À retenir

- Horvitz-Thompson : une unité observée avec probabilité π représente $1/\pi$ unités ; $\mathrm{E}\{RY/\pi(V)\} = \mathrm{E}(Y)$ sous MAR et positivité.
- Les GEE ordinaires exigent MCAR (ou une non-réponse dépendant des seules covariables du modèle) ; les GEE pondérées de Robins, Rotnitzky et Zhao sont valides sous MAR si le modèle de non-réponse est correct.
- Poids IPCW : produit cumulatif des inverses des probabilités conditionnelles de rester, estimées par régressions logistiques sur les ensembles à risque ; poids stabilisés (numérateur sans historique de la réponse) et normalisés.
- Avec des poids variant dans le temps et les logiciels usuels, seule la structure de travail d'indépendance donne des résultats valides.
- L'erreur type sandwich qui ignore l'estimation des poids est conservatrice ; le bootstrap qui réestime les poids est correct.
- Les poids extrêmes signalent une quasi-violation de la positivité ; la troncature échange du biais contre de la variance.
- L'AIPW est doublement robuste : convergent si le modèle de non-réponse *ou* le modèle de régression est correct.

12 Exercices

Exercice 12.1 (Exercice d'imputation H26, volet pondération)**.** On reprend les données de croissance dentaire avec valeurs manquantes (`dentaire_manquant.txt`, exercice 8.2).

a) Vérifiez que le patron de données manquantes est monotone. Pour chaque âge $t \in \{10{,}12{,}14\}$, comparez la distribution du sexe et des mesures antérieures entre les enfants observés et non observés à l'âge t.

b) Ajustez, pour chaque âge $t \in \{10{,}12{,}14\}$, une régression logistique de l'indicatrice d'observation à l'âge t en fonction du sexe et de la mesure à l'âge précédent, sur les enfants observés à l'âge précédent.

c) Calculez les poids par probabilité inverse de censure (produit cumulatif) et estimez le modèle `y ~ age * sexe` par GEE pondérées avec `geeglm`. Quelle structure de corrélation de travail devez-vous utiliser ? Pourquoi ?

d) Comparez avec les données pleines, l'analyse des cas complets et l'imputation multiple de l'exercice 8.2. Commentez les erreurs types.

Exercice 12.2 (Diapositives H26 sur la pondération ; complété)**.** On reprend l'exemple 5.1 : 100 femmes et 100 hommes au départ, 75 femmes et 25 hommes à la visite 2, puis 60 femmes et 15 hommes à la visite 3.

a) Calculez les poids IPCW de chaque sujet aux visites 2 et 3 en supposant que la probabilité de rester ne dépend que du sexe.

b) Calculez les poids stabilisés avec un numérateur $\mathbb{P}(C_k = 0 \mid C_{k-1} = 0)$ sans covariable, puis les

poids normalisés. Vérifiez que la somme des poids normalisés égale n_t à chaque visite.

c) Si la réponse moyenne vaut 10 chez les femmes et 14 chez les hommes à chaque visite, calculez la moyenne non pondérée et la moyenne pondérée à la visite 3. Interprétez.

Exercice 12.3 (Examen final H26, question 2 c)–d))**.** Dans les *Linxian nutrition intervention trials*, 431 sujets (phase 1) ont subi une biopsie à l'entrée dans l'étude, et 81 d'entre eux ont développé un cancer de l'œsophage pendant 16 ans de suivi. Faute de moyens, la concentration de zinc a été mesurée par fluorescence aux rayons X seulement chez un sous-ensemble (phase 2) de 56 cas et 67 sujets sans cancer, sélectionnés de façon à apparier les fréquences des degrés de dysplasie (normale, légère, modérée, sévère) entre cas et témoins. Les antécédents familiaux de cancer n'ont été enregistrés qu'en phase 2. On veut estimer le rapport de cotes de cancer associé à une augmentation d'un pour cent de la concentration de zinc.

a) Identifiez les données manquantes dans cette étude. Quel est leur mécanisme ?

b) Vous souhaitez une méthode d'estimation simple à implanter, rapide et qui donne une estimation convergente du rapport de cotes sans modéliser la loi des variables manquantes. Laquelle des méthodes évaluées par Breslow et Chatterjee (1999) répond le mieux à ces exigences ? Écrivez l'équation d'estimation correspondante.

Exercice 12.4 (Complément théorique (nouveau))**.** a) Montrez que l'estimateur de Hájek

$$\widehat{\mu}_{\mathrm{H}} = \frac{\sum_i R_i Y_i / \pi(V_i)}{\sum_i R_i / \pi(V_i)}$$

est convergent sous les hypothèses de la proposition 3.1, et calculez sa variance asymptotique. Comparez-la à celle de $\widehat{\mu}_{\mathrm{IPW}}$ lorsque μ est grand.

b) Montrez que $\widehat{\mu}_{\mathrm{AIPW}}$ avec π connue et $\widehat{m}$ quelconque (même mal spécifiée, mais convergeant vers m^*) est convergent, et que sa variance asymptotique est minimale lorsque $m^* = m$.

c) Dans le modèle de la section 6, expliquez pourquoi un numérateur de poids stabilisé dépendant de $Y_{i,t-1}$ serait incorrect.

Remarque 12.1. La question 1 du devoir 2 (données de pleuropneumonie bovine `cbppw`) applique ces méthodes ; elle n'est pas traitée dans ces notes.

Références

— Breslow, N. E. et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, 48, 457–468.

— Cole, S. R. et Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168, 656–664.

— Fitzmaurice, G. M., Laird, N. M. et Ware, J. H. (2011). *Applied Longitudinal Analysis*, 2e éd., ch. 17–18. Wiley.

— Horvitz, D. G. et Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *JASA*, 47, 663–685.

— Little, R. J. A. et Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 3e éd., ch. 3. Wiley.

— Preisser, J. S., Lohman, K. K. et Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21, 3035–3054.

— Robins, J. M., Rotnitzky, A. et Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89, 846–866.

— Robins, J. M., Rotnitzky, A. et Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*, 90, 106–121.

— Seaman, S. R. et White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22, 278–295.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, ch. 6–10. Springer.

— van Buuren, S. (2018). *Flexible Imputation of Missing Data*, 2e éd. Chapman & Hall/CRC.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*, § 3.3. Chapman & Hall/CRC.