Section 6

Équations d'estimation généralisées, fonction d'influence et robustesse

Séance du 19 février 2027

Lectures : Wu (2010), passages sur les GEE (dont la p. 343) ; Fitzmaurice et coll. (2011), ch. sur les GEE ; Huber et Ronchetti (2009), ch. 1 ;
niveau doctorat : Tsiatis (2006), ch. 4 ; van der Laan et Rose (2011), ch. 5 et annexe A.1

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. formuler un modèle marginal pour des données longitudinales ou en grappes et écrire les équations d'estimation généralisées (GEE) correspondantes ;
2. choisir une structure de corrélation de travail, estimer ses paramètres par la méthode des moments et décrire l'algorithme itératif d'estimation ;
3. démontrer la convergence de $\widehat{\beta}$ lorsque seule la moyenne est bien spécifiée et obtenir la variance sandwich par linéarisation ;
4. définir la fonction d'influence d'une fonctionnelle, la calculer pour des estimateurs usuels et l'utiliser pour obtenir une variance asymptotique ;
5. construire des estimateurs M robustes (Huber, Tukey, loi de Student), y compris dans les modèles mixtes ;
6. distinguer la robustesse *à la mauvaise spécification de la covariance* (GEE) de la robustesse *aux valeurs aberrantes* ;
7. (doctorat) situer les GEE dans la théorie semi-paramétrique : espace tangent de nuisance, score efficace, borne d'efficacité, modèle de moments restreints, et comprendre les idées de base du TMLE.

Cette séance prolonge directement la section 1 (estimateurs M, linéarisation, variance sandwich) et la section 3 (modèles mixtes). Elle prépare la séance 7 (TMLE), les séances 8 et 9 (données manquantes, GEE pondérées) et la séance 10 (échantillonnage à deux phases).

1 Mise en contexte : un essai clinique avec mesures répétées binaires

Considérons un essai clinique randomisé comparant un traitement actif à un placebo chez des patients souffrant d'un problème respiratoire (données `respiratory`). Les patients de deux centres ont été répartis au hasard entre les deux bras. Le statut respiratoire (1 = bon, 0 = mauvais) a été évalué avant le traitement, puis à quatre visites de suivi. On dispose aussi du sexe et de l'âge.

La question clinique est populationnelle : *dans la population des patients, de combien le traitement actif augmente-t-il la cote d'un bon statut respiratoire ?* Les quatre mesures d'un même patient sont manifestement corrélées : un patient en bon état à la visite 1 a de bonnes chances de l'être encore à la visite 2. Trois approches viennent à l'esprit.

1. Ignorer la corrélation et ajuster une régression logistique ordinaire aux 444 observations. L'estimation des coefficients reste raisonnable, mais les erreurs types supposent 444 observations indépendantes alors qu'on n'a que 111 patients : elles sont trop petites.
2. Ajuster un modèle linéaire mixte généralisé (GLMM, section 3) avec un effet aléatoire par patient. Les coefficients ont alors une interprétation *conditionnelle au patient* (effet du traitement pour un patient donné), et l'inférence repose sur une loi postulée pour les effets aléatoires.
3. Modéliser seulement la moyenne marginale $\mathrm{E}[Y_{it} \mid X_{it}]$, traiter la corrélation comme une nuisance et corriger l'inférence en conséquence : c'est l'approche des GEE (Liang et Zeger, 1986).

Cette séance développe la troisième approche, puis en analyse les propriétés à l'aide d'un outil général, la fonction d'influence, qui sert aussi à parler de robustesse aux valeurs aberrantes.

2 Équations d'estimation généralisées

2.1 Modèle marginal

Pour le sujet (ou la grappe) $i = 1, \ldots, n$, on observe aux occasions $t = 1, \ldots, n_i$ une réponse Y_{it} et un vecteur ligne de p covariables $X_{it} = (x_{it1}, \ldots, x_{itp})$. On note $Y_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ et X_i la matrice $n_i \times p$ dont la ligne t est X_{it}. Les vecteurs (Y_i, X_i) sont indépendants d'un sujet à l'autre ; la taille des grappes n_i est bornée et $n \to \infty$.

Définition 2.1 (Modèle marginal)**.** Le modèle marginal spécifie seulement les deux premiers moments de chaque réponse :

$$\mathrm{E}[Y_{it} \mid X_{it}] = \mu_{it} = h(X_{it}\beta), \qquad \mathrm{Var}(Y_{it} \mid X_{it}) = v(\mu_{it})/\phi,$$

où h est l'inverse de la fonction de lien, v une fonction de variance connue et $\phi > 0$ un paramètre de précision. Aucune loi conjointe pour Y_i n'est postulée.

Pour une réponse binaire, $h(u) = e^u/(1+e^u)$, $v(\mu) = \mu(1-\mu)$ et $\phi = 1$; pour des comptes, $h = \exp$ et $v(\mu) = \mu$, avec $\phi < 1$ en présence de surdispersion. On suit ici la convention du matériel H26, où ϕ est une *précision* ; les logiciels (`geepack`, `gee`) rapportent plutôt le paramètre d'échelle (dispersion) $1/\phi$.

Remarque 2.2. Le paramètre β a une interprétation *marginale* (ou populationnelle) : pour une réponse binaire et un lien logit, e^{β_j} est le rapport de la cote moyenne dans la population entre deux groupes de sujets qui diffèrent d'une unité sur x_j. Ce n'est pas le même paramètre que le coefficient d'un GLMM (voir la section 2.9).

2.2 Des équations de score indépendantes aux GEE

Si les observations étaient indépendantes et issues d'une famille exponentielle, l'équation de score de β serait (section 1)

$$U_I(\beta) = \sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{\partial \mu_{it}}{\partial \beta} \frac{\phi}{v(\mu_{it})} (Y_{it} - \mu_{it}) = \sum_{i=1}^{n} D_i^\top A_i^{-1} \phi \, (Y_i - \mu_i) = 0,$$

avec

$$\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top, \quad D_i = \frac{\partial \mu_i}{\partial \beta^\top} \ \ (n_i \times p, \text{ ligne } t : h'(X_{it}\beta) X_{it}), \quad A_i = \text{diag}\{v(\mu_{i1}), \ldots, v(\mu_{in_i})\}.$$

Cette équation a deux propriétés remarquables : elle ne dépend de la loi qu'à travers les deux premiers moments (c'est une équation de quasi-vraisemblance), et elle est sans biais, $\mathrm{E}[U_I(\beta)] = 0$, dès que la moyenne est correcte.

Liang et Zeger (1986) proposent de remplacer la matrice diagonale A_i/ϕ par une matrice de covariance plausible pour Y_i.

Définition 2.3 (GEE)**.** Les équations d'estimation généralisées sont

$$U(\beta) = \sum_{i=1}^{n} D_i^\top V_i^{-1} (Y_i - \mu_i) = 0, \qquad V_i = \frac{1}{\phi} A_i^{1/2} R_i(\alpha) A_i^{1/2},$$

où $R_i(\alpha)$ est une *matrice de corrélation de travail* de dimension $n_i \times n_i$ qui dépend d'un petit nombre de paramètres α, et $A_i^{1/2} = \text{diag}\{v(\mu_{it})^{1/2}\}$.

Attention : racines carrées dans V_i

La matrice de covariance de travail est $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}/\phi$, et non $A_i R_i(\alpha) A_i/\phi$. En effet, l'élément (t,t') de $A_i^{1/2} R_i A_i^{1/2}/\phi$ vaut $\sqrt{v(\mu_{it})}\sqrt{v(\mu_{it'})}\,[R_i]_{tt'}/\phi$: c'est bien une covariance $\text{Corr}(Y_{it},Y_{it'}) \times$ é.-t.$(Y_{it}) \times$ é.-t.$(Y_{it'})$, et la diagonale redonne $v(\mu_{it})/\phi$. Avec $A_i R_i A_i$, la diagonale serait $v(\mu_{it})^2/\phi$, ce qui contredit le modèle marginal (une partie du matériel H26 écrivait $A_i R_i A_i$).

Lorsque $R_i = I_{n_i}$, on retrouve exactement U_I (puisque $V_i^{-1} = \phi A_i^{-1}$). Le paramètre ϕ multiplie toute l'équation et n'affecte donc pas sa solution : il ne compte que pour l'estimation de α et de la variance.

2.3 Structures de corrélation de travail

Les choix les plus courants, illustrés pour $n_i = 4$, sont :

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{indépendance}} \quad \underbrace{\begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}}_{\text{échangeable}} \quad \underbrace{\begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}}_{\text{AR(1)}} \quad \underbrace{\begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{pmatrix}}_{\text{non structurée}}$$

— *Indépendance* : aucun paramètre ; c'est le point de départ de l'algorithme.

— *Échangeable* (ou symétrie composée) : une seule corrélation α pour toutes les paires ; naturelle pour des grappes sans ordre (familles, cliniques) et proche de la structure induite par un intercept aléatoire.

— *AR(1)* : $[R_i]_{tt'} = \alpha^{|t-t'|}$; la corrélation décroît avec l'écart entre les occasions ; adaptée à des mesures équidistantes dans le temps.

— *Non structurée* : $n_i(n_i-1)/2$ paramètres ; flexible si les occasions sont peu nombreuses et communes à tous les sujets, mais instable sinon.

Pour une réponse binaire, la corrélation est une mesure d'association mal adaptée : ses bornes dépendent des probabilités marginales μ_{it} et $\mu_{it'}$. On peut alors paramétrer l'association par le *log du rapport de cotes* entre Y_{it} et $Y_{it'}$, ce qui mène aux équations de type « régression logistique alternée » (Wu, 2010, p. 343 ; Carey, Zeger et Diggle, 1993). Nous n'en dirons pas plus ici.

2.4 Estimation des paramètres de nuisance α et ϕ

Les paramètres α et ϕ sont estimés par la méthode des moments à partir des *résidus de Pearson*

$$r_{it} = \frac{Y_{it} - \mu_{it}}{\sqrt{v(\mu_{it})}}, \qquad \mathrm{E}[r_{it}] = 0, \quad \mathrm{Var}(r_{it}) = 1/\phi.$$

Si $R_i(\alpha)$ est la vraie corrélation, $\mathrm{E}[r_{it}r_{it'}] = [R_i(\alpha)]_{tt'}/\phi$. En égalant moments théoriques et empiriques, avec $N = \sum_i n_i$:

$$\begin{aligned}
\text{précision :} \quad & \widehat{\phi}^{-1} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{t=1}^{n_i} r_{it}^2 ; \\
\text{échangeable :} \quad & \widehat{\alpha} = \widehat{\phi}\, \frac{\sum_i \sum_{t \neq t'} r_{it} r_{it'}}{\sum_i n_i(n_i-1) - 2p} \quad \text{(paires ordonnées)} ; \\
\text{AR(1) :} \quad & \widehat{\alpha} = \frac{\widehat{\phi}}{n} \sum_{i=1}^{n} \frac{1}{n_i - 1} \sum_{t=1}^{n_i - 1} r_{it} r_{i,t+1} ; \\
\text{non structurée :} \quad & \widehat{\alpha}_{tt'} = \frac{\widehat{\phi}}{n} \sum_{i=1}^{n} r_{it} r_{it'}.
\end{aligned}$$

Les corrections $-p$ et $-2p$ au dénominateur imitent la correction de degrés de liberté de la variance résiduelle ; on peut les omettre (dénominateurs N et $\sum_i n_i(n_i-1)$, ou moyenne des moyennes par sujet) sans changer les propriétés asymptotiques. Ce qui compte pour la théorie est seulement que $\widehat{\alpha}$ et $\widehat{\phi}$ convergent vers *une* limite α^*, ϕ^* à la vitesse $\sqrt{n}$, même si la structure de travail est fausse (auquel cas α^* n'a pas d'interprétation particulière).

2.5 Algorithme itératif

Comme α et ϕ dépendent de β par les résidus, on résout $U\{\beta,\widehat{\alpha}(\beta,\widehat{\phi}(\beta))\} = 0$ en alternant :

1. Initialisation : $\widehat{\beta}^{(0)}$ solution de $U_I(\beta) = 0$ (GLM ordinaire).
2. À l'itération $j+1$:
   (a) calculer μ_{it} et les résidus de Pearson r_{it} en $\widehat{\beta}^{(j)}$;
   (b) calculer $\widehat{\phi}(\widehat{\beta}^{(j)})$, puis $\widehat{\alpha}\{\widehat{\beta}^{(j)},\widehat{\phi}(\widehat{\beta}^{(j)})\}$;
   (c) mettre à jour D_i, μ_i et V_i ;
   (d) faire un pas de Fisher (Gauss-Newton) :

$$\widehat{\beta}^{(j+1)} = \widehat{\beta}^{(j)} + \Big[\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\Big]^{-1} \sum_{i=1}^{n} D_i^\top V_i^{-1}(Y_i - \mu_i).$$

3. Répéter jusqu'à ce que $\|\widehat{\beta}^{(j+1)} - \widehat{\beta}^{(j)}\|$ soit négligeable.

Le pas (d) utilise $-\partial U/\partial\beta^\top \approx \sum_i D_i^\top V_i^{-1} D_i$, qui néglige les termes faisant intervenir $Y_i - \mu_i$ (d'espérance nulle). Le signe est $+$ parce que U est un gradient ascendant : $\partial U/\partial\beta^\top \approx -\sum_i D_i^\top V_i^{-1} D_i$.

2.6 Propriétés asymptotiques

Notons $\psi_i(\beta,\alpha) = D_i^\top V_i^{-1}(Y_i - \mu_i)$ la contribution du sujet i, de sorte que $U(\beta) = \sum_i \psi_i$. L'idée clé est la suivante.

Proposition 2.4 (Sans biais sous la seule spécification de la moyenne)**.** *Si* $\mathrm{E}[Y_i \mid X_i] = \mu_i(\beta_0)$ *et si* V_i *ne dépend que de* X_i, β *et* α, *alors* $\mathrm{E}[\psi_i(\beta_0,\alpha)] = 0$ *pour toute valeur de* α *et toute structure* R_i.

Démonstration. Par l'espérance itérée, $\mathrm{E}[\psi_i(\beta_0,\alpha)] = \mathrm{E}\big[D_i^\top V_i^{-1}\ \mathrm{E}\{Y_i - \mu_i(\beta_0) \mid X_i\}\big] = \mathrm{E}[D_i^\top V_i^{-1} \cdot 0] = 0$, puisque D_i et V_i sont des fonctions de X_i. □

Attention : covariables variant dans le temps

L'hypothèse utilisée est $\mathrm{E}[Y_{it} \mid X_i] = \mathrm{E}[Y_{it} \mid X_{it}]$: la moyenne de Y_{it} ne dépend pas des covariables des *autres* occasions. Avec des covariables fixes dans le temps ou contrôlées par le protocole (traitement randomisé, temps), c'est automatique. Avec une covariable variant dans le temps et possiblement influencée par les réponses passées, cette hypothèse peut être fausse ; seule la matrice de travail *d'indépendance* garantit alors un estimateur convergent (Pepe et Anderson, 1994), car chaque terme ψ_{it} ne fait intervenir que X_{it}.

Théorème 2.5 (Liang et Zeger, 1986)**.** *Supposons que la moyenne marginale soit bien spécifiée, que* $\widehat{\alpha} \xrightarrow{\mathbb{P}} \alpha^*$ *avec* $\sqrt{n}(\widehat{\alpha} - \alpha^*) = O_p(1)$, $\widehat{\phi} \xrightarrow{\mathbb{P}} \phi^*$, *et les conditions de régularité usuelles des estimateurs M. Alors, même si* $R_i(\alpha)$ *est mal spécifiée,*

$$\widehat{\beta} \xrightarrow{\mathbb{P}} \beta_0, \qquad \sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0,\ \mathcal{V}_1^{-1}\mathcal{V}_2\mathcal{V}_1^{-1}),$$

où, avec $V_i^* = V_i(\beta_0,\alpha^*,\phi^*)$,

$$\mathcal{V}_1 = \mathrm{E}[D_i^\top V_i^{*-1} D_i], \qquad \mathcal{V}_2 = \mathrm{E}[D_i^\top V_i^{*-1}\,\mathrm{Cov}(Y_i \mid X_i) V_i^{*-1} D_i].$$

Démonstration (esquisse détaillée). Étape 1 : convergence. Par la loi des grands nombres, $n^{-1}U(\beta)$ converge uniformément vers $\mathrm{E}[\psi_i(\beta,\alpha^*)]$, qui s'annule en β_0 par la proposition 2.4. Sous une condition d'identifiabilité (zéro unique), le zéro de $n^{-1}U$ converge vers β_0 : c'est l'argument standard de convergence des estimateurs M de la section 1.

Étape 2 : linéarisation. Un développement de Taylor de U autour de β_0 (avec, composante par composante, un point intermédiaire β^* entre $\widehat\beta$ et β_0) donne

$$0 = \frac{1}{\sqrt n} U(\widehat\beta,\widehat\alpha) = \frac{1}{\sqrt n} U(\beta_0,\widehat\alpha) + \Big[\frac{1}{n}\frac{\partial U}{\partial \beta^\top}(\beta^*,\widehat\alpha)\Big]\sqrt n(\widehat\beta - \beta_0).$$

La dérivée vaut

$$\frac{\partial \psi_i}{\partial \beta^\top} = -D_i^\top V_i^{-1} D_i + \Big[\frac{\partial}{\partial \beta^\top}(D_i^\top V_i^{-1})\Big](Y_i - \mu_i).$$

Le second terme a une espérance nulle en β_0 (même argument que la proposition), donc

$$\frac{1}{n}\frac{\partial U}{\partial \beta^\top}(\beta^*,\widehat\alpha) \xrightarrow{\mathbb{P}} -\mathcal{V}_1.$$

Étape 3 : effet de l'estimation de α. Un second développement en α donne

$$\frac{1}{\sqrt n} U(\beta_0,\widehat\alpha) = \frac{1}{\sqrt n} U(\beta_0,\alpha^*) + \Big[\frac{1}{n}\frac{\partial U}{\partial \alpha^\top}(\beta_0,\tilde\alpha)\Big]\sqrt n(\widehat\alpha - \alpha^*).$$

Or $\partial\psi_i/\partial\alpha^\top = [\partial(D_i^\top V_i^{-1})/\partial\alpha^\top](Y_i-\mu_i)$ est d'espérance nulle en β_0 : le crochet tend vers 0 en probabilité, tandis que $\sqrt n(\widehat\alpha - \alpha^*) = O_p(1)$. Le produit est $o_p(1)$: *estimer* α *ne change pas la loi asymptotique de* $\widehat\beta$. Le même raisonnement vaut pour ϕ (qui, de toute façon, se simplifie).

Étape 4 : représentation linéaire. En combinant,

$$\sqrt n(\widehat\beta - \beta_0) = \frac{1}{\sqrt n}\sum_{i=1}^n \underbrace{\mathcal{V}_1^{-1} D_i^\top V_i^{*-1}\{Y_i - \mu_i(\beta_0)\}}_{\mathrm{IF}_i} + o_p(1). \tag{1}$$

Les termes IF_i sont i.i.d., centrés, de variance $\mathcal{V}_1^{-1}\,\mathrm{E}[D_i^\top V_i^{*-1}(Y_i-\mu_i)(Y_i-\mu_i)^\top V_i^{*-1} D_i]\mathcal{V}_1^{-1} = \mathcal{V}_1^{-1}\mathcal{V}_2\mathcal{V}_1^{-1}$ (espérance itérée sur X_i). Le théorème central limite et le lemme de Slutsky concluent. □

L'expression (1) dit que $\widehat\beta$ est *asymptotiquement linéaire* de fonction d'influence $\mathrm{IF}(Y_i,X_i) = \mathcal{V}_1^{-1} D_i^\top V_i^{*-1}(Y_i-\mu_i)$: c'est la forme générale $A^{-1}\psi$ d'un estimateur M, que nous retrouverons à la section 3.

Estimateur sandwich. On estime la variance de $\widehat\beta$ en remplaçant les espérances par des moyennes empiriques et la covariance inconnue $\mathrm{Cov}(Y_i \mid X_i)$ par le produit des résidus :

$$\widehat{\mathrm{Var}}(\widehat\beta) = \frac{1}{n}\widehat V_1^{-1}\widehat V_2 \widehat V_1^{-1}, \quad \widehat V_1 = \frac{1}{n}\sum_{i=1}^n D_i^\top V_i^{-1} D_i, \quad \widehat V_2 = \frac{1}{n}\sum_{i=1}^n D_i^\top V_i^{-1}(Y_i-\widehat\mu_i)(Y_i-\widehat\mu_i)^\top V_i^{-1} D_i,$$

toutes les quantités étant évaluées en $(\widehat{\beta},\widehat{\alpha},\widehat{\phi})$. Le produit $(Y_i-\widehat{\mu}_i)(Y_i-\widehat{\mu}_i)^\top$ est un très mauvais estimateur de $\mathrm{Cov}(Y_i \mid X_i)$ pris individuellement, mais la *moyenne* $\widehat{V}_2$ converge vers $\mathcal{V}_2$ par la loi des grands nombres, parce que les sujets sont indépendants. C'est pourquoi l'inférence par GEE exige un nombre de grappes n grand, et non des grappes de grande taille.

Corollaire 2.6 (Variance « naïve » ou basée sur le modèle)**.** *Si la matrice de travail est correcte,* $\mathrm{Cov}(Y_i \mid X_i) = V_i^*$*, alors* $\mathcal{V}_2 = \mathcal{V}_1$ *et la variance asymptotique se réduit à* $\mathcal{V}_1^{-1}$*. L'estimateur* $n^{-1}\widehat{V}_1^{-1}$ *est appelé variance* naïve *(ou basée sur le modèle) : il n'est valide que si* R_i *est correcte.*

2.7 Efficacité et choix de la matrice de travail

La convergence ne dépend pas de R_i, mais la variance oui. On montre (section 4) que, parmi toutes les équations de la forme $\sum_i W(X_i)(Y_i-\mu_i)=0$, la variance asymptotique est minimale lorsque $W(X_i) = D_i^\top \mathrm{Cov}(Y_i \mid X_i)^{-1}$, c'est-à-dire lorsque la matrice de travail est la vraie covariance. Une structure mal choisie entraîne une *perte d'efficacité*, parfois importante lorsque des covariables varient à l'intérieur des grappes. Elle est en revanche faible, voire nulle, pour des covariables constantes dans la grappe : avec des grappes de même taille, sans données manquantes et des covariables toutes constantes dans la grappe, les GEE indépendance et échangeable donnent exactement la même estimation (on le verra dans l'exemple).

Exemple 2.7 (Simulation : perte d'efficacité)**.** On simule $n=60$ grappes de taille 4 selon $Y_{it} = 1 + 0{,}5\,x_{it} + b_i + \varepsilon_{it}$, avec $x_{it}\sim\mathcal{N}(0{,}1)$ variant dans la grappe, $b_i\sim\mathcal{N}(0;0{,}6)$ et $\varepsilon_{it}\sim\mathcal{N}(0;0{,}4)$, de sorte que la vraie corrélation est échangeable avec $\alpha=0{,}6$. On répète $B=500$ fois (cf. section 2 pour la planification d'une étude de simulation).

```r
library(geepack)
set.seed(7360); n <- 60; m <- 4; rho <- 0.6; B <- 500; id <- rep(1:n, each=m)
res <- t(replicate(B, {
  x <- rnorm(n*m); b <- rep(rnorm(n, sd=sqrt(rho)), each=m)
  y <- 1 + 0.5*x + b + rnorm(n*m, sd=sqrt(1-rho)); d <- data.frame(y, x, id)
  sapply(c("independence","exchangeable","ar1"), function(cs){
    fit <- geeglm(y ~ x, id=id, data=d, corstr=cs)
    c(coef(fit)[2], summary(fit)$coefficients[2,2])}) |> as.vector()
}))
```

```
               indep echang   ar1
moyenne        0.504  0.502 0.502
et.empirique   0.066  0.049 0.053
se.robuste.moy 0.062  0.045 0.050
couverture     0.926  0.928 0.934
ERE indep/echang: 0.54
```

Les trois estimateurs sont sans biais, mais l'efficacité relative de l'estimateur « indépendance » par rapport à l'estimateur « échangeable » (correct) n'est que de 0,54 : il faudrait presque deux fois plus de sujets pour la même précision. La structure AR(1), fausse mais proche, perd peu. On note aussi que l'erreur type sandwich sous-estime légèrement l'écart type réel avec 60 grappes (couverture d'environ 93 %), ce qui motive les corrections de la section suivante.

En pratique, on choisit R_i selon le plan (ordre temporel, espacement des visites), on vérifie la corrélation empirique des résidus de Pearson et on compare les erreurs types robustes sous quelques structures. Des critères comme le QIC (Pan, 2001) existent, mais leur utilité pour choisir R_i est discutée.

2.8 Corrections pour petits échantillons

L'estimateur sandwich est biaisé vers le bas lorsque le nombre de grappes est petit (disons $n < 50$), parce que les résidus $Y_i - \widehat{\mu}_i$ sont « plus petits » que les erreurs $Y_i - \mu_i$: la grappe i a contribué à l'ajustement. Plusieurs corrections ont été proposées :

— *Mancl et DeRouen (2001)* : remplacer $Y_i - \widehat{\mu}_i$ par $(I - H_{ii})^{-1}(Y_i - \widehat{\mu}_i)$, où

$$H_{ii} = D_i\Big(\sum_j D_j^\top V_j^{-1} D_j\Big)^{-1} D_i^\top V_i^{-1}$$

est le bloc diagonal du « chapeau » ;

— *Kauermann et Carroll (2001)* : même idée avec $(I - H_{ii})^{-1/2}$;

— *Fay et Graubard (2001)* et l'usage de lois t ou F de référence plutôt que normale ou khi-deux ;

— le *jackknife* par grappe (`std.err = "jack"` ou `"fij"` dans `geeglm`).

Le module `geesmv` implante plusieurs de ces corrections ; nous ne les détaillons pas.

2.9 GEE ou modèle mixte ?

Effets marginaux et effets individuels. Dans un GLMM logistique $\operatorname{logit}\mathbb{P}(Y_{it}=1 \mid X_{it}, b_i) = X_{it}\beta^C + b_i$, $b_i \sim \mathcal{N}(0,\sigma_b^2)$, le coefficient β^C est *conditionnel* au sujet. La moyenne marginale $\mathrm{E}[Y_{it} \mid X_{it}] = \int \operatorname{expit}(X_{it}\beta^C + b)\,\mathrm{d}F(b)$ n'est plus logistique en X_{it}, mais elle l'est approximativement avec des coefficients atténués :

$$\beta^M \approx \beta^C / \sqrt{1 + c^2\sigma_b^2}, \qquad c = \tfrac{16\sqrt{3}}{15\pi} \approx 0{,}588$$

(Zeger, Liang et Albert, 1988). Les deux paramètres sont légitimes mais répondent à des questions différentes : β^M (GEE) pour une question de santé publique sur la population, β^C (GLMM) pour l'effet chez un individu ou pour des prédictions individuelles. Pour un lien identité (modèle linéaire), les deux coïncident.

Données manquantes. Les GEE ne reposent pas sur une vraisemblance. L'argument de la proposition 2.4 exige que, parmi les observations disponibles, $\mathrm{E}[Y_{it} \mid X_i, \text{observé}] = \mu_{it}$; cela est garanti si les données manquent complètement au hasard (MCAR), ou au hasard conditionnellement aux seules covariables. Les méthodes de vraisemblance (GLMM, modèle linéaire mixte) restent valides sous l'hypothèse plus faible MAR (manquant au hasard, conditionnellement aux réponses observées), pourvu que le modèle soit bien spécifié. Sous MAR, on peut rétablir la validité des GEE en pondérant par l'inverse de la probabilité d'observation (GEE pondérées) : ce sera l'objet des séances 8 et 9.

En résumé. GEE : moyenne marginale seulement, inférence robuste à la covariance, calcul simple, mais MCAR requis et pas de prédiction individuelle. GLMM : modèle complet, effets individuels, MAR suffisant, mais dépendance à la loi des effets aléatoires et calcul plus lourd (intégration, section 3).

2.10 Exemple : données `respiratory`

Les données sont disponibles dans `geepack` (le matériel H26 utilisait la version de `HSAUR2`). Chaque ligne est une visite ; `outcome` est le statut à la visite, `baseline` le statut initial, `treat` vaut `A` (actif) ou `P` (placebo).

Piège : l'identifiant des sujets

Dans `geepack::respiratory`, la variable `id` recommence à 1 dans chaque centre : il y a 111 patients mais seulement 56 valeurs distinctes de `id`. Utiliser `id` tel quel fusionnerait deux patients différents en une même grappe de taille 8. Il faut construire un identifiant unique, et trier les données par grappe (exigence de `geeglm`).

```
library(geepack)
data(respiratory, package="geepack")
resp <- respiratory
resp$id <- as.integer(interaction(resp$center, resp$id, drop=TRUE))  # identifiant
    unique
resp <- resp[order(resp$id, resp$visit), ]
resp$center <- factor(resp$center); resp$visit <- factor(resp$visit)
resp$treat <- relevel(resp$treat, ref="P")
length(unique(resp$id))
with(resp, round(tapply(outcome, list(treat, visit), mean), 2))
```

```
[1] 111
     1    2    3    4
P 0.49 0.39 0.46 0.44
A 0.69 0.70 0.72 0.61
```

La proportion de bons statuts est stable dans le temps dans chaque bras, nettement plus élevée sous traitement actif. On ajuste d'abord une régression logistique naïve, puis des GEE sous trois structures (avec $\phi = 1$ fixé, ce qui est naturel pour une réponse binaire).

```
f <- outcome ~ center + treat + sex + baseline + age
m.glm <- glm(f, family=binomial, data=resp)
m.ind <- geeglm(f, family=binomial, data=resp, id=id, corstr="independence", scale.fix=
    TRUE)
m.exc <- geeglm(f, family=binomial, data=resp, id=id, corstr="exchangeable", scale.fix=
    TRUE)
m.ar1 <- geeglm(f, family=binomial, data=resp, id=id, corstr="ar1", scale.fix=TRUE)
summary(m.ar1)
```

```
 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept) -0.78525  0.63219  1.543 0.214191
center2      0.73020  0.35251  4.291 0.038316 *
treatA       1.19095  0.34832 11.691 0.000628 ***
sexM        -0.13461  0.44779  0.090 0.763711
baseline     1.87103  0.34660 29.141 6.73e-08 ***
age         -0.01740  0.01288  1.825 0.176691
Correlation structure = ar1
Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.4595 0.05518
Number of clusters:   111  Maximum cluster size: 4
```

Le tableau suivant rassemble les estimations et les erreurs types (é.-t.) ; « naïf » désigne l'erreur type de `glm`, les autres sont des erreurs types sandwich.

	GLM naïf		GEE indép.		GEE échang.		GEE AR(1)
	Est.	é.-t.	Est.	é.-t.	Est.	é.-t.	Est. (é.-t.)
Ordonnée	−0,719	0,438	−0,719	0,631	−0,719	0,631	−0,785 (0,632)
Centre 2	0,649	0,238	0,649	0,353	0,649	0,353	0,730 (0,353)
Traitement A	1,265	0,235	1,265	0,347	1,265	0,347	1,191 (0,348)
Sexe M	−0,137	0,293	−0,137	0,440	−0,137	0,440	−0,135 (0,448)
Statut initial	1,846	0,239	1,846	0,346	1,846	0,346	1,871 (0,347)
Âge	−0,019	0,009	−0,019	0,013	−0,019	0,013	−0,017 (0,013)

Plusieurs leçons se dégagent.

1. Les estimations de `glm` et du GEE indépendance sont identiques (même équation), mais les erreurs types naïves sont environ 30 % trop petites (0,235 contre 0,347 pour le traitement). Les conclusions changent pour le centre et l'âge, « significatifs » à tort avec `glm`.
2. Les GEE indépendance et échangeable donnent exactement les mêmes estimations : toutes les covariables sont constantes dans la grappe et toutes les grappes sont complètes de taille 4 (cf. section 2.7).
3. La structure AR(1) modifie légèrement les estimations (elle pondère différemment les visites), sans gain de précision visible.
4. La corrélation empirique des résidus de Pearson du modèle d'indépendance aide à choisir R_i :

```
r <- residuals(m.glm, type="pearson"); round(cor(matrix(r, ncol=4, byrow=TRUE)), 2)
```

```
     [,1] [,2] [,3] [,4]
[1,] 1.00 0.34 0.23 0.30
[2,] 0.34 1.00 0.43 0.31
[3,] 0.23 0.43 1.00 0.39
[4,] 0.30 0.31 0.39 1.00
```

Les corrélations ne décroissent pas clairement avec l'écart entre visites : une structure échangeable ($\widehat{\alpha} = 0{,}33$) est plus plausible qu'un AR(1) ($\widehat{\alpha} = 0{,}46$ impliquerait 0,10 au décalage 3).

Erreurs types naïves et robustes d'un même GEE. Le module `gee` affiche les deux :

```
library(gee)
g <- gee(f, id=id, data=resp, family=binomial, corstr="exchangeable",
         scale.fix=TRUE, scale.value=1)
round(summary(g)$coefficients, 3)
```

```
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.719      0.617  -1.166       0.631   -1.140
center2        0.649      0.335   1.937       0.353    1.839
treatA         1.265      0.331   3.826       0.347    3.650
sexM          -0.137      0.413  -0.331       0.440   -0.311
baseline       1.846      0.337   5.481       0.346    5.335
age           -0.019      0.012  -1.512       0.013   -1.447
```

Ici, la variance naïve $n^{-1}\widehat{V}_1^{-1}$ du GEE échangeable est proche de la variance sandwich : la structure échangeable décrit raisonnablement la dépendance.

Vérification du sandwich « à la main ». Pour le GEE indépendance avec lien logit canonique, $D_i^\top V_i^{-1} = X_i^\top$, donc $\widehat{V}_1 \propto \sum_i X_i^\top \widehat{W}_i X_i$ et $\widehat{V}_2 \propto \sum_i X_i^\top (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)^\top X_i$:

```
X <- model.matrix(m.glm); mu <- fitted(m.glm); e <- resp$outcome - mu
V1 <- crossprod(X, mu*(1-mu)*X)
U  <- rowsum(X*e, resp$id)           # contributions psi_i sommées par sujet
Vs <- solve(V1) %*% crossprod(U) %*% solve(V1)
round(cbind(manuel=sqrt(diag(Vs)), geepack=summary(m.ind)$coefficients[,2]), 4)
```

```
            manuel geepack
(Intercept) 0.6312  0.6312
center2     0.3532  0.3532
treatA      0.3467  0.3467
sexM        0.4402  0.4402
baseline    0.3460  0.3460
age         0.0130  0.0130
```

Tests de Wald avec `anova`. La méthode `anova` de `geepack` produit des tests de Wald généralisés (basés sur la variance sandwich), séquentiels ou emboîtés. On teste ici l'interaction traitement × visite (le modèle du matériel H26) :

```
m.int <- geeglm(outcome ~ center + treat*visit + sex + baseline + age, family=binomial,
                data=resp, id=id, corstr="exchangeable", scale.fix=TRUE)
anova(m.int)
anova(m.int, m.exc)
```

```
Analysis of 'Wald statistic' Table
Terms added sequentially (first to last)
            Df    X2 P(>|Chi|)
center       1  7.65    0.0057 **
treat        1 10.60    0.0011 **
visit        3  3.60    0.3076
sex          1  0.60    0.4403
baseline     1 28.84   7.9e-08 ***
age          1  2.06    0.1517
treat:visit  3  3.16    0.3676

Model 1 outcome ~ center + treat * visit + sex + baseline + age
Model 2 outcome ~ center + treat + sex + baseline + age
  Df  X2 P(>|Chi|)
1  6 6.7      0.35
```

Rien n'indique que l'effet du traitement ou le niveau moyen varient selon la visite ($p = 0{,}35$ pour les six paramètres). On retient le modèle sans visite : le rapport de cotes marginal du traitement actif est $e^{1{,}265} = 3{,}54$ (IC à 95 % robuste : 1,80 à 6,99). Comme les tests de Wald séquentiels dépendent de l'ordre des termes, on préfère les comparaisons emboîtées explicites.

Grille méthodologique

1. **Objet d'intérêt** : β, paramètres de la moyenne marginale $\mathrm{E}[Y_{it} \mid X_{it}]$.
2. **Nuisance** : toute la loi conjointe de Y_i au-delà de la moyenne, en particulier la covariance intra-sujet ; données manquantes (MCAR requis).
3. **Approche** : ne pas modéliser la loi ; remplacer la covariance par une matrice de travail $V_i(\alpha)$ dont la mauvaise spécification n'introduit pas de biais.
4. **Critère** : équations d'estimation $\sum_i D_i^\top V_i^{-1}(Y_i - \mu_i) = 0$ (estimateur M).
5. **Calcul** : algorithme itératif alternant Fisher scoring pour β et méthode des moments pour (α,ϕ).
6. **Inférence** : normalité asymptotique via la fonction d'influence ; variance sandwich $\mathcal{V}_1^{-1}\mathcal{V}_2\mathcal{V}_1^{-1}$; tests de Wald ; corrections si peu de grappes.

À retenir

— Les GEE estiment les paramètres d'une moyenne *marginale* ; $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}/\phi$.

— $\widehat{\beta}$ converge si la moyenne est correcte, quelle que soit R_i : les équations sont sans biais par l'espérance itérée. Estimer α n'affecte pas la loi limite de $\widehat{\beta}$.

— La variance sandwich est valide même si R_i est fausse ; la variance naïve ne l'est que si R_i est correcte. Un bon choix de R_i améliore l'efficacité, surtout pour les covariables qui varient dans la grappe.

— Il faut beaucoup de grappes ; avec peu de grappes, corriger (Mancl-DeRouen, jackknife, lois t/F).

— GEE : effets populationnels et MCAR ; GLMM : effets individuels et MAR.

3 Fonction d'influence et robustesse

3.1 Fonctionnelles et dérivée de Gâteaux

Beaucoup d'estimateurs s'écrivent $\widehat{\theta} = T(F_n)$, où F_n est la fonction de répartition empirique et T une *fonctionnelle* définie sur un ensemble de lois, avec $T(F) = \theta$ à la vraie loi F (convergence de Fisher). Par exemple, la moyenne $T(F) = \int y \,\mathrm{d}F(y)$, la variance $T(F) = \int \{y - \int z \,\mathrm{d}F(z)\}^2 \,\mathrm{d}F(y)$, la médiane $T(F) = F^{-1}(1/2)$, ou la solution $T(F)$ de $\int \psi(y,T(F)) \,\mathrm{d}F(y) = 0$ pour un estimateur M.

Définition 3.1 (Fonction d'influence)**.** Soit Δ_y la loi qui place une masse 1 au point y, et $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_y$ la loi contaminée. La fonction d'influence de T en F est la dérivée de Gâteaux de T dans la direction $\Delta_y - F$:

$$\mathrm{IF}(y;T,F) = \lim_{\varepsilon \downarrow 0} \frac{T\{(1-\varepsilon)F + \varepsilon\Delta_y\} - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_\varepsilon)\Big|_{\varepsilon=0}.$$

L'interprétation est directe : $\mathrm{IF}(y)$ mesure l'effet, normalisé par la proportion ε, de l'ajout d'une petite fraction d'observations au point y. Pour un échantillon de taille n, ajouter une observation en y correspond à $\varepsilon \approx 1/(n+1)$, de sorte que $T(F_{n+1}) - T(F_n) \approx \mathrm{IF}(y)/(n+1)$: c'est la base des diagnostics d'influence empiriques (distance de Cook, jackknife).

3.2 Estimateurs asymptotiquement linéaires

Définition 3.2. Un estimateur $\widehat{\theta}$ est *asymptotiquement linéaire* de fonction d'influence φ si

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(O_i)+o_p(1),\qquad \mathrm{E}[\varphi(O)]=0,\quad \mathrm{E}[\varphi(O)\varphi(O)^\top]<\infty.$$

Proposition 3.3. *Si $\widehat{\theta}$ est asymptotiquement linéaire de fonction d'influence φ, alors $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mathrm{E}[\varphi\varphi^\top])$, et $\widehat{\mathrm{Var}}(\widehat{\theta})=n^{-2}\sum_i\widehat{\varphi}(O_i)\widehat{\varphi}(O_i)^\top$ est un estimateur convergent de sa variance (à la normalisation près), si $\widehat{\varphi}$ converge convenablement vers φ.*

Démonstration. Théorème central limite pour la somme i.i.d. et lemme de Slutsky pour le reste $o_p(1)$. □

Lien heuristique avec la définition 3.1 : si T est suffisamment régulière (Hadamard-différentiable), un développement de von Mises donne $T(F_n)-T(F)=\int \mathrm{IF}(y;T,F)\,\mathrm{d}(F_n-F)(y)+\text{reste}=\frac{1}{n}\sum_i \mathrm{IF}(Y_i;T,F)+\text{reste}$, car $\int \mathrm{IF}\,\mathrm{d}F=0$. La fonction d'influence de la définition 3.1 est donc la fonction φ de la représentation linéaire : c'est la *méthode delta fonctionnelle*. La variance sandwich des GEE en est un cas particulier, avec $O_i=(Y_i,X_i)$.

3.3 Exemples

Exemple 3.4 (Moyenne)**.** $T(F_\varepsilon)=(1-\varepsilon)\mu+\varepsilon y$, donc $\mathrm{IF}(y)=y-\mu$, et $\mathrm{E}[\mathrm{IF}^2]=\sigma^2$: on retrouve $\mathrm{Var}(\bar{Y})\approx\sigma^2/n$.

Exemple 3.5 (Variance)**.** $T(F)=\int y^2\,\mathrm{d}F-\mu(F)^2$. On a $\frac{\partial}{\partial\varepsilon}\int y^2\,\mathrm{d}F_\varepsilon=y^2-\int z^2\,\mathrm{d}F$ et $\frac{\partial}{\partial\varepsilon}\mu(F_\varepsilon)^2=2\mu(y-\mu)$. Donc $\mathrm{IF}(y)=y^2-\mathrm{E}\,Y^2-2\mu(y-\mu)=(y-\mu)^2-\sigma^2$, et $\mathrm{E}[\mathrm{IF}^2]=\mu_4-\sigma^4$ (qui vaut $2\sigma^4$ sous la normalité).

Exemple 3.6 (Médiane)**.** Soit $m=F^{-1}(1/2)$, F de densité f continue et positive en m. La médiane m_ε de F_ε vérifie $(1-\varepsilon)F(m_\varepsilon)+\varepsilon\mathbf{1}\{y\le m_\varepsilon\}=1/2$. En dérivant en $\varepsilon=0$: $-F(m)+f(m)\,\dot{m}+\mathbf{1}\{y\le m\}=0$, d'où

$$\mathrm{IF}(y)=\frac{\frac{1}{2}-\mathbf{1}\{y\le m\}}{f(m)}=\frac{\mathrm{sgn}(y-m)}{2f(m)},\qquad \mathrm{E}[\mathrm{IF}^2]=\frac{1}{4f(m)^2}.$$

Cette fonction est *bornée* : une observation très grande n'a pas plus d'effet qu'une observation à peine supérieure à la médiane.

Exemple 3.7 (Estimateur M général)**.** Soit $T(F)$ défini par $\int\psi(y,T(F))\,\mathrm{d}F(y)=0$, avec ψ de dimension p. En remplaçant F par F_ε et en dérivant en $\varepsilon=0$:

$$\psi(y,\theta)-\int\psi(z,\theta)\,\mathrm{d}F(z)+\Big[\int\frac{\partial\psi(z,\theta)}{\partial\theta^\top}\,\mathrm{d}F(z)\Big]\mathrm{IF}(y)=0,$$

et comme le deuxième terme est nul,

$$\boxed{\mathrm{IF}(y;T,F)=A^{-1}\psi(y,\theta),\qquad A=-\,\mathrm{E}_F\Big[\frac{\partial\psi(Y,\theta)}{\partial\theta^\top}\Big].}$$

La variance asymptotique est $A^{-1}\,\mathrm{E}[\psi\psi^\top]A^{-\top}$: c'est la variance sandwich de la section 1. Pour les GEE, $\psi=D_i^\top V_i^{-1}(Y_i-\mu_i)$ et $A=\mathcal{V}_1$.

Exemple 3.8 (Estimateur du maximum de vraisemblance)**.** Avec $\psi = S_\theta(y) = \partial \log f(y;\theta)/\partial\theta$, on a $A = \mathcal{I}(\theta)$ (information de Fisher) et $\mathrm{IF}(y) = \mathcal{I}(\theta)^{-1} S_\theta(y)$; $\mathrm{E}[\mathrm{IF}\,\mathrm{IF}^\top] = \mathcal{I}(\theta)^{-1}$ si le modèle est correct. Pour la loi normale de variance connue, $S_\mu(y) = (y-\mu)/\sigma^2$ et $\mathcal{I} = 1/\sigma^2$: $\mathrm{IF}(y) = y - \mu$. Pour une régression linéaire, $\mathrm{IF}(y,x) = \mathrm{E}[x^\top x]^{-1} x^\top (y - x\beta)$: l'influence est le produit d'un *résidu* et d'un *levier*, tous deux non bornés.

3.4 Sensibilité aux valeurs extrêmes et point de rupture

La *sensibilité aux erreurs grossières* est $\gamma^* = \sup_y \|\mathrm{IF}(y;T,F)\|$ (éventuellement après standardisation). Un estimateur est dit *B-robuste* si $\gamma^* < \infty$. La moyenne, la variance, l'EMV normal et les GEE ont une fonction d'influence non bornée en y : une seule valeur extrême peut déplacer l'estimation arbitrairement loin (exemple 3.11 ci-dessous).

La fonction d'influence est une notion *locale* (contamination infinitésimale). La notion globale complémentaire est le *point de rupture* : la plus petite fraction de contamination qui peut entraîner l'estimateur arbitrairement loin. Il vaut 0 pour la moyenne ($1/n$ en échantillon fini), 1/2 pour la médiane, et 1/2 pour un estimateur M de localisation à ψ bornée muni d'une échelle robuste (MAD). En régression, les estimateurs M ordinaires restent sensibles aux points de levier ; on renvoie à Maronna et coll. (2019) pour les estimateurs à haut point de rupture (S, MM).

3.5 Estimateurs M robustes

Considérons le modèle de localisation $Y = \mu + \sigma Z$, σ connu (ou estimé de façon robuste), et l'estimateur M défini par $\sum_i \psi\{(Y_i - \mu)/\sigma\} = 0$. Par l'exemple 3.7,

$$\mathrm{IF}(y;\mu) = \frac{\sigma\,\psi\{(y-\mu)/\sigma\}}{\mathrm{E}[\psi'(Z)]}, \qquad \text{variance asymptotique} = \sigma^2 \frac{\mathrm{E}[\psi(Z)^2]}{\{\mathrm{E}[\psi'(Z)]\}^2}.$$

L'EMV normal correspond à $\psi(z) = z$. Pour borner l'influence, il suffit de choisir une fonction ψ bornée : *on modifie directement la fonction d'influence standardisée.*

Huber. $\psi_k(z) = \max\{-k, \min(k,z)\}$: identité au centre, troncature en $\pm k$. Huber (1964) a montré que cet estimateur minimise la variance asymptotique maximale sur un voisinage de contamination de la loi normale. Avec $k = 1{,}345$, l'efficacité sous la loi normale est $\{2\Phi(k) - 1\}^2 / \mathrm{E}[\psi_k^2] = 0{,}95$. L'estimateur est monotone (solution unique).

Bisquare de Tukey. $\psi_c(z) = z\{1 - (z/c)^2\}^2 \mathbf{1}\{|z| \le c\}$: fonction *redescendante*, nulle au-delà de c ; les observations très éloignées sont complètement rejetées. Avec $c = 4{,}685$, l'efficacité normale est 0,95. L'équation peut avoir plusieurs solutions : il faut un bon point de départ (médiane, ou estimateur de Huber).

Loi de Student. L'EMV de μ sous une loi t_ν de paramètre d'échelle σ a pour fonction $\psi_\nu(z) = (\nu+1)z/(\nu + z^2)$, bornée et redescendante (plus doucement que Tukey).

Poids. En écrivant $\psi(z) = w(z)\,z$, l'équation devient $\sum_i w(z_i)(Y_i - \mu) = 0$: l'estimateur est une moyenne pondérée, avec des poids qui dépendent de l'estimation. On le calcule par *moindres carrés repondérés itérativement*. Huber : $w(z) = \min(1, k/|z|)$; Tukey : $w(z) = \{1 - (z/c)^2\}_+^2$; Student : $w(z) = (\nu + 1)/(\nu + z^2)$.

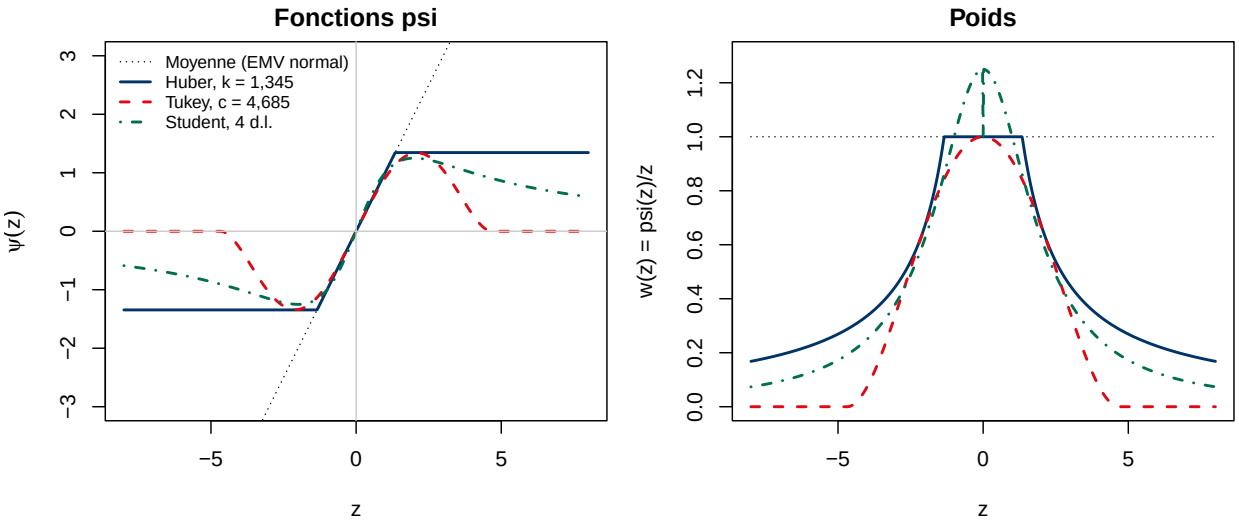


FIGURE 1 – Fonctions ψ (à gauche) et poids $w(z) = \psi(z)/z$ (à droite) pour la moyenne, l'estimateur de Huber, le bisquare de Tukey et l'EMV sous une loi de Student à 4 degrés de liberté.

Exemple 3.9 (Localisation avec deux valeurs aberrantes)**.**

```
library(MASS)
set.seed(7360); y <- c(rnorm(20, mean=10, sd=1), 25, 30)
c(moyenne=mean(y), mediane=median(y),
  huber=coef(rlm(y ~ 1, psi=psi.huber, k=1.345)),
  bisquare=coef(rlm(y ~ 1, psi=psi.bisquare, c=4.685)))
round(tail(rlm(y ~ 1, psi=psi.huber)$w, 2), 3); round(tail(rlm(y ~ 1, psi=psi.bisquare)
    $w, 2), 3)
```

```
          moyenne            mediane   huber.(Intercept) bisquare.(Intercept)
         11.044914           9.293826           9.566037           9.381831
[1] 0.108 0.082
[1] 0.000 0.000
```

Les deux valeurs aberrantes tirent la moyenne de plus d'une unité. L'estimateur de Huber leur donne un poids d'environ 0,1 ; le bisquare les élimine complètement. (La fonction `rlm` estime l'échelle par la MAD.)

Exemple 3.10 (Question d'examen H23)**.** *La fonction d'influence de l'EMV de la moyenne μ d'une loi normale d'écart type σ (connu) est* $\mathrm{IF}(y,\mu) = y - \mu$. *Donnez deux façons de modifier cette fonction pour obtenir une estimation robuste de μ et écrivez la nouvelle fonction d'influence dans les deux cas.*

Solution. On standardise : $z = (y - \mu)/\sigma$, et $\mathrm{IF}/\sigma = z$.

1. *Borner* la fonction d'influence standardisée (Huber) : $\psi_k(z) = -k$ si $z \le -k$, z si $|z| < k$, k si $z \ge k$. La nouvelle fonction d'influence est $\mathrm{IF}(y,\mu) = \sigma\psi_k(z)/\{2\Phi(k) - 1\}$.

2. *Remplacer la loi normale par une loi de Student* à ν degrés de liberté : $\psi_\nu(z) = (\nu+1)z/(\nu+z^2)$; $\mathrm{IF}(y,\mu) = \sigma\psi_\nu(z)/\,\mathrm{E}[\psi'_\nu(Z)]$, l'espérance étant prise sous le modèle postulé. Le solutionnaire H23 écrit $\psi(z) = z/(z^2+k)$ avec k le nombre de degrés de liberté : c'est la même fonction à une constante multiplicative près, qui ne change ni l'estimateur ni sa fonction d'influence normalisée.

Le bisquare de Tukey serait une troisième réponse acceptable.

3.6 Robustesse dans les modèles mixtes

Dans le modèle linéaire mixte $Y_i = X_i\beta + Z_i b_i + \varepsilon_i$ (section 3), les équations de vraisemblance de β sont $\sum_i X_i^\top \Sigma_i^{-1}(Y_i - X_i\beta) = 0$, $\Sigma_i = Z_i G Z_i^\top + \sigma^2 I$: la fonction d'influence est linéaire en Y_i et donc non bornée. Un sujet aberrant (effet aléatoire extrême) ou des mesures aberrantes (erreurs extrêmes) peuvent déplacer les effets fixes et gonfler les composantes de variance.

Remplacer la loi normale par une loi de Student. On peut postuler $b_i \sim t_\nu(0,G)$ et/ou $\varepsilon_{it} \sim t_\nu(0,\sigma^2)$ (Lange, Little et Taylor, 1989 ; Pinheiro, Liu et Wu, 2001). La loi de Student est un mélange d'échelle de lois normales : si $w_i \sim \mathrm{Gamma}(\nu/2,\nu/2)$ et $b_i \mid w_i \sim \mathcal{N}(0,G/w_i)$, alors $b_i \sim t_\nu(0,G)$. Dans la version où b_i et ε_i partagent le même poids latent w_i (Pinheiro, Liu et Wu, 2001), Y_i suit une loi de Student multivariée et l'étape E d'un algorithme EM utilise

$$\mathrm{E}[w_i \mid Y_i] = \frac{\nu + n_i}{\nu + \delta_i^2}, \qquad \delta_i^2 = (Y_i - X_i\beta)^\top \Sigma_i^{-1}(Y_i - X_i\beta),$$

où δ_i^2 est une distance de Mahalanobis : c'est l'analogue multivarié du poids de Student de la section 3.5. Les sujets éloignés sont donc automatiquement sous-pondérés. Le nombre de degrés de liberté ν est le seul paramètre supplémentaire ; on peut le fixer (par exemple $\nu = 4$) ou l'estimer. La même idée s'applique aux GLMM (question d'examen H26 sur le mélanome : effets aléatoires régionaux de Student).

Exemple 3.11 (Influence d'un sujet aberrant sur `lmer`)**.** Les données `sleepstudy` (`lme4`) contiennent les temps de réaction de 18 sujets privés de sommeil pendant 10 jours. On ajuste le modèle à ordonnée et pente aléatoires indépendantes, puis on mesure l'influence empirique de chaque sujet sur la pente fixe par jackknife, $\widehat{\mathrm{IF}}_i = (n-1)(\widehat\beta - \widehat\beta_{(-i)})$.

```
library(lme4)
data("sleepstudy", package="lme4")
fit <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), data=sleepstudy)
subj <- levels(sleepstudy$Subject); n <- length(subj); b <- fixef(fit)["Days"]
bm <- sapply(subj, function(s)
  fixef(update(fit, data=subset(sleepstudy, Subject != s)))["Days"])
round(sort(setNames((n-1)*(b - bm), subj)), 1)
```

```
  335   309   330   331   310   351   333   371   332   372   369   334   349
-13.3  -8.2  -7.5  -5.2  -4.4  -4.0  -1.3  -1.3  -0.9   0.8   0.9   1.8   3.0
  352   370   337   350   308
  3.1   7.6   8.6   9.0  11.3
```

La pente fixe estimée est 10,47 ms/jour (é.-t. 1,56). On contamine ensuite le sujet 335 en ajoutant δ ms par jour à ses temps de réaction, et on compare la pente fixe de `lmer` à la médiane des pentes individuelles (estimées par moindres carrés sujet par sujet), un estimateur robuste rudimentaire.

```
contamine <- function(delta) {
  sc <- sleepstudy; k <- sc$Subject == "335"
  sc$Reaction[k] <- sc$Reaction[k] + delta*sc$Days[k]
  f <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), data=sc)
  pentes <- sapply(split(sc, sc$Subject), function(d) coef(lm(Reaction ~ Days, d))[2])
  c(delta=delta, pente.fixe=unname(fixef(f)[2]),
    et.pente=attr(VarCorr(f)$Subject.1, "stddev"), mediane=median(pentes))
}
round(t(sapply(c(0, 25, 50, 100, 200), contamine)), 2)
```

```
     delta pente.fixe et.pente.Days mediane
[1,]     0      10.47          5.99   10.43
[2,]    25      11.86          5.71   11.32
[3,]    50      13.25          9.95   11.32
[4,]   100      16.02         20.98   11.32
[5,]   200      21.58         44.22   11.32
```

La pente fixe croît linéairement avec δ (environ $\delta/18$, car le plan est équilibré) et l'écart type des pentes aléatoires explose : c'est la signature d'une fonction d'influence non bornée. La médiane des pentes individuelles cesse de bouger dès que le sujet 335 passe au-dessus de la médiane : son influence est bornée.

Le matériel H26 illustrait le modèle de Student avec `brms` (estimation bayésienne par Stan). Ce module ne pouvait pas compiler de modèles dans l'environnement de préparation de ces notes : le code ci-dessous est donné **sans sortie**, à exécuter de votre côté.

```
library(brms)
set.seed(7360)
# Erreurs normales : reproduit essentiellement le maximum de vraisemblance de lmer
modeleb  <- brm(Reaction ~ Days + (1|Subject) + (0+Days|Subject), data=sleepstudy)
# Erreurs de Student : nu est estimé (loi a priori gamma(2; 0,1) par défaut)
modelebt <- brm(Reaction ~ Days + (1|Subject) + (0+Days|Subject),
                family=student(), data=sleepstudy)
summary(modelebt)
# Écart type résiduel implicite de la loi de Student : sigma*sqrt(nu/(nu-2))
s <- summary(modelebt)$spec_pars; s["sigma","Estimate"]*sqrt(s["nu","Estimate"]/(s["nu",
    "Estimate"]-2))
loo(modeleb, modelebt)   # comparaison des modèles (avertissements possibles)
```

Selon le matériel H26, les effets fixes du modèle à erreurs de Student sont semblables à ceux du modèle normal, avec des erreurs types un peu plus grandes. Pour des *effets aléatoires* de Student, `brms` permet de préciser la loi des effets d'un groupe avec `gr()` :

```
modelebtt <- brm(Reaction ~ Days + (1 + Days | gr(Subject, dist="student")),
                 family=student(), data=sleepstudy)
```

(ici, ordonnée et pente aléatoires corrélées, de loi de Student bivariée). Le module `robustlmm` (Koller, 2016) propose une autre approche, par estimation M robuste des effets fixes, des effets aléatoires et des composantes de variance.

3.7 Deux sens du mot « robuste »

Ne pas confondre

— **Robustesse à la mauvaise spécification** (sens des GEE, des « erreurs types robustes ») : la variance sandwich reste valide si la matrice de covariance de travail est fausse. Mais l'équation $\sum_i D_i^\top V_i^{-1}(Y_i - \mu_i) = 0$ a une fonction d'influence *non bornée* en Y_i : une seule grappe aberrante peut biaiser $\widehat{\beta}$ arbitrairement, et l'estimateur sandwich lui-même (qui contient des carrés de résidus) y est encore plus sensible. La moyenne doit aussi être bien spécifiée.

— **Robustesse aux valeurs aberrantes** (sens de Huber et Hampel) : l'estimateur a une fonction d'influence bornée, et sa loi change peu sous une petite contamination de la loi postulée.

On peut combiner les deux : des GEE robustes remplacent $Y_i - \mu_i$ par une version tronquée des résidus de Pearson et ajoutent des poids sur les leviers, avec une correction de biais pour conserver la convergence de Fisher (Preisser et Qaqish, 1999 ; Cantoni et Ronchetti, 2001, pour les GLM ; Cantoni, 2004, pour les données longitudinales).

Grille méthodologique

1. **Objet** : une fonctionnelle $T(F)$ (moyenne, coefficient, composante de variance).
2. **Nuisance / biais** : contamination de la loi postulée par des valeurs aberrantes ; erreur de modèle.
3. **Approche** : borner (ou faire redescendre) la fonction d'influence ; loi à queues lourdes (Student).
4. **Critère** : équation d'estimation M $\sum_i \psi(O_i,\theta) = 0$ avec ψ bornée.
5. **Calcul** : moindres carrés repondérés itérativement ; EM pour la loi de Student ; MCMC (`brms`).
6. **Inférence** : variance $A^{-1}\,\mathrm{E}[\psi\psi^\top]A^{-\top}$, i.e. $\mathrm{E}[\mathrm{IF}\,\mathrm{IF}^\top]$.

À retenir

— $\mathrm{IF}(y;T,F)$ est la dérivée de T dans la direction d'une masse ponctuelle en y ; pour un estimateur asymptotiquement linéaire, la variance asymptotique est $\mathrm{E}[\mathrm{IF}\,\mathrm{IF}^\top]$.

— Estimateur M : $\mathrm{IF} = A^{-1}\psi$ avec $A = -\,\mathrm{E}[\partial\psi/\partial\theta^\top]$; EMV : $\mathcal{I}^{-1}S_\theta$.

— Moyenne, EMV normal, GEE : IF non bornée. Huber (troncature), Tukey (redescendante), Student : IF bornée, au prix d'une petite perte d'efficacité sous le modèle.

— Dans les modèles mixtes, remplacer la loi normale des effets aléatoires et/ou des erreurs par une loi de Student sous-pondère les sujets ou mesures aberrants.

— « Erreurs types robustes » (GEE) $\neq$ « estimateur robuste aux valeurs aberrantes ».

4 Niveau doctorat : estimation semi-paramétrique

Niveau doctorat – cadre de cette section

Cette section suit Tsiatis (2006, ch. 4) et van der Laan et Rose (2011, ch. 5 et annexe A.1). Elle explique *pourquoi* les GEE ont la forme $D^\top V^{-1}(Y-\mu)$, quand elles sont efficaces, et introduit les idées qui mènent au TMLE (séance 7). Les démonstrations sont des esquisses ; les conditions de régularité sont omises.

4.1 Espace de Hilbert et rappel paramétrique

Soit O l'observation (par exemple $O=(Y,X)$) et $\mathcal{H}$ l'espace des fonctions $h(O)$ à valeurs dans $\mathbb{R}^q$, d'espérance nulle et de variance finie, muni du produit scalaire $\langle h_1,h_2\rangle = \mathrm{E}[h_1^\top h_2]$. Pour un sous-espace fermé Λ, la projection $\Pi(h \mid \Lambda)$ est l'unique élément de Λ tel que $h-\Pi(h\mid\Lambda)\perp\Lambda$.

Dans un modèle paramétrique $p(o;\beta,\eta)$, avec paramètre d'intérêt $\beta\in\mathbb{R}^q$ et nuisance $\eta\in\mathbb{R}^r$, les scores sont S_β et S_η. L'*espace tangent de nuisance* est $\Lambda=\{BS_\eta : B\in\mathbb{R}^{q\times r}\}$. On sait (Tsiatis, ch. 3) que :

- toute fonction d'influence φ d'un estimateur régulier et asymptotiquement linéaire (RAL) de β vérifie $\mathrm{E}[\varphi S_\beta^\top]=I_q$ et $\varphi\perp\Lambda$;
- le *score efficace* est $S_{\mathrm{eff}}=S_\beta-\Pi(S_\beta\mid\Lambda)$, la *fonction d'influence efficace* est $\varphi_{\mathrm{eff}}=\{\mathrm{E}[S_{\mathrm{eff}}S_{\mathrm{eff}}^\top]\}^{-1}S_{\mathrm{eff}}$ et la borne d'efficacité est $\{\mathrm{E}[S_{\mathrm{eff}}S_{\mathrm{eff}}^\top]\}^{-1}$;
- toute autre fonction d'influence s'écrit $\varphi=\varphi_{\mathrm{eff}}+\ell$ avec ℓ orthogonal à l'espace tangent complet, d'où $\mathrm{Var}(\varphi)=\mathrm{Var}(\varphi_{\mathrm{eff}})+\mathrm{Var}(\ell)$.

4.2 Modèle semi-paramétrique et sous-modèles paramétriques

Un modèle *semi-paramétrique* est une famille de lois $\{p(o;\beta,\eta)\}$ où $\beta\in\mathbb{R}^q$ est de dimension finie et la nuisance η est de dimension infinie (une fonction, une loi). Par exemple, le modèle de moments restreints $\mathrm{E}[Y\mid X]=\mu(X,\beta)$ laisse libres la loi de X et toute la loi conditionnelle de Y sachant X hormis sa moyenne.

Définition 4.1 (Sous-modèle paramétrique ; estimateur semi-paramétrique)**.** Un *sous-modèle paramétrique* est une famille régulière $\{p(o;\beta,\gamma):\gamma\in\mathbb{R}^r\}$ contenue dans le modèle semi-paramétrique et qui contient la vraie loi pour $\gamma=\gamma_0$. Un estimateur est *semi-paramétrique* (RAL dans le modèle) s'il est RAL dans tous les sous-modèles paramétriques.

Par exemple, si p_0 est la vraie densité de Z, la famille $p(z;\gamma)=p_0(z)\{1+\gamma^\top a(z)\}$ avec a bornée et $\mathrm{E}_0[a(Z)]=0$ est un sous-modèle de lois (pour γ petit) dont le score en $\gamma=0$ est $a(Z)$. Un estimateur semi-paramétrique étant RAL dans chaque sous-modèle, sa variance ne peut être inférieure à la borne paramétrique du sous-modèle le plus défavorable : la borne semi-paramétrique est le supremum des bornes paramétriques.

Définition 4.2 (Espace tangent de nuisance)**.** L'espace tangent de nuisance Λ du modèle semi-paramétrique est la fermeture (en moyenne quadratique) de l'ensemble des éléments BS_γ de $\mathcal{H}$, où S_γ parcourt les scores de nuisance de tous les sous-modèles paramétriques.

Les résultats paramétriques se transposent alors mot pour mot (Tsiatis, théorèmes 4.2 et 4.3) : les fonctions d'influence des estimateurs semi-paramétriques sont les $\varphi \in \Lambda^\perp$ telles que $\mathrm{E}[\varphi S_\beta^\top] = I$; le score efficace est $S_{\text{eff}} = S_\beta - \Pi(S_\beta \mid \Lambda)$; la fonction d'influence efficace $\varphi_{\text{eff}} = \{\mathrm{E}[S_{\text{eff}} S_{\text{eff}}^\top]\}^{-1} S_{\text{eff}}$ s'obtient aussi en projetant une fonction d'influence quelconque sur l'espace tangent complet. La stratégie pratique est donc : (i) caractériser $\Lambda^\perp$, ce qui suggère une *famille* d'équations d'estimation $\sum_i \varphi(O_i;\beta) = 0$; (ii) projeter S_β sur $\Lambda^\perp$ pour trouver l'élément efficace.

4.3 Modèle de moments restreints

Soit $O = (Y,X)$, $Y \in \mathbb{R}^d$, et le modèle $\mathrm{E}[Y \mid X] = \mu(X,\beta)$, sans autre restriction. Notons $\varepsilon = Y - \mu(X,\beta_0)$, $D(X) = \partial\mu(X,\beta_0)/\partial\beta^\top$ $(d \times q)$ et $V(X) = \mathrm{Var}(Y \mid X)$, supposée inversible.

Théorème 4.3 (Tsiatis, 2006, § 4.5)**.** *Dans le modèle de moments restreints :*

1. *l'espace tangent de nuisance est* $\Lambda = \{a(Y,X) \in \mathcal{H} : \mathrm{E}[a(Y,X)\varepsilon^\top \mid X] = 0\}$ *;*
2. *son orthogonal est* $\Lambda^\perp = \{A(X)\varepsilon : A(X)$ *matrice* $q \times d$ *quelconque*$\}$ *;*
3. *le score efficace est* $S_{\text{eff}}(Y,X) = D(X)^\top V(X)^{-1}\varepsilon$ *et la borne d'efficacité est*

$$\{\mathrm{E}[D(X)^\top V(X)^{-1} D(X)]\}^{-1}.$$

Esquisse. (1) Factorisons $p(y,x) = p(y \mid x)p(x)$. Un sous-modèle paramétrique fait varier $p(x;\gamma_2)$ librement et $p(y \mid x;\beta_0,\gamma_1)$ sous les contraintes $\int p(y \mid x)\,\mathrm{d}y = 1$ et $\int y\, p(y \mid x;\beta_0,\gamma_1)\,\mathrm{d}y = \mu(x,\beta_0)$. En dérivant ces contraintes en γ_1, le score $S_{\gamma_1}(Y,X)$ vérifie $\mathrm{E}[S_{\gamma_1} \mid X] = 0$ et $\mathrm{E}[S_{\gamma_1}\varepsilon^\top \mid X] = 0$; le score $S_{\gamma_2}(X)$ vérifie $\mathrm{E}[S_{\gamma_2}] = 0$. Inversement, tout élément borné vérifiant ces conditions est le score d'un sous-modèle du type $p_0\{1 + \gamma^\top a\}$. Comme $\mathrm{E}[\alpha(X)\varepsilon^\top \mid X] = 0$ pour toute fonction $\alpha(X)$, l'espace des fonctions de X est inclus dans l'ensemble décrit en (1), qui est donc la somme des deux espaces.

(2) Si $h = A(X)\varepsilon$ et $a \in \Lambda$, alors $\mathrm{E}[h^\top a] = \mathrm{E}[\mathrm{tr}\{A(X)\,\mathrm{E}(\varepsilon a^\top \mid X)\}] = 0$: donc $A(X)\varepsilon \in \Lambda^\perp$. Réciproquement, pour $h \in \mathcal{H}$, la fonction $h - \mathrm{E}[h\varepsilon^\top \mid X]V(X)^{-1}\varepsilon$ appartient à Λ (vérifier $\mathrm{E}[\cdot\,\varepsilon^\top \mid X] = 0$), d'où

$$\Pi(h \mid \Lambda^\perp) = \mathrm{E}[h\varepsilon^\top \mid X]\, V(X)^{-1}\varepsilon.$$

Si $h \in \Lambda^\perp$, $h = \Pi(h \mid \Lambda^\perp)$ est bien de la forme $A(X)\varepsilon$.

(3) $S_{\text{eff}} = \Pi(S_\beta \mid \Lambda^\perp) = \mathrm{E}[S_\beta\varepsilon^\top \mid X]V(X)^{-1}\varepsilon$. Pour calculer $\mathrm{E}[S_\beta\varepsilon^\top \mid X]$, on dérive en β l'identité $\int y\, p(y \mid x;\beta,\gamma_1)\,\mathrm{d}y = \mu(x,\beta)$:

$$\int y\, S_\beta^\top(y,x)\, p_0(y \mid x)\,\mathrm{d}y = \mathrm{E}[Y S_\beta^\top \mid X = x] = \frac{\partial\mu(x,\beta_0)}{\partial\beta^\top} = D(x).$$

Comme $\mathrm{E}[S_\beta \mid X] = 0$, $\mathrm{E}[\varepsilon S_\beta^\top \mid X] = \mathrm{E}[Y S_\beta^\top \mid X] = D(X)$, d'où $S_{\text{eff}} = D(X)^\top V(X)^{-1}\varepsilon$. Enfin $\mathrm{E}[S_{\text{eff}} S_{\text{eff}}^\top] = \mathrm{E}[D^\top V^{-1}\,\mathrm{E}(\varepsilon\varepsilon^\top \mid X)V^{-1}D] = \mathrm{E}[D^\top V^{-1} D]$. □

Conséquences.

1. *Toutes* les équations d'estimation semi-paramétriques du modèle sont (asymptotiquement) de la forme $\sum_i A(X_i)\{Y_i - \mu(X_i,\beta)\} = 0$; la fonction d'influence correspondante est $\varphi_A =$

$\{\mathrm{E}[A(X)D(X)]\}^{-1}A(X)\varepsilon$. Les GEE, avec $A = D^\top V_w^{-1}$ pour une variance de travail V_w, en sont un cas particulier : c'est la version semi-paramétrique de la proposition 2.4.

2. L'équation *efficace* est la GEE construite avec la *vraie* variance conditionnelle $V(X)$. On le vérifie directement : avec $M = \mathrm{E}[D^\top V^{-1} D]$, on a $\mathrm{E}[\varphi_A \varphi_{\text{eff}}^\top] = \{\mathrm{E}[AD]\}^{-1}\,\mathrm{E}[A\ \mathrm{E}(\varepsilon\varepsilon^\top \mid X)V^{-1}D]M^{-1} = M^{-1} = \mathrm{Var}(\varphi_{\text{eff}})$, donc $\varphi_A - \varphi_{\text{eff}} \perp \varphi_{\text{eff}}$ et

$$\mathrm{Var}(\varphi_A) = \mathrm{Var}(\varphi_{\text{eff}}) + \mathrm{Var}(\varphi_A - \varphi_{\text{eff}}) \succeq M^{-1}.$$

   C'est la justification théorique de la perte d'efficacité observée dans l'exemple 2.7.

3. Comme $V(X)$ est inconnue, on la remplace par une variance de travail $V(X;\alpha)$ (Tsiatis, § 4.6). Si le modèle de travail est correct, l'estimateur est efficace ; sinon, il reste convergent et asymptotiquement normal, de variance sandwich : on parle d'estimateur *localement efficace*.

4. Pour des données en grappes, $Y = Y_i$ est un vecteur et $V(X)$ la covariance intra-grappe : la GEE efficace utilise la vraie matrice de covariance non diagonale, ce qui motive la matrice de travail des GEE.

Cas Bernoulli. Si $Y \in \{0,1\}$ est univarié, la loi conditionnelle de Y sachant X est entièrement déterminée par sa moyenne : $V(X) = \mu(X,\beta)\{1-\mu(X,\beta)\}$. Il n'y a plus de variance inconnue, l'équation efficace est calculable, et elle coïncide avec l'équation de vraisemblance du modèle binomial à lien h : le modèle de moments restreints *est* alors un modèle paramétrique pour $Y \mid X$ (Tsiatis, p. 96). Cela vaut pour n'importe quelle fonction de lien, mais avec un lien non canonique (identité, log), les contraintes $0 < \mu < 1$ peuvent causer des problèmes de convergence. Pour des vecteurs binaires corrélés, en revanche, les corrélations ne sont pas déterminées par les moyennes et la matrice de travail redevient une approximation.

Exemple 4.4 (Question d'examen doctorat H26 : modèle de probabilité linéaire)**.** On observe un échantillon i.i.d. de (Y_i,X_i), Y_i binaire, avec $\mathrm{E}[Y_i \mid X_i] = X_i\beta$ (X_i vecteur ligne).

(a) Équation efficace. Ici $D(X_i) = X_i$ et $V(X_i) = X_i\beta(1 - X_i\beta)$; par le théorème 4.3,

$$\sum_{i=1}^{n} \frac{X_i^\top (Y_i - X_i\beta)}{X_i\beta\,(1 - X_i\beta)} = 0.$$

On la résout par moindres carrés pondérés itérés (poids $1/\{\widehat{\pi}_i(1-\widehat{\pi}_i)\}$, $\widehat{\pi}_i = X_i\widehat{\beta}$) ; la variance asymptotique estimée est $\big[\sum_i X_i^\top X_i/\{\widehat{\pi}_i(1-\widehat{\pi}_i)\}\big]^{-1}$.

(b) Estimateur explicite, convergent mais non efficace. En prenant $A(X_i) = X_i^\top$ (variance de travail constante) : $\sum_i X_i^\top(Y_i - X_i\beta) = 0$, soit $\widehat{\beta} = (X^\top X)^{-1}X^\top Y$, l'estimateur des moindres carrés ordinaires.

(c) Variance robuste. Dans la forme générale $n^{-1}\widehat{V}_1^{-1}\widehat{V}_2\widehat{V}_1^{-1}$ avec $D_i = X_i$ et $V_i = \sigma^2$ (la constante se simplifie) :

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \Big(\sum_i X_i^\top X_i\Big)^{-1}\Big(\sum_i (Y_i - X_i\widehat{\beta})^2 X_i^\top X_i\Big)\Big(\sum_i X_i^\top X_i\Big)^{-1},$$

l'estimateur de White (HC0), qui tient compte de l'hétéroscédasticité $\mathrm{Var}(Y_i \mid X_i) = \pi_i(1-\pi_i)$.

```
set.seed(7360); n <- 500
x <- runif(n); X <- cbind(1, x); y <- rbinom(n, 1, 0.1 + 0.7*x)
```

```
XtX.inv <- solve(crossprod(X))
b.mco <- drop(XtX.inv %*% crossprod(X, y)); e <- drop(y - X %*% b.mco)
V.sand <- XtX.inv %*% crossprod(X * e) %*% XtX.inv           # (c)
b <- b.mco                                                   # (a) par MCP itérés
for (it in 1:20) { p <- drop(X %*% b); w <- 1/(p*(1-p))
  b <- drop(solve(crossprod(X, w*X), crossprod(X, w*y))) }
p <- drop(X %*% b); V.eff <- solve(crossprod(X, X/(p*(1-p))))
round(rbind(MCO=b.mco, ET.sandwich=sqrt(diag(V.sand)), efficace=b,
            ET.efficace=sqrt(diag(V.eff))), 4)
```

```
                      x
MCO         0.0731 0.8322
ET.sandwich 0.0323 0.0567
efficace    0.0679 0.8385
ET.efficace 0.0308 0.0542
```

Les deux estimateurs sont proches de $\beta = (0{,}1;\ 0{,}7)$ compte tenu de l'erreur d'échantillonnage ; l'estimateur efficace gagne environ 5 % d'erreur type (environ 9 % en variance). Le gain serait plus grand si les probabilités s'approchaient davantage de 0 ou de 1, mais c'est aussi là que les poids explosent et que $X_i\widehat{\beta} \notin (0{,}1)$ devient possible.

4.4 Aperçu de l'estimation ciblée (TMLE)

Le modèle de moments restreints est un cas où $\Lambda^{\perp}$ se calcule explicitement. Van der Laan et Rose (2011) considèrent un paramètre défini comme une *application* de la loi des données, dans un modèle non paramétrique. Nous en donnons ici l'essentiel ; l'étude de cas de la séance 7 le développera.

Paramètre cible. Avec $O = (W,A,Y)$, W des covariables, $A \in \{0,1\}$ un traitement et $Y \in \{0,1\}$, soit $\bar{Q}(A,W) = \mathrm{E}[Y \mid A,W]$, $g(a \mid W) = \mathbb{P}(A = a \mid W)$ et Q_W la loi de W. La différence de risque causale (sous les hypothèses d'identification usuelles) est

$$\Psi(P) = \mathrm{E}_W\big[\bar{Q}(1,W) - \bar{Q}(0,W)\big].$$

L'estimateur par substitution applique Ψ à une estimation de P : $\Psi(\widehat{P}) = n^{-1}\sum_i \{\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)\}$.

Fonction d'influence efficace. Dans le modèle non paramétrique, l'unique fonction d'influence (donc efficace) de Ψ est

$$D^*(O) = \underbrace{\Big(\frac{A}{g(1 \mid W)} - \frac{1-A}{g(0 \mid W)}\Big)}_{H_g(A,W)}\{Y - \bar{Q}(A,W)\} \;+\; \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P).$$

On peut l'obtenir par la méthode delta fonctionnelle appliquée à l'estimateur non paramétrique de Ψ lorsque W est discrète (van der Laan et Rose, annexe A.1 ; exercice 5.7 ci-dessous). Le premier terme est la composante dans l'espace tangent de $P_{Y|A,W}$, le second celle de Q_W ; la composante dans l'espace tangent de $g = P_{A|W}$ est nulle, mais g apparaît quand même dans D^* et doit donc être estimée.

L'idée du TMLE. Un estimateur initial $\bar{Q}_n$ (souvent obtenu par apprentissage automatique) ne résout pas l'équation $\sum_i D^*(O_i) = 0$, et l'estimateur par substitution hérite de son biais. On le « fluctue » le long d'un sous-modèle paramétrique qui passe par $\bar{Q}_n$ et dont le score contient la composante $Y \mid A,W$ de D^* :

$$\operatorname{logit} \bar{Q}_n^{(\epsilon)}(A,W) = \operatorname{logit} \bar{Q}_n(A,W) + \epsilon\, H_{g_n}(A,W).$$

H_{g_n} est la *covariable astucieuse* (*clever covariate*). On estime ϵ par une régression logistique de Y sur H_{g_n} avec $\operatorname{logit} \bar{Q}_n$ en décalage (*offset*), sans ordonnée à l'origine. L'équation de score de cette régression est exactement $\sum_i H_{g_n}(A_i,W_i)\{Y_i - \bar{Q}_n^*(A_i,W_i)\} = 0$, avec $\bar{Q}_n^* = \bar{Q}_n^{(\widehat{\epsilon})}$. La loi empirique de W n'a pas à être fluctuée (elle résout déjà sa composante). Le TMLE est alors $\Psi(\widehat{P}^*) = n^{-1}\sum_i \{\bar{Q}_n^*(1,W_i) - \bar{Q}_n^*(0,W_i)\}$, qui résout $\sum_i D^*(O_i) = 0$ tout en restant un estimateur par substitution (il respecte les bornes de Ψ).

Double robustesse et efficacité. Un calcul direct donne, pour toute paire $(\bar{Q},g)$,

$$\begin{aligned}\Psi(\bar{Q},Q_{W,0}) - \Psi(P_0) + \mathrm{E}_0\big[D^*(\bar{Q},g)(O)\big] \\ = \mathrm{E}_0\Big[\sum_{a\in\{0,1\}} (-1)^{1-a}\frac{g_0(a \mid W) - g(a \mid W)}{g(a \mid W)}\{\bar{Q}_0(a,W) - \bar{Q}(a,W)\}\Big] =: R_2.\end{aligned}$$

Le reste R_2 est un *produit* des erreurs sur g et sur $\bar{Q}$. Comme le TMLE résout l'équation empirique de D^*,

$$\Psi(\widehat{P}^*) - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} D^*(\bar{Q}_n^*,g_n)(O_i) - \mathrm{E}_0[D^*(\bar{Q}_n^*,g_n)] + R_2.$$

— *Double robustesse* : $R_2 = 0$ dès que $\bar{Q}_n \to \bar{Q}_0$ *ou* $g_n \to g_0$; l'estimateur est convergent si l'un des deux est bien estimé.

— *Efficacité locale* : si les deux sont estimés de façon convergente, le TMLE est asymptotiquement linéaire de fonction d'influence $D^*(\bar{Q}_0,g_0)$, donc efficace.

— *Condition de taux* : il faut que $R_2 = o_p(n^{-1/2})$. Par Cauchy-Schwarz (avec g borné loin de 0), $|R_2| \lesssim \|g_n - g_0\|_2\, \|\bar{Q}_n - \bar{Q}_0\|_2$: il suffit que le *produit* des taux soit plus rapide que $n^{-1/2}$, par exemple deux taux en $o_p(n^{-1/4})$. Avec des modèles paramétriques corrects (taux $n^{-1/2}$), c'est automatique ; avec de l'apprentissage automatique, ce n'est pas garanti (il faut aussi contrôler le terme empirique, par des conditions de Donsker ou par validation croisée).

Niveau doctorat – à retenir

— Fonctions d'influence des estimateurs semi-paramétriques = éléments de $\Lambda^\perp$ normalisés ; $S_{\text{eff}} = S_\beta - \Pi(S_\beta \mid \Lambda)$; borne $= \{\mathrm{E}[S_{\text{eff}} S_{\text{eff}}^\top]\}^{-1}$.

— Moments restreints : $\Lambda^\perp = \{A(X)(Y - \mu(X,\beta))\}$; $S_{\text{eff}} = D^\top V^{-1}(Y - \mu)$ avec la vraie variance ; les GEE avec variance de travail sont localement efficaces.

— Bernoulli univarié : la variance est déterminée par la moyenne, l'estimateur efficace est calculable.

— TMLE : estimateur initial + fluctuation par la covariable astucieuse pour résoudre l'équation de la fonction d'influence efficace ; doublement robuste ; efficace si le produit des taux est $o(n^{-1/2})$.

5 Exercices

Exercice 5.1 (Fonction d'influence dans l'ANOVA à un effet aléatoire ; source : exercices H26, A. Bureau)**.** Considérons le modèle $Y_{ij} = \mu + b_i + \epsilon_{ij}$, $i = 1,\ldots,n$, $j = 1,\ldots,J$, avec $b_i \sim \mathcal{N}(0,\sigma_b^2)$ et $\epsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$ indépendants (ANOVA équilibrée à un effet aléatoire). Soit

$$S_1^2 = \frac{J}{n-1}\sum_{i=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \qquad S_2^2 = \frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i\cdot})^2.$$

On estime σ_b^2 par une combinaison linéaire $\widehat{\zeta} = c_1 S_1^2 + c_2 S_2^2$, où c_1 et c_2 sont des constantes.

(a) Montrez que $c_1 = 1/J$, $c_2 = -1/J$ donne un estimateur sans biais de σ_b^2.

(b) Obtenez la fonction d'influence de $\widehat{\zeta}$ (l'unité d'observation étant le vecteur $Y_i = (Y_{i1},\ldots,Y_{iJ})$, avec $n \to \infty$ et J fixe), ce qui démontre que $\widehat{\zeta}$ est asymptotiquement linéaire.

Note : l'énoncé H26 écrivait $\widehat{\zeta} = cS_1^2 + S_2^2$; la forme à deux constantes est nécessaire pour estimer σ_b^2.

Exercice 5.2 (Robustification ; source : examen final H23, question 5)**.** La fonction d'influence de l'EMV de la moyenne d'une loi normale d'écart type σ connu est $\mathrm{IF}(y,\mu) = y - \mu$. Donnez deux façons de la modifier pour obtenir une estimation robuste de μ et écrivez la nouvelle fonction d'influence dans chaque cas. (Solution : exemple 3.10.)

Exercice 5.3 (Robustesse dans un GLMM ; source : examen final H26, question 1c)**.** Dans un modèle de Poisson mixte du nombre de cas de mélanome par localité, avec effets aléatoires normaux des régions et des provinces et le rayonnement UVB régional comme covariable, décrivez une approche pour obtenir une estimation du coefficient de l'UVB robuste aux valeurs extrêmes d'incidence dans une région ou une province, tout en conservant ces sources de variation. Précisez les paramètres additionnels introduits.

Exercice 5.4 (GEE ; source : matériel H26, fichier `gee.R`, adapté)**.** Avec les données `respiratory` de `geepack` (identifiant unique construit comme à la section 2.10) :

(a) ajustez le modèle `outcome ~ center + treat*visit + sex + baseline + age` par GEE avec les structures indépendance, échangeable, AR(1) et non structurée ; comparez estimations et erreurs types robustes ;

(b) pour la structure échangeable, comparez les erreurs types naïves et robustes (module `gee`) ;

(c) ajustez un GLMM logistique à intercept aléatoire (`glmer`) et comparez le coefficient du traitement à celui du GEE ; vérifiez l'approximation $\beta^M \approx \beta^C / \sqrt{1 + 0{,}346\,\sigma_b^2}$.

Exercice 5.5 (Niveau doctorat ; source : exercices doctorat H26, question 1)**.** Dérivez l'équation d'estimation du maximum de vraisemblance des effets fixes β d'un modèle linéaire mixte généralisé donnée par l'équation (3.10) de Wu (2010) : $\sum_{i=1}^{n} \mathrm{E}\big[\partial \log f(y_i \mid X_i,b_i;\beta,\sigma)/\partial\beta \mid y_i,X_i;\theta\big] = 0$. Expliquez en quoi il s'agit d'une forme de fonction d'influence.

Exercice 5.6 (Niveau doctorat ; source : exercices doctorat H26, question 2 ; Tsiatis, 2006, exercice 4.3)**.** Résolvez les parties b) à d) de l'exercice 4.3 de Tsiatis (2006), qui porte sur un modèle semi-paramétrique où la variance conditionnelle $V(X,\beta)$ de Y est aussi spécifiée : projection du score sur

l'orthogonal de l'espace tangent de nuisance, score efficace, et procédure d'estimation adaptative faisant intervenir un modèle de travail pour le moment conditionnel d'ordre 4. *Indication : pour des restrictions de moments conditionnels* $\mathrm{E}[m(Y,X;\beta) \mid X] = 0$*, l'orthogonal de l'espace tangent de nuisance est* $\{A(X)\, m(Y,X;\beta)\}$*,* $A(X)$ *quelconque.*

Exercice 5.7 (Niveau doctorat ; source : exercices doctorat H26, question 3)**.** Écrivez explicitement les dérivées fonctionnelles (dérivées par rapport à la loi de probabilité) qui permettent de linéariser le risque contrefactuel chez les sujets traités $\psi_n(1) = \Phi(P_n)$ à la page 528 de van der Laan et Rose (2011), lorsque W est discrète. Déduisez-en la fonction d'influence de $\psi_n(1)$ et comparez-la à D^* (section 4.4).

Exercice 5.8 (Niveau doctorat ; source : examen final H26, question 4)**.** On modélise $\mathrm{E}[Y \mid X] = X\beta$ pour Y binaire, à partir d'un échantillon i.i.d. (a) À partir de la formule générale du score efficace de Tsiatis, obtenez l'équation d'estimation semi-paramétrique efficace de β. (b) Proposez une équation d'estimation convergente, non efficace, à solution explicite, et donnez cette solution. (c) Donnez un estimateur robuste de la variance de l'estimateur obtenu en (b). (Solution : exemple 4.4.)

Références

— Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canadian Journal of Statistics*, 32, 169–180.

— Cantoni, E. et Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96, 1022–1030.

— Carey, V., Zeger, S. L. et Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80, 517–526.

— Fay, M. P. et Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57, 1198–1206.

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

— Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. et Stahel, W. A. (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley.

— Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

— Huber, P. J. et Ronchetti, E. M. (2009). *Robust Statistics*, 2e éd. Wiley.

— Kauermann, G. et Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387–1396.

— Koller, M. (2016). robustlmm : an R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(6).

— Lange, K. L., Little, R. J. A. et Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881–896.

— Liang, K.-Y. et Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

— Mancl, L. A. et DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57, 126–134.

— Maronna, R. A., Martin, R. D., Yohai, V. J. et Salibián-Barrera, M. (2019). *Robust Statistics : Theory and Methods (with R)*, 2e éd. Wiley.

— Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120–125.

— Pepe, M. S. et Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics – Simulation and Computation*, 23, 939–951.

— Pinheiro, J. C., Liu, C. et Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10, 249–276.

— Preisser, J. S. et Qaqish, B. F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics*, 55, 574–579.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer. (Chapitres 3 et 4.)

— van der Laan, M. J. et Rose, S. (2011). *Targeted Learning : Causal Inference for Observational and Experimental Data*. Springer. (Chapitre 5 et annexe A.1.)

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

— Zeger, S. L., Liang, K.-Y. et Albert, P. S. (1988). Models for longitudinal data : a generalized estimating equation approach. *Biometrics*, 44, 1049–1060.