Modèle de médiation à réponse binaire
Approche non paramétrique pour tester l'effet direct

Jean-Michel Galharret
Anne Philippe
Paul Rochet

Laboratoire Jean Leray, Nantes

11 décembre 2019

La médiation dans le modèle linéaire

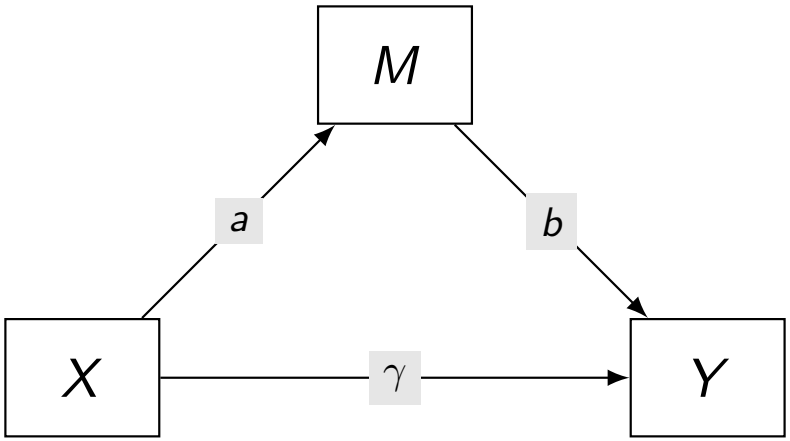


FIGURE – Baron & Kenny (1986)

$$\mathbb{E}(Y|X,M) = \beta + \gamma X + bM \quad (1)$$

L'effet direct de X entre deux valeurs x, x^* sur Y.

$$\begin{aligned} d &= \mathbb{E}(Y|X=x, M=m) - \mathbb{E}(Y|X=x^*, M=m) \\ &= \gamma(x - x^*) \end{aligned}$$

Test effet direct

Tester l'effet direct dans le contexte linéaire revient à tester

$$\mathcal{H}_0 : \gamma = 0 \text{ vs } \mathcal{H}_1 : \gamma \neq 0$$

L'effet direct dans le modèle logistique

$E(Y|X,M) = \beta_1 + \gamma X + bM$ (1) s'écrit dans le modèle logistique

$$\text{logit}\big(\mathbb{P}(Y=1|X,M)\big) = \beta_1 + \gamma X + bM$$

On définit dans ce cadre :

$$d = \mathbb{P}(Y=1|X=x,M=m) - \mathbb{P}(Y=1|X=x^*,M=m)$$

Test effet direct

Tester l'effet direct dans le modèle logistique revient à tester

$$\mathcal{H}_0 : \gamma = 0 \text{ vs } \mathcal{H}_1 : \gamma \neq 0$$

En effet, en considérant l'odds ratio OR^d de l'effet direct de X entre x, x^* : On a l'équivalence immédiate :

$$d = 0 \Longleftrightarrow OR^d = 1 \Longleftrightarrow \gamma = 0$$

.

Test de l'effet direct dans le modèle logistique

Pour tester $\mathcal{H}_0 : \gamma = 0$ on peut utiliser deux tests :

- Soit $\widehat{\gamma}$ l'estimateur du maximum de vraisemblance de γ et $\widehat{\sigma}_{\widehat{\gamma}}$ l'écart-type estimé de cet estimateur.La statistique de test $z = \dfrac{\widehat{\gamma}}{\widehat{\sigma}_{\widehat{\gamma}}}$ qui sous $\mathcal{H}_0$ est asymptotiquement gaussien (test de Wald).
- Soient $\widehat{L_0}$ et $\widehat{L_1}$ les vraisemblances estimées respectivement des modèles $\text{logit}(p) = \beta_1 + bM$ et $\text{logit}(p) = \beta_1 + \gamma X + bM$.
  La statistique de test $D = 2[\log(\widehat{L_1}) - \log(\widehat{L_0})]$, qui sous H_O suit approximativement une loi du $\chi^2(1)$ (LR test : Likelihood Ratio test).

L'approche non paramétrique de l'effet direct

- L'absence d'effet direct peut simplement se traduire en terme de lois de probabilité, soit A une partie Y-mesurable :

$$\mathcal{H}_0 : P(Y \in A \,|X, M) \overset{ps}{=} P(Y \in A \,|M)$$

- Cette condition est difficile à tester sans hypothèse sur le lien entre Y, X, M.
- On suppose que Y, X, M vérifient un modèle de régression non paramétrique :

$$Y \quad = \quad \rho(X, M) + \varepsilon \tag{2}$$

où ε est l'erreur résiduelle indépendante de X, M et ρ est une fonction mesurable inconnue.

L'hypothèse du test non paramétrique

Hypothèse : $Y \in L^1$
Dans ce cas on a :

$$\rho(X, M) = \mathbb{E}(Y|X, M)$$

Pour un modèle de régression non paramétrique l'hypothèse $\mathcal{H}_0$ est équivalente à :

$$\mathbb{E}(Y|X, M) \overset{ps}{=} \mathbb{E}(Y|M)$$

que l'on peut écrire, avec $\phi(M) := \mathbb{E}(Y|M)$:

$$\theta := \mathbb{E}\Big|\rho(X, M) - \phi(M)\Big| = 0$$

Test non paramétrique

L'absence d'effet direct pour le modèle non paramétrique revient à tester

$$\mathcal{H}_0 : \theta = 0 \text{ vs } \mathcal{H}_1 : \theta \neq 0$$

Statistique de test :

- On peut estimer ces leux espérances conditionnelles par la méthode de Nadaraya et Watson

$$\widehat{\rho}(x,m) := \frac{\sum_{i=1}^{n} Y_i K_h(X_i - x) K_h(M_i - m)}{\sum_{i=1}^{n} K_h(X_i - x) K_h(M_i - m)}$$

$$\widehat{\phi}(m) := \frac{\sum_{i=1}^{n} Y_i K_h(M_i - m)}{\sum_{i=1}^{n} K_h(M_i - m)}.$$

- Et par suite la statistique de test s'écrit :

$$\widehat{\theta} := \frac{1}{n} \sum_{j=1}^{n} \left| \widehat{\rho}(X_j, M_j) - \widehat{\phi}(M_j) \right|.$$

Distribution de $\widehat{\theta}$ sous $\mathcal{H}_0$ dans le cadre binaire

Estimation de la distribution de $\widehat{\theta}$ sous $\mathcal{H}_0$ par une procédure bootstrap :

- On simule B échantillons de taille n : $\forall (i,b)$

$$(X_i^{(b)}, M_i^{(b)}) \overset{i.i.d}{\sim} \mathcal{U}\left((X_j, M_j)_{j=1,\dots,n}\right)$$

- Sous $\mathcal{H}_0$: $P(Y=1|M,X) = P(Y=1|M) = \phi(M)$.
  Ainsi, $\forall (i,b)$, on sait approximativement simuler $Y_i^{(b)}$:

$$Y_i^{(b)} \overset{i.i.d}{\sim} \mathcal{B}\left(\hat{\phi}\left(M_i^{(b)}\right)\right).$$

- On calcule la statistique $\widehat{\theta}_b^*$ sur chaque $(Y_i^{(b)}, X_i^{(b)}, M_i^{(b)})_{\substack{i=1,\dots,n \\ b=1,\dots,B}}$.

p-value

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\widehat{\theta}_b^* > \widehat{\theta}}.$$

où $\widehat{\theta}$ est l'estimation de θ calculée sur l'échantillon $(X_i, M_i)_{i=1,\dots,n}$.

Comparaison de la puissance par simulations

- Les simulations sont basées sur un couple $(X_i, M_i)_{i=1,\dots,244}$ de valeurs issues d'une étude du CREN dans le cadre de l'étude sur le bien-être des élèves à l'école et au collège.
- Pour $n = 20, 30, 50, 100$, $\gamma \in [0, 1]$ on va simuler $k = 10000$ échantillons

$$Y_i \sim \mathcal{B}\big(\rho_\gamma(X_i, M_i)\big)$$

  - Cas 1 : modèle logistique

$$\rho_\gamma(X, M) = \frac{1}{1 + \exp\left(\beta + 2\gamma X + (1-\gamma)M\right)}$$

  - Cas 2 : modèle monotone

$$\rho_\gamma(X, M) = 0.5\gamma \times \mathbb{1}_{X>3} + 0.25(1-\gamma) \times \mathbb{1}_{M>4}$$

  - Cas 3 : modèle non monotone

$$\rho_\gamma(X, M) = \gamma(aX^2 + bX + c) + 0.1(1-\gamma)M$$

Remarque

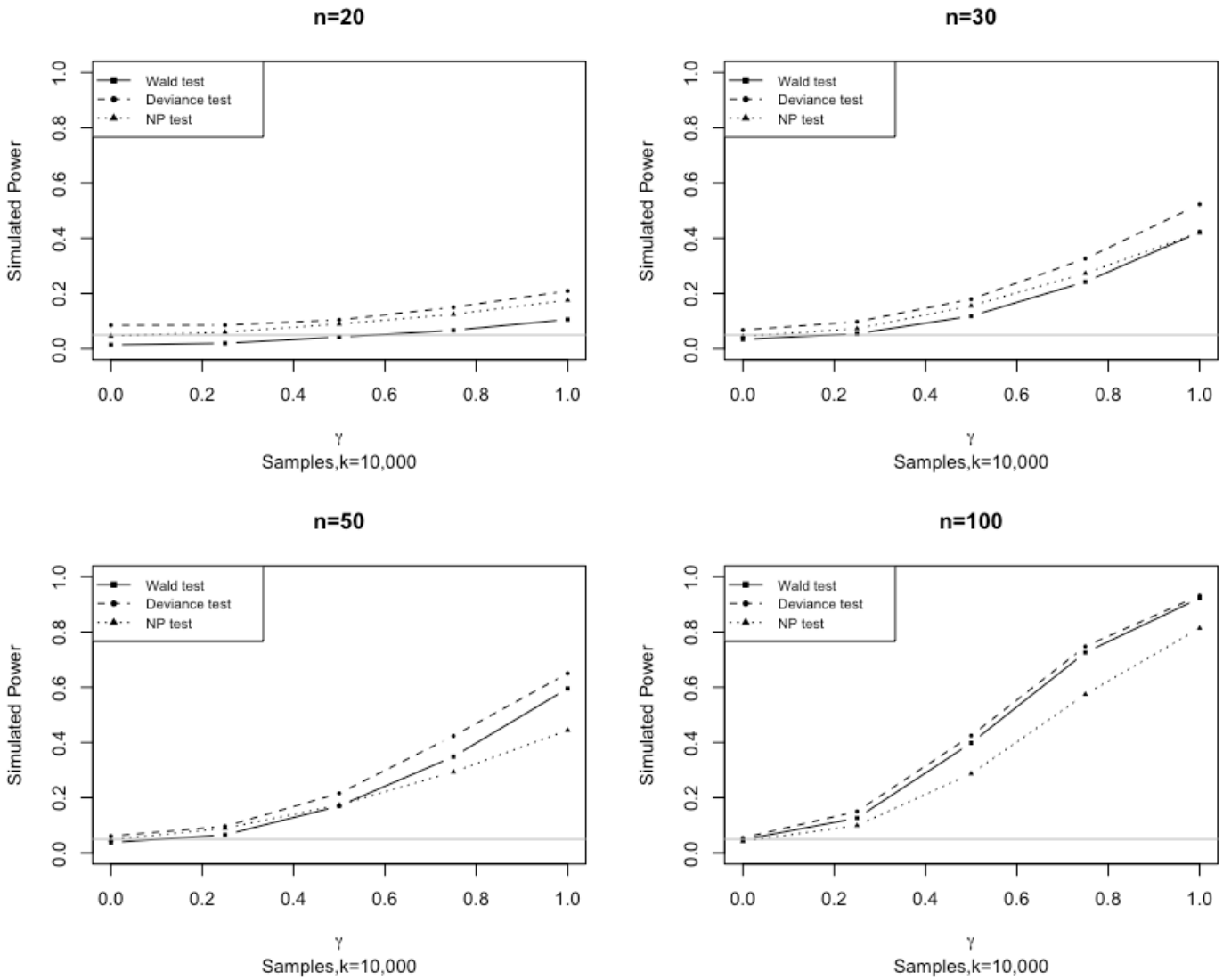
$\gamma = 0$ est équivalent à $\mathcal{H}_0$, et l'effet de X sur Y augmente avec γ (effet maximal pour $\gamma = 1$).

Niveau empirique des tests

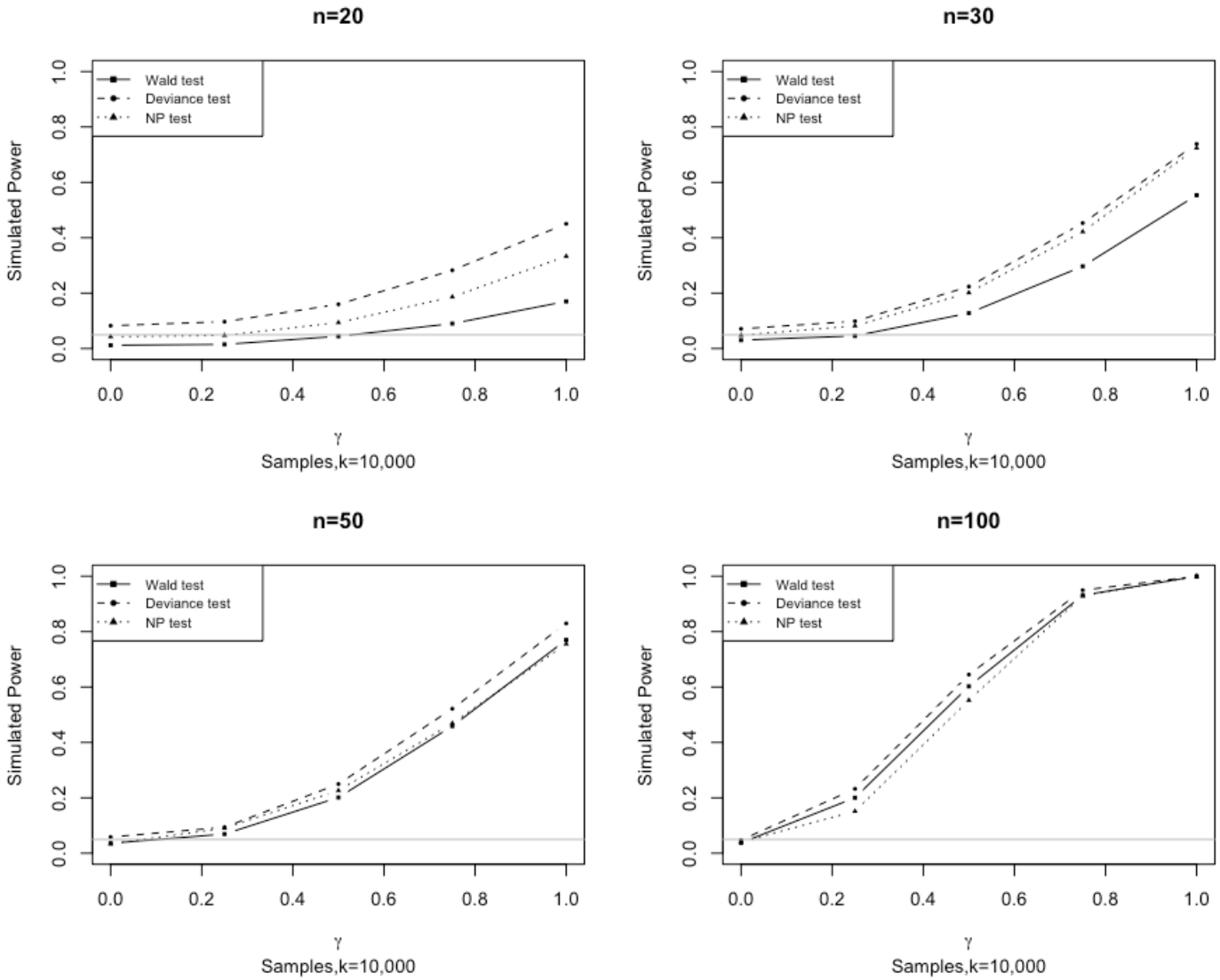
Soit $\pi_\bullet(\gamma)$ la probabilité de rejeter $\mathcal{H}_0$ pour les tests.

		20	30	50	100
logistique	π_W	0.014	0.034	0.038	**0.048**
	π_{LR}	0.085	0.068	0.061	0.055
	π_{NP}	**0.046**	**0.045**	**0.048**	0.042
indicatrice	π_W	0.012	0.030	0.036	0.037
	π_{LR}	0.083	0.071	**0.058**	**0.047**
	π_{NP}	**0.043**	**0.048**	0.034	0.040
quadratique	π_W	0.027	0.037	0.041	0.048
	π_{LR}	0.078	0.064	**0.058**	0.055
	π_{NP}	**0.054**	**0.049**	**0.042**	**0.050**

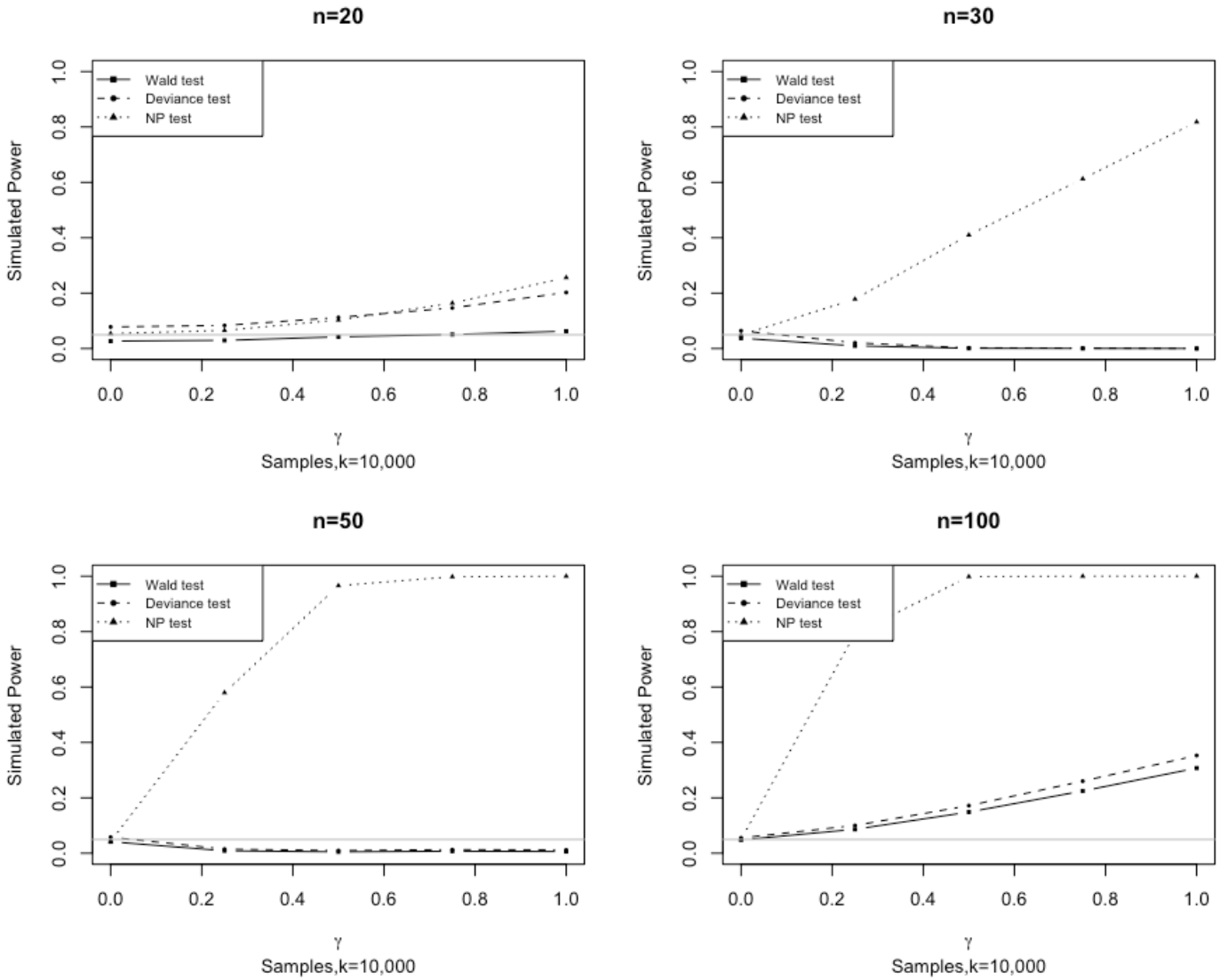
Modèle logistique



Modèle monotone



Modèle non monotone



Conclusion

- Pour un niveau $\alpha = 5\%$, le test non paramétrique a systématiquement un meilleur niveau empirique que les tests paramétriques pour les petits échantillons ($n = 20, 30$).
- Dans un cadre logistique, le test non paramétrique donne des résultats satisfaisants relativement au deux autres tests et s'avère meilleur que le test de Wald lorsque $n = 20, 30$
- Dans le cadre monotone, les trois tests ont des performances comparables.
- Dans le cadre non monotone, la plus value apportée par le test non paramétrique est incontestable : les tests paramétriques sont alors totalement inadaptés.