Demi-journée Data Science n°2 :Evaluation des méthodes de sélection de variable (Protocole de simulation)

M1 MIASHS

12 Novembre 2025

Sommaire

01	Méthodes utilisées
02	Résultats : Métriques
03	Méthodologie : échantillonnage
04	Visualisation
05	Conclusion

I. MÉTHODES UTILISÉES :

- Recherche exhaustive avec critère CP Mallows
- Méthode stepwise forward avec critères AIC
- Méthode stepwise backward avec critères BIC
- Méthode test de significativité des coefficients (test de student)
- Méthode oracle

II. RÉSULTATS PAR MÉTHODES ET PAR JEUX DE DONNÉES : MÉTRIQUES

Xp indépendantes

method	precision	recall	specificity	rmse	prediction
Cp	1	0.75	1.0	0.0585	0.6995
AIC	1	0.75	1.0	0.0585	0.6995
BIC	0	*NA*	0.7	0.5477	0.6995
OLS	1	1.00	1.0	0.0359	0.6995
Oracle	1	1.00	1.0	0.0403	0.6995

Xp avec dépendance par blocs

	method	precision	recall	specificity	rmse	prediction
1	Cp	0.75	1	0.8571	0.0585	0.7256
2	AIC	0.75	1	0.8571	0.0585	0.7305
3	BIC	1.00	1	1.0000	0.0403	0.7200
4	OLS	1.00	1	1.0000	0.0359	0.7176
5	Oracle	1.00	1	1.0000	0.0403	0.7150

III. MÉTHODOLOGIE

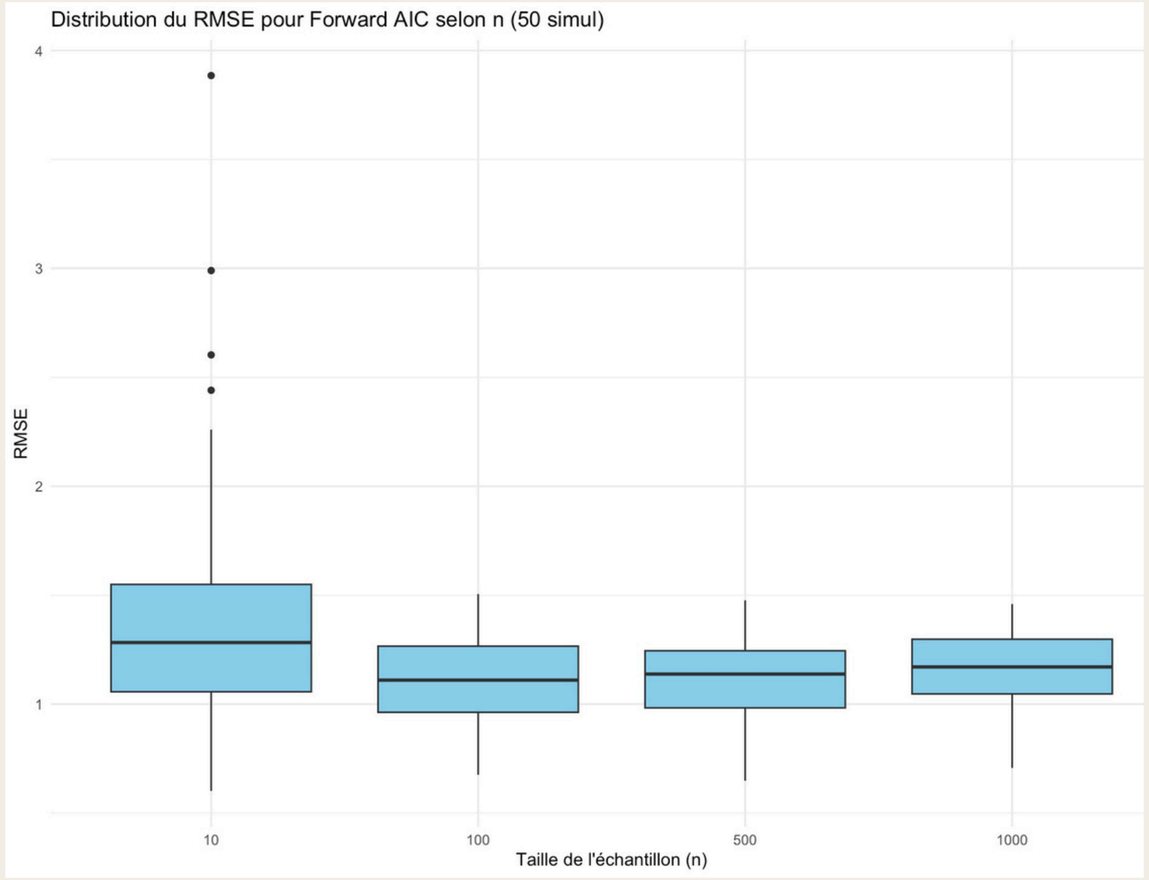
On regarde les métriques RMSE sur les méthodes de sélection de variables CP de Mallows et AIC (Forward) pour les données générées avec les Xp variables indépendantes et pareillement pour les Xp variables avec dépendance par blocs :

- On fixe p à 10, p étant le nombre de variables
- On teste pour différentes valeurs de n : **10, 100, 500, 1000**
- On réalise un échantillonage en faisant 50 simulations (50 RMSE pour chaque boxplot).

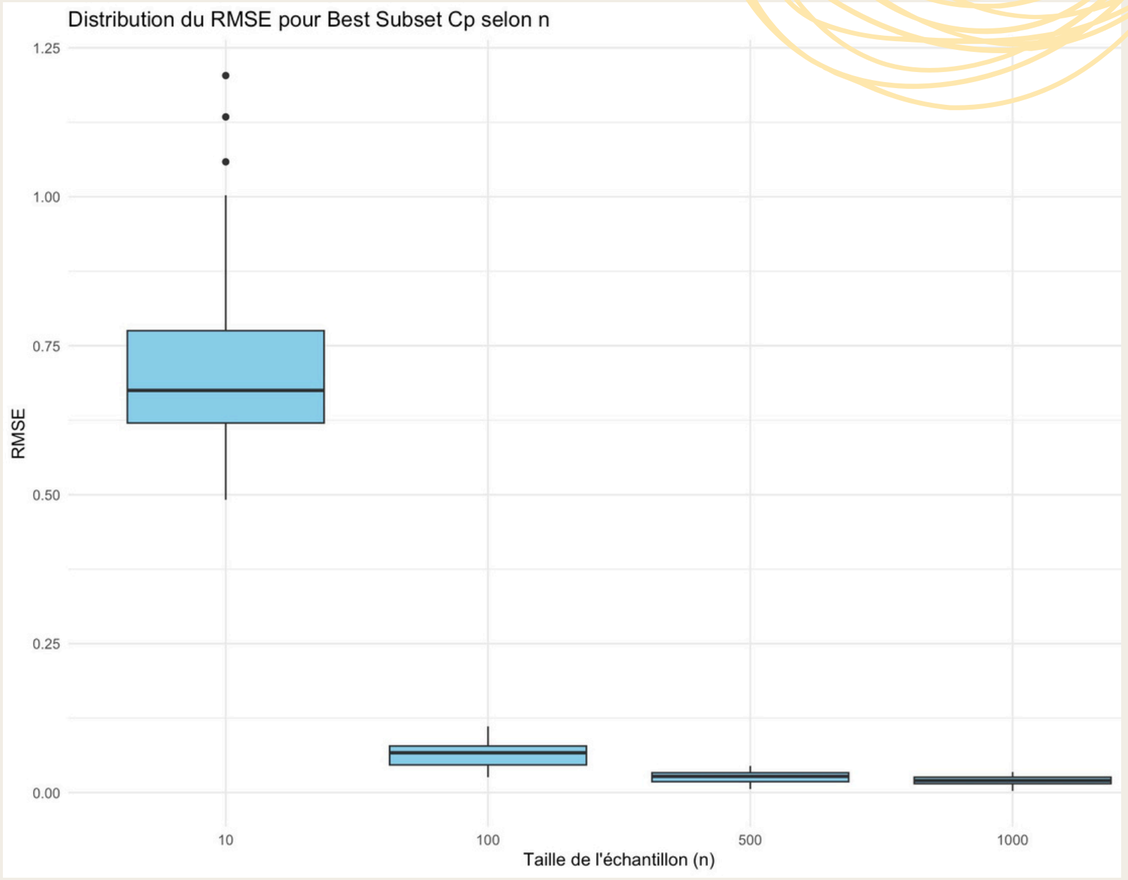
IV. VISUALISATIONS

Xp indépendantes

Boxplots des RMSE pour la méthode : AIC (Forward)

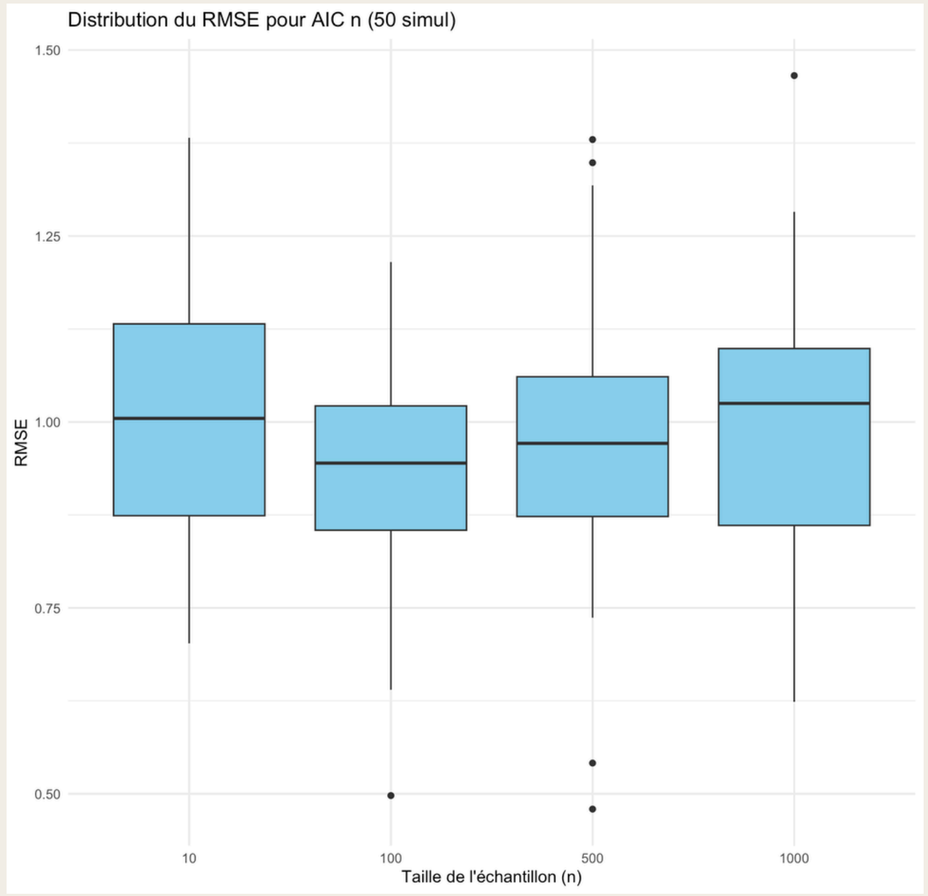


Pour la méthode : Cp de Mallows

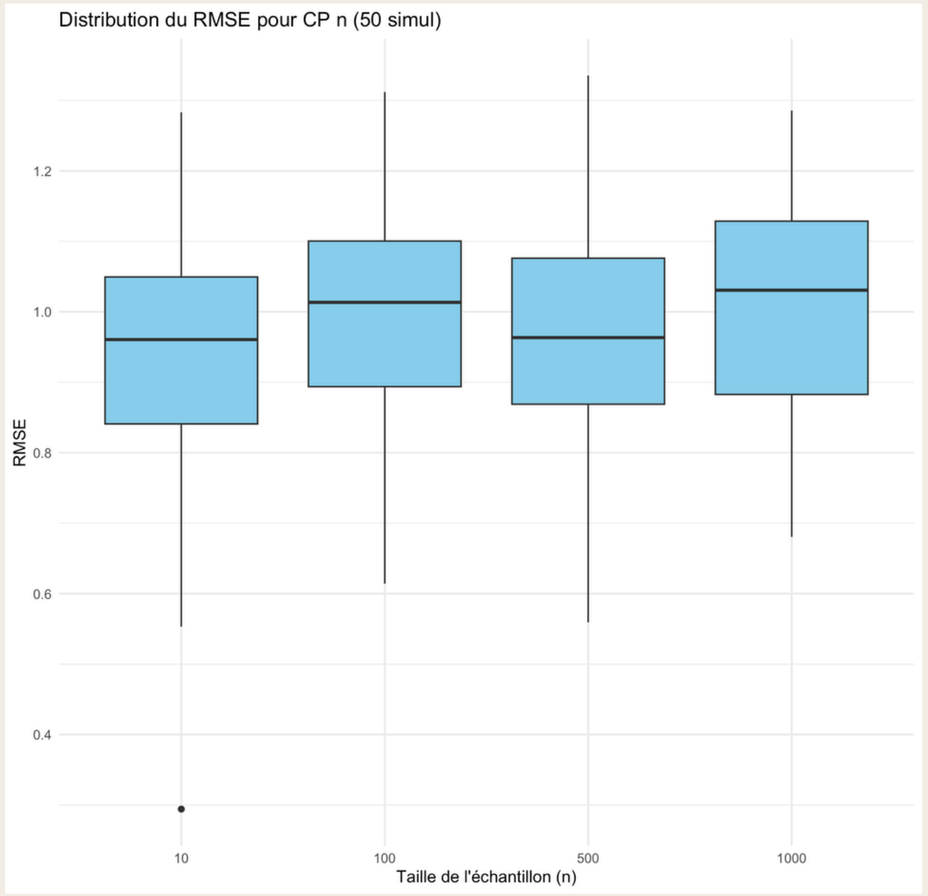


Xp avec dépendance par blocs

Boxplots des RMSE pour la méthode : AIC (Forward)



Pour la méthode : Cp de Mallows)



Conclusion

Méthode de sélection de variables :

- OLS et Oracles semblent meilleurs pour estimer les coeffs (Xp Independantes)
- BIC (Backward) / OLS et Oracle semblent plus performants (Xp dépendant par boc) et AIC pour prédire

Régresseurs dépendants entre eux ou non :

- De manière générale nous avons observé sur les boxplots que le fait de simuler des données avec des régresseurs indépendants entre eux, permet une meilleure estimation des paramètres grâce à des valeurs plus faibles du RMSE.

Effet du ratio n/p :

- Nous observons clairement que plus le nombre d’individus est grand, plus l’ajustement des coefficients du modèle est précis d'où l’importance de travailler sur des jeux de données ou le nombre d’individus est d’un autre ordre de grandeur que p. (n>>p)