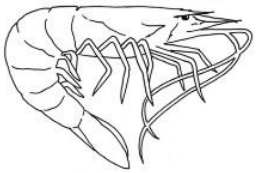
Les modèles statistiques classiques d'analyse des données binaires, source de biais d'estimation importants?

Ugoline Godeau, Frédéric Gosselin

Présentation donnée le 24/01/2020 en visio aux

Premières Journées Scientifiques Annuelles des sites INRAE du Loiret



GAMBAS

Région Centre-Val de Loire

Les modèles statistiques classiques d'analyse des données binaires, source de biais d'estimation importants?

Ugoline Godeau & Frédéric Gosselin
INRAE - UR Ecosystèmes Forestiers (EFNO)
Domaine des Barres – Nogent-sur-Vernisson
godeau.ugoline@gmail.com & frederic.gosselin@inrae.fr

INRAE

24/11/2020

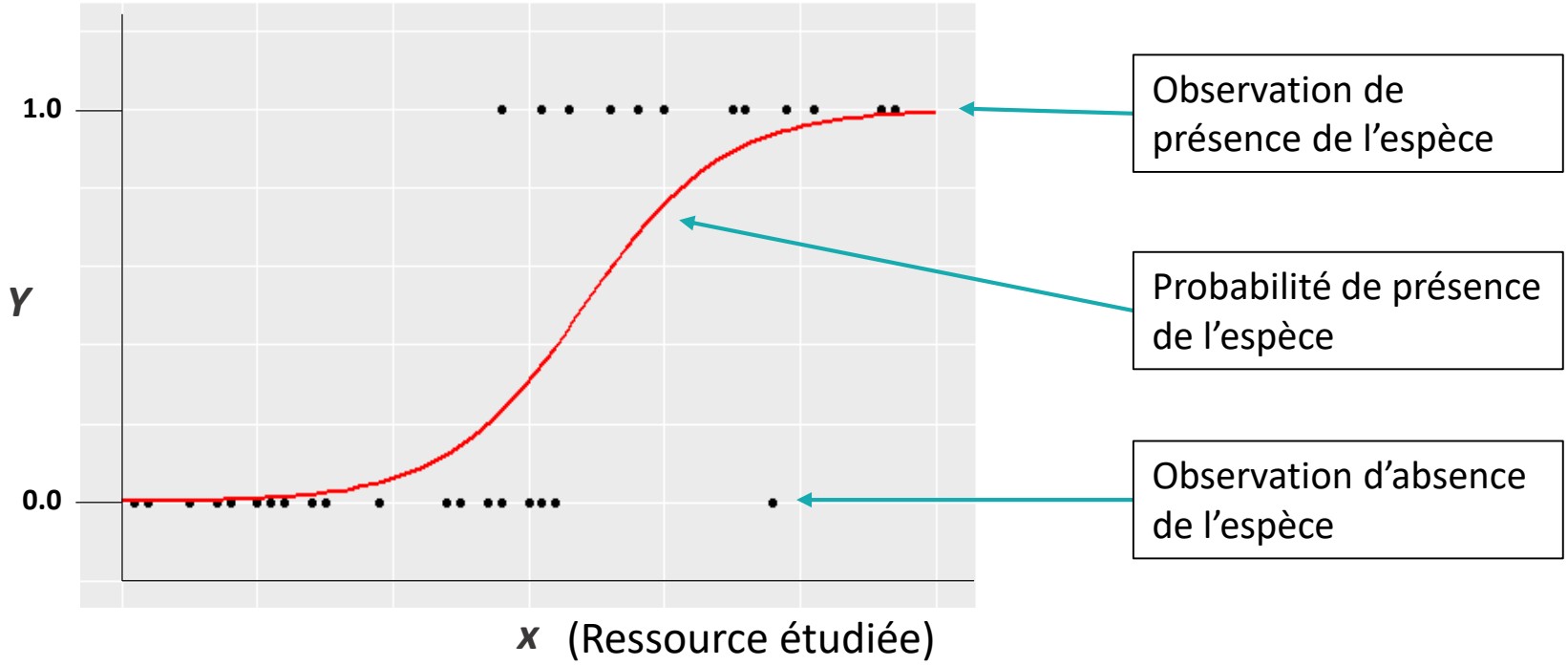
JSA2020

Données binaires

- Modélise Y (avec deux états) en fonction de X
- Y : Succès (1) ou échec (0)
- **Domaines d'application très variés, e.g. :**
  - Analyses de survie (physiologie, teste de traitement …)
  - Présence absence (apparition d'agents pathogènes en épidémiologie, espèces en écologie, maladies en médecine, comportement en sociologie ou en éthologie…)

Données binaires : SDMs

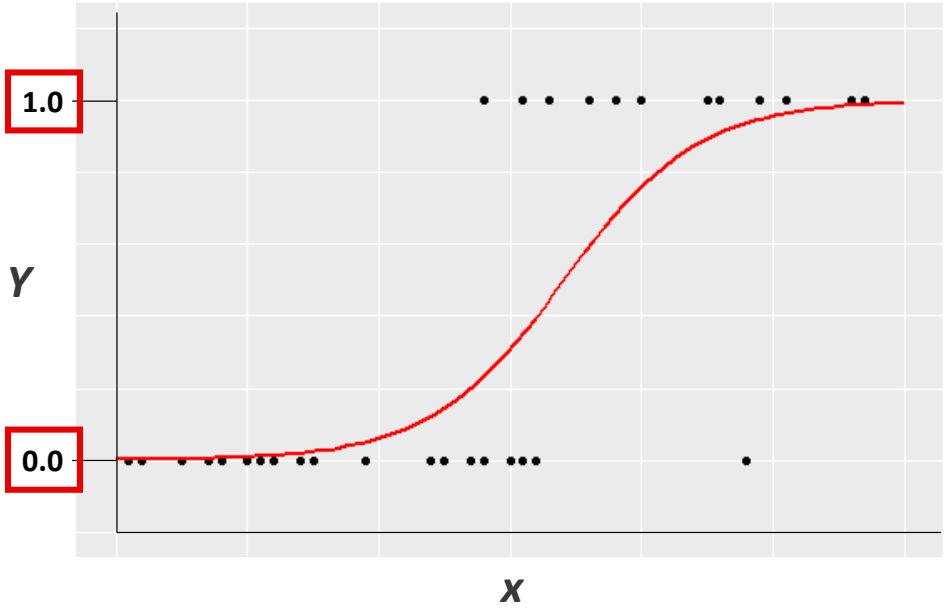
- Modélise la distribution d'une espèce dans l'espace (géographique ou écologique)
- Repose sur des données d'abondance ou **présence/absence**



Données binaires : modèle canonique: GLM

- Approche de référence: GLM avec distribution de Bernoulli
- Logit: Fonction de lien canonique → issue de l'extension du modèle linéaire (LM)
- Inverse logit → logistique classique
- Asymptote basse = 0
- Asymptote haute = 1

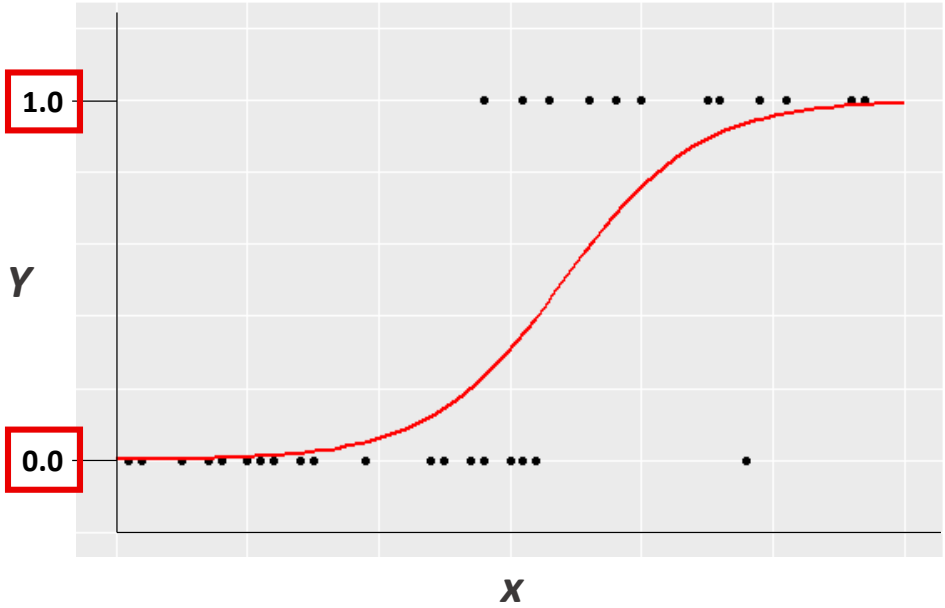
⚠ → **Hypothèse auxiliaire** forte



Données binaires : modèle canonique: GLM

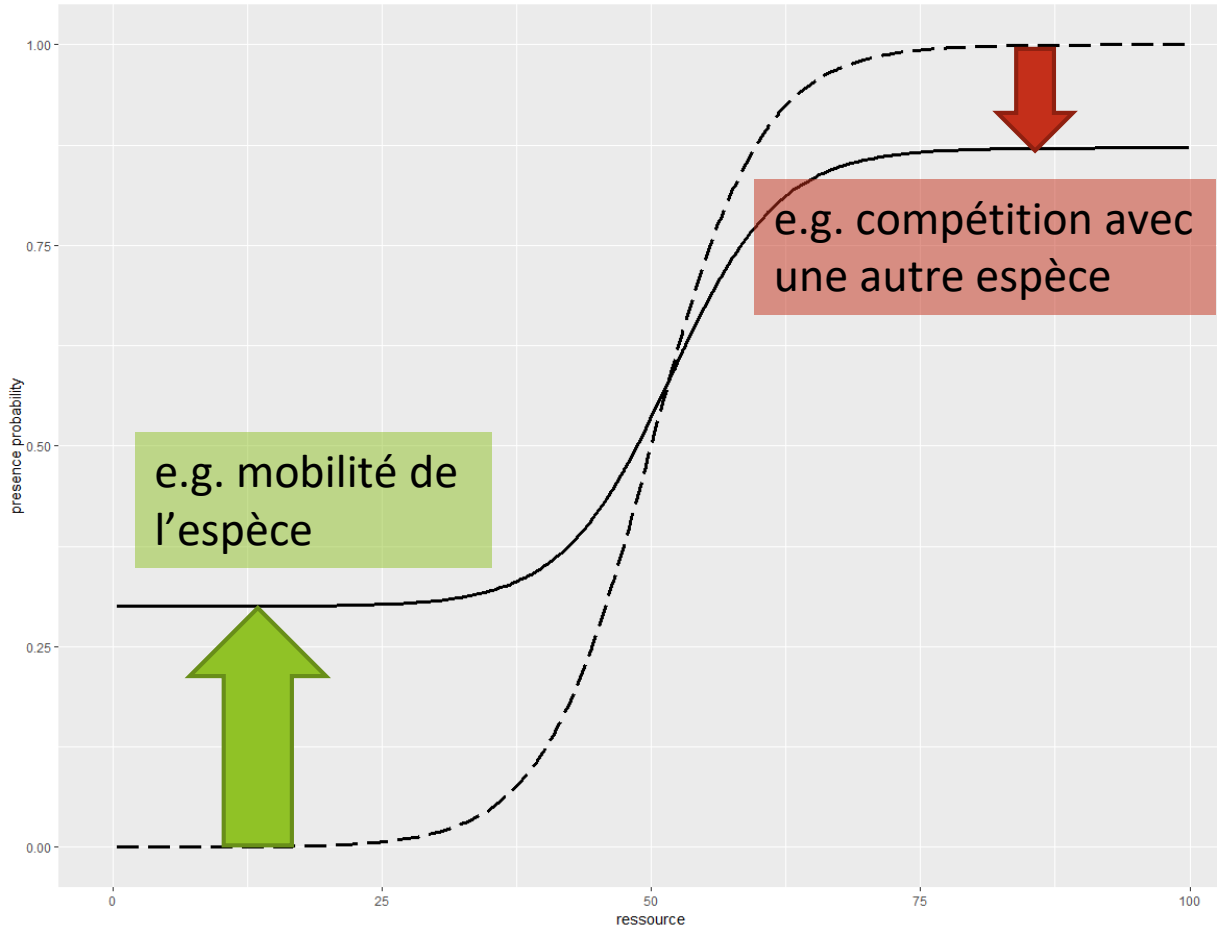
- Approche de référence: GLM avec distribution de Bernoulli
- Logit: Fonction de lien canonique → issue de l'extension du modèle linéaire (LM)
- Inverse logit → logistique classique
- Asymptote basse = 0
- Asymptote haute = 1

⚠ → **Hypothèse auxiliaire** forte



↳ Test par simulations des impacts de cette hypothèse auxiliaire

SDM: exemple de mécanismes

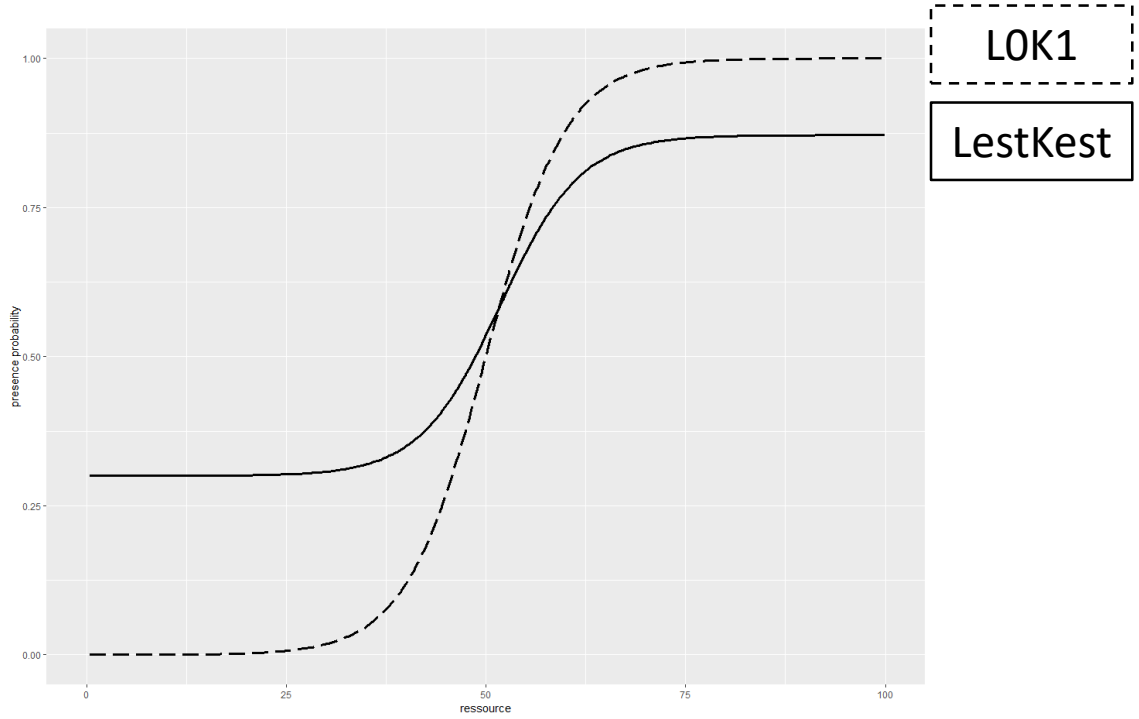


Logistique classique

Logistique avec asymptotes estimées

Modèles étudiés

- L0K1 : Logistique classique (asymptotes = 0 et 1)
- LestKest : Logistique ELUA (“estimated lower and upper asymptotes”)
- GAM : semi-paramétrique



TMB R-package

Données simulées

3 Scenarios univariés :

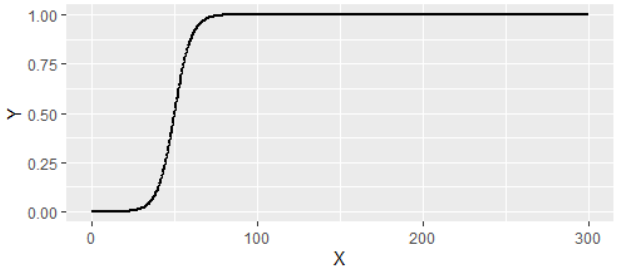
- 2 Scénarios : 10,000 jeux de données
- 1 Scénario (VarRand) : 1,000 jeux de données

Paramètres fixes :
- **ip** (point d'inflexion) = 50
- **sl** (pente) = 0.2

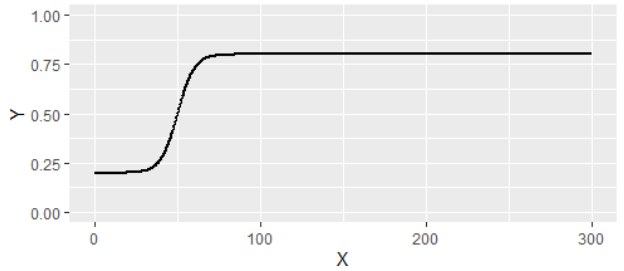
Paramètres variables entre Scénarios :
- FixDet : asymptotes fixées à 0 et 1
- VarDet : asymptotes variables
- VarAléa : asymptotes variables et données aléatoirement distribuées sur le gradient

Paramètres variables entre datasets au sein d'un Scénario :
- **Nobs** variable (entre ≈403 et ≈2981)
- **L et K** aléatoires dans les gammes définies

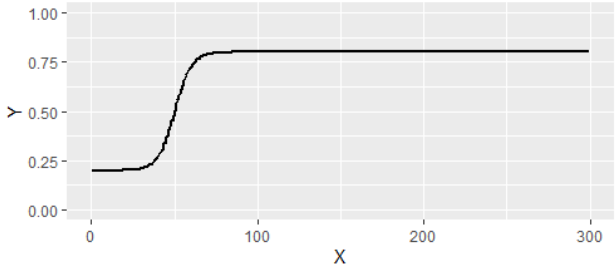
FixDet



VarDet



VarRand



NSL (pente normalisée) = sl * (K – L)

Analyse des simulations

- **AICc (critère d'information d'Akaike corrigé) :**

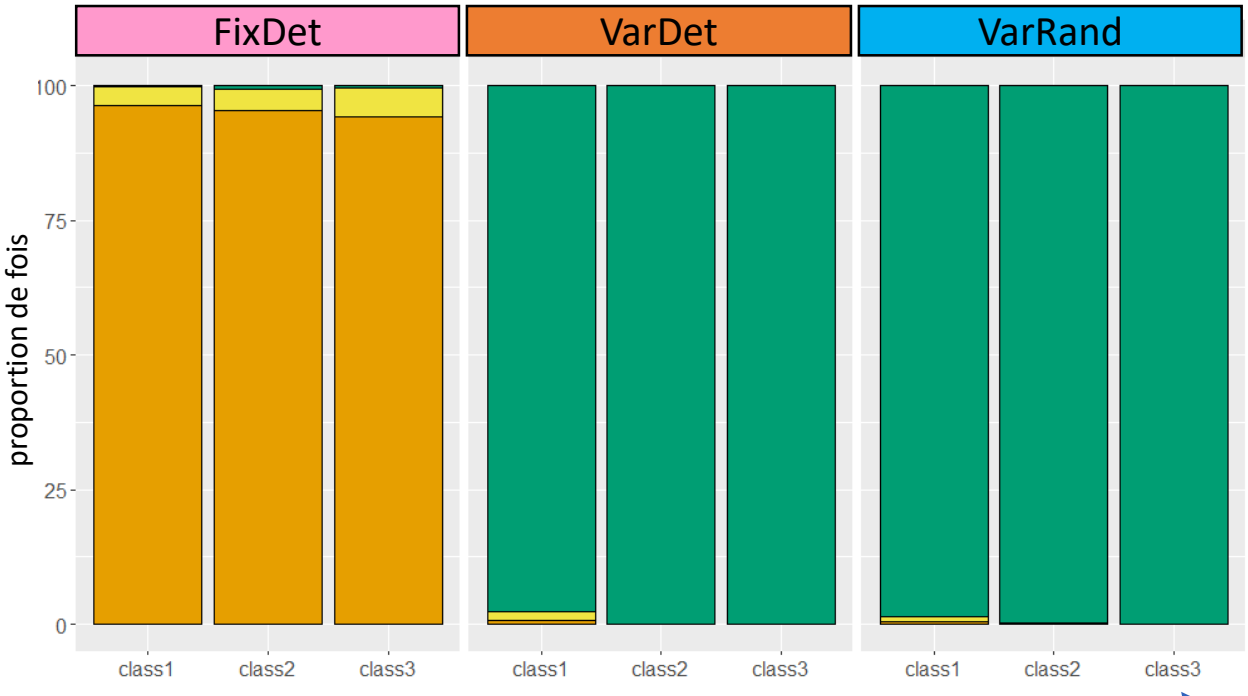
  plus faible AICc → meilleure capacité prédictive du modèle
  différence de moins de 2,0 points → les modèles sont équivalents

Paramètre d'intérêt majeur : **pente** (donne la magnitude de la relation)

- Estimation du paramètre :
  - comparaison de la précision de l'estimation entre les modèles (précision)
  - comparaison avec le paramètre réel (biais)

Comparaison AICc : LestKest / L0K1

Modèles univariés



Classe 1 : nobs ∈ [147;1092]
Classe 2 : nobs ∈]1092;2037]
Classe 3 : nobs ∈]2037;2982]

LestKest mieux que L0K1

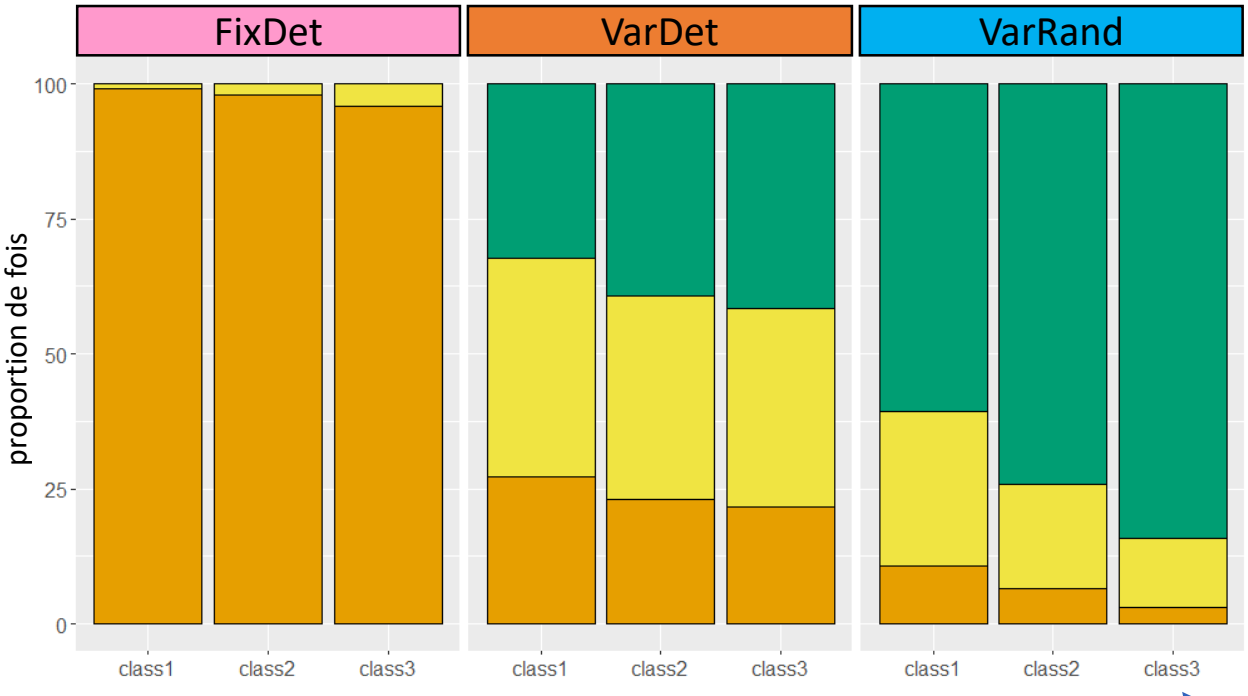
La proportion de cas où LestKest est mieux que L0K1 augmente avec le nombre d'observations (nobs)

Sauf lorsque :

- les asymptotes sont égales à 0 et 1 dans les données (Scénario FixDet)

Comparaison AICc : LestKest / GAM

Modèles univariés



Classe 1 : nobs ∈ [147;1092]
Classe 2 : nobs ∈]1092;2037]
Classe 3 : nobs ∈]2037;2982]

GAM mieux ou equivalent à LestKest

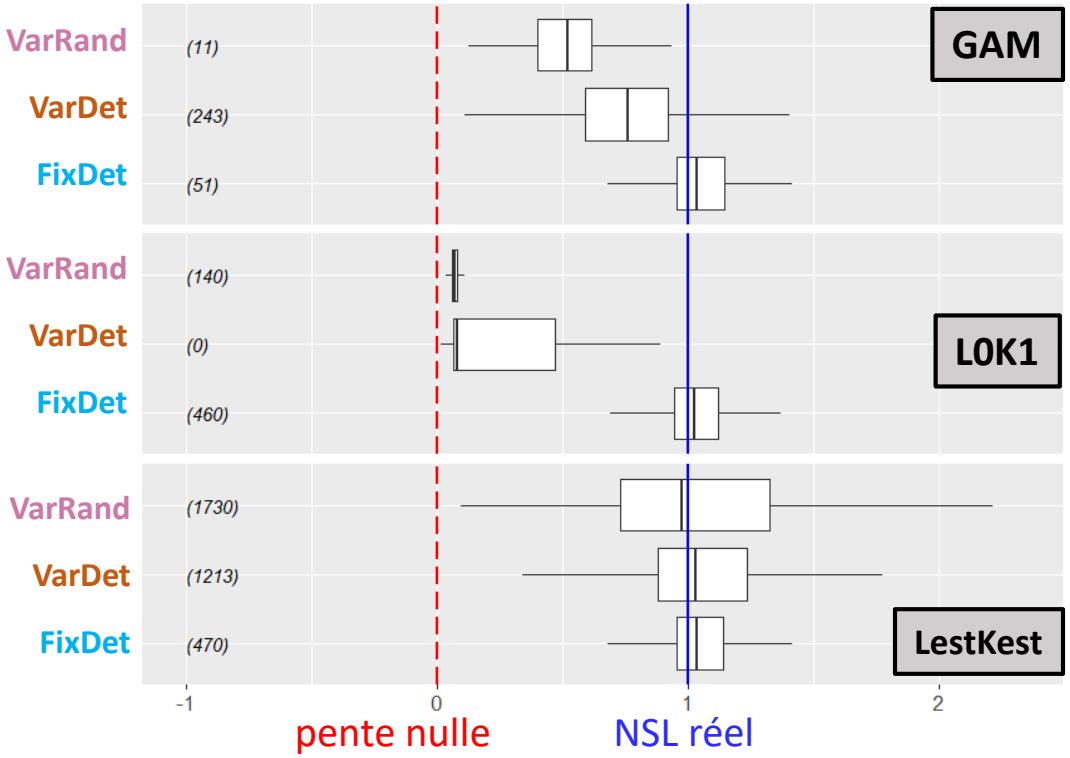
La proportion de cas où LestKest est mieux que GAM augmente avec le nombre d'observations (nobs)

Sauf lorsque :

- les données sont aléatoirement réparties sur le gradient (Scénario VarRand)

Estimation: Paramètre NSL (réel / estimé)

Modèles univariés



Estimation de la pente : LestKest **moins biaisé** que **L0K1** et que **GAM**

→ PENTE SOUS-ESTIMEE

Surtout lorsque :

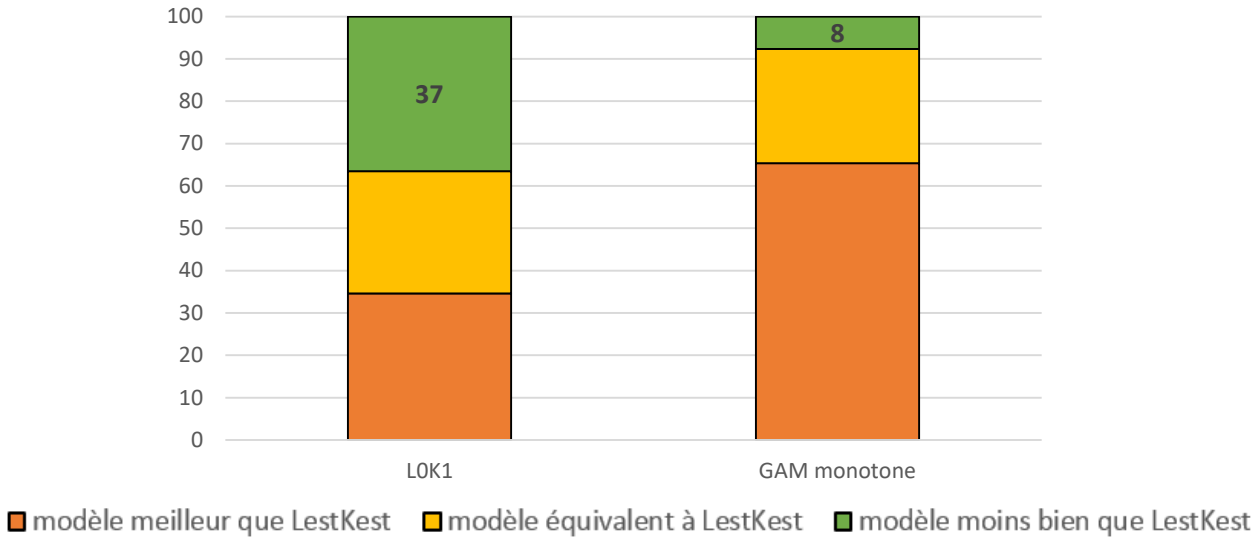
- les données sont aléatoirement réparties sur le gradient (VarRand)

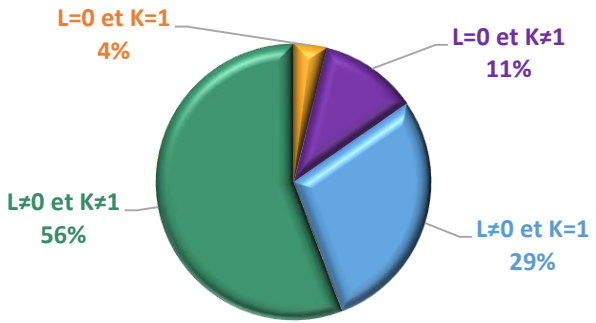
Sauf lorsque :

- les asymptotes sont égales à 0 et 1 dans les données (FixDet)

Données réelles multiples

- 52 jeux de données de ≈ 15 domaines scientifiques





K ≠ 1 si K ≤ 0,99
L ≠ 0 si L ≥ 0,01

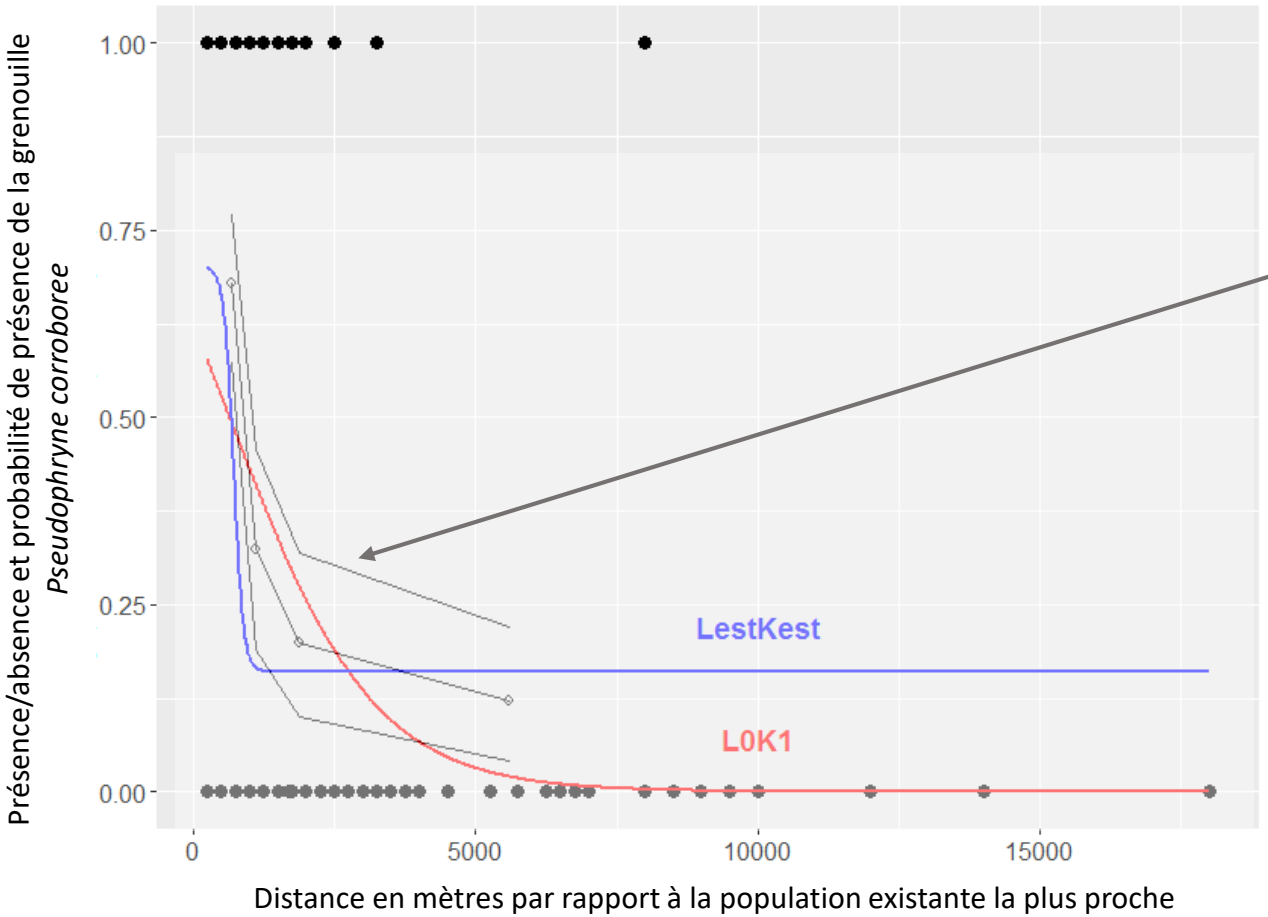
Données réelles SDM (pres/abs)

- Exemple sur un jeu de données : Présence/absence de la grenouille *Pseudophryne corroboree*

		L0K1	**GAM (monotone)**	**LestKest**
Univarié	Variable explicative	Distance en mètres par rapport à la population existante la plus proche		
	AICc	244.80	235.05	235.15
	L estimé			L = 0.16 ± 0.036
	K estimé			K = 0.70 ± 0.085

Données réelles SDM (pres/abs)

- Exemple sur un jeu de données : Présence/absence de l'amphibien *Pseudophryne corroboree* (212 observations)



Centiles des valeurs de présence / absence de sur la base d'une approximation normale (médiane et intervalle de confiance à 95%, cf. Harrell, 2001)

Si au moins une des deux asymptotes est atteinte **ET** différente de **0** ou **1**, alors :

- GLM canonique (L0K1) → Mauvaises capacité predictive et estimation de la pente
- GAM canonique → Mauvaises capacité prédictive et estimation des paramètres sur un gradient aléatoire (et en bivarié)
- Modèle ELUA (LestKest) → Corrige ces problèmes
- Modèle LestKest à stabiliser (stabilité numérique notamment)

Proposition que la fonction ELUA soit intégrée à la trousse à outils de l'analyse des données binaires

Merci pour votre attention

Des questions ?

Godeau, U. & Gosselin, F. (Soumis). GLM for Generalized Linear Misleading? The need for a logistic function with estimated asymptotes to complement canonical logit link functions for binomial GLMs

- Godeau, U., Bouget, C., Piffady, J., Pozzi, T., Gosselin, F., In Press. Lack of definition of mathematical terms in ecology: The case of the sigmoid class of functions in macro-ecology. *Ecology & Evolution*.

- Godeau, U., Bouget, C., Piffady, J., Pozzi, T., Gosselin, F., 2020. The importance of being radom! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models reveals the relationship between deadwood and the species richness of saproxylic beetles. *Forest Ecology and Management* 465, 118064.

- Harrell, F. E. Jr., 2001. *Regression Modeling Strategies*, ed. Springer-Verlag New York, New York.
- Huisman, J., H. Olff, and L. F. M. Fresco. 1993b. A Hierarchical Set of Models for Species Response Analysis. *Journal of Vegetation Science,* 4(1):37–46.
- Jansen, F. and Oksanen, J. (2013) 'How to model species responses along ecological gradients - Huisman-Olff-Fresco models revisited', *Journal of Vegetation Science*. Edited by J. Podani, 24(6):1108–1117.
- Mackay, A. W. *et al.* (2006) 'Assessing the vulnerability of endemic diatom species in Lake Baikal to predicted future climate change: a multivariate approach', *Global Change Biology*, 12(12), pp. 2297–2315. Suchrow, S. and Jensen, K. (2010) 'Plant Species Responses to an Elevational Gradient in German North Sea Salt Marshes', *Wetlands*, 30(4):735–746.
- Oksanen, J. and Minchin, P. R. (2002) 'Continuum theory revisited: What shape are species responses along ecological gradients?', *Ecological Modelling*, 157(2):119–129.
- Michaelis, J., Pannek, A. and Diekmann, M. (2016) 'Soil pH limits of forest vascular plants determine range size and threat level', *Journal of Vegetation Science*. Edited by S. Roxburgh, 27(3):535–544
- Richard, E. (2004) *Réponse des communautés de coléoptères carabiques à la conservation en futaie régulière de chêne : aspects écologique et méthodologiques*.
- Savić, A. et al. (2020) 'Assessing environmental response of gastropod species in karst springs: what species response curves say us about niche characteristic and extinction risk?', Biodiversity and Conservation, 29(3):695–708.

8 Scenarios univariés :

- Scénarios 1 à 7 : 10,000 jeux de données
- Scénario 8 : 1,000 jeux de données

Paramètres fixes :

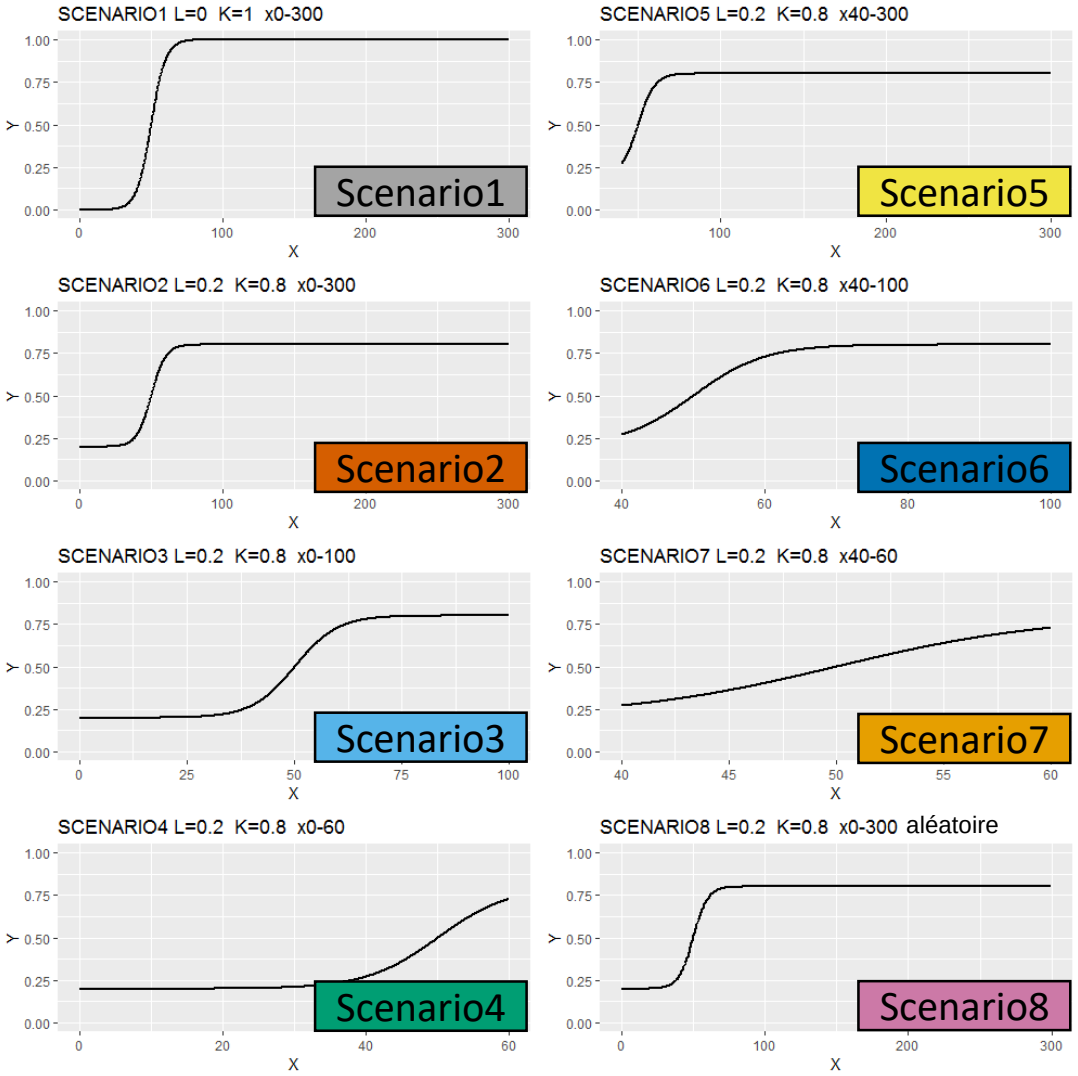
- **ip** (point d'inflexion) = 50
- **sl** (pente) = 0.2

Paramètres variables entre Scénarios :

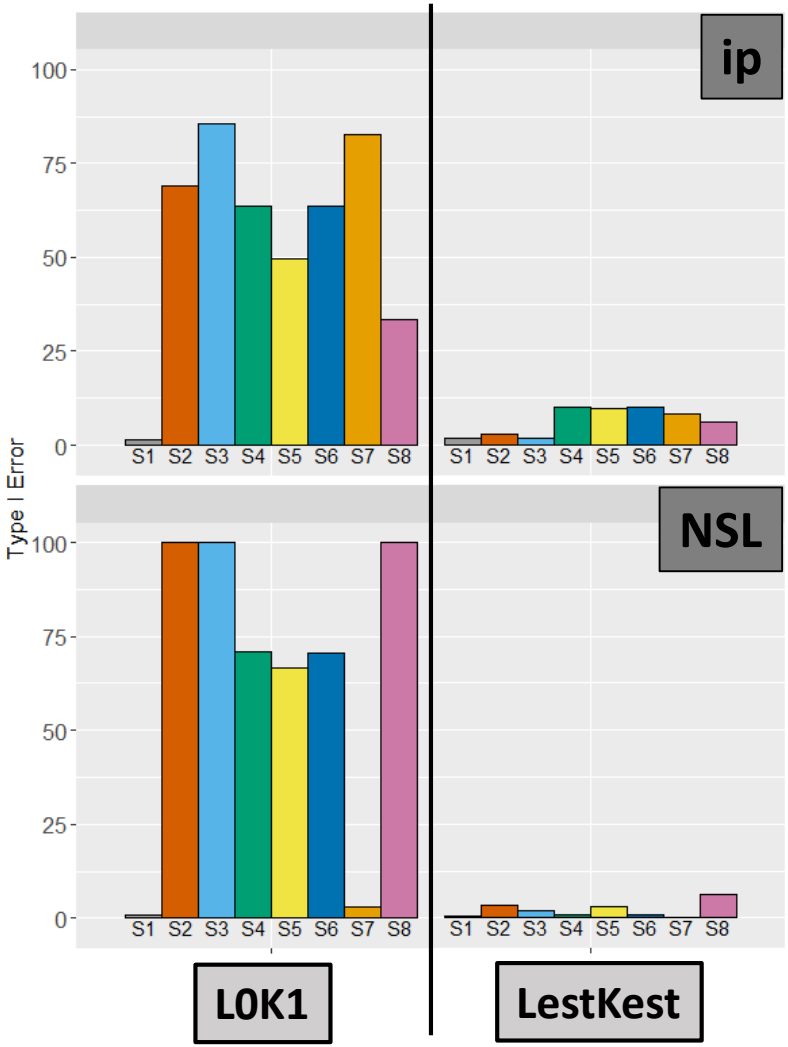
- gamme de variabilité de l'asymptote haute **K**
- gamme de variabilité de l'asymptote basse **L**
- gradient de **X**

Paramètres variables entre datasets au sein d'un Scénario :

- **Nobs** variable (entre ≈403 et ≈2981)
- **L et K** aléatoires dans les gammes définies pour le Scénario



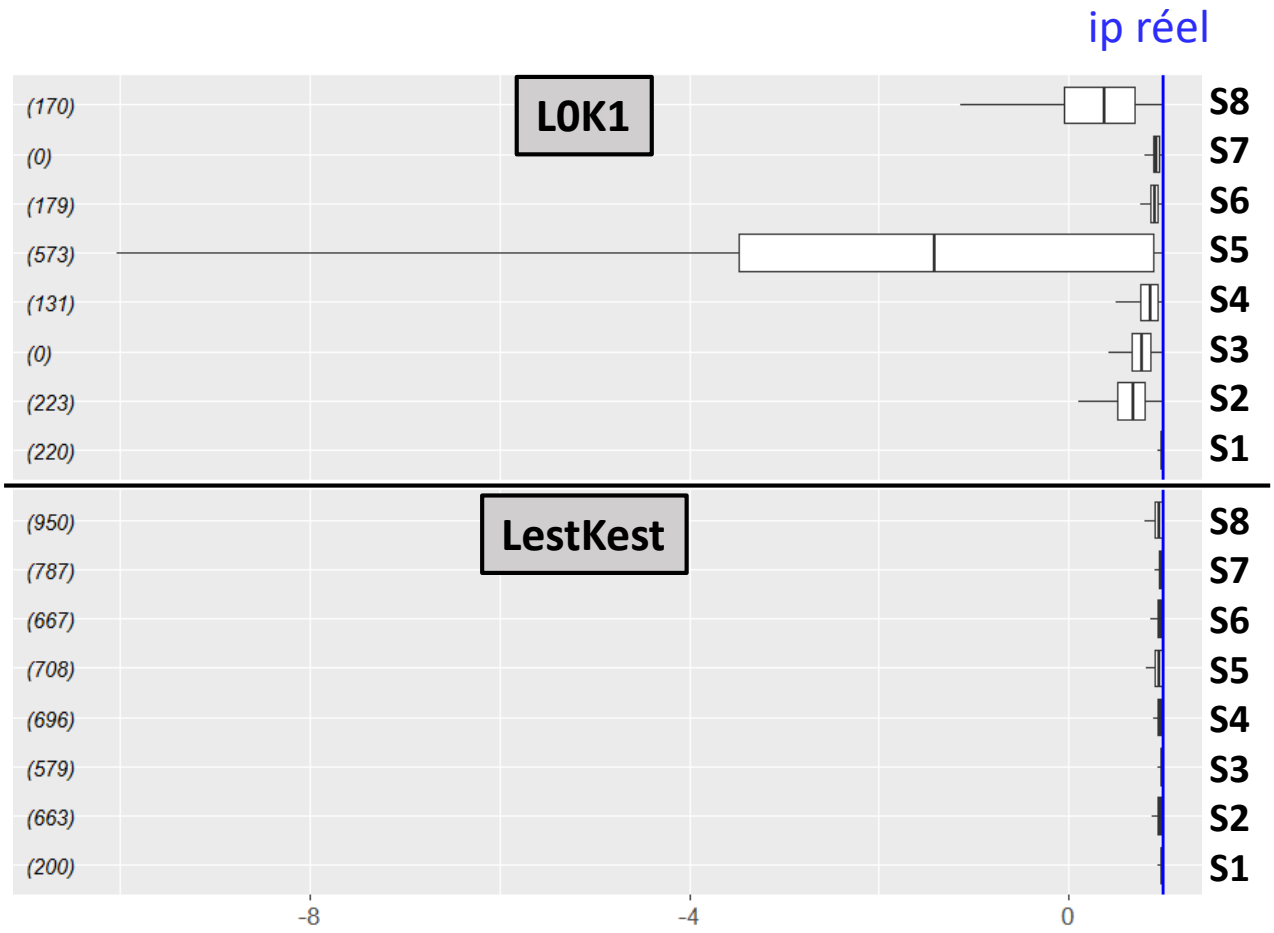
NSL (pente normalisée) = sl * (K – L)



Erreur de type I de LestKest beaucoup plus **faible** que LOK1 pour ip et NSL

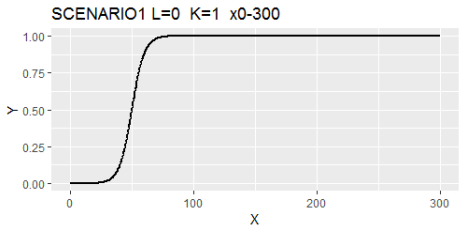
Erreur de type I LestKest pour ip reste trop élevée

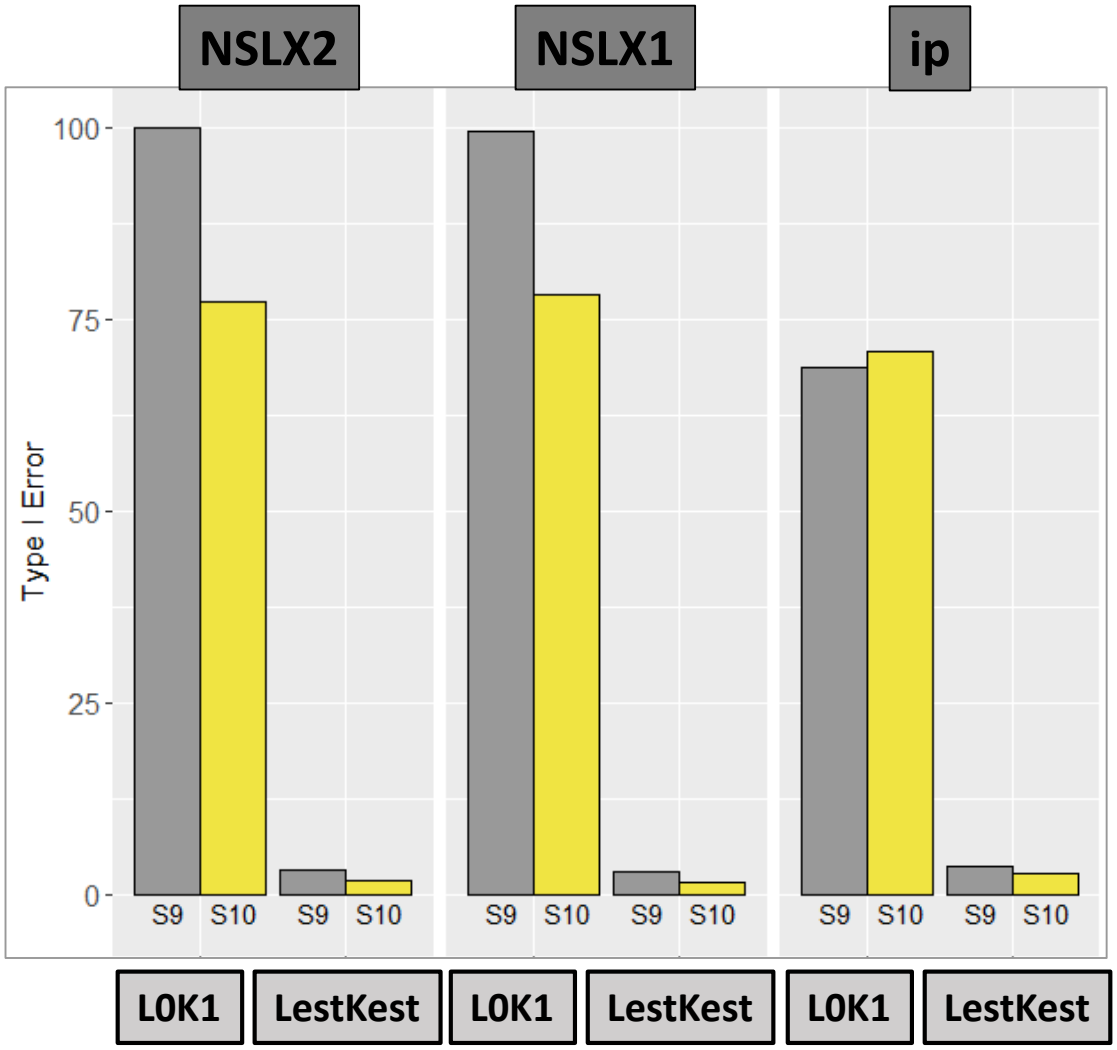
ip estimé / ip réel



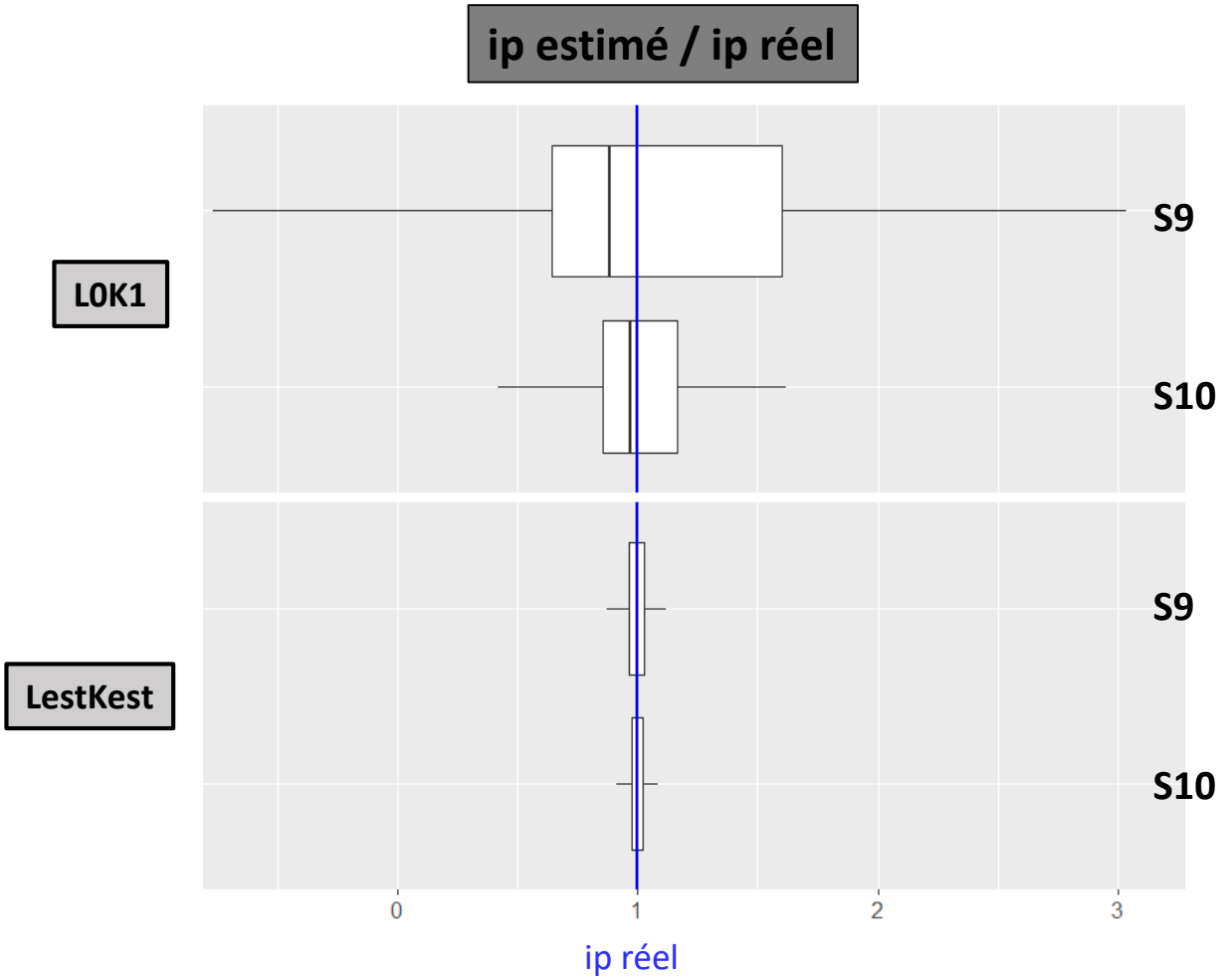
Estimation de ip :
LestKest **moins biaisé** et plus **précis** que L0K1

Sauf S1, LestKest et L0L1 équivalents





Beaucoup **moins** d'erreur de type I que LOK1 pour ip, NSLX1 et NSLX2 pour LestKest



Estimation de ip :
LestKest moins **biaisé**
et plus **précis** que LOK1