

# Une alternative à la régression pas-à-pas

Application à la prédiction de la créativité managériale

Nils Myszkowski  
Communication à MODEVAIIA 2013



# Modèles linéaires généralisés

- GLM
  - Expliquer une variable dépendante...
    - Numérique
    - Catégorielle
  - ...par un ensemble de variables prédictrices
    - Numériques
    - Catégorielles
  - Modèles largement utilisés dans de nombreux domaines scientifiques

# Sélection de modèle

- Perspective de parcimonie
  - Dans certains contextes, il peut être utile de sélectionner un sous ensembles de prédicteurs
    - ex: recruter le meilleur candidat en utilisant le moins possible d'outils de recrutement
- Modèle optimal/pertinent/économique de prédiction d'une VD
  - Précision
  - Parcimonie

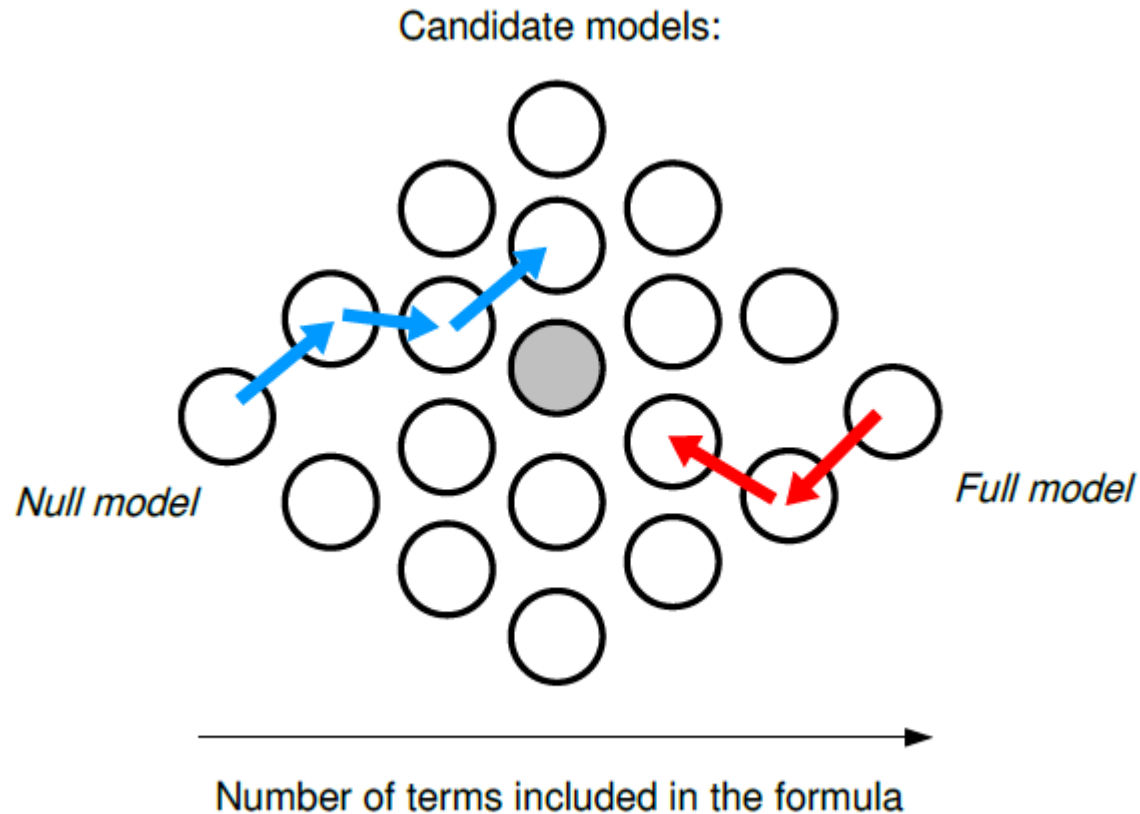
# Régression pas-à-pas

- Méthode usuelle de sélection de modèles
  - Régression pas-à-pas
    - Forward : ajouter un à un des prédicteurs au modèle nul
    - Backward : retirer un à un des prédicteurs au modèle complet

# Régression pas-à-pas

- Problèmes
  - Point de départ (modèle nul ou complet) ?  
Règle d'arrêt?
  - Convergence backward/forward?  
Convergence avec le modèle optimal?
  - Test d'un sous-ensemble des modèles possibles (or souvent la conclusion porte sur l'ensemble des modèles possibles)

# Différentes démarches



Calcagno & Mazancourt, 2010

# Une proposition de solution

- Une proposition de solution
  - Ajuster tous les modèles possibles
  - Classer les modèles en fonction de la valeur d'un critère d'information (AIC, BIC, etc.)
  - → autre problème: explosion combinatoire du nombre de modèles possibles lorsque le nombre de prédicteurs possibles augmente
    - $2^n$  (sans interaction)

# Une proposition de solution

- Automatisation de la procédure
  - Package r glmulti (Calcagno & Mazancourt, 2010)
  - Package r MuMIn (Barton, 2009)
  - Package r bestglm (McLeod and Xu 2009).
  - Classement de tous les modèles possibles



# Une proposition de solution

- Automatisation de la procédure
  - 1) Génération de tous les modèles possibles sur la base du modèle complet
    - Format « glm » classique sous r
    - Ex :  $y \sim f1 + c1$
  - 2) Ajustement de tous les modèles, calcul de la valeur du critère d'information
  - 3) Classement de tous les modèles possibles en fonction de leur critère d'information

# Application

- Créativité managériale
  - N=110
  - VD
    - Pensée divergente -> fluidité
    - Pensée convergente -> C.A.T.
  - VI
    - BFI
    - SPM

# Application

- Créativité managériale
  - Modèle complet :
  - $VI \sim \text{Intelligence} + O + C + E + A + N$
  - Sans interaction :  $2^6 = 64$  modèles possibles
  - En stepwise: 21 modèles testés seulement au maximum

# Application – Fluidité

		Full model	BIC-based optimal model
	Intelligence	.07	
	Conscientiousness	.03	
Beta	Agreeableness	-.03	
estimate	Neuroticism	.22*	.24**
	Openness	.30**	.33***
	Extraversion	.13	
Model fit	BIC	548.85	532.70
index	Multiple R <sup>2</sup>	0.15**	.14***

\*\*\* significant at  $p < .001$  ; \*\* significant at  $p < .01$  ; \* significant at  $p < .05$

# Application – Pensée convergente

		Full model	BIC-based Optimal model
Beta estimate	Intelligence	.42***	.39***
	Conscientiousness	-.17	
	Agreeableness	.22*	
	Neuroticism	.08	
	Openness	.08	
	Extraversion	.15	
Model fit index	BIC	390.34	377.74
	Multiple R <sup>2</sup>	0.23***	.15***

\*\*\* significant at  $p < .001$  ; \*\* significant at  $p < .01$  ; \* significant at  $p < .05$

# Problèmes

- Questions
  - Démarche habituelle de régression pas à pas justifiée, ou historique?
    - *“This is an unusual choice. Was it made for practical reasons? Please confirm that the results are robust to stepwise selection.”*
    - Ajout ou retranchement séquentiel plus justifié que d'autres algorithmes qui ne testent qu'un sous ensemble de modèles? (Calcagno & Mazancourt, 2010)
    - Mettre en compétition les différents algorithmes en fonction de leur convergence avec le modèle optimal

# Problèmes

- Questions
  - Démarche confirmatoire?
    - Vérifier que le modèle hypothétique est non seulement mieux ajusté aux données que le modèle nul mais aussi mieux ajusté à l'ensemble des modèles possibles
  - Explosion computationnelle: le moindre mal?
    - Exclure les termes d'interactions VS. se rabattre sur la régression stepwise avec interaction ?
    - Algorithmes génétiques?

# Problèmes

- Questions
  - Retenir 1 modèle? Un sous-ensemble de modèles?
  - Démarche « athéorique » de prédiction → Oublier la règle de marginalité?
    - N'inclure les termes d'interaction qu'avec les effets simples?