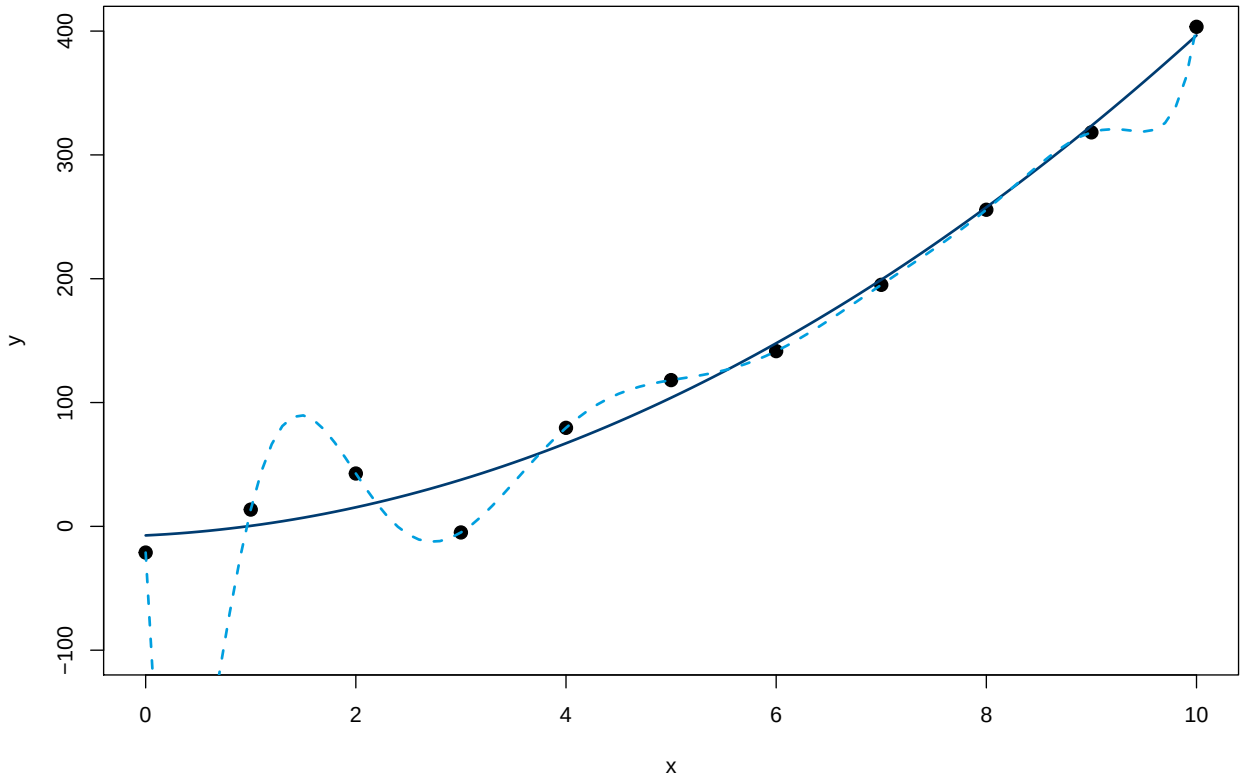
MATH 60602 - Récapitulatif (Sélection de modèles)

Léo Belzile

23 février 2021

Surajustement



Séparation des données

Ne pas utiliser les données employés pour ajuster un modèle pour **prédire la performance**

- échantillons d'apprentissage/validation/test (fixes)
- validation croisée (avec $K = 5, 10$ groupes), mais **résultat aléatoire**

Sélection de modèles

- Flexibilité est la clef
- Trop de modèles à explorer!
- Algorithmes gloutons pour créer un ensemble de “bons candidats”
- Critère pour la qualité de l’ajustement: erreur moyenne quadratique (EMQ), critères d’information (BIC et AIC)

Méthodes de sélection classique

- Recherche exhaustive: regarder le modèle avec une variable de plus à la fois (parmi tous les modèles)
- Sélection ascendante: à chaque étape, on rajoute la meilleur variable selon un critère en partant du modèle avec l'ordonnée à l'origine
- Sélection descendante: en partant du modèle avec toutes les variables, enlever une variable à la fois
- Sélection séquentielle: comme la sélection ascendante, mais possibilité d'enlever des variables à chaque étape.

Options **SAS**

- `stop`: critère d'arrêt pour la recherche,
- `select`: pour sélectionner quel variable ajouter à une étape, de sortie et de sélection
- `choose`: critère pour choisir le modèle final

Lasso

Pénaliser l'erreur moyenne quadratique pour faire simultanément de la sélection et du rétrécissement.

- données standardisées (centrées et réduites)
- pénalité $\lambda|\beta_j|$ sur chaque coefficient
- selon λ, coefficients exactement zéro

Moyenne de modèles

- Créer de nouveaux échantillons (tirage avec remise) de même taille que l'original (**autoamorçage**)
- Répéter sur chacun une procédure de sélection
- Faire la moyenne des coefficients obtenus