# TP6
# Sélection de variables

## Objectifs du TP

— Manipuler des variables explicatives qualitatives
— savoir sélectionner un ensemble de variables pertinentes par recherche exhaustive,
— interpréter les résultats obtenus via des critères différents ($R^2_a$, $C_p$, BIC, tests),
— mettre en œuvre une sélection stepwise,
— évaluer une erreur de prévision sur des données test.

## Étude d'une cohorte de patients atteints de diabète

Le jeu de données `diabetes` met en relation une mesure quantitative du degré d'avancement de la maladie et 10 variables cliniques dans une cohorte de 442 patients. Chacun des 442 patients est décrit par les variables suivantes :

— `prog` : l'indice de progression de la maladie
— `age` : l'âge du patient
— `sex` : le sexe du patient, codé numériquement (1 ou 2)
— `bmi` : l'indice de masse corporelle
— `map` : la pression artérielle moyenne
— `ser1`–`ser6` : 6 mesures sérologiques

On cherche à construire un modèle linéaire avec de bonnes propriétés prédictives pour la variable `prog`. Simultanément, on souhaite identifier parmi les covariables mesurées lesquelles sont directement liées à la réponse. Dans ce double but, on va faire de la sélection de variables.

On a séparé au préalable le jeu de donnée `diabetes` en deux sous-ensembles : `diabetes1` et `diabetes2`. Dans un premier temps, *on utilisera uniquement* `diabetes1` ; le jeu de données `diabetes2` ne sera utilisé qu'à la dernière question.

Un fichier d'aide `TP6-2021_aide.R` est disponible sur ecampus.

## Prise en compte des variables qualitatives

*Note : on pourra consulter utilement le livre « Régression avec R » de Pierre-André Cornillon et Eric Matzner-Lober (Springer, 2011), notamment le chapitre 5 au sujet des variables qualitatives.*

1. Importer les données du fichier `diabetes1.txt` et faire une analyse descriptive rapide (matrice de corrélations ; graphes paire à paire) en se limitant aux variables quantitatives.
2. Quel est le type de la variable `sex` ?
3. Analysez le lien entre les différentes variables explicatives et la variable `sex`. On pourra notamment utiliser la commande `boxplot`.

On propose le modèle suivant afin d'expliquer la variable `prog` en fonction des autres variables.

$$\texttt{prog}_i = \varepsilon_i + \mu_0 + \mu_{\texttt{age}} \cdot \texttt{age}_i + \mu_{\texttt{bmi}} \cdot \texttt{bmi}_i + \mu_{\texttt{map}} \cdot \texttt{map}_i + \sum_{k=1}^{6} \mu_{\texttt{ser(k)}} \cdot \texttt{ser(k)}_i + \begin{cases} \mu_F & \text{si } \texttt{sex}_i = F \\ \mu_M & \text{si } \texttt{sex}_i = M \end{cases}, \qquad i = 1, \ldots, n, \quad (1)$$

où les $\varepsilon_i$ sont des variables de bruit (sur lesquelles on fait les hypothèses habituelles)

4. Montrer que le modèle (1) peut s'écrire comme un modèle linéaire de la forme

$$Y = X\beta + \varepsilon \ .$$

   On précisera en particulier ce que valent $X$ et $\beta$ en fonction des quantités apparaissant dans (1). Quel est le rang de la matrice $X$ ?

5. Ajuster le modèle (1) incluant toutes les covariables (on l'appelera « complet »).
   Effectuer (rapidement) les diagnostics d'usage.
   Comparer au modèle n'incluant que l'intercept (on l'appelera `nul`).

   *Attention ! Le modèle étant surparamétré, il est nécessaire d'imposer des contraintes sur les $\mu_k$. Par défaut, `R` utilise le premier niveau du facteur comme niveau de référence, et lui associe un coefficient $\mu_k = 0$. L'ordre des niveaux a donc une importance ! On observera comment `R` choisit l'ordre des niveaux par défaut. La fonction `factor` permet de réordonner les niveaux comme on le souhaite.*

## Sélection de variables

6. **Recherche exhaustive**.
   Si besoin, installer le package `leaps` (par exemple, dans R studio, avec la commande « Install packages » de l'onglet « Tools »).
   Pour sélectionner les meilleurs modèles parmi tous les modèles possibles (contenant l'intercept), on pourra utiliser la fonction `regsubset` du package `leaps` (voir le code R fourni dans le fichier `TP7-2018_aide.R`).
   (a) Pour $D = 1, \ldots, 10$, déterminer le meilleur modèle à $D$ paramètres (en plus de l'intercept). La réponse dépend-elle du critère utilisé (parmi SCR, $R^2_a$, $C_p$ et BIC) ?
   (b) Pour les 500 meilleurs modèles, tracer en fonction du nombre de covariables :
   — la somme des carrés résiduelle,
   — le coefficient de détermination ajusté $R^2_a$,
   — le critère $C_p$,
   — le critère BIC.
   (c) Déterminer le meilleur modèle selon chacun de ces critères (sans contrainte sur le nombre de paramètres). Comparer, puis interpréter les résultats.
7. **Recherche stepwise** manuelle.
   *À la main*, mettre en œuvre la sélection backward à base de tests de comparaison de sous-modèles : partir du modèle complet et supprimer les covariables une à une (à partir des p-valeurs des tests de comparaison de sous-modèles : on supprime celle dont le test individuel fournit la plus grande p-valeur), tant qu'au moins un résultat de test justifie la suppression d'une covariable.
8. **Recherche stepwise** automatique.
   Mettre en œuvre la sélection forward. Comparer au résultat obtenu à la question 4.
   Faire ensuite de même avec la sélection backward. Comparer les résultats.
9. **Évaluation de l'erreur de prévision avec des données test.**
   Utiliser les données test (fichier `diabetes2.txt`) pour estimer l'erreur de prévision avec le modèle choisi par le critère $C_p$. (On pourra utiliser la fonction `predict`).
   Faire de même pour tous les modèles obtenus aux questions précédentes (c'est-à-dire, le modèle complet, le modèle nul, les modèles choisis par $R^2_a$, $C_p$ et BIC, et les modèles obtenus par recherche stepwise).
   Quel modèle choisir au final ?
   Analyser les résidus de ce modèle.