# SY19 – A18
# TP 2: Régression et classification linéaires

## 1 Régression linéaire

### 1.1 Pratique de la régression linéaire

Appliquez la régression linéaire sur les données `prostate`, en prenant la variable `lpsa` comme variable à expliquer.

1. Quels coefficients sont significativement non nuls ? La régression est-elle globalement significative ?
2. Calculer les intervalles de confiance à 95% sur les coefficients.
3. Tracer les valeurs prédites $\widehat{y}_i = \widehat{f}(x_i)$ en fonctions des $y_i$.
4. Tracer les résidus bruts et standardisés en fonction de $y_i$ et des variables quantitatives.
5. Vérifiez la normalité des résidus.
6. Etudiez la stabilité de la régression (certaines observations ont-elles une grande influence sur les résultats) ?
7. Reprendre l'analyse avec différents sous-ensembles de prédicteurs. Qu'observe-t-on ?
8. Essayez quelques transformations non linéaires de certains prédicteurs. Améliore-t-on les résultats ?
9. Estimer les coefficients uniquement sur les données d'apprentissage (train=TRUE), et prédire la valeur de `lpsa` sur les données de test. Représenter les valeurs prédites et les intervalles de prédiction à 95% en fonction des valeurs observées.

### 1.2 Intervalles de confiance et de prédiction

On considère un modèle de régression linéaire avec deux variables explicatives $X_1$ et $X_2$ et un terme d'erreur gaussien $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

En générant un grand nombre d'ensembles d'apprentissage, vérifiez par simulation que les intervalles de confiance à 95% sur les paramètres $\beta_j$ contiennent bien la vraie valeur des paramètres dans environ 95% des cas.

Estimez la probabilité que les trois intervalles de confiance sur les paramètres $\beta_j$ contiennent *simultanément* les vraies valeurs des paramètres.

Calculer les intervalles de confiance et de prévision à 95% sur $Y_0$ pour une nouvelle valeur $x_0 = (x_{10}, x_{20})$, et vérifiez expérimentalement les propriétés théoriques de ces intervalles.

# 2 Classification linéaire

## 2.1 Analyse des données `spam`

Chargez les données `spam`. Séparez-les en un ensemble d'apprentissage (2/3 des exemples) et un ensemble de test. Appliquez sur ces données l'analyse discriminate linéaire et la régression logistique. Calculer la matrice de confusion et le taux d'erreur pour chacune des méthodes. Tracer sur le même graphique les courbes COR. Quels prédicteurs ont un coefficient significativement non nuls (régression logistique) ?

## 2.2 Comparaison ADL-ADQ sur des données simulées

Dans cet exercice, on génère des données gaussiennes avec $K = 2$, $p = 3$, et les paramètres suivants :

$$\pi_1 = \pi_2 = 0.5, \quad \mu_1 = (0, 0, 0)^T, \quad \mu_2 = (1, 1, 1)^T$$

$$\Sigma_1 = I_3, \quad \Sigma_2 = 0.8 I_3.$$

Générer des échantillons d'apprentissage de taille $n \in \{30, 100, 1000, 10000\}$ et un échantillon de test de taille 10000. Etudiez le taux d'erreur de test pour l'ADL et l'ADQ, en fonction de $n$. Commentez les résultats.