

# SY19 – A16

## TP 2: regression linéaire

### 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  $\hat{f}(x_0)$  en fonctions des  $y_i$ .
4. Tracer les résidus bruts, standardisés, studentisés, 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 ?

### 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.

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.