Modélisation statistique - TD4
Régression linéaire simple avec R

1 Exercice : Ventes selon les budgets publicitaires

Dans le fichier `Advertising.csv` vous trouverez les ventes d'un produit réalisés sur 200 marchés différents, ainsi que les budgets de publicités alloués aux média : télévision, radio et journaux.

1. Importer dans **R**, le fichier de données et créez un *data.frame* que vous nommerez `Ad`.
2. Dans un premier temps, vous observerez le lien entre les couples de variables `Sales` et `TV`, `Sales` et `Radio`, et enfin `Sales` et `Newspaper`.
   (a) Pour chaque couple de variable, réaliser un nuage de points et calculer la corrélation.
   (b) Y a t-il une relation linéaire entre les variables de chaque couple ?
   (c) Pour chaque couple, à partir du nuage de point uniquement, faites une estimation grossière de l'écart-type lorsque les dépenses publicitaires sont modérées et lorsqu'elles sont élevées. Que remarque vous ?
3. Estimer un modèle de régression linéaire simple entre les ventes et les dépenses publicitaires (un par couple).
4. Rappeler les hypothèses faites dans le cadre de la régression linéaire simple vous semblent-elles vérifiées ?
5. Est-ce le cas pour le modèle `Sales~Newspaper` ? Comment interprétez-vous le résultat de la régression ?
6. On s'intéresse maintenant au modèle `Sales~Radio`. Pour tenter de supprimer l'hétéroscédasticité des résultats nous allons utiliser une méthode dite de "Stabilisation de la variance". Elle consiste à transformer la variable Y par $ln(Y)$ ou $\sqrt{Y}$ et à modéliser ces variables transformées plutôt que Y.
   (a) Créez deux nouvelles colonnes au data.frame `Ad`. Une avec le logarithme de `Sales` et l'autre avec la racine carrée.
   (b) Estimer un modèle de régression linéaire simple `SqrtSales~Radio` et observez comment se comportent les résidus. Si vous estimez que les résidus sont encore hétéroscédastiques passez au log.
   (c) Interpréter les résultats du modèle choisi. Attention au fait que l'espérance du log n'est pas égal au log de l'espérance. Si $\ln(Y) \sim \mathcal{N}(\mu, \sigma^2)$ alors $E(Y) = e^{\mu+\sigma^2/2} \neq \exp(E[ln(Y)]) = e^{\mu}$.
7. On s'intéresse maintenant au modèle `Sales~TV`. Quelles transformations du modèles pouvez-vous proposez pour que les hypothèses de la régression semblent valides ?