6. RÉGRESSION LINÉAIRE

6.1. **Salaires.** Le jeu des données *Income2* (disponible ici `http://www-bcf.usc.edu/~gareth/ISL/Income2.csv`) contient un échantillon de triplets (niveau d'éducation, ancienneté au travail, salaire) pour 30 personnes. Les noms des champs sont *Education*, *Seniority* et *Income*. Le jeu peut être chargé en *R* avec la commande *Income=read.csv("Income2.csv")*.

Une régression linéaire avec le salaire comme réponse et le niveau d'éducation comme variable explicative donne le resultat suivant avec *R* (la ligne de commande correspondante est *summary(lm(Income Education,data=Income))*):

```
Call:
lm(formula = Income ~ Education, data = Income)
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.9166     9.7689  -4.291 0.000192 ***
Education     6.3872     0.5812  10.990 1.15e-11 ***

Residual standard error: 11.93 on 28 degrees of freedom
Multiple R-squared:  0.8118,    Adjusted R-squared:  0.8051
F-statistic: 120.8 on 1 and 28 DF,  p-value: 1.151e-11
```

(1) Estimer les intervalles de confiance à 95% pour les paramètres de la régression linéaire.
(2) Dire si l'education explique le salaire avec un risque de première espèce de 0.1%.

On passe à une régression linéaire multiple avec le salaire comme réponse et les autres données comme variables explicatives. Le coefficient de détermination R^2 passe de 0.8118 à 0.9341.

(3) Dire si la plus grande complexité du deuxième modèle est justifiée.

6.2. **Ventes.** Le jeu des données *Advertising* (disponible ici `http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv`) contient un échantillon de 200 quadruplets avec les dépenses pour la publicité sur télévisions, radios et journaux et les ventes correspondantes (*TV*, *radio*, *newspaper*, *sales*).

Le tableau suivant montre le coefficient de détermination ajusté pour des modèles de régression linéaire avec les ventes comme réponse et différentes variables explicatives.

Adjusted R^2	TV	radio	newspaper
0.6099	oui	no	no
0.3287	no	oui	no
0.04733	no	no	oui
0.8962	oui	oui	no
0.6422	oui	no	oui
0.3259	no	oui	oui
0.8956	oui	oui	oui

(1) Quels modèles sont considérés par la méthode de sélection *forward* qui se base sur le coefficient de détermination ajusté et quel modèle est retenu comme le meilleur ?
(2) Et par la méthode de sélection *backward* ?

6.3. **Ventes [R].** Étudier le jeu des données *Advertising* en essayant de répondre àux questions suivantes :

(1) Existe-t-il une relation entre le budget publicitaire et les ventes ?
(2) Quantifier la dépendance entre le budget publicitaire et les ventes.
(3) Quels médias contribuent aux ventes ?
(4) Avec quelle précision pouvons-nous estimer l'effet de chaque moyen de communication sur les ventes ?
(5) Avec quelle précision pouvons-nous prévoir les ventes futures ?
(6) La relation est-elle linéaire ?
(7) Existe-t-il une synergie entre les médias publicitaires ?