10. Quelles conclusions peut-on extraire de la commande *summary* ?
11. Retrouver la valeur Std. error pour l'intercept à partir d'une autre grandeur du *summary*.
12. Expliquer entièrement à quoi correspondent les commandes ci-après ? Justifier le calcul du fit.

```
> nd <- as.data.frame(matrix(c(4, "Y"),1,2))
> names(nd) = c('Lotcreme','Marque')
> pred = predict(mod2, newdata = nd, interval = "confidence")
> pred = predict(mod2, newdata = nd, interval = "prediction")
> predict(mod2, newdata = nd, interval = "confidence")
     fit      lwr      upr
1 9.2088 8.260148 10.15745
> predict(mod2, newdata = nd, interval = "prediction")
     fit      lwr      upr
1 9.2088 7.364952 11.05265
```

3 Valeur des logements autour de Boston

Le jeu de données étudié ici concerne la valeur des logements des villes aux alentours de Boston. On cherche à identifier les variables dont la valeur des logements dépend.

Les variables utilisées et leur signification sont les suivantes :
- CRIM taux de criminalité par habitant
- ZN proportion de terrains résidentiels
- INDUS proportion de terrains industriels
- CHAS 1 si ville en bordure de la rivière Charles 0 sinon
- NOX concentration en oxydes d'azote
- RM nombre moyen de pièces par logement
- AGE proportion de logements construit avant 1940
- DIS distance du centre de Boston
- RAD accessibilité aux autoroutes de contournement
- TAX taux de l'impôt foncier
- PTRATIO rapport élèves-enseignant par ville
- B $1000(B_k - 0,63)^2$ où B_k est la proportion de Noirs par ville
- LSTAT % de la population à faibles revenus
- class valeur du logement en 1000$

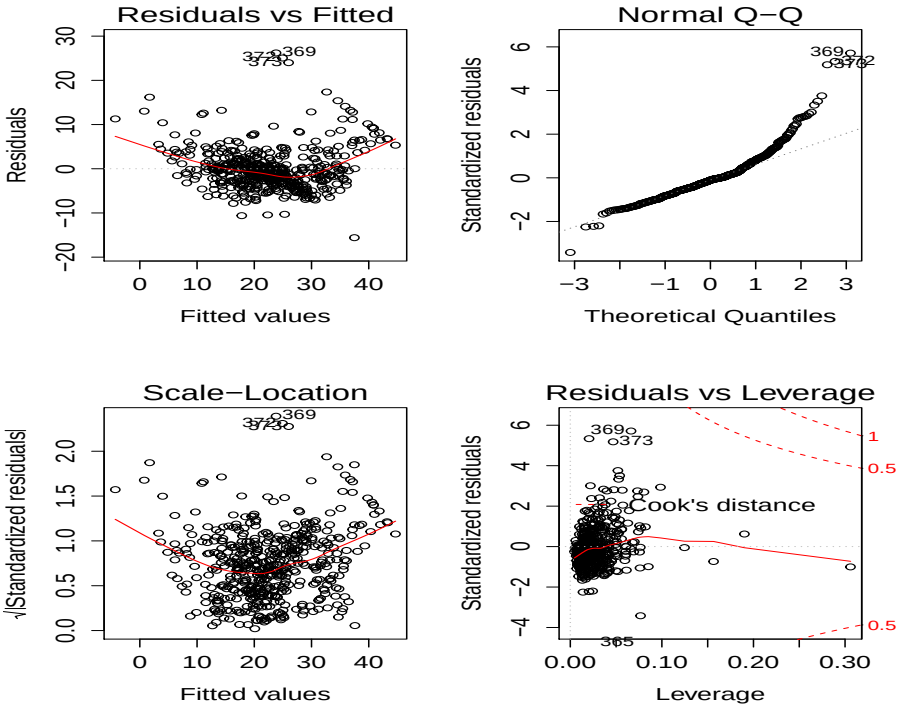
On commence par faire l'étude suivante :

```
> mod1 <- lm(class ~., data = housing)
> summary(mod1)
Call:
lm(formula = class ~ ., data = housing)
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.287e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.381 0.000779 ***
INDUS        2.055e-02  6.150e-02   0.334 0.738352
CHAS         2.687e+00  8.616e-01   3.119 0.001924 **
NOX         -1.777e+01  3.820e+00  -4.651 4.24e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.915e-04  1.321e-02   0.052 0.958274
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.528e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.311e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,        Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16

> plot(mod1)
```



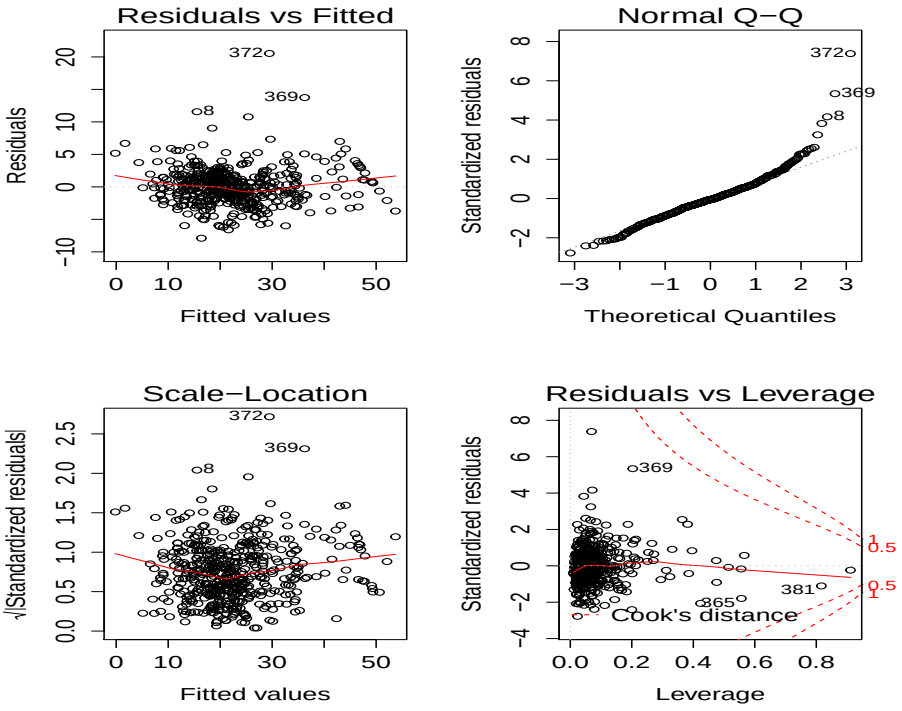
1. Quelle est la part de variance expliquée par ce modèle ?
2. Le modèle de régression est-il significatif dans son ensemble (prendre un risque de première espèce $\alpha = 1\%$) ? Donner l'hypothèse H_0, la statistique du test et la conclusion.
3. Quelles sont les variables significatives (prendre un risque de première espèce $\alpha = 1\%$) ? Donner l'hypothèse H_0, la statistique du test et la conclusion pour l'une d'elles.
4. Est-on sûr qu'il n'y en a pas d'autres ? Proposer un test pour répondre à la dernière question. On explicitera H_0 et la statistique à considérer.
5. Pour un tel modèle, comment s'interprète le signe des coefficients obtenus ? Illustrer votre réponse.
6. Analyser la sortie de $> plot(mod1)$. Chaque graphique est à commenter avec soin.

On se propose maintenant d'ajouter les interactions. Les commandes R sont les suivantes :

```
> mod2 <- lm(class ~ .^2, data = housing)
> mod3 <- step(mod2,k=log(nrow(housing)))
```

7. A quoi correspond la commande $> mod3 < -step(mod2, k = log(nrow(housing)))$. ? Pourquoi a-t-elle été utilisée ? Expliquer le principe de la méthode. Donner deux méthodes alternatives.

```
Le bas de la table du $summary$ est le suivant :
....
Residual standard error: 2.882 on 466 degrees of freedom
Multiple R-squared:  0.9094,        Adjusted R-squared:  0.9018
F-statistic: 119.9 on 39 and 466 DF,  p-value: < 2.2e-16
....
> plot(mod3)
```



8. Le modèle *mod*3 comporte 13 variables principales, combien d'interactions comporte-t-il ? Ce modèle est-il meilleur que le modèle *mod*1 ?