5.1. Multiple regression with permutation tests

The function 'lmorigin' {ape} computes a multiple linear regression and performs tests of significance of *R*-square (*F*-test) and of the equation parameters (*t*-tests) using permutations. A special permutational procedure has been developed to carry out permutation tests in regression through the origin.

The default option of this function is to carry out regressnion through the origin, which is fixed to zero. To obtain an ordinary multiple regression with intercept, call the function with the option 'origin=FALSE'.

```
library(ape)
```

Example 1 from Sokal & Rohlf (1995) Table 16.1: SO_2 air pollution in 41 cities of the USA.

```
data(lmorigin.ex1)
summary(lmorigin.ex1)
head(lmorigin.ex1)

out <- lmorigin(SO2 ~ ., data=lmorigin.ex1, origin=FALSE, nperm=999)
out
```

Example 2: Contrasts computed on the phylogenetic tree of *Lamellodiscus* parasites. Response variable: non-specificity index (NSI); explanatory variable: maximum host size. Data from Table 1 of Legendre & Desdevises (2009).

```
data(lmorigin.ex2)
summary(lmorigin.ex2)
head(lmorigin.ex2)

out <- lmorigin(NSI ~ MaxHostSize, data=lmorigin.ex2, origin=TRUE, nperm=99)
out
```

Example 3: random numbers. No significant result is expected except by chance (type I error).

```
y <- rnorm(50)
X <- as.data.frame(matrix(rnorm(250),50,5))
out <- lmorigin(y ~ ., data=X, origin=FALSE, nperm=999)
out
```

5.2. Sélection pas à pas en régression multiple

Premier exemple : fichier de données 'lmorigin.ex1' disponible dans {ape}

```
library(ape)
data(lmorigin.ex1)
summary(lmorigin.ex1)
head(lmorigin.ex1)
```

Il y a 4 variables dans ce fichier:
SO2, Mean_Annual_Temp_F, No_manufactures, Population_x_1000

5.2.1. Calcul du R-carré ajusté et des VIF

La méthode de Blanchet et al. (2008) demande de calculer d'abord le R-carré ajusté du modèle contenant toutes les variables (sans sélection), puis d'utiliser ce R-carré ajusté comme critère d'arrêt de la sélection pas à pas.

lm.out = lm(SO2 ~ ., data=lmorigin.ex1)

summary(lm.out)

Quel est le R-carré ajusté du modèle global dans ce 'summary' ?

Calculer les vif à partir de l'objet de sortie de 'lm': fonction 'vif' {car}

```
library(car)

vif1 = vif(lm.out)
vif1
```

Comparer ce résultat avec ceux du calcul décrit en classe

```
vif2 = diag(solve(cor(lmorigin.ex1[,2:4])))
vif2
```

5.2.2. Sélection pas à pas par la fonction 'forward.sel' de {adespatial}

```
library(adespatial)
?forward.sel
sel.out = forward.sel(lmorigin.ex1[,1], lmorigin.ex1[,2:4], adjR2thresh =
###Inscrivez ici le R-carré ajusté du modèle global###)

sel.out
```

Recalcular la régression avec les variables sélectionnées

```
lm.out2 = lm(SO2 ~ No_manufactures + Population_x_1000, data=lmorigin.ex1)

summary(lm.out2)
```

Comparez le R-carré ajusté calculé par 'lm' à celui obtenu à la fin de la sélection pas à pas.

5.2.3. Sélection pas à pas par la fonction ‘step’ {stats}

Illustration de la méthode de sélection à l’aide du fichier de données ‘swiss’ disponible dans R {datasets}. Six variables sont fournies pour 47 régions de Suisse. On cherche à calculer un modèle linéaire parcimonieux de la variable ‘Fertility’.

```
?swiss
dim(swiss)
head(swiss)
```

5.2.3.1 Sélection progressive (« forward »)

Il faut créer deux modèles de régression. Le premier modèle, ‘m0’, ne contient que les variables qui doivent obligatoirement faire partie du modèle ; dans le cas le plus général et dans l’exemple ci-dessous, on n’inclut que l’ordonnée à l’origine qui est égale à la moyenne des valeurs de la variable-réponse. Le second modèle, ‘mtot’, contient toutes les variables auxquelles on veut appliquer la sélection progressive ascendante.

```
m0 = lm(Fertility ~ 1, data = swiss)  # Modèle contenant l’ordonnée à l’origine seulement
mtot = lm(Fertility ~ ., data=swiss)  # Modèle contenant toutes les variables explicatives
```

Sélection progressive des variables significatives

```
res.for = step(m0, scope=formula(mtot), direction="forward")
```

Voici les détails de la sélection, fournis à l’écran.
Le signe + indique qu’on tente d’ajouter la variable en question au modèle. La valeur de AIC est celle du modèle contenant la variable en question.
<none> réfère au modèle précédent. Point de départ dans cet exemple : aucune variable.

```
Start:  AIC=238.35
Fertility ~ 1

                  Df Sum of Sq    RSS    AIC  # Étape 1
+ Education        1    3162.7 4015.2 213.04  # AIC le plus bas, < AIC(none)
+ Examination      1    2994.4 4183.6 214.97
+ Catholic         1    1543.3 5634.7 228.97
+ Infant.Mortality 1    1245.5 5932.4 231.39
+ Agriculture      1     894.8 6283.1 234.09
<none>                         7178.0 238.34  # AIC du modèle avec 0 var.

Step:  AIC=213.04                   # Premier modèle : 1 var. explicative
Fertility ~ Education

                  Df Sum of Sq    RSS    AIC  # Étape 2
+ Catholic         1    961.07 3054.2 202.18  # AIC le plus bas, < AIC(none)
+ Infant.Mortality 1    891.25 3124.0 203.25
+ Examination      1    465.63 3549.6 209.25
<none>                         4015.2 213.04  # AIC du modèle avec 1 var.
+ Agriculture      1     61.97 3953.3 214.31

Step:  AIC=202.18                   # Second modèle : 2 var. explicatives
Fertility ~ Education + Catholic

                  Df Sum of Sq    RSS    AIC
+ Infant.Mortality 1    631.92 2422.2 193.29  # Étape 3
```

```
+ Agriculture       1    486.28 2567.9 196.03  # AIC le plus bas, < AIC(none)
<none>                          3054.2 202.18  # AIC du modèle avec 2 var.
+ Examination       1      2.46 3051.7 204.15

Step:  AIC=193.29                     # Troisième modèle : 3 var. explicatives
Fertility ~ Education + Catholic + Infant.Mortality

              Df Sum of Sq    RSS    AIC       # Étape 4
+ Agriculture  1   264.176 2158.1 189.86       # AIC le plus bas, < AIC(none)
<none>                     2422.2 193.29       # AIC du modèle avec 3 var.
+ Examination  1     9.486 2412.8 195.10

Step:  AIC=189.86                     # Quatrième modèle : 4 var. explicatives
Fertility ~ Education + Catholic + Infant.Mortality + Agriculture

              Df Sum of Sq    RSS    AIC       # Étape 5
<none>                     2158.1 189.86       # AIC du modèle avec 4 var.
+ Examination  1    53.027 2105.0 190.69       # AIC > AIC(none)
                                               # Variable non sélectionnée
```

Analyse ‘lm’ du modèle sélectionné

```
summary(res.for)                      # Modèle de régression après sélection

Call:
lm(formula = Fertility ~ Education + Catholic + Infant.Mortality +
    Agriculture, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared: 0.6993,           Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Comparez ces résultats avec ceux de la fonction ‘forward.sel’ de {adespatial}

5.2.3.2 Sélection régressive (« backward »)

Dans ce cas, il suffit de fournir le modèle de régression calculé avec toutes les variables
Voici les détails de la sélection fournis à l’écran
Le signe – indique qu’on tente d’éliminer la variable en question du modèle. La valeur de AIC est celle du modèle sans la variable en question.
<none> réfère au modèle précédent. Point de départ : toutes les variables.

```
res.back = step(mtot, direction="backward")

Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                  Df Sum of Sq    RSS    AIC  # Étape 1
- Examination      1     53.03 2158.1 189.86  # AIC le plus bas, < AIC(none)
<none>                         2105.0 190.69  # AIC modèle avec 5 var.
- Agriculture      1    307.72 2412.8 195.10
- Infant.Mortality 1    408.75 2513.8 197.03
- Catholic         1    447.71 2552.8 197.75
- Education        1   1162.56 3267.6 209.36

Step:  AIC=189.86                     # Premier modèle : 1 var. éliminée
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                  Df Sum of Sq    RSS    AIC  # Étape 2
<none>                         2158.1 189.86  # AIC modèle avec 4 var.
- Agriculture      1    264.18 2422.2 193.29
- Infant.Mortality 1    409.81 2567.9 196.03
- Catholic         1    956.57 3114.6 205.10
- Education        1   2249.97 4408.0 221.43
                                      # Aucune var. avec AIC < AIC(none)
                                      # donc, aucune variable n'est éliminée
```

Analyse 'lm' du modèle sélectionné

```
summary(res.back)                     # Modèle de régression après sélection

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared: 0.6993,           Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

5.2.3.3 Sélection par étapes (ou « stepwise », `direction="both")`

Si on veut commencer avec le modèle ne contenant que l'ordonnée à l'origine, il faut fournir deux modèles à 'step' : m0 et mtot.
Le signe + indique qu'on tente d'ajouter la variable en question au modèle.
Le signe – indique qu'on tente d'éliminer la variable en question du modèle.
La valeur AIC est celle du modèle contenant (+) ou ne contenant pas (–) la variable.
<none> réfère au modèle précédent. Point de départ dans cet exemple : aucune variable.

```
res.step = step(m0, scope=formula(mtot), direction="both")

Start:  AIC=238.35
Fertility ~ 1

                  Df Sum of Sq    RSS    AIC  # Étape 1 « forward »
+ Education        1    3162.7 4015.2 213.04  # AIC le plus bas, < AIC(none)
+ Examination      1    2994.4 4183.6 214.97
+ Catholic         1    1543.3 5634.7 228.97
+ Infant.Mortality 1    1245.5 5932.4 231.39
+ Agriculture      1     894.8 6283.1 234.09
<none>                        7178.0 238.34  # AIC modèle avec 0 var.

Step:  AIC=213.04
Fertility ~ Education

                  Df Sum of Sq    RSS    AIC  # Étape 1 « backward »
+ Catholic         1     961.1 3054.2 202.18  # AIC le plus bas, < AIC(none)
+ Infant.Mortality 1     891.2 3124.0 203.25  # On tente d'inclure
+ Examination      1     465.6 3549.6 209.25  # On tente d'inclure
<none>                        4015.2 213.04  # AIC modèle avec 1 var.
+ Agriculture      1      62.0 3953.3 214.31  # On tente d'inclure
- Education        1    3162.7 7178.0 238.34  # On tente d'exclure cette var.

Step:  AIC=202.18
Fertility ~ Education + Catholic

                  Df Sum of Sq    RSS    AIC  # Étape 2 « forward »
+ Infant.Mortality 1    631.92 2422.2 193.29  # AIC le plus bas, < AIC(none)
+ Agriculture      1    486.28 2567.9 196.03  # On tente d'inclure
<none>                        3054.2 202.18  # AIC modèle avec 2 var.
+ Examination      1      2.46 3051.7 204.15  # On tente d'inclure
- Catholic         1    961.07 4015.2 213.04  # On tente d'exclure cette var.
- Education        1   2580.50 5634.7 228.97  # On tente d'exclure cette var.

Step:  AIC=193.29
Fertility ~ Education + Catholic + Infant.Mortality

                  Df Sum of Sq    RSS    AIC  # Étape 2 « backward »
+ Agriculture      1    264.18 2158.1 189.86  # AIC le plus bas, < AIC(none)
<none>                        2422.2 193.29  # AIC modèle avec 3 var.
+ Examination      1      9.49 2412.8 195.10  # On tente d'inclure
- Infant.Mortality 1    631.92 3054.2 202.18  # On tente d'exclure cette var.
- Catholic         1    701.74 3124.0 203.25  # On tente d'exclure cette var.
- Education        1   2380.38 4802.6 223.46  # On tente d'exclure cette var.
```

```
Step:  AIC=189.86
Fertility ~ Education + Catholic + Infant.Mortality + Agriculture

                   Df Sum of Sq    RSS    AIC  # Étape 3 « forward »
<none>                          2158.1 189.86  # AIC modèle avec 4 var.
+ Examination       1     53.03 2105.0 190.69  # On tente d'inclure
- Agriculture       1    264.18 2422.2 193.29  # On tente d'exclure cette var.
- Infant.Mortality  1    409.81 2567.9 196.03  # On tente d'exclure cette var.
- Catholic          1    956.57 3114.6 205.10  # On tente d'exclure cette var.
- Education         1   2249.97 4408.0 221.43  # On tente d'exclure cette var.
                                   # Aucune variable incluse ou excluse
```

Analyse 'lm' du modèle sélectionné

```
summary(res.step)                    # Modèle de régression après sélection

Call:
lm(formula = Fertility ~ Education + Catholic + Infant.Mortality +
    Agriculture, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared: 0.6993,          Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10


-------
```

Réalisez les trois formes de sélection pour les données du fichier 'lmorigin.ex1' {ape}.
Comparez ces résultats avec ceux de la fonction 'forward.sel' {adespatial} obtenus plus haut.

Voici les modèles pour le point de départ :

```
m.0 = lm(SO2 ~ 1, data=lmorigin.ex1)  # Modèle contenant l'ordonnée à l'origine seulement
m.tot = lm(SO2 ~ .,data=lmorigin.ex1) # Modèle contenant toutes les variables explicatives
```

Écrivez les commandes R pour les différentes formes de sélection pas à pas.

Exemple de sélection impliquant une variable quantitative ainsi que des facteurs.

Réalisez la sélection progressive (`direction="forward")` pour les données du fichier 'CO2' {datasets} qui contient 5 variables : 'Plant' (facteur ordonné : no de la plante et concentration), 'Type' (facteur, 2 régions), 'Treatment' (facteur, 2 niveaux), 'conc' (ambient carbon dioxide concentrations, variable quantitative, 7 niveaux), et 'uptake' (carbon dioxide uptake rates, variable réponse. N'utilisez pas le facteur ordonné 'Plant' (variable composite) dans la régression.

```
?CO2
dim(CO2)
summary(CO2)
```

Voici les modèles pour le point de départ :

```
m.0 = lm(uptake ~ 1, data=CO2)          # Modèle contenant l'ordonnée à l'origine seulement
m.3 = lm(uptake ~ Type+Treatment+conc, data=CO2) # Modèle contenant 3 var. explicatives

res.CO2.for = step(m.0, scope=formula(m.3), direction="forward")

summary(res.CO2.for)
```

========