Chapitre II : Régression linéaire multiple

M. Champion

Université de Paris

2020-2021

I. Le modèle linéaire multiple

1.1 Objectifs

Un modèle linéaire multiple :

- permet de **décrire** et **modéliser** la relation entre une variable aléatoire quantitative continue Y (dite à expliquer) et p variables quantitatives contrôlées $X^1, ..., X^p$ (dites explicatives),

- utilise des observations $(x_i^1, ..., x_i^p, y_i)_{i=1,\cdots,n}$ d'un échantillon de taille n où :
  - $x_1^j, \cdots, x_n^j$: valeurs connues et fixées (non aléatoires) de X^j $(1 \leq j \leq p)$,
  - $y_1, \cdots, y_n$: réponses obtenues considérées comme n réalisations de Y.

I. Le modèle linéaire multiple

1.2 Description du modèle

On appelle **modèle linéaire multiple gaussien** un modèle statistique qui peut s'écrire sous la forme :

$$\forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + \varepsilon_i,$$

où les $(\varepsilon_i)_{1 \leq i \leq n}$ sont des termes d'erreur non observés, i.i.d :

$$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

- $\beta_0, ..., \beta_p$ (paramètres d'espérance) sont inconnus et à estimer pour comprendre l'effet des variables $X^1, ..., X^p$ sur la réponse Y.
- σ^2 (paramètre de variance) est également inconnu et à estimer.

$\rightarrow$ La **dimension** du modèle est $p + 1$.

I. Le modèle linéaire multiple

1.2 Description du modèle

Le modèle

$$\forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \beta_1 x_i^1 + ... + \beta_p x_i^p + \varepsilon_i,$$

avec

$$\forall i \in [\![1, n]\!], \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

est équivalent à

$$\forall i = 1, \cdots, n, \quad \text{les } Y_i \text{ sont indépendants et } Y_i \sim \mathcal{N}(\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j; \sigma^2).$$

I. Le modèle linéaire multiple

1.3 Ecriture matricielle du modèle linéaire multiple

Le modèle se réécrit matriciellement :

$$Y = X\beta + \varepsilon,$$

où X est une matrice de taille $n \times (p+1)$ de terme général x_i^j, excepté la première colonne qui contient uniquement des 1, $Y = {}^t(Y_1, ..., Y_n)$, $\varepsilon = {}^t(\varepsilon_1, ..., \varepsilon_n)$ et $\beta = {}^t(\beta_0, ..., \beta_p)$.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^1 & \dots & x_1^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \dots & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

I. Le modèle linéaire multiple

1.3 Ecriture matricielle du modèle linéaire multiple (R)

On utilisera le jeu de données `state.x77` tout au long de ce cours. Ce jeu de données contient des stats sur les 50 états des Etats-Unis dans les années 70.

```
data <- as.data.frame(state.x77)
str(data)
```

```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

On souhaite expliquer l'espérance de vie `Life.Exp` en fonction des 7 autres variables `Population`, `Income`, `Illiteracy`, `Murder`, `HS.Grad`, `Frost`, `Area`.

II. Estimation des paramètres

2.1 Estimation de β par moindres carrés

La **méthode des moindres carrés** consiste à estimer $\beta \in \mathbb{R}^{p+1}$ en minimisant la somme des carrés des erreurs :

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i^1 - ... - \beta_p x_i^p)^2,$$

qui se réécrit aussi :

$$\min_{\beta \in \mathbb{R}^{p+1}} \|Y - X\beta\|^2.$$

Théorème

Si ${}^tXX$ est inversible, l'estimateur des moindres carrés B de β est :

$$B = ({}^tXX)^{-1}{}^tXY.$$

II. Estimation des paramètres

2.2 Estimation de la variance σ^2

On note

- $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (Somme des Carrés Totale),
- $SCM = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$ (Somme des Carrés du Modèle),
- $SCR = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$ (Somme des Carrés Résiduelle),

de telle sorte que :

$$SCT = SCM + SCR.$$

Théorème

Un estimateur s^2 de σ^2 est donné par :

$$s^2 = \frac{SCR}{n-(p+1)}.$$

II. Estimation des paramètres

2.3 Propriétés et lois des estimateurs

Théorème

Sous les hypothèses du modèle linéaire gaussien multiple, s^2 est un estimateur sans biais de σ^2 et on a :

$$\frac{(n-p-1)s^2}{\sigma^2} = \frac{\text{SCR}}{\sigma^2} \sim \chi^2(n-p-1).$$

De plus, s^2 est indépendant de B.

II. Estimation des paramètres

2.3 Propriétés et lois des estimateurs

Théorème

$b_0, ..., b_p$ sont des estimateurs sans biais de $\beta_0, ..., \beta_p$. $B = {}^t(b_0, ..., b_p)$ est un vecteur gaussien d'espérance β et de matrice de covariance $\sigma^2({}^tXX)^{-1}$:

$$B \sim \mathcal{N}(\beta, \sigma^2({}^tXX)^{-1}).$$

En particulier, si on note pour tout $j = 0, ..., p$, c_j le $(j+1) \times (j+1)$ élément diagonal de la matrice $({}^tXX)^{-1}$:

$$\forall j \in [\![0, p]\!], \quad b_j \sim \mathcal{N}(\beta_j, \sigma^2 c_j).$$

On a aussi :

$$\forall j \in [\![0, p]\!], \quad \frac{b_j - \beta_j}{\sqrt{s^2 c_j}} \sim St(n - p - 1).$$

II. Estimation des paramètres

2.4 Qualité d'ajustement du modèle

La qualité d'ajustement du modèle est mesurée par le coefficient de détermination, défini par :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y}_i)^2}.$$

Il détermine à quel point l'équation de régression est adaptée pour décrire la distribution des points.

Remarque : En règle général, on utilise le R^2 ajusté :

$$R^2_{\text{ajusté}} = 1 - \frac{SCR\,/(n-p-1)}{SCT\,/(n-1)}.$$

II. Estimation des paramètres

2.5 Application sur R

```
modele <- lm(Life.Exp~.,data=data)
coefficients(modele)
```

```
##   (Intercept)    Population        Income    Illiteracy          Mu
##  7.094322e+01  5.180036e-05 -2.180424e-05  3.382032e-02 -3.011232
##       HS.Grad         Frost          Area
##  4.892948e-02 -5.735001e-03 -7.383166e-08
```

```
summary(modele)$sigma
```

```
## [1] 0.7447777
```

```
summary(modele)$r.squared
```

```
## [1] 0.7361563
```

III. Inférence sur les paramètres

3.1 Intervalle de confiance

Les **intervalles de confiance** de niveau de confiance $1-\delta$ sont établis à partir des lois donnés dans le théorème précédent :

$$\mathrm{IC}_{1-\delta}(\beta_j) = [b_j - t_{\delta/2}\sqrt{s^2 c_j}, b_j + t_{\delta/2}\sqrt{s^2 c_j}].$$

où $t_{\delta/2}$ est le $(1-\delta/2)$-quantile de la distribution de Student $St(n-p-1)$:

$$\mathbb{P}\left(St(n-p-1) \leq t_\delta\right) = 1-\delta/2.$$

En pratique, t_δ est lu dans table de $St(n-p-1)$.

III. Inférence sur les paramètres

3.2 Prévision

Etant donnée une nouvelle observation $(x_{n+1}^1, ..., x_{n+1}^p)$ de $(X^j)_{1\leq j\leq p}$, une prévision $\hat{y}_{n+1}$ de y_{n+1} (non disponible!) est donnée par :

$$\begin{aligned} \hat{y}_{n+1} &= b_0 + b_1 x_{n+1}^1 + ... + b_p x_{n+1}^p \\ &= X_{n+1} B, \end{aligned}$$

où $B = {}^t(b_0, ..., b_p)$ est le vecteur de taille $p+1$ contenant les estimations des paramètres $(\beta_0, ..., \beta_p)$ du modèle et $X_{n+1} = (1, x_{n+1}^1, ..., x_{n+1}^p)$ est le vecteur ligne de taille $p+1$ contenant les nouvelles observations.

A t'on une idée de la précision de cette prévision?

- Construction d'un intervalle de prédiction de Y_{n+1}
- Construction d'un intervalle de confiance de $\mathbb{E}(Y_{n+1})$

III. Inférence sur les paramètres

3.2 Prévision

Théorème

Dans le cadre du MLGM, on a les deux résultats suivants :

$$\frac{\hat{Y}_{n+1} - \mathbb{E}(Y_{n+1})}{\sqrt{s^2 X_{n+1}({}^tXX)^{-1}X_{n+1}}} \sim St(n-p-1),$$

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{\sqrt{s^2(1 + X_{n+1}({}^tXX)^{-1}X_{n+1})}} \sim St(n-p-1).$$

On en déduit les intervalles de prédiction et de confiance suivants :

$$\mathrm{IP}_{1-\delta}(Y_{n+1}) = [\hat{y}_{n+1} \pm c_\delta s\sqrt{1 + X_{n+1}({}^tXX)^{-1}X_{n+1}}].$$

$$\mathrm{IC}_{1-\delta}(\mathbb{E}(Y_{n+1})) = [\hat{y}_{n+1} \pm c_\delta s\sqrt{X_{n+1}({}^tXX)^{-1}X_{n+1}}].$$

III. Inférence sur les paramètres

3.3 Application sur R

```
confint(modele)
```

```
##                    2.5 %        97.5 %
## (Intercept)  6.741567e+01  7.447078e+01
## Population  -7.101457e-06  1.107022e-04
## Income      -5.150751e-04  4.714666e-04
## Illiteracy  -7.053624e-01  7.730031e-01
## Murder      -3.952076e-01 -2.070387e-01
## HS.Grad      1.861199e-03  9.599776e-02
## Frost       -1.207830e-02  6.082932e-04
## Area        -3.440321e-06  3.292657e-06
```

III. Inférence sur les paramètres

3.3 Application sur R

```
newdata <- data.frame(Population=100000,Income=8000,
                      Illiteracy = 0.5, Murder = 3,
                      HS.Grad = 50,Frost=150,Area=1500)
predict(modele, newdata,interval="prediction")
```

```
##        fit      lwr      upr
## 1 76.64848 71.03262 82.26434
```

```
predict(modele, newdata,interval="confidence")
```

```
##        fit      lwr      upr
## 1 76.64848 71.23749 82.05947
```

IV. Tests dans le modèle linéaire multiple

4.1 Test de la contribution globale

Il s'agit d'une généralisation du test de Fisher à p variables explicatives :

- compare les modèles M_1 (constant) et M_{p+1} (complet)

$$M_1 \quad : \quad \forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

$$M_{p+1} \quad : \quad \forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

- revient à tester, au risque δ fixé, l'hypothèse nulle

$$(H_0) : \text{ modèle } M_1 \text{ (aucune variable n'a d'influence)}$$

contre l'alternative

$$(H_1) : \text{ modèle } M_{p+1} \text{ (au moins une des variables a de l'influence)}$$

IV. Tests dans le modèle linéaire multiple

4.1 Test de la contribution globale

La **statistique de test** est

$$T_n = \frac{\text{SCM}/p}{\text{SCR}/(n-p-1)} \sim_{H_0} F(p, n-p-1),$$

ou encore :

$$T_n = \frac{(\text{SCR}_{\text{M}_1} - \text{SCR})/p}{\text{SCR}/(n-p-1)} \sim_{H_0} F(p, n-p-1),$$

où SCR_{M_1} est la Somme des Carrés Résiduels pour le modèle M_1.

IV. Tests dans le modèle linéaire multiple

4.1 Test de la contribution globale

Notons c_δ le quantile d'ordre δ de $F(p, n-p-1)$ tel que :

$$\mathbb{P}(T_n \leq c_\delta) = 1 - \delta$$

La **règle de décision** est alors la suivante :

- si $T_n > c_\delta$, on rejette H_0,
- si $T_n \leq c_\delta$, on ne rejette pas H_0.

En pratique, on prend un échantillon de taille n, on calcule la réalisation t_n de T_n sur cet échantillon et on compare sa valeur à c_δ, qui est lu sur la table de Fisher. On peut aussi calculer la p-valeur$= \mathbb{P}_{H_0}(T_n > t_n)$ que l'on compare au risque δ. Si la p-valeur est inférieure à δ, le test est significatif.

IV. Tests dans le modèle linéaire multiple

4.2 Test du modèle réduit

On teste si un ensemble de q variables explicatives ne suffit pas à expliquer Y :

- **Hypothèses** :

$$(\mathrm{H}_0) : \text{modèle réduit}$$

$$\text{Modèle } M_{q+1} : \forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_i^j + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

contre

$$(\mathrm{H}_1) : \text{modèle complet}$$

$$\text{Modèle } M_{p+1} : \forall i \in [\![1, n]\!], \quad Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2).$$

Ou encore :

$$(H_0) : \forall j = q+1, ..., p, \quad \beta_j = 0 \text{ contre } (H_1) : \exists j \in \{q+1, ..., p\}, \quad \beta_j \neq 0.$$

IV. Tests dans le modèle linéaire multiple

4.2 Test du modèle réduit

Théorème

On a le résultat suivant :

$$\frac{\text{SCR}}{\sigma^2} \sim \chi^2(n-p-1).$$

On en déduit donc que sous H_0 *:*

$$\frac{\text{SCR}_{M_{q+1}}}{\sigma^2} \sim_{H_0} \chi^2(n-q-1).$$

Si on note $\text{SCE} = \text{SCR}_{M_{q+1}} - \text{SCR}$, SCE *et* SCR *sont indépendants et*

$$\frac{\text{SCE}}{\sigma^2} \sim_{H_0} \chi^2(p-q).$$

La **statistique de test** est alors définie par

$$T_n = \frac{(\text{SCR}_{M_{q+1}} - \text{SCR})/(p-q)}{\text{SCR}\,/(n-p-1)} \sim_{H_0} F(p-q, n-p-1).$$

IV. Tests dans le modèle linéaire multiple

4.2 Test du modèle réduit

Notons c_δ le quantile d'ordre δ de $F(p-q, n-p-1)$ tel que :

$$\mathbb{P}(T_n \leq c_\delta) = 1 - \delta$$

- La **règle de décision** est alors la suivante :
  - si $T_n > c_\delta$, on rejette H_0,
  - si $T_n \leq c_\delta$, on ne rejette pas H_0.

En pratique, on mesure T_n sur un échantillon de taille n. Si $c_n > t_\delta$, on conserve le modèle complet, on considère que le passage de M_{q+1} à M_{p+1} est significatif : au moins l'une des variables $X^{q+1}, ..., X^p$ a une influence significative sur Y (en plus de $X^1, ..., X^q$). Si $t_n \leq c_\delta$, on ne rejette pas (H_0), on considère que le passage de M_{q+1} à M_{p+1} n'est pas significatif : l'influence des variables explicatives $X^{q+1}, ..., X^p$ n'est pas significative pour expliquer Y.

IV. Tests dans le modèle linéaire multiple

4.3 Table de l'analyse de la variance de la régression

IV. Tests dans le modèle linéaire multiple

4.4 Application sous R

```
modele1 <- lm(Life.Exp~1,data=data)
anova(modele,modele1)
```

```
## Analysis of Variance Table
##
## Model 1: Life.Exp ~ Population + Income + Illiteracy + Murder + F
##     Frost + Area
## Model 2: Life.Exp ~ 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     42 23.297
## 2     49 88.299 -7   -65.002 16.741 2.534e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

IV. Tests dans le modèle linéaire multiple

4.4 Application sous R

```
modele4 <- lm(Life.Exp~Population+Frost+Area+Income,data=data)
anova(modele,modele4)
```

```
## Analysis of Variance Table
##
## Model 1: Life.Exp ~ Population + Income + Illiteracy + Murder + F
##     Frost + Area
## Model 2: Life.Exp ~ Population + Frost + Area + Income
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     42 23.297
## 2     45 68.941 -3   -45.644 27.429 5.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

IV. Tests dans le modèle linéaire multiple

4.4 Application sous R

Quelques remarques :

- on peut toujours comparer des modèles emboîtés l'un dans l'autre,
- on peut toujours comparer tous les modèles au modèle complet (structure la plus riche),
- on peut toujours comparer tous les modèles au modèle constant (structure la moins riche),
- on ne peut jamais comparer les modèles non emboîtés à l'aide de ce test.

V. Sélection de modèles/variables

Si le test global (H_0) : M_1 contre (H_1) : M_{p+1} est significatif, au moins une des variables contribue à expliquer Y. Mais lesquelles?

- Première idée :
  - tester la nullité des p coefficients de régression avec le test de Student $\forall j = 1, \cdots, p$, $(H_0) : \beta_j = 0$ contre $(H_1) : \beta_j \neq 0$,
  - éliminer toutes les variables X^j telles que le test de Student associé n'est pas significatif,

V. Sélection de modèles/variables

Si le test global (H_0) : M_1 contre (H_1) : M_{p+1} est significatif, au moins une des variables contribue à expliquer Y. Mais lesquelles?

- Première idée :
  - tester la nullité des p coefficients de régression avec le test de Student $\forall j = 1, \cdots, p$, $(H_0) : \beta_j = 0$ contre $(H_1) : \beta_j \neq 0$,
  - éliminer toutes les variables X^j telles que le test de Student associé n'est pas significatif,
  - **démarche fausse** car chaque test est effectué alors que les autres variables sont fixées, on ne prend pas en compte les possibles effets conjoints!

V. Sélection de modèles/variables

Si le test global (H_0) : M_1 contre (H_1) : M_{p+1} est significatif, au moins une des variables contribue à expliquer Y. Mais lesquelles?

- Première idée :
  - tester la nullité des p coefficients de régression avec le test de Student $\forall j = 1, \cdots, p$, $(H_0) : \beta_j = 0$ contre $(H_1) : \beta_j \neq 0$,
  - éliminer toutes les variables X^j telles que le test de Student associé n'est pas significatif,
  - **démarche fausse** car chaque test est effectué alors que les autres variables sont fixées, on ne prend pas en compte les possibles effets conjoints!

- Deuxième idée : sélectionner les variables pertinentes par des méthodes de recherche exhaustive :
  - nécessite de comparer 2^p modèles,
  - si p pas trop élevé, on peut comparer tous les modèles possibles et choisir "le meilleur" modèle à partir d'un critère statistique de sélection de modèles.

V. Sélection de modèles/variables

5.1. Critères de sélection : critère du R^2

On peut se baser sur le coefficient de détermination R^2, défini pour tout modèle M_{q+1} par :

$$R^2(q) = 1 - \frac{SCR_{M_{q+1}}}{SCT}.$$

- $0 \leq R^2 \leq 1$ donne la pourcentage de la variation totale de Y expliquée par le modèle de régression linéaire,
- R^2 augmente avec le nombre q de variables explicatives qui entrent dans le modèle,
- R^2 atteint son maximum si toutes les variables disponibles sont incluses, c'est-à-dire pour le modèle complet M_{p+1},
- Défaut : ne permet pas de comparer deux modèles ayant des nombres de variables explicatives différents.

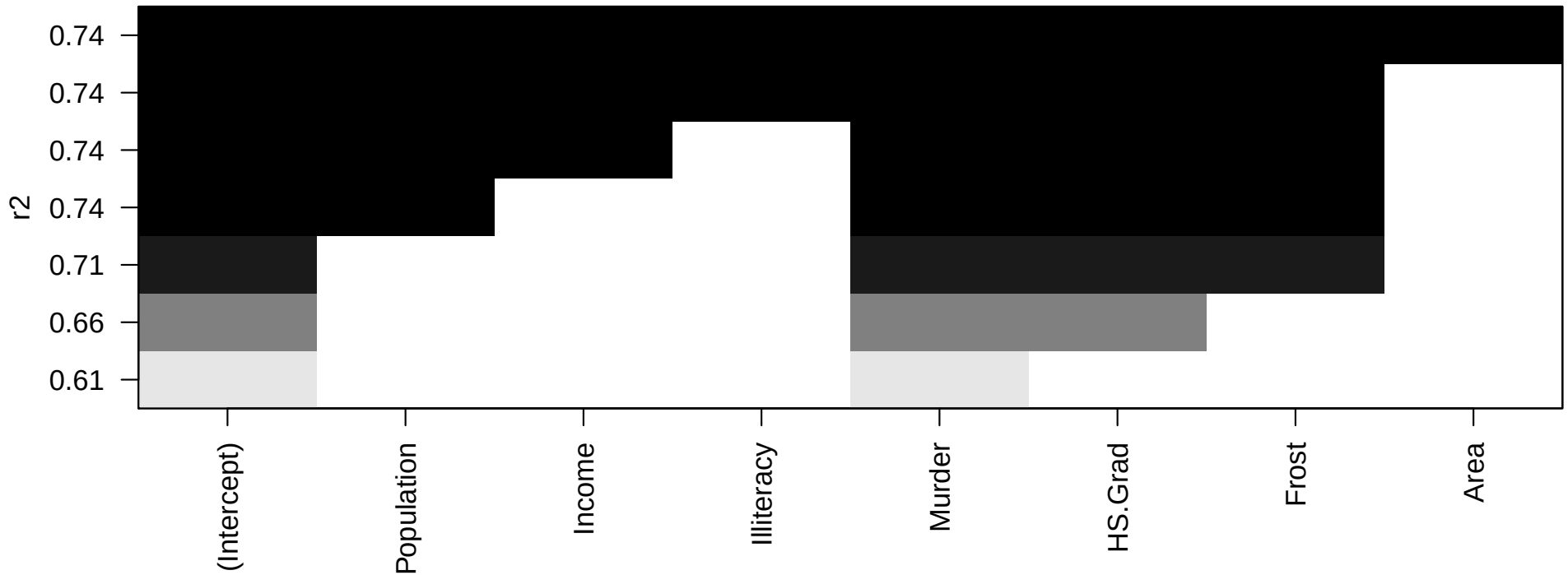
V. Sélection de modèles/variables

5.1. Critères de sélection : critère du R^2 (R)

```
leaps <- regsubsets(Life.Exp~.,data=data,nbest=1,nvmax=10)
round(summary(leaps)$rsq,3)
```

```
## [1] 0.610 0.663 0.713 0.736 0.736 0.736 0.736
```

```
plot(leaps, scale="r2")
```



V. Sélection de modèles/variables

5.1. Critères de sélection : critère du R^2 ajusté

On choisit le modèle qui optimise le R^2 ajusté :

$$\max_{q \leq p} \quad R^2_{\text{ajusté}}(q) = \max_{q \leq p} \quad 1 - \frac{\text{SCR}_{M_{q+1}} / (n - q - 1)}{\text{SCT} / (n - 1)}.$$

- $R^2_{\text{ajusté}}$ n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle,
- comparaison possible de modèles n'ayant pas le même nombre de variables explicatives.

V. Sélection de modèles/variables

5.1. Critères de sélection : critère du R^2 ajusté (R)

```
summary(modele)$adj.r.squared
```

```
## [1] 0.6921823
```

```
summary(modele1)$adj.r.squared
```

```
## [1] 0
```

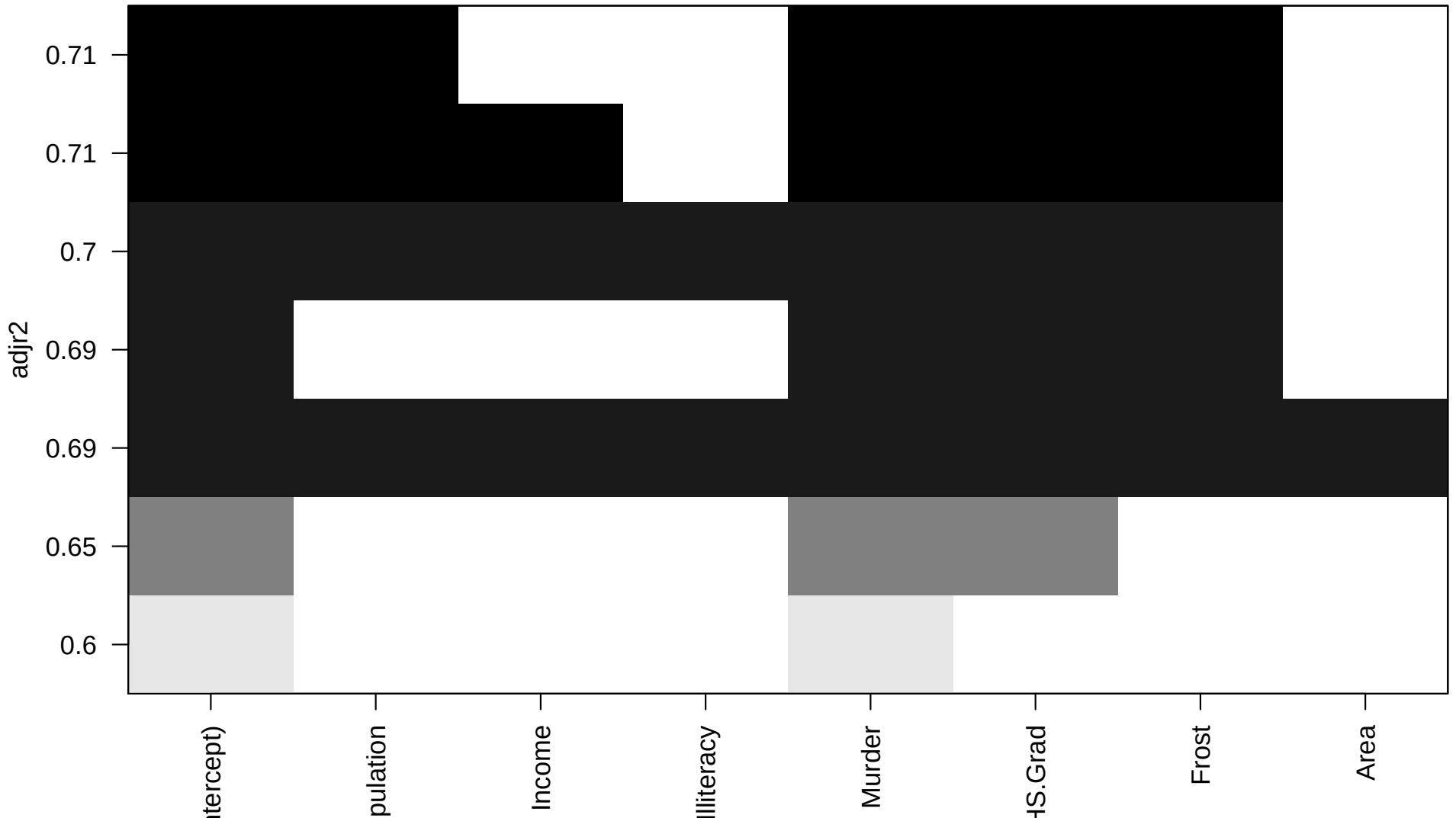
```
round(summary(leaps)$adjr2,3)
```

```
## [1] 0.602 0.648 0.694 0.713 0.706 0.699 0.692
```

V. Sélection de modèles/variables

5.1. Critères de sélection : critère du R^2 ajusté (R)

```
plot(leaps,scale="adjr2")
```



V. Sélection de modèles/variables

5.1. Critères de sélection : C_p de Mallows

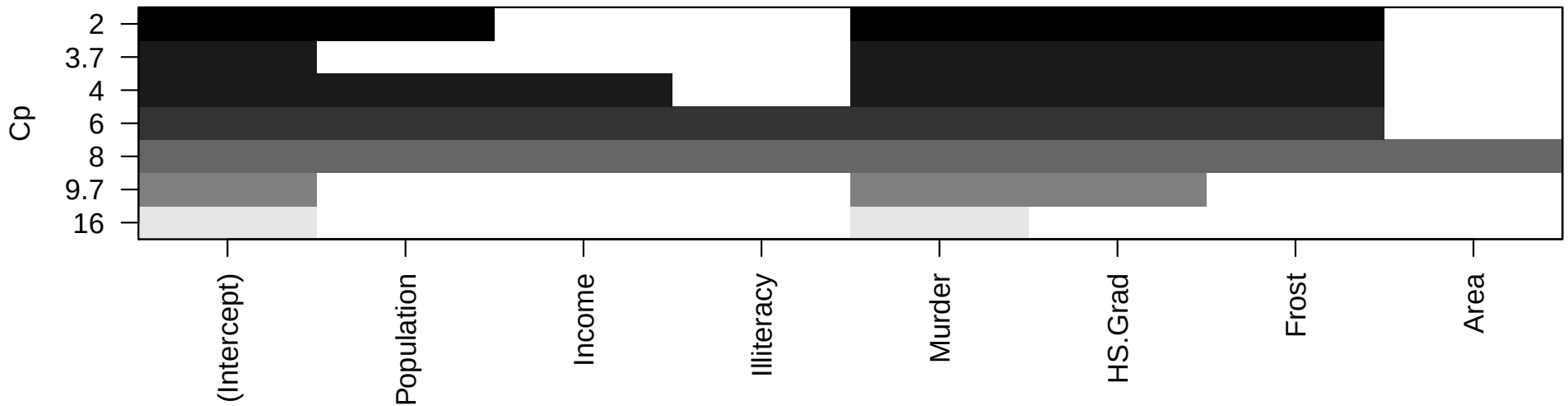
On choisit le modèle qui minimise le coefficient C_p de Mallows :

$$\min_{q \leq p} C_p(q) = \min_{q \leq p} \quad \frac{SCR_{M_{q+1}}}{s^2} - n + 2(q+1).$$

```
round(summary(leaps)$cp,3)
```

```
## [1] 16.127  9.670  3.740  2.020  4.009  6.002  8.000
```

```
plot(leaps,scale="Cp")
```



V. Sélection de modèles/variables

5.1. Critères de sélection : PRESS

Si on note $\hat{y}_{(i)}$ la prédiction de y_i calculée sans tenir compte de la i-ème observation $(y_i, x_i^1, ..., x_i^p)$, le PRESS est défini par :

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

et permet donc de comparer les capacités prédictives de deux modèles.

```
library(qpcR)
PRESS(modele4)$stat
```

```
## .........10.........20.........30.........40.........50

## [1] 125.0406
```

```
PRESS(modele1)$stat
```

```
## .........10.........20.........30.........40.........50

## [1] 91.93982
```

V. Sélection de modèles/variables

5.2. Critères de vraisemblance pénalisée

Certains critères se ramènent à l'utilisation d'une version pénalisée de la vraisemblance du modèle, afin de favoriser un modèle parcimonieux.

- Critère AIC :

$$\min_{q \leq p} \mathrm{AIC}(q) = \min_{q \leq p} -2 \log(L(M_{q+1})) + 2(q+1),$$

où $\log(L(M_{q+1}))$ est la log-vraisemblance du modèle M_{q+1} ($\sim \log \frac{\mathrm{SCR}(M_{q+1})}{n}$).

- Critère BIC : on choisit le modèle qui minimise le critère BIC suivant :

$$\min_{q \leq p} \mathrm{BIC}(q) = \min_{q \leq p} -2 \log(L(M_{q+1})) + (q+1) \log n.$$

Le critère BIC, contrairement au critère AIC, fait intervenir la taille de l'échantillon n.

V. Sélection de modèles/variables

5.2. Critères de vraisemblance pénalisée (R)

```
AIC(modele)
```

```
## [1] 121.7092
```

```
AIC(modele1)
```

```
## [1] 174.3291
```

```
BIC(modele)
```

```
## [1] 138.9174
```

```
BIC(modele1)
```

```
## [1] 178.1532
```

V. Sélection de modèles/variables

5.3. Algorithmes de sélection

Lorsque p est grand, on préfère les méthodes de sélection pas à pas qui consiste à introduire ou à supprimer les variables les unes après les autres :

Sélection ascendante "forward"

On part du modèle M_1 : $\forall i \in [\![1, n]\!], \quad y_i = \beta_0 + \varepsilon_i$

1. choisir la variable X^{j_1} qui contribue le plus à expliquer Y
   - celle qui garantit le coefficient de détermination R^2 le plus élevé,
   - celle pour laquelle le test de Fisher du modèle M_1 contre modèle M_2 : $\forall i \in [\![1, n]\!], \quad y_i = \beta_0 + \beta_{j_1} x_i^{j_1} + \varepsilon_i$ est le plus significatif...

1'. tester la nullité du coefficient de régression β_{j_1} : retenir X^{j_1} si le test est significatif (p-valeur du test de student $\leq 5\%$)

V. Sélection de modèles/variables

5.3. Algorithmes de sélection

Lorsque p est grand, on préfère les méthodes de sélection pas à pas qui consiste à introduire ou à supprimer les variables les unes après les autres :

Sélection ascendante "forward"

2. choisir la variable X^{j_2} qui apporte le plus d'information en plus de X^{j_1}
   - celle pour laquelle le test de Fisher du modèle M_2 contre modèle M_3 : $\forall i \in [\![1, n]\!], \quad y_i = \beta_0 + \beta_{j_1} x_i^{j_1} + \beta_{j_2} x_i^{j_2} + \varepsilon_i$ est le plus significatif possible,
   - celle qui fait progresser le plus le R^2 ...

2'. tester la nullité du coefficient de régression associé à X^{j_2} que l'on retient si le test de student est significatif.

...

V. Sélection de modèles/variables

5.3. Algorithmes de sélection

Lorsque p est grand, on préfère les méthodes de sélection pas à pas qui consiste à introduire ou à supprimer les variables les unes après les autres :

Sélection ascendante "forward"

Arrêt lorsque

- le nombre maximal de variables fixé à l'avance est atteint
- ou lorsque lorsque la valeur de R^2 fixée à l'avance est atteinte
- ou quand le test de nullité du coefficient de régression de la dernière variable introduite n'est pas significatif ...

En pratique, cette méthode est mal fondée théoriquement et donc déconseillée.

V. Sélection de modèles/variables

5.3. Algorithmes de sélection

Lorsque p est grand, on préfère les méthodes de sélection pas à pas qui consiste à introduire ou à supprimer les variables les unes après les autres :

Sélection descendante "backward" Il s'agit de la version symétrique du "forward" :

- on part du modèle complet M_{p+1}
- à chaque étape k de l'algorithme, on enlève la variable j_k
  - qui donne le R^2 le plus faible
  - ou qui donne le test de Student $H_0 : \beta_{j_k} = 0$ contre $H_1 : \beta_{j_k} \neq 0$ le moins significatif (p-valeur la plus grande)
  - ou qui donne le test de Fisher du modèle sans X^{j_k} contre le modèle comprenant X^{j_k} le moins significative...
- arrêt : lorsque toutes les p-valeurs des tests de student/Fisher sont inférieures à 10% (ou 5% ...)

V. Sélection de modèles/variables

5.3. Algorithmes de sélection

Lorsque p est grand, on préfère les méthodes de sélection pas à pas qui consiste à introduire ou à supprimer les variables les unes après les autres :

Sélection mixte "stepwise" Ascendante avec remise en cause à chaque étape des variables déjà introduites, ce qui permet d'éliminer les variables qui ne sont plus informatives compte tenu de celle qui vient d'être ajoutée.

- étape 1 : identique à sélection ascendante
- étape 2 : choisir X^{j_2} qui apporte le plus d'information en plus de X^{j_1}.
  - tester le modèle M_1 contre le modèle M_3 : $Y_i = \beta + \beta_{j_1} x_i^{j_1} + \beta_{j_2} x_i^{j_2} + \varepsilon_i$
  - tester la nullité du coefficient de régression associé à X^{j_2} (dans le modèle M_3)
  - remettre alors en cause X^{j_1} en testant la nullité de β_{j_1}, Si le test est significatif on conserve M_3, sinon on enlève X^{j_1} du modèle

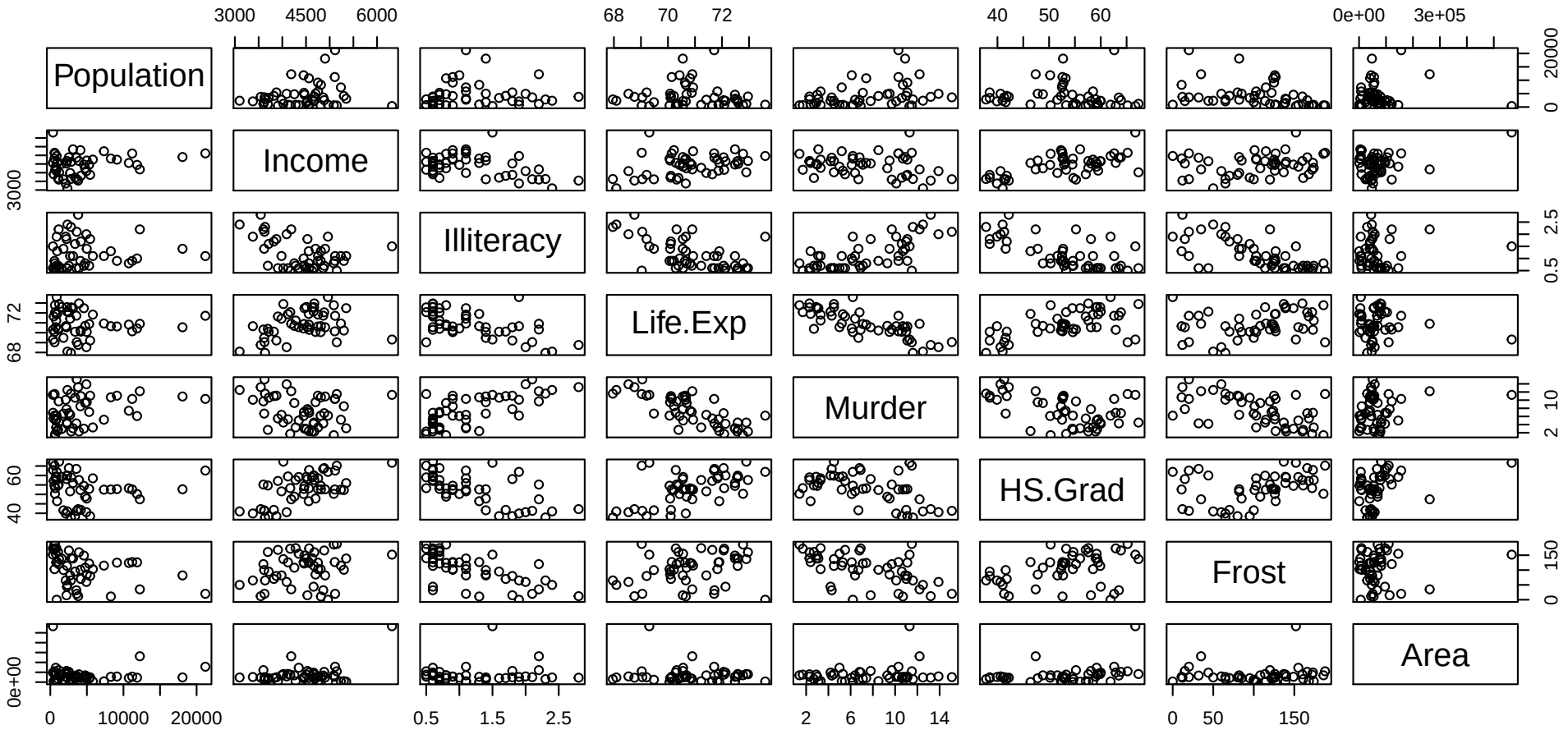
VI. Problèmes de colinéarité

- Inverse de la matrice ${}^tXX$ non définie (problème de précision numérique),
- Estimateurs de variance très (trop) grande
- Difficulté de la sélection de variables : on ne peut enlever les variables X^j dont l'effet n'est pas significatif (test de Student non significatif) à cause des problèmes de colinéarité de variables explicatives
- Situation classique
  - $r(X^1, X^2)$ élevé : les 2 variables explicatives sont fortement corrélées
  - $r(X^1, Y)$ et $r(X^2, Y)$: les 2 variables explicatives sont corrélées avec la variable à expliquer
  - dans le modèle avec X^1 seul, X^1 est significatif
  - dans le modèle avec X^2 seul, X^2 est significatif
  - dans le modèle avec les deux variables, ni X^1, ni X^2 ne sont significatifs
  - quel modèle choisir ? celui avec X^1 seul ou celui avec X^2 seul ? les 2 conviennent!

VI. Problèmes de colinéarité

Comment diagnostiquer une situation de colinéarité critique?

```
plot(data)
```



VI. Problèmes de colinéarité

Comment diagnostiquer une situation de colinéarité critique?

```
cor(data)
```

```
##                Population     Income  Illiteracy    Life.Exp      Mur
## Population  1.00000000  0.2082276  0.10762237 -0.06805195  0.3436
## Income      0.20822756  1.0000000 -0.43707519  0.34025534 -0.2300
## Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793  0.7029
## Life.Exp   -0.06805195  0.3402553 -0.58847793  1.00000000 -0.7808
## Murder      0.34364275 -0.2300776  0.70297520 -0.78084575  1.0000
## HS.Grad    -0.09848975  0.6199323 -0.65718861  0.58221620 -0.4879
## Frost      -0.33215245  0.2262822 -0.67194697  0.26206801 -0.5388
## Area        0.02254384  0.3633154  0.07726113 -0.10733194  0.2283
##                HS.Grad      Frost         Area
## Population -0.09848975 -0.3321525  0.02254384
## Income      0.61993232  0.2262822  0.36331544
## Illiteracy -0.65718861 -0.6719470  0.07726113
## Life.Exp    0.58221620  0.2620680 -0.10733194
## Murder     -0.48797102 -0.5388834  0.22839021
## HS.Grad     1.00000000  0.3667797  0.33354187
```

VI. Problèmes de colinéarité

Comment diagnostiquer une situation de colinéarité critique?

On définit le **VIF** (Variance Inflation Factor) qui mesure qualitativement la dépendance linéaire d'une variable X^j par rapport aux autres variables :

$$\text{VIF}_j = \frac{1}{1 - \text{R}_j^2},$$

où R_j^2 est le coefficient de détermination de la régression de la variable X^j sur les autres variables.

VI. Problèmes de colinéarité

Pour le calculer :

- effectuer la régression de X^j sur les autres variables explicatives :

$$\forall i \in [\![1, n]\!], \quad x_i^j = \beta'_0 + \sum_{k \neq j} \beta'_k x_i^k + \varepsilon'_i,$$

- estimer les coefficients $(\beta'_j)_{j \neq k}$,
- calculer le R_k^2 associé,
- en déduire le VIF_k.

Interprétation du VIF :

- si $R_j^2 \sim 0$, X^j n'est pas colinéaire aux autres variables et $\text{VIF}_j = 1$,
- si $R_j^2 \sim 1$, X^j est fortement liée aux autres variables et VIF_j est grand,
- on considère souvent que $\text{VIF}_j \geq 10$ est un signe de colinéarité importante.