MAT7381 Solution exercice 5.15

Un chercheur a tenté de déterminer si la grosseur du cerveau est liée à l'intelligence. Il a donc prélevé des données sur la grosseur du cerveau (mesurée par l'IRM : imagerie par résonance magnétique) et une série de tests d'aptitude. Les variables sont les suivantes :

1. Sexe: Masculin (=1) et féminin (=0)
2. Total: Le QI basé sur les quatre sous-tests de Wechsler (1981)
3. Verbal: Le QI basé sur les quatre sous-tests d'aptitude verbale de Wechsler (1981)
4. Perf: Le QI basé sur les quatre sous-tests de performance de Wechsler (1981)
5. poids: Le poids du sujet, en livres
6. taille: La taille du sujet, en pouces
7. irm: Le nombre total de pixels obtenus à l'IRM.

Inverser les rôles intuitivement normaux des variables et considérez IRM comme variable endogène et toutes les autres comme variables exogènes. Déterminer le meilleur choix possible des variables exogènes et déterminer si votre choix permet de conclure en une dépendance réelle entre les mesures d'intelligence (ajustées pour tenir compte du sexe, du poids et de la taille).

Voici le tableau de données (nous avons divisé la variable **irm** par 1000 et conservé une seule décimale)

	sexe	total	verbal	perf	poids	taille	irm
1	F	133	132	124	118	64.5	816.9
2	M	139	123	150	143	73.3	1038.4
3	M	133	129	128	172	68.8	965.4
4	F	137	132	134	147	65.0	951.5
5	F	99	90	110	146	69.0	928.8
6	F	138	136	131	138	64.5	991.3
7	F	92	90	98	175	66.0	854.3
8	M	89	93	84	134	66.3	904.9
9	M	133	114	147	172	68.8	955.5
10	F	132	129	124	118	64.5	833.9
11	M	141	150	128	151	70.0	1079.5
12	M	135	129	124	155	69.0	924.1
13	F	140	120	147	155	70.5	856.5
14	F	96	100	90	146	66.0	878.9
15	F	83	71	96	135	68.0	865.4
16	F	132	132	120	127	68.5	852.2
17	M	100	96	102	178	73.5	945.1
18	F	101	112	84	136	66.3	808.0
19	M	80	77	86	180	70.0	889.1
20	M	97	107	84	186	76.5	905.9
21	F	135	129	134	122	62.0	790.6
22	M	139	145	128	132	68.0	955.0
23	F	91	86	102	114	63.0	831.8
24	M	141	145	131	171	72.0	935.5
25	F	85	90	84	140	68.0	798.6
26	M	103	96	110	187	77.0	1062.5
27	F	77	83	72	106	63.0	793.5
28	F	130	126	124	159	66.5	866.7

Matrice de corrélations :

```
> round(cor(X),3)
        total verbal  perf  poids taille   irm
total   1.000  0.938 0.920 -0.023  0.009 0.371
verbal  0.938  1.000 0.739 -0.072 -0.027 0.325
perf    0.920  0.739 1.000  0.045  0.058 0.425
poids  -0.023 -0.072 0.045  1.000  0.754 0.504
taille  0.009 -0.027 0.058  0.754  1.000 0.597
irm     0.371  0.325 0.425  0.504  0.597 1.000
```

Première question : la variable **total** — qui est une somme de plusieurs scores, *dont* **verbal** *et* **perf** — devrait-elle être exclue de l'analyse ? Les corrélations très fortes entre **verbal** et **perf**, d'une part, et **total** d'autre part, peuvent donner des résultats inexplicables dans une régression multiple avec ces trois variables. Nous décidons donc d'exclure **total** de nos analyses.

L'objectif de l'analyse étant d'établir une relation entre les résultats des tests et **irm**, s'il en existe une, l'inclusion des variables **poids** et **taille** n'a pour but que d'ajuster la mesure du cerveau de façon à tenir compte de la grandeur de la personne. Il est sage pour l'instant de les maintenir dans la régression.

```
> a<-lm(irm~verbal+perf+poids+taille)
> summary(a)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.4332   245.1016  -0.075   0.9407
verbal        0.4564     0.7606   0.600   0.5543
perf          1.0482     0.7688   1.363   0.1859
poids         0.4855     0.7618   0.637   0.5302
taille        9.9619     4.5737   2.178   0.0399 *

Residual standard error: 59.51 on 23 degrees of freedom
Multiple R-squared: 0.5231,     Adjusted R-squared: 0.4402
F-statistic: 6.308 on 4 and 23 DF,  p-value: 0.001405
```

La relation n'est pas très forte mais elle est certainement significative. Le poids ne semble pas significatif, la raison principale étant qu'elle est fortement liée à la taille. Lorsqu'on omet la taille, le poids se révèle significatif avec une valeur p de 0,009 ; lorsqu'on omet le poids, la taille se révèle encore significative avec une valeur p de 0,000446.

Il est donc clair, qu'on peut se contenter de l'une ou de l'autre. Mais il est nécessaire d'en tenir compte car sans elles, le coefficient de détermination dégringole :

```
> summary(lm(irm~verbal+perf))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 729.19024   79.00650   9.229 1.58e-09 ***
verbal        0.08287    0.94427   0.088    0.931
perf          1.44850    0.95613   1.515    0.142

Residual standard error: 74.83 on 25 degrees of freedom
Multiple R-squared: 0.1805,     Adjusted R-squared: 0.1149
F-statistic: 2.753 on 2 and 25 DF,  p-value: 0.0831
```

Nous choisissons d'inclure la taille (plutôt que le poids) étant donné qu'elle est plus fortement significative. On a alors ceci :

```
> b<-lm(irm~verbal+perf+taille)
> summary(b)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -94.0707   211.7763  -0.444 0.660879
verbal        0.3979     0.7456   0.534 0.598475
perf          1.0930     0.7560   1.446 0.161158
taille       12.1478     2.9878   4.066 0.000446 ***

Residual standard error: 58.77 on 24 degrees of freedom
Multiple R-squared: 0.5147,     Adjusted R-squared: 0.4541
F-statistic: 8.485 on 3 and 24 DF,  p-value: 0.0005104
```

Ici aussi, la taille demeure le seul effet significatif.

Nous n'avons pas tenu compte du sexe et bien sûr il faut le faire. Nous définissons la variable dichotomique **homme** qui prend la valeur 1 pour un homme et 0 pour une femme. Nous ajoutons **homme** comme variable exogène.

```
> d<-lm(irm~verbal+perf+taille+homme)
> summary(d)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 325.65764  264.85522   1.230   0.2313
verbal        0.03967    0.70181   0.057   0.9554
perf          1.21674    0.69636   1.747   0.0939 .
taille        5.95106    3.81757   1.559   0.1327
homme        67.83397   29.05285   2.335   0.0286 *

Residual standard error: 53.98 on 23 degrees of freedom
Multiple R-squared: 0.6077,     Adjusted R-squared: 0.5395
F-statistic: 8.907 on 4 and 23 DF,  p-value: 0.0001694
```

Évidemment, les variables **homme** et **taille** sont corrélées, de sorte qu'il serait raisonnable de n'en conserver qu'une seule. Si on élimine la taille, voici ce qu'on obtient :

```
> e<-lm(irm~verbal+perf+homme)
> summary(e)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 728.8594    58.6624  12.425 6.06e-12 ***
verbal       -0.2064     0.7039  -0.293 0.771903
perf          1.3641     0.7102   1.921 0.066707 .
homme        99.3196    21.4966   4.620 0.000109 ***

Residual standard error: 55.56 on 24 degrees of freedom
Multiple R-squared: 0.5663,     Adjusted R-squared: 0.512
F-statistic: 10.44 on 3 and 24 DF,  p-value: 0.0001383
```

Si on élimine **homme** on a ceci :

```
> f<-lm(irm~verbal+perf+taille)
> summary(f)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -94.0707   211.7763  -0.444 0.660879
verbal        0.3979     0.7456   0.534 0.598475
perf          1.0930     0.7560   1.446 0.161158
taille       12.1478     2.9878   4.066 0.000446 ***

Residual standard error: 58.77 on 24 degrees of freedom
Multiple R-squared: 0.5147,     Adjusted R-squared: 0.4541
F-statistic: 8.485 on 3 and 24 DF,  p-value: 0.0005104
```

Nous éliminons donc **taille** et conservons **homme**. Dans un tel modèle, **verbal** n'est nullement significatif alors que **perf** l'est presque. De plus ces deux variables sont assez fortement corrélées (r = 0,739). On élimine donc **verbal**, ce qui donne ceci :

```
> summary(lm(irm~perf+homme))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 723.1159    54.2745  13.323 7.39e-13 ***
perf          1.2114     0.4738   2.557   0.0170 *
homme        98.7591    21.0163   4.699 8.13e-05 ***

Residual standard error: 54.53 on 25 degrees of freedom
Multiple R-squared: 0.5647,     Adjusted R-squared: 0.5299
F-statistic: 16.22 on 2 and 25 DF,  p-value: 3.054e-05
```

C'est donc le modèle qui doit théoriquement être retenu. On s'abstiendra, cependant, de considérer ses conclusions (les hommes ont un plus gros cerveau et pour un sexe donné, la grosseur du cerveau est liée au score **perf**) comme étant définitives.