5.7 [Mead et Curnow (Mead, R. and Curnow, R.N. (1983) *Statistical Methods in Agriculture and Experimental Biology*, Chapman and Hall, London] présentent les données suivantes sur une expérience dans laquelle on compare quatre traitements hormonaux sur 16 bœufs. La variable *Y* est le poids du gras dans le foie (en grammes) et la covariable *X* est le poids initial du veau (en kilos). Faites une étude de l'effet des traitements hormonaux sur *Y*. La variable concomitante *X* est-elle utile dans le modèle ? Voici les données. La première donnée est *X*, la deuxième *Y*:

	Traitement				
Bloc	1	2	3	4	All
1	56	44	53	69	55.500
	133	128	129	134	131
2	47	44	51	42	46
	132	127	130	125	128.50
3	41	36	38	43	39.500
	127	127	124	126	126
4	50	46	50	54	50
	132	128	129	131	130
All	48.5	42.5	48	52	47.750
	131	127.5	128	129	128.87

Voici les données

```
> cbind(y,x,bloc,traitement)
       y  x bloc traitement
 [1,] 133 56    1          1
 [2,] 128 44    1          2
 [3,] 129 53    1          3
 [4,] 134 69    1          4
 [5,] 132 47    2          1
 [6,] 127 44    2          2
 [7,] 130 51    2          3
 [8,] 125 42    2          4
 [9,] 127 41    3          1
[10,] 127 36    3          2
[11,] 124 38    3          3
[12,] 126 43    3          4
[13,] 132 50    4          1
[14,] 128 46    4          2
[15,] 129 50    4          3
[16,] 131 54    4          4
```

Le modèle habituel pour ces données est une analyse de variance à deux facteurs additifs (sans interactions) . Le modèle s'exprime formellement comme ceci :

$$E(y) = \beta_o+\beta_1 x+\beta_2 b_2+\beta_3 b_3+\beta_4 b_4+\beta_5 t_2+\beta_6 t_3+\beta_7 t_3$$

où b_j est une variable dont la i^e composante est «1» si l'unité i est dans le bloc j; et t_j est une variable dont la i^e composante est «1» si l'unité i subit le traitement j (et 0 sinon). Le traitement de référence est le traitement 1 et le bloc de référence est le bloc 1.

Voici une analyse de variance à deux facteurs sans interactions :

```
> a<-lm(y~x+bloc+traitement)
> anova(a)
Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 89.743  89.743 59.0093 5.842e-05 ***
bloc        3  1.782   0.594  0.3907   0.76311
traitement  3 24.058   8.019  5.2729   0.02676 *
Residuals   8 12.167   1.521
```

Sans surprise, le poids initial est un facteur important. Il y a des différences entre les traitements mais pas entre les blocs. Il y a un effet de traitement significatif. Quels résultats aurions-nous obtenus si nous n'avions pas fait intervenir le poids initial ? Voici :

```
> anova(lm(y~bloc+traitement))
Response: y
```

```
           Df Sum Sq Mean Sq F value  Pr(>F)
bloc        3  56.75 18.9167  4.0296 0.04517 *
traitement  3  28.75  9.5833  2.0414 0.17860
Residuals   9  42.25  4.6944
```

Nos conclusions concernant les blocs et les traitements auraient été contraires à celles que nous avions tirées avant. Comment expliquer ce paradoxe ? Nous n'avons pas les détails sur les critères du blocage. Mais si les blocs sont liés au poids initial *x*, il est possible qu'en présence de *x* les blocs se révèlent non significatives parce qu'elles n'ajoutent pas beaucoup d'information qui ne soit déjà contenue dans *x*. Nous pouvons d'ailleurs vérifier notre hypothèse sur le lien entre les blocs et le poids initial :

```
> anova(lm(x~bloc))
Response: x
          Df Sum Sq Mean Sq F value  Pr(>F)
bloc       3    545 181.667  5.0935 0.01675 *
Residuals 12    428  35.667
```

Effectivement, il semble bien qu'en séparant les bœufs en blocs, on a dans une certaine mesure réunis dans un bloc des bœufs de poids comparables.

Il est également normal que le poids initial ne soit pas affecté par le traitement comme on le voit ci-dessous :

```
> anova(lm(x~bloc+traitement))
Response: x
           Df Sum Sq Mean Sq F value  Pr(>F)
bloc        3    545 181.667  6.7284 0.01123 *
traitement  3    185  61.667  2.2840 0.14776
Residuals   9    243  27.000
```

Le modèle traité jusqu'ici est totalement additif. Il est possible d'y ajouter certaines interactions, comme dans l'analyse suivante, qui admet des interactions entre *x* et le traitement :

```
> anova(lm(y~x+bloc+traitement+bloc*x))
Response: y
           Df Sum Sq Mean Sq  F value   Pr(>F)
x           1 89.743  89.743 128.4010 9.36e-05 ***
bloc        3  1.782   0.594   0.8501  0.52329
traitement  3 24.058   8.019  11.4736  0.01113 *
x:bloc      3  8.672   2.891   4.1359  0.08025 .
Residuals   5  3.495   0.699
```

Que signifie l'interaction `bloc*x` ? Formellement le modèle présenté ci-dessus s'écrit comme ceci

$E(y) = \beta_o+\beta_1 x+\beta_2 b_2+\beta_3 b_3+\beta_4 b_4+\beta_5 t_2+\beta_6 t_3+\beta_7 t_4+\beta_8 (bx)_2+\beta_9 (bx)_3+\beta_{10} (bx)_4$ où $(bx)_j$ est une variable dont la i^e composante est x_i si l'unité *i* est dans le bloc *j*, 0 sinon. Le modèle stipule donc que la pente de la droite de régression diffère d'un bloc à l'autre. L'hypothèse testée par la 3e ligne (`x:bloc`) est l'hypothèse que $\beta_8=\beta_9=\beta_{10} = 0$, c'est-à-dire, que la pente de la droite est la même dans les quatre blocs. Il n'est pas évident qu'on peut la rejeter. Par contre, on peut sans trop de crainte accepter l'hypothèse que les interactions `x:traitement` sont faibles, sinon nulles :

```
> anova(lm(y~x+bloc+traitement+traitement*x))
Response: y
             Df Sum Sq Mean Sq F value    Pr(>F)
x             1 89.743  89.743 73.7150 0.0003534 ***
bloc          3  1.782   0.594  0.4880 0.7055114
traitement    3 24.058   8.019  6.5870 0.0345208 *
x:traitement  3  6.079   2.026  1.6646 0.2881959
Residuals     5  6.087   1.217
```

La question qui reste est de savoir si on retient les interactions `x:bloc`. Le modèle avec interactions est un peu fragile du fait qu'il ne nous laisse que 5 degrés de liberté pour estimer la variance. Par ailleurs, le coefficient de détermination passe de 0,9 à 0,95 lorsqu'on ajoute les interactions, ce qui n'est pas énorme.

Un modèle particulièrement parcimonieux omettrait le blocs et donnerait le résultat suivant (avec une diminution négligeable du coefficient de détermination :

```
> summary(lm(y~x+traitement))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.0590     2.0198  56.965 6.02e-15 ***
x             0.3287     0.0400   8.216 5.07e-06 ***
traitement2  -1.5279     0.8296  -1.842  0.09259 .
traitement3  -2.8357     0.7943  -3.570  0.00440 **
traitement4  -3.1504     0.8063  -3.907  0.00245 **

Residual standard error: 1.123 on 11 degrees of freedom
Multiple R-squared: 0.8914,     Adjusted R-squared: 0.8519
F-statistic: 22.58 on 4 and 11 DF,  p-value: 2.936e-05
```

On peut alors décrire la situation simplement. Par exemple, l'estimation (ponctuelle et par intervalle de confiance) de la moyenne pour chaque traitement et un poids initial $x = 40$, est donnée par les commandes suivantes :

```
predict(a,data.frame(x=40,traitement=c("1","2","3","4")),interval="confidence")
       fit      lwr      upr
1 128.2062 126.7614 129.6510
2 126.6783 125.4230 127.9336
3 125.3706 123.9481 126.7930
4 125.0558 123.4299 126.6818
```