Modèles Linéaires Appliqués

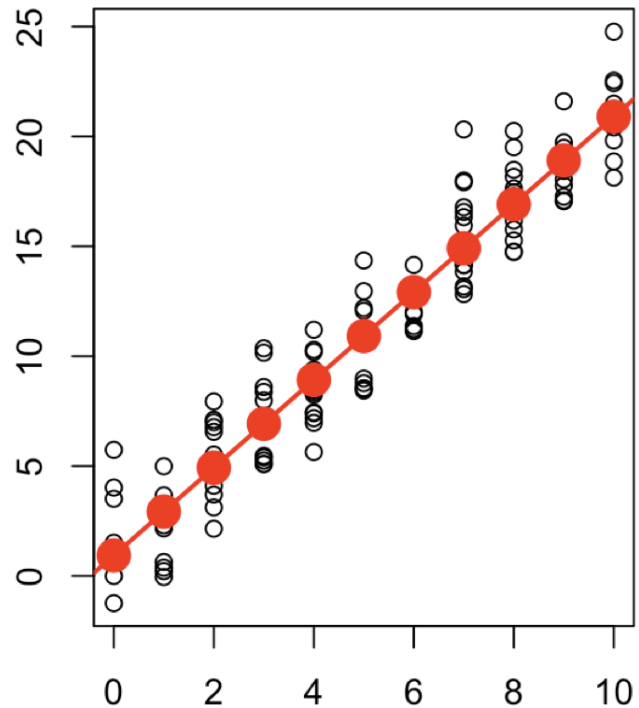
Arthur Charpentier

Automne 2Q20

OLS #24 (example)

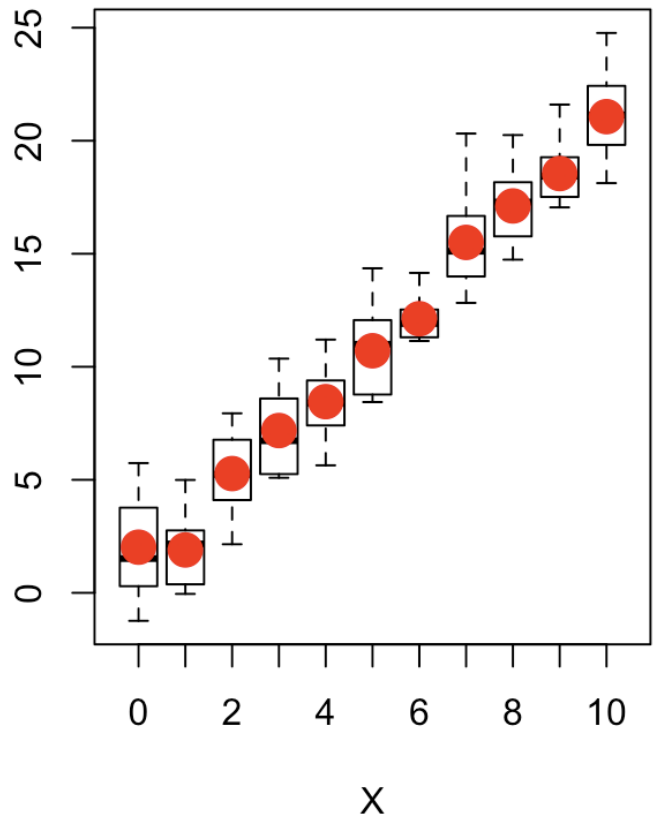
Numérique vs. Facteur

```
> n = 123
> set.seed(1)
> df = data.frame(X = sample(0:10,size=n,
  replace=TRUE))
> y = 1+2*df$X + rnorm(n)*2
> df$Y = y
> head(df)
  X         Y
1 2  3.118702
2 4  8.768349
3 6 11.370063
4 9 19.484527
5 2  2.149803
6 9 19.731882
> str(df)
'data.frame': 123 obs. of  2 variables:
 $ X: int  2 4 6 9 2 9 10 7 6 0 ...
 $ Y: num  3.12 8.77 11.37 19.48 2.15 ...
> mean(df$X)
[1] 5.186992
> plot(df)
> abline(lm(Y~X,data=df),col="red",lwd=2)
> u=seq(0,10)
> reg = lm(Y~X,data=df)
> v = predict(reg, newdata=data.frame(X=u))
> points(u,v,pch=19,cex=2,col="red")
```



Numérique vs. Facteur

```
> df$X = as.factor(df$X)
> head(df, 3)
  X         Y
1 2  3.118702
2 4  8.768349
3 6 11.370063
> str(df)
'data.frame': 123 obs. of  2 variables:
 $ X: Factor w/ 11 levels "0","1","2","3",..: 3 5 7
      10 3 10 11 8 7 1 ...
 $ Y: num  3.12 8.77 11.37 19.48 2.15 ...
> mean(X)
[1] NA
Warning message:
In mean.default(X) : argument is not numeric or
      logical: returning NA
> mean(as.numeric(df$X))
[1] 6.186992
> plot(df)
> u=seq(0,10)
> v= predict(lm(Y~X,data=df),newdata=data.frame(X=u)
      )
Error: variable 'X' was fitted with type "factor"
      but type "numeric" was supplied
> v= predict(lm(Y~X,data=df),newdata=data.frame(X=as
      .character(u)))
> points(1+u,v,pch=19,cex=2,col="red")
```



Génération de données

```
> n = 123
> df = data.frame(A = runif(n), B=runif(n), C=runif(n))
> y = 1.5 + 2*df$A + rnorm(n)/4
> df$Y = y
> head(df)
          A         B         C        Y
1 0.2655087 0.1734423 0.4287504 1.994980
2 0.3721239 0.7548209 0.1326900 2.181179
3 0.5728534 0.4538955 0.4600964 2.707884
4 0.9082078 0.5111698 0.9429571 3.744899
5 0.2016819 0.2075451 0.7619739 1.633830
6 0.8983897 0.2286581 0.9329098 3.018593
> str(df)
'data.frame': 123 obs. of  4 variables:
 $ A: num  0.266 0.372 0.573 0.908 0.202 ...
 $ B: num  0.173 0.755 0.454 0.511 0.208 ...
 $ C: num  0.429 0.133 0.46 0.943 0.762 ...
 $ Y: num  1.99 2.18 2.71 3.74 1.63 ...
```

Génération de données

```
> n = 123
> set.seed(1)
> dff = data.frame(A = cut(runif(n),breaks = c(0,.3,.8,1),labels = letters[1:3])
      , B = cut(runif(n),breaks = c(0,.5,1),labels = letters[10+1:2]), C = cut(
      runif(n),breaks = c(0,.3,1),labels = letters[20+1:2]))
> y = 1.5 + 4*(dff$A==letters[3])+2*(dff$B==letters[11])*(dff$C==letters[21]) +
      rnorm(n)/4
> dff$Y = y
> head(dff)
  A B C        Y
1 a k v 1.463962
2 b l u 1.436931
3 b k v 1.562177
4 c l v 5.928483
5 a k v 1.230466
6 c k v 5.221814
> str(dff)
'data.frame': 123 obs. of  4 variables:
 $ A: Factor w/ 3 levels "a","b","c": 1 2 2 3 1 3 3 2 2 1 ...
 $ B: Factor w/ 2 levels "k","l": 1 2 1 2 1 1 2 2 1 1 ...
 $ C: Factor w/ 2 levels "u","v": 2 1 2 2 2 2 2 2 2 1 ...
 $ Y: num  1.46 1.44 1.56 5.93 1.23 ...
```

Régression (sur des variables continues)

```
> reg = lm(Y ~ A+B+C, data=df)
> summary(reg)

Call:
lm(formula = Y ~ A + B + C, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56390 -0.27185 -0.02698  0.23174  0.54676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.39397    0.09082  15.349   <2e-16 ***
A            2.17039    0.10246  21.182   <2e-16 ***
B            0.05681    0.10349   0.549    0.584
C           -0.01479    0.09294  -0.159    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3009 on 119 degrees of freedom
Multiple R-squared:  0.7956, Adjusted R-squared:  0.7905
F-statistic: 154.4 on 3 and 119 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A, data=df))

Call:
lm(formula = Y ~ A, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55915 -0.26126 -0.01895  0.23361  0.52663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42132    0.05813   24.45   <2e-16 ***
A            2.15950    0.09967   21.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2988 on 121 degrees of freedom
Multiple R-squared:  0.7951, Adjusted R-squared:  0.7934
F-statistic: 469.4 on 1 and 121 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A+A^2, data=df))

Call:
lm(formula = Y ~ A + A^2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55915 -0.26126 -0.01895  0.23361  0.52663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42132    0.05813   24.45   <2e-16 ***
A            2.15950    0.09967   21.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2988 on 121 degrees of freedom
Multiple R-squared:  0.7951, Adjusted R-squared:  0.7934
F-statistic: 469.4 on 1 and 121 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A^2, data=df))

Call:
lm(formula = Y ~ A^2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55915 -0.26126 -0.01895  0.23361  0.52663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42132    0.05813   24.45   <2e-16 ***
A            2.15950    0.09967   21.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2988 on 121 degrees of freedom
Multiple R-squared:  0.7951, Adjusted R-squared:  0.7934
F-statistic: 469.4 on 1 and 121 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A+B, data=df))

Call:
lm(formula = Y ~ A + B, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57053 -0.26828 -0.03214  0.23511  0.55160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38856    0.08388  16.554   <2e-16 ***
A            2.16843    0.10130  21.405   <2e-16 ***
B            0.05591    0.10291   0.543    0.588
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2997 on 120 degrees of freedom
Multiple R-squared:  0.7956, Adjusted R-squared:  0.7922
F-statistic: 233.5 on 2 and 120 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A+B+A*B, data=df))

Call:
lm(formula = Y ~ A + B + A * B, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5898 -0.2747 -0.0137  0.2326  0.5212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.45914    0.12966  11.254  < 2e-16 ***
A            2.03258    0.21542   9.435 4.07e-16 ***
B           -0.08029    0.21661  -0.371    0.712
A:B          0.27439    0.38377   0.715    0.476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared:  0.7964, Adjusted R-squared:  0.7913
F-statistic: 155.2 on 3 and 119 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A+B+I(A*B), data=df))

Call:
lm(formula = Y ~ A + B + I(A * B), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.5898 -0.2747 -0.0137  0.2326  0.5212

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.45914    0.12966  11.254  < 2e-16 ***
A            2.03258    0.21542   9.435 4.07e-16 ***
B           -0.08029    0.21661  -0.371    0.712
I(A * B)     0.27439    0.38377   0.715    0.476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared:  0.7964, Adjusted R-squared:  0.7913
F-statistic: 155.2 on 3 and 119 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ A*B, data=df))

Call:
lm(formula = Y ~ A * B, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5898 -0.2747 -0.0137  0.2326  0.5212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.45914    0.12966  11.254  < 2e-16 ***
A            2.03258    0.21542   9.435 4.07e-16 ***
B           -0.08029    0.21661  -0.371    0.712
A:B          0.27439    0.38377   0.715    0.476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared:  0.7964, Adjusted R-squared:  0.7913
F-statistic: 155.2 on 3 and 119 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> summary(lm(Y ~ poly(cbind(A,B),2), data=df))

Call:
lm(formula = Y ~ poly(cbind(A, B), 2), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55514 -0.22850 -0.03089  0.24460  0.53581

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.53999    0.02752  92.280   <2e-16 ***
poly(cbind(A, B), 2)1.0    6.50765    0.30583  21.278   <2e-16 ***
poly(cbind(A, B), 2)2.0   -0.36416    0.30350  -1.200    0.233
poly(cbind(A, B), 2)0.1    0.16903    0.30620   0.552    0.582
poly(cbind(A, B), 2)1.1    1.99218    3.47584   0.573    0.568
poly(cbind(A, B), 2)0.2    0.03692    0.30560   0.121    0.904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.301 on 117 degrees of freedom
Multiple R-squared:  0.799, Adjusted R-squared:  0.7904
F-statistic: 92.99 on 5 and 117 DF,  p-value: < 2.2e-16
```

Régression (sur des variables continues)

```
> data.frame(df,model.matrix(lm(Y ~ A+B+A*B, data=df)))
          A      B      C      Y X.Intercept.    A.1    B.1    A.B
1    0.2655 0.1734 0.4288 1.9950             1 0.2655 0.1734 0.0461
2    0.3721 0.7548 0.1327 2.1812             1 0.3721 0.7548 0.2809
3    0.5729 0.4539 0.4601 2.7079             1 0.5729 0.4539 0.2600
4    0.9082 0.5112 0.9430 3.7449             1 0.9082 0.5112 0.4642
5    0.2017 0.2075 0.7620 1.6338             1 0.2017 0.2075 0.0419
6    0.8984 0.2287 0.9329 3.0186             1 0.8984 0.2287 0.2054
7    0.9447 0.5957 0.4707 3.3096             1 0.9447 0.5957 0.5628
8    0.6608 0.5749 0.6036 2.6551             1 0.6608 0.5749 0.3799
9    0.6291 0.0771 0.4850 3.1230             1 0.6291 0.0771 0.0485
10   0.0618 0.0355 0.1088 1.3008             1 0.0618 0.0355 0.0022
11   0.2060 0.6428 0.2477 1.9053             1 0.2060 0.6428 0.1324
12   0.1766 0.9286 0.4985 1.7828             1 0.1766 0.9286 0.1640
13   0.6870 0.5981 0.3729 2.9383             1 0.6870 0.5981 0.4109
```

Régression (sur des variables continues)

```
> summary(aov(Y ~ A*B, data=df))
             Df  Sum Sq  Mean Sq F value  Pr(>F)
A             1   41.92   41.92 464.790 <2e-16 ***
B             1    0.03    0.03   0.294  0.589
A:B           1    0.05    0.05   0.511  0.476
Residuals   119   10.73    0.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(Y ~ A*B*C, data=df))
             Df  Sum Sq  Mean Sq F value  Pr(>F)
A             1   41.92   41.92 452.547 <2e-16 ***
B             1    0.03    0.03   0.286  0.594
C             1    0.00    0.00   0.025  0.875
A:B           1    0.04    0.04   0.485  0.488
A:C           1    0.05    0.05   0.578  0.449
B:C           1    0.00    0.00   0.037  0.849
A:B:C         1    0.02    0.02   0.239  0.626
Residuals   115   10.65    0.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Régression (sur des variables factorielles)

```
> reg = lm(Y ~ A+B+C, data=dff)
> summary(reg)

Call:
lm(formula = Y ~ A + B + C, data = dff)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08387 -0.45257  0.05352  0.43494  1.13125

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.855847   0.130754  21.841  < 2e-16 ***
Ab          -0.001228   0.122478  -0.010    0.992
Ac           4.021432   0.160949  24.986  < 2e-16 ***
Bl          -0.666923   0.102213  -6.525 1.78e-09 ***
Cv          -0.965636   0.106048  -9.106 2.60e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5608 on 118 degrees of freedom
Multiple R-squared:  0.8956, Adjusted R-squared:  0.8921
F-statistic: 253.1 on 4 and 118 DF,  p-value: < 2.2e-16
```

Régression (sur des variables factorielles)

```
> summary(lm(Y ~ A, data=dff))

Call:
lm(formula = Y ~ A, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8931 -0.5333 -0.2455  0.0471  2.0934

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.90850    0.14989  12.732   <2e-16 ***
Ab          -0.01354    0.17841  -0.076     0.94
Ac           3.96146    0.23359  16.959   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.821 on 120 degrees of freedom
Multiple R-squared:  0.7725, Adjusted R-squared:  0.7687
F-statistic: 203.7 on 2 and 120 DF,  p-value: < 2.2e-16
```

Régression (sur des variables factorielles)

```
> summary(lm(Y ~ A+A^2, data=dff))

Call:
lm(formula = Y ~ A + A^2, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8931 -0.5333 -0.2455  0.0471  2.0934

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.90850    0.14989  12.732   <2e-16 ***
Ab          -0.01354    0.17841  -0.076     0.94
Ac           3.96146    0.23359  16.959   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.821 on 120 degrees of freedom
Multiple R-squared:  0.7725, Adjusted R-squared:  0.7687
F-statistic: 203.7 on 2 and 120 DF,  p-value: < 2.2e-16

> table(dff$A,dff$A)

    a  b  c
  a 30  0  0
  b  0 72  0
  c  0  0 21
```

Régression (sur des variables factorielles)

```
> summary(lm(Y ~ A+B, data=dff))

Call:
lm(formula = Y ~ A + B, data = dff)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26417 -0.45817 -0.08168  0.35256  1.73016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.36613    0.15486  15.279  < 2e-16 ***
Ab          -0.07922    0.15875  -0.499    0.619
Ac           3.86702    0.20797  18.594  < 2e-16 ***
Bl          -0.76272    0.13211  -5.774 6.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7287 on 119 degrees of freedom
Multiple R-squared:  0.8223, Adjusted R-squared:  0.8178
F-statistic: 183.5 on 3 and 119 DF,  p-value: < 2.2e-16

> table(dff$A,dff$B)

    k  l
  a 12 18
  b 35 37
  c 11 10
```

Régression (sur des variables factorielles)

```
> summary(lm(Y ~ A+B+A*B, data=dff))

Call:
lm(formula = Y ~ A + B + A * B, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3222 -0.3949 -0.1018  0.3826  1.8041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5527     0.2106  12.123  < 2e-16 ***
Ab           -0.3065     0.2440  -1.256 0.211554
Ac            3.6065     0.3045  11.845  < 2e-16 ***
Bl           -1.0737     0.2718  -3.950 0.000134 ***
Ab:Bl         0.3902     0.3217   1.213 0.227565
Ac:Bl         0.4662     0.4189   1.113 0.267986
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7294 on 117 degrees of freedom
Multiple R-squared:  0.8249, Adjusted R-squared:  0.8174
F-statistic: 110.2 on 5 and 117 DF,  p-value: < 2.2e-16
```

Régression (sur des variables factorielles)

```
> summary(lm(Y ~ A*B, data=dff))

Call:
lm(formula = Y ~ A * B, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3222 -0.3949 -0.1018  0.3826  1.8041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5527     0.2106  12.123  < 2e-16 ***
Ab           -0.3065     0.2440  -1.256 0.211554
Ac            3.6065     0.3045  11.845  < 2e-16 ***
Bl           -1.0737     0.2718  -3.950 0.000134 ***
Ab:Bl         0.3902     0.3217   1.213 0.227565
Ac:Bl         0.4662     0.4189   1.113 0.267986
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7294 on 117 degrees of freedom
Multiple R-squared:  0.8249, Adjusted R-squared:  0.8174
F-statistic: 110.2 on 5 and 117 DF,  p-value: < 2.2e-16
```

Régression (sur des variables factorielles)

```
> data.frame(dff,model.matrix(lm(Y ~ A+B+A*B, data=dff)))
    A B C        Y X.Intercept. Ab Ac Bl Ab.Bl Ac.Bl
1   a k v 1.463962             1  0  0  0     0     0
2   b l u 1.436931             1  1  0  1     1     0
3   b k v 1.562177             1  1  0  0     0     0
4   c l v 5.928483             1  0  1  1     0     1
5   a k v 1.230466             1  0  0  0     0     0
6   c k v 5.221814             1  0  1  0     0     0
7   c l v 5.420216             1  0  1  1     0     1
8   b l v 1.333521             1  1  0  1     1     0
9   b k v 1.864808             1  1  0  0     0     0
10  a k u 3.177195             1  0  0  0     0     0
11  a l u 1.493319             1  0  0  1     0     0
12  a l v 1.429713             1  0  0  1     0     0
13  b l v 1.564264             1  1  0  1     1     0
14  b l v 1.656162             1  1  0  1     1     0
15  b l v 1.978554             1  1  0  1     1     0
16  b l v 1.232161             1  1  0  1     1     0
17  b l u 1.240812             1  1  0  1     1     0

> summary(aov(Y ~ A*B, data=dff))
             Df Sum Sq Mean Sq F value   Pr(>F)
A             2 274.61  137.31 258.074  < 2e-16 ***
B             1  17.70   17.70  33.266 6.68e-08 ***
A:B           2   0.94    0.47   0.879    0.418
Residuals   117  62.25    0.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Régression (sur des variables factorielles)

```
> summary(aov(Y ~ B*C, data=dff))
             Df Sum Sq Mean Sq F value   Pr(>F)
B             1  24.74   24.74  10.343 0.001674 **
C             1  14.35   14.35   5.997 0.015784 *
B:C           1  31.73   31.73  13.263 0.000402 ***
Residuals   119 284.68    2.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm(Y ~ B*C, data=dff))

Call:
lm(formula = Y ~ B * C, data = dff)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2937 -0.8984 -0.5776 -0.2778  3.9573

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0933     0.3157  12.965  < 2e-16 ***
Bl           -2.1712     0.4622  -4.698 7.11e-06 ***
Cv           -1.7768     0.4124  -4.309 3.40e-05 ***
Bl:Cv         2.1183     0.5817   3.642 0.000402 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.547 on 119 degrees of freedom
Multiple R-squared:  0.1992, Adjusted R-squared:  0.179
F-statistic: 9.868 on 3 and 119 DF,  p-value: 7.332e-06
```