devoir MAN 2016

sarah FELDMAN

21 octobre 2016

# 

# ATTENTION SUPPRIMER TOUTES LES LIGNES DE CODE

# Q1/Dans un premier temps décrivez vos variables.

## Nombre de variables

dim(rt)

## [1] 315 14

J'ai 14 variables.

## Type de variables

## age sexe tabac bmi   
## Min. :19.00 Min. :1.000 Min. :1.000 Min. :16.33   
## 1st Qu.:39.00 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:21.80   
## Median :48.00 Median :2.000 Median :2.000 Median :24.74   
## Mean :50.15 Mean :1.867 Mean :1.638 Mean :26.16   
## 3rd Qu.:62.50 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:28.85   
## Max. :83.00 Max. :2.000 Max. :3.000 Max. :50.40   
## vitamine calories graisses fibres   
## Min. :1.000 Min. : 445.2 Min. : 14.40 Min. : 3.10   
## 1st Qu.:1.000 1st Qu.:1338.0 1st Qu.: 53.95 1st Qu.: 9.15   
## Median :2.000 Median :1666.8 Median : 72.90 Median :12.10   
## Mean :1.965 Mean :1796.7 Mean : 77.03 Mean :12.79   
## 3rd Qu.:3.000 3rd Qu.:2100.4 3rd Qu.: 95.25 3rd Qu.:15.60   
## Max. :3.000 Max. :6662.2 Max. :235.90 Max. :36.80   
## alcool cholesterol betadiet retdiet   
## Min. : 0.000 Min. : 37.7 Min. : 214 Min. : 30.0   
## 1st Qu.: 0.000 1st Qu.:155.0 1st Qu.:1116 1st Qu.: 480.0   
## Median : 0.300 Median :206.3 Median :1802 Median : 707.0   
## Mean : 3.279 Mean :242.5 Mean :2186 Mean : 832.7   
## 3rd Qu.: 3.200 3rd Qu.:308.9 3rd Qu.:2836 3rd Qu.:1037.0   
## Max. :203.000 Max. :900.7 Max. :9642 Max. :6901.0   
## betaplasma retplasma   
## Min. : 0.0 Min. : 179.0   
## 1st Qu.: 90.0 1st Qu.: 466.0   
## Median : 140.0 Median : 566.0   
## Mean : 189.9 Mean : 602.8   
## 3rd Qu.: 230.0 3rd Qu.: 716.0   
## Max. :1415.0 Max. :1727.0

Les variables sont reconnues comme quantitatives mais en regardant plus attentivement les min et max on peut voir que sexe a un minimum de 1 et un maximum de 2 c'est donc une variables qualitative binaire. Je la recode en 0(homme)/1(femme) pour une interprétation plus facile par la suite. Les autres variables sont quantitatives: soit continues : age, bmi, calories, graisses, fibres, alcool, cholestérol, betadiet,retdiet,betaplasma,retplasma soit discètes : tabac, vitamine (Je recode vitamine : 1=non, 2=oui pas souvent, 3= oui souvent) NB: tabac est limite entre qualitatif à 3 classes et quantitatif...

|  |  |  |
| --- | --- | --- |
| variables.quantitatives.continues | variables.quantitatives.discretes | variable.binaire |
| age | tabac | sexe |
| bmi | vitamine |  |
| calories |  |  |
| graisses |  |  |
| fibres |  |  |
| alcool |  |  |
| cholestérol |  |  |
| betadiet |  |  |
| retdiet |  |  |
| betaplasma |  |  |
| retplasma |  |  |

## Données manquantes

summary ne montre pas de données manquantes mais je préfère m'en assurer en demandant spécifiquement de sortir les NA:

## variables NA.  
## 1 age 0  
## 2 sexe 0  
## 3 tabac 0  
## 4 bmi 0  
## 5 vitamine 0  
## 6 calories 0  
## 7 graisses 0  
## 8 fibres 0  
## 9 alcool 0  
## 10 cholesterol 0  
## 11 betadiet 0  
## 12 retdiet 0  
## 13 betaplasma 0  
## 14 retplasma 0

# Données abberrantes

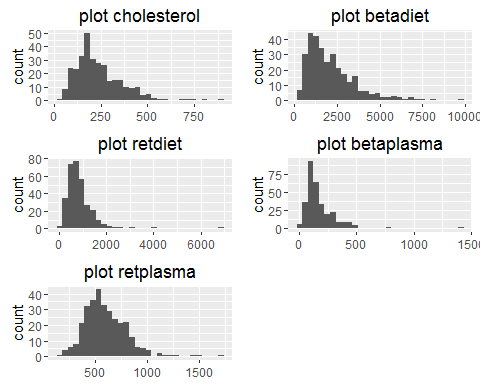
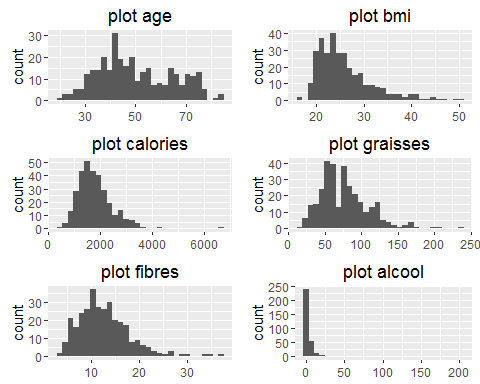
Observer ses variables sert également à dépister les "bizarreries". Par exemple, j'ai un doute concernant le max d'alcool, je vais donc regarder plus attentivement la variable, par exemple avec une table

table(rt$alcool)

##   
## 0 0.1 0.2 0.3 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5   
## 111 25 19 6 10 3 8 2 5 12 1 4 3 1 3   
## 1.7 2 2.1 2.2 2.3 2.4 2.5 2.6 2.9 3 3.1 3.2 3.4 3.5 3.7   
## 2 6 1 2 1 1 1 1 1 4 1 4 1 1 2   
## 4 4.1 4.2 4.5 4.7 4.9 5 5.2 5.6 5.7 6.1 6.2 6.4 6.5 6.7   
## 3 3 1 2 1 1 5 1 1 1 1 2 1 2 1   
## 6.8 7 7.1 7.2 7.3 8 8.3 8.4 8.5 9 10 10.5 11 14 14.1   
## 1 5 2 5 1 3 1 1 1 1 2 2 3 2 3   
## 14.2 15 15.5 17 18 18.2 20 21 22 35 203   
## 1 2 1 1 2 1 1 1 1 2 1

en effet, je passe de 35 à 203 verres...203/7=29 verres par jour, il y a probablement un erreur de codage

## Distribution des variables



## [[1]]  
## NULL  
##   
## [[2]]  
## NULL

|  |  |
| --- | --- |
| variables.normales | variables.non.normales |
| retplasma | age |
| bmi | sexe |
| cholesterol | tabac |
| retdiet | vitamine |
| graisses | alcool |
| betadiet |  |
| betaplasma |  |
| calories |  |
| fibres |  |

-bmi, graisses, cholestérol, betadiet, retdiet et betaplasme on des distribution à peu près normales mais asymétriques. -calories, fibres, ret plasma ont de distributions d'allure normales -age a une distribution irrégulière avec 2 cloches à 40 et 75 ans. -alcool a une distribution très asymétrique qui ne semble pas normale

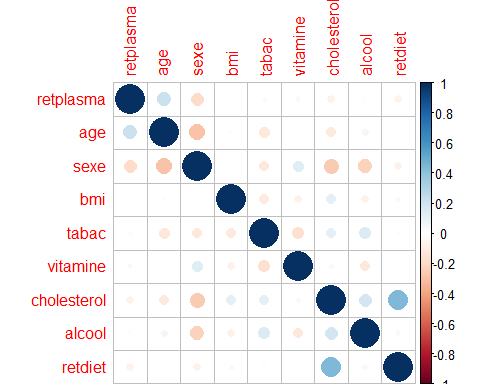
# Q2/Etudiez les relations existant entre toutes les paires possibles de variables.

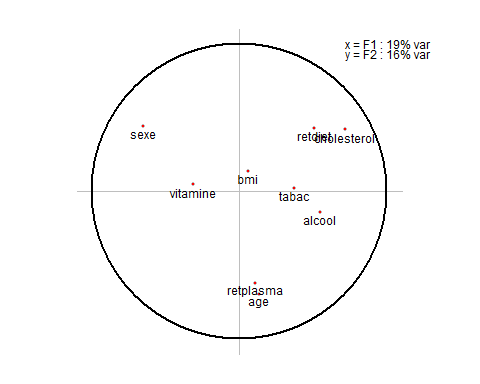
NB : ce ne sont que les 9 variables qui sont concernées

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | retplasma | age | sexe | bmi | tabac | vitamine | cholesterol | alcool | retdiet |
| retplasma | 1.000 | 0.212 | -0.184 | 0.013 | 0.029 | 0.037 | -0.070 | 0.017 | -0.063 |
| age | 0.212 | 1.000 | -0.280 | -0.017 | -0.126 | 0.012 | -0.114 | 0.052 | -0.010 |
| sexe | -0.184 | -0.280 | 1.000 | -0.007 | -0.121 | 0.135 | -0.255 | -0.228 | -0.074 |
| bmi | 0.013 | -0.017 | -0.007 | 1.000 | -0.113 | -0.074 | 0.110 | -0.073 | 0.032 |
| tabac | 0.029 | -0.126 | -0.121 | -0.113 | 1.000 | -0.167 | 0.119 | 0.160 | -0.020 |
| vitamine | 0.037 | 0.012 | 0.135 | -0.074 | -0.167 | 1.000 | -0.033 | -0.121 | 0.012 |
| cholesterol | -0.070 | -0.114 | -0.255 | 0.110 | 0.119 | -0.033 | 1.000 | 0.182 | 0.443 |
| alcool | 0.017 | 0.052 | -0.228 | -0.073 | 0.160 | -0.121 | 0.182 | 1.000 | 0.045 |
| retdiet | -0.063 | -0.010 | -0.074 | 0.032 | -0.020 | 0.012 | 0.443 | 0.045 | 1.000 |

Je peux faire une matrice de corrélation. Aucune condition n'est nécessaire pour faire des coefficients de corrélation. C'est pour les tester que nous avons besoin de vérifier les confidions de validité

scatter plot? Il faut ensuite interprêter la matrice. Pour cela je peux faire des schéma





bmi, tabac et vitamine sont très proches du centre du cercle donc non interprétables Par contre on peut voir que les paires retdiet-cholestérol, retplasma-age sont fortement associés (corrélation positive), et ces deux groupes de variables sont indépendants l'un de l'autre car forme un angle droit avec le centre. sexe est également indépendant de retdiet et cholestérol. Alcool-sexe est négativement corrélée

Je peux aussi sélectionner dans la matrice les valeurs absolues supérieures ou égales à certain niveau de corrélation

## [[1]]  
## var1 var2 r  
## 1 age retplasma 0.212  
## 2 sexe age -0.280  
## 3 cholesterol sexe -0.255  
## 4 alcool sexe -0.228  
## 5 retdiet cholesterol 0.443  
##   
## [[2]]  
## var1 var2 r  
## 1 retdiet cholesterol 0.443

Est-ce pertinent de faire de tests de corrélation pour chaque variable? mat\_test <-round(cor.test(rt[,var]),3) #ne marche pas

Attention aux conditions de validité: une des 2 variables du couple testé doit être normale

## var1 var2 r testvalid test  
## 1 age retplasma 0.212 TRUE 0.00015  
## 2 sexe age -0.280 FALSE 0.00000  
## 3 cholesterol sexe -0.255 TRUE 0.00000  
## 4 alcool sexe -0.228 FALSE 0.00005  
## 5 retdiet cholesterol 0.443 TRUE 0.00000

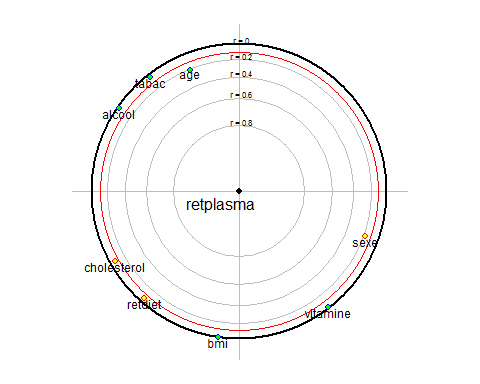
Pour r>=0.2, 5 couples ressortent: age-retplasma, sexe\_age, cholesterol-sexe, alcool-sexe,retdiet-cholestérol

Il n'y a qu'un couple qui ressort pour r>=0.4 : retdiet-cholestérol

Faut il faire des tests?

# Q3/Effectuez ensuite une régression linéaire où la variable à expliquer sera la concentration en rétinol plasmatique, les autres variables étant explicatives. Recherchez des interactions entre les variables explicatives.

## ACP focalisée

Je peux commencer par regarder les intéraction enter retplasma et les variables explicatives avec une acp focalisée:  La seule variable qui semble significativement liée à retplasma est âge. Et en effet c'est un couple qui ressortait dans la matrice de corrélation avec un r=0.212 significativement différent de 0.

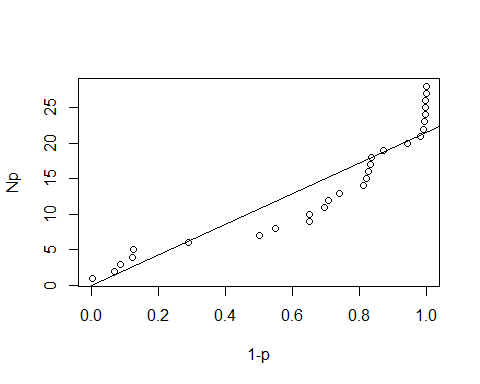
## régression linéaire multiple

La variable à expliquer retplasma étant une variable Quantitative, je peux faire une régression linéaire. Et dans la mesure où j'introduis plusieurs variables explicatives, ce sera une régression linéaire multiple.

##   
## Call:  
## lm(formula = my\_formula, data = rt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -472.46 -123.23 -27.77 110.06 1019.45   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 527.85168 100.92343 5.230 3.15e-07 \*\*\*  
## age 2.32277 0.84875 2.737 0.00657 \*\*   
## sexe -102.73716 37.90234 -2.711 0.00710 \*\*   
## bmi 1.23654 1.95205 0.633 0.52691   
## tabac 16.20647 16.99551 0.954 0.34105   
## vitamine 15.95859 13.72995 1.162 0.24601   
## cholesterol -0.13406 0.10437 -1.284 0.19994   
## alcool -0.17916 0.97944 -0.183 0.85498   
## retdiet -0.01294 0.02185 -0.592 0.55411   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 203.2 on 306 degrees of freedom  
## Multiple R-squared: 0.07769, Adjusted R-squared: 0.05358   
## F-statistic: 3.222 on 8 and 306 DF, p-value: 0.001561

## Recherche d'intéractions entre les variables explicatives:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Df | Sum.of.Sq | RSS | AIC | Pr..Chi. | signif |
| vitamine:alcool | 1 | 6.108792e+05 | 12026667 | 3343.270 | 0.0001 | \* |
| cholesterol:alcool | 1 | 5.852187e+05 | 12052327 | 3343.942 | 0.0001 | \* |
| bmi:alcool | 1 | 4.083928e+05 | 12229153 | 3348.530 | 0.0013 | \* |
| age:alcool | 1 | 3.972297e+05 | 12240316 | 3348.817 | 0.0015 | \* |
| alcool:retdiet | 1 | 3.302681e+05 | 12307278 | 3350.536 | 0.0039 | \* |
| tabac:alcool | 1 | 2.993610e+05 | 12338185 | 3351.326 | 0.0060 | \* |
| sexe:alcool | 1 | 2.669188e+05 | 12370627 | 3352.153 | 0.0095 | \* |
| sexe:bmi | 1 | 2.250680e+05 | 12412478 | 3353.217 | 0.0174 | \* |
| sexe:cholesterol | 1 | 1.445861e+05 | 12492960 | 3355.253 | 0.0569 |  |
| sexe:retdiet | 1 | 9.230802e+04 | 12545238 | 3356.568 | 0.1286 |  |
| cholesterol:retdiet | 1 | 7.755493e+04 | 12559991 | 3356.938 | 0.1638 |  |
| bmi:cholesterol | 1 | 7.708315e+04 | 12560463 | 3356.950 | 0.1651 |  |
| bmi:tabac | 1 | 7.464118e+04 | 12562905 | 3357.011 | 0.1719 |  |
| age:bmi | 1 | 7.262306e+04 | 12564923 | 3357.062 | 0.1779 |  |
| tabac:vitamine | 1 | 6.984812e+04 | 12567698 | 3357.131 | 0.1864 |  |
| vitamine:cholesterol | 1 | 5.105495e+04 | 12586491 | 3357.602 | 0.2588 |  |
| age:retdiet | 1 | 4.427747e+04 | 12593268 | 3357.772 | 0.2930 |  |
| age:sexe | 1 | 4.221337e+04 | 12595333 | 3357.823 | 0.3046 |  |
| vitamine:retdiet | 1 | 3.520266e+04 | 12602343 | 3357.999 | 0.3486 |  |
| sexe:tabac | 1 | 3.511267e+04 | 12602433 | 3358.001 | 0.3492 |  |
| tabac:cholesterol | 1 | 2.281310e+04 | 12614733 | 3358.308 | 0.4506 |  |
| tabac:retdiet | 1 | 1.839588e+04 | 12619150 | 3358.418 | 0.4982 |  |
| age:vitamine | 1 | 5.562105e+03 | 12631984 | 3358.739 | 0.7096 |  |
| age:cholesterol | 1 | 1.012963e+03 | 12636533 | 3358.852 | 0.8737 |  |
| bmi:vitamine | 1 | 9.596472e+02 | 12636586 | 3358.853 | 0.8771 |  |
| age:tabac | 1 | 5.018884e+02 | 12637044 | 3358.865 | 0.9109 |  |
| sexe:vitamine | 1 | 3.025970e+02 | 12637243 | 3358.870 | 0.9308 |  |
| bmi:retdiet | 1 | 5.761194e-01 | 12637545 | 3358.877 | 0.9970 |  |
|  | NA | NA | 12637546 | 3356.877 | NA | NA |

Je trace la représentation graphique des pvalue des 28 intéractions possibles 

# Q4/Transformez la variable "rétinol plasmatique" en une variable binaire (en la coupant en deux au niveau de la médiane). Refaites les calculs précédents en ayant recours cette fois à une régression logistique.

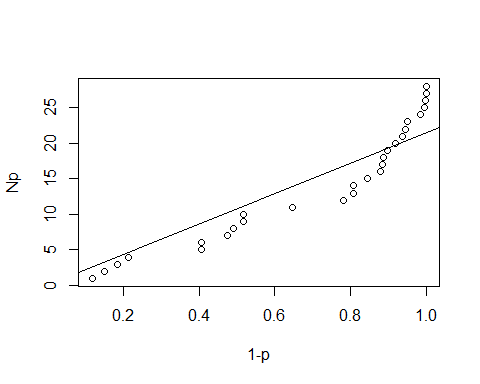
## régression logistique : construction du modèle

La variable à expliquer est binaire, il faut donc faire une régression logistique

##   
## Call:  
## glm(formula = retplasma.bin ~ age + sexe + bmi + tabac + vitamine +   
## cholesterol + alcool + retdiet, family = "binomial", data = rt)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7798 -1.1104 0.6976 1.1460 1.5373   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.757e+00 1.026e+00 -1.712 0.08681 .   
## age 2.811e-02 8.726e-03 3.221 0.00128 \*\*  
## sexe -2.913e-01 3.878e-01 -0.751 0.45255   
## bmi 1.066e-02 1.971e-02 0.541 0.58851   
## tabac 8.030e-02 1.712e-01 0.469 0.63911   
## vitamine 1.430e-01 1.389e-01 1.030 0.30318   
## cholesterol -3.050e-04 1.055e-03 -0.289 0.77247   
## alcool 3.478e-05 9.930e-03 0.004 0.99721   
## retdiet -2.132e-05 2.209e-04 -0.096 0.92314   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 436.68 on 314 degrees of freedom  
## Residual deviance: 420.87 on 306 degrees of freedom  
## AIC: 438.87  
##   
## Number of Fisher Scoring iterations: 4

## Recherche d'intéractions entre les variables explicatives:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Df | Deviance | AIC | LRT | Pr..Chi. | signif |
| age:alcool | 1 | 408.8005 | 428.8005 | 12.0743187 | 0.0005 | \* |
| vitamine:alcool | 1 | 409.6070 | 429.6070 | 11.2678189 | 0.0008 | \* |
| sexe:alcool | 1 | 412.4177 | 432.4177 | 8.4570858 | 0.0036 | \* |
| cholesterol:alcool | 1 | 413.0898 | 433.0898 | 7.7850366 | 0.0053 | \* |
| tabac:alcool | 1 | 414.9763 | 434.9763 | 5.8985620 | 0.0152 | \* |
| sexe:cholesterol | 1 | 417.0602 | 437.0602 | 3.8146637 | 0.0508 |  |
| tabac:vitamine | 1 | 417.1975 | 437.1975 | 3.6773574 | 0.0552 |  |
| alcool:retdiet | 1 | 417.3985 | 437.3985 | 3.4763050 | 0.0623 |  |
| cholesterol:retdiet | 1 | 417.8293 | 437.8293 | 3.0454871 | 0.0810 |  |
| bmi:alcool | 1 | 418.2299 | 438.2299 | 2.6448815 | 0.1039 |  |
| vitamine:retdiet | 1 | 418.3752 | 438.3752 | 2.4996307 | 0.1139 |  |
| age:sexe | 1 | 418.4170 | 438.4170 | 2.4578169 | 0.1169 |  |
| tabac:retdiet | 1 | 418.4787 | 438.4787 | 2.3961750 | 0.1216 |  |
| sexe:retdiet | 1 | 418.8645 | 438.8645 | 2.0103659 | 0.1562 |  |
| age:vitamine | 1 | 419.1755 | 439.1755 | 1.6993472 | 0.1924 |  |
| age:bmi | 1 | 419.1770 | 439.1770 | 1.6978068 | 0.1926 |  |
| sexe:bmi | 1 | 419.3705 | 439.3705 | 1.5043174 | 0.2200 |  |
| bmi:vitamine | 1 | 420.0103 | 440.0103 | 0.8645717 | 0.3525 |  |
| age:cholesterol | 1 | 420.3824 | 440.3824 | 0.4924505 | 0.4828 |  |
| sexe:tabac | 1 | 420.3836 | 440.3836 | 0.4912630 | 0.4834 |  |
| vitamine:cholesterol | 1 | 420.4377 | 440.4377 | 0.4371170 | 0.5085 |  |
| tabac:cholesterol | 1 | 420.4721 | 440.4721 | 0.4027624 | 0.5257 |  |
| bmi:retdiet | 1 | 420.5895 | 440.5895 | 0.2852959 | 0.5933 |  |
| age:retdiet | 1 | 420.5918 | 440.5918 | 0.2830402 | 0.5947 |  |
| sexe:vitamine | 1 | 420.8011 | 440.8011 | 0.0737779 | 0.7859 |  |
| age:tabac | 1 | 420.8202 | 440.8202 | 0.0546390 | 0.8152 |  |
| bmi:tabac | 1 | 420.8385 | 440.8385 | 0.0363025 | 0.8489 |  |
| bmi:cholesterol | 1 | 420.8531 | 440.8531 | 0.0217518 | 0.8827 |  |
|  | NA | 420.8748 | 438.8748 | NA | NA | NA |

Je trace la représentation graphique des pvalue des 28 intéractions possibles