Rapport CYS

14 mars, 2020

Contents

Etude des données

Les données

```
CYS3 = read.csv("Semestre3_Complet.csv")
summary(CYS3)
```

```
##        prénom          Semestre   Filière     snapshot.1    snapshot.2
##  ALEXANDRE:  4   S3 2017-18: 54   EEA:242          : 30   Min.   : 1.75
##  ALEXIS   :  3   S3 2018-19:188             5.5    : 11   1st Qu.: 7.00
##  GUILLAUME:  3                              7.75   : 11   Median : 8.75
##  HUGO     :  3                              3.5    : 10   Mean   : 8.84
##  LUCAS    :  3                              7      : 10   3rd Qu.:10.75
##  NICOLAS  :  3                              7.5    : 10   Max.   :15.75
##  (Other)  :223                              (Other):160   NA's   :27
##  Snapshot.2...4m CYS.S3    CYS.S4    TP.S3    TP.S4     CMI        Groupe.S3
##         :169     ?  :  1      :185     : 23     :192      : 19   Siuban :59
##  -      : 21     non:165   non: 28   FR:155   FR: 21   ?  :  2   Akane  :44
##  11     :  4     oui: 76   oui: 29   GB: 64   GB: 29   non:184   Steven :40
##  11.75  :  4                                          oui: 37   Alba   :38
##  14.5   :  4                                                    Nadia  :29
##  6.75   :  4                                                    Yolanda:20
##  (Other): 36                                                    (Other):12
##     Groupe.S4     Prof.TP
##          :210         :220
##  Nadia   : 17   Didier: 10
##  Virginia: 15   Pierre: 12
##
##
##
```

```
##
```

```
N=242
table(CYS3$CYS.S3)/N*100
```

```
##
##          ?        non        oui
##  0.4132231 68.1818182 31.4049587
```

```
#pie(table(CYS3$CYS.S3))
```

Commentaire:

On trouve que presque un tier des étudiants utilisent l'outil CheckYourSmile en semestre 3.

```
table(CYS3$CMI)/N*100
```

```
##
##                     ?        non        oui
##  7.8512397  0.8264463 76.0330579 15.2892562
```

```
#pie(table(CYS3$CMI))
```

Commentaire:

On trouve que presque 15% des étudiants sont CMI en semestre 3. Or, presque 9% des étudiants ne donnent pas de réponse.

```
table(CYS3$TP.S3)/N*100
```

```
##
##                  FR        GB
##  9.504132 64.049587 26.446281
```

```
#pie(table(CYS3$TP.S3))
```

Commentaire:

On trouve que presque 64% des étudiants pratiquent les TP en français en semestre 3 et 26% pratiquent les TP en anglais. Or, presque 10% des étudiants ne donnent pas de réponse.

Commentaire:

On trouve que presque un tier

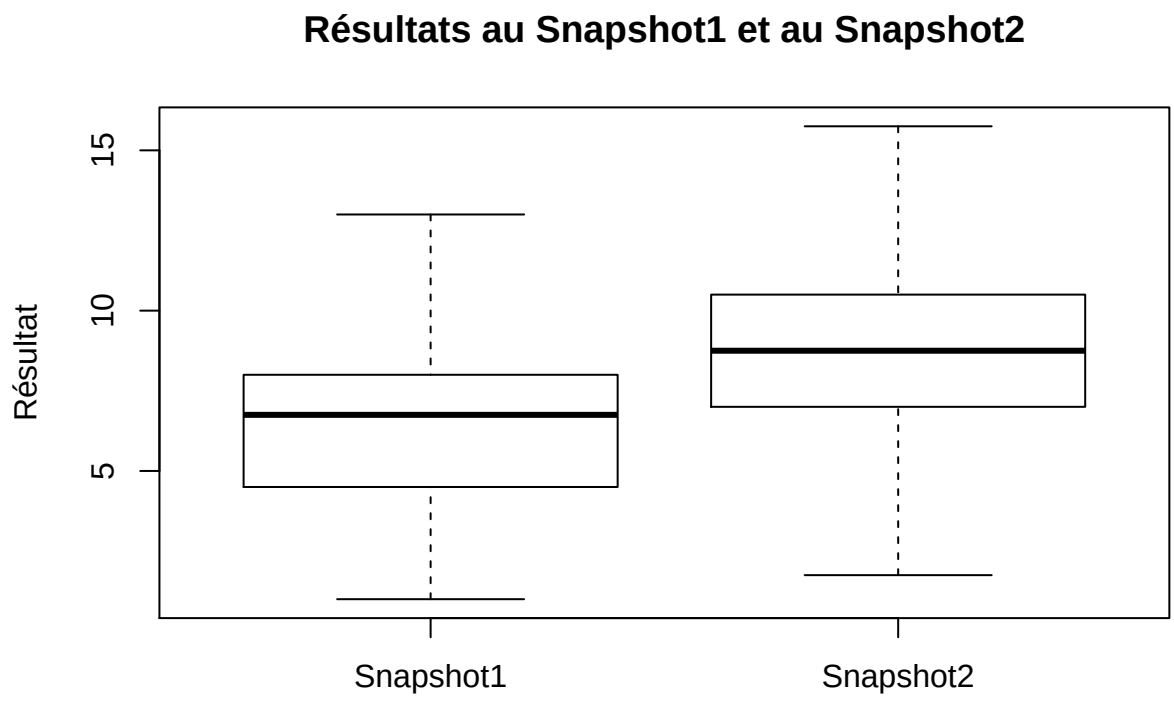
```
library(gmodels)
CrossTable(CYS3$CYS.S3,CYS3$CMI)
CrossTable(CYS3$CYS.S3,CYS3$TP.S3)
CrossTable(CYS3$CMI,CYS3$TP.S3)
```

```
CYS = read.csv("DonneeS3_filtre.csv")
summary(CYS)
```

```
##        prénom          Semestre   Filière     snapshot.1       snapshot.2
##  ALEXANDRE:  3   S3 2017-18: 38   EEA:181   Min.   : 1.000   Min.   : 1.75
##  ALEXIS   :  3   S3 2018-19:143             1st Qu.: 4.500   1st Qu.: 7.00
##  HUGO     :  3                              Median : 6.750   Median : 8.75
##  LUCAS    :  3                              Mean   : 6.442   Mean   : 8.82
##  NICOLAS  :  3                              3rd Qu.: 8.000   3rd Qu.:10.50
##  VINCENT  :  3                              Max.   :13.000   Max.   :15.75
##  (Other)  :163
##  Snapshot.2...4m CYS.S3    CYS.S4    TP.S3    TP.S4     CMI        Groupe.S3
```

```
##          :127     non:120      :137   FR:132     :143   non:150   Siuban :49
##  -       : 15     oui: 61   non: 21   GB: 49   FR: 15   oui: 31   Akane  :34
##  11      :  3               oui: 23            GB: 23             Alba   :29
##  11.75   :  3                                                     Nadia  :24
##  14.5    :  3                                                     Steven :23
##  6.75    :  3                                                     Yolanda:15
##  (Other): 27                                                      (Other): 7
##     Groupe.S4     Prof.TP
##           :154          :168
##  Nadia    : 15   Didier:  6
##  Virginia: 12    Pierre:  7
##
##
##
##
```

```
#  Stat. descriptives à completer
boxplot(CYS$snapshot.1,CYS$snapshot.2,names=c("Snapshot1","Snapshot2"), ylab="Résultat",
        main="Résultats au Snapshot1 et au Snapshot2")
```



Au regard de boxplot, on constate que le Snapshot2 prend souvent les valeurs plus grandes que le Snapshot1 d'où la progresson obtenue en résultat.

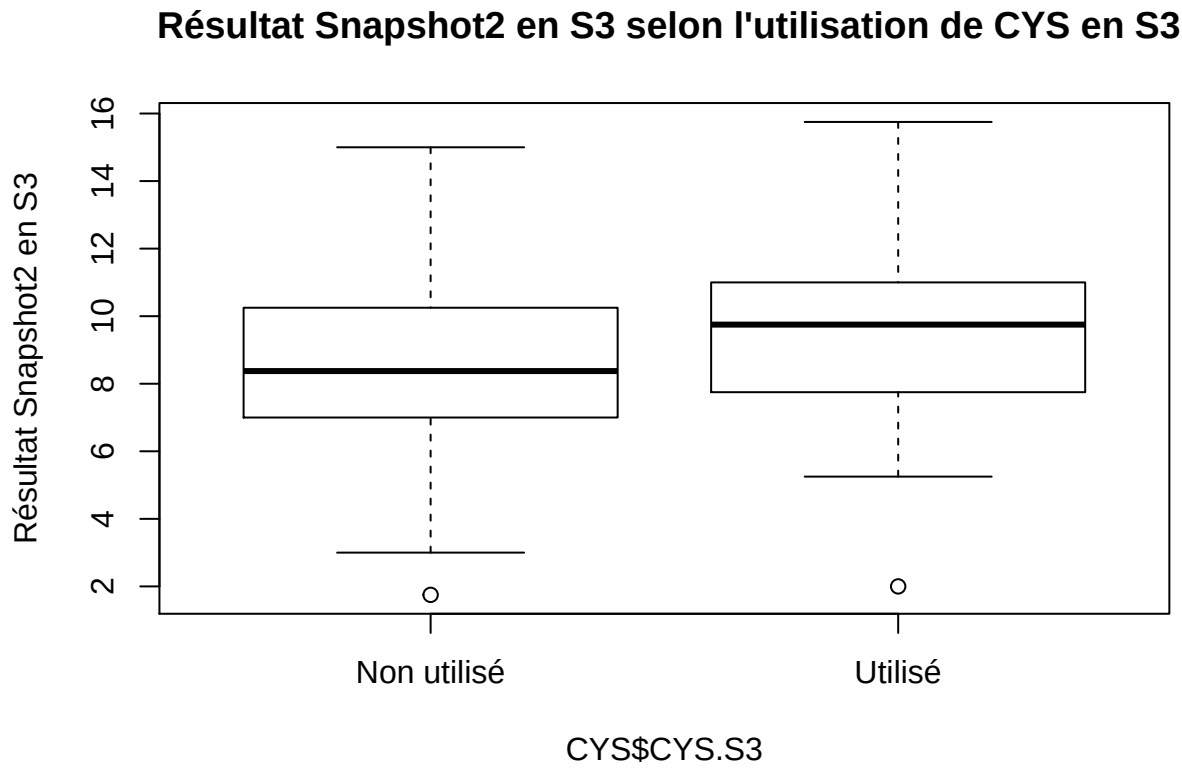
```
#  Stat. descriptives à completer
ks.test(CYS$snapshot.1,CYS$snapshot.2,alternative = "greater")
```

```
## Warning in ks.test(CYS$snapshot.1, CYS$snapshot.2, alternative = "greater"): p-
## value will be approximate in the presence of ties
```

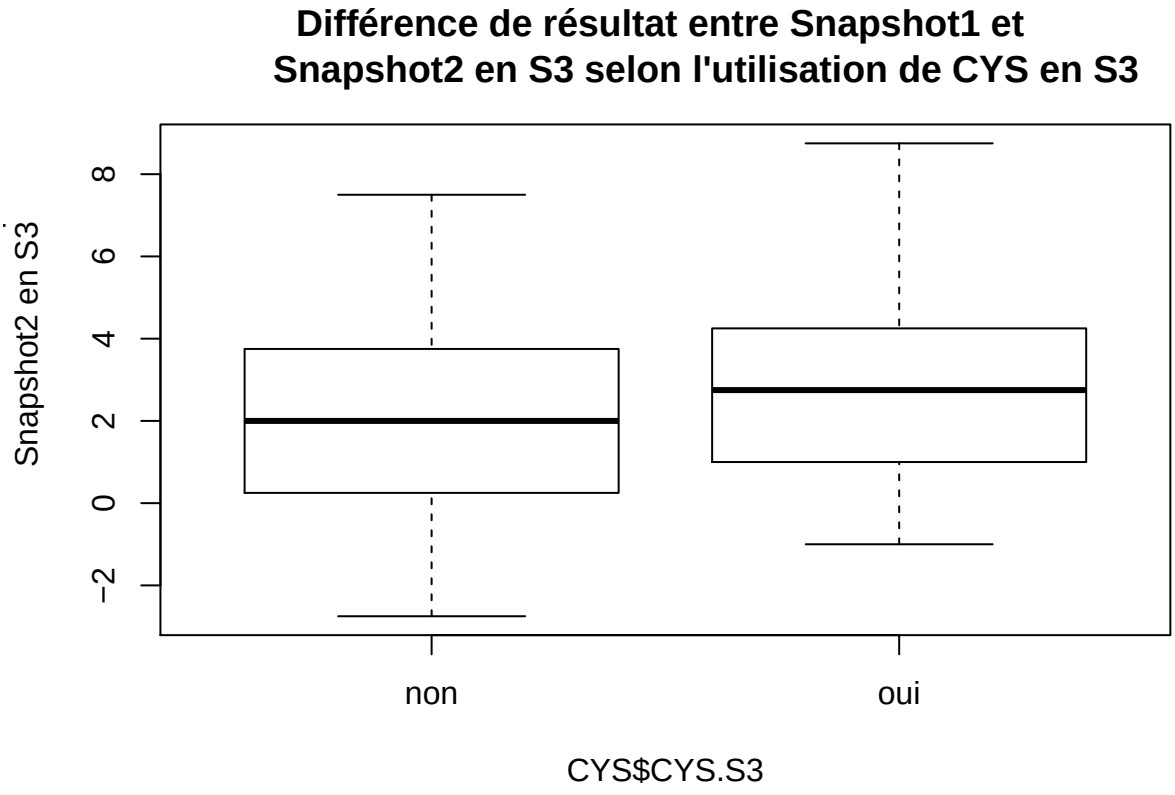
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  CYS$snapshot.1 and CYS$snapshot.2
## D^+ = 0.35912, p-value = 7.286e-11
## alternative hypothesis: the CDF of x lies above that of y
```

La p_valeur associée au test Kolmogrov est inférieure à 0,05 donc on accepte que les étudiants ont des notes de Snapshot2 plus élevé que Snapshot1.

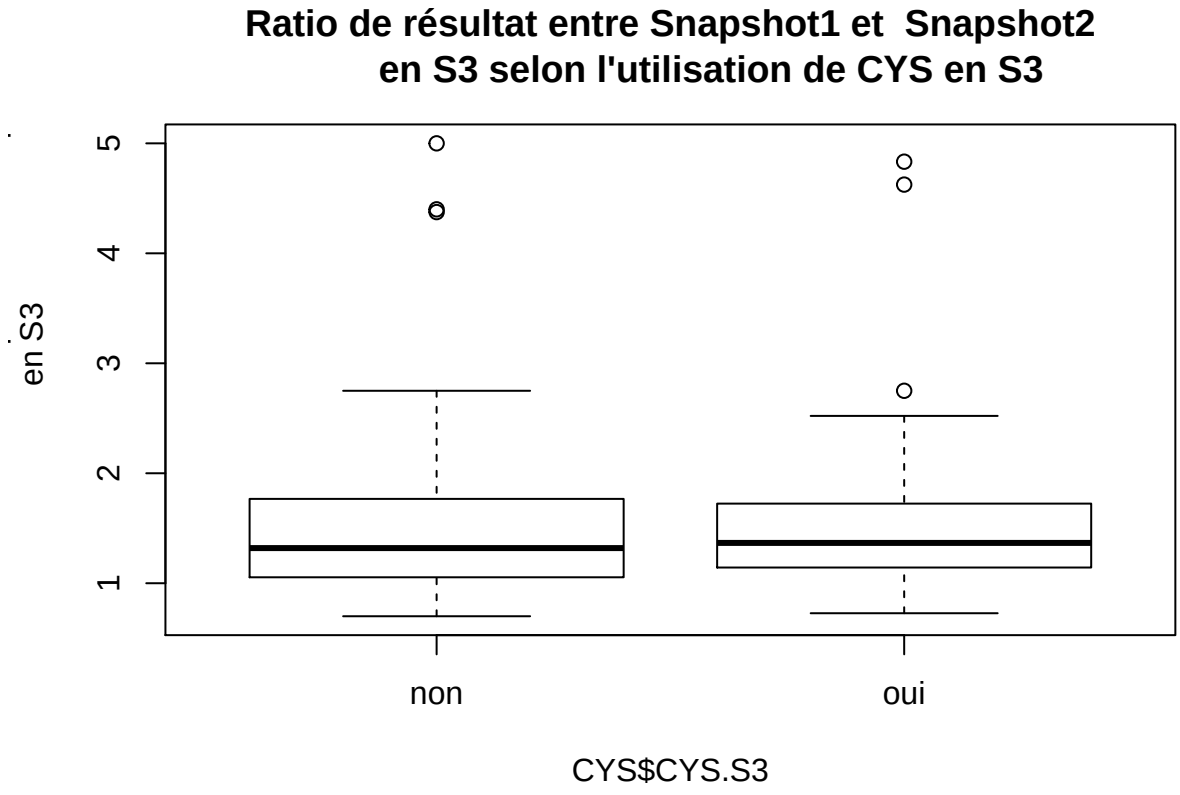
```
#  Stat. descriptives à completer
boxplot(CYS$snapshot.2~CYS$CYS.S3,names=c("Non utilisé","Utilisé"),
        ylab="Résultat Snapshot2 en S3",
        main=" Résultat Snapshot2 en S3 selon l'utilisation de CYS en S3")
```



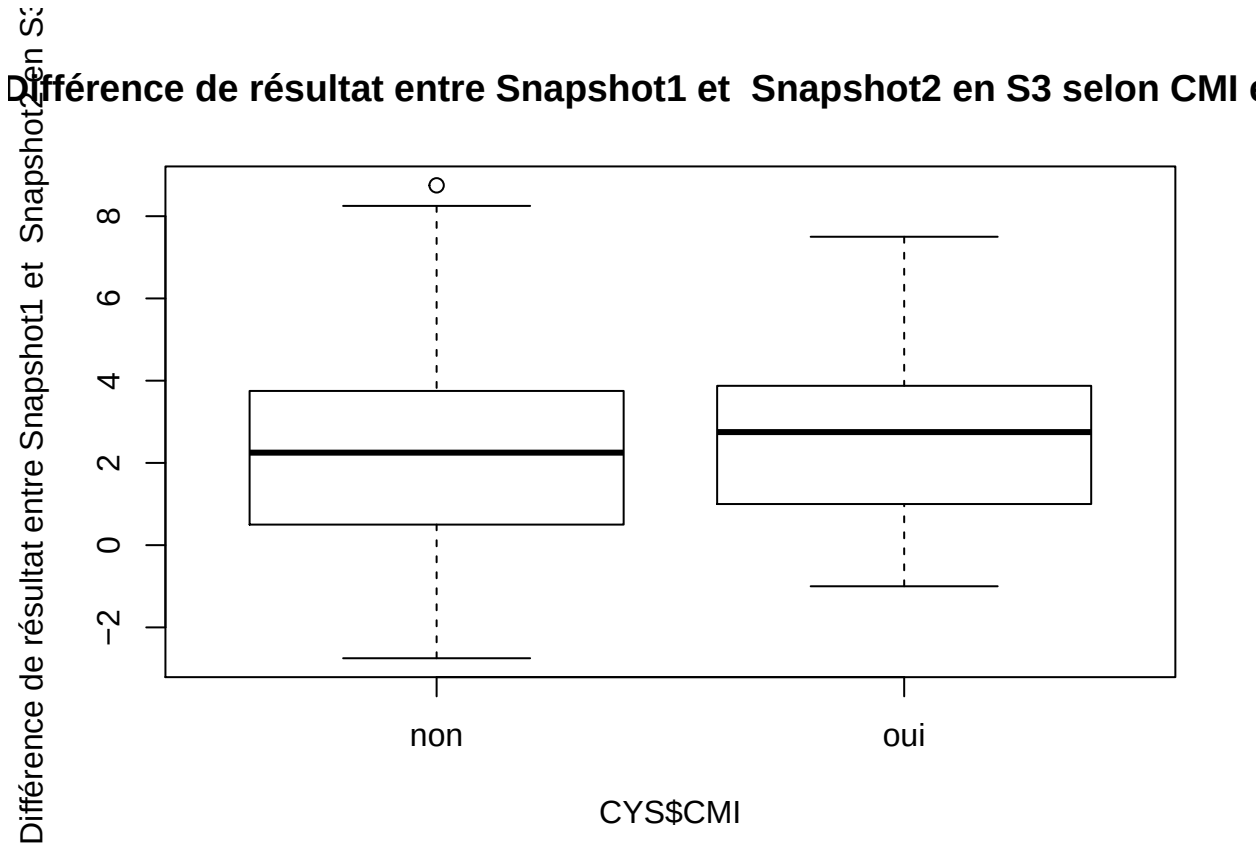

```
#  Stat. descriptives à completer
dif_snap=CYS$snapshot.2-CYS$snapshot.1
ratio_snap=CYS$snapshot.2/CYS$snapshot.1
boxplot(dif_snap~CYS$CYS.S3, ylab="Différence de résultat entre Snapshot1 et
        Snapshot2 en S3",
        main=" Différence de résultat entre Snapshot1 et
        Snapshot2 en S3 selon l'utilisation de CYS en S3")
```



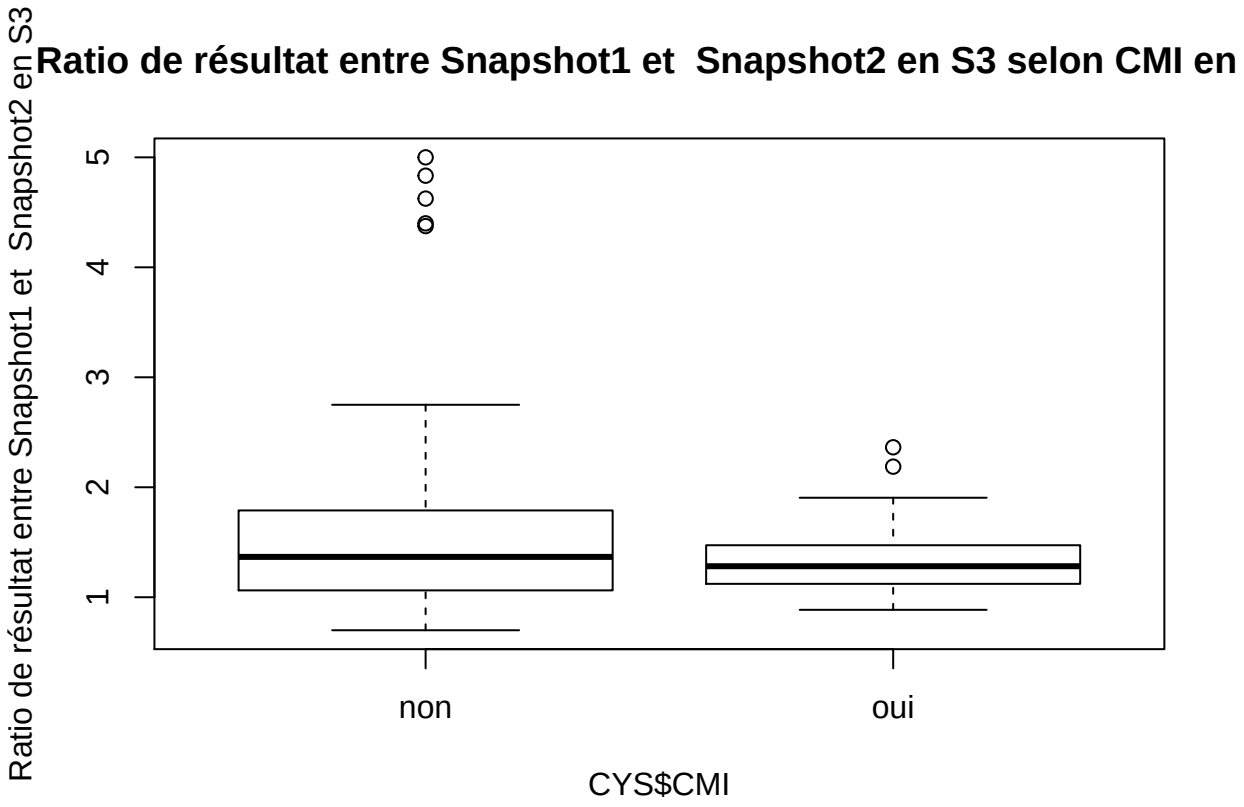

```
boxplot(ratio_snap~CYS$CYS.S3, ylab="Ratio de résultat entre Snapshot1 et  Snapshot2
        en S3",
        main=" Ratio de résultat entre Snapshot1 et  Snapshot2
        en S3 selon l'utilisation de CYS en S3")
```




```
#  Stat. descriptives à completer
boxplot(dif_snap~CYS$CMI, ylab="Différence de résultat entre Snapshot1 et  Snapshot2 en S3",
        main=" Différence de résultat entre Snapshot1 et  Snapshot2 en S3 selon CMI en S3")
```




```
#  Stat. descriptives à completer
boxplot(ratio_snap~CYS$CMI, ylab="Ratio de résultat entre Snapshot1 et  Snapshot2 en S3",
        main=" Ratio de résultat entre Snapshot1 et  Snapshot2 en S3 selon CMI en S3")
```



Remarque: On observe peu d'évolution entre les résultats de Snapshot1 et Snapshot2. Il vaut donc mieux considérer un modèle avec plusieurs variables.

Test d'un modèle ANOVA de 3 facteurs(CYS S3, CMI, TP S3)

```
# A completer
mod1=lm(dif_snap~(CYS$CYS.S3+CYS$TP.S3+ CYS$CMI)^2,data=CYS)
summary(mod1)
```

```
##
## Call:
## lm(formula = dif_snap ~ (CYS$CYS.S3 + CYS$TP.S3 + CYS$CMI)^2,
##     data = CYS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.875 -1.739  0.000  1.255  5.005
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.9954     0.2063   9.675   <2e-16 ***
## CYS$CYS.S3oui                  0.7437     0.4941   1.505    0.134
## CYS$TP.S3GB                   -0.9954     2.1632  -0.460    0.646
## CYS$CMIoui                     3.1250     2.2584   1.384    0.168
## CYS$CYS.S3oui:CYS$TP.S3GB      2.4328     2.2702   1.072    0.285
```

```
## CYS$CYS.S3oui:CYS$CMIoui   -5.5515     2.3652  -2.347    0.020 *
## CYS$TP.S3GB:CYS$CMIoui           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.153 on 175 degrees of freedom
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.09876
## F-statistic: 4.945 on 5 and 175 DF,  p-value: 0.0002929
```

```
# A completer - fonction lm
#step.backward = step(mod1)
```

Selon le test d'AIC, on trouve le meilleur modèle modAIC1:

$$dif\_snap \sim CYS\$CYS.S3 + CYS\$TP.S3 + CYS\$CMI + CYS\$CYS.S3 : CYS\$CMI$$

```
# A completer
#modAIC1=lm(dif_snap ~ CYS$CYS.S3 + CYS$TP.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,data=CYS)
#summary(modAIC1)
#anova(modAIC1,mod1)
```

La p_valeur de Test Fisher est 0,2854 supérieure que 0,05 donc on accepte le modèle modAIC1.

```
#step.backward = step(mod1,direction="backward",k=log(nrow(CYS)))
```

Selon le test d'BIC, on trouve le meilleur modèle:

$dif\_snap \sim CYS\$CYS.S3 + CYS\$CMI + CYS\$CYS.S3 : CYS\$CMI$

```
# A completer
modBIC1=lm(dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,data=CYS)
anova(modBIC1,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI
## Model 2: dif_snap ~ (CYS$CYS.S3 + CYS$TP.S3 + CYS$CMI)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    177 832.61
## 2    175 811.43  2    21.177 2.2835  0.105
```

La p_valeur de Test Fisher est 0,105 supérieure que 0,05 donc on accepte le modèle modBIC1.

```
# A completer
#anova(modBIC1,modAIC1)
```

La p_valeur de Test Fisher est 0,105 supérieure que 0,05 donc on accepte le modèle modBIC1.

```
# A completer
mod2=lm(ratio_snap~(CYS$CYS.S3+CYS$TP.S3+ CYS$CMI)^2,data=CYS)
summary(mod1)
```

```
##
## Call:
## lm(formula = dif_snap ~ (CYS$CYS.S3 + CYS$TP.S3 + CYS$CMI)^2,
##     data = CYS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -4.875 -1.739  0.000  1.255  5.005
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.9954     0.2063   9.675   <2e-16 ***
## CYS$CYS.S3oui                0.7437     0.4941   1.505    0.134
## CYS$TP.S3GB                 -0.9954     2.1632  -0.460    0.646
## CYS$CMIoui                   3.1250     2.2584   1.384    0.168
## CYS$CYS.S3oui:CYS$TP.S3GB    2.4328     2.2702   1.072    0.285
## CYS$CYS.S3oui:CYS$CMIoui    -5.5515     2.3652  -2.347    0.020 *
## CYS$TP.S3GB:CYS$CMIoui           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.153 on 175 degrees of freedom
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.09876
## F-statistic: 4.945 on 5 and 175 DF,  p-value: 0.0002929
```

```
#step.backward = step(mod2,direction="backward",k=log(nrow(CYS)))
```

Selon le test d'BIC, on trouve le meilleur modèle modBIC2:

$ratio\_snap \sim CYS\$CMI$

De même façon, on trouve le meilleur modèle pour modéliser le ratio_snap:

```
# A completer
modBIC2=lm(ratio_snap ~ CYS$CMI,data=CYS)
summary(modBIC2)
```

```
##
## Call:
## lm(formula = ratio_snap ~ CYS$CMI, data = CYS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8666 -0.4505 -0.1435  0.1981  3.4334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.56657    0.05769  27.153   <2e-16 ***
## CYS$CMIoui  -0.20431    0.13941  -1.466    0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7066 on 179 degrees of freedom
## Multiple R-squared:  0.01186,    Adjusted R-squared:  0.006336
## F-statistic: 2.148 on 1 and 179 DF,  p-value: 0.1445
```

```
anova(modBIC2,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: ratio_snap ~ CYS$CMI
## Model 2: ratio_snap ~ (CYS$CYS.S3 + CYS$TP.S3 + CYS$CMI)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    179 89.376
```

```
## 2    175 86.978  4    2.3978 1.2061   0.31
```

La p_valeur de Test Fisher est 0,31 supérieure que 0,05 donc on accepte le modèle modBIC1.

Remarque: Grâce au test ANOVA on trouve que:

- Si on considère la différence entre les deux snapshots on obtient le modèle

$dif\_snap \sim CYS\$CYS.S3 + CYS\$CMI + CYS\$CYS.S3 : CYS\$CMI$

Cela montre l'impact de CMI et CYS S3 sur l'évolution de résultat.

- SI on considère la ratio entre les deux snapshots on obtient le modèle

$ratio\_snap \sim CYS\$CMI$

Cela montre l'impact de CMI S3 sur l'évolution de résultat.

Dans ces deux cas on ne trouve pas l'effet de la variable TP S3.

Test d'un modèle ANCOVA de 3 facteurs qualitatives (CYS S3, CMI, TP S3) et 1 facteur quantitative(Snapshot1) pour modéliser le Snapshot2

On souhaite expliquer le Snapshot2 en fonction du Snapshot1, des choix (de la langue de TP, l'utilisation de CMI et celle de CYS) en S3. On met en place un modèle d'analyse de la covariance.

```
# A completer
modR=lm(CYS$snapshot.2 ~ (CYS$snapshot.1+CYS$TP.S3+CYS$CYS.S3+ CYS$CMI)^2,data=CYS)
summary(modR)
```

```
##
## Call:
## lm(formula = CYS$snapshot.2 ~ (CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 +
##     CYS$CMI)^2, data = CYS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9027 -1.3388 -0.0402  1.1452  5.7622
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.52851    0.50089   9.041 3.29e-16 ***
## CYS$snapshot.1                 0.58867    0.07554   7.792 6.10e-13 ***
## CYS$TP.S3GB                    0.36310    2.94340   0.123  0.90197
## CYS$CYS.S3oui                  0.21415    1.12806   0.190  0.84966
## CYS$CMIoui                     3.96204    3.15382   1.256  0.21073
## CYS$snapshot.1:CYS$TP.S3GB    -0.18738    0.33981  -0.551  0.58207
## CYS$snapshot.1:CYS$CYS.S3oui   0.01915    0.19788   0.097  0.92300
## CYS$snapshot.1:CYS$CMIoui      0.04565    0.34210   0.133  0.89400
## CYS$TP.S3GB:CYS$CYS.S3oui      3.03385    2.06231   1.471  0.14310
## CYS$TP.S3GB:CYS$CMIoui              NA         NA      NA       NA
## CYS$CYS.S3oui:CYS$CMIoui      -6.05498    2.16507  -2.797  0.00575 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.938 on 171 degrees of freedom
## Multiple R-squared:  0.4853, Adjusted R-squared:  0.4582
```

```
## F-statistic: 17.91 on 9 and 171 DF,  p-value: < 2.2e-16
```

```
# A completer
#step.backward = step(modR,direction="backward",k=log(nrow(CYS)))
```

```
# A completer
modbest=lm(CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 + CYS$CMI +
    CYS$CYS.S3:CYS$CMI,data=CYS)
summary(modbest)
```

```
##
## Call:
## lm(formula = CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 +
##     CYS$CMI + CYS$CYS.S3:CYS$CMI, data = CYS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8965 -1.2725 -0.0205  1.1728  5.8232
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.62384    0.44064  10.493  < 2e-16 ***
## CYS$snapshot.1             0.56912    0.06511   8.740 1.85e-15 ***
## CYS$TP.S3GB                1.90361    0.59802   3.183  0.00172 **
## CYS$CYS.S3oui              0.43506    0.43744   0.995  0.32132
## CYS$CMIoui                 1.28154    0.86629   1.479  0.14084
## CYS$CYS.S3oui:CYS$CMIoui -3.02652    0.85983  -3.520  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.932 on 175 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4615
## F-statistic: 31.85 on 5 and 175 DF,  p-value: < 2.2e-16
```

```
anova(modbest,modR)
```

```
## Analysis of Variance Table
##
## Model 1: CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 + CYS$CMI +
##     CYS$CYS.S3:CYS$CMI
## Model 2: CYS$snapshot.2 ~ (CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 + CYS$CMI)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    175 653.30
## 2    171 642.27  4    11.031 0.7343 0.5698
```

Selon le test d'BIC, on trouve le meilleur modèle modBIC2:

$CYS\$snapshot.2 \sim CYS\$snapshot.1 + CYS\$TP.S3 + CYS\$CYS.S3 + CYS\$CMI + CYS\$CYS.S3 : CYS\$CMI$

En fait, on veut estimer le résultat de Snapshot2 par le modèle:

$$(modbest) : Snapshot2_{ijkl} = \mu + \alpha Snapshot1_{ijkl} + \beta_i + \gamma_j + \theta_k + \delta_{jk} + \varepsilon_{ijkl}, \forall i = 1,\ 2,\ \forall j = 1, 2, \forall k = 1, 2$$

où:

i,j,k sont les indices de modalité pour les variables qualitatives TP.S3, CYS.S3 et CMI, respectivement.(1 pour la réponse Non et 2 pour la réponse Oui, dans le cas de TP 1 pour FR et 2 pour GB)

L'indice ijkl est pour indiquer l'individu l-ième ayant des modalités i,j,k pour TP.S3, CYS.S3 et CMI, respectivement. ε_{ijkl} est des erreurs de l'estimation de l'individu ayant l'indice ijkl.

D'où:

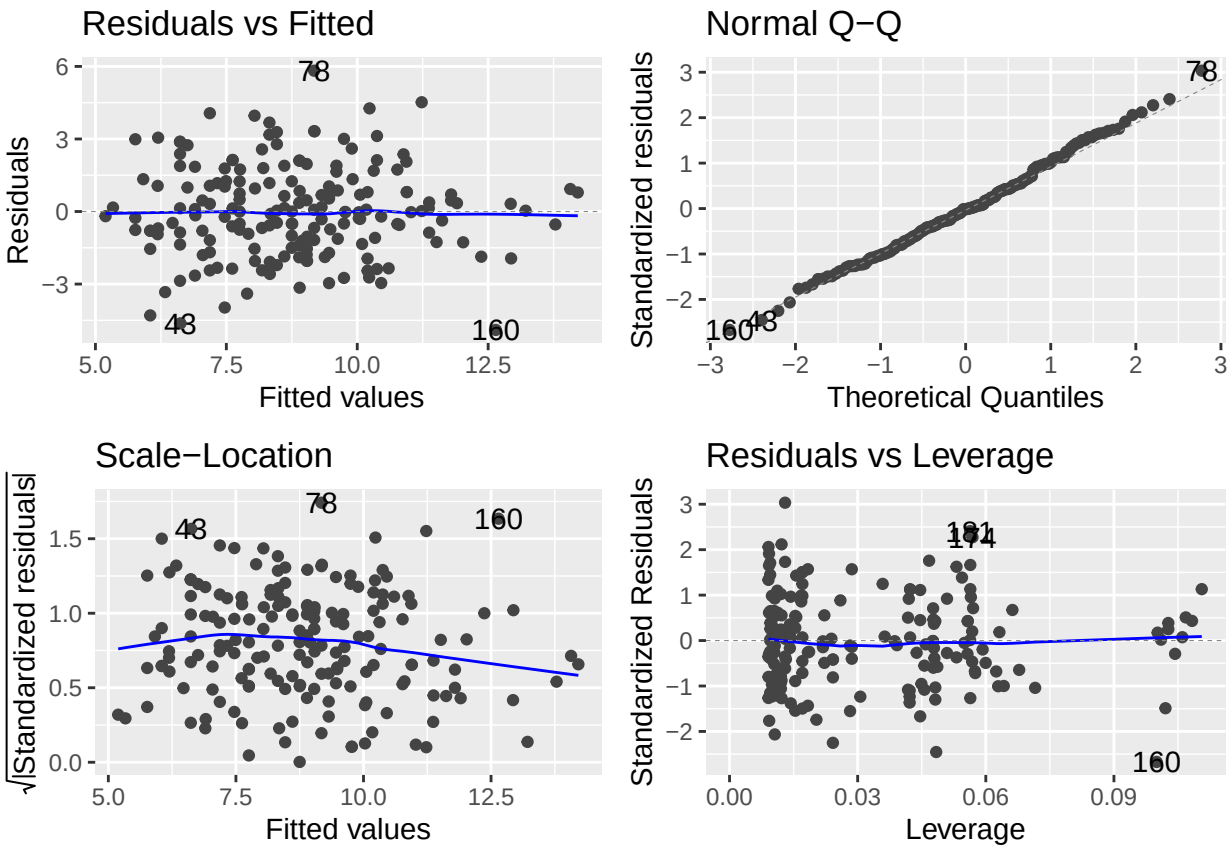
$$\mu = 4.62384 \alpha = 0.56912 \beta_1 = \gamma_1 = \theta_1 = \delta_{11} = \delta_{12} == \delta_{21} = 0 \beta_2 = 1.90361 \gamma_2 = 0.43506 \delta_{22} = -3.02652$$

Train+ Test par le modèle ANCOVA

```
all.err=numeric(0)
K=9
n=nrow(CYS)
taille=n%/%K
set.seed(5)
alea<-runif(n)
rang=rank(alea)
bloc=(rang-1)%/%taille+1
bloc=as.factor(bloc)
err=0
for (k in 1:K){
  dt=CYS[bloc==k,]
  modk=lm(snapshot.2 ~ snapshot.1 +TP.S3 + CYS.S3 + CMI +
    CYS.S3:CMI,data=CYS[bloc!=k,])
  pred=predict(modk,newdata=dt)
  xerr=sum((dt$snapshot.2-pred)^2)
  err= err+xerr
}
print(err)
```

```
## [1] 697.8652
```

```
autoplot(modbest)
```



Vérifions les 4 hypothèses pour ce modèle:

Hypothèse H1 : Les erreurs sont centrées

Hypothèse H2 : La variance des erreurs est constante

Hypothèse H3 : Les résidus sont indépendantes

Hypothèse H4 : Les données suivent des lois gaussiennes

- Les résidus sont situés aléatoirement sur les deux côtés de 0. Or, on ne trouve pas la forme bannane ni la forme trompette, donc on peut déduire que H1 et H2 sont vérifiées
- On observe dans le graphe Residuals vs Fitted aucun phénomène de pacquets sur les côtés de 0 donc H3 est vérifiée.
- Au regard du graphe Normal Q-Q, on constate que les résidus suivent une loi gaussienne. Ainsi, H4 est vérifiée.

Abre binaire de régression

```
library(rpart) # chargement de la librairie
data3=CYS[,c(4,5,7,9,11)]
data5=CYS[,c(4,5,7,9,11)]
data3[,"CYS_CMI"]=(data3[,"CYS.S3"]=="oui")*(data3[,"CMI"]=="oui")
data3[,"CYS_CMI"]=as.factor(data3[,"CYS_CMI"])
data3[,"CYS_TP"]=(data3[,"CYS.S3"]=="oui")*(data3[,"TP.S3"]=="GB")
data3[,"CYS_TP"]=as.factor(data3[,"CYS_TP"])
```

```
data3[,"TP_CMI"]=(data3[,"TP.S3"]=="GB")*(data3[,"CMI"]=="oui")
data3[,"TP_CMI"]=as.factor(data3[,"TP_CMI"])
data3[,"CYS_CMInon"]=(data3[,"CYS.S3"]=="oui")*(data3[,"CMI"]=="non")
data3[,"CYS_CMInon"]=as.factor(data3[,"CYS_CMInon"])
data3[,"CYS_TPFR"]=(data3[,"CYS.S3"]=="oui")*(data3[,"TP.S3"]=="FR")
data3[,"CYS_TPFR"]=as.factor(data3[,"CYS_TPFR"])
data4=data3
data4[,"eval"]=data4[,"snapshot.2"]-data4[,"snapshot.1"]
data4=data4[,c(3,4,5,6,7,8,9,10,11)]
summary(data3)
```

```
##    snapshot.1       snapshot.2    CYS.S3    TP.S3     CMI      CYS_CMI CYS_TP
##  Min.   : 1.000   Min.   : 1.75   non:120   FR:132   non:150   0:160   0:143
##  1st Qu.: 4.500   1st Qu.: 7.00   oui: 61   GB: 49   oui: 31   1: 21   1: 38
##  Median : 6.750   Median : 8.75
##  Mean   : 6.442   Mean   : 8.82
##  3rd Qu.: 8.000   3rd Qu.:10.50
##  Max.   :13.000   Max.   :15.75
##  TP_CMI  CYS_CMInon CYS_TPFR
##  0:150   0:141      0:158
##  1: 31   1: 40      1: 23
##
##
##
##
```

```
summary(data4)
```

```
##  CYS.S3    TP.S3     CMI      CYS_CMI CYS_TP  TP_CMI  CYS_CMInon CYS_TPFR
##  non:120   FR:132   non:150   0:160   0:143   0:150   0:141      0:158
##  oui: 61   GB: 49   oui: 31   1: 21   1: 38   1: 31   1: 40      1: 23
##
##
##
##
##       eval
##  Min.   :-2.750
##  1st Qu.: 0.500
##  Median : 2.250
##  Mean   : 2.378
##  3rd Qu.: 3.750
##  Max.   : 8.750
```

```
tree.reg3=rpart(snapshot.2~.,data=data3)
tree.reg4=rpart(eval~.,data=data4)
```

```
#plot(tree.reg3)
#text(tree.reg3)
```

```
library(partykit)
```

```
## Warning: package 'partykit' was built under R version 3.6.3
```

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Warning: package 'libcoin' was built under R version 3.6.3
```

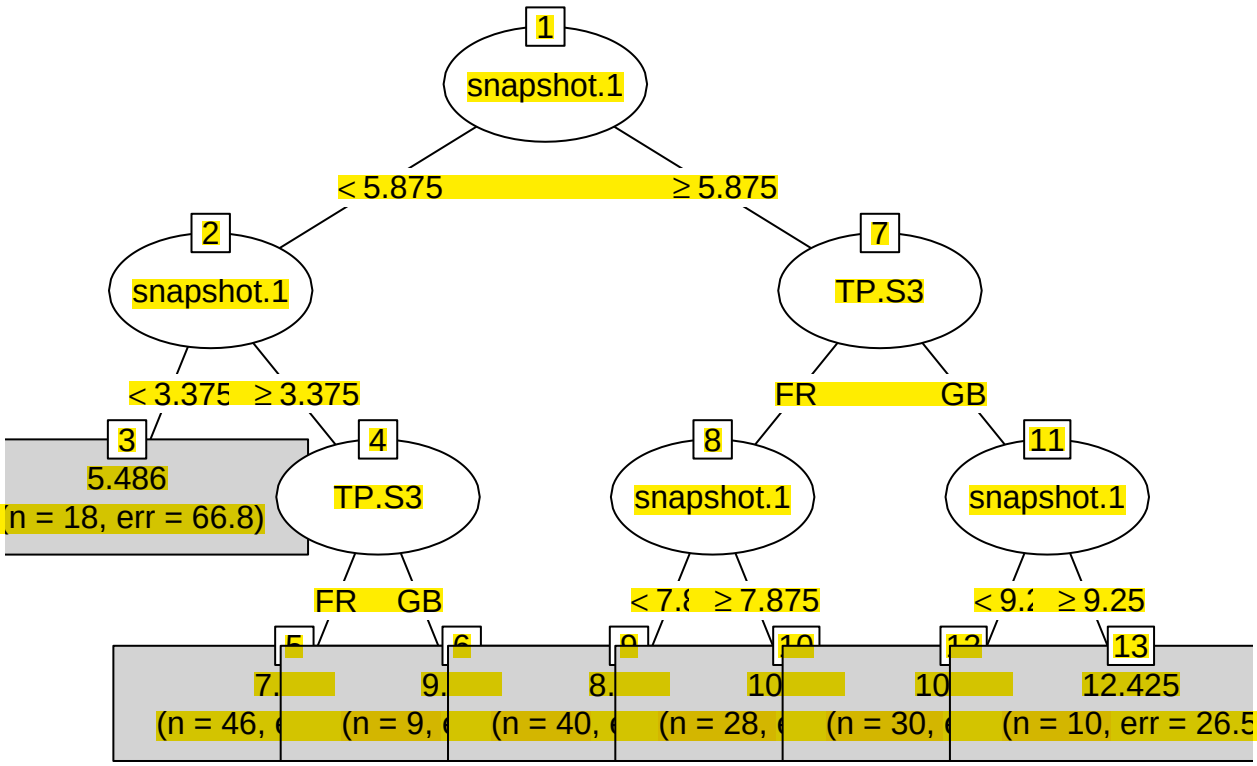
```
## Loading required package: mvtnorm
```

```
xmat3=xpred.rpart(tree.reg3)
xerr3=(xmat3-data3[,"snapshot.2"])^2
CVerr3=apply(xerr3,2,sum)
xmat4=xpred.rpart(tree.reg4)
xerr4=(xmat4-data4[,"eval"])^2
CVerr4=apply(xerr4,2,sum)
#tree.reg3=rpart(snapshot.2~.,data=data3,
#  control=rpart.control(cp=as.numeric(attributes(which.min(CVerr3))$names)))
#plot(as.party(tree.reg3), type="simple")
#tree.reg4=rpart(eval~.,data=data4,control=rpart.control(cp=as.numeric(attributes(which.min(CVerr4))$na
#plot(as.party(tree.reg4), type="simple")
```

```
tree.reg5=rpart(snapshot.2~.,data=data5)
xmat5=xpred.rpart(tree.reg5)
xerr5=(xmat5-data5[,"snapshot.2"])^2
CVerr5=apply(xerr5,2,sum)
```

```
data6=data5
data6[,"eval"]=data6[,"snapshot.2"]-0.55*data6[,"snapshot.1"]
data6=data6[,-c(1,2)]
```

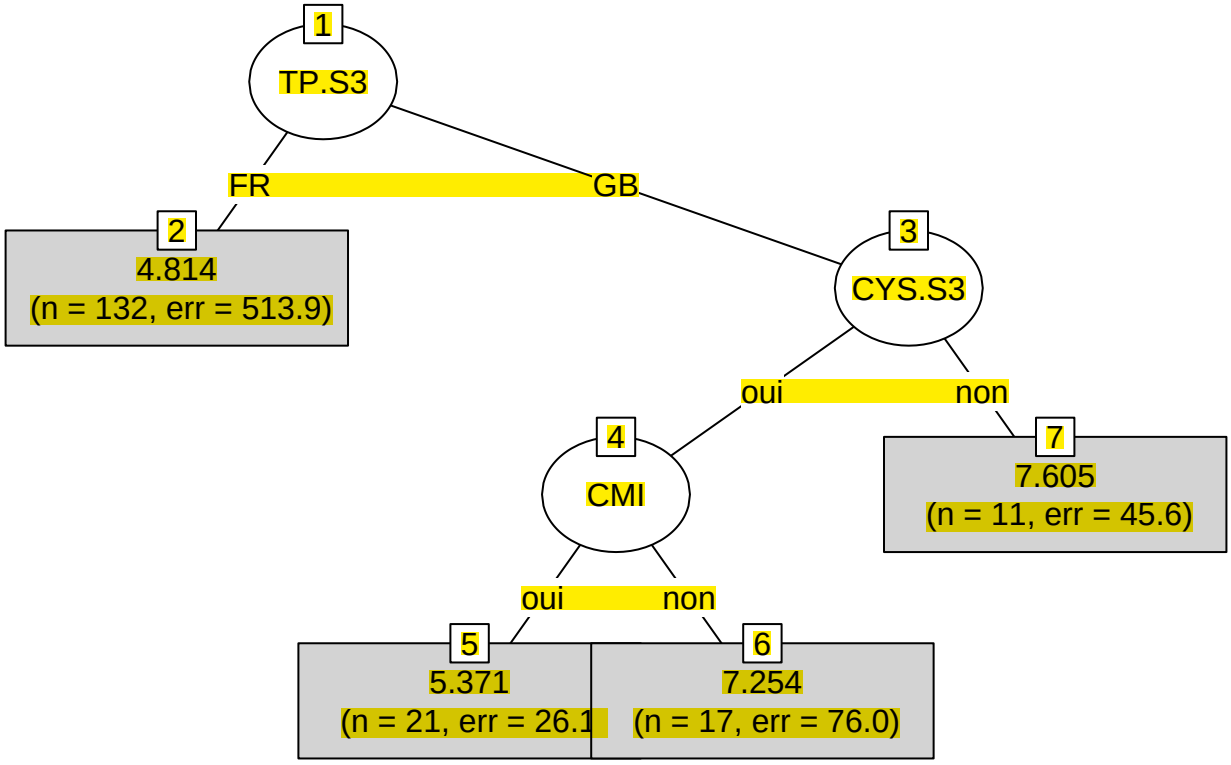
```
tree.reg5=rpart(snapshot.2~.,data=data5,
                control=rpart.control(cp=as.numeric(attributes(which.min(CVerr5))$names)))
plot(as.party(tree.reg5), type="simple")
```



```
for (alpha in seq(0.1,2,0.1)){
  data7=data5
  data7[,"eval"]=data7[,"snapshot.2"]-alpha*data7[,"snapshot.1"]
  data7=data7[,-c(1,2)]
  tree.reg7=rpart(eval~.,data=data7)
  xmat7=xpred.rpart(tree.reg7)
  xerr7=(xmat7-data7[,"eval"])^2
  CVerr7=apply(xerr7,2,sum)
  #print(min(CVerr7))
}
```

```
tree.reg6=rpart(eval~.,data=data6)
xmat6=xpred.rpart(tree.reg6)
xerr6=(xmat6-data6[,"eval"])^2
CVerr6=apply(xerr6,2,sum)
```

```
tree.reg6=rpart(eval~.,data=data6,
                control=rpart.control(cp=as.numeric(attributes(which.min(CVerr6))$names)))
plot(as.party(tree.reg6), type="simple")
```




```
#min(CVerr3)
```

```
#min(CVerr4)
```

```
min(CVerr5)
```

```
## [1] 777.6372
```

```
min(CVerr6)
```

```
## [1] 686.4129
```

```
#residuals=data5$snapshot.2-predict(tree.reg5,data5)
#SSE=sum(residuals^2)
#SST=sum((data5$snapshot.2-mean(data5$snapshot.2))^2)
#SSE/SST
```

Commentaire:

Au regard de l'arbre de régression, on trouve les résultats suivants:

- On considère 2 modèles: l'un modélise snapshot2 en fonction d'autres variables et l'autre modélise l'écart snapshot2-α×snapshot1 en variant α de 0.1 à 2.
- On a fait une boucle en variant les valeurs de α et on trouve le α "optimal" pour rendre l'erreur plus petite possible. $\alpha = 0.55$
- On a essayé plusieurs fois pour étudier le comportement des erreurs de validation croisée de ces deux types de l'arbre et a gardé le second.
- Parmi les étudiants faisant les TP en anglais, un étudiant utilisant l'outil CYS progresse moins qu'un autre n'utilisant pas CYS. (cf des feuilles 5,6,7). De plus, parmi les étudiants faisant les TP en anglais et utilisant l'outil CYS, un étudiant en CMI progresse moins qu'un autre non CMI.

Conclusion:

On a décidé de modéliser la note de Snapshot2 en fonction de Snapshot1, l'utilisation de l'outil CheckYourSmile, la langue de TP et le fait que l'étudiant est en CMI ou pas.

Alors, sous le modèle ANCOVA on a trouvé que les trois facteurs qualitatives ont des impactes sur le résultat Snapshot2. Or, le Snaphot1 a un gros effet sur le résultat de Snapshot2. On y trouve aussi un terme d'interaction entre la variable CYS.S3 et la variable CMI et celui dernier a un effet important négatif sur le Snapshot2.

ie, un étudiant en CMI utilisant l'outil Check Your Smile a tendance de dégrader environ 2,6 points (-3,02652+0,43506) et un étudiant non CMI utilisant l'outil Check Your Smile a tendance de progresser environ 0,4 points (0,43506)

On voit que le modèle ANCOVA nous donne un R-ajusté bien meilleur que le modèle ANOVA. (0,4615»0,006 et 0,4615»0,09)

L'arbre binaire de régression nous permet d'observer l'effet de l'outil CYS dans la progression des étudiants n'est pas remarquable.

L'erreur de validation croisé du modèle ANCOVA est presque égale à celle du l'arbre de décision donc on trouve le même phénomène sur l'effet de CYS sur la progression des étudiants.

Pourtant, l'analyse des donnée peut s'améliorer car on n'a gardé que 181 données en semestre 3 et 9 en semestre 4 sur 459 étudiants.

En semestre 4, les données ne sont pas assez bonnes:

- 126 individus n'ont pas de note de Snapshot1
- 51 individus n'ont pas de note de Snapshot2
- 28 individus ont des réponses invalidées si ils utilisent ou pas l'outil CYS (la valeur de réponse est soit BLANK soit ?)
- 86 individus ont des réponses invalidées de la langue pour TP (la valeur de réponse est soit non soit BLANK)
- 43 individus ont des réponses invalidées si ils sont CMI (la valeur de réponse est soit BLANK soit ?)

S3	Snapshot1	Snapshot2	CYS S3	TP S3	CMI
Manquant	38	27	1	23	21
Total		242			
Restant		181			
Nb négligé		61			
S4	Snapshot1	Snapshot2	CYS S4	TP S4	CMI
Manquant	126	51	28	86	43
Total		217			
Restant		50			
Nb négligé		167			

Figure 1: Données Manquantes