A4 - Anàlisi de la variància i repàs del curs

Xavier Vizcaino Gascon

10 de junio, 2022

Contents

1. Lectura del fitxer i preparació de les dades

Llegiu el fitxer gpa.csv i guardeu les dades en un objecte denominat gpa. A continuació, verifiqueu el tipus de cada variable. Quines variables són de tipus numèric? Quines variables són de tipus qualitatiu?

Llegim el fitxer amb l'opció *stringsAsFactors = TRUE* i posteriorment corregim el tipus de dades de la variable *tothors* a *char*.

```
gpa <- read.csv("gpa.csv", stringsAsFactors=TRUE)
gpa$tothrs<-as.character(gpa$tothrs)
str(gpa)
```

```
## 'data.frame':    4137 obs. of  10 variables:
##  $ sat    : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs : chr  "43h" "18h" "14h" "40h" ...
##  $ hsize  : num  0.1 9.4 1.19 5.71 2.14 ...
##  $ hsrank : int  4 191 42 252 86 41 161 101 161 3 ...
##  $ hsperc : num  40 20.3 35.3 44.1 40.2 ...
##  $ colgpa : num  2.04 4 1.78 2.42 2.61 ...
##  $ athlete: logi  TRUE FALSE TRUE FALSE FALSE FALSE ...
##  $ female : logi  TRUE FALSE FALSE FALSE FALSE TRUE ...
##  $ white  : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
##  $ black  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

1.1. Preparació de les dades

La variable tothrs està classificada com a character. Per a poder treballar amb ella cal convertir-la en numèrica, eliminant el text "h" de les dades.

Substituïm la "h" per " " i posteriorment convertim les dades a *integer*.

```
gpa_orignal<-gpa
gpa$tothrs<-as.integer(sub("h","",gpa$tothrs))
str(gpa$tothrs)
```

```
##  int [1:4137] 43 18 14 40 18 114 78 55 18 17 ...
```

1.2. Valors absents

Comproveu quantes observacions tenen valors absents i traieu conclusions sobre com de preocupant és el problema de valors absents en aquestes dades.

Obtenim les dimensions del *DataFrame* i estudiem quants NA tenim i en quines variables.

```
dim (gpa)
```

```
## [1] 4137   10
```

```
sapply(gpa,function(x) sum(is.na(x)))
```

```
##     sat  tothrs   hsize  hsrank  hsperc  colgpa athlete  female   white   black
##       0       0       0       0       0      41       0       0       0       0
```

Observem que només tenim 41 NA i tots són en la variable *colgpa*. Si tenim present que el dataset té 4137 registres; els NA representen tan sols el 0.99% de les dades i conseqüentment arribem a la conclusió que el nombre de NAs no suposa un problema.

Elimineu els valors absents del conjunt de dades. Denomineu al nou conjunt de dades 'gpaclean'.

```
gpa<-gpa[!is.na(gpa$colgpa),]
dim (gpa)
```

```
## [1] 4096   10
```

```
gpaclean<-gpa
```

1.3. Equivalència de la nota en lletres

La variable colgpa conté la nota numèrica de l'estudiant. Creeu una variable categòrica anomenada gpaletter, que indiqui la nota en lletra de cada estudiant de la següent forma: A, de 3.50 a 4.00; B, de 2.50 a 3.49; C, de 1.50 a 2.49; D, de 0 a 1.49.

```
gpaclean$gpaletter<-as.factor(ifelse(gpaclean$colgpa<=1.49,"D",
                              ifelse(gpaclean$colgpa<=2.49,"C",
                              ifelse(gpaclean$colgpa<=3.49,"B",
                                                           "A"))))
summary(gpaclean$gpaletter)
```

```
##    A    B    C    D
##  458 1973 1521  144
```

2. Estadística descriptiva i visualització

2.1. Anàlisi descriptiva

Realitzeu una anàlisi descriptiva numèrica de les dades (resumiu els valors de les variables numèriques i categòriques). Mostreu el nombre d'observacions i el nombre de variables.

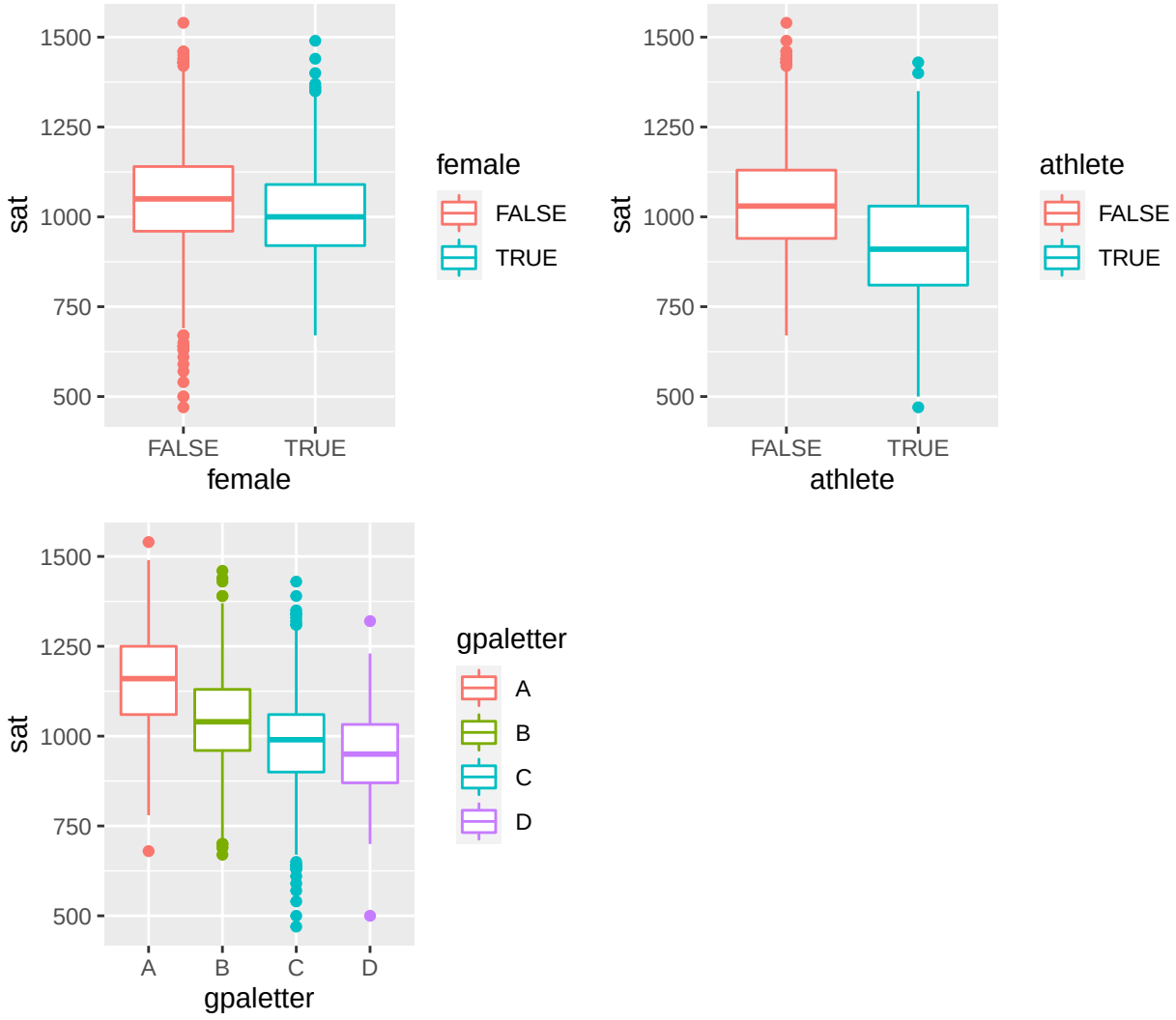
```
summary(gpaclean)
```

```
##       sat           tothrs           hsize           hsrank
##  Min.   : 470   Min.   :  6.00   Min.   :0.030   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 17.00   1st Qu.:1.647   1st Qu.: 11.00
##  Median :1030   Median : 47.00   Median :2.510   Median : 30.00
##  Mean   :1031   Mean   : 52.78   Mean   :2.795   Mean   : 52.74
##  3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.660   3rd Qu.: 70.00
```

```
##  Max.   :1540   Max.   :137.00   Max.   :9.400   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4252   1st Qu.:2.210   FALSE:3905      FALSE:2253
##  Median :14.5833   Median :2.660   TRUE :191       TRUE :1843
##  Mean   :19.2227   Mean   :2.655
##  3rd Qu.:27.6755   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white             black          gpaletter
##  Mode :logical   Mode :logical   A: 458
##  FALSE:304       FALSE:3871      B:1973
##  TRUE :3792      TRUE :225       C:1521
##                                  D: 144
##
##
```

2.2. Visualització

Mostreu amb diversos diagrames de caixa (boxplot) la distribució de la variable 'sat' segons la variable 'female', segons 'athlete', i segons 'gpaletter'.



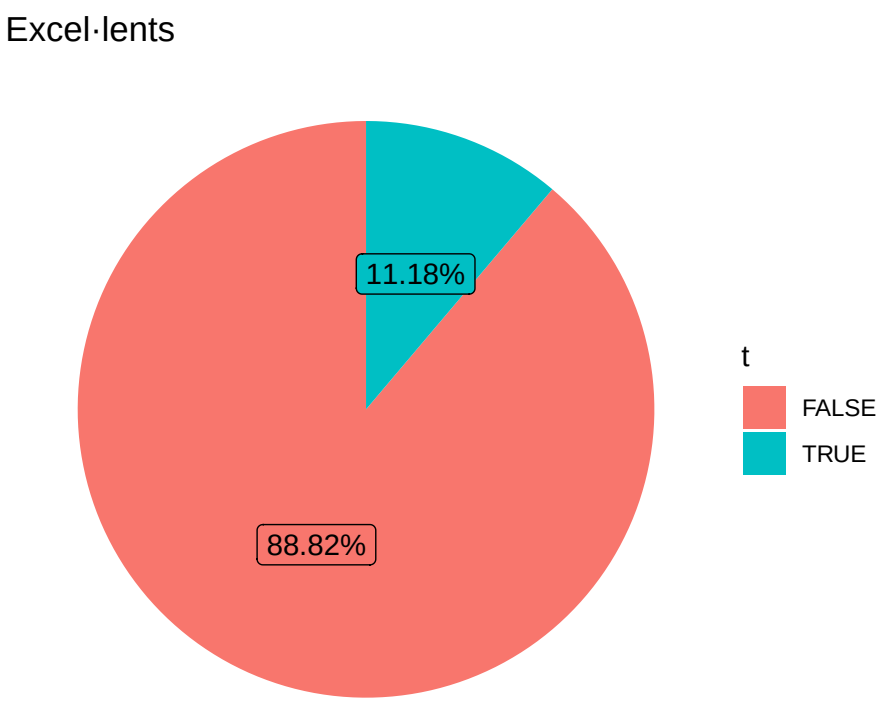
Creeu una variable denominada 'excelente' que indiqui si l'estudiant ha obtingut una A de nota mitjana al final del semestre. Aquesta nova variable s'ha de codificar com una variable dicotòmica que pren el valor 1 quan l'estudiant ha obtingut una A i el valor 0 en cas contrari.

```
gpaclean$excelente<-as.factor(ifelse(gpaclean$gpaletter == "A", 1, 0))
summary(gpaclean$excelente)
```

```
##    0    1
## 3638  458
```

Dibuixeu un gràfic que mostri el percentatge d'estudiants excel · lents.

```
n<-summary(gpaclean$excelente)
p<-paste(round(prop.table(summary(gpaclean$excelente))*100,2), "%", sep = "")
t<-c("FALSE", "TRUE")

df<-data.frame(n,p,t)

ggplot(df, aes(x = "", y = n, fill = t)) +
  geom_col() +
  geom_label(aes(label = p),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
  labs(title = "Excel·lents") +
  coord_polar(theta = "y") +
  theme_void()
```



Interpreteu els gràfics breument.

- *sat* vs *female*: S'observen diferències en les medianes de *sat* segons la variable *female*. La mediana per al grup *female* = *False* (homes) és superior tot i que la dispersió, especialment per valors baixos de *sat* és superior.

- *sat* vs *athlete*: També s'observen diferències en les distribucions de *sat* segons la variable *athlete*. En general, la distribució per al grup *athlete = False* (no esportistes) té tots els valors (min, max, quartils i mediana) més elevats.
- *sat* vs *gpaletter*: S'observa un cert nivell d'ordre en les distribucions de *sat* segons la variable *gpaletter*. En general, tots els valors (min, max, quartils i mediana) estan endreçats seguint A > B > C > D. Més enllà d'aquesta afirmació general caldria destacar que el grup *gpaletter = C* podria presentar una dispersió superior a la resta de grups especialment per valors baixos de *sat*
- excel · lents: Només el 11.18% dels alumnes pertanyen al grup *excelente*.

3. Estadística inferencial

3.1. Interval de confiança de la mitjana poblacional de la variable sat

Calculeu manualment l'interval de confiança al 95% de la mitjana poblacional de la variable sat dels estudiants. Per fer-ho, definiu una funció IC que rebi la variable, la confiança, i que retorni un vector amb els valors de l'interval de confiança.

```
IC<-function(x, NC){
  alfa<-1-NC
  sd<-sd(x)
  n<-length(x)
  SE<-sd/sqrt(n)
  # Distribució t-student doncs no coneixem la variança poblacional
  z<-qt(alfa/2, df=n-1, lower.tail = FALSE)
  L<-mean(x)-z*SE
  U<-mean(x)+z*SE
  round(c(L,U),3)
}

IC_sat<-IC(gpaclean$sat, 0.95)
IC_sat
```

```
## [1] 1026.637 1035.174
```

A partir del resultat obtingut, expliqueu com s'interpreta l'interval de confiança.

La interpretació dels resultats indica que el NC% (en el nostres cas el 95%) de les mostres aleatòries obtingudes de la població donen lloc a un interval que conté el valor real de la mitjana poblacional. En el cas que estem estudiant, podem afirmar que si obtinguéssim infinites mostres de la població, el **95%** de les mostres, contindrien el valor real de la **mitjana poblacional** en l'interval [1026.637, 1035.174].

Calculeu els intervals de confiança al 95% de la mitjana poblacional de la variable sat, en funció de si els estudiants son homes o dones.

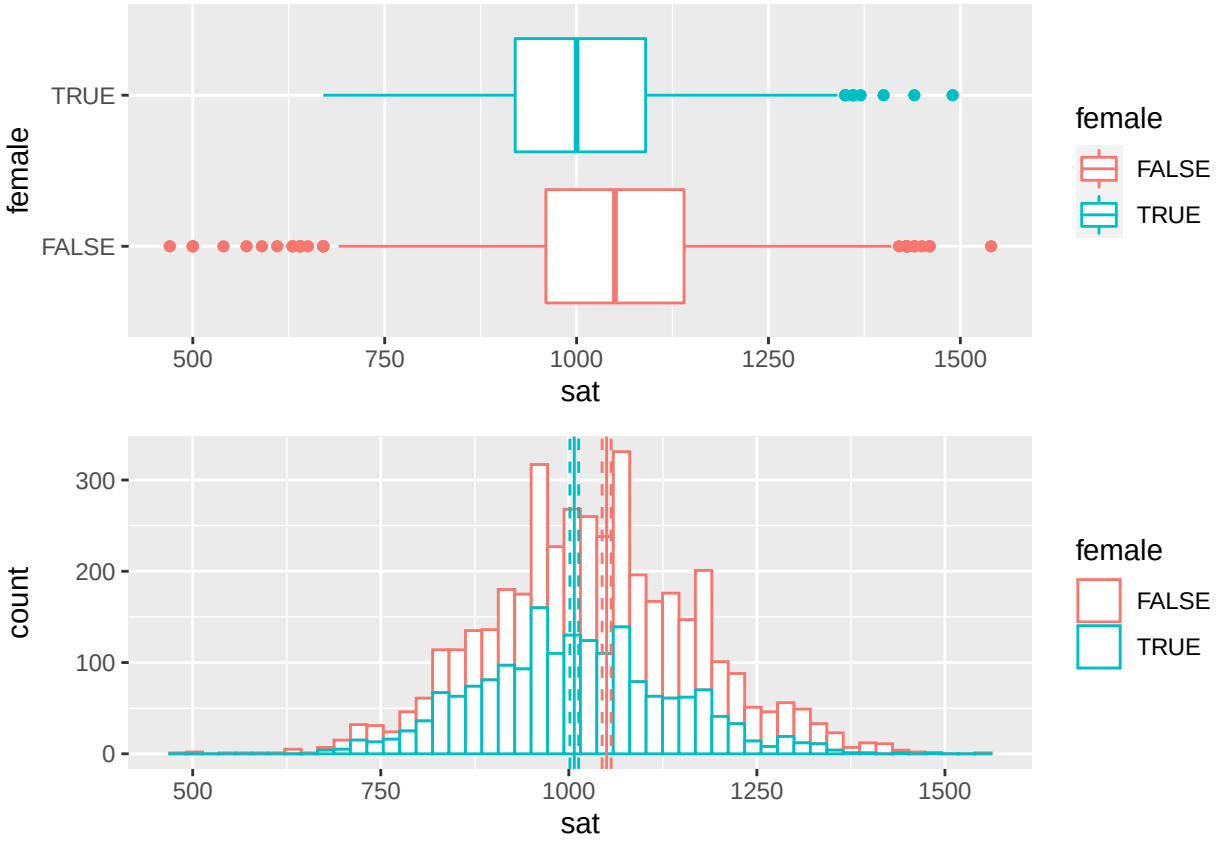
```
IC_sat_dones<-IC(gpaclean$sat[gpaclean$female == TRUE], 0.95)
IC_sat_homes<-IC(gpaclean$sat[gpaclean$female == FALSE], 0.95)

IC_sat_dones
```

```
## [1] 1001.409 1013.110
```

```
IC_sat_homes
```

```
## [1] 1044.253 1056.244
```

A continuació es representen gràficament les distribucions de la variable *sat* per als grups *homes* i *dones* utilitzant boxplot i histogrames. Adicionalment en els histogrames s'han afegit els valors de la mitjana i l'IC al 95% de la mitjana poblacional per cada grup.



Cal destacar que l'interval de confiança s'ha calculat per la mitjana mentre que el gràfic boxplot assenyala la mediana. En el cas dels homes, mitjana i mediana son força semblants (mitjana = 1050.25 i mediana = 1050), per contra en el cas de les dones aquests dos paràmetres son lleugerament diferents (mitjana = 1007.26 i mediana = 1000).

Quina conclusió es pot extreure de la comparació dels dos intervals, en relació a si existeix solapament o no en els intervals de confiança? Justifiqueu la resposta.

Observant valors i representacions gràfiques es pot concloure que amb un nivell de confiança del 95% les mitjanes poblacionals per a homes i dones **NO** es solapen.

3.2. Contrast d'hipòtesi per a la diferència de mitjanes de colgpa

Volem analitzar si la nota mitjana del primer semestre és diferent per a les dones i els homes utilitzant un nivell de confiança 95%.

3.2.1. Pregunta de recerca

Formuleu la pregunta de recerca.

La mitja de notes mitjanes de les *dones* és diferent a la mitja de notes mitjanes dels *homes*?

3.2.2. Escriviu la hipòtesi nul · la i l'alternativa.

$$H_0 : \mu_{dones} = \mu_{homes}$$

$$H_1 : \mu_{dones} \neq \mu_{homes}$$

3.2.3. Justificació del test a aplicar

El test a aplicar és per a **dues mostres independents, sobre la mitjana amb variàncies desconegudes**.

En aquest moment, però, no sabem si les variàncies son desconegudes però iguals o bé, desconegudes i diferents. Per aquest motiu, aplicarem inicialment un test d'igualtat de variàncies.

Test d'igualtat de variàncies

$$H_0 : \sigma^2_{dones} = \sigma^2_{homes}$$
$$H_1 : \sigma^2_{dones} \neq \sigma^2_{homes}$$

Funció var_test per a calcular el test de variança:

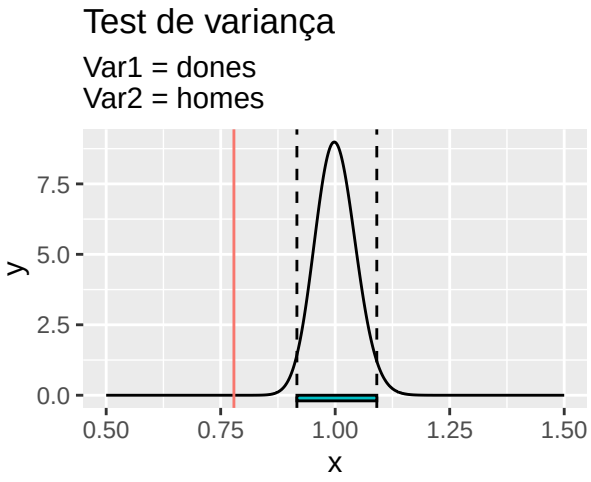
```
var_test<-function(m1, m2, NC){
  alfa<-1-NC
  meanX<-mean(m1); meanY<-mean(m2)
  nX<-length(m1); nY<-length(m2)
  sX<-sd(m1); sY<-sd(m2)

  fobs<-sX^2/sY^2
  fcritL<-qf(alfa/2, df1 = nX-1, df2 = nY-2)
  fcritU<-qf(1-alfa/2, df1 = nX-1, df2 = nY-2)
  pvalue<-2*min(pf(fobs, df1 = nX-1, df2 = nY-2, lower.tail = FALSE),
                pf(fobs, df1 = nX-1, df2 = nY-2))
  return(data.frame(fobs, fcritL, fcritU, pvalue, nX, nY))
}
```

Aplicació de la funció definida anteriorment:

```
var_sat_HD<-var_test(gpaclean$sat[gpaclean$female == TRUE],
                     gpaclean$sat[gpaclean$female == FALSE],
                     0.95)
kable(var_sat_HD)
```

fobs	fcritL	fcritU	pvalue	nX	nY
0.7788633	0.916399	1.090809	0	1843	2253

Graficant els valors resultat obtenim:



S'obté que la f_{obs} està fora de l'interval d'acceptació d'H_0, i que el valor p és inferior al nivell de significança; per tant rebutjem la hipòtesi nul · la en favor de l'alternativa i confirmem que les variàncies son diferents.

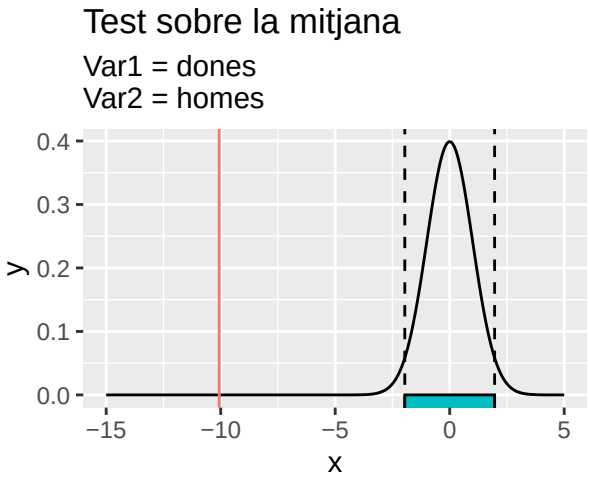
Així doncs, el test a aplicar és un **test bilateral** per a **dues mostres independents, sobre la mitjana amb variàncies desconegudes i diferents**.

Realitzeu els càlculs de l'estadístic de contrast, valor crític i p valor a un nivell de confiança del 95%.

Funció e_contrast per a calcular el test bilateral sobre la mitjana de dues mostres independents amb variances desconegudes i independents:

```
e_contrast<-function(m1, m2, NC){

  alfa<-1-NC
  meanX<-mean(m1); meanY<-mean(m2)
  nX<-length(m1); nY<-length(m2)
  sX<-sd(m1); sY<-sd(m2)

  v<-((((sX^2)/nX)+((sY^2)/nY))^2)/(((((sX^2)/nX)^2)/(nX-1))+((((sY^2)/nY)^2)/(nY-1)))

  tobs<-(meanX-meanY)/sqrt((sX^2/nX)+(sY^2/nY))

  tcritL<-qt(alfa/2, v)
  tcritU<-qt(1-alfa/2, v)
  pvalue<-pt(abs(tobs), df = v, lower.tail = FALSE)*2

  tobs<-round(tobs,4)
  tcritL<-round(tcritL,4)
  pvalue<-round(pvalue,4)

  return(data.frame(tobs,tcritL,tcritU,pvalue,v))
}

sat_HD<-e_contrast(gpaclean$sat[gpaclean$female == TRUE],
                gpaclean$sat[gpaclean$female == FALSE],
                0.95)
kable(sat_HD)
```

tobs	tcritL	tcritU	pvalue	v
-10.0644	-1.9605	1.960547	0	4070.484



3.2.5. Interpretació del test

En la població, la mitja de notes mitjanes de les *dones* **SI** és diferent a la mitja de notes mitjanes dels *homes* amb un nivell de confiança del 95%, donat que la t_{obs} està fora de l'interval d'acceptació d'H_0 i el valor p és inferior al nivell de significança α per tant podem rebutjar la hipòtesi nul · la en favor de l'alternativa. t_{obs}=-10.0644, interval d'acceptació d'H_0=[-1.9605, 1.9605].

4. Model de regressió lineal

Estimeu un model de regressió lineal múltiple que tingui com a variables explicatives: sat, female, tothrs, athlete, i hsperc, i com a variable dependent colgpa.

```
lm_colgpa<-lm(formula=colgpa~sat+female+tothrs+athlete+hsperc, data=gpaclean)
summary(lm_colgpa)
```

```
##
## Call:
## lm(formula = colgpa ~ sat + female + tothrs + athlete + hsperc,
##     data = gpaclean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64634 -0.36187  0.02472  0.38901  1.91689
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.034e+00  7.728e-02  13.375  < 2e-16 ***
## sat          1.637e-03  6.685e-05  24.488  < 2e-16 ***
## femaleTRUE   1.522e-01  1.805e-02   8.435  < 2e-16 ***
## tothrs       1.893e-03  2.460e-04   7.694 1.77e-14 ***
## athleteTRUE  1.479e-01  4.248e-02   3.480 0.000506 ***
## hsperc      -1.259e-02  5.637e-04 -22.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5531 on 4090 degrees of freedom
## Multiple R-squared:  0.299,  Adjusted R-squared:  0.2981
## F-statistic: 348.8 on 5 and 4090 DF,  p-value: < 2.2e-16
```

4.1. Interpretació del model

Interpreteu el model lineal ajustat:

Com es pot observar en el sumari anterior, totes les variables explicatives son significatives ja que el valor p associat a l'estadístic és molt inferior al nivell se significança α.

Quina és la qualitat de l'ajust?

```
summary(lm_colgpa)$adj.r.squared
```

```
## [1] 0.2981
```

La qualitat de l'ajust és pobre doncs tan sols el 29.81% de la variabilitat de *colgpa* és explicada pel model.

Expliqueu la contribució de les variables explicatives

```
coefficients(lm_colgpa)
```

```
##  (Intercept)          sat   femaleTRUE       tothrs  athleteTRUE       hsperc
##  1.033556328  0.001637099  0.152209955  0.001893107  0.147854502 -0.012589624
```

Com es pot observar en la taula de coeficients, les variables *sat*, *female=TRUE*, *tothrs* i *athlete=TRUE* tenen una correlació positiva amb la variable *colgpa*, és a dir un increment en aquestes variable suposa un increment en la resposta del model i per tant un valor de *colgpa* predit més elevat. Per altra banda, la variable *hsperc* té una correlació negativa, és a dir un increment en el valor d'aquesta variables suposa un decrement en el valor predit de *colgpa*.

Així doncs, si tenim en compte que la variable *sat* és la nota d'accés, la variable *tothors* és el total d'hores cursades i la variable *hsperc* és el rànquing relatiu de l'alumne en percentatge; els signes dels coeficients del model semblen lògics.

4.2. Predicció

Independentment del R2 obtingut a l'apartat previ, apliqueu el model de regressió per a predir la nota mitjana d'un estudiant home, atleta, amb una nota d'entrada de 800, un total d'hores al semestre de 60 i una posició relativa al rànquing del 60%.

```
predict(lm_colgpa, newdata = data.frame(sat = 800,
                                        female = FALSE,
                                        tothrs = 60,
                                        athlete = TRUE,
                                        hsperc = 60))
```

```
##        1
## 1.849299
```

D'acord al model l'estudiant obtindrà una nota mitjana de 1.85

5. Regressió logística

5.1. Estimació del model

Estimeu un model logístic per a predir la probabilitat de ser un estudiant excel · lent al final del primer semestre a la universitat en funció de les variables: female, athlete, sat, tothrs, black, white i hsperc.

Inicialment es comproven els nivells assignats a les variables categòriques:

```
contrasts(gpaclean$female)
```

```
##       TRUE
## FALSE    0
```

```
## TRUE     1
```

```
contrasts(gpaclean$athlete)
```

```
##       TRUE
## FALSE    0
## TRUE     1
```

```
contrasts(gpaclean$black)
```

```
##       TRUE
## FALSE    0
## TRUE     1
```

```
contrasts(gpaclean$white)
```

```
##       TRUE
## FALSE    0
## TRUE     1
```

I posteriorment s'estima el model de regressió logística:

```
logi_excelent<-glm(formula = excelente~female+athlete+sat+tothrs+black+white+hsperc,
                   data = gpaclean,
                         family = binomial(link = logit))
summary(logi_excelent)
```

```
##
## Call:
## glm(formula = excelente ~ female + athlete + sat + tothrs + black +
##     white + hsperc, family = binomial(link = logit), data = gpaclean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8086  -0.4554  -0.2381  -0.0884   3.4703
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.5496578  0.7080393 -10.663  < 2e-16 ***
## femaleTRUE   0.4248932  0.1185659   3.584 0.000339 ***
## athleteTRUE -0.0067452  0.4194483  -0.016 0.987170
## sat          0.0062877  0.0004915  12.794  < 2e-16 ***
## tothrs      -0.0050991  0.0016394  -3.110 0.001868 **
## blackTRUE   -0.9086047  0.5345492  -1.700 0.089176 .
## whiteTRUE   -0.0666162  0.4017860  -0.166 0.868314
## hsperc      -0.1041352  0.0084606 -12.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2869.6  on 4095  degrees of freedom
## Residual deviance: 2100.0  on 4088  degrees of freedom
## AIC: 2116
##
## Number of Fisher Scoring iterations: 7
```

5.2. Interpretació del model estimat

Interpreteu els resultats obtinguts. Concretament, analitzeu la significativitat de les variables explicatives i expliqueu la seva contribució per predir la probabilitat de ser un estudiant excel · lent.

Observant els valor p de la taula sumari obtenim que les variables *athlete=TRUE* i *white=TRUE* no son significatives doncs el seu p valor és clarament superior a un nivell de significança α=0.05. Adicionalment podem dir que la variable *black=TRUE* tampoc és significativa doncs el seu p valor és lleugerament superior al nivell de significança citat anteriorment.

En relació als signes dels coeficients estimats i dels valors de l'estadístic (z value) podem afirmar que les condicions *female=TRUE* i/o l'increment de *sat* contribueixen positivament (incrementen) la probabilitat de ser un estudiant excel · lent, mentre que la resta de variables en redueixen la probabilitat doncs tenen signes negatius.

Les variables amb un pes més elevat son *sat* i *hsperc* ja que tenen els valors de l'estadístic (z value) més alts. La primera incrementa la probabilitat de ser excel · lent (coeficient positiu) mentre que la segona en redueix la probabilitat (coeficient negatiu).

5.3. Importància de ser dona

Al model anterior, interpreteu els nivells de la variable female a partir del odds ratio. En quin percentatge es veu augmentada la probabilitat de ser un estudiant excel · lent si ets dona? Proporcioneu intervals de confiança del 95% dels odds ratio.

```
round(exp(coefficients(logi_excelent)),4)
```

```
## (Intercept)  femaleTRUE athleteTRUE         sat      tothrs   blackTRUE
##      0.0005      1.5294      0.9933      1.0063      0.9949      0.4031
##   whiteTRUE      hsperc
##      0.9356      0.9011
```

El fet de ser dona increment la probabilitat de ser un estudiant excel · lent en un 52.94% respecte a ser home.

Els intervals de confiança del 95% per als *odds ratio* són:

```
round(exp(confint(logi_excelent, level = 0.95)),4)
```

```
## Waiting for profiling to be done...
```

```
##                2.5 % 97.5 %
## (Intercept) 0.0001 0.0021
## femaleTRUE  1.2132 1.9314
## athleteTRUE 0.4068 2.1420
## sat         1.0053 1.0073
## tothrs      0.9917 0.9981
## blackTRUE   0.1399 1.1643
## whiteTRUE   0.4444 2.1804
## hsperc      0.8858 0.9157
```

Continuant amb la importància de ser dona, podem afirmar amb un nivell de confiança del 95% que la probabilitat de ser un estudiant excel · lent si l'estudiant és dona incrementa entre un 21.32% i un 93.14% respecte a si l'estudiant és home.

5.4. Predicció

Amb quina probabilitat una estudiant dona, no atleta, amb un sat de 1200 punts, 50 hores cursades, de raça negra i amb un rànquing relatiu (hsperc) del 10% serà excel · lent?

```
predict(logi_excelent, newdata = data.frame(female = TRUE,
                                            athlete = FALSE,
                                            sat = 1200,
                                            tothrs = 50,
                                            black = TRUE,
                                            white = FALSE,
                                            hsperc = 10),
        type="response")
```

```
##         1
## 0.1437584
```

La probabilitat que l'alumne amb les condicions citades sigui excel · lent és del 14.38%.

6. Anàlisi de la variància (ANOVA) d'un factor

Realitzarem un ANOVA per a contrastar si existeixen diferències a la variable colgpa en funció de la raça dels estudiants.

En primer lloc, a partir de les variables black i white creeu una variable categòrica denominada race, que indiqui la raça de l'estudiant en una d'aquestes tres categories: black, white i other (per a estudiants que no són de raça negra ni blanca).

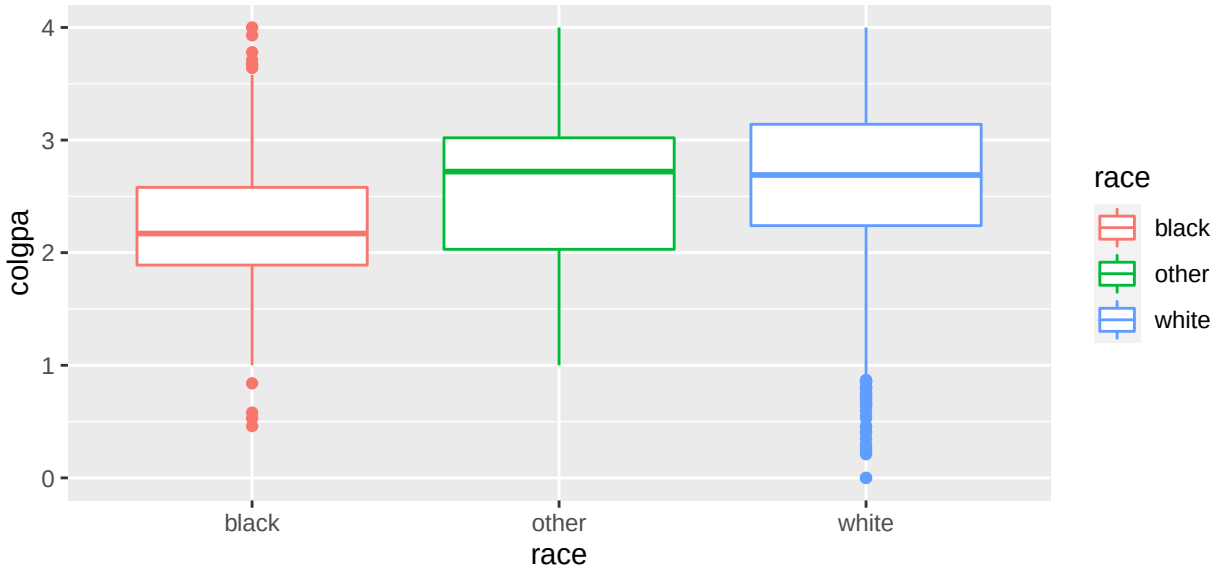
```
gpaclean$race<-as.factor(ifelse(gpaclean$white,"white",
                                ifelse(gpaclean$black,"black","other")))
summary(gpaclean$race)
```

```
## black other white
##   225    79  3792
```

6.1. Visualització gràfica

Mostreu gràficament la distribució de colgpa segons els valors de race.

```
ggplot(gpaclean, aes(x=race, y=colgpa, color=race)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 0))
```



```
mean(gpaclean$colgpa)
```

```
## [1] 2.654546
```

```
tapply(gpaclean$colgpa,gpaclean$race,mean)
```

```
##    black    other    white
## 2.248444 2.635190 2.679045
```

6.2. Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa.

$$H_0 : \alpha_{black} = \alpha_{other} = \alpha_{white}$$

$$H_1 : \alpha_i \neq \alpha_j \text{ per algun } i \neq j \text{ amb } i, j \in [black, white, other]$$

6.3. Model

Calculeu l'anàlisi de variància, fent servir la funció aov o lm. Interpreteu el resultat de l'anàlisi, tenint en compte els valors: Sum Sq, Mean SQ, F i Pr (> F).

```
mod<-aov(colgpa~race, data=gpaclean)
taov<-anova(mod)
taov
```

```
## Analysis of Variance Table
##
## Response: colgpa
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## race         2   39.41 19.7061  46.215 < 2.2e-16 ***
## Residuals 4093 1745.26  0.4264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En la taula sumari obtenim els següents valors tan per als tractament com per a l'error: els graus de llibertat (a-1 i N-a), les sumes de quadrats (SSA i SSE), els quadrats mitjans (MSA i MSE) el valor estadístic F i el valor p associat.

Observem que el valor de l'estadístic F és força més gran que la unitat fet que indica que MSA > MSE i per tant existeix algun $\alpha_i \neq 0$. De manera adicional, el valor p és clarament inferior a un nivell de significança α del 5% i conseqüentment acceptem la hipòtesi alternativa per concloure que el factor *race* és significatiu.

6.4. Efectes dels nivells del factor

Proporcioneu l'estimació de l'efecte dels nivells del factor race. Calculeu també la part de la variabilitat de colgpa explicada per l'efecte dels nivells.

```
m<-model.tables(mod, type="means")
m
```

```
## Tables of means
## Grand mean
##
## 2.654546
##
##  race
##       black  other    white
```

```
##       2.248  2.635    2.679
## rep 225.000 79.000 3792.000
```

```
e<-model.tables(mod, type="effects")
e
```

```
## Tables of effects
##
##  race
##        black    other     white
##      -0.4061 -0.01936    0.0245
## rep 225.0000 79.00000 3792.0000
```

En la taula de mitjanes obtenim els valors mitjans de la variable *colgpa* per a cada un dels tractaments.

En la taula d'efectes obtenim la diferencia en la mitjana de cada un dels tractaments α_i. Així doncs podem afirmar que l'efecte del tractament *black* en la mitjana de *colgpa* és -0.4061. Anàlogament l'efecte pel tractament *white* és 0.0245.

Per tal de conèixer la variabilitat de *colgpa* explicada per l'efecte dels nivells dividirem el valor de la suma de quadrats de la variable *race* per la suma de quadrats totals (variable *race* + residus).

```
(taov$"Sum Sq"[1]/(taov$"Sum Sq"[1]+taov$"Sum Sq"[2]))*100
```

```
## [1] 2.20838
```

Així doncs, el 2.21% de la variabilitat és explicada per l'efecte dels nivells.

6.5. Conclusió dels resultats del model ANOVA

Extraieu conclusions de l'ANOVA realitzat.

```
LSD.test(mod,"race",group=T,p.adj="bonferroni",console=T)
```

```
##
## Study: mod ~ "race"
##
## LSD t Test for colgpa
## P value adjustment method: bonferroni
##
## Mean Square Error:  0.4264006
##
## race,  means and individual ( 95 %) CI
##
##         colgpa       std    r      LCL      UCL  Min Max
## black 2.248444 0.6173364  225 2.163096 2.333793 0.46   4
## other 2.635190 0.6841556   79 2.491154 2.779226 1.00   4
## white 2.679045 0.6543850 3792 2.658256 2.699835 0.00   4
##
## Alpha: 0.05 ; DF Error: 4093
## Critical Value of t: 2.394964
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##         colgpa groups
## white 2.679045      a
## other 2.635190      a
```

```
## black 2.248444      b
```

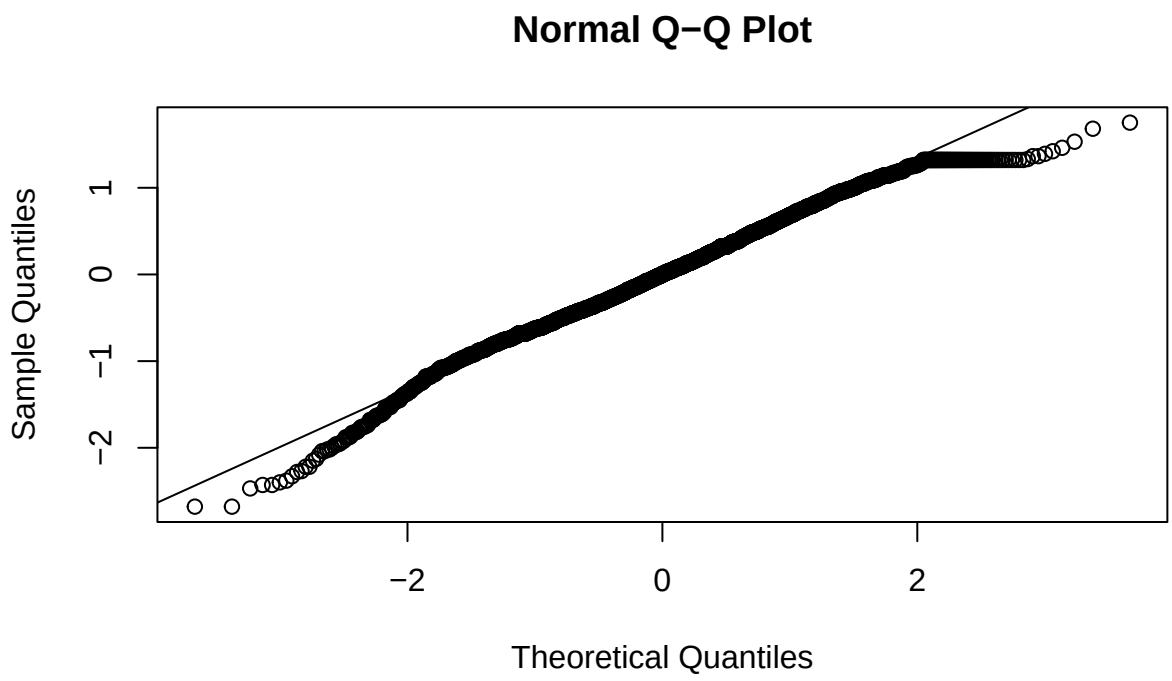
Amb l'anàlisi realitzat es pot concloure que:

- El factor *race* és significatiu.
- Els efectes dels tractaments son:
  - Black = -0.4061
  - White = 0.0245
  - Other = -0.0194
- El test LSD amb ajust de Bonferroni indica que els tractaments **white** i **other** no son significativament diferents entre ells (grup a), mentre que si ho son amb el tractament **black** (grup b).

6.6. Normalitat dels residus

Feu servir el gràfic Normal Q-Q i el test Shapiro-Wilk per avaluar la normalitat dels residus. Podeu fer servir les funcions de R corresponents per fer el gràfic i el test.

```
qqnorm(residuals(mod))
qqline(residuals(mod))
```



Observem que la majoria dels residus s'ajusten a la recta, especialment en la zona central del gràfic. De totes formes no sembla prou evidència per afirmar o rebutjar el supòsit de normalitat. Així doncs contrastarem la normalitat mitjançant el test de Shapiro-Wilk.

```
shapiro.test(residuals(mod))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod)
## W = 0.99175, p-value = 1.121e-14
```

En el test Shapiro-Wilk la hipòtesi nul · la afirma que la distribució és normal. En el càlcul s'obté un valor p molt petit (inferior a 0.05) per tant podem rebutjar la hipòtesi nul · la en favor de la alternativa i dir que la distribució no és normal.

Sota aquest supòsit hauriem d'escollir el test no-paramètric de Kruskal-Wallis doncs no necessita la assumpció de normalitat.

```
kruskal.test(colgpa~race, data=gpaclean)
```

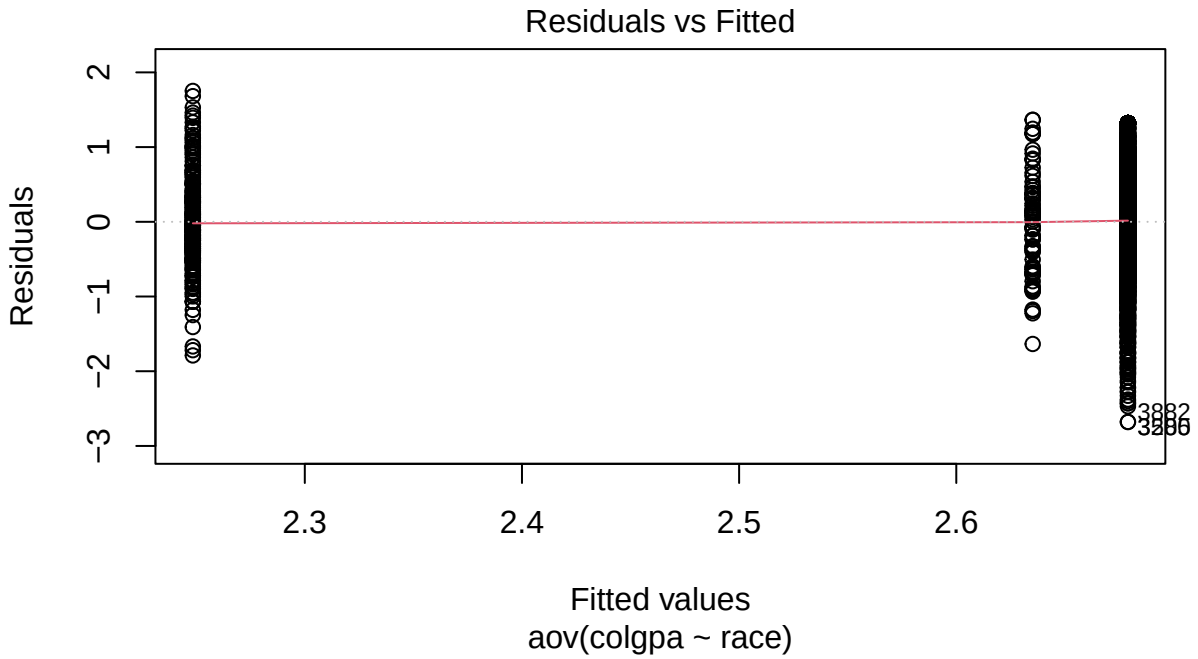
```
##
##  Kruskal-Wallis rank sum test
##
## data:  colgpa by race
## Kruskal-Wallis chi-squared = 98.671, df = 2, p-value < 2.2e-16
```

En aquest cas, obtenim un valor p molt petit, inferior al nivell de significança i per tant acceptem que hi ha diferències significatives en la variable *colgpa* segons el nivell del factor *race*.

6.6.1. Homoscedasticitat dels residus

El gràfic "Residuals vs Fitted" proporciona informació sobre la homoscedasticitat dels residus. Mostreu i interpreteu aquest gràfic.

```
plot(mod,which=1)
```



Observem que els residus estan alineats sobre 3 linies corresponents a les mitjanes de cada un dels tractaments. Pel que fa a la homoscedasticitat, la visualització del gràfic no resulta concloent, així doncs apliquem el test de Bartlett.

```
bartlett.test(colgpa~race, data=gpaclean)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  colgpa by race
## Bartlett's K-squared = 1.7404, df = 2, p-value = 0.4189
```

Obtenim un p-valor superior al nivell de significança, per tant no rebutjem la hipòtesi nul · la i acceptem que les variances son iguals.

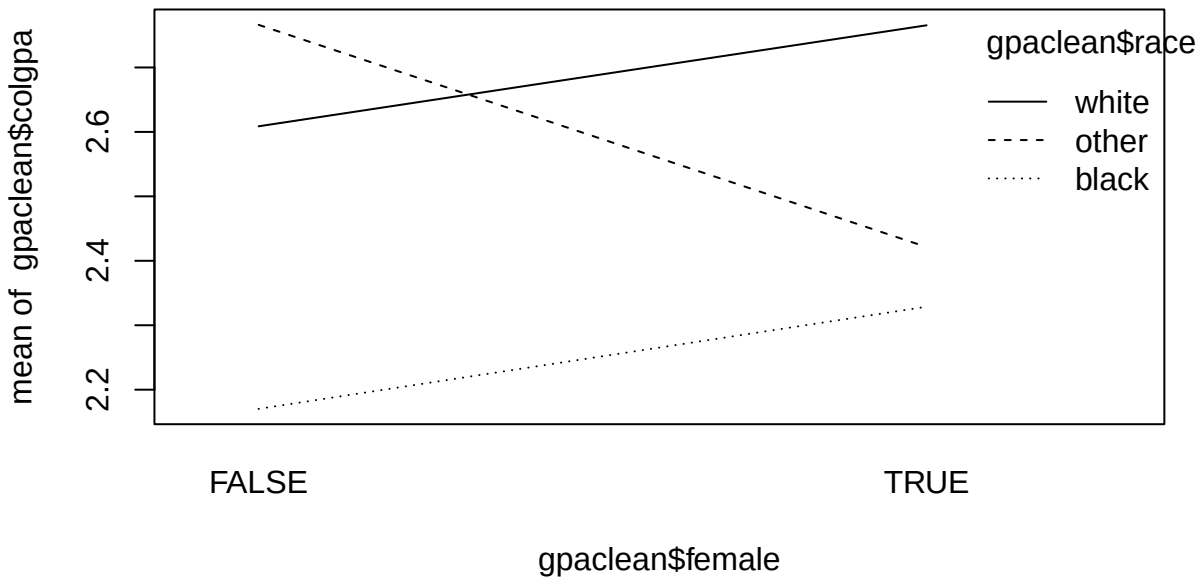
7. ANOVA multifactorial

A continuació, es desitja avaluar l'efecte sobre colgpa de la raça del estudiant combinada amb el factor gènere de l'estudiant (female). Seguiu els passos que s'indiquen a continuació.

7.1. Anàlisi visual dels efectes principals i possibles interaccions

Representeu la interacció dels dos factors race i female i comenteu els gràfics resultants.

```
gpaclean$female<-as.factor(gpaclean$female)
interaction.plot(gpaclean$female,gpaclean$race,gpaclean$colgpa)
```



En el gràfic s'observa que existeix interacció en els efectes dels factors *female* i *race* sobre la variable *colgpa* ja que la el nivell *TRUE* de la variable *female* incrementa la resposta de *colgpa* per als nivells *white* i *black*, en canvi disminueix la resposta de *colgpa* per al nivell *others* de la variable *race*.

7.2. Càlcul del model

Calculeu el model ANOVA multifactorial. Podeu fer servir la funció aov.

```
mod_m <- aov(colgpa~race*female, data = gpaclean)
anova(mod_m)
```

```
## Analysis of Variance Table
##
## Response: colgpa
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## race           2   39.41 19.7061 46.8976 < 2.2e-16 ***
## female         1   22.06 22.0627 52.5058 5.102e-13 ***
## race:female    2    4.60  2.2978  5.4685  0.004249 **
## Residuals   4090 1718.60  0.4202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.3. Interpretació dels resultats

Interpreteu els resultats obtinguts.

Observem en els resultat de l'ANOVA que els estadístics F dels factors principals (*race* i *female*) son significatius (p valor molt petit i inferior al nivell de significança), per tant acceptem que hi ha efecte de la raça i del sexe. També observem que la interacció *race:female* és significativa, fet que està alineat amb el que s'havia observat en els gràfics d'interacció.

Les mitjanes per nivells dels factors i interaccions són les següents:

```
model.tables(mod_m, type = "means")
```

```
## Tables of means
## Grand mean
##
## 2.654546
##
##  race
##       black  other    white
##       2.248  2.635    2.679
## rep 225.000 79.000 3792.000
##
##  female
##        FALSE     TRUE
##        2.588    2.736
## rep 2253.000 1843.000
##
##  race:female
##        female
## race    FALSE  TRUE
##   black    2.2    2.3
##   rep    114.0  111.0
##   other    2.8    2.4
##   rep     49.0   30.0
##   white    2.6    2.8
##   rep   2090.0 1702.0
```

Els efectes dels factors principals i de les interaccions són:

```
model.tables(mod_m, type = "effects")
```

```
## Tables of effects
##
##  race
##        black    other     white
##      -0.4061 -0.01936    0.0245
## rep 225.0000 79.00000 3792.0000
##
##  female
##          FALSE      TRUE
##       -0.06635 8.111e-02
## rep 2253.00000 1.843e+03
##
##  race:female
##        female
## race    FALSE  TRUE
##   black    0.0    0.0
##   rep    114.0  111.0
##   other    0.2   -0.3
##   rep     49.0   30.0
##   white    0.0    0.0
##   rep   2090.0 1702.0
```

Com que els factors principals són significatius realitzarem comparacions per parelles. Començarem per el factor *race*.

```
HSD.test(mod_m, "race", console = TRUE)
```

```
##
## Study: mod_m ~ "race"
##
## HSD Test for colgpa
##
## Mean Square Error:  0.4201954
##
## race,  means
##
##         colgpa       std    r  Min Max
## black 2.248444 0.6173364  225 0.46   4
## other 2.635190 0.6841556   79 1.00   4
## white 2.679045 0.6543850 3792 0.00   4
##
## Alpha: 0.05 ; DF Error: 4090
## Critical Value of Studentized Range: 3.315703
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##         colgpa groups
## white 2.679045      a
## other 2.635190      a
## black 2.248444      b
```

Observem que els nivells *white* i *other* son significativament poc diferents entre si i per tant formen un grup (a), mentre que si son significativament diferents amb el nivell *black* (group b). Aquestes són les mateixes conclusions que s'han obtingut en el model d'un factor.

No es necessari realitzar l'estudi per a la variable *female* ja que només té 2 nivells i com anteriorment hem conclòs que és significativa, cada un dels seus nivells serà significativament diferent de l'altre.

Finalment realitzem l'estudi de les interaccions doncs hem obtingut anteriorment que aquestes són significatives.

```
inter <- with(gpaclean, interaction(race, female))
mod_m_i<-aov(colgpa~inter, data = gpaclean)
HSD.test(mod_m_i, "inter", console = TRUE)
```

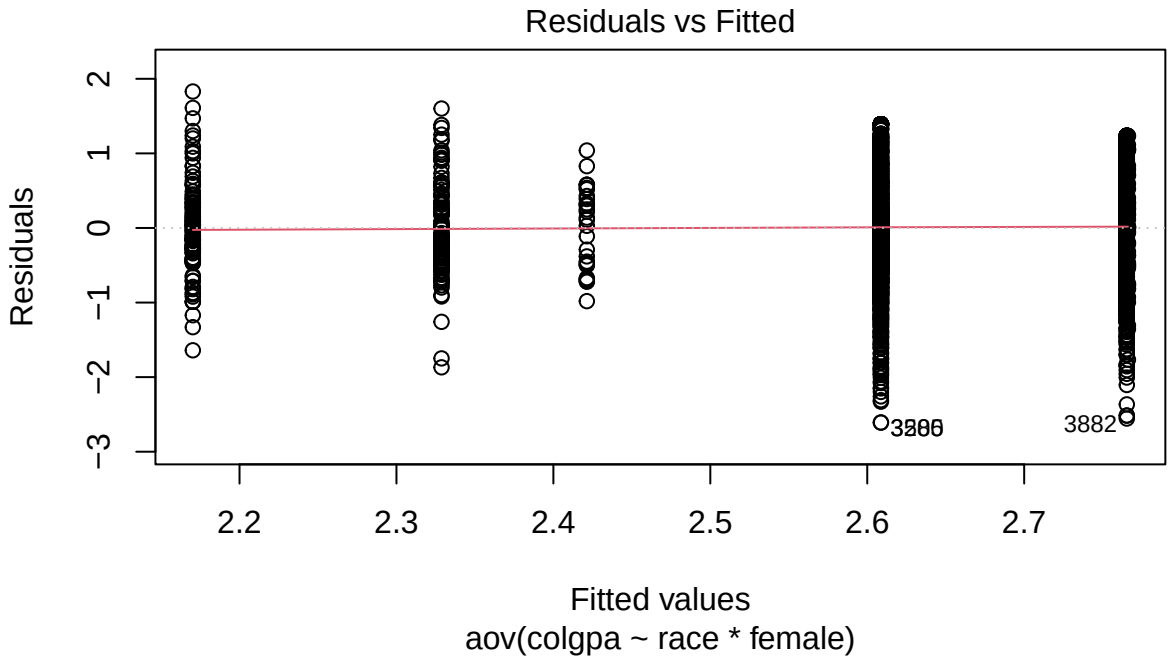
```
##
## Study: mod_m_i ~ "inter"
##
## HSD Test for colgpa
##
## Mean Square Error:  0.4201954
##
## inter,  means
##
##                colgpa       std    r  Min  Max
## black.FALSE 2.170263 0.6077777  114 0.53 4.00
## black.TRUE  2.328739 0.6194824  111 0.46 3.93
## other.FALSE 2.766122 0.7308067   49 1.00 4.00
## other.TRUE  2.421333 0.5465108   30 1.44 3.46
## white.FALSE 2.608713 0.6785843 2090 0.00 4.00
## white.TRUE  2.765411 0.6126297 1702 0.21 4.00
##
## Alpha: 0.05 ; DF Error: 4090
## Critical Value of Studentized Range: 4.032008
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##                colgpa groups
## other.FALSE 2.766122      a
## white.TRUE  2.765411      a
## white.FALSE 2.608713      a
## other.TRUE  2.421333     ab
## black.TRUE  2.328739      b
## black.FALSE 2.170263      b
```

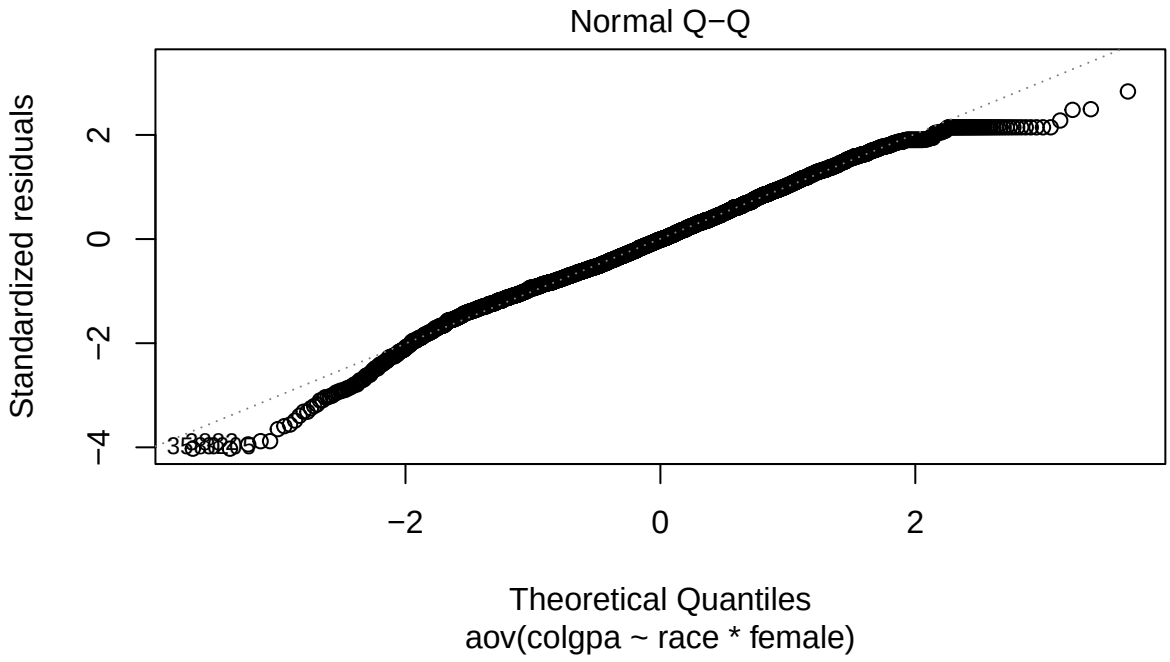
Observem que l'HSD test detecta 2 grups homogenis: el grup a format pels tractaments *other.male*, *white.female*, *white.male* i *other.female*; i el grup b format pels tractaments *other.female*, *black.female* i *black.male*. Observem també que el tractament *other.female* no és significativament diferent dels tractaments del grup a ni dels tractaments del grup b; tot i que els tractaments del grup a si que son significativament diferents als tractaments del grup b.

7.4. Adequació del model

Interpreteu l'adequació del model ANOVA obtingut utilitzant els gràfics dels residus.

```
plot(mod_m, which = c(1,2))
```





Analitzem la normalitat dels residus amb el test de Shapiro-Wilk

```
shapiro.test(residuals(mod_m))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod_m)
## W = 0.99267, p-value = 1.172e-13
```

Novament, obtenim un valor p molt petit (inferior a 0.05), rebutjem la hipòtesi nul · la en favor de la alternativa i podem dir que la distribució no és normal, per tant el model anova no és adequat.

8. Conclusions

Resumiu les conclusions principals de l'anàlisi (apartats 3 a 7). Per a això, podeu resumir les conclusions de cadascun dels apartats.

Estadística descriptiva i visualització

De la informació gràfica podem extreure:

- *sat* vs *female*: S'observen diferències en les medianes de *sat* segons la variable *female*. La mediana per al grup *female = False* (homes) és superior tot i que la dispersió, especialment en el rang baix de *sat* és superior.
- *sat* vs *athlete*: També s'observen diferències en les distribucions de *sat* segons la variable *athlete*. En general, la distribució per al grup *athlete = False* (no esportistes) té tots els valors (min, max, quartils i mediana) superiors.
- *sat* vs *gpaletter*: S'observa un cert nivell d'ordre en les distribucions de *sat* segons la variable *gpaletter*. En general, tots els valors (min, max, quartils i mediana) estan endreçats seguint A > B > C > D. Més enllà caldria destacar que el grup *gpaletter = C* podria presentar una dispersió superior a la resta de grups especialment en el rang inferior de *sat*

Estadística inferencial

Podem concloure (amb un nivell de confiança del 95%) que:

- La mitjana poblacional de la variable *sat* estarà dins l'interval [1026.637, 1035.174].
- La mitjana poblacional per a les dones de la variable *sat* estarà dins l'interval [1001.409, 1013.11].
- La mitjana poblacional per als homes de la variable *sat* estarà dins l'interval [1044.253, 1056.244].
- En la població, la mitja de notes mitjanes de les dones **SI** és diferent a la mitja de les notes mitjanes dels homes donat que la t_{obs} està fora de l'interval d'acceptació d'H_0 i el valor p és inferior al nivell de significança α. Conseqüentment podem rebutjar la hipòtesi nul · la en favor de l'alternativa.

Model de regressió lineal

- Les variables explicatives *sat*, *female*, *tothrs*, *athlete* i *hsperc* són significatives per explicar la variable dependent *colgpa*
- La qualitat del model és pobre ja que només el 29.81% de la variabilitat de *colgpa* és explicada pel model.
- D'acord al model l'estudiant amb les característiques proposades obtindrà una nota mitjana de 1.85

Regressió logística

- Les variables *athlete=TRUE* i *white=TRUE* no són significatives doncs el seu p valor és clarament superior a un nivell de significança α=0.05. Adicionalment, la variable *black=TRUE* tampoc és significativa doncs el seu p valor és lleugerament superior al nivell de significança citat anteriorment.

- Les condicions *female=TRUE* i/o l'increment de *sat* contribueixen positivament (incrementen) la probabilitat de ser un estudiant excel · lent, mentre que la resta de variables redueixen la probabilitat doncs tenen signes negatius.
- Les variables amb un pes més elevat son *sat* i *hsperc* doncs tenen els valors de l'estadístic (z value) més alts.
- El fet de ser dona incrementa la probabilitat de ser un estudiant excel · lent en un 52.94% respecte a ser home.
- Podem afirmar amb un nivell de confiança del 95% que la probabilitat de ser un estudiant excel · lent si l'estudiant és dona incrementa entre un 21.32% i un 93.14% respecte a si l'estudiant és home.
- La probabilitat que l'alumne amb les condicions citades sigui excel · lent és del 14.38%.

Anàlisi de la variància (ANOVA) d'un factor

- Podem concloure que el factor *race* és significatiu ja que l'estadístic F és força més gran que la unitat, indicant que MSA > MSE i per tant existeix algun $\alpha_i \neq 0$. De manera adicional, el valor P és clarament inferior a un nivell de significança α del 5%.
- Els efectes dels tractaments son:
  - Black = -0.4061
  - White = 0.0245
  - Other = -0.0194
- El test LSD amb ajust de Bonferroni indica que els tractaments **white** i **other** no son significativament diferents entre ells (grup a), mentre que si ho son amb el tractament **black** (grup b).
- En l'anàlisi de normalitat de residus (Shapiro-Wilk) obtenim un valor p molt petit fet que ens indica que la distribució no és normal, per tant no es compleixen els supòsits per al test anova i procedim a l'anàlisi a través del test de Kruskal-Wallis. Aquest darrer ens dona un p-valor molt petit confirmant que existeixen diferencies significatives en la variable *colgpa* segons el nivell del factor *race*.
- Realitzem l'anàlisi d'homoscedasticitat a través del test de Bartlett i concloem que les variances son iguals.

ANOVA multifactorial

- Gràficament s'observa interacció en els efectes dels factors *female* i *race*.
- En el model ANOVA s'observa que tant els factors principals com la interacció són significatius.
- Analitzant per parelles obtenim:
  - Els nivells del factor *race*: *white* i *other* son significativament poc diferents entre si (formen el grup a), alhora són significativament diferents amb el nivell *black* (grup b).
  - Els 2 nivells del factor *female* són significativament diferents.
  - En l'anàlisi d'interacció entre els dos factors obtenim 2 grups homogenis:
    - El grup a format pels tractaments *other.male*, *white.female*, *white.male* i *other.female*
    - El grup b format pels tractaments *other.female*, *black.female* i *black.male*.
- Aplicant el test de normalitat de Shapiro-Wilk als residus obtenim un valor p molt petit (inferior a 0.05). Llavors rebutjem la hipòtesi nul · la en favor de la alternativa i afirmem que la distribució dels residus no és normal, per tant el model anova no és adequat.