A4 - Análisis de la varianza y repaso del curso

Solución

Semestre 2021.2

Índice

Introducción

El conjunto de datos es `gpa.csv`. Este conjunto de datos contiene la nota media de estudiantes universitarios tras el primer semestre de clases (GPA: grade point average, en inglés), así como información sobre la nota de acceso, la cohorte de graduación en el instituto y algunas características de los estudiantes.

Este conjunto de datos surge de una encuesta realizada a una muestra representativa de estudiantes de una universidad de EEUU (por razones de confidencialidad el conjunto de datos no incluye el nombre de la universidad). Las variables incluidas en el conjunto de datos son:

- sat: nota de acceso (medida en escala de 400 a 1600 puntos)
- tothrs: horas totales cursadas en el semestre
- colgpa: nota media del estudiante al final del primer semestre (medida en escala de 0 a 4 puntos)
- athlete: indicador de si el estudiante practica algún deporte en la universidad
- hsize: numero total de estudiantes en la cohorte de graduados del bachillerato (en cientos)
- hsrank: ranking del estudiante, dado por la nota media del bachillerato, en su cohorte de graduados del bachillerato
- hsperc: ranking relativo del estudiante, en porcentaje (hsrank/hsize)
- female: indicador de si el estudiante es mujer
- white: indicador de si el estudiante es de raza blanca o no
- black: indicador de si el estudiante es de raza negra o no

El objetivo de esta actividad final es doble. En primer lugar, consolidar los conocimientos y competencias de preprocesado, análisis descriptivo, inferencia estadística y análisis de regresión. En segundo lugar, adquirir los conocimientos y competencias para llevar a cabo un análisis tipo ANOVA (análisis de la varianza).

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

---

1. Lectura del fichero y preparación de los datos

Leed el fichero `gpa.csv` y guardad los datos en un objeto denominado `gpa`. A continuación, verificad el tipo de cada variable. ¿Qué variables son de tipo numérico? ¿Qué variables son de tipo cualitativo?

Solución:

```
gpa<-read.csv("gpa.csv", sep=",", stringsAsFactors=TRUE, fileEncoding = "UTF-8")
```

```
  str(gpa)
```

```
## 'data.frame':    4137 obs. of  10 variables:
##  $ sat    : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs : Factor w/ 125 levels "100h","101h",..: 67 46 42 64 46 16 103 79 46 45 ...
##  $ 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 ...
```

Las variables `athlete`, `female`, `white` y `black` son cualitativas. El resto son cuantitativas.

1.1. Preparación de los datos

La variable `tothrs` está clasificada como character. Para poder trabajar con ella hay que convertirla en numérica, eliminando el texto "h" de los datos.

Solución:

```
head(gpa$tothrs,10)
```

```
##  [1] 43h  18h  14h  40h  18h  114h 78h  55h  18h  17h
## 125 Levels: 100h 101h 102h 103h 104h 105h 106h 107h 108h 109h 10h 110h ... 9h
```

```
gpa$tothrs <- as.numeric( trimws( sub('h', "", gpa$tothrs ) ) )
head(gpa$tothrs,10)
```

```
##  [1]  43  18  14  40  18 114  78  55  18  17
```

```
# Comprobación
class(gpa$tothrs)
```

```
## [1] "numeric"
```

```
summary(gpa$tothrs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   17.00   47.00   52.83   80.00  137.00
```

1.2. Valores ausentes

- Comprobad cuántas observaciones tienen valores ausentes y sacad conclusiones sobre cómo de serio es el problema de valores ausentes en estos datos.
- Eliminad los valores ausentes del conjunto de datos. Denominad al nuevo conjunto de datos 'gpaclean'. Nota: En el resto de apartados se usará el nuevo conjunto de datos 'gpaclean'.

Solución:

```
str(gpa)
```

```
## 'data.frame':    4137 obs. of  10 variables:
##  $ sat    : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs : num  43 18 14 40 18 114 78 55 18 17 ...
##  $ 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 ...
```

```
colSums(is.na(gpa))
```

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

```
compl2 <- complete.cases(gpa)
table(compl2)
```

```
## compl2
## FALSE  TRUE
##    41  4096
```

```
gpaclean<-gpa %>% drop_na

nrow(gpa)
```

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

```
nrow(gpaclean)
```

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

Respuesta:

El nuevo conjunto de datos tiene 4096 observaciones, mientras que el conjunto inicial tiene 4137 observaciones. Hay 41 observaciones con valores ausentes en `colgpa` (1 % de la muestra inicial) y, por tanto, eliminar estas observaciones no generará problemas serios de selección del conjunto de datos.

1.3. Equivalencia de la nota en letras

La variable `colgpa` contiene la nota numérica del estudiante. Cread una variable categórica denominada `gpaletter`, que indique la nota en letra de cada estudiante de la siguiente 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.

Solución:

```
gpanum <- gpaclean$colgpa
gpa_level<-c("D","C","B", "A")
classif <- ifelse( gpanum<=1.49, gpa_level[1],
                   ifelse(gpanum<=2.49, gpa_level[2],
                   ifelse(gpanum<=3.49, gpa_level[3],
                   gpa_level[4])))

gpaclean$gpaletter <- factor( classif, order=TRUE, levels=gpa_level)
```

```
#Comprobación
table(gpaclean$gpaletter)
```

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

```
sum(table(gpaclean$gpaletter))
```

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

```
sum(table(gpaclean$colgpa))
```

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

---

2. Estadística descriptiva y visualización

2.1. Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas). Mostrad el número de observaciones y el número de variables.

Solución:

```
# número de observaciones
  nrow(gpaclean)
```

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

```
# número de variables
  ncol(gpaclean)
```

```
## [1] 11
```

```
# nombre de variables
  colnames(gpaclean)
```

```
##  [1] "sat"       "tothrs"    "hsize"     "hsrank"    "hsperc"    "colgpa"
##  [7] "athlete"   "female"    "white"     "black"     "gpaletter"
```

```
# resumen
  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   D: 144
##  FALSE:304       FALSE:3871      C:1521
##  TRUE :3792      TRUE :225       B:1973
##                                  A: 458
##
##
```

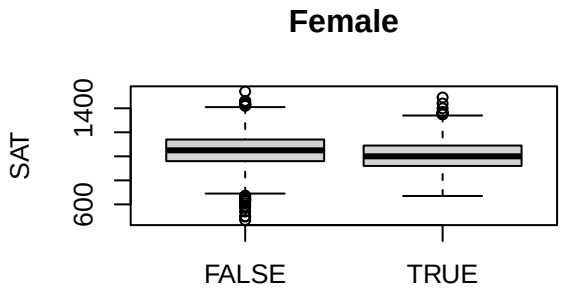
Hay 4096 observaciones y 11 variables.

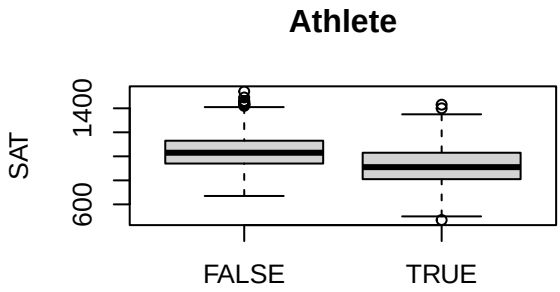
2.2. Visualización

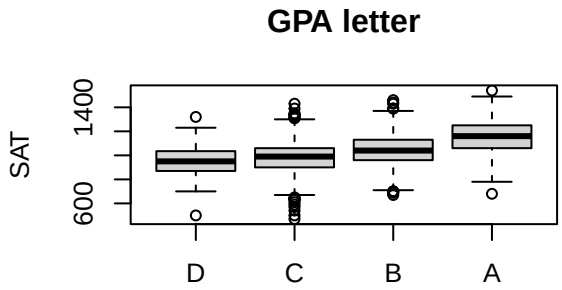
1. Mostrad con diversos diagramas de caja (boxplot) la distribución de la variable 'sat' según la variable 'female', según 'athlete', y según 'gpaletter'.
2. Cread una variable denominada 'excelente' que indique si el estudiante ha obtenido una A de nota media al final del semestre. Esta nueva variable debe codificarse como una variable dicotómica que toma el valor 1 cuando el estudiante ha obtenido una A, y el valor 0 en caso contrario. Realizad un gráfico que muestre el porcentaje de estudiantes excelentes.
3. Interpretad los gráficos brevemente.

Solución:

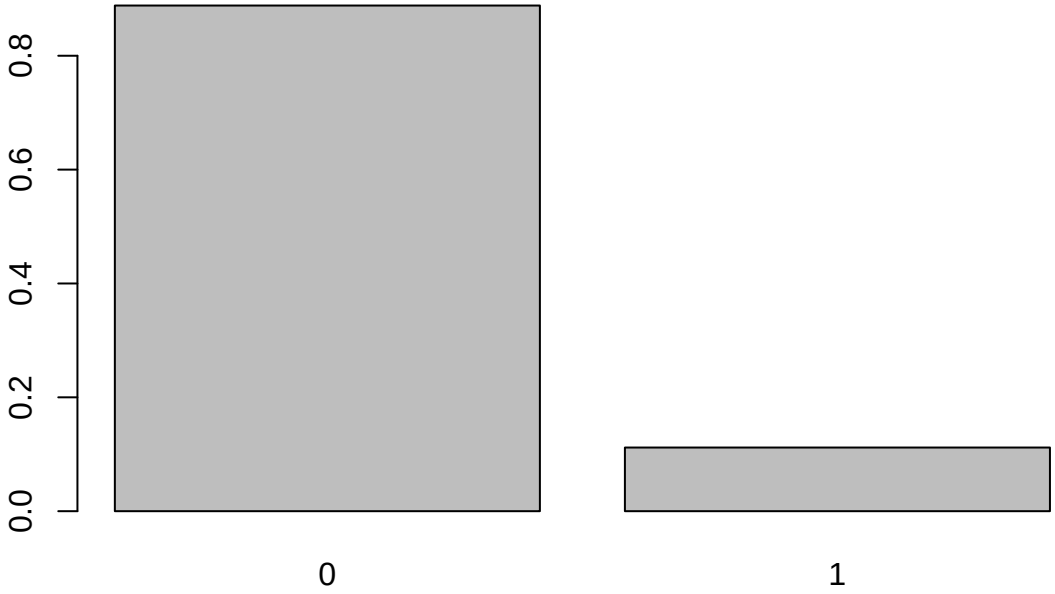
```
par(mfrow=c(2,2))
#female
boxplot(gpaclean$sat~gpaclean$female, main="Female",xlab="",ylab="SAT")

# athlete
boxplot(gpaclean$sat~gpaclean$athlete, main="Athlete", xlab="",ylab="SAT")

# gpaletter
boxplot(gpaclean$sat~droplevels(gpaclean$gpaletter), main="GPA letter", xlab="",ylab="SAT")
```



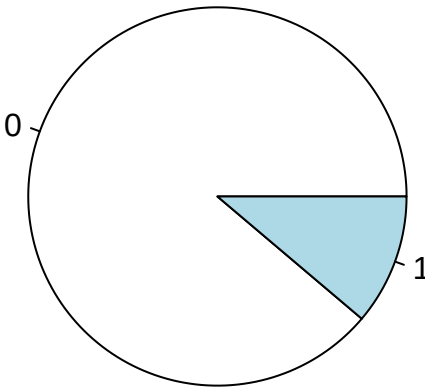





```
#excelente
gpaclean$excelente<-as.integer(ifelse(gpaclean$gpaletter=="A",1,0))
barplot( table(gpaclean$excelente)/length(gpaclean$gpaletter))
```




```
pie(table(gpaclean$excelente))
```



Breve interpretación de los gráficos:

- El valor mediano de la nota de acceso de las mujeres es algo inferior al de los hombres, aunque no se observan diferencias pronunciadas.
- Se observan diferencias en la nota de acceso en función de la variable 'athlete'. El valor mediano de la nota de acceso es mayor cuando el estudiante no practica ningún deporte en la universidad.
- Hay una asociación positiva entre la nota de acceso y la nota obtenida en el primer semestre: tendencia creciente del valor mediano de la nota de acceso con la nota en letra obtenida al final del semestre.
- En torno al 11 % de los estudiantes son excelentes.

---

3. Estadística inferencial

Utilizamos el conjunto de datos `gpaclean`.

3.1. Intervalo de confianza de la media poblacional de la variable `sat`

a) Calculad manualmente el intervalo de confianza al 95 % de la media poblacional de la variable `sat` de los estudiantes. Para ello, definid una función IC que reciba la variable, la confianza, y que devuelva un vector con los valores del intervalo de confianza. No se pueden utilizar funciones como `t.test` o `z.test` para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

Respuesta:

```
IC <- function( x, NC ){
  n <- length(x)
  alfa <- 1-(NC/100)
  sd <- sd(x)
  SE <- sd / sqrt(n)

  t <- qt( alfa/2, df=n-1, lower.tail=FALSE )
  L <- mean(x) - t*SE
  U <- mean(x) + t*SE
  return (c(L, U))
}
```

```r
ic95<-IC(gpaclean$sat, 95)
ic95
```

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

```r
# Comprobación
t.test(gpaclean$sat)$conf.int
```

```
## [1] 1026.637 1035.174
## attr(,"conf.level")
## [1] 0.95
```

El intervalo de confianza para `sat` es (1026.6, 1035.2), siendo la media muestral: 1030.91 puntos. La interpretación del intervalo de confianza es que si repitiésemos en un número elevado de muestras el mismo procedimiento, el 95 % de los intervalos obtenidos contendrían el valor de la media poblacional de la variable `sat`.

Nota: Por el teorema del límite central, podemos asumir que la media muestral sigue una distribución normal, puesto que tenemos una muestra de tamaño grande n=4096, y se podría usar la distribución normal para calcular el IC.

b) Calculad los intervalos de confianza al 95 % de la media poblacional de la variable `sat`, en función de si los estudiantes son hombres o mujeres. ¿Qué conclusión se puede extraer de la comparación de los dos intervalos, en relación a si existe solapamiento o no en los intervalos de confianza? Justificad la respuesta.

Respuesta:

```r
I <- which( gpaclean$female=="TRUE"  )
icfemale<-IC( gpaclean[I,"sat"], 95 )
icmale<-IC( gpaclean[-I,"sat"], 95 )
icfemale;icmale
```

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

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

```r
#Comprobación
t.test( gpaclean[gpaclean$female=="TRUE","sat"], conf.level=0.95 )$conf.int
```

```
## [1] 1001.409 1013.110
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test( gpaclean[gpaclean$female=="FALSE","sat"], conf.level=0.95 )$conf.int
```

```
## [1] 1044.253 1056.244
## attr(,"conf.level")
## [1] 0.95
```

Como los intervalos de confianza no están solapados podemos concluir que en promedio las mujeres tienen un `sat` menor que los hombres, con un nivel de confianza del 95 %.

3.2. Contraste de hipótesis para la diferencia de medias de `colgpa`

Queremos analizar si la nota media del primer semestre es diferente para las mujeres que para los hombres utilizando un nivel de confianza del 95 %.

Nota: se deben realizar los cálculos manualmente. No se pueden usar funciones de `R` que calculen directamente el contraste como `t.test` o similar. Sí se puede usar `var.test` y funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`.

Seguid los pasos que se detallan a continuación.

3.2.1. Pregunta de investigación

Formulad la pregunta de investigación.

Respuesta: ¿La nota media poblacional de las mujeres es diferente de la nota media poblacional de los hombres?

3.2.2. Escribid la hipótesis nula y la alternativa

Respuesta: Se trata de una comparación de medias en poblaciones independientes:

$H_0 : \mu_{female} = \mu_{male}$

$H_1 : \mu_{female} \neq \mu_{male}$

3.2.3. Justificación del test a aplicar

Respuesta:

Es un test de dos muestras sobre la media con varianzas desconocidas. Por el teorema del límite central, podemos asumir normalidad. Comprobamos igualdad de varianzas:

```
df_1<-gpaclean$colgpa[gpaclean$female=="TRUE"]
df_2<-gpaclean$colgpa[gpaclean$female=="FALSE"]

var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 0.82788, num df = 1842, denom df = 2252, p-value = 2.305e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7589643 0.9033950
## sample estimates:
## ratio of variances
##          0.8278771
```

El test var.test de R nos muestra un valor p menor de 0.05 por lo que rechazamos la hipótesis nula de igualdad de varianzas.

En consecuencia, aplicamos un test sobre la media de dos muestras independientes con varianza desconocida y diferente. Es un test bilateral.

3.2.4. Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95 %.

Respuesta:

```r
n1<-length(df_1)
n2<-length(df_2)


m1<-mean(df_1)
m2<-mean(df_2)


s1<-sd(df_1)
s2<-sd(df_2)

dfree<-(s1^2/n1+s2^2/n2)^2/(s1^4/(n1^2*(n1-1))+s2^4/(n2^2*(n2-1)))
dfree
```

```
## [1] 4047.939
```

```r
s<- sqrt( s1^2/n1 + s2^2/n2 )
t <- (m1-m2) / s
t
```

```
## [1] 7.029779
```

```r
pvalue<-pt( abs(t), df=dfree, lower.tail=FALSE )*2 #two sided
pvalue
```

```
## [1] 2.416939e-12
```

```r
#con nivel de confianza del 95%
t.crit95 <- qt( 0.05/2, df=dfree, lower.tail=FALSE ) #two sided
t.crit95
```

```
## [1] 1.96055
```

3.2.5. Interpretación del test

Respuesta:

El pvalor del test ($2.4169386 \times 10^{-12}$) es inferior al nivel de significación (0.05). Además el valor observado 7.0297786 es mayor que el valor crítico 1.9605502. Por tanto, podemos rechazar la hipótesis nula a favor de la alternativa y podemos concluir que en promedio la nota del semestre de las mujeres es diferente de la de los hombres.

Para comprobarlo podemos usar la función R t.test:

```r
t.test( df_1, df_2, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_1 and df_2
## t = 7.0298, df = 4047.9, p-value = 2.417e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1035218 0.1835970
## sample estimates:
## mean of x mean of y
##  2.733511  2.589951
```

4. Modelo de regresión lineal

Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: `sat`, `female`, `tothrs`, `athlete`, y `hsperc`, y como variable dependiente `colgpa`.

Solución:

```
modreg1 <- lm(colgpa~sat + female + tothrs + athlete + hsperc, gpaclean )

summary(modreg1)
```

```
##
## 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. Interpretación del modelo

Interpretad el modelo lineal ajustado:

– ¿Cuál es la calidad del ajuste?

– Explicad la contribución de las variables explicativas.

Respuesta:

```
summary(modreg1)
```

```
##
## 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
```

El valor del R^2 es 0.299. Es decir, el modelo explica un 30 % de la variación observada en la nota media de los estudiantes. Al ser un valor relativamente bajo, el modelo estimado no sería demasiado adecuado para predecir la nota media de los estudiantes. No obstante, en los modelos estimados con datos micro el R^2 suele ser bastante bajo pero los modelos estimados son igualmente válidos para analizar el efecto causal de variables explicativas en la dependiente. En este caso, tendremos que fijarnos en la significatividad de las variables independientes.

El pvalue de todas las variables explicativas es menor de 0.001 por lo que podemos rechazar la hipótesis nula en los tests de significatividad individual al 0.1 %. Concluimos que todas las variables son significativas. El pvalue del test de significatividad conjunta también es menor de 0.001 por lo que el conjunto de variables explicativas contribuye significativamente a explicar la nota media de los estudiantes.

Las variables `sat` y `tothrs` tienen una correlación positiva con la nota, indicando que cuanto mayores sean los valores de estas variables, mayor es la nota media en la universidad (manteniéndose constante el resto de variables incluidas en la regresión). Los coeficientes estimados de `female` y `athlete` también son positivos, indicando que si un estudiante es mujer o atleta, mayor será la nota media. La variable `hsperc` tiene una correlación negativa. Para interpretar esta variable hay que tener en cuenta que mayores valores del ranking indican peor desempeño en el instituto, de ahí que `hsperc` tenga una correlación negativa con `colgpa`.

4.2. Predicción

Independientemente del R^2 obtenido en el apartado previo, aplicad el modelo de regresión para predecir la nota media de un estudiante hombre, atleta, con una nota de entrada de 800, un total de horas en el semestre de 60 y una posición relativa en el ranking del 60 %.

Solución:

```
new<-data.frame(sat=800,tothrs=60, hsperc=60,
                female = FALSE, athlete= TRUE)

predict(modreg1,new,type="response")
```

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

El modelo predice que un estudiante con esas características tendrá una nota media en la universidad de 1.85 puntos. Esta predicción, no obstante, debe tomarse con cautela debido a que el R^2 es algo bajo.

---

5. Regresión logística

5.1. Estimación del modelo

Estimad un modelo logístico para predecir la probabilidad de ser un estudiante excelente al final del primer semestre en la universidad en función de las variables: `female`, `athlete`, `sat`, `tothrs`, `black`, `white` y `hsperc`.

Solución:

```
mod.log.1<-glm(excelente~female+athlete+sat+tothrs+black+white+hsperc, family = binomial(),data=gpaclean
summary(mod.log.1)
```

```
##
## Call:
## glm(formula = excelente ~ female + athlete + sat + tothrs + black +
##     white + hsperc, family = binomial(), 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. Interpretación del modelo estimado

Interpretad los resultados obtenidos. Concretamente, analizad la significatividad de las variables explicativas y explicad su contribución para predecir la probabilidad de ser un estudiante excelente.

Respuesta:

`female`, `sat` y `hsperc` son significativas al 0.1 %; `tothrs` es significativa al 1 %; `black` es significativa al 10 %. `athlete` y `white` no son variables significativas.

- La probabilidad de ser un estudiante excelente aumenta con la nota de acceso y si el estudiante es mujer.

- La probabilidad de ser un estudiante excelente disminuye con el número de horas cursadas, si es de raza negra y cuanto más lejos esté en el ranking relativo de su cohorte de graduados del instituto.
- La probabilidad de ser un estudiante excelente no se ve afectada significativamente por ser de raza blanca y por ser atleta en la universidad.

5.3. Importancia de ser mujer

En el modelo anterior, interpretad los niveles de la variable `female` a partir del **odds ratio**. ¿En qué porcentaje se ve aumentada la probabilidad de ser un estudiante excelente si se es mujer? Proporcionad intervalos de confianza del 95 % de los odds ratio.

Respuesta:

```
tt<-summary(mod.log.1)

tt$coefficients[2,c(1,2)]
```

```
##   Estimate Std. Error
##  0.4248932  0.1185659
```

```
#odds ratio e IC para female
od<-exp(tt$coefficients[2,1])
od
```

```
## [1] 1.529427
```

```
od.i<-exp(tt$coefficients[2,1]-1.96*tt$coefficients[2,2])
od.s<-exp(tt$coefficients[2,1]+1.96*tt$coefficients[2,2])
cbind(od.i,od,od.s)
```

```
##          od.i       od     od.s
## [1,] 1.212281 1.529427 1.929541
```

```
#Check
exp(coefficients(mod.log.1))
```

```
##  (Intercept)   femaleTRUE  athleteTRUE          sat       tothrs    blackTRUE
## 0.0005262902 1.5294270174 0.9932774752 1.0063075240 0.9949138729 0.4030862446
##    whiteTRUE       hsperc
## 0.9355542277 0.9011035003
```

```
exp(confint(mod.log.1))
```

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

```
##                    2.5 %      97.5 %
## (Intercept) 0.0001273844 0.002053433
## femaleTRUE  1.2131611597 1.931381653
## athleteTRUE 0.4068028772 2.141978981
## sat         1.0053495687 1.007289066
## tothrs      0.9916973267 0.998094205
## blackTRUE   0.1398572854 1.164258869
## whiteTRUE   0.4444296954 2.180390215
## hsperc      0.8857969193 0.915674886
```

Se observa que el odds ratio de la variable `female` es mayor que uno, con lo que la probabilidad de ser un estudiante excelente es mayor entre las mujeres que entre los hombres. En concreto, ser un estudiante excelente entre las mujeres es entre un 21 % y un 93 % más probable que entre los hombres.

5.4. Predicción

¿Con que probabilidad una estudiante mujer, no atleta, con un sat de 1200 puntos, 50 horas cursadas, de raza negra y con un ranking relativo (`hsperc`) del 10 % será excelente?

Respuesta:

```
new2<-data.frame(sat=1200,tothrs=50, hsperc=10,
                 female = TRUE, athlete= FALSE, black = TRUE, white= FALSE)
p1<-predict(mod.log.1,new2,type="response")
p1
```

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

La probabilidad de ser excelente es 0.1437584.

---

6. Análisis de la varianza (ANOVA) de un factor

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable `colgpa` en función de la raza de los estudiantes. Seguid los pasos que se indican.

En primer lugar, a partir de las variables `black` y `white` cread una variable categórica denominada `race`, que indique la raza del estudiante en una de estas tres categorías: `black`, `white` y `other` (para estudiantes que no son de raza negra ni blanca).

6.1. Visualización gráfica

Mostrad gráficamente la distribución de `colgpa` según los valores de `race`.

Solución:

```
# Variable race
race_level<-c("black","white","other")
race2 <- ifelse( gpaclean$black==TRUE, race_level[1],
                 ifelse(gpaclean$white==TRUE, race_level[2],
                 race_level[3]))

gpaclean$race <- factor( race2, levels=race_level)

#Comprobación
table(gpaclean$race)
```

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

```
sum(table(gpaclean$race))
```

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

```
nrow(gpaclean)
```

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

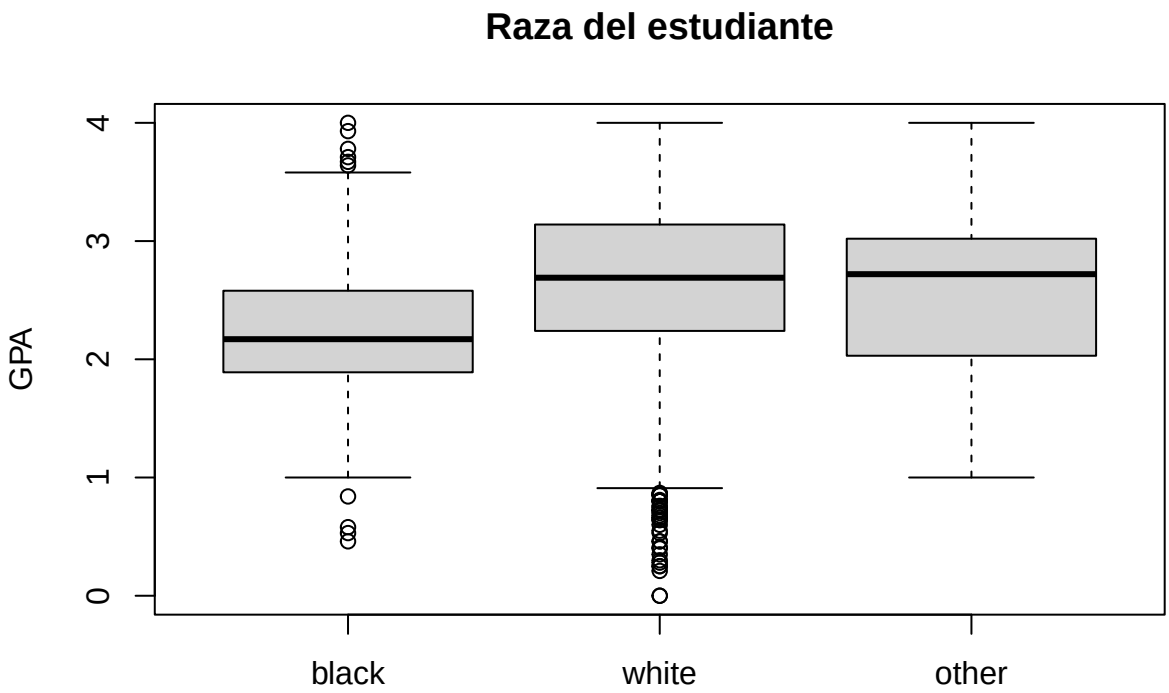
```
table(gpaclean$black)
```

```
##
## FALSE  TRUE
##  3871   225
```

```
table(gpaclean$white)
```

```
##
## FALSE  TRUE
##   304  3792
```

```
# Plotting
boxplot(gpaclean$colgpa~factor(gpaclean$race, order=F), main="Raza del estudiante",xlab="",ylab="GPA")
```



6.2. Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

Respuesta:

El factor `race` tiene 3 niveles: 1 es `black`, 2 es `white` y 3 es `other`. Las hipótesis son:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i \neq \mu_j \quad$ para algún i, j

donde μ_1, μ_2 y μ_3 denotan, respectivamente, la media poblacional de `colgpa` para los grupos de estudiantes de raza negra, blanca y otra.

6.3. Modelo

Calculad el análisis de varianza, usando la función `aov` o `lm`. Interpretad el resultado del análisis, teniendo en cuenta los valores: Sum Sq, Mean SQ, F y Pr (> F).

Respuesta:

```
#Usando aov
mod1 <- aov(colgpa ~ race, gpaclean)
kk <- summary( mod1 )
kk
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## race           2   39.4  19.706   46.22 <2e-16 ***
## Residuals   4093 1745.3   0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kk[[1]][1,4]
```

```
## [1] 46.21509
```

```
# Usando lm
mod2<-lm(colgpa ~ race,data=gpaclean)
anova(mod2)
```

```
## 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
```

Valores del contraste: Sum Sq = 39.41; Mean Sq = 19.71; estadístico F = 46.22; pvalor = $1.4200745 \times 10^{-20}$. El pvalor es menor que 0.05 y la conclusión es, por tanto, que el factor analizado es significativo. En conclusión, en este caso, rechazamos la hipótesis nula de igualdad de medias entre los tres grupos.

6.4. Efectos de los niveles del factor

Proporcionad la estimación del efecto de los niveles del factor `race`. Calculad también la parte de la variabilidad de `colgpa` explicada por el efecto de los niveles.

Solución:

El factor `race` tiene 3 niveles: black, white y other. Estimamos el efecto de cada uno de los niveles:

```
#Media global
mu<-mean(gpaclean$colgpa)
mu
```

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

```
#Efecto estimado de cada nivel de `race`:
alpha1<-mean(gpaclean$colgpa[gpaclean$race=="black"])-mu
alpha2<-mean(gpaclean$colgpa[gpaclean$race=="white"])-mu
alpha3<-mean(gpaclean$colgpa[gpaclean$race=="other"])-mu
alpha1 ; alpha2; alpha3
```

```
## [1] -0.4061015
```

```
## [1] 0.02449946
```

```
## [1] -0.01935603
```

```
#Alternativamente:
model.tables(mod1, type = "effects")
```

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

Calculamos la parte de la variabilidad de `colgpa` explicada por el efecto de los niveles:

```
eta<-kk[[1]][1,2]/(kk[[1]][2,2]+kk[[1]][1,2])
eta
```

```
## [1] 0.0220838
```

El efecto de los niveles explica el 2 % de la variabilidad observada en `colgpa`. Es decir, el 2 % de la variabilidad total observada en `colgpa` se debe a la variabilidad observada entre las tres categorías.

6.5. Conclusión de los resultados del ANOVA

Sacad conclusiones del ANOVA realizado.

Respuesta:

De los resultados obtenidos podemos concluir que no se puede aceptar la hipótesis nula de igualdad de medias entre los grupos dados por el factor `race`. `race` tiene un efecto significativo sobre `colgpa` (pvalor menor que 0.05) y, por tanto, la raza del estudiante determina diferencias en la nota media de los estudiantes. El efecto de `race` es positivo para los estudiantes de raza blanca mientras que es negativo para los de raza negra y para los de otras razas.

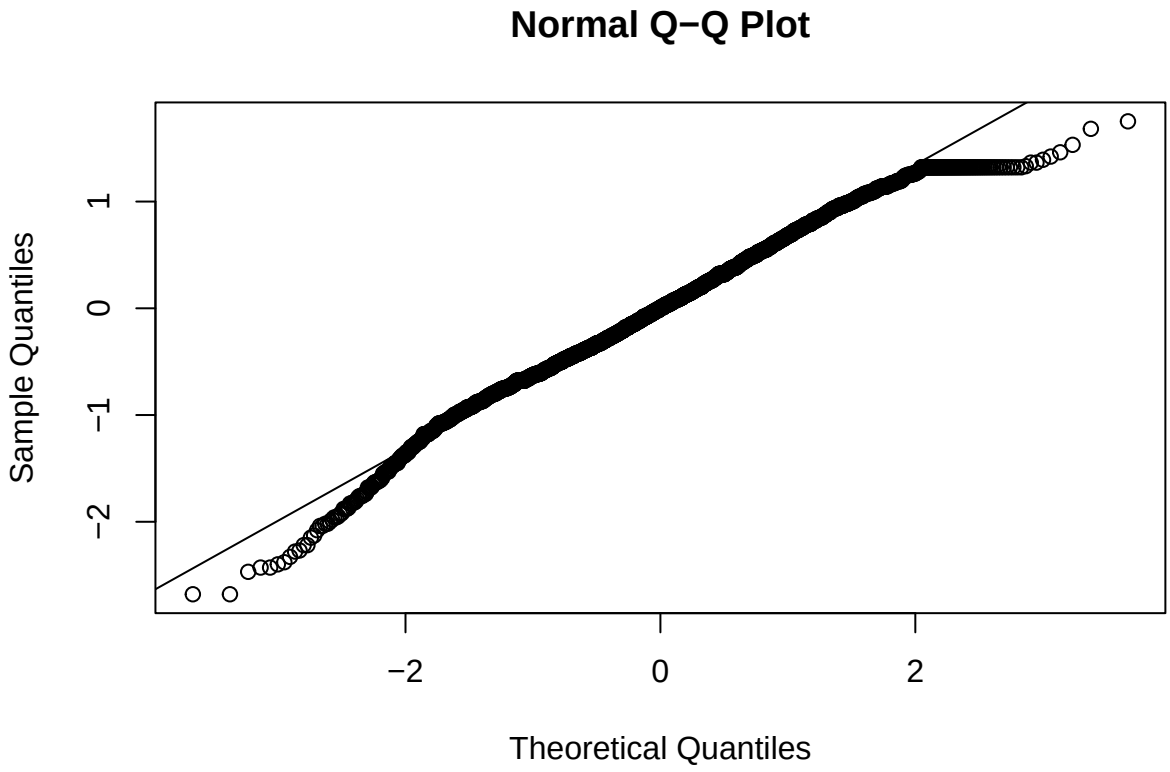
Por otro lado, el factor explica solo un 2 % de la variabilidad observada en `colgpa`. El resto (98 %) es la parte no explicada por el modelo.

6.6. Normalidad de los residuos

Usad el gráfico Normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de los residuos. Podéis usar las funciones de R correspondientes para hacer el gráfico y el test.

Solución:

```
qqnorm(residuals(mod1))
qqline(residuals(mod1))
```




```
#Test Shapiro-Wilk
shapiro.test(residuals(mod1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod1)
## W = 0.99175, p-value = 1.121e-14
```

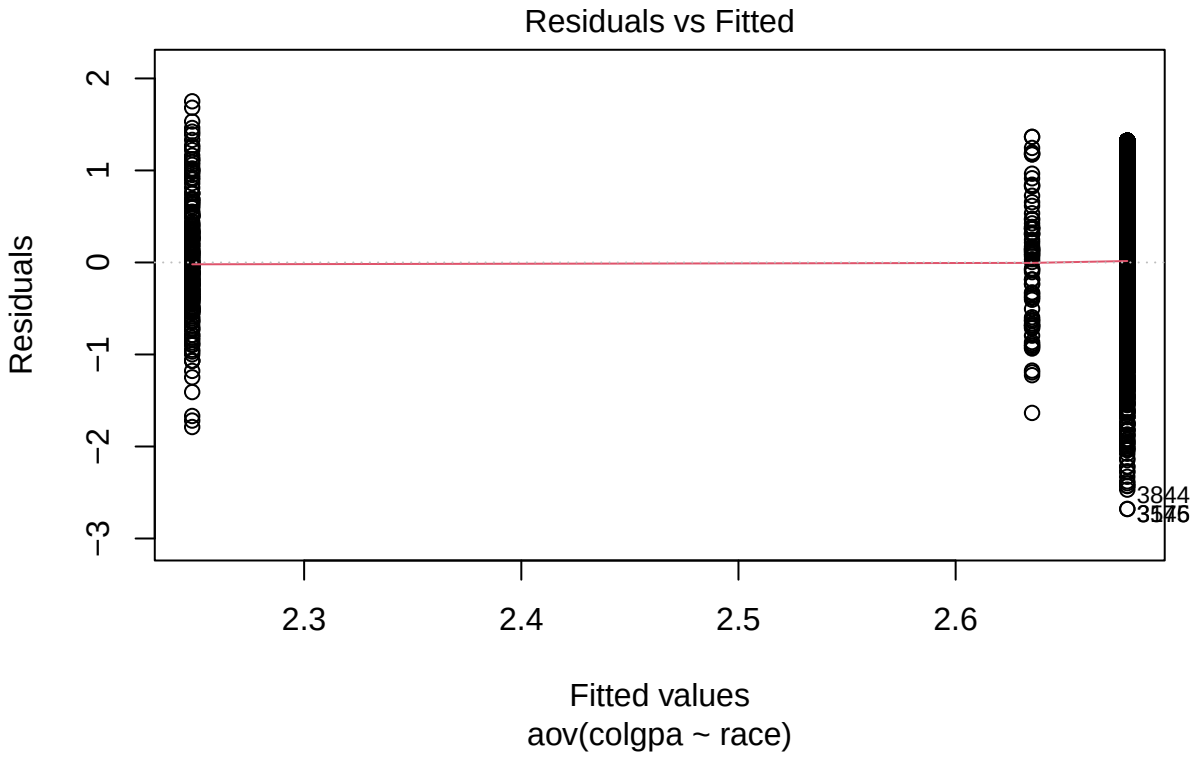
Los extremos del Normal Q-Q muestran una cierta desviación con respecto a los cuantiles teóricos lo que apunta al no cumplimiento de la normalidad. El test Shapiro-Wilk tiene un pvalue menor de 0.05 por lo que rechazamos la hipótesis nula de normalidad.

6.7. Homocedasticidad de los residuos

El gráfico "Residuals vs Fitted" proporciona información sobre la homcedasticidad de los residuos. Mostrad e interpretad este gráfico.

Solución:

```
plot(mod1,which=1)
```



A nivel visual, parece que el supuesto de homocedasticidad se mantiene. Para las tres niveles la dispersión de los residuos es bastante similar.

---

7. ANOVA multifactorial

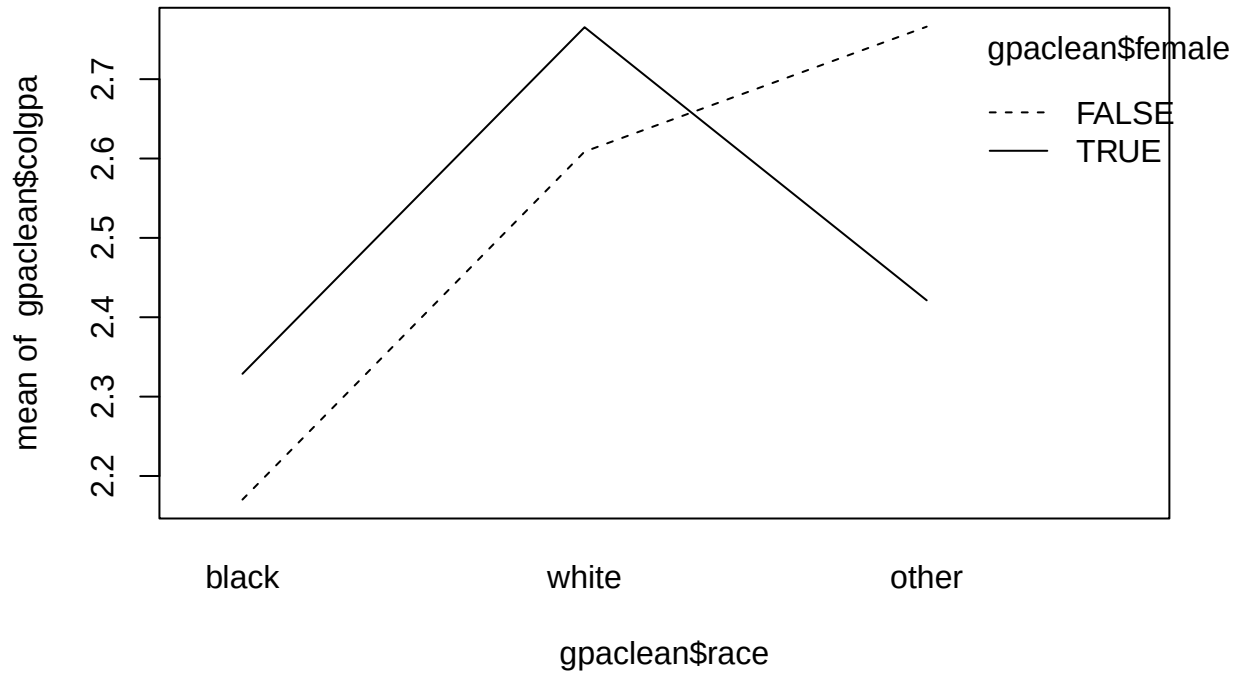
A continuación, se desea evaluar el efecto sobre `colgpa` de la raza del estudiante combinada con el factor género del estudiante (`female`). Seguid los pasos que se indican a continuación.

7.1. Análisis visual de los efectos principales y posibles interacciones

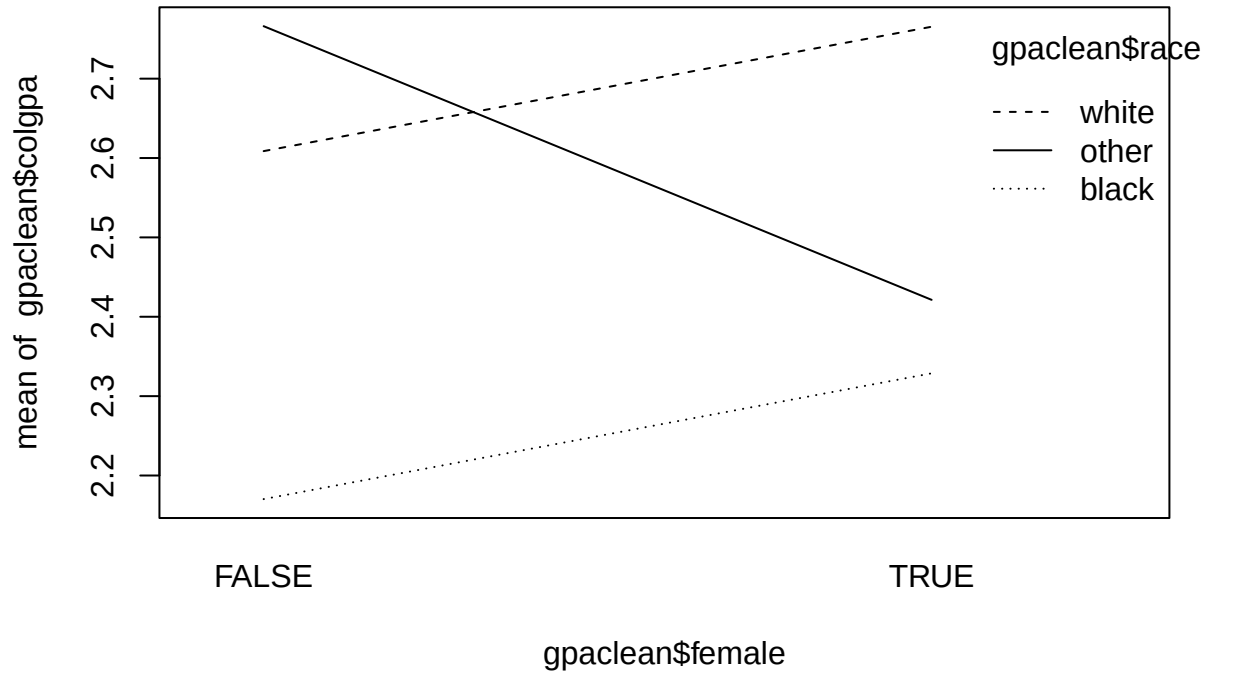
Representad la interacción de los dos factores `race` y `female` y comentad los gráficos resultantes.

Solución:

```
interaction.plot(gpaclean$race,gpaclean$female,gpaclean$colgpa)
```




```
interaction.plot(gpaclean$female,gpaclean$race,gpaclean$colgpa)
```



Como conclusión de los gráficos, podemos decir que se observa interacción entre los factores. El grupo `other` en el factor `race` presenta una nota media menor entre las mujeres que entre los hombres. Las rectas no son paralelas.

Con el análisis ANOVA multifactorial se comprobará si esta interacción es significativa.

7.2. Cálculo del modelo

Calculad el modelo ANOVA multifactorial. Podéis usar la función **aov**.

Solución:

```
mod4<-lm(colgpa~race*female,data=gpaclean)
anova(mod4)
```

```
## 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. Interpretación de los resultados

Respuesta:

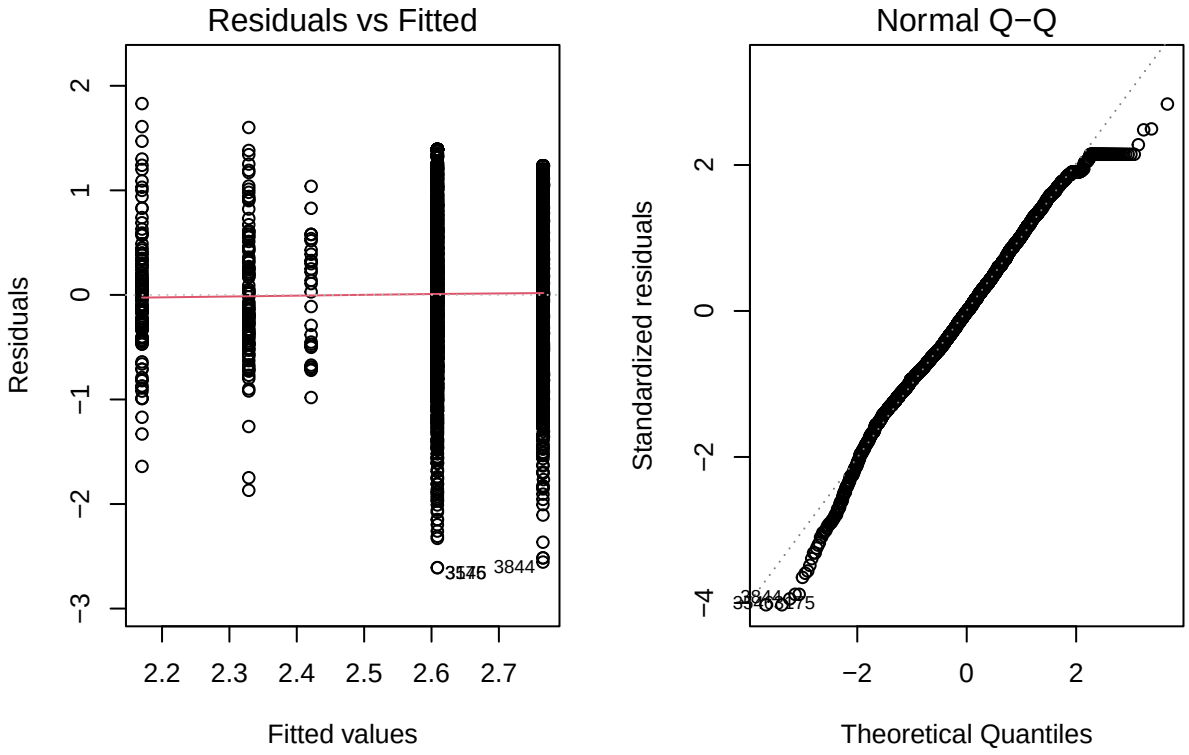
Tanto los factores principales como la interacción entre factores son significativos. Por tanto, la nota media de los estudiantes en función de la raza, es diferente según si el estudiante es mujer u hombre.

7.4. Adecuación del modelo

Interpretad la adecuación del modelo ANOVA obtenido usando los gráficos de los residuos.

Solución:

```
par(mfrow=c(1,2),cex=0.8)
plot(mod4,which=1)
plot(mod4,which=2)
```



Visualmente se observa que se cumple la homoscedasticidad pero no la normalidad. Al igual que en el apartado anterior del ANOVA con un factor, los extremos del Normal Q-Q muestran una cierta desviación con respecto a los cuantiles teóricos.

---

8. Conclusiones

Resumid las conclusiones principales del análisis (apartados 3 a 7). Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Podemos concluir (con un nivel de confianza del 95 %) que:

- El IC para la nota media de acceso es (1026.6, 1035.2) y en promedio las mujeres tienen un ‘sat‘ menor que el de los hombres.
- La nota del semestre ‘colgpa‘ es en promedio diferente entre hombres y mujeres.
- Las variables ‘sat‘, ‘female‘, ‘tothrs‘, ‘athlete‘, y ‘hsperc‘ contribuyen a explicar de manera significativa la nota del semestre. Las cuatro primeras presentan una correlación positiva con ‘colgpa‘ y la última tiene una correlación negativa.
- Las variables ‘sat‘, ‘female‘, ‘hsperc‘, ‘tothrs‘ y ‘black‘ inciden significativamente en la probabilidad de ser un estudiante excelente, mientras que las variables ‘athlete‘ y ‘white‘ no están significativamente asociadas con esa probabilidad. El análisis del odds ratio muestra que la probabilidad de ser un estudiante excelente es mayor entre las mujeres que entre los hombres.
- El factor ‘race‘ así como su interacción con ‘female‘, tienen un efecto significativo sobre ‘colgpa‘. No obstante, hay que interpretar con precaución estos resultados ya que el análisis muestra que no hay cumplimiento de la normalidad.

Puntuación de la actividad

- Apartados 1 y 2 (10 %)
- Apartado 3 (10 %)
- Apartado 4 (10 %)
- Apartado 5 (15 %)
- Apartado 6 (20 %)
- Apartado 7 (15 %)
- Apartado 8 (10 %)
- Calidad del informe dinámico (10 %)