A4 - Análisis de la varianza y repaso del curso

Yésica Fernández

2022-06-13

Antes de empezar con la práctica, vamos a cargar los paquetes necesarios para la realización del ejercicio.

```
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/graphics/index.html
if (!require('graphics')) install.packages('graphics'); library('graphics')
# https://cran.r-project.org/web/packages/ResourceSelection/index.html
if (!require('ResourceSelection')) install.packages('ResourceSelection'); library('ResourceSelection')
# https://cran.r-project.org/web/packages/lessR/index.html
if (!require('lessR')) install.packages('lessR'); library('lessR')
```

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).

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? ¿Que variables son de tipo cualitavivo?

```
gpa <-  read.csv("gpa.csv", sep = ",")
summary(gpa)
```

```
##       sat          tothrs              hsize          hsrank
##  Min.   : 470   Length:4137        Min.   :0.03   Min.   :  1.00
##  1st Qu.: 940   Class :character   1st Qu.:1.65   1st Qu.: 11.00
##  Median :1030   Mode  :character   Median :2.51   Median : 30.00
##  Mean   :1030                      Mean   :2.80   Mean   : 52.83
##  3rd Qu.:1120                      3rd Qu.:3.68   3rd Qu.: 70.00
##  Max.   :1540                      Max.   :9.40   Max.   :634.00
##
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4328   1st Qu.:2.210   FALSE:3943      FALSE:2277
##  Median :14.5833   Median :2.660   TRUE :194       TRUE :1860
##  Mean   :19.2371   Mean   :2.655
##  3rd Qu.:27.7108   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##                    NA's   :41
##    white           black
##  Mode :logical   Mode :logical
##  FALSE:308       FALSE:3908
##  TRUE :3829      TRUE :229
##
##
##
##
```

Las variables **sat, hsize, hsrank, hsperc, colgpa** son variables numéricas. Por otra parte, las variables **athlete, female, white, black** son variables cualitativas binarias, sólo tienen dos posibles valores. Por último, la variable **tothrs** es una variable numérica, pero viene acompañada de la letra "h" que nos indica que son horas, por lo que para poder trabajar con esta variable como numérica debemos eliminar esa "h" de los valores.

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.

```
gpa$tothrs <- as.numeric(substr(gpa$tothrs, 1, nchar(gpa$tothrs)-2))
summary(gpa)
```

```
##       sat           tothrs          hsize          hsrank
##  Min.   : 470   Min.   : 1.00   Min.   :0.03   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 1.00   1st Qu.:1.65   1st Qu.: 11.00
##  Median :1030   Median : 4.00   Median :2.51   Median : 30.00
```

```
##  Mean   :1030   Mean   : 4.81   Mean   :2.80   Mean   : 52.83
##  3rd Qu.:1120   3rd Qu.: 8.00   3rd Qu.:3.68   3rd Qu.: 70.00
##  Max.   :1540   Max.   :13.00   Max.   :9.40   Max.   :634.00
##                 NA's   :2
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4328   1st Qu.:2.210   FALSE:3943      FALSE:2277
##  Median :14.5833   Median :2.660   TRUE :194       TRUE :1860
##  Mean   :19.2371   Mean   :2.655
##  3rd Qu.:27.7108   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##                    NA's   :41
##    white           black
##  Mode :logical   Mode :logical
##  FALSE:308       FALSE:3908
##  TRUE :3829      TRUE :229
##
##
##
##
```

Se puede observar que la variable **tothrs** ya es numérica, y que toma valores entre 1 y 13.

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'.*

Cómo se ve en los datos resumidos del apartado anterior, hay dos variables que contienen valores ausentes o **NA**: **tothrs** y **colgpa**. La primera tiene dos valores ausentes y la segunda 41. Para ver el impacto de estos valores ausentes vamos a mirar cuantas observaciones contiene el conjunto de datos.

```
nrow(gpa)
```

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

El conjunto de datos, por lo tanto, tiene 4137 observaciones. Los valores faltantes de sólo suponen el 1% de los datos totales, por lo que eliminarlos no influiría en los resultados de los análisis. Se procede a eliminarlos del conjunto de datos.

```
gpaclean <- na.omit(gpa)
summary(gpaclean)
```

```
##       sat           tothrs           hsize           hsrank
##  Min.   : 470   Min.   : 1.000   Min.   :0.030   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 1.000   1st Qu.:1.643   1st Qu.: 11.00
##  Median :1030   Median : 4.000   Median :2.510   Median : 30.00
##  Mean   :1031   Mean   : 4.805   Mean   :2.794   Mean   : 52.71
##  3rd Qu.:1120   3rd Qu.: 8.000   3rd Qu.:3.660   3rd Qu.: 70.00
```

```
##  Max.   :1540   Max.   :13.000   Max.   :9.400   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4286   1st Qu.:2.210   FALSE:3903      FALSE:2253
##  Median :14.5833   Median :2.660   TRUE :191       TRUE :1841
##  Mean   :19.2208   Mean   :2.654
##  3rd Qu.:27.6694   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white           black
##  Mode :logical   Mode :logical
##  FALSE:304       FALSE:3869
##  TRUE :3790      TRUE :225
##
##
##
```

1.3.- Equivalencia de la nota en letras

La variable **colgp** contiene la nota numérica del estudiante. Crear 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

Se procede a crear la variable cualitativa.

```
gpaclean$gpaletter <- as.factor(ifelse(gpaclean$colgpa > 3.49 & gpaclean$colgpa <= 4.00, "A", ifelse(gpa
summary(gpaclean)
```

```
##       sat           tothrs          hsize           hsrank
##  Min.   : 470   Min.   : 1.000   Min.   :0.030   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 1.000   1st Qu.:1.643   1st Qu.: 11.00
##  Median :1030   Median : 4.000   Median :2.510   Median : 30.00
##  Mean   :1031   Mean   : 4.805   Mean   :2.794   Mean   : 52.71
##  3rd Qu.:1120   3rd Qu.: 8.000   3rd Qu.:3.660   3rd Qu.: 70.00
##  Max.   :1540   Max.   :13.000   Max.   :9.400   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4286   1st Qu.:2.210   FALSE:3903      FALSE:2253
##  Median :14.5833   Median :2.660   TRUE :191       TRUE :1841
##  Mean   :19.2208   Mean   :2.654
##  3rd Qu.:27.6694   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white           black          gpaletter
##  Mode :logical   Mode :logical   A: 458
##  FALSE:304       FALSE:3869      B:1971
##  TRUE :3790      TRUE :225       C:1521
##                                  D: 144
##
##
```

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.

Se procede a ver, en primer lugar, las dimensiones del conjunto de datos.

```
dim(gpaclean)
```

```
## [1] 4094   11
```

Por lo tanto, el conjunto de datos cuenta con 4094 observaciones y 11 variables. Se va a realizar ahora el análisis descriptivo de los datos.

```
summary(gpaclean)
```

```
##       sat           tothrs           hsize           hsrank
##  Min.   : 470   Min.   : 1.000   Min.   :0.030   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 1.000   1st Qu.:1.643   1st Qu.: 11.00
##  Median :1030   Median : 4.000   Median :2.510   Median : 30.00
##  Mean   :1031   Mean   : 4.805   Mean   :2.794   Mean   : 52.71
##  3rd Qu.:1120   3rd Qu.: 8.000   3rd Qu.:3.660   3rd Qu.: 70.00
##  Max.   :1540   Max.   :13.000   Max.   :9.400   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4286   1st Qu.:2.210   FALSE:3903      FALSE:2253
##  Median :14.5833   Median :2.660   TRUE :191       TRUE :1841
##  Mean   :19.2208   Mean   :2.654
##  3rd Qu.:27.6694   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white           black          gpaletter
##  Mode :logical   Mode :logical   A: 458
##  FALSE:304       FALSE:3869      B:1971
##  TRUE :3790      TRUE :225       C:1521
##                                  D: 144
##
##
```

En base a los resultados mostrados, vamos a analizar cada variable:

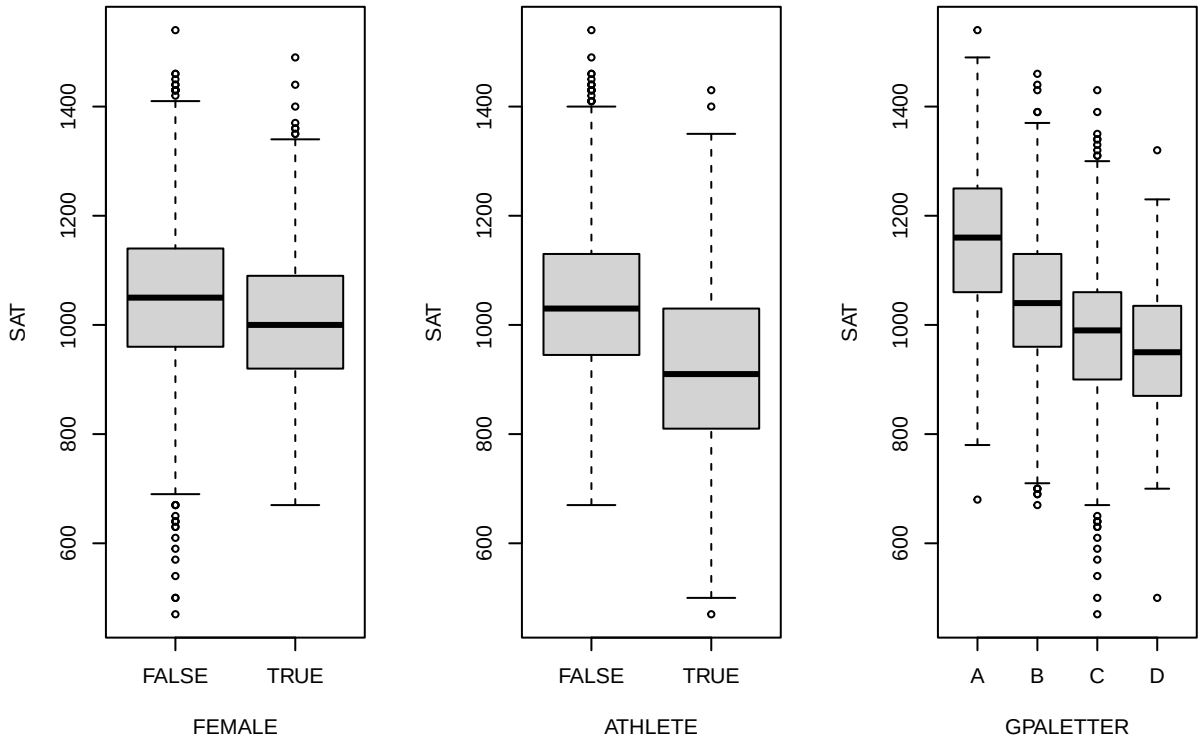
- **sat**: Se trata de una variable numérica que toma valores entre 470 y 1540. La media de esta variable es de 1031.
- **tothrs**: Variable numérica que toma valores entre 1 y 13 y cuya media se encuentra en 4.805.
- **hsize**: Variable numérica que toma valores entre 0.030 y 9.40 y cuya media se encuentra en 2.794.
- **hsrank**: Variable numérica que toma valores entre 1 y 634 y cuya media se encuentra en 52.71.
- **hsperc**: Variable numérica que toma valores entre 0.1 y 92 y cuya media se encuentra en 19.22.
- **colgpa**: Variable numérica que toma valores entre 0 y 4 y cuya media se encuentra en 2.654.
- **athlete**: Variable categórica que toma valores TRUE o FALSE. La mayoría son FALSE, siendo TRUE solo 191 de las observaciones.

- **female**: Varaible categórica que toma valores TRUE o FALSE. En este caso, 2253 observaciones son FALSE y 1841 son TRUE.
- **white**: Variable categórica que toma valores TRUE o FALSE. La mayoría son TRUE, siendo FALSE solo 304 de las observaciones.
- **black**: Variable categórica que toma valores TRUE o FALSE. La mayoría de las observaciones son FALSe, siendo TRUE solo 225.
- **gpaletter**: Variable categórica que puede tomar 4 valores: A, B, C o D. Las que menos observaciones tienen son la A y la D, con 458 y 144 observaciones respectivamente.

2.2.- Visualización

Se va a mostrar con diagramas de cajas (boxplot) la distribución de la variable **sat** según la variable **female**, según **athlete** y según **gpaletter**.

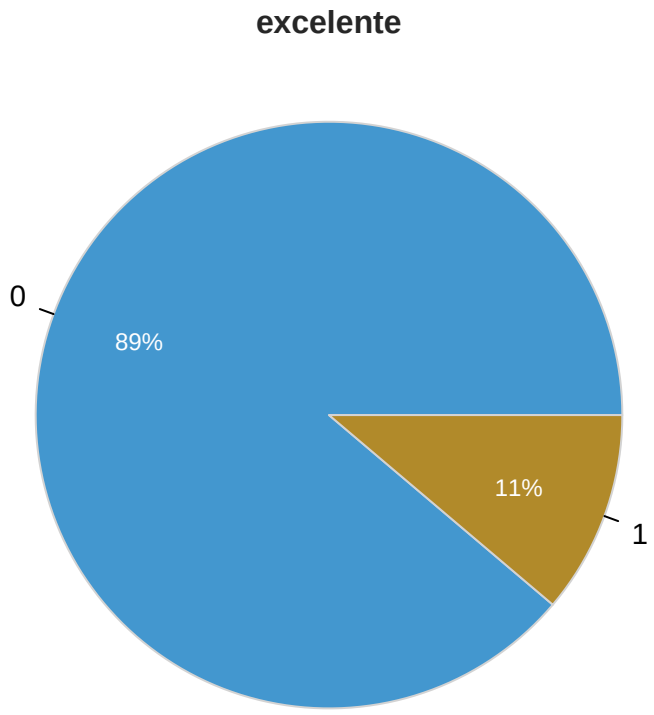
```
par(mfrow = c(1,3))
boxplot(gpaclean$sat ~ gpaclean$female, xlab = "FEMALE", ylab = "SAT" )
boxplot(gpaclean$sat ~ gpaclean$athlete, xlab = "ATHLETE", ylab = "SAT")
boxplot(gpaclean$sat ~ gpaclean$gpaletter, xlab = "GPALETTER", ylab = "SAT")
```



Se observa en los gráficos de cajas que existen outliers en la variable sat. Estos varían dependiende de como la estemos agrupando en base a las otras variables. También observamos que la media de los valores de SAT en mujeres es menor que en hombres. En la segunda gráfica se observa que la media es menor si el estudiante práctica deporte que si no lo hace. En la última gráfica se puede ver la correspondencia en las notas medias de acceso y las notas del primer semestre, y vemos que las notas bajas del primer semestre se corresponden con notas bajas de acceso por regla general.

Ahora se va a crear una variable denominada excelente que indique si el estudiante ha obtenido una A de nota media al final del semestre. Esta nueva variable se codificará como una variable dicotómica que toma el valor 1 cuando el estudiante ha obtenido una !, y el valor 0 en caso contrario. Se va a realizar también un gráfico que muestre el porcentaje de estudiantes excelentes.

```
gpaclean$excelente <- ifelse(gpaclean$gpaletter == "A", 1, 0)
PieChart(excelente, hole = 0, values = "%", data = gpaclean,)
```




```
## >>> Suggestions
## PieChart(excelente, hole=0)  # traditional pie chart
## PieChart(excelente, values="%")  # display %'s on the chart
## PieChart(excelente)  # bar chart
## Plot(excelente)  # bubble plot
## Plot(excelente, values="count")  # lollipop plot
##
## --- excelente ---
##
##                    0      1      Total
## Frequencies:    3636    458       4094
## Proportions:   0.888  0.112      1.000
##
## Chi-squared test of null hypothesis of equal probabilities
##   Chisq = 2466.948, df = 1, p-value = 0.000
```

Por lo tanto, sólo el 11% de los estudiantes es excelente, es decir, que tienen una ! de nota media al final del semestre.

3.- Estadística inferencial

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

a) Calcular el intervalo de confianza al 95% de la media poblacional de la variable sat de los estudiantes.

Para realizar el ejercicio, se define una función IC que recibe la variable, la confianza y que devuelve un vector con los valores del intervalo de confianza.

```
IC <- function(x, NC){
  alpha <- 1 - NC
  n <- length(x)
  desviacion <- sd(x)
  media <- mean(x)
  cuartil <- qt(alpha/2, df=n-1,  lower.tail=FALSE)
  margen_error = abs((cuartil*desviacion) / sqrt(n))
  lim_inf <- media - margen_error
  lim_sup <- media + margen_error
  a <- c(lim_inf, lim_sup)
  return(a)
}
```

Una vez definida la función se calcula el intervalo de confianza al 95%.

```
Resultado_1 <- IC(gpaclean$sat, 0.95)
Resultado_1
```

```
## [1] 1026.651 1035.191
```

Por lo tanto, podemos decir que la media poblacional de **sat** se encuentra entre 1026.651 y 1035.191 con un nivel de confianza del 95%.

b) Calcular el intervalo de confianza al 95% de la media poblacional de la variable sat en función de si los estudiantes son hombres o mujeres.

Se calcula en primer lugar el intervalo de confianza de la media poblacional de la varaible **sat** para las mujeres.

```
Resultado_female <- IC(filter(gpaclean, female == TRUE)$sat, 0.95)
Resultado_female
```

```
## [1] 1001.413 1013.123
```

Ahora se va a calcular el mismo intervalo en el caso de los hombres.

```
Resultado_male <- IC(filter(gpaclean, female == FALSE)$sat, 0.95)
Resultado_male
```

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

No existe solapamiento en los intervalos de confianza de la media poblacional de la variable **sat** entre hombres y mujeres. Se observa que el intervalo de confianza es menor en el caso de las mujeres que de los hombres. Por lo tanto, se deduce con una confianza del 95% que la media poblacional de la variable **sat** es menor en mujeres que en hombres.

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

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

3.2.1.- Pregunta de investigación

¿ La media de colgpa es la misma para las mujeres que para los hombres?

3.2.2. Hipótesis nula y alternativa

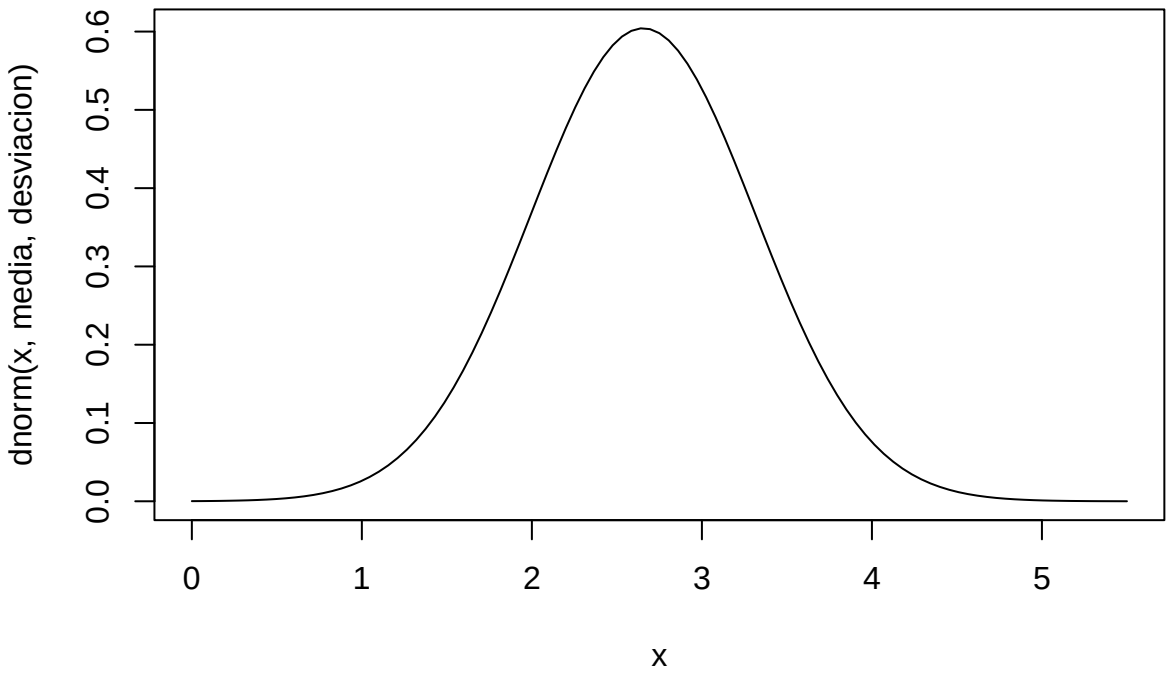
$$\begin{cases} H_0: & \mu_0 \neq \mu_1 \\ H_1: & \mu_0 = \mu_1 \end{cases}$$

Siendo μ_0 la media de colgpa de las mujeres y μ_1 la media de colgpa de los hombres.

3.2.3.- Justificación del test a aplicar

En primer lugar, para saber que test podemos utilizar, vamos a observar si la variable sigue una distribución normal.

```
media <- mean(gpaclean$colgpa)
desviacion <- sd(gpaclean$colgpa)
curve(dnorm(x, media, desviacion), xlim = c(0, 5.5))
```



Podemos asumir normalidad en colgpa y por lo tanto se puede calcular la probabilidad de que varios valores ocurran en un cierto intervalo dada la confianza, por lo que se puede afirmar que se trata de un test paramétrico. Por otra parte, es un contraste de dos muestras independientes.

3.2.4.- Cálculos

Para el cálculo del test se va a implementar una función propia.

```
my_test_3 <- function(x, y, NC){
  media_x <- mean(x)
  media_y <- mean(y)
  var_x <- var(x)
  var_y <- var(y)
  alpha <- 1 - NC
  n_x <- length(x)
  n_y <- length(y)
  #Estadístico de contraste
  zobs <- (media_x - media_y) / sqrt((var_x/n_x) + (var_y/n_y))
  #Región de aceptación
  zcrit.L <- qnorm(1-alpha, lower.tail=TRUE)
  #Cálculo del p valor
  pvalue <- pnorm(zobs, lower.tail=TRUE,)
  return(data.frame(L=zcrit.L, U="INF", zobs, pvalue))
}
```

Se va a calcular el contraste para un nivel de confianza del 95%.

```
Resultado <- my_test_3(filter(gpaclean, female == TRUE )$colgpa, filter(gpaclean, female == FALSE)$colg
Resultado$NivelConfianza <- "95%"
knitr::kable(Resultado)
```

L	U	zobs	pvalue	NivelConfianza
1.644854	INF	7.004054	1	95%

3.2.5.- Interpretación del test

Con un nivel de confianza del 95%, dado que p-valor $> \alpha$, aceptamos la hipótesis nula y se concluye que la media de colgpa no es la misma para hombres que para mujeres.

4.- Modelo de regresión lineal

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

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

```
##
## Call:
## lm(formula = colgpa ~ sat + female + tothrs + athlete + hsperc,
##     data = gpaclean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64899 -0.36004  0.02526  0.39035  1.90896
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)  1.04319551  0.07707040  13.536 < 0.0000000000000002 ***
## sat          0.00163925  0.00006686  24.518 < 0.0000000000000002 ***
## femaleTRUE   0.15167410  0.01805106   8.403 < 0.0000000000000002 ***
## tothrs       0.01845950  0.00243555   7.579   0.0000000000000428 ***
## athleteTRUE  0.14861728  0.04248356   3.498              0.000473 ***
## hsperc      -0.01262509  0.00056372 -22.396 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5531 on 4088 degrees of freedom
## Multiple R-squared:  0.2991, Adjusted R-squared:  0.2982
## F-statistic: 348.9 on 5 and 4088 DF,  p-value: < 0.00000000000000022
```

4.1.- Interpretación del modelo

Para ver la calidad del ajuste se mira el R-squared, que en este caso tiene un valor de 0.2991. La calidad del ajuste es mejor cuanto más cerca se encuentra este valor de 1. Por lo tanto, la calidad en este caso es baja.

El p-value nos indica como influyen las variables en el modelo. En este caso, podemos ver que todas las variables son sifnificativas, siendo la que menos que un estudiante sea atleta. Aunque las variables influyen en el modelo no son suficientes para predecir variable dependiente.

4.2.- Predicción

Independientemente del R2 obtenido en el apartado previo, se va a aplicar 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 %.

```
datos <- data.frame(sat = 800, female = FALSE, tothrs = 60, athlete = TRUE, hsperc = 60)
p <- predict(modelo_lineal, datos)
p
```

```
##        1
## 2.853273
```

Según el modelo, el valor de colgpa es 2.8532.

5.- Regresión logística

5.1.- Estimación del modelo

Estimar 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**.

```
modelo_logistica <- glm(excelente ~ female + athlete + sat + tothrs + black + white + hsperc, data = gpa
summary(modelo_logistica)
```

```
##
## Call:
## glm(formula = excelente ~ female + athlete + sat + tothrs + black +
##     white + hsperc, data = gpaclean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.45928  -0.16136  -0.07924   0.01537   1.05675
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -0.46177712  0.05239260  -8.814 < 0.0000000000000002 ***
## femaleTRUE   0.03028651  0.00949927   3.188              0.00144 **
## athleteTRUE  0.07550732  0.02261822   3.338              0.00085 ***
## sat          0.00064945  0.00003621  17.933 < 0.0000000000000002 ***
## tothrs      -0.00374275  0.00128154  -2.921              0.00351 **
## blackTRUE   -0.03073443  0.03837720  -0.801              0.42326
## whiteTRUE   -0.02124251  0.03313513  -0.641              0.52150
## hsperc      -0.00383368  0.00029976 -12.789 < 0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08463426)
##
##     Null deviance: 406.76  on 4093  degrees of freedom
## Residual deviance: 345.82  on 4086  degrees of freedom
## AIC: 1518.5
##
## Number of Fisher Scoring iterations: 2
```

5.2.- Interpretación del modelo estimado

Podemos observar en el resultado anterior que las variables son significativas para el modelo, exceptuando las variables **black** y **white**, cuyo p-value es más elevado y por lo tanto no se tratan de variables significativas para el modelo. De las otras variables, influyen más **athlete, sat y hsperc** que **female y tothrs**.

5.3.- Importancia de ser mujer

En base al modelo anterior, interpretar 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% del odds ratio.

```
exp(coefficients(modelo_logistica))
```

```
## (Intercept)  femaleTRUE athleteTRUE         sat      tothrs   blackTRUE
##   0.6301628   1.0307498   1.0784311   1.0006497   0.9962642   0.9697331
##   whiteTRUE      hsperc
##   0.9789815   0.9961737
```

Se obtiene un OR para la variable female de 1.030, con lo que la probabilidad de ser un estudiante excelente si se es mujer es 1.030 veces mayor.

Se va a calcular el intercalo de confianza del odds ratio.

```
exp(confint(modelo_logistica))
```

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

```
##                 2.5 %    97.5 %
## (Intercept) 0.5686645 0.6983118
## femaleTRUE  1.0117366 1.0501203
## athleteTRUE 1.0316675 1.1273145
## sat         1.0005786 1.0007207
## tothrs      0.9937650 0.9987698
## blackTRUE   0.8994675 1.0454877
## whiteTRUE   0.9174234 1.0446701
## hsperc      0.9955886 0.9967591
```

Con estos resultados, podemos decir que la probabilidad de ser un estudiante excelente está entre 1.01 y 1.05 veces más que los hombres.

5.4.- Predicción

¿Con que probabilidad una mujer, no atleta, son un sat de 120 puntos, 50 horas cursadas, de raza negra y con un ranking relativo del 10% será excelente?

```
datos <- data.frame(female = TRUE, athlete = FALSE, sat = 120, tothrs = 50, black = TRUE, white = FALSE
p_logistica <- predict(modelo_logistica, datos)
p_logistica
```

```
##          1
## -0.6097655
```

La probabilidad de que se trate de un estudiante excelente es de -0.61.

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

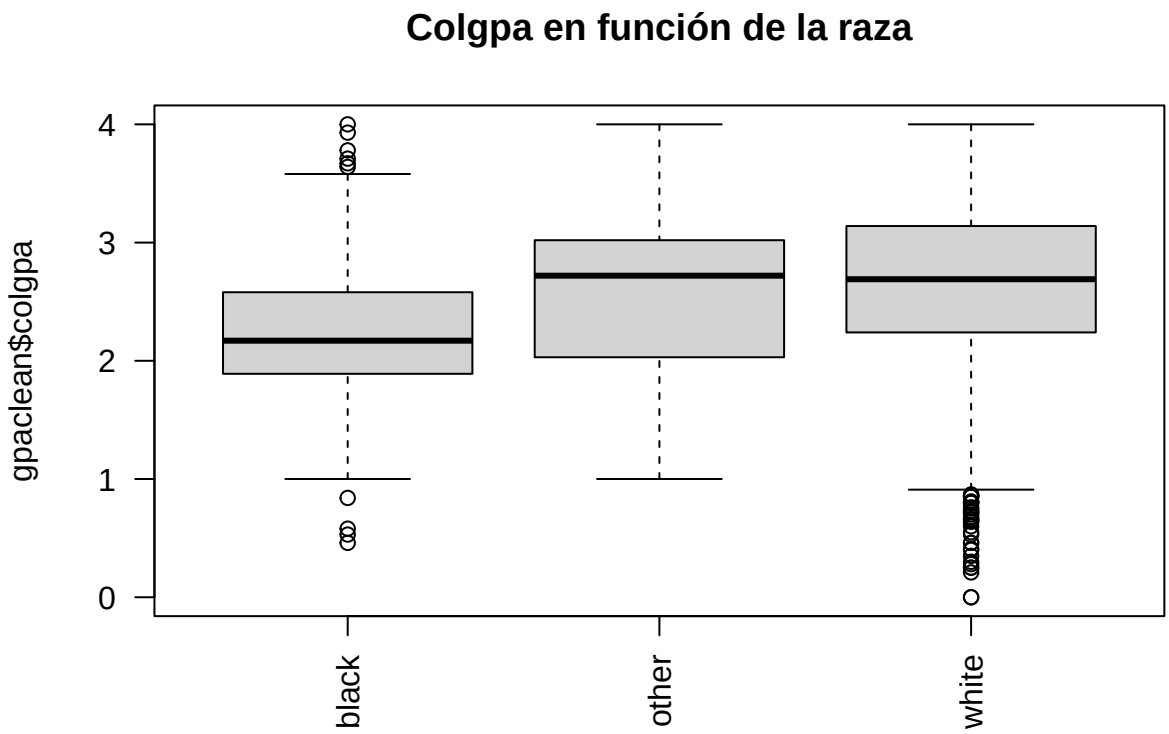
Se va a realizar un ANOVA para contrastar si existen diferencias en la variable **colgpa** en función de la raza de los estudiantes. En primer lugar, a partir de las variables **black** y **white** se crea una variable categórica denominada **race** que indica la raza del estudiante en una de las tres categorías siguientes: **black, white, other**.

```
gpaclean$race <- ifelse(gpaclean$black == TRUE, "black",ifelse(gpaclean$white == TRUE, "white", "other"
```

6.1.- Visualización gráfica

Se va a mostrar gráficamente la distribución de colgpa según los valores de race.

```
boxplot(gpaclean$colgpa ~ gpaclean$race, main = "Colgpa en función de la raza",las=2, xlab = "")
```



6.2.- Hipótesis nula y alternativa

$$\begin{cases} H_0 : \mu_0 = \mu_1 = \mu_2 \\ H_1 : \text{Existen diferencias significativas} \end{cases}$$

Siendo μ la variable colgpa de los tres posibles valores de las razas.

6.3.- Modelo

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

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

```
##               Df Sum Sq Mean Sq F value              Pr(>F)
## race           2   39.4  19.682   46.15 <0.0000000000000002 ***
## Residuals   4091 1744.7   0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La primer columna indica el número de grados de libertad. Sum Sq la suma de cuadrados, que determina el estadístico de contraste (F de Snedecor), determinado con F value. Mean SQ indica a media de cuadrados y el p-valor está determinado por Pr(>F).

Los grados de libertad entre grupos del factor raza son 2 ya que está determinado por n-1, siendo n = 3(el número de razas diferentes), mientras que dentro de dichos grupos el grado de libertad es 4091 al ser el total de observaciones 4094 - 3 (número de grupos diferentes). La suma de los cuadrados entre los grupos de razas está dada por la formula $SCE = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$, siendo k = 3. Mientas que la suma de cuadrados de dentro de los grupos se realiza con la formula $SCD = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_{ij} - \bar{x}_j)^2$. La media de ambos cuadrados constituye la división de sus respectivas sumas, computadas con las formulas SCE y SCD, entre el número de grados de libertad. Finalmente, el estadístico de contraste F se computa únicamente en la fuente de variación entre grupos, en este caso, educativos, ya que es la división de su media de cuadrados y la media de los cuadrados de dentro de los grupos. $fvalue = \frac{19.682}{0.426} = 46.15$.

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

La media general μ se estima con la media de los datos:

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

```
## [1] 3
```

Los efectos de los tratamientos se estiman mediante $\hat{\alpha}_i = \bar{y}_i - \bar{y}$. Por lo tanto:

```
alpha_black <- mean(filter(gpaclean, race == "black")$colgpa) - 3
alpha_black
```

```
## [1] -0.7515556
```

```
alpha_white <-  mean(filter(gpaclean, race == "white")$colgpa) - 3
alpha_white
```

```
## [1] -0.3212111
```

```
alpha_other <-  mean(filter(gpaclean, race == "other")$colgpa) - 3
alpha_other
```

```
## [1] -0.3648101
```

Falta la estimación de la varianza del error, que podemos obtenerla a partir de la tabla del análisis de la varianza. Este viene proporcionado por el valor de MSE, en este caso $\hat{\omega^2} = 0.426$.

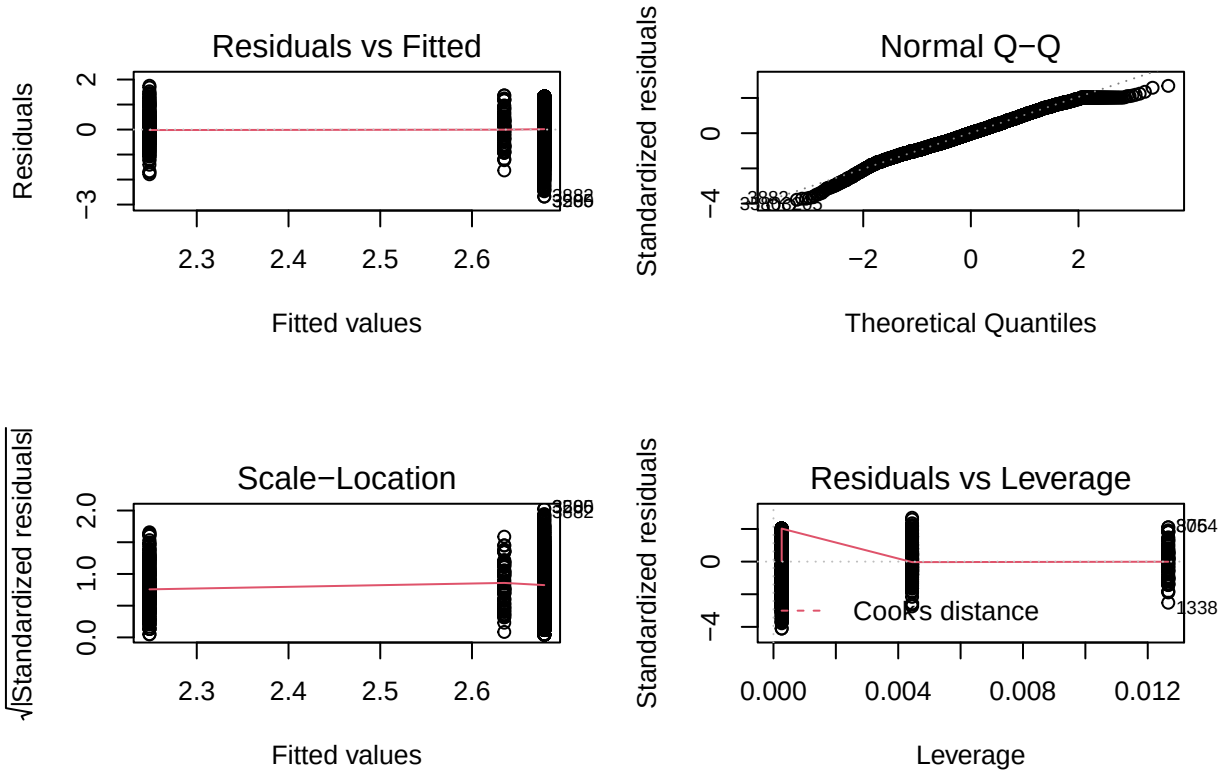
6.5.- Conclusiones de los resultados del ANOVA

Puesto que se obtiene un p-valor de <2e-16, claramente inferior a un nivel de significación del 5% se acepta la hipótesis alternativa y se concluye que la raza influye en la variable colgpa.

6.6.- Normalidad de los residuos

Se usa el gráfico Normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de los residuos. Se realizan los gráficos sobre el modelo ANOVA.

```
par(mfrow = c(2, 2))
plot(anova)
```



En la gráfica Normal Q-Q se puede ver que los residuos siguen una distribución normal, ya que están notablemente localizados cerca de la recta.

Ahora realizamos el test de Shapiro-Wilk.

```
results <- lm(colgpa~race, data = gpaclean)
shapiro.test(residuals(results))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(results)
## W = 0.99175, p-value = 0.00000000000001128
```

A la vista del p-valor obtenido, se puede aceptar que la variable sigue una distribución normal.

6.7.- Homocedasticidad de los residuos

Según el gráfico "Residuals vs Fitted" proporciona información sobre la homocedasticidad de los residuos, se muestra en el apartado anterior. La gráfica nos indica que no existe relación entre los residuos y la media de valores de cada grupo (se muestran 3 líneas verticales, correspondientes a cada raza). Si hubiese algún otro patrón en los residuos, como una línea horizontal, indicaría que podría haber algún otro predictor que no está siendo incluido en el modelo. Por eso podemos asumir la homocedasticidad de las varianzas.

7.- ANOVa multifactorial

Se debe evaluar el efecto sobre colgpa de la raza del estudiante combinada con el factor de género del estudiante (female).

7.1.- Análisis visual de los efectos principales y posibles interacciones

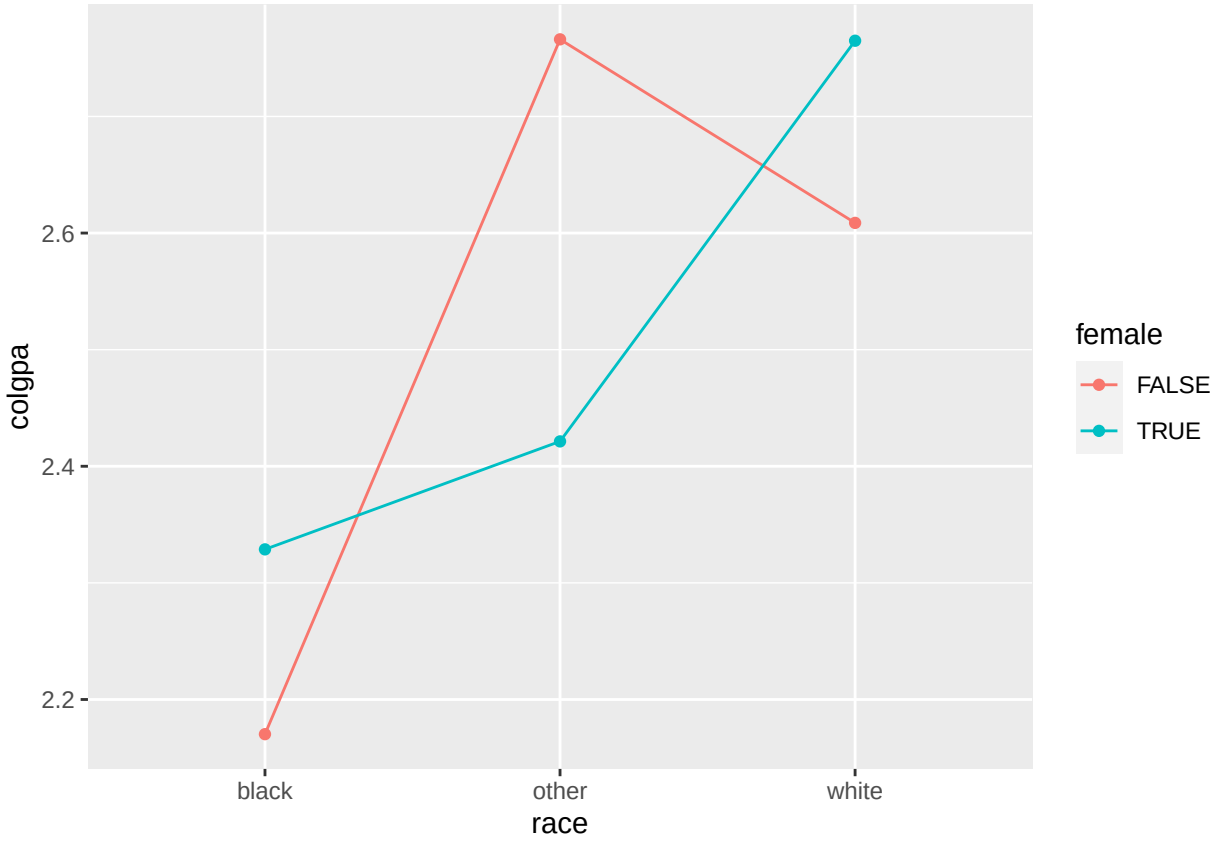
```
t <- as.data.frame(gpaclean %>%
                     group_by(race, female) %>%
                     summarise(mean(colgpa)))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
names(t) <- c("race", "female", "colgpa")
t
```

```
##    race female   colgpa
## 1 black  FALSE 2.170263
## 2 black   TRUE 2.328739
## 3 other  FALSE 2.766122
## 4 other   TRUE 2.421333
## 5 white  FALSE 2.608713
## 6 white   TRUE 2.764941
```

```
ggplot(data = t, aes(race, colgpa, color=female, group = female)) + geom_point() + geom_line()
```



Los efectos de estos factores se pueden apreciar en el gráfico. La raza negra determina una nota menor, mientras que la blanca predomina con notas mejores. Se observa también que las mujeres negras tienen notas mayores que los hombres negros.También observamos que las mujeres suelen tener mejores notas, exceptuando en la raza **other** que se ve claramente que son los hombres los que consiguen notas mayores.

7.2.- Cálculo del modelo

```
anova_multi <- aov(colgpa ~ race * female, data = gpaclean)
summary(anova_multi)
```

```
##               Df Sum Sq Mean Sq F value               Pr(>F)
## race           2   39.4  19.682  46.827 < 0.0000000000000002 ***
## female         1   21.9  21.918  52.147    0.000000000000611 ***
## race:female    2    4.6   2.294   5.457               0.0043 **
## Residuals   4088 1718.2   0.420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.3.- Interpretación de los resultados

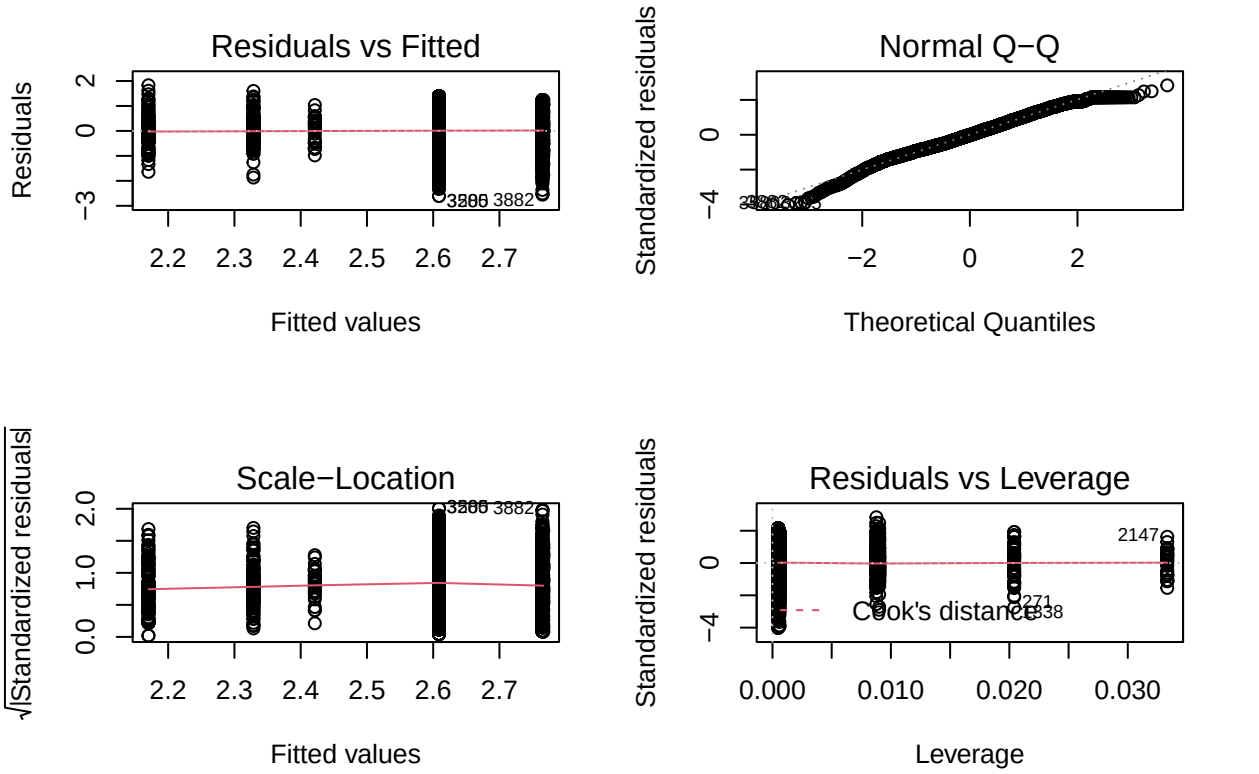
Como se puede observar, el p-valor de todos los factores, y su interacción, son significativos dado que es menor que 0.05. Lo más determinante es la raza frente al sexo o la interacción de ambos.

7.4.- Adecuación del modelo

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

Realizamos por lo tanto los gráficos de los residuos del modelo ANOVA generado.

```
par(mfrow = c(2, 2))
plot(anova_multi)
```



La gráfica *Normal Q-Q* indica que los residuos siguen una distribución normal, dada la cercanía de los puntos a la recta. La gráfica *Residuals vs Fitted* muestra como se distribuyen los residuos de cada grupo en una línea vertical sobre el 0, generando una línea roja suficientemente recta. Podemos afirmar que las varianzas de los errores son iguales, y que se observa homocedasticidad. Además, en la gráfica *Residuals vs Leverage* no se observa ningún outlier que influya en el resultado. Estos serían los que se encontrarían en alguna de las esquinas superiores, fuera de la línea de puntos que marcaría la distancia de Cook. En la gráfica *Scale-Location* observamos que la línea roja, que marca la dispersión en la predicción de los residuos es bastante recta, si bien es cierto que se observa una ligera subida debido al incremento en la dispersión de los residuos predecidos del último grupo.