A4: Análisis de varianza y repaso del curso

Estadística Avanzada, Universitat Oberta de Catalunya

Paula Muñoz Lago

26 diciembre 2019

Contents

1. Lectura del fichero **2**

1.1. Variables categóricas 2

1.2. Variables numéricas 3

2. Estadística descriptiva y visualización **3**

2.1. Análisis descriptivo 3

2.2. Visualización 4

2.3. Comprobación de normalidad 8

3. Estadística inferencial **9**

3.1. Intervalo de confianza de la variable age 9

3.2. Contraste de hipótesis para la diferencia de medias 11

3.3. Contraste no paramétrico 13

4. Regresión logística **14**

4.1. Modelo predictivo 14

4.2. Interpretación 16

4.3. Importancia del nivel de estudios 17

4.4. Predicción 18

5. Análisis de la varianza de un factor (ANOVA) **19**

5.1. Nivel de educación y salario 19

5.2. Adecuación al modelo 21

5.3. ANOVA no paramétrico 23

6. ANOVA multifactorial **24**

6.1. Factores: Raza y tipo de trabajo 24

6.2. Factores: raza y nivel de educación 26

7. Comparaciones múltiples **28**

8. Conclusiones **30**

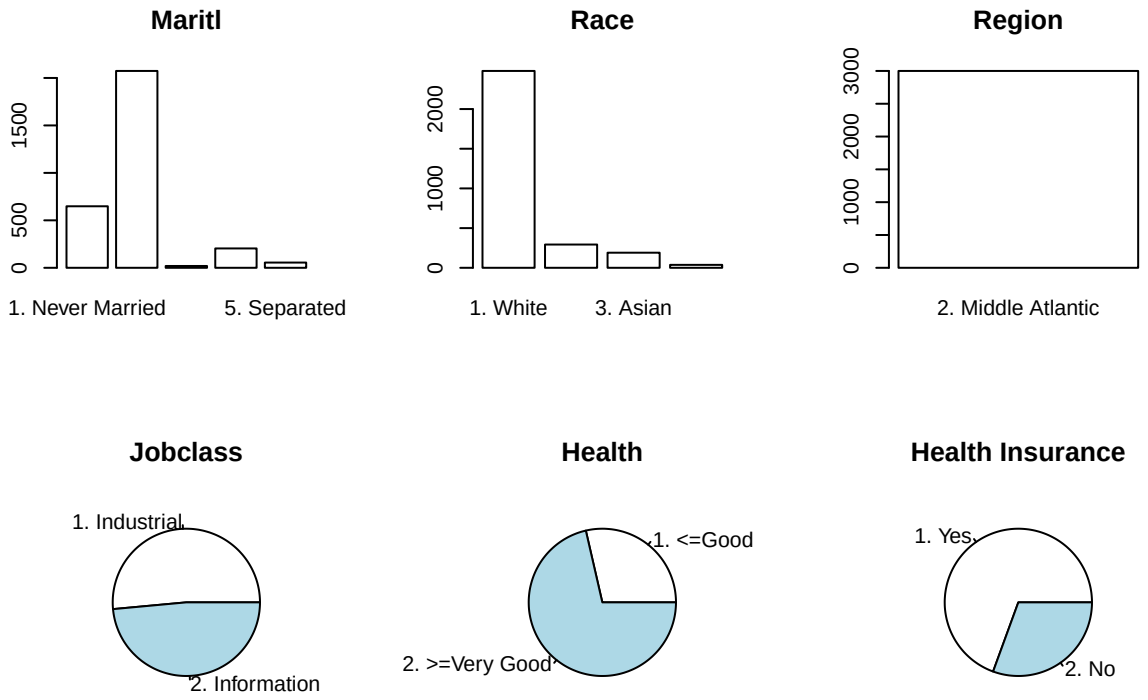
1. Lectura del fichero

Leed el fichero Wage.csv el cual contiene los datos del estudio Mid-Atlantic Wage Data. A continuación, verificad el tipo de cada variable. ¿Qué variables son de tipo numérico? ¿Qué variables son de tipo categórico? Realizad conversiones de tipo si es necesario.

```
current_working_directory <- getwd()
data <- read.csv(paste(current_working_directory,"/Wage.csv", sep = ""))
sapply(data, class)
```

```
##       year        age     maritl       race  education     region   jobclass
##  "integer"  "integer"   "factor"   "factor"   "factor"   "factor"   "factor"
##     health health_ins    logwage       wage
##   "factor"   "factor"  "numeric"  "numeric"
```

1.1. Variables categóricas



1.2. Variables numéricas

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).

2.1.1. Variables categóricas

```
summary(data$maritl)
```

```
## 1. Never Married       2. Married       3. Widowed      4. Divorced
##              648              2074              19              204
##     5. Separated
##               55
```

```
summary(data$race)
```

```
## 1. White 2. Black 3. Asian 4. Other
##     2480      293      190       37
```

```r
summary(data$region)
```

```
## 2. Middle Atlantic
##               3000
```

```r
summary(data$jobclass)
```

```
##  1. Industrial 2. Information
##           1544           1456
```

```r
summary(data$health)
```

```
##      1. <=Good 2. >=Very Good
##            858           2142
```

```r
summary(data$health_ins)
```

```
## 1. Yes  2. No
##   2083    917
```

2.1.2. Variables numéricas

```r
summary(data$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2003    2004    2006    2006    2008    2009
```

```r
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.75   42.00   42.41   51.00   80.00
```

```r
summary(data$logwage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.447   4.653   4.654   4.857   5.763
```

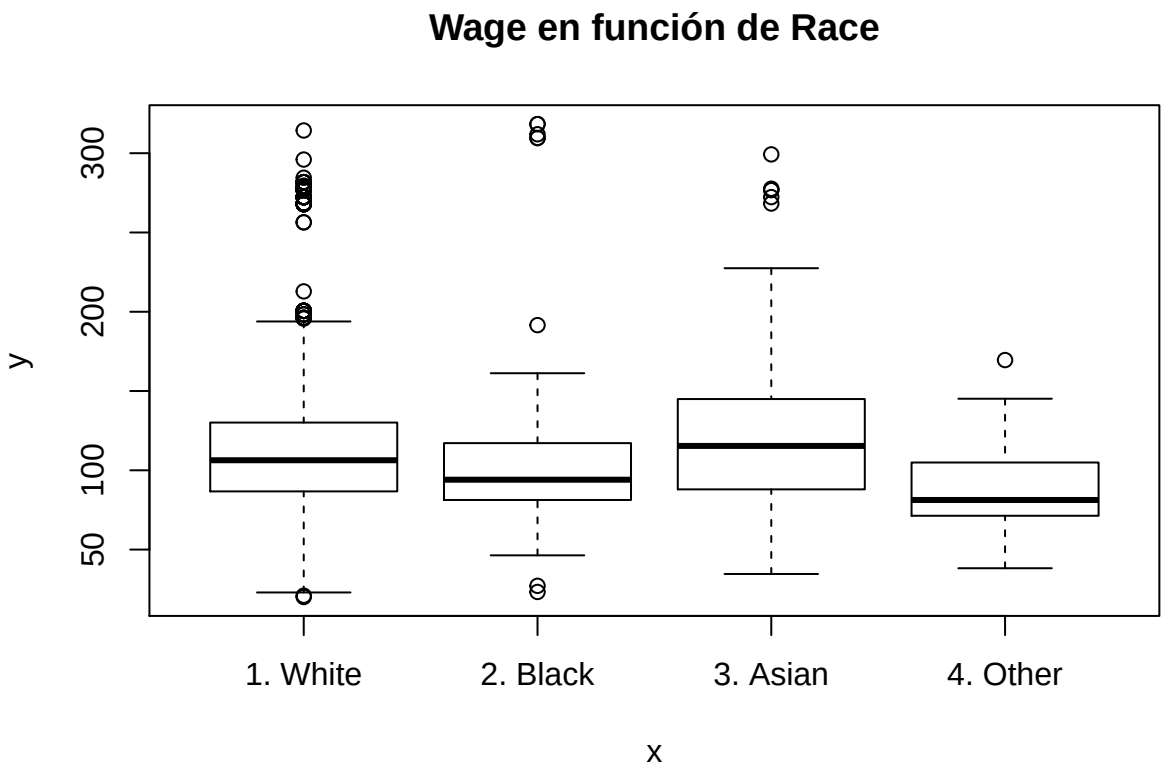
```r
summary(data$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.09   85.38  104.92  111.70  128.68  318.34
```

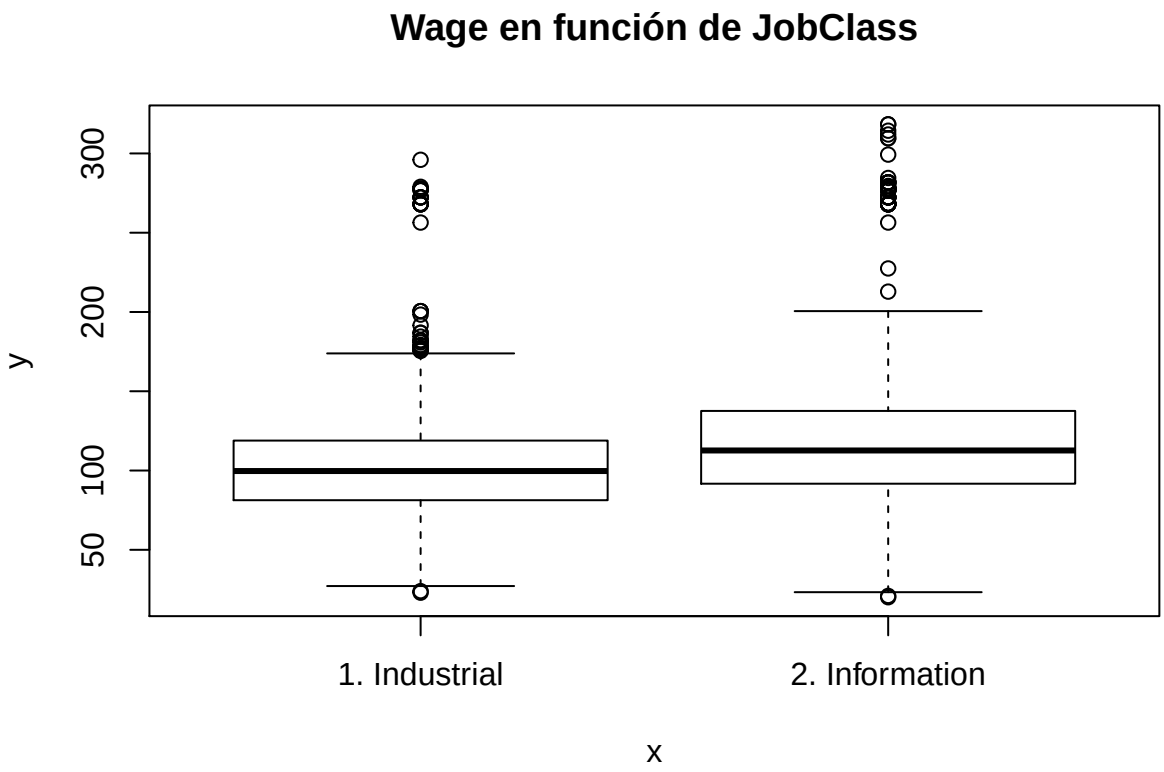
2.2. Visualización

Mostrad con diversos diagramas de caja la distribución de la variable wage según: race, jobclass, health y health_ins. Interpretar los gráficos brevemente.

```r
plot(x = data$race, y=data$wage, main = "Wage en función de Race")
```




```r
plot(x = data$jobclass, y=data$wage, main = "Wage en función de JobClass")
```




```
plot(x = data$health, y=data$wage, main = "Wage en función de health")
```


```
plot(x = data$health_ins, y=data$wage, main = "Wage en función de health_ins")
```

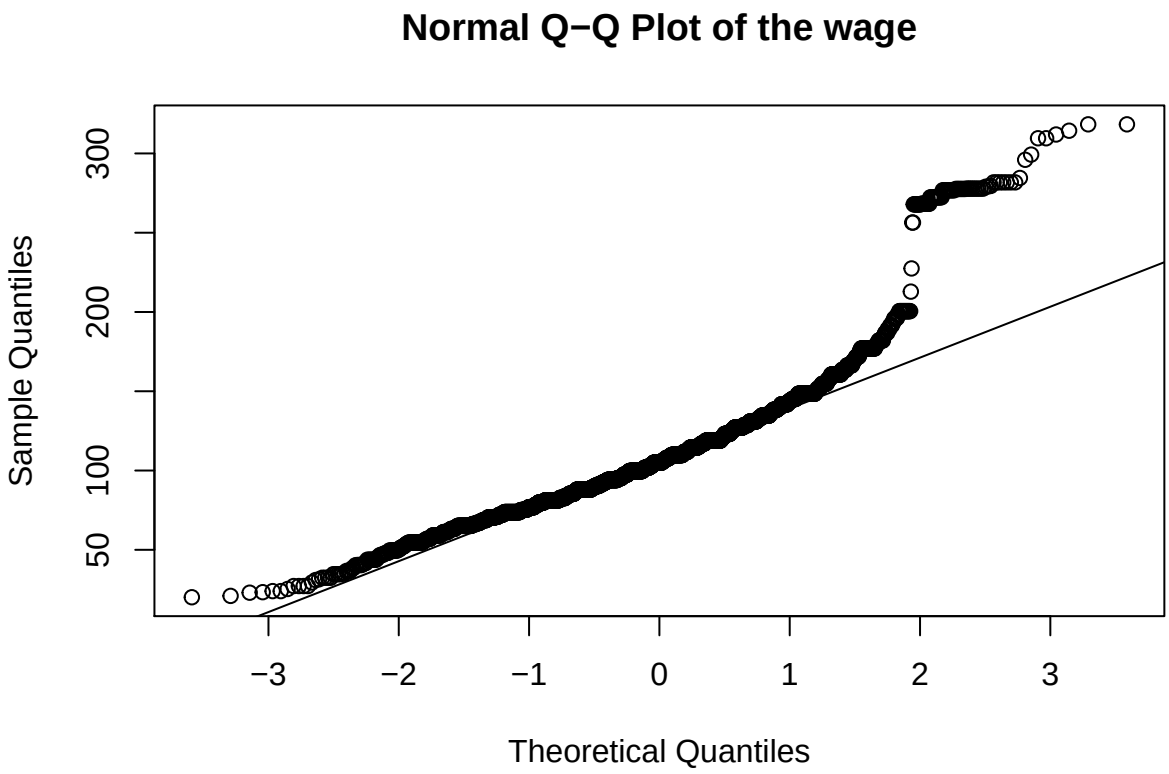
Se puede obserbar que el salario es ligeramente superior para personas de raza asiática, aunque el rango intercuartílico (el segundo cuartil, corresponiente a la caja) es algo más grande que el resto de razas, por lo que, a pesar de que la mediana del salario en personas asiáticas es superior a la del resto de razas, existe más variedad de salario dentro de su propia raza. También, el salario máximo es superior al del resto de razas, obviando los *outliers*. En el caso de la raza blaca observamos una mayor cantidad de *outliers* superiores. En el caso del tipo de trabajo, vemos que por lo general el salario es mayor para trabajos del tipo *Information*, así como para los casos en los que la salud es muy buena. Finalmente también denota un sueldo mayor que el individuo tenga contratado un seguro de salud.

2.3. Comprobación de normalidad

¿Podemos asumir que la variable wage tiene una distribución normal? Justificar la respuesta a partir de métodos visuales.

Utilizaremos la función *qqnorm* para visualizar la linea que forman los datos sobre la recta de la normal.

```
qqnorm(y = data$wage, main = "Normal Q-Q Plot of the wage")
qqline(y = data$wage)
```



Como vemos, exceptuando los últimos cuantiles, los valores se encuentran muy próximos a la recta. Por ello, asumiremos que la variable tiene una distribución normal.

3. Estadística inferencial

3.1. Intervalo de confianza de la variable age

a) Calcular el intervalo de confianza al 95% de la variable age de los trabajadores. A partir del valor obtenido, explicad como se interpreta el resultado del intervalo de confianza b) Calcular los intervalos de confianza al 95% de la variable age, segregando los trabajadores por la variable jobclass. ¿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.

3.1.1. Intervalo al 95%

Para comenzar con los cálculos, necesitamos disponer de la media de edad.

```
X <- mean(data$age)
X
```

```
## [1] 42.41467
```

S (desviación típica para una distribución de Student) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2}$

```r
compute_S <- function(vector){
  # S
  # (xi-xmean)^2
  sum = 0
  m = mean(vector)
  for(a in vector){
    x = (a - m)^2
    sum = sum + x
  }
  #1/n - 1
  y = 1 / (length(vector) - 1)
  #sqrt
  S = sqrt(sum / (length(vector) - 1))
}

S <- compute_S(data$age)
S
```

```
## [1] 11.54241
```

$\alpha = 0.05$ (indicado en el enunciado)

Para conocer el intervalo de confianza, deberemos realizar la siguiente operación:

$$P(-|t_{n-1}| <= \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} <= |t_{n-1}|) = 0.95$$

```r
tn1 = qt(p=(1 - 0.95)/2, df = length(data$age) - 1)
tn1
```

```
## [1] -1.960755
```

$$P(-1.64 <= \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} <= 1.64) = 0.95$$

$$P(\bar{X} - 1.64\frac{S}{\sqrt{n}} <= \mu <= \bar{X} + 1.64\frac{S}{\sqrt{n}}) = 0.95$$

```r
margen_error <- abs((tn1 * S)/sqrt(length(data$age)))
limite_superior <- X + margen_error
limite_inferior <- X - margen_error
limite_inferior
```

```
## [1] 42.00147
```

```r
limite_superior
```

```
## [1] 42.82787
```

En esta muestra, la edad abarca desde 42 hasta 43 con un intervalo de confianza del 95%.

3.1.2. Intervalo al 95% en función de jobclass

En primer lugar, obtendremos el intervalo de confianza de los trabajadores cuyo oficio es industrial.

```
data_industrial <- data$age[which(data$jobclass == levels(data$jobclass)[1])]
X <- mean(data_industrial)
S <- compute_S(data_industrial)

tn1 <- qt(p=(1 - 0.95)/2, df = length(data_industrial) - 1)

margen_error <- abs((tn1 * S)/sqrt(length(data_industrial)))
limite_superior <- X + margen_error
limite_inferior <- X - margen_error
limite_inferior
```

```
## [1] 40.81469
```

```
limite_superior
```

```
## [1] 41.98195
```

Y a continuación, los que tienen un trabajo "Informativo".

```
data_information <- data$age[which(data$jobclass == levels(data$jobclass)[2])]
X <- mean(data_information)
S <- compute_S(data_information)

tn1 <- qt(p=(1 - 0.95)/2, df = length(data_information) - 1)

margen_error <- abs((tn1 * S)/sqrt(length(data_information)))
limite_superior <- X + margen_error
limite_inferior <- X - margen_error
limite_inferior
```

```
## [1] 42.91223
```

```
limite_superior
```

```
## [1] 44.07266
```

Podemos concluir que la edad de los trabajadores del sector industrial es ligeramente inferior que los del sector informativo, ya que tienen intervalos de confianza de, aproximadamente (41, 42) y (43, 44) respectivamente.

3.2. Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que los trabajadores que tienen contratado un seguro médico variable (health_ins) tienen un salario (wage) que supera en más de 20$ (en miles de dólares) el salario de los que no tienen seguro médico? Calcularlo para un nivel de confianza del 95%.

3.2.1. Escribir la hipótesis nula y alternativa

$$\begin{cases} H_0: & \mu_0 - \mu_1 = 0 \\ H_1: & \mu_0 - \mu_1 > 0 \end{cases}$$

Siendo μ_0 el salario de los empleados que tienen contratado el seguro médico, y μ_1 el salario de los que no tienen dicho seguro contratado.

3.2.2. Método

Se trata de un contraste unilateral de datos independientes. Dado que no conocemos la varianza, procederemos con la ley t de Student con n-1 grados de libertad. Para calcular el estadístico de contraste, necesitaremos previamente aproximar la desviación estándar por la desviación estándar muestral, calculando S. Para ello disponemos de una función definida previamente.

3.2.3. Cálculos

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

```
salary_insurance <- data$wage[which(data$health_ins == levels(data$health_ins)[1])]
salary_not_insurance <- data$wage[which(data$health_ins == levels(data$health_ins)[2])]
```

T de Student Para decidir si rechazamos la hipótesis nula o no, calcularemos el estadístico de contraste con la siguiente formula.

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Seguiremos la ley de t de Student con n-1 grados de libertad, dado que no conocemos la varianza (σ)

- $\bar{X}$

```
X1 <- mean(salary_insurance)
X2 <- mean(salary_not_insurance)
```

- **S** Dado que no conocemos la desviación típica, la calcularemos con la siguiente fórmula:

$$S = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x_1})^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x_2})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

```
S1 <- compute_S(salary_insurance)
S2 <- compute_S(salary_not_insurance)
```

```
S <- sqrt(((length(salary_insurance) - 1) * S1^2 + (length(salary_not_insurance) - 1) * S2^2) / (length
S
```

```
## [1] 39.70245
```

- **t**

Procedemos a calcular el estadístico de contraste t, con la siguiente fórmula.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

```
error_estandar <- (S*sqrt((1 / length(salary_insurance)) + (1 / length(salary_not_insurance))))
t <- (X1 - X2) / error_estandar
t
```

```
## [1] 17.74565
```

- **Valor crítico**

```
vc <- qt(p=0.05, df=2999)
vc
```

```
## [1] -1.645362
```

Este valor determinará el intervalo de confianza.

```
intervalo <- c((X1-X2)+(vc*error_estandar), (X1-X2)-(vc*error_estandar))
intervalo
```

```
## [1] 25.33274 30.51048
```

- **P-valor**

```
1 - pt(t, df = 2999)
```

```
## [1] 0
```

3.3.4. Interpretación

Puesto que el p-valor $< \alpha$, establecido en el enunciado a 0.05, rechazaremos la hipótesis nula, y aceptaremos la alternativa. Además, el intervalo de confianza al 95% indica que la diferencia de salarios es entre 25.3k y 30.5k, que es más de 20k, lo indicado en la hipótesis alternativa.

3.3. Contraste no paramétrico

A la hora de escoger entre los tests no-paramétricos estudiados: *"Suma de rangos (o test U)"* y *"test de rangos y signos del Wilcoxon"*, partimos de la base de que el primero establece la comparación entre medias de poblaciones independientes, mientras que el segundo se trata de un test para comparar medias entre muestras dependientes. Dado que en nuestro caso trabajamos con muestras independientes, aplicaremos el test U.

3.3.1. Aplicación del test

Aplicad un contraste no paramétrico para responder la misma pregunta anterior. Podéis usar funciones R.

```r
wilcox.test(data$wage ~ data$health_ins, alternative = "greater", paired = FALSE, conf.int = 0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data$wage by data$health_ins
## W = 1401710, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
## 95 percent confidence interval:
##  23.75359      Inf
## sample estimates:
## difference in location
##               25.91366
```

3.3.2. Interpretación

Otra vez, encontramos que el p-valor es menor que α, por lo que aceptamos la hipótesis alternativa, como indica el test realizado.

3.3.3. Paramétrico V.S. no paramétrico

Justificad qué tipo de contraste (paramétrico/no paramétrico) se debería aplicar en este caso.

La principal diferencia entre estas pruebas es que en las paramétricas deben cumplirse previamente algunas condiciones, como que la muestra tenga una distribución normal (asuncion que hemos realizado en el enunciado del ejercicio). Por otra parte las pruebas no paramétricas no deben ajustarse a ninguna distribución, por lo que son más robustas. Dado que no conocemos la varianza ni la distribución de los datos (pese a que la estamos asumiendo), considero más seguro utilizar un test no paramétrico.

4. Regresión logística

4.1. Modelo predictivo

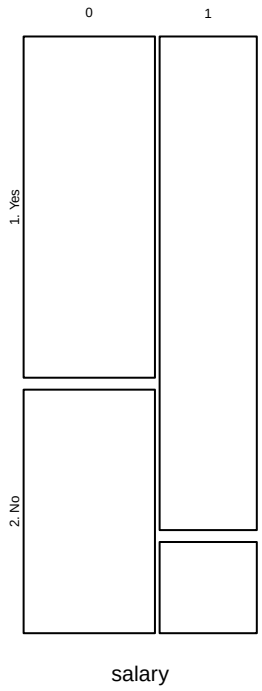
Ajustad un modelo predictivo basado en regresión logística para predecir la probabilidad de tener un salario superior a la media en función de las variables: health_ins, jobclass y age. Tomad como salario medio el valor de la media muestral de la variable wage. Podéis codificar como 0 cuando el salario es inferior a la media y 1 cuando el salario es superior o igual.

Para correlacionar si el salario está por encima de la media en función de ciertas variables explicativas, debemos estipular un vector en el que marcaremos con 1 si el salario de ese individuo está por encima de la media y 0 en caso contrario.

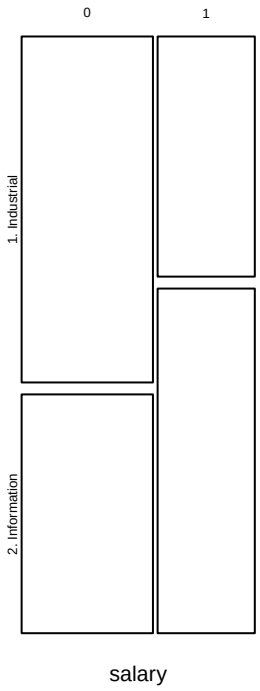
```r
salary_mean <- mean(data$wage)
salary <- c()
salary[which(data$wage >= salary_mean)] <- 1
salary[which(data$wage < salary_mean)] <- 0
```

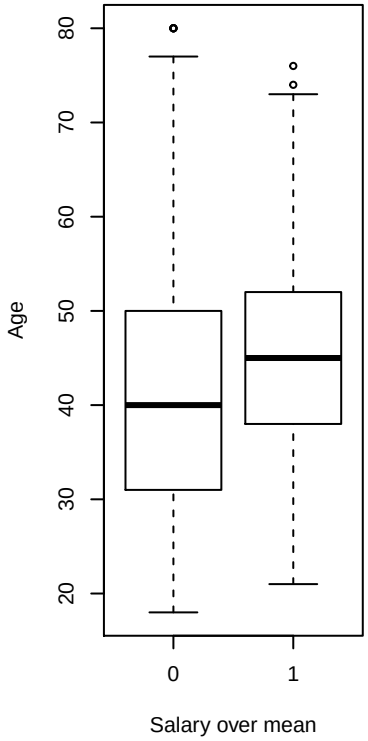
Antes de comenzar, estudiaremos con el *Test Chi Square* si ambas variables están relacionadas o no, tras visualizar la relación entre el salario con respecto a la media y las variables explicativas.

Salary and health insurance



Salary and job class






```
chisq.test(table(salary, data$health_ins))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(salary, data$health_ins)
## X-squared = 233.26, df = 1, p-value < 2.2e-16
```

```
chisq.test(table(salary, data$jobclass))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(salary, data$jobclass)
## X-squared = 95.439, df = 1, p-value < 2.2e-16
```

```
chisq.test(table(salary, data$age))
```

```
## Warning in chisq.test(table(salary, data$age)): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(salary, data$age)
## X-squared = 248.53, df = 60, p-value < 2.2e-16
```

Dado que en todos los casos el p-valor es menor que 0.05, podemos concluir que están asociadas.

A continuación, podemos ajustar el modelo predictivo basado en regresión logística.

```
model <- glm(salary ~ health_ins + jobclass + age, data=data, family="binomial")
summary(model)
```

```
##
## Call:
## glm(formula = salary ~ health_ins + jobclass + age, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.752  -1.042  -0.586   1.104   2.062
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.373068   0.163067  -8.420  < 2e-16 ***
## health_ins2. No        -1.224173   0.093342 -13.115  < 2e-16 ***
## jobclass2. Information  0.586006   0.078709   7.445 9.68e-14 ***
## age                     0.025984   0.003509   7.405 1.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4091.7  on 2999  degrees of freedom
## Residual deviance: 3726.2  on 2996  degrees of freedom
## AIC: 3734.2
##
## Number of Fisher Scoring iterations: 4
```

```
odds <- exp(coef(model))
odds
```

```
##            (Intercept)        health_ins2. No jobclass2. Information
##              0.2533284              0.2940007              1.7967984
##                    age
##              1.0263243
```

4.2. Interpretación

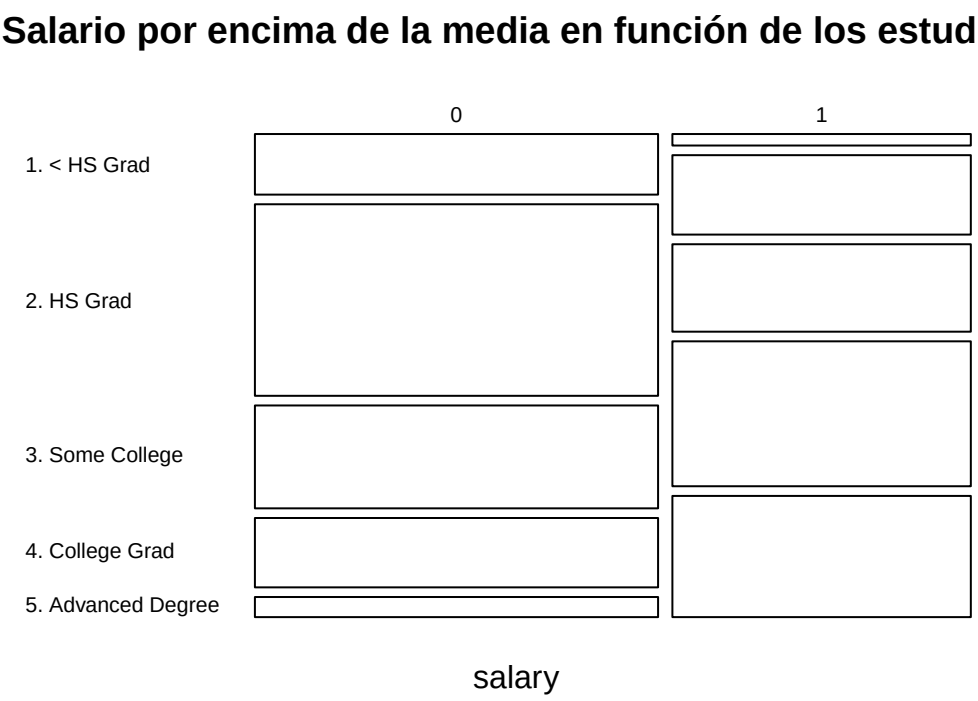
Interpretad el modelo ajustado. Concretamente, explicad la contribución de las variables explicativas para predecir si el salario es superior o inferior a la media.

Como podemos observar, todas las variables influyen en que el salario esté por encima de la media, ya que el p-value de cada una de ellas es menor que 0.05. Concretamente, que el tipo de trabajo sea "Informativo" es una de las características que más influyen en que su sueldo esté por encima de la media, aumentando

en 1.79 unidades la probabilidad de tener un sueldo mayor. Sin embargo, es más notable la influencia de que el individuo tenga contratado un seguro de salud, siendo esta la variable que más hace aumentar la probabilidad de que el sueldo se encuentre por encima de la media, hasta en 3.4 puntos, que es la inversa de 0.294.

4.3. Importancia del nivel de estudios

Añadid al modelo anterior la variable education. Interpretad los niveles de la variable education a partir del odds ratio. ¿En qué porcentaje se ve incrementada la probabilidad de tener un salario superior al salario medio según el nivel educativo? Proporcionad intervalos de confianza del 95% de los odds ratio.



De la variable *education* lo más destacado es la determinación de que el sueldo esté por debajo de la media si individuo tiene el graduado del instituto o una titulación inferior. Sin embargo, si dispone de titulación universitaria o un título superior, se dan más casos en los que el salario es superior a la media que casos en los que es inferior. Sin embargo, puesto que tenemos que ajustar el modelo junto con otras variables explicativas, estos datos no tienen porqué ser significativos.

```
model_edu <- glm(salary ~ health_ins + jobclass + age + education, data=data, family="binomial")
summary(model_edu)
```

```
##
## Call:
## glm(formula = salary ~ health_ins + jobclass + age + education,
##     family = "binomial", data = data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2149  -0.8726  -0.4314   0.8940   2.5503
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.788611   0.264697 -10.535  < 2e-16 ***
## health_ins2. No             -1.108971   0.102322 -10.838  < 2e-16 ***
## jobclass2. Information       0.105917   0.088599   1.195  0.23191
## age                          0.028543   0.003862   7.390 1.47e-13 ***
## education2. HS Grad          0.647155   0.206820   3.129  0.00175 **
## education3. Some College     1.363390   0.210033   6.491 8.51e-11 ***
## education4. College Grad     2.214955   0.209626  10.566  < 2e-16 ***
## education5. Advanced Degree  3.190188   0.233075  13.687  < 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: 4091.7  on 2999  degrees of freedom
## Residual deviance: 3242.6  on 2992  degrees of freedom
## AIC: 3258.6
##
## Number of Fisher Scoring iterations: 4
```

```
odds <- exp(coef(model_edu))
odds
```

```
##                 (Intercept)             health_ins2. No
##                  0.06150659                  0.32989831
##      jobclass2. Information                         age
##                  1.11172958                  1.02895467
##         education2. HS Grad    education3. Some College
##                  1.91009903                  3.90942444
##    education4. College Grad education5. Advanced Degree
##                  9.16100049                 24.29299592
```

Como podemos ver, el tipo de trabajo deja de ser relevante en este caso, ya que el p-valor > 0.05. Sin embargo, la edad y si dispone de seguro de salud se mantienen influyentes y pasa a cobrar mucha más relevancia el tipo de estudios que haya obtenido esa persona. Concretamente, si tiene un *Advanced Degree*, aumenta en 24.29 puntos la probabilidad de obtener un salario mayor que la media.

4.4. Predicción

¿Superaría el salario medio un trabajador con seguro médico, que trabaja en el ámbito de la información y con 42 años de edad y formación de graduado? ¿Y si se trata de un trabajador del ámbito industrial?

```
trabajador <- data.frame(health_ins = levels(data$health_ins)[1], jobclass = levels(data$jobclass)[2], a

predict(model_edu, trabajador, type="response")
```

```
##         1
## 0.6750432
```

```
trabajador <- data.frame(health_ins = levels(data$health_ins)[1], jobclass = levels(data$jobclass)[1], a

predict(model_edu, trabajador, type="response")
```

```
##         1
## 0.6513929
```

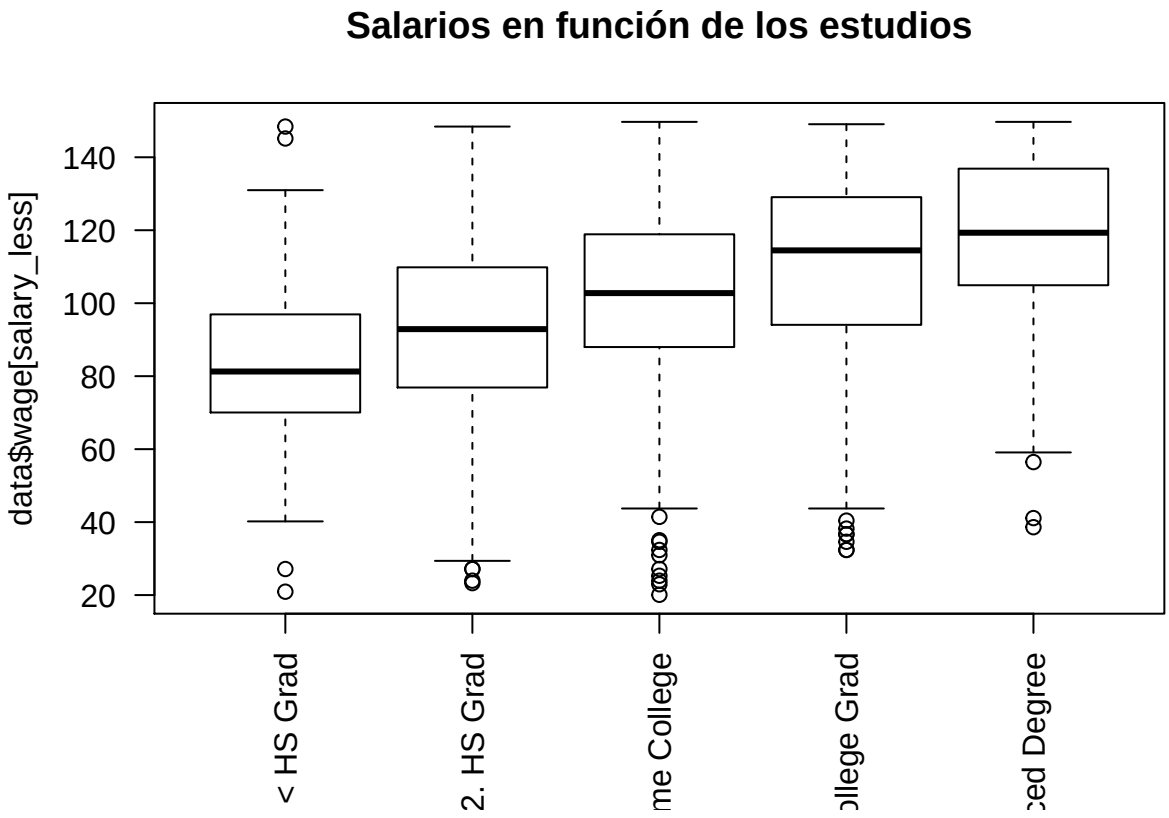
La probabilidad de que esta persona supere el salario medio es de 67,5% si trabaja en el ámbito de la Información, mientras que si trabaja en el ámbito industrial, su probabilidad es ligéramente inferior, del 65,13%.

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

5.1. Nivel de educación y salario

Seleccionar las observaciones que tengan un salario inferior a 150000$. Para este grupo de trabajadores realizad un Anova para contrastar si existen diferencias en el salario según el nivel de educación.

```
salary_less <- which(data$wage < 150.000)
```



5.1.1. Hipótesis nula y alternativa

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ H_1: & No\ todos\ los\ salarios\ son\ iguales \end{cases}$$

```
levels(data$education)
```

```
## [1] "1. < HS Grad"       "2. HS Grad"         "3. Some College"
## [4] "4. College Grad"    "5. Advanced Degree"
```

Siendo μ_1 el salario de personas con el tipo de educación "1. < HS Grad", μ_2 de 2. HS Grad" y así sucesivamente.

5.1.2. Modelo

Calcular 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(wage ~ education, data = data[salary_less,])
summary(anova)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## education      4  267049   66762   120.5 <2e-16 ***
## Residuals   2652 1468949     554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Siendo la primera columna el número de grados de libertad, *Sum Sq* la suma de cuadrados, que determinan el estadístico de contraste (F de Snedecor), determinado con *F value*. *Mean Sq* indica la media de cuadrados y el p-valor está determinado por *Pr(>F)*.

Los grados de libertad entre grupos del factor educación son 4 ya que está determinado por n-1, siendo n = 5 (el número de tipos de educación diferentes), mientras que dentro de dichos grupos el grado de libertad es 2652 al ser el total de observaciones 2657 - 5 (número de grupos diferentes). La suma de los cuadrados entre los grupos educativos está dada por la formula $SCE = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$, siendo k = 5. 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. $fvalue = \frac{66762}{554} = 120.5$.

5.1.3. Cálculos

Para profundizar en la comprensión del modelo ANOVA, calcular manualmente la suma de cuadrados intra y la suma de cuadrados entre grupos. Los resultados deben coincidir con el resultado del modelo ANOVA. Como referencia, podéis obtener las fórmulas de López-Roldán i Fachelli (2015), páginas 29-33.

En este punto, realizaremos los cálculos manuales de las formulas representadas en el punto anterior, de suma de cuadrados entre grupos e intra grupal.

En primer lugar, calcularemos el SCE, siendo la j del bucle cada nivel educativo.

```r
mean_all_groups <- aggregate(data$wage[salary_less], by=list(data$education[salary_less]), FUN=mean)[,"
length_all_groups <- aggregate(data$wage[salary_less], by=list(data$education[salary_less]), FUN=length
mean_all <- sum(length_all_groups*mean_all_groups) / sum(length_all_groups)
mean_all
```

```
## [1] 100.8741
```

```r
sum_sq <- 0
for (j in levels(data$education)){
  data_j <- data$wage[intersect(salary_less, which(data$education == j))]
  nj <- length(data_j)
  c <- (mean(data_j) - mean_all)^2
  sum_sq <- sum_sq + nj*c
}
sum_sq
```

```
## [1] 267049.5
```

Para calcular la suma de cuatrados dentro de los grupos, la formula varía ligeramente.

```r
sum_sq <- 0
for (j in levels(data$education)){
  data_j <- data$wage[intersect(salary_less, which(data$education == j))]
  sum_group <- 0
  for (xij in data_j){
    sum_group <- sum_group + (xij - mean(data_j))^2
  }
  sum_sq <- sum_sq + sum_group
}
sum_sq
```

```
## [1] 1468949
```

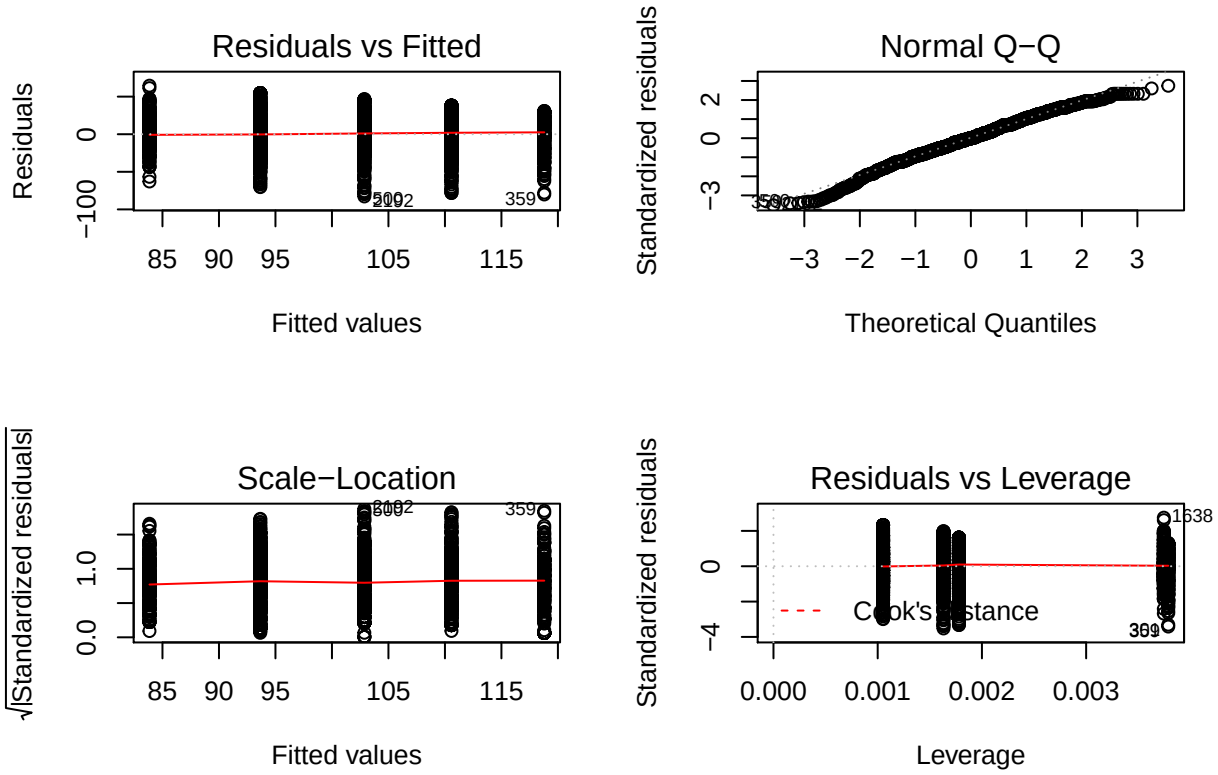
Como podemos observar, calculandolo manualmente se obtiene el mismo resultado que con la función *anov*.

5.1.4. Interpretación

Como podemos ver, el p-valor muestra $P(F > fvalue) = P(F > 120.5) \approx 0$. Por tanto, dado que el p-valor es inferior a 0.05, rechazamos la hipótesis nula, y aceptamos que la media de los salarios entre los diferentes niveles educativos es diferente.

5.2. Adecuación al modelo

Mostrad visualmente la adecuación del modelo ANOVA. Podéis usar plot sobre el modelo ANOVA calculado. En los apartados siguientes, realizad la interpretación de estos gráficos.



5.2.1. Normalidad de los residuos

Interpretar la normalidad de los residuos a partir del gráfico Normal Q-Q que habéis mostrado en el apartado anterior.

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

5.2.2. Homocedasticidad de los residuos

Los gráficos "Residuals vs Fitted", "Scale-Location" y "Residuals vs Factor levels" proporcionan información sobre la homcedasticidad de los residuos. Interpretad estos gráficos.

Para ver si nuestro modelo tiene la propiedad de la homocedasticidad, es decir, la varianza de sus errores es constante, observaremos diferentes gráficos. La gráfica *Residuals vs Fitted* muestra que no existe relación entre los residuos y la media de valores de cada grupo (se aprecian 5 líneas de puntos verticales, correspondientes a cada nivel de estudios). 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. De forma similar a esta gráfica, la *Scale-Location* muestra si los residuos se incrementan con los valores predecidos, sin embargo en este caso no lo hacen. La línea roja indica la dispersión de las predicciónes de cada nivel educativo, y al ser prácticamente plana, podemos asumir homocedasticidad. Finalmente en la gráfica *Residuals vs Leverage* muestra, con la distancia de Cook, la influencia de cada punto en el cómputo total, es decir, si algún outlier está modificando el resultado. En nuestro caso no tenemos nungún outlier que sea influyente, aunque los hay, como el 1638 o 359. De ser influyente sería recomendable quitarlo de la muestra y realizar el estudio otra vez.

5.3. ANOVA no paramétrico

Si las asunciones de normalidad y homocedasticidad no se cumplen, se puede aplicar un contraste no paramétrico como el test de Kruskal-Wallis.

5.3.1. Test Kruskal-Wallis

Aplicad el test de Kruskal-Wallis para contrastar si hay diferencias en el salario según la raza (race). Como en el apartado anterior, seleccionad las observaciones con salario inferior a 150 y además, descartad los casos en que la variable race tenga el valor "4. Other". Podéis usar funciones R que calculen el test Kruskal-Wallis.

Test no paramétrico que se presenta como alternativa al ANOVA para datos no pareados, como es nuestro caso. La finalidad de esta prueba es contrastar si las muestras están equidistribuidas, y si pertenecen a una misma población.

```
salary_less_race <- intersect(which(data$wage < 150.000), which(data$race != levels(data$race)[4]))
data_no_others <- data[salary_less_race,]
data_no_others$race <- droplevels(data_no_others$race, levels(data$race)[4])
```


```
kruskal.test(wage ~ race, data = data_no_others)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  wage by race
## Kruskal-Wallis chi-squared = 11.661, df = 2, p-value = 0.002937
```

5.3.2 Interpretación de los resultados

Interpretad los resultados del test Kruskal-Wallis.

La hipótesis principal es que el sueldo de cada raza humana es igual, sin embargo, el test devuelve un p-valor inferior a 0.05, que indica que la hipótesis nula es falsa, por lo que concluimos que el sueldo varía según la raza.

6. ANOVA multifactorial

A continuación, se desea evaluar el efecto de la raza combinado con otro factor. Primero se realizará el análisis con el factor tipo de trabajo (jobclass) y posteriormente, con el factor nivel de educación (education). En este apartado seleccionad las observaciones con un salario inferior a 150 y además, descartad las observaciones en que la variable race tengan el valor "4. Other", como se ha hecho anteriormente.

6.1. Factores: Raza y tipo de trabajo

6.1.1. Análisis visual de los efectos principales y posibles interacciones

Dibujar en un gráfico la variable wage en función de la raza (race) y en función del tipo de trabajo (jobclass). El gráfico ha de permitir evaluar si hay interacción entre los dos factores

```
##
## Attaching package: 'dplyr'
```

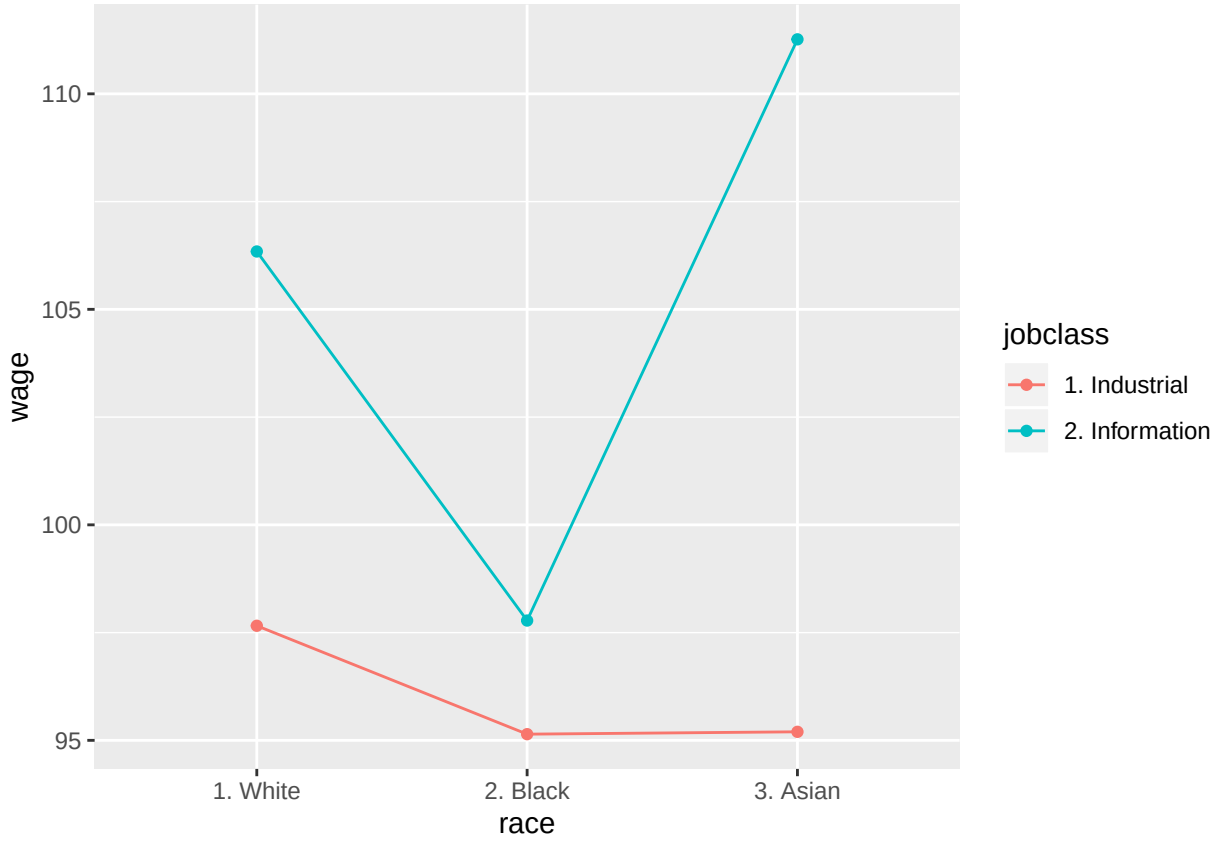
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
t <- as.data.frame(data_no_others %>%
                     group_by(race, jobclass) %>%
                     summarise(mean(wage)))
names(t) <- c("race", "jobclass", "wage")
t
```

```
##       race       jobclass      wage
## 1 1. White  1. Industrial  97.65866
## 2 1. White 2. Information 106.34269
## 3 2. Black  1. Industrial  95.14360
## 4 2. Black 2. Information  97.77960
## 5 3. Asian  1. Industrial  95.19800
## 6 3. Asian 2. Information 111.26418
```

```
ggplot(data = t, aes(race, wage, color=jobclass, group = jobclass)) + geom_point() + geom_line()
```



Los efectos de estos factores se pueden apreciar en el gráfico. El factor tipo de trabajo *Industrial* determina un sueldo menor que el tipo *Information*. Por otra parte, la raza negra determina un sueldo menor en ambos tipos de trabajo.

6.1.2. Modelo ANOVA

Aplicar un modelo anova con estos factores y su posible interacción. A continuación, analizar si la interacción es significativa.

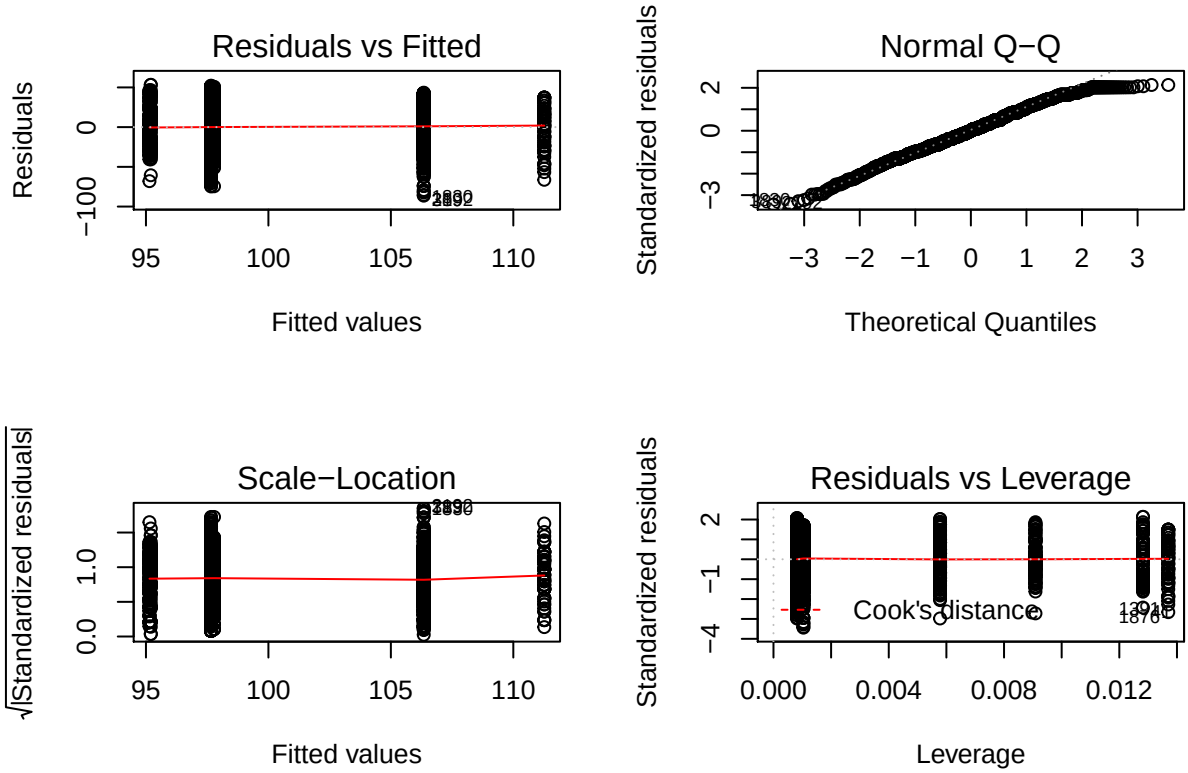
```
anova_rj <- aov(wage ~ race * jobclass, data = data_no_others)
summary(anova_rj)
```

```
##                 Df  Sum Sq Mean Sq F value  Pr(>F)
## race             2    6176    3088   4.898 0.00753 **
## jobclass         1   46347   46347  73.515 < 2e-16 ***
## race:jobclass    2    4489    2245   3.560 0.02857 *
## Residuals     2615 1648621     630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como se puede apreciar, el p-valor de todos los factores, y su interacción, son significativos, dado que es menor que 0.05. Sin embargo, lo más determinante es el tipo de trabajo, frente a la raza o la interacción entre ambos.

6.1.3 Adecuación del modelo

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



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 cómo se distribuyen los residuos de gada grupo en una línea vertical sobre el 0, generando una línea roja suficientemente recta. Podremos 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 en entre las dos últimas agrupaciones de puntos verticales apreciamos una ligera subida, dada a causa del incremento en la dispersión de los residuos predecidos del último grupo.

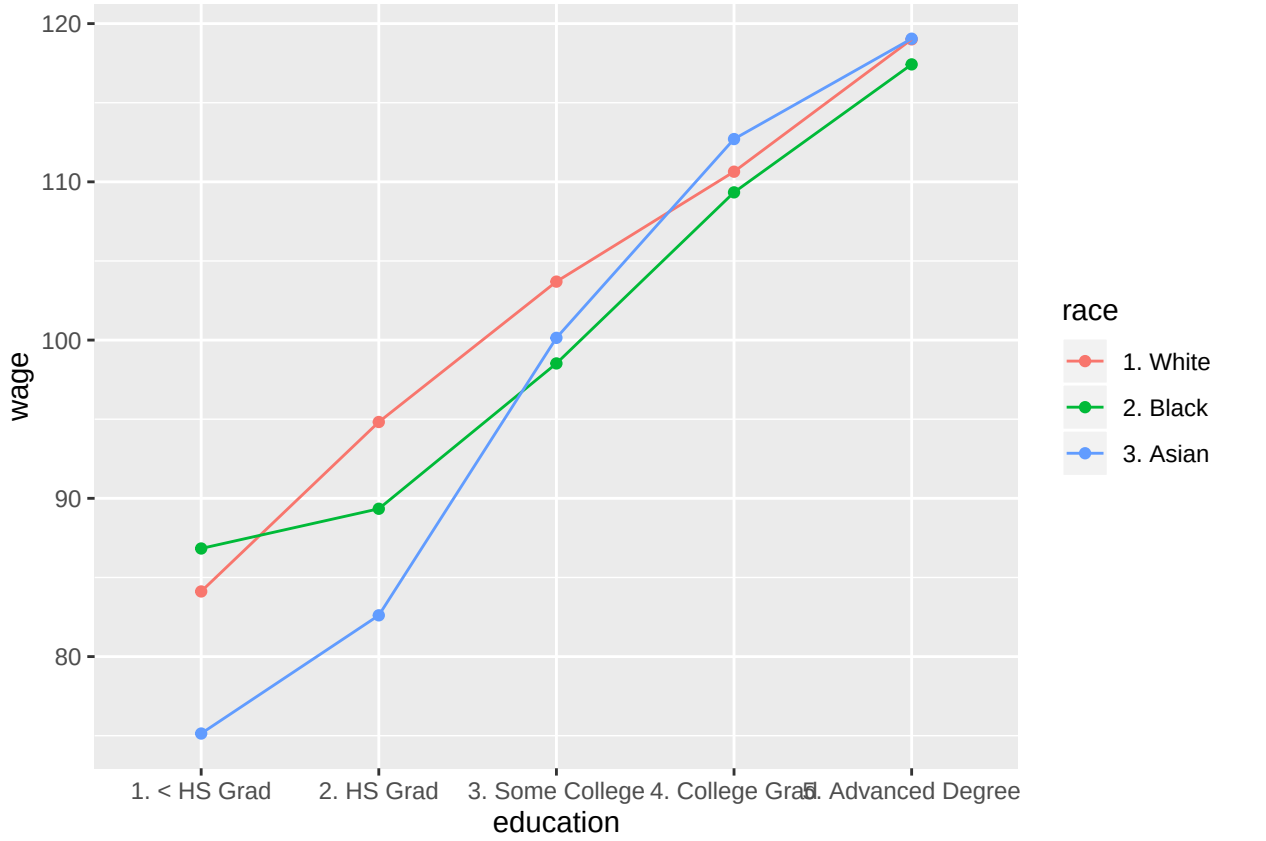
6.2. Factores: raza y nivel de educación

Seguid los mismos pasos que el apartado anterior para aplicar un modelo ANOVA con los factores raza y nivel de educación.

6.2.1. Análisis visual de los efectos principales y posibles interacciones

```
t <- as.data.frame(data_no_others %>%
                     group_by(race, education) %>%
                     summarise(mean(wage)))
```

```r
names(t) <- c("race", "education", "wage")
ggplot(data = t, aes(education, wage, color=race, group = race)) + geom_point() + geom_line()
```



Al visualizar el salario según el nivel de estudios y la raza del individuo se puede apreciar que a medida que el nivel de estudios es superior, el sueldo se incrementa, independientemente de la raza. Si bien es cierto que se aprecian algunas diferencias entre razas dentro de cada nivel de estudios.

6.2.2. Modelo ANOVA

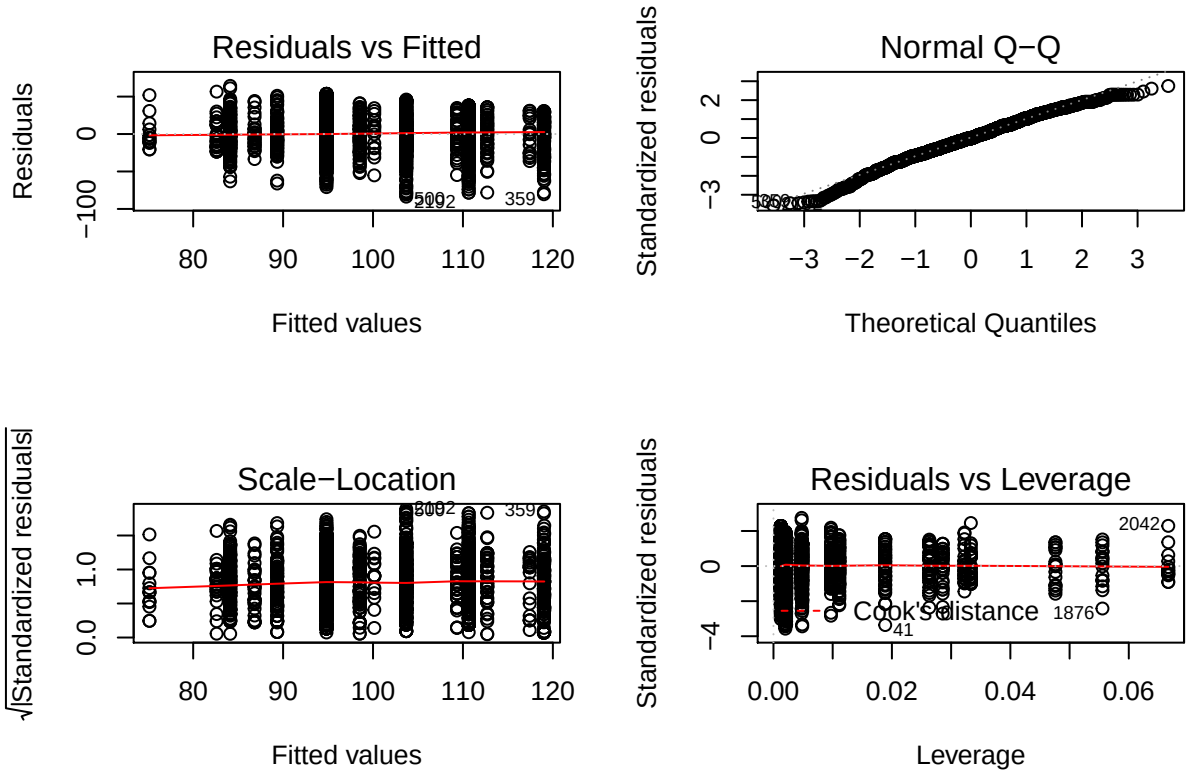
```r
anova_re <- aov(wage ~ race * education, data = data_no_others)
summary(anova_re)
```

```
##                  Df  Sum Sq Mean Sq F value  Pr(>F)
## race              2    6176    3088   5.615 0.00369 **
## education         4  259984   64996 118.180 < 2e-16 ***
## race:education    8    6241     780   1.418 0.18340
## Residuals      2606 1433232     550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como se puede observar, el factor más significativo es el que determina el nivel educativo de la persona, mientras que la raza se sigue manteniendo por debajo de 0.05, es decir, sigue siendo influyente, aunque en menor medida que el factor anterior. Sin embargo, cabe destacar que no existe interacción entre ambos

factores. El p-valor de la interacción raza y educación está por encima de 0.05, por lo que habrá que descartarlo.

6.2.3. Adecuación del modelo



La gráfica *Normal Q-Q* indica que los residuos siguen una distribución normal, dada la cercanía de los puntos a la recta. Si bien es cierto que no se muestra tan recta como en ejemplos anteriores, seguimos manteniendo la normalidad de los residuos. En *Residuals vs Fitted* observamos cómo se distribuyen los residuos de gada grupo en una línea vertical sobre el 0, generando una línea roja sobre el 0, lo cual indica que podremos 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. Sin embargo, aunque no existan outliers influyentes, en este caso se aprecian residuos más alejados de su agrupación, como es el 41, 1876 o 2042. En la gráfica *Scale-Location* observamos que la línea roja, es bastante recta, sin embargo, las "subidas" o "bajadas" de esta linea indican el aumento o disminución de la dispersión de los residuos.

7. Comparaciones múltiples

Tomando como referencia el modelo ANOVA multifactorial, con los factores raza y tipo de trabajo, aplicar el test de comparación múltiple Scheffé. Interpretar el resultado del test e indicar qué grupos son significativamente diferentes entre si.

```
ScheffeTest(anova_rj)
```

```
##
```

```
##   Posthoc multiple comparisons of means : Scheffe Test
##     95% family-wise confidence level
##
## $race
##                         diff     lwr.ci     upr.ci   pval
## 2. Black-1. White -4.723515 -10.005219  0.5581897 0.1149
## 3. Asian-1. White  1.486575  -5.548239  8.5213894 0.9923
## 3. Asian-2. Black  6.210090  -2.215618 14.6357983 0.3043
##
## $jobclass
##                                   diff  lwr.ci   upr.ci    pval
## 2. Information-1. Industrial 8.387704 5.11149 11.66392 4.5e-14 ***
##
## $`race:jobclass`
##                                                        diff     lwr.ci    upr.ci
## 2. Black:1. Industrial-1. White:1. Industrial   -2.5150517 -10.836894  5.806791
## 3. Asian:1. Industrial-1. White:1. Industrial   -2.4606545 -12.224032  7.302723
## 1. White:2. Information-1. White:1. Industrial   8.6840362   5.082299 12.285774
## 2. Black:2. Information-1. White:1. Industrial   0.1209395  -6.669623  6.911502
## 3. Asian:2. Information-1. White:1. Industrial  13.6055261   3.532705 23.678347
## 3. Asian:1. Industrial-2. Black:1. Industrial    0.0543972 -12.321552 12.430346
## 1. White:2. Information-2. Black:1. Industrial  11.1990878   2.784037 19.614139
## 2. Black:2. Information-2. Black:1. Industrial   2.6359912  -7.559709 12.831691
## 3. Asian:2. Information-2. Black:1. Industrial  16.1205778   3.499077 28.742079
## 1. White:2. Information-3. Asian:1. Industrial  11.1446906   1.301745 20.987636
## 2. Black:2. Information-3. Asian:1. Industrial   2.5815940  -8.821160 13.984348
## 3. Asian:2. Information-3. Asian:1. Industrial  16.0661806   2.451013 29.681348
## 2. Black:2. Information-1. White:2. Information -8.5630967 -15.467570 -1.658623
## 3. Asian:2. Information-1. White:2. Information  4.9214900  -5.228473 15.071453
## 3. Asian:2. Information-2. Black:2. Information 13.4845867   1.815784 25.153390
##                                                      pval
## 2. Black:1. Industrial-1. White:1. Industrial    0.9615
## 3. Asian:1. Industrial-1. White:1. Industrial    0.9827
## 1. White:2. Information-1. White:1. Industrial  2.1e-12 ***
## 2. Black:2. Information-1. White:1. Industrial   1.0000
## 3. Asian:2. Information-1. White:1. Industrial   0.0012 **
## 3. Asian:1. Industrial-2. Black:1. Industrial    1.0000
## 1. White:2. Information-2. Black:1. Industrial   0.0015 **
## 2. Black:2. Information-2. Black:1. Industrial   0.9806
## 3. Asian:2. Information-2. Black:1. Industrial   0.0029 **
## 1. White:2. Information-3. Asian:1. Industrial   0.0145 *
## 2. Black:2. Information-3. Asian:1. Industrial   0.9894
## 3. Asian:2. Information-3. Asian:1. Industrial   0.0088 **
## 2. Black:2. Information-1. White:2. Information  0.0045 **
## 3. Asian:2. Information-1. White:2. Information  0.7603
## 3. Asian:2. Information-2. Black:2. Information  0.0114 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Este test nos muestra las comparaciones entre todas las combinaciones posibles, para establecer así cuales son los grupos más significativamente diferentes entre sí. Se puede observar que los grupos diferenciados por el tipo de trabajo (*Information* o *Industrial*), devuelve la diferencia más amplia, como habíamos venido viendo, teniendo el p-value más cercano a 0. Sin embargo lo más interesante de este test son las comparaciones entre

la raza y el tipo de trabajo que realiza. En este caso, observamos una significación mayor en la diferencia entre una persona blanca que trabaja en el sector de la información con otro del mismo color en el sector industrial, aunque también tendría significación la diferencia entre personas asiáticas trabajando en sectores diferentes. La última columna de la tabla, llamada *pval*, indica si la diferencia es significativa o no, en función de si el valor es menor que α, que se ha asumido a 0.05.

8. Conclusiones

Resumid las conclusiones principales del análisis. Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Para resumir la práctica, se realizó el ejercicio 3 partiendo del previo establecimiento de la normalidad de la variable *wage*, a partir de la cual se basarán algunos de los ejercicios siguientes. En el tercer ejercicio se calculó el intervalo de confianza de la edad de los trabajadores, y a continuación se dividió la muestra según su tipo de trabajo. En el primer apartado se obtuvo un intervalo de confianza de cerca de 42 a 43 años, y en el segundo apartado observamos que los trabajadores de la información tienen una edad entre 43 y 44, mientras que los del sector industrial son más jóvenes, entre 41 y 42 años. A continuación comparamos los sueldos de trabajadores que tienen contratado un seguro médico, con el fin de entender si los salarios de los que sí lo tienen contratados son más altos, en concreto, hasta 20 mil dólares más alto. En éste estudio obtuvimos que efectivamente, los trabajadores que tienen un seguro contratado cobran más que los que no lo tienen, en concreto la diferencia es aproximadamente de entre 25 y 30 mil dólares. A continuación realizamos un test no-paramétrico, con el test U, dado que nuestros datos no son dependientes. En este caso volvemos a obtener que los sueldos de personas que tienen contratado un seguro médico son mayores. El apartado 4 consistía en un repaso de la regresión logística, en él observamos la relación del salario (que fuese más o menos que la media), con otras variables como la presencia de un seguro contratado, el tipo de trabajo o la edad. En primer lugar, interpretamos que las tres son influyentes en el salario de la persona, siendo el tipo de trabajo lo más concluyente. Después, añadimos al modelo el tipo de estudios de los que disponía el trabajador. Vimos que este atributo pasó a ser mucho más influyente en el salario que lo estudiado anteriormente. A continuación realizamos una predicción salarial para un ejemplo de trabajador.

A partir del apartado 5 comenzó el temario nuevo, analizamos la varianza con el método ANOVA. Para ver la relación entre el nivel de estuios y su rango salarial (más o menos de 150k), definímos un contraste de hipótesis. Como hipótesis nula establecimos que los salarios eran iguales independientemente de su nivel de estudios, y como hipótesis alternativa lo contratio. Con el modelo ANOVA obtubimos la importancia del nivel educativo, además de hacerlo manualmente. Además realizamos un test no paramétrico, para establecer relación del salario, esta vez con la raza. Otra vez, obtuvimos que la raza es determinante en el sueldo.

El punto 6 se basa en el modelo ANOVA, pero añadiéndo más factores a estudiar, por ello, estudiamos la influencia del factor "raza" con el tipo de trabajo o el nivel de estudios. La primera comparación obtuvo un resultado claro en cuanto a la mayor relevancia del tipo de trabajo frente a la raza, al igual que en el segundo caso, en el que obtenemos que el nivel educativo es más determinante.

Para realizar comparaciones múltiples y estudiar toda la casuística posible, hemos estudiado en el último punto el test de Scheffé. En él hemos visto las combinaciones más derteminantes o relevantes para ver qué combinaciones son más diferentes entre sí. Otra vez, las diferencias más grandes se dan según el tipo de trabajo, para personas de una misma raza.