

A4 - Análisis estadístico avanzado: Análisis de varianza y repaso del curso

Leonardo Segovia Vilchez

Enero 2022

Contents

Lectura del fichero y preparación de los datos	2
Preparación de los datos	3
Análisis visual	5
Estadística inferencial	10
Contrastes de hipótesis	10
Modelo de regresión lineal	13
Estimación de modelos	13
Interpretación de los modelos	14
Análisis de residuos	14
Predicción	15
Regressión logística	15
Generación de los conjuntos de entrenamiento y de test	16
Modelo predictivo	16
Interpretación	18
Matriz de confusión	18
Predicción	19
Análisis de la varianza (ANOVA) de un factor	20
Visualización	
ANOVA multifactorial	24
Estudio visual de la interacción.	24
Conclusiones	27

Introducción El conjunto de datos CensusIncomedata.txt se inspira (ha sido modificado por motivos académicos) en un elemento de la base de datos disponible en la web Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Adult>.

Este conjunto de datos contiene información de una muestra extraída a partir de un censo, en el que para cada persona, se registran los salarios aparte de información personal adicional. El conjunto de datos contiene 32.560 registros y 9 variables.

Las variables de esta muestra son:

Age: Edad del individuo. Workclass: Categorización del individuo en base al perfil laboral. *Education_num: Número de años de formación educativa del individuo.* Marital_status: Estado civil del individuo. *Occupation: Categorización del individuo en base a la tipología de trabajo.* Race: Grupo racial al que pertenece el individuo. *Sex: Género del individuo.* hours_per_week: Horas por semana trabajadas por el individuo. *income: Salario (anual) del individuo, en k€.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, preprocesamiento, análisis descriptivo e inferencia estadística, así como la regresión (lineal y logística) y el Análisis de Variancia (ANOVA). Verás que, en relación a estos datos, pondremos el foco en el estudio de la probabilidad de no alcanzar cierto umbral de retribución económica en base a las características descritas en el conjunto de datos.

Nota importante a tener en cuenta para entregar la actividad:

```
# Librerías
# Paquetes y librerías..
#install.packages("car")
#install.packages("agricolae")
library(agricolae)
library(pROC)
library(tibble)
library(dplyr)
library(ggplot2)
library(scales)
library(knitr)
library(tidyr)
library(car)
library(stringr)
library(grid)
library(gridExtra)
library(nortest)
library(BSDA)
library(RColorBrewer)
#library(Ecdat)
library(caret)
library(corrplot)
library(ResourceSelection)
```

Lectura del fichero y preparación de los datos

Leer el archivo CensusIncomedada.txt y guardar los datos en un objeto con identificador denominado adult. A continuación, verifica que los datos se han cargado correctamente.

```
# Cargamos el fichero de datos.
adult <- read.csv('CensusIncomedata.txt',stringsAsFactors = FALSE, sep=" ", strip.white=TRUE)
# Mostramos si el dataset se ha cargado correctamente.
head(adult,2)
```

```

##   age      workclass education_num marital_status occupation race sex
## 1 50      Self-Employed          13       Married White-Collar White Male
## 2 38      Private             9       Divorced Blue-Collar White Male
##   hours_per_week income
## 1                 13 53.80
## 2                 40 51.67

```

Examinamos la interpretación que hace R de cada una de las variables.

```
# Valores resumenes.
str(adult)
```

```

## 'data.frame': 32560 obs. of 9 variables:
## $ age : int 50 38 53 28 37 49 52 31 42 37 ...
## $ workclass : chr "Self-Employed" "Private" "Private" "Private" ...
## $ education_num : int 13 9 7 13 14 5 9 14 13 10 ...
## $ marital_status: chr "Married" "Divorced" "Married" "Married" ...
## $ occupation : chr "White-Collar" "Blue-Collar" "Blue-Collar" "Professional" ...
## $ race : chr "White" "White" "Black" "Black" ...
## $ sex : chr "Male" "Male" "Male" "Female" ...
## $ hours_per_week: int 13 40 40 40 40 16 45 50 40 80 ...
## $ income : num 53.8 51.7 50.1 44.2 48.9 ...

```

```
# Valores resumenes.
summary(adult)
```

```

##      age      workclass      education_num      marital_status
## Min.   :17.00  Length:32560  Min.   : 1.00  Length:32560
## 1st Qu.:28.00  Class :character 1st Qu.: 9.00  Class :character
## Median :37.00  Mode  :character Median :10.00  Mode  :character
## Mean   :38.58                   Mean   :10.08
## 3rd Qu.:48.00                   3rd Qu.:12.00
## Max.   :90.00                   Max.   :16.00
##      occupation      race      sex      hours_per_week
## Length:32560  Length:32560  Length:32560  Min.   : 1.00
## Class :character  Class :character  Class :character  1st Qu.:40.00
## Mode  :character  Mode  :character  Mode  :character  Median :40.00
##                           Mean   :40.44
##                           3rd Qu.:45.00
##                           Max.   :99.00
##      income
## Min.   :22.54
## 1st Qu.:43.22
## Median :49.71
## Mean   :48.75
## 3rd Qu.:54.32
## Max.   :68.37

```

No observamos valores nulos

Preparación de los datos

- Fíjate en los valores de las variables categóricas para identificar y proceder a quitar los molestos espacios en blanco al inicio de los valores.

```
# Eliminación de espacios en blanco.
adult$workclass <- str_trim(adult$workclass)
```

```

adult$marital_status <- str_trim(adult$marital_status)
adult$occupation <- str_trim(adult$occupation)
adult$race <- str_trim(adult$race)
adult$sex <- str_trim(adult$sex)

```

- Corrige el error en el nombre de la séptima variable, ya que realmente nos queremos referimos al rol social o percepción individual del género propia del individuo. (https://en.wikipedia.org/wiki/Sex_and_gender_distinction)

```

names(adult)[names(adult)=="sex"] <- "gender"
colnames(adult)

```

```

## [1] "age"           "workclass"       "education_num"   "marital_status"
## [5] "occupation"    "race"          "gender"         "hours_per_week"
## [9] "income"

```

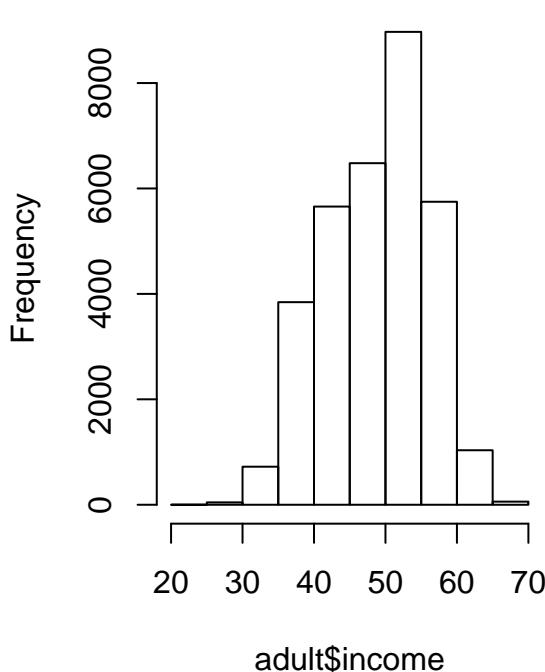
- ¿Qué podemos afirmar sobre la normalidad de la variable salario? Ayúdate de la inspección visual y el test conocido de Lilliefors.

```

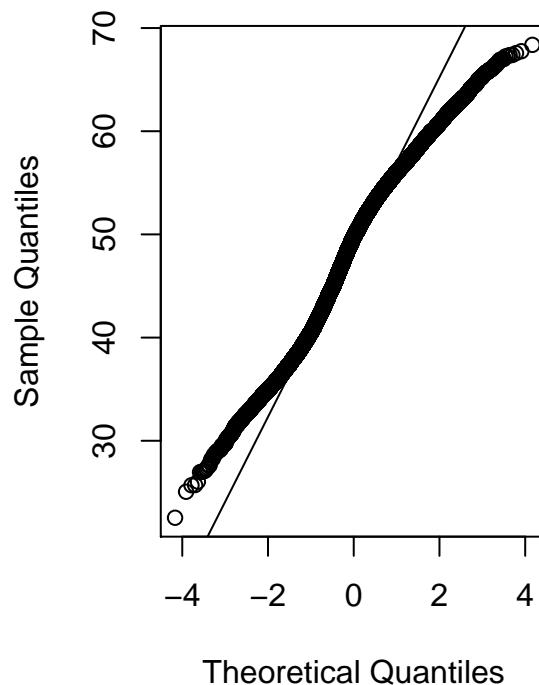
#Gráficos
par(mfrow=c(1,2))
hist(adult$income) # histograma
qqnorm(adult$income) # gráfico de cuantiles
qqline(adult$income)

```

Histogram of adult\$income



Normal Q-Q Plot



Observamos que la distribución no es completamente simétrica, teniendo esta más pendiente en la parte alta de la income.

```

#contraste de normalidad
lillie.test(adult$income) #contraste
## 

```

```

##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: adult$income
## D = 0.061672, p-value < 2.2e-16

```

No se distribuye según una distribución normal

Genera una variable denominada ‘Less50’ que clasifique binariamente los salarios dado el límite de 50 k€. Como hemos dicho antes, focalizamos sobre tener un ingreso menor en esta cantidad (‘Less50’), y por tanto, codificaremos la variable ‘Less50’ con el valor 1 cuando el salario sea inferior a 50k€, e igual a 0 en caso contrario.

```

adult$Less50[(50 > adult$income)] <- 1
adult$Less50[(50 <= adult$income)] <- 0
adult$Less50 <-as.factor(adult$Less50)
levels(adult$Less50)

## [1] "0" "1"

```

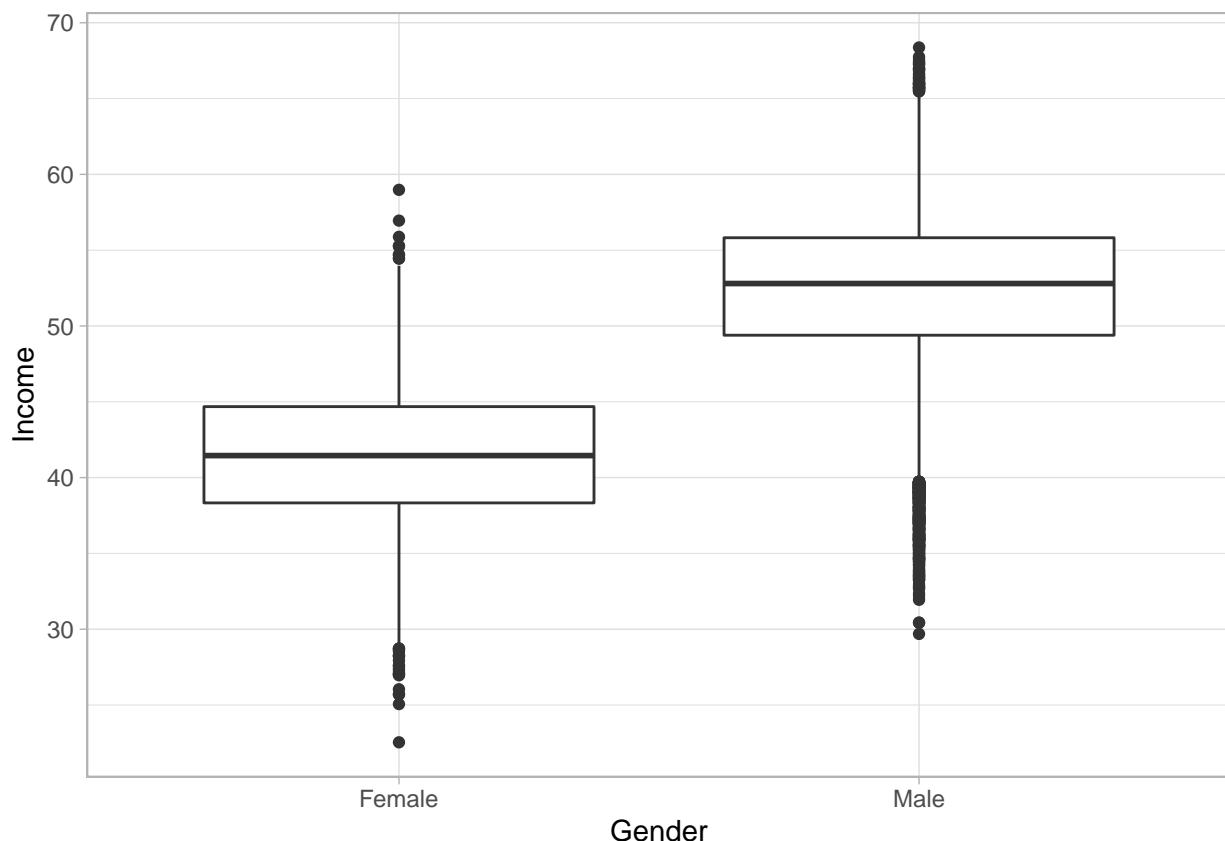
Análisis visual

1. Muestra con varios diagramas de caja la distribución de la variable income según las variables gender, race, workclass, marital_status y occupation.

```

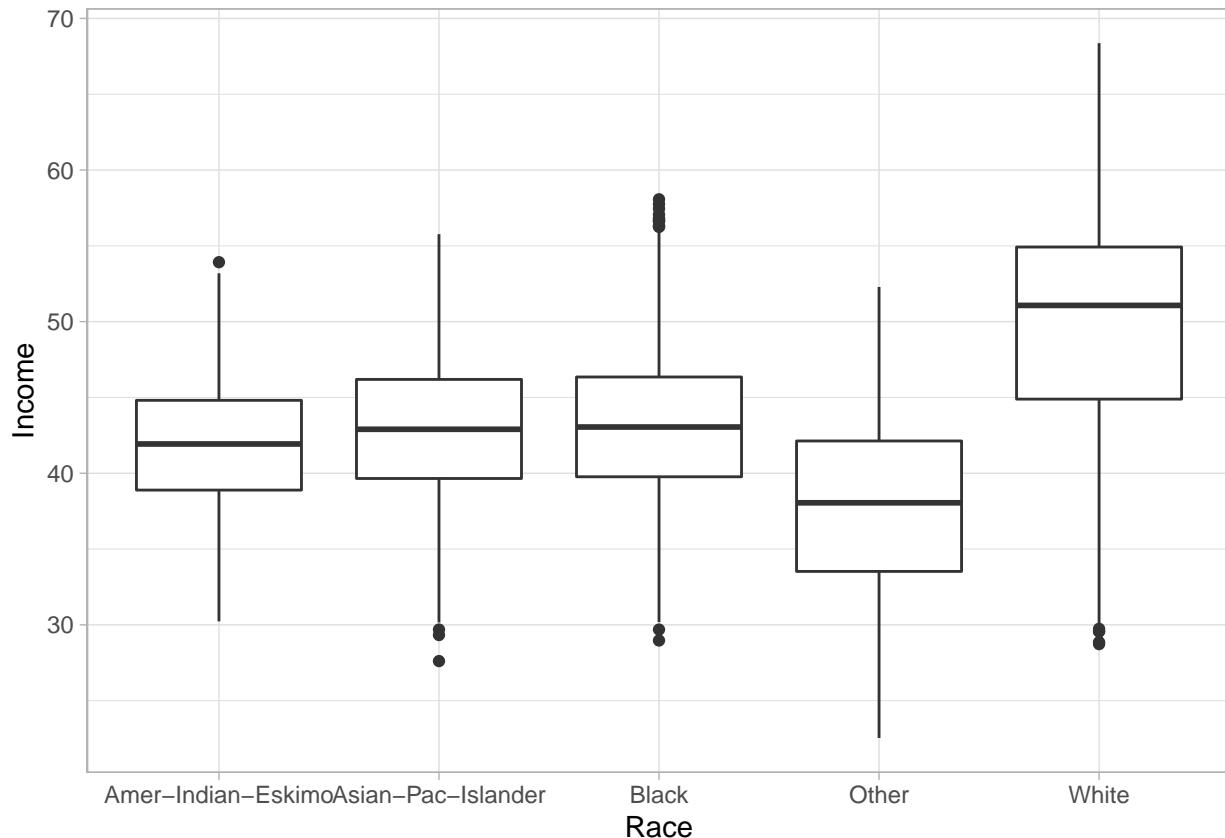
# Gender
ggplot(adult, aes(gender, income)) +
  geom_boxplot() +
  labs(x = "Gender", y = "Income") +
  theme_light()

```



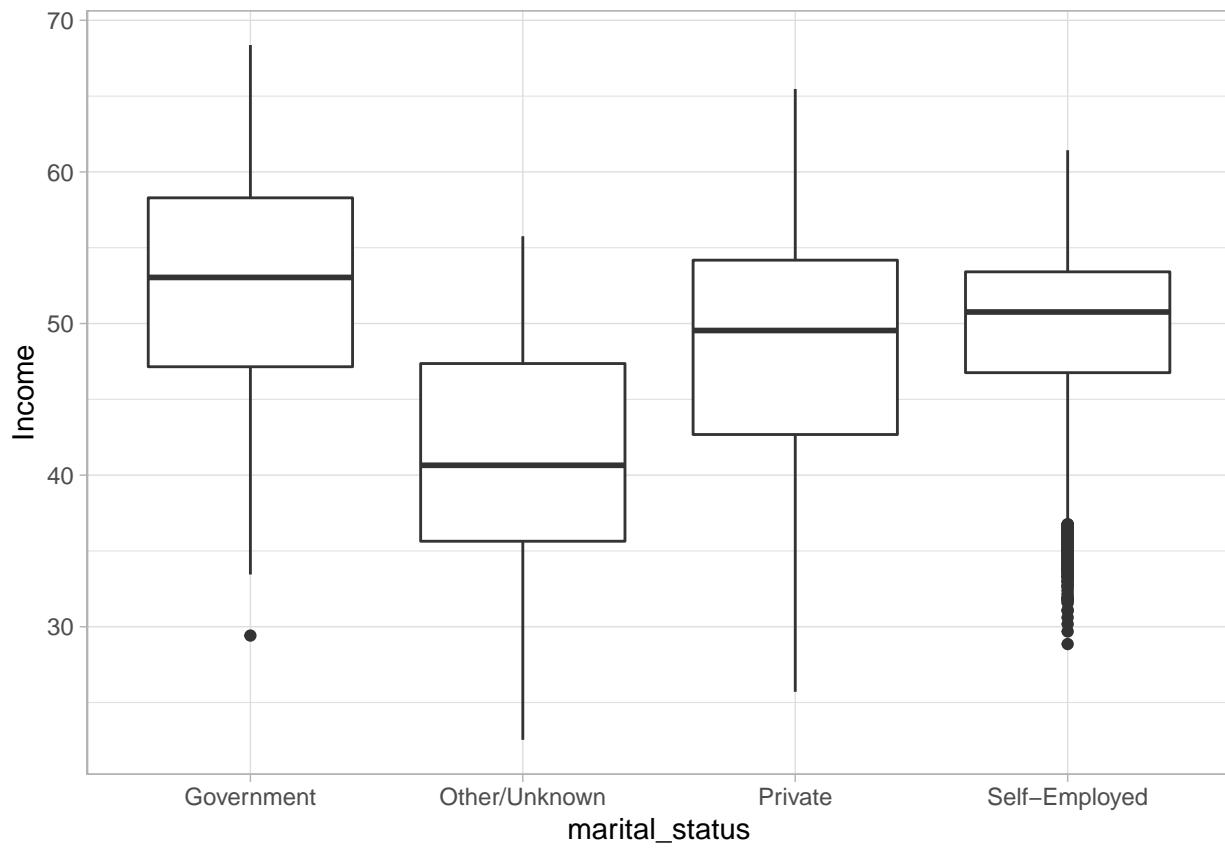
Observamos un income claramente superior para el genero masculino. Las muestras de cada uno de los grupos parecen estar bien distribuidas en su mediana.

```
# Race  
ggplot(adult, aes(race, income)) +  
  geom_boxplot() +  
  labs(x = "Race", y = "Income") +  
  theme_light()
```



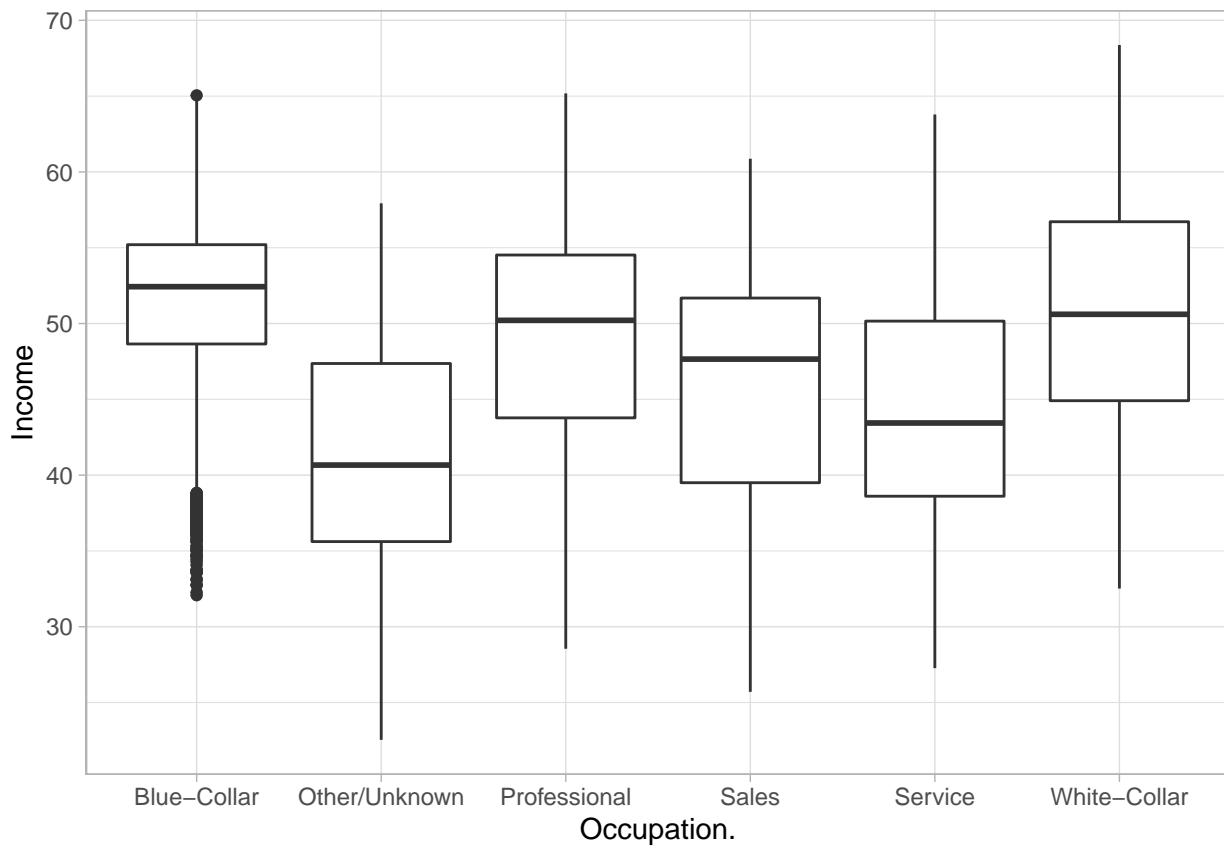
Observamos un income significativamente superior para la raza blanca con respecto a las demás.

```
# Workclass  
ggplot(adult, aes(workclass, income)) +  
  geom_boxplot() +  
  labs(x = "marital_status", y = "Income") +  
  theme_light()
```



El sector público es ligeramente superior al resto.

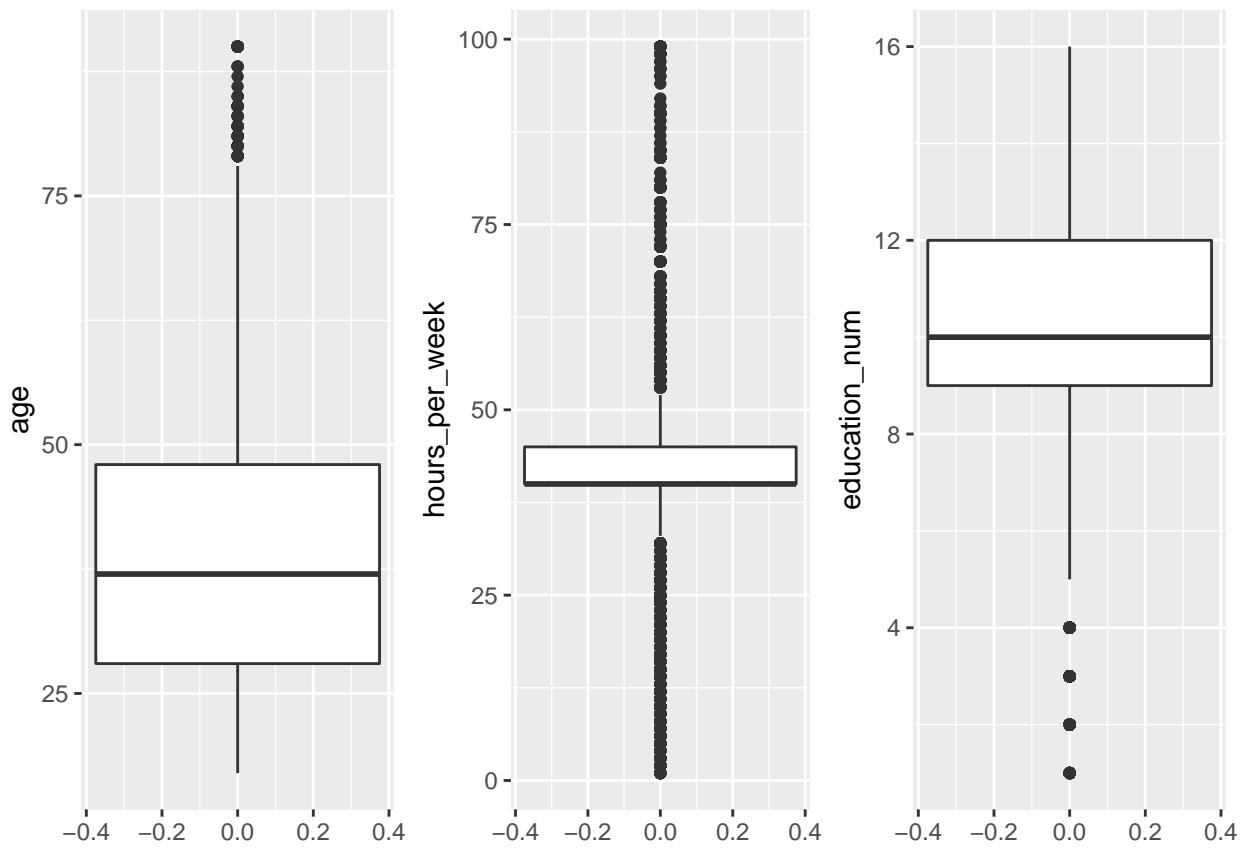
```
# Occupation.
ggplot(adult, aes(occupation, income)) +
  geom_boxplot() +
  labs(x = "Occupation.", y = "Income") +
  theme_light()
```



Observamos todas las medianas de los diferentes grupos se distribuyen aproximadamente en un rango de 10k, siendo Bluecollar el superior y Other el inferior.

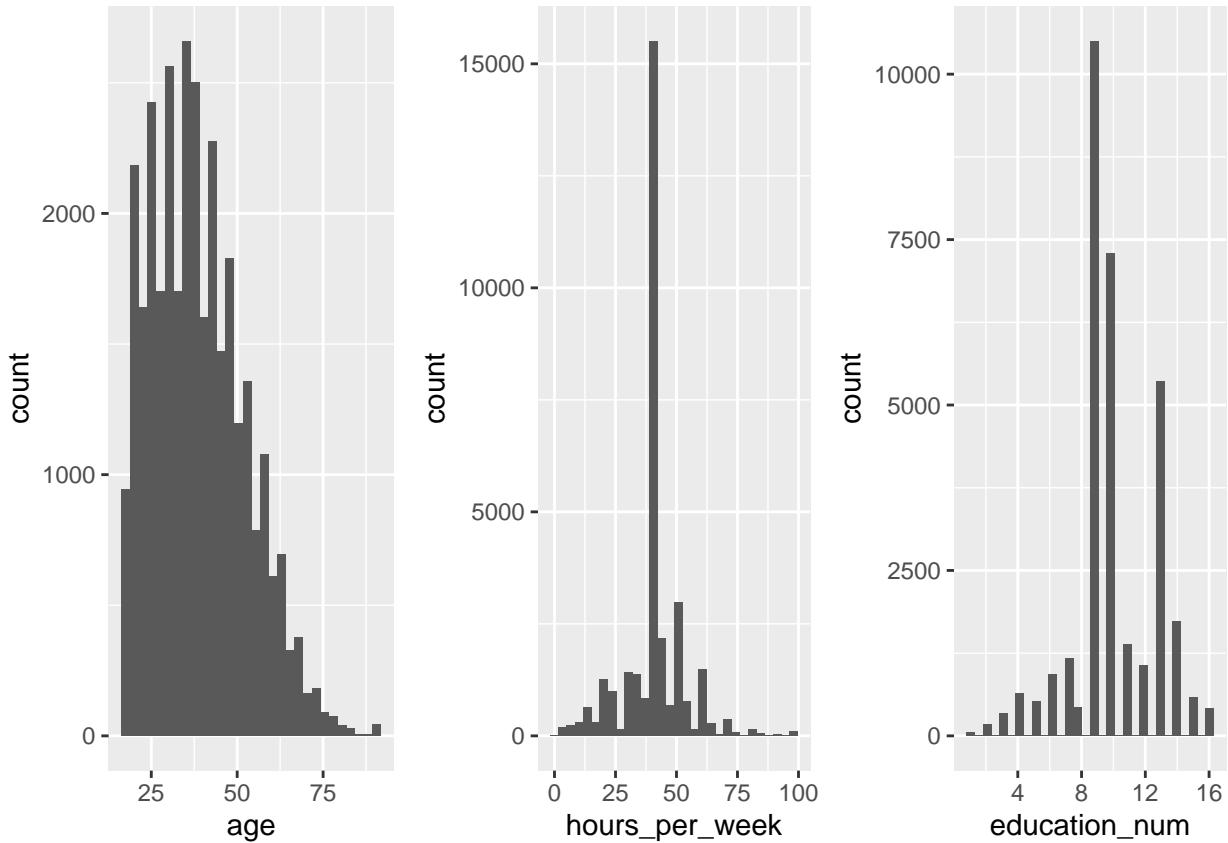
2. Interesa visualizar también las variables age, hours_per_week y education_num.

```
g1 <- ggplot(adult, aes(y=age)) + geom_boxplot()
g2 <- ggplot(adult, aes(y=hours_per_week)) + geom_boxplot()
g3 <- ggplot(adult, aes(y=education_num)) + geom_boxplot()
grid.arrange(g1,g2,g3, nrow=1)
```



Las muestras del grupo de horas trabajadas no están distribuidas en su mediana, ya que la mayoría de sus muestras están en la parte alta sobre el valor 30h. También se observa un número alto de outliers.

```
g1b<-ggplot(adult, aes(x=age)) + geom_histogram()
g2b<-ggplot(adult, aes(x=hours_per_week)) + geom_histogram()
g3b<-ggplot(adult, aes(x=education_num)) + geom_histogram()
grid.arrange(g1b,g2b,g3b, nrow=1)
```



Se observa la edad laboral sobre los 20 a 35 y un descenso paulatino en el lado derecho de la distribución.

3. Interpreta los gráficos brevemente. Aprovecha que las últimas variables son continuas para interpretar su tendencia.

Respuesta a pie de gráfico.

Estadística inferencial

Contrastes de hipótesis

Nos interesamos ahora por las potenciales diferencias en el salario de los individuos para diferentes grupos, en particular, las mujeres y los hombres, y los grupos raciales blanco y negro.

- ¿Cobran los hombres más que las mujeres? Responde a la pregunta con un nivel de confianza del 95%.
- ¿Cobra la gente blanca 6450€ más al año que la gente negra? Responde a la pregunta con un nivel de confianza del 95%.

Nota: Valora la conveniencia de crear funciones que le permitan no repetir cálculos. Sigue la siguiente estructura de apartados: ### Hipótesis nula y alternativa

RESPUESTA: por genero

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

$$H_0 : \mu_1 = \mu_2 = 0$$

$$H_1 : \mu_1 > \mu_2$$

donde μ_1 denota la media de salario semanal de los hombres y μ_2 la media de salario de las mujeres.

RESPUESTA: por raza

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

$$H_0 : \mu_1 = \mu_2 = 0$$

$$H_1 : \mu_1 > \mu_2 > 6450\text{€}$$

donde μ_1 denota la media de salario semanal de los hombres y μ_2 la media de salario de las mujeres.

Justificación del test a aplicar (por el género y por el caso racial)

Por el teorema del límite central, podemos asumir normalidad, puesto que tenemos una muestra de tamaño grande y se desea realizar un test sobre la media. Por tanto, aplicamos un test de hipótesis de dos muestras sobre la media. Aplicaremos la distribución t, dado que no se conoce la varianza de la población.

#separamos la variable salario del conjunto de datos entre hombres y mujeres.

```
Mi <- adult$income[adult$gender=="Male"]
Fi <- adult$income[adult$gender=="Female"]
var.test(Mi, Fi)
```

```
##
## F test to compare two variances
##
## data: Mi and Fi
## F = 1.331, num df = 21788, denom df = 10770, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.288172 1.375099
## sample estimates:
## ratio of variances
## 1.331048
#separamos la variable salario del conjunto de datos entre raza blanca y negra.
Wi <- adult$income[adult$race=="White"]
Bi <- adult$income[adult$race=="Black"]
var.test(Wi, Bi)
```

```
##
## F test to compare two variances
##
## data: Wi and Bi
## F = 2.1516, num df = 27814, denom df = 3123, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 2.040794 2.265905
## sample estimates:
## ratio of variances
## 2.151597
```

Aplicación, interpretación y comprobación del test (por el género y por el caso racial)

```
myttest <- function( x1, x2, d=0, CL=0.95, equalvar=TRUE, alternative="bilateral" ){
mean1<-mean(x1) - d
n1<-length(x1)
sd1<-sd(x1)
mean2<-mean(x2)
n2<-length(x2)
sd2<-sd(x2)
```

```

if (equalvar==TRUE){
  s <-sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2 )/(n1+n2-2) )
  Sb <- s*sqrt(1/n1 + 1/n2)
  df<-n1+n2-2
}
else{ #equalvar==FALSE
  Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
  denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1) )
  df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / denom
}
alfa <- (1-CL)
t<- (mean1-mean2) / Sb
if (alternative=="bilateral"){
  tcritical <- qt( alfa/2, df, lower.tail=FALSE ) #two sided
  pvalue<-pt( abs(t), df, lower.tail=FALSE )*2 #two sided
}
else if (alternative=="less"){
  tcritical <- qt( alfa, df, lower.tail=TRUE )
  pvalue<-pt( t, df, lower.tail=TRUE )
}
else{ #(alternative=="greater")
  tcritical <- qt( alfa, df, lower.tail=FALSE )
  pvalue<-pt( t, df, lower.tail=FALSE )
}

#Guardamos el resultado en un data frame
info<-data.frame(t,tcritical,pvalue,df)
return (info)
}

```

```

info<-myttest( Mi, Fi, equalvar=FALSE, alternative = "greater")
info

```

```

##          t tcritical pvalue      df
## 1 194.1064  1.644916    0 24386.74

```

Se puede afirmar que el income en los hombres es superior al de las mujeres con un 95% de nivel de confianza. El valor crítico para un nivel de confianza del 95% es 1.644916 y el valor observado es 194.1064. Por tanto, nos encontramos en la zona de rechazo de la hipótesis nula. Se concluye lo mismo con el valor p muy inferior a alfa=0.05.

```

info<-myttest( Wi, Bi, equalvar=FALSE, d=6.450, alternative = "greater")
info

```

```

##          t tcritical      pvalue      df
## 1 2.037051  1.645172 0.02084993 4783.375

```

```

# Validación
t.test(Wi, Bi, alternative="greater", mu = 6.450, var.equal=FALSE)

```

```

##
##  Welch Two Sample t-test
##
## data:  Wi and Bi
## t = 2.0371, df = 4783.4, p-value = 0.02085
## alternative hypothesis: true difference in means is greater than 6.45

```

```

## 95 percent confidence interval:
## 6.48668     Inf
## sample estimates:
## mean of x mean of y
## 49.78750 43.14683

```

Se puede afirmar que el income de los hombres blancos es 6450€ más al año que la gente negra con un 95% de nivel de confianza. El valor crítico para un nivel de confianza del 95% es 1.645172 y el valor observado es 2.037051 Por tanto, nos encontramos en la zona de rechazo de la hipótesis nula. Se concluye lo mismo con el valor p inferior a alfa=0.05.

Modelo de regresión lineal

Estimación de modelos

- Estima un modelo de regresión lineal múltiple que tenga como variables explicativas: age, education_num, hours_per_week y gender, y como variable dependiente el Income.

```

Model.1 <- lm(income~age+education_num+hours_per_week+gender, data=adult)
summary(Model.1)

```

```

##
## Call:
## lm(formula = income ~ age + education_num + hours_per_week +
##     gender, data = adult)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4319  -2.7990   0.2859   3.1140  17.0719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.315777  0.139088 225.15  <2e-16 ***
## age          0.082515  0.001858  44.41  <2e-16 ***
## education_num 0.446817  0.009913  45.07  <2e-16 ***
## hours_per_week 0.073834  0.002123  34.77  <2e-16 ***
## genderMale    10.108404  0.055211 183.09  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.548 on 32555 degrees of freedom
## Multiple R-squared:  0.5892, Adjusted R-squared:  0.5891
## F-statistic: 1.167e+04 on 4 and 32555 DF,  p-value: < 2.2e-16

```

- Genera un segundo modelo pero esta vez añadiendo la variable race.

```

Model.2 <- lm(income~age+education_num+hours_per_week+gender+race, data=adult)
summary(Model.2)

```

```

##
## Call:
## lm(formula = income ~ age + education_num + hours_per_week +
##     gender + race, data = adult)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4319  -2.7990   0.2859   3.1140  17.0719
## 
```

```

## -15.264 -2.817  0.008  2.775 16.073
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           26.183312  0.260285 100.595 <2e-16 ***
## age                  0.078247  0.001673  46.768 <2e-16 ***
## education_num        0.419036  0.008971  46.710 <2e-16 ***
## hours_per_week       0.071274  0.001911  37.301 <2e-16 ***
## genderMale            9.780936  0.049976 195.712 <2e-16 ***
## raceAsian-Pac-Islander -0.214519  0.264932 -0.810  0.418
## raceBlack              2.382315  0.243405  9.787 <2e-16 ***
## raceOther             -3.562672  0.340136 -10.474 <2e-16 ***
## raceWhite              6.681488  0.233491  28.616 <2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.092 on 32551 degrees of freedom
## Multiple R-squared:  0.6674, Adjusted R-squared:  0.6674
## F-statistic:  8166 on 8 and 32551 DF,  p-value: < 2.2e-16

```

Interpretación de los modelos

Interpreta los modelos lineales ajustados y valora la calidad del ajuste: * Valora la significación de las variables explicativas.

Primero observamos que todas las estimaciones contiene valores significations (p-value: < 2.2e-16). Esto quiere decir que la probabilidad de que el valor obtenido se deba al azar es aceptable. Las variables explicativas explican el 58.91% de la variacion del income y tienen una corelacion positiva todas ellas cercanas a cero.

- Explica la contribución de las variables explicativas en el modelo.

La variable gender en el caso de los hombre tiene una correlacion por lo que el income aumenta con cada una de las variables del modelo especialmente en el caso de los hombres, gender (male).

- ¿La inclusión de la variable race ha supuesto una mejora del segundo modelo respecto al primero?

El model.1 sin race tiene un R2 ajustado de 0.5891 y el modelo.2 con race presenta un R2 ajustado de 0.6674. Observamos una mejora significativa. Y no se observan cambios drásticos en las demás variables por lo que podemos descartar que sea una variable de confusión.

Análisis de residuos

Por último, para profundizar en la calidad del ajuste deben analizarse los residuos que nos indicarán realmente cómo se ajusta nuestro modelo a los datos muestrales. Lo haremos sólo por el segundo de los modelos lineales obtenidos.

- La salida de ‘summary()’ presenta los principales estadísticos de la distribución de los residuos. Analizalos valores estimados de los estadísticos.

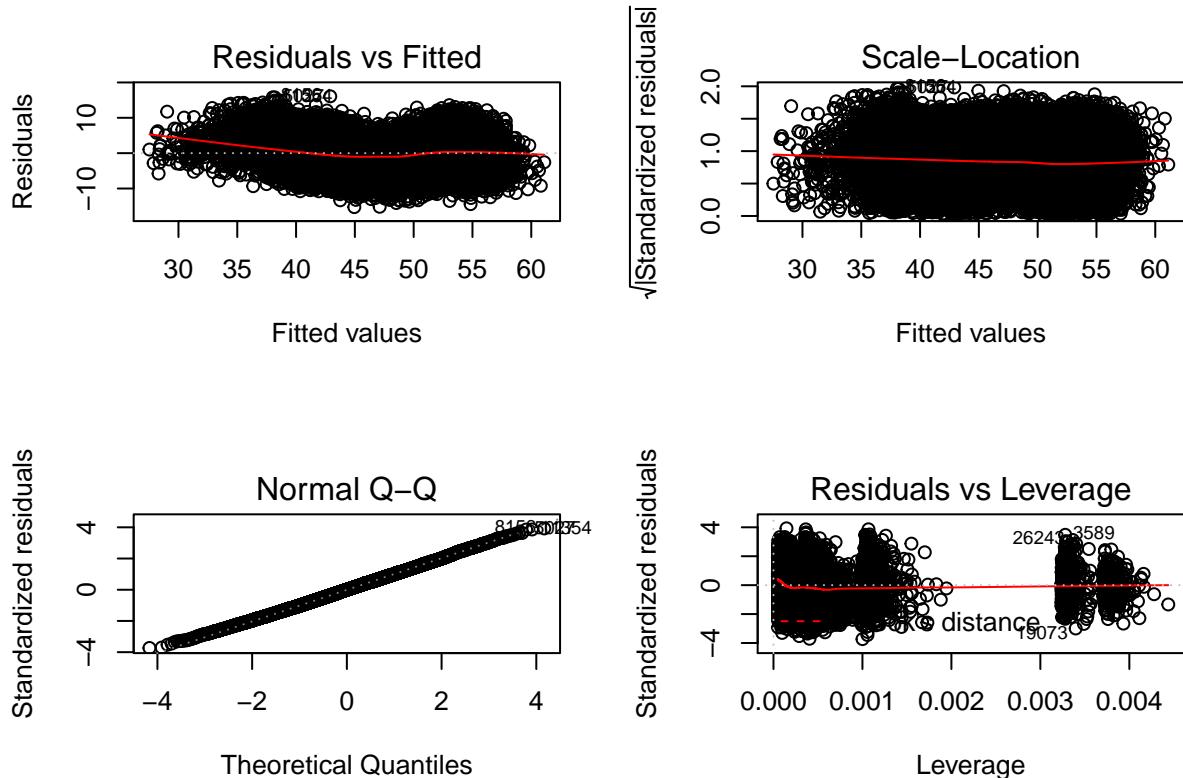
Los residuos están bastante bien distribuidos sobre la mediana (cercana a cero) ya que 1Q y 3Q tienen valores parecidos.

- Realiza ahora un análisis visual de los residuos. ¿Qué podemos decir sobre la bondad de la adecuación del modelo?

```

layout(matrix(c(1,2,3,4),2,2))
plot(Model.2)

```



El estudio de los residuos nos permite extraer información acerca del cumplimiento de las suposiciones de modelización. De la gráfica Normal Q–Q obtenemos la siguiente información; la mayoría de datos centrales siguen la distribución normal, detectando algunas observaciones extremas que se desbían. Por otro lado, el gráfico de residuos frente valores ajustados, muestra un patrón “aleatorio” de los residuos, excepto en los extremos. Por lo que aceptaremos el cumplimiento de las suposiciones.

Predicción

De nuevo, sólo por el segundo modelo estimado, realiza la predicción del income esperado para las siguientes características: age=24, education_num= “4”, hours_per_week=“40”, gender=“ Female”, race=“Black”. Proporciona, además, el intervalo de confianza del 95%.

```
pre_data_d<-data.frame(age=c(24), education_num= c(4), hours_per_week=c(40), gender='Female', race='Blac
predict(Model.2,pre_data_d ,type="response", interval = "confidence")
```

```
##          fit      lwr      upr
## 1 34.97064 34.785 35.15627
```

El income estimado es de 34.97064

Regressión logística

Utilizando las variables explicativas posibles, ajusta un modelo predictivo basado en la regresión logística para predecir la probabilidad de tener un salario menor de 50 k€. Por eso, usaremos la variable dicotómica Less50 que ha creado en el primer apartado, que será nuestra variable dependiente del modelo. Para poder estimar de forma más objetiva la precisión del modelo, separaremos el conjunto de datos en dos partes: el conjunto de entrenamiento (training) y el conjunto de prueba (test). Ajustaremos el modelo de regresión logística con el conjunto de entrenamiento, y evaluaremos la precisión con el conjunto de prueba. Siga los pasos que se especifican a continuación.

- * Generar los conjuntos de train y test
- * Entrena el modelo
- * Interprete el modelo entrenado
- * Evalúe la calidad del modelo sobre los datos de test
- * Predicción

Generación de los conjuntos de entrenamiento y de test

Genere los conjuntos de datos para entrenar el modelo y para testarlo. Puedes fijar el tamaño de la muestra de entrenamiento a un 80% del original.

```
# 50% de los datos.
smp_size <- floor(0.5 * nrow(adult))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(adult)), size = smp_size)

train <- adult[train_ind, ]
test <- adult[-train_ind, ]
```

Modelo predictivo

Entrene el modelo con el conjunto que acaba de generar. Utilice, como valores de referencia, el valor mayoritario de cada variable. Por ejemplo, para race, utilizaremos White.

```
train$workclass <-as.factor(train$workclass)
train$workclass<-relevel(train$workclass, ref="Private")

train$race <-as.factor(train$race)
train$race<-relevel(train$race, ref="White")

train$marital_status <-as.factor(train$marital_status)
train$marital_status<-relevel(train$marital_status, ref="Married")

train$occupation <-as.factor(train$occupation)
train$occupation<-relevel(train$occupation, ref="Blue-Collar")

train$gender <-as.factor(train$gender)
train$gender<-relevel(train$gender, ref="Male")

train$Less50 <-as.factor(train$Less50)
train$Less50 <-relevel(train$Less50, ref="1")

modlog<-glm(Less50~age+workclass+education_num+marital_status+occupation+race+gender+hours_per_week, family = binomial())
summary(modlog)

## Call:
## glm(formula = Less50 ~ age + workclass + education_num + marital_status +
##       occupation + race + gender + hours_per_week, family = binomial(),
##       data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.1572 -0.0860 -0.0003  0.1908  4.3933
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.200061   0.224778  -0.890  0.3734
## age          0.025273   0.003175   7.961 1.71e-15 ***
## workclassGovernment 3.520080   0.148165  23.758 < 2e-16 ***
```

```

## workclassOther/Unknown -3.836568 1.604496 -2.391 0.0168 *
## workclassSelf-Employed -2.507279 0.108870 -23.030 < 2e-16 ***
## education_num 0.293929 0.015627 18.809 < 2e-16 ***
## marital_statusDivorced -2.665165 0.115243 -23.127 < 2e-16 ***
## marital_statusSeparated -3.674711 0.218976 -16.781 < 2e-16 ***
## marital_statusSingle -3.523750 0.104565 -33.699 < 2e-16 ***
## marital_statusWidowed -2.726532 0.289267 -9.426 < 2e-16 ***
## occupationOther/Unknown -1.671363 1.596285 -1.047 0.2951
## occupationProfessional -1.852249 0.137449 -13.476 < 2e-16 ***
## occupationSales -2.560997 0.114823 -22.304 < 2e-16 ***
## occupationService -2.671119 0.116618 -22.905 < 2e-16 ***
## occupationWhite-Collar 1.707516 0.136397 12.519 < 2e-16 ***
## raceAmer-Indian-Eskimo -7.829967 0.580403 -13.491 < 2e-16 ***
## raceAsian-Pac-Islander -7.985330 0.295195 -27.051 < 2e-16 ***
## raceBlack -6.273345 0.186620 -33.616 < 2e-16 ***
## raceOther -9.907970 1.183013 -8.375 < 2e-16 ***
## genderFemale -8.281638 0.193372 -42.828 < 2e-16 ***
## hours_per_week 0.025427 0.002843 8.944 < 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: 22557.8 on 16279 degrees of freedom
## Residual deviance: 5851.7 on 16259 degrees of freedom
## AIC: 5893.7
##
## Number of Fisher Scoring iterations: 8
info<-summary(modlog)

exp(cbind(coef(modlog),confint(modlog)))

##                                     2.5 %      97.5 %
## (Intercept) 8.186810e-01 5.267915e-01 1.271652e+00
## age          1.025595e+00 1.019257e+00 1.032022e+00
## workclassGovernment 3.378715e+01 2.537439e+01 4.536572e+01
## workclassOther/Unknown 2.156748e-02 9.832134e-04 3.070084e-01
## workclassSelf-Employed 8.148966e-02 6.572903e-02 1.007271e-01
## education_num 1.341689e+00 1.301447e+00 1.383678e+00
## marital_statusDivorced 6.958784e-02 5.544408e-02 8.711541e-02
## marital_statusSeparated 2.535674e-02 1.647444e-02 3.889345e-02
## marital_statusSingle 2.948865e-02 2.396626e-02 3.611209e-02
## marital_statusWidowed 6.544583e-02 3.682464e-02 1.144864e-01
## occupationOther/Unknown 1.879906e-01 1.339326e-02 4.031518e+00
## occupationProfessional 1.568839e-01 1.197718e-01 2.053097e-01
## occupationSales 7.722773e-02 6.156624e-02 9.657343e-02
## occupationService 6.917476e-02 5.494065e-02 8.678928e-02
## occupationWhite-Collar 5.515245e+00 4.239395e+00 7.238136e+00
## raceAmer-Indian-Eskimo 3.976387e-04 1.175085e-04 1.152389e-03
## raceAsian-Pac-Islander 3.404202e-04 1.881098e-04 5.990800e-04
## raceBlack 1.885910e-03 1.299255e-03 2.700883e-03
## raceOther 4.977640e-05 2.334378e-06 3.437088e-04
## genderFemale 2.531223e-04 1.719527e-04 3.670231e-04
## hours_per_week 1.025753e+00 1.020074e+00 1.031508e+00

```

Interpretación

Interpreta el modelo ajustado. Concretamente, explica la contribución de las variables explicativas con coeficiente estadísticamente significativo para predecir el salario de los individuos. Se observa que las variables explicativas son significativas menos la dummy variable occupationOther/Unknown.

- La probabilidad que el income sea inferior a los 50k aumenta con la edad.
- La probabilidad que el income sea inferior a los 50k aumenta con los años de formación educativa.
- La probabilidad que el income sea inferior a los 50k aumenta muy significativamente para la raza negra en comparación con la raza blanca.
- La probabilidad que el income sea inferior a los 50k aumenta muy significativamente para el género femenino en comparación con el masculino.
- La probabilidad que el income sea inferior a los 50k aumenta para trabajos públicos en comparación con los trabajos del sector privado.
- La probabilidad que el income sea superior a los 50k aumenta para white-collar en comparación con blue-collar.
- La probabilidad que el income sea inferior a los 50k aumenta para separados en comparación con los matrimonios.

Matriz de confusión

A continuación analiza la precisión del modelo, comparando la predicción del modelo contra el conjunto de prueba (testing_set). Asumiremos que la predicción del modelo es 1 (salario por debajo de 50k€) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 de lo contrario. Analice la matriz de confusión y las medidas de sensibilidad y especificidad. Nota: Toma como categoría de interés que el salario esté por debajo de 50k€. Por tanto, Less50 igual a 1 será el caso positivo en la matriz de confusión y 0 el caso negativo.

```
response<- ifelse( predict(modlog, type="response") >= 0.5, "1", "0")
actual <- ifelse( train$Less50==1, "1", "0")
confusionMatrix( table(response, actual ), positive="1" )

## Confusion Matrix and Statistics
##
##           actual
## response      0      1
##           0 508 7665
##           1 7420  687
##
##           Accuracy : 0.0734
##             95% CI : (0.0694, 0.0775)
##   No Information Rate : 0.513
##   P-Value [Acc > NIR] : 1.00000
##
##           Kappa : -0.853
##
##   Mcnemar's Test P-Value : 0.04696
##
##           Sensitivity : 0.08226
##           Specificity  : 0.06408
##   Pos Pred Value  : 0.08474
##   Neg Pred Value  : 0.06216
##           Prevalence  : 0.51302
```

```

##          Detection Rate : 0.04220
##    Detection Prevalence : 0.49797
##    Balanced Accuracy : 0.07317
##
##    'Positive' Class : 1
##

```

Modelo ajustado con baja especificidad y baja sensibilidad. Es decir, el modelo predice correctamente cuando se supera los 50k del income y no predice adecuadamente cuando un es inferior a los 50k. El modelo clasifica correctamente más del 8.226% de los positivos y clasifica correctamente el 6.408% de los negativos. Una posibilidad para mejorar la sensibilidad del test es cambiar el punto de corte de 0.5 en la identificación de positivos. Pero esta estrategia conducirá a aumentar el número de falsos positivos y por tanto, bajará la especificidad del test.

Predicción

Utiliza el modelo anterior para realizar predicciones. Haga el cálculo de la predicción manualmente, y use la función predict para validar. * ¿Con qué probabilidad el salario de un individuo será menor a 50k€ para un hombre blanco de 20 años de edad, autónomo (self-employed), con 3 años de estudios, soltero, trabajando en el sector profesional, y trabajando actualmente unas 25 horas semanales?

```

#manualmente
# age + workclass + education_num + marital_status + occupation + race + gender + hours_per_week,
expresion <-  info$coefficients[1,1] +
               info$coefficients[2,1]*20+      #age
               info$coefficients[5,1]+       #workclass
               info$coefficients[6,1]*3+    #education_num
               info$coefficients[9,1]+     #marital_status
               info$coefficients[12,1]+    #occupation
               info$coefficients[21,1]*25   #hours_per_week

p1.num<-exp( expresion )
p1.den<-1+exp(expresion )
p1<-p1.num/p1.den; p1

## [1] 0.002328004

new <- data.frame(race='White', age=20, workclass="Self-Employed",education_num=3, occupation='Professional')
p1<-predict(modlog,new,type="response")
p1

##           1
## 0.002328004

```

La probabilidad de que el income sea inferior es de 0.002328004

- ¿Con qué probabilidad el salario de un individuo será menor a 50k€ para un hombre negro de 60 años de edad, con trabajo gubernamental, con 15 años de estudios, casado, trabajando como ‘white-collar’ y trabajando actualmente unas 35 horas semanales?

```

#manualmente
# age + workclass + education_num + marital_status + occupation + race + gender + hours_per_week,
expresion <-  info$coefficients[1,1] +
               info$coefficients[2,1]*60+      #age
               info$coefficients[3,1]+       #workclass
               info$coefficients[18,1]+     #race
               info$coefficients[6,1]*15+    #education_num
               info$coefficients[15,1]+     #occupation

```

```

info$coefficients[21,1]*35      #hours_per_week

p1.num<-exp( expresion )
p1.den<-1+exp(expresion )
p1<-p1.num/p1.den; p1

## [1] 0.9962019

new<-data.frame(race='Black', age=60, workclass="Government", education_num=15, occupation='White-Collar'
p2<-predict(modlog,new,type="response")
p2

##          1
## 0.9962019

```

La probabilidad de que el income sea inferior es de 0.9962019

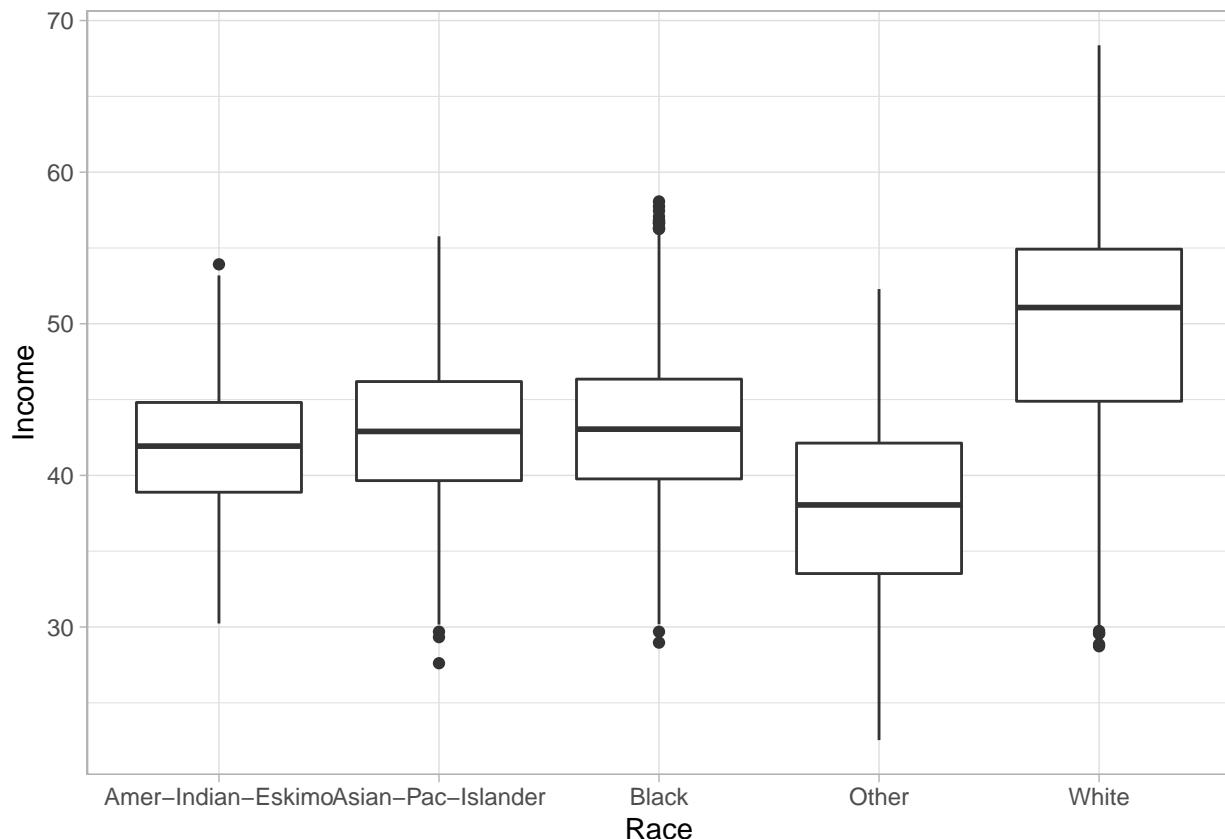
Análisis de la varianza (ANOVA) de un factor Visualización

En este apartado, nos centraremos en analizar la existencia de diferencias significativas de income entre los diferentes grupos raciales. Tomaremos siempre un nivel de significación del 5%. – Haga un análisis visual de esta dependencia.

```

# Race
ggplot(adult, aes(race, income)) +
  geom_boxplot() +
  labs(x = "Race", y = "Income") +
  theme_light()

```



Modelo ANOVA Completa los siguientes apartados: ### Formula el modelo Explica el modelo que se plantea en el ANOVA.

Queremos comprobar si los distintos grupos de race y ocupacion afectan al income total anual de un trabajador.

Indica las hipótesis nula y alternativa

Escrivid las hipótesis nula y alternativa.

Las hipótesis son:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \mu_i \neq \mu_j$$

para algún i, j donde $\mu_1, \mu_2, \mu_3, \mu_4$ y μ_5 denotan, respectivamente, la media poblacional del Income para los diferentes grupos raciales.

Estima la signifcación del factor grupo racial

Calculad la variabilidad explicada per la variable race sobre la variable income mediante la función anova().

```
#Usando aov
adult$race <-as.factor(adult$race)
adult$race<-relevel(adult$race, ref="White")
levels(adult$race)

## [1] "White"           "Amer-Indian-Eskimo" "Asian-Pac-Islander"
## [4] "Black"            "Other"

# Usando lm
modlm<-lm(income~race,data=adult)
anova(modlm)

## Analysis of Variance Table
##
## Response: income
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## race          4  211909   52977  1208.4 < 2.2e-16 ***
## Residuals  32555 1427206      44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estima los efectos de los niveles de factor

Interpretad los resultados del modelo generado en el apartado anterior.

Valores del contraste: Sum Sq = 211909 ; Mean Sq = 52977; estadístico F = 1208.4 pvalor = 0. 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.

###Realiza los contrastes dos-a-dos Para los contrastes dos-a-dos, podeis usar, por ejemplo, la función HSD.test() del paquete agricolae.

```
modelo <- aov(income~race, data = adult)
HSD.test(modelo, "race", console = TRUE)

##
## Study: modelo ~ "race"
##
## HSD Test for income
```

```

## 
## Mean Square Error: 43.83985
## 
## race, means
## 
##           income      std      r   Min   Max
## Amer-Indian-Eskimo 41.88682 4.498709  311 30.23 53.92
## Asian-Pac-Islander 42.89898 4.805233 1039 27.61 55.77
## Black            43.14683 4.694985 3124 28.97 58.07
## Other             37.60339 5.766821  271 22.54 52.29
## White            49.78750 6.886755 27815 28.73 68.37
## 
## Alpha: 0.05 ; DF Error: 32555
## Critical Value of Studentized Range: 3.857656
## 
## Groups according to probability of means differences and alpha level( 0.05 )
## 
## Treatments with the same letter are not significantly different.
## 
##           income groups
## White          49.78750     a
## Black          43.14683     b
## Asian-Pac-Islander 42.89898   bc
## Amer-Indian-Eskimo 41.88682     c
## Other          37.60339     d

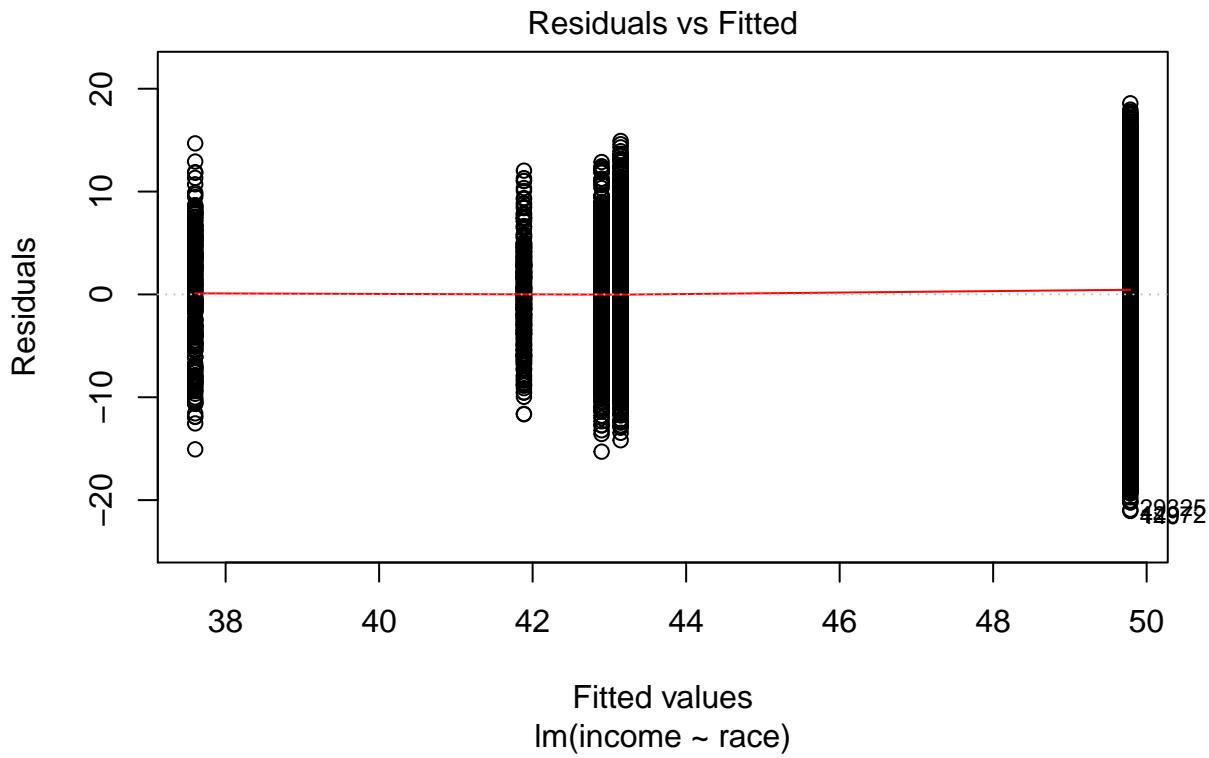
```

Detectamos un grupo homogéneo, formado por los niveles de raza blanca (grupo A), un segundo grupo formado por la raza negra (grupo B), un tercero por la raza Amer-Indian-Eskimo (grupo C) y un ultimo por otras (grupo D). Una de las razas Asian-Pac-Islander podría pertenecer a dos grupos: B y C

Adecuación del modelo

Mostrad la adecuación del modelo ANOVA en los dos siguientes sub-apartados. ##### Homocedasticidad de los residuos El gráfico “Residuals vs Fitted” proporciona información sobre la homocedasticidad de los residuos. Mostrad e interpretad este gráfico.

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



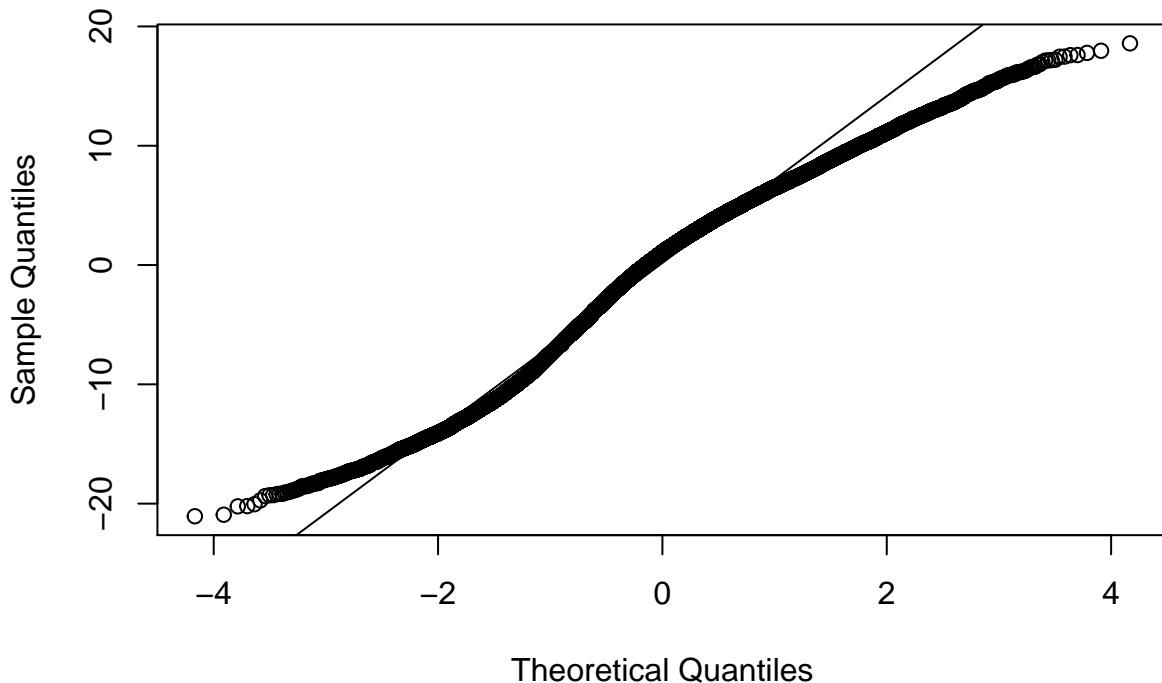
Observamos cinco tiras verticales de puntos que están situadas en las medias de cada grupo. Estas corresponden a los valores ajustados de las observaciones. La disposición de los residuos muestra una dispersión parecida en cada tira. Por lo tanto, no se aprecia efecto de embudo.

Normalidad de los residuos

Se puede comprobar el supuesto de normalidad de los residuos con los gráficos usuales. Aplicad también el test de Kruskal-Wallis e interpretad los resultados.

```
qqnorm(residuals(modlm))
qqline(residuals(modlm))
```

Normal Q-Q Plot



Observamos que la mayoría de los residuos se ajustan a la recta, por lo que no hay evidencia en contra del supuesto de normalidad.

```
kruskal.test(income~race,data=adult)
```

```
##  
## Kruskal-Wallis rank sum test  
##  
## data: income by race  
## Kruskal-Wallis chi-squared = 4210.3, df = 4, p-value < 2.2e-16
```

p-valor de 0.005, inferior al nivel de significación. Por lo tanto, aceptamos que existen diferencias significativas en el income según el grupo racial.

ANOVA multifactorial

La modelización con ANOVA facilita la inclusión de múltiples factores. Estamos interesados en incluir el factor occupation para saber si existen diferencias en los ingresos entre los empleos, a la vez que estimar la existencia de interacción significativa entre ambos factores: grupo racial y empleo.

Estudio visual de la interacción.

- Calcula la tabla cruzada entre razas y empleos para saber cuántas observaciones hay por condición. ¿Se trata de un escenario balanceado? Valora los posibles inconvenientes de la modelización basada en anova en caso de un escenario no balanceado.

```
# Agrupamos los dos factores a estudiar  
adult %>% group_by(race, occupation) -> DS2  
  
# Aplicamos las distintas funciones.  
DS3 <- summarise( DS2, m=mean(income), sd=sd(income), n=length(income))
```

DS3

```
## # A tibble: 30 x 5
## # Groups:   race [5]
##   race       occupation     m     sd     n
##   <fct>     <chr>      <dbl>  <dbl>  <int>
## 1 White     Blue-Collar  52.5  4.63  8714
## 2 White     Other/Unknown 42.5  6.50  1522
## 3 White     Professional 50.0  6.55  3651
## 4 White     Sales        46.8  6.47  3237
## 5 White     Service      45.6  7.07  3962
## 6 White     White-Collar 51.7  6.81  6729
## 7 Amer-Indian-Eskimo Blue-Collar  42.7  3.21  120
## 8 Amer-Indian-Eskimo Other/Unknown 36.2  3.79  26
## 9 Amer-Indian-Eskimo Professional 42.9  4.38  33
## 10 Amer-Indian-Eskimo Sales      38.1  3.02  26
## # ... with 20 more rows
```

Sistema no balanceado, ya que n es diferente en los distintos grupos

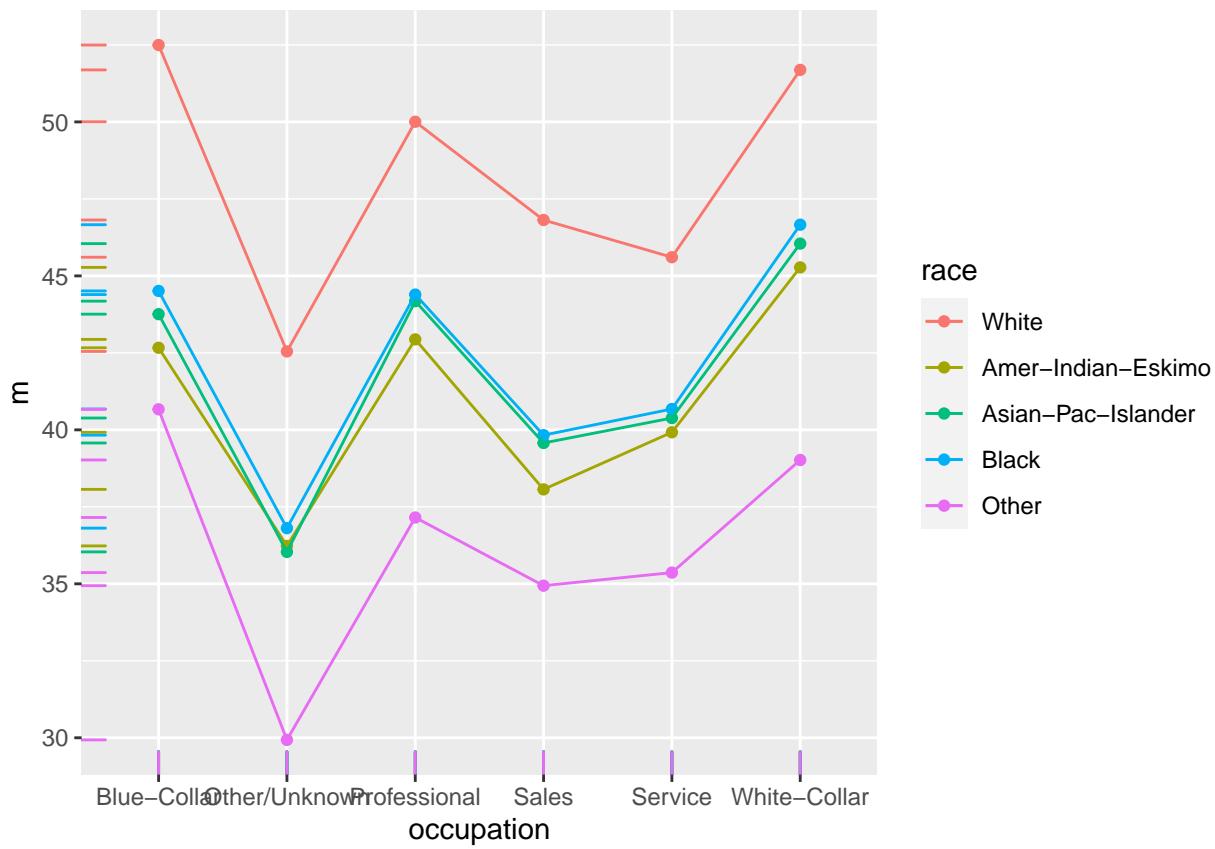
```
modlm<-lm(income~race*occupation,data=adult)
anova(modlm)
```

```
## Analysis of Variance Table
##
## Response: income
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## race                  4  211909  52977 1538.650 < 2.2e-16 ***
## occupation            5  300041  60008 1742.854 < 2.2e-16 ***
## race:occupation       20    7125    356   10.347 < 2.2e-16 ***
## Residuals             32530 1120040          34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

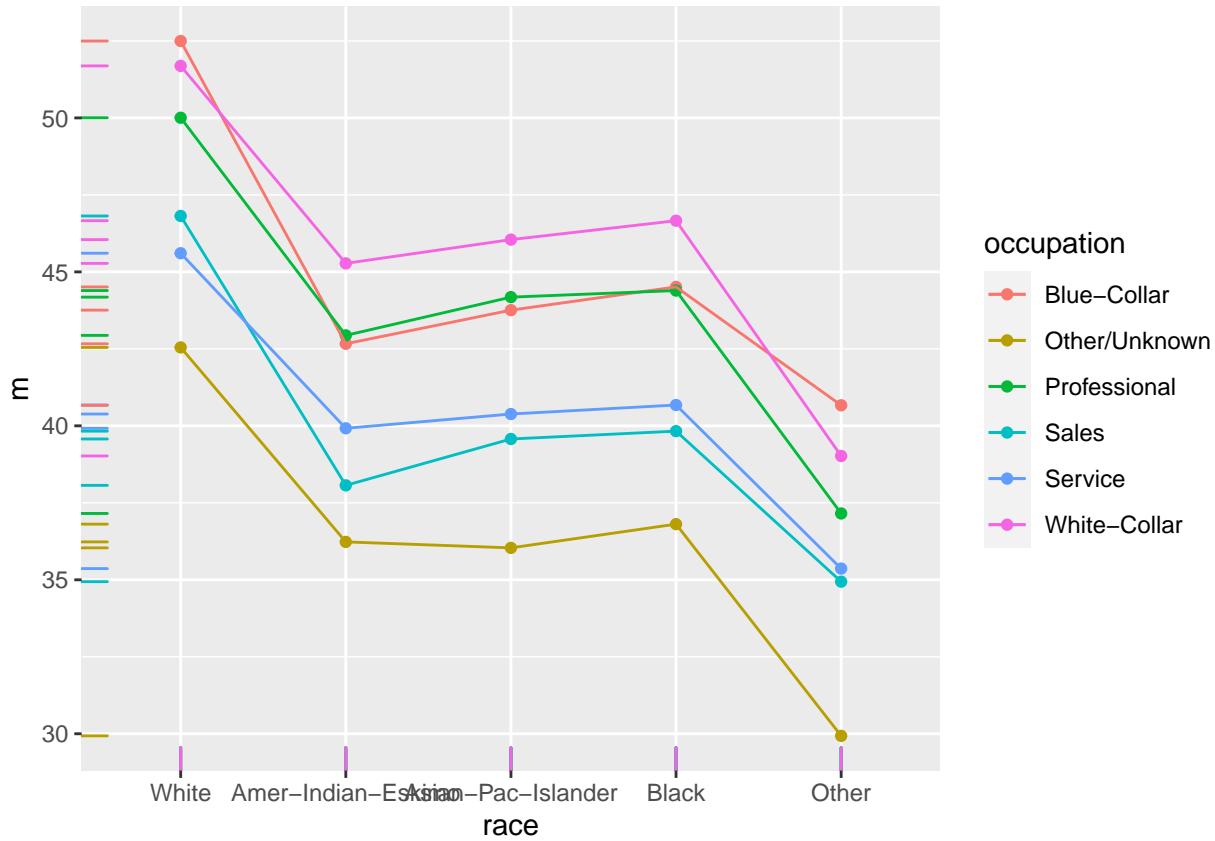
Tanto los factores principales como la interacción entre factores son significativos. Por tanto, el $income$ en función de la raza (race), es diferente según el tipo de trabajo que realiza.

- Representa la interacción entre ambos factores y comenta los gráficos resultantes.

```
ggplot(DS3, aes(x=occupation, y=m, group=race, color=race)) +
  geom_point() + geom_line() + geom_rug()
```



```
ggplot(DS3, aes(x=race, y=m, group=occupation, color=occupation)) +
  geom_point() + geom_line() + geom_rug()
```



Observamos las siguientes interacciones:

- Entre White/Sales y Service/Amer-Indian
- Entre White/Blue-collar y Profesional/Amer-Indian
- Entre Black/White-collar y Other/Blue-Collar

Por lo que el income medio más alto cambia dependiendo de la combinación de estos dos grupos.

Conclusiones

Resuma las principales conclusiones del análisis. Para ello, puede resumir las conclusiones de cada uno de los apartados.

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

- El income en los hombres es superior al de las mujeres.
- El income de los hombres blancos es como mínimo superior a 6450€ más al año que la gente de raza negra.
- El income esperado para las siguientes características: $age=24$, $education_num=4$, $hours_per_week=40$, $gender="Female"$, $race="Black"$. Proporciona, además, el intervalo de confianza del 95% es de 34.97064k .
- La probabilidad que el income sea inferior a los 50k aumenta con la edad.
- La probabilidad que el income sea inferior a los 50k aumenta con los años de formación educativa.
- La probabilidad que el income sea inferior a los 50k aumenta muy significativamente para la raza negra en comparación con la raza blanca.
- La probabilidad que el income sea inferior a los 50k aumenta muy significativamente para el género femenino en comparación con el masculino.

- La probabilidad que el income sea inferior a los 50k aumenta para trabajos publicos en comparación con los trabajos del sector privado.
 - La probabilidad que el income sea superior a los 50k aumenta para white-collar en comparación con blue-collar.
 - La probabilidad que el income sea inferior a los 50k aumenta para separados en comparación con los matrimonios.
 - La probabilidad de que el salario de un individuo sea menor a 50k€ para un hombre blanco de 20 años de edad, autónomo (self-employed), con 3 años de estudios, soltero, trabajando en el sector profesional, y trabajando actualmente unas 25 horas semanales es de 0.002328004
 - El income cambia en función de la raza (race) y del tipo de sector.
-
-