

A4 - Análisis de varianza y repaso del curso

Solución

Semestre 2020.2

Índex

1 Lectura del fichero y preparación de los datos	3
1.1 Preparación de los datos	4
1.2 Clasificación de tiempo	5
1.3 Valores ausentes	6
1.4 Salud mental	7
1.5 Análisis visual	8
1.6 Comprobación de normalidad	11
2 Estadística inferencial	13
2.1 Intervalo de confianza de la media poblacional de la variable CosteFinal	13
2.2 Contraste de hipótesis para la diferencia de medias	14
3 Modelo de regresión lineal	17
3.1 Interpretación del modelo	17
3.2 Análisis residuos	18
3.3 Predicción	20
4 Regresión logística	21
4.1 Modelo predictivo	21
4.2 Interpretación	22
4.3 Matriz de confusión	22
4.4 Predicción	24
5 Análisis de la varianza (ANOVA) de un factor	24
5.1 Hipótesis nula y alternativa	25
5.2 Modelo	25
5.3 Efectos de los niveles del factor	26
5.4 Contraste dos-a-dos	26
5.5 Adecuación del modelo	27
6 ANOVA multifactorial	30
6.1 Análisis de los efectos principales y posibles interacciones	30
6.2 Cálculo del modelo	31
6.3 Interpretación de los resultados	33
7 Conclusiones	33

Introducción

El conjunto de datos trainCLEAN.csv se inspira (ha sido modificado por motivos académicos) en la base de datos disponible en la plataforma Kaggle: <https://www.kaggle.com/c/actuarial-loss-estimation>.

Este conjunto de datos contiene información de una muestra de indemnizaciones otorgadas por una compañía de seguros por el tiempo que ha estado de baja laboral el trabajador. El conjunto de datos contiene 54,000 registros y 15 variables.

Las principales variables que se usarán en esta actividad son:

- ClaimNumber: Identificador de la póliza.
- DateTimeOfAccident: Fecha del accidente.
- DateReported: Fecha que se comunica a la compañía y ésta abre un expediente del siniestro.
- Age: Edad del trabajador.
- Gender: Sexo.
- MaritalStatus: Estado civil, (M)arried, (S)ingle, (U)nknown.
- DependentChildren: Número de hijos dependientes.
- DependentsOther: Número de dependientes excluyendo hijos
- WeeklyWages: Salario semanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWorkedPerWeek: Número horas por semana.
- DaysWorkedPerWeek: Número de días por semana.
- ClaimDescription: Descripción siniestros.
- InitialIncurredClaimCost: Estimación inicial del coste realizado por la compañía.
- UltimateIncurredClaimCost: Coste total pagado por siniestro.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, preprocesado, análisis descriptivo e inferencia estadística.

Nota importante a tener en cuenta para entregar la actividad:

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

1 Lectura del fichero y preparación de los datos

Leed el fichero `trainCLEAN.csv` y guardad los datos en un objeto con identificador denominado `claim`. A continuación, verificad que los datos se han cargado correctamente.

Solución:

```
# Cargamos el conjunto de datos
claim<-read.csv("trainCLEAN.csv", sep=",", stringsAsFactors=TRUE)

# Obtenemos las dimensiones del conjunto de datos
dim(claim)

## [1] 54000    15

# Verificamos que no ha habido errores de lectura
problems(claim)

## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)

# Listado de variables con sus tipos de datos
str(claim)

## 'data.frame': 54000 obs. of 15 variables:
##   $ ClaimNumber          : Factor w/ 54000 levels "WC1592889","WC1592905",...: 42983 34374 24826 5...
##   $ DateTimeOfAccident   : Factor w/ 36673 levels "1988-01-01T09:00:00Z",...: 28968 22374 16652 35...
##   $ DateReported         : Factor w/ 6653 levels "1988-01-08T00:00:00Z",...: 5278 4020 3009 6390 9...
##   $ Age                  : int  48 43 30 41 36 50 39 56 49 30 ...
##   $ Gender               : Factor w/ 3 levels "F","M","U": 2 1 2 2 2 2 2 2 2 2 ...
##   $ MaritalStatus        : Factor w/ 4 levels "", "M", "S", "U": 2 2 4 3 2 2 2 2 2 3 ...
##   $ DependentChildren    : int  0 0 0 0 0 0 0 0 0 0 ...
##   $ DependentsOther      : int  0 0 0 0 0 0 0 0 0 0 ...
##   $ WeeklyWages          : num  500 509 709 555 377 ...
##   $ PartTimeFullTime     : Factor w/ 2 levels "F","P": 1 1 1 1 1 1 1 1 1 1 ...
##   $ HoursWorkedPerWeek   : num  38 37.5 38 38 38 38 38 40 38 37 ...
##   $ DaysWorkedPerWeek    : int  5 5 5 5 5 5 5 5 5 5 ...
##   $ ClaimDescription      : Factor w/ 28114 levels "A BACK AND HIT HAND CRUSHED INJURY LEFT HAND",...
##   $ InitialIncurredClaimsCost: int  1500 5500 1700 15000 2800 500 500 500 925 1500 ...
##   $ UltimateIncurredClaimCost: num  4303 6106 2099 16283 3772 ...

# Inspeccionamos 10 observaciones elegidas de manera aleatoria
set.seed(1)
claim[sample(nrow(claim), 10), ]

##      ClaimNumber DateTimeOfAccident       DateReported Age Gender
## 24388 WC6275162 1997-01-28T17:00:00Z 1997-03-04T00:00:00Z 22     F
## 43307 WC4085066 1993-05-27T16:00:00Z 1993-07-10T00:00:00Z 39     M
## 4050  WC8202666 2002-04-06T05:00:00Z 2002-04-18T00:00:00Z 47     F
## 11571 WC5284670 1995-08-22T11:00:00Z 1995-09-19T00:00:00Z 16     M
## 25173 WC8654211 2003-05-26T12:00:00Z 2003-06-15T00:00:00Z 26     F
## 32618 WC8480757 2002-06-04T11:00:00Z 2002-06-11T00:00:00Z 43     M
## 50951 WC6773896 1999-07-20T10:00:00Z 1999-10-05T00:00:00Z 22     M
## 13903 WC5352408 1996-07-23T10:00:00Z 1996-08-07T00:00:00Z 19     M
## 8229  WC2074383 1989-04-25T12:00:00Z 1989-06-15T00:00:00Z 20     M
## 25305 WC3805664 1992-10-21T08:00:00Z 1992-11-05T00:00:00Z 22     M
##      MaritalStatus DependentChildren DependentsOther WeeklyWages
## 24388           S                  0                   0      200.00
```

```

## 43307      M      0      0      655.74
## 4050       M      0      0      500.00
## 11571      S      0      0      200.00
## 25173      M      0      0      500.00
## 32618      M      0      0      500.00
## 50951      S      0      0      399.00
## 13903      S      0      0      610.32
## 8229       S      0      0      293.40
## 25305      S      0      0      378.29
##          PartTimeFullTime HoursWorkedPerWeek DaysWorkedPerWeek
## 24388       P      14      2
## 43307       F      40      5
## 4050        P      30      5
## 11571        F      38      5
## 25173        F      38      5
## 32618        F      38      5
## 50951        F      38      5
## 13903        F      38      5
## 8229         F      40      5
## 25305        F      38      5
##                               ClaimDescription
## 24388 KNIFE SLIPPED WHILST USING KNIFE LACERATED LEFT MIDDLE FINGER
## 43307                      STRAINED LEANING OVER BACK STRAIN
## 4050          SPRAINED WRIST LEFT WRIST PAIN SWELLING LEFT KNEE
## 11571          STRUCK KNIFE LACERATION THUMB FINGER RIGHT THUMB
## 25173          SLIPPED ON ROLLER TENDONITIS RIGHT SHOULDER
## 32618          FINGER STRUCK DOOR FRAME LACERATED LEFT INDEX FINGER
## 50951          FINGER SEVERED FINGER RIGHT FINGER
## 13903          JARRED BACK DRIVING FORKLIFT STRAINED SHOULDER UPPER BACK
## 8229           HIT CHEST ON BOTTOM RUNG STRAINED LEFT CALF
## 25305          NATURE AND CONDITIONS RIGHT KNEE AND STRAINED KNEE
##          InitialIncurredClaimsCost UltimateIncurredClaimCost
## 24388            200      264.76554
## 43307            7500     5951.51745
## 4050             33000    20119.47663
## 11571             300      76.66139
## 25173             1000     2204.90575
## 32618             2000     1460.36620
## 50951             19000    11083.54117
## 13903             200      2656.80052
## 8229              600      646.56496
## 25305             2500     3249.06594

```

1.1 Preparación de los datos

Cambiamos el nombre de las variables a castellano. En concreto, se pide que se denominen de la siguiente forma: Id, Ocurrencia, Apertura, Edad, Sexo, Estado, Dependientes, OtrosDepend, Salario, Jornada, CosteInicio, CosteFinal, HorasSemana, DiasSemana y Descripcion.

Solución:

```

claim<-claim %>%
  rename(
  Id=ClaimNumber,
  Ocurrencia=DateTimeOfAccident,

```

```

Apertura=DateReported,
Edad=Age,
Sexo=Gender,
Estado=MaritalStatus,
Dependientes=DependentChildren,
OtrosDepend=DependentsOther,
Salario=WeeklyWages,
Jornada=PartTimeFullTime,
CosteInicio=InitialIncurredClaimsCost,
CosteFinal=UltimateIncurredClaimCost,
HorasSemana=HoursWorkedPerWeek,
DiasSemana=DaysWorkedPerWeek,
Descripcion=ClaimDescription
)
names(claim)

```

```

## [1] "Id"           "Ocurrencia"    "Apertura"     "Edad"         "Sexo"
## [6] "Estado"        "Dependientes"  "OtrosDepend"  "Salario"       "Jornada"
## [11] "HorasSemana"   "DiasSemana"    "Descripcion"  "CosteInicio"   "CosteFinal"

```

- Las variables ‘Ocurrencia’ y ‘Apertura’ están clasificadas como factor. Para poder trabajar con ellas hay que convertirlas en fechas.
- Crear una variable denominada ‘tiempo’ que contabilice en días el tiempo que tarda en abrirse un siniestro por la compañía desde su ocurrencia.

Solución:

```

claim$Ocurrencia<-as.Date(claim$Ocurrencia)
claim$Apertura<-as.Date(claim$Apertura)

# Tiempo en abrirse siniestro desde ocurrencia

claim$tiempo<-as.numeric(claim$Apertura-claim$Ocurrencia)

```

1.2 Clasificación de tiempo

La variable `tiempo` indica la duración de apertura del siniestro de la siguiente forma: “Muy rápido” si se apertura en 15 días o menos, “Rápido” si se apertura entre 16 y 30 días, “Lento” si se apertura entre 31 y 89 días, y “Muy lento” si tarda 90 días o más en aperturarse el siniestro. Cread una variable categórica denominada `Clasificacion`, que clasifique el siniestro según estas categorías.

Solución:

```

# Recodificamos los valores de tiempo en una nueva variable
claim <- claim %>% mutate(Clasificacion = case_when(
  tiempo >= 90 ~ "Muy lento",
  (tiempo>=31) & (tiempo<=89) ~ "Lento",
  (tiempo>=16) & (tiempo<=30) ~ "Rápido",
  (tiempo<=15) ~ "Muy rápido"
))

#comprobación
head(cbind(claim$tiempo, claim$Clasificacion))

##      [,1] [,2]
## [1,] "87"  "Lento"

```

```

## [2,] "13" "Muy rápido"
## [3,] "20" "Rápido"
## [4,] "30" "Rápido"
## [5,] "29" "Rápido"
## [6,] "80" "Lento"
#summary(claim)

# Convertimos la variable a factor ordenado con 4 categorías
claim$Clasificacion <- factor(claim$Clasificacion, levels = c("Muy rápido", "Rápido", "Lento", "Muy lento"))

```

1.3 Valores ausentes

- Analizad el número de categorías distintas en las variables ‘Descripcion’, ‘Sexo’ y ‘Estado’. ¿Cuántas descripciones distintas hay de los siniestros?
- Representad las observaciones con la categoría “U” (U=unknown) en las variables ‘Sexo’ y ‘Estado’ como missings.
- Comprobad la proporción de observaciones que 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. Denominamos al conjunto de datos claimNet.

Solución:

```

# Obtenemos el número de atributos distintos de la variable 'Descripcion', 'Sexo' y 'Estado'
apply(claim, 2, function(x) length(unique(x)))

##          Id    Ocurrencia     Apertura      Edad       Sexo
##      54000        6326        6653        68         3
##      Estado  Dependientes OtrosDepend   Salario    Jornada
##           4            9            5        13211        2
##      HorasSemana    DiasSemana  Descripcion CosteInicio CosteFinal
##        424            7        28114        1989        53897
##      tiempo Clasificacion
##        574            4

## Las variables `Sexo` y `Estado` están clasificadas como factor. Analizar categorías

table(claim$Sexo)

##
##      F      M      U
## 12338 41660     2
table(claim$Estado)

##
##      M      S      U
## 29 22516 26161 5294

#U=unknown y categoría "" se clasifican como missing
claim$Sexo[claim$Sexo=="U"] <- NA
claim$Estado[claim$Estado=="U" | claim$Estado==""] <- NA

#elimina categorías vacías
claim$Sexo <- droplevels(claim$Sexo)
claim$Estado <- droplevels(claim$Estado)

```

```

summary(claim$Sexo)

##      F      M  NA's
## 12338 41660     2

summary(claim$Estado)

##      M      S  NA's
## 22516 26161   5323

# Analizamos la proporción en tanto por ciento de datos vacíos
sort(round(colMeans(is.na(claim))*100,3), decreasing = TRUE)

##          Estado            Sexo           Id    Ocurrencia       Apertura
## 9.857        0.004        0.000        0.000        0.000
##          Edad    Dependientes  OtrosDepend     Salario      Jornada
## 0.000        0.000        0.000        0.000        0.000
## HorasSemana    DiasSemana Descripcion  CosteInicio  CosteFinal
## 0.000        0.000        0.000        0.000        0.000
##          tiempo Clasificacion
## 0.000        0.000

# Valores perdidos (NA) en las variables relacionadas
length(which(is.na(claim$Estado)))

## [1] 5323

length(which(is.na(claim$Sexo)))

## [1] 2

#eliminamos valores perdidos. Conjunto de datos claimNet
claimNet <- claim %>% filter(!is.na(claim$Estado) & !is.na(claim$Sexo))

```

Respuesta:

El nuevo conjunto de datos tiene 48675 observaciones, mientras que el conjunto inicial tiene 54000 observaciones. Esto se debe principalmente a que la variable **Estado** tiene un 9.86% de missings. La otra variable con missings es **Sexo**, pero su incidencia es mucho menor.

1.4 Salud mental

La compañía está preocupada por las bajas por salud mental. Por este motivo, quiere monitorizar las bajas que incluyan las palabras **Stress**, **Anxiety**, **Harassment** o **Depression**. Se pide:

- Crear la variable dicotómica denominada ‘RiesgoSM’ si la variable ‘Descripcion’ incluye alguna de estas palabras.

Solución:

```

# Generar la variable 'RiesgoSM' como binaria.

Descripcion<-as.character(claimNet$Descripcion)

claimNet$RiesgoSM<-as.factor(ifelse(grepl("STRESS", Descripcion)==TRUE, TRUE,
ifelse(grepl("ANXIETY", Descripcion)==TRUE, TRUE,
ifelse(grepl("HARASSMENT", Descripcion)==TRUE, TRUE,
grepl("DEPRESSION", Descripcion)))))


```

```
summary(claimNet$RiesgoSM)
```

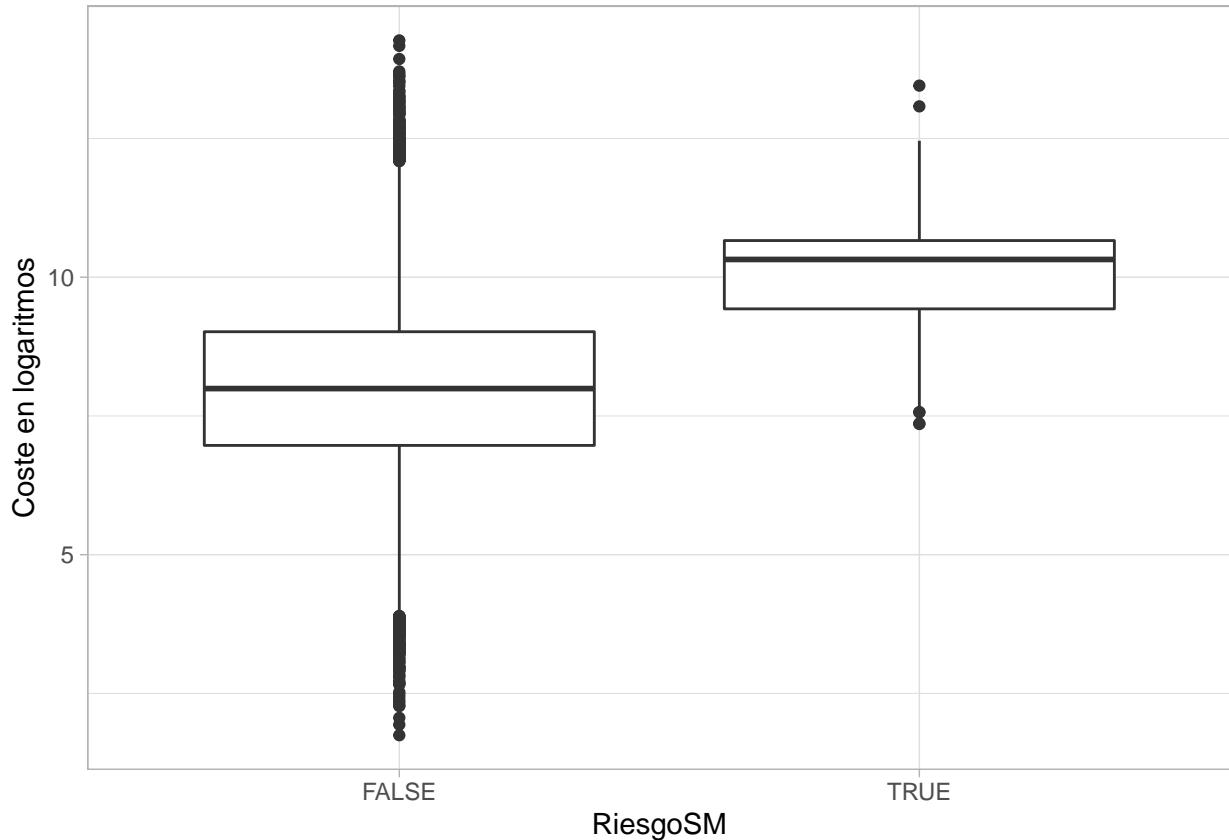
```
## FALSE TRUE  
## 48459 216
```

1.5 Análisis visual

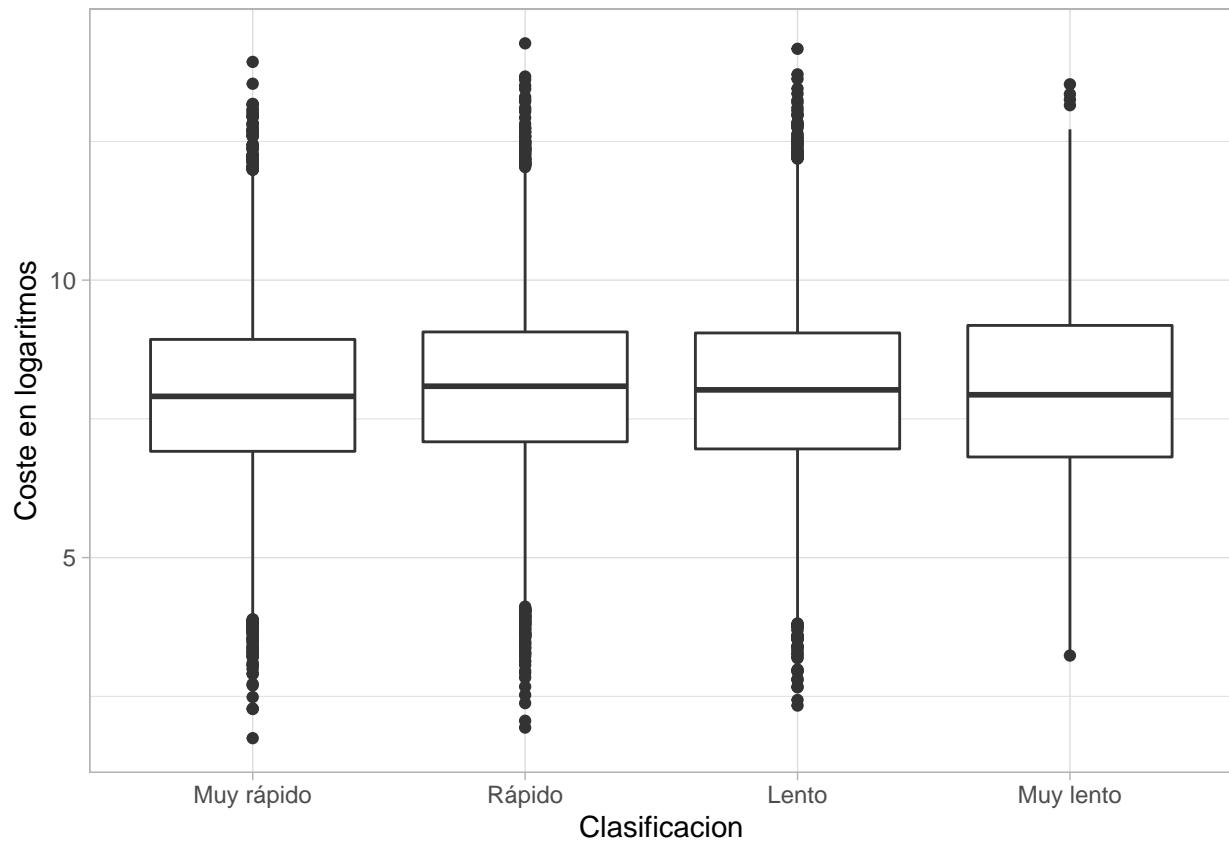
1. Mostrad con diversos diagramas de caja la distribución de la variable ‘CosteFinal’ en escala logarítmica según la variable ‘Sexo’, según ‘Estado’, según ‘Clasificacion’ y según ‘RiesgoSM’.
2. Interpretad los gráficos brevemente.

Solución:

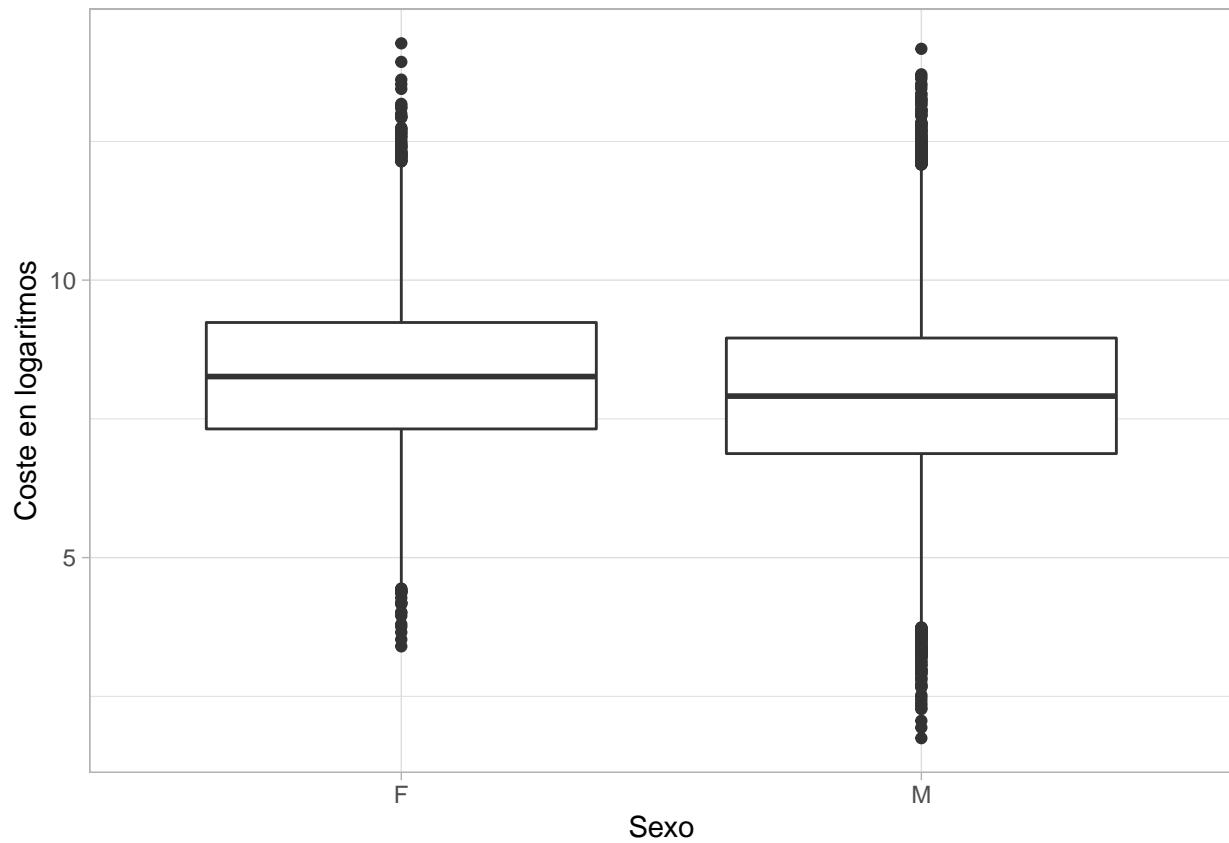
```
par(mfrow=c(2, 2))  
#RiesgoSM  
ggplot(claimNet, aes(RiesgoSM, log(CosteFinal))) +  
  geom_boxplot() +  
  labs(x = "RiesgoSM", y = "Coste en logaritmos") +  
  theme_light()
```



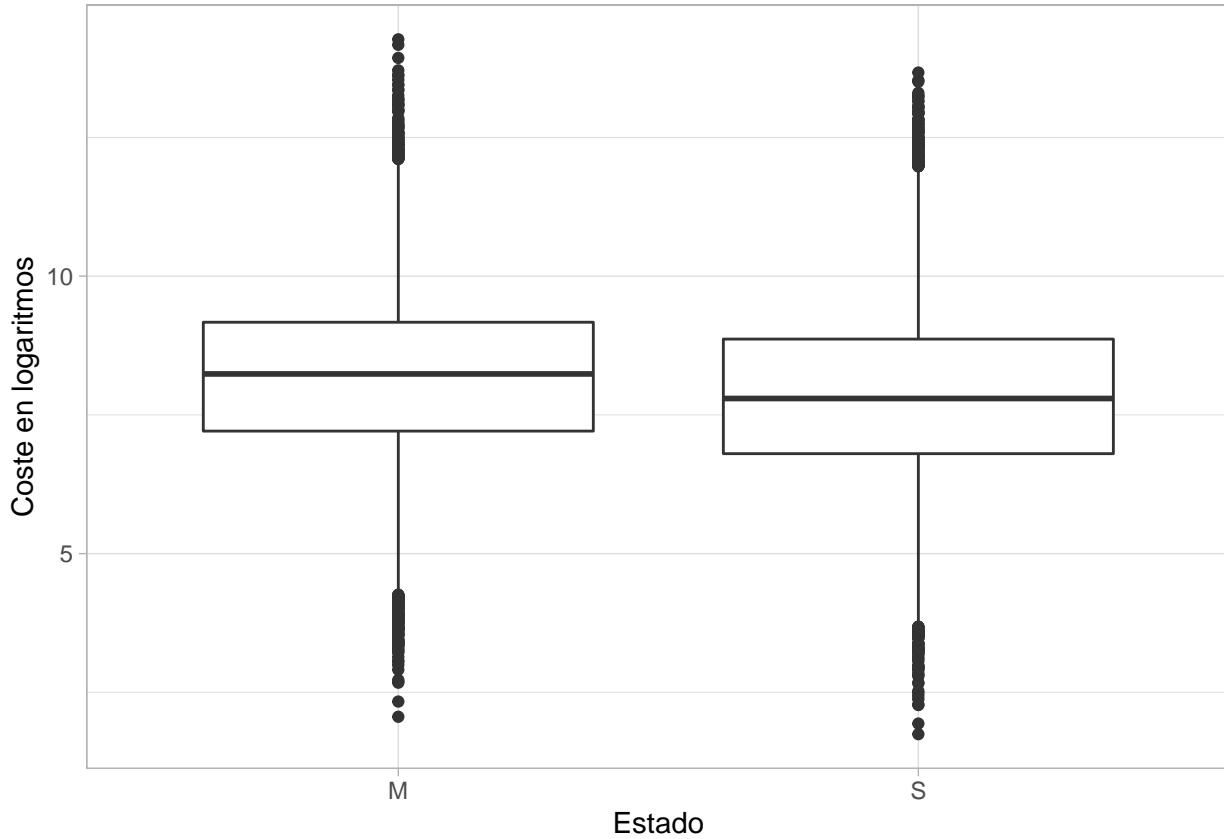
```
#Clasificacion  
ggplot(claimNet, aes(Clasificacion, log(CosteFinal))) +  
  geom_boxplot() +  
  labs(x = "Clasificacion", y = "Coste en logaritmos") +  
  theme_light()
```



```
#sexo  
ggplot(claimNet, aes(Sexo, log(CosteFinal))) +  
  geom_boxplot() +  
  labs(x = "Sexo", y = "Coste en logaritmos") +  
  theme_light()
```



```
#Clasificacion
ggplot(claimNet, aes(Estado, log(CosteFinal))) +
  geom_boxplot() +
  labs(x = "Estado", y = "Coste en logaritmos") +
  theme_light()
```



Breve interpretación de los gráficos:

- En los boxplots en escala normal no se puede apreciar nada debido a la escala de las observaciones extremas (no mostrado).
- Cuando analizamos las observaciones en escala logarítmica, las mayores diferencias en costes se observan para la variable 'RiesgoSM'. En el caso de 'Sexo' y 'Estado' también se aprecian diferencias pero más sutiles. En cambio, en relación a 'Clasificacion' no se observan diferencias destacables.

1.6 Comprobación de normalidad

¿Podemos asumir que la variable `CosteFinal` tiene una distribución normal? Debéis justificar la respuesta a partir de métodos visuales y contrastes.

- Realizad inspección visual de normalidad.
- Realizad contraste de normalidad de Lilliefors (p.ej. con función `lillie.test` de la librería `nortest`).
- Realizad inspección visual y contraste de normalidad a la variable `Coste Final` en escala logarítmica.

Respuesta:

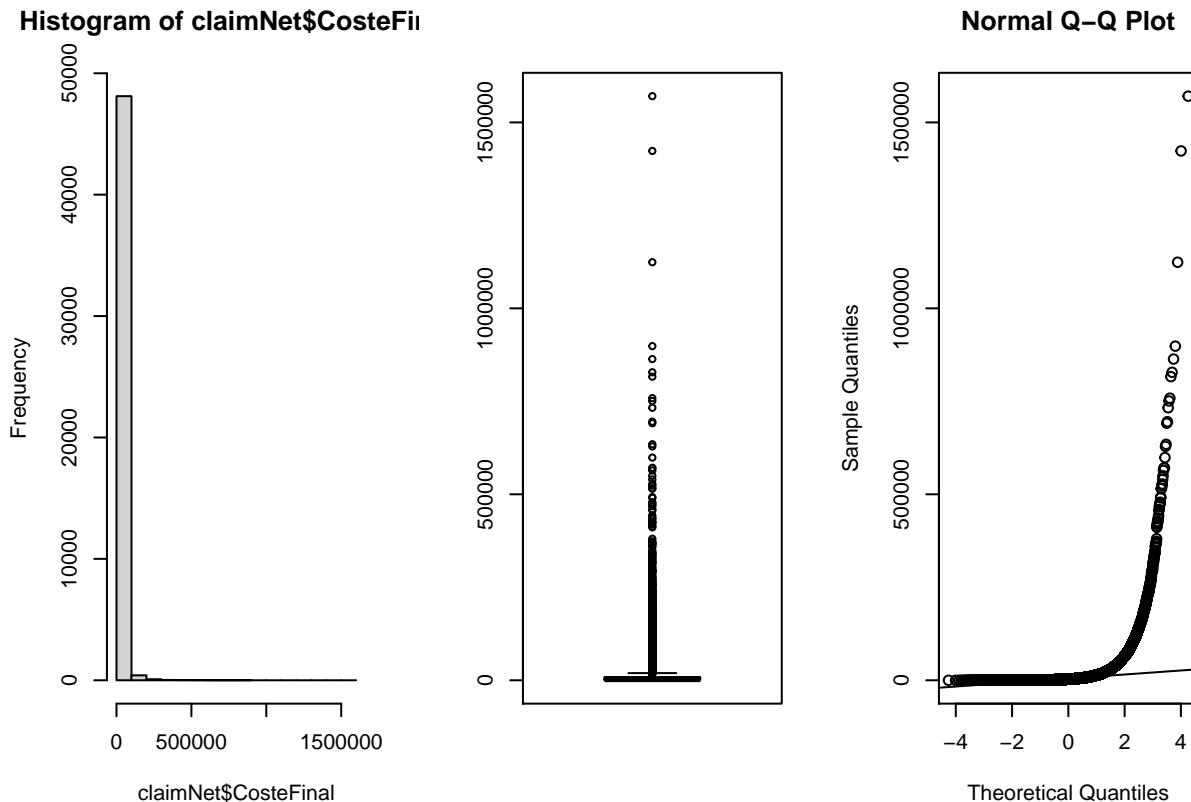
```
lillie.test(claimNet$CosteFinal) #contraste
```

```
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
##  data:  claimNet$CosteFinal
##  D = 0.36734, p-value < 2.2e-16
```

```

par(mfrow=c(1,3))
hist(claimNet$CosteFinal) # histograma
boxplot(claimNet$CosteFinal) # diagrama de cajas
qqnorm(claimNet$CosteFinal) # gráfico de cuantiles
qqline(claimNet$CosteFinal)

```



Respuesta: Los resultados indican claramente que `CosteFinal` no se distribuye según una distribución normal

```

## Comprobación de normalidad CosteFinal en escala logarítmica
CosteFinalLog<-log(claimNet$CosteFinal)

```

```

lillie.test(CosteFinalLog) #contraste normalidad

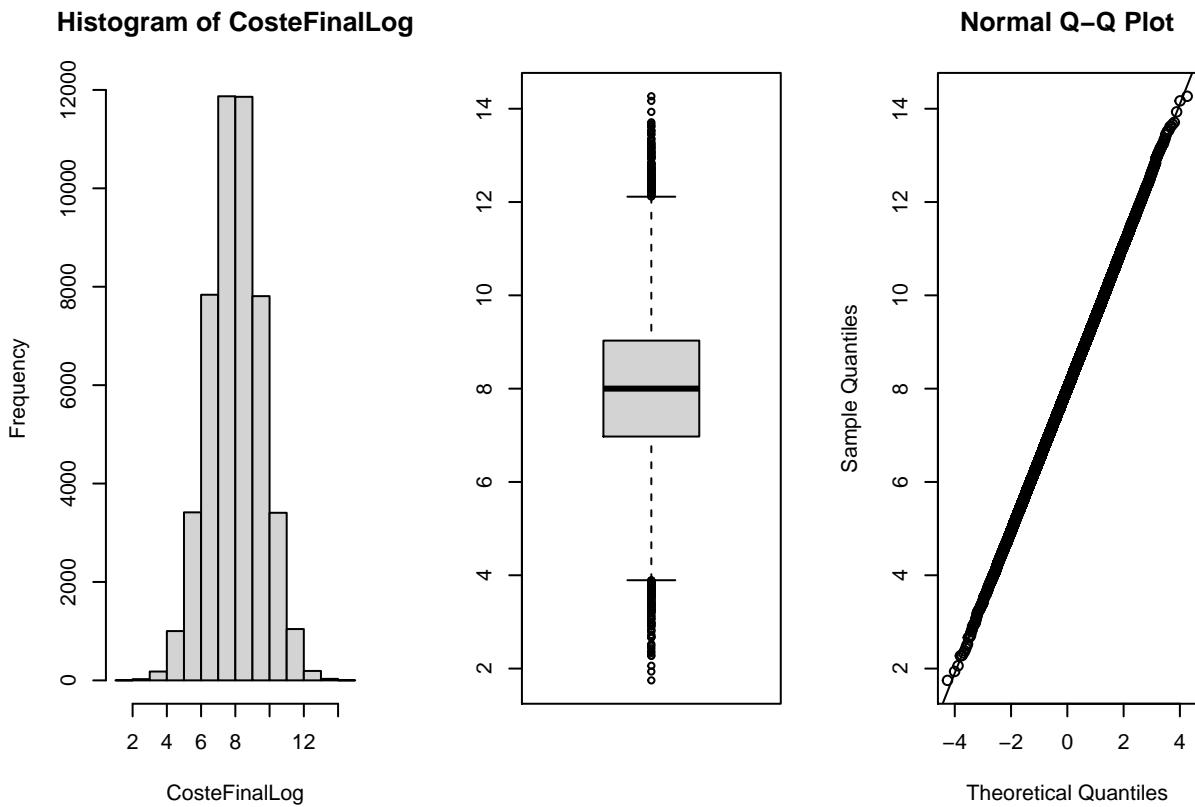
```

```

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: CosteFinalLog
## D = 0.0018307, p-value = 0.9579

par(mfrow=c(1,3))
hist(CosteFinalLog) # histograma
boxplot(CosteFinalLog) # diagrama de cajas
qqnorm(CosteFinalLog) # gráfico de cuantiles
qqline(CosteFinalLog)

```



Respuesta: Cuando realizamos contraste de normalidad no se puede descartar con un nivel de significación del 5% que el coste en escala logarítmica se distribuya según una distribución normal.

2 Estadística inferencial

Utilizamos el conjunto de datos claimNet.

2.1 Intervalo de confianza de la media poblacional de la variable CosteFinal

- Calculad manualmente el intervalo de confianza al 95% de la media poblacional de la variable ‘CosteFinal’ en escala normal (No se pueden utilizar funciones como t.test o z.test para el cálculo).
- A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

```
# Definimos una función para calcular el intervalo de confianza de una variable que sigue una distribución normal
InterConf_tS <- function(x, alfa){
  n <- length(x)
  sd <- sd(x)
  SE <- sd / sqrt(n)
  z <- qt(1 - alfa/2, df=n-1, lower.tail=FALSE )
  L <- mean(x) - z*SE
  U <- mean(x) + z*SE
  c(L, U)
}

# Calculamos el intervalo de confianza al 95%, es decir, alfa = 0.05
```

```

icNor<-claimNet %>% .$CosteFinal %>% InterConf_tS(0.05)
icNor

## [1] 9450.34 9958.82
# Comprobación
t.test(claimNet$CosteFinal)

##
## One Sample t-test
##
## data: claimNet$CosteFinal
## t = 74.815, df = 48674, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 9450.34 9958.82
## sample estimates:
## mean of x
## 9704.58

```

Respuesta:

- Por el teorema del límite central, podemos asumir que la media muestral sigue una distribución normal, puesto que tenemos una muestra de tamaño grande n=48675.
 - El intervalo de confianza para ‘CosteFinal’ es: (9450.34, 9958.82). La interpretación del intervalo de confianza es que si repitiésemos en un número elevado de muestras el mismo procedimiento, el 95% de los intervalos obtenidos contendrían el valor de la media poblacional de la variable ‘CosteFinal’.
-

2.2 Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que la indemnización a las mujeres supera en más de 1000 EUR la de los hombres?

Responded a la pregunta utilizando un nivel de confianza del 95%.

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

Seguid los pasos que se detallan a continuación.

2.2.1 Escribid la hipótesis nula y la alternativa

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

$$H_0 : \mu_1 - \mu_2 \leq 1000$$

$$H_1 : \mu_1 - \mu_2 > 1000$$

donde μ_1 denota la media de la indemnización a las mujeres y μ_2 la media de la indemnización a los hombres.

2.2.2 Justificación del test a aplicar

Respuesta:

- Por el teorema del límite central, podemos asumir normalidad, puesto que tenemos una muestra de tamaño grande n=48675 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.
- Comparamos las varianzas de las dos muestras.

```

F <- claimNet$CosteFinal[claimNet$Sexo=="F"] # Mujeres
M <- claimNet$CosteFinal[claimNet$Sexo=="M"] # Hombres

# Definimos nuestra propia función para hacer el test de homocedasticidad
# basado en la distribución F de Snedecor
varianzasTest <- function(F, M, alfa){
  n1 <- length(F)
  n2 <- length(M)
  s1 <- sd(F)
  s2 <- sd(M)
  fobs <- s1^2 / s2^2
  fcritL <- qf( alfa/2, df1=n1-1, df2=n2-1)
  fcritU <- qf( 1- alfa/2, df1=n1-1, df2=n2-1)
  pvalue <- min(pf(abs(fobs), df1=n1-1, df2=n2-1, lower.tail=FALSE ),
                 pf( fobs, df1=n1-1, df2=n2-1))*2
  c(fobs, fcritL, fcritU, pvalue)
}

# Obtenemos el valor observado, los umbrales de la región de aceptación y el p-valor

Resul<-cbind(varianzasTest(F, M, 0.05))
rownames(Resul)<-c("fobs", "fcritL", "fcritU", "pvalue")
Resul

##          [,1]
## fobs    1.575073e+00
## fcritL  9.705235e-01
## fcritU  1.030123e+00
## pvalue  1.298263e-211
#comprobación
var.test(F, M)

##
##  F test to compare two variances
##
## data: F and M
## F = 1.5751, num df = 11256, denom df = 37417, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.529015 1.622911
## sample estimates:
## ratio of variances
##          1.575073


- El contraste de varianzas nos muestra un valor p menor de 0.05 por lo que no se puede asumir igualdad de varianzas en las dos poblaciones.
- En consecuencia, aplicamos un test sobre la media de dos muestras independientes con varianza desconocida y diferente. Es un test unilateral por la derecha.

```

2.2.3 Cálculos

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

Respuesta:

```

#Contraste de media

alfa <- 0.05      # Nivel de significación
n1 <- length(F)  # Tamaño muestra 1
n2 <- length(M)  # Tamaño muestra 2
s1 <- sd(F)       # Desviación típica muestra 1
s2 <- sd(M)       # Desviación típica muestra 2
mean1 <- mean(F) # Media muestra 1
mean2 <- mean(M) # Media muestra 2
d0 <- 1000         # Diferencia entre las medias

# Calculamos los grados de libertad
df <- ((s1^2/n1 + s2^2/n2)^2) / (((s1^2/n1)^2/(n1-1)) + ((s2^2/n2)^2/(n2-1)))

# Calculamos el valor observado, el estadístico de contraste
tobs <- (mean1-mean2-d0) / sqrt(s1^2/n1 + s2^2/n2)

# Obtenemos el valor crítico que define la región de aceptación
tcritU <- qt(alfa, df, lower.tail=FALSE)

# Calculamos el p-valor
pvalue <- pt(abs(tobs), df, lower.tail=FALSE)

# Mostramos los resultados

Result<-cbind(c(tobs, tcritU, pvalue, df))
rownames(Result)<-c("tobs", "tcritU", "p-value", "df")
Result

##          [,1]
## tobs    3.591609e+00
## tcritU  1.644950e+00
## p-value 1.648202e-04
## df      1.579319e+04

```

2.2.4 Interpretación del test

Respuesta:

El pvalor del test (2×10^{-4}) es inferior al nivel de significación (0.05). Además el valor observado 3.592 es mayor que el valor crítico 1.645. Por tanto, podemos rechazar la hipótesis nula a favor de la alternativa y podemos concluir que en promedio la indemnización a los mujeres es superior en más de 1000EUR a la indemnización de los hombres.

Para comprobarlo podemos usar la función R t.test:

```
#Comprobamos los resultados con los proporcionados con la funcion t.test()
t.test(F, M, alternative="greater", mu = d0, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data: F and M
## t = 3.5916, df = 15793, p-value = 0.0001648
## alternative hypothesis: true difference in means is greater than 1000
```

```

## 95 percent confidence interval:
## 1675.27      Inf
## sample estimates:
## mean of x mean of y
## 11431.059   9185.178
t.test(F, M, alternative="greater", mu = d0, var.equal=FALSE)$p.value
## [1] 0.0001648202

```

3 Modelo de regresión lineal

Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: Edad, Sexo, Estado, Dependientes, OtrosDepend, Salario, Jornada, HorasSemana, DiasSemana, Clasificacion, RiesgoSM, CosteInicio y como variable dependiente el CosteFinal en escala logarítmica (Nota: se recomienda transformar también a escala logarítmica la variable explicativa CosteInicio)

Solución:

```
regresionLineal <- lm(log(CosteFinal) ~ Edad+Sexo+Estado+Dependientes+OtrosDepend+Salario+Jornada+Horas
```

3.1 Interpretación del modelo

Interpretad el modelo lineal ajustado:

- ¿Cuál es la calidad del ajuste?
- Explicad la contribución de las variables explicativas en el modelo.

Respuesta:

```
summary(regresionLineal)
```

```

##
## Call:
## lm(formula = log(CosteFinal) ~ Edad + Sexo + Estado + Dependientes +
##     OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
##     Clasificacion + RiesgoSM + log(CosteInicio), data = claimNet)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -5.1107 -0.4414 -0.0821  0.3568  7.3049 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           1.509e+00  4.867e-02 31.002 < 2e-16 ***
## Edad                  4.466e-03  3.628e-04 12.310 < 2e-16 ***
## SexoM                -1.423e-01  8.871e-03 -16.041 < 2e-16 ***
## EstadoS               -3.836e-02  8.788e-03 -4.365 1.28e-05 ***
## Dependientes          4.265e-02  6.950e-03  6.136 8.51e-10 ***
## OtrosDepend          5.834e-02  3.251e-02  1.794 0.072773  
## Salario                6.697e-04  1.602e-05 41.802 < 2e-16 ***
## JornadaP              5.649e-02  1.638e-02  3.448 0.000565 ***
## HorasSemana            4.032e-04  3.109e-04  1.297 0.194678  
## DiasSemana             -5.559e-02  8.758e-03 -6.347 2.21e-10 ***
## ClasificacionRápido  2.561e-02  9.005e-03  2.844 0.004462 ** 
## 
```

```

## ClasificacionLento      1.583e-02  9.430e-03   1.679  0.093255 .
## ClasificacionMuy lento 1.540e-03  1.405e-02   0.110  0.912672
## RiesgoSMTRUE            1.892e-01  5.415e-02   3.494  0.000476 ***
## log(CosteInicio)       8.273e-01  2.562e-03  322.927 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7901 on 48660 degrees of freedom
## Multiple R-squared:  0.7333, Adjusted R-squared:  0.7332
## F-statistic:  9555 on 14 and 48660 DF,  p-value: < 2.2e-16

```

El valor de R^2 ajustado es 0.733. Es decir, el modelo explica un 73.3% de la varianza de las indemnizaciones, lo cual indica que la capacidad explicativa del modelo es buena para estimar el coste. El pvalor del estadístico es menor de 0.05, lo cual indica que el conjunto de variables explicativas contribuyen significativamente a el coste.

En cuanto al análisis por separado de las variables explicativas, se observa que las todas variables incluidas en el modelo son significativas con la excepción de `OtrosDepend` y `HorasSemana` que no son significativas al 5% de nivel de significación. Los coeficientes positivos indican que cuanto mayores sean los valores de estas variables, mayor es coste. Por el contrario, los coeficientes negativos indican que a mayor valor, menor coste esperado. Por ejemplo, la variable `Sexo` tiene una correlación negativa: si es un hombre, comparado con ser mujer, la indemnización se reduce en 0,1423 (en escala logarítmica).

3.2 Análisis residuos

Por último, para profundizar en la calidad del ajuste se deben analizar los residuos que nos indicarán realmente como se ajusta nuestro modelo a los datos muestrales.

- La salida de ‘summary()’ presenta los principales estadísticos de la distribución de los residuos. Analizad los valores estimados de los estadísticos.
- Realizad un análisis visual de los residuos

Respuesta:

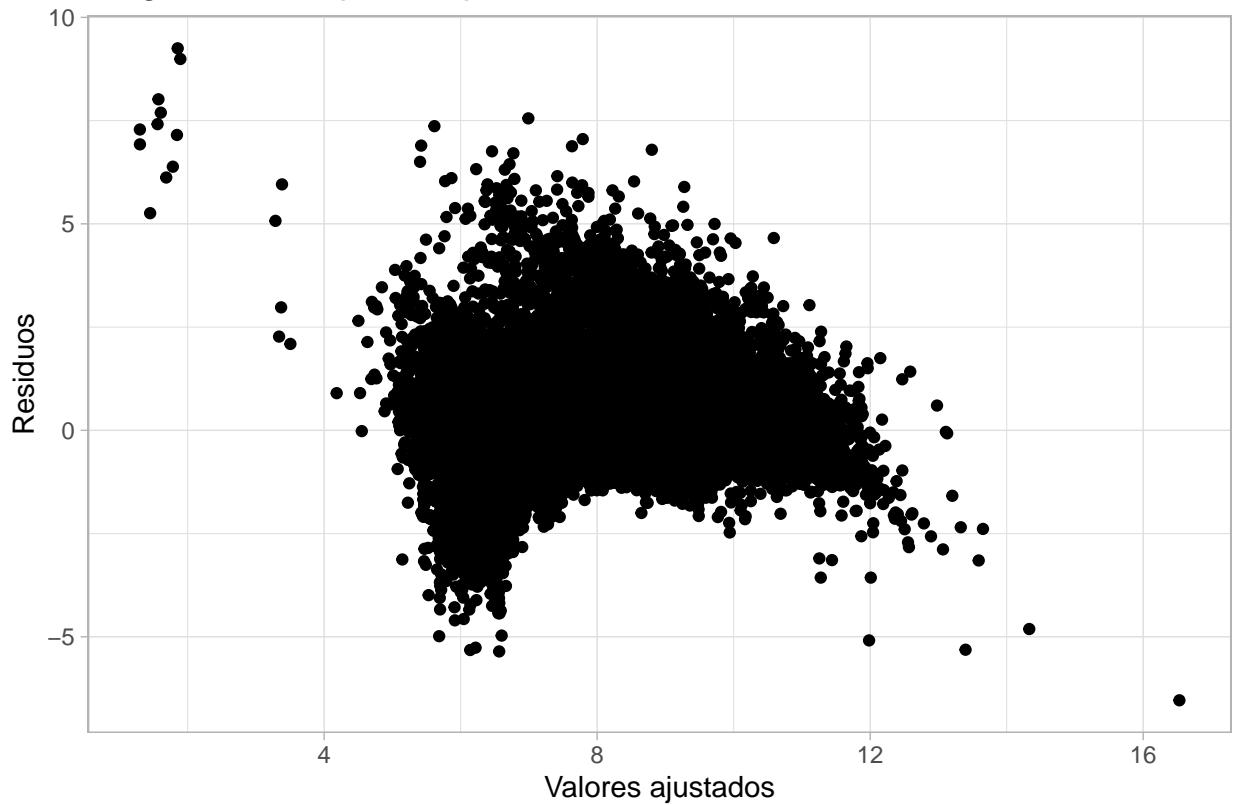
```

residuos <- rstandard(regresionLineal)
valores.ajustados <- fitted(regresionLineal)

ggplot(regresionLineal, aes(x=valores.ajustados, y=residuos)) + geom_point() +
  ggtitle("Diagrama de dispersion para los residuos") + xlab("Valores ajustados") +
  ylab("Residuos") + theme_light()

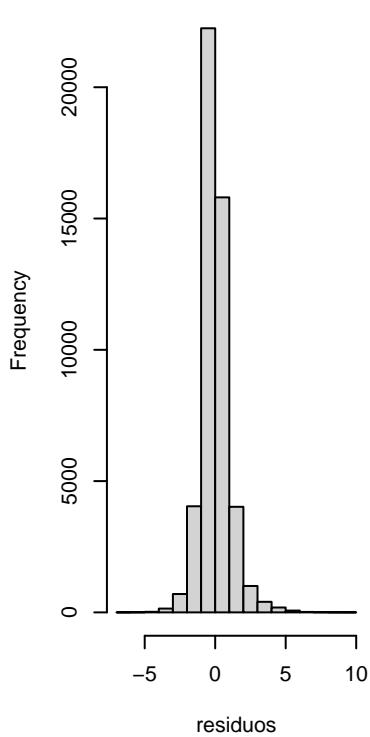
```

Diagrama de dispersion para los residuos

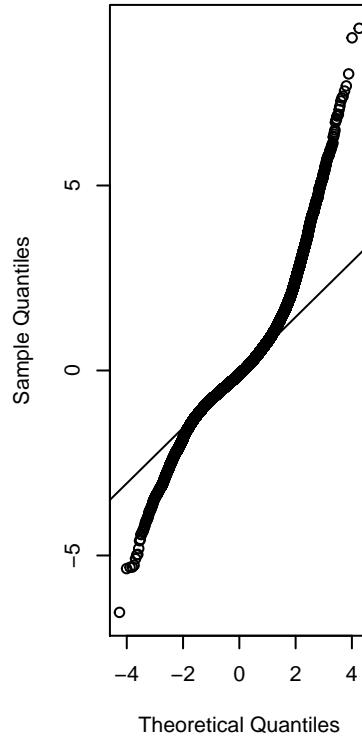


```
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
qqline(residuos)
```

Histogram of residuos



Normal Q-Q Plot



Los 5 estadísticos principales de la distribución de los residuos mostrados por la salida de `summary()` : mínimo, primer cuartil, mediana, tercer cuartil y máximo muestran que la distribución de los residuos es simétrica entorno a 0.

Los gráficos de los residuos muestran que no se puede aceptar que los residuos se distribuyan según una normal, con mal ajuste en las colas de la distribución (no normalidad).

3.3 Predicción

Predecir el coste esperado para las siguientes características: Edad=24, Sexo= “F”, Estado=“S”, Dependientes=1, OtrosDepend=0, Salario=500, Jornada=“F”, HorasSemana=40, DiasSemana=5, Clasificacion=“Lento”, RiesgoSM=“TRUE” y “CosteInicio”=10000.

(Nota: Debes tener en cuenta que el valor esperado de una variable aleatoria que su logaritmo se distribuye según una normal, i.e. distribución lognormal, es $\exp(\mu + \sigma^2/2)$ donde μ y σ^2 son la media y la varianza de la transformación logarítmica).

Solución:

```
#sin transformación
pre1<-predict(regresionLineal, data.frame(Edad=24, Sexo= "F", Estado="S", Dependientes=1, OtrosDepend=0))

#varianza de los residuos
SSE<-sum((regresionLineal$residuals)^2)
n<-nrow(claimNet)
k<-length(regresionLineal$coef)
sigma<-SSE/(n-(k+1))
```

```
#valor esperado en escala original
preL<-exp(pre1+sigma/2)
#valor mediano
preMe<-exp(pre1)
```

El coste esperado es 1.858492×10^4 .

4 Regresión logística

4.1 Modelo predictivo

Utilizando las mismas características como variables explicativas, ajustad un modelo predictivo basado en la regresión logística para predecir la probabilidad de que la compañía cuantifique inicialmente el coste del siniestro de forma insuficiente.

Para ello, cread una variable **Deficit** que indique si la valoración inicial del coste del siniestro (**CosteInicio**) es inferior a la indemnización finalmente pagada por la compañía (**CosteFinal**). La variable **Deficit** debe codificarse como una variable dicotómica, que toma el valor 0 cuando la valoración inicial ha sido suficiente y 1 cuando la valoración inicial ha sido insuficiente.

La variable **Deficit** será la variable dependiente del modelo. Analizad la calidad del modelo y las variables que son relevantes.

Solución:

```
claimNet <- claimNet %>% mutate(Deficit = if_else(claimNet$CosteFinal>claimNet$CosteInicio , 1, 0))

claimNet$Deficit<-as.factor(claimNet$Deficit)
levels(claimNet$Deficit)<-c("No","Si")
#ajuste
model.log <- glm(claimNet, formula= Deficit~ Edad+Sexo+Estado+Dependientes+OtrosDepend+Salario+Jornada+
family=binomial(link=logit))

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(model.log)

##
## Call:
## glm(formula = Deficit ~ Edad + Sexo + Estado + Dependientes +
##     OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
##     Clasificacion + RiesgoSM + CosteInicio, family = binomial(link = logit),
##     data = claimNet)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -5.0322   -1.1967    0.7727   1.0054    6.5446
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.228e-01  1.271e-01   3.325  0.000883 ***
## Edad                      9.298e-03  9.872e-04   9.419  < 2e-16 ***
## SexoM                     -2.982e-01  2.400e-02 -12.427  < 2e-16 ***
## EstadoS                  -2.820e-02  2.378e-02  -1.186  0.235728
## Dependientes              1.322e-01  2.022e-02   6.538 6.22e-11 ***
```

```

## OtrosDepend          1.994e-01  9.764e-02   2.042  0.041110 *
## Salario              2.060e-03  5.343e-05  38.558 < 2e-16 ***
## JornadaP             1.515e-01  4.497e-02   3.368  0.000756 ***
## HorasSemana          -4.978e-04 8.408e-04  -0.592  0.553876
## DiasSemana           -1.404e-01 2.433e-02  -5.770 7.93e-09 ***
## ClasificacionRápido  6.660e-02  2.412e-02   2.761  0.005756 **
## ClasificacionLento   4.988e-02  2.526e-02   1.975  0.048313 *
## ClasificacionMuy lento 5.160e-02  3.797e-02   1.359  0.174176
## RiesgoSMTRUE          7.017e-01  1.667e-01   4.209  2.56e-05 ***
## CosteInicio            -5.724e-05 1.183e-06  -48.383 < 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: 66399  on 48674  degrees of freedom
## Residual deviance: 61716  on 48660  degrees of freedom
## AIC: 61746
##
## Number of Fisher Scoring iterations: 5

```

Las variables explicativas que no muestran coeficientes significativos al 5% son **Estado** y **HorasSemana**. Una forma de medir la calidad del modelo es comparando null deviance con residual deviance, de la siguiente forma, $(\text{null.deviance} - \text{deviance})/\text{null.deviance}$. Un valor de 1 indica un ajuste perfecto y un valor de cero sin ajuste. En nuestro caso, se obtiene: 0.0705347

4.2 Interpretación

Interpretad el modelo ajustado. Concretamente, explicad la contribución de las variables explicativas con coeficiente estadísticamente significativo para predecir si la valoración inicial es insuficiente para cubrir el coste del siniestro.

Respuesta: Las variables explicativas que muestran coeficientes significativos al 5% son **Sexo**, **Edad**, **Dependientes**, **OtrosDepen**, **Salario**, **Clasificacion**, **Jornada**, **DiasSemana**, **RiesgoSM** y **CosteInicio**.

Se observa que:

- La probabilidad que la valoración inicial del coste sea insuficiente disminuye si es hombre, con el número de días de la semana que trabaja y con el montante de la valoración inicial.
- La probabilidad aumenta con el salario, número de dependientes, si la baja es por depresión o si la jornada de trabajo es parcial.

4.3 Matriz de confusión

A continuación analizad la precisión del modelo, comparando la predicción del modelo sobre los mismos datos del conjunto de datos. Asumiremos que la predicción del modelo es 1 (valoración inicial del coste insuficiente) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Analizad la matriz de confusión y las medidas de ‘sensitivity’ y ‘specificity’.

Nota: Tomad como categoría de interés que haya déficit en la valoración inicial del coste. Por tanto, déficit igual a 1 será el caso positivo en la matriz de confusión y 0 el caso negativo.

Respuesta:

```
prob <- model.log$fitted.values # Obtenemos las predicciones del modelo
```

```
# Mediante un umbral de probabilidad, 0,5 creamos dos grupos
```

```

prediccion <- as.factor(ifelse(prob >= 0.5, "Si", "No"))
# Calculamos la matriz de confusión
confusionMatrix(prediccion, claimNet$Deficit, positive="Si")

## Confusion Matrix and Statistics
##
##             Reference
## Prediction      No      Si
##           No  7582   3596
##           Si 13139  24358
##
##                   Accuracy : 0.6562
##                   95% CI : (0.6519, 0.6604)
##       No Information Rate : 0.5743
##       P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2523
##
## McNemar's Test P-Value : < 2.2e-16
##
##                   Sensitivity : 0.8714
##                   Specificity : 0.3659
##       Pos Pred Value : 0.6496
##       Neg Pred Value : 0.6783
##       Prevalence : 0.5743
##       Detection Rate : 0.5004
## Detection Prevalence : 0.7704
##       Balanced Accuracy : 0.6186
##
##       'Positive' Class : Si
##

```

La diagonal principal de la matriz obtenida contiene la suma de todas las predicciones correctas, la otra diagonal representa los errores de predicción del modelo, los falsos positivos y los falsos negativos. En concreto, los datos que representa la matriz de confusión corresponden a los siguientes conceptos:

- Verdaderos positivos (VP): Valores predichos como positivos por el modelo y que efectivamente corresponden a un valor positivo, para los datos observados.
- Verdaderos negativos (VN): Valores predichos como negativos por el modelo y que corresponden a un valor negativo para los datos observados.
- Falsos negativos(FN): Valores predichos por el modelo como negativos de forma incorrecta, ya que corresponden a un valor positivo para los datos observados.
- Falsos positivos(FP): Valores predichos como positivos por el modelo, pero que en los datos observados corresponden a valores negativos.

Una de las métricas que se puede utilizar para evaluar un modelo es la exactitud o “accuracy”, que es la proporción entre las predicciones correctas que ha hecho el modelo y el total de predicciones. Como vemos en la matriz de confusión, para nuestro modelo es el 0.6562 con un intervalo de confianza de (0.6519, 0.6604).

Exactitud= (Predicciones correctas)/(Número total de Predicciones)

Si vemos la métrica Kappa que calcula el porcentaje de aciertos más allá del que se podría conseguir haciendo predicciones al azar observamos que este valor no es muy alto, 0.2523.

Disponemos de otras métricas que también debemos considerar:

- La sensibilidad: 0.8714. Proporcion de casos positivos correctamente clasificados.

Sensibilidad = (Verdaderos positivos) / (Verdaderos positivos + Falsos negativos)

- La especificidad: 0.3659. Proporcion de casos negativos correctamente clasificados.

Especificidad = (Verdaderos negativos) / (Verdaderos negativos + Falsos positivos)

La proporción de casos positivos correctamente clasificados es del 87% mientras que la de los negativos es prácticamente del 36%. No es muy buen modelo, como ya se intuía con la métrica kappa.

4.4 Predicción

¿Con qué probabilidad la valoración inicial del siniestro será insuficiente para un hombre de 20 años de edad, soltero, sin hijos ni otros dependientes, con un salario semanal de 300 EUR, jornada partida, con 30 horas semanales y cinco días a la semana, una clasificación del tiempo hasta la apertura del siniestro de “Muy lento”, una baja que no es por depresión y una valoración inicial de 10000EUR?

Respuesta:

```
p1<-predict(model.log, newdata= data.frame(Edad=20, Sexo= "M", Estado="S", Dependientes=0, OtrosDepend=0, Insuficiencia=1), type="response")  
  
exp(p1)/(1+exp(p1))  
  
## [1] 0.4537046  
  
#manualmente  
tt<-summary(model.log)  
expresion <- tt$coefficients[1,1]+  
tt$coefficients[2,1]*20+  
tt$coefficients[3,1] +  
tt$coefficients[4,1] +  
tt$coefficients[5,1]*0+  
tt$coefficients[6,1]*0+  
tt$coefficients[7,1]*300 +  
tt$coefficients[8,1] +  
tt$coefficients[9,1]*30+  
tt$coefficients[10,1]*5+  
tt$coefficients[11,1]*0 +  
tt$coefficients[12,1]*0+  
tt$coefficients[13,1] +  
tt$coefficients[14,1]*0+  
tt$coefficients[15,1]*10000  
  
p2.num<-exp( expresion )  
p2.den<-1+exp(expresion )  
p2<-p2.num/p2.den; p2  
  
## [1] 0.4537046
```

La probabilidad de ser insuficiente la valoración inicial es 45.4%.

5 Análisis de la varianza (ANOVA) de un factor

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable CosteFinal en escala logarítmica en función de la clasificación del siniestro en relación al tiempo transcurrido hasta la apertura. Seguid los pasos que se indican.

5.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

Respuesta:

El factor `Clasificacion` tiene 4 niveles: 1 es el nivel Muy rápido, 2 el Rápido, 3 el Lento y 4 el Muy lento.
Las hipótesis son:

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

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

donde μ_1, μ_2, μ_3 y μ_4 denotan, respectivamente, la media poblacional de Coste final (en escala logarítmica) del siniestro para el grupo de siniestros Muy rápido, Rápido, Lento y Muy lento.

5.2 Modelo

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

Respuesta:

```
#Usando aov
claimNet$CosteFinalL<-log(claimNet$CosteFinal)
claimNet %>% group_by(Clasificacion) %>% summarise(Media=mean(CosteFinalL), Desviacion=sd(CosteFinalL))

## `summarise()` ungrouping output (override with ` `.groups` argument)

## # A tibble: 4 x 3
##   Clasificacion Media Desviacion
##   <fct>        <dbl>     <dbl>
## 1 Muy rápido    7.92      1.50
## 2 Rápido         8.07      1.51
## 3 Lento          8.01      1.54
## 4 Muy lento      8.00      1.63

#anova
results <- aov(CosteFinalL~Clasificacion, claimNet)
kk <- summary( results)
kk

##           Df Sum Sq Mean Sq F value Pr(>F)
## Clasificacion     3    190   63.29  27.09 <2e-16 ***
## Residuals       48671 113707    2.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
kk[[1]][1,4]

## [1] 27.09167

# Usando lm
results2<-lm(CosteFinalL~Clasificacion,data=claimNet)
anova(results2)

## Analysis of Variance Table
##
## Response: CosteFinalL
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Clasificacion     3    190   63.293  27.092 < 2.2e-16 ***
## Residuals       48671 113707    2.336
```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Valores del contraste: Sum Sq = 189.88; Mean Sq = 63.29; estadístico F = 27.09; pvalor = $1.6899999 \times 10^{-17}$. 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 cuatro grupos. **Clasificacion** tiene un efecto significativo sobre **CosteFinal** (p-valor menor que 0.05) y, por tanto, el tiempo de apertura del siniestro influye en el coste final del siniestro (en escala logarítmica).

5.3 Efectos de los niveles del factor

Calculad la variabilidad explicada por la variable **Clasificacion** sobre la variable **CosteFinal** mediante la métrica eta squared. Interpretad los resultados.

Solución:

```

## Calculamos la variabilidad explicada por la variable mediante la métrica eta squared.

```

```

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

```

```

## [1] 0.001667102

```

#alternativa

```

etaSquared(results)

```

```

##          eta.sq eta.sq.part
## Clasificacion 0.001667102 0.001667102

```

El efecto de los niveles es 0.0017, lo cual significa que el 0.17% de la variabilidad observada en **CosteFinal** se explica por el efecto de los niveles. Es decir, el 0.17% de la variabilidad total observada en **CosteFinal** se debe a la variabilidad observada entre las cuatro categorías. El resto (99.83%) es la parte no explicada por el modelo.

5.4 Contraste dos-a-dos

Como los factores han resultado significativos hay que hacer los contrastes de las comparaciones múltiples. Se puede utilizar la prueba de Tukey-Kramer que compara dos-a-dos las diferentes categorías de la variable. (Nota: por ejemplo, con la función **HSD.test()** del paquete **agricolae**).

```

HSD.test(results, "Clasificacion", unbalance=TRUE, console =TRUE)

```

```

##
## Study: results ~ "Clasificacion"
##
## HSD Test for CosteFinalL
##
## Mean Square Error: 2.336238
##
## Clasificacion, means
##
##          CosteFinalL      std      r      Min      Max
## Lento      8.014673 1.542670 13589 2.334284 14.16842
## Muy lento  7.996111 1.634199 4060 3.234919 13.52911
## Muy rápido 7.917190 1.503789 14789 1.746615 13.93241
## Rápido     8.072725 1.511404 16237 1.937128 14.26693
##
## Alpha: 0.05 ; DF Error: 48671

```

```

## Critical Value of Studentized Range: 3.63316
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##          CosteFinalL groups
## Rápido      8.072725    a
## Lento       8.014673    ab
## Muy lento   7.996111    b
## Muy rápido  7.917190    c

```

La categoría “Muy rápido” se diferencia del resto.

5.5 Adecuación del modelo

Mostrad la adecuación del modelo ANOVA. Se pide lo siguiente:

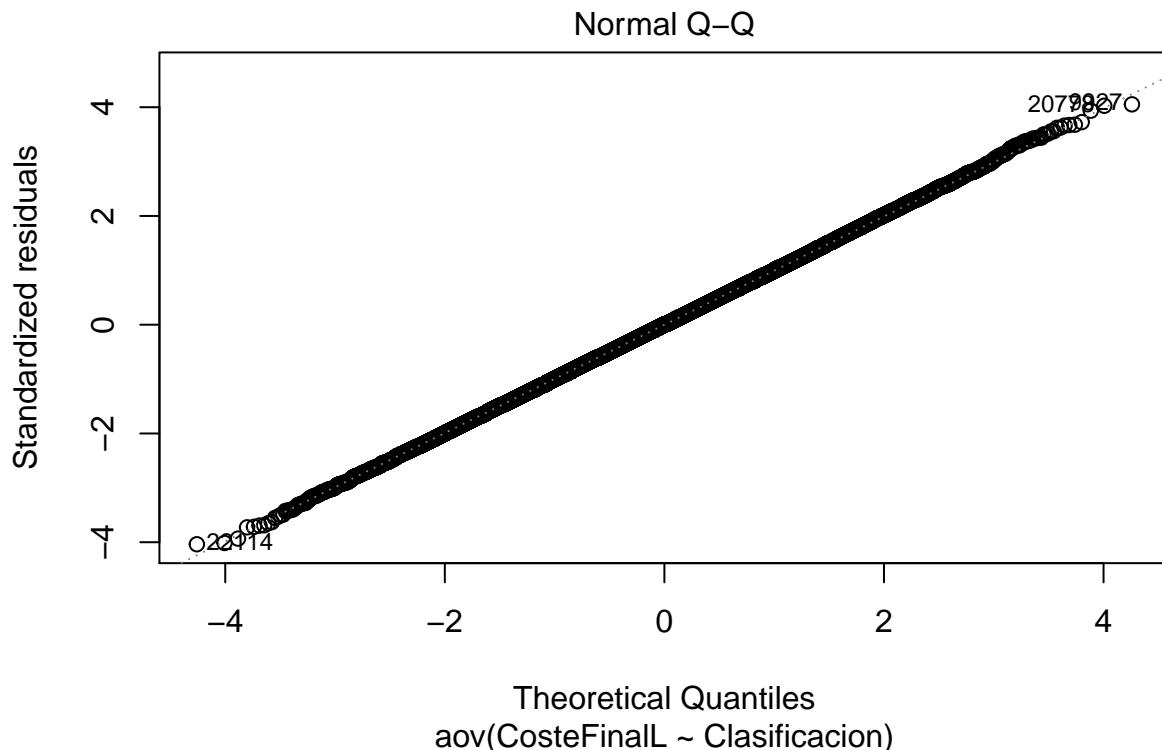
- Análisis visual de normalidad de los residuos. Podéis usar la función plot sobre el modelo ANOVA calculado.
- Análisis visual de homocedasticidad de los residuos. Podéis usar plot sobre el modelo ANOVA calculado.
- Contraste de normalidad y homocedasticidad.

5.5.1 Normalidad de los residuos

El análisis visual de la normalidad de los residuos se puede hacer a partir del gráfico Normal Q-Q. Mostrad e interpretad este gráfico.

Solución:

```
plot(results,which=2)
```

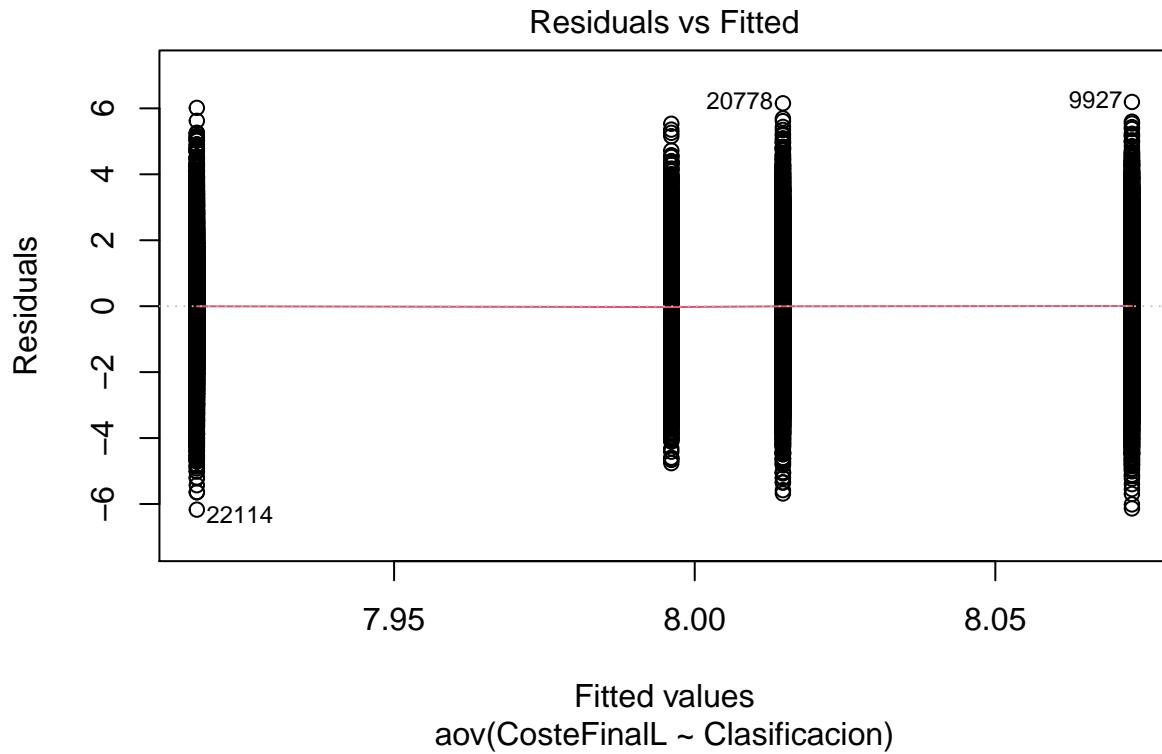


5.5.2 Homocedasticidad de los residuos

El gráfico “Residuals vs Fitted” proporciona información sobre la homocedasticidad de los residuos. Mostrad e interpretad este gráfico.

Solución:

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



5.5.3 Contraste de normalidad

Se puede comprobar el supuesto de normalidad de los residuos con las pruebas estadísticas de Shapiro-Wilk o Lilliefors, entre otros. El supuesto de homocedasticidad se puede comprobar a partir de la prueba de Bartlett.

Solución:

```
# contraste normalidad
# muestra muy grande para shapiro.test(residuals(results))
lillie.test(residuals(results))

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: residuals(results)
## D = 0.0018258, p-value = 0.959
#contraste de igualdad de varianzas

bartlett.test(CosteFinalL~Clasificacion, claimNet)

##
##  Bartlett test of homogeneity of variances
##
## data: CosteFinalL by Clasificacion
## Bartlett's K-squared = 52.145, df = 3, p-value = 2.789e-11
```

A nivel visual, parece que el supuesto de normalidad y homocedasticidad no se descartan. El contraste de normalidad nos lo confirma. En cambio, el supuesto de igualdad de varianzas no se puede aceptar al 95% de

nivel de confianza.

6 ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre `CosteFinal` en escala logarítmica según `Sexo` combinado con el factor `RiesgoSM`. Seguid los pasos que se indican a continuación.

6.1 Análisis de los efectos principales y posibles interacciones

Dibujad en un gráfico la variable `CosteFinal` en escala logarítmica en función de `Sexo` y en función de `RiesgoSM`. El gráfico debe permitir evaluar si hay interacción entre los dos factores. Por ello, se recomienda seguir estos pasos:

1. Agrupad el conjunto de datos por `Sexo` y por `RiesgoSM`. Calculad el número de casos disponibles de cada combinación de factores.
2. Calculad la media de coste (en log) para cada grupo.
3. Mostrad en un gráfico el valor medio de la variable `CosteFinal` en escala logarítmica para cada factor.
4. Interpretad el resultado sobre si sólo hay efectos principales o hay interacción entre los factores. Si hay interacción, explicad cómo se observa esta interacción en el gráfico.

Solución:

```
# Número de casos disponibles de cada combinación de factores.

claimNet %>% count(Sexo,RiesgoSM)

##   Sexo RiesgoSM     n
## 1   F   FALSE 11173
## 2   F    TRUE   84
## 3   M   FALSE 37286
## 4   M    TRUE   132

## Media para cada grupo según coste (escala logaritmica)

claimNetGrouped <- claimNet %>% group_by(Sexo, RiesgoSM) %>% summarise(MeanCoste=mean(CosteFinalL))

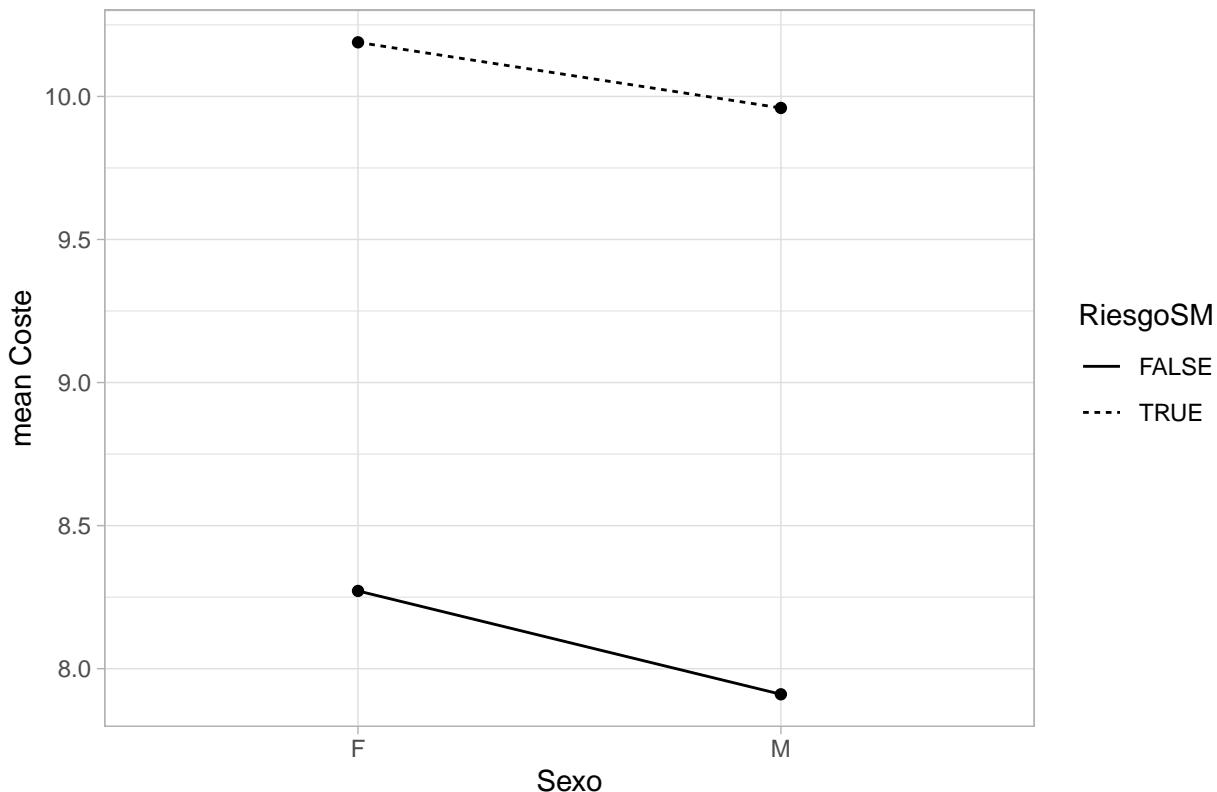
## `summarise()`` regrouping output by 'Sexo' (override with `.`groups` argument)
claimNetGrouped

## # A tibble: 4 x 3
## # Groups:   Sexo [2]
##   Sexo RiesgoSM MeanCoste
##   <fct> <fct>     <dbl>
## 1 F     FALSE      8.27
## 2 F     TRUE       10.2 
## 3 M     FALSE      7.91 
## 4 M     TRUE       9.96 

## Análisis visual de los efectos principales y posibles interacciones (escala logaritmica)

ggplot(claimNetGrouped, aes(x=Sexo, y=MeanCoste, group=RiesgoSM)) +
  geom_line(aes(linetype=RiesgoSM)) + geom_point() + ggtitle("Grafico de perfil ANOVA") +
  xlab("Sexo") + ylab("mean Coste") + theme_light()
```

Grafico de perfil ANOVA



Parece que no hay interacción entre los factores **Sexo** y **RiesgoSM** en relación a la variable **CosteFinal** (en escala logarítmica). Se observa paralelismo entre las dos líneas.

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

6.2 Cálculo del modelo

- Calculad el modelo incluyendo la interacción entre los factores.
- Medid el efecto de los factores sobre la variabilidad explicada del Coste final (en escala logarítmica).
- Analizad dos-a-dos las diferencias de medias entre los distintos factores.
- Adecuación del modelo. Realizar análisis visual de normalidad y homocedasticidad.

Solución:

```
#anova con interacción
results3 <- aov(CosteFinal ~ Sexo * RiesgoSM, claimNet)
summary(results3)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## Sexo                      1   1174   1173.9 510.771 <2e-16 ***
## RiesgoSM                   1     858    857.9 373.276 <2e-16 ***
## Sexo:RiesgoSM              1      1      0.9   0.386  0.534
## Residuals                 48671 111864      2.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

# efecto de los factores
etaSquared(results3)

##                                eta.sq   eta.sq.part
## Sexo          9.865646e-03 9.945024e-03
## RiesgoSM     7.532493e-03 7.611000e-03
## Sexo:RiesgoSM 7.791747e-06 7.933273e-06

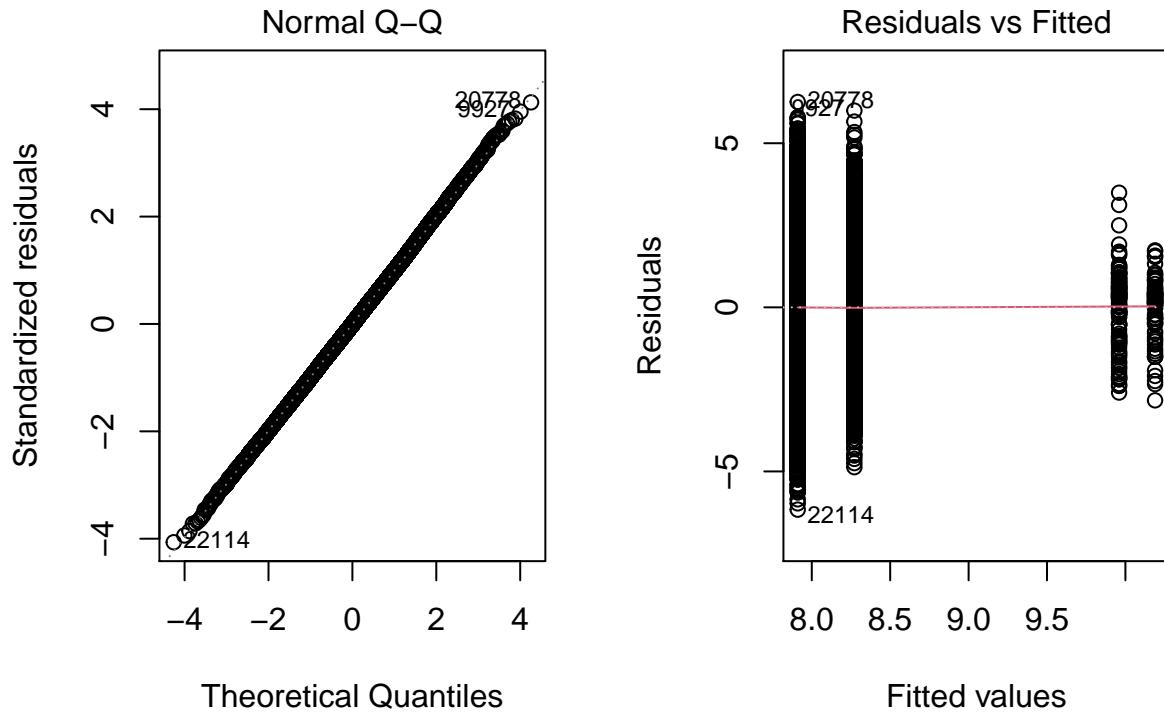
# Comparación múltiple

condition <- with(claimNet, interaction(Sexo, RiesgoSM))
results4 <- aov(CosteFinalL~condition, claimNet)
HSD.test(results4, "condition", unbalance=TRUE, console = TRUE)

##
## Study: results4 ~ "condition"
##
## HSD Test for CosteFinalL
##
## Mean Square Error:  2.298374
##
## condition,  means
##
##           CosteFinalL      std      r      Min      Max
## F.FALSE    8.271703 1.4123165 11173 3.401110 14.26693
## F.TRUE     10.189007 0.9333191   84 7.355029 11.92483
## M.FALSE    7.910462 1.5479010 37286 1.746615 14.16842
## M.TRUE     9.959641 1.1491186   132 7.363679 13.45198
##
## Alpha: 0.05 ; DF Error: 48671
## Critical Value of Studentized Range: 3.63316
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##           CosteFinalL groups
## F.TRUE     10.189007     a
## M.TRUE     9.959641     a
## F.FALSE    8.271703     b
## M.FALSE    7.910462     b

## Adecuación del modelo
par(mfrow=c(1,2))
plot(results4, which = 2)
plot(results4, which = 1)

```



6.3 Interpretación de los resultados

Los factores principales son significativos pero la interacción entre factores no es significativa. Por tanto, el coste en función de `RiesgoSM`, no es diferente según si es hombre o mujer.

7 Conclusiones

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

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

- La indemnización a las mujeres es en promedio superior a la de los hombres.
- El coste de las indemnizaciones no se distribuye según una distribución normal. Cuando realizamos la transformación logarítmica, no podemos rechazar que el coste en escala logarítmica se distribuye según una normal.
- La capacidad explicativa de los regresores sobre la variable dependiente es alta cuando la variable dependiente está expresada en escala logarítmica.
- Las variables ‘Edad’, ‘Dependientes’, ‘Salario’ y ‘CosteInicio’ tienen una relación positiva con el coste (en escala logarítmica). En cambio, el número de días trabajado tiene una relación negativa.
- Las variables ‘Sexo’, ‘Dependientes’, ‘Salario’, ‘Clasificacion’, ‘Jornada’, ‘DiasSemana’, ‘RiesgoSM’ y ‘CosteInicio’ inciden significativamente en la probabilidad de la valoración inicial del coste del siniestro sea insuficiente para cubrir el coste final.

- Las variables explicativas tienen una capacidad limitada para predecir la probabilidad que la valoración inicial del coste sea insuficiente para cubrir el coste final.
- Los factores ‘Sexo’ y ‘RiesgoSM’ tienen un efecto significativo sobre la media del ‘CosteFinal’ en escala logarítmica. La interacción, en cambio, no es significativa.

Puntuación de la actividad

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