

A4 - Análisis de varianza y repaso del curso

Pablo A. Delgado

18 de June, 2021

Contents

Introducción	2
1 Lectura del fichero y preparación de los datos	3
1.1 Preparación de los datos	8
1.2 Clasificación de tiempo	10
1.3 Valores ausentes	10
1.4 Salud mental	12
1.5 Análisis visual	12
1.6 Comprobación de normalidad	14
2 Estadística inferencial	17
2.1 Intervalo de confianza de la media poblacional de la variable CosteFinal	17
2.2 Contraste de hipótesis para la diferencia de medias	19
2.2.1 Escribid la hipótesis nula y la alternativa	19
2.2.2 Justificación del test a aplicar	19
2.2.3 Cálculos	19
2.2.4 Interpretación del test	21
3 Modelo de regresión lineal	21
3.1 Interpretación del modelo	22
3.2 Análisis residuos	22
3.3 Predicción	24
4 Regresión logística	25
4.1 Modelo predictivo	25
4.2 Interpretación	28
4.3 Matriz de confusión	29
4.4 Predicción	31

5 Análisis de la varianza (ANOVA) de un factor	32
5.1 Hipótesis nula y alternativa	32
5.2 Modelo	32
5.3 Efectos de los niveles del factor	33
5.4 Contraste dos-a-dos	34
5.5 Adecuación del modelo	35
6 ANOVA multifactorial	38
6.1 Análisis de los efectos principales y posibles interacciones	38
6.1.1. Agrupacion Conjunto de datos	38
6.1.2. Media por grupo	38
6.1.3. Visualizacion del valor medio	38
6.1.4. Interpretacion	40
6.2 Cálculo del modelo	40
6.3 Interpretación de los resultados	46
7 Conclusiones	46
Recursos	47

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 (apertura).
- 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.

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.

Como primer step instalamos y cargamos todas las librerías que utilizaremos.

```
packages <- c("ggplot2", "dplyr", "gridExtra", "ResourceSelection",
            "eeprotools", "pROC", "data.table", "nortest", "caret",
            "gmodels", "lsr", "agricolae", "doBy")
new <- packages[!(packages %in% installed.packages()[, "Package"])]
if(length(new)) install.packages(new)
foo=lapply(packages, require, character.only=TRUE)
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

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

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

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 4.0.5

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
## 
##     combine
```

```
## Loading required package: ResourceSelection

## Warning: package 'ResourceSelection' was built under R version 4.0.5

## ResourceSelection 0.3-5 2019-07-22

## Loading required package: eeptools

## Warning: package 'eeptools' was built under R version 4.0.5

## Loading required package: pROC

## Warning: package 'pROC' was built under R version 4.0.5

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
## cov, smooth, var

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 4.0.5

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
## between, first, last

## Loading required package: nortest

## Warning: package 'nortest' was built under R version 4.0.3

## Loading required package: caret

## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: gmodels

## Warning: package 'gmodels' was built under R version 4.0.5
```

```

## 
## Attaching package: 'gmodels'

## The following object is masked from 'package:pROC':
## 
##     ci

## Loading required package: lsr

## Warning: package 'lsr' was built under R version 4.0.3

## Loading required package: agricolae

## Warning: package 'agricolae' was built under R version 4.0.5

## Loading required package: doBy

## Warning: package 'doBy' was built under R version 4.0.5

## 
## Attaching package: 'doBy'

## The following object is masked from 'package:dplyr':
## 
##     order_by

```

Carguemos el fichero y realicemos una primera inspección del dataframe.

```

claim = read.csv('trainCLEAN.csv', header=TRUE, sep=',')
str(claim)

```

```

## 'data.frame':      54000 obs. of  15 variables:
## $ ClaimNumber          : chr  "WC8285054" "WC6982224" "WC5481426" "WC9775968" ...
## $ DateTimeOfAccident   : chr  "2002-04-09T07:00:00Z" "1999-01-07T11:00:00Z" "1996-
03-25T00:00:00Z" "2005-06-22T13:00:00Z" ...
## $ DateReported         : chr  "2002-07-05T00:00:00Z" "1999-01-20T00:00:00Z" "1996-
04-14T00:00:00Z" "2005-07-22T00:00:00Z" ...
## $ Age                  : int  48 43 30 41 36 50 39 56 49 30 ...
## $ Gender                : chr  "M" "F" "M" "M" ...
## $ MaritalStatus         : chr  "M" "M" "U" "S" ...
## $ 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      : chr  "F" "F" "F" "F" ...
## $ HoursWorkedPerWeek    : num  38 37.5 38 38 38 38 40 38 37 ...
## $ DaysWorkedPerWeek     : int  5 5 5 5 5 5 5 5 5 5 ...
## $ ClaimDescription       : chr  "LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY" "STEPPED AROUND
A DOG" ...
## $ InitialIncurredClaimsCost: int  1500 5500 1700 15000 2800 500 500 500 925 1500 ...
## $ UltimateIncurredClaimCost: num  4303 6106 2099 16283 3772 ...

```

```
summary(claim)
```

```
##   ClaimNumber      DateTimeOfAccident DateReported      Age
##   Length:54000      Length:54000      Length:54000      Min.   :13.00
##   Class :character  Class :character  Class :character  1st Qu.:23.00
##   Mode  :character  Mode  :character  Mode  :character  Median  :32.00
##                               Mean   :33.84
##                               3rd Qu.:43.00
##                               Max.   :81.00
##   Gender          MaritalStatus DependentChildren DependentsOther
##   Length:54000      Length:54000      Min.   :0.00000  Min.   :0.000000
##   Class :character  Class :character  1st Qu.:0.00000  1st Qu.:0.000000
##   Mode  :character  Mode  :character  Median  :0.00000  Median  :0.000000
##                               Mean   :0.1192   Mean   :0.009944
##                               3rd Qu.:0.00000 3rd Qu.:0.000000
##                               Max.   :9.00000  Max.   :5.000000
##   WeeklyWages     PartTimeFullTime HoursWorkedPerWeek DaysWorkedPerWeek
##   Min.   : 1.0    Length:54000      Min.   : 0.00    Min.   :1.000
##   1st Qu.: 200.0   Class :character  1st Qu.: 38.00   1st Qu.:5.000
##   Median  : 392.2   Mode  :character  Median  : 38.00   Median  :5.000
##   Mean    : 416.4                    Mean   : 37.74   Mean   :4.906
##   3rd Qu.: 500.0                    3rd Qu.: 40.00   3rd Qu.:5.000
##   Max.   :7497.0                    Max.   :640.00   Max.   :7.000
##   ClaimDescription InitialIncurredCalimsCost UltimateIncurredClaimCost
##   Length:54000      Min.   :     1      Min.   : 5.7
##   Class :character  1st Qu.: 700      1st Qu.:1128.7
##   Mode  :character  Median  : 2000     Median  :3179.1
##                               Mean   : 7841     Mean   :10195.5
##                               3rd Qu.: 9500     3rd Qu.: 8900.1
##                               Max.   :2000000   Max.   :1570535.8
```

```
head(claim)
```

```
##   ClaimNumber      DateTimeOfAccident      DateReported Age Gender
##   1   WC8285054 2002-04-09T07:00:00Z 2002-07-05T00:00:00Z 48   M
##   2   WC6982224 1999-01-07T11:00:00Z 1999-01-20T00:00:00Z 43   F
##   3   WC5481426 1996-03-25T00:00:00Z 1996-04-14T00:00:00Z 30   M
##   4   WC9775968 2005-06-22T13:00:00Z 2005-07-22T00:00:00Z 41   M
##   5   WC2634037 1990-08-29T08:00:00Z 1990-09-27T00:00:00Z 36   M
##   6   WC6828422 1999-06-21T11:00:00Z 1999-09-09T00:00:00Z 50   M
##   MaritalStatus DependentChildren DependentsOther WeeklyWages PartTimeFullTime
##   1           M                 0                  0    500.00        F
##   2           M                 0                  0    509.34        F
##   3           U                 0                  0    709.10        F
##   4           S                 0                  0    555.46        F
##   5           M                 0                  0    377.10        F
##   6           M                 0                  0    200.00        F
##   HoursWorkedPerWeek DaysWorkedPerWeek
##   1                   38.0                5
##   2                   37.5                5
##   3                   38.0                5
##   4                   38.0                5
##   5                   38.0                5
```

```

## 6          38.0          5
##                                         ClaimDescription
## 1          LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2          STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
## 3          CUT ON SHARP EDGE CUT LEFT THUMB
## 4          DIGGING LOWER BACK LOWER BACK STRAIN
## 5          REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
## 6          STRUCK HEAD ON HEAD LACERATED HEAD
##   InitialIncurredCalimsCost UltimateIncurredClaimCost
## 1          1500          4303.1880
## 2          5500          6105.8729
## 3          1700          2098.6300
## 4          15000         16282.9408
## 5          2800          3771.7326
## 6          500           746.6213

```

```
tail(claim)
```

```

##      ClaimNumber  DateTimeOfAccident      DateReported Age Gender
## 53995  WC7006507  1999-07-19T11:00:00Z 2001-03-04T00:00:00Z 35   M
## 53996  WC9370727  2004-08-21T18:00:00Z 2004-09-08T00:00:00Z 32   F
## 53997  WC8396269  2002-04-28T09:00:00Z 2002-09-03T00:00:00Z 20   F
## 53998  WC3609528  1992-02-28T09:00:00Z 1992-03-18T00:00:00Z 19   M
## 53999  WC5038565  1995-01-10T07:00:00Z 1995-01-31T00:00:00Z 24   M
## 54000  WC2542601  1990-10-24T14:00:00Z 1990-11-03T00:00:00Z 22   M
##   MaritalStatus DependentChildren DependentsOther WeeklyWages
## 53995        M            0            0           200
## 53996        S            0            0           500
## 53997        S            0            0           500
## 53998        S            0            0           283
## 53999        S            0            0           200
## 54000        S            0            0           200
##   PartTimeFullTime HoursWorkedPerWeek DaysWorkedPerWeek
## 53995        F             40            5
## 53996        F             38            5
## 53997        F             40            5
## 53998        F             40            5
## 53999        F             38            5
## 54000        F             38            5
##                                         ClaimDescription
## 53995          FELL STAIRS BRUISE RIGHT ANKLE AND RIGHT LEG
## 53996          STRUCK KNIFE LACERATED LEFT MIDDLE FINGER LEFT HAND
## 53997          LEFT HAND LACERATION LEFT SIDE BACK AND LEFT LEG
## 53998          METAL SLIPPED ACROSS METAL CUT FINGER
## 53999          BURN WHILST USING SPANNER LACERATION RIGHT MIDDLE FINGER
## 54000          CUT WITH BREAD KNIFE LACERATION LEFT INDEX AND MIDDLE FINGERS
##   InitialIncurredCalimsCost UltimateIncurredClaimCost
## 53995        15000          12767.7593
## 53996        1000           518.9856
## 53997        1000           1005.4221
## 53998        210            488.0741
## 53999        7500           2630.2124
## 54000        550            1090.5865

```

1.1 Preparación de los datos

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

- 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.

```
# vemos los nombres actuales
names(claim)
```

```
## [1] "ClaimNumber"           "DateTimeOfAccident"
## [3] "DateReported"          "Age"
## [5] "Gender"                 "MaritalStatus"
## [7] "DependentChildren"      "DependentsOther"
## [9] "WeeklyWages"            "PartTimeFullTime"
## [11] "HoursWorkedPerWeek"    "DaysWorkedPerWeek"
## [13] "ClaimDescription"       "InitialIncurredCalimsCost"
## [15] "UltimateIncurredClaimCost"
```

```
# traducimos los nombres, teniendo en cuenta el orden correcto de columnas
names(claim) = c ('Id', 'Ocurrencia', 'Apertura', 'Edad', 'Sexo', 'Estado',
                  'Dependientes', 'OtrosDepend', 'Salario', 'Jornada',
                  'HorasSemana', 'DiasSemana', 'Descripcion', 'CosteInicio',
                  'CosteFinal')

# Ahora vamos a convertir a fecha los campos de este tipo, pero primero
# verifiquemos que sea posible la conversion:
which(is.na(as.Date(claim$Ocurrencia, "%Y-%m-%d")))
```

```
## integer(0)
```

```
which(is.na(as.Date(claim$Apertura, "%Y-%m-%d")))
```

```
## integer(0)
```

```
# dado que es posible convertir sin problemas, hagamos efectiva la
# transformacion
claim$Ocurrencia = as.Date(claim$Ocurrencia, "%Y-%m-%d")
claim$Apertura = as.Date(claim$Apertura, "%Y-%m-%d")

# creamos la variable tiempo
claim$tiempo = as.integer(age_calc(claim$Ocurrencia,
                                    enddate = claim$Apertura, units = "days",
                                    precise = TRUE))

# vemos los resultados de las transformaciones
names(claim)
```

```

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

```

```
str(claim)
```

```

## 'data.frame': 54000 obs. of 16 variables:
## $ Id       : chr  "WC8285054" "WC6982224" "WC5481426" "WC9775968" ...
## $ Ocurrencia : Date, format: "2002-04-09" "1999-01-07" ...
## $ Apertura  : Date, format: "2002-07-05" "1999-01-20" ...
## $ Edad      : int  48 43 30 41 36 50 39 56 49 30 ...
## $ Sexo      : chr  "M" "F" "M" "M" ...
## $ Estado     : chr  "M" "M" "U" "S" ...
## $ Dependientes: int  0 0 0 0 0 0 0 0 0 ...
## $ OtrosDepend : int  0 0 0 0 0 0 0 0 0 ...
## $ Salario    : num  500 509 709 555 377 ...
## $ Jornada    : chr  "F" "F" "F" "F" ...
## $ HorasSemana: num  38 37.5 38 38 38 38 38 40 38 37 ...
## $ DiasSemana : int  5 5 5 5 5 5 5 5 5 ...
## $ Descripcion: chr  "LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY" "STEPPED AROUND CRATES AND"
## $ CosteInicio : int  1500 5500 1700 15000 2800 500 500 500 925 1500 ...
## $ CosteFinal  : num  4303 6106 2099 16283 3772 ...
## $ tiempo     : int  87 13 20 30 29 80 10 38 33 15 ...

```

```
head(claim)
```

```

##          Id Ocurrencia Apertura Edad Sexo Estado Dependientes OtrosDepend
## 1 WC8285054 2002-04-09 2002-07-05  48   M     M         0         0
## 2 WC6982224 1999-01-07 1999-01-20  43   F     M         0         0
## 3 WC5481426 1996-03-25 1996-04-14  30   M     U         0         0
## 4 WC9775968 2005-06-22 2005-07-22  41   M     S         0         0
## 5 WC2634037 1990-08-29 1990-09-27  36   M     M         0         0
## 6 WC6828422 1999-06-21 1999-09-09  50   M     M         0         0
##          Salario Jornada HorasSemana DiasSemana
## 1      500.00     F       38.0        5
## 2      509.34     F       37.5        5
## 3      709.10     F       38.0        5
## 4      555.46     F       38.0        5
## 5      377.10     F       38.0        5
## 6     200.00     F       38.0        5
##                                     Descripcion
## 1             LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY
## 2             STEPPED AROUND CRATES AND TRUCK TRAY FRACTURE LEFT FOREARM
## 3             CUT ON SHARP EDGE CUT LEFT THUMB
## 4             DIGGING LOWER BACK LOWER BACK STRAIN
## 5 REACHING ABOVE SHOULDER LEVEL ACUTE MUSCLE STRAIN LEFT SIDE OF STOMACH
## 6             STRUCK HEAD ON HEAD LACERATED HEAD
##          CosteInicio CosteFinal tiempo
## 1          1500     4303.1880     87
## 2          5500     6105.8729     13
## 3          1700     2098.6300     20
## 4         15000    16282.9408    30

```

```

## 5      2800  3771.7326    29
## 6      500   746.6213    80

```

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.

Pasemos a crear la nueva variable:

```

claim = claim %>% mutate(Clasificacion = case_when
  (tiempo <= 15 ~ 'Muy rápido',
  tiempo %in% c(16:30) ~ 'Rápido',
  tiempo %in% c(31:89) ~ 'Lento',
  tiempo >= 90 ~ '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?

```

# hay 28114 descripciones distintas
length(table(claim$Descripcion))

```

```

## [1] 28114

```

```

# mientras que de sexo y estado tenemos estas cantidades:
table(claim$Sexo)

```

```

##
##      F      M      U
## 12338 41660      2

```

```

table(claim$Estado)

```

```

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

```

```

# o sea tenemos estas cantidades distintas de categorias para esas variables:
length(table(claim$Sexo))

```

```

## [1] 3

```

```

length(table(claim$Estado))

```

```
## [1] 4
```

Como vemos arriba hay valores U y vacios en las variables sexo y estado.

- Representad los observaciones con la categoría “U” (U=unknown) en las variables ‘Sexo’y ‘Estado’ como missings.

```
claim[claim$Sexo == 'U', 'Sexo'] = NA  
claim[claim$Estado %in% c(' ', 'U'), 'Estado'] = NA
```

Veamos el resultado de la transformacion:

```
claim$Sexo = factor(claim$Sexo)  
table(claim$Sexo, exclude = FALSE)
```

```
##  
##      F      M  <NA>  
## 12338 41660      2
```

```
table(claim$Estado, exclude = FALSE)
```

```
##  
##      M      S  <NA>  
## 22516 26161  5323
```

- 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.

```
# Estadísticas de valores vacíos  
colSums(is.na(claim))
```

```
##          Id    Ocurrencia     Apertura       Edad       Sexo  
##          0        0          0        0        2  
## Estado Dependientes OtrosDepend     Salario     Jornada  
## 5323        0          0        0        0        0  
## HorasSemana DiasSemana Descripcion CosteInicio CosteFinal  
##          0        0          0        0        0  
## tiempo Clasificacion  
##          0        0
```

Solo tenemos NA en las variables que convertimos antes. No tiene sentido mantener observaciones sin saber el sexo de la persona ni tampoco el estado civil, preferimos tener todos los datos completos para crear un modelo lo mas preciso posible, y no queremos vernos afectados por missing values. Y dado que representan solo un 9.8% del total de las observaciones no vemos problema en eliminarlos. Ya que todavía contaremos con 48675 observaciones.

- Eliminad los valores ausentes del conjunto de datos. Denominamos al conjunto de datos claimNet.

```

claimNet = claim[-which(is.na(claim$Sexo) | is.na(claim$Estado)), ]
str(claimNet)

## 'data.frame': 48675 obs. of 17 variables:
## $ Id      : chr "WC8285054" "WC6982224" "WC9775968" "WC2634037" ...
## $ Ocurrencia : Date, format: "2002-04-09" "1999-01-07" ...
## $ Apertura : Date, format: "2002-07-05" "1999-01-20" ...
## $ Edad     : int 48 43 41 36 50 39 56 49 30 20 ...
## $ Sexo     : Factor w/ 2 levels "F","M": 2 1 2 2 2 2 2 2 2 2 ...
## $ Estado   : chr "M" "M" "S" "M" ...
## $ Dependientes : int 0 0 0 0 0 0 0 0 0 0 ...
## $ OtrosDepend : int 0 0 0 0 0 0 0 0 0 0 ...
## $ Salario   : num 500 509 555 377 200 ...
## $ Jornada   : chr "F" "F" "F" "F" ...
## $ HorasSemana : num 38 37.5 38 38 38 38 40 38 37 40 ...
## $ DiasSemana : int 5 5 5 5 5 5 5 5 5 5 ...
## $ Descripcion : chr "LIFTING TYRE INJURY TO RIGHT ARM AND WRIST INJURY" "STEPPED AROUND CRATES AND PALLETS" ...
## $ CosteInicio : int 1500 5500 15000 2800 500 500 500 925 1500 3500 ...
## $ CosteFinal  : num 4303 6106 16283 3772 747 ...
## $ tiempo    : int 87 13 30 29 80 10 38 33 15 11 ...
## $ Clasificacion: chr "Lento" "Muy rápido" "Rápido" "Rápido" ...

```

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.

```

condition = (tolower(claimNet$Descripcion) %like% 'stress' |
             tolower(claimNet$Descripcion) %like% 'anxiety' |
             tolower(claimNet$Descripcion) %like% 'harassment' |
             tolower(claimNet$Descripcion) %like% 'depression'
           )
claimNet$RiesgoSM = ifelse(condition, 1, 0)

table (claimNet$RiesgoSM)

##
##      0      1
## 48459   216

```

Como vemos solo 216 observaciones cumplen la condicion.

1.5 Análisis visual

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

```

a= ggplot(claimNet,aes(x=log(CosteFinal),fill=Sexo)) +
  geom_boxplot() +
  theme_bw()

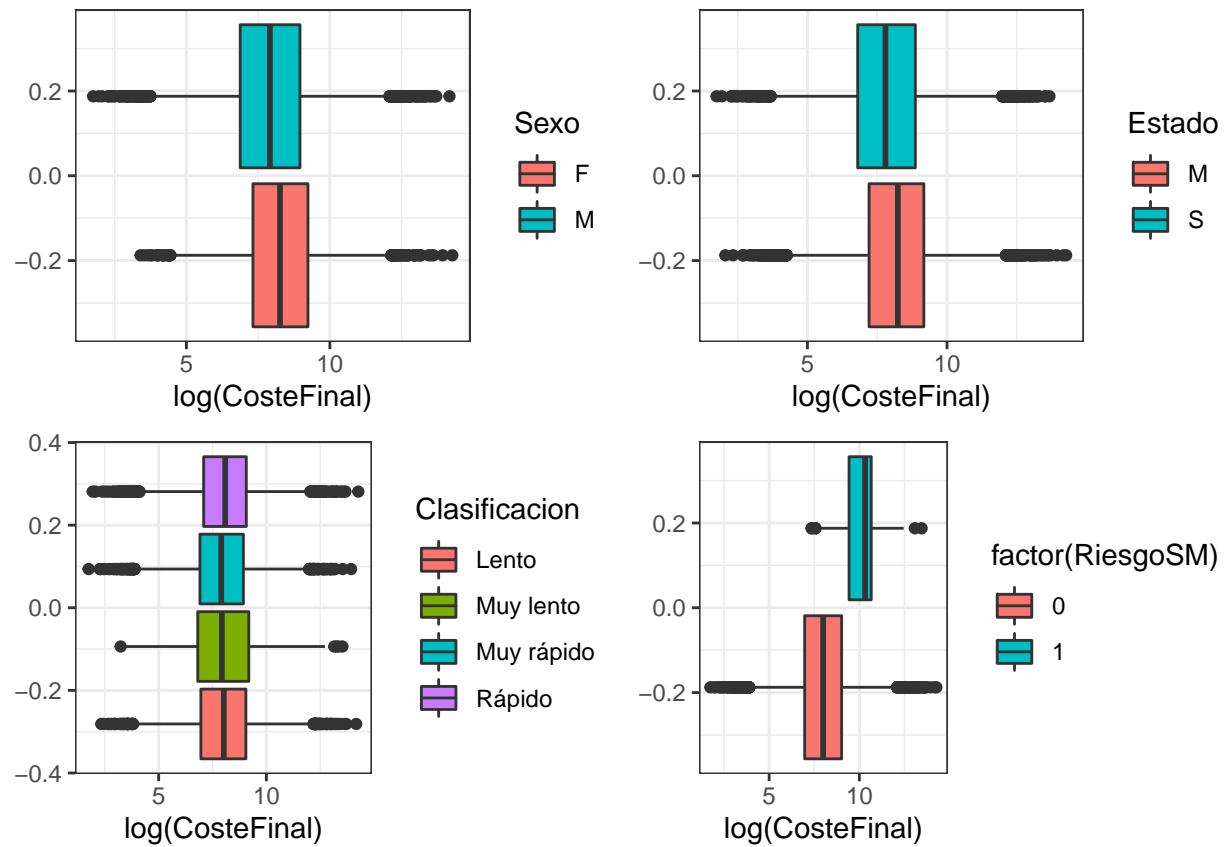
b=ggplot(claimNet,aes(x=log(CosteFinal),fill=Estado)) +
  geom_boxplot() +
  theme_bw()

c=ggplot(claimNet,aes(x=log(CosteFinal),fill=Clasificacion)) +
  geom_boxplot() +
  theme_bw()

d=ggplot(claimNet,aes(x=log(CosteFinal),fill=factor(RiesgoSM))) +
  geom_boxplot() +
  theme_bw()

grid.arrange(a,b, c,d, nrow=2, ncol=2)

```



Como puede visualizarse la distribucion e IQR del logaritmo de Coste final para Sexo, Estado y Clasificacion son similares, y en todos los casos se ven outliers en ambos extremos de la distribucion. Se observa una distribucion distinta para el nivel de riesgo respecto a las demas variables e incluso la distribucion entre los grupos con y sin riesgo es bastante diferente, de hecho la mediana de los con riesgo tiene una tendencia hacia la derecha de la distribucion.

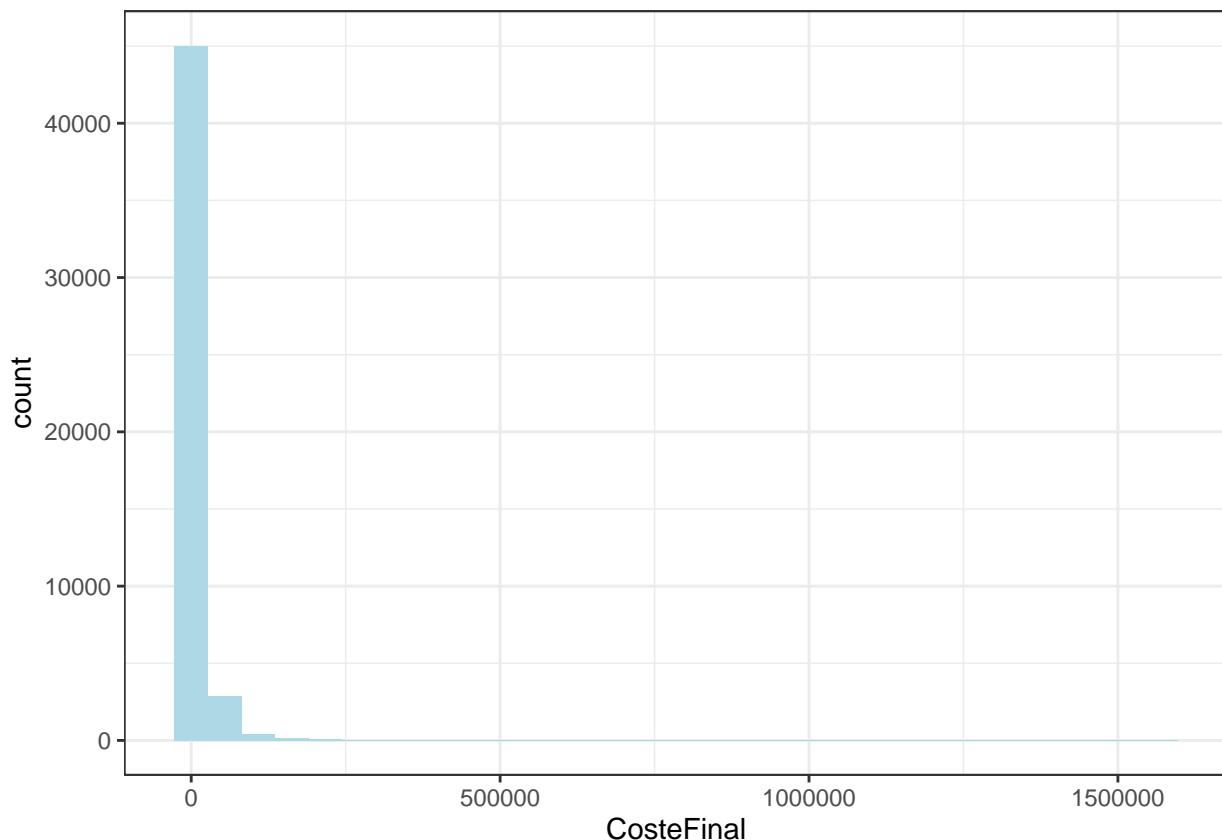
1.6 Comprobación de normalidad

¿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.

```
ggplot(claimNet,aes(x=CosteFinal)) +  
  geom_histogram(fill="lightblue") +  
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

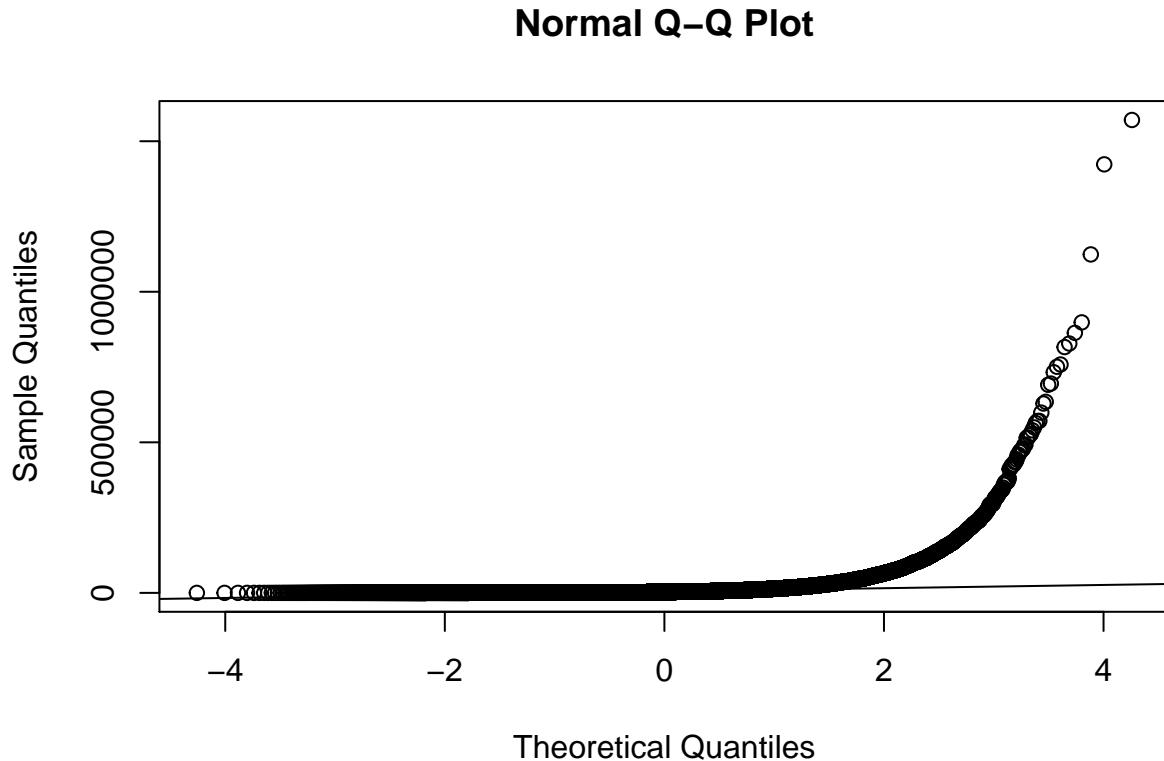


- Realizad contraste de normalidad de Lilliefors (p.ej. con función lillie.test de la librería nortest).

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

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

```
qqnorm(claimNet$CosteFinal)
qqline(claimNet$CosteFinal)
```

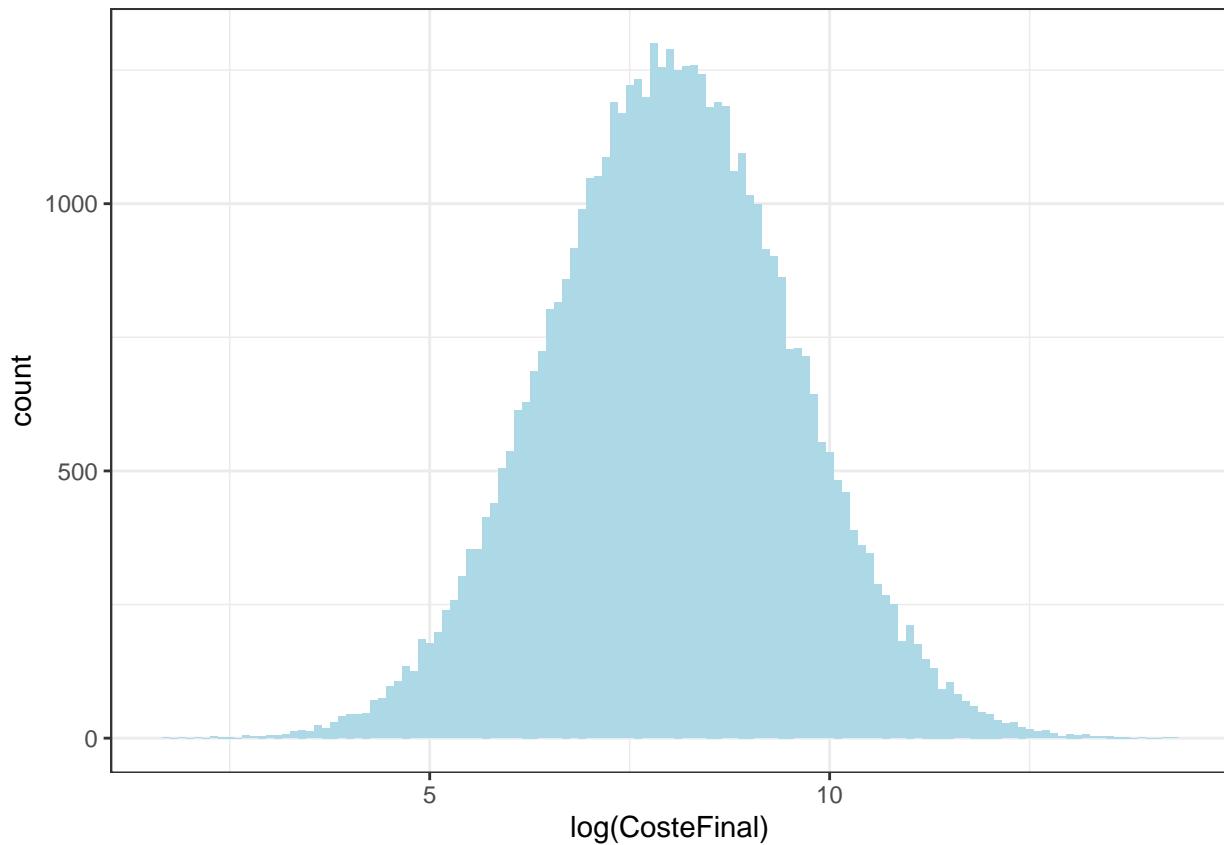


Como vemos tanto a nivel grafico como con el p-value < 2.2e-16 se puede afirmar el rechazo de la hipotesis nula de normalidad, osea, no tenemos distribucion normal para CosteFinal.

Pero veamos que pasa cuando aplicamos el logaritmo a nuestra variable objetivo:

- Realizad inspección visual y contraste de normalidad a la variable Coste Final en escala logarítmica.

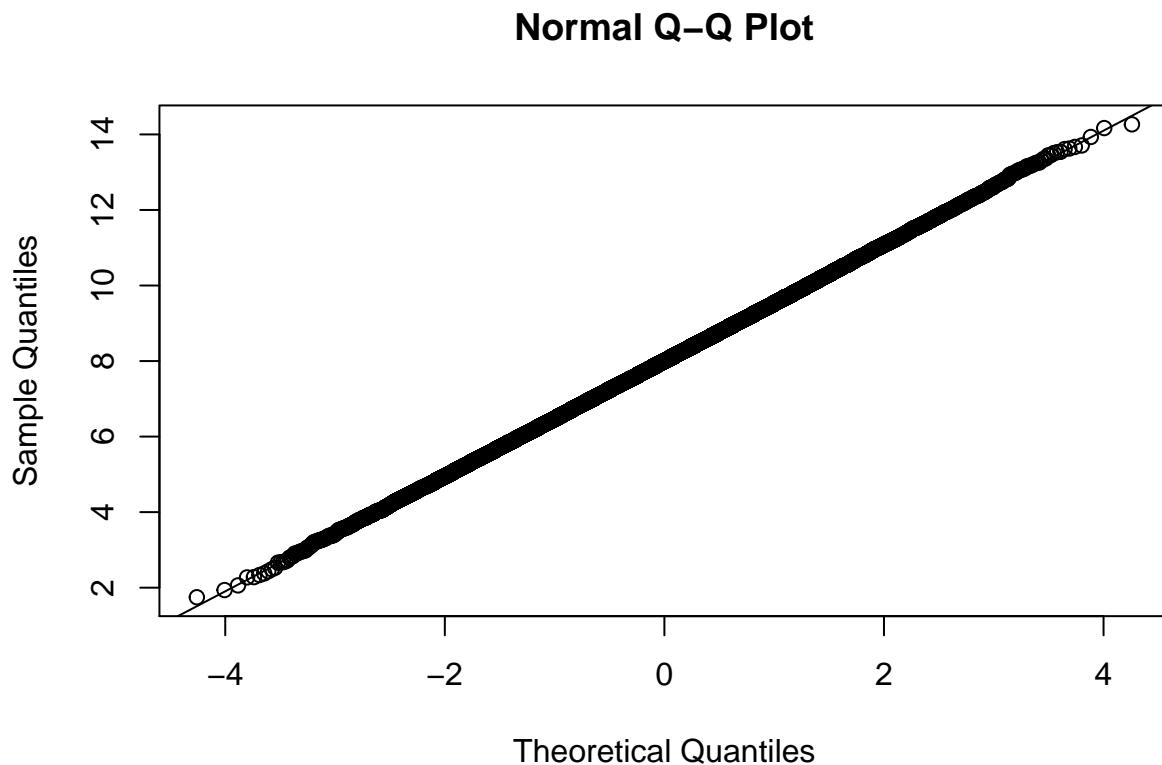
```
ggplot(claimNet,aes(x=log(CosteFinal))) +
  geom_histogram(fill="lightblue", binwidth=0.1) +
  theme_bw()
```



```
lillie.test(log(claimNet$CosteFinal))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(claimNet$CosteFinal)  
## D = 0.0018307, p-value = 0.9579
```

```
qqnorm(log(claimNet$CosteFinal))  
qqline(log(claimNet$CosteFinal))
```



Esta vez tanto con las graficas QQ como con el contraste de normalidad obteniendo un p-value mayor a 0.05 podemos decir que nuestros datos siguen una distribución normal luego de aplicar el logaritmo.

2 Estadística inferencial

Utilicemos a partir de aqui 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).

Asumiendo un intervalo de confianza del 95% y siendo que no tenemos previamente calculada la varianza de la poblacion y debemos estimarla a partir de la desviacion de la muestra, nuestra variable sigue de esta forma una distribucion t de Student con $n-1$ grados de libertad.

Por lo que calcularemos el intervalo de confianza para la media de la variable CosteFinal cuando la varianza es desconocida previamente siguiendo estos calculos.

Nota: Dado que el intervalo de confianza a calcular es sobre la media, y apoyandonos en el TLC aplicaremos el estadistico mencionado en el parrafo anterior (mas alla que sepamos por el apartado 1.6 que esta variable en escala normal no esta normalmente distribuida).

```

alfa <- 1-0.95
sd <- sd(claimNet$CosteFinal)
n <- nrow(claimNet)
SE <- sd / sqrt(n)
# Para obtener las probabilidades o cuantiles de la distribución t
# se usan las funciones pt y qt de R, que son análogas a las
# funciones pnorm y qnorm de la distribución normal
z <- qt( alfa/2, df=n-1, lower.tail=FALSE )
peso_medio = mean(claimNet$CosteFinal)
L <- peso_medio - z*SE
U <- peso_medio + z*SE

data.frame(L, peso_medio, U, z )

```

```

##           L peso_medio      U      z
## 1 9450.34    9704.58 9958.82 1.960013

```

Por lo tanto, el intervalo de confianza del 95% del CosteFinal es: [9450.34 , 9958.82].

La función t.test de R realiza este cálculo automáticamente. Veámoslo:

```
t.test( claimNet$CosteFinal, conf.level = 0.95)
```

```

##
##  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

```

Dado que no se conocía la varianza de la población y se ha estimado a partir de la muestra, hemos usado la distribución t de Student en lugar de la distribución normal. La consecuencia de esto es que el intervalo de confianza calculado con la distribución t es más ancho que el equivalente calculado con distribución normal. Por lo tanto, para un mismo nivel de confianza, hay más incertidumbre en el valor del parámetro de la población cuando se usa la distribución t. Pero tambien es cierto que, para tamaños de muestra grandes, la distribución t de Student se aproxima a la distribución normal.

De hecho solo a modo de chequeo si hubiesemos aplicado qnorm, vemos que el intervalo de confianza es casi identico, ya que como mencionamos para muestras grandes t student tiende a una distribucion normal.

```

z_qnorm <- qnorm( alfa/2, lower.tail=FALSE )
L_qnorm <- peso_medio - z_qnorm*SE
U_qnorm <- peso_medio + z_qnorm*SE
data.frame(L_qnorm, peso_medio, U_qnorm, z_qnorm )

```

```

##      L_qnorm peso_medio  U_qnorm  z_qnorm
## 1 9450.346    9704.58 9958.814 1.959964

```

2.2 Contraste de hipótesis para la diferencia de medias

En este caso queremos verificar la siguiente pregunta de investigacion:

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

2.2.1 Escribid la hipótesis nula y la alternativa

Planteemos entonces la hipótesis nula y alternativa para esta pregunta:

H0: Indemnización Mujeres \leq (Indemnización Hombres + 1000)

H1: Indemnización Mujeres $>$ (Indemnización Hombres + 1000)

Entonces segun lo que comprobemos en las siguientes secciones podremos llegar a decir:

- Rechazo la H0 a favor de la H1 y por tanto si que hay evidencias que demuestran que la media de indemnización de las mujeres supera en mas de 1000 euros a las de los hombres- Y por tanto la respuesta a la pregunta de investigacion es SI, con el 95 de confianza

o

- No hay evidencia que permita rechazar la hipótesis nula por lo que no puede afirmarse que la media de indemnización de las mujeres sea mayor en mas de 1000 euros que la de los hombres para la muestra seleccionada.

2.2.2 Justificación del test a aplicar

Dicho eso aplicaremos un contraste de dos muestras independientes, la muestra de mujeres y la de hombres hombres no tiene relacion entre ellos, con media y varianzas desconocidas.

Se trata de un test parametrico, porque contamos con distribuciones donde podemos asumir la normalidad de datos, porque aplica el Teorema del Limite Central al tener muestras de mas de 30 elementos y porque estamos aplicando contrastes de hipótesis sobre las medias de las muestras.

Y dada las hipótesis planteadas se aplicara un test unilateral por la derecha, con un nivel de confianza del 95%, por lo que el nivel de significancia (α) sera de 0.05.

2.2.3 Cálculos

Dados que asumimos que no se conocen las varianzas de la población y, por lo tanto, hay que estimarlas a partir de las muestras, para aplicar el estadístico adecuado, hay que comprobar si las varianzas de las dos poblaciones son iguales o diferentes. Para ello, aplicamos primero el test de igualdad de varianzas, tambien denominado test de homoscedasticidad.

```
alfa <- 0.05
mujeres <- claimNet[claimNet$Sexo=='F',]$CosteFinal
hombres <- claimNet[claimNet$Sexo=='M',]$CosteFinal
mean1 <- mean(mujeres); n1 <- length(mujeres); s1 <- sd(mujeres)
mean2 <- mean(hombres); n2 <- length(hombres); s2 <- sd(hombres)

fobs<-s1^2 / s2^2
fcritL <- qf(alfa/2, df1=n1-1, df2=n2-2 )
fcritU <- qf(1-alfa/2, df1=n1-1, df2=n2-2)
```

```

pvalue <- min(pf( fobs, df1=n1-1, df2=n2-2, lower.tail=FALSE ),
               pf( fobs, df1=n1-1, df2=n2-2)) *2
data.frame(fobs, fcritL, fcritU, pvalue)

```

```

##      fobs    fcritL    fcritU      pvalue
## 1 1.575073 0.9705234 1.030123 1.303136e-211

```

Como se puede observar, las funciones qf y pf devuelven el cuantil y la probabilidad de la distribución F respectivamente.

Como vemos aqui el valor observado cae fuera de la zona de aceptacion de la hipotesis nula, por ende se puede rechazar la hipotesis nula ($H_0 = \text{igualdad de varianzas}$). Sucede lo mismo si analizamos el valor p , basandonos en el, tambien podemos rechazar H_0 . Todo esto implica que estamos ante varianzas distintas.

Tambien se puede aplicar simplemente el `var.test` para evaluar estos conceptos o sea si podemos asumir varianzas parecidas o no. y segun eso aplicar una formula u otra.

En R, la función `var.test` calcula el test de igualdad de varianzas:

```
var.test(mujeres, hombres, conf.level = 0.95 )
```

```

##
##  F test to compare two variances
##
## data: mujeres and hombres
## 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

```

Una vez que podemos asumir que las varianzas son distintas, aplicaremos el metodo correspondiente a la media de dos poblaciones independientes con varianza desconocida distinta para obtener: el valor critico y el p -value.

Hay dos formas de rechazar la H_0 para la pregunta de investigacion, con el valor critico o evaluando valor P (que es el error que estaria cometiendo) contra α osea si valor $P < \alpha$, osea si el error es menor que el error maximo permitido, si esto se cumple se puede rechazar la H_0 .

Dada este breve explicacion calculemos cada uno de los valores y evaluemos los resultados:

```

mean1 <- mean(mujeres); n1 <- length(mujeres); s1 <- sd(mujeres)
mean2 <- mean(hombres)+1000; n2 <- length(hombres); s2 <- sd(hombres)

tobs <- (mean1-mean2) / sqrt((s1**2)/n1 + (s2**2)/n2)
numerador <- ((s1**2)/n1 + (s2**2)/n2)**2
denominador <- ((s1**2)/n1)**2/(n1-1) + ((s2**2)/n2)**2/(n2-1)
grados_libertad <- numerador / denominador
tcritL <- "-INF"
tcritU <- qt( 1-alfa, df=grados_libertad)
pvalue <- pt( tobs, df=grados_libertad, lower.tail=FALSE)

data.frame(tobs, tcritL, tcritU, pvalue)

```

```
##      tobs tcritL   tcritU      pvalue
## 1 3.591609 -INF 1.64495 0.0001648202
```

Por lo que vemos aqui el valor observado (tobs: t observado) esta por fuera de la zona de aceptacion, mayor al limite superior (t critico). Y esto sumado a que el pvalue es mucho menor que el nivel de significancia podemos rechazar la hipotesis nula. Esto implica que rechazamos la H₀ a favor de aceptar la Hipotesis alternativa.

Reconfirmemos estos calculos con la ayuda de la funcion t.test:

```
t.test(mujeres,hombres+1000, alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data: mujeres and hombres + 1000
## t = 3.5916, df = 15793, p-value = 0.0001648
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## 675.2701     Inf
## sample estimates:
## mean of x mean of y
## 11431.06 10185.18
```

2.2.4 Interpretación del test

Como hemos visto en el apartado anterior, si el p-value es menor que el valor de seleccionado, existen evidencias suficientes para rechazar H₀ en favor de H₁, osea:

H₀: *Indemnizacion Mujeres* \leq (*Indemnizacion Hombres* + 1000) => RECHAZADA

H₁: *Indemnizacion Mujeres* > (*Indemnizacion Hombres* + 1000) => ACEPTADA

Dado que el p-value es menor que y que el valor observado esta fuera de la zona de aceptacion de la hipotesis nula, podemos afirmar que se dispone de evidencia suficiente para considerar que las indemnización de las mujeres superan en mas de 1000 euros a la de los hombres con un nivel confianza del 95%

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 tambien a escala logarítmica la variable explicativa CosteInicio.

```
Clasif_rel=relevel(factor(claimNet$Clasificacion), ref = 'Muy lento')
Model_3 = lm(formula = log(CosteFinal) ~ Edad + Sexo + Estado + Dependientes +
             OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
             Clasif_rel + RiesgoSM + log(CosteInicio), data = claimNet)
summary(Model_3)
```

```

## 
## Call:
## lm(formula = log(CosteFinal) ~ Edad + Sexo + Estado + Dependientes +
##      OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
##      Clasif_rel + 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)           1.510e+00  4.979e-02 30.331 < 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 *** 
## Clasif_relLento       1.429e-02  1.414e-02   1.010 0.312277  
## Clasif_relMuy rápido -1.540e-03  1.405e-02  -0.110 0.912672  
## Clasif_relRápido      2.407e-02  1.388e-02   1.733 0.083046 .  
## RiesgoSM              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

```

3.1 Interpretación del modelo

Observando el p-value obtenido para cada variable explicativa podemos decir que las mas significativas para el modelo son: Edad, Sexo, y Salario en mayor medida, seguidas por Estado, Dependientes, la Jornada, DiasSemana, el RiesgoSM y el CosteInicio. Todas estas variables influyen con mayor significancia sobre la variable dependiente. Mientras que Clasificacion, OtrosDependientes y HorasSemanas, no tienen ninguna implicancia sobre el modelo.

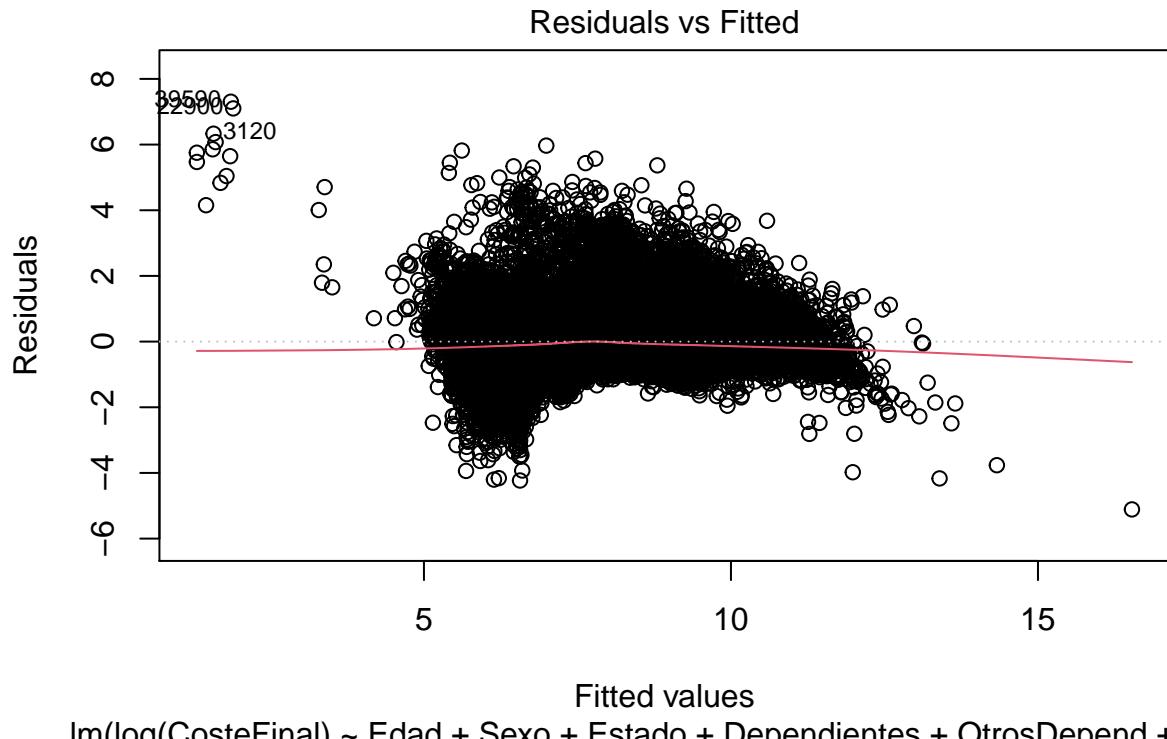
En cuanto al coeficiente de determinacion R^2 , el cual mide la proporción de variación de la variable dependiente explicada por la(s) variable(s) independiente(s), es de 0.7333. Tal vez no sea tan elevado, pero eso no necesariamente implique que el modelo no se ajuste a los datos. Esta claro que cuando mas alto sea mejor se ajustara, pero por el mismo no puede determinarse si las estimaciones y predicciones de los coeficientes estan sesgados o no, y es por eso que se deben examinar tambien las graficas de residuos como realizaremos a continuacion.

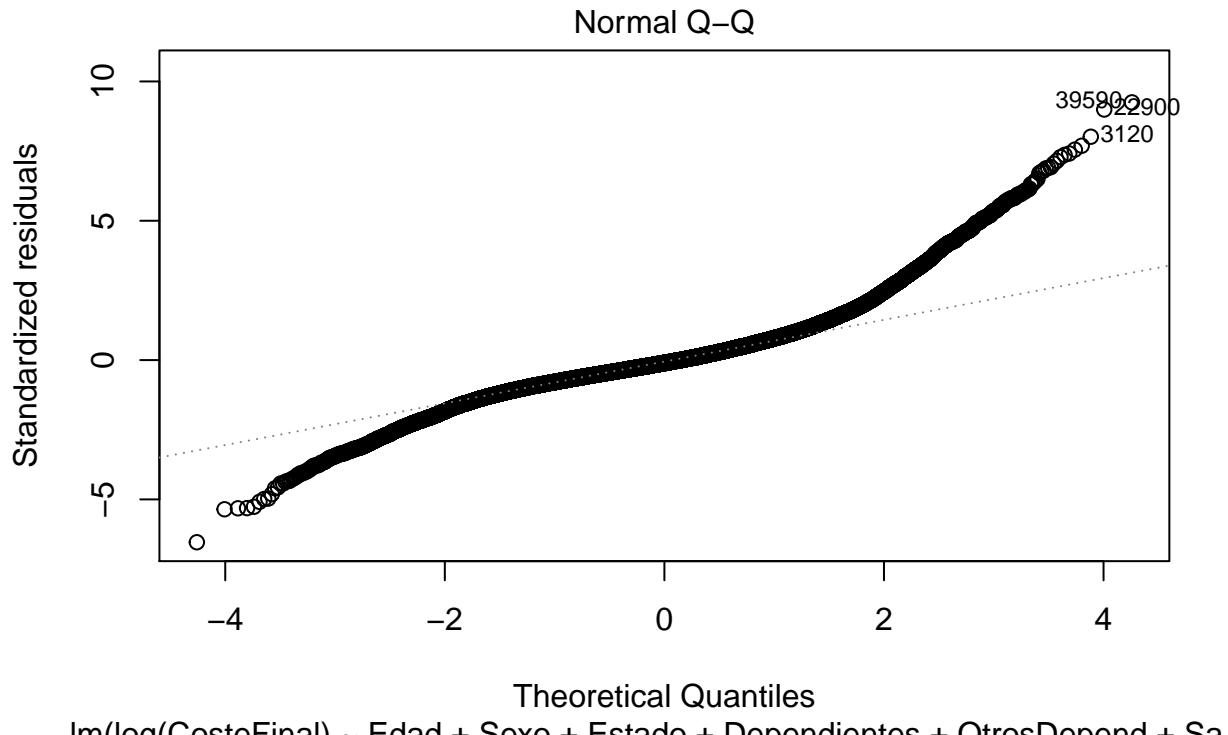
3.2 Análisis residuos

Como hemos dicho antes, 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. Recordemos que los residuos (o

errores) son la diferencia entre los valores observados y los valores que predice el modelo. Es por eso que utilizaremos los siguientes graficos de residuos para entender nuestra regresion:

```
plot(Model_3, which = c(1,2), caption = list("Residuals vs Fitted", "Normal Q-Q"))
```





A la vista del gráfico se observa un patrón de dispersión irregular. Es decir no es un patrón aleatorio de los residuos. Ademas se ven bastantes outliers. Esto indica que no se cumple el supuesto de varianza constante en los errores del modelo. Aqui las alternativas son probar con tests de igualdad de varianzas (complementarias a los graficos) y ver la posibilidad de transformar variables. o incluso para los outliers entender si son errores de medicion, tratarlos, o considerar realizar analisis robustos.

Mientras que si vemos el grafico QQ, claramente nuestros residuos no siguen una distribucion normal. Vemos bastantes outliers tanto a la izquierda como a la derecha. Si hicieramos una segunda iteracion se deberia analizar si tratando los outliers en nuestros datos podemos normalizar los residuos y de esta manera mejorar el modelo.

Que los residuos de un modelo de regresión lineal se distribuyan de forma normal es una condición necesaria para que la significancia (p-value) y los intervalos de confianza asociados a los predictores (calculados a partir de modelos teóricos) sean precisos.

En resumen si queremos determinar si encontramos modelos mejor ajustados o validados, tendriamos que profundizar mas en las variables para entender los outliers, o incluso si hay relaciones no lineales, osea exponenciales o logaritmicas, para ajustar mejor el modelo.

3.3 Predicción

Predecid 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).

```

newdata = data.frame(Edad=24, Sexo= 'F', Estado='S', Dependientes=1,
                     OtrosDepend=0, Salario=500, Jornada='F',
                     HorasSemana=40,DiasSemana=5, Clasif_rel='Lento',
                     RiesgoSM=1, CosteInicio=10000)
result=predict(Model_3, newdata)

exp(result)

```

```

##      1
## 13601.47

```

La predicción a partir del modelo definido nos dice que la indemnización de la persona será de 13601.47 euros.

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.

Creamos la nueva variable

```
claimNet$Deficit = as.integer(claimNet$CosteInicio < claimNet$CosteFinal)
```

Creamos el modelo

```

logit_model_4 = glm(formula=Deficit~Edad + Sexo + Estado + Dependientes +
                     OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
                     Clasif_rel + RiesgoSM + log(CosteInicio),
                     data=claimNet,
                     family=binomial)
summary(logit_model_4)

```

```

##
## Call:
## glm(formula = Deficit ~ Edad + Sexo + Estado + Dependientes +
##       OtrosDepend + Salario + Jornada + HorasSemana + DiasSemana +
##       Clasif_rel + RiesgoSM + log(CosteInicio), family = binomial,
##       data = claimNet)

```

```

## 
## Deviance Residuals:
##      Min       1Q   Median      3Q     Max 
## -5.4153 -1.1179  0.6951  0.9877  2.3301 
## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            3.832e+00 1.397e-01 27.423 < 2e-16 ***
## Edad                  1.122e-02 9.980e-04 11.241 < 2e-16 ***
## SexoM                -3.755e-01 2.436e-02 -15.414 < 2e-16 ***
## EstadoS              -6.871e-02 2.397e-02 -2.866 0.004152 **  
## Dependientes          1.090e-01 1.999e-02  5.450 5.03e-08 *** 
## OtrosDepend          1.719e-01 9.432e-02  1.822 0.068407 .   
## Salario               2.272e-03 5.416e-05 41.945 < 2e-16 *** 
## JornadaP              1.821e-01 4.528e-02  4.022 5.76e-05 *** 
## HorasSemana           -6.938e-04 8.648e-04 -0.802 0.422397  
## DiasSemana            -1.535e-01 2.448e-02 -6.270 3.61e-10 *** 
## Clasif_rellento       9.411e-02 3.846e-02  2.447 0.014407 *  
## Clasif_relMuy rápido  4.382e-02 3.821e-02  1.147 0.251520  
## Clasif_relRápido      1.401e-01 3.777e-02  3.708 0.000209 *** 
## RiesgoSM              7.214e-02 1.475e-01  0.489 0.624804  
## log(CosteInicio)     -4.902e-01 7.504e-03 -65.332 < 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: 60617  on 48660  degrees of freedom 
## AIC: 60647 
## 
## Number of Fisher Scoring iterations: 4

```

Calidad del Modelo

Se dice que un modelo presenta un buen ajuste a los datos si los valores que predice reflejan de manera adecuada los valores observados. Si el modelo presenta un mal ajuste, este no puede ser utilizado para extraer conclusiones ni efectuar predicciones.

Un modo de medir la adecuación de un modelo es proporcionando medidas globales de **bondad de ajuste** mediante test estadísticos.

Existen varias medidas de ajuste global para comparar la diferencia entre valores predichos y valores observados. Tres de las más utilizadas son:

- 1- el test basado en la devianza D,
- 2- el estadístico 2 (chi cuadrado) de Pearson y
- 3- el test de Hosmer-Lemeshow.

Los dos primeros se basan en los patrones de las covariables y pueden ser usados en los modelos lineales generalizados (MLG) en general.

El tercero se basa en probabilidades estimadas y se aplica en el caso de un MLG con distribución binomial, es decir, un modelo de regresión logística, que es justamente nuestro caso. Si una de las variables explicativas es continua (DISTANCE), no deben usarse los test 1 y 2, sino el test de Hosmer-Lemeshow. Este test consiste en comparar los valores previstos (esperados) por el modelo con los valores observados.

Veamoslo:

```
hoslem.test(claimNet$Deficit,fitted(logit_model_4))
```

```
##  
##  Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: claimNet$Deficit, fitted(logit_model_4)  
## X-squared = 282.78, df = 8, p-value < 2.2e-16
```

Este test se basa en las siguientes hipótesis:

H0: no hay diferencias entre las frecuencias observadas y las predichas (buen ajuste).

H1: sí hay diferencias (mal ajuste).

Por lo tanto dado que nuestro p-value es significativo, lo que implica el rechazo de H0 y por lo tanto que el modelo no ajusta bien a los datos. Tengamos en cuenta que aquí estamos usando por default todas las mismas variables que usamos para la regresión lineal, tal vez debería realizar un análisis más profundo o una 2da iteración para entender qué variables se adecuan más a los datos y de esta manera mejorar este valor obtenido con el test de Hosmer-Lemeshow.

Pero validemos también con el análisis de ROC y veamos qué obtenemos.

Curva de ROC

El análisis ROC proporciona un modo de seleccionar modelos posiblemente óptimos y subóptimos basado en la calidad de la clasificación a diferentes niveles o umbrales. Para tener una regla objetiva de comparación de las curvas ROC, se calcula el área bajo la curva, simplemente llamada AUROC (area under the ROC).

El modelo cuya área sea superior es el mejor.

En general:

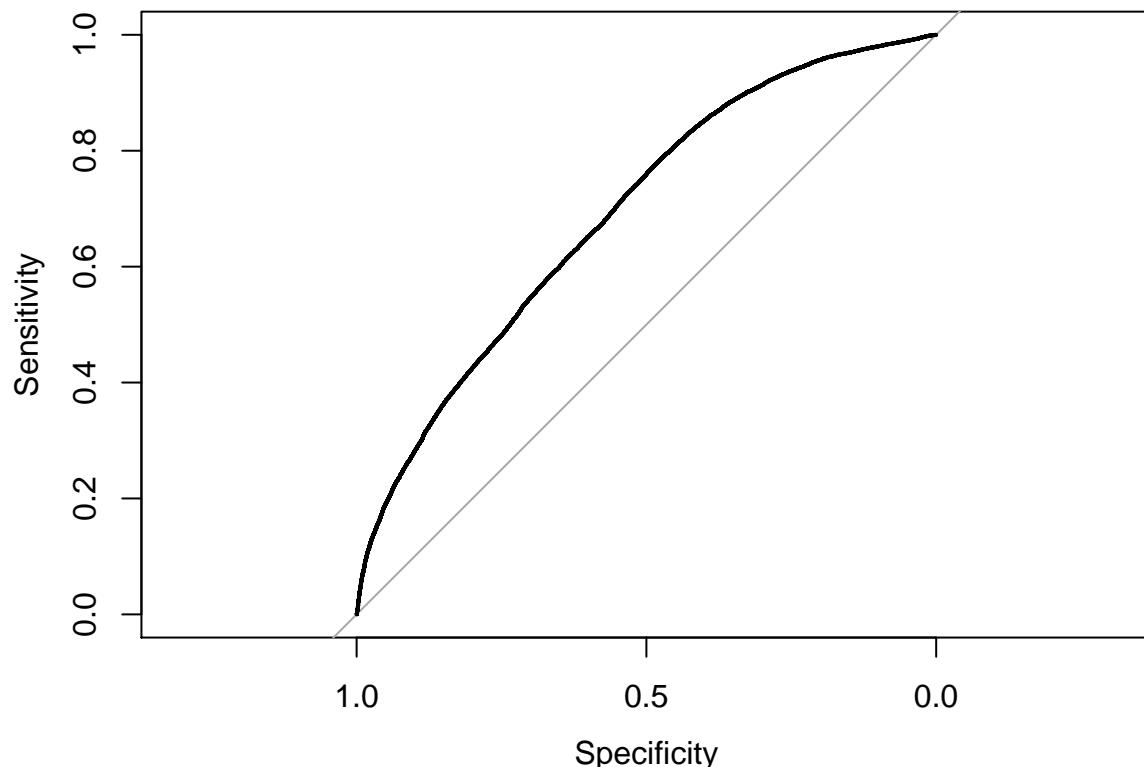
- Si AUROC $\leq 0,5$, el modelo no ayuda a discriminar.
- Si $0,6 \leq \text{AUROC} < 0,8$, el modelo discrimina de manera adecuada.
- Si $0,8 \leq \text{AUROC} < 0,9$, el modelo discrimina de forma excelente.
- Si $\text{AUROC} \geq 0,9$, el modelo discrimina de modo excepcional.

```
prob_low=predict(logit_model_4, claimNet, type="response")  
r=roc(claimNet$Deficit,prob_low, data=claimNet)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(r)
```



```
auc(r)
```

```
## Area under the curve: 0.6915
```

Por lo tanto segun el valor obtenido, podemos decir que nuestro modelo discrimina de manera adecuada.

4.2 Interpretación

Si comparamos la contribucion de cada variable respecto al modelo de regresion multiple del apartado anterior vemos que con regresion logistica algunas variables mejoraron su significancia, tal como es el caso de Clasificacion.

Aunque claro esta que en este caso nuestra variable dependiente es otra, y es mas probable que el tiempo en que se aperture un siniestro, como es el caso, tenga mayor significancia. Y todo lo contrario sucede por ej con el RiesgoSM que deja de ser significativa o tener efecto sobre el Deficit, cuando si era muy infuyente sobre el CosteFinal en la regresion lineal multiple.

Ademas obtuvimos valores un poco opuestos entre el test de Hosmer-Lemeshow y la curva de ROC, claramente se deberia realizar una 2da iteracion para mejorar el modelo.

Pero veamos tambien que nos dicen los odd ratios

```
exp(coefficients(logit_model_4))
```

##	(Intercept)	Edad	SexoM
----	-------------	------	-------

```

##          46.1473857      1.0112821      0.6869197
##          EstadoS      Dependientes      OtrosDepend
##          0.9335960      1.1151210      1.1875296
##          Salario       JornadaP       HorasSemana
##          1.0022742      1.1997821      0.9993064
##          DiasSemana    Clasif_relLento Clasif_relMuy rápido
##          0.8577338      1.0986766      1.0447924
##          Clasif_relRápido RiesgoSM      log(CosteInicio)
##          1.1503377      1.0748035      0.6124767

```

En base a las OR ajustadas saquemos algunas conclusiones de nuestro modelo.

Si focalizamos por ej en Sexo, observamos que para el sexo masculino, ajustado por el resto de las variables, tiene un valor menor 1. Esto significa que la posibilidad que exista Deficit cuando se trate de un hombre sera menor que cuando se trate de una mujer. Y un caso similar sucede con por el estado civil, donde el deficit siendo soltero sera menor que para un casado. Mientras que si observamos Clasificación, donde todos sus odds son mayores a 1 (por muy poco claro esta), quiere decir que es mas probable tener deficit en esos casos que cuando el siniestros se haya aperturado MuyLento.

4.3 Matriz de confusión

A continuación analizaremos 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. Analizaremos 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.

Para que un modelo de regresión logística con fines predictivos tenga éxito, el número de casos que se clasifican correctamente tiene que ser alto, mientras que el número de casos que se clasifican incorrectamente debe ser bajo.

Dicho eso analicemoslo para nuestro modelo con la matriz de confusión mediante el uso de funciones R.

La función confusionMatrix de la librería caret nos da muchas métricas, pero en este caso práctico nos concentraremos en explicar Accuracy, sensitivity, specificity y los porcentajes de falsos positivos y negativos analizando primero a partir de los resultados de la función CrossTable de la librería gmodels.

Veamos:

```

predicted <- predict(logit_model_4, newdata = claimNet, type = "response")

CrossTable( factor(as.numeric(predicted>0.5)), factor(claimNet$Deficit),
            prop.chisq = FALSE, prop.c = FALSE, prop.r =FALSE,
            dnn = c('Prediction', 'Reality'))

```

```

##
##
##      Cell Contents
##      |-----|-----|
##      |                   N |           |
##      |           N / Table Total |           |
##      |-----|-----|
##
```

```

## 
## Total Observations in Table: 48675
##
##          | Reality
##  Prediction |      0 |      1 | Row Total |
##  -----|-----|-----|-----|
##  0 |    9759 |    5900 |    15659 |
##  |    0.200 |    0.121 |        |
##  -----|-----|-----|-----|
##  1 |   10962 |   22054 |   33016 |
##  |    0.225 |    0.453 |        |
##  -----|-----|-----|-----|
## Column Total | 20721 | 27954 | 48675 |
##  -----|-----|-----|-----|
## 
## 
```

El número de resultados denominados falsos positivos corresponde a casos en los que la predicción de la probabilidad de la respuesta afirmativa es elevada, pero la respuesta observada es negativa. En nuestro caso existen 10962 falsos positivos. Que representan el 22.5 % del total de casos. Veamoslo caso por caso:

- True positives (TP): correct positive prediction: 22054
- False positives (FP): incorrect positive prediction: 10962
- True negatives (TN): correct negative prediction: 9759
- False negatives (FN): incorrect negative prediction: 5900
- Positivos (P) = (FN + TP): 27954
- Negativos (N) = (TN + FP): 20721

Pero que hay de la precision del modelo? tambien denominada Accuracy

- La misma se calcula como el total de predicciones correctas sobre el total de casos, positivos y negativos.

$$Accuracy = (TP + TN)/(P + N) = 0.6535$$

- Sensitivity es calculada como el numero correcto de predicciones positivas (TP) sobre el total de casos positivos (P). La mejor sensitivity es 1 mientras que la peor es 0.

$$Sensitivity = (TP)/(P) = 0.7889$$

- Specificity es calculada como el numero correcto de predicciones negativas (TN) sobre el total de casos negativos (N). La mejor Specificity es 1 mientras que la peor es 0.

$$Specificity = (TN)/(N) = 0.4709$$

O sea a nuestro modelo con una Sensitivity de 0.7889 demuestra buena capacidad para detectar los casos positivos mientras que con un Specificity de 0.4709 parece no detectar los negativos de forma tan precisa.

Todo esto es calculado directamente ejecutando la funcion que hemos comentado antes:

```

confusionMatrix(data = factor(as.numeric(predicted>0.5)),
                 reference = factor(claimNet$Deficit), positive='1')

## Confusion Matrix and Statistics
##
##             Reference
## Prediction      0      1
##               0 9759  5900
##               1 10962 22054
##
##                  Accuracy : 0.6536
##                  95% CI : (0.6493, 0.6578)
##      No Information Rate : 0.5743
##      P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.2684
##
## Mcnemar's Test P-Value : < 2.2e-16
##
##                  Sensitivity : 0.7889
##                  Specificity : 0.4710
##      Pos Pred Value : 0.6680
##      Neg Pred Value : 0.6232
##      Prevalence : 0.5743
##      Detection Rate : 0.4531
##      Detection Prevalence : 0.6783
##      Balanced Accuracy : 0.6300
##
##      'Positive' Class : 1
##

```

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?

```

predict_44 = data.frame(Edad=20, Sexo= 'M', Estado='S', Dependientes=0,
                      OtrosDepend=0, Salario=300, Jornada='F',
                      HorasSemana=30,DiasSemana=5, Clasif_rel='Muy lento',
                      RiesgoSM=0, CosteInicio=10000)
result = predict(logit_model_4, newdata=predict_44, type ="response")
result

##          1
## 0.2669856

```

El modelo nos predice una probabilidad del 26,7% de que la valoración inicial será insuficiente para un hombre con los atributos descriptos.

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.

Recordemos que clasificacion nos indicaba la “velocidad” con que los siniestros era abiertos desde su ocurrencia.

```
table(claimNet$Clasificacion)
```

```
##          Lento    Muy lento    Muy rápido      Rápido
##      13589         4060       14789       16237
```

5.1 Hipótesis nula y alternativa

Queremos preguntarnos si la clasificacion del tiempo de apertura de los siniestros es significativa, en el sentido si el CosteFinal de la indemnizacion de los empleados es afectado por ese factor. Por lo tanto, planteemos la hipotesis nulas y alternativa para nuestra pregunta de investigacion.

En ANOVA como primer paso queremos analizar si existira diferencias significativas entre los grupos del factor Clasificacion. Por lo que nuestras hipotesis serian

- H0: no hay diferencias significativas entre los niveles de clasificacion
- H1: Al menos hay una diferencia significativa entre los niveles de clasificacion

Pero de que diferencias hablamos? Este analisis consiste en un analisis de varianzas.

Si denominamos Varianza como V y a cada grupo de la siguiente forma:

- Nivel 1 = Muy Lento = ml
- Nivel 2 = Lento = l
- Nivel 3 = Rápido = r
- Nivel 4 = Muy Rapido = mr

Tendremos las hipotesis escritas de la siguiente forma.

- H0: $V_{ml} = V_l = V_r = V_{mr} = 0$
- H1: $V_i \neq V_j$, para algun i, j de: c(ml, l, r, mr)

5.2 Modelo

Para calcular la tabla ANOVA asociada a este problema debemos primero ajustar el modelo mediante la función aov:

```
claimNet$CosteFinal.log = log(claimNet$CosteFinal)
model_52<-aov(CosteFinal.log~Clasificacion,data=claimNet)
model_52
```

```

## Call:
##   aov(formula = CosteFinal.log ~ Clasificacion, data = claimNet)
##
## Terms:
##           Clasificacion Residuals
## Sum of Squares      189.88 113707.02
## Deg. of Freedom       3      48671
##
## Residual standard error: 1.528476
## Estimated effects may be unbalanced

```

y luego usamos la funcion anova para analizar los resultados, obteniendo aqui la Tabla ANOVA:

```

taov_52<-anova(model_52)
print(taov_52)

```

```

## Analysis of Variance Table
##
## Response: CosteFinal.log
##             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

```

Obtenemos un p-valor 2.2e-16, claramente inferior a un nivel de significación del 5 %. Por lo tanto, aceptamos la hipótesis alternativa y concluimos que el factor Clasificacion es significativo. Es decir, el tipo de apertura de un siniestro afecta al CosteFinal de la indemnizacion otorgada. La tarea a continuación será determinar en qué niveles del factor están esas diferencias.

Como vemos en la tabla ANOVA y en concreto el valor de MSE nos proporciona una estimación de la varianza del error $\hat{\sigma}^2 = 2.336$.

Tambien tenemos que la suma de cuadrados es 190, la media de la suma de cuadrados 63.293, el F 27.092 y el p-value que ya hemos mencionado.

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.

Esto es posible calcularlo con la funcion etaSquared de R. Pero que nos dice esta funcion? Basicamente obtendremos un porcentaje el cual indicadara la medida en que la variabilidad de CosteFinal es explicada por el factor Clasificacion. Osea hasta que punto este factor es capaz de explicar la variabilidad de los datos.

```

#Effect size
etaSquared(model_52)

```

```

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

```

Como vemos el valor es menor a 1%. Es decir menos del 1% de la variabilidad se explica por la variable Clasificacion, osea, tiene poco impacto sobre coste final. Pero ahora vemos como es explicada la variabilidad por cada nivel del factor.

```

model.tables(model_52,type="effects")

## Tables of effects
##
## Clasificacion
##      Lento Muy lento Muy rápido Rápido
## 1.180e-02 -6.761e-03 -8.568e-02 6.985e-02
## rep 1.359e+04 4.060e+03 1.479e+04 1.624e+04

```

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.

La función HSD de la prueba Tukey-Kramer nos permite analizar a partir del estadístico calculado cual es la diferencia mínima que tienen q tener dos grupos para que pueda considerarse distintos entre si.

Pero antes por ejemplos usemos la función pairwise.t.test del paquete stat para obtener los p-valores de todas las comparaciones por parejas.

```

pairwise.t.test(claimNet$CosteFinal.log,
                claimNet$Clasificacion,
                p.adj=c("none"))

##
## Pairwise comparisons using t tests with pooled SD
##
## data: claimNet$CosteFinal.log and claimNet$Clasificacion
##
##          Lento Muy lento Muy rápido
## Muy lento 0.4971 -       -
## Muy rápido 8e-08 0.0036 -       -
## Rápido     0.0011 0.0043 <2e-16
##
## P value adjustment method: none

```

Vemos que hay diferencias entre los niveles Lento y Muy Lento con el nivel Rapido o Muy Rapido. No obstante, cabe remarcar que, en el caso de un diseño con a tratamientos, hay $a(a-1)/2$ comparaciones por parejas de tratamientos distintas. Si estamos interesados en hacer todos los contrastes por parejas, hay que emplear técnicas de inferencia simultánea para ajustar los p-valores obtenidos mediante el test anterior.

Para ello realicemos las comparaciones múltiples con la prueba inicialmente mencionada:

```
HSD.test(model_52, "Clasificacion", group=T, console = TRUE)
```

```

##
## Study: model_52 ~ "Clasificacion"
##
## HSD Test for CosteFinal.log
##
## Mean Square Error: 2.336238
##
```

```

## Clasificacion, means
##
##          CosteFinal.log      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.
##
##          CosteFinal.log groups
## Rápido        8.072725    a
## Lento         8.014673    b
## Muy lento     7.996111    b
## Muy rápido    7.917190    c

```

Observamos que las aperturas de siniestros lentas y muy lentas forman un grupo homogéneo que provoca un nivel de Coste Final similar, mientras que las aperturas Muy Rapida y Rapida forman dos grupos mas, siendo el primero el que menos CosteFinal genera y el segundo el de mayor impacto en el coste final de la indemnización.

5.5 Adecuación del modelo

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

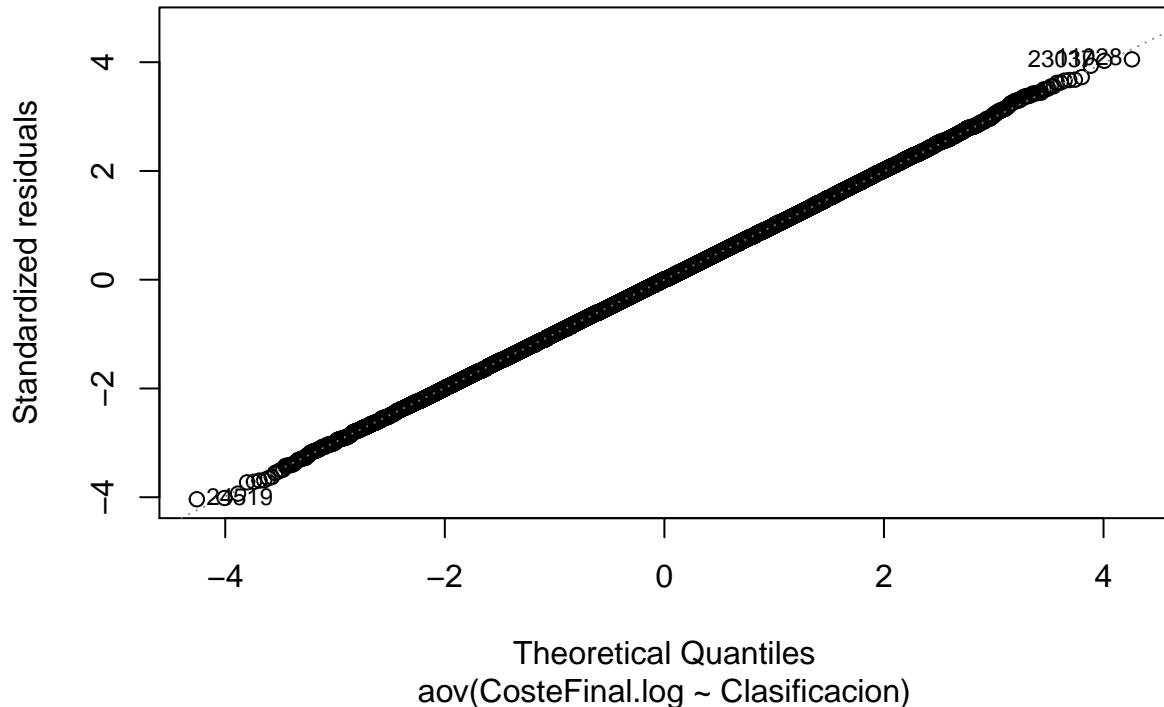
- Análisis visual de normalidad de los residuos. Podéis usar la función plot sobre el modelo ANOVA calculado.
- Análisis visual de homocedasticidad de los residuos. Podéis usar plot sobre el modelo ANOVA calculado.
- Contraste de normalidad y homocedasticidad.

Veamos cada punto: la normalidad y homocedasticidad analizada visualmente como tambien con estadistica inferencial.

- **Normalidad de los residuos**

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

```
plot(model_52, which = 2,
      caption = list( "Normal Q-Q"))
```



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

Adicionalmente, podemos contrastar la normalidad mediante el test de Lilliefors (Kolmogorov-Smirnov), donde la hipótesis nula es aquella que afirma que la distribución es normal:

```
lillie.test(residuals(model_52))
```

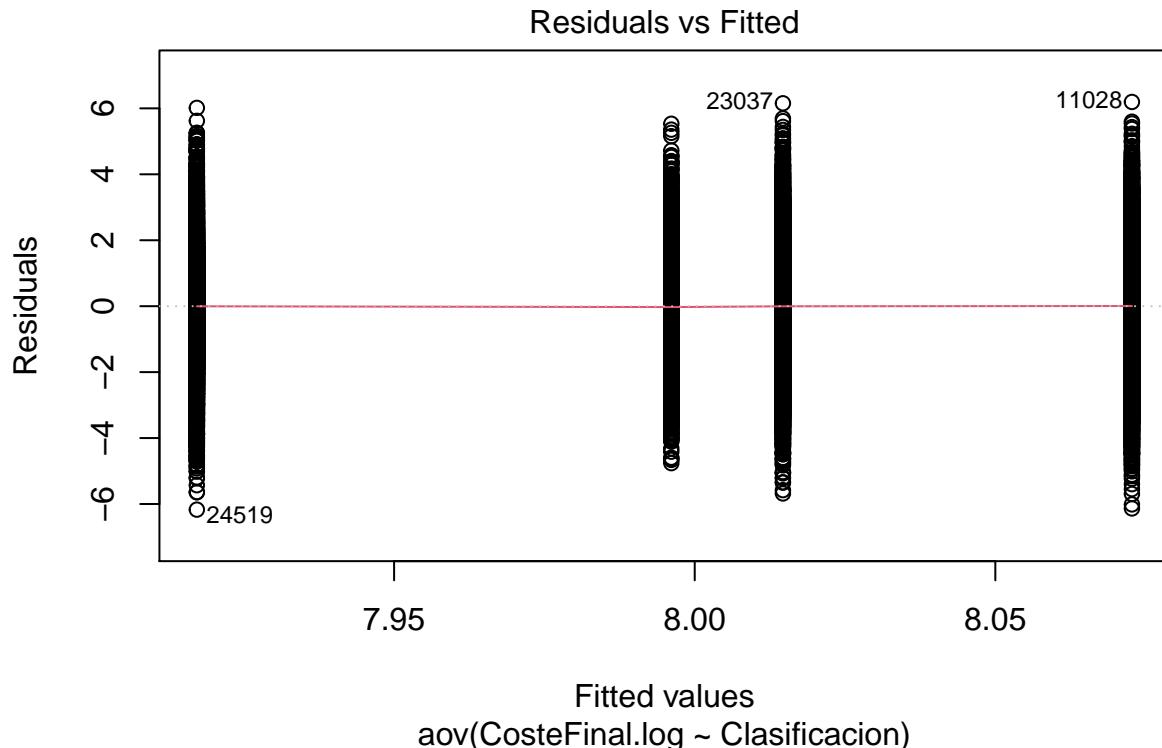
```
##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data:  residuals(model_52)  
## D = 0.0018258, p-value = 0.959
```

A la vista del p-valor obtenido, mantenemos la hipótesis nula y podemos aceptar que la variable aleatoria e_{ij} sigue una distribución normal.

- **Homocedasticidad**

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

```
plot(model_52, which = 1, caption = list( "Residuals vs Fitted"))
```



Observamos 4 tiras verticales de puntos que están situadas en las medias de cada grupo. Como hemos dicho, estas corresponden a los valores ajustados de las observaciones. La disposición de los residuos muestra una dispersión parecida en cada tira.

Tener en cuenta que la visualización de las varianzas nos da una indicación visual orientativa en lugar de una confirmación estadística de homogeneidad. Para confirmarla o no debemos realizar un test estadístico.

Por eso mismo vamos a testear la homogeneidad de varianzas mediante el test de Bartlett, que es aplicable en caso de normalidad. La hipótesis nula es la de homogeneidad de varianzas a través de los niveles. Esto es

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$H_1: \alpha_i \neq \alpha_j \text{ para algun } i \neq j$$

Sigamos analizando los datos del ejemplo. Evaluamos la homogeneidad de las varianzas.

```
bartlett.test( CosteFinal.log~Clasificacion, data=claimNet)
```

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

A la vista del p-valor, rechazamos la hipótesis nula de que todas las varianzas son iguales. Con que una no sea igual, razon suficiente para rechazar la hipótesis nula. Esto de alguna forma confirma lo que se vio visualmente ya que si bien a "la vista" eran las 4 líneas muy parecidas, la 2da mostró una mayor compresión de los puntos.

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:

6.1.1. Agrupacion Conjunto de datos

Agrupad el conjunto de datos por Sexo y por RiesgoSM. Calculad el número de casos disponibles de cada combinación de factores.

```
summaryBy(CosteFinal.log ~ Sexo+RiesgoSM, data=claimNet, FUN=c(NROW))
```

```
##   Sexo RiesgoSM CosteFinal.log.NROW
## 1   F      0          11173
## 2   F      1           84
## 3   M      0          37286
## 4   M      1           132
```

6.1.2. Media por grupo

Calculad la media de coste (en log) para cada grupo.

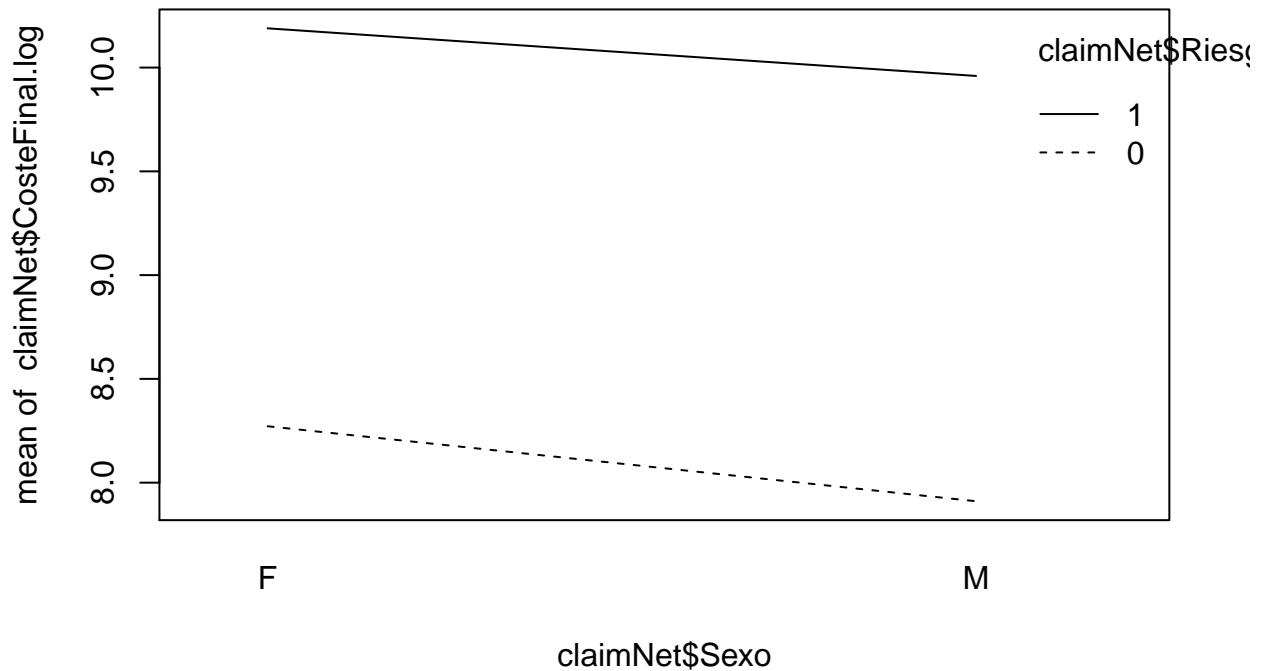
```
claimNet$RiesgoSM = factor(claimNet$RiesgoSM)
summaryBy(CosteFinal.log ~ Sexo+RiesgoSM, data=claimNet, FUN=c(mean))
```

```
##   Sexo RiesgoSM CosteFinal.log.mean
## 1   F      0        8.271703
## 2   F      1       10.189007
## 3   M      0        7.910462
## 4   M      1       9.959641
```

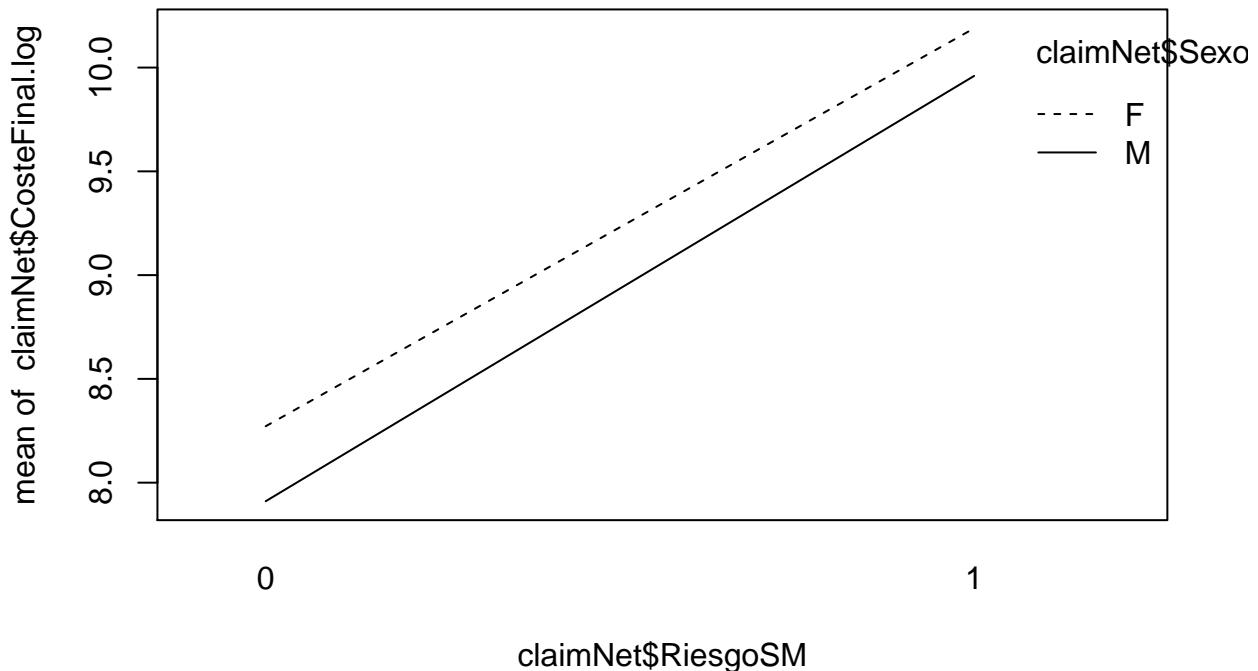
6.1.3. Visualizacion del valor medio

Mostrad en un gráfico el valor medio de la variable CosteFinal en escala logarítmica para cada factor.

```
interaction.plot(claimNet$Sexo, claimNet$RiesgoSM, claimNet$CosteFinal.log)
```



```
interaction.plot(claimNet$RiesgoSM, claimNet$Sexo, claimNet$CosteFinal.log)
```



6.1.4. Interpretacion

Como vemos en las graficas se ve que hay efectos sobre la variable respuesta pero no hay interaccion entre las variables condicion. Veamos en detalle:

En el primer grafico en el eje X aparecen los dos niveles del factor Sexo, en el eje Y aparece la media de la variable respuesta (CosteFinal.log). Estrictamente, el gráfico se basa en las cuatro medias para las cuatro condiciones experimentales, pero se unen mediante líneas las medias que corresponden al mismo nivel del factor que no está representado en el eje X (RiesgoSM). Vemos que tanto para los empleados con o sin Riesgo el valor de CosteFinal es mayor en el grupo de Mujeres. También observamos que los empleados con Riesgo (1) tienen un CosteFinal mayor que los Sin Riesgo (0) sean hombres o mujeres. Constatamos que el descenso en el CosteFinal entre los dos Niveles de Riesgo es aproximadamente el mismo en ambos Sexos, las líneas parecen paralelas y diremos que no se aprecia interacción.

Y como vemos al intercambiar los factores en la representación de la interacción llegamos a conclusiones similares.

6.2 Cálculo del modelo

Dado que no existe interaccion entre nuestros factores podríamos calcular el modelo de la siguiente forma:

```
modelo_62 <- aov(CosteFinal.log ~ Sexo + RiesgoSM, data = claimNet)
```

Pero creemoslo forzandola para verificar que esa interaccion no es significativa:

```

modelo_62_interaccion_forzada=aov(CosteFinal.log~Sexo+RiesgoSM+Sexo:RiesgoSM,
                                    data = claimNet)
anova(modelo_62)

```

```

## Analysis of Variance Table
##
## Response: CosteFinal.log
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Sexo          1   1174  1173.94  510.78 < 2.2e-16 ***
## RiesgoSM      1     858   857.93  373.28 < 2.2e-16 ***
## Residuals  48672 111865      2.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

anova(modelo_62_interaccion_forzada)

```

```

## Analysis of Variance Table
##
## Response: CosteFinal.log
##           Df Sum Sq Mean Sq  F value Pr(>F)
## Sexo          1   1174  1173.94 510.7705 <2e-16 ***
## RiesgoSM      1     858   857.93 373.2760 <2e-16 ***
## Sexo:RiesgoSM 1      1     0.89   0.3861 0.5343
## Residuals  48671 111864      2.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Como vemos en el anova forzando la interaccion nada cambia y se observa que esa interaccion NO es significativa. Tal como habiamos validado visualmente con el plot de interaccion de grupos.

Vemos que los factores principales son significativas, y aceptamos por lo tanto que hay efecto del Sexo y RiesgoSM.

Ademas observamos que la variable Sexo es la que mayor efecto tiene sobre el CosteFinal, dado que por ej si nos basamos en la suma de cuadrados entre grupos de Sexo, 1174, esta es la diferencia entre grupos de sexo. Mientras que para el RiesgoSM esa diferencia entre grupos es de 858, menor a la de Sexo. Y como hemos dicho antes, la interaccion Sexo:RiesgoSM la suma de cuadrados es casi cero, por lo que no tiene efectos sobre la variabilidad del CosteFinal.

Por ultimo la variabilidad no explicada, los residuos, es la que se observa en la ultima fila de la tabla ANOVA. Y ahí mismo tenemos que la estimacion de la varianza del error a partir de los MSE es $\hat{\sigma}^2 = 2.3$

Con la funcion model.tables podemos tener un mayor detalle de los efectos de las variables condicion sobre el CosteFinal en relacion a la media de todos los datos:

```

model.tables(modelo_62, type = "effects")

```

```

## Tables of effects
##
##   Sexo
##       F         M
## 2.831e-01 -8.518e-02
## rep 1.126e+04  3.742e+04

```

```

##  

##   RiesgoSM  

##      0      1  

## -8.861e-03  1.988  

## rep  4.846e+04 216.000

```

Dado que los factores principales han resultado significativos cada uno por su lado sin haber interaccion, deberemos realizar comparaciones por parejas. Por ejemplo, mediante el test de Tukey:

```

# Aqui vemos que se forman dos grupos para cada sexo  

HSD.test(modelo_62, c("Sexo"), group=T, console = TRUE)

```

```

##  

## Study: modelo_62 ~ c("Sexo")  

##  

## HSD Test for CosteFinal.log  

##  

## Mean Square Error:  2.298345  

##  

## Sexo,  means  

##  

##   CosteFinal.log      std      r      Min      Max  

## F       8.286010 1.418945 11257 3.401110 14.26693  

## M       7.917691 1.551428 37418 1.746615 14.16842  

##  

## Alpha: 0.05 ; DF Error: 48672  

## Critical Value of Studentized Range: 2.771808  

##  

## Groups according to probability of means differences and alpha level( 0.05 )  

##  

## Treatments with the same letter are not significantly different.  

##  

##   CosteFinal.log groups  

## F       8.286010     a  

## M       7.917691     b

```

```

# caso similar para el riesgo que son dos grupos claramente distintos  

HSD.test(modelo_62, c("RiesgoSM"), group=T, console = TRUE)

```

```

##  

## Study: modelo_62 ~ c("RiesgoSM")  

##  

## HSD Test for CosteFinal.log  

##  

## Mean Square Error:  2.298345  

##  

## RiesgoSM,  means  

##  

##   CosteFinal.log      std      r      Min      Max  

## 0       7.993752 1.525308 48459 1.746615 14.26693  

## 1      10.048839 1.073968 216 7.355029 13.45198  

##  

## Alpha: 0.05 ; DF Error: 48672

```

```

## Critical Value of Studentized Range: 2.771808
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
## CosteFinal.log groups
## 1      10.048839      a
## 0      7.993752      b

```

Mientras que este analisis se hubiera hecho si hubiese habido interaccion entre Sexo y RiesgoSM. De forma de determinar la combinacion de interaccion mas optima

```
HSD.test(modelo_62, c("Sexo", "RiesgoSM"), group=T, console = TRUE)
```

```

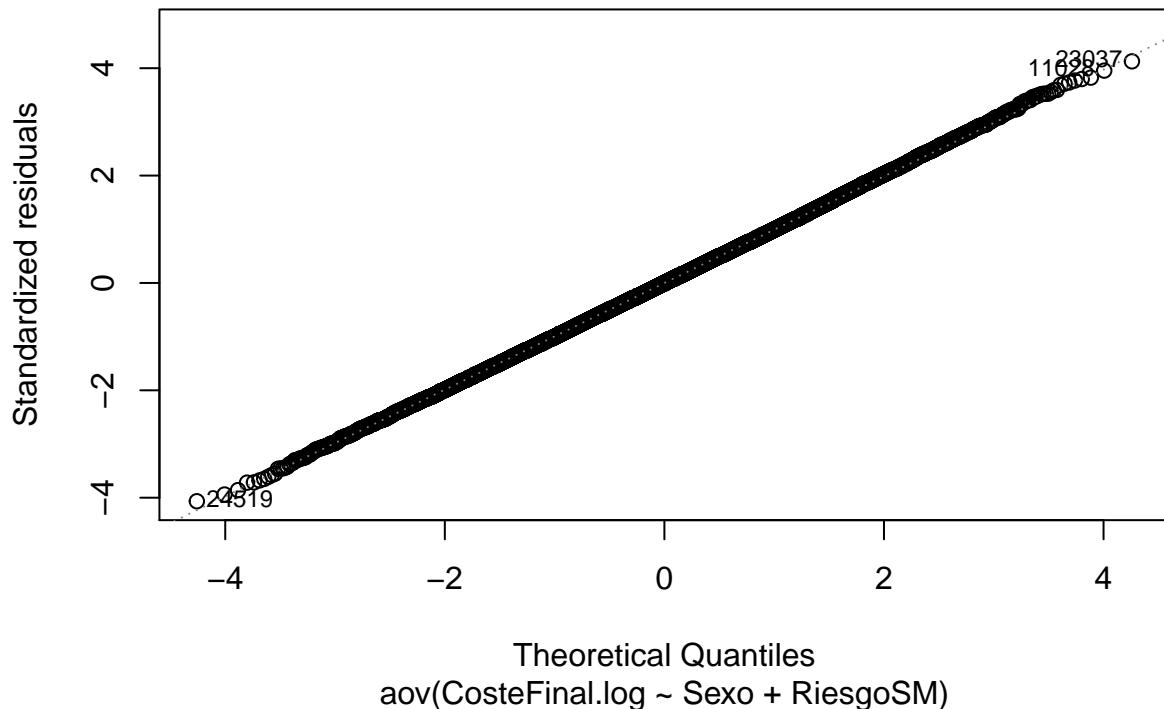
##
## Study: modelo_62 ~ c("Sexo", "RiesgoSM")
##
## HSD Test for CosteFinal.log
##
## Mean Square Error:  2.298345
##
## Sexo:RiesgoSM,  means
##
##      CosteFinal.log      std      r      Min      Max
## F:0      8.271703 1.4123165 11173 3.401110 14.26693
## F:1      10.189007 0.9333191    84 7.355029 11.92483
## M:0      7.910462 1.5479010 37286 1.746615 14.16842
## M:1      9.959641 1.1491186   132 7.363679 13.45198
##
## Alpha: 0.05 ; DF Error: 48672
## 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.
##
## CosteFinal.log groups
## F:1      10.189007      a
## M:1      9.959641      a
## F:0      8.271703      b
## M:0      7.910462      c

```

Donde hubiesemos tenido 3 grupos bien definidos. Donde la combinacion Riesgo con cualquiera de los sexos conforman uno, y otros dos grupos mas conformados por las observaciones sin riesgo para mujeres y otra sin riesgo para hombres.

Adecuación del modelo.. Analicemos por ultimo la adecuacion del modelo

```
plot(modelo_62, which = 2,
      caption = list( "Normal Q-Q"))
```



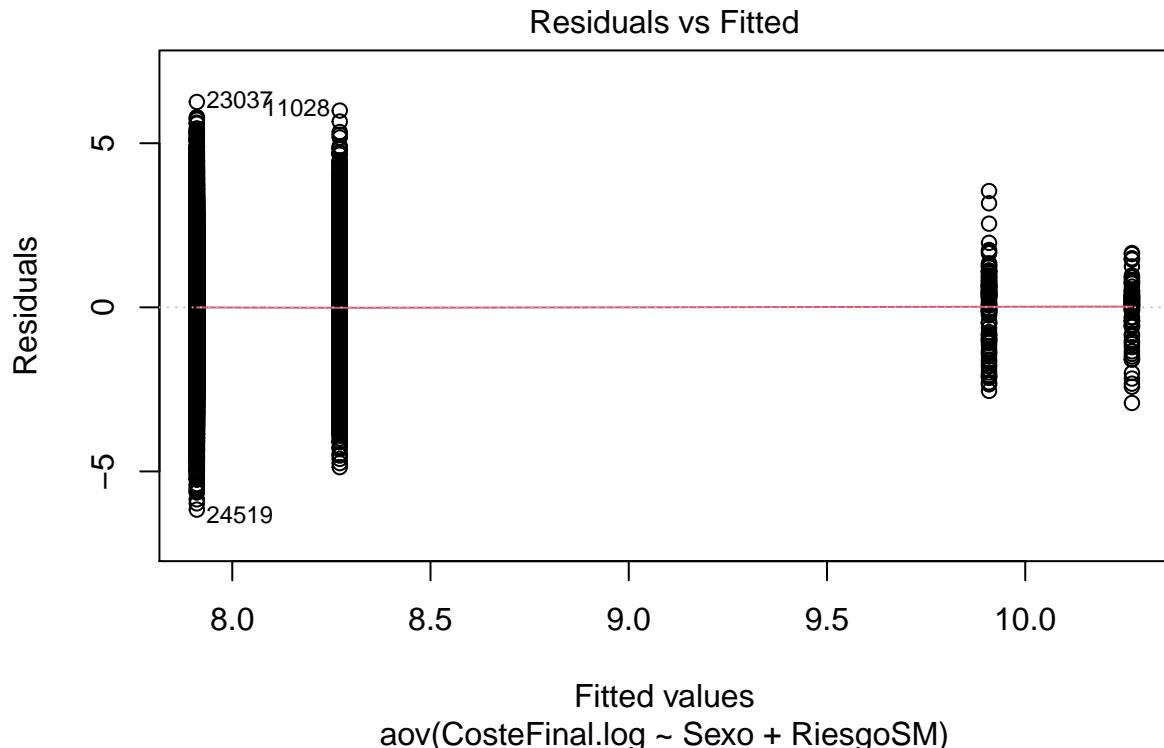
Claramente los residuos siguen una distribución normal.

```
lillie.test(residuals(modelo_62))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: residuals(modelo_62)  
## D = 0.0032984, p-value = 0.2247
```

Tal como es comprobado estadísticamente con Lilliefors. Osea a la vista de la disposición de los cuantiles y del p-valor del test de Lilliefors, aceptamos la hipótesis de normalidad.

```
plot(modelo_62, which = 1, caption = list("Residuals vs Fitted"))
```



Aquí vemos justamente el patrón habitual cuando no se cumple la suposición de varianza constante, que es aquel que aparece en caso de que la varianza depende de la media del grupo. Por lo general, en estos casos hay un aumento de la varianza a medida que la media del grupo aumenta o disminuye. En consecuencia, el gráfico de residuos muestra una apertura a la derecha o a la izquierda (tal como es nuestro caso) en forma de embudo.

```
condition <- with(claimNet, interaction(RiesgoSM, Sexo))
bartlett.test(CosteFinal.log~condition, data=claimNet)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  CosteFinal.log by condition
## Bartlett's K-squared = 185.26, df = 3, p-value < 2.2e-16
```

Hecho que se comprueba con el test de Barlett donde el p-value obtenido nos marca un rechazo de la hipótesis nula de homocedasticidad.

La forma habitual de solucionar la no constancia de la varianza es por la transformación de la variable respuesta. Para algunas distribuciones hay transformaciones estándar que igualan o estabilizan la varianza. Hay una teoría general de las transformaciones estabilizadoras de la varianza que se aplica a las distribuciones donde la varianza depende de la media. Mientras que en casos donde no se cuenta con una transformación conocida, se suele utilizar el método de Box-Cox el cual establece un procedimiento para determinar a partir de los datos la transformación de potencia. Pero la aplicación del mismo excede el objetivo de esta práctica.

6.3 Interpretación de los resultados

Como hemos podido analizar en el paso a paso del anova multifactorial es que las variables condicion Sexo y RiesgoSM son significativas y tienen gran influencia sobre la variable independiente, pero sin interaccionar entre ellas. Y si bien los residuos siguen una distribucion normal (tras la aplicacion del log), no tenemos homogeneidad entre grupos, lo cual obligaria a realizar transformacion, por ej aplicando el metodo de Box-Cox para lograr la homocedasticidad. Y asi, hacer cumplir las condiciones de aplicacion de Anova.

7 Conclusiones

En esta practica hemos realizado varios analisis estadisticos sobre el conjunto de datos provisto y del cual hemos obtenido interesantes resultado que pasamos a detallar a continuacion

- **Normalidad de la variable CosteFinal**

Hemos podido comprobar que la variable CosteFinal no sigue una distribucion normal, pero que al aplicarle la funcion logaritmo se logra normalizar dicha variable. Dato interesante, dado que es el objetivo a analizar y/o predecir en los apartados siguientes de esta practica.

- **Contrastes de hipotesis**

Se pudo inferir a traves de contrastes de hipotesis que la indemnización de las mujeres supera a los hombres en 1000 euros con una confianza del 95%

- **Regresion Lineal**

En este punto se busco predecir la variable CosteFinal a partir de un grupo de variables de nuestro dataset, pero no se ha conseguido un muy buen ajuste del mismo a los datos ($R^2 = 0.7333$) y de hecho se ha comprobado que los residuos no siguen una distribucion normal ni existe homogeneidad de la varianza. Condiciones necesarias para una buena adecuacion de un modelo.

- **Regresion Logistica**

En este caso buscamos predecir la variable deficit, la cual indica probabilidad de que la compañia cuantifique inicialmente el coste del siniestro de forma insuficiente.

Aqui hemos generado un modelo con una precision de solo el 65%. No obtuviendo tampoco buenos resultados con los test de calidad del modelo como ser Hosmer-Lemeshow, y digamos que la AUROC tampoco ha sido muy elevado.

Por lo que podemos decir que este modelo de regresion logistica asi tal como fue planteado no logro una gran adecuacion a los datos.

Pero si se ha podido obtener interesantes conclusiones de los datos para siguientes iteraciones, ya que se ha logrado identificar variables significativas que influyen sobre el Deficit.

- **ANOVA de un factor**

Se aplico ANOVA para contrastar si existen diferencias en la variable CosteFinal en función de la clasificación del siniestro en relación al tiempo transcurrido hasta la apertura, determinando que **si** existen diferencias entre las varianzas de cada nivel del factor y dada esa situación, se obtuvieron luego los grupos mas significativos del factor sobre CosteFinal.

- **ANOVA Multifactorial**

Por ultimo se evaluo el efecto sobre CosteFinal según el factor Sexo combinado con el factor RiesgoSM, donde se pudo tambien comprobar que ambas variables influyen sobre el Coste, pero sin interaccionar entre ellos (los factores).

Recursos

- Análisis de la varianza (ANOVA), Ferran Reverter.
- Modelos de regresión logística, Montserrat Guillén Estany y María Teresa Alonso Alonso.
- https://rdrr.io/cran/eeprotools/man/age_calc.html
- <https://picandoconr.wordpress.com/2016/08/30/normalidad-shapiro-test-y-lillie-test>
- <https://classeval.wordpress.com/introduction/basic-evaluation-measures/>
- <https://www.rdocumentation.org/packages/lsr/versions/0.5/topics/etaSquared>