Práctica 2 - Limpieza y validación de los datos

Irene Calvo Cuesta - icalvocu

Contents

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset elegido se encuentra disponible en este enlace de kaggle. Recoge un análisis sobre los gastos médicos individuales facturados por el seguro de salud de Estados Unidos, y una serie de características personales de los asegurados. Para realizar el análisis se han recogido datos de 1338 asegurados. El objetivo de dicho análisis es investigar si se puede predecir la prima del seguro a partir de la edad del asegurado, su género, su índice de masa corporal, el número de hijos que tiene y por tanto que están cubiertos por el seguro de salud, si es fumador o no, y la región de Estados Unidos en la que vive; y que grado de influencia tienen dichas características sobre los costos del seguro.

2. Integración y selección de los datos de interés a analizar.

En primer lugar se debe realizar la carga de los datos, para ello se inspecciona el tipo de formato csv. Se puede comprobar que se usa la coma (,) como separador de valores y el punto (.) como separador decimal, por tanto se usará la funcion **read.csv()** para la lectura del fichero.

```
#Lectura de datos
asegurados <- read.csv('insurance.csv')
#Se agiliza la manipulacion de los datos,
#para que no sea necesario especificar el nombre del dataframe
attach(asegurados)

#Se comprueba la carga correcta del archivo
str(asegurados)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
#Se observan las primeras filas del conjunto
head(asegurados)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Como he comentado, el fichero que tenemos recoge información sobre una previsión del coste de la prima de los seguros médicos en Estados Unidos y una serie de datos personales de los asegurados.

Podemos observar que el fichero contiene 7 variables que corresponden a la edad de los asegurados (**age**), al género (**sex**), al índice de masa corporal (**bmi**), el número de hijos (**children**), si fuman o no (**smoker**), la

región a la que pertenecen (**region**), y los cargos de la prima del seguro (**charges**). Este fichero cuenta con 1338 observaciones.

Visualizando las 5 primeras filas del archivo se comprueba que se ha cargado correctamente.

Por otro lado voy a observar las principales características que tienen las variables del conjunto de datos:

```
#Se observan las principales características de las variables del conjunto
summary(asegurados)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Observamos que la muestra presenta una media de edad de 39 años, un índice de masa corporal medio de 30.66 kg/m2 (el ideal se encuentra entre 18.5 y 24.9kg/m2), una media de hijos entre 1 y 2, y unas primas de seguro medias de 13270$.

Comprobamos también que los niveles de las variables categóricas son los correctos, y que no es necesario realizar alguna estandarización en los nombre. La variable **sex** tiene dos niveles (female y male), la variables **smoker** tiene dos niveles (no y yes) en función de si el asegurado fuma o no, y la variable región tiene cuatro niveles correspondientes a la división por regiones de EEUU en noreste, noroeste, sureste, suroeste.

Si que voy a estandarizar los valores de la variable **bmi** y los de la variable **charges** en dos cifras decimales para una mejor interpretación, ya que suelen tratarse de esta forma:

```
#Estandarizacion variable bmi
asegurados$bmi<- round(asegurados$bmi, 2)
#Estandarizacion variable charges
asegurados$charges<- round(asegurados$charges, 2)

#Se comprueba que se ha realizado el cambio
head(asegurados, 3)
```

```
##   age    sex   bmi children smoker    region  charges
## 1  19 female 27.90        0    yes southwest 16884.92
## 2  18   male 33.77        1     no southeast  1725.55
## 3  28   male 33.00        3     no southeast  4449.46
```

A continuación se va a comprobar si se cumple el tipo de variable estadística que debe tener asociada cada variable:

Las variables **sex**, **smoker**, y **region** deben ser de tipo factor (cualitativa nominal), ya que no tienen un criterio de orden; las variables **age** y **children** deben ser de tipo integer, ya que contienen valores discretos; y las variables **bmi** y **charges** deben ser de tipo numeric, ya que la naturaleza de estas variables es continua.

```r
#Se muestran los tipos de variables
sapply(asegurados, class)
```

```
##       age       sex       bmi  children    smoker    region   charges
## "integer"  "factor" "numeric" "integer"  "factor"  "factor" "numeric"
```

Se observa que todas las variables tienen asignado el tipo apropiado, y por lo tanto no necesitan una conversión para conseguir que el tipo final sea el adecuado.

En cuanto a la selección de variables, todas son de interés para el estudio por lo que no prescindiré de ninguna.

3. Limpieza de los datos.

A continuación se va a llevar a cabo la limpieza de los datos. Este paso es muy importante en cualquier análisis de datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En primer lugar se va a buscar si los datos contienen elementos vacíos:

```r
#Se busca que variable tiene elementos vacios
sapply(asegurados, function(x) sum(is.na(x)))
```

```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

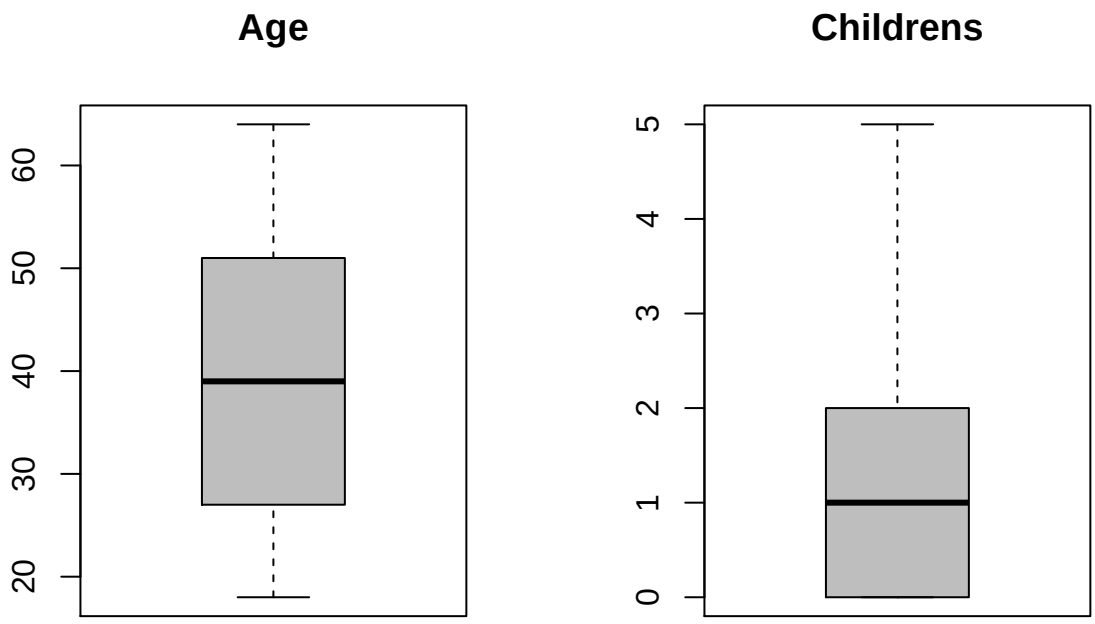
Se observa que ninguna de las variables contiene elementos vacíos, por tanto no habría que seguir ningún procedimiento adicional.

En el caso de que si que hubiese registros desconocidos una decisión sería eliminar dichos registros, siempre y cuando fuese una cantidad que no afectase a nuestra investigación; otro de los procedimientos, que sería el que yo utilizaría para gestionarlo, sería imputar los valores faltantes a partir de los k-vecinos más cercanos, utilizando por ejemplo la distancia de Gower.

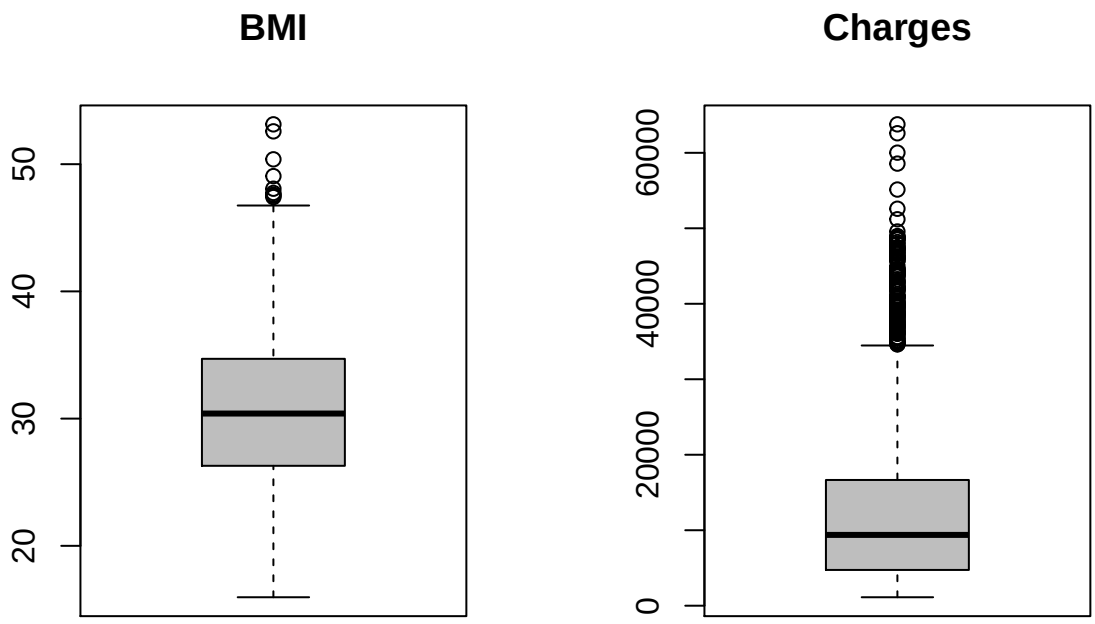
3.2 Identificación y tratamiento de valores extremos

Para identificar los valores extremos voy a presentar un boxplot para cada variable cuantitativa. En este dataset las variables cuantitativas que tenemos son: **age**, **children**, **bmi**, y **charges**.

```r
par(mfrow = c(1,2))
boxplot(asegurados$age, main='Age', col = 'grey')
boxplot(asegurados$children, main='Childrens', col = 'grey')
```




```
boxplot(asegurados$bmi, main='BMI', col = 'grey')
boxplot(asegurados$charges, main='Charges', col = 'grey')
```



Se encuentran varios valores atípicos en las variables **bmi** y **charges**. Por tanto voy a inspeccionar dichos valores de manera más concreta:

```
#listado de valores atípicos en bmi
boxplot.stats(asegurados$bmi)$out
```

```
## [1] 49.06 48.07 47.52 47.41 50.38 47.60 52.58 47.74 53.13
```

```
length(boxplot.stats(asegurados$bmi)$out)
```

```
## [1] 9
```

```
#listado de valores atípicos en charges
boxplot.stats(asegurados$charges)$out
```

```
##   [1] 39611.76 36837.47 37701.88 38711.00 35585.58 51194.56 39774.28
##   [8] 48173.36 38709.18 37742.58 47496.49 37165.16 39836.52 43578.94
##  [15] 47291.06 47055.53 39556.49 40720.55 36950.26 36149.48 48824.45
##  [22] 43753.34 37133.90 34779.61 38511.63 35160.13 47305.31 44260.75
##  [29] 41097.16 43921.18 36219.41 46151.12 42856.84 48549.18 47896.79
##  [36] 42112.24 38746.36 42124.52 34838.87 35491.64 42760.50 47928.03
##  [43] 48517.56 41919.10 36085.22 38126.25 42303.69 46889.26 46599.11
##  [50] 39125.33 37079.37 35147.53 48885.14 36197.70 38245.59 48675.52
##  [57] 63770.43 45863.21 39983.43 45702.02 58571.07 43943.88 39241.44
```

```
##  [64] 42969.85 40182.25 34617.84 42983.46 42560.43 40003.33 45710.21
##  [71] 46200.99 46130.53 40103.89 34806.47 40273.65 44400.41 40932.43
##  [78] 40419.02 36189.10 44585.46 43254.42 36307.80 38792.69 55135.40
##  [85] 43813.87 39597.41 36021.01 45008.96 37270.15 42111.66 40974.16
##  [92] 46113.51 46255.11 44202.65 48673.56 35069.37 39047.29 47462.89
##  [99] 38998.55 41999.52 41034.22 36580.28 35595.59 42211.14 44423.80
## [106] 37484.45 39725.52 44501.40 39727.61 48970.25 39871.70 34672.15
## [113] 41676.08 44641.20 41949.24 36124.57 38282.75 46661.44 40904.20
## [120] 36898.73 52590.83 40941.29 39722.75 37465.34 36910.61 38415.47
## [127] 41661.60 60021.40 47269.85 49577.66 37607.53 47403.88 38344.57
## [134] 34828.65 62592.87 46718.16 37829.72 36397.58 43896.38
```

```
length(boxplot.stats(asegurados$charges)$out)
```

```
## [1] 139
```

Se observa que hay 9 valores extremos en la variable **bmi**, y 139 valores extremos de la variable **charges**, de un total de 1338 observaciones. Dichos registros no parecen erroneos ya que presentan valores posibles y coherentes de distintos asegurados, como son índices de masa corporal muy elevados (que pueden darse en un individuo con obesidad), y primas de seguro muy elevadas (que es posible que se den en individuos con factores muy marcados).

Una vez realizada la limpieza de los datos se va a proceder a su análisis.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Voy a realizar una agrupacion por condicion de fumador y género, que me será util a la hora de analizar los datos.

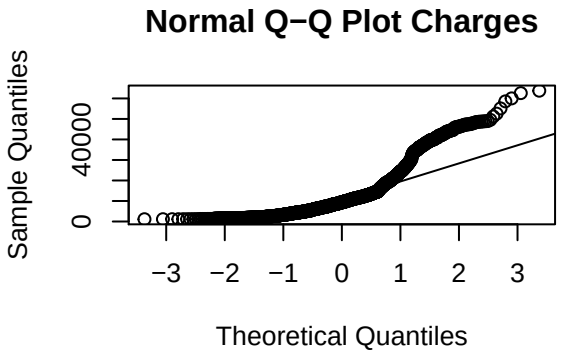
```
asegurados.fumadores <- asegurados[asegurados$smoker=='yes',]
asegurados.nofumadores <- asegurados[asegurados$smoker=='no',]

asegurados.mujer <- asegurados[asegurados$sex=='female',]
asegurados.hombre <- asegurados[asegurados$sex=='male',]
```

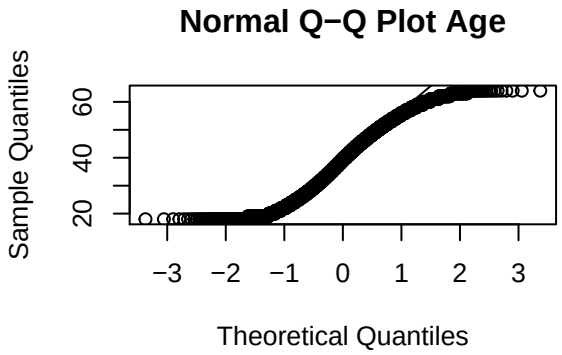
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

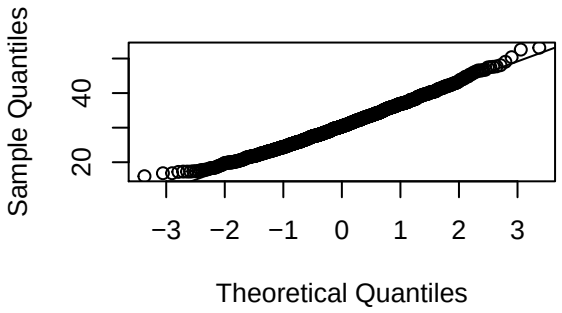
Para comprobar la normalidad de la varianza, en primer lugar voy a representar los gráficos **cuantil-cuantil** para obtener una primera idea de como son las distribuciones y, a continuación, voy a applicar el **test de Shapiro Wilk** en las variables cuantitativas (**age**, **children**, **bmi**, y **charges**) para comprobarlo de forma definitiva.

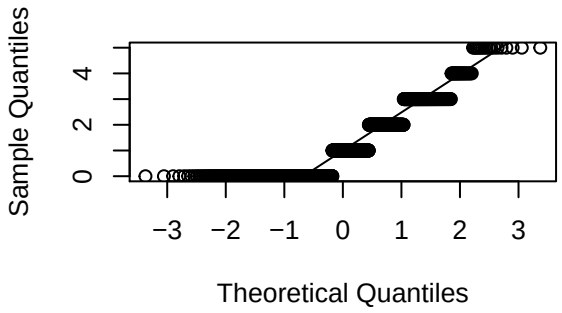
```
par(mfrow = c(2,2))
qqnorm(asegurados$charges, main='Normal Q-Q Plot Charges')
qqline(asegurados$charges)
qqnorm(asegurados$age, main='Normal Q-Q Plot Age')
qqline(asegurados$age)
```

```
qqnorm(asegurados$bmi, main='Normal Q-Q Plot BMI')
qqline(asegurados$bmi)
qqnorm(asegurados$children, main='Normal Q-Q Plot Children')
qqline(asegurados$children)
```










```
asegurados.cuantit <- asegurados[, c(1,3,4,7)]

alpha = 0.05
for(i in 1:length(asegurados.cuantit)){
  if(shapiro.test(asegurados.cuantit[,i])$p.value<alpha){
    cat('Se rechaza la hipótesis nula. La variable', names(asegurados.cuantit)[i],
        'no sigue una distribución normal.\n')
  }
  else{
    cat('No se rechaza la hipótesis nula. La variable', names(asegurados.cuantit)[i],
        'sigue una distribución normal.\n')
  }
}
```

```
## Se rechaza la hipótesis nula. La variable age no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable bmi no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable children no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable charges no sigue una distribución normal.
```

Aunque en algún gráfico se observa que la distribución no se aleja mucho de la normal, tras las pruebas de

Shapiro Wilk podemos comprobar que ninguno de nuestros datos sigue una distribución normal, por tanto ya debemos aplicar pruebas no paramétricas para realizar el análisis.

A continuación voy a comprobar también la homogeneidad de la varianza respecto a la prima del seguro, y para ello voy a utilizar el **test Fligner-Killeen**:

```
fligner.test(charges ~ age, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by age
## Fligner-Killeen:med chi-squared = 78.585, df = 46, p-value =
## 0.001952
```

```
fligner.test(charges ~ sex, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by sex
## Fligner-Killeen:med chi-squared = 9.4445, df = 1, p-value =
## 0.002118
```

```
fligner.test(charges ~ bmi, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by bmi
## Fligner-Killeen:med chi-squared = 610.72, df = 536, p-value =
## 0.01375
```

```
fligner.test(charges ~ children, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by children
## Fligner-Killeen:med chi-squared = 22.198, df = 5, p-value =
## 0.0004801
```

```
fligner.test(charges ~ smoker, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by smoker
## Fligner-Killeen:med chi-squared = 238.15, df = 1, p-value <
## 2.2e-16
```

```r
fligner.test(charges ~ region, data=asegurados)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by region
## Fligner-Killeen:med chi-squared = 19.233, df = 3, p-value =
## 0.0002447
```

En todos los casos rechazamos la hipótesis nula de que las varianzas sean iguales, al tener un p-valor inferior a 0.05.

Como he comentado, se deben aplicar pruebas no paramétricas para realizar el análisis de los datos. Aún así como el tamaño de nuestra muestra es superior a 30 observaciones, por el Teorema Central del Límite, se podría asumir la aproximación a una distribución normal.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

En primer lugar voy a analizar la correlación entre variables cuantitativas a partir del **coeficiente de Spearman**. De esta forma podemos identificar cuáles con las variables más correlacionadas con el importe de la prima del seguro.

```r
cor.test(asegurados$charges, asegurados$age, method = 'spearman')
```

```
## Warning in cor.test.default(asegurados$charges, asegurados$age, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  asegurados$charges and asegurados$age
## S = 185880000, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5343921
```

```r
cor.test(asegurados$charges, asegurados$bmi, method = 'spearman')
```

```
## Warning in cor.test.default(asegurados$charges, asegurados$bmi, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  asegurados$charges and asegurados$bmi
## S = 351560000, p-value = 1.192e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1193996
```

```r
cor.test(asegurados$charges, asegurados$children, method = 'spearman')
```

```
## Warning in cor.test.default(asegurados$charges, asegurados$children, method
## = "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  asegurados$charges and asegurados$children
## S = 345990000, p-value = 9.847e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1333389
```

Se observa que la variable más relevante en la fijación de la prima del seguro es la edad.

A continuación voy a realizar dos contraste de hipótesis para las dos agrupaciones realizadas anteriormente (fumadores-NoFumadores, Hombres-Mujeres). Como se ha visto, nuestras muestras no provienen de poblaciones normales, pero como los tamaños de los grupos son superiores a 30, por el Teorema Central del Límite, podremos contrastar la diferencia de medias a partir del estadístico t.

Por un lado voy a contrastar el hecho de que el valor medio de las primas del seguro de los fumadores y no fumadores sea igual, con el de que el valor medio de las primas del seguro de los no fumadores sea inferior al valor medio de las primas del seguro de los fumadores.

$$\begin{cases} H_0 : \mu_{fumNo} = \mu_{fumSi} \rightarrow \mu_{fumNo} - \mu_{fumSi} = 0 \\ H_1 : \mu_{fumNo} < \mu_{fumSi} \rightarrow \mu_{fumNo} - \mu_{fumSi} < 0 \end{cases}$$

```r
t.test(asegurados.nofumadores$charges, asegurados.fumadores$charges, mu=0,
       conf.level = 0.95, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  asegurados.nofumadores$charges and asegurados.fumadores$charges
## t = -32.752, df = 311.85, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -22426.4
## sample estimates:
## mean of x mean of y
##  8434.268 32050.232
```

Se rechaza la hipótesis nula de que las primas del seguro médico para los fumadores y los no fumadores sean iguales, con un nivel de confianza del 95%, a favor de la hipótesis alternativa de que la prima del seguro es inferior para los no fumadores.

Por otro lado se contrasta el hecho de que el valor medio de las primas del seguro para los hombres y para las mujeres sea igual, con el hecho de que el valor medio de las primas del seguro para los hombres sea inferior al valor medio de las primas del seguro para las mujeres.

$$\begin{cases} H_0 : \mu_H = \mu_M \rightarrow \mu_H - \mu_M = 0 \\ H_1 : \mu_H < \mu_M \rightarrow \mu_H - \mu_M < 0 \end{cases}$$

```r
t.test(asegurados.hombre$charges, asegurados.mujer$charges, mu=0,
       conf.level = 0.95, alternative ='less')
```

```
##
##  Welch Two Sample t-test
##
## data:  asegurados.hombre$charges and asegurados.mujer$charges
## t = 2.1009, df = 1313.4, p-value = 0.9821
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2474.002
## sample estimates:
## mean of x mean of y
##  13956.75  12569.58
```

En este caso, como el p-valor es superior al nivel de significación fijado, no se puede rechazar la hipótesis nula de que las primas del seguro médico para hombres y mujeres sean iguales, con un nivel de confianza del 95%.

Por último, como el objetivo del análisis es predecir la prima del seguro a partir de las variables que tenemos, voy a estimar por mínimos cuadrados ordinarios un modelo lineal que explique el importe de la prima de los seguros médicos a partir de diferentes regresores.

```r
#Se definen nuevas variables dicotómicas fijando la categoría de referencia
asegurados$sexR <- relevel(asegurados$sex, ref = 'female')
asegurados$smokerR <- relevel(asegurados$smoker, ref = 'yes')
asegurados$regionR <- relevel(asegurados$region, ref = 'southwest')

#Se construyen los modelos de regresión lineal
regModel.1 <- lm(charges~age+factor(sexR)+bmi+children+factor(smokerR)+factor(regionR),
                 data = asegurados)
summary(regModel.1)
```

```
##
## Call:
## lm(formula = charges ~ age + factor(sexR) + bmi + children +
##     factor(smokerR) + factor(regionR), data = asegurados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11305.2  -2848.3   -982.2   1392.6  29992.7
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                10950.49    1069.97  10.234  < 2e-16 ***
## age                          256.86      11.90  21.587  < 2e-16 ***
## factor(sexR)male            -131.30     332.95  -0.394 0.693390
## bmi                          339.18      28.60  11.859  < 2e-16 ***
## children                     475.51     137.81   3.451 0.000577 ***
## factor(smokerR)no         -23848.60     413.16 -57.723  < 2e-16 ***
## factor(regionR)northeast     960.00     477.94   2.009 0.044778 *
## factor(regionR)northwest     606.99     477.21   1.272 0.203610
## factor(regionR)southeast     -74.92     470.64  -0.159 0.873538
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
regModel.2 <- lm(charges~age+bmi+children+factor(smokerR),
                 data = asegurados)
summary(regModel.2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smokerR),
##     data = asegurados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11898.1  -2920.6   -986.6   1392.5  29509.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11709.11     991.37  11.811  < 2e-16 ***
## age                  257.85      11.90  21.674  < 2e-16 ***
## bmi                  321.84      27.38  11.755  < 2e-16 ***
## children             473.51     137.79   3.436 0.000608 ***
## factor(smokerR)no -23811.47     411.22 -57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

```
regModel.3 <- lm(charges~age+children+factor(smokerR)+factor(regionR),
                 data = asegurados)
summary(regModel.3)
```

```
##
## Call:
## lm(formula = charges ~ age + children + factor(smokerR) + factor(regionR),
##     data = asegurados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15656.5  -1939.2  -1302.9   -360.2  28887.2
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               20533.42     710.25  28.910  < 2e-16 ***
## age                         273.29      12.42  22.012  < 2e-16 ***
## children                    496.86     144.77   3.432 0.000617 ***
## factor(smokerR)no        -23780.51     432.89 -54.934  < 2e-16 ***
```

```
## factor(regionR)northeast    484.17     500.44   0.967 0.333478
## factor(regionR)northwest    138.22     499.72   0.277 0.782142
## factor(regionR)southeast    873.98     487.35   1.793 0.073144 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6370 on 1331 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.7233
## F-statistic: 583.5 on 6 and 1331 DF,  p-value: < 2.2e-16
```

```
regModel.4 <- lm(charges~bmi+children+factor(smokerR)+factor(regionR),
                 data = asegurados)
summary(regModel.4)
```

```
##
## Call:
## lm(formula = charges ~ bmi + children + factor(smokerR) + factor(regionR),
##     data = asegurados)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15149  -4634   -839   3583  31989
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                18532.99    1157.98  16.005  < 2e-16 ***
## bmi                          410.08      32.96  12.441  < 2e-16 ***
## children                     595.39     159.92   3.723 0.000205 ***
## factor(smokerR)no         -23629.82     478.43 -49.391  < 2e-16 ***
## factor(regionR)northeast    1030.31     555.14   1.856 0.063682 .
## factor(regionR)northwest     639.70     554.31   1.154 0.248687
## factor(regionR)southeast    -378.71     546.44  -0.693 0.488406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7042 on 1331 degrees of freedom
## Multiple R-squared:  0.6634, Adjusted R-squared:  0.6619
## F-statistic: 437.2 on 6 and 1331 DF,  p-value: < 2.2e-16
```

```
#Se comprueba cuál es el mejor modelo
cat("R2 modelo 1:", summary(regModel.1)$r.squared, "\nR2 modelo 2:",
    summary(regModel.2)$r.squared, "\nR2 modelo 3:", summary(regModel.3)$r.squared,
    "\nR2 modelo 4:", summary(regModel.4)$r.squared)
```

```
## R2 modelo 1: 0.7509089
## R2 modelo 2: 0.7496906
## R2 modelo 3: 0.7245445
## R2 modelo 4: 0.6634081
```

Para decidir cuál es el mejor modelo nos debemos fijar en el coeficiente R cuadrado. En este caso el mejor modelo sería el **RegModel.1** al tener el coeficiente de determinación más elevado. Por tanto, voy a visualizar de nuevo las características del modelo:

```r
summary(regModel.1)
```

```
##
## Call:
## lm(formula = charges ~ age + factor(sexR) + bmi + children +
##     factor(smokerR) + factor(regionR), data = asegurados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11305.2  -2848.3   -982.2   1392.6  29992.7
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                10950.49    1069.97  10.234  < 2e-16 ***
## age                          256.86      11.90  21.587  < 2e-16 ***
## factor(sexR)male            -131.30     332.95  -0.394 0.693390
## bmi                          339.18      28.60  11.859  < 2e-16 ***
## children                     475.51     137.81   3.451 0.000577 ***
## factor(smokerR)no         -23848.60     413.16 -57.723  < 2e-16 ***
## factor(regionR)northeast    960.00     477.94   2.009 0.044778 *
## factor(regionR)northwest    606.99     477.21   1.272 0.203610
## factor(regionR)southeast    -74.92     470.64  -0.159 0.873538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

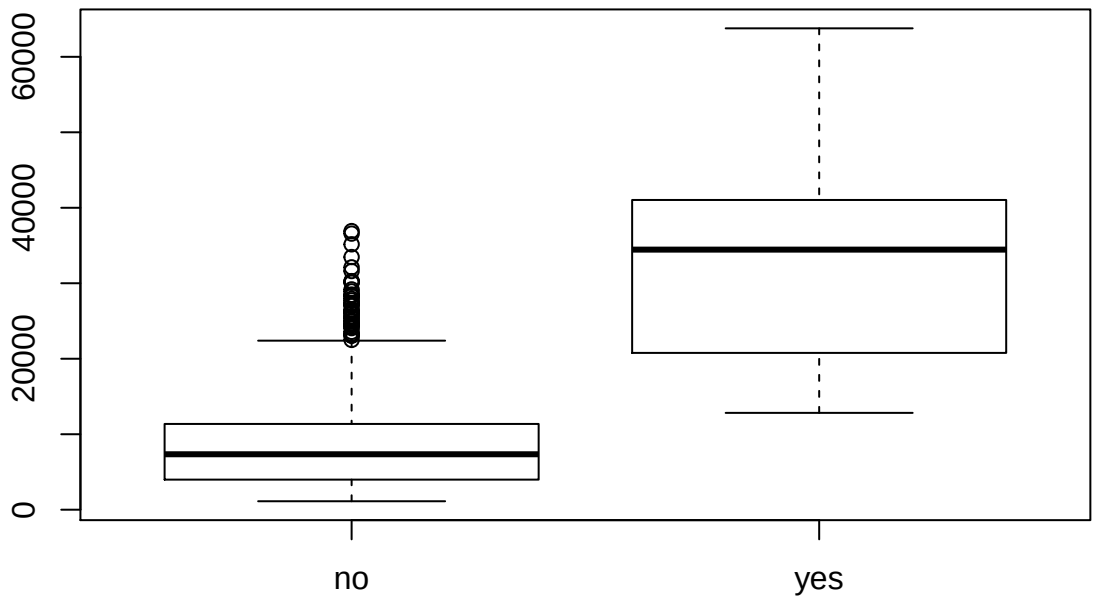
En cuanto a la interpretación de los coeficientes, el intercepto (10887.21) representa la estimación del importe de la prima del seguro médico cuando todas las varibales independientes sean 0. Por otro lado, los coeficientes asociados a las variables independientes representan el incremento/reducción del importe del seguro cuando dicha variable aumenta/disminuye una unidad y las demás se mantienen constantes.

Centrandonos en el modelo 1 se puede decir que al aumentar la edad (y el resto de variables constantes), aumenta el importe del seguro. Lo mismo ocurre con el índice de masa corporal, y el número de hijos. Por otro lado, el importe de la prima es superior en fumadores que en no fumadores, y se puede decir que es el factor más influyente. Por último, las variables asociadas a la región y al sexo no son significativas por lo que estadísticamente no podemos afirmar ninguna conclusión respecto a ellas.

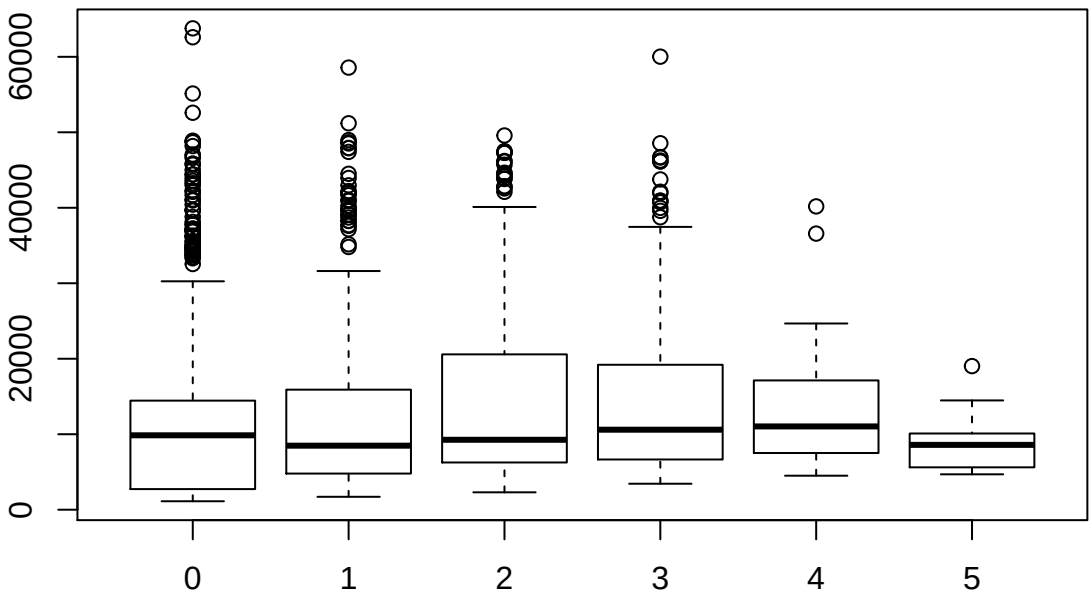
5. Representación de los resultados a partir de tablas y gráficas.

Mediante las siguientes gráficas podemos observar algunas de las conclusiones del modelo:

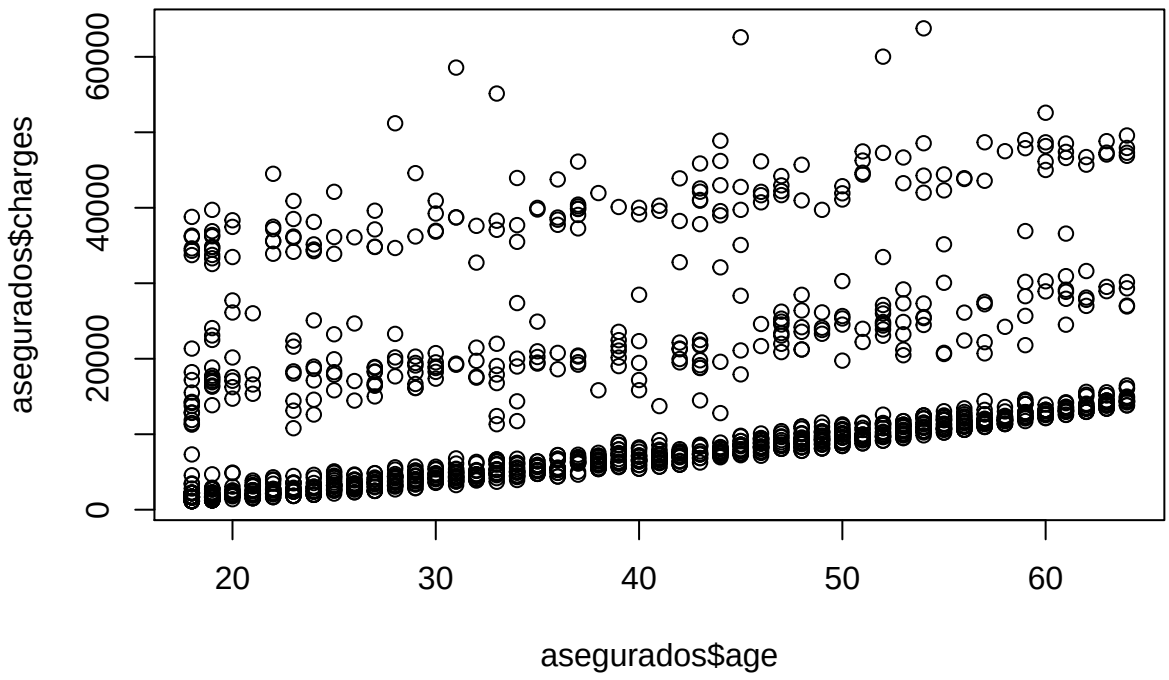
```r
boxplot(asegurados$charges~asegurados$smoker)
```



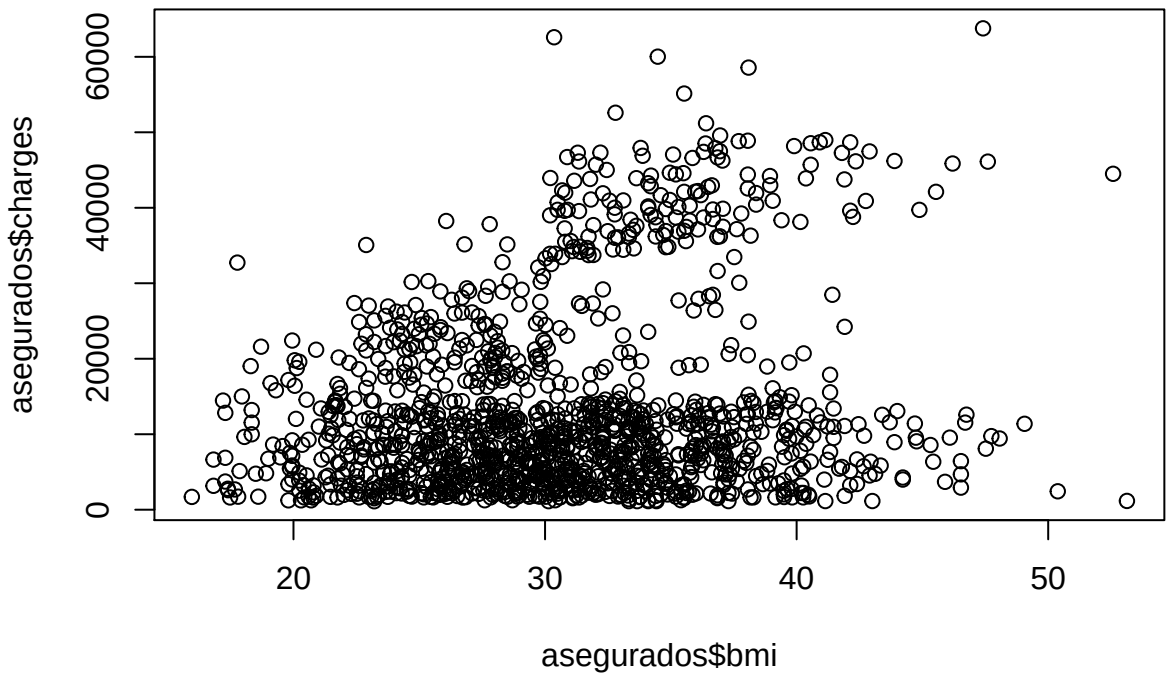

```
boxplot(asegurados$charges~asegurados$children)
```




```
plot(asegurados$age, asegurados$charges)
```




```
plot(asegurados$bmi, asegurados$charges)
```



Tras visualizar todas las graficas de los factores más influyentes podemos decir que el factor más marcado es la condición de fumador, y que las primas de los seguros se ajustan mucho a cada individuo particular en función de sus características.

Antes de finalizar se va a crear el archivo de datos corregido, con todos los cambios hasta el momento:

```
write.csv(asegurados, file = 'insurance_clean.csv', row.names = FALSE)
```

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Después de la limpieza, análisis, y representación de los datos se ha podido comprobar que las variables que más influyen en las primas de los seguros médicos son la edad, el índice de masa corporal, el número de hijos, y la condición de fumador o no fumador. Es posible pensar que un individuo fumador con un índice de masa corporal elevado tenga más riesgo de sufrir algún problema de salud, y por tando tenga que pagar un seguro más elevado. El análisis realizado lo confirma.

Se ha visto que el modelo de regresión lineal nos permite, mediante una serie de factores concretos de un individuo, predecir el importe de la prima de sus seguros de salud. Por tanto se puede concluir que el importe de los seguros de salud en EEUU dependen mucho de las caracteristicas personales de cada asegurado, sobre todo de la condición de fumador, y se ajustan mucho a ellos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código se encuentra en el archivo 'insurance.Rmd', que está disponible en este enlace de github