PRACTICA 2 - TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS

Jose Alvaro Rodriguez Botero

07/01/2020

# Detalles de la Practica

## Descripcion

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas.

Tendréis que entregar un solo archivo con el enlace Github <https://github.com> donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

* Ejemplo: <https://github.com/Bengis/nba-gap-cleaning>
* Ejemplo complejo (archivo adjunto).

## Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

* Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
* Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

## Objetivos

Los objetivos concretos de esta práctica son:

* Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
* Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
* Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
* Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
* Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
* Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
* Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

## Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle <https://www.kaggle.com>. Algunos ejemplos de dataset con los que podéis trabajar son:

* Red Wine Quality <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>
* Titanic: Machine Learning from Disaster <https://www.kaggle.com/c/titanic>

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos.
   * ¿Los datos contienen ceros o elementos vacíos?
   * ¿Cómo gestionarías cada uno de estos casos?
   * Identificación y tratamiento de valores extremos.
4. Análisis de los datos.
   * Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).
   * Comprobación de la normalidad y homogeneidad de la varianza.
   * 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. Aplicar al menos tres métodos de análisis diferentes.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
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.

# Desarrollo de la practica

## Descripción del dataset

El dataset objeto de análisis ha sido obtenido de <https://www.kaggle.com/mirichoi0218/insurance> y está constituido por 7 columnas (características) que presentan 1338 personas (registros).

Entre los campos de este dataset, encontramos los siguientes:

* age: Edad del asegurado.
* sex: Genero (female, male).
* bmi: Indice masa corporal. Idealmente entre 18.5 to 24.9
* children: Número de hijos cubiertos por el seguro de salud / Número de dependientes.
* smoker: Fumador
* region: Región donde vive el asegurado (northeast, southeast, southwest, northwest).
* charges: Cargos (prima del seguro).

A partir de este conjunto de datos se plantea el problema de determinar que variables influyen mas sobre el cargo (prima del seguro).

Con este conjunto de datos también se podrán realizar modelos de regresión que permitan predecir el valor de los cargos en funcion de sus características.

Finalmente, también se podrán utilizar contrastes de hipótesis que nos permitan identificar las caracteristicas más interesantes en las muestras.

Este ánalisis es muy relevante para el sector salud debido a que se pueden detectar características y patrones por grupos de usuarios y su relacion con los gastos medicos.

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

### Integración

Para este conjunto de datos no es necesario hacer ninguna fusión de datos, ya que nos provee la información suficiente para realizar los análisis

### Selección de datos

Inicialmente se consideran todos los atributos del conjunto de datos, solamente se realizaran (si son necesarias) algunas transformaciones.

## Limpieza de los datos.

Antes de realizar la limpieza de los datos, se hace la lectura del archivo (dataset) que está en formato CSV. Para ello se utiliza la función .

### Lectura del dataset

# Librerias a utilizar  
library(nortest)  
library(MVar)  
library(dplyr)  
library(pspearman)  
library(ggpubr)  
library(devtools)

insuranceDS <- read.csv('insurance.csv', stringsAsFactors = FALSE)  
head(insuranceDS)

## 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

### Tamaño del dataset

dim(insuranceDS)

## [1] 1338 7

### Tipo de dato asignado a cada campo

sapply(insuranceDS, class)

## age sex bmi children smoker region   
## "integer" "character" "numeric" "integer" "character" "character"   
## charges   
## "numeric"

### Resumen de datos

summary(insuranceDS)

## age sex bmi children   
## Min. :18.00 Length:1338 Min. :15.96 Min. :0.000   
## 1st Qu.:27.00 Class :character 1st Qu.:26.30 1st Qu.:0.000   
## Median :39.00 Mode :character 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   
## smoker region charges   
## Length:1338 Length:1338 Min. : 1122   
## Class :character Class :character 1st Qu.: 4740   
## Mode :character Mode :character Median : 9382   
## Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

### Ceros y elementos vacios

sapply(insuranceDS, function(x) sum(is.na(x)))

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

Se puede observar que este conjunto de datos no contiene elementos vacios, por lo que no es necesario realizar ningún tipo de reemplazo.

### Cambio de tipo de datos a atributos

En este punto lo que se desea hacer es factorizar las variables , y .

insuranceDS$sex <- as.factor(insuranceDS$sex)  
insuranceDS$smoker <- as.factor(insuranceDS$smoker)  
insuranceDS$region <- as.factor(insuranceDS$region)  
sapply(insuranceDS, function(x) sum(is.na(x)))

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

En este punto tambien nos puede interesar crear una categoria para clasificar a las personas segun la edad. Para ello es necesario crear una clasificación como la que se muestra en la siguiente tabla:

|  |  |
| --- | --- |
| Nombre | Rango de edad |
| Adulto Joven | >= 18 y < 40 |
| Adulto Maduro | >= 40 y < 60 |
| Adulto Mayor | >= 60 |

insuranceDS['adult\_type']<-ifelse(insuranceDS$age>=18 & insuranceDS$age<45, 1, ifelse(insuranceDS$age>=45 & insuranceDS$age<60, 2, 3))  
insuranceDS<-insuranceDS[,c("age", "adult\_type", "sex", "bmi", "children", "smoker", "region", "charges")]

### Normalizar / Estandarizar variables cuantitativas

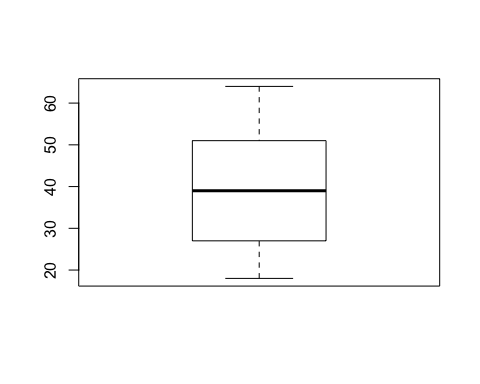
En este punto buscamos estandarizar el numero de decimales de las variables cuantitativas.

insuranceDS$bmi<-round(insuranceDS$bmi, digits = 3)  
insuranceDS$charges <- round(insuranceDS$charges, digits = 3)

### Valores extremos

Los valores extremos o son aquellos valores que no son congruentes con el resto de los valores de un atributo. A continuacion se analizará el dataset con el fin de identificar estos valores.

# Age  
boxplot(insuranceDS$age)

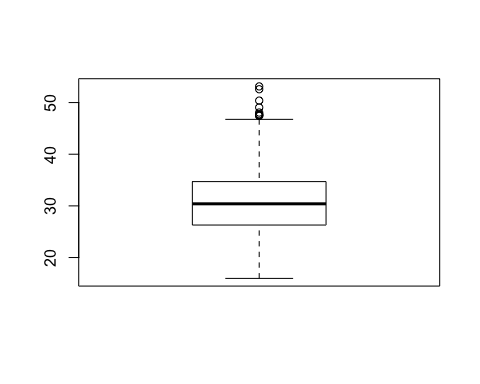


boxplot.stats(insuranceDS$age)$out

## integer(0)

El atributo no presenta outliers.

# bmi  
boxplot(insuranceDS$bmi)

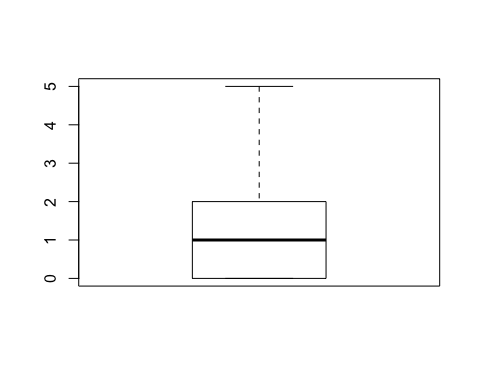


boxplot.stats(insuranceDS$bmi)$out

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

El atributo presentan 9 outliers, pero estos valores se pueden dar perfectamente en la realidad (valores del indice masa corporal mayores a 40). Por este motivo se decide dejar los valores tal como fueron recogidos.

# children  
boxplot(insuranceDS$children)



boxplot.stats(insuranceDS$children)$out

## integer(0)

El atributo no presenta outliers.

### Exportación de los datos preprocesados

Despues de realizar el proceso de limpieza de los datos del conjunto de datos, se procede a guardar en un nuevo conjunto de datos

write.csv(insuranceDS, "insurance\_data\_clean.csv")

## Análisis de los datos.

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

A continuación, se realizan algunas agrupaciones en nuestro conjunto de datos que pueden ser interesantes para anlizar.

#### Agrupacion por tipo de adulto

insuranceDS.adult\_1 <- insuranceDS[insuranceDS$adult\_type==1,]  
insuranceDS.adult\_2 <- insuranceDS[insuranceDS$adult\_type==2,]  
insuranceDS.adult\_3 <- insuranceDS[insuranceDS$adult\_type==3,]

#### Agrupacion por genero

insuranceDS.female <- insuranceDS[insuranceDS$sex=="female",]  
insuranceDS.male <- insuranceDS[insuranceDS$sex=="male",]

#### Agrupacion por fumador

insuranceDS.smoker\_y <- insuranceDS[insuranceDS$smoker=="yes",]  
insuranceDS.smoker\_n <- insuranceDS[insuranceDS$smoker=="no",]

#### Agrupacion por región

insuranceDS.region\_se <- insuranceDS[insuranceDS$region=="southeast",]  
insuranceDS.region\_sw <- insuranceDS[insuranceDS$region=="southwest",]  
insuranceDS.region\_ne <- insuranceDS[insuranceDS$region=="northeast",]  
insuranceDS.region\_nw <- insuranceDS[insuranceDS$region=="northwest",]

### Comprobación de la normalidad

Para comprobar que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de .

Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado . Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

|  |  |
| --- | --- |
| Atributo | p-valor |
| age | 5.692045710^{-22} |
| adult\_type | 8.789705110^{-44} |
| bmi | 2.604683410^{-5} |
| children | 5.066436410^{-36} |

Como los valores son menores que 5% entonces podemos rechazar la hipotesis nula.

### Homogeneidad de la varianza.

Ahora bien, pasamos a revisar la homogeneidad de varianzas mediante la aplicación del test de Fligner-Killeen. En los test la hipótesis nula consiste en que ambas varianzas son iguales.

Age

fligner.test(charges ~ age, data = insuranceDS)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: charges by age  
## Fligner-Killeen:med chi-squared = 78.583, df = 46, p-value = 0.001953

Sex

fligner.test(charges ~ sex, data = insuranceDS)

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

bmi

fligner.test(charges ~ bmi, data = insuranceDS)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: charges by bmi  
## Fligner-Killeen:med chi-squared = 623.92, df = 547, p-value = 0.01239

children

fligner.test(charges ~ children, data = insuranceDS)

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

smoker

fligner.test(charges ~ smoker, data = insuranceDS)

##   
## 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

region

fligner.test(charges ~ region, data = insuranceDS)

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

Puesto que obtenemos en todos los casos p-valor inferiores a 0,05, rechazamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

### Aplicación de pruebas estadísticas para comparar los grupos de datos.

#### ¿Que variables cuantitativas influyen mas en los gastos medicos?

Primero se realiza un un análisis de correlación entre las distintas variables para determinar cuales de ellas ejercen una mayor influencia sobre los gastos medicos. Para ello se usara el coeficiente de correlación de Spearman.

corr\_matrix <- matrix(nc = 2, nr = 0)  
colnames(corr\_matrix) <- c("estimate", "p-value")  
# Calcular el coeficiente de correlación para cada variable cuantitativa  
# con respecto al campo "charges"  
for (i in 1:(ncol(insuranceDS) - 1)) {  
 if (is.integer(insuranceDS[,i]) | is.numeric(insuranceDS[,i])) {  
 spearman\_test = cor.test(insuranceDS[,i],  
 insuranceDS[,length(insuranceDS)],  
 method = "spearman")  
 corr\_coef = spearman\_test$estimate  
 p\_val = spearman\_test$p.value  
 # Add row to matrix  
 pair = matrix(ncol = 2, nrow = 1)  
 pair[1][1] = corr\_coef  
 pair[2][1] = p\_val  
 corr\_matrix <- rbind(corr\_matrix, pair)  
 rownames(corr\_matrix)[nrow(corr\_matrix)] <- colnames(insuranceDS)[i]  
 }  
}  
  
print(corr\_matrix)

## estimate p-value  
## age 0.5343921 1.130692e-99  
## adult\_type 0.4698395 2.024420e-74  
## bmi 0.1193959 1.192606e-05  
## children 0.1333389 9.846806e-07

Así, identificamos cuáles son las variables más correlacionadas con el gasto medico en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, se puede observar que la variable más relevante es la edad ().

#### ¿El gasto medico es superior para los fumadores?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si los gastos medicos son superiores dependiendo si la persona fuma o no.

Para ello, tendremos dos muestras:

* Gastos fumadores
* Gastos No fumadores

insuranceDS.smy.charges <-insuranceDS[insuranceDS$smoker == "yes",]$charges  
insuranceDS.smn.charges <-insuranceDS[insuranceDS$smoker == "no",]$charges

Así, se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

donde μ1 es la media de la población de la que se extrae la primera muestra y es la media de la población de la que extrae la segunda. Así, tomaremos .

t.test(insuranceDS.smy.charges, insuranceDS.smn.charges, alternative="two.sided",conf.level=0.95, correct = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: insuranceDS.smy.charges and insuranceDS.smn.charges  
## t = 32.752, df = 311.85, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 22197.21 25034.71  
## sample estimates:  
## mean of x mean of y   
## 32050.232 8434.268

Puesto que obtenemos un p-valor menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, efectivamente, los gastos hospitalarios de una persona son superiores esta fuma.

#### Modelo de regresión lineal

Ahora apuntando a uno de los objetivos de la actividad, resulta de mucho interes poder hacer predicciones sobre los gastos medicos dadas sus caracteristicas.

Por lo tanto se calculara un modelo de regresión lineal utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar las predicciones de los gastos medicos. Para obtener un modelo de regresión lineal eficiente.

Se utilizaran las variables mas correlacionadas obtenidas anteriormente y se escogerá el mejor modelo utilizando como criterio aquel que presente un mayor coeficiente de determinación (R^2).

# Regresores cuantitativos con mayor coeficiente  
# de correlación con respecto al precio  
age = insuranceDS$age  
adult\_type = insuranceDS$adult\_type  
children = insuranceDS$children  
bmi = insuranceDS$bmi

# Regresores cualitativos  
sex = insuranceDS$sex  
smoker = insuranceDS$smoker  
region = insuranceDS$region

# Variable a predecir  
charges = insuranceDS$charges

# Generación de varios modelos  
modelo1 <- lm(charges ~ smoker + sex + region + age + children + bmi, data = insuranceDS)  
modelo2 <- lm(charges ~ smoker + region + age + children + bmi, data = insuranceDS)  
modelo3 <- lm(charges ~ smoker + sex + age + children + bmi, data = insuranceDS)  
modelo4 <- lm(charges ~ sex + region + age + children + bmi, data = insuranceDS)  
modelo5 <- lm(charges ~ region + age + children + bmi, data = insuranceDS)  
modelo6 <- lm(charges ~ smoker + sex + age, data = insuranceDS)  
modelo7 <- lm(charges ~ smoker + region + age, data = insuranceDS)  
modelo8 <- lm(charges ~ smoker + age, data = insuranceDS)  
  
tabla.coeficientes <- matrix(c(1,summary(modelo1)$r.squared,2,summary(modelo2)$r.squared,3,summary(modelo3)$r.squared,4, summary(modelo4)$r.squared,5, summary(modelo5)$r.squared,6, summary(modelo6)$r.squared,7, summary(modelo7)$r.squared,8, summary(modelo8)$r.squared),ncol = 2, byrow = TRUE)  
  
colnames(tabla.coeficientes) <- c("Modelo", "R^2")  
tabla.coeficientes

## Modelo R^2  
## [1,] 1 0.7509130  
## [2,] 2 0.7508839  
## [3,] 3 0.7497225  
## [4,] 4 0.1264211  
## [5,] 5 0.1234960  
## [6,] 6 0.7214122  
## [7,] 7 0.7221068  
## [8,] 8 0.7214008

En este caso, tenemos que el primer modelo es el más conveniente dado que tiene un mayor coeficiente de determinación.

## Representación de los resultados a partir de tablas y gráficas

## Resolución del problema. A partir de los resultados obtenidos

Ahora, empleando este modelo, podemos proceder a realizar predicciones de gastos medicos como se muetra en los siguientes ejemplos:

# Ejemplo 1   
insuranceDS\_New1 <- data.frame(  
age = 25,  
sex = "male",  
bmi = 26.220,  
children = 0,  
smoker = "no",  
region = "northeast"  
)

# Ejemplo 2   
insuranceDS\_New2 <- data.frame(  
age = 25,  
sex = "male",  
bmi = 26.220,  
children = 0,  
smoker = "yes",  
region = "northeast"  
)

# Ejemplo 3   
insuranceDS\_New3 <- data.frame(  
age = 52,  
sex = "female",  
bmi = 30.780,  
children = 1,  
smoker = "no",  
region = "northeast"  
)

# Ejemplo 4   
insuranceDS\_New4 <- data.frame(  
age = 52,  
sex = "female",  
bmi = 30.780,  
children = 1,  
smoker = "yes",  
region = "northeast"  
)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No. Pred | age | sex | bmi | children | smoker | region | Pred |
| 1 | 25 | male | 26.220 | 0 | no | northeast | 3245.2082422 |
| 1 | 25 | male | 26.220 | 0 | yes | northeast | 2.709374310^{4} |
| 1 | 52 | female | 30.780 | 0 | no | northeast | 1.233386710^{4} |
| 1 | 52 | female | 30.780 | 0 | yes | northeast | 3.618240110^{4} |

NOTA: a lo largo del desarrollo de está practica se han estado elaborando tablas y graficos para los respectivos análisis.

### Conclusiones

Inicialmente, se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Para el primer caso no se encontraron valores perdidos (vacios o ceros) por lo que no hubo necesidad de realizar ningún tipo de conversion o ajuste. Para el segundo caso, se ha optado por incluir los valores extremos en los análisis dado que parecen no resultar del todo atípicos si los comparamos con los valores que toman las correspondientes variables, de hecho solo se dan valores atipicos en la variable (índice masa corporal).

En la presente practica, se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas a información de asegurados al sistema de salud y en el que se planteo principalmente como objetivo predecir los cargos al seguro medico.

Para cada una de estás pruebas, se ha podido observar cuáles son los resultados que arrojan (entre otros, mediante tablas) y qué conocimientos pueden extraerse a partir de ellas como se describe a continuación:

* Análisis de correlación y contraste de hipótesis: Nos ha permitido conocer cuáles de estas variables ejercen una mayor influencia sobre los cargos al seguro medico
* Modelo de regresión lineal: Resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas. En este punto se encontro que la variable es determinante a la hora de elegir el modelo y al momento de las predicciones.

### Los resultados permiten responder al problema?

Si, el modelo permite predecir a partir de las caracteristicas los gastos medicos. Se pudo observar ademas que la variable es determinante al momento de elegir el modelo y al momento de hacer las predicciones.

# Recursos

1. Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.
2. Vegas, E. (2017). Preprocesamiento de datos. Material UOC.
3. Gibergans, J. (2017). Regresión lineal múltiple. Material UOC.
4. Rovira, C. (2008). Contraste de hipótesis. Material UOC.