Tipología y ciclo de vida de los datos

Práctica 2

Universitat Oberta de Catalunya

Javier Martínez Arellano

Diciembre
2018

Índice

PRÁCTICA 2

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?
El conjunto de datos que se usará este caso práctico, obtenido del siguiente enlace *https://www.kaggle.com/spscientist/students-performance-in-exams*, recoge las notas de distintas pruebas (matemáticas, lectura y escritura) de un total de 1000 personas, así como información complementaria de cada individuo como género, raza/etnia, etc.

El dataset se compone de los siguientes características:
Gender: Género (male/female)
race/ethnicity: raza o etnia
parental level of education: Nivel de estudios de los padres (high school, master's degree,…)
lunch: si tienen costeada la comida (free/reduced, standard)
test preparation course: Si han realizado un test de preparación (completed/none)
math score: nota en la prueba de matemáticas.
reading score: nota en la prueba de lectura.
writing score: nota en la prueba escrita

Mediante el tratamiento de estos datos se pretende responder a si las mujeres obtienen mejores resultados en esta prueba de matemáticas así como averiguar si la nota de esta prueba guarda alguna relación con la notas de alguna de las otras dos, intentando encontrar un modelo para predecir la nota de matemáticas mediante las otras dos (previamente realizadas: lectura y escritura).
Este tipo de información podría ser requerida por el centro que realiza las pruebas a modo de control de los estudiantes, de las materias examinadas y control de becas, por ejemplo.

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

Usaremos este dataset como única fuente de datos. Comenzaremos con la lectura de los datos, previsualizando algunos valores de cada variable (columnas), comprobando el tipo de cada variable y pasaremos los enteros a numéricos.

```
datos <- read.csv("Students.csv")
summary(datos)
```

```
   gender     race.ethnicity     parental.level.of.education          lunch
female:518   group A: 89    associate's degree:222          free/reduced:355
male  :482   group B:190    bachelor's degree :118          standard    :645
             group C:319    high school       :196
             group D:262    master's degree   : 59
             group E:140    some college      :226
                            some high school  :179
test.preparation.course   math.score     reading.score    writing.score
completed:358           Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
none     :642           1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
                        Median : 66.00   Median : 70.00   Median : 69.00
                        Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
                        3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
                        Max.   :100.00   Max.   :100.00   Max.   :100.00
```

Con este resumen de los datos, vemos los distintos valores de las variables cualitativas y los valores extremos de las cuantitativas. Así, sin profundizar, ya vemos un nota = 0 en la prueba de matemáticas que nos indica que tendremos que revisar este valor para ver qué significado tiene, y una posible actuación, si se realiza alguna, ya que no debería dar esta nota si se presenta a la prueba.

```
library(knitr)
tipos <- sapply(datos, class)
kable(data.frame(variables=names(tipos),clase=as.vector(tipos)))
```

```
|variables                   |clase   |
|:---------------------------|:-------|
|gender                      |factor  |
|race.ethnicity              |factor  |
|parental.level.of.education |factor  |
|lunch                       |factor  |
|test.preparation.course     |factor  |
|math.score                  |integer |
|reading.score               |integer |
|writing.score               |integer |
```

```
datos[6:8] <- lapply(datos[6:8], as.numeric)
kable(data.frame(variables=names(tipos),clase=as.vector(tipos)))
```

```
|variables                   |clase   |
|:---------------------------|:-------|
|gender                      |factor  |
|race.ethnicity              |factor  |
|parental.level.of.education |factor  |
|lunch                       |factor  |
|test.preparation.course     |factor  |
|math.score                  |numeric |
|reading.score               |numeric |
|writing.score               |numeric |
```

A continuación, se eliminarán las variables que no se usarán en el análisis de los datos: Educación de los padres, comida y realización del test

```
#Eliminación de datos irrelevantes
datos <- datos[,-(3:5)]
summary(datos)
```

```
   gender    race.ethnicity   math.score     reading.score    writing.score
female:518   group A: 89    Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
male  :482   group B:190    1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
             group C:319    Median : 66.00   Median : 70.00   Median : 69.00
             group D:262    Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
             group E:140    3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
                            Max.   :100.00   Max.   :100.00   Max.   :100.00
```

3. Limpieza de los datos.

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

```
#CEROS Y ELEMENTOS VACIOS
> sapply(datos, function(x) sum(is.na(x)))
gender race.ethnicity     math.score  reading.score  writing.score
   0              0              0              0              0
```

La muestra no contiene elementos vacíos, en cambio, sí contiene valores 0 en las pruebas que no se han realizado o no se ha informado la nota. Como se veía en la vista previa de los datos, existía algún registro en la prueba de matemáticas con valor 0.

En este caso, se procederá a eliminar estos registros ya que una nota no informada no es útil para este análisis.

```
➢ #Borrar registros con resultado = 0 en alguna prueba
➢ nrow(datos)
[1] 1000
➢ datos <- datos[!datos$math.score==0,]
➢ nrow (datos)
[1] 999
```

Se ha eliminado 1 registro.

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos son aquellos que se distancian mucho del resto dando a pensar que no se han obtenido de la misma manera que el resto o que pueden ser incorrectos. Vamos a analizar estos valores en las variables cuantitativas.

```
#OUTLIERS
➢ boxplot.stats(datos$math.score)$out
[1] 18 22 24 26 19 23  8
➢ boxplot.stats(datos$reading.score)$out
[1] 17 26 28 23 24 24
➢ boxplot.stats(datos$writing.score)$out
[1] 10 22 19 15 23
```

Una vez revisado los valores 0 en los resultados de las notas, el resto de valores son perfectamente válidos aunque haya valores más distanciados de los más comunes, por lo que no se procederá a su tratamiento.

Después del tratamiento de los datos, antes del análisis, podemos guardar el conjunto resultante en el mismo formato que el dataset original.

```
➢ write.csv(datos, file = "ResultadoNotas.csv" )
```

4. Análisis de los datos.

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

Se procede a seleccionar los datos que pueden ser útiles en el análisis..

```
#selección de datos para el análisis
Mates     <- datos$math.score
Lectura   <- datos$reading.score
Escritura<- datos$writing.score
Genero    <- datos$gender
Raza      <- datos$race.ethnicity
```

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

Para comprobar la normalidad de las variables cuantitativas, se usará el test de Shapiro Wilk con un alpha de 0.05, donde la hipótesis nula es que sigue una distribución normal:

```
NombreColumnas <- c("Matemáticas","Lectura","Escritura")
NormalidadSp <-matrix(c(shapiro.test(Mates)$p.value,
+                       shapiro.test(Lectura)$p.value,
+                       shapiro.test(Escritura)$p.value), ncol = 3,
byrow = T)
colnames(NormalidadSp) <- NombreColumnas
NormalidadSp
```

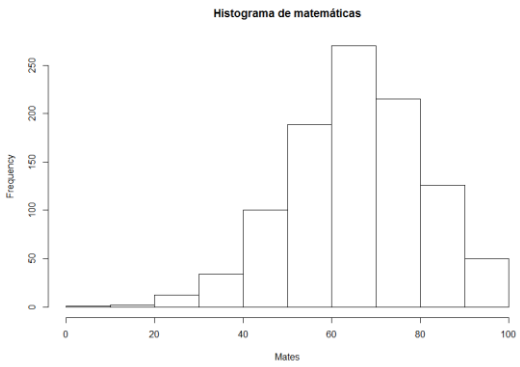
```
     Matemáticas           Lectura          Escritura
[1,] 0.002366662      0.0001895538       8.816204e-05
```

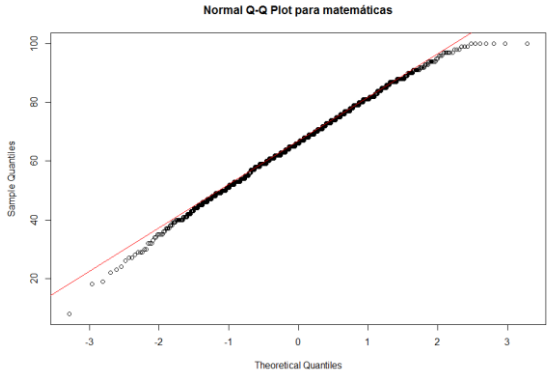
Según estos resultados, las 3 devuelve un p-value inferior a alpha con lo que no siguen una distribución normal.

A continuación se presentan los gráficos de histograma y quantile-quantile Plot de cada una de las pruebas

Matemáticas:

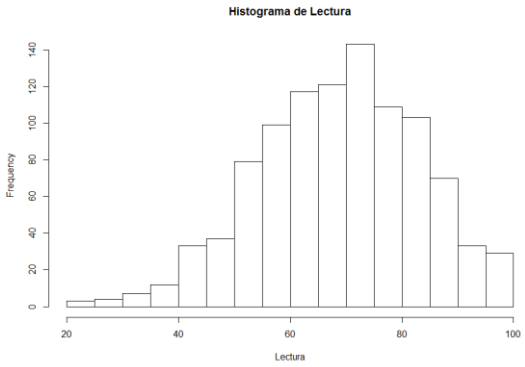
```
hist(Mates,main = paste("Histograma de matemáticas"))
qqnorm(Mates,main = paste("Normal Q-Q Plot para matemáticas"))
qqline(Mates,col="red")
```

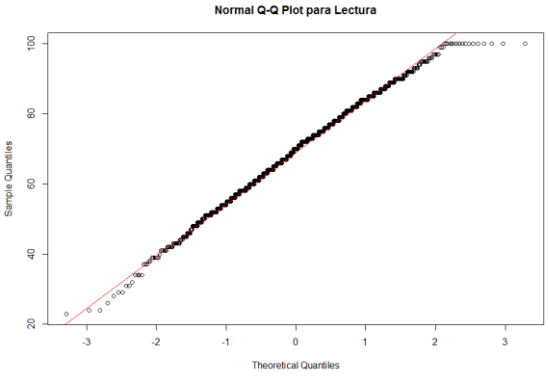




Lectura:

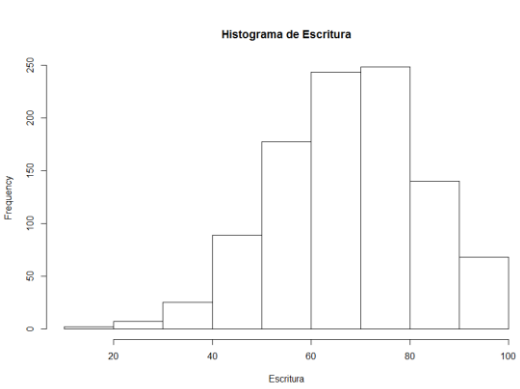
- `hist(Lectura, main = paste("Histograma de Lectura"))`
- `qqnorm(Lectura,main = paste("Normal Q-Q Plot para Lectura"))`
- `qqline(Lectura,col="red")`

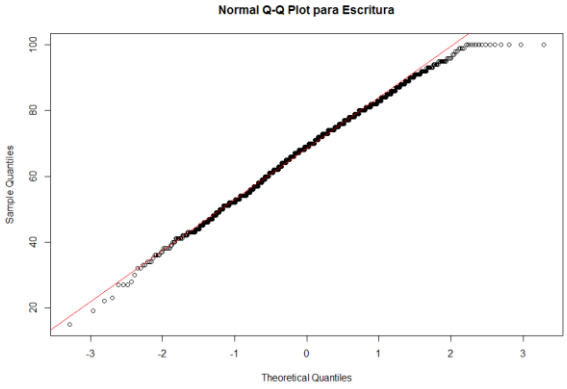




Escritura:

- `hist(Escritura, main = paste("Histograma de Escritura"))`
- `qqnorm(Escritura,main = paste("Normal Q-Q Plot para Escritura"))`
- `qqline(Escritura,col="red")`





Comparando las gráficas y los resultados del test de Shapiro wilk nos lleva a hacer alguna prueba más para confirmar los resultados. Estos resultados pueden deberse a que hay muchos valores que se repiten.

Vamos a analizar los resultados, separándolos por género:

```
➢ tapply(Mates,Genero,shapiro.test)

$female
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.99458, p-value = 0.06441

$male
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.99356, p-value = 0.03802
```

Vemos que para las mujeres, sí sigue una distribución normal.

```
➢ tapply(Lectura,Genero,shapiro.test)

$female
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98797, p-value = 0.000292


$male
        Shapiro-Wilk normality test
```

```
data:  X[[i]]
W = 0.99462, p-value = 0.08965
```

En este caso, es el grupo de los hombres el que sigue una distribución normal

```
> tapply(Escritura,Genero,shapiro.test)
$female
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.9829, p-value = 9.171e-06


$male
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.99481, p-value = 0.104
```

Y es esta prueba, el grupo de los hombres el que sigue una distribución normal.

Es conocido que al haber muchos valores repetidos, en el test de Shapiro puede haber algunas imprecisiones, por ello, podemos comprobar, estos datos con el test de Lillie:

```
#Test de Lillie
> NormalidadLi <-matrix(c(lillie.test(Mates)$p.value,
+                         lillie.test(Lectura)$p.value,
+                         lillie.test(Escritura)$p.value), ncol = 3,
byrow = T)
> colnames(NormalidadLi) <- NombreColumnas
> NormalidadLi
     Matemáticas      Lectura    Escritura
[1,]  0.05583007 0.0001390798 0.0005397114
```

En este caso nos indica que las notas de la prueba de matemáticas sí siguen una distribución normal.

A continuación se comprobará la homogeneidad de la varianza de las notas de la prueba de matemáticas por género.

```
> fligner.test(Mates ~ Genero, data = datos)

Fligner-Killeen test of homogeneity of variances

data:  Mates by Genero
Fligner-Killeen:med chi-squared = 0.18773, df = 1, p-value = 0.6648
```

Con este resultado con un p-value mayor a 0.05, podemos aceptar la homogeneidad de ambas varianzas

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.

Correlación de variables

Vamos a analizar qué variable influye en la nota de la prueba de matemáticas, para ellos calculamos la correlación con el resto de las pruebas:

```
> corr_matrix <- matrix(nc = 2, nr = 0)
> colnames(corr_matrix) <- c("Estimación", "p-value")
> for (i in 4:5) {
+   spearman_test = cor.test(datos[,i],
+                            datos[,"math.score"],
+                            method = "spearman")
+   corr_coef = spearman_test$estimate
+   p_val = spearman_test$p.value
+   # incluiomos el valor en la matriz de correlaciones
+   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(datos)[i]
+ }

> print(corr_matrix)
              Estimación       p-value
reading.score  0.8034746 8.656027e-227
writing.score  0.7776720 3.459858e-203
```

Ambas variables tienen una correlación significativa con la prueba de mátemáticas, siendo la prueba de lectura la que tiene valores más próximos a -1 o 1

¿Quién saca mejor nota media en la prueba de matemáticas: hombres o mujeres?

Para ello realizaremos un contraste de hipótesis, tomando un alpha de 0,05, usando la media de las dos poblaciones: hombres y mujeres.

$$H0 : \mu1 - \mu2 <= 0$$
$$H1 : \mu1 - \mu2 > 0$$

Donde μ1 es la media de los hombres y μ2, la media de las mujeres. Como hipótesis nula se establece que las mujeres sacan mejores notas

```
> MatesHombre <- datos[datos$gender=="male",]$math.score
> MatesMujer  <- datos[datos$gender=="female",]$math.score
> t.test(MatesHombre, MatesMujer, alternative = "greater")

        Welch Two Sample t-test

data:  MatesHombre and MatesMujer
t = 5.3077, df = 996.91, p-value = 6.842e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.429699      Inf
sample estimates:
mean of x mean of y
 68.72822  63.75629
```

Con un p-value inferior al nivel de significación fijado, rechazamos la hipótesis nula. Podemos concluir, por tanto, que las notas de los hombres son mejores.

Estimación de notas de matemáticas por medio de una regresión lineal

Regresores cuantitativos:

- Lectura
- Escritura

Regresores cualitativos:

- Género
- Raza

Variable a predecir:

- Mates

```
> modeloLectura   <- lm(Mates ~ Lectura, data = datos)
> modeloEscritura <- lm(Mates ~ Escritura, data = datos)
> modeloAmbas     <- lm(Mates ~ Lectura + Escritura, data = datos)
> modeloLER     <- lm(Mates ~ Lectura + Escritura + Raza, data = dato
  s)
> modeloLEG     <- lm(Mates ~ Lectura + Escritura + Genero, data =dat
  os)
> modeloLEGR     <- lm(Mates ~ Lectura + Escritura + Genero + Raza, d
  ata = datos)

> tablaCoef  <- matrix(c(1, summary(modeloLectura)$r.squared,
+                        2, summary(modeloEscritura)$r.squared,
+                        3, summary(modeloAmbas)$r.squared,
+                        4, summary(modeloLER)$r.squared,
+                        5, summary(modeloLEG)$r.squared,
+                        6, summary(modeloLEGR)$r.squared
+                        ),
+                                ncol = 2, byrow = TRUE)
> colnames(tablaCoef) <- c("Modelo", "R^2")
> tablaCoef
     Modelo       R^2
[1,]      1 0.6641399
[2,]      2 0.6390615
[3,]      3 0.6695021
[4,]      4 0.6867489
[5,]      5 0.8377030
[6,]      6 0.8496834
```

El modelo 6, modeloLEGR, que usa las notas de las otras dos pruebas, el género y la raza/etnia es que tiene mayor coeficiente de determinación y con él intentaremos predecir el valor de la nota de matemáticas del registro que se eliminó al principio del tratamiento de datos porque no tenía informada la nota.

Si recuperamos los valores de este registro, antes de ser eliminado, sus valores son:

```
> Datos[datos$math.score==0,]
   gender race.ethnicity parental.level.of.education       lunch test
.preparation.course
60 female        group C            some high school free/reduced
none
   math.score reading.score writing.score
60          0            17            10
```

Nota en Lectura = 17 y en Escritura = 10, mujer y del grupo c, con que lo que podemos intentar predecir el valor de la prueba de Matemáticas:

```
> registro <- data.frame(
+   Lectura = 17,
+   Escritura = 10,
+   Genero = "female",
+   Raza = "group C"
```

```
+ )
➢ predict(modeloLEGR, registro)
       1
6.048428
```

Al ser notas enteras, redondeamos obteniendo un valor de 6

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

Durante el tratamiento y el análisis de los datos se ha incluido apoyo gráfico y tablas que muestren los resultados obtenidos de los distintos comandos pasos que se han ido ejecutando.

6. Resolución del problema.

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

En este caso práctico se ha realizado un tratamiento de los datos, pre-procesamiento, en el que se han comprado los valores nulos, ceros y extremos. Una vez preparados los datos, se han exportado a un nuevo fichero csv, dando paso al análisis. En el cuál se ha orientado a responder a las preguntar fijadas al comienzo. Para ello se han realizado varias pruebas estadísticas.
Se ha mostrado la relación de las distintas variables cuantitativas con la nota de la prueba de matemáticas y, mediante contrastes de hipótesis, hemos podido comprobar si las mujeres obtenían mejores notas en esta prueba.
Al final, por medio de regresión lineal, hemos conseguido un modelo mediante el cual estimar el valor de la nota de la prueba de matemáticas usando, como ejemplo, un registro descartado para el análisis por no disponer de esta información.

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 usado está en R y se ajunta el repositorio de Github, en la siguiente dirección: https://github.com/jmartinezare/TyCVD_Practica2/tree/master/code

Referencias

Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.

Squire, Megan (2015). Clean Data. Packt Publishing Ltd.

Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.