Aprendizaje Supervisado en R

Estudio de análisis de regresión

Ejercicio 1: Regresión lineal simple: datos de una población de kanguros
Se le proporciona un fichero de nombre "kanguro.xls" [1] que contiene la medición de la anchura y largo de la nariz de una población aleatoria de kanguros grises. El nombre de las variables son:

- nose_width (mm)
- nose_length (mm)

Estos datos representan las variables de 45 kanguros. El objetivo es familiarizarnos con el concepto de regresión simple[2].

1. Cargue los datos en una variable de nombre kang_nose que será un dataframe con 2 variables o atributos y 45 observaciones (kanguros) que vienen codificados en el fichero.

```
kang_nose <-read.delim("kanguros.csv", sep = "\t", head = TRUE)
```

   Inspeccione el conjunto de entrenamiento (funciones head, str, dim).

2. Modifique el nombre de las variables X e Y por nombres más intuitivos: nose_length y nose_width.

```
colnames(kang_nose) <- c('nose_length', 'nose_width')
colnames(kang_nose)
```

3. El objetivo es describir la relación lineal entre las dos variables con la función lm() en caso de que exista, para ello exploraremos previamente los datos:
   - Dibuje las observaciones en el plano, de manera que el eje X sea la anchura de la nariz, y el eje Y el largo. Para ello use la función plot.
   - Cree una función lineal que aproxime la longitud de la nariz en función del ancho. Para ello utilice la función lm con dos parámetros: el primero indica la variable a predecir mediante una fórmula y el segundo el conjunto de datos.
   - Puede predecir el valor de un nuevo canguro utilizando la función predict. Para ello cree una variable nueva utilizando el primero del dataset. ¿Qué observa?

```
nose_width_new <- kang_nose[1,]
```

   - Para dibujar la recta de regresión escriba el siguiente código:

```
abline(lm_kang$coefficients, col = "red")
```

[1] http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/frame.html
[2] Introduction to Machine Learning. Data Camp.

- Calcule como medida de rendimiento el error RMSE (Root Mean Squared Error) definido a través de:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{res}_i^2}$$

$$\mathrm{res}_i = y_i - \hat{y}_i$$

Para ello realice los siguientes pasos:

- Llame la función predict para obtener la predicción en función de los datos de entrenamiento.
- Calcule el residuo y almacénalo en la variable res. El residuo es la diferencia entre los valores reales y los estimados con el modelo de regresión lineal simple.
- Finalmente, calcule el RMSE aplicando la fórmula anterior.

4. El RMSE es una medida difícil de interpretar. Con el valor obtenido anteriormente, ¿podrías decir si el modelo es bueno o malo? Por ello, vamos a utilizar otra medida de rendimiento que es el R-squared.

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

Esta medida varía entre 0 y 1, de manera que cuanto más cerca esté a 1 mayor será el grado de asociación lineal ente la variable predictor y la variable respuesta. Calcule la medida, para ello:

- Calcule la suma de los residuos al cuadrado y asígnelo a la variable ss_res
- Calcule la variable ss_tot
- Y almacene en la variable r_sq el resultado de aplicar la ecuación anterior.
- Este cálculo no hace falta realizarlo ya que lo calcula la función lm, para visualizarlo escriba:

```
summary(lm_kang)$r.squared
```

<u>Ejercicio 2</u>: Ejemplo del Banco Mundial

Sean los siguientes datos, proporcionado por el Banco Mundial, en donde aparece el PIB y el porcentaje de población urbana de varios países de la ONU para el año 2014. El conjunto de datos se proporciona en el fichero world_bank_train.csv y tiene dos variables:

- PIB, el nombre en el dataset es cgdp.
- urb_pop.

También se proporciona el PIB para Afganistan en 2014 de 413 dólares, pero su población urbana se desconoce, ¿puedes predecir este valor? Para ello llame a la función lm tal como hizo en el ejercicio anterior y observa el valor R2, ¿el valor observado es bueno?

Para mejorar el modelo observe los datos, tenemos que dar un paso atrás, y observarás que la

variable predictora es numérica y la variable respuesta está expresada en percentiles. Tendría más sentido analizar si hay una relación lineal entre variables medidas en percentiles. Para ello, vamos a realizar un cambio de medición y para ello tomamos logaritmo del PIB al dibujar los datos y calcular el modelo lineal:

```
plot(urb_pop ~ log(cgdp), data = world_bank_train,
     xlab = "log(GDP per Capita)",
     ylab = "Percentage of urban population")

# Linear model: change the formula
lm_wb <- lm(urb_pop ~ log(cgdp),
       data = world_bank_train)
```

¿Qué el modelo es mejor?

Ejercicio 3: Regresión multivariable

Vamos a trabajar con un conjunto de datos de más de dos variables y vamos a generar un modelo de regresión multivariable. El conjunto de datos es de ventas netas de un negocio y tiene las siguientes variables:

- Publicidad (nombre de la variable advertisement).
- Competencia (competition).
- Inventario (inv).
- Tamaño de distrito (size_dist).
- Tamaño de la tienda (sq_ft).
- Ventas (sales), es la variable a predecir o respuesta.

En este ejercicio debe generar un modelo multivariable, para ello:

- Dibuje las ventas en función del resto de variables para observar si hay una relación lineal.
- Construya un modelo lineal para predecir las ventas en función del resto de variables.
- Muestre el modelo y observe las variables de rendimiento.

Una vez construido el modelo, observe las medidas de rendimiento, y ¿todas las variables predictoras son relevantes? Observe que cada variable predictora viene acompañada de un valor de p-value. ¿Interesa eliminar las variables predictoras con un valor p no bueno? Para responder esta pregunta hay que responder las siguientes preguntas:

- ¿Existe un patrón si dibujamos los valores estimados frente a los residuos (distancia entre valor estimado y valor real)? Para que el modelo sea bueno no se debe observar ningún patrón. Para visualizar escriba el siguiente código:

```
plot(lm_shop$fitted.values,  lm_shop$residuals,
     xlab = "Fitted values", ylab = "Residuals")
```

- ¿Existe un patrón en los residuos? Para que el modelo sea bueno se debe observar una línea. Para visualizar escriba el siguiente código:

```
qqnorm(lm_shop$residuals, ylab = "Residual Quantiles")
```