

Sesión 3:

El Lenguaje R:
Regresión



Objetivos

- Construir modelos básicos de regresión en R.
- Evaluar la eficacia de la regresión.

Palabras claves: variable, regresión, optimización, SVR, arboles.

Introducción

Una herramienta útil en el análisis estadístico es la regresión. En esta interesa conocer la relación entre un conjunto de variables independientes (predictoras) y una variable que se considera dependiente (o variable de respuesta).

El análisis trata de explicar a la variable dependiente en términos de las otras, buscando posibles relaciones entre ellas y a partir de estas predecir luego el valor de la variable dependiente. Dicho análisis tiene sentido mientras los valores de las variables predictoras no salgan de un cierto margen.

Dicho proceso puede aplicarse en la toma de decisiones si consideramos que los datos que usamos para construir la regresión no modifican su rango de valores de una manera muy fuerte a lo largo del tiempo.

Regresión lineal

Existen distintos tipos de regresión, de distinta complejidad y para distintas situaciones. Entre ellas, la más sencilla es la regresión lineal. De manera intuitiva, y suponiendo un conjunto de puntos (x_i, y_i) de observaciones de dos variables, la idea de la regresión lineal es tratar de conseguir una recta que aproxime los posibles datos de la mejor manera.

La figura 1 muestra los posibles datos que serán objeto de la aproximación de la recta.

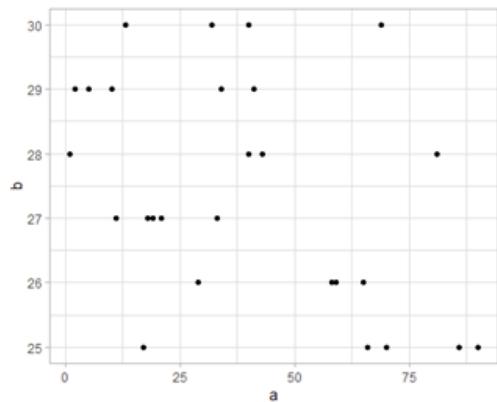


Figura 1. Posibles Datos. Fuente: Elaboración propia.

Si es posible conseguir la ecuación de la recta, se puede usar para predecir la respuesta de la variable dependiente (en este caso y) respecto a un valor específico de x . La ecuación de la recta se denomina curva de regresión:

$$y = mx + b$$

El ajuste de la recta se puede realizar de distintas maneras, pero una de las más utilizadas es el método de mínimos cuadrados ordinarios, donde se busca que la suma de los cuadrados de las distancias en vertical de cada punto a la recta sea mínimo.

$$SMC = \sum_{i=0}^n (y_i - f(x_i))^2 = \sum_{i=0}^n (y_i - mx_i - b)^2$$

El valor $y_i - f(x_i)$ representa la diferencia entre el valor real de la variable dependiente y su valor predicho por medio de la recta de regresión. La función SMC depende no de (x_i, y_i) (que son puntos muestrales) si no de los parámetros m, b . Para encontrar los valores de m, b se usa un proceso de optimización basado en cálculo diferencial, que en R está implementado en la función `lm()`.

```
Lineal<-lm(b~a)
Lineal
##
## Call:
## lm(formula = b ~ a)
##
## Coefficients:
## (Intercept)           a
##      28.58087     -0.02891
```

En esta función, se coloca la variable dependiente antes de la variable predictora. Los coeficientes son el intercepto (b) y la pendiente de la recta (m).

Con estos ya es posible trazar una recta de regresión. La figura 2 muestra la recta de la regresión colocada de manera que cada dato posible aproximado tenga una distancia mínima entre el dato y la recta.

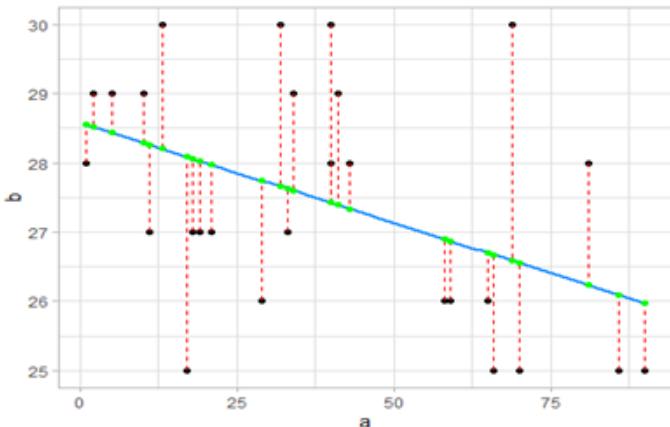


Figura 2. Recta de una regresión. Fuente: Elaboración propia.

Las rectas verticales miden la distancia de cada punto a la recta, mientras los puntos verdes son los valores estimados por la recta de regresión.

Si se quiere predecir valores nuevos, se usa la función `predict()`.

```
c<-data.frame (a=c(10,75,99))  
predict.lm (object=Lineal,newdata=c)  
## 1 2 3  
## 28.29182 26.41298 25.71925
```

En esta función, los datos nuevos que se usan para la predicción deben llevar el mismo nombre de las variables originales.

Si se quiere más información acerca del modelo, se puede usar la función `summary()`.

Actividad 1: De la página de datos abiertos de su ciudad, escoja una base de datos con al menos dos variables cuantitativas (para esta sesión, entre más variables cuantitativas tenga la base va a ser más útil), cárguela al entorno de R. Escoja dos de las variables cuantitativas y realice un análisis de regresión lineal.

Regresión Lineal Múltiple

Si se tienen observaciones de múltiples variables predictoras, se puede realizar una regresión lineal múltiple. Suponiendo que x_1, x_2, \dots, x_k son las variables predictoras independientes, se intenta conseguir una expresión

$$y = b + a_1 x_1 + a_2 x_2 + \dots + a_k x_k$$

que aproxime el valor de la variable dependiente a partir de los valores de las predictores. La figura 3 demuestra las k-dimensiones de una regresión lineal múltiple conforme el número de variables predictoras aumenten.

```
## Loading required package: scatterplot3d
```

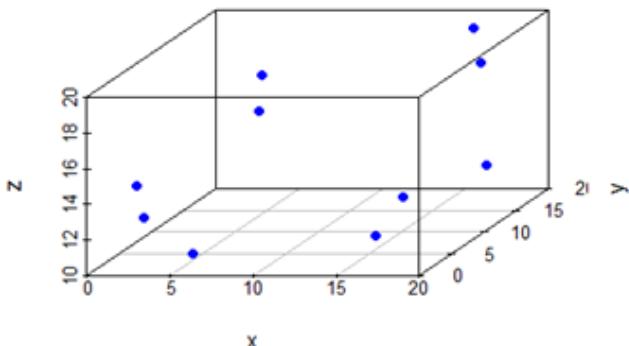


Figura 3. Datos en k-dimensiones conforme la regresión lineal múltiple. Fuente: Elaboración propia.

La ecuación $y = b + a_1x_1 + a_2x_2 + \dots + a_kx_k$, cuando $k=2$, corresponde a la ecuación de un plano, con mas variables se denomina hiperplano. Este debe aproximar lo mejor posible a los puntos.

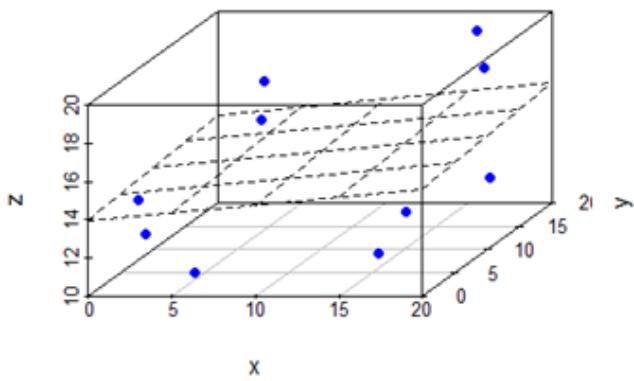


Figura 4. Recta de una regresión lineal múltiple. Fuente: Elaboración propia.

Una vez más, se busca que la distancia entre los puntos y el hiperplano (distancia perpendicular a las variables predictoras) sea lo menor posible como lo muestra la figura 4.

Con un proceso similar al de la regresión lineal, se pueden conseguir los coeficientes, y en R la función `lm()` se encargará también de la regresión multilíneaal.

La variable dependiente se coloca primero, seguida de las variables predictoras que se quieren usar.

```

data
##      x  y  z
## 1  17  1 12
## 2   3  0 15
## 3   7  9 19
## 4   3  1 13
## 5   6  1 11
## 6  12 18 10
## 7  17 16 20
## 8  10  1 19
## 9  19 13 13
## 10 19 12 19

hiper <- lm(z ~ x + y,data=data)
hiper

##
## Call:
## lm(formula = z ~ x + y, data = data)
##
## Coefficients:
## (Intercept)          x              y
## 13.93129       0.08205       0.03355

```

La función de regresión es $z=16.6147-0.1924x+0.1976y$. De esta se puede decir que se estima una variación negativa de -0.1924 unidades en el valor de z por cada unidad de cambio en x .

En cambio, una variación positiva de 0.1976 unidades por cada unidad de cambio en y .

Con los coeficientes se puede construir la ecuación del plano y realizar predicciones, igual que en la regresión lineal.

```

pr=data.frame(x=c(4,10,8),y=c(10,5,15))
predict (object = hiper,newdata = pr)

##           1           2           3
## 14.59497 14.91953 15.09091

```

También es posible usar la función `summary()` para extraer más información sobre el modelo.

Actividad 2: Usando una base con más de dos variables cuantitativas, realice un análisis de regresión múltiple.

Es posible que una o varias de las variables predictoras tengan más efecto sobre la variable dependiente que otras. Una manera sencilla de buscar indicios de esta relación es con la matriz de correlación, que calcula las covarianzas entre pares de variables.

```
if (!require('corrplot'))  
  install.packages('corrplot')  
## Loading required package: corrplot  
## corrplot 0.84 loaded  
library(corrplot)  
matrizcor<-cor(data)
```

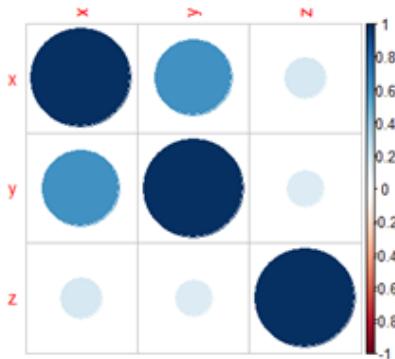


Figura 5. Gráfico de correlación por calor. Fuente: Elaboración propia.

Valores más cercanos a 1 indican una correlación creciente, y cercanos a -1 una correlación decreciente. Valores cercanos a 0 muestran la ausencia de una correlación. Valga la pena anotar que correlación no es lo mismo que causalidad.

Si se tienen muchas variables a veces es conveniente escoger para el modelo solo aquellas que tengan mayor correlación con la variable dependiente.

La figura 5 es la manera gráfica de representar las correlaciones entre variables, a medida que el círculo sea más grande y de un tono más intenso, indicaría la existencia de una correlación más grande.

Actividad 3: Realice un análisis de correlación entre las variables cuantitativas presentes en su base de datos.

Regresión Polinómica

Volviendo al caso de la regresión lineal simple sobre un conjunto de observaciones (x,y), si la correlación entre las variables no es fuerte, una regresión lineal no es adecuada. En ese caso, una regresión polinómica es recomendable. En esta, se busca construir una curva de regresión de la forma

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots a_kx^k$$

donde k es el grado de la regresión. El proceso es una regresión multilínea, donde las variables extras son potencias de la variable x . Por ejemplo, para realizar una regresión de orden 3 se necesitan potencias de x hasta ese orden.

```
ggplot(data.frame(a,b), aes(x=a, y=b)) +  
  geom_point() + theme_light()
```

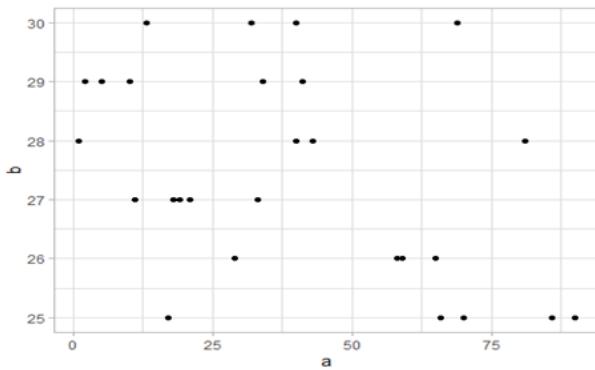


Figura 6. Gráfico del dataframe. Fuente: Elaboración propia.

Se prepara un dataframe con columnas extra por cada potencia como lo muestra la figura 6.

```
datos<-data.frame(x=a,y=b)  
datos$x2<-datos$x**2  
datos$x3<-datos$x**3  
datos
```

```

##      x   y   x2      x3
## 1  40 28 1600 64000
## 2  65 26 4225 274625
## 3   5 29   25    125
## 4  13 30  169   2197
## 5  40 30 1600 64000
## 6  58 26 3364 195112
## 7  81 28 6561 531441
## 8  17 25  289   4913
## 9  70 25 4900 343000
## 10 2 29    4     8
## 11 58 26 3364 195112
## 12 10 29  100   1000
## 13 18 27  324   5832
## 14 43 28 1849  79507
## 15 11 27  121   1331
## 16  1 28    1     1
## 17 86 25 7396 636056
## 18 32 30 1024  32768
## 19 34 29 1156  39304
## 20 59 26 3481 205379
## 21 66 25 4356 287496
## 22 33 27 1089  35937
## 23 40 28 1600 64000
## 24 19 27  361   6859
## 25 21 27  441   9261
## 26 41 29 1681  68921
## 27 40 28 1600 64000
## 28 69 30 4761 328509
## 29 90 25 8100 729000
## 30 29 26  841   24389

```

y se realiza un modelo lineal múltiple

```

cub <- lm(y ~ x +x2+x3,data=datos)
cub
##
## Call:
## lm(formula = y ~ x + x2 + x3, data = datos)
##
## Coefficients:
## (Intercept)          x           x2          x3
## 2.833e+01 -2.565e-02  4.533e-04 -6.146e-06

```

y se realiza un modelo lineal múltiple

```
datos$yp<-predict(cub)
ggplot(datos, aes(x=a, y=b)) +
  geom_point() +geom_smooth(aes(x=x,y=yp),
  col='dodgerblue1') +geom_segment(aes(xend=a,
  yend=yp), col='red', lty='dashed')
+geom_point(aes(y=yp), col='green') + theme_light()
## `geom_smooth()` using method = 'loess' and formula
'y ~ x'
```

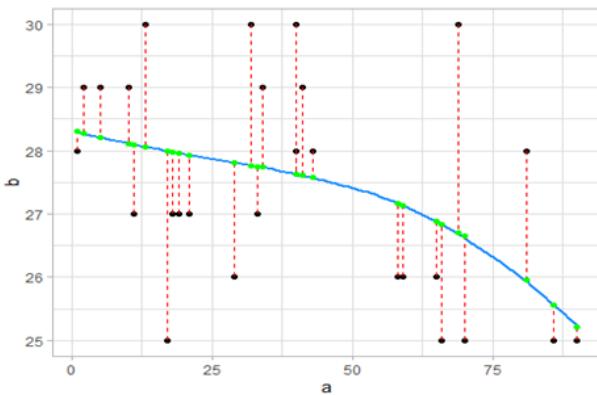


Figura 7. Curva de regresión polinómica. Fuente: Elaboración propia.

Actividad 4: Sobre las variables que uso para realizar la regresión lineal en la actividad 1, construya regresiones polinómicas de orden 2, 3 y 4. Cual ajusta mejor los datos?

Regresión con máquinas de soporte vectorial (SVM)

Volviendo al caso de la regresión lineal simple sobre un conjunto de observaciones (x, y) , si la correlación entre las variables no es fuerte, una regresión lineal no es adecuada.

En ese caso, una regresión polinómica es recomendable. En esta, se busca construir una curva de regresión de la forma

```

if (!require('e1071')) install.packages('e1071')

## Loading required package: e1071

library(e1071)
svr = svm(y~x,datos)
datos$ypsvr = predict(svr)
ggplot(datos, aes(x=x, y=y)) +
  geom_point() +geom_smooth(aes(x=x,y=ypsvr),
  col='dodgerblue1') +geom_segment(aes(xend=x,
  yend=ypsvr), col='red', lty='dashed') +
  geom_point(aes(y=ypsvr), col='green') +
  theme_light()

## `geom_smooth()` using method = 'loess' and formula
'y ~ x'

```

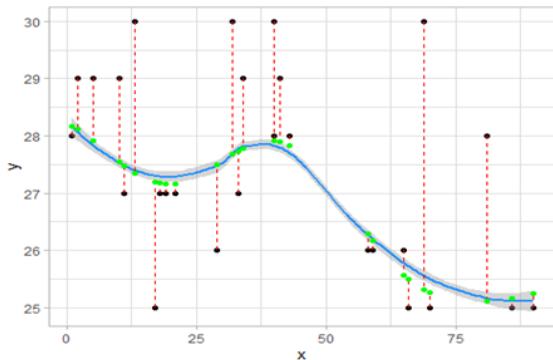


Figura 8. Curva de regresión en SVM. Fuente: Elaboración propia.

También es posible usar las herramientas de SVM para realizar una afinación de parámetros.

Actividad 5: Realice una regresión con SVM para las primeras variables que utilizo en la regresión lineal.

Regresión con árboles de decisión.

Análogo al proceso con el SVM, sin tener que pensar en categorías, el árbol de decisión genera una función predictora que también se puede usar en regresión (ver figura 9). La implementación también es bastante similar

```

if (!require('rpart')) install.packages('rpart')

## Loading required package: rpart

library(rpart)
dt<-rpart(y ~ x, method = "anova", data = datos )
datos$ypdt = predict(dt)
ggplot(datos, aes(x=x, y=y)) +
  geom_point() +geom_smooth(aes(x=x,y=ypdt),
  col='dodgerblue1',type='l') +geom_segment(aes(xend=x,
  yend=ypdt), col='red', lty='dashed')+
  geom_point(aes(y=ypdt), col='green') + theme_light()

## `geom_smooth()` using method = 'loess' and formula
'y ~ x'

```

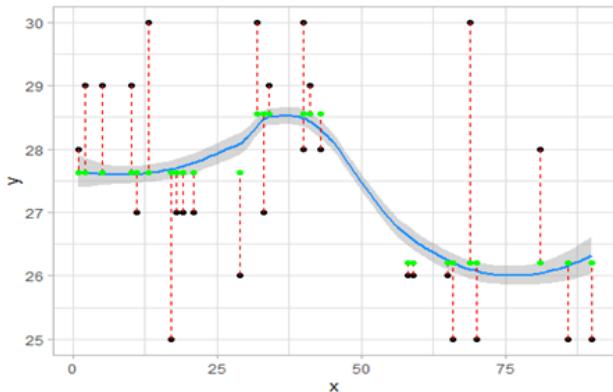


Figura 9. Curva de regresión con árbol de desición. Fuente: Elaboración propia.

Es importante anotar que tanto la regresión con SVR como con los árboles de decisión escalan de manera automática para cualquier cantidad de variables predictoras.

Actividad 6: Usando las variables que escogió para la regresión lineal múltiple, realice una regresión por SVR y una por árboles de decisión.

Referencias

Ggplot2(s.f.).Create Elegant Data Visualisations Using the Grammar of Graphics. Recuperado de <https://ggplot2.tidyverse.org/>.

Kreyszig, Erwin, Herbert Kreyszig, y E. J. Norminton. Advanced engineering mathematics. 10th ed. Hoboken, NJ: John Wiley, 2011.

R.(s.f.).Welcome | R for Data Science. Recuperado de <https://r4ds.had.co.nz/>.



¡ideas que crean valor!