Actividad 4: Regresión Lineal Múltiple

Especialidad en Métodos Estadísticos a Distancia - CIMAT

Edgar Anuar Sánchez Hernández

2022-12-17

Ejercicio

Los datos de una encuesta sobre la satisfacción de los pacientes en un hospital se muestran en la siguiente tabla. Las variables independientes son la edad del paciente, un índice de gravedad de la enfermedad (los valores más altos indican mayor gravedad), una variable indicadora que denota si el paciente es un paciente médico (0) o un paciente quirúrgico (1) y, un índice de ansiedad (los valores más altos indican mayor ansiedad). Los datos también se encuentran en el archivo denominado "Datos_actividad.xlsx".

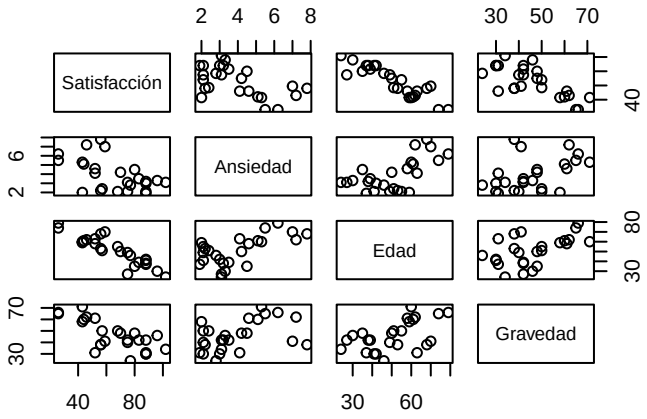
Solución

Importando los datos

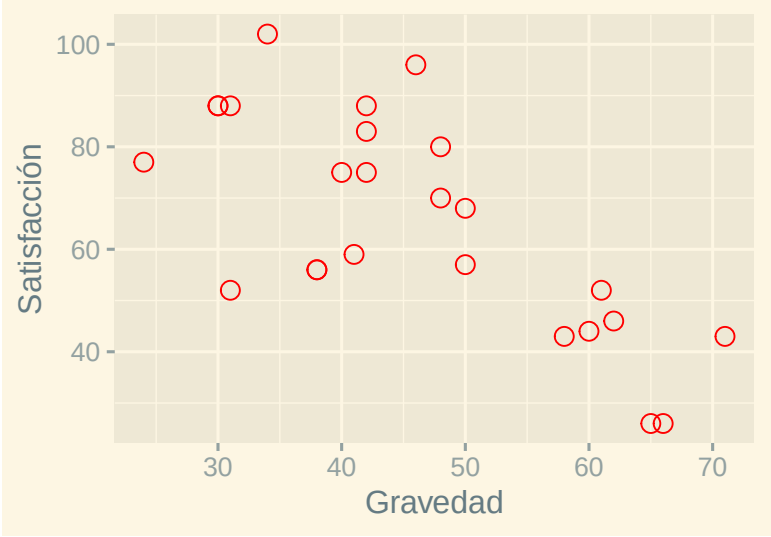
```
library(readxl)
library(ggplot2)
library("ggthemes")

Datos_actividad <- read_excel("Datos_actividad.xlsx",
    col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric"))
```

a) Graficas de dispersión de las variables numéricas

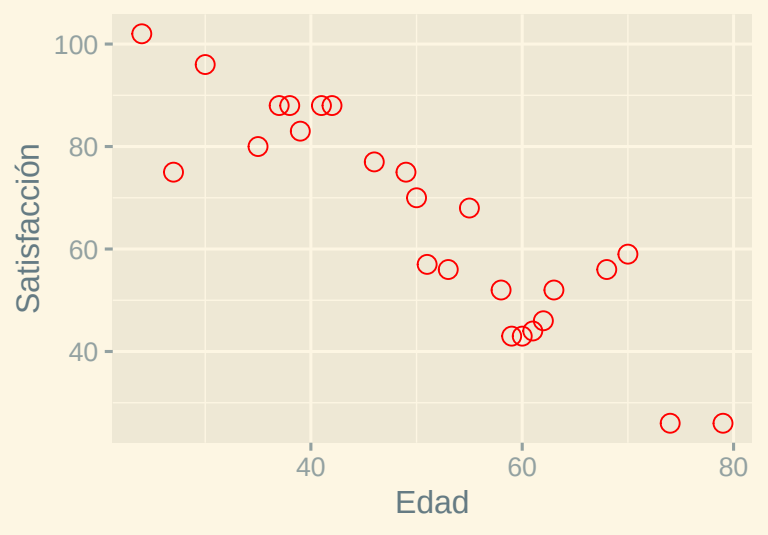
```
pairs(~Satisfacción + Ansiedad + Edad + Gravedad, data =  Datos_actividad)
```



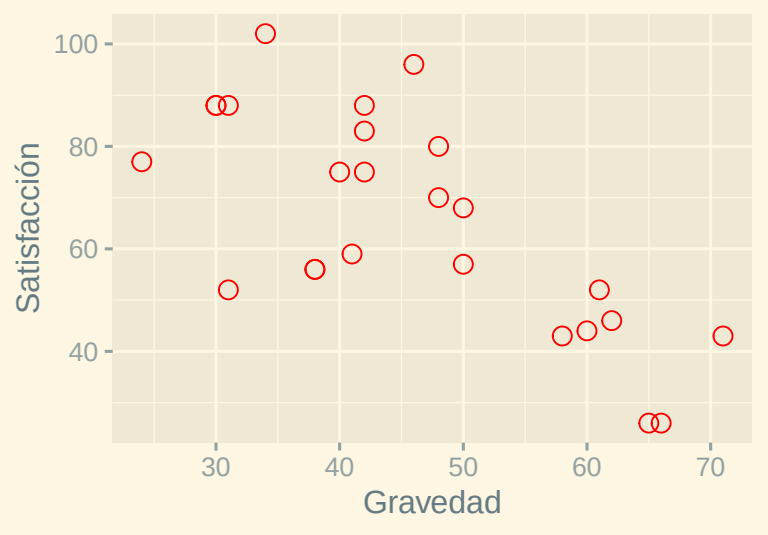

```
ggplot(data = Datos_actividad) +
  geom_point(mapping = aes(x = Gravedad, y = Satisfacción),color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```




```
ggplot(data = Datos_actividad) +
  geom_point(mapping = aes(x = Edad, y = Satisfacción),color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```




```
ggplot(data = Datos_actividad) +
  geom_point(mapping = aes(x = Gravedad, y = Satisfacción),color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```



b) Modelo lineal múltiple

```
modelo <- lm(Satisfacción ~ Ansiedad + Gravedad + Edad + Médico_Quirúrgico, data = Datos_actividad)
summary(modelo)
```

```
##
## Call:
## lm(formula = Satisfacción ~ Ansiedad + Gravedad + Edad + Médico_Quirúrgico,
##     data = Datos_actividad)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -18.1322  -3.5662   0.5655   4.7215  12.1448
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       143.8672     6.0437  23.804 3.80e-16 ***
## Ansiedad            1.3064     1.0841   1.205 0.242246
## Gravedad           -0.5862     0.1356  -4.324 0.000329 ***
## Edad               -1.1172     0.1383  -8.075 1.01e-07 ***
## Médico_Quirúrgico   0.4149     3.0078   0.138 0.891672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.207 on 20 degrees of freedom
## Multiple R-squared:  0.9036, Adjusted R-squared:  0.8843
## F-statistic: 46.87 on 4 and 20 DF,  p-value: 6.951e-10
```

La ecuación estimada del modelo de regresión lineal mútiple es:

$$S = 143.86 + 1.30a - 0.58g - 1.11e + 0.41m$$

Donde S = Satisfacción, a = índice de ansiedad, g = índice de gravedad de la enfermedad, e = edad, m = tipo de paciente (médico o quirúrgico).

c) R^2 ajustado

Observando la información obtenida por el modelo lineal, se tiene que el valor de R^2 ajustado es de 0.8843.

d) Método Stepwise:

```
# Stepwise

modelo_completo<-lm(Satisfacción~.,data = Datos_actividad)
modelo_reducido<-lm(Satisfacción~1,data = Datos_actividad)
####Seleccion de var stepwise
step(modelo_reducido,scope=list(lower=modelo_reducido,
                                upper=modelo_completo),
     direction = "both")
```

```
## Start:  AIC=153.66
## Satisfacción ~ 1
##
##                     Df Sum of Sq     RSS    AIC
## + Edad               1    8756.7  2021.6 113.82
## + Gravedad           1    5634.3  5143.9 137.17
## + Ansiedad           1    3100.2  7678.0 147.18
## <none>                            10778.2 153.66
## + Médico_Quirúrgico  1     589.9 10188.4 154.25
##
## Step:  AIC=113.82
```

```
## Satisfacción ~ Edad
##
##                     Df Sum of Sq     RSS    AIC
## + Gravedad           1     907.0  1114.5 100.93
## <none>                           2021.6 113.82
## + Ansiedad           1       9.8  2011.8 115.70
## + Médico_Quirúrgico  1       1.8  2019.8 115.80
## - Edad               1    8756.7 10778.2 153.66
##
## Step:  AIC=100.93
## Satisfacción ~ Edad + Gravedad
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           1114.5 100.93
## + Ansiedad           1      74.6 1039.9 101.20
## + Médico_Quirúrgico  1       0.2 1114.4 102.93
## - Gravedad           1     907.0 2021.6 113.82
## - Edad               1    4029.4 5143.9 137.17
```

```
##
## Call:
## lm(formula = Satisfacción ~ Edad + Gravedad, data = Datos_actividad)
##
## Coefficients:
## (Intercept)         Edad     Gravedad
##     143.472       -1.031       -0.556
```

Se obtiene que el mejor modelo para estas variables incluye solamente Gravedad y Edad como variables independientes:

```
# Mejor modelo
mejor_modelo <- lm(Satisfacción ~ Gravedad + Edad, data = Datos_actividad)
summary(mejor_modelo)
```

```
##
## Call:
## lm(formula = Satisfacción ~ Gravedad + Edad, data = Datos_actividad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2800  -5.0316   0.9276   4.2911  10.4993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.4720     5.9548  24.093  < 2e-16 ***
## Gravedad     -0.5560     0.1314  -4.231 0.000343 ***
## Edad         -1.0311     0.1156  -8.918 9.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.118 on 22 degrees of freedom
## Multiple R-squared:  0.8966, Adjusted R-squared:  0.8872
## F-statistic: 95.38 on 2 and 22 DF,  p-value: 1.446e-11
```

Y la ecuación de este mejor modelo es:

$$S = 143.4720 - 0.5560g - 1.0311e$$

Donde S = Satisfacción, g = índice de gravedad de la enfermedad y e = edad.

El valor de R^2 ajustado para este mejor modelo es de 0.8872, que es ligeramente mejor que el valor del modelo completo: 0.8843. Es ligeramente mayor pero contiene dos variables menos. Es decir, es una mejora consirable del modelo.

e) Analizando multicolinealidad (Calculando el VIF)

Para este mejor modelo se tiene lo siguiente:

```
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```
car::vif(mejor_modelo)
```

```
## Gravedad     Edad
## 1.388632 1.388632
```

Es decir, como los valores del VIF de cada variable son cercanos a 1, se puede concluir que el modelo no tiene problemas de multicolinealidad.

f) Estimaciones

Estimar la media del nivel de satisfacción para un paciente quirúrgico de edad de 60 años, gravedad de 45 y un nivel de ansiedad de 3. Utilice la ecuación de regresión establecida en el inciso d) y solamente las variables involucradas en este. Proporcione la estimación puntual y mediante un intervalo de confianza. Utilizar un nivel de confianza del 95%.

Utilizando el mejor modelo, se tiene:

```
dato <- data.frame(Gravedad = 45, Edad = 60)
predict(mejor_modelo, newdata = dato,
        interval = 'confidence', level = 0.95, se.fit = TRUE)
```

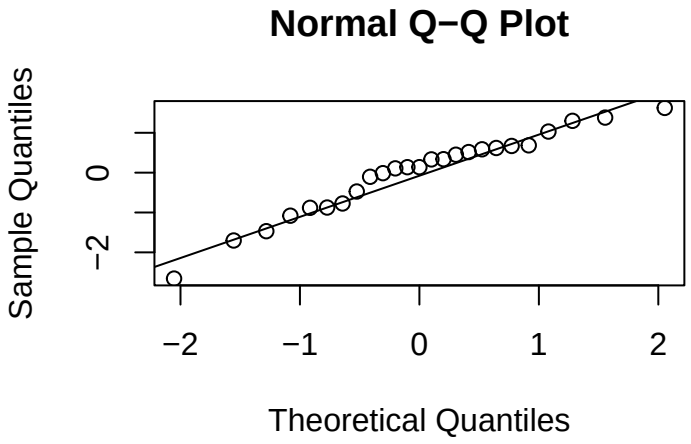
```
## $fit
##        fit      lwr      upr
## 1 56.58711 52.82087 60.35334
##
## $se.fit
## [1] 1.81604
##
## $df
## [1] 22
##
## $residual.scale
## [1] 7.117667
```

El modelo predice que el nivel de satisfacción para un paciente quirúrgico de edad de 60 años, gravedad de 45 está entre 52.82 y 60.35 con un nivel de confianza del 95%.

g) Verifique los supuestos del modelo

Normalidad

```
resi_rlm6<-rstandard(mejor_modelo) #Se obtienen los residuales estandarizados
qqnorm(resi_rlm6) #Gráfico QQ
qqline(resi_rlm6) #Línea del gráfico QQ
```



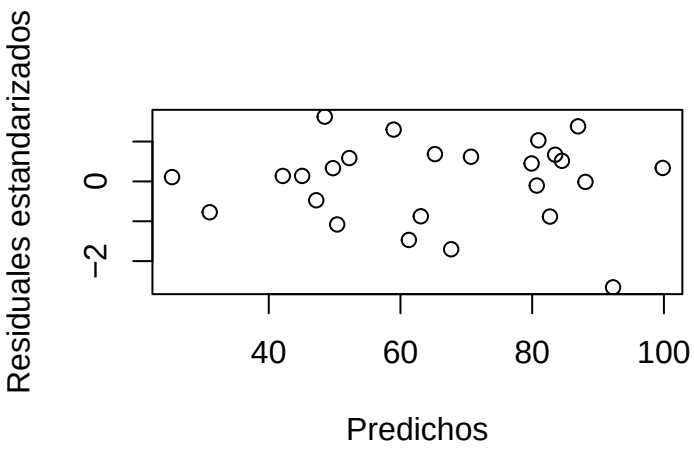

```
shapiro.test(resi_rlm6) #Prueba de Shapiro-Wilks usando residuales estandarizados
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resi_rlm6
## W = 0.95333, p-value = 0.2977
```

Se concluiye que, con un nivel de significancia de 0.05, el p-value=0.29>0.05, por lo que NO se rechaza la hipótesis nula. Es decir, los residuales estandarizados provienen de una distrinución Normal.

Varianza constante

```
predichos_rlm6<-predict(mejor_modelo) #Se obtienen los valores predichos
# Gráfico para observar varianza de los residuales
plot(predichos_rlm6,resi_rlm6,xlab = "Predichos",ylab="Residuales estandarizados")
```

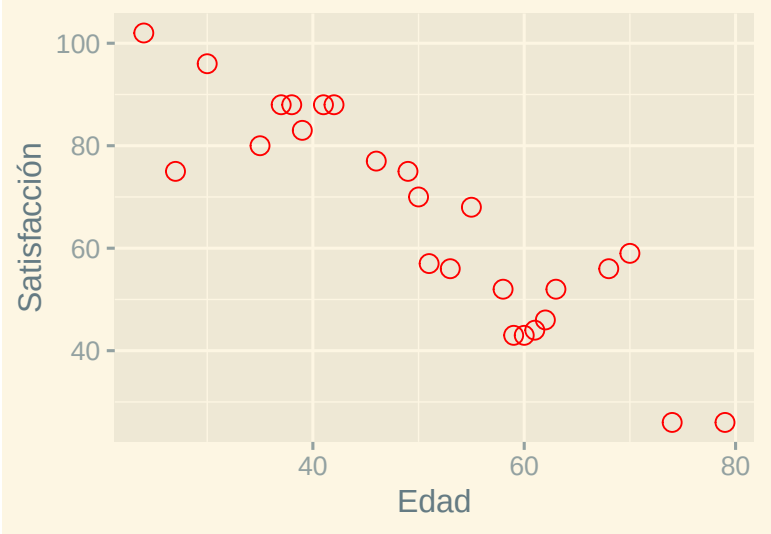


De la gráfica se obsberva cierta homogeneidad en las varianzas, aunque se recomienda obtener más evidencia para valores predichos menores a 50.

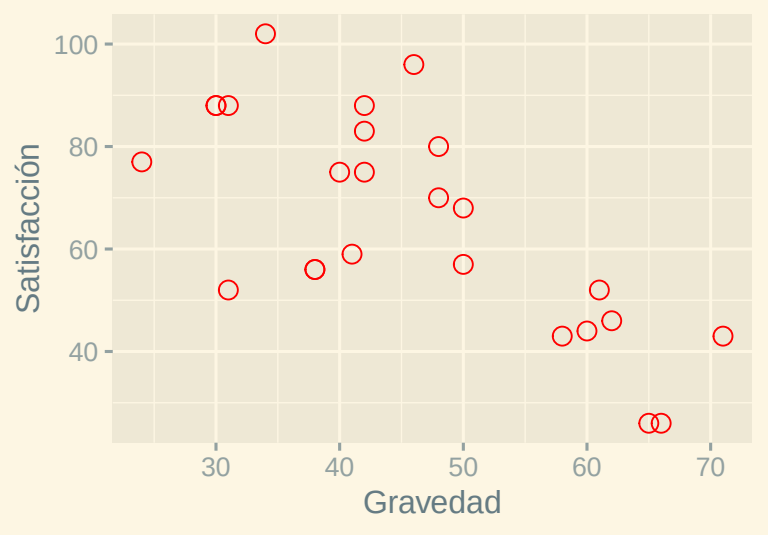
Outliers

Elaboramos gráficos de dispersión para las variables Satisfacción vs. Edad y Satisfacción vs. Gravedad.

```
ggplot(data = Datos_actividad) +
  geom_point(mapping = aes(x = Edad, y = Satisfacción),color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```




```
ggplot(data = Datos_actividad) +
  geom_point(mapping = aes(x = Gravedad, y = Satisfacción),color = "red", size = 3, shape = 1) +
  theme_solarized_2()
```



En las gráficas, no se observa algún punto en los diagramas de dispersión que se separe del resto de puntos, por lo que se concluye que no existen puntos atípicos.