Modelo_Estadistico_Base

María del Carmen Vargas Villarreal

5/9/2022

Momento de Retroalimentación: Módulo 1 Construcción de un modelo estadístico base (Portafolio Implementación)

EL PROBLEMA

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio.

X1 = número de indentificación X2 = nombre del lago X3 = alcalinidad (mg/l de carbonato de calcio) X4 = PH X5 = calcio (mg/l) X6 = clorofila (mg/l) X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago X8 = número de peces estudiados en el lago X9 = mínimo de la concentración de mercurio en cada grupo de peces X10 = máximo de la concentración de mercurio en cada grupo de peces X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

1) EXPLORACIÓN DE LA BASE DE DATOS

1.- Accede a la base de datos

```
datos_mercurio = read.csv(file = 'mercurio.csv')

head(datos_mercurio)
```

```
##   X1           X2    X3  X4   X5    X6   X7 X8   X9  X10  X11 X12
## 1  1    Alligator   5.9 6.1  3.0   0.7 1.23  5 0.85 1.43 1.53   1
## 2  2        Annie   3.5 5.1  1.9   3.2 1.33  7 0.92 1.90 1.33   0
## 3  3       Apopka 116.0 9.1 44.1 128.3 0.04  6 0.04 0.06 0.04   0
## 4  4 Blue Cypress  39.4 6.9 16.4   3.5 0.44 12 0.13 0.84 0.44   0
## 5  5        Brick   2.5 4.6  2.9   1.8 1.20 12 0.69 1.50 1.33   1
## 6  6       Bryant  19.6 7.3  4.5  44.1 0.27 14 0.04 0.48 0.25   1
```

2.- Explora las variables

```
print("Cantida de renglones")
```

```
## [1] "Cantida de renglones"
```

```
nrow(datos_mercurio)
```

```
## [1] 53
```

```
print("Cantidad de columnas")
```

```
## [1] "Cantidad de columnas"
```

```
ncol(datos_mercurio)
```

```
## [1] 12
```

```
sapply(datos_mercurio, class)
```

```
##          X1          X2          X3          X4          X5          X6
##   "integer" "character"   "numeric"   "numeric"   "numeric"   "numeric"
##          X7          X8          X9         X10         X11         X12
##   "numeric"   "integer"   "numeric"   "numeric"   "numeric"   "integer"
```

#3.- Exploración de la base de datos

3.1) Calcula medidas estadísticas: Variables cuantitativas

```
# Medidas de tendencia central: promedio, media y mediana
# Medidas de dispersión: rango: máximo - mínimo, varianza

summary(datos_mercurio)
```

```
##        X1          X2                  X3               X4
##  Min.   : 1   Length:53          Min.   :  1.20   Min.   :3.600
##  1st Qu.:14   Class :character   1st Qu.:  6.60   1st Qu.:5.800
##  Median :27   Mode  :character   Median : 19.60   Median :6.800
##  Mean   :27                      Mean   : 37.53   Mean   :6.591
##  3rd Qu.:40                      3rd Qu.: 66.50   3rd Qu.:7.400
##  Max.   :53                      Max.   :128.00   Max.   :9.100
##        X5             X6              X7               X8
##  Min.   : 1.1   Min.   :  0.70   Min.   :0.0400   Min.   : 4.00
##  1st Qu.: 3.3   1st Qu.:  4.60   1st Qu.:0.2700   1st Qu.:10.00
##  Median :12.6   Median : 12.80   Median :0.4800   Median :12.00
##  Mean   :22.2   Mean   : 23.12   Mean   :0.5272   Mean   :13.06
##  3rd Qu.:35.6   3rd Qu.: 24.70   3rd Qu.:0.7700   3rd Qu.:12.00
##  Max.   :90.7   Max.   :152.40   Max.   :1.3300   Max.   :44.00
##        X9              X10              X11              X12
```

```
##  Min.   :0.0400   Min.   :0.0600   Min.   :0.0400   Min.   :0.0000
##  1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500   1st Qu.:1.0000
##  Median :0.2500   Median :0.8400   Median :0.4500   Median :1.0000
##  Mean   :0.2798   Mean   :0.8745   Mean   :0.5132   Mean   :0.8113
##  3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000   3rd Qu.:1.0000
##  Max.   :0.9200   Max.   :2.0400   Max.   :1.5300   Max.   :1.0000
```

```
# Medidas de dispersión: desviación estándar

library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_long <- gather(datos_mercurio, factor_key=TRUE)

data_long%>% group_by(key)%>%
  summarise(sd= sd(value))
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introducidos por coerción

## # A tibble: 12 x 2
##    key       sd
##    <fct>  <dbl>
##  1 X1    15.4
##  2 X2    NA
##  3 X3    38.2
##  4 X4     1.29
##  5 X5    24.9
##  6 X6    30.8
##  7 X7     0.341
##  8 X8     8.56
##  9 X9     0.226
## 10 X10    0.522
## 11 X11    0.339
## 12 X12    0.395
```

3.2) Explora los datos usando herramientas de visualización

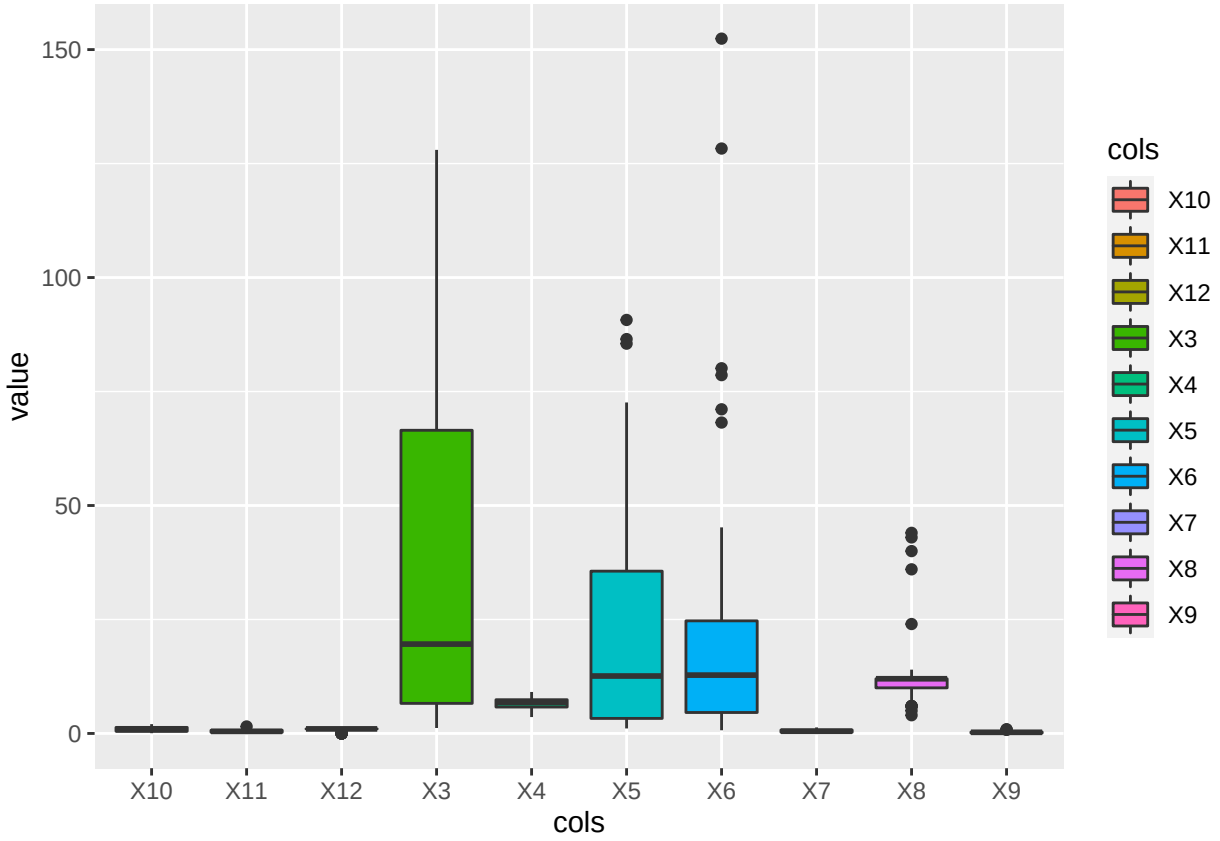
Medidas de posición: cuartiles, outlier (valores atípicos), boxplots. Análisis de Outliers

```
df_numericas = subset(datos_mercurio, select = -c(X1,X2)) # X1 es integer y X2 es character
data_long <- df_numericas %>%
  pivot_longer(colnames(df_numericas)) %>%
  as.data.frame()
head(data_long)
```

```
##   name value
## 1   X3  5.90
## 2   X4  6.10
## 3   X5  3.00
## 4   X6  0.70
## 5   X7  1.23
## 6   X8  5.00
```

Se separan solo las variables númericas

```
library(tidyr)
library(ggplot2)
df_tidy <- gather(df_numericas, cols, value)
ggplot(df_tidy, aes(x = cols, y=value)) +
  geom_boxplot(aes(fill=cols))
```



Para X6 “Clorofila” (mg/l)

```r
upper_bound <- quantile(datos_mercurio$X6, 0.99)
upper_bound
```

```
##     99%
## 139.868
```

```r
outlier_ind <- which(datos_mercurio$X6 > upper_bound)

datos_mercurio[outlier_ind, ]
```

```
##    X1     X2 X3  X4   X5    X6   X7 X8   X9  X10  X11 X12
## 38 38 Parker 53 8.4 45.6 152.4 0.04  4 0.04 0.06 0.04   0
```

Para X8 "Número de peces estudiados en el lago"

```r
upper_bound <- quantile(datos_mercurio$X8, 0.99)
upper_bound
```

```
##   99%
## 43.48
```

```r
outlier_ind <- which(datos_mercurio$X8 > upper_bound)

datos_mercurio[outlier_ind, ]
```

```
##    X1           X2   X3  X4   X5   X6   X7 X8  X9 X10  X11 X12
## 47 51 Tohopekaliga 25.6 6.2 12.6 27.7 0.65 44 0.3 1.1 0.58   1
```

Análisis de distribución de datos (Histogramas)

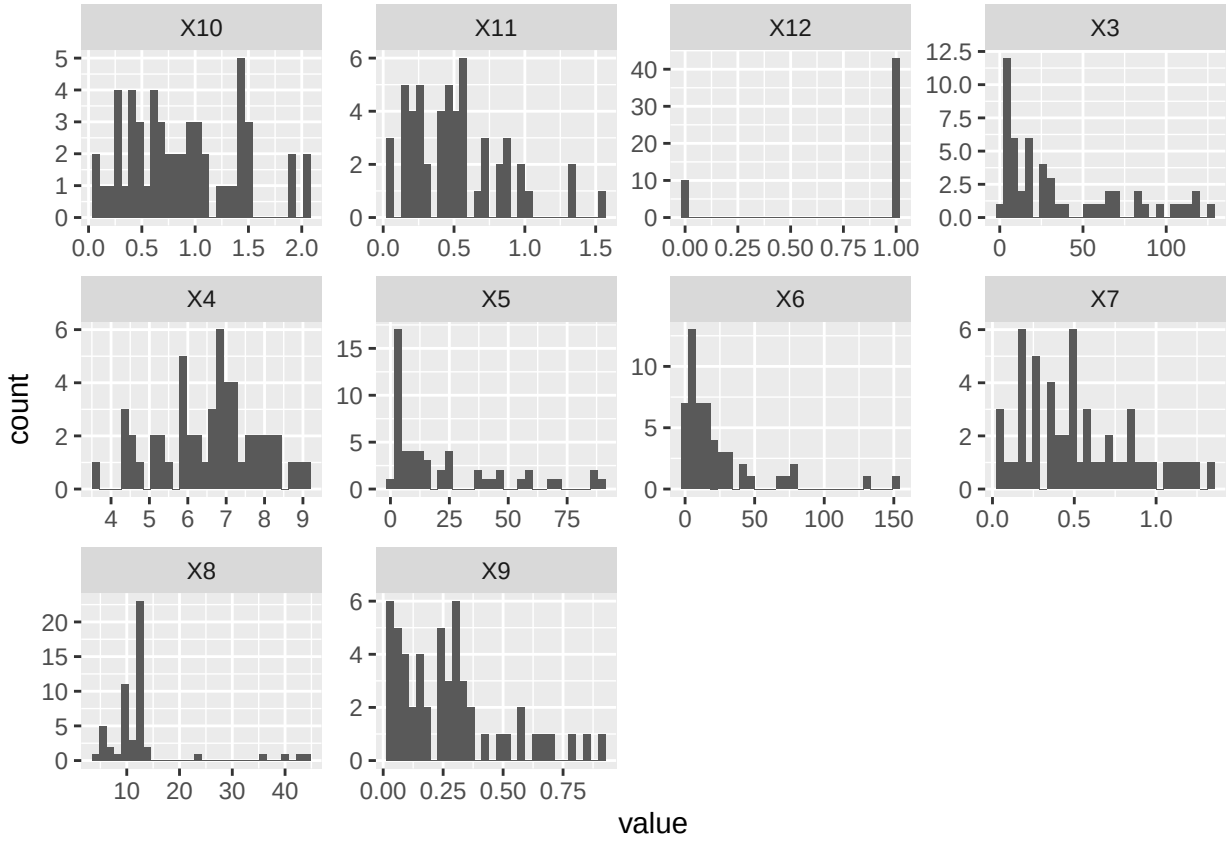
Separación de variables númericas

```r
df_numericas = subset(datos_mercurio, select = -c(X1,X2)) # X1 es integer y X2 es character
data_long <- df_numericas %>%
  pivot_longer(colnames(df_numericas)) %>%
  as.data.frame()
head(data_long)
```

```
##   name value
## 1   X3  5.90
## 2   X4  6.10
## 3   X5  3.00
## 4   X6  0.70
## 5   X7  1.23
## 6   X8  5.00
```

```r
library(ggplot2)
ggp1 <- ggplot(data_long, aes(x = value)) +    # Se imprime cada columna como histograma
  geom_histogram() +
  facet_wrap(~ name, scales = "free")
ggp1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



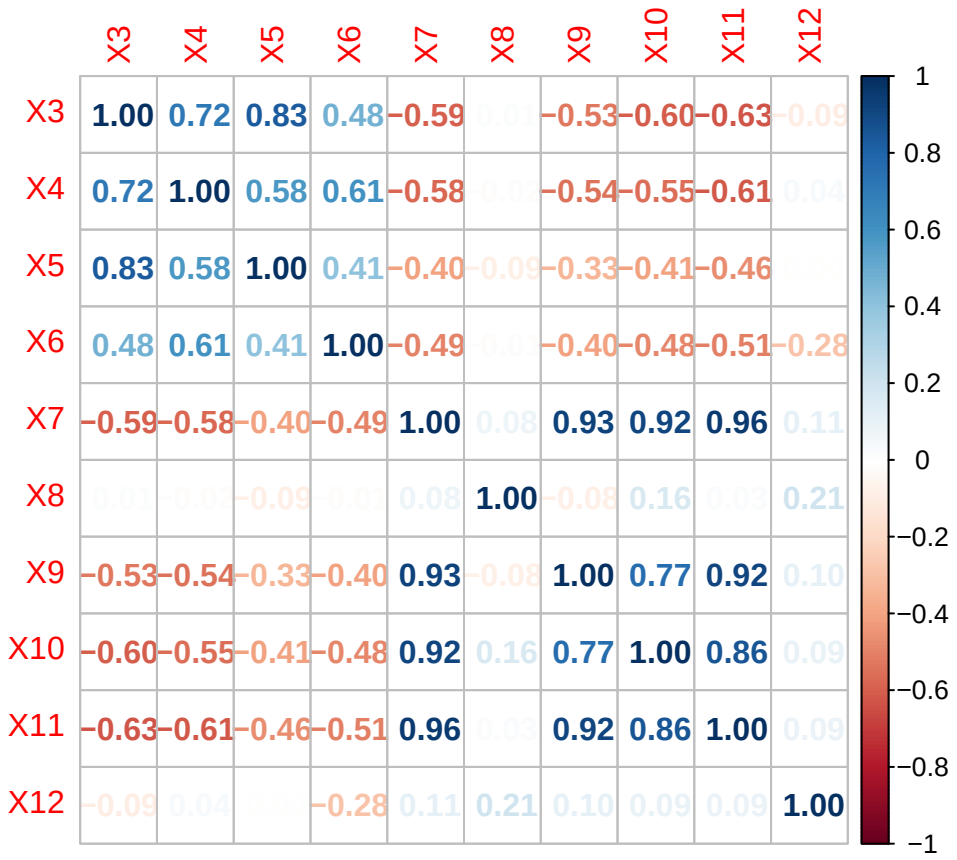
Se puede observar que la variable X4 = PH es el histograma que representa mayor simetría, en contraste con el resto que se refleja simetría

3.3) Explora la correlación entre las variables.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
v = cor(df_numericas)
corrplot(v, method = 'number')
```



Interpretación: Tomando en cuenta que las siguientes variables:

X9 = mínimo de la concentración de mercurio en cada grupo de peces X10 = máximo de la concentración de mercurio en cada grupo de peces X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

Están muy correlacionadas sí y con respecto a la variable dependiente, siendo X7. Las correlaciones lineales entre pares de las variables mencionadas no deberían de exceder 0.5.

A este problema se le conoce colinealidad o multicolinealidad, que es cuando las variables predictoras del modelo se encuentran relacionadas. Es por esto que, para el modelo de regresión, se tomará la decisión de descartar algunas de estas variables para que no alteren los resultados del modelo, o la otra opción sería emplear un PCA y utilizar los componentes generados como nuevas variables independientes. Sin embargo, por motivos prácticos solo se recurrirá a eliminar las variables X9 y X10, y mantener la variable X10 ya que es la refleja mayor correlación con respecto a X7, la variable dependiente seleccionada.

4.- Preguntas base

¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Primero, se realizará un análisis de regresión, independientemente de lo descubierto en la gráfica de correlación. De todas estas variables:

X1 = número de indentificación X2 = nombre del lago X3 = alcalinidad (mg/l de carbonato de calcio) X4 = PH X5 = calcio (mg/l) X6 = clorofila (mg/l) X7 = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago X8 = número de peces estudiados en el lago

X9 = mínimo de la concentración de mercurio en cada grupo de peces X10 = máximo de la concentración de mercurio en cada grupo de peces X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Se tomará a X7 como la variable dependiente, ya que se busca encontrar qué factores/variables influyen o explican en el nivel de contaminación/concentración por mercurio que describen dicha variable.

Modelo Inicial, el cual toma en cuenta todas las variables como las variables independientes, a excepción de X7 que es la dependiente

```
modelo_1=lm(X7~.,data=df_numericas)
summary(modelo_1)
```

```
##
## Call:
## lm(formula = X7 ~ ., data = df_numericas)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.177686 -0.026022 -0.001994  0.029491  0.112589
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0746097  0.0706873  -1.055  0.29710
## X3          -0.0002395  0.0005521  -0.434  0.66653
## X4           0.0115200  0.0115096   1.001  0.32247
## X5           0.0004408  0.0006752   0.653  0.51729
## X6          -0.0004808  0.0004012  -1.199  0.23725
## X8           0.0021498  0.0011529   1.865  0.06907 .
## X9           0.5873587  0.1041611   5.639 1.22e-06 ***
## X10          0.2282644  0.0348793   6.544 5.90e-08 ***
## X11          0.3054867  0.0870650   3.509  0.00107 **
## X12         -0.0155862  0.0253791  -0.614  0.54236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06018 on 43 degrees of freedom
## Multiple R-squared:  0.9742, Adjusted R-squared:  0.9689
## F-statistic: 180.8 on 9 and 43 DF,  p-value: < 2.2e-16
```

Interpretación: R^2 = 0.9742. Evidentemente el modelo es muy bueno a juzgar por el valor de R^2, la cual indica la cercanía de los datos a la línea de regresión ajustada, debido a que las variables X9, X10 Y X11 ya cuentan con la información directa de los niveles mínimos y máximos de concentración de mercurio, además de la estimación de (mediante regresión) de la concentración.

A continuación, se utiliza la siguiente linea de código que obtiene una serie de modelos con diferentes combinaciones de variables hasta dar con la combinación que maximiza la eficiencia, es decir, el modelo más óptimo.

```
step(modelo_1,direction="both",trace=1)
```

```
## Start:  AIC=-288.98
## X7 ~ X3 + X4 + X5 + X6 + X8 + X9 + X10 + X11 + X12
##
##        Df Sum of Sq     RSS     AIC
## - X3    1  0.000682 0.15643 -290.75
## - X12   1  0.001366 0.15711 -290.52
## - X5    1  0.001544 0.15729 -290.46
## - X4    1  0.003629 0.15937 -289.76
## - X6    1  0.005203 0.16095 -289.24
## <none>              0.15575 -288.98
## - X8    1  0.012593 0.16834 -286.86
## - X11   1  0.044591 0.20034 -277.64
## - X9    1  0.115171 0.27092 -261.64
## - X10   1  0.155128 0.31087 -254.35
##
## Step:  AIC=-290.75
## X7 ~ X4 + X5 + X6 + X8 + X9 + X10 + X11 + X12
##
##        Df Sum of Sq     RSS     AIC
## - X5    1  0.000902 0.15733 -292.44
## - X12   1  0.000912 0.15734 -292.44
## - X4    1  0.002949 0.15938 -291.76
## - X6    1  0.004629 0.16106 -291.20
## <none>              0.15643 -290.75
## + X3    1  0.000682 0.15575 -288.98
## - X8    1  0.012026 0.16845 -288.82
## - X11   1  0.044567 0.20100 -279.46
## - X9    1  0.114637 0.27106 -263.61
## - X10   1  0.182554 0.33898 -251.76
##
## Step:  AIC=-292.44
## X7 ~ X4 + X6 + X8 + X9 + X10 + X11 + X12
##
##        Df Sum of Sq     RSS     AIC
## - X12   1  0.000920 0.15825 -294.13
## - X6    1  0.004583 0.16191 -292.92
## - X4    1  0.004921 0.16225 -292.81
## <none>              0.15733 -292.44
## + X5    1  0.000902 0.15643 -290.75
## - X8    1  0.011813 0.16914 -290.61
## + X3    1  0.000040 0.15729 -290.46
## - X11   1  0.044460 0.20179 -281.25
## - X9    1  0.127243 0.28457 -263.03
## - X10   1  0.184291 0.34162 -253.35
##
## Step:  AIC=-294.13
## X7 ~ X4 + X6 + X8 + X9 + X10 + X11
##
##        Df Sum of Sq     RSS     AIC
## - X6    1  0.003671 0.16192 -294.92
## - X4    1  0.004107 0.16236 -294.78
## <none>              0.15825 -294.13
## - X8    1  0.010893 0.16914 -292.61
## + X12   1  0.000920 0.15733 -292.44
```

```
## + X5    1  0.000911 0.15734 -292.44
## + X3    1  0.000162 0.15809 -292.19
## - X11   1  0.046058 0.20431 -282.60
## - X9    1  0.127701 0.28595 -264.78
## - X10   1  0.186249 0.34450 -254.91
##
## Step:  AIC=-294.92
## X7 ~ X4 + X8 + X9 + X10 + X11
##
##        Df Sum of Sq     RSS     AIC
## - X4    1  0.001774 0.16370 -296.34
## <none>              0.16192 -294.92
## + X6    1  0.003671 0.15825 -294.13
## - X8    1  0.010011 0.17193 -293.74
## + X5    1  0.000858 0.16106 -293.20
## + X3    1  0.000249 0.16167 -293.00
## + X12   1  0.000009 0.16191 -292.92
## - X11   1  0.052293 0.21422 -282.09
## - X9    1  0.124057 0.28598 -266.77
## - X10   1  0.192747 0.35467 -255.36
##
## Step:  AIC=-296.34
## X7 ~ X8 + X9 + X10 + X11
##
##        Df Sum of Sq     RSS     AIC
## <none>              0.16370 -296.34
## - X8    1  0.010273 0.17397 -295.12
## + X5    1  0.001956 0.16174 -294.98
## + X4    1  0.001774 0.16192 -294.92
## + X6    1  0.001338 0.16236 -294.78
## + X3    1  0.001299 0.16240 -294.76
## + X12   1  0.000003 0.16369 -294.34
## - X11   1  0.050769 0.21446 -284.02
## - X9    1  0.126611 0.29031 -267.98
## - X10   1  0.191228 0.35492 -257.33


##
## Call:
## lm(formula = X7 ~ X8 + X9 + X10 + X11, data = df_numericas)
##
## Coefficients:
## (Intercept)           X8           X9          X10          X11
##    -0.01651      0.00176      0.57486      0.23673      0.29777
```

Interpretación: Evidentemente, el "mejor modelo" contiene las variables X9, X10 y X11, las cuales, como se ha mencionado anteriormente, ya dan información exacta sobre las concentraciones de mercurio en grupos de peces y la estimación en el pez de 3 años.

De todas formas, a continuación se genera el modelo correspondiente a la combinación de variables obtenidas en el paso anterior:

```
modelo_1=lm(X7~X8+X9+X10+X11,data=df_numericas)
summary(modelo_1)
```

```
##
## Call:
## lm(formula = X7 ~ X8 + X9 + X10 + X11, data = df_numericas)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.188438 -0.028797 -0.003159  0.030784  0.117510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.016513   0.019166  -0.862  0.39321
## X8           0.001760   0.001014   1.736  0.08906 .
## X9           0.574864   0.094347   6.093 1.81e-07 ***
## X10          0.236728   0.031613   7.488 1.32e-09 ***
## X11          0.297773   0.077177   3.858  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0584 on 48 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9707
## F-statistic: 431.3 on 4 and 48 DF,  p-value: < 2.2e-16
```

Interpretación: El modelo es casi perfecto, obteniéndose una valor R^2 = 0.9729

Segundo Modelo con variables de interés (ahora ya no tomando en cuenta X9 ni X10), para encontrar realmente qué factores influyen en el nivel de concentración de mercurio sin tomar en consideración las variable que contengan cualquier tipo de información previa sobre el mercurio, a excepción de la variable X11, de la cual la explicación del por qué se mantiene dentro del modelo ya fue anteriormete mencionada.

Segundo Modelo con variables de interés

```
modelo_2=lm(X7~X3+X4+X5+X6+X8+X11+X12,data=df_numericas)
summary(modelo_2)
```

```
##
## Call:
## lm(formula = X7 ~ X3 + X4 + X5 + X6 + X8 + X11 + X12, data = df_numericas)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.245347 -0.050039 -0.005074  0.037133  0.301195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0574307  0.1136244  -0.505   0.6157
## X3          -0.0013055  0.0008355  -1.563   0.1252
## X4           0.0118741  0.0180885   0.656   0.5149
## X5           0.0021560  0.0010319   2.089   0.0424 *
## X6          -0.0004071  0.0006168  -0.660   0.5125
## X8           0.0029527  0.0016633   1.775   0.0826 .
```

```
## X11          0.9555246  0.0550944  17.343   <2e-16 ***
## X12         -0.0148500  0.0395735  -0.375   0.7092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09678 on 45 degrees of freedom
## Multiple R-squared:  0.9303, Adjusted R-squared:  0.9195
## F-statistic: 85.82 on 7 and 45 DF,  p-value: < 2.2e-16
```

Ahora buscaremos la combinación de variables que aporten mayor valor al modelo:

```
step(modelo_2,direction="both",trace=1)
```

```
## Start:  AIC=-240.22
## X7 ~ X3 + X4 + X5 + X6 + X8 + X11 + X12
##
##        Df Sum of Sq    RSS     AIC
## - X12   1   0.00132 0.4228 -242.05
## - X4    1   0.00404 0.4255 -241.71
## - X6    1   0.00408 0.4256 -241.71
## <none>              0.4215 -240.22
## - X3    1   0.02287 0.4443 -239.42
## - X8    1   0.02951 0.4510 -238.63
## - X5    1   0.04089 0.4624 -237.31
## - X11   1   2.81730 3.2388 -134.14
##
## Step:  AIC=-242.05
## X7 ~ X3 + X4 + X5 + X6 + X8 + X11
##
##        Df Sum of Sq    RSS     AIC
## - X6    1   0.00289 0.4257 -243.69
## - X4    1   0.00296 0.4258 -243.68
## <none>              0.4228 -242.05
## - X3    1   0.02160 0.4444 -241.41
## - X8    1   0.02834 0.4511 -240.61
## + X12   1   0.00132 0.4215 -240.22
## - X5    1   0.03972 0.4625 -239.29
## - X11   1   2.82321 3.2460 -136.02
##
## Step:  AIC=-243.69
## X7 ~ X3 + X4 + X5 + X8 + X11
##
##        Df Sum of Sq    RSS     AIC
## - X4    1   0.00134 0.4270 -245.52
## <none>              0.4257 -243.69
## - X3    1   0.02070 0.4464 -243.17
## - X8    1   0.02801 0.4537 -242.31
## + X6    1   0.00289 0.4228 -242.05
## + X12   1   0.00013 0.4256 -241.71
## - X5    1   0.03820 0.4639 -241.13
## - X11   1   2.99381 3.4195 -135.26
##
## Step:  AIC=-245.52
```

```
## X7 ~ X3 + X5 + X8 + X11
##
##        Df Sum of Sq    RSS     AIC
## <none>              0.4270 -245.52
## - X3    1    0.0197 0.4468 -245.13
## - X8    1    0.0277 0.4547 -244.20
## + X4    1    0.0013 0.4257 -243.69
## + X6    1    0.0013 0.4258 -243.68
## + X12   1    0.0000 0.4270 -243.53
## - X5    1    0.0379 0.4649 -243.01
## - X11   1    3.2315 3.6585 -133.68
```

```
##
## Call:
## lm(formula = X7 ~ X3 + X5 + X8 + X11, data = df_numericas)
##
## Coefficients:
## (Intercept)           X3           X5           X8          X11
##   -0.003915    -0.001073     0.002008     0.002744     0.956615
```

Interpretación: Como se muestra en la última tabla, junto a las variables X3, X5, X8 y X11 se encuentra un signo de (-), lo cual indica que al descartar dichas variables el modelo empeorará en términos de eficencia.

Las variables X3 = alcalinidad (mg/l de carbonato de calcio), X5 = calcio (mg/l) y X8 = número de peces estudiados en el lago y X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible), son las que se tomarán en cuenta para el segundo modelo:

```
segundo_modelo=lm(X7~X3+X5+X8+ X11,data=df_numericas)
summary_segundo_modelo = summary(segundo_modelo)
summary_segundo_modelo
```

```
##
## Call:
## lm(formula = X7 ~ X3 + X5 + X8 + X11, data = df_numericas)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.243120 -0.060464 -0.004841  0.038591  0.300959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0039152  0.0442001  -0.089   0.9298
## X3          -0.0010727  0.0007200  -1.490   0.1428
## X5           0.0020078  0.0009727   2.064   0.0444 *
## X8           0.0027438  0.0015560   1.763   0.0842 .
## X11          0.9566154  0.0501934  19.059   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09432 on 48 degrees of freedom
## Multiple R-squared:  0.9294, Adjusted R-squared:  0.9235
## F-statistic: 157.9 on 4 and 48 DF,  p-value: < 2.2e-16
```

Llevar a cabo una prueba de Hipotesis para verificar que B1 sea significativa con un nivel de confianza de 0.3 no es necesario debido a que la columna Pr(>|t|) de la tabla obtenida ya asocia el p-value junto el t-value. Si el p-value es menor al nivel de significancia, por ejemplo, alfa = 0.03, entonces la variable predictora tiene una relación significativa con la variable de respuesta del modelo.

X3: Pr(>|t|) = 0.1428 > 0.03, por lo que esta variable (X3 = alcalinidad mg/l de carbonato de calcio) no es significativa para el modelo

X5: Pr(>|t|) = 0.0444 > 0.03, esta variable (calcio mg/l) por muy poco no es significativa para el modelo

X8: Pr(>|t|) = 0.0842 > 0.03, por lo que esta variable (número de peces estudiados en el lago) no es significativa para el modelo

X11: Pr(>|t|) = <2e-16 < 0.03, si es muy significativa para el modelo. Tiene sentido ya que es la variable con más correlación.

Analizando la tabla más detalladamente, se puede observar que:

- X11 es muy significativa ya que las *** indican que entra con todos los niveles de significancia.
- X8 no es significativa ya que el (.) indica que entra con el nivel de significancia = 0.1, el cual es nivel bastante ineficiente.
- X5 es significativa que el * indica ue entra con el nivel de significancia = 0.05, el cual todavía es estadísticamente significativo.
- X3, por último, no es signifcativa porque ni si quiera forma de ningún nivel de significancia.

Por estas razones, descartar solamente X3 y X8 podría aportar mayormente al modelo.

A continuación se realiza un 3er modelo considerando solo X5 y X11:

```
tercer_modelo=lm(X7~X5+ X11,data=df_numericas)
summary_tercer_modelo = summary(tercer_modelo)
summary_tercer_modelo
```

```
##
## Call:
## lm(formula = X7 ~ X5 + X11, data = df_numericas)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.291283 -0.044776 -0.009983  0.023924  0.273074
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0004969  0.0342415   0.015    0.988
## X5          0.0007811  0.0006075   1.286    0.204
## X11         0.9924463  0.0447139  22.195   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09673 on 50 degrees of freedom
## Multiple R-squared:  0.9227, Adjusted R-squared:  0.9196
## F-statistic: 298.2 on 2 and 50 DF,  p-value: < 2.2e-16
```

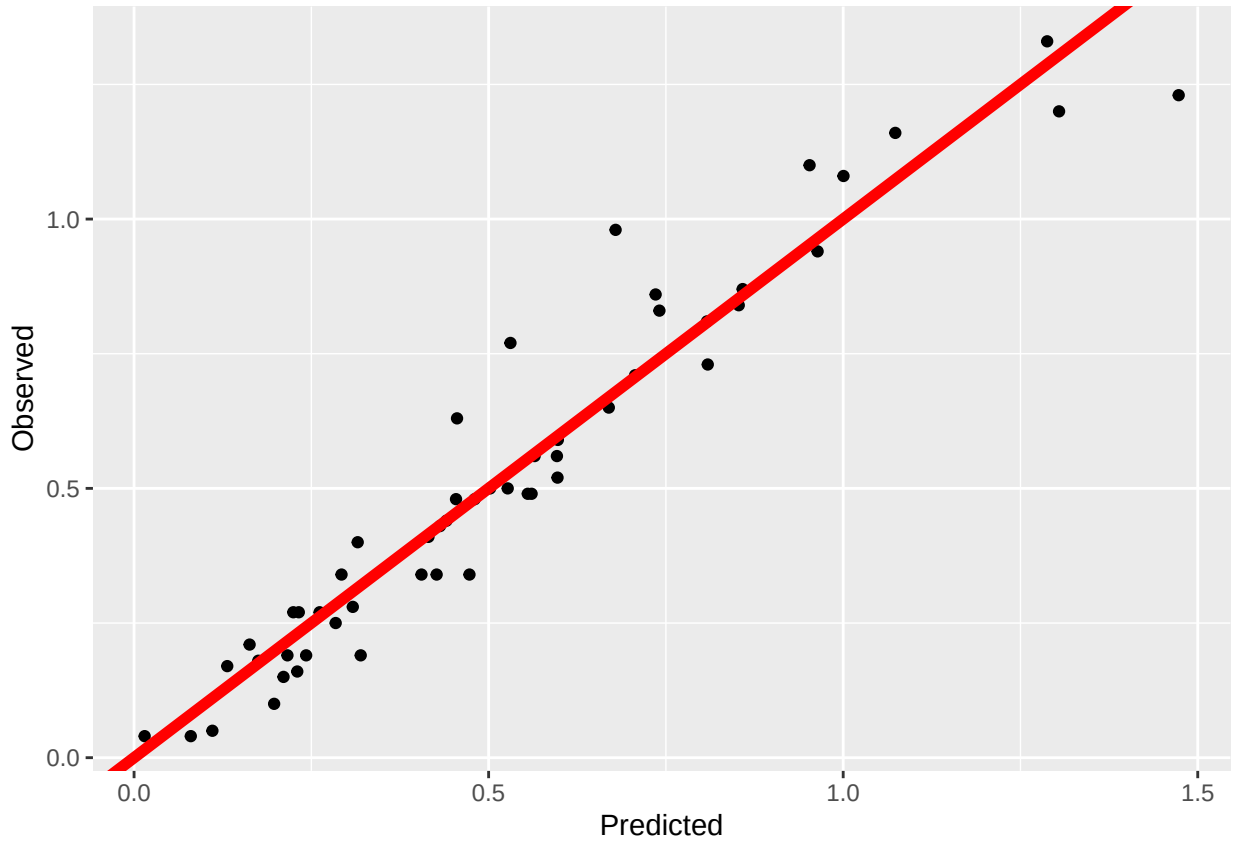
Se obtuvo un R^2 = 0.9227, el cual no es tan diferente del modelo que utiliza X3 + X5 + X8 + X11 (Segundo Modelo). Esto me permite concluir que el modelo obtenido por el comando "step" utilizando el criterio estadístico AIC (donde se busca el valor más bajo posible), es mucho mejor que verificar la significancia de

las variables y que además, este método sí toma en cuenta variables que verdaderamente aportan al modelo independientemente de las "contradicciones" que surjan a partir de la significancia.

Por esta razón, se continuará con la interpretación de los gráficos utilizando el Segundo Modelo.

```
data_modelo2 <- data.frame(Predicted = predict(segundo_modelo),
                           Observed = df_numericas$X7)
```

```
ggplot(data_modelo2,
       aes(x = Predicted,
           y = Observed)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 2)
```



1. ¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?

Las normativas de referencia para evaluar los niveles máximos de Hg (Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995) establecen que la concentración promedio de mercurio en productos de la pesca no debe superar los 0.5 mg de Hg/kg.

Se realiza una prueba de hipótesis para verificar si hay evidencia significativa para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana con un nivel de significancia de 0.03.

B_0: Mu $>$ 0.5 B_1: Mu $\leq$ 0.5

```
alpha = 0.03
t0 = qt(alpha,length(df_numericas$X3))
cat("t0 = ", t0)
```

```
## t0 =  -1.921914
```

```
m = mean(df_numericas$X7)
s = sd(df_numericas$X7)
n = length(df_numericas$X7)

sm = s/sqrt(n)

t = (m-0.5)/sm
p = pt(t,n-2)

cat("te = ", t, "\n")
```

```
## te =  0.5799957
```

```
cat("p = ", p)
```

```
## p =  0.7177644
```

Debido a que el valor de |t|= 0.5799957 es mayor a |t0| = -1.921914, y el p-value = 0.7177644 es mayor a alpha = 0.03:

No se rechaza la hipótesis nula, hay evidencia significativa para suponer que la concentración promedio de mercurio en los lagos es dañina para la salud humana.

2. ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?

```
jovenes = df_numericas[df_numericas$X12 == 0,]
maduros = df_numericas[df_numericas$X12 == 1,]

m1 = mean(jovenes$X7)
m2 = mean(maduros$X7)

cat("Promedio de concentración media de mercurio en los peces jóvenes: ",m1, "\n")
```

```
## Promedio de concentración media de mercurio en los peces jóvenes:  0.451
```

```r
cat("Promedio de concentración media de mercurio en los peces maduros: ", m2)
```

```
## Promedio de concentración media de mercurio en los peces maduros:  0.5448837
```

Se realiza una prueba de hipótesis para verificar si no hay una diferencia significativa entre la concentración de mercurio por la edad de los peces con un nivel de significancia de 0.03:

B_0: $Mu_1 = Mu_2$ B_1: $Mu_1 \neq Mu_2$

```r
alpha = 0.03
t0 = qt(alpha/2,length(df_numericas))
cat("t0 = ", t0)
```

```
## t0 =  -2.527484
```

```r
t.test(jovenes$X7, maduros$X7,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.97)
```

```
##
##  Welch Two Sample t-test
##
## data:  jovenes$X7 and maduros$X7
## t = -0.67932, df = 11.831, p-value = 0.51
## alternative hypothesis: true difference in means is not equal to 0
## 97 percent confidence interval:
##  -0.4346009  0.2468335
## sample estimates:
## mean of x mean of y
## 0.4510000 0.5448837
```

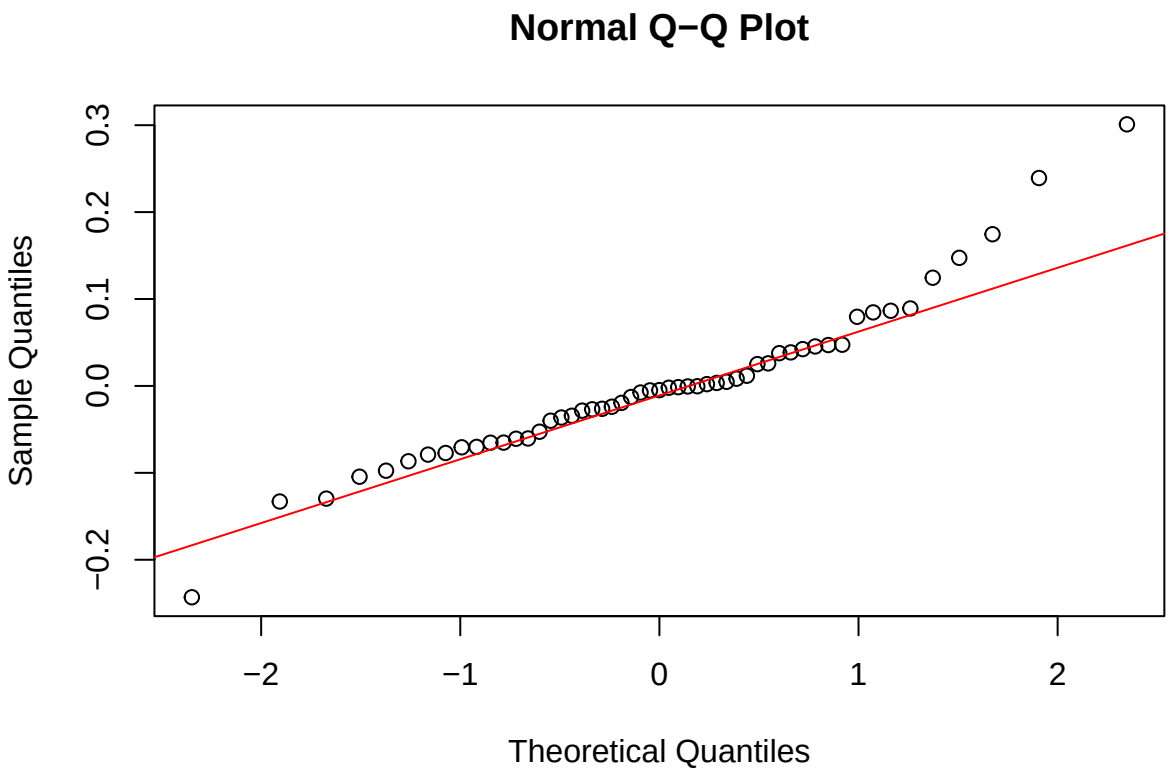
Debido a que el valor de |t|= -0.67932 es mayor a |t0| = -2.230086, y el p-value = 0.51 es mayor a alpha = 0.03:

No se rechaza, no hay una diferencia significativa entre la concentración de mercurio por la edad de los peces con un nivel de significancia de 0.03

Verificación de Supuestos

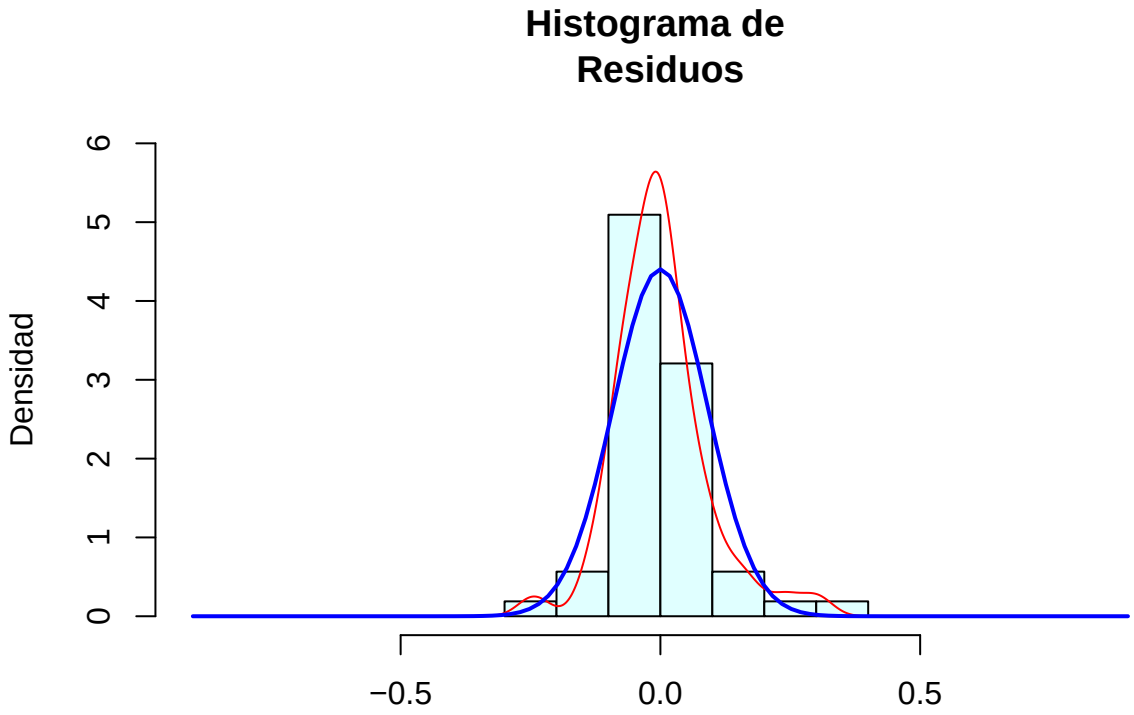
```r
E=segundo_modelo$residuals  # Error: residuals
Y=segundo_modelo$fitted.values # Y predicted

qqnorm(E)
qqline(E,col="red")
```



Interpretación: La mayoría de los valores estimados tienden a ajustarse a la recta roja diagonal, a excepción de aquellos puntos en los extremos que se desvían de la linea. De forma general, no se puede afirmar distribución normal de los datos ya que al final se forma una cola lateral en la parte derecha. Término: skewness

```
hist(E,col="lightcyan",freq=FALSE,main="Histograma de
Residuos",ylim=c(0,6),xlim = c(-0.9,0.9),xlab="",ylab="Densidad")
lines(density(E),col="red")
curve(dnorm(x,mean=mean(E),sd=sd(E)), add=TRUE, col="blue",lwd=2)
```



Interpretación: Claramente la asimetría se refleja en este histograma, por lo que se reitera una distribución no normal de los datos del modelo.

```
shapiro.test(E)
```

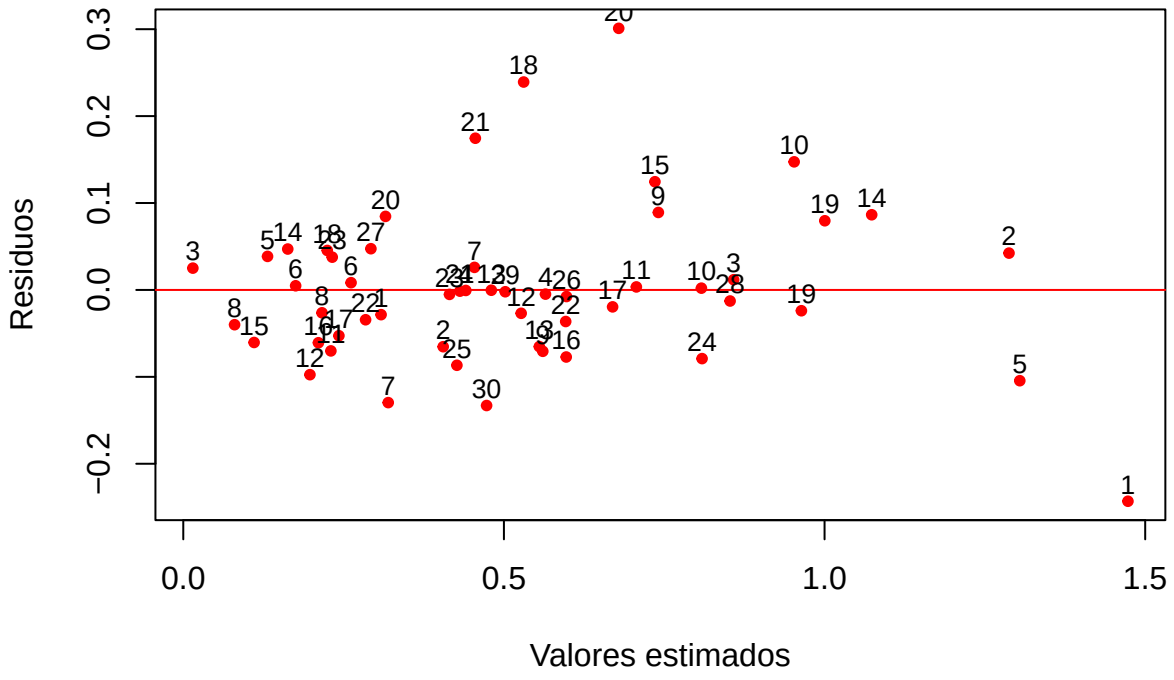
```
##
##  Shapiro-Wilk normality test
##
## data:  E
## W = 0.93772, p-value = 0.008241
```

La hipótesis nula del test Shapiro-Wilk es que la población representa una distribución normal. Por lo tanto, un valor de $p < 0.05$ indica que se debe rechazar la hipótesis nula. En otras palabras, los datos no poseen distribución normal.

Para este caso, nuestro valor p-value = 0.008241 es menor a $p = 0.005$, por lo que se rechaza la hipótesis nula.

Homocedasticidad y modelo apropiado

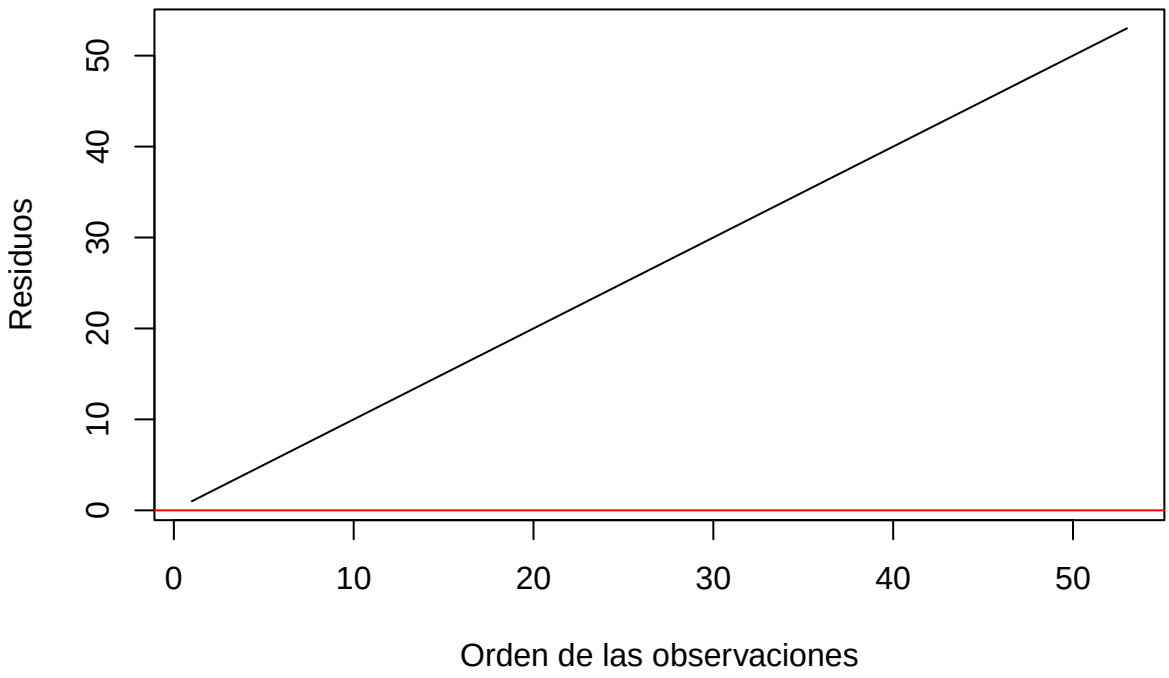
```
plot(Y,E,ylab="Residuos",xlab="Valores estimados",pch=20,col="red")
abline(h=0,col="red")
text(Y[],E[],1:30,cex=0.8,pos=3,offset=0.2)
```



Los puntos no mantienen la misma dispersión en las distintas zonas de la variable X, por lo que se descarta el cumplimiento del parámetro de la homocedasticidad en el modelo. La varianza del error de la variable que está siendo predicha (o desviación estándar de la variable dependiente) no se mantiene uniforme con respecto a las variables independientes.

Independencia

```
n=length(df_numericas$X6)
plot(c(1:n),df_numericas$residuals,type="l",xlab="Orden de las observaciones",ylab="Residuos")
abline(h=0,col="red")
```



Interpretación: En esta gráfica no se detectan patrones por lo tanto. Lo que si se observa es que hay cambios de patrones ascendentes y descendentes en los residuales.

Conclusión final:

El modelo final fue, con una evaluación alta de 0.92 de r^2, indicando que el modelo es lo suficientemente bueno y se ajusta a la variable que se busca explicar –> X7 = -0.003915 - 0.001073 X3 + 0.002008 X5 + 0.002744 X8 + 0.956615 X11

Recordando que:

X3 = alcalinidad (mg/l de carbonato de calcio)

X5 = calcio (mg/l)

X8 = número de peces estudiados en el lago

X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

Esto quiere decir que la concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago disminuirá -0.003915 unidades si la alcalinidad, el calcio, número de peces estudiados y la estimación de la concentración en el pez de 3 años son cero.

b1 = - 0.001073 X3, quiere decir que la concentración media de mercurio disminuirá - 0.001073 unidades de concentración media de mercurio si se aumenta la alcalinidad mg/l en una unidad.

b2 = 0.002008 X5 quiere decir que la concentración media de mercurio aumentará 0.002008 unidades de concentración media de mercurio si se aumenta el calcio mg/l en una unidad.

b3 = 0.002744 X8, quiere decir que la concentración media de mercurio aumentará aumentará 0.002744 unidades de concentración media de mercurio si se aumenta el número de peces estudiados en una unidad.

Por último, b4 = 0.956615 X11 , quiere decir que la concentración media de mercurio aumentará aumentará 0.956615 unidades de concentración media de mercurio si se aumenta la estimación de laconcentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) en una unidad.

Cada una de estas interpretaciones individuales son válidas cuando el resto de los coeficientes se mantienen constantes.

Con este proyecto se logró construir un modelo estadístico a partir del set de datos mercurio.csv, utilizando variables que fueron seleccionadas con sustento de significancia estadística.

SMA0101A

Construye un modelo estadístico base a partir de un set de datos, seleccionando las variables a utilizar. Explica correctamente cada una de las variables seleccionadas en el modelo y su utilidad en el modelo. Explica correctamente como funciona el modelo que utiliza y valida los supuestos del modelo.