

Construcción de un modelo estadístico base

Jesús David Núñez Rodríguez A01634928

2022-09-09

Descripción del problema:

Detectar los factores que afectan el nivel de contaminación por mercurio en los peces de agua dulce comestibles. Los datos recolectados son de un estudio realizado en 53 lagos de florida.

Preguntas a responder:

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

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

Si el muestreo se realizó lanzando una red y analizando los peces que la red encontraba ¿Habrá influencia del número de peces encontrados en la concentración de mercurio en los peces?

¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?

Consideraciones

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.

Introducción

Hay evidencia de que tiene efectos adversos en la salud de los humanos, especialmente en el desarrollo neuronal en fetos y algunos organos como hígado y riñón(Li et al. 2010). Los grupos más susceptibles de intoxicación por mercurio son fetos en desarrollo, mujeres que amamanten y niños pequeños. además hay evidencia que en adultos también tiene efectos neurotoxicos y afecta al sistema inmune y cardiovascular, el consumo de MeHg incluso en pequeñas cantidades es dañino si se está expuesto durante periodos prolongados de tiempo(USEPA 2013).

Etaapa de exploración

| ## | 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 |

El tamaño del dataframe es de 53 filas con 12 variables, de las cuales la variable dependiente es X7, Variables que sirven de identificador X1, X2 y finalmente el resto de variables son independientes(se validará más adelante si efectivamente son independientes).

Descripción de variables(nombre, descripción, tipo de variable):

X1 = número de indentificación, variable cuantitativa discreta

X2 = nombre del lago, variable cualitativa nominal

X3 = alcalinidad (mg/l de carbonato de calcio), variable cuantitativa continua

X4 = PH, variable cuantitativa continua

X5 = calcio (mg/l), variable cuantitativa continua

X6 = clorofila (mg/l), variable cuantitativa continua

X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago, variable cuantitativa continua

X8 = número de peces estudiados en el lago, variable cuantitativa continua

X9 = mínimo de la concentración de mercurio en cada grupo de peces, variable cuantitativa continua

X10 = máximo de la concentración de mercurio en cada grupo de peces, variable cuantitativa continua

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), variable cuantitativa continua

X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros), variable cuantitativa discreta

Descripción de los datos.

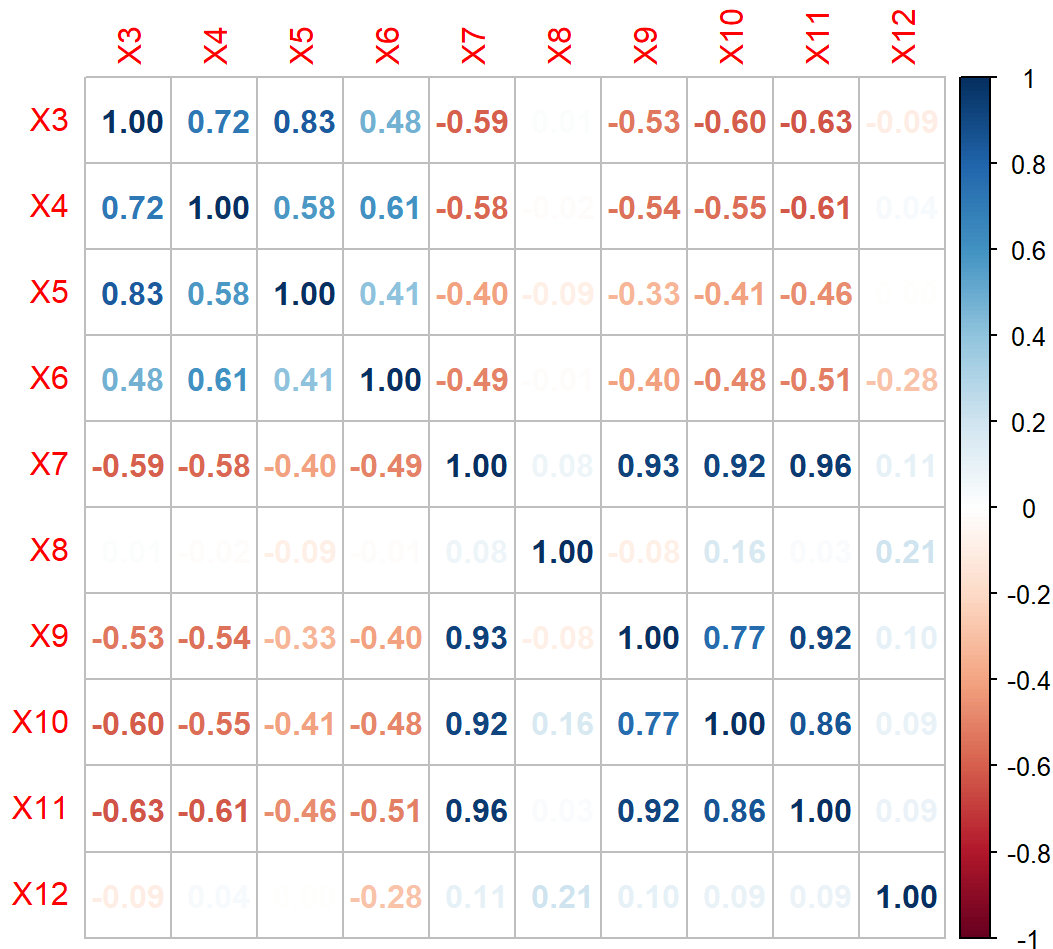
```
##           X1           X2           X3           X4
## Min.      : 1   Length:53   Min.      : 1.20   Min.      :3.600
## 1st Qu.:14   Class :character 1st Qu.: 6.60   1st Qu.:5.800
## Median :27   Mode  :character Median : 19.60   Median :6.800
## Mean    :27                                Mean    : 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
```

```
## [1] "sd: 15.4434" "sd: 38.2035" "sd: 1.2884" "sd: 24.9326" "sd: 30.8163"
## [6] "sd: 0.341" "sd: 8.5607" "sd: 0.2264" "sd: 0.522" "sd: 0.3387"
## [11] "sd: 0.395"
```

| | | | | |
|----|-------------------|--------------|--------------|--------------|
| ## | | | | |
| ## | Alligator | Annie | Apopka | Blue Cypress |
| ## | 1 | 1 | 1 | 1 |
| ## | Brick | Bryant | Cherry | Crescent |
| ## | 1 | 1 | 1 | 1 |
| ## | Deer Point | Dias | Dorr | Down |
| ## | 1 | 1 | 1 | 1 |
| ## | East Tohopekaliga | Eaton | Farm-13 | George |
| ## | 1 | 1 | 1 | 1 |
| ## | Griffin | Harney | Hart | Hatchineha |
| ## | 1 | 1 | 1 | 1 |
| ## | Iamonia | Istokpoga | Jackson | Josephine |
| ## | 1 | 1 | 1 | 1 |
| ## | Kingsley | Kissimmee | Lochloosa | Louisa |
| ## | 1 | 1 | 1 | 1 |
| ## | Miccasukee | Minneola | Monroe | Newmans |
| ## | 1 | 1 | 1 | 1 |
| ## | Ocean Pond | Ocheese Pond | Okeechobee | Orange |
| ## | 1 | 1 | 1 | 1 |
| ## | Panasoffkee | Parker | Placid | Puzzle |
| ## | 1 | 1 | 1 | 1 |
| ## | Rodman | Rousseau | Sampson | Shipp |
| ## | 1 | 1 | 1 | 1 |
| ## | Talquin | Tarpon | Tohopekaliga | Trafford |
| ## | 1 | 1 | 1 | 1 |
| ## | Trout | Tsala Apopka | Weir | Wildcat |
| ## | 1 | 1 | 1 | 1 |
| ## | Yale | | | |
| ## | 1 | | | |

Como se puede observar en la tabla de distribución de frecuencia de la variable cualitativa, solamente hay un registro por cada valor, por lo tanto no es relevante la moda de esta variable.

Matriz de correlación

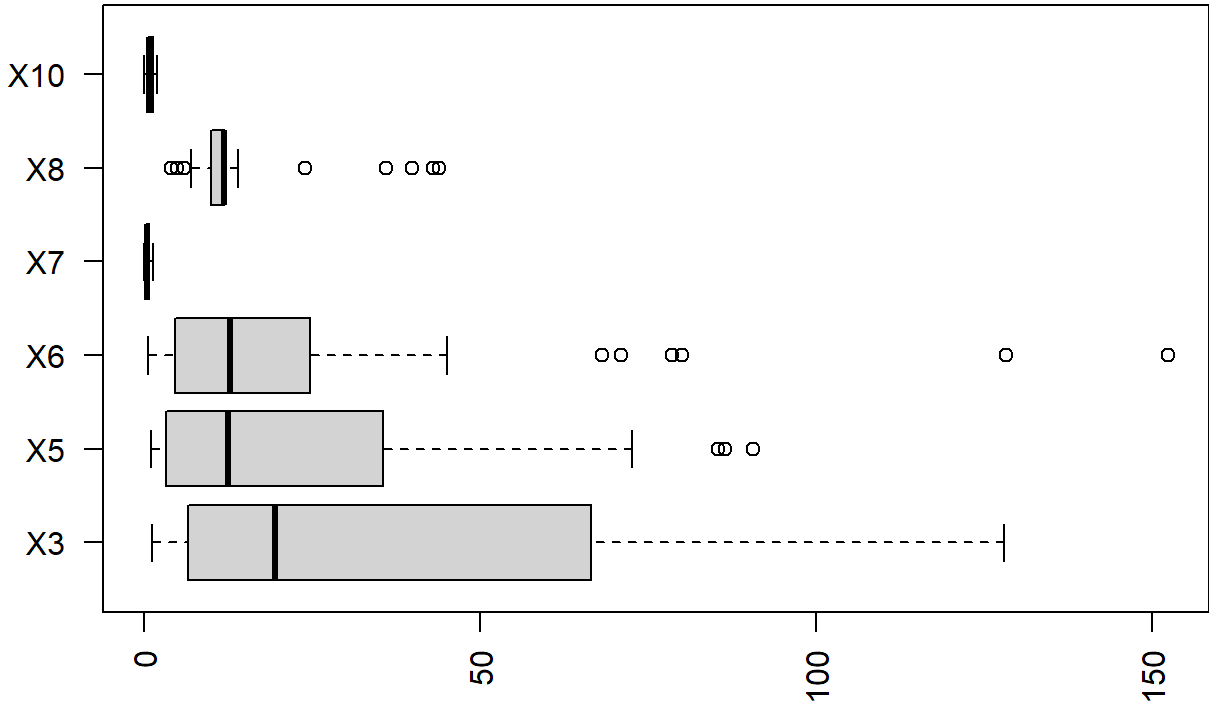


Posterior a analizar la matriz de correlación, se identifica dependencia en algunas variables independientes, por lo que se eliminarán.

Se observa dependencia entre X9, X10, X11 y se tomará la variable X10. Pese a que no es la variable con la mayor correlación de las 3 respecto a la variable dependiente (X7), pero es esencial para reponder a la pregunta “¿Hay evidencia para suponer que la concentración máxima de mercurio en los lagos es dañino para la salud humana?” y ya que su correlación del 92% se tomará esta.

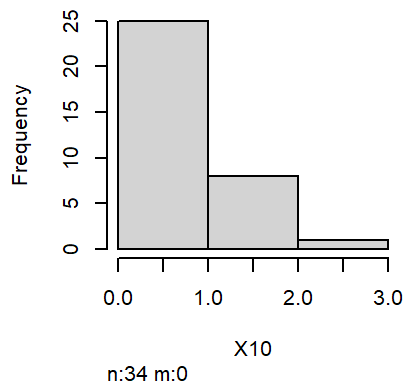
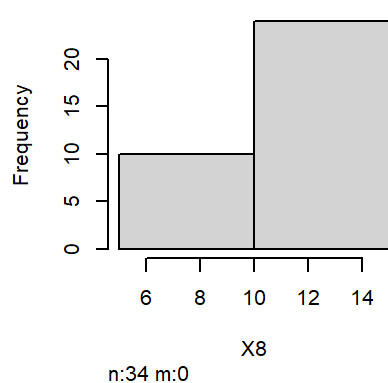
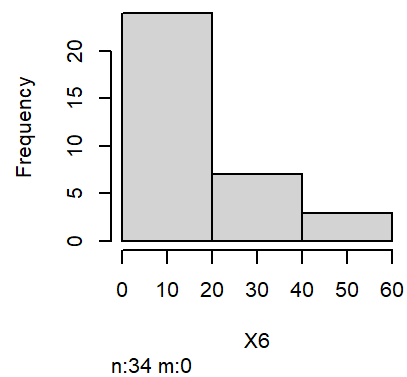
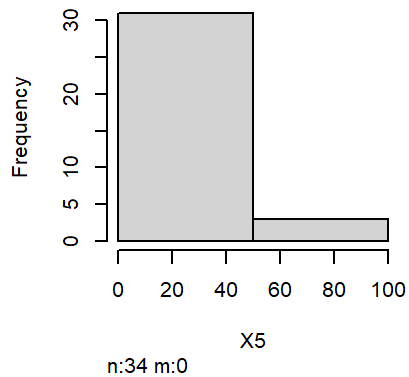
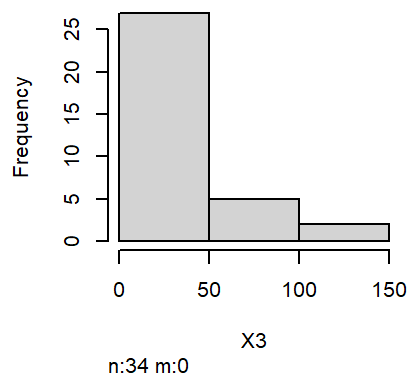
Tambien se observa dependencia entre la variable X3 y X4 por lo que se tomará X3 ya que es la que tiene mayor correlación con la variable dependiente(X7) Sumado a esto es importante notar con esta gráfica que la edad de los peces(X11) no tiene correlación con la concentración de mercurio en los peces(X7)

Boxplots de variables para exploración



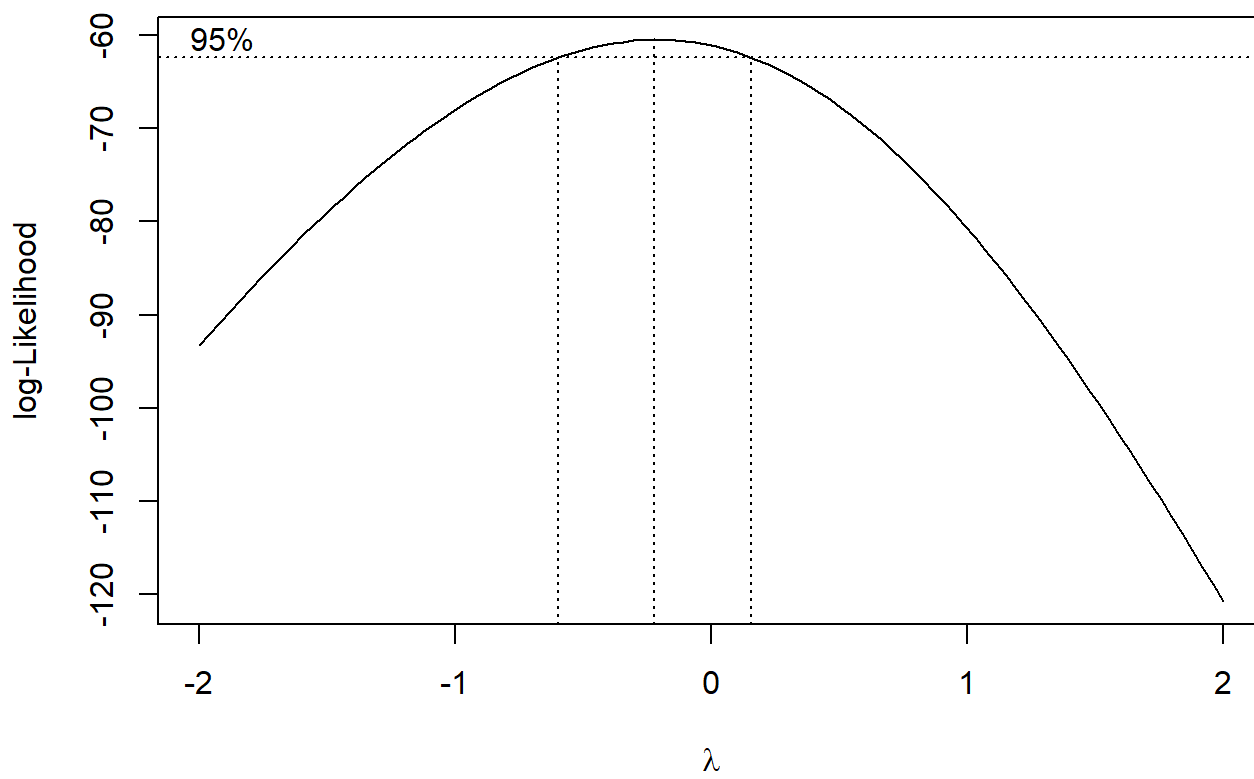
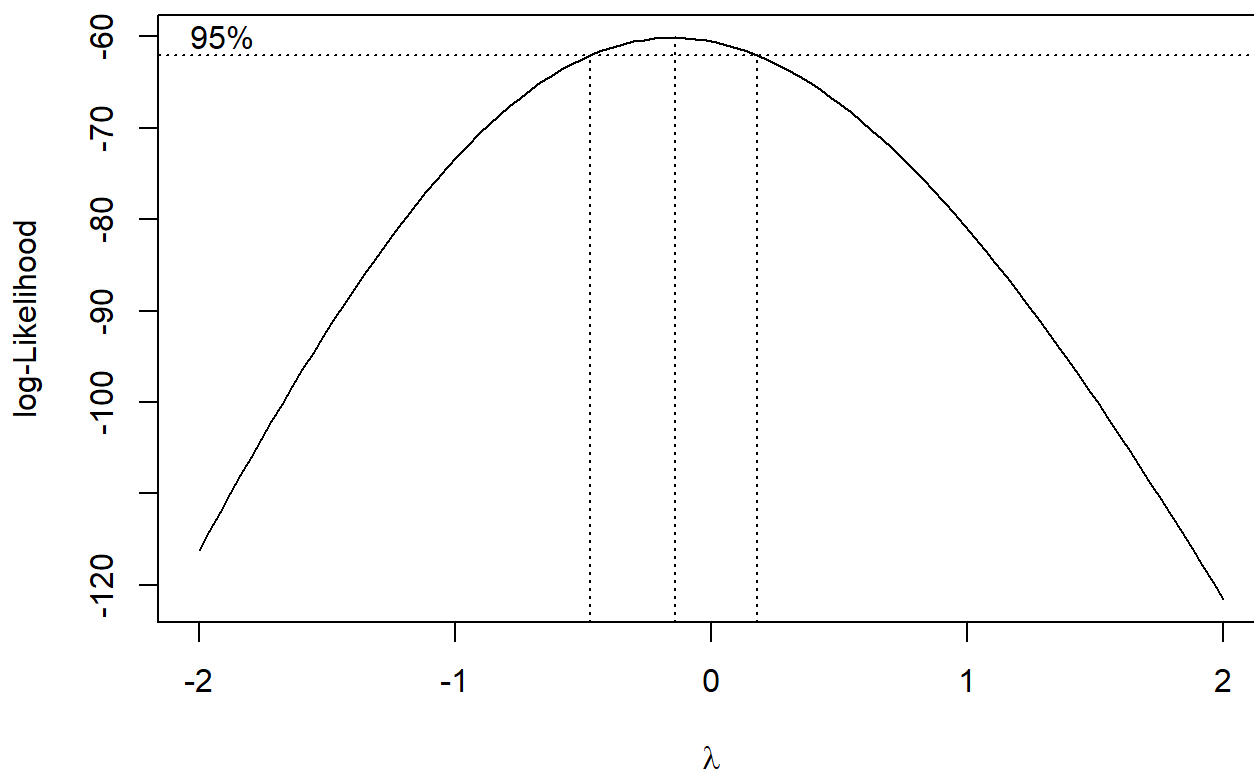
Ya que las variables cuentan con escalas muy diferentes será necesario hacer un rescalamiento de los datos más adelante para poder analizarlos, pero primero se eliminarán los outliers.

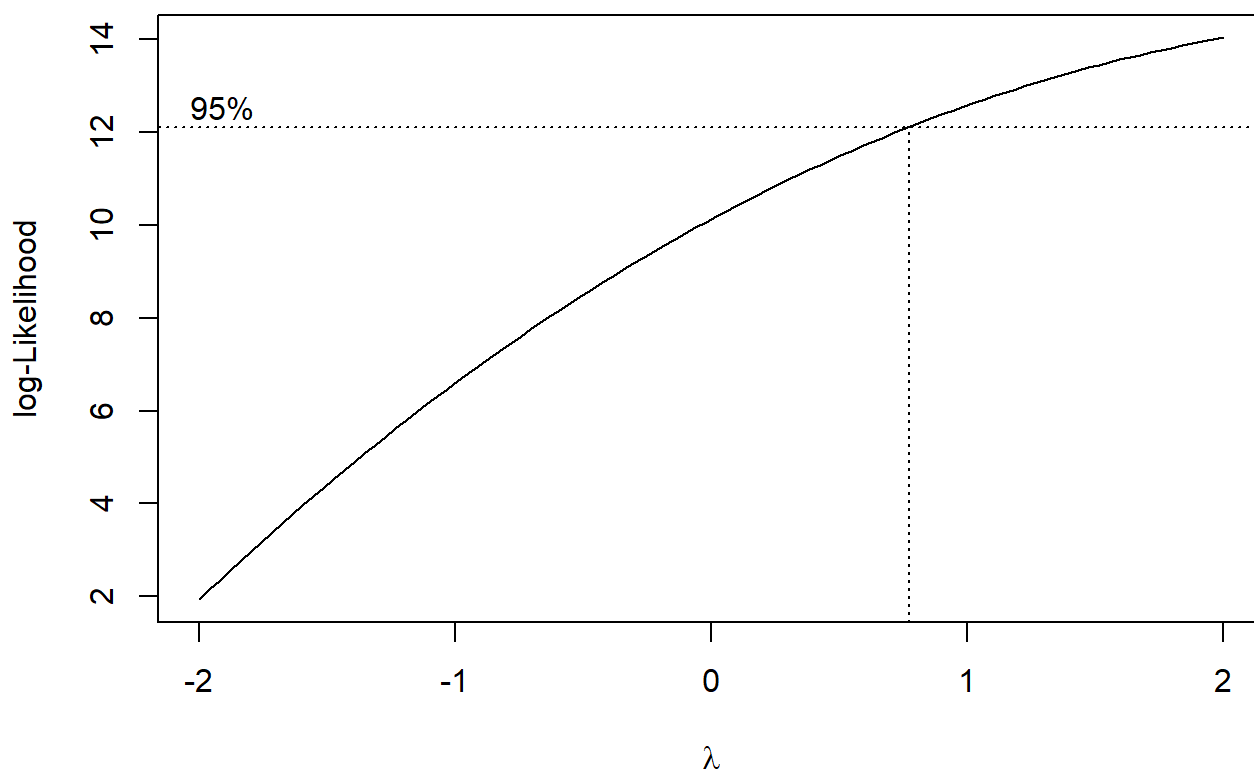
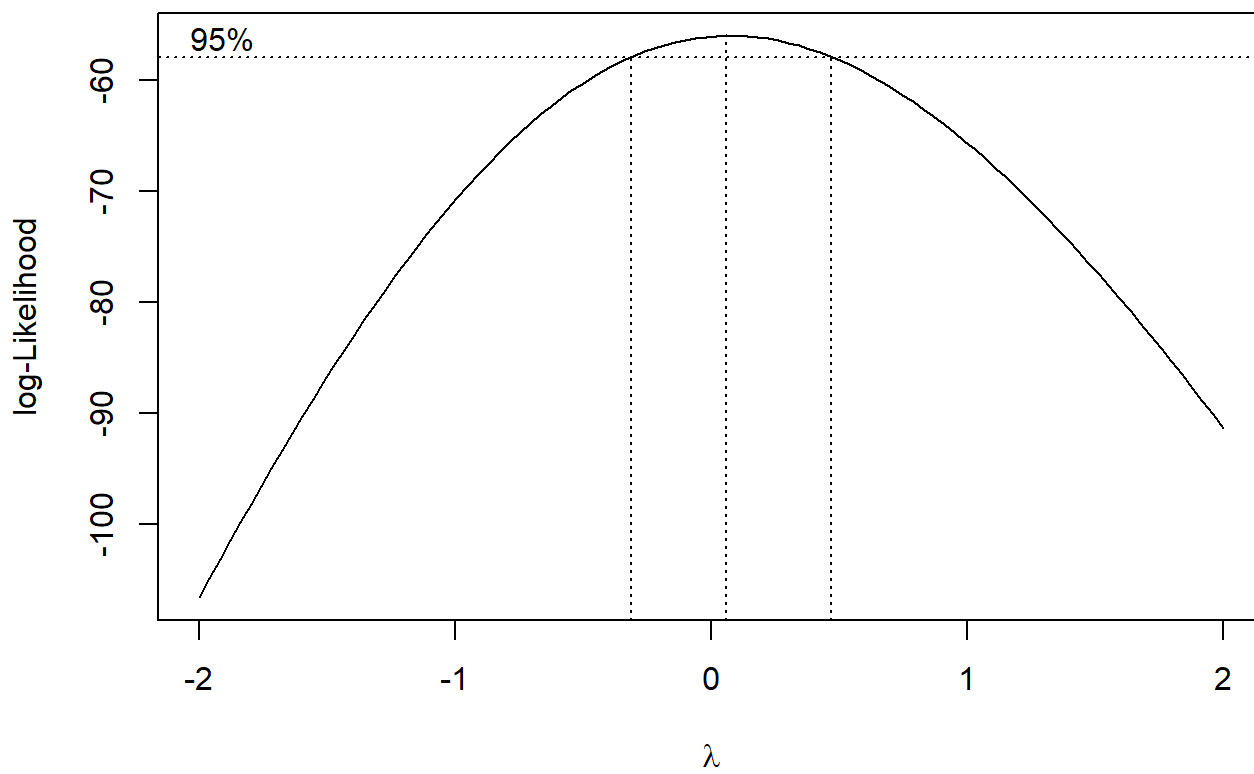
Despues de eliminar outliers de las variables con mayor numero de los mismos siguen apareciendo outliers, esto debido a que la nueva dimensión de los datos es mucho menor. No se realizará un segundo filtro de outliers, ya que actualmente n se volvió de tamaño 34, disminuir el valor de n a un numero menor a 30 comprometería nuestro análisis.

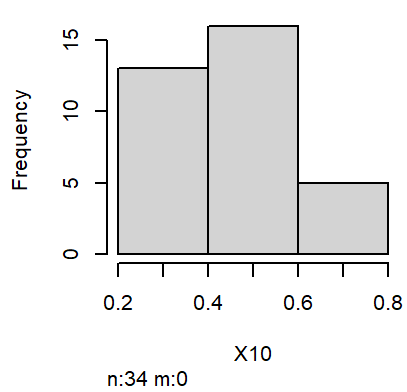
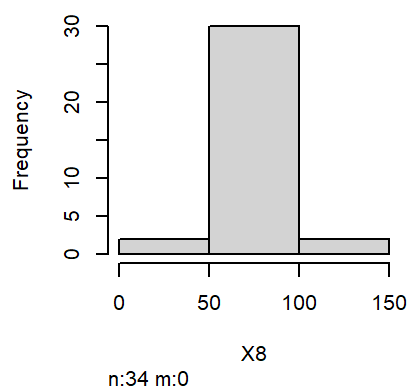
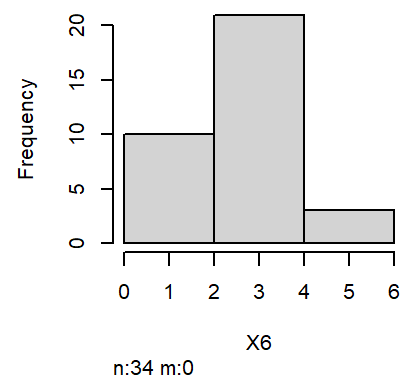
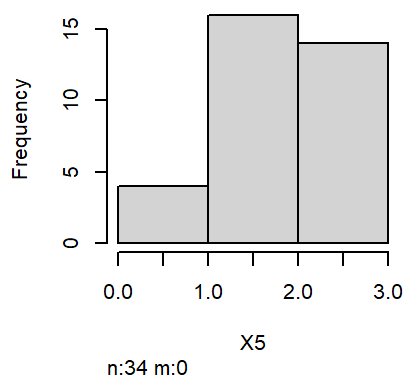
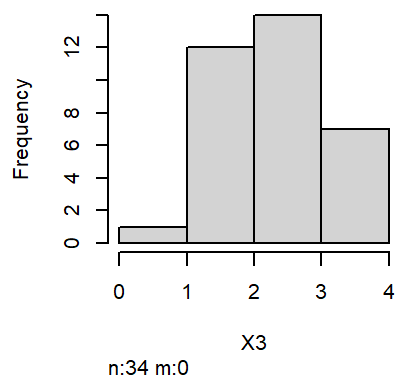
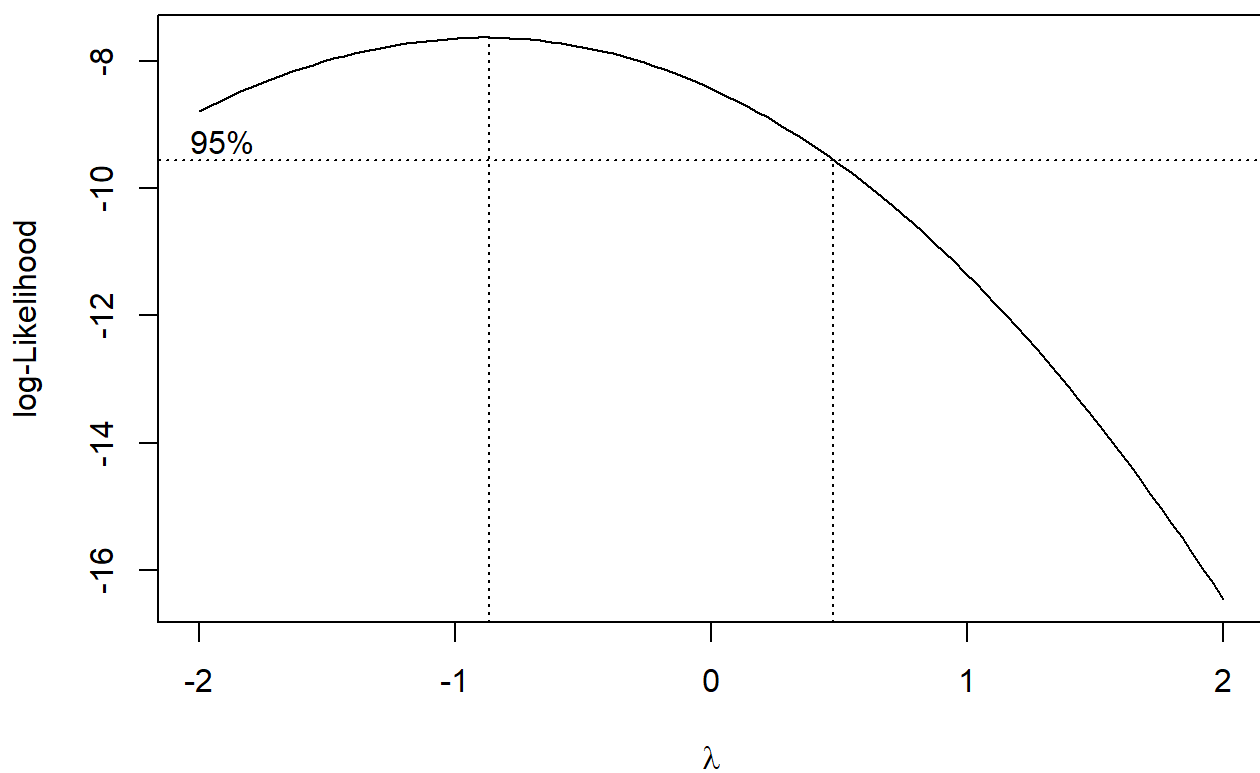


Las variables X3, X5, X6 y X10 tienen sesgo positivo a la derecha. La variable X8 tiene sesgo negativo a la izquierda.

Transformación por medio de Box-Cox





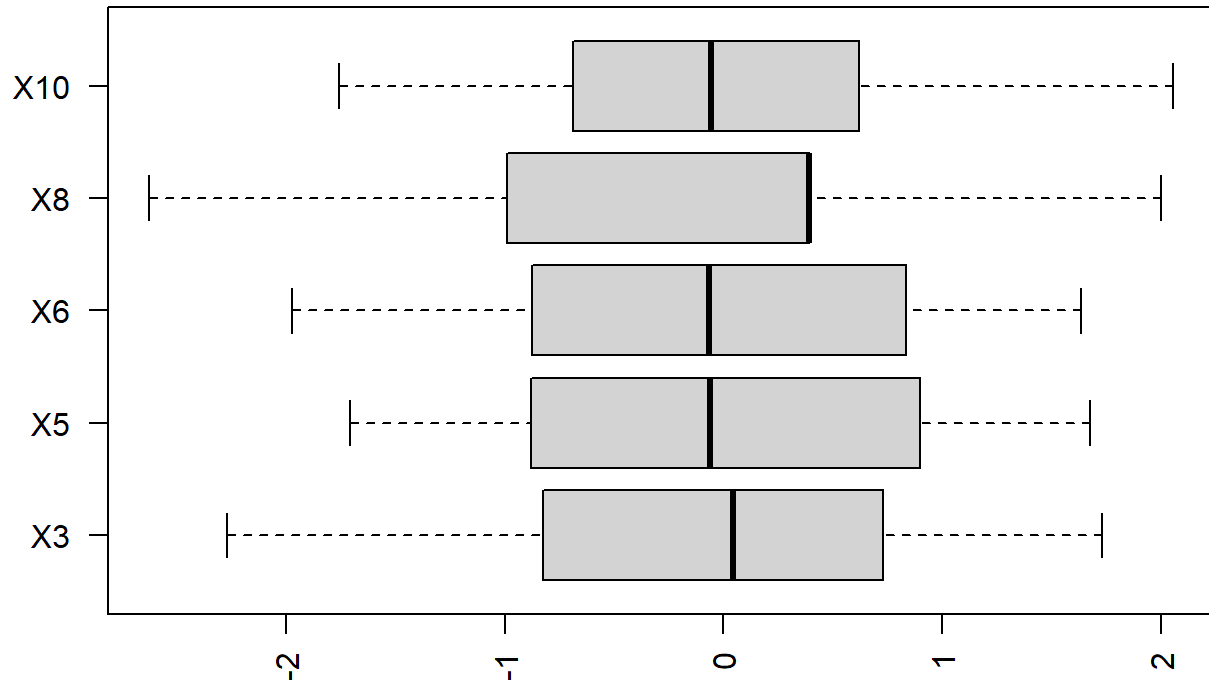


Como se puede observar se ha reducido el sesgo de los datos, obteniendo una forma aparentemente normalizada, posterior se validará eso.

Ahora se realizará un escalamiento de los datos para que sean más comparables por medio del tipo de escalamiento Estandar-Max.

Como siguiente paso se realiza un reescalamiento

Boxplots de variables reescaladas



Modelo

Ahora que los datos están listos para ser analizados, se crearán los modelos para evaluar cuales factores son los relevantes para el problema.

```
##
## Call:
## lm(formula = X7 ~ X3 + X5 + X6 + X8 + X10, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.152915 -0.053760 -0.002637  0.051244  0.159307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.467059   0.014285  32.696 < 2e-16 ***
## X3          -0.076544   0.033175  -2.307  0.0286 *
## X5           0.010715   0.025721   0.417  0.6802
## X6           0.018636   0.015288   1.219  0.2330
## X8           0.009988   0.015666   0.638  0.5289
## X10          0.163296   0.022003   7.422 4.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08329 on 28 degrees of freedom
## Multiple R-squared:  0.8848, Adjusted R-squared:  0.8642
## F-statistic: 43.01 on 5 and 28 DF,  p-value: 2.741e-12
```

En una primera instancia vemos como el p-value de X3, y X10 son los que se mantienen por debajo de 0.03 que será nuestro valor de alpha.

Ahora obtendremos cual es el mejor modelo en función de estas 5 variables.

Tenemos Hipotesis nula siendo:

$$H2_0: \mu X3 = \mu X5 = \mu X6 = \mu X8 = \mu X10$$

$H2_1$: Alguna beta es diferente

La cual se rechaza ya que las betas si son significativamente diferentes.

```
## Start: AIC=-163.61
## X7 ~ X3 + X5 + X6 + X8 + X10
##
##      Df Sum of Sq    RSS    AIC
## - X5   1   0.00120 0.19546 -165.40
## - X8   1   0.00282 0.19708 -165.12
## - X6   1   0.01031 0.20457 -163.85
## <none>                0.19426 -163.61
## - X3   1   0.03693 0.23119 -159.69
## - X10  1   0.38213 0.57639 -128.63
##
## Step: AIC=-165.4
## X7 ~ X3 + X6 + X8 + X10
##
##      Df Sum of Sq    RSS    AIC
## - X8   1   0.00222 0.19768 -167.01
## - X6   1   0.01057 0.20604 -165.61
## <none>                0.19546 -165.40
## + X5   1   0.00120 0.19426 -163.61
## - X3   1   0.06803 0.26350 -157.24
## - X10  1   0.43794 0.63340 -127.42
##
## Step: AIC=-167.01
## X7 ~ X3 + X6 + X10
##
##      Df Sum of Sq    RSS    AIC
## <none>                0.19768 -167.01
## - X6   1   0.01234 0.21003 -166.95
## + X8   1   0.00222 0.19546 -165.40
## + X5   1   0.00060 0.19708 -165.12
## - X3   1   0.06842 0.26610 -158.91
## - X10  1   0.44484 0.64252 -128.94
```

```
##
## Call:
## lm(formula = X7 ~ X3 + X6 + X10, data = df)
##
## Coefficients:
## (Intercept)          X3          X6          X10
##    0.46706    -0.06603     0.02015     0.16364
```

Este fue el mejor modelo con las variables disponibles. Cabe señalar que la variable X8 (cantidad de peces estudiados) fue descartada del modelo más efectivo, incluso en el primer modelo donde se tomó en cuenta su p-value fue de 0.52 el cual es significativamente más grande que $\alpha(0.03)$.

H_{2_0} : La cantidad de peces encontrados no tiene influencia en la concentración de mercurio en los peces.

H_{2_1} : La cantidad de peces encontrados tiene influencia en la concentración de mercurio en los peces.

No se rechaza H_{2_0} por lo que se concluye que el numero de peces encontrados no tiene correlación con la concentración de mercurio.

```
##
## Call:
## lm(formula = X7 ~ X3 + X6 + X10, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.146240 -0.064432 -0.006656  0.045867  0.165723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.46706     0.01392  33.550 < 2e-16 ***
## X3          -0.06603     0.02049  -3.222  0.00306 **
## X6           0.02015     0.01472   1.369  0.18126
## X10          0.16364     0.01992   8.216 3.59e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08118 on 30 degrees of freedom
## Multiple R-squared:  0.8828, Adjusted R-squared:  0.871
## F-statistic: 75.29 on 3 and 30 DF,  p-value: 4.593e-14
```

A pesar de que la variable X6 no tiene correlación significativamente con la variable dependiente, es necesaria para mejorar la precisión del modelo de acuerdo con lo obtenido.

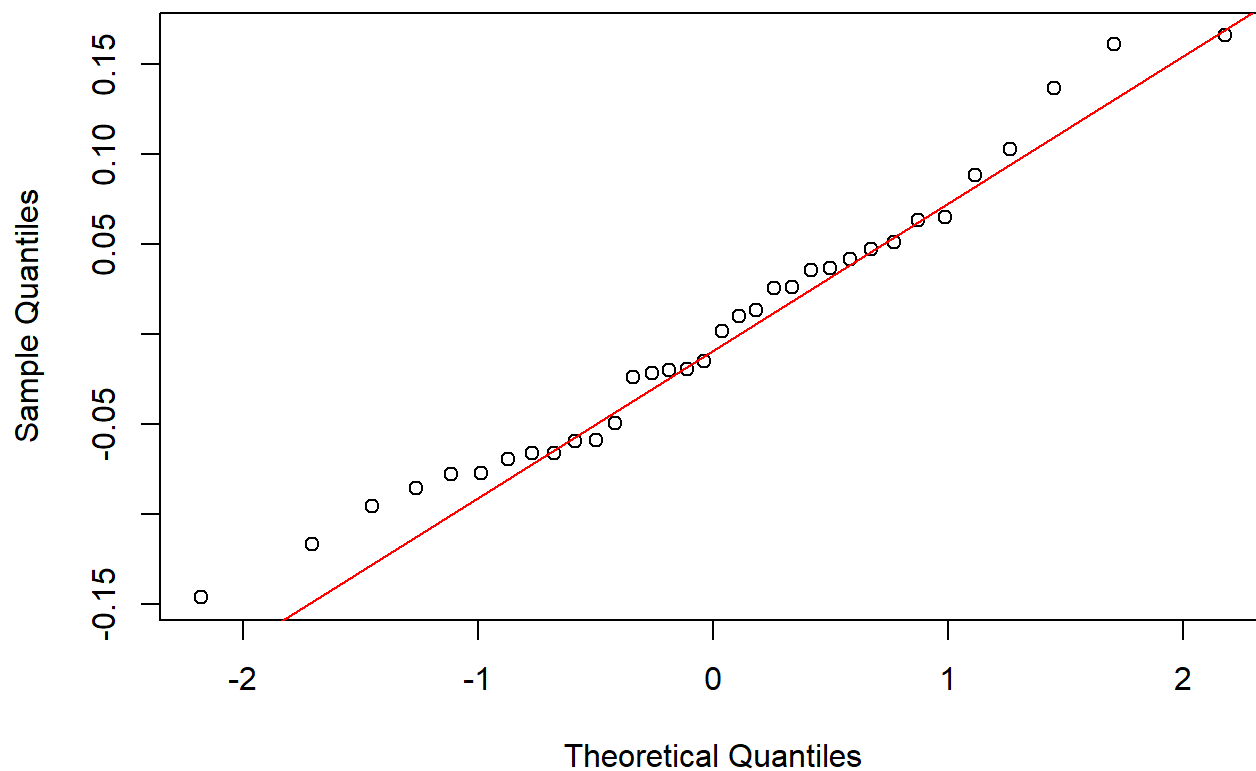
Así mismo se da respuesta a la pregunta: Las concentraciones de alcalinidad(X3), clorofila(X6), calcio(X5) en el agua del lago influyen en la concentración de mercurio de los peces? ya que el modelo nos valida que alcalinidad(X3) influye en la concentración de mercurio en los peces, sin embargo la clorofila(X6) y el calcio(X5) no influyen significativamente en la concentración de mercurio en los peces.

```
##              2.5 %      97.5 %
## (Intercept)  0.438627426  0.49549022
## X3          -0.107886837 -0.02418186
## X6          -0.009916901  0.05021922
## X10          0.122961597  0.20430857
```

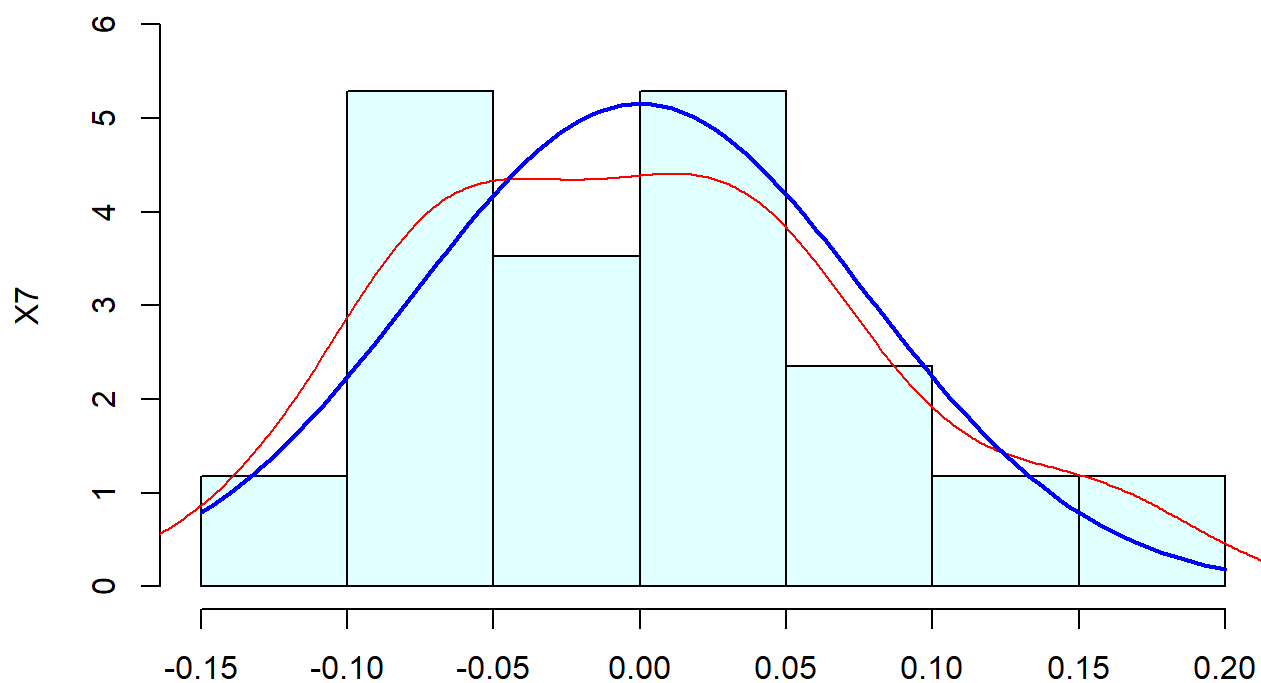
Verificación de supuestos

Normalidad

Normal Q-Q Plot



Histograma de Resíduos

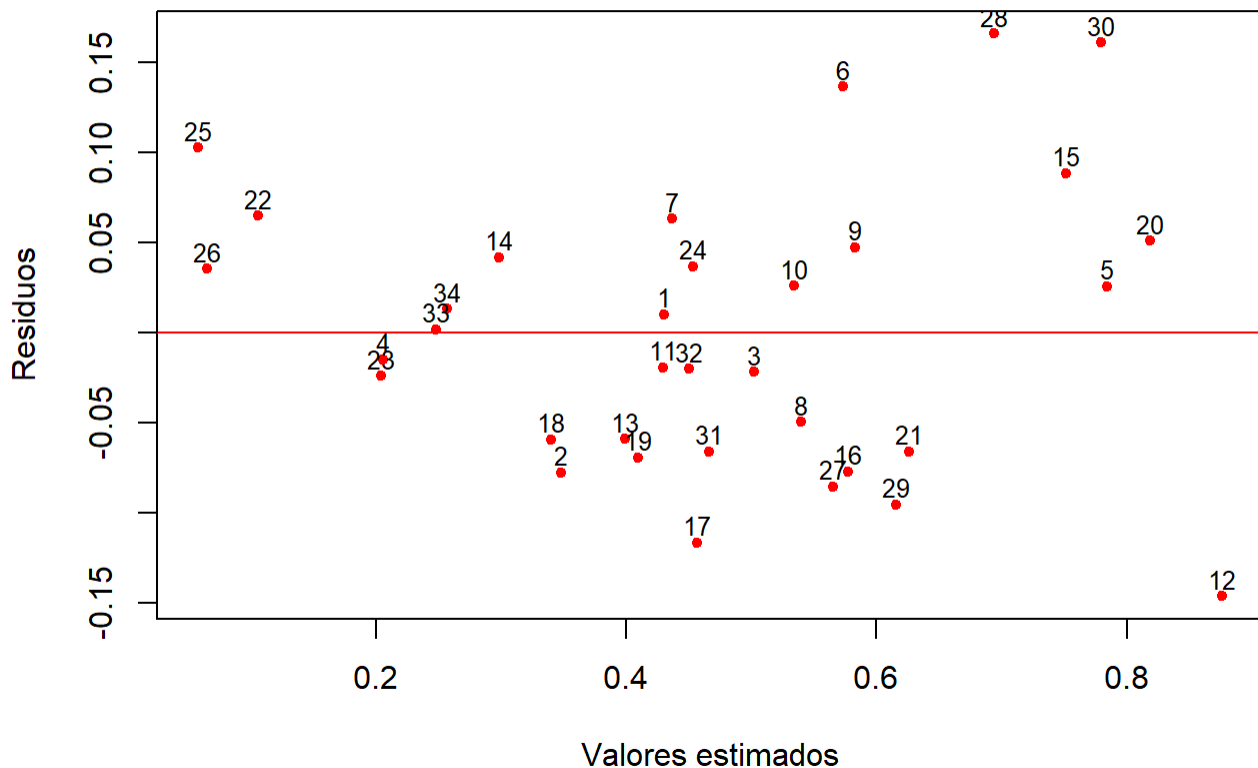


```
##
## Shapiro-Wilk normality test
##
## data: E
## W = 0.97365, p-value = 0.569
```

Se puede concluir que los datos vienen de una fuente normal ya que los residuos cumplen con normalidad. En la prueba de normalidad se puede apreciar que los datos están distribuidos uniformemente alrededor de la pendiente, solamente en la cola inferior están un poco dispersos los datos, sin embargo no es significativo. Seguido se observa que la distribución de los datos tiene forma de campana y por ultimo se confirma mediante la prueba de Shapiro-Wilk ya que el valor de p es superior a 0.3 que es el valor de alpha establecido.

Homocedasticidad y modelo apropiado

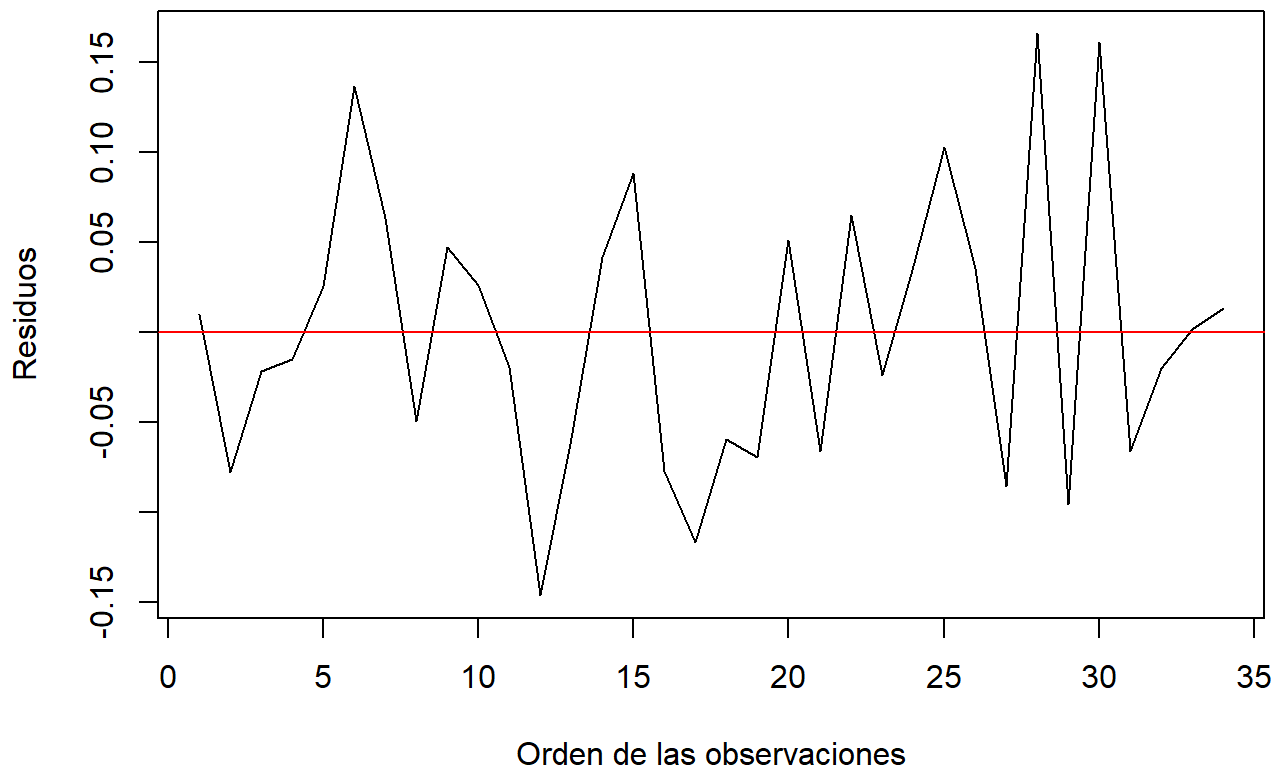
Gráfica Valores estimados vs Residuos



Se puede apreciar gráficamente homocedasticidad ya que los valores se encuentran dispersos uniformemente sin ningún patrón significativo aparente.

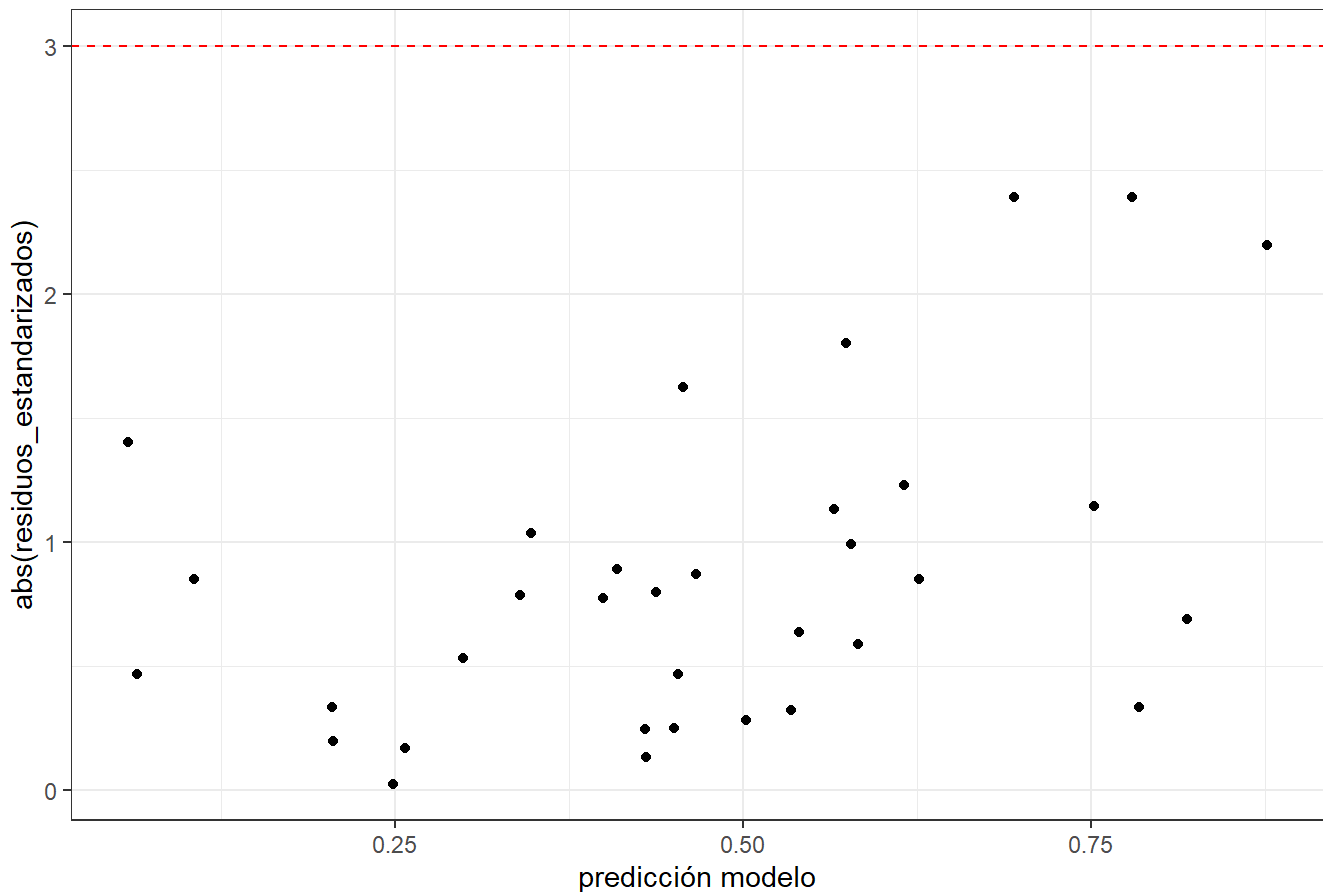
Independencia

Errores vs Orden de observación



De igual manera se puede confirmar que no hay dependencia.

Distribución de los residuos estandarizados



```
## integer(0)
```

En la gráfica anterior se observa que no hay datos atípicos en la predicción.

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

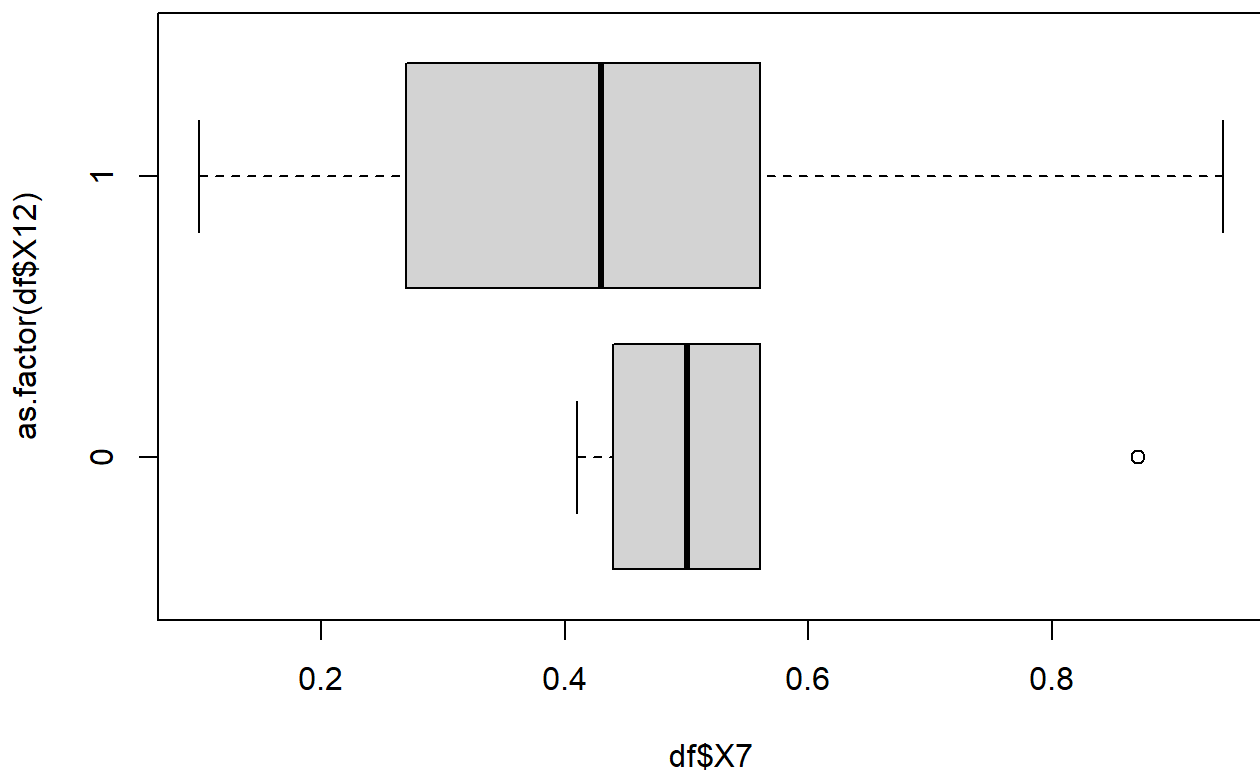
De acuerdo con el análisis realizado se puede concluir tomando como referencia la tabla de correlación que no hay evidencia suficiente para asumir que la concentración de mercurio en los peces(X7) varía conforme a la edad(X12). Para validar esto se muestra la siguiente gráfica.

H_{10} : μ concentración de mercurio(X7) de los peces viejos = μ concentración de mercurio(X7) de los peces jóvenes
 H_{11} : μ concentración de mercurio(X7) de los peces viejos \neq μ concentración de mercurio(X7) de los peces jóvenes

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(X12)  1  0.0464  0.04637    0.905   0.349
## Residuals      32  1.6397  0.05124
```

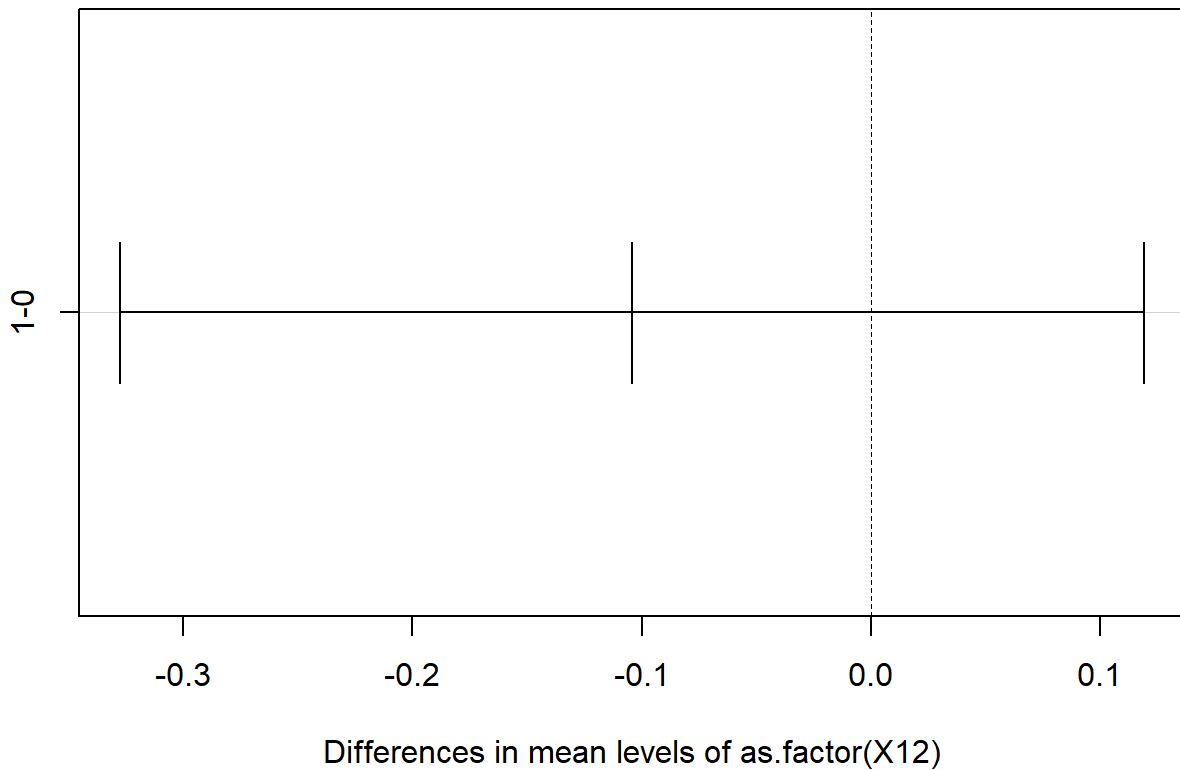
```
##           0           1
## 0.5560000 0.4517241
```

```
## [1] 0.4670588
```



```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = X7 ~ as.factor(X12), data = df)
##
## $`as.factor(X12)`
##      diff      lwr      upr    p adj
## 1-0 -0.1042759 -0.3275528 0.1190011 0.3485835
```

95% family-wise confidence level



Como se puede observar no hay tendencia aparente además el f-value del modelo es significativamente mayor a $\alpha(0.03)$. Con lo cual se concluye que no se rechaza la hipótesis nula, los peces no presentan diferencia en la concentración de mercurio. además de validarlo con el anova, se puede observar en la gráfica de intervalos de confianza de los peces jóvenes vs peces viejos que sus valores de sus medias se superponen sobre el intervalo de confianza del otro.

Conclusión del análisis

En conclusión el modelo obtenido es: $X7 = X3 - 0.06603 + X6 - 0.02015 + X10 * -0.16364 + 0.46706$

La variabilidad explicada por el modelo (coeficiente de determinación) es: 0.8828

Significancia del modelo: Valor p del modelo es $4.593e-14$

Esto nos deja ver que el valor p del modelo está muy por debajo de α siendo 0.03, por lo que es preciso. El modelo satisface todos los supuestos, ya que sus residuos son normales, no hay sesgo aparente y tiene homocedasticidad.

Un factor importante para el desarrollo del modelo fue la preparación correcta de los datos, ya que se realizó limpieza de datos atípicos, transformación de Box-Cox para normalizar los datos y escalarlos por método desviación-máx para que los datos fueran comparables.

Referencias

Li P, Feng X, Qiu G (2010) Methylmercury exposure and health effects from rice and fish consumption: a review. Int J Environ Res Public Health 7:2666–2691. <https://doi.org/10.3390/ijerph7062666> (<https://doi.org>)

/10.3390/ijerph7062666)

USEPA (2013) Mercury: health effects. Retrieved from: <http://www.epa.gov/hg/effects.htm> (<http://www.epa.gov/hg/effects.htm>).

Liga a código y documentos

https://github.com/a01634928/TC3006C_101_A01634928/tree/main/modulo_1/Los_peces_y_el_mercurio

(https://github.com/a01634928/TC3006C_101_A01634928/tree/main/modulo_1/Los_peces_y_el_mercurio)