Modelos Semiparamétricos

Magister en Matemáticas con Mención en Estadística

Bladimir Morales Torrez

Julio 2022

Contents

1 Datos e Información **3**

1.1 Análisis exploratorio 3

2 Modelos **5**

2.1 Modelo Lineal LM 5

2.2 Modelo no Paramétrico GAM 7

2.3 Modelo Semiparamétrico o Parcial PLM 9

3 Selección del modelo **11**

3.1 Primer criterio 11

3.2 Criterio de Información de Akaike 12

3.3 Prueba F 12

3.4 Conclusión 13

4 Análisis de Diagnóstico **13**

4.1 Puntos de apalancamiento 13

4.2 Puntos influyentes 14

5 Análisis de residuos **15**

5.1 Supuesto de normalidad 15

5.2 Supuesto de homocedasticidad 18

5.3 Supuesto de autocorrelación 19

6 Conclusión **19**

1 Datos e Información

El conjunto de datos es `Caschool` del paquete `Ecdat`. Este conjunto de datos contiene 420 observaciones transversales recogidas durante el año escolar 1998 − 1999 en los distritos escolares de California.

Las variables a ser utilizadas son: `mathscr`: puntuaciones medias de matemáticas (variable de respuesta) las siguientes cuatro serán las covariables: `calwpct`: porcentaje de niños que cumplen los requisitos para recibir CalWORKs, `log.avginc`: logaritmo natural de la renta media del distrito, `compstu`: número de computadoras por alumno, `expnstu`: gasto por alumno

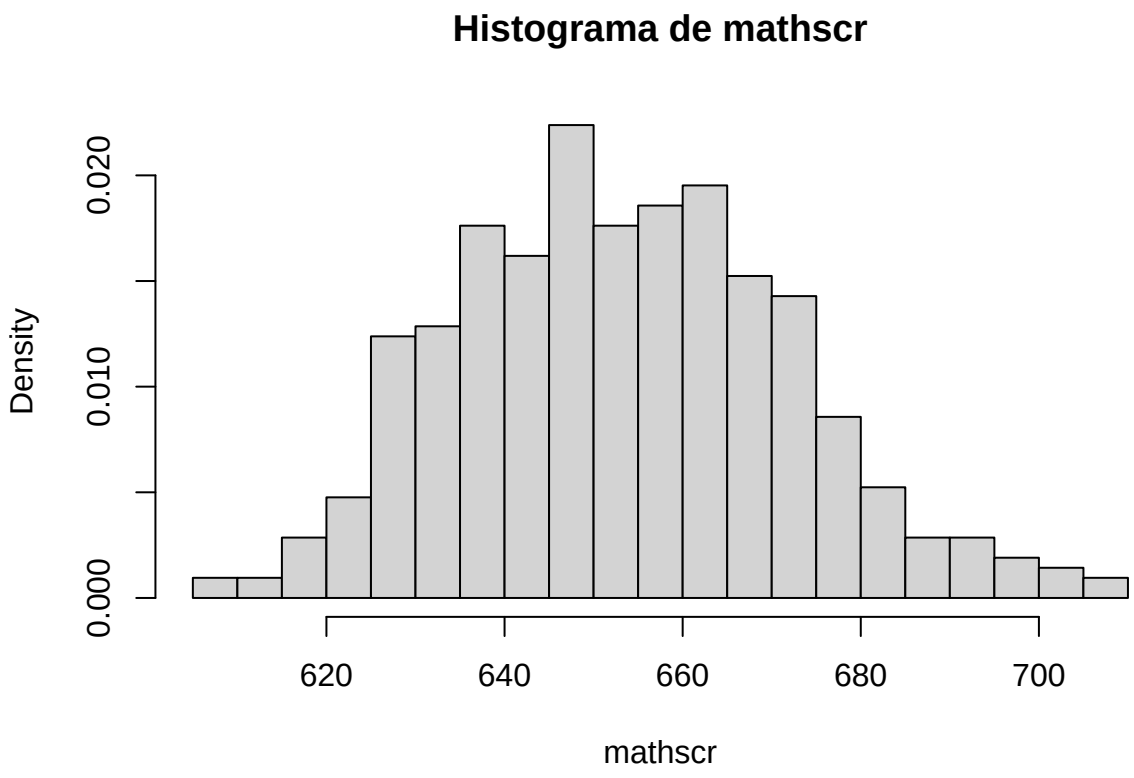
CalWORKs es un *programa de asistencia social que ofrece ayuda en efectivo y servicios a las familias necesitadas de California que reúnen los requisitos.*

Primero se realizará un análisis exploratorio de las variables de estudio, luego se estimará un modelo de regresión lineal, un modelo no paramétrico y un modelo parcial y evaluar sus diferentes características para encontrar el modelo que mejor ajuste, para luego realizar el diagnóstico del modelo, la predicción de datos y detallar las conclusiones.

1.1 Análisis exploratorio

Primero se visualizará el comportamiento de la variable de respuesta `mathscr` (puntuaciones medias de matemáticas):

```
hist(mathscr,main = "Histograma de mathscr",freq = FALSE,breaks = 20)
```



Sus estadísticos son los siguientes:

```
bd %>%
  summarise(obs=n(),
            media=mean(mathscr),
            sd=sd(mathscr),
            asimetria=skewness(mathscr),
            curtosis=kurtosis(mathscr),
            min=min(mathscr),
            max=max(mathscr),
            mediana=median(mathscr))
```

```
##   obs    media      sd asimetria curtosis   min   max mediana
## 1 420 653.3426 18.7542 0.2550819 2.840189 605.4 709.5  652.45
```

Se presume tener una distribución simétrica con una leve tendencia hacia la positividad (asimetría=0.25) y platicúrtica (Curtosis<3).

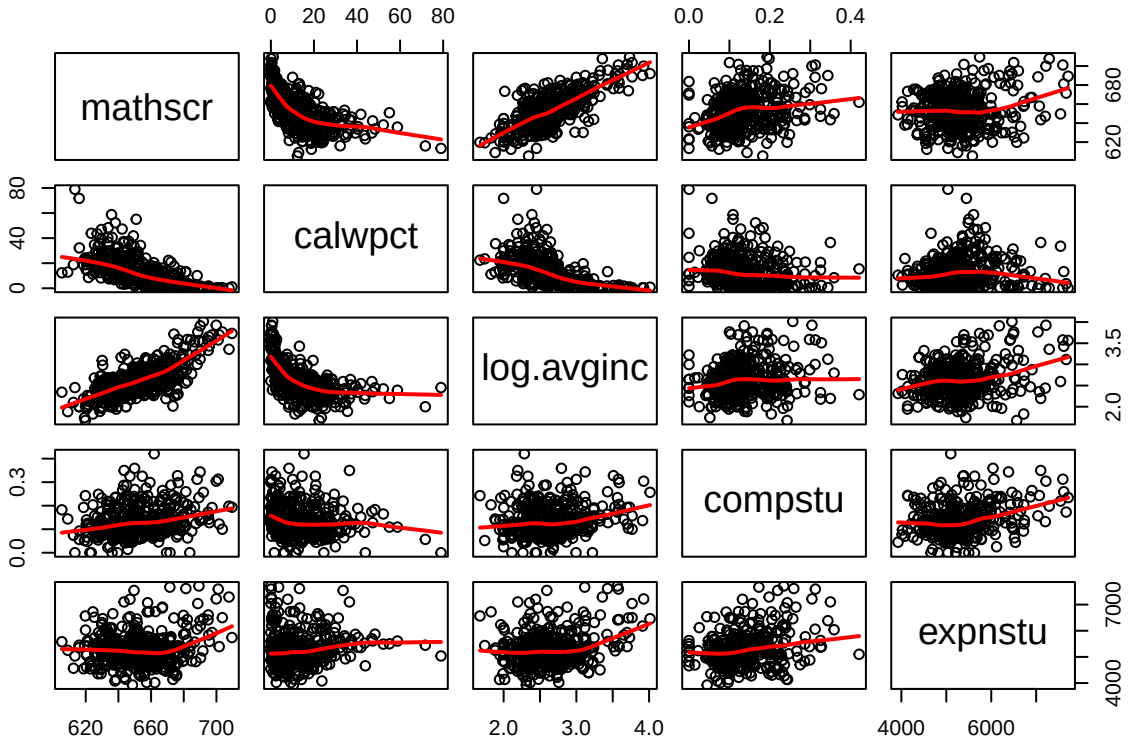
Ahora vemos las correlaciones de las variables en estudio .

```
cor(bd)
```

```
##               mathscr     calwpct log.avginc    compstu    expnstu
## mathscr     1.0000000 -0.61769076  0.7252397  0.2485891 0.15498949
## calwpct    -0.6176908  1.00000000 -0.5687013 -0.1519675 0.06788857
## log.avginc  0.7252397 -0.56870132  1.0000000  0.1593155 0.25113384
## compstu     0.2485891 -0.15196750  0.1593155  1.0000000 0.28655958
## expnstu     0.1549895  0.06788857  0.2511338  0.2865596 1.00000000
```

Las correlaciones más altas con la variable `mathscr` son las de `calwpct` y `log.avginc`. En el siguiente gráfico se observa visualmente las relaciones entre variables con una curva suavizada.

```
pairs(bd,panel = function(x,y){
  points(x,y)
  lines(lowess(x,y),lwd=2,col="red")
})
```



La variable `calwpct`, `compstu` y `expnstu` al parecer tienen una relación no lineal con la variable de respuesta, mientras que `log.avginc` podría tener una relación lineal positiva.

2 Modelos

Se planteara tres diferentes modelos para esta sección:

- Modelo lineal clásico LM
- Modelo no paramétrico GAM
- Modelo semiparamétrico o parcial PLM

2.1 Modelo Lineal LM

Se utilizará el paquete `mgcv` para estimar el modelo lineal asumiendo que la distribución de los errores es Normal (Gaussiana) en nuestras variables de estudio. Es así que se tiene:

```
modlm <- mgcv::gam(mathscr ~ calwpct +log.avginc+compstu+ expnstu,family = gaussian)
summary(modlm)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## mathscr ~ calwpct + log.avginc + compstu + expnstu
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.862e+02  6.189e+00  94.719  < 2e-16 ***
## calwpct     -4.829e-01  6.471e-02  -7.462 5.06e-13 ***
## log.avginc   2.568e+01  1.924e+00  13.346  < 2e-16 ***
## compstu      3.356e+01  9.491e+00   3.536 0.000452 ***
## expnstu      2.003e-04  1.025e-03   0.195 0.845233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.598   Deviance explained = 60.2%
## GCV = 143.11  Scale est. = 141.41    n = 420
```

Al ver la prueba t las covariables `calwpct`, `log.avginc` y `compstu` son significativas para el modelo, vale decir que si aportan a explicar el modelo propuesto, mientras que `expnstu` no es significativo para el modelo. Se tiene un $R^2 - ajust$ de 0.598, un ajuste relativamente bueno para los datos presentados.

Ahora se eliminará la covariable `expnstu` del modelo, teniendo:

```
modlm <- mgcv::gam(mathscr ~ calwpct +log.avginc+compstu,family = gaussian)
summary(modlm)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## mathscr ~ calwpct + log.avginc + compstu
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 586.79111    5.36451 109.384  < 2e-16 ***
## calwpct      -0.47913    0.06176  -7.758 6.75e-14 ***
## log.avginc   25.81018    1.80520  14.298  < 2e-16 ***
## compstu      34.09414    9.07484   3.757 0.000197 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.599   Deviance explained = 60.2%
## GCV = 142.44  Scale est. = 141.08    n = 420
```

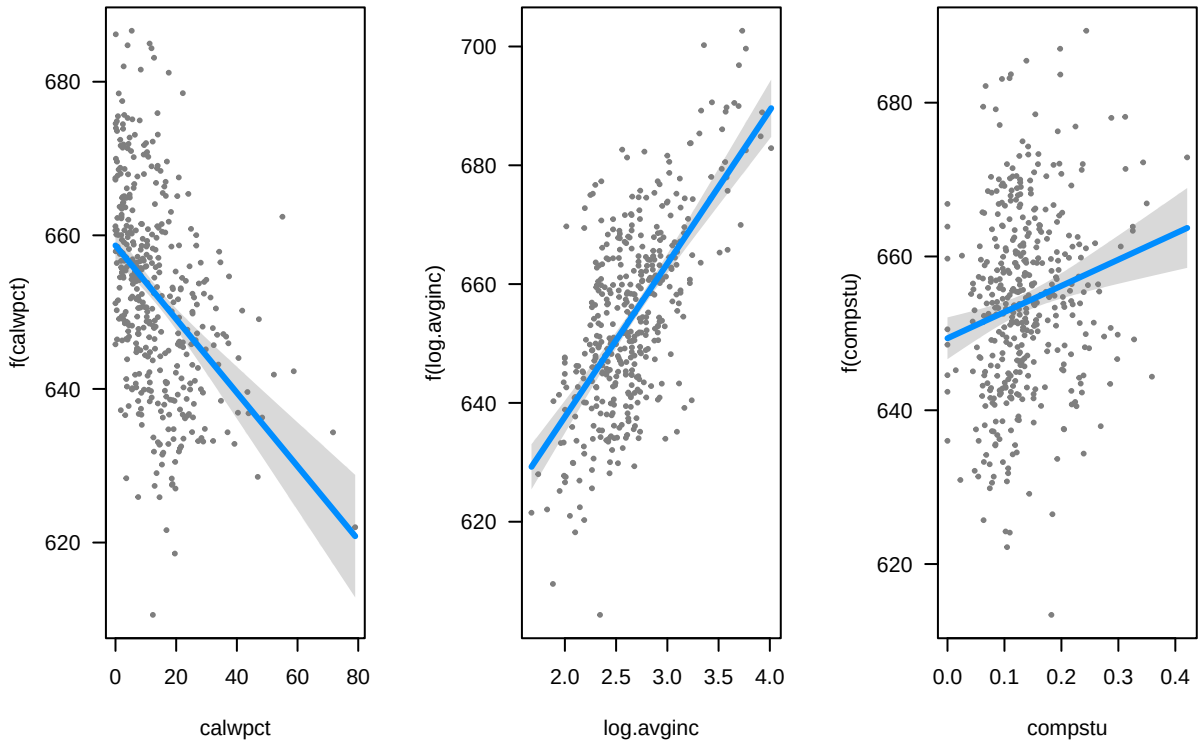
Ahora todas las covariables son significativas incluyendo el intercepto. Con el nuevo modelo se tiene un $R^2 - ajust$ de 0.599, un ajuste levemente mayor al anterior modelo.

En el siguiente gráfico se puede ver los valores ajustados estimados bajo el modelo propuesto.

```
library(visreg)
```

```
## Warning: package 'visreg' was built under R version 4.1.3
```

```
par(mfrow=c(1,3))
visreg(modlm)
```




```
dev.off()
```

```
## null device
##           1
```

2.2 Modelo no Paramétrico GAM

Ahora bajo el análisis exploratorio se pudo observar que existe una relación no lineal entre las covariables y la variable de respuesta excepto la `log.avginc`, pero por motivos prácticos asumiremos que todas las covariables no tienen relación lineal, es así que se estimará un modelo no paramétrico con errores gaussianos, con suavizamiento spline cúbico natural y el grado del polinomio se elegirá en base al método de validación cruzada generalizada.

```
modgam <- mgcv::gam(mathscr ~ s(calwpct)+s(log.avginc)+s(compstu)+ s(expnstu),family = gaussian, method
summary(modgam)
```

```
##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## mathscr ~ s(calwpct) + s(log.avginc) + s(compstu) + s(expnstu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 653.3426     0.5495    1189   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F p-value
## s(calwpct)    3.826  4.751 19.976 < 2e-16 ***
## s(log.avginc) 3.942  4.938 19.829 < 2e-16 ***
## s(compstu)    3.480  4.396  3.290 0.00971 **
## s(expnstu)    4.311  5.336  0.847 0.48557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.639   Deviance explained = 65.3%
## GCV = 132.02  Scale est. = 126.82    n = 420
```

De la misma manera que el modelo lineal, la covariable `expnstu` bajo la función no paramétrica no es significativa para el modelo, por tal motivo se excluira del modelo no paramétrico.

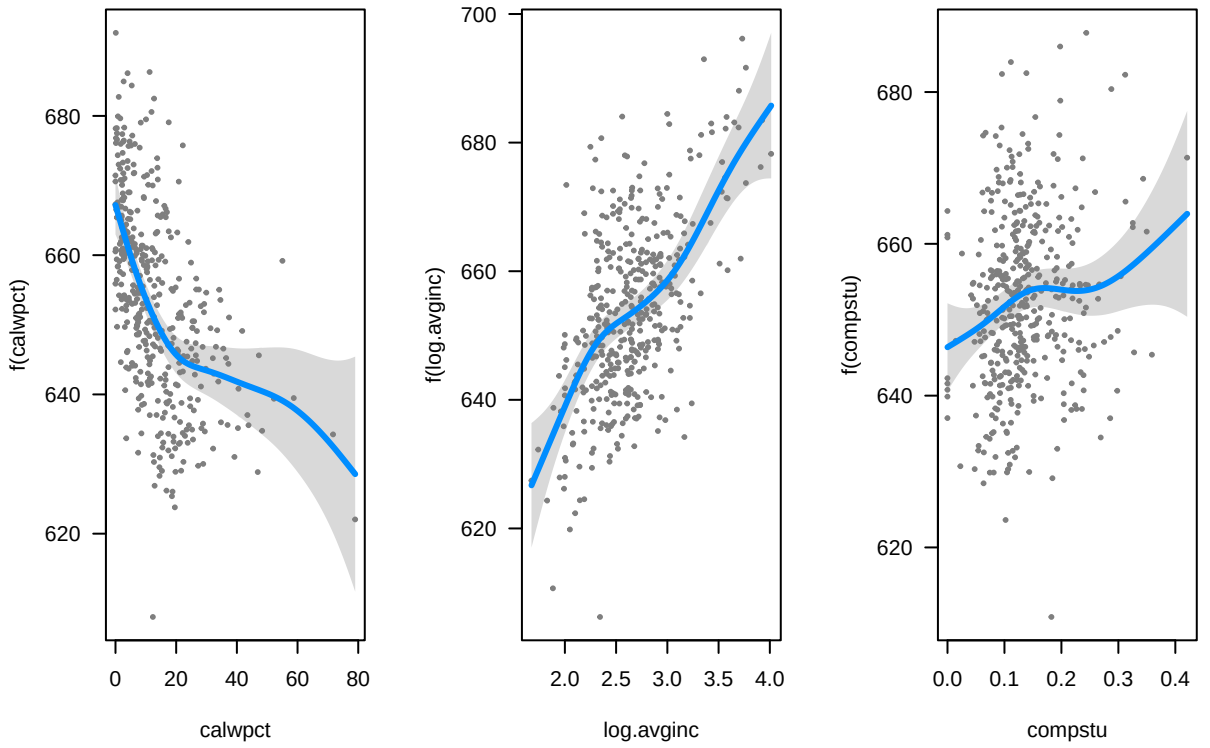
```
modgam <- mgcv::gam(mathscr ~ s(calwpct)+s(log.avginc)+s(compstu),family = gaussian, method = "GCV.Cp")
summary(modgam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## mathscr ~ s(calwpct) + s(log.avginc) + s(compstu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 653.3426     0.5516    1184   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F p-value
## s(calwpct)    3.778  4.698 21.884 < 2e-16 ***
## s(log.avginc) 4.077  5.100 23.132 < 2e-16 ***
## s(compstu)    3.322  4.206  3.545 0.00687 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.637   Deviance explained = 64.6%
## GCV = 131.61  Scale est. = 127.79    n = 420
```

Ahora todas las covariables son significativas bajo funciones suaves no paramétricas y así también el intercepto. Con el nuevo modelo se tiene un $R^2 - ajust$ de 0.637, un ajuste bueno para los datos y una *Deviance* de 64.6%.

En el siguiente gráfico se puede ver los valores ajustados estimados bajo el modelo propuesto.

```
par(mfrow=c(1,3))
visreg(modgam)
```




```
dev.off()
```

```
## null device
##           1
```

Se puede observar que la covariable `log.avginc` bajo el modelo propuesto no presentaría una relación lineal.

2.3 Modelo Semiparamétrico o Parcial PLM

Ahora se estimará un modelo parcial, ya que bajo el análisis exploratorio la covariable `log.avginc` presentaba aparentemente una relación lineal y las demás no lineales. Para el componente no paramétrico se asumira errores con distribución gaussiana, suavizamiento spline cúbico natural y el grado del polinomio se ajustará mediante validación cruzada generalizada. Bajo esa descripción se ajustara el siguiente modelo:

```
modplm <- mgcv::gam(mathscr ~ log.avginc+s(calwpct)+s(compstu)+ s(expnstu),family = gaussian,method = "
summary(modplm)
```

```
##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## mathscr ~ log.avginc + s(calwpct) + s(compstu) + s(expnstu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  601.093      5.749 104.554   <2e-16 ***
## log.avginc    19.755      2.164   9.131   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F p-value
## s(calwpct) 4.169  5.153 16.605  <2e-16 ***
## s(compstu) 3.549  4.483  3.394  0.0075 **
## s(expnstu) 4.344  5.377  0.914  0.4455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.631   Deviance explained = 64.2%
## GCV = 134.36  Scale est. = 129.87    n = 420
```

El componente paramétrico es significativo para el modelo y también el intercepto. En el componente no paramétrico al igual que los anteriores modelos la covariable `expnstu` es no significativa, es así que se excluira la misma para proponer el siguiente modelo parcial:

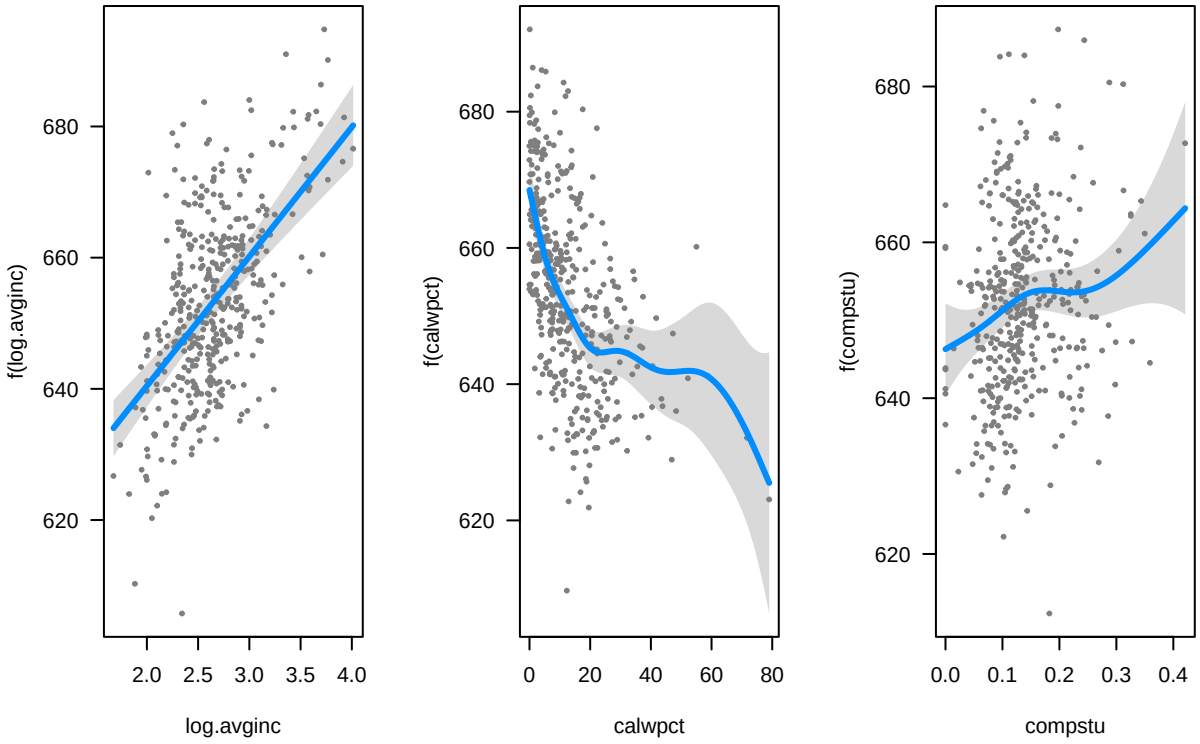
```
modplm <- mgcv::gam(mathscr ~ log.avginc+s(calwpct)+s(compstu),family = gaussian,method = "GCV.Cp")
summary(modplm)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## mathscr ~ log.avginc + s(calwpct) + s(compstu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  601.146      5.466 109.977   <2e-16 ***
## log.avginc    19.735      2.056   9.599   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(calwpct) 5.733  6.882 14.25 < 2e-16 ***
## s(compstu) 3.321  4.208  3.58 0.00636 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.629   Deviance explained = 63.8%
## GCV = 134.01  Scale est. = 130.48    n = 420
```

Ahora todas las covariables del componente no paramétrico son significativas bajo funciones suaves y así también el componente paramétrico y el intercepto. Con el nuevo modelo se tiene un $R^2 - ajust$ de 0.629, un ajuste bueno para los datos y una *Deviance* de 63.8%.

En el siguiente gráfico se puede ver los valores ajustados estimados bajo el modelo propuesto.

```
par(mfrow=c(1,3))
visreg(modplm)
```




```
dev.off()
```

```
## null device
##           1
```

Se muestra claramente las covariables que están siendo ajustadas por el componente paramétrico (lineal) y no paramétrico (funciones suaves).

3 Selección del modelo

3.1 Primer criterio

Para evaluar qué modelo se esta ajustando mejor a los datos una opción es utilizar el $R^2 - ajust$ o la *Deviance*. Bajo este criterio de los tres modelos presentados el mayor $R^2 - ajust$ y *Deviance* es el modelo no paramétrico con 0.637 y 64.6%.

3.2 Criterio de Información de Akaike

Se presenta el AIC de los tres modelos propuestos:

```
modlm$aic
```

```
## [1] 3276.616
```

```
modgam$aic
```

```
## [1] 3243.069
```

```
modplm$aic
```

```
## [1] 3250.722
```

Bajo el AIC el modelo que presenta el menor número es el modelo no paramétrico con 3243.069.

3.3 Prueba F

La prueba F es otro criterio de comparación de deviance para ver si el modelo propuesto es el que presenta más ajuste. Es así que, primero se evaluará el modelo lineal clásico con el modelo no paramétrico.

```
anova(modlm,modgam,test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: mathscr ~ calwpct + log.avginc + compstu
## Model 2: mathscr ~ s(calwpct) + s(log.avginc) + s(compstu)
##   Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
## 1       416      58691
## 2       405      52116 11.005   6574.9 4.6753 9.629e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como se puede ver sale que el modelo propuesto es significativo, vale decir que el modelo no paramétrico presenta mejor ajuste que el modelo lineal clásico. De la misma manera se compara el modelo lineal clásico con el modelo parcial.

```
anova(modlm,modplm,test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: mathscr ~ calwpct + log.avginc + compstu
## Model 2: mathscr ~ log.avginc + s(calwpct) + s(compstu)
##   Resid. Df Resid. Dev   Df Deviance      F    Pr(>F)
## 1    416.00      58691
## 2    406.91      53359 9.09   5331.9 4.4955 1.106e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede observar que el modelo parcial presenta mejor ajuste que el nodelo lineal bajo este tipo de pruebas.

3.4 Conclusión

Como se pudo observar el modelo que presenta mejor ajuste a los datos es un modelo no paramétrico con suavizadores spline cúbico natural, teniendo en cuenta que es el que presenta mayor $R^2 - ajust$, $Deviance$, menor AIC y bajo la prueba F es significativo.

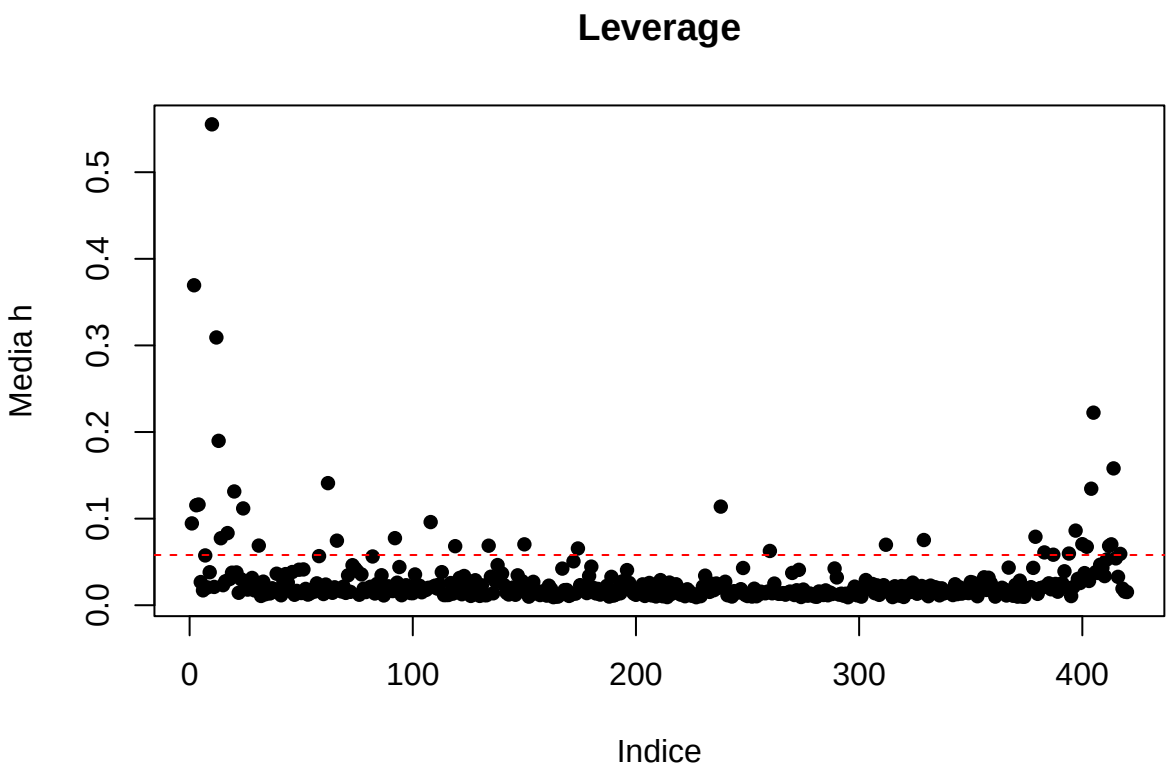
4 Análisis de Diagnóstico

4.1 Puntos de apalancamiento

Bajo el modelo no paramétrico que presenta mejor ajuste a los datos, se tiene, el análisis de apalancamiento (leverage).

```
n=length(modgam$y)
ri=modgam$residuals
h=modgam$hat
phi=modgam$sig2
varr=(1-h)*phi
tdf=ri/sqrt(varr)
di=(h*(ri^2))/(((1-h)^2)*phi*sum(h))#Distancia de Cox
a=max(tdf)
b=min(tdf)
cut=(2*sum(h))/n
```

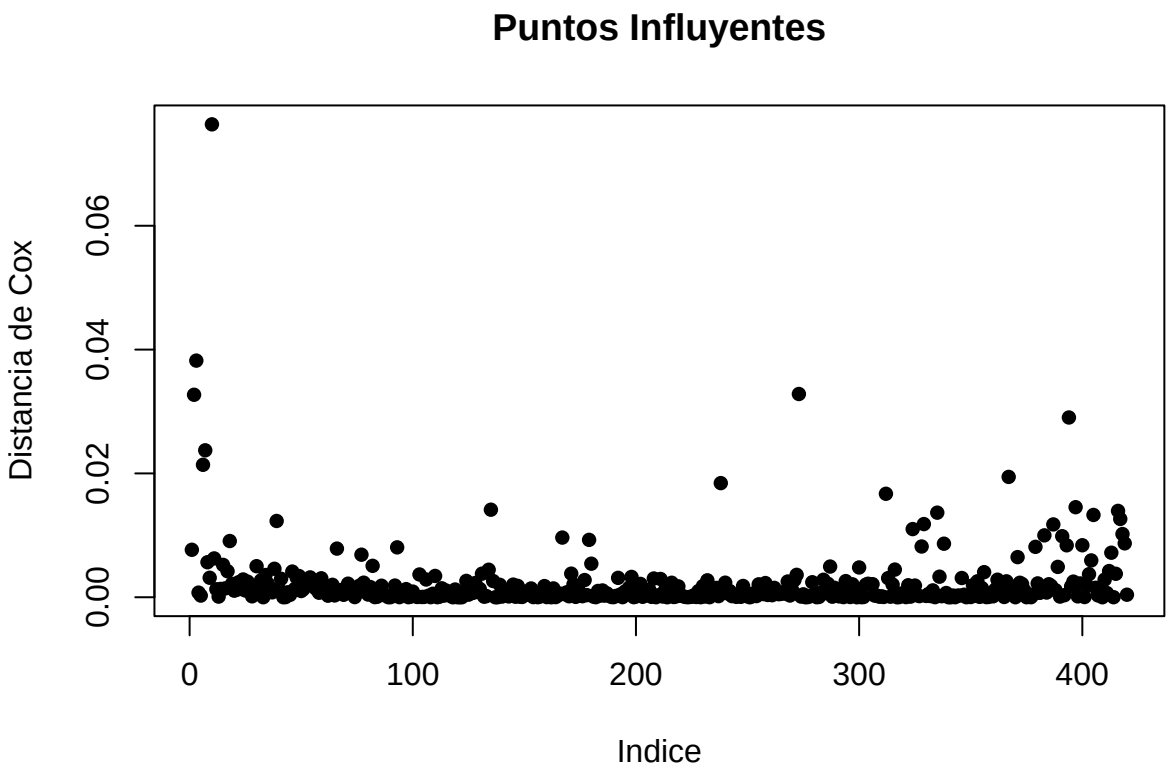
```
plot(h,xlab="Indice",ylab="Media h",main="Leverage",pch=16)
abline(cut,0,lty=2,col="red")
```



Si presenta puntos de apalancamiento.

4.2 Puntos influyentes

```
plot(di,xlab="Indice",ylab="Distancia de Cox",main="Puntos Influyentes",pch=16)
```



De la misma manera existen puntos u observaciones influyentes para el modelo.

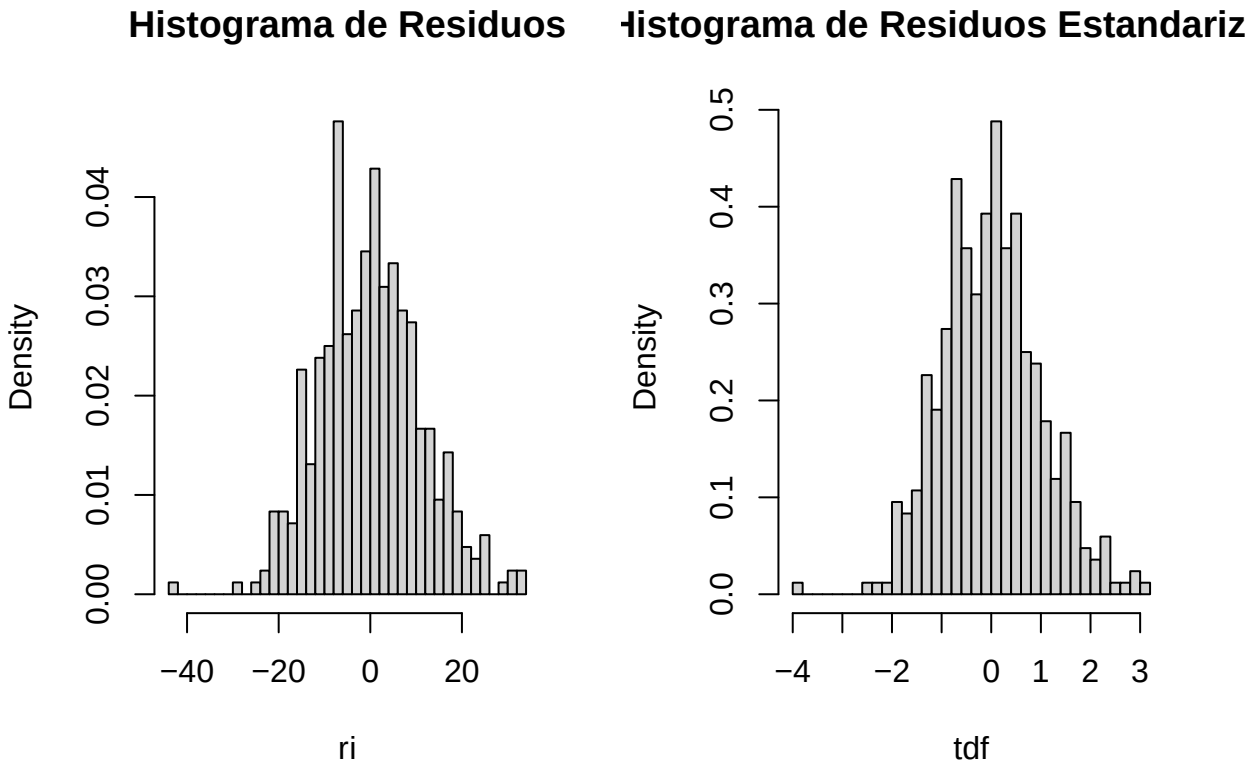
5 Análisis de residuos

Se realizará el análisis de residuos para verificar los supuestos de normalidad, para este cometido se utilizarán los residuos obtenidos por la regresión no paramétrica y así también los estandarizados.

5.1 Supuesto de normalidad

Se grafica el histograma de los residuos de la regresión no paramétrica y los estandarizados, se observa que posiblemente tengan un comportamiento con distribución normal.

```
#Normalidad
par(mfrow=c(1,2))
hist(ri,main="Histograma de Residuos",freq = FALSE,breaks = 30)
hist(tdf,main="Histograma de Residuos Estandarizados",freq = FALSE,breaks = 30)
```



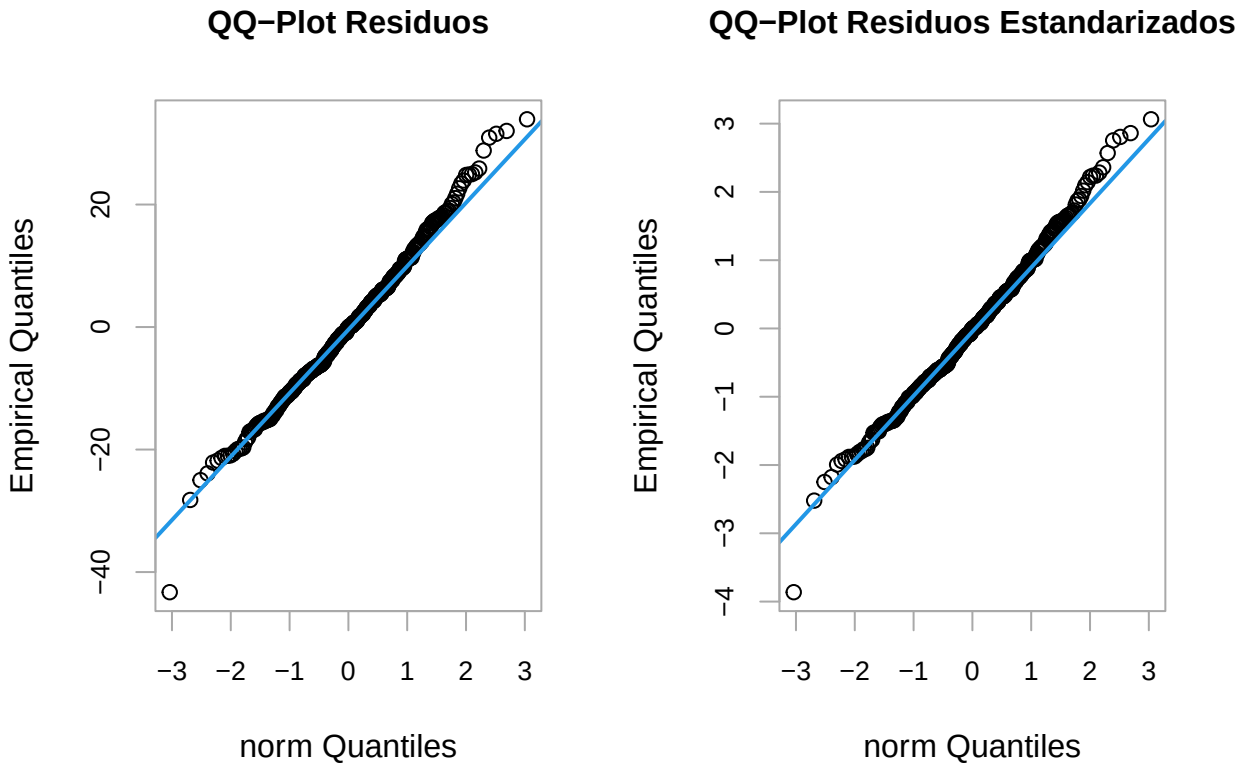
También se puede mostrar que existen datos atípicos en la cola izquierda en ambos histogramas.

Se grafica el qq-plot, donde se puede observar que el diagrama es casi lineal lo cual muestra un posible comportamiento normal

```
## Warning: package 'PerformanceAnalytics' was built under R version 4.1.3

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'PerformanceAnalytics'

## The following objects are masked from 'package:moments':
##
##     kurtosis, skewness

## The following object is masked from 'package:graphics':
##
##     legend
```

```
par(mfrow=c(1,2))
chart.QQPlot(ri,distribution = "norm",main = "QQ-Plot Residuos")
chart.QQPlot(tdf,distribution = "norm",main = "QQ-Plot Residuos Estandarizados")
```



Para tener evidencia estadística se realizará la dócima de Jarque-Bera para modelos de regresión.

H_0 : residuos son normales

```
library(normtest)
jb.norm.test(ri)
```

```
##
##  Jarque-Bera test for normality
```

```
##
## data:  ri
## JB = 3.7109, p-value = 0.1395
```

```
jb.norm.test(tdf)
```

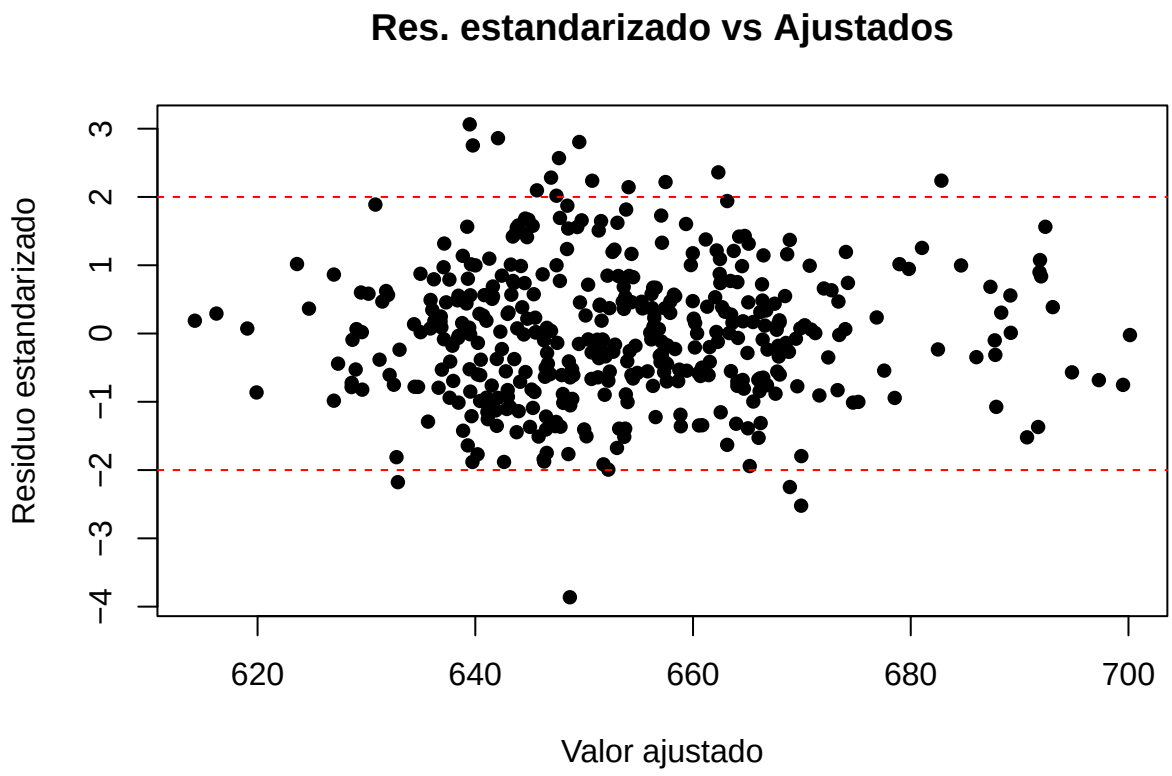
```
##
##  Jarque-Bera test for normality
##
## data:  tdf
## JB = 3.5682, p-value = 0.1335
```

Tanto para los residuos de la regresión no paramétrica como los estandarizados el p-valor es mayor a 0.05 teniendo evidencia estadística para no rechazar H_0, así los residuos son normales.

5.2 Supuesto de homocedasticidad

Primero se hará un análisis gráfico de los residuos estandarizados con los indices y luego con los valores ajustados:

```
#Gráficos de Reisudos Estandarizados vs valores ajustados
plot(modgam$fitted.values,tdf,xlab="Valor ajustado",main="Res. estandarizado vs Ajustados",ylab="Residu
abline(2,0,lty=2,col="red")
abline(-2,0,lty=2,col="red")
```

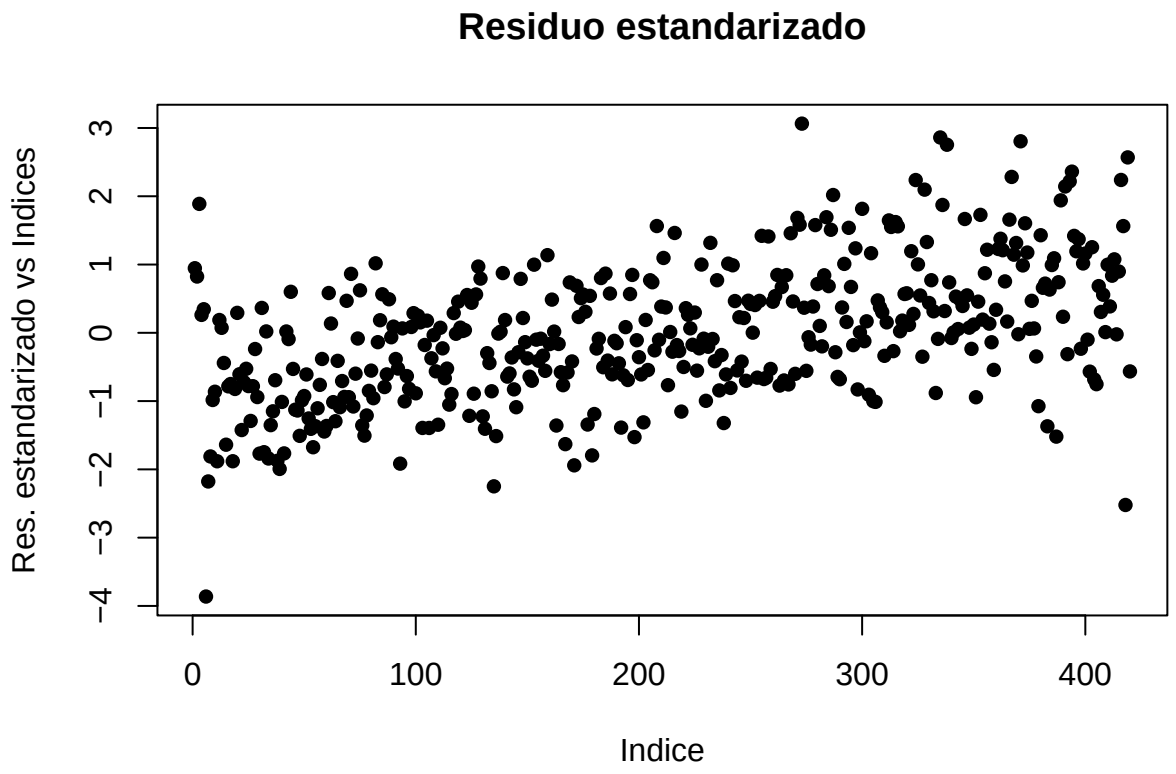


Como se puede ver existen algunos puntos que salen de las bandas, lo cual concluiria que posiblemente los residuos sean heterocedasticos.

5.3 Supuesto de autocorrelación

Se gráfica los residuos estandarizados, los cuales siguen un patron lineal lo cual podria concluir que estamos bajo errores autocorrelacionados.

```
#Gráficos de Reisudos Estandarizados vs Indices
plot(tdf,xlab="Indice",ylab = "Res. estandarizado vs Indices",main="Residuo estandarizado",pch=16)
```



6 Conclusión

Se pudo evaluar tres tipos de modelos el lineal clásico, no paramétrico y el parcial donde se pudo evidenciar que la covariable `expnstu` es no significativa tomando la decisión de eliminar la misma ya que no aporta a la explicación del modelo. Por otro lado se comparo los tres tipos de modelos resultando el mejor el modelo de regresión no paramétrico bajo los criterios de $R^2 - ajust$, $Deviance$, AIC y prueba F.

En el análisis de diagnóstico mediante puntos de apalancamiento e influyentes globales se determino que existen puntos que influyen bastante al modelo y en el análisis de residuos se pudo verificar el supuesto de normalidad, pero no así el de homocedasticidad y autocorrelación.