Índice general

Capítulo 1

Aplicación de los MAG al análisis del cambio climático

En esta sección nos proponemos el aplicar el contexto teórico visto hasta ahora sobre los modelos aditivos generalizados al análisis del cambio climático. Para ello principalmente utilizaremos el paquete de R: "mgcv" (siglas en inglés de "Vehículo de Computación para MAG Mixtos"), en particular haremos uso de su funcion "gam", la cual permite ajustar modelos aditivos generalizados, entre otros tipos de modelos, mediante splines de regresión penalizados (u smoothers similares) donde los parámetros de suavizado pueden ser estimados por distintos métodos, como por ejempli: mínima validación cruzada generalizada, mínimo AIC, por máxima verosimilitud o por REML (que es la opción por defecto).

Además del método de estimación, esta función también admite otras entradas que indican qué familia de distribuciones exponenciales se utiliza, si las observaciones toman distintos pesos, el método de optimización numérica utilizado, otros parámetros de control de estos métodos para el caso de que los habituales no converjan, etc.

Dividiremos las aplicaciones prácticas de los MAG en tres partes: la primera se centra en modelar la temperatura media mensual según una serie de variables climáticas y ver cómo ha variado con los años, en la siguiente veremos cómo han evolucionado a lo largo del tiempo las concentraciones de gases de efecto invernadero en la atmósfera y algunos de sus efectos, por último estudiaremos la media variacional del nivel del mar respecto del año 1993.

1.1. Modelización de la temperatura media mensual

1.1.1. Datos

Para esta primera aplicación de los modelos aditivos generalizados utilizaremos datos de elementos climáticos proporcionados por la Agencia Estatal de Meteorología española (AEMET), para ello utilizaremos la libreria "climaemet" **?**:

```
#install.packages('climaemet')
library(climaemet)
```

Como hemos dicho antes, el conjunto de datos sobre el que trabajaremos a lo largo de esta sección está formado por variables climáticas, estas se definen como elementos que caracterizan el tiempo atmosférico y que interactuan entre sí en la troposfera. Aunque son elementos relacionados con el campo de la meteorología, su estudio a largo plazo, fundamenta las bases científicas de la climatología. En particular, el conjunto de datos mensuales que nos proporciona la anterior librería contiene más de 40 variables climáticas, por comodidad y para una mejor interpretación de los modelos nos quedaremos con las variables que representen la temperatura media mensual, la humedad relativa, la media mensual de precipitaciones y la velocidad media del viento. Otras variables climáticas de interés pueden ser la presión atmosférica y la nubosidad.

Estos datos provienen de una base de datos *open source* que ofrece la AEMET y en particular utilizamos las mediciones tomadas por la estación situada en el aeropuerto de Sevilla, se puede leer más sobre ellos en: https://opendata.aemet.es/centrodedescargas/inicio. Procedamos con la lectura y limpieza de los datos mensuales desde 1960 a 2023:

```
library(tidyr)
library(dplyr)
Clima <- aemet_monthly_period(station = "5783", start = 1960, end = 2023)
Clima <- Clima %>% separate(fecha, into = c("Año", "Mes"), sep = "-")
Clima$Año <- as.numeric(Clima$Año)
Clima$Mes <- factor(Clima$Mes, levels = as.character(1:12))
Clima <- Clima[,c(1,2,6,11,27,29,32)] # Seleccionamos las variables que nos interesa
colnames(Clima) <- c('Año','Mes','HR','PresM','Prec','WMed','TMedM')
Clima <- Clima %>% arrange(Año, Mes) # Ordenamos por año y mes
Clima <- Clima[complete.cases(Clima$Mes),] # Retiramos las medias anuales
```

Hagamos ahora la primera visualización de ellos observando su estructura y un resumen:

```
str(Clima)
```

```
## tibble [768 x 7] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:768] 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ HR   : num [1:768] 88 85 87 76 73 73 48 45 47 75 ...
##  $ PresM: num [1:768] 1011 1013 1015 1013 1012 ...
##  $ Prec : num [1:768] 81.3 205.2 108.7 18.4 45 ...
##  $ WMed : num [1:768] 9 12 16 12 12 12 11 12 13 15 ...
##  $ TMedM: num [1:768] 10.3 13 14.4 17.4 21 25.9 27.2 25.9 24.3 17.1 ...
```

```
summary(Clima)
```

```
##       Año             Mes             HR            PresM           Prec
```

```
##  Min.   :1960   1       : 64   Min.   :31.00   Min.   :1004   Min.   :  0.00
##  1st Qu.:1976   2       : 64   1st Qu.:51.00   1st Qu.:1011   1st Qu.:  2.00
##  Median :1992   3       : 64   Median :61.00   Median :1013   Median : 26.40
##  Mean   :1992   4       : 64   Mean   :60.74   Mean   :1014   Mean   : 45.81
##  3rd Qu.:2007   5       : 64   3rd Qu.:71.00   3rd Qu.:1016   3rd Qu.: 65.25
##  Max.   :2023   6       : 64   Max.   :90.00   Max.   :1028   Max.   :361.10
##                 (Other):384   NA's   :9       NA's   :5
##       WMed            TMedM
##  Min.   : 5.00   Min.   : 8.40
##  1st Qu.: 9.00   1st Qu.:13.50
##  Median :11.00   Median :18.10
##  Mean   :11.09   Mean   :18.93
##  3rd Qu.:12.00   3rd Qu.:24.65
##  Max.   :22.00   Max.   :30.70
##  NA's   :17      NA's   :5
```

- HR: la humedad relativa media es un valor porcentual de la cantidad de vapor de agua presente en el aire con respecto a la máxima posible para unas condiciones dadas de presión y temperatura.
- PresM: la presión media mensual al nivel de la estación.
- Prec: la precipitación total mensual medida en milímetros.
- WMed: la velocidad media del aire se mide en metros por segundo.
- TMedM: la temperatura media mensual viene dada en grados centígrados.

1.1.2. Descripción del modelo

Tomaremos a la variable *TMedM* como variable de respuesta y al resto como variables explicativas. Antes de definir el modelo debemos notar que la variable *Año* no se ha definido como variable categórica, como sí se hizo para la variable *Mes*, sino que se define como variable numérica para luego poder tener una mejor representación de los resultados obtenidos. Además, si se hubiera definido de tal forma, resultaría que la mayoría de factores son no significativos. Dicho esto definimos el modelo como:

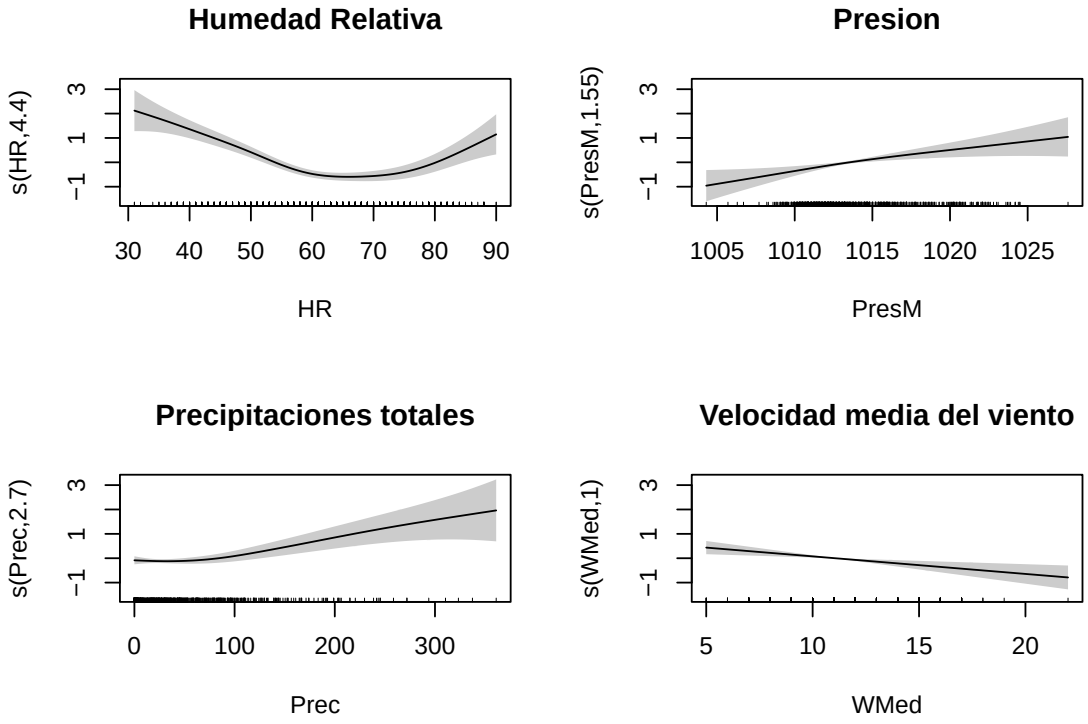
```
#install.packages('mgcv')
library(mgcv)
```

```
mag1 <- gam(TMedM~ s(HR)+s(PresM)+s(Prec)+s(WMed)+Año+Mes,data = Clima)
summary(mag1)
```

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

```
## Formula:
## TMedM ~ s(HR) + s(PresM) + s(Prec) + s(WMed) + Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -55.026379   5.890706  -9.341  < 2e-16 ***
## Año           0.032945   0.002917  11.294  < 2e-16 ***
## Mes2          1.923150   0.232581   8.269 6.56e-16 ***
## Mes3          4.537181   0.272233  16.667  < 2e-16 ***
## Mes4          6.970016   0.318541  21.881  < 2e-16 ***
## Mes5         10.287254   0.340623  30.201  < 2e-16 ***
## Mes6         13.898675   0.351116  39.584  < 2e-16 ***
## Mes7         16.644447   0.378536  43.971  < 2e-16 ***
## Mes8         16.817040   0.373449  45.032  < 2e-16 ***
## Mes9         14.188786   0.329066  43.118  < 2e-16 ***
## Mes10         9.765255   0.283822  34.406  < 2e-16 ***
## Mes11         4.157867   0.238799  17.412  < 2e-16 ***
## Mes12         0.798865   0.218239   3.661  0.00027 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df      F  p-value
## s(HR)    4.396  5.468 19.901  < 2e-16 ***
## s(PresM) 1.551  1.945  6.211 0.001852 **
## s(Prec)  2.704  3.410  6.083 0.000262 ***
## s(WMed)  1.000  1.000 10.396 0.001320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.962   Deviance explained = 96.3%
## GCV = 1.4701  Scale est. = 1.4251    n = 740
```

Como podemos ver en el resumen del modelo, se toma por defecto que la variable dependiente sigue la distribución normal y que la función de enlace es la identidad. Se obtiene que el modelo es capaz de explicar el 96.3 % de la varianza con $R^2_{adj} = 0.962$. Se tiene también que todas las variables predictoras son significativas. Observemos qué efecto tienen las variables predictoras sobre la temperatura media mensual:



De las gráficas de la derecha se puede interpretar que los efectos de *PresM* y *WMed* sobre la temperatura media mensual son lineales. Ajustamos entonces un nuevo modelo aditivo generalizado del mismo modo que antes pero imponiendo que el efecto de estas variables sea lineal:

```
mag2 <- gam(TMedM~ s(HR)+PresM+s(Prec)+WMed+Año+Mes,data = Clima)
summary(mag2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## TMedM ~ s(HR) + PresM + s(Prec) + WMed + Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.388e+02  2.690e+01  -5.158 3.22e-07 ***
## PresM        8.417e-02  2.400e-02   3.507 0.000481 ***
## WMed        -7.201e-02  2.235e-02  -3.222 0.001332 **
## Año          3.253e-02  2.886e-03  11.272  < 2e-16 ***
## Mes2         1.943e+00  2.315e-01   8.394 2.50e-16 ***
## Mes3         4.559e+00  2.713e-01  16.804  < 2e-16 ***
## Mes4         6.972e+00  3.184e-01  21.899  < 2e-16 ***
## Mes5         1.029e+01  3.402e-01  30.245  < 2e-16 ***
## Mes6         1.391e+01  3.507e-01  39.648  < 2e-16 ***
```

```
## Mes7         1.664e+01  3.782e-01  44.011  < 2e-16 ***
## Mes8         1.681e+01  3.730e-01  45.064  < 2e-16 ***
## Mes9         1.420e+01  3.285e-01  43.242  < 2e-16 ***
## Mes10        9.791e+00  2.827e-01  34.635  < 2e-16 ***
## Mes11        4.188e+00  2.366e-01  17.702  < 2e-16 ***
## Mes12        8.095e-01  2.181e-01   3.711 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(HR)   4.469  5.551 20.250  < 2e-16 ***
## s(Prec) 2.610  3.294  5.994 0.000342 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.962   Deviance explained = 96.3%
## GCV = 1.4708  Scale est. = 1.4269    n = 740
```

Podemos ver que tanto la desviación explicada como la estimación del error por validación cruzada coinciden con las del modelo anterior, sin embargo utilizaremos la función *anova* para dar una prueba de razón de verosimilitud que determine si el modelo más complejo mejora significativamente el ajuste.

```
anova(mag1,mag2,test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: TMedM ~ s(HR) + s(PresM) + s(Prec) + s(WMed) + Año + Mes
## Model 2: TMedM ~ s(HR) + PresM + s(Prec) + WMed + Año + Mes
##   Resid. Df Resid. Dev       Df Deviance      F Pr(>F)
## 1    715.18     1022.3
## 2    716.15     1024.4 -0.97706  -2.1327 1.5316 0.2158
```

Como el p-valor resultante del contraste de hipótesis es > 0.05, no tenemos evidencias significativas como para rechazar la hipótesis nula, es decir, se acepta que ambos modelos tienen el mismo ajuste.

También es posible compararlos mediante otros criterios, por ejemplo el AIC (Akaike Information Criterion) que tiene en cuenta el número de parámetros a estimar y el valor objetivo de la función de log-verosimilitud:

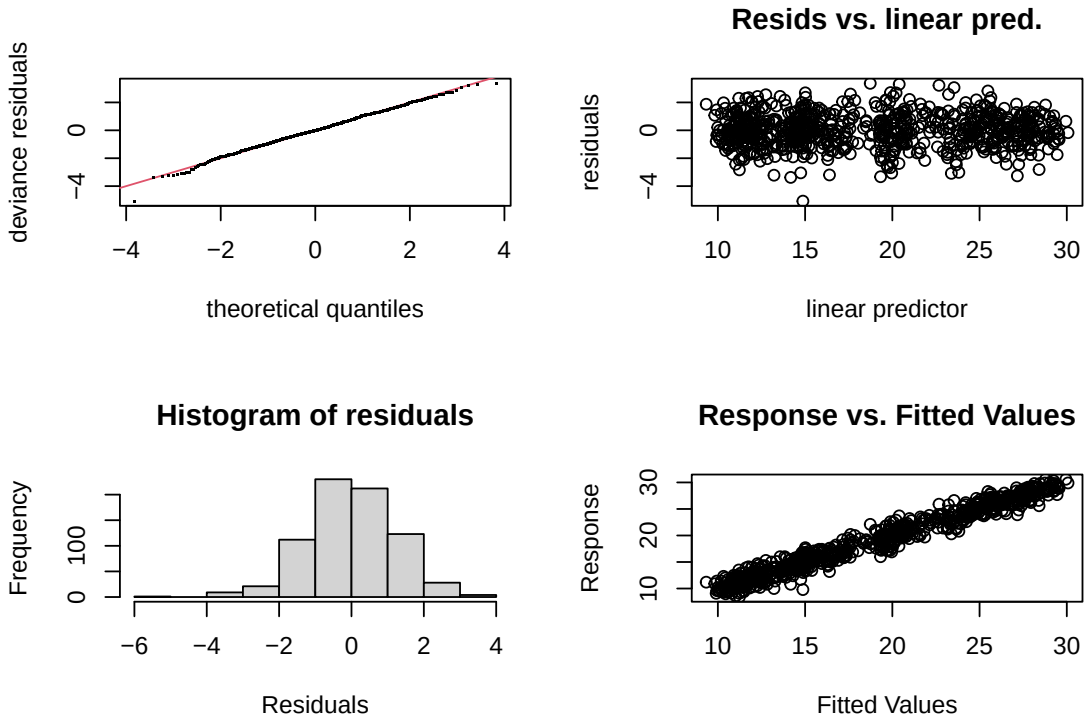
```
AIC(mag1,mag2)
```

```
##            df      AIC
## mag1 23.65122 2386.466
## mag2 23.07915 2386.864
```

En este caso son casi idénticos, aunque el primer modelo tiene menor AIC.

Para obtener más información sobre el modelo planteado se utiliza la siguiente rutina de diagnósticos que nos proporciona información y gráficos útiles para evaluar la calidad del ajuste del modelo.

```
gam.check(mag1)
```



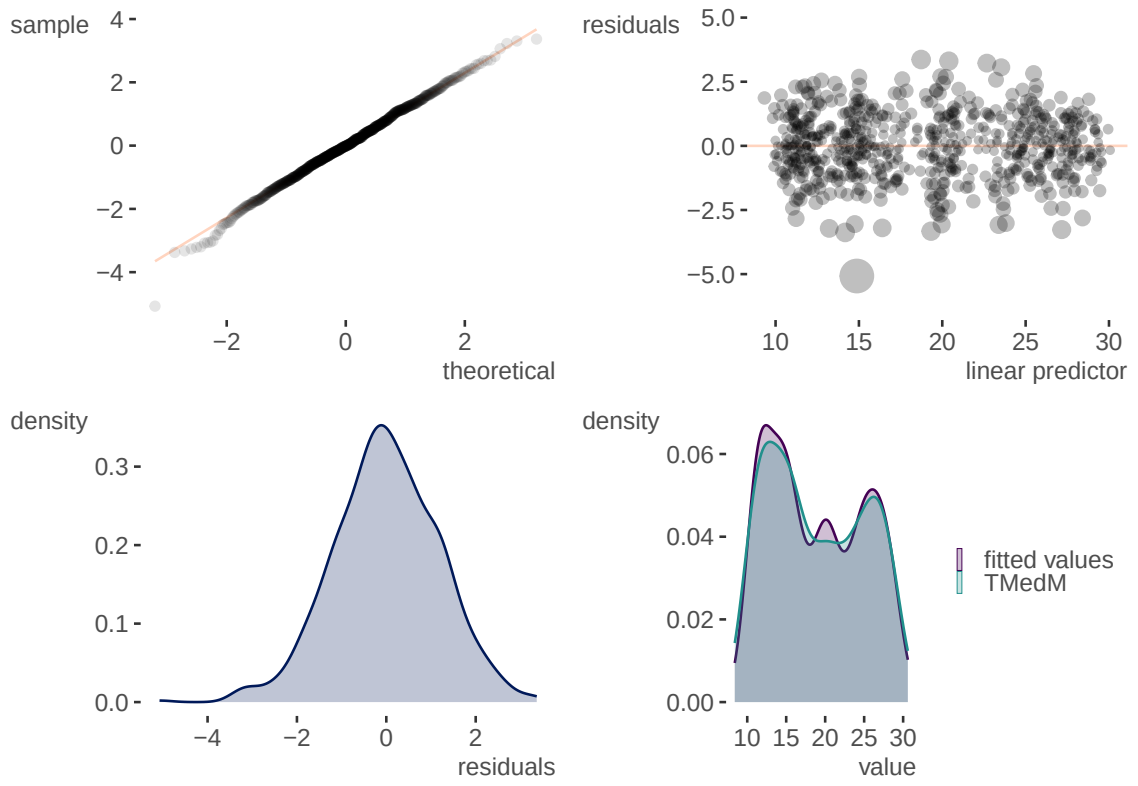

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 10 iterations.
## The RMS GCV score gradient at convergence was 1.49324e-07 .
## The Hessian was positive definite.
## Model rank =  49 / 49
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'  edf k-index p-value
## s(HR)    9.00 4.40    0.93   0.020 *
## s(PresM) 9.00 1.55    0.92   0.010 **
## s(Prec)  9.00 2.70    0.94   0.075 .
## s(WMed)  9.00 1.00    0.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por un lado, la salida por consola nos informa de que se obtiene la convergencia por optimización del GCV y que el modelo es de rango completo. Los p-valores que aparecen se corresponden a los tests de residuos aleatorios correspondientes para cada predictor, en este caso todos excepto el asociado a *Prec* son < 0.05, lo que indica que los residuos no están distribuidos aleatoriamente y que se necesitaría una base de funciones de mayor dimensión. Por otro lado, observemos qué representa cada una de las gráficas generadas y cuál sería el caso ideal para cada una de ellas:

- Q-Q Plot: compara la distribución de los residuos con una distribución normal. Lo ideal es que los puntos se alineen aproximadamente en una línea recta.
- Resids vs. linear pred: representa los residuos contra el predictor lineal, ayuda a verificar si los residuos se distribuyen aleatoriamente, que sería lo ideal.
- Histogram of residuals: se trata de un histograma de los residuos, en este caso lo ideal es que muestre una distribución aproximadamente normal, centrada en cero, esto indicaría que los residuos no presentan sesgos significativos.
- Response vs. Fitted Values: representa los valores observados frente a los valores ajustados, lo ideal sería que los puntos resultantes se aconglomerasen en torno a la recta $x = y$.

Por lo general este modelo se comporta de manera decente, ya que se aproxima mucho a los casos ideales de cada gráfica. Podemo utilizar la librería *visibly* de **?** para hacer esta representación de forma más clara:

```
#install.packages('visibly')
library(visibly)
plot_gam_check(mag1)
```



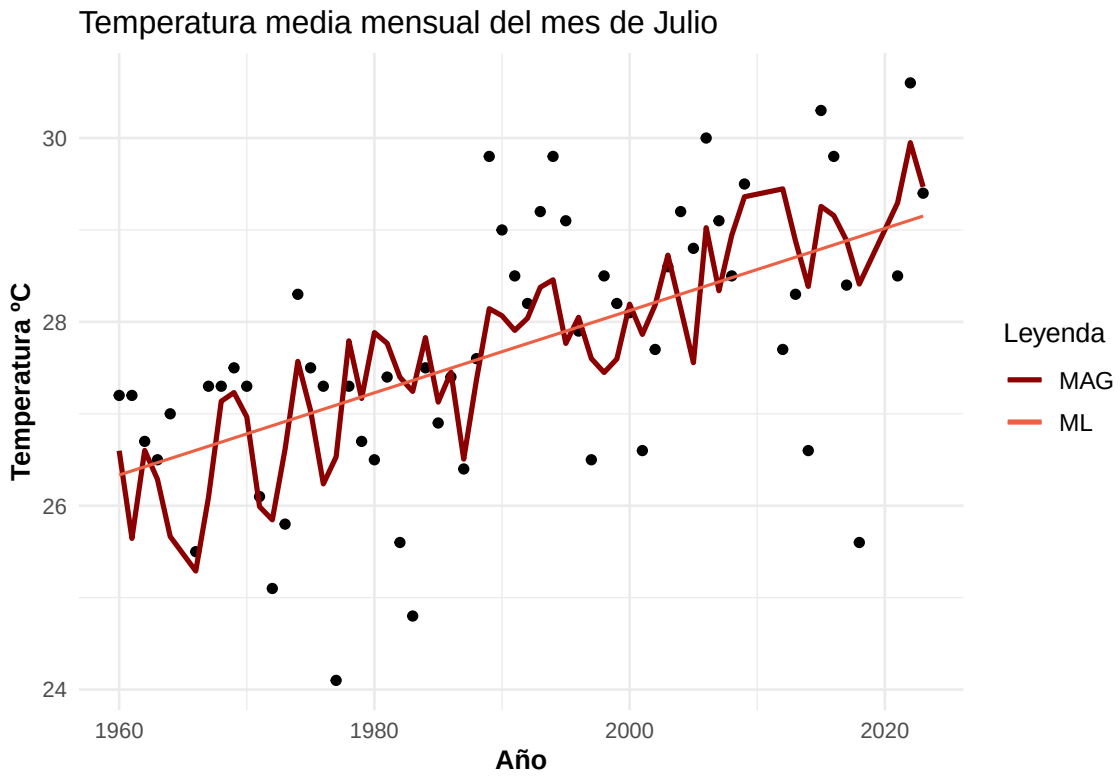
1.1.3. Visualización de los resultados

Una vez generado el modelo y comprobado que se tiene una bondad de ajuste decente, veamos cómo ajusta los datos para poder visualizar si se ha producido un cambio significativo en la temperatura media mensual a lo largo de los años.

```
Julio <- filter(Clima, Clima$Mes == 7)
Julio$Preds <- predict(mag1, newdata = Julio)
Julio <- Julio[complete.cases(Julio$Preds),]

lm1 <- gam(TMedM ~ Año,data = Julio)
Julio$LPreds <- predict(lm1,Julio)

library(ggplot2)
ggplot(Julio,aes(x=Año))+
  geom_point(aes(y=TMedM),size=1.5, col = 'black')+
  theme_minimal()+
  geom_line(aes(y=Preds, color = 'MAG'),linewidth=1)+
  geom_line(aes(y=LPreds, color = 'ML'),linewidth=0.6)+
  labs(title ="Temperatura media mensual del mes de Julio",x="Año",y="Temperatura ºC"
  scale_color_manual(values = c('MAG' = 'darkred', 'ML' = '#EB6146'), name = "Leyenda
  theme(axis.title = element_text(face = "bold"),
        legend.text = element_text( size = 10,colour = "black"),
        legend.position = 'right')
```



Con esta gráfica podemos apreciar claramente como la temperatura media en el mes de Julio ha ido aumentado con el paso de los años en la estación meteorológica del aeropuerto

de San Pablo. Hemos introducido la recta proporcionada por el modelo lineal para los datos de los meses de Julio para tener una mejor apreciación de tal incremento.

1.2. Modelización de gases de efecto invernadero

1.2.1. Descripción de los datos

En esta sección ajustaremos modelos aditivos generalizados para estudiar la concentración atmosférica de gases de efecto invernadero, en particular analizaremos las medias mensuales globales de las concentraciones de dióxido de carbono (CO_2), Metano (CH_4) y óxido nitroso (N_2O). Para ello consideraremos los datos proporcionados por United Nations Environment Programme (UNEP), para el CO_2 se obtienen en https://wesr.unep.org/climate/essential-climate-variables-ecv/atmospheric-co2-concentration y para los dos siguientes en https://wesr.unep.org/climate/essential-climate-variables-ecv/atmospheric-ch4-n2o-sf6-concentration.

Ya hablamos sobre las emisiones de gases contaminantes en **??** pero no se llego a definir en qué consistían. Estos gases son capaces de absorber y emitir radiación dentro del espectro inflarrojo, por tanto son capaces de retener el calor del Sol, lo que permite que el clima terrestre sea habitable para la humanidad. Sin embargo, desde el inicio de la revolución industrial, la actividad humana ha producido un desequilibrio en los niveles de concentración de estos gases en la atmósfera. En particular estudiaremos las concentraciones de los tres tipos de gases antes mencionados por ser los que se emiten en mayor cantidad o por ser las más potentes (en términos de contribución al efecto invernadero). Por ejemplo, las emisiones de CO_2 se corresponden aproximadamente con tres cuartas partes del total de emisiones de GEI, sin embargo el CH_4 y el N_2O representan una parte mucho menor que el dióxido de carbono pero por unidad son mucho más potentes como gases de efecto invernadero.

Hagamos ahora una primera visualización de los datos. Para tener una mejor lectura de ellos, primero debemos transformar el archivo a una hoja de excel *.xlsx*, luego utilizaremos la librería *readxl* para leerlo y la librería *lubridate* para obtener la fecha como las variable *Año* y *Mes*:

```
library(readxl)
library(lubridate)
```

- CO_2:

```
CO2 <- read_excel('trends-in-atmospheric-carbon-dioxide-concentration.xlsx')

CO2$DateTime <- as.Date(CO2$DateTime) # Creamos las variables año y mes
CO2$Año <- as.numeric(year(CO2$DateTime))
CO2$Mes <- factor(month(CO2$DateTime),levels = as.character(1:12))
CO2$Tmes <- as.numeric(CO2$'Monthly Data')
CO2$Trend <- as.numeric(CO2$'Trend')
CO2 <- CO2[,c(4,5,6,3)]
CO2 <- CO2 %>% arrange(Año, Mes)
```

De este modo nos queda un data frame con 794 observaciones, correspondientes a los meses desde marzo del 1958 hasta abril de 2024, y con las variables *Año*, *Mes*, *Tmes* que se corresponde con la concentración media de CO_2 a nivel global medida en partes por millón (ppm) y *Trend* que es la media anual de las anteriores. 1 ppm de un gas significa que existe una molécula de ese gas por cada millón de moléculas de aire.

```
str(CO2)
```

```
## tibble [794 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:794] 1958 1958 1958 1958 1958 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 3 4 5 6 7 8 9 10 11 12 ...
##  $ Tmes : num [1:794] 316 317 318 317 316 ...
##  $ Trend: num [1:794] NA NA NA NA NA NA NA NA NA NA ...
```

```
summary(CO2)
```

```
##       Año             Mes           Tmes           Trend      
##  Min.   :1958   3      : 67   Min.   :312.4   Min.   :316.0  
##  1st Qu.:1974   4      : 67   1st Qu.:330.4   1st Qu.:331.1  
##  Median :1991   1      : 66   Median :355.0   Median :355.7  
##  Mean   :1991   2      : 66   Mean   :359.0   Mean   :359.3  
##  3rd Qu.:2007   5      : 66   3rd Qu.:384.6   3rd Qu.:384.0  
##  Max.   :2024   6      : 66   Max.   :426.6   Max.   :421.1  
##                 (Other):396                   NA's   :729
```

- CH_4:

```
CH4 <- read_excel('trends-in-atmospheric-methane-concentration.xlsx')

CH4$DateTime <- as.Date(CH4$DateTime) # Creamos las variables año y mes
CH4$Año <- as.numeric(year(CH4$DateTime))
CH4$Mes <- factor(month(CH4$DateTime),levels = as.character(1:12))
CH4$Trend <- as.numeric(CH4$'Trend')
CH4 <- CH4[,c(3,4,2)]
CH4 <- CH4 %>% arrange(Año, Mes)
```

En este caso disponemos de 487 observaciones correspondientes a meses entre 1983 y 2024, las variables de tiempo *Año* y *Mes* y la variable *Trend* la cual representa la media mensual de concentración de metano a nivel global medida en partes por billón (ppb).

```
str(CH4)
```

```
## tibble [487 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:487] 1983 1983 1983 1983 1983 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 7 8 9 10 11 12 1 2 3 4 ...
##  $ Trend: num [1:487] 1635 1636 1636 1637 1638 ...
```

```r
summary(CH4)
```

```
##       Año             Mes          Trend
##  Min.   :1983   1      : 41   Min.   :1635
##  1st Qu.:1993   7      : 41   1st Qu.:1737
##  Median :2003   8      : 41   Median :1775
##  Mean   :2003   9      : 41   Mean   :1778
##  3rd Qu.:2013   10     : 41   3rd Qu.:1816
##  Max.   :2024   11     : 41   Max.   :1928
##                 (Other):241
```

- N_2O:

```r
N2O <- read_excel('trends-in-atmospheric-nitrous-oxide-concentration.xlsx')

N2O$DateTime <- as.Date(N2O$DateTime) # Creamos las variables año y mes
N2O$Año <- as.numeric(year(N2O$DateTime))
N2O$Mes <- factor(month(N2O$DateTime),levels = as.character(1:12))
N2O$Trend <- as.numeric(N2O$'Trend')
N2O <- N2O[,c(3,4,2)]
N2O <- N2O %>% arrange(Año, Mes) # Ordenamos por año y mes
```

Para el caso del óxido nitroso sólo disponemos datos desde el 2001, por lo que obtenemos un conjunto de 277 observaciones para las variables *Año*, *Mes* y *Trend* que representa la media mensual de concentración de N_2O a nivel global medida en partes por billón (ppb).

```r
str(N2O)
```

```
## tibble [277 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Año  : num [1:277] 2001 2001 2001 2001 2001 ...
##  $ Mes  : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Trend: num [1:277] 316 316 316 316 316 ...
```

```r
summary(N2O)
```

```
##       Año             Mes          Trend
##  Min.   :2001   1      : 24   Min.   :316.0
##  1st Qu.:2006   2      : 23   1st Qu.:320.0
##  Median :2012   3      : 23   Median :325.1
##  Mean   :2012   4      : 23   Mean   :325.6
##  3rd Qu.:2018   5      : 23   3rd Qu.:330.7
##  Max.   :2024   6      : 23   Max.   :337.3
##                 (Other):138
```

Solo con los resúmenes de los datos para los tres gases, teniendo en cuenta las medidas en las que vienen dados, ya se puede ver la gran diferencia que hay entre sus proporciones en la atmósfera.

1.2.2. Descripción de los modelos

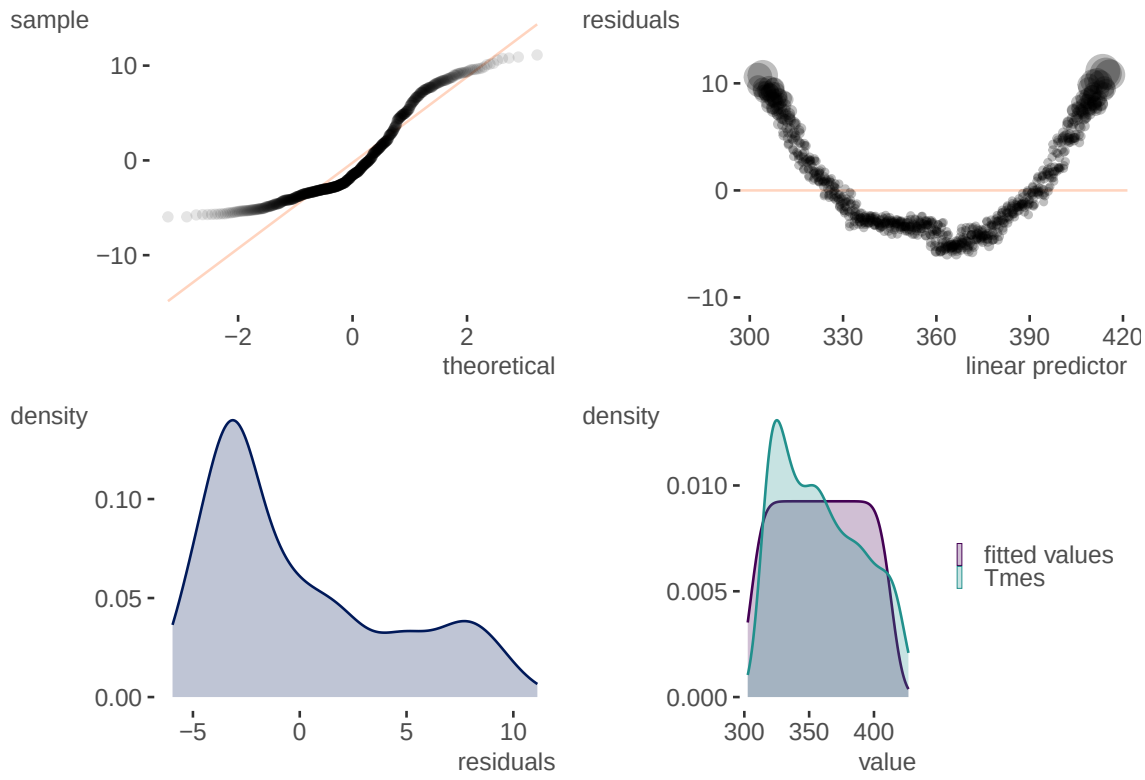
Partiremos definiendo un modelo lineal para los datos de CO_2 que tenga como variable de respuesta la media mesual global de la concentración de este gas medida en ppm y como predictoras las variables *Año* y *Mes*:

```
magCO2 <- gam(Tmes ~ Año + Mes, data = CO2)
summary(magCO2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Tmes ~ Año + Mes
##
## Parametric coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.897e+03  1.637e+01 -176.948  < 2e-16 ***
## Año          1.635e+00  8.215e-03  199.022  < 2e-16 ***
## Mes2         7.917e-01  7.697e-01    1.029 0.303997
## Mes3         1.770e+00  7.668e-01    2.308 0.021232 *
## Mes4         3.071e+00  7.668e-01    4.004 6.81e-05 ***
## Mes5         3.486e+00  7.697e-01    4.530 6.84e-06 ***
## Mes6         2.925e+00  7.697e-01    3.800 0.000156 ***
## Mes7         1.390e+00  7.697e-01    1.806 0.071250 .
## Mes8        -6.210e-01  7.697e-01   -0.807 0.420035
## Mes9        -2.164e+00  7.697e-01   -2.812 0.005046 **
## Mes10       -2.111e+00  7.697e-01   -2.743 0.006228 **
## Mes11       -7.661e-01  7.697e-01   -0.995 0.319867
## Mes12        5.563e-01  7.697e-01    0.723 0.470079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =   0.98   Deviance explained = 98.1%
## GCV = 19.874  Scale est. = 19.549    n = 794
```

Podemos observar que representa un 98.1 % de la desviación explicada, por lo que en principio no parece un mal ajuste, comprobémoslo con el diagnóstico de residuos como se hizo en el apartado anterior:

```
plot_gam_check(magCO2)
```



Gracias a estos gráfios se puede ver cómo el modelo falla en varios aspectos:

- Se puede ver claramente que el Q-Q plot no se adapta a la recta, es decir, la distribución de los residuos no es normal.
- En la gráfica de arriba a la derecha se puede ver como los residuos toman un patrón claro respecto de los predictores, lo que implica que la hipótesis de homocedasticidad de los residuos no es cierta.
- En la de abajo a la izquierda se ve facilmente no es una distribución normal centrada en el 0, lo que confirma lo observado en el Q-Q plot.
- Por último, en el gráfico Response vs. Fitted también se puede intuir la falta de homocedasticidad para los residuos.

Luego, aunque el modelo representanse un gran porcentaje de la desviación explicada, presenta errores significativos en el análisis de los residuos por lo que inferimos que el modelo no es adecuado. Para definir un modelo que se ajuste mejor nos referiremos a dichas gráficas. En primer lugar, que la gráfica de residuos frente predictores se asemeje a una función cuadrática nos indica que puede existir una relación no lineal entre la varible de respuesta y las predictoras, por lo que será conveniente añadir funciones de suavizado al modelo. Además, también puede implicar que la familia de distribuciones exponenciales para la variable de respuesta que estemos utilizando, la normal en este caso, no sea la adecuada. Podemos utilizar el test de normalidad univariante Shapiro-Wilk para comprobarlo:

```
shapiro.test(CO2$Tmes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CO2$Tmes
## W = 0.9407, p-value < 2.2e-16
```

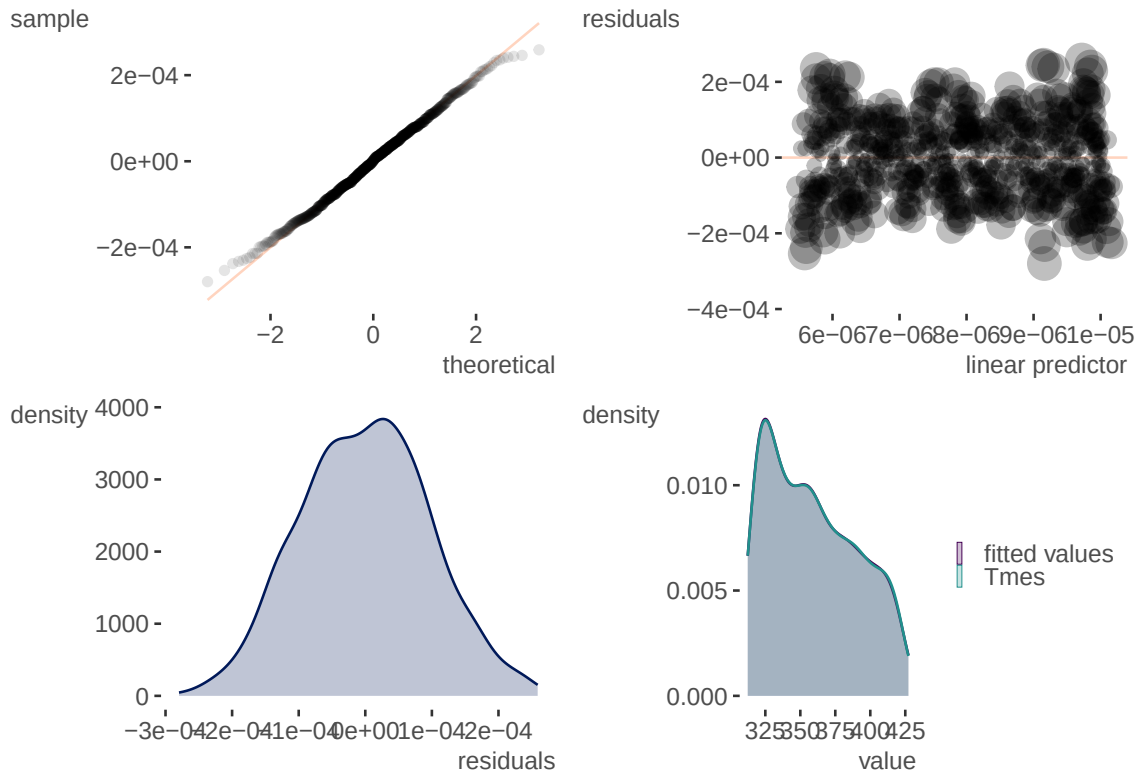
Como el p-valor es muy cercano a 0, se rechaza la hipótesis nula de normalidad de la muestra. Lo que haremos entonces será razonar que como los datos tratados son niveles de concentraciones positivas, quizás nos interese utizar la distribución *Gamma* o la *Inverse Gaussian*. Para determinar cuál de las dos es más conveniente compararemos los modelos definidos para ambas familias con sus respectivas funciones de enlace.

```
magCO2b <- gam(Tmes ~ s(Año) + Mes, data = CO2,
               family = inverse.gaussian(link = "1/mu^2"))
summary(magCO2b)
```

```
##
## Family: inverse.gaussian
## Link function: 1/mu^2
##
## Formula:
## Tmes ~ s(Año) + Mes
##
## Parametric coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  7.962e-06  3.493e-09 2279.492  < 2e-16 ***
## Mes2        -3.316e-08  4.911e-09   -6.754 2.83e-11 ***
## Mes3        -6.782e-08  4.890e-09  -13.869  < 2e-16 ***
## Mes4        -1.217e-07  4.877e-09  -24.956  < 2e-16 ***
## Mes5        -1.462e-07  4.902e-09  -29.820  < 2e-16 ***
## Mes6        -1.229e-07  4.908e-09  -25.040  < 2e-16 ***
## Mes7        -5.866e-08  4.924e-09  -11.913  < 2e-16 ***
## Mes8         2.682e-08  4.945e-09    5.423 7.86e-08 ***
## Mes9         9.343e-08  4.962e-09   18.828  < 2e-16 ***
## Mes10        9.112e-08  4.962e-09   18.364  < 2e-16 ***
## Mes11        3.304e-08  4.947e-09    6.679 4.58e-11 ***
## Mes12       -2.339e-08  4.933e-09   -4.742 2.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df      F p-value
## s(Año) 8.947  8.999 196755  <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 9.7546e-09  Scale est. = 9.4978e-09  n = 794
```

```
plot_gam_check(magCO2b)
```

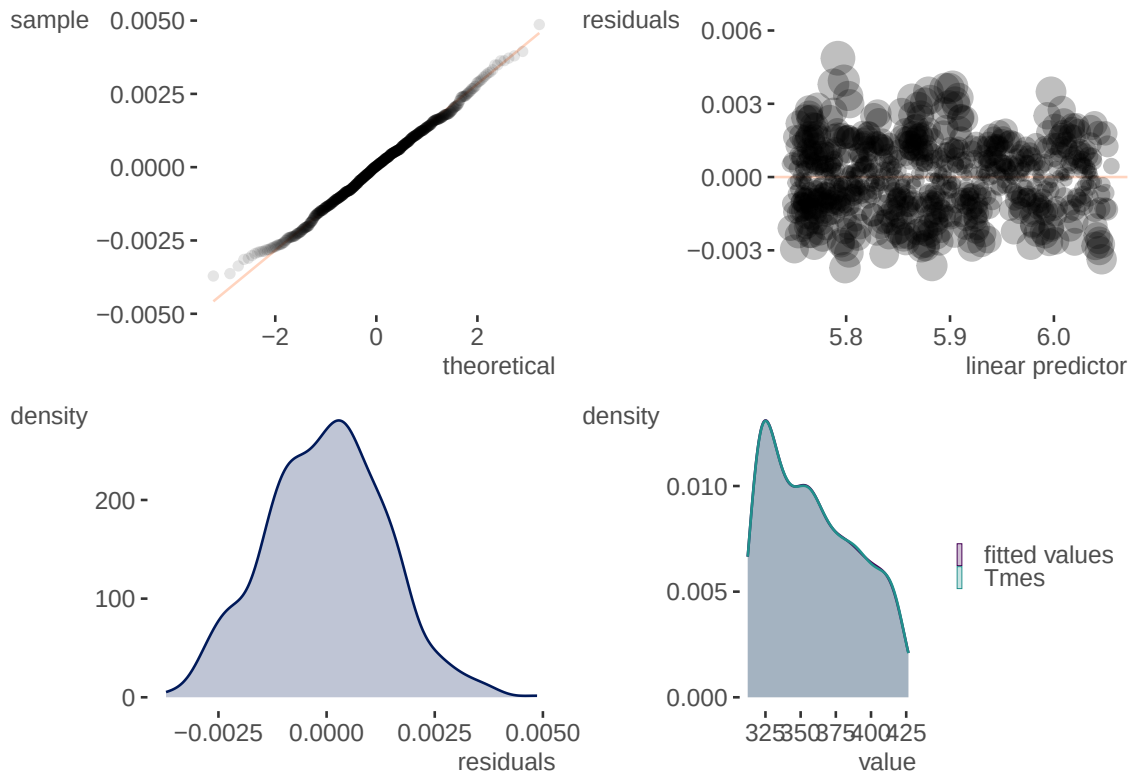



```
magCO2c <- gam(Tmes ~ s(Año) + Mes, data = CO2, family = Gamma(link = "log"))
summary(magCO2c)
```

```
##
## Family: Gamma
## Link function: log
##
## Formula:
## Tmes ~ s(Año) + Mes
##
## Parametric coefficients:
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  5.8777562  0.0001721 34160.995  < 2e-16 ***
## Mes2         0.0021993  0.0002432     9.042  < 2e-16 ***
## Mes3         0.0045205  0.0002424    18.652  < 2e-16 ***
## Mes4         0.0081229  0.0002424    33.516  < 2e-16 ***
## Mes5         0.0097221  0.0002433    39.956  < 2e-16 ***
## Mes6         0.0081689  0.0002433    33.572  < 2e-16 ***
```

```
## Mes7          0.0039223  0.0002433    16.120  < 2e-16 ***
## Mes8         -0.0017190  0.0002433    -7.065 3.59e-12 ***
## Mes9         -0.0061035  0.0002433   -25.084  < 2e-16 ***
## Mes10        -0.0059939  0.0002433   -24.634  < 2e-16 ***
## Mes11        -0.0022295  0.0002433    -9.163  < 2e-16 ***
## Mes12         0.0014633  0.0002433     6.014 2.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df      F p-value
## s(Año) 8.965      9 341186  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =      1   Deviance explained =  100%
## GCV = 2.0052e-06  Scale est. = 1.9524e-06  n = 794
```

```
plot_gam_check(magCO2c)
```



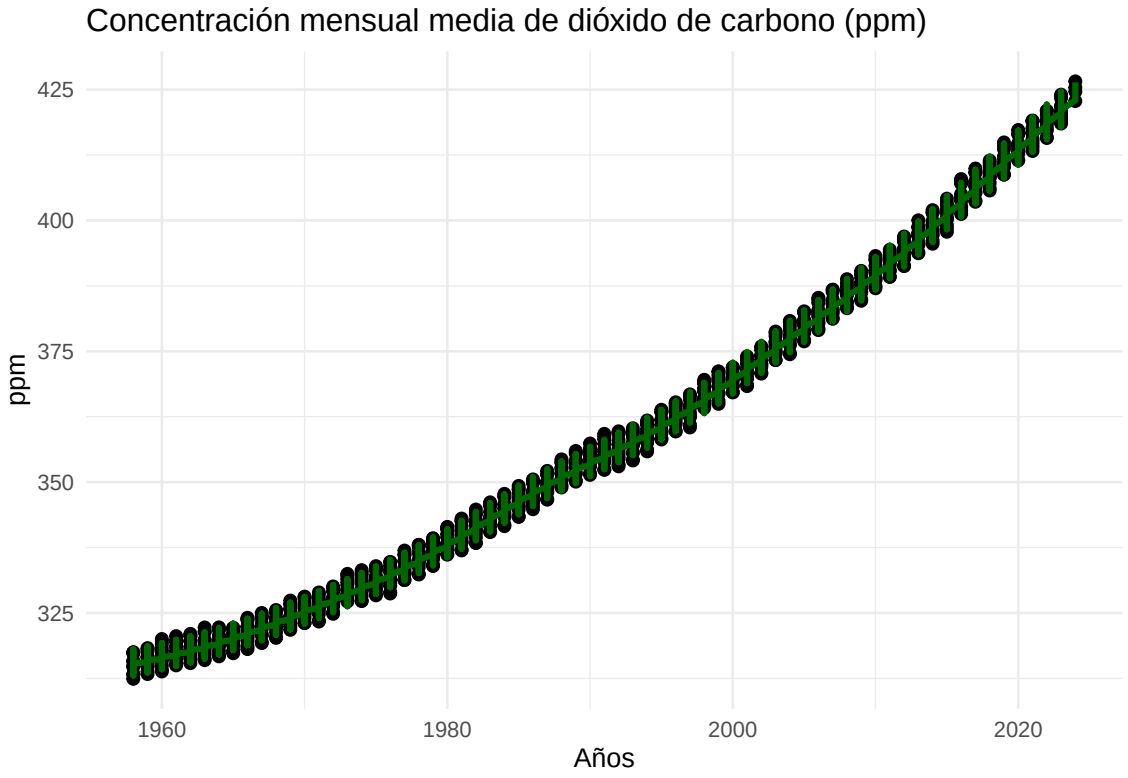
Vemos ahora como las gráficas para ambos diagnósticos ofrecen resultados mucho mejores y que incluso los R^2_{adj} llegan a 1 para los dos modelos. Comparémoslos:

```
##                      AIC          GCV
## Inv. Gauss (b) 1635.459 9.754596e-09
## Gamma (c)      1174.703 2.005172e-06
```

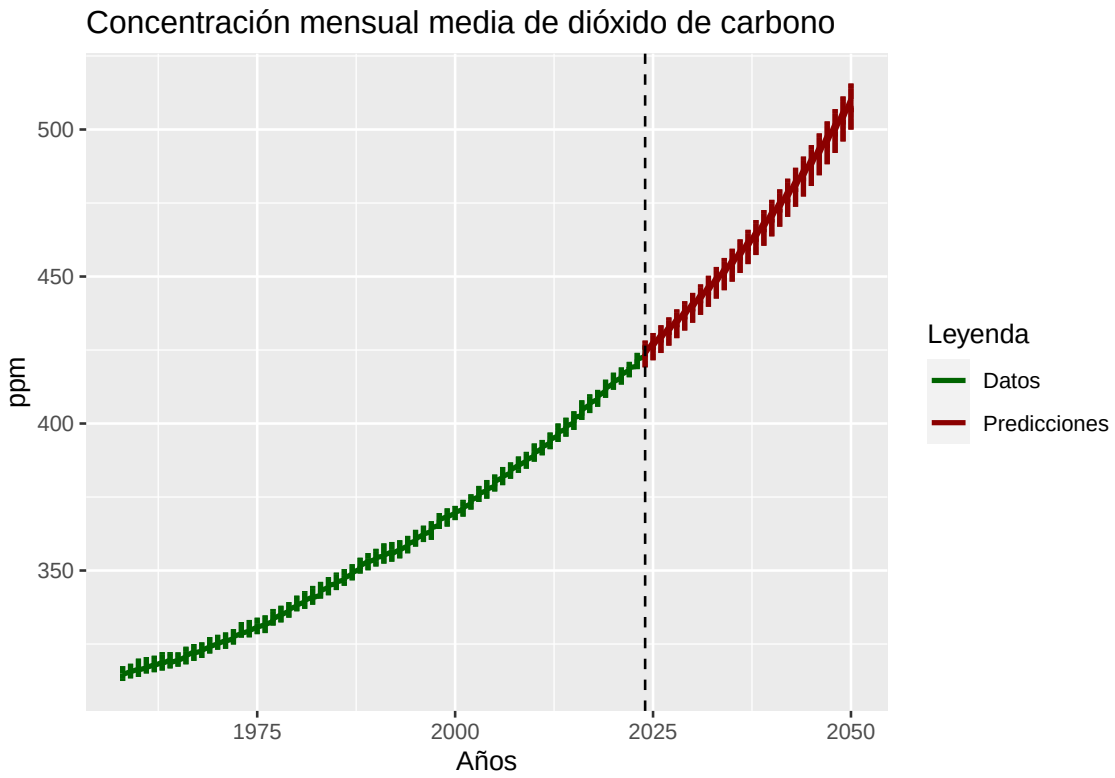
Por un lado, el modelo c tiene un AIC significativamente menor que el modelo b. Esto sugiere que el modelo c es mejor en términos de bondad del ajuste relativo, teniendo en cuenta la penalización de la complejidad. Por el otro lado, el modelo b tiene un GCV significativamente menor que el modelo c, por lo que tedrá una mejor capacidad predictiva y un mejor equilibrio entre ajuste y penalización por complejidad. Con cuál quedarnos ya dependerá del objetivo que nos propongamos, para dar un buen ajuste de los datos preferiremos el que utiliza la familia *Gamma* y para predecir nuevas observaciones preferiremos el modelo con la familia *Inverse Gaussian*. Apliquemos estos dos casos:

1.2.3. Visualización de los resultados

Ajuste de la concentración del dióxido de carbono
Como hemos indicado, el MAG que utiliza la distribución *Gamma* como familia de distribución exponencial proporciona una mejor bondad de ajuste teniendo en cuenta su complejidad, así que utilizaremos ese para representar el ajuste de los datos. Se debe tener en cuenta que sobre las predicciones se debe invertir la función de enlace utilizada, en el este caso la función logarítmica.



Predicciones de la concentración del dióxido de carbono
Para representar las predicciones utilizaremos el modelo que utilizaba la distribución *Inverse Gaussian* pues tenía un menor valor del error estimado por validación cruzada generalizada.

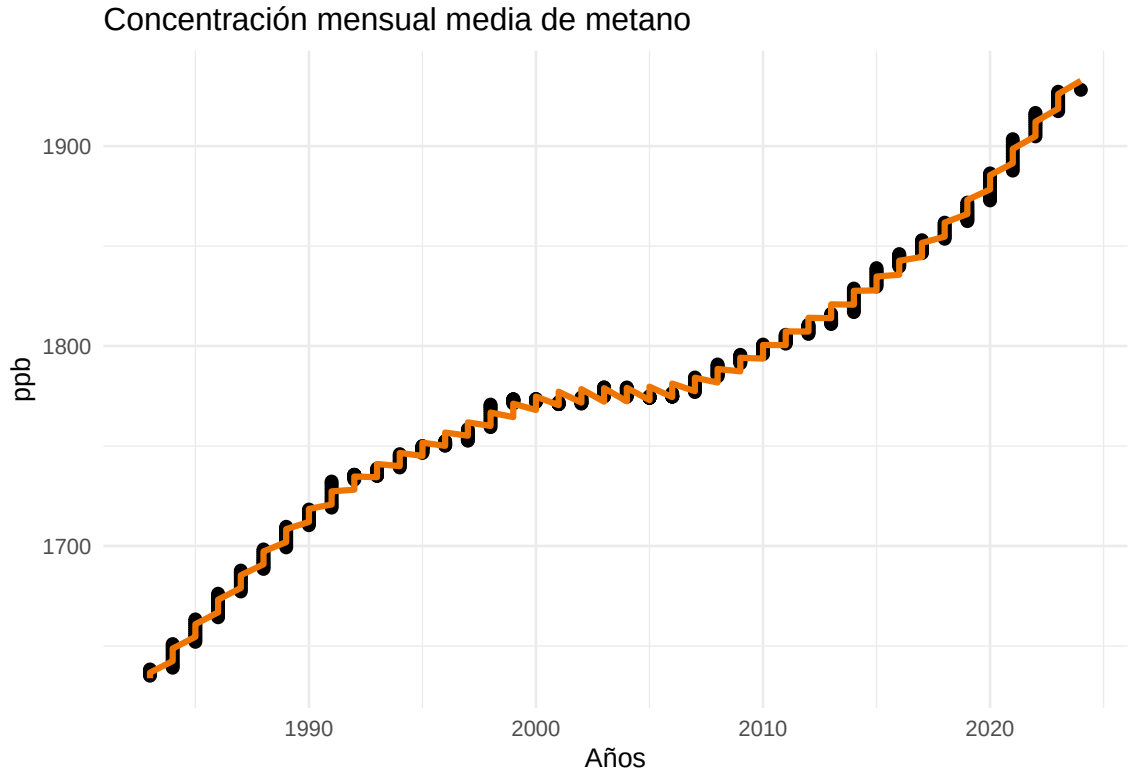


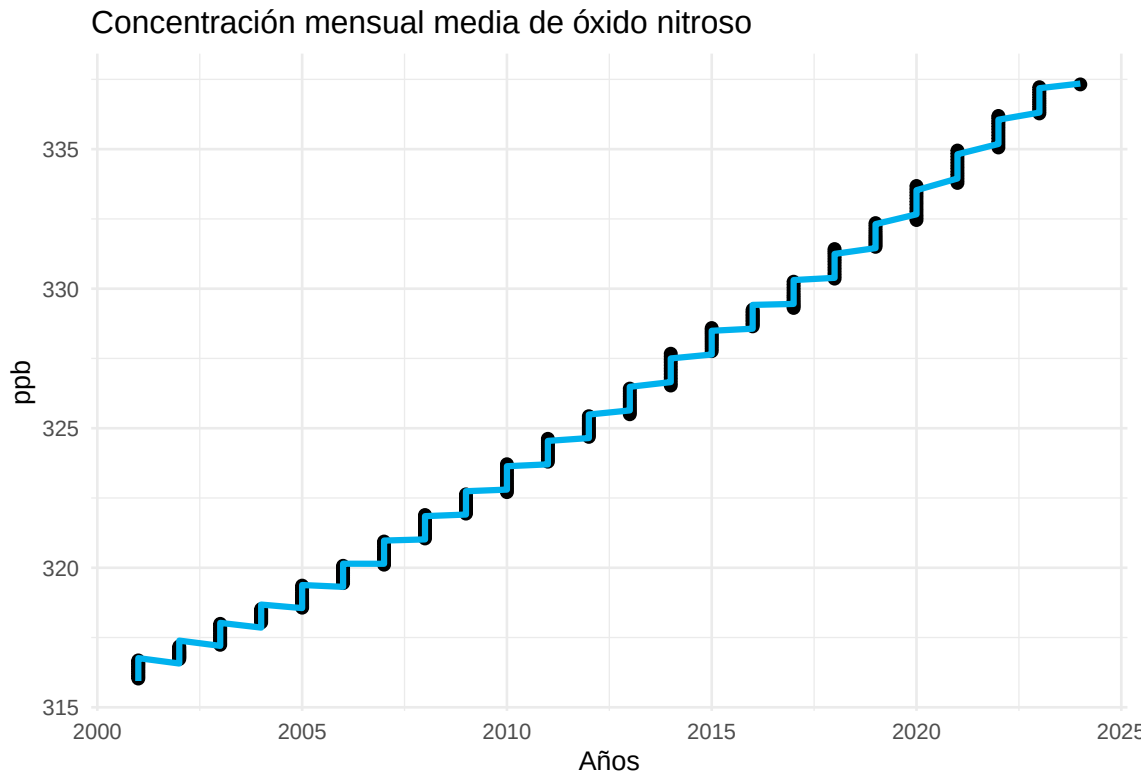

```
## List of 3
##  $ axis.title     :List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.text    :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : chr [1:2] "darkgreen" "darkred"
##   ..$ size         : num 10
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
```

```
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Obviamente esta es una predicción basta que solamente tiene en cuenta el paso del tiempo y los niveles de concentración medidos hasta ahora. Para poder hacer predicciones más exactas se necesitarían datos relativos a las emisiones de CO_2, al clima, al crecimiento de la población y crecimiento económico y se deberían tener en cuenta políticas sobre energías, economía y tecnología.

Seguimos el mismo razonamiento para los datos con las concentraciones de CH_4 y N_2O para obtener la siguientes representaciones de los ajustes para cada modelo:





Se puede apreciar en la primera gráfica que la media mensual de concentración atmosférica de metano ha aumentado en más de 200 ppb desde 1983 y que la correspondiente al N_2O ha aumentado en más de 20 ppb desde el 2001. Comparado con el incremento que se vio anteriormente para el CO_2 parece poco llamativo, pero como se comentó al principio de la sección, se debe tener en cuenta que estos dos gases captan mucha mas radiación que el primero, lo que hace que su aumento, por poco que sea, también genere preocupación respecto al cambio climático.

1.3. Modelización del aumento del nivel del mar

Ya comentamos en la sección **??** que la subida del nivel del mar dada en el último siglo es relevante comparada con la de cualquier siglo anterior. Como se argumenta en **?** , si este aumento tan pronunciado se progonla a medio-largo plazo pueden darse grandes consecuencias tales como el incremento de la frecuencia y la importancia de las inundaciones en zonas costeras, el aumento de amenazas por fenómenos climáticos extremos como huracanes y grandes tormentas, la erosión del suelo y las costas que puede implicar la pérdida del hábitat de peces, pájaros y plantas, etc. Esto conlleva a que esta sea una de las causas del cambio climático con mayor interés de estudio. Sin embargo, para poder realizar un análisis aceptable de los datos asociados a este hecho primero se debe distinguir si se quiere hacer a nivel global o a nivel local y se necesita disponer de información relacionada con múltiples factores como: la sanilidad, la geología del terreno, el deshielo de los polos, las oscilaciones oceánicas, eventos climáticos extremos que puedan ocurrir o hayan ocurrido...

En nuestro caso no disponemos de tanta información así que nos proponemos un objetivo más simple como es el definir un modelo aditivo generalizado que tenga como variable

de respuesta la media mensual global del nivel del mar (GMSL sus siglas en inglés) y como variables predictoras utilizaremos la media mensual de la temperatura de la superficie marítima global, la media de concentración atmosférica de CO_2 (la que utilizamos en la sección anterior como variable dependiente) y los datos temporales de meses y años.

1.3.1. Descripción de los datos

Como acabamos de comentar, las medidas de concentraciones de CO_2 que utilizaremos son las mismas que en la sección anterior así que no entraremos en más detalles para esos datos.

Con respecto a las medidas del GMSL utilizamos la misma fuente de datos que para los gases de efecto invernadero: la UNEP https://wesr.unep.org/climate/essential-climate-variables/sea-level-rise, lo único que debemos tener en cuenta es que se ha modificado el excel de cierta forma para generar las columnas *Year* y *Month*, correspondientes a la fecha en la que se obtuvo la medición, y la columna *GMSL* que hasta 1993 se corresponde con las medidas reconstruidas por la UNEP de la media global del nivel del mar en milímetros y a partir de ese año es la media de una serie de distintas medidas satelitales. La carga y limpieza de los datos es la siguiente:

```
SeaL <- read_excel('SeaLevelv2.xlsx')

# Retiramos datos con otro formato y nos quedamos con las variables necesarias.
SeaL <- SeaL[-(1:240),c(8,9,11)]
# La variable Año es numérica para la representación como ya se ha indicado otras v
# y la mensual es categórica
SeaL$Year <- as.numeric(SeaL$Year)
SeaL$Month <- factor(SeaL$Month,levels = 1:12)
colnames(SeaL) <- c('Año','Mes','GMSL')

# Como en algunos casos se tienen varias mediciones por mes lo que hacemos es tomar
# GMSL mensual la media de todas ellas.
SeaL <- SeaL %>%
  group_by(Año, Mes) %>%
  summarise(GMSL = mean(GMSL, na.rm = TRUE))
summary(SeaL)
```

```
##       Año             Mes           GMSL
##  Min.   :1900   1      :123   Min.   :-160.10
##  1st Qu.:1930   2      :123   1st Qu.:-123.20
##  Median :1961   3      :123   Median : -67.65
##  Mean   :1961   4      :123   Mean   : -63.39
##  3rd Qu.:1992   5      :123   3rd Qu.: -17.02
##  Max.   :2022   (Other):860   Max.   :  79.31
##  NA's   :1      NA's   :  1   NA's   :2
```

Por otro lado, para obtener los datos respectivos a la temperatura media global nos referiremos a los datos ofrecidos por la NASA en https://data.giss.nasa.gov/gistemp/. La

preparación de estos datos para poder ser manejados con comodidad es más compleja y se añade como un anexo.

```
##       Año            Mes             Temp
##  Min.   :1880   Min.   : 1.00   Min.   :-0.81000
##  1st Qu.:1916   1st Qu.: 3.75   1st Qu.:-0.22000
##  Median :1952   Median : 6.50   Median :-0.03000
##  Mean   :1952   Mean   : 6.50   Mean   : 0.07082
##  3rd Qu.:1988   3rd Qu.: 9.25   3rd Qu.: 0.29250
##  Max.   :2024   Max.   :12.00   Max.   : 1.48000
##  NA's   :12                     NA's   :20
```

Tras haber cargado los tres conjuntos de datos que necesitaremos para aplicar el modelo, los unimos en un solo data frame con observaciones desde 1959 hasta 2015:

```
SeaL <- (SeaL[(2015 > SeaL$Año) &(SeaL$Año >1958),])[1:672,]
SeaT <- (SeaT[(2015 > SeaT$Año) & (SeaT$Año > 1958),])[1:672,]
CO2 <-  CO2[(2015 > CO2$Año) & (CO2$Año > 1958),]

Sea <- cbind(SeaL,SeaT,CO2)
Sea <- Sea[,c(1,2,3,6,9)]
colnames(Sea) <- c('Año','Mes','GMSL','Temp','CO2')
str(Sea)
```

```
## tibble [672 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Año : num [1:672] 1959 1959 1959 1959 1959 ...
##  $ Mes : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ GMSL: num [1:672] -65.4 -68.7 -70.8 -70.1 -69.7 -68.3 -68 -65.6 -66.6 -65 ...
##  $ Temp: num [1:672] 0.08 0.07 0.18 0.16 0.04 0.03 0.03 -0.01 -0.06 -0.07 ...
##  $ CO2 : num [1:672] 316 316 317 318 318 ...
```

```
summary(Sea)
```

```
##       Año            Mes           GMSL              Temp
##  Min.   :1959   1      : 56   Min.   :-80.700   Min.   :-0.3500
##  1st Qu.:1973   2      : 56   1st Qu.:-48.675   1st Qu.: 0.0400
##  Median :1986   3      : 56   Median :-24.250   Median : 0.2650
##  Mean   :1986   4      : 56   Mean   :-22.486   Mean   : 0.2803
##  3rd Qu.:2000   5      : 56   3rd Qu.:  1.978   3rd Qu.: 0.5200
##  Max.   :2014   6      : 56   Max.   : 47.987   Max.   : 1.0200
##                 (Other):336
##       CO2
##  Min.   :313.3
##  1st Qu.:328.3
##  Median :348.7
##  Mean   :350.9
##  3rd Qu.:370.8
##  Max.   :402.0
##
```

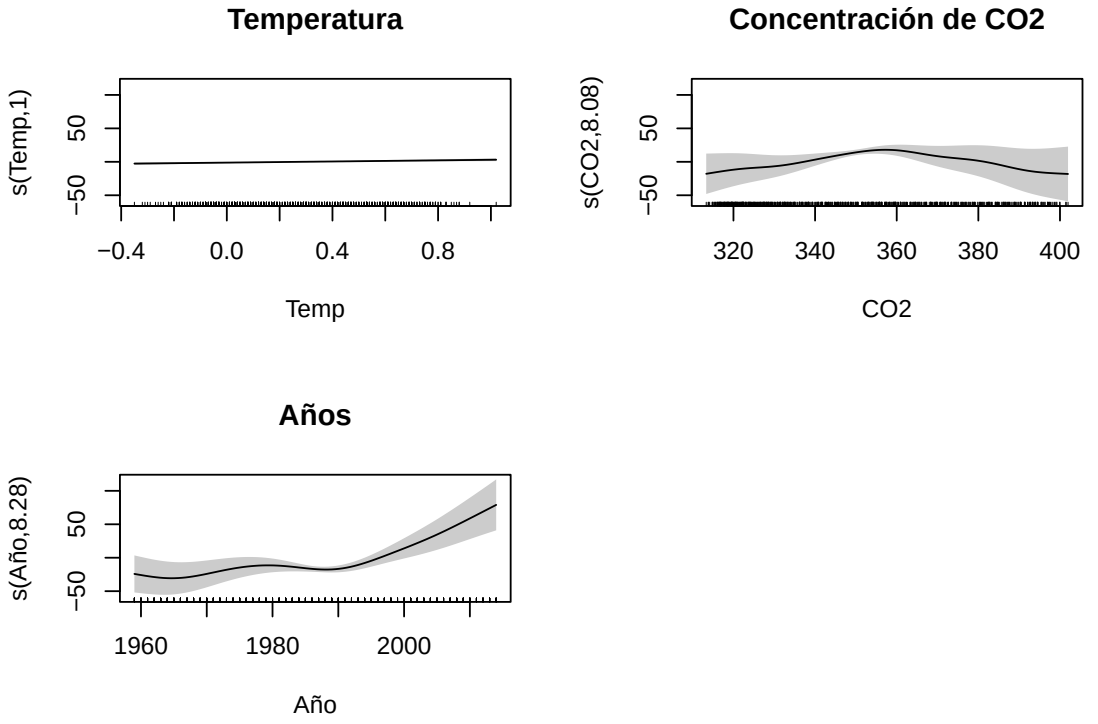
1.3.2. Descripción del modelo

Procedemos de forma similar a las secciones anteriores:

```
magSL <- gam(GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes, data = Sea)
summary(magSL)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.9481     0.6614 -34.695  < 2e-16 ***
## Mes2         -0.6542     0.9125  -0.717 0.473643
## Mes3         -1.6037     1.0743  -1.493 0.135988
## Mes4         -2.0515     1.4304  -1.434 0.152006
## Mes5         -2.3487     1.6201  -1.450 0.147629
## Mes6         -1.7603     1.4375  -1.225 0.221208
## Mes7         -0.9714     1.0356  -0.938 0.348583
## Mes8          0.8961     0.8931   1.003 0.316067
## Mes9          2.9220     1.2168   2.401 0.016616 *
## Mes10         4.0794     1.2121   3.366 0.000809 ***
## Mes11         3.9769     0.9256   4.297    2e-05 ***
## Mes12         3.0620     0.8820   3.472 0.000552 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F p-value
## s(Temp) 1.000  1.000  8.967 0.00285 **
## s(CO2)  8.083  8.768 10.085 < 2e-16 ***
## s(Año)  8.279  8.820 18.497 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.98   Deviance explained = 98.1%
## GCV = 21.716  Scale est. = 20.768    n = 672
```

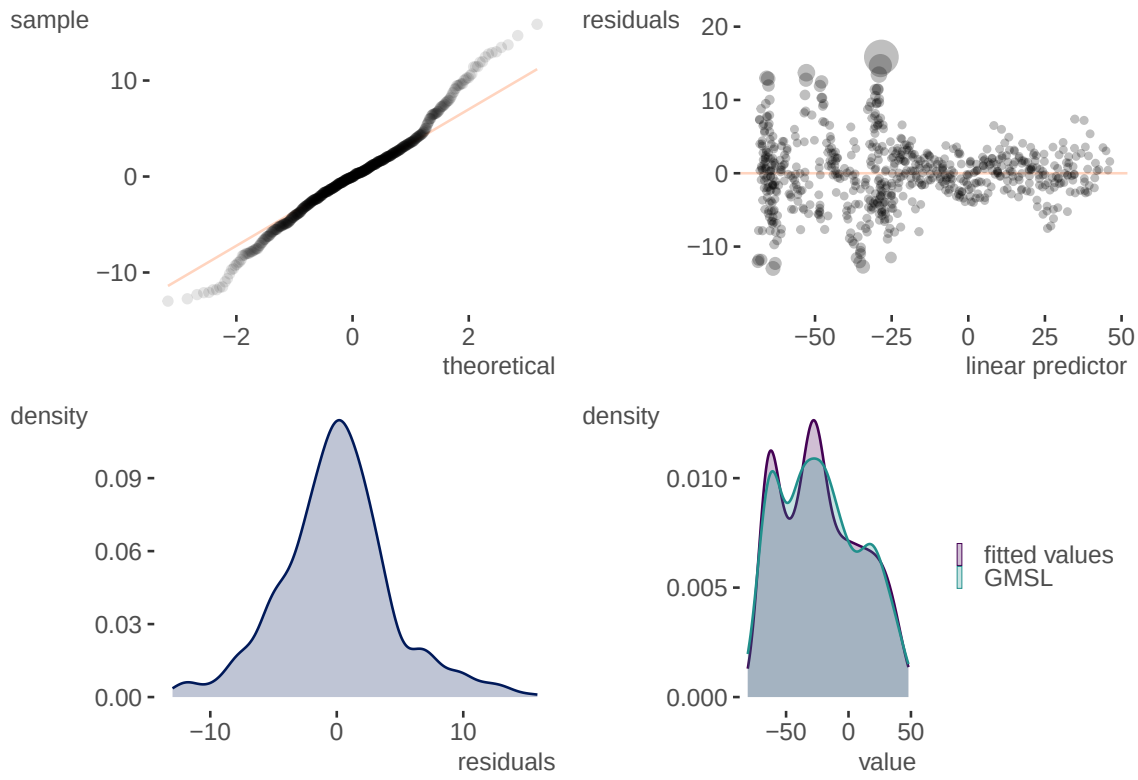
La desviación explicada por el modelo es del 98.1 % y el error estimado por validación cruzada generalizada es de 21.716.



Hay indicios de que la relación de la variable *Temp* tenga un efecto lineal sobre la variable de respuesta. Lo comprobamos como hicimos en **??**.

```
## Analysis of Deviance Table
##
## Model 1: GMSL ~ s(Temp) + s(CO2) + s(Año) + Mes
## Model 2: GMSL ~ Temp + s(CO2) + s(Año) + Mes
##   Resid. Df Resid. Dev          Df   Deviance      F    Pr(>F)
## 1    641.41      13346
## 2    641.41      13346 -1.7322e-05 -0.0016203 4.5043 8.293e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso como el p-valor es cercano a 0 se rechaza la hipótesis nula, por lo que se trata de modelos distintos. Por comodidad trabajaremos con el primero, partimos estudiando las gráficas de diagnóstico:

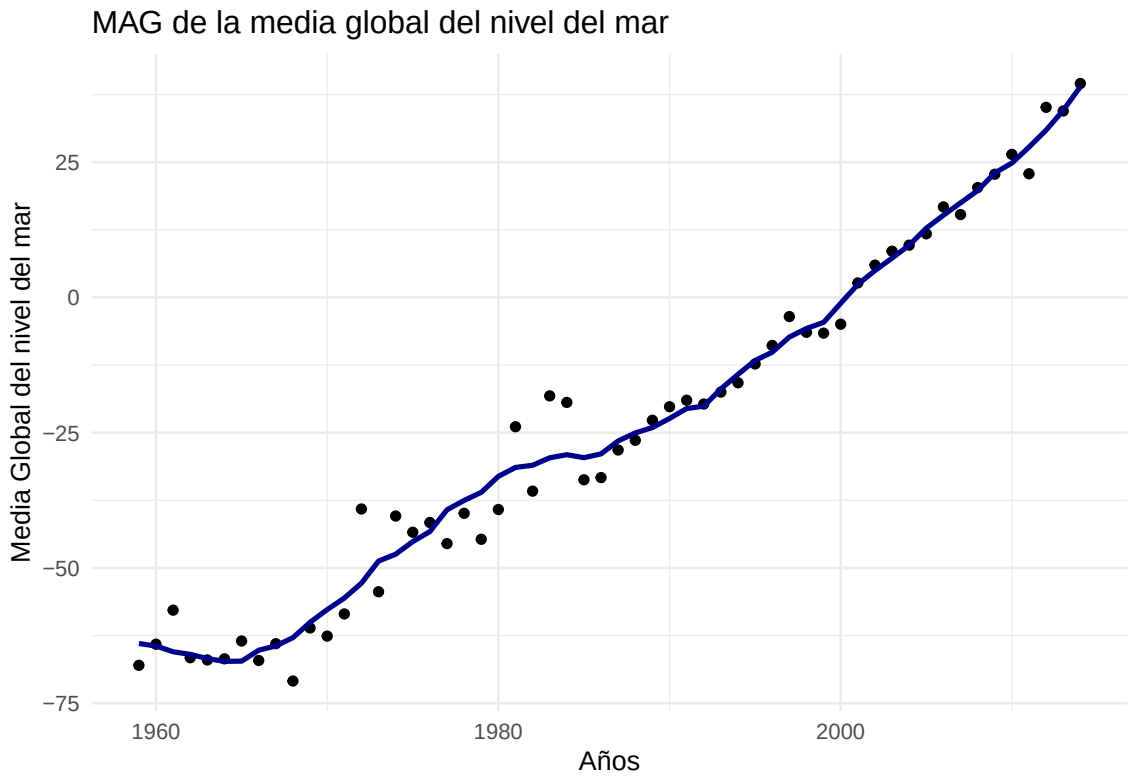


En este caso nos encontramos principalmente frente a dos problemas. Por una parte, en la gráfica Q-Q plot podemos observar que en los extremos los puntos no se ajustan correctamente a la recta, se puede interpretar que se tienen colas más pesadas, es decir, los valores extremos se alejan de seguir una distribución normal. Por otra parte, en la gráfica de arriba a la derecha se puede intuir un patrón en los residuos, lo que sugiere que no se verifique la hipótesis de varianza constante.

Representemos los datos ajustados por el modelo:

```
Julio <- Sea[Sea$Mes == 7,]
Julio$preds <- predict(magSL, newdata = Julio)

ggplot(data = Julio, aes(x = Año, y = GMSL)) +
  geom_point() +
  geom_line(aes(x = Año, y = preds), color = "darkblue", linewidth = 1) +
  labs(x = "Años", y = "Media Global del nivel del mar",
       title = "MAG de la media global del nivel del mar") +
  theme_minimal()
```



Es preferible ver cómo varía el nivel del mar fijando un mes que representando todos los datos a la vez pues entonces se tiene el ruido de las variaciones entre estaciones.