Tema 05.01: El Modelo de Regresión Lineal con residuos autocorrelados

@umh1465: Análisis estadístico de series económicas

Xavi Barber

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

2018-05-11

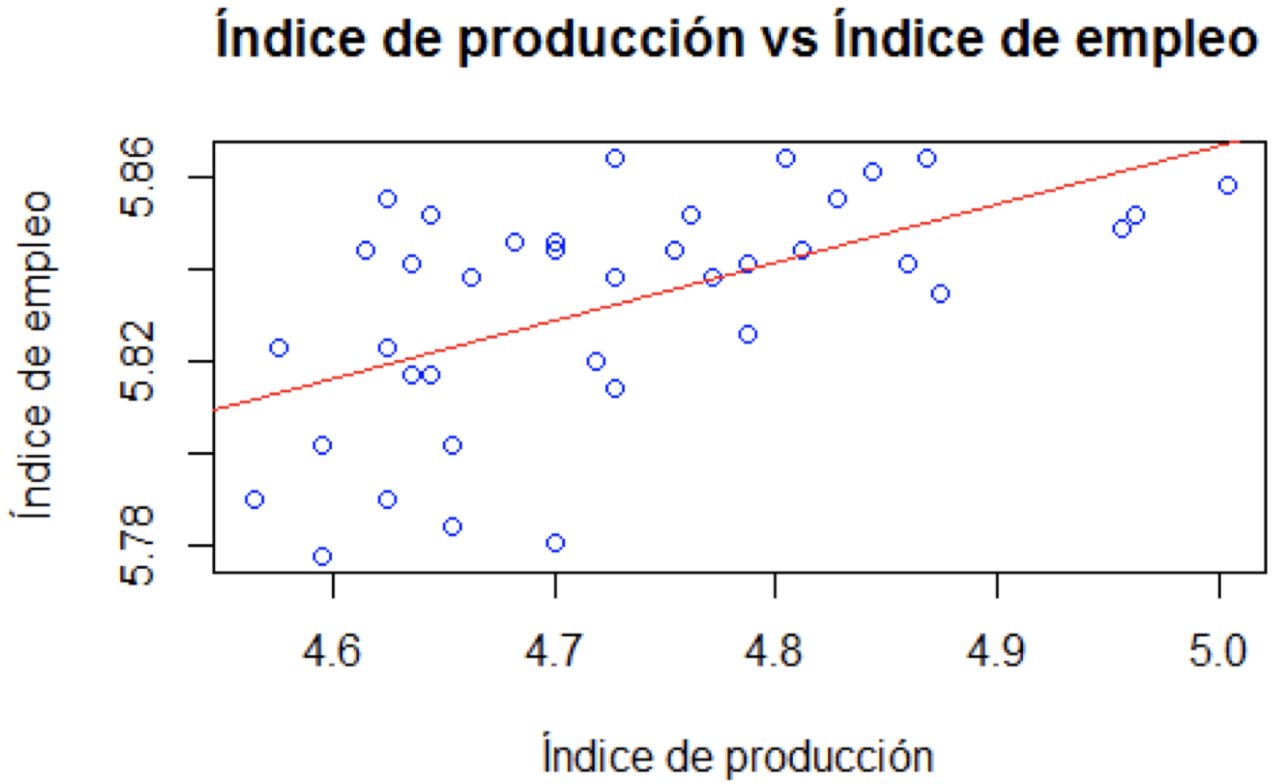
1. Análisis Estadístico de una un modelo
2. Algoritmo de ajuste de Modelo
3. Regresión con residuos autocorrelados en R
4. Modelo regresión y ARIMA

Análisis Estadístico de una un modelo

El modelo lineal

- Habitualmente se intenta modelizar la relación existente entre una variable "ndependiente" y diversas variables "explicativas" (independientemente de la naturaleza contínua o categórica de estas), mediante Modelos de Regresión.
- En concreto nos centraremos en el Modelo Lineal:

$$y_i = \beta_0 + \beta_1 \times x_1 + \cdots + \beta_p \times x_p + \varepsilon_i$$



1

[1] Ejemplo 9.8.1, Fuller,pág. 557-558

Figure 1

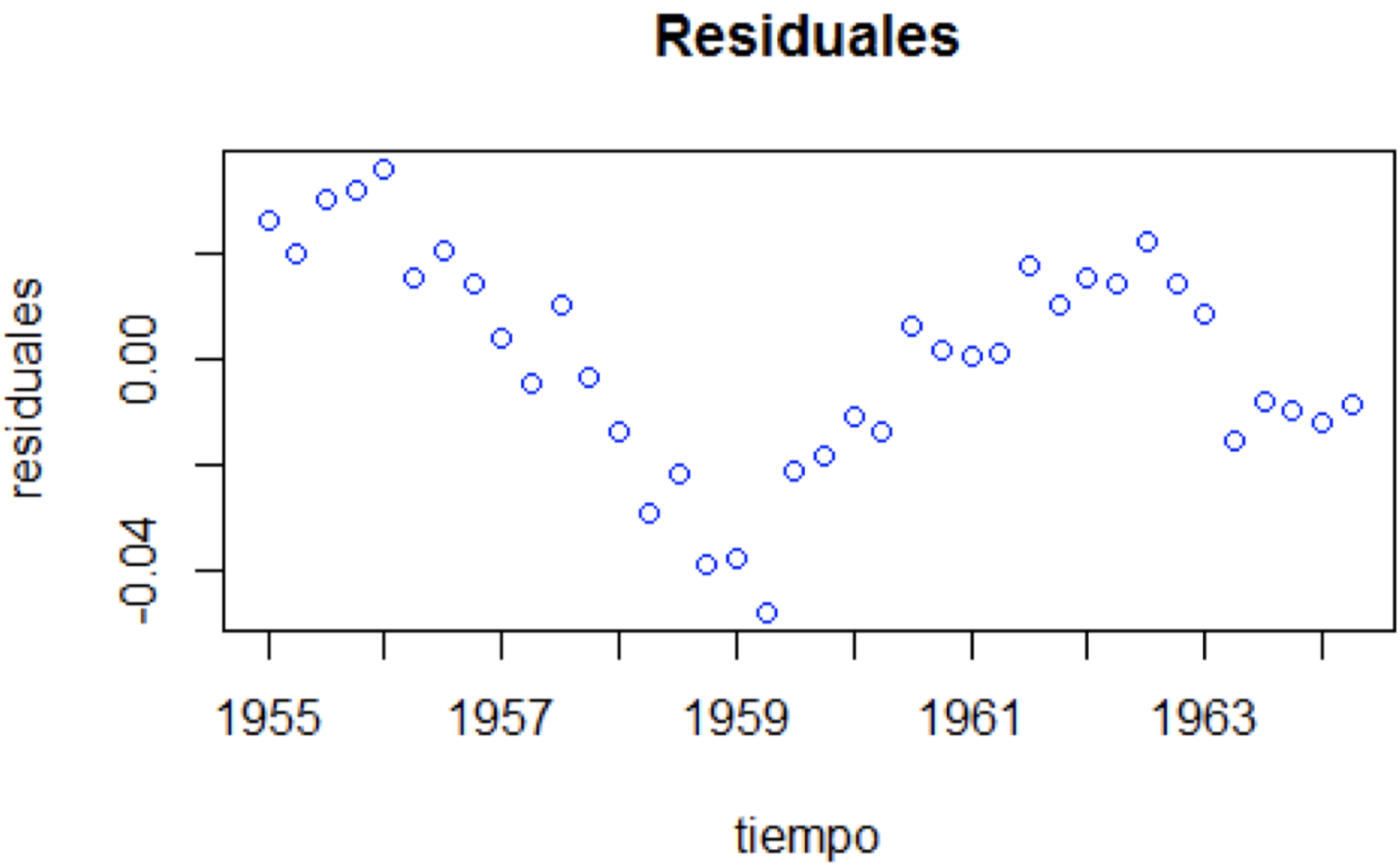

Residuales
residuales
0.00
-0.04
1955
1957
1959
1961
1963
tiempo

Figure 2

- Si consideramos el modelo lineal habitual donde lo serrores son independientes, y vista la gráfica anterior, nuestro resultados no serían **válidos**.
- Por lo que necesitaremos reparametrizar nuestro modelo, con el fin de encontrar una distribución para esos errores correlados, ε_i, que nos permita predecir de forma adecuada.

Residuos como $AR(p)$

- Cambiamos y_i por y_t por estar los errores correlados.

Si ε_t siguen un proceso $AR(p)$, entonces reescribimos los residuos como:

$$\phi(B)\varepsilon_t = w_t, w_t \sim \mathcal{WN}(0, \sigma^2) \quad .$$

De esta manera (y suponiendo que solo tenemos un variable "x")

$$\mu_t = \phi(B)y_y = \beta \times x_t + wt$$

Residuos como $ARMA(p, q)$

Si ε_t siguen un proceso $ARMA(p, q)$, entonces reescribimos los residuos como:

$$\phi(\theta)\varepsilon_t = \frac{\theta(B)}{\phi(B)} w_t \quad ,$$

donde B es el operador de ratardo que nos dará polinomios de orden p y q para el AR y el MA.

Por lo que (suponiendo una sola variable "x")

$$y_t = \beta_1 \times x_1 + \frac{\theta(B)}{\phi(B)} w_t$$

Algoritmo de ajuste de Modelo

Paso 1 y 2

Paso 1

Aplica una regresión ordinaria de y_t para obtener los residuos ε_t.

Paso 2

Ajustar un modelo ARMA a los residuos tal que

$$\hat{\varepsilon}_t = y_t - \hat{\beta}' x_t \times y_t$$

y

$$\hat{\phi}(B)\hat{\varepsilon}_t = \hat{\theta}(B) w_t$$

Paso 3

Aplicando la trasnformación *ARMA* a

$$y_t = \beta' x_t + \varepsilon_t \quad ,$$

se obtiene un nuevo modelo

$$u_t = \beta' v_t + w_t$$

Paso 4 y 5

Paso 4

Aplicamos mínimos cuadrados a este nuevo modelo u_t

Paso 5

Se buscará el mejor modelo: "bondadoso" y "válido" donde ya tengamos unos residuos con distribución normal, homocedásticos y de media cero.

Diversos tipos de ajuste

- Se pueden utilizar los Mínimos cuadrados ordinarios (OLS), si se cumplen todas las hipótesis de usuabilidad, es decir, que los resoduos tengan media cero, normalidad y homocedasticidad
- Cuando exite heterocedasticidad en los residuos, utilizaremos los *Generalized Least Squares*.
- Si la serie presenta "perturbaciones" se recomienda el uso de los *Estimated generalized Least Squares*.

Regresión con residuos autocorrelados en R

Modelo Lineal

```
library(car)
data(Hartnagel)
```

Se trata de un conjunto de datos con 38 observaciones y 7 variables. Los datos corresponden a las tasas de delincuencia desde 1931 a 1968.

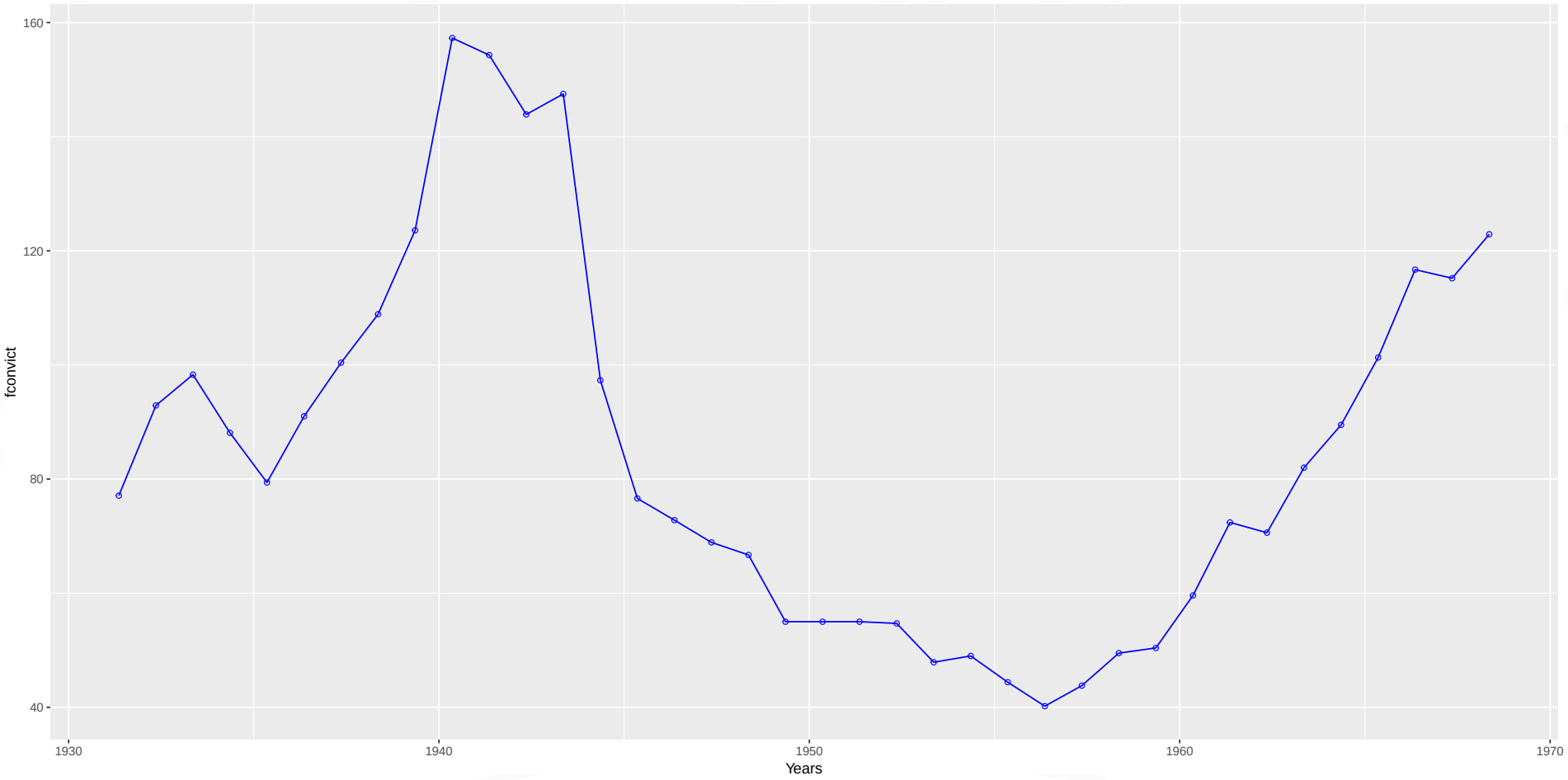
Con variables X relacionadas con delitos penales tanto del hombre como de la mujer y nivel de estudios de la mujer.

year	tfr	partic	degrees	fconvict	ftheft	mconvict	mtheft
1931	3200	234	12.4	77.1	NA	778.7	NA
1932	3084	234	12.9	92.9	NA	745.7	NA
1933	2864	235	13.9	98.3	NA	768.3	NA
1934	2803	237	13.6	88.1	NA	733.6	NA
1935	2755	238	13.2	79.4	20.4	765.7	247.1
1936	2696	240	13.2	91.0	22.1	816.5	254.9

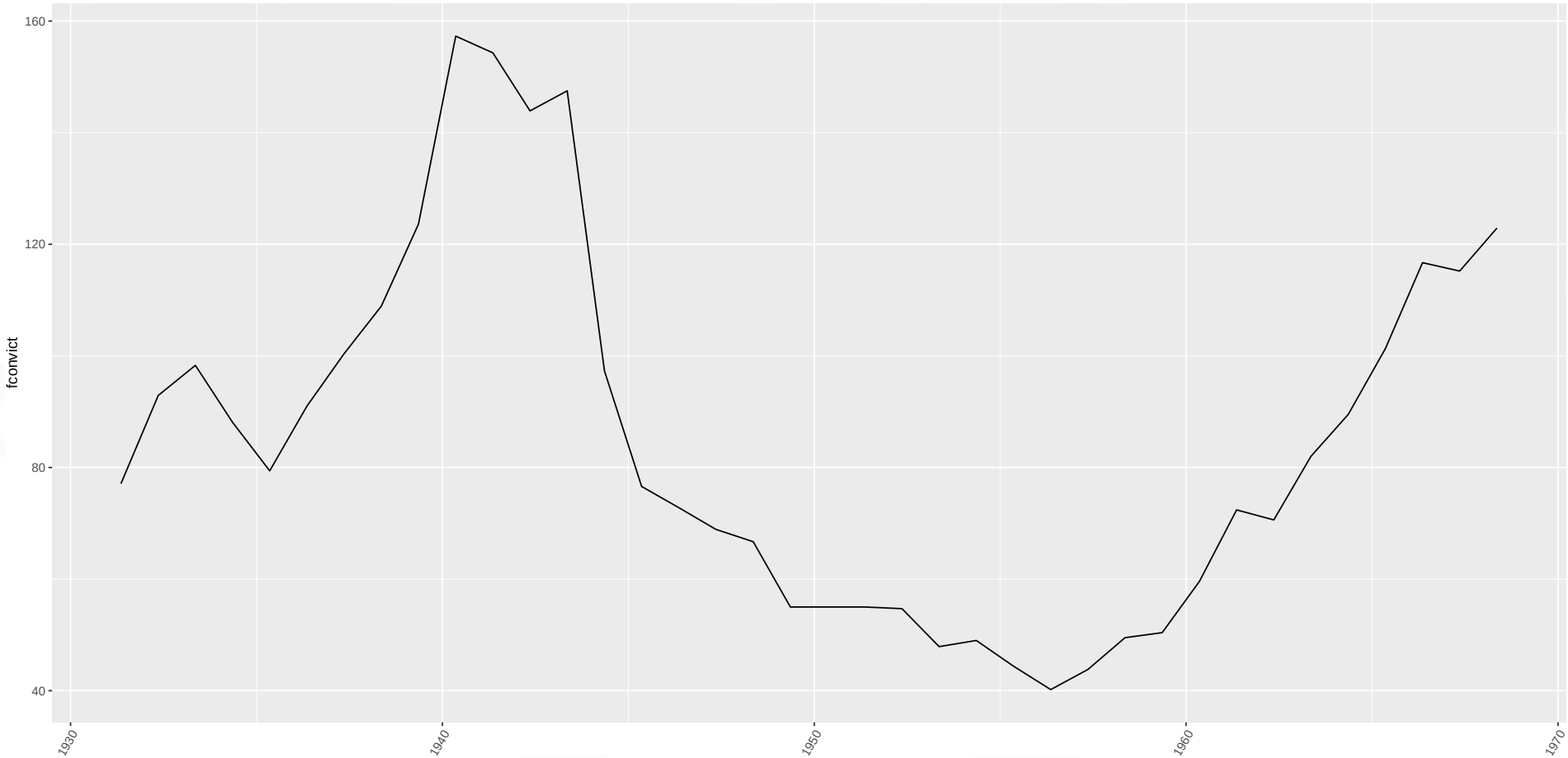
Deseamos realizar un modelo que ajuste *fconvict* con *tfr* ,*partic* , *degrees* y *mconvict*

Gráficamente

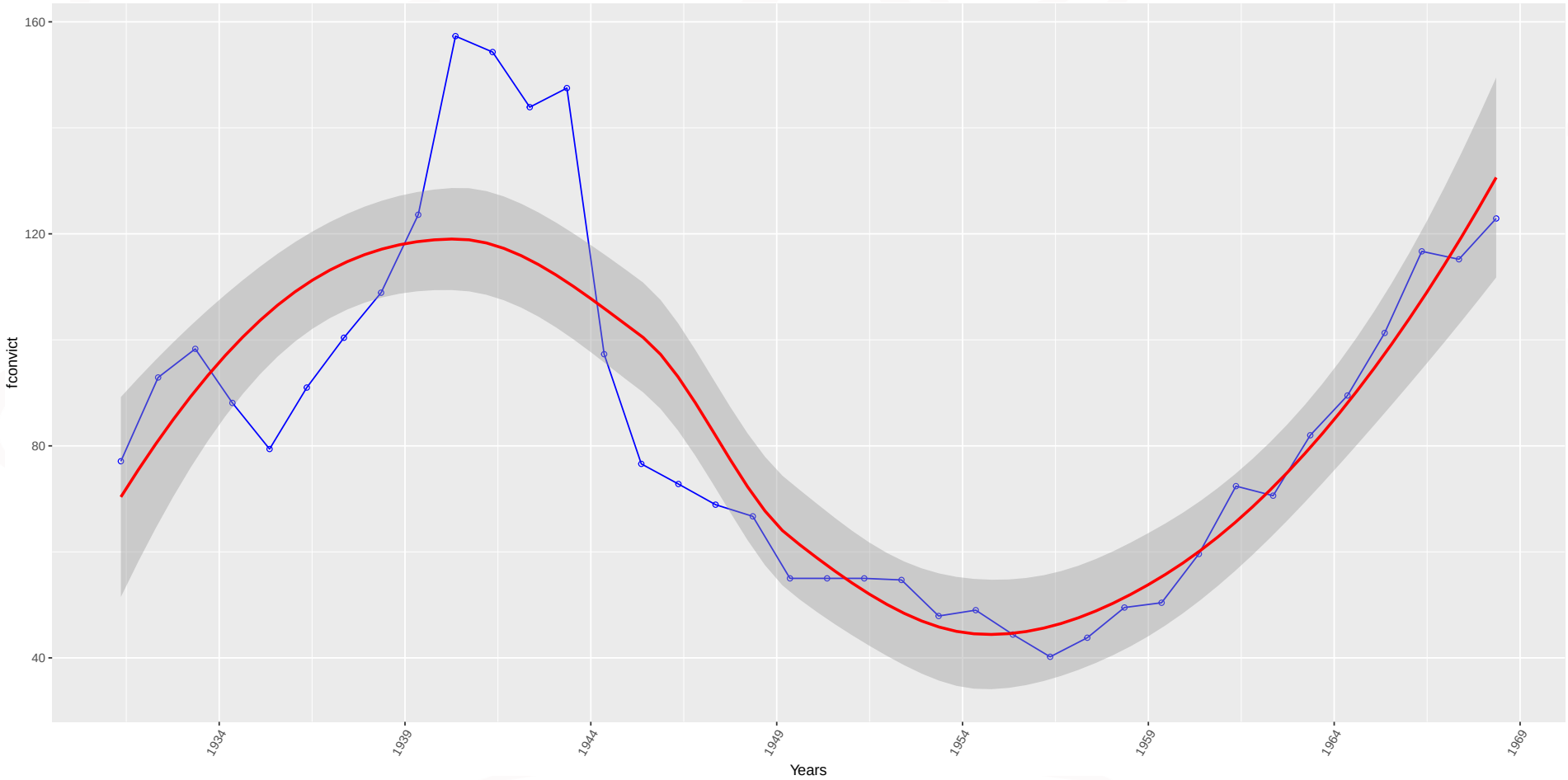
```
Hartnagel$year2<-as.Date(as.character(Hartnagel$year), "%Y")

p<-ggplot(Hartnagel, aes(x=year2, y=fconvict)) +
                  geom_line(col="blue")          +
                  geom_point(col="blue", pch=1) +
                   xlab("Years")
p
```

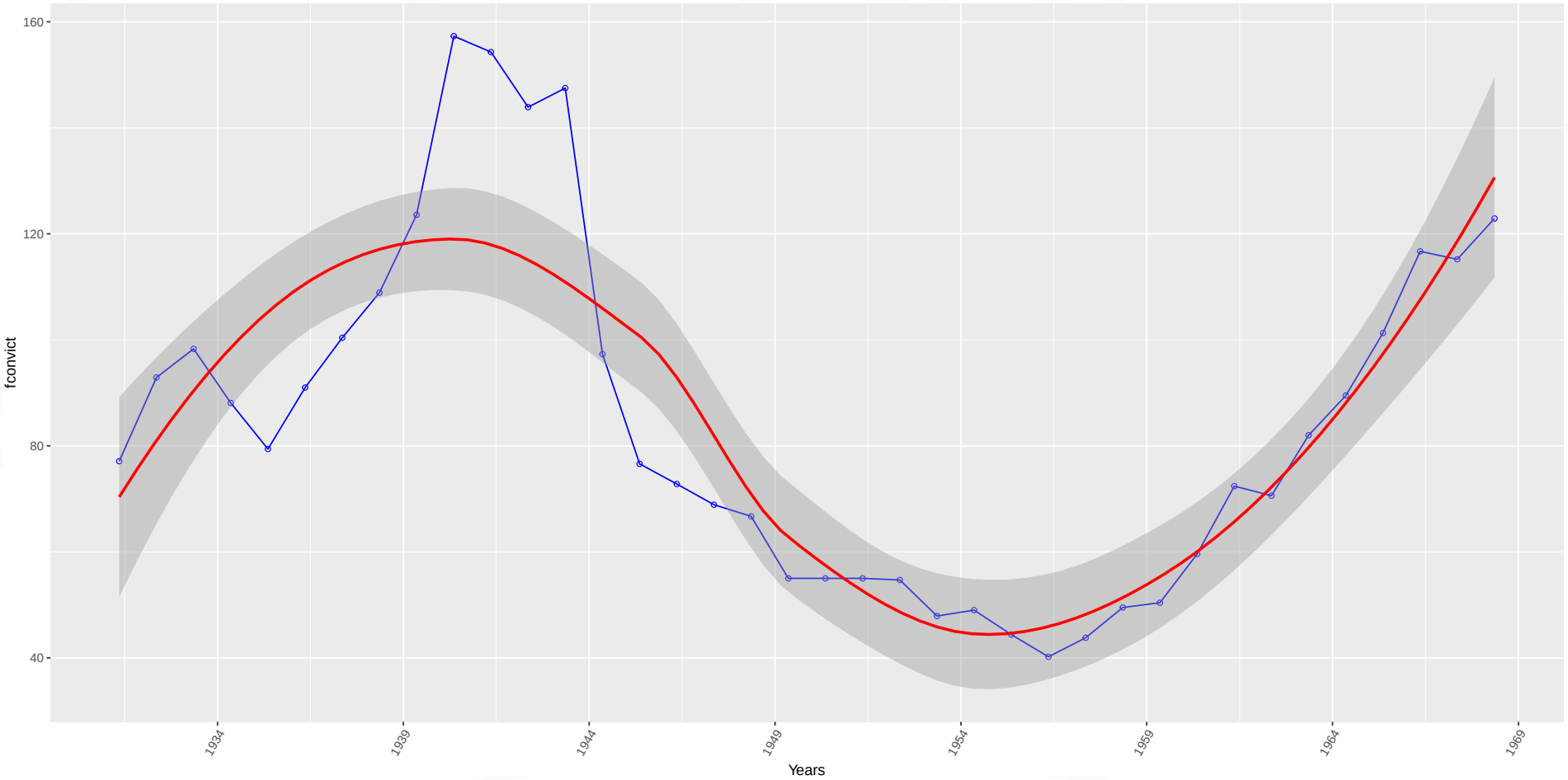


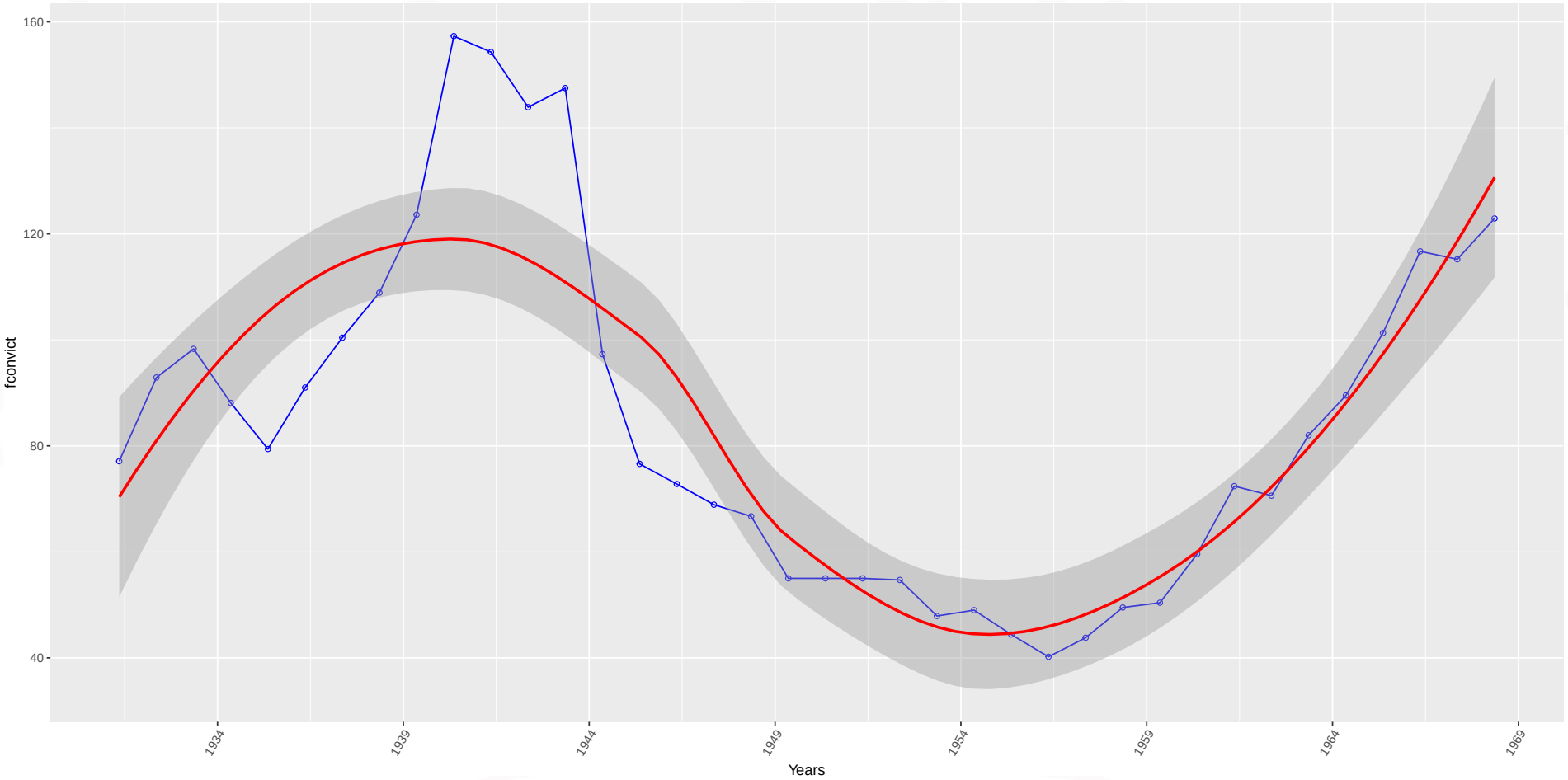
```
Hartnagel$year2 <- as.Date(as.character(Hartnagel$year), "%Y")
p <- ggplot(Hartnagel, aes(x = year2, y = fconvict)) + geom_line() +
    xlab("")
p + theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



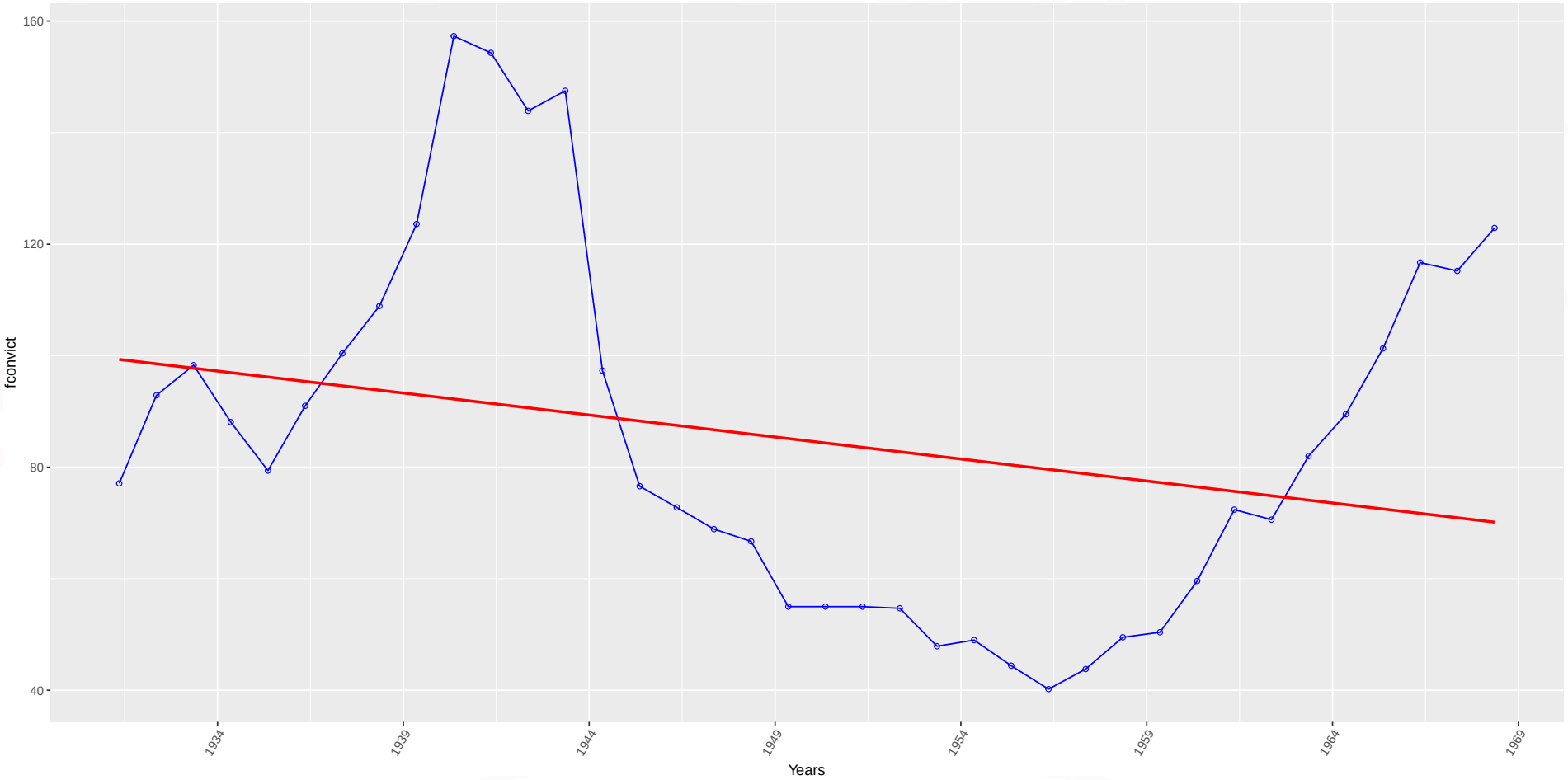
```
p <- ggplot(Hartnagel, aes(x = year2, y = fconvict)) + geom_line(col = "blue") +
    geom_point(col = "blue", pch = 1) + xlab("Years")
p <- p + scale_x_date(date_breaks = "5 year", date_labels = "%Y")
p + theme(axis.text.x = element_text(angle = 60, hjust = 1))
```







```
p <- ggplot(Hartnagel, aes(x = year2, y = fconvict)) + geom_line(col = "blue") +
    geom_point(col = "blue", pch = 1) + xlab("Years") + geom_smooth(method = lm,
    se = FALSE, col = "red")
p <- p + scale_x_date(date_breaks = "5 year", date_labels = "%Y")
p + theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Mínimos Cuadrados Ordinalios (OLS)

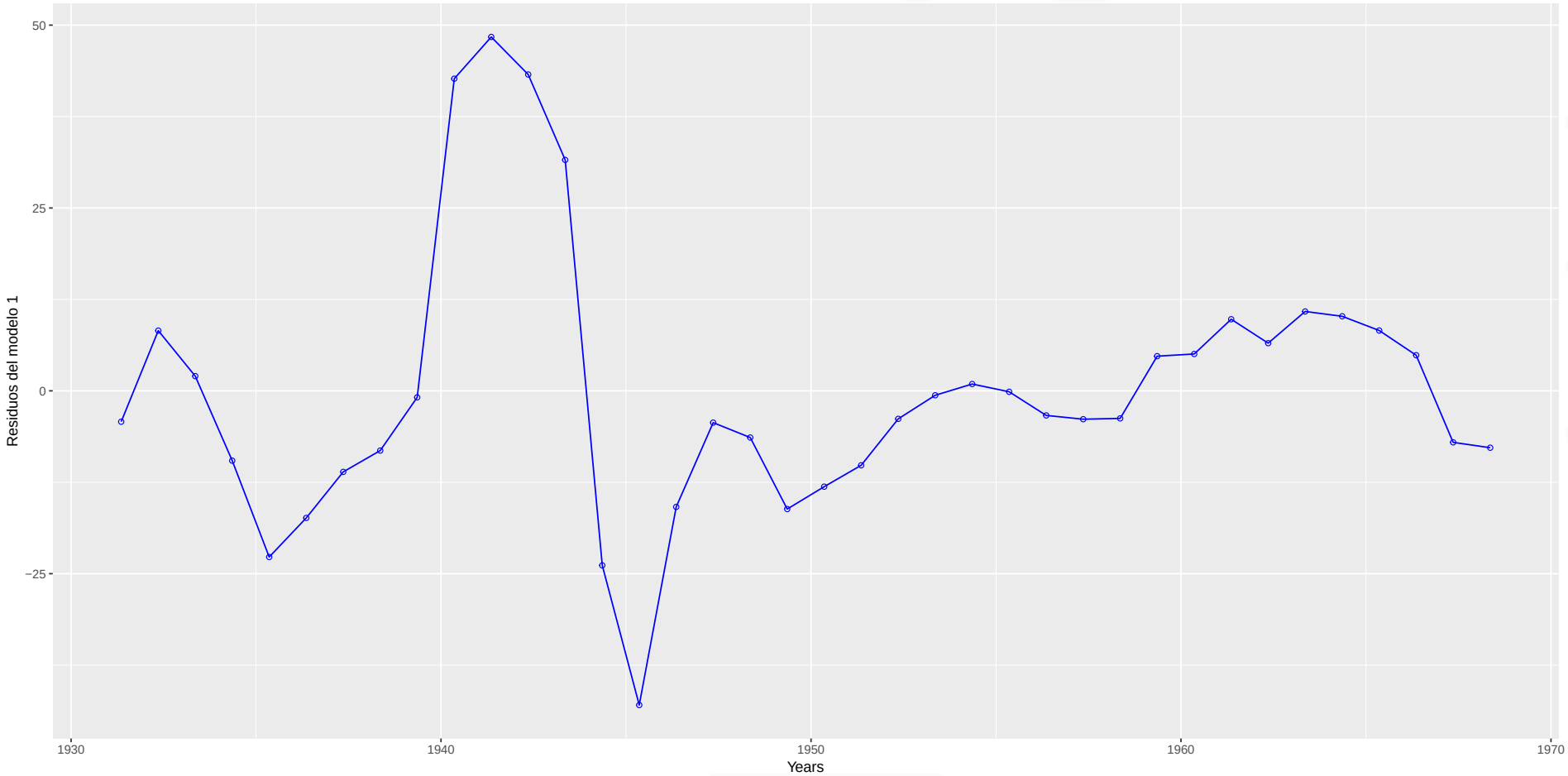
```
fit1<-lm( fconvict ~ tfr + partic +
            degrees + mconvict, data=Hartnagel)
```

Table 2

	Dependent variable:
	fconvict
tfr	-0.047^{***}
	(0.008)
partic	0.253^{**}
	(0.115)
degrees	-0.212
	(0.211)
mconvict	0.059
	(0.045)
Constant	127.640^{**}
	(59.957)
Observations	38
R^2	0.695

Residuos por año

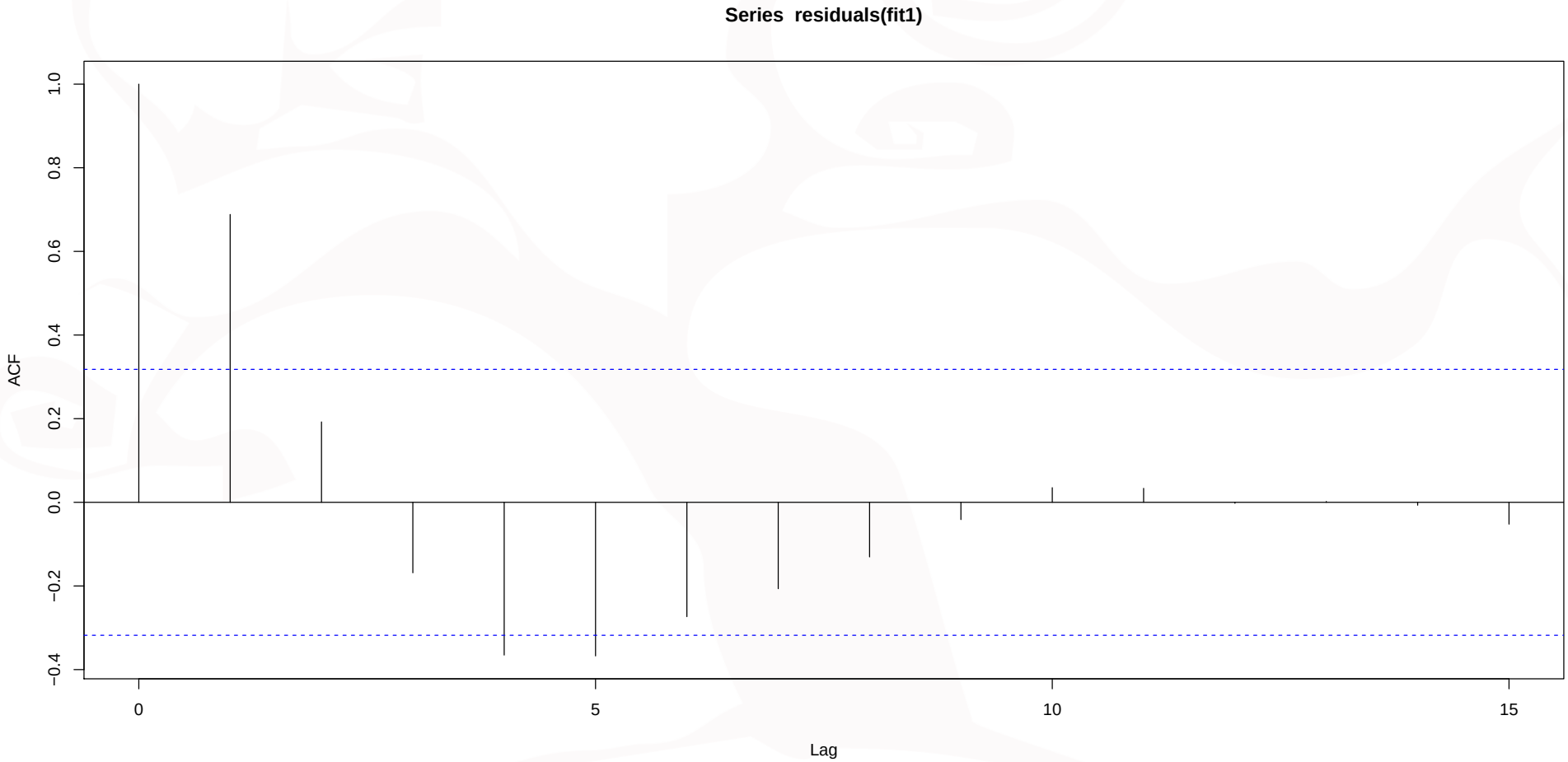
```
ggplot(Hartnagel, aes(x = year2, y = fit1$residuals)) + geom_line(col = "blue") +
    geom_point(col = "blue", pch = 1) + xlab("Years") + geom_hline(yintercept = 0)
```

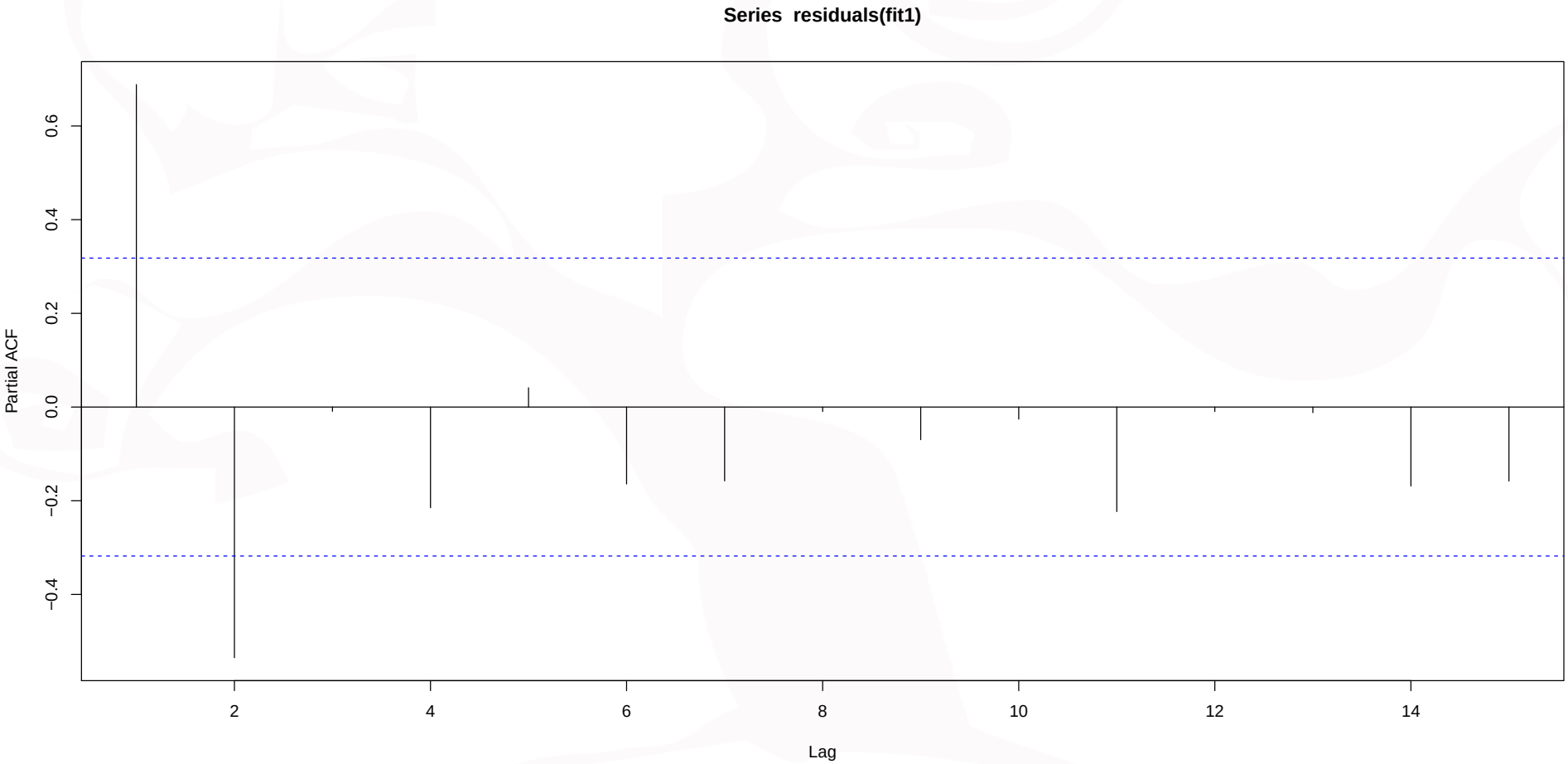



```
mapping: yintercept = yintercept
geom_hline: na.rm = FALSE
stat_identity: na.rm = FALSE
position_identity
```

¿Los residuos estas correlados?

```
library(tseries)
acf(residuals(fit1))
acf(residuals(fit1), type = "partial")
```





Test de Durbin Watson

```
durbinWatsonTest(fit1, max.lag = 5)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.6883450     0.6168636   0.000
   2       0.1922665     1.5993563   0.124
   3      -0.1685699     2.3187448   0.300
   4      -0.3652775     2.6990538   0.010
   5      -0.3673240     2.6521103   0.008
 Alternative hypothesis: rho[lag] != 0
```

Dado que existe una clara autocorrelación debemos de probar otros métodos de ajuste.

Mínimos cuadrados generalizados (GLS)

```
library(nlme)
mod.gls <- gls(fconvict ~ tfr + partic +
               degrees + mconvict, data=Hartnagel,
               correlation=corARMA(p=2), method='ML')
```

Le vamos a indicar que necesitamos un método robusto para el ajuste y que según los ACF-PACF pensamos que es un $AR(2)$.

```
- Generalized least squares fit by maximum likelihood
-   Model: fconvict ~ tfr + partic + degrees + mconvict
-   Data: Hartnagel
-        AIC      BIC    logLik
-   305.4145 318.5152 -144.7073
-
- Correlation Structure: ARMA(2,0)
-  Formula: ~1
-  Parameter estimate(s):
-       Phi1       Phi2
-  1.0683473 -0.5507269
-
- Coefficients:
-                Value Std.Error   t-value p-value
- (Intercept) 83.34028  59.47084  1.401364  0.1704
- tfr         -0.03999   0.00928 -4.308632  0.0001
- partic       0.28761   0.11201  2.567653  0.0150
- degrees     -0.20984   0.20658 -1.015757  0.3171
- mconvict     0.07569   0.03501  2.161899  0.0380
-
-  Correlation:
-          (Intr) tfr    partic degres
- tfr      -0.773
- partic   -0.570  0.176
- degrees   0.093  0.033 -0.476
- mconvict -0.689  0.365  0.047  0.082
-
- Standardized residuals:
-        Min         Q1        Med         Q3        Max
- -2.4991516 -0.3716988 -0.1494540  0.3372409  2.9094711
-
- Residual standard error: 17.70228
- Degrees of freedom: 38 total; 33 residual
```

¿es este modelo necesario?

```
mod.gls.3 <- update(mod.gls, correlation = corARMA(p = 3))
mod.gls.1 <- update(mod.gls, correlation = corARMA(p = 1))
mod.gls.0 <- update(mod.gls, correlation = NULL)
anova(mod.gls, mod.gls.1)
```

```
r  anova(mod.gls, mod.gls.1)

Model df      AIC      BIC    logLik   Test  L.Ratio p-value    mod.gls       1  8 305.4145
318.5152 -144.7073                                 mod.gls.1     2  7 312.4234 323.8865 -149.2117
1 vs 2 9.008881  0.0027
```

¿Es mejor el *mod.gls* o el *mod.gls*?

```
anova(mod.gls, mod.gls.0)
```

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.gls       1  8 305.4145 318.5152 -144.7073
mod.gls.0     2  6 339.0011 348.8266 -163.5006 1 vs 2 37.58658  <.0001
```

¿Es mejor el *mod.gls* o el *mod.gls.0*?

```
anova(mod.gls.3, mod.gls)
```

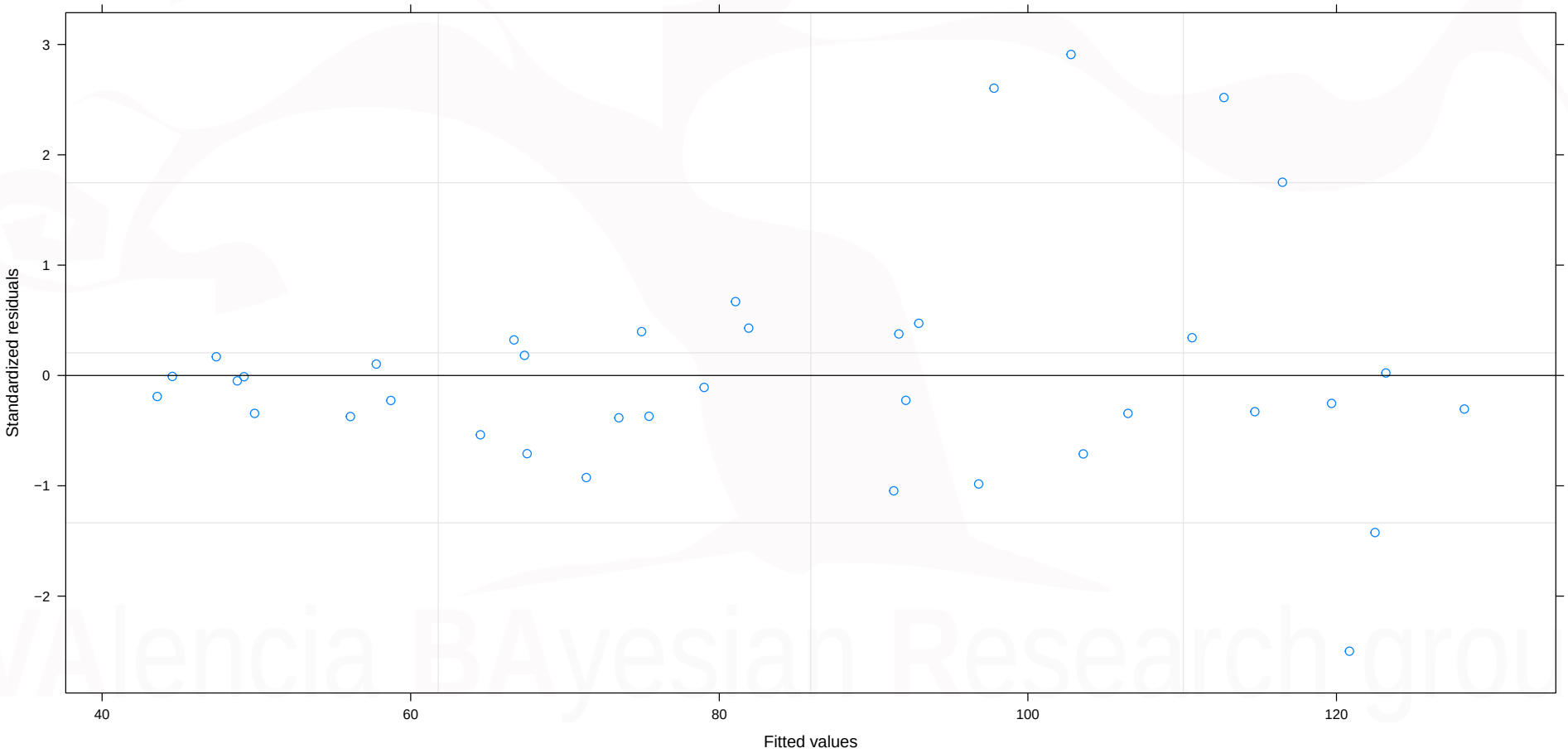
```
          Model df      AIC      BIC    logLik   Test    L.Ratio p-value
mod.gls.3     1  9 307.3961 322.1343 -144.6980
mod.gls       2  8 305.4145 318.5152 -144.7073 1 vs 2 0.01846713  0.8919
```

¿Es mejor el *mod.gls* o el *mod.gls.3*?

Un AR(3), no mejora al modelo con AR(2)

Residuos del GLS

```
plot(mod.gls)
```



Predicción

Vamos a predecir para las mujeres la tasa de criminalidad:

year	tfr	partic	degrees	fconvict	ftheft	mconvict
	2755	238	13.2	???	20.4	765.7

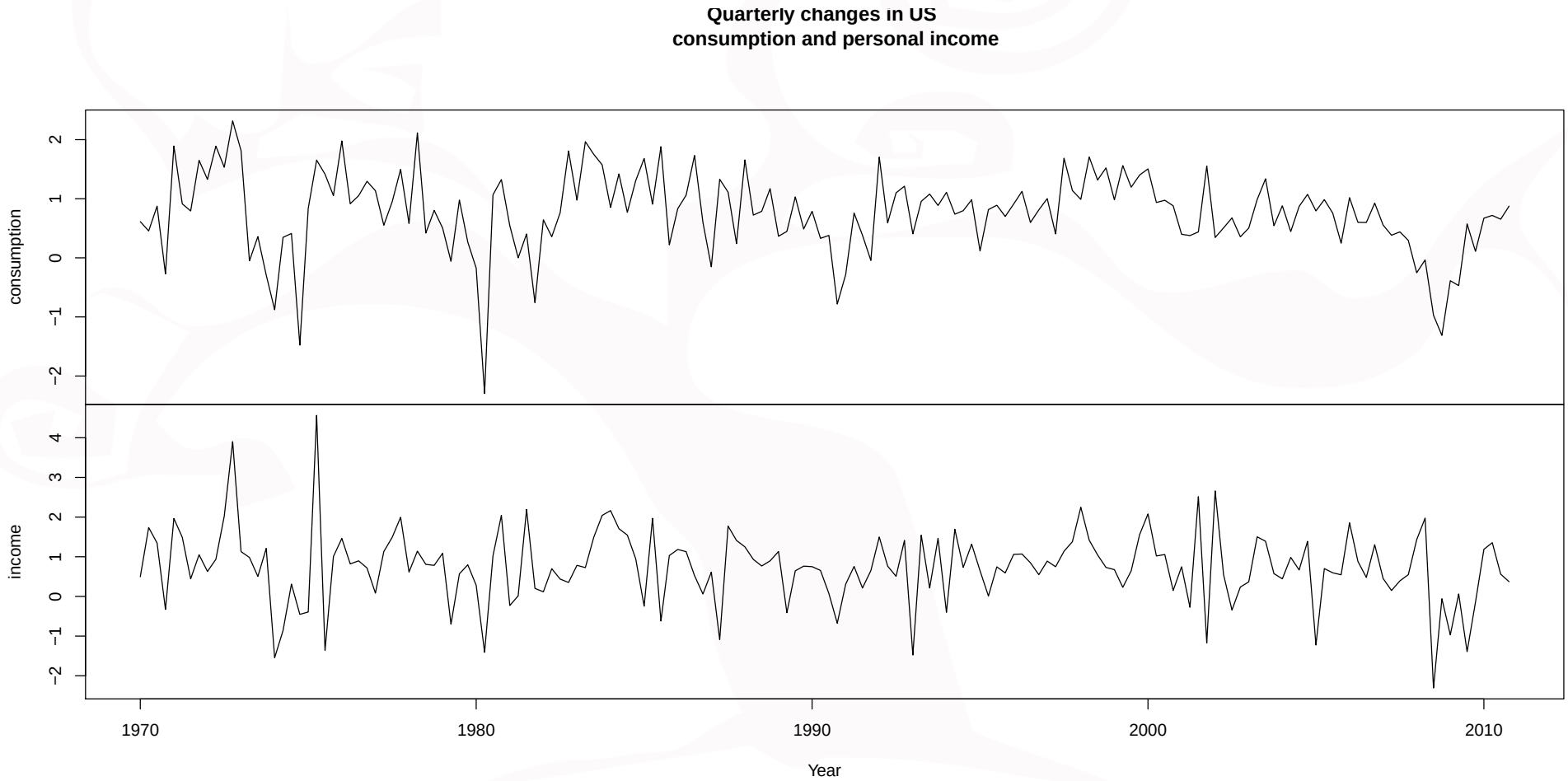
```
new.data<-data.frame( tfr=2755, partic=238,
                      degrees=13.2,
                      ftheft=20.4,
                      mconvict= 765.7 )
predict( mod.gls, new.data)
```

```
  [1] 96.81063
  attr(,"label")
  [1] "Predicted values"
```

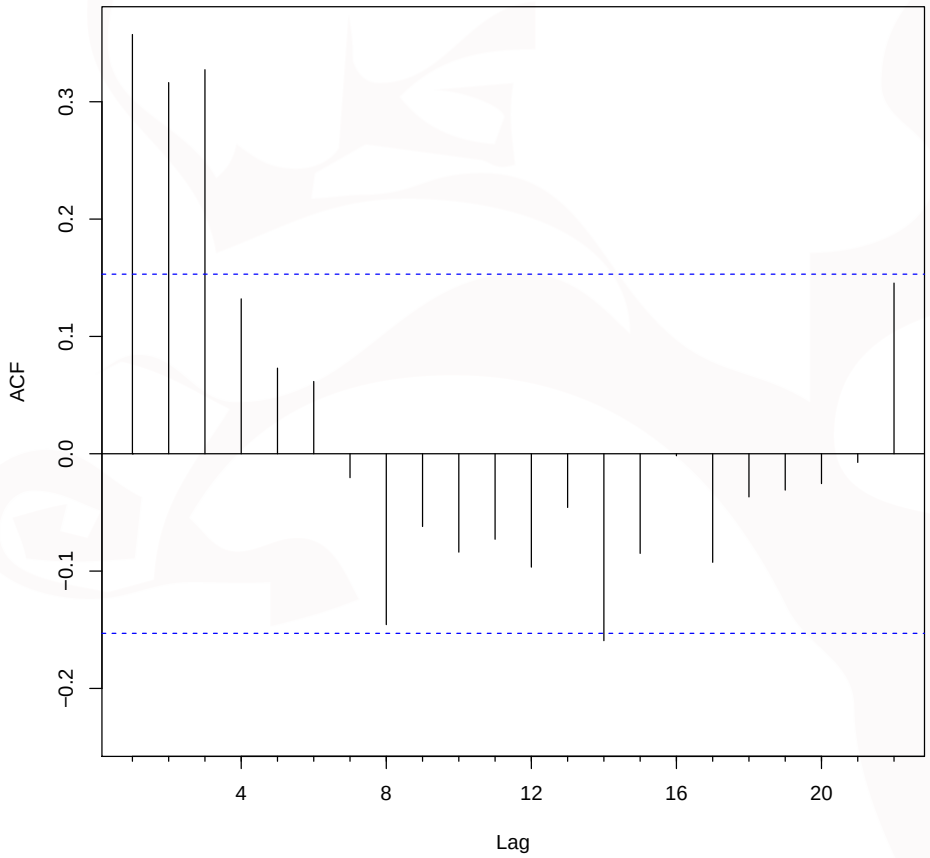
Modelo regresión y ARIMA

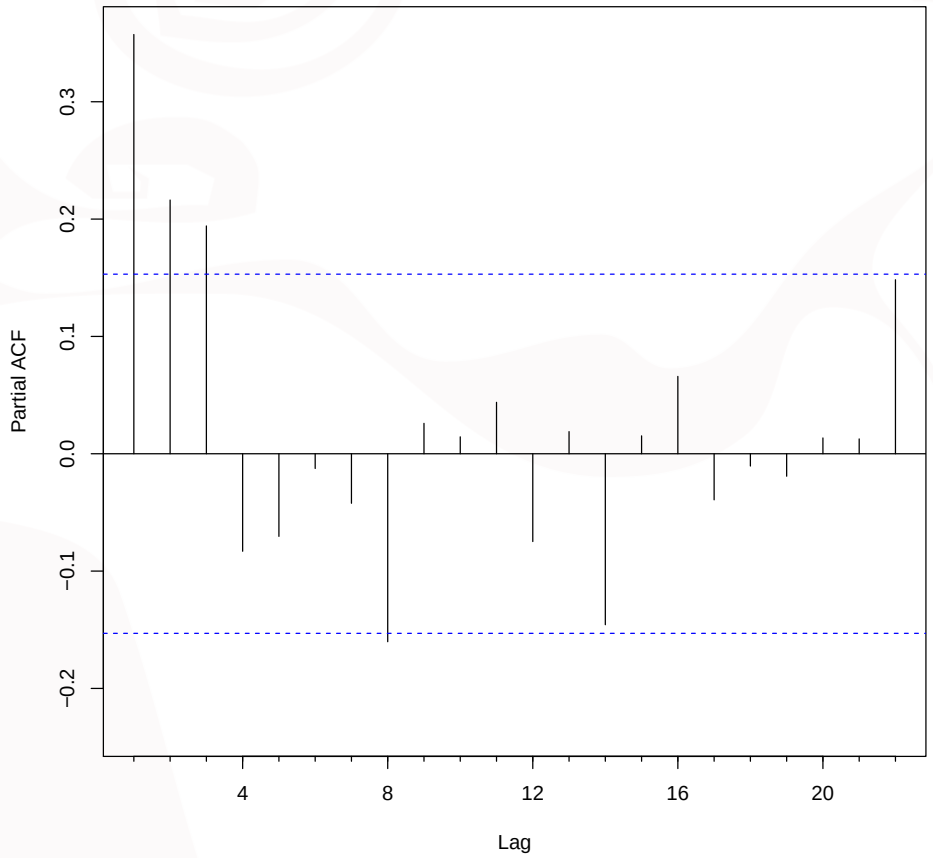
Utilizando el comando *xreg* y *ARIMA*

```
library(fpp) # Forecasting: principles and practice
plot(usconsumption, xlab="Year",
  main="Quarterly changes in US
    consumption and personal income")
```



```
par(mfrow = c(1, 2))
Acf(usconsumption[, 1], main = "")
Pacf(usconsumption[, 1], main = "")
```





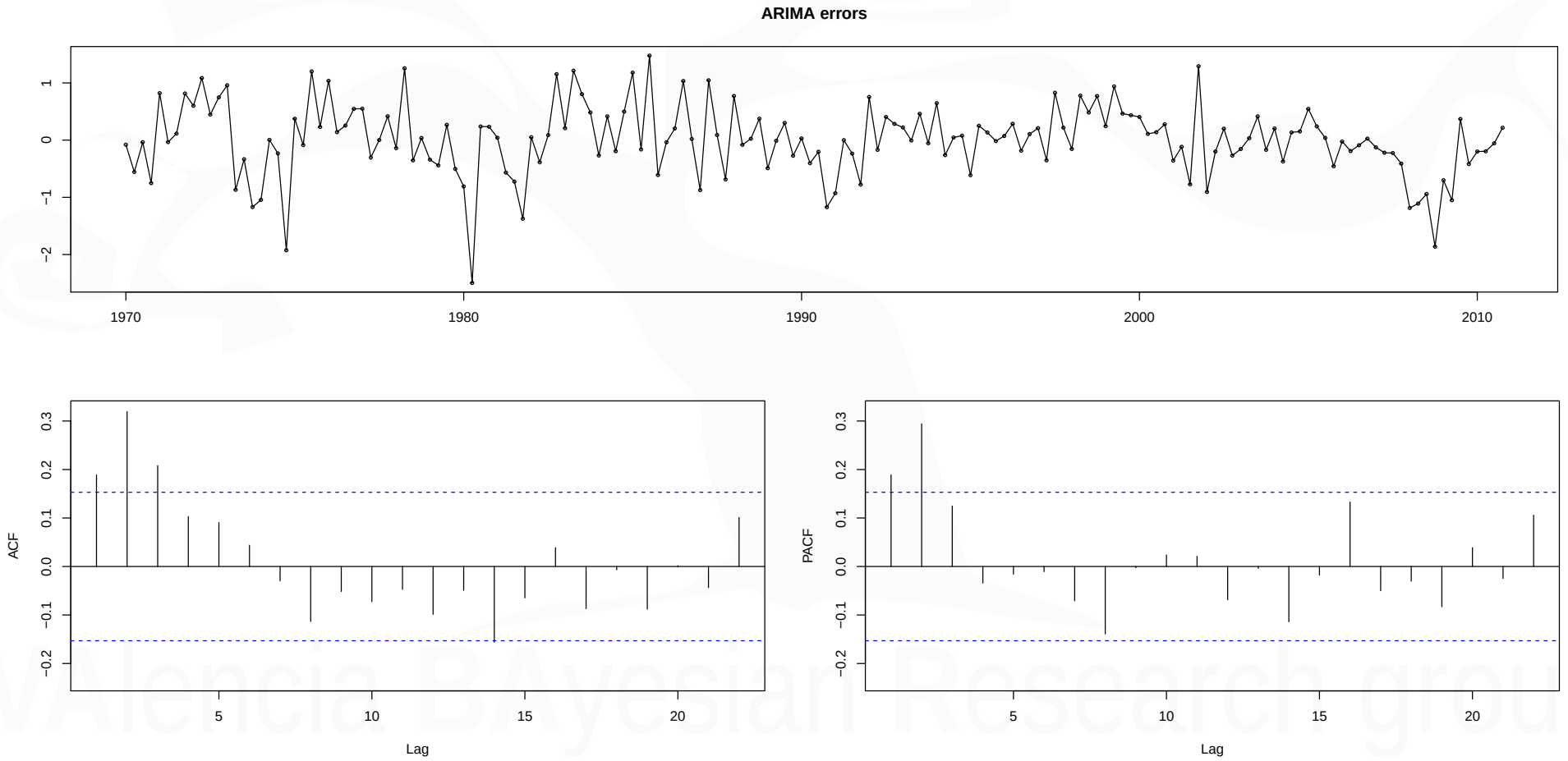
¿Podría ser un $AR(2)$ o $AR(3)$?

Dynamic regression models

Los datos

consumption	income
0.6122769	0.496540
0.4549298	1.736460
0.8746730	1.344881
-0.2725144	-0.328146
1.8921870	1.965432
0.9133782	1.490757

```
fit <- Arima(usconsumption[,1],
             xreg=usconsumption[,2],
             order=c(2,0,0))
tsdisplay(arima.errors(fit), main="ARIMA errors")
```



```
fit2 <- Arima(usconsumption[, 1], xreg = usconsumption[, 2],
    order = c(1, 0, 2))
summary(fit2)
```

```
  Series: usconsumption[, 1]
  Regression with ARIMA(1,0,2) errors

  Coefficients:
          ar1      ma1     ma2  intercept  usconsumption[, 2]
        0.6516  -0.5440  0.2187     0.5750              0.2420
  s.e.  0.1468   0.1576  0.0790     0.0951              0.0513

  sigma^2 estimated as 0.3502:  log likelihood=-144.27
  AIC=300.54   AICc=301.08   BIC=319.14

  Training set error measures:
                       ME      RMSE       MAE  MPE MAPE      MASE
  Training set 0.001835782 0.5827238 0.4375789 -Inf  Inf 0.6298752
                     ACF1
  Training set 0.000656846
```

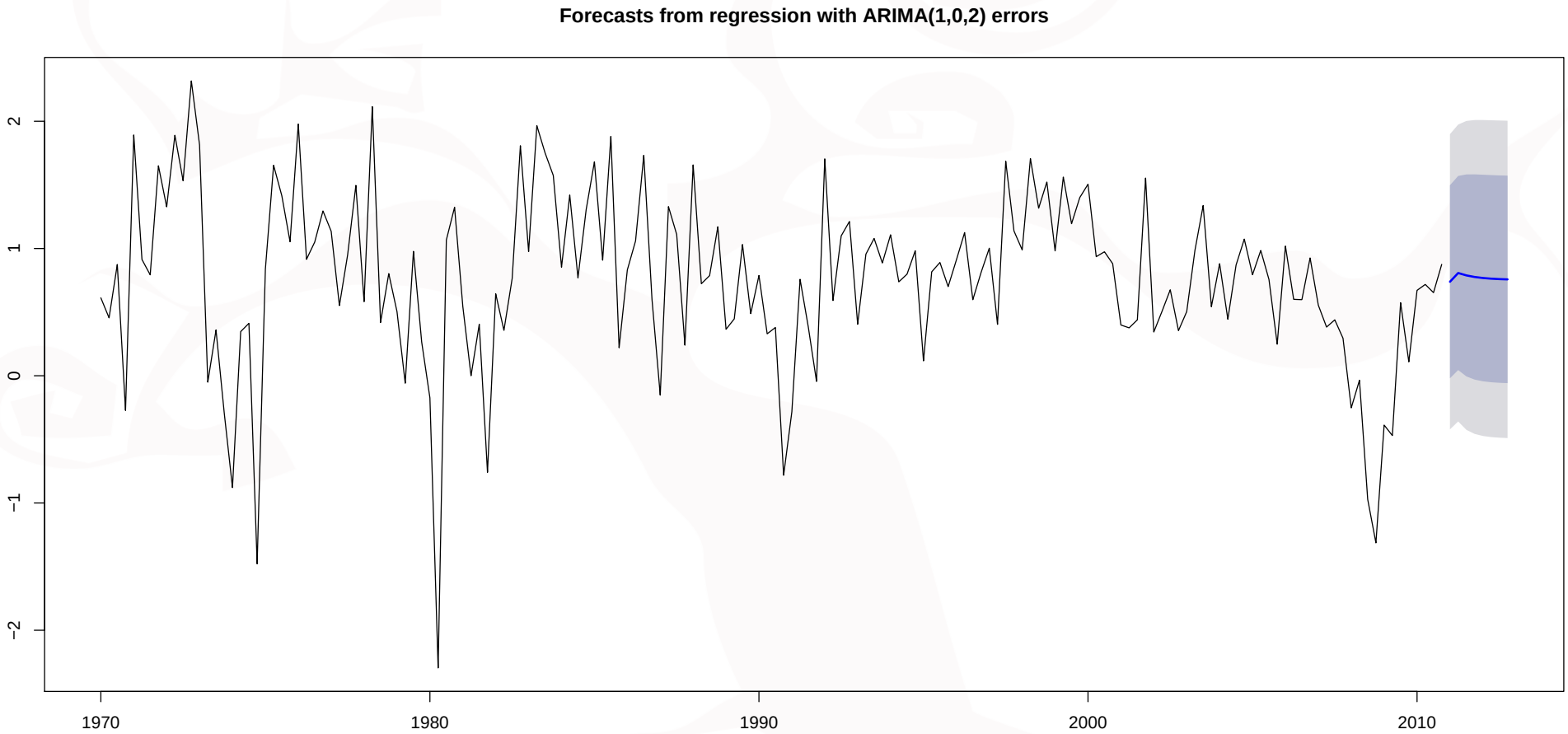
Parece que este modelo funciona mejor ARIMA(1,0,2)

```
Box.test(residuals(fit2), fitdf = 5, lag = 10, type = "Ljung")
```

```
    Box-Ljung test

data:  residuals(fit2)
X-squared = 4.5948, df = 5, p-value = 0.4673
```

¿Podemos predecir?

```
fcast <- forecast(fit2, xreg = rep(mean(usconsumption[, 2]),
    8), h = 8)
# h= periodos a predecir
plot(fcast, main = "Forecasts from regression with ARIMA(1,0,2) errors")
```



Ejemplo de la violencia mujeres

```
sel <- c("tfr", "partic", "degrees", "mconvict")
new.data <- data.frame(tfr = 2755, partic = 238, degrees = 13.2,
    mconvict = 765.7)
fit.xreg <- Arima(Hartnagel$fconvict, xreg = Hartnagel[, sel],
    order = c(2, 0, 0))
Box.test(residuals(fit.xreg), fitdf = 5, lag = 10, type = "Ljung")
new.data <- fcast <- forecast(fit.xreg, xreg = new.data, h = 8)
plot(fcast, main = "Forecasts from regression with ARIMA(2,0,0) errors")
```