Tema 06: El modelo de regresión lineal con residuos autocorrelados

Análisis estadístico de series económicas

Xavier Barber

Departamento de Estadística, Matemáticas e Informática
Centro de Investigación Operativa
Universitas Miguel Hernández de Elche

07/May/2019

Ajuste por mínimos cuadrados y la tendencia

Serie temporal Y_t como modelo lineal

- El modelo lineal ha sido la herramienta predominante en el contexto de la estadística clásica.
- La primera idea se basa en estudiar el "crecimiento" de la serie y su tendencia, es decir, expresar la relación de Y_t para $t = 1, 2, , q$.

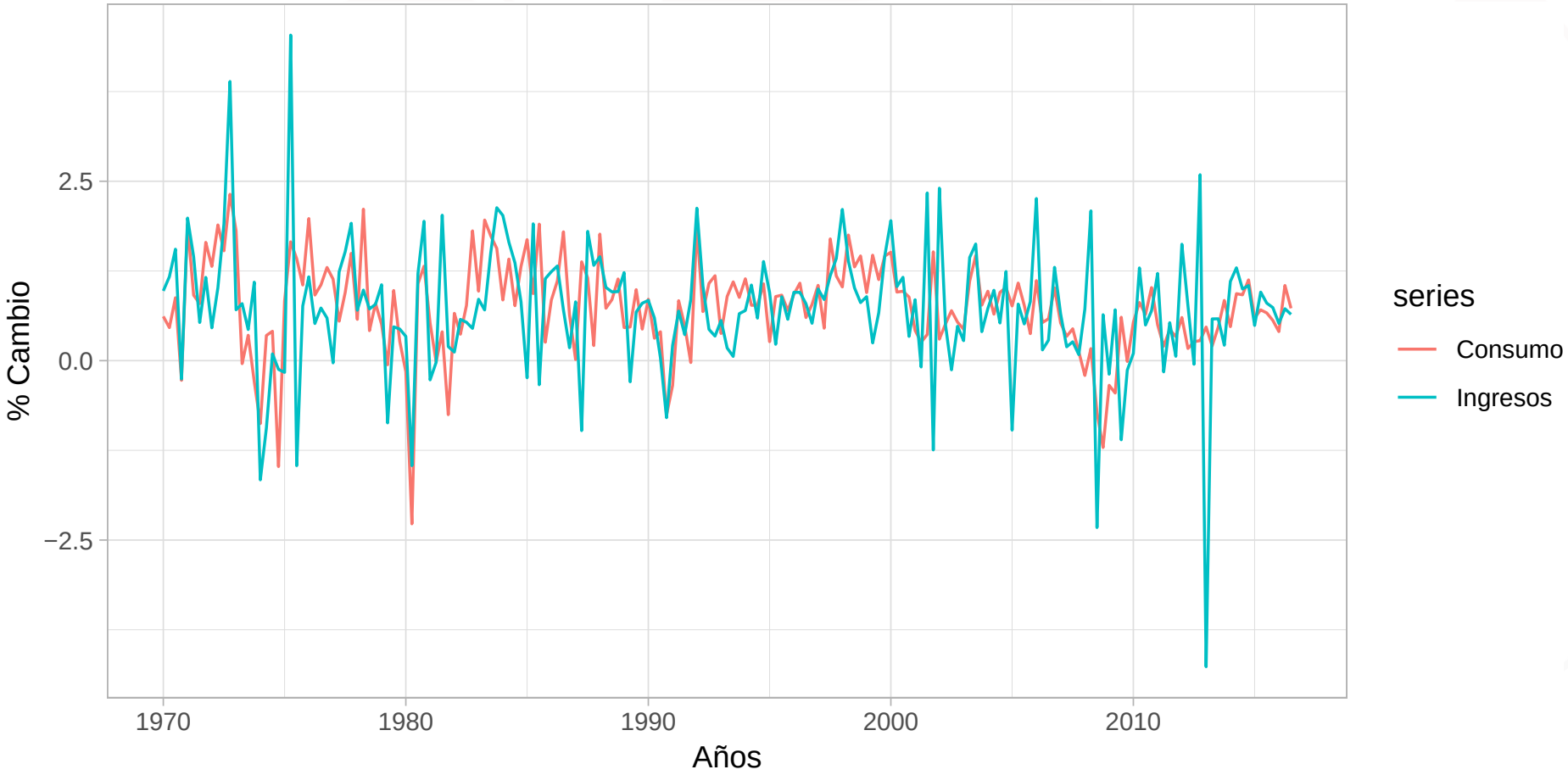
Modelo Lineal

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_q + Y_+ e_t$$

donde $\beta_1, \beta_2, \ldots, \beta_q$ son los coeficientes fijos de la regresión, y siendo $e_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ el error aleatorio o proceso de ruido.

Ejemplo: Gasto en consumo en USA

- Porcentaje de cambio trimestral (tasa de crecieminto) del gasto personal real en consumo y su disponibilidad de ingresos en USA desde 1970 a 2016.



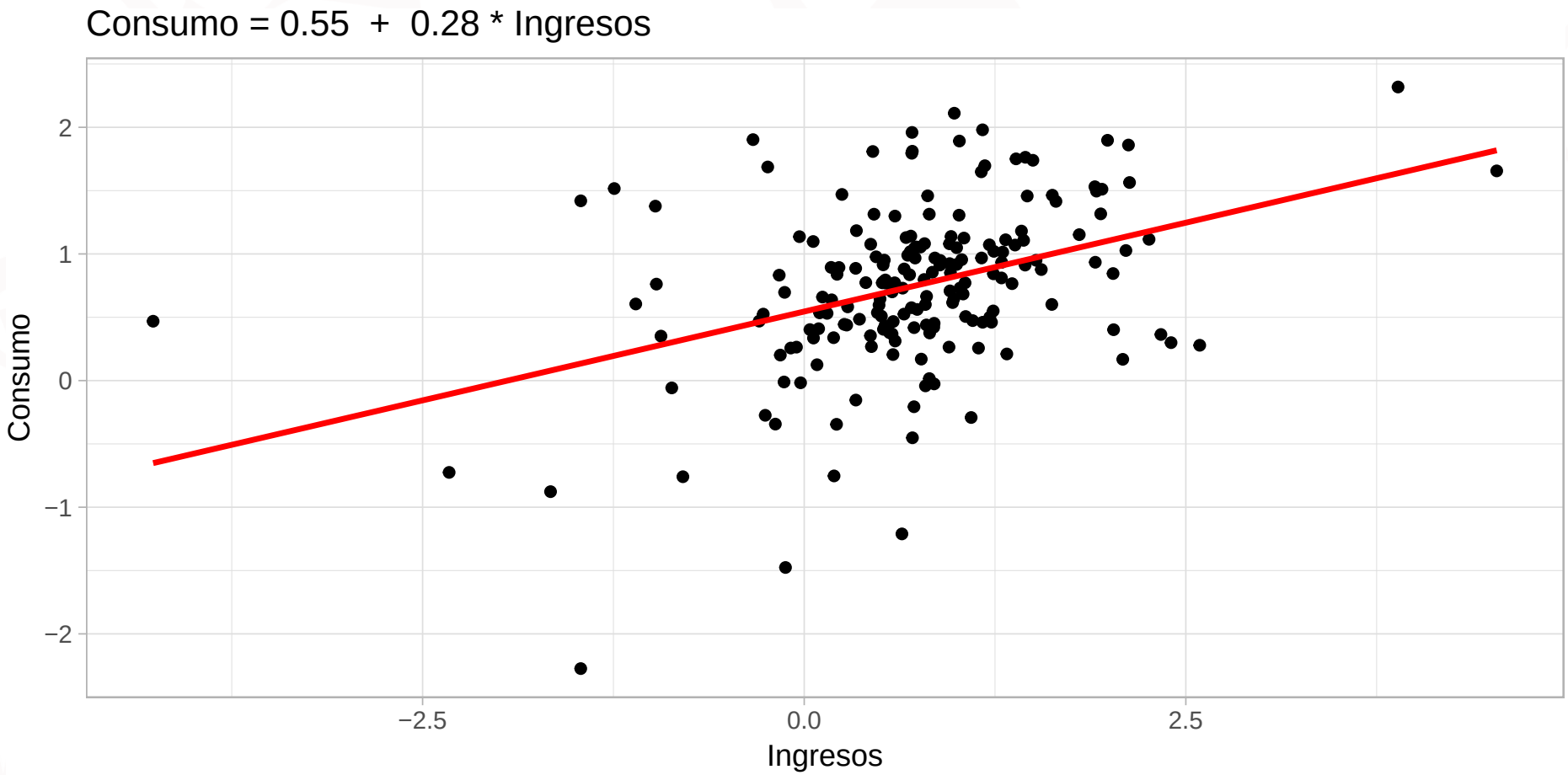
Ejemplo: Gasto en consumo en USA

```
tslm(Consumo ~ Ingresos, data = uschange) %>% summary
```

```
Call:
tslm(formula = Consumo ~ Ingresos, data = uschange)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40845 -0.31816  0.02558  0.29978  1.45157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54510    0.05569   9.789  < 2e-16 ***
Ingresos     0.28060    0.04744   5.915 1.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6026 on 185 degrees of freedom
Multiple R-squared:  0.159, Adjusted R-squared:  0.1545
F-statistic: 34.98 on 1 and 185 DF,  p-value: 1.577e-08
```

Ajuste modelo lineal

- ¿Cuál es la relación entre Consumo e ingresos?

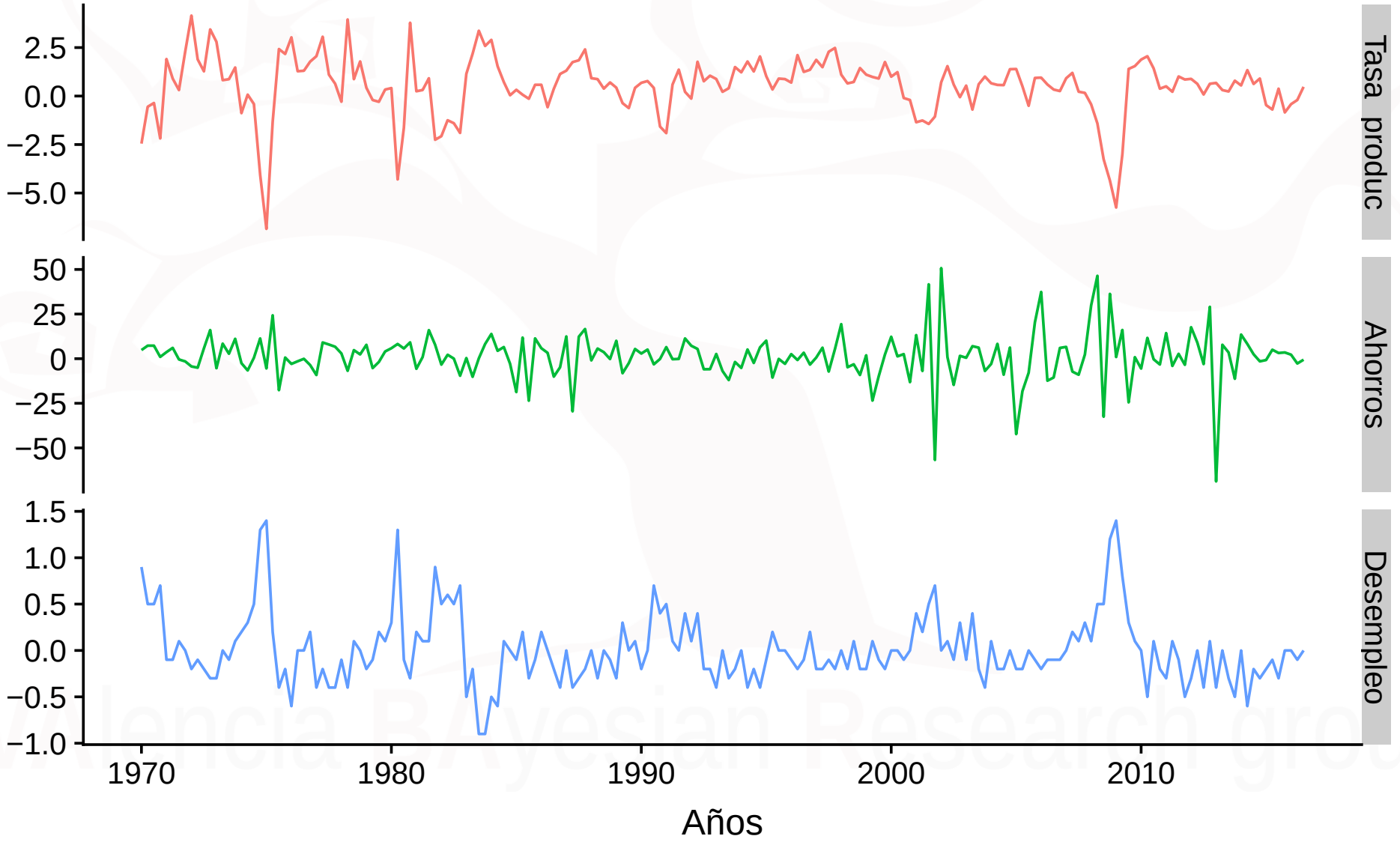


Por cada cambio de un 1% en los ingresos hay un cambio de 0.28 en el Consumo

Ajuste modelo lineal > 2 variables

```
autoplot(uschange[,3:5], facets = TRUE,
         colour=TRUE) +
         ylab("") + xlab("Años") +
         guides(colour="none")
```

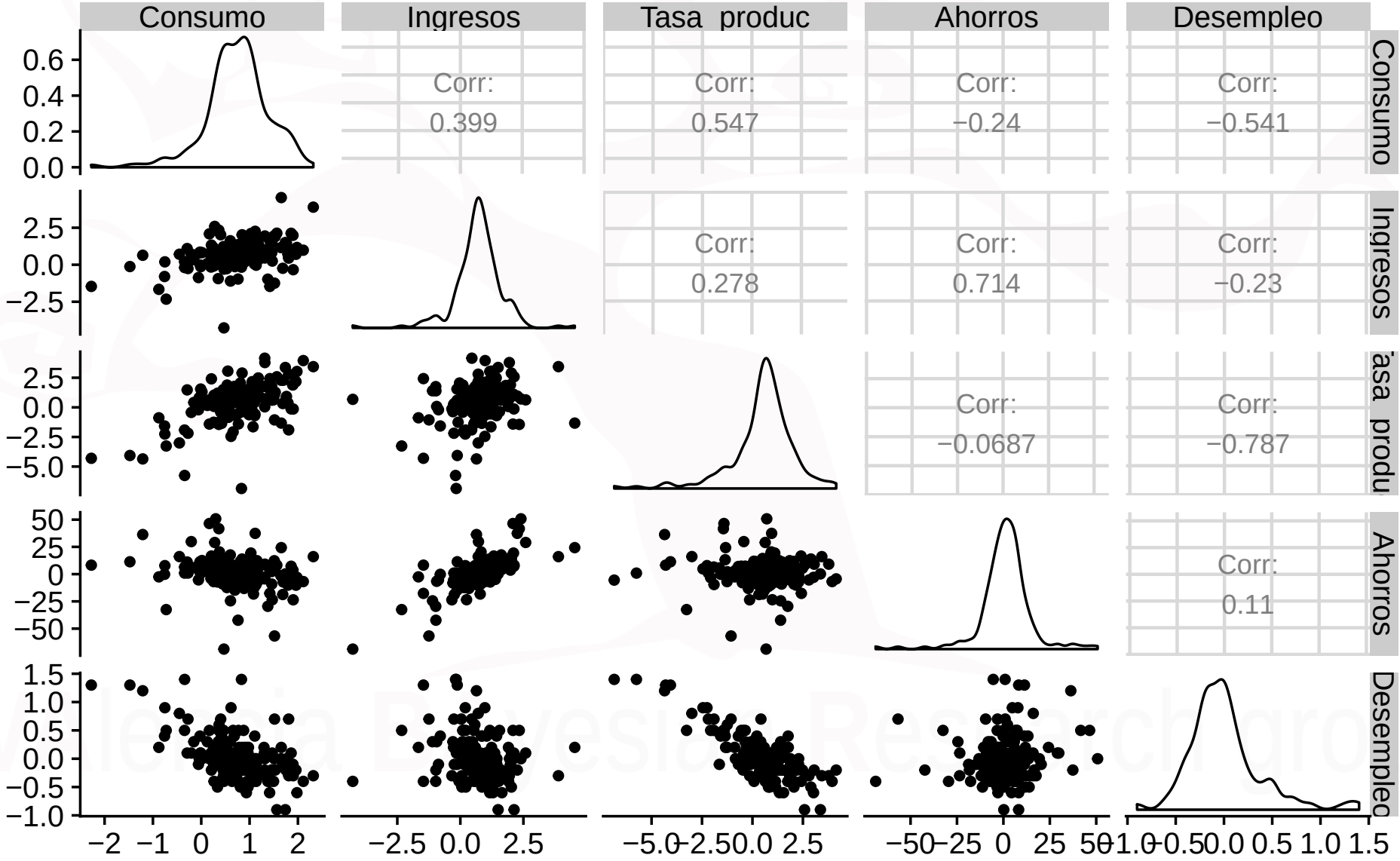
Ajuste modelo lineal > 2 variables



Ajuste modelo lineal > 2 variables

```
uschange %>%
  as.data.frame %>%
  GGally::ggpairs()
```

Ajuste modelo lineal > 2 variables



Ajuste modelo lineal > 2 variables

```
fit.consMR<-tslm(Consumo~Ingresos+Tasa_produc+Desempleo+Ahorros,  d
summary(fit.consMR)
```

```
Call:
tslm(formula = Consumo ~ Ingresos + Tasa_produc + Desempleo +
    Ahorros, data = uschange)

Residuals:
     Min       1Q   Median       3Q      Max
-0.88296 -0.17638 -0.03679  0.15251  1.20553

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26729    0.03721   7.184 1.68e-11 ***
Ingresos     0.71449    0.04219  16.934  < 2e-16 ***
Tasa_produc  0.04589    0.02588   1.773   0.0778 .
Desempleo   -0.20477    0.10550  -1.941   0.0538 .
Ahorros     -0.04527    0.00278 -16.287  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3286 on 182 degrees of freedom
Multiple R-squared:  0.754, Adjusted R-squared:  0.7486
F-statistic: 139.5 on 4 and 182 DF,  p-value: < 2.2e-16
```

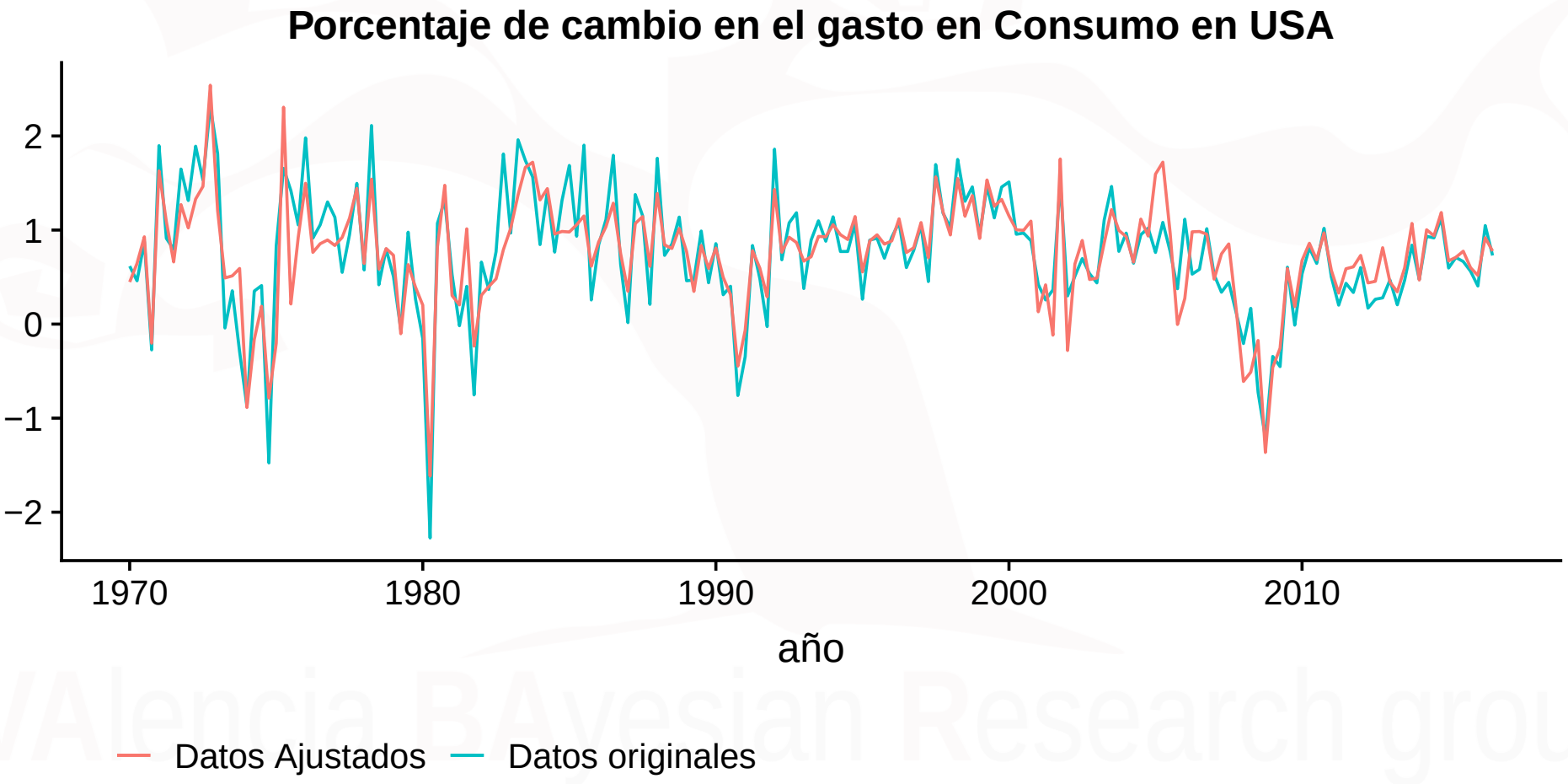
Ejemplo: Gasto en consumo de USA

- ¿Cómo ajusta este modelo?

```
autoplot(uschange[,'Consumo'], series="Datos originales") +
  autolayer(fitted(fit.consMR), series="Datos Ajustados") +
  xlab("año") + ylab("") +
  ggtitle("Porcentaje de cambio en el gasto en Consumo en USA") +
  guides(colour=guide_legend(title=" "))+
  theme(legend.position = "bottom")
```

Ejemplo: Gasto en consumo de USA

- ¿Cómo ajusta este modelo?



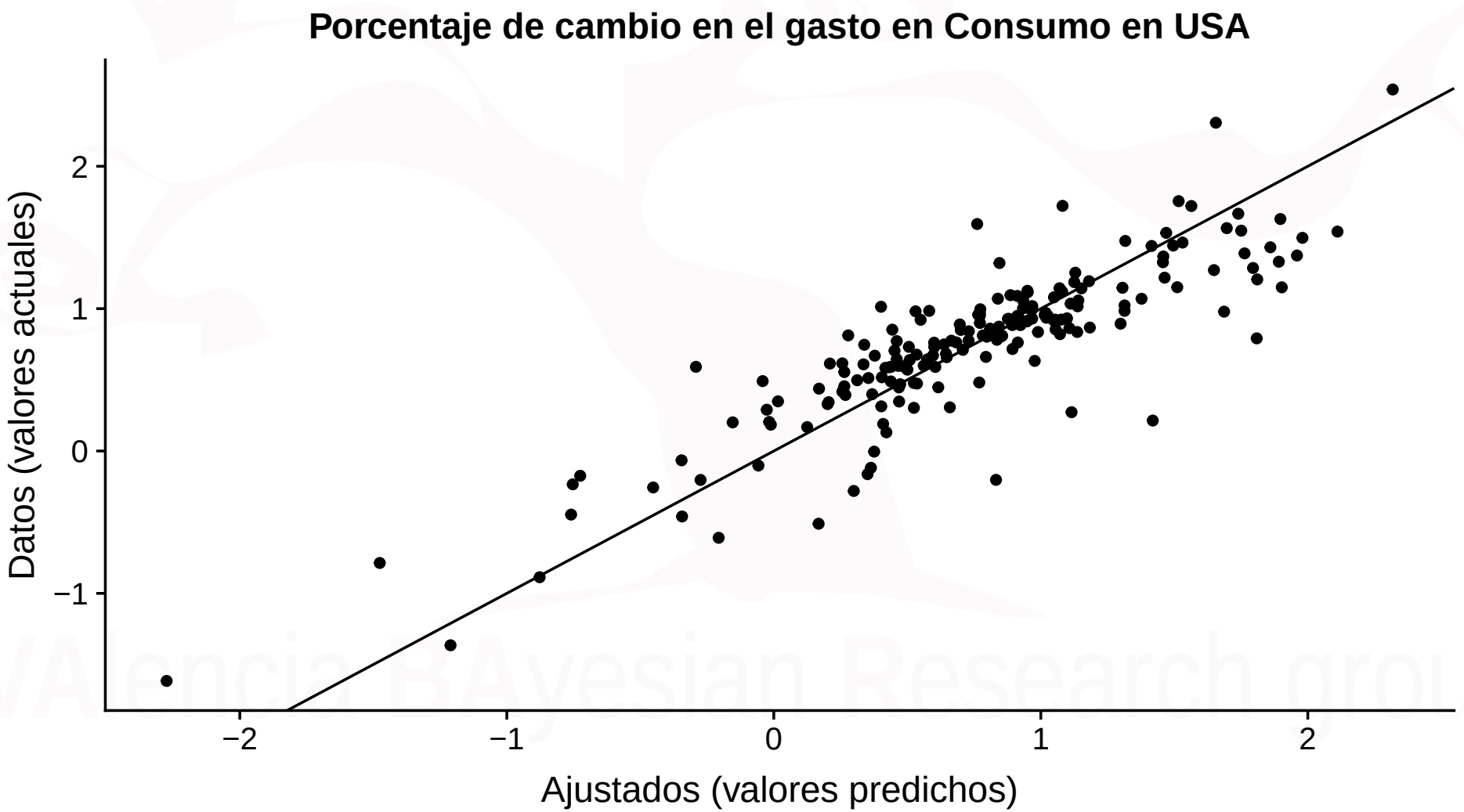
Ejemplo: Gasto en consumo de USA

- ¿Cómo ajusta este modelo?

```
cbind(Data=uschange[,"Consumo"], Fitted=fitted(fit.consMR)) %>%
  as.data.frame %>%
  ggplot(aes(x=Data, y=Fitted)) +
  geom_point() +
  xlab("Ajustados (valores predichos)") +
  ylab("Datos (valores actuales)") +
  ggtitle("Porcentaje de cambio en el gasto en Consumo en USA") +
  geom_abline(intercept=0, slope=1)
```

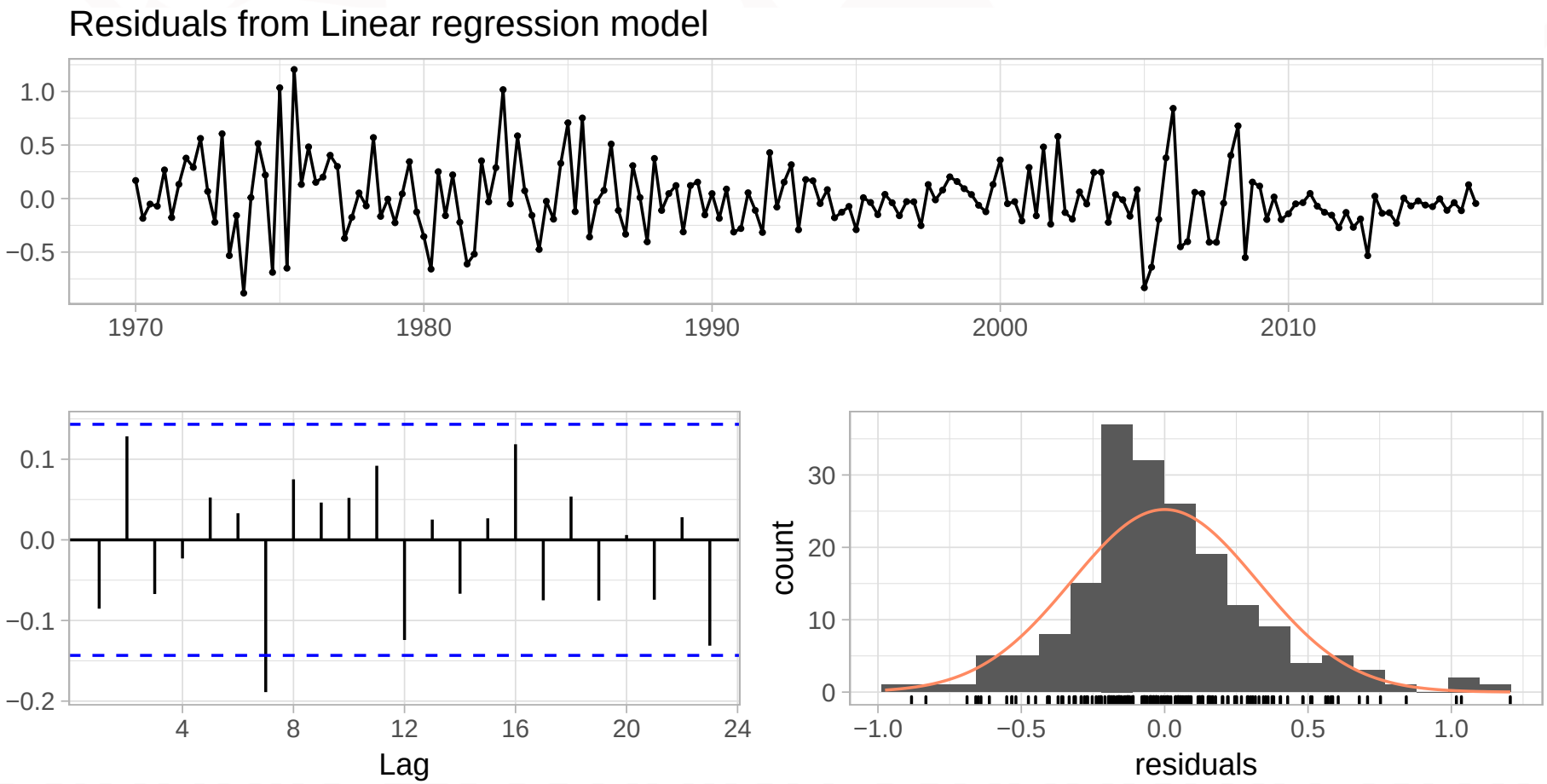
Ejemplo: Gasto en consumo de USA

- ¿Cómo ajusta este modelo?



Ejemplo: Gasto en consumo de USA

- ¿Es válido este modelo?



Autocorrelación serial

Introducción

- En el entorno de las series temporales la principal asunción que no suele cumplirse en la independencia de los residuos.
- La correlación en los residuos representa una estructura en los datos, y el modelo de regresión lineal "inicialmente" no lo tiene en cuenta.
- Cuando las observaciones tienen un orden secuencial, y como resultado una estructura de correlación relacionada con ese orden, esta correlación se le llama **autocorrelación**.

Residuos autocorrelados

- ¿Y si al mirar la validez del modelo tenemos autocorelación? (gráfico ACF)
- Este es un caso habitual cuando se estudian series temporales mediante modelos de regresión
- Hay situaciones en las que es relativamente sencillo solucionarlo introduciendo alguna variable más en el modelo
- En otras ocasiones se necesitan de métodos de ajuste alternativos a los mínimos cuadrados ordinarios.

Efectos de la autocorrelación

- La autocorrelación puede tener varios efectos sobre el ajuste del modelo:

1. Ajuste insesgado, pero no eficiente, en el sentido de no tener la mínima varianza posible. Un proceso **autorregresivo de orden uno**, AR(1), puede ser suficiente

Modelos lineal con residuos $\sim AR(1)$

$$Y_t = \beta_0 + \beta_1 \times X_1 + \ldots + \beta_p \times X_p + w_t$$
$$w_t = \rho e_{t-1} + z_i, \quad |\rho| < 1$$

Efectos de la autocorrelación

La autocorrelación puede tener varios efectos sobre el ajuste del modelo:

2. Las estimaciones de los errores estándar de los coeficientes de la regresión y de σ_e^2 son sesgados. Esto ocurre muy a menudo en los datos de series económicas, donde la correlación es positiva y el coeficiente de dicho AR(1) también es positivo.

3. Como resultado de ese sesgo, los intervalos de confianza, los contrastes de hipótesis y los intervalos de predicción no son válidos (muy anchos).

Identificando la autocorrelación

Un forma habitual de identificar la autocorrelación, $\rho_k \equiv corr(e_t, e_{t-1}) = \rho^k$, es mediante el estadístico de **Durbin-Watson**.

$H_0 : \rho = 0$; $H_a : \rho > 0$

$$d = \frac{\sum_{t=2} n(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- **Región de rechazo**:
  - Si $d < d_L$, se rechaza H_0.
  - Si $d > d_U$, no se rechaza H_0.
  - Si $d \in (d_U, d_L)$, el test no puede concluir nada.

Identificando la autocorrelación

- En el caso de R:

```
dwtest(formula, order.by = NULL,
       alternative = c("greater", "less",
                       "two.sided"),
       iterations = 15, exact = NULL,
       tol = 1e-10,  data = list())
```

Test de autocorrelación serial

- Otra alternativa para el estudio de la autocorrelación: *rank von Neumann ratio test*, que es una aproximación a la normal basada en la ecuaciones de Yule-Walker que se vieron en temas anteriores.

***Nota**: Muchos de los paquetes estadísticos del mercado utilizan por defecto los residuos estudentizados para eliminar el efecto de la magnitud de las variables a estudio:*

$$e^{\star t} = \frac{e_t - \bar{e_t}}{\sigma_e}$$

Lógicamente estos residuos estudendizados siempre tendrán media cero, pero tienen numerosas ventajas para tratarlos en otros tipos de contrastes.

Test de autocorrelación serial

- *Rank von Neumann ratio test*

$$H_0 : \rho_1 = 0$$
$$H_a : \rho_1 > 0; \quad H_a : \rho_1 < 0; \quad H_a : \rho_1 \neq 0$$

donde ρ_1 denota la verdadera autocorrelación con retardo-1.

El test se basa en la aporximación a la normal de las ecuaciones de Yule-Walker (`test="AR1.yw"`) que viene dada por:

$$\hat{\rho}_1 = \frac{\hat{\gamma}_1}{\hat{\gamma}_0}\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

donde $\hat{\gamma}_k$ es la estimación de la autocovarianza de retardo-k

Test de autocorrelación serial

Bajo la hipótesis nula:

- El valor estimado de retardo-1 se distribuye como una variable aleatoria Gaussiana[1] con media 0 y varianza dada por

$$Var(\hat{\rho}_1) \approx \frac{1}{n}.$$

- Entonces, la hipótesis nula puede contrastarse con el estadístico:

$$z = \sqrt{n}\hat{\rho}_1$$

El cual se distribuye aproximadamente como una normal estándar bajo la hipótesis nula, esto es, que la autocorrelación de retardo-1 es 0.

[1] Box and Jenkins, 1976, pp.34-35

Breusch-Godfrey test

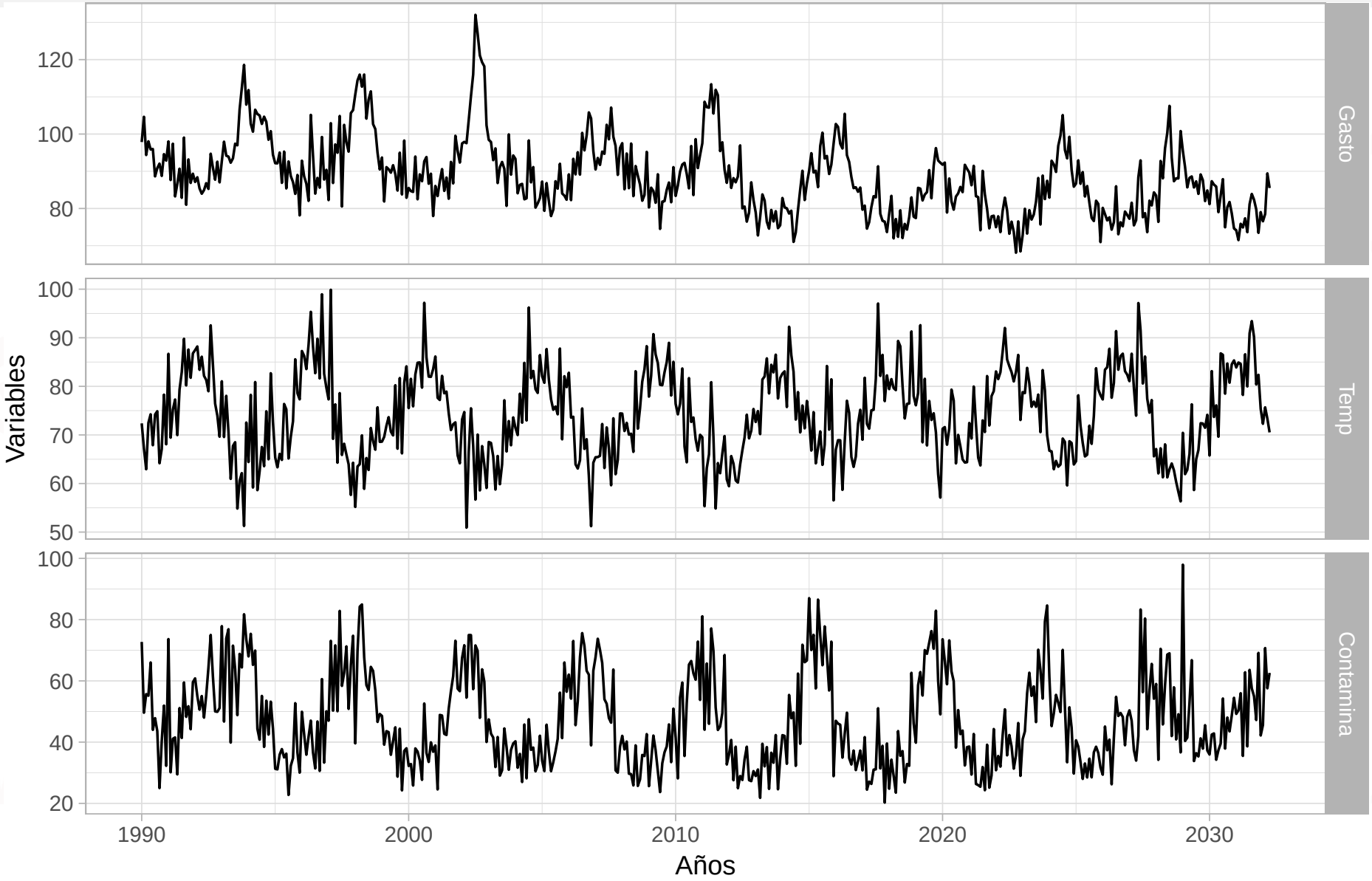
- El contraste de Godfrey-Breusch para la autocorrelación serial es un contraste basado en los multiplicadores de Lagrange, cuya hipótesis de trabajo son:

$$H_0 : \rho = 0$$
$$H_a : e_t \sim AR(p);\ H_a : e_t \sim MA(q)$$

- En *R* se utiliza el comando:

```
# del paquete lmtest
bgtest(formula, order = 1, order.by = NULL,
       type = c("Chisq", "F"),data = list(), fill = 0)
# o del paquete EnvStats el comando
checkresiduals(object, lag, df = NULL, test, plot = TRUE, ...)
```

Ejemplo: Gasto.farmacia∼Temp.+Conta.



Ejemplo: Gasto en farmacia

- Diferentes modelos y estudiar cuál se adecua mejor a los datos a estudio.

$$GF_t = \beta_0 + \beta_1 t + e_t \quad (1)$$

$$GF_t = \beta_0 + \beta_1 t + \beta_2 (T_t - \bar{T}) + e_t \quad (2)$$

$$GF_t = \beta_0 + \beta_1 t + \beta_2 (T_t - \bar{T}) + \beta_3 (T_t - \hat{T})^2 + e_t \quad (3)$$

$$GF_t = \beta_0 + \beta_1 t + \beta_2 (T_t - \bar{T}) + \beta_3 (T_t - \bar{T})^2 + \beta_4 C_t + e_t \quad (4)$$

Ejemplo: Gasto en farmacia

```
##Creación de las variables centradas y temepratura al cuadrado
temp = temp-mean(temp)
temp2 = temp^2
trend = time(gasto)          #  tiempo como tendencia ( != beta_0)

## ajuste del modelo con  tendencia ;  temp. ; temp^2  y cont.
regre1 = lm(gasto~  trend, na.action=NULL)
regre2 = lm(gasto~  trend + temp, na.action=NULL)
regre3 = lm(gasto~  trend + temp + temp2 , na.action=NULL)
regre4 = lm(gasto~ trend + temp + temp2 + part, na.action=NULL)
```

Ejemplo: Gasto en farmacia

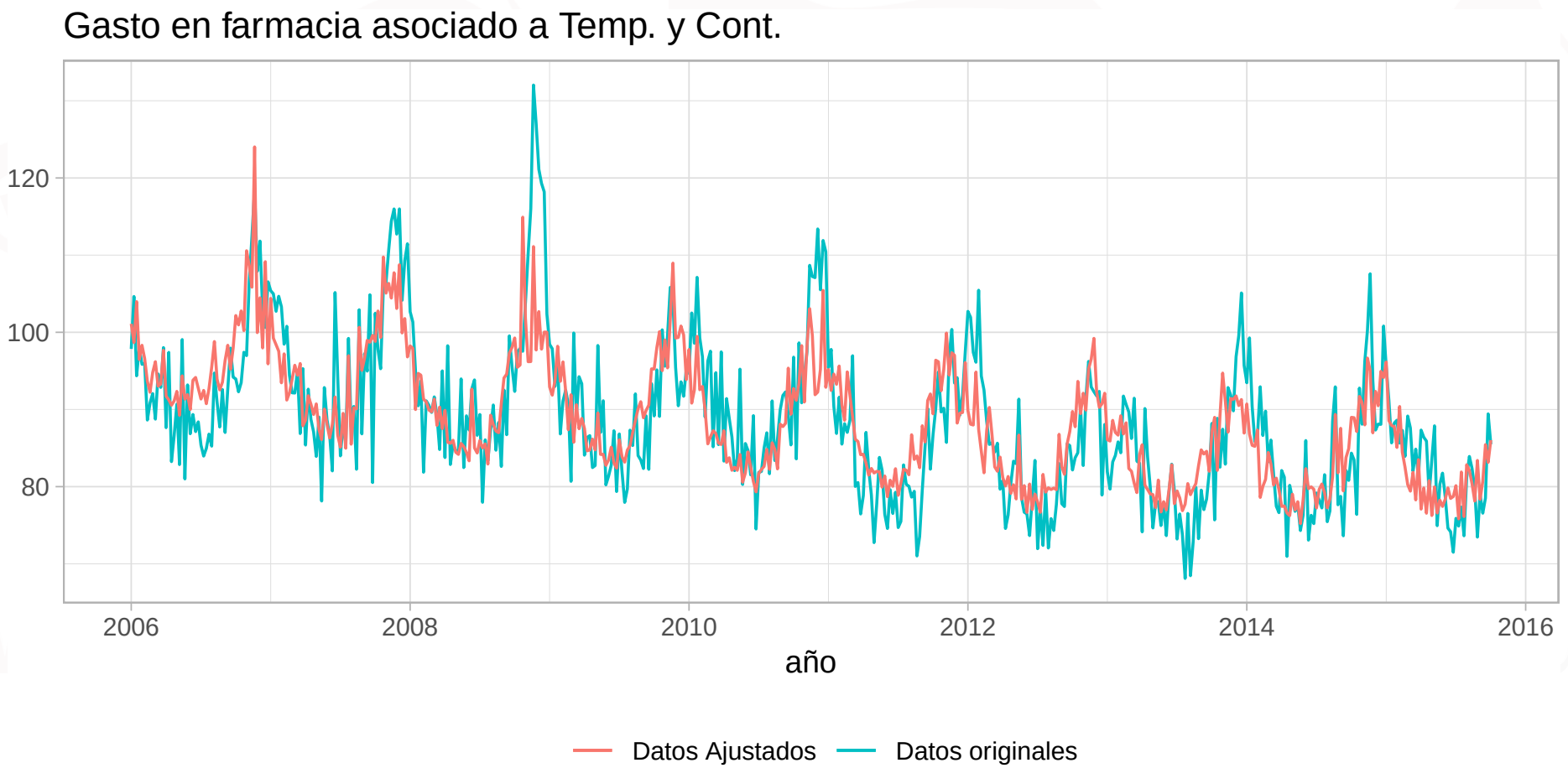
	Modelo-1	Modelo-2	Modelo-3	Modelo-4
β_0	3356.10^{***}	3181.12^{***}	3091.59^{***}	2881.74^{***}
	(281.35)	(249.96)	(236.43)	(203.19)
Tendencia	-1.62^{***}	-1.54^{***}	-1.49^{***}	-1.40^{***}
	(0.14)	(0.12)	(0.12)	(0.10)
Temperatura		-0.46^{***}	-0.48^{***}	-0.47^{***}
		(0.04)	(0.04)	(0.03)
Temperatura^2			0.03^{***}	0.02^{***}
			(0.00)	(0.00)
Contaminación				0.26^{***}
				(0.02)
R^2	0.21	0.38	0.45	0.60
Adj. R^2	0.21	0.38	0.44	0.59
Num. obs.	508	508	508	508
RMSE	8.89	7.89	7.45	6.39

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Ejemplo: Gasto en farmacia

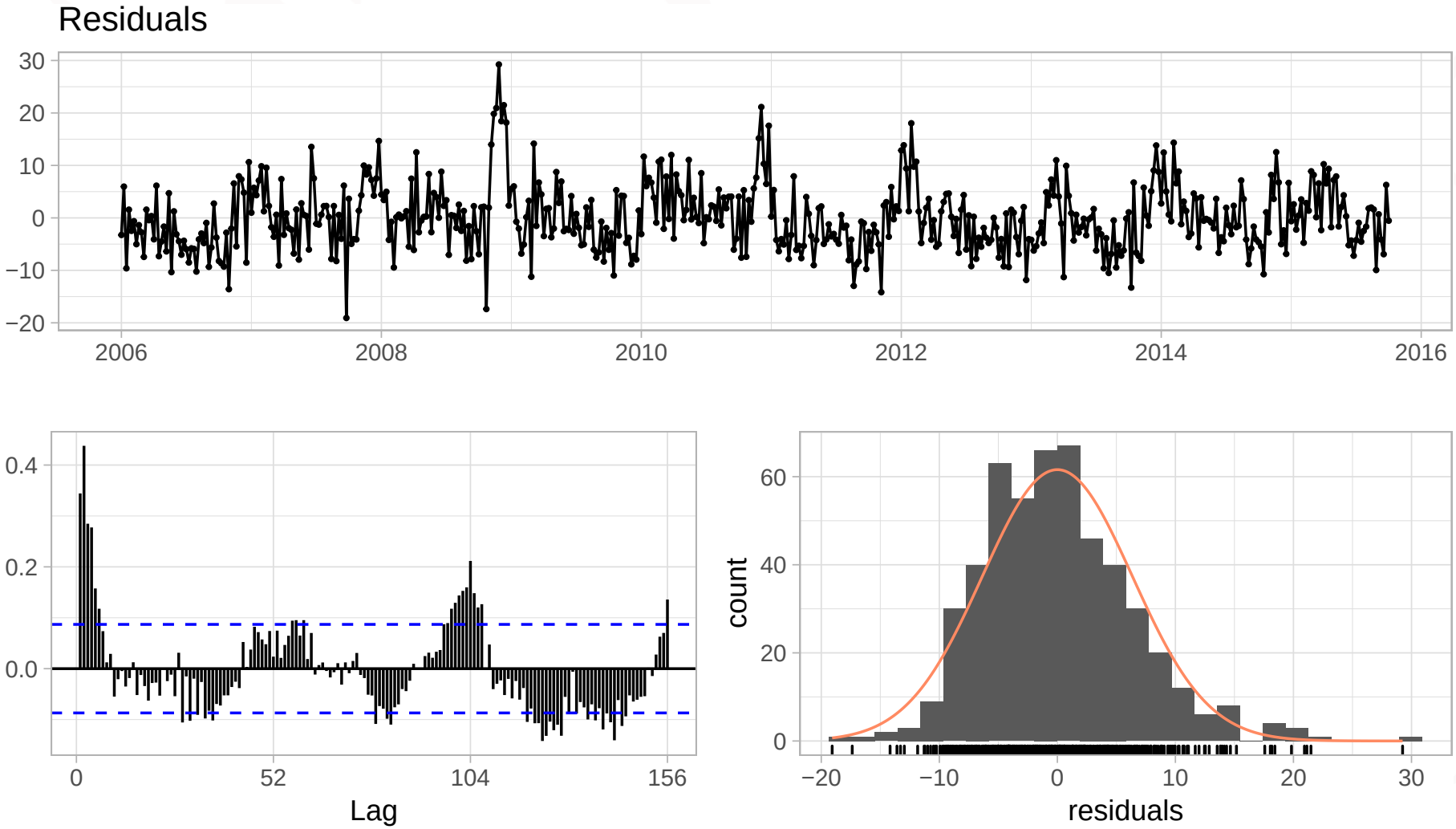
- El modelo más bondadoso:

$$\text{Gasto} = \beta_0 - 1.40 \times t - 0.47 \times \text{Temp.} + 0.02 \times \text{Temp.}^2 + 0.26 \times \text{Cont.} + e_t$$



Ejemplo: Gasto en farmacia

¿Es válido el modelo?



Ejemplo: Gasto en farmacia

¿Es válido el modelo?

```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:

Test Name: Breusch-Godfrey test for serial correlation of
order up to 101

Data: Residuals

Test Statistic: LM test = 198.3097

Test Statistic Parameter: df = 101

P-value: 2.599885e-08
```

Ejemplo: Gasto en farmacia

- ¿Existe autocorrelación en los residuos

```
    Durbin-Watson test

data:  regre4
DW = 1.3109, p-value = 1.523e-15
alternative hypothesis: true autocorrelation is greater than 0
```

Estos modelos van a producir estimaciones sesgadas y no es un modelo válido.

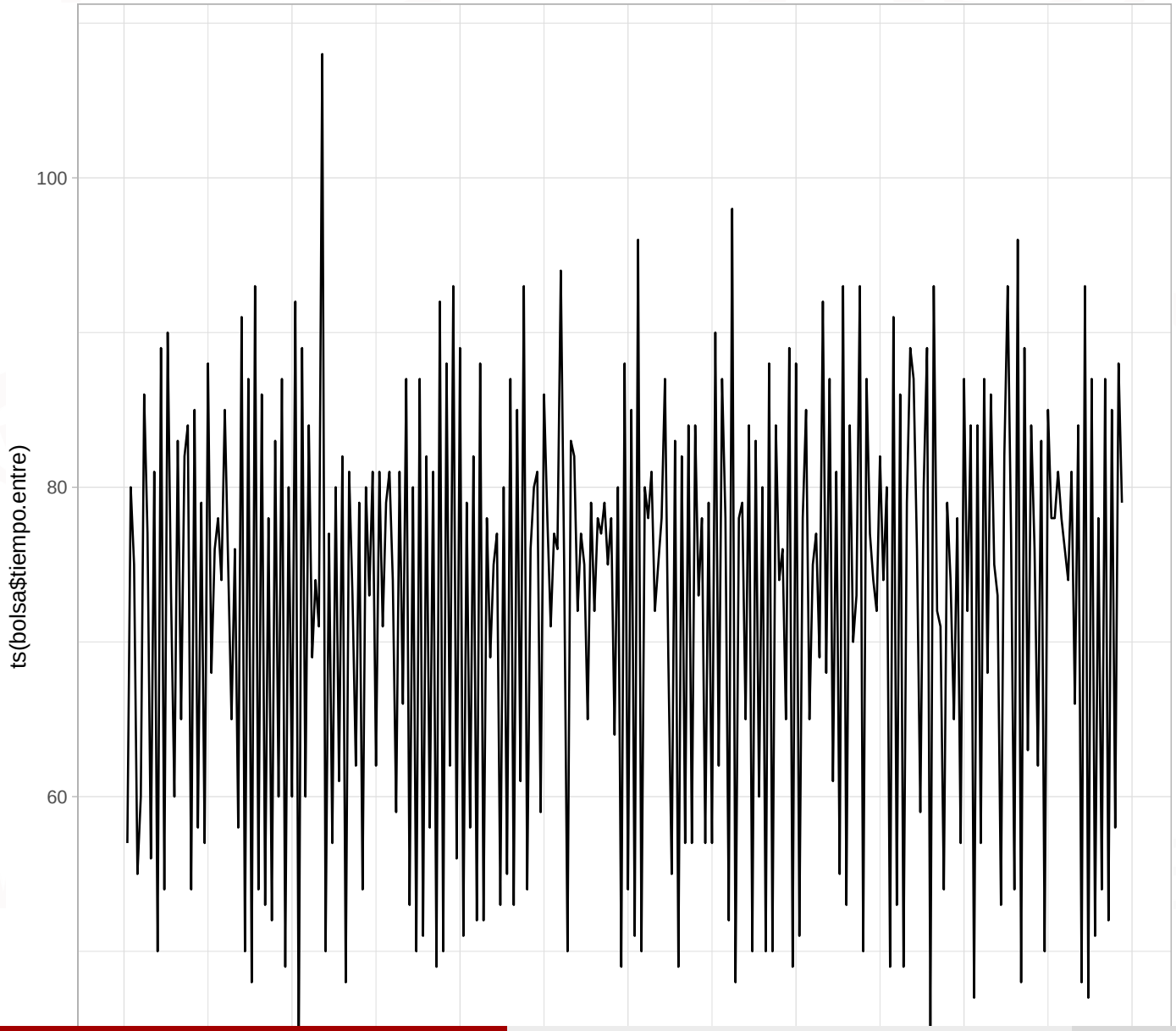
Se deben probar diferentes alternativas para analizar la relación existente entre el gasto en farmacia y las temperaturas.

Ejemplo: Gasto en farmacia

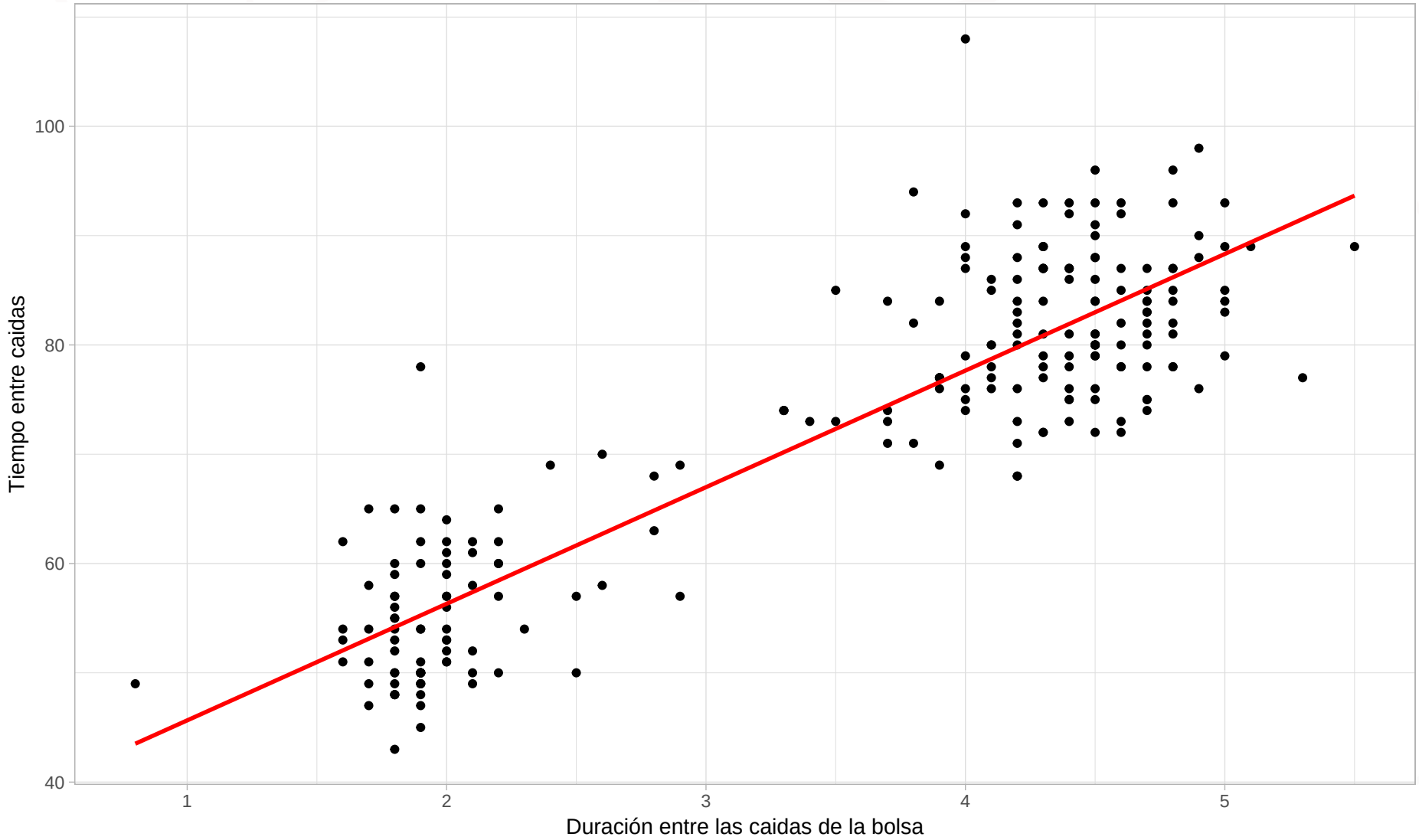
- Para el *Rank von Neumann ratio test*:

```
  z-Test for Lag-1 Autocorrelation (Normal Approximation)

data:  Residuals
z = 7.76, p-value = 8.438e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.344296
```

Ejemplo: Tiempo entre "caidas" en bolsa



Ejemplo: Tiempo entre "caidas" en bolsa



Ejemplo: Tiempo entre "caidas" en bolsa

- Los modelos que se van a probar son:

$$Tiempo.entre_t = \beta_0 + \beta_1 \times Dura_t + e_t$$
$$Tiempo.entre_t = \beta_0 + \beta_1 \times Dura_t + e_t$$
$$Tiempo.entre_t = \beta_0 + \beta_1 \times Dura_t + \beta_2 \times Dura_t^2 + e_t$$
$$Tiempo.entre_t = \beta_0 + \beta_1 \times Dura_t + \beta_2 \times Dura_t^2 + \beta_3 \times Dura_t^3 + e_t$$

- Para realizar estos ajustes en *R* es recomentable utilizar el comando *poly(variable, grado)*.

```
regreBolsa3b <- lm(tiempo.entre ~ poly(Duracion, 3))
```

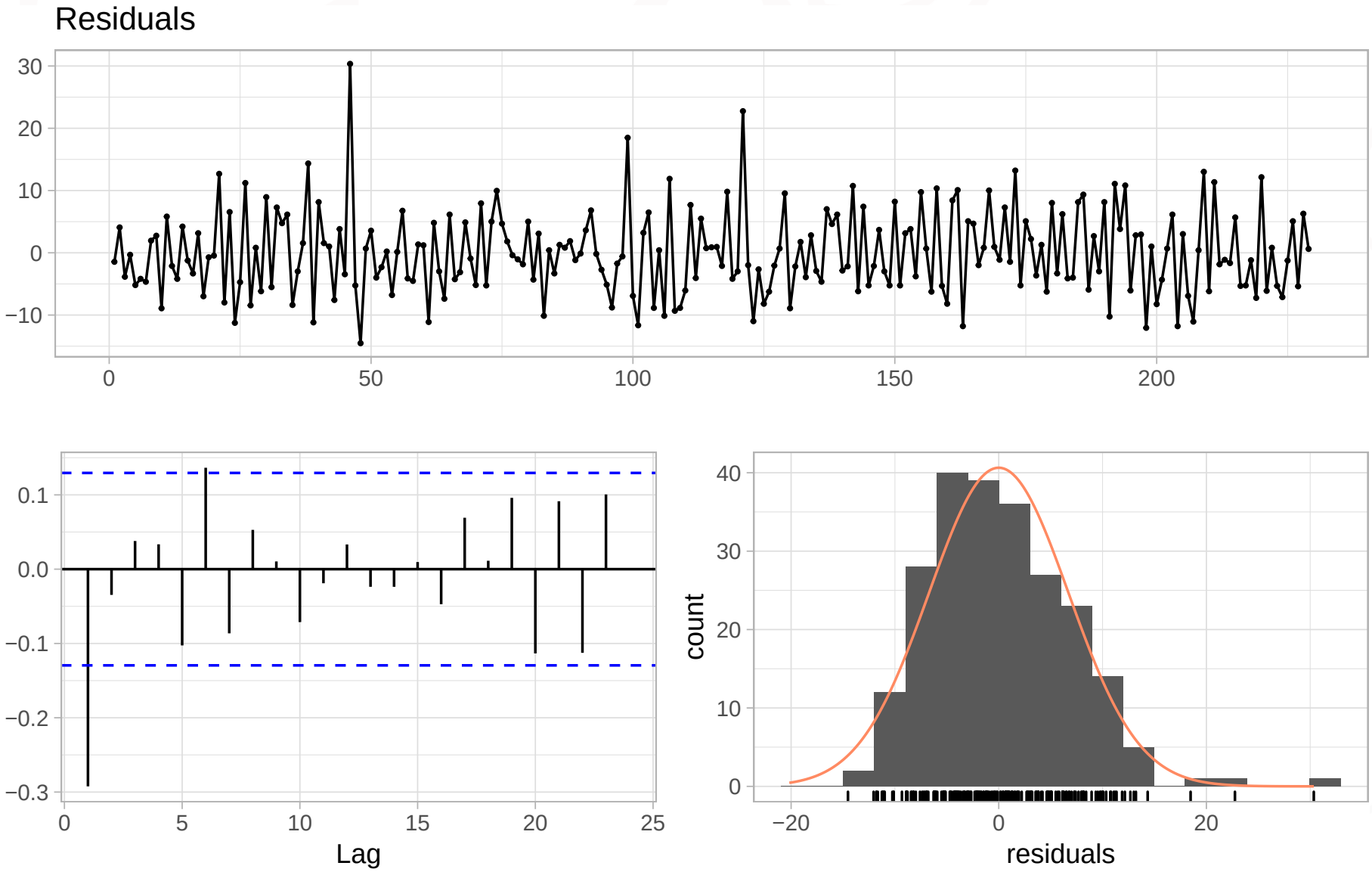
Ejemplo: Tiempo entre "caidas" en bolsa

	Mod.Lineal	Mod.Cuadrático	Mod.Cúbico
(Intercept)	34.98^{***}	20.88^{***}	54.80^{***}
	(1.35)	(4.83)	(12.74)
Duracion	10.67^{***}	20.89^{***}	-14.41
	(0.37)	(3.39)	(12.75)
dura2		-1.58^{**}	9.61^{*}
		(0.52)	(3.93)
dura3			-1.11^{**}
			(0.39)
R^2	0.79	0.80	0.80
Adj. R^2	0.79	0.80	0.80
Num. obs.	229	229	229
RMSE	6.74	6.62	6.52

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- Dados los valores del $Adj.R^2$ y del $RMSE$, el Modelo 1 sería el más parsimonioso (mínimo número de variables en el modelo)

Ejemplo: Tiempo entre "caidas" en bolsa

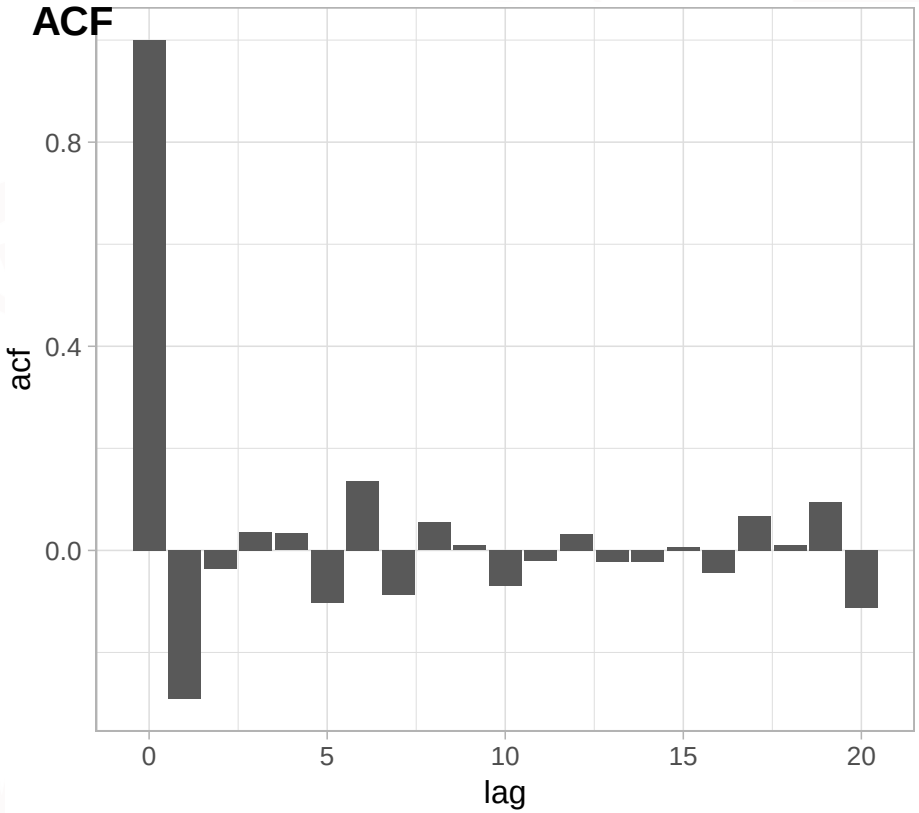


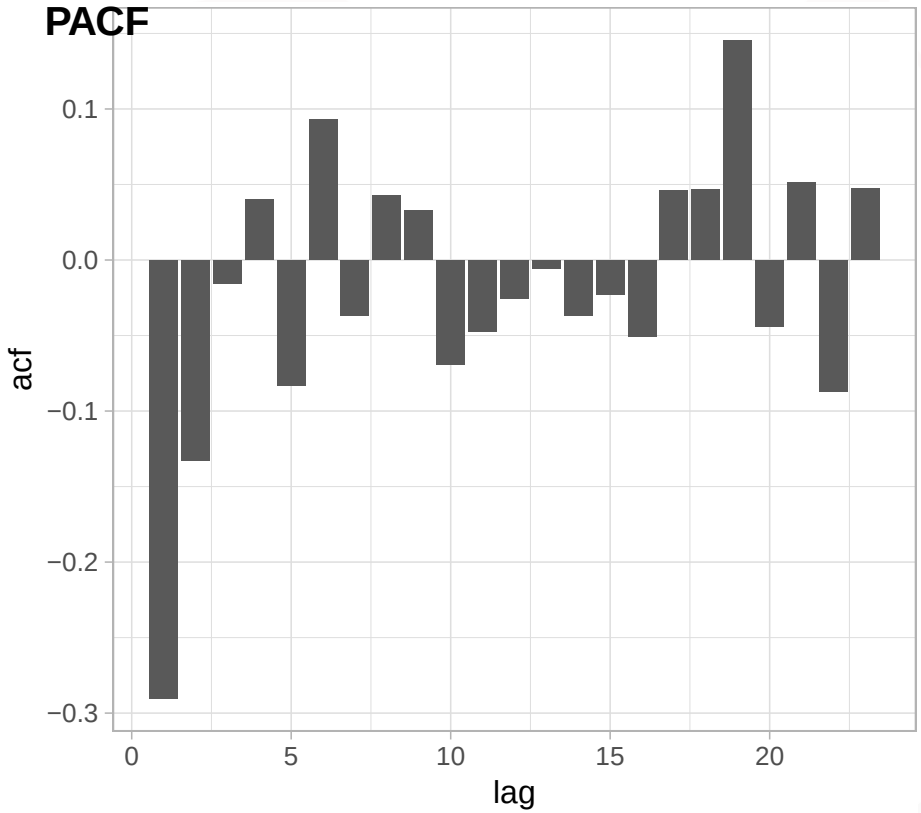
Ejemplo: Tiempo entre "caidas" en bolsa

```
Results of Hypothesis Test
--------------------------

Alternative Hypothesis:

Test Name: Breusch-Godfrey test for serial correlation of
order up to 10

Data: Residuals

Test Statistic: LM test = 28.89081

Test Statistic Parameter: df = 10

P-value: 0.001297776
```

Ejemplo: Tiempo entre "caidas" en bolsa

- A la vista del gráfico anterior, creemos que es neceario profundizar en el estudio de la autocorrelación





Procedimiento de Cochrane-Orcutt

Procedimiento de Cochrane-Orcutt

- El procedimiento indica que los residuos debe de seguir una proceso AR(1):

$$y_t = \beta_0 + \beta_1 x_t + e_t, \quad e_t = \rho e_{t-1} + w_t$$

ahora consideraremos la trasnformación:

$$y_t^\star = y_t - \rho y_{t-1}$$

Procedimiento de Cochrane-Orcutt

$$
\begin{aligned}
y_t^\star &= y_t - \rho y_{t-1} \\
&= \beta_0 + \beta_1 x_t + e_t - \rho(\beta_0 + \beta_1 x_{t-1} + e_{t-1}) \\
&= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + e_t - \rho e_{t-1} \\
&= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + w_i \\
&\equiv \beta_0^\star + \beta_1 x_t^\star + w_t
\end{aligned}
$$

donde $\beta_0^\star = \beta_0(1-\rho)$ *y* $x_t^\star = x_t - \rho x_{t-1}$. *Estas trasnformaciones harán que* ***la regresión sobre*** $y_t^\star$ ***cumpla las asunciones y los nuevos errores*** w_t ***sean independientes e identicamente distribuidos***.

Procedimiento de Cochrane-Orcutt

1. Determinar una estimación para ρ (retardo 1 del ACF)
2. Trasnformar en $y_t^\star = y_t - \rho y_{t-1}$ y $x_t^\star = x_t - \rho x_{t-1}$.
3. Ajustar la regresión de $y_t^\star$ en $x_t^\star$ usando OLS. El intervalo de predicción vendrá dado por $y_t^\star \pm 2\tilde{\sigma}$, donde $\tilde{\sigma} = \hat{\sigma}/\sqrt{1-\hat{\rho})}$ y $\hat{\sigma}$ es el error standar de la estimación en el ajuste de Cochrane-Orcutt.

Cochrane-Orcutt en R

- Existe un paquete que se llama *orcutt*,
- Resuelve los problemas de autocorrelación de primer orden de forma iterativa.

```
cochrane.orcutt(reg, convergence = 8)
```

Ejemplo: Gasto en farmacia

```
Call:
lm(formula = gasto ~ trend + temp + temp2 + part, na.action = NULL)

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.1613e+03  5.6946e+02   5.551 4.599e-08 ***
trend       -1.5323e+00  2.8316e-01  -5.411 9.709e-08 ***
temp        -1.6203e-02  4.2458e-02  -0.382    0.7029
temp2        1.4833e-02  2.1747e-03   6.821 2.612e-11 ***
part         1.5587e-01  2.4082e-02   6.472 2.300e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.6905 on 502 degrees of freedom
Multiple R-squared:  0.2194 ,  Adjusted R-squared:  0.2132
F-statistic: 35.3 on 4 and 502 DF,  p-value: < 5.571e-26

Durbin-Watson statistic
(original):    1.31088 , p-value: 1.523e-15
(transformed): 2.58152 , p-value: 1e+00
```

Ajuste por mínimos cuadrados generalizados (GLS)

Introducción a los GLS

- Del inglés *Generalized Least Squares* que no debe confundirse con los modelos lineales generalizados (GLM).
- Se trata de una correción al algoritmo de los *OLS* propuesta por Cochrane-Orcutt (1949) que es óptima para cuando los errores tienen una estructura AR(1).
- La estimación vía GLS puede ser una de las soluciones al problema de la autocorrelación.

Estimación de los parámetros

- Usa una matriz diagonal para ayudar a corregir la varianza no constante. Sin embargo, en un modelo con errores correlacionados, los errores tienen una estructura de varianza-covarianza más complicada.
- Por lo tanto, la matriz de ponderación para la estructura de varianza-covarianza más complicada no es diagonal y utiliza el método de **mínimos cuadrados generalizados**, de los cuales los mínimos cuadrados ponderados son casos especiales.

Estimación de los parámetros

- Dado el modelo lineal en forma matricial

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{e}\ E[\boldsymbol{e}] = 0\ Var[\boldsymbol{e}] = \sigma^2 \boldsymbol{I} \tag{5}$$

- Los cuadrados mínimos ordinarios (OLS) se estimaban minimizando la suma del cuadrado de los errores.

$$RSS = \sum_{t=1}^{n} (y_t - \boldsymbol{\beta x}_t)^2,$$

y se obtiene

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}\boldsymbol{y}$$

Modelo con GLS

- Dado el modelo lineal en forma matricial, los cuadrados mínimos generalizados

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{e}\ E[\boldsymbol{e}] = 0\ Var[\boldsymbol{e}] = \sigma^2 \boldsymbol{V}$$

donde $\boldsymbol{V}$ *es una matriz desconocida de orden* $n \times n$.

Si $\boldsymbol{V}$ *es diagonal pero con elementos distintos, las observaciones* y *están incorreladas pero tienen varianza desigual.*

Si $\boldsymbol{V}$ *tiene elementos distintos de cero fuera de la diagonal, las observaciones están correladas.*

Estimación por GLS

- Se descompone la matriz de la varianza en $\mathbf{V} = \mathbf{K}'\mathbf{K}$

  donde $\mathbf{K}$ *se la conoce como la raiz cuadrada de* $\mathbf{V}$

- Se define pues

$$\mathbf{g} = \mathbf{K}^{-1}\mathbf{e} \Rightarrow \mathbf{z} = \mathbf{B}\beta + g$$

con lo que se obtiene, utilizando resultados de álgebra matricial

$$E[\mathbf{g}] = \mathbf{K}^{-1}E[\mathbf{e}] = 0$$
$$Var[\mathbf{g}] = Var[\mathbf{K}^{-1}\mathbf{e}] = \sigma^2\mathbf{I}$$

Estimación por GLS

- Y entonces bajo las asunciones de los cuadrados mínimos ordinarios, las estimaciones aquí de los cuadrados mínimos generalizados serán:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{y}$$

que es insesgado y con

$$Var\left(\hat{\beta}\right) = \sigma^2\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

y su verosimilitud vendrá dada por

$$L \propto -ln(\sigma^2) - \frac{1}{2}ln|\mathbf{V}| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta)$$

Desempleo en USA (1947-1962)

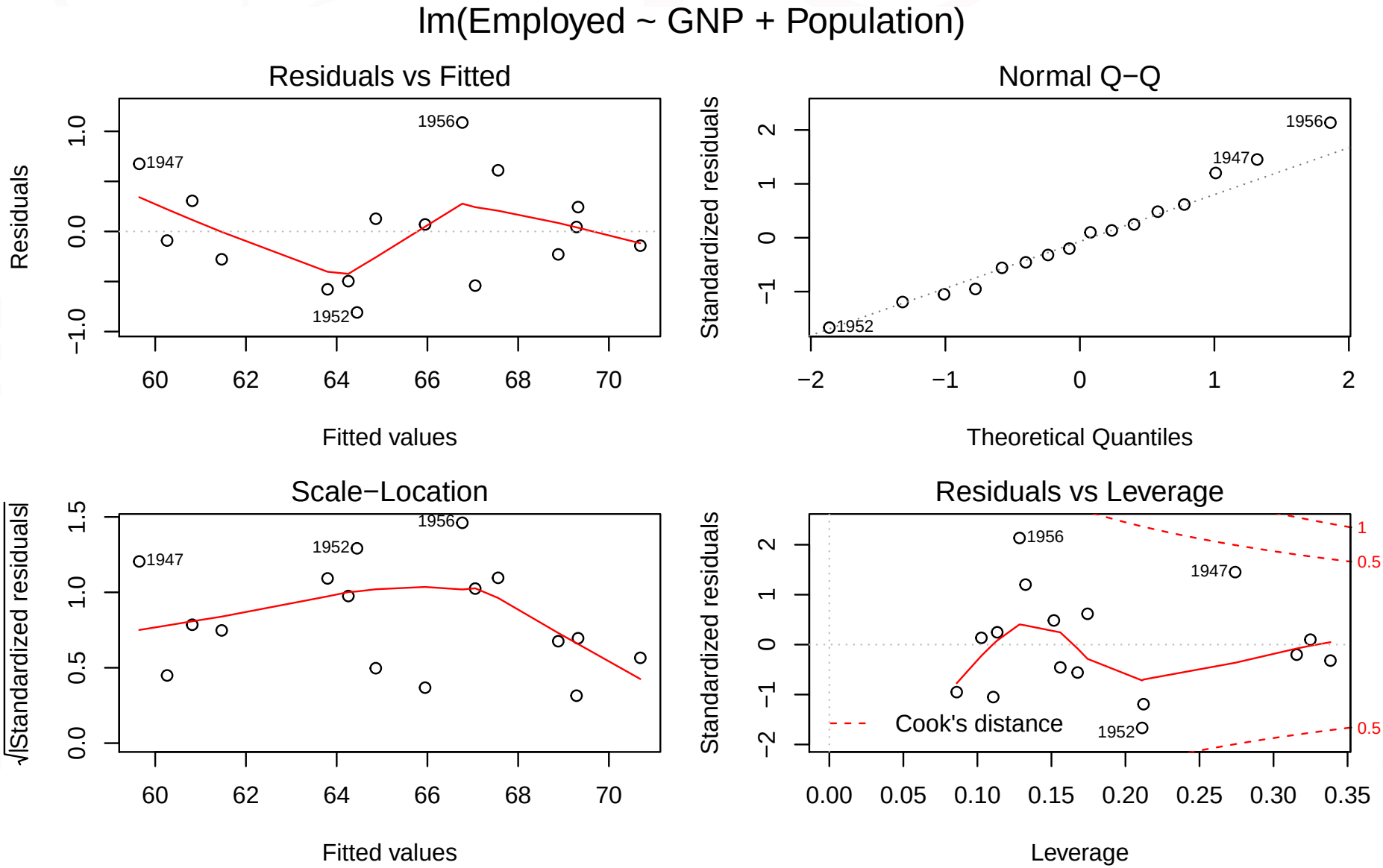
- Se dispone para el periodo 1947-1962 de 7 variables económicas:
  - **GNP**: Producto interior bruto (PIB)
  - **GNP.deflator**: PIB con deflacción de precio (1954=100)
  - **Unemployed**: Numeros desempleados
  - **Armed.Forces**: Número de personas en el ejercito
  - **Population**: población en edad laboral
  - **Year**: año
  - **Employed**: número de gente empleada

Desempleo en USA (1947-1962)

- Este ejemplo se utiliza en diversos libros de texto para mostrar un claro ejemplo de colinealidad y de autocorrelación. Sólo se utilizarán: *GNP*, *Population* y *Employed*.

```
Call:
lm(formula = Employed ~ GNP + Population, data = longley)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80899 -0.33282 -0.02329  0.25895  1.08800

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.93880   13.78503   6.452 2.16e-05 ***
GNP          0.06317    0.01065   5.933 4.96e-05 ***
Population  -0.40974    0.15214  -2.693   0.0184 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5459 on 13 degrees of freedom
Multiple R-squared:  0.9791,    Adjusted R-squared:  0.9758
F-statistic: 303.9 on 2 and 13 DF,  p-value: 1.221e-11

Correlation of Coefficients:
           (Intercept) GNP
GNP         0.98
Population -1.00       -0.99
```

Desempleo en USA (1947-1962)



Desempleo en USA (1947-1962)

- Habitualmente cuando los datos proviene de una serie temporal, los errores suelen tomar forma autorregresiva:

$$e_{t+1} = \rho e_t + w_t,\ w_t \sim \mathcal{N}(0, \tau^2)$$

se puede estimar ρ utilizando la correlación muestral de los residuos:

```
[1] 0.3104092
```

Desempleo en USA (1947-1962)

- Un modelo con errores autorregresivos tiene la matriz de covarianza de la siguiente forma

$$\mathbf{V}_{ij} = \rho^{|i-j|}$$

- Asuminedo que ρ es conocido e igual a 0.3104092. Entonces, $\mathbf{V}$ se calcula como

Desempleo en USA (1947-1962)

- Y la estimación por mínimos cuadrados generalizados será pues

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{y}$$

```
                   [,1]
(Intercept) 94.89887752
GNP          0.06738948
Population  -0.47427391
```

Desempleo en USA (1947-1962)

- El error estandar de $\hat{\beta}$ es ,

$$\sqrt{Var(\hat{\beta})} = \sqrt{\sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{y}}$$

```
## B(Intercept)         BGNP  BPopulation
##  94.89887752   0.06738948  -0.47427391
```

Desempleo en USA (1947-1962)

- El modelo GLS en *R* se escribe como parte del paquete de los modelos no-lineales:

```
library(nlme)
g<-gls(Employed~GNP+Population,
       correlation=corAR1(form=~Year),
       data=longley)
```

Desempleo en USA (1947-1962)

- Donde *correlation=* es la autocorrelación que le queremos dar a los errores:
  - ***corAR1*** : autoregressive process of order 1.
  - ***corARMA***: autoregressive moving average process, with arbitrary orders for the autoregressive and moving average components.
  - ***corCAR1***: continuous autoregressive process (AR(1) process for a continuous time covariate).
  - ***corCompSymm***: compound symmetry structure corresponding to a constant correlation.
  - ***corSymm***: general correlation matrix, with no additional structure.

Desempleo en USA (1947-1962)

```
Generalized least squares fit by REML
  Model: Employed ~ GNP + Population
  Data: longley
       AIC      BIC    logLik
  44.66377 47.48852 -17.33188

Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
      Phi
0.6441692

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 101.85813 14.198932  7.173647  0.0000
GNP           0.07207  0.010606  6.795485  0.0000
Population   -0.54851  0.154130 -3.558778  0.0035

 Correlation:
            (Intr) GNP
GNP          0.943
Population -0.997 -0.966

Standardized residuals:
      Min         Q1        Med         Q3        Max
-1.5924564 -0.5447822 -0.1055401  0.3639202  1.3281898

Residual standard error: 0.689207
Degrees of freedom: 16 total; 13 residual
```

Desempleo en USA (1947-1962)

```
    Durbin-Watson test

data:  fit.lm.as.gls
DW = 1.6152, p-value = 0.1198
alternative hypothesis: true autocorrelation is greater than 0


    Breusch-Godfrey test for serial correlation of order up to 6

data:  Residuals
LM test = 5.7272, df = 6, p-value = 0.4544
```

Ejemplo: Gasto en farmacia

```
    Durbin-Watson test

data:  regre4
DW = 1.3109, p-value = 1.523e-15
alternative hypothesis: true autocorrelation is greater than 0


    Breusch-Godfrey test for serial correlation of order up to 101

data:  Residuals
LM test = 198.31, df = 101, p-value = 2.6e-08
```

Ejemplo: Gasto en farmacia

- El modelo GLS en *R*se escribe como parte del paquete de los modelos no-lineales:

```
library(nlme)
regre4.gls<-gls(gasto~ trend + temp + temp2 + part,
                na.action=NULL ,
                correlation = corAR1(form = ~ 1))
```

Frente al modelo anterior (OLS)

```
regre4<-lm(gasto~ trend + temp + temp2 + part, na.action=NULL)
```

Ejemplo: Gasto en farmacia

```
Generalized least squares fit by REML
  Model: gasto ~ trend + temp + temp2 + part
  Data: NULL
       AIC      BIC    logLik
  3237.099 3266.643 -1611.55

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.6959633

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 3155.9776  584.6619  5.397953  0.0000
trend         -1.5295    0.2907 -5.260982  0.0000
temp          -0.0028    0.0426 -0.065110  0.9481
temp2          0.0148    0.0022  6.833991  0.0000
part           0.1498    0.0241  6.214143  0.0000

 Correlation:
      (Intr) trend  temp   temp2
trend -1.000
temp   0.049 -0.047
temp2 -0.013  0.013 -0.070
part  -0.053  0.051 -0.571  0.005

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.1195215 -0.7008765 -0.1425260  0.5490998  4.1999770

Residual standard error: 7.917382
Degrees of freedom: 508 total; 503 residual
```

Ejemplo: Gasto en farmacia

	OLS	GLS
(Intercept)	2881.743^{***}	3155.978^{***}
	(203.193)	(584.662)
trend	-1.396^{***}	-1.529^{***}
	(0.101)	(0.291)
temp	-0.472^{***}	-0.003
	(0.032)	(0.043)
temp2	0.023^{***}	0.015^{***}
	(0.003)	(0.002)
part	0.255^{***}	0.150^{***}
	(0.019)	(0.024)
R^2	0.595	
Adj. R^2	0.592	
Num. obs.	508	508
RMSE	6.385	
AIC		3237.099
BIC		3266.643
Log Likelihood		-1611.550

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Ejemplo: tpo. entre "caidas" en bolsa

```
regreBolsa1.gls<-gls( tiempo.entre ~ Duracion,  data=bolsa,
                  na.action=na.omit   ,
                  correlation = corAR1(form = ~ 1))
```

Frente al modelo anterior (OLS)

```
regre4<-lm(tiempo.entre~ Duracion, data=bolsa, na.action=na.omit )
```

Ejemplo: tpo. entre "caidas" en bolsa

```
Generalized least squares fit by REML
  Model: tiempo.entre ~ Duracion
  Data: bolsa
       AIC      BIC    logLik
  1505.775 1519.475 -748.8876

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
       Phi
-0.3550814

Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 38.33386 1.5372250 24.93706       0
Duracion     9.70735 0.4315217 22.49563       0

 Correlation:
         (Intr)
Duracion -0.979

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.8737852 -0.8196553 -0.1180854  0.5827529  4.5116691

Residual standard error: 6.834881
Degrees of freedom: 229 total; 227 residual
```

Ejemplo: tpo. entre "caidas" en bolsa

	OLS	GLS
(Intercept)	34.978^{***}	38.334^{***}
	(1.352)	(1.537)
Duracion	10.669^{***}	9.707^{***}
	(0.366)	(0.432)
R^2	0.789	
Adj. R^2	0.788	
Num. obs.	229	229
RMSE	6.740	
AIC		1505.775
BIC		1519.475
Log Likelihood		-748.888

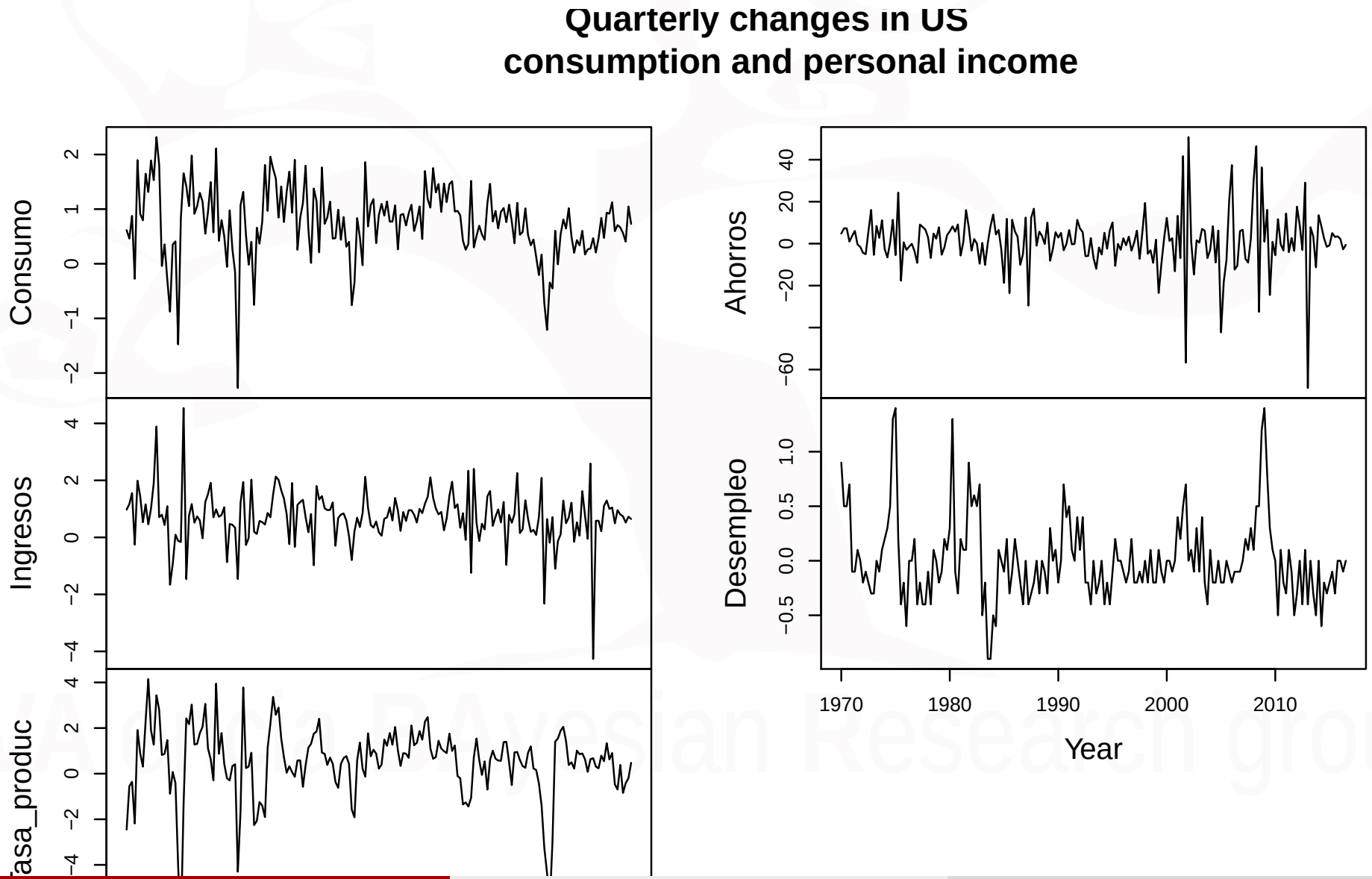
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Modelos ARIMA incluyendo covariables

Utilizando *xreg* dentro de *ARIMA*

```
#All data sets required for the examples and
# exercises in the book
# "Forecasting: principles and practice" by
# Rob J Hyndman and George Athanasopoulos

library(fpp) # Forecasting: principles and practice
library(fpp2) # Forecasting: principles and practice

plot(usconsumption, xlab="Year",
  main="Quarterly changes in US
    consumption and personal income")
```

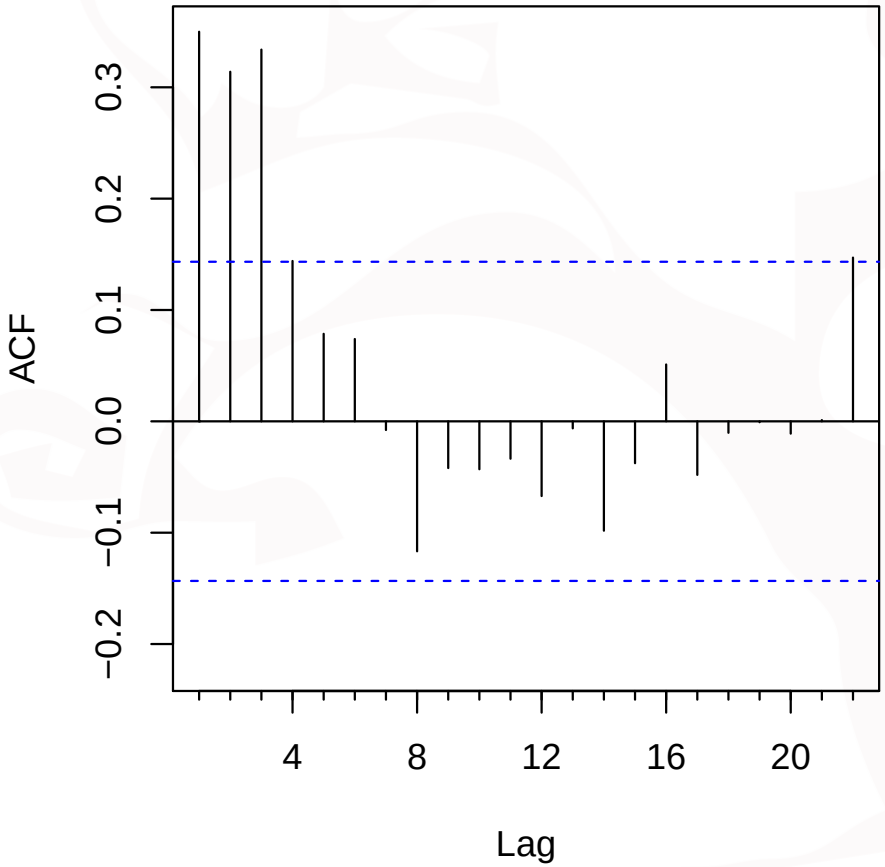
Utilizando *xreg* dentro de *ARIMA*

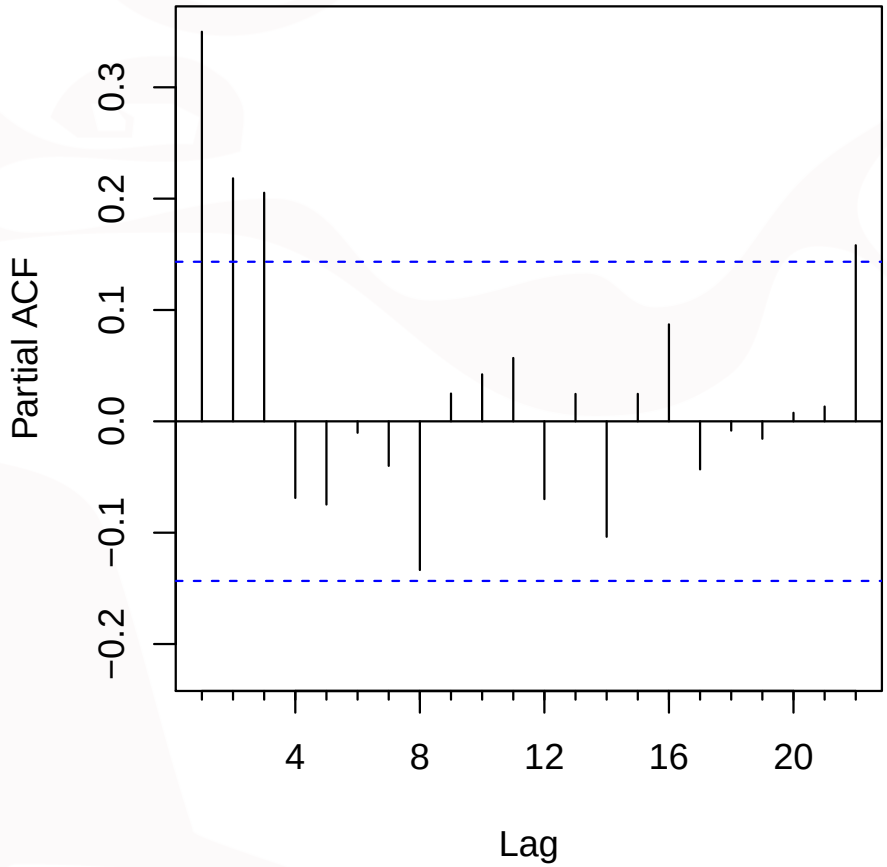


Correlogramas

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

Correlogramas





¿Podría ser un $AR(2)$ o $AR(3)$, o un $ARMA(1,1)$?

Ejemplo: US Personal Consumption and Income

Los datos

Consumo	Ingresos	Tasa_produc	Ahorros	Desempleo
0.6159862	0.9722610	-2.4527003	4.8103115	0.9
0.4603757	1.1690847	-0.5515251	7.2879923	0.5
0.8767914	1.5532705	-0.3587079	7.2890131	0.5
-0.2742451	-0.2552724	-2.1854549	0.9852296	0.7
1.8973708	1.9871536	1.9097341	3.6577706	-0.1
0.9119929	1.4473342	0.9015358	6.0513418	-0.1

US Comsumption: ajuste del modelo

```
fit <- auto.arima(uschange[,"Consumo"], # Serie Y_t
                  xreg=uschange[,"Ingresos"]) # covariable X
```

```
fit
```

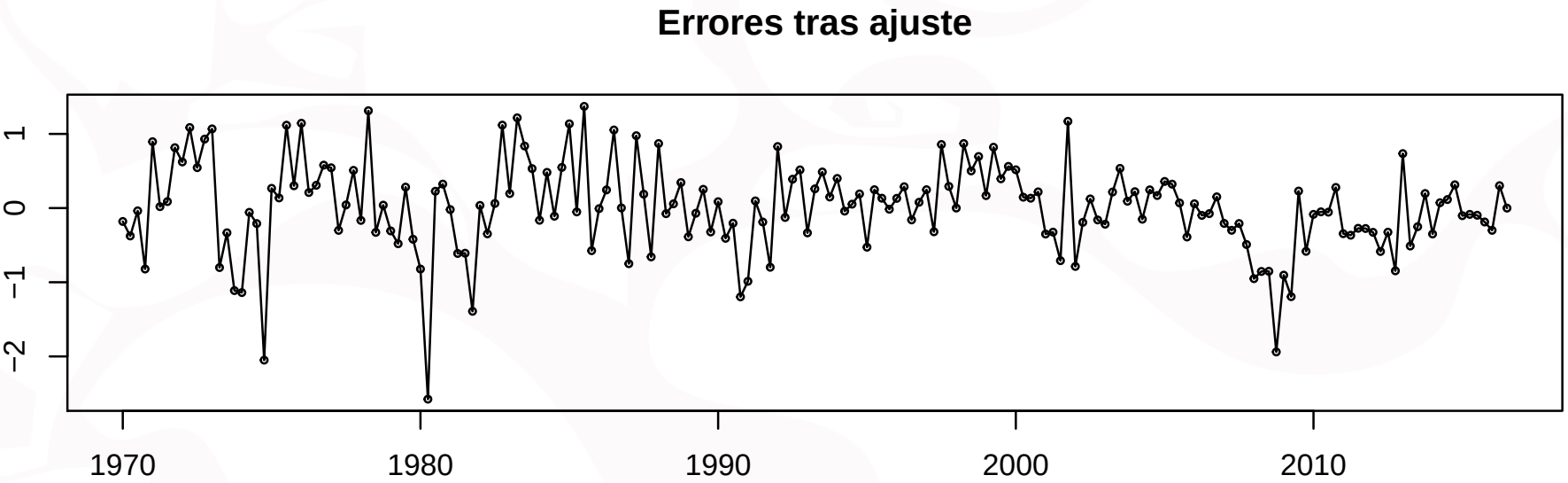
```
  Series: uschange[, "Consumo"]
  Regression with ARIMA(1,0,2) errors

  Coefficients:
           ar1      ma1     ma2  intercept    xreg
        0.6922  -0.5758  0.1984     0.5990  0.2028
  s.e.  0.1159   0.1301  0.0756     0.0884  0.0461

  sigma^2 estimated as 0.3219:  log likelihood=-156.95
  AIC=325.91   AICc=326.37   BIC=345.29
```

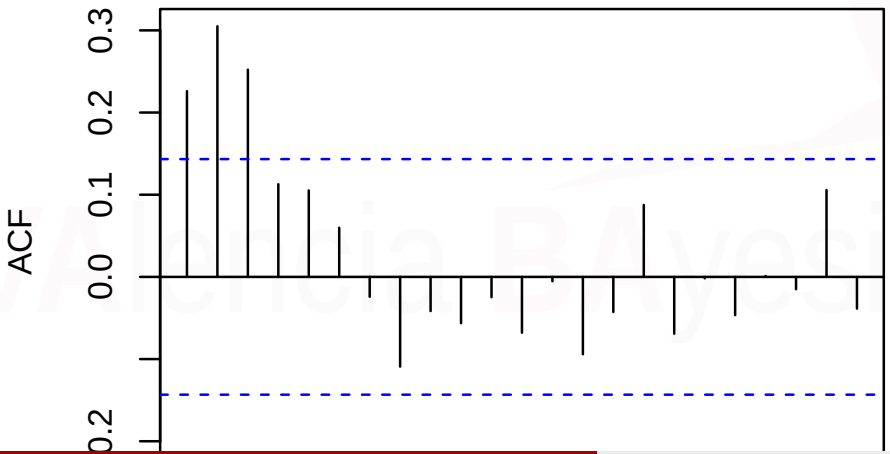
US Comsumption: ajuste del modelo

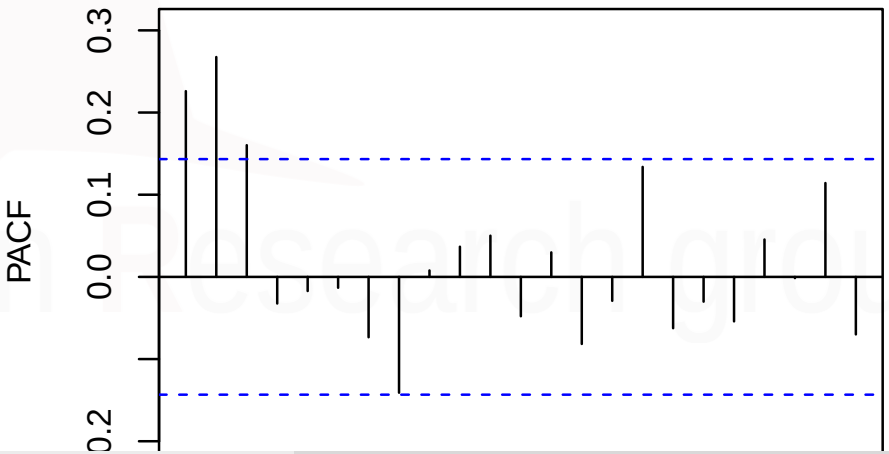
$$y_t = 0.599 + 0.203x_t + \eta_t,$$
$$\eta_t = 0.692\eta_{t-1} + \varepsilon_t - 0.576\varepsilon_{t-1} + 0.198\varepsilon_{t-2},$$
$$\varepsilon_t \sim \mathsf{NID}(0, 0.322).$$

]

US Comsumption: los residuos







US Comsumption: ¿residuos autocorrelados?

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

```
##
##  Box-Ljung test
##
## data:  residuals(fit)
## X-squared = 6.0465, df = 5, p-value = 0.3017
```

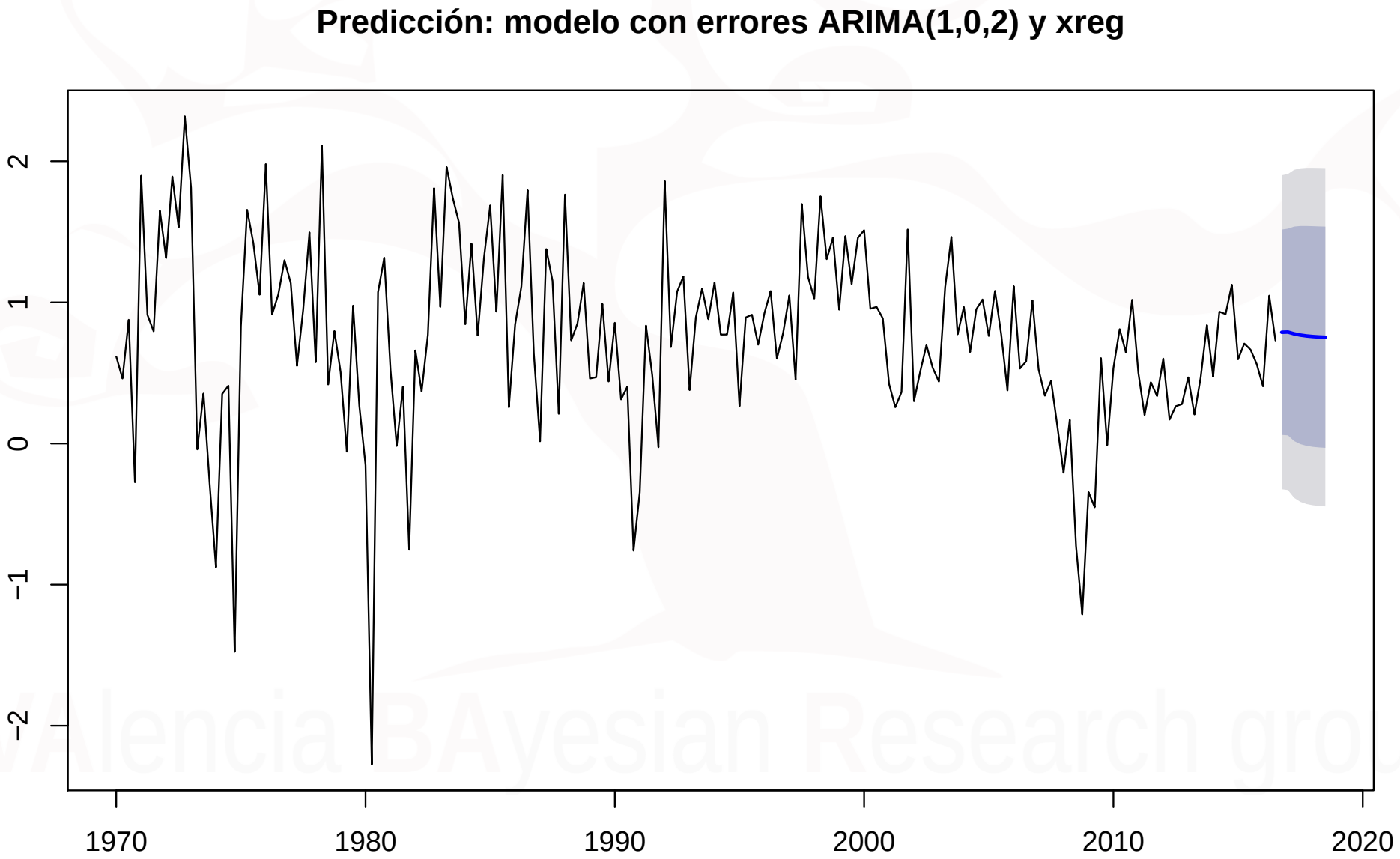
If the p value is greater than 0.05 then the residuals are independent which we want for the model to be correct.

¿Podemos predecir?

US Comsumption: predicción

```
fcast <- forecast(fit,
                  xreg=rep(mean(usconsumption[,2]),8), h=8)
#h= periodos a predecir
plot(fcast,
  main="Predicción: modelo con errores ARIMA(1,0,2) y xreg")
```

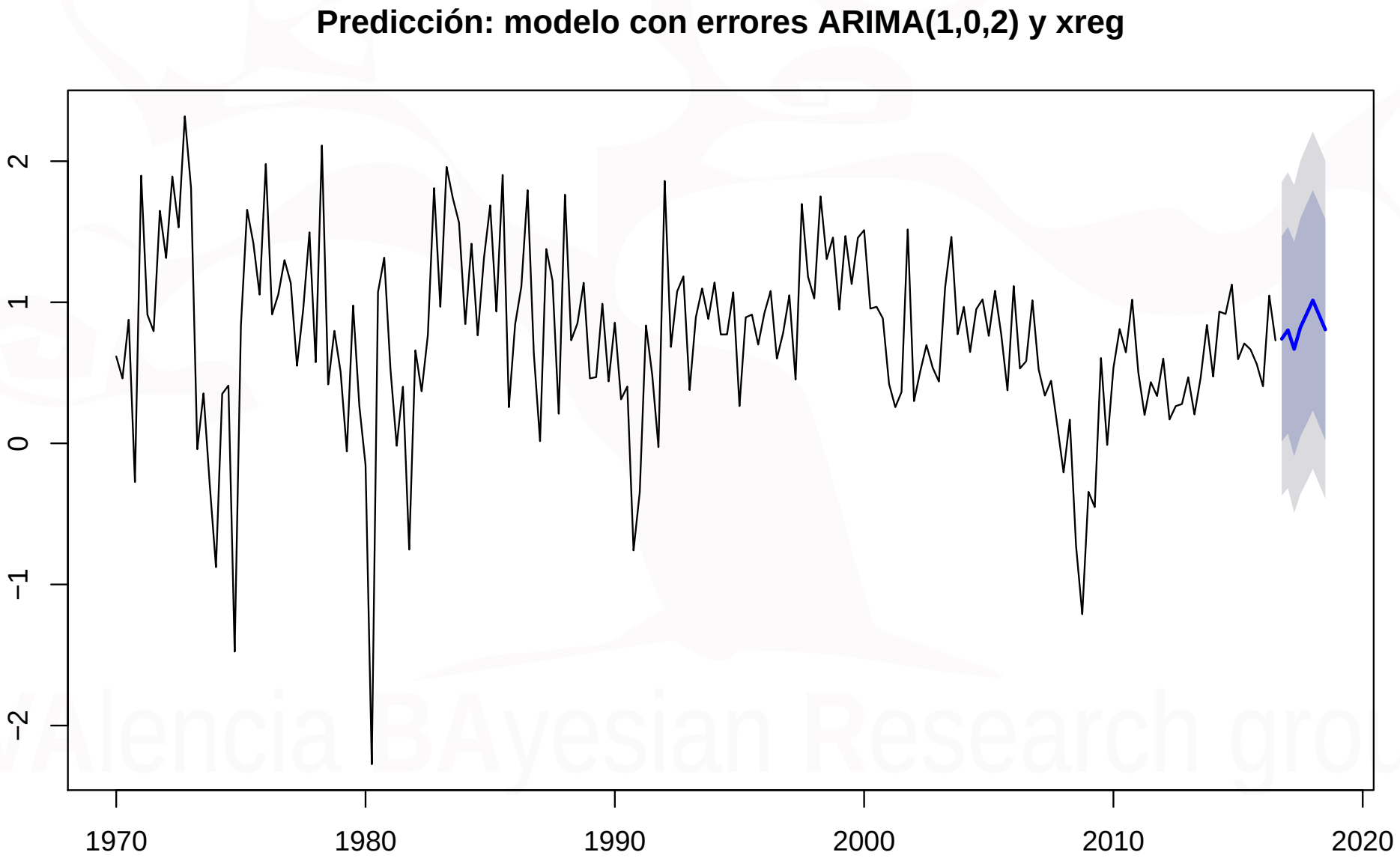
US Comsumption: predicción



US Comsumption: predicción

```
fcast <- forecast(fit,
                  xreg=c(0.5,0.8,0.2,1,1.5,2,1.5,1), h=8)
#h= periodos a predecir
plot(fcast,
  main="Predicción: modelo con errores ARIMA(1,0,2) y xreg")
```

US Comsumption: predicción



Covariables “no contínuas”

Datos: Compras de navidad

	Customer_Visit	Weekday	Christmas	Day
1	3000	1	0	1
2	2985	2	0	2
3	3112	3	0	3
4	3001	4	0	4
5	3003	5	0	5
6	2978	6	0	6
7	2904	7	0	7
8	3062	1	0	8
9	2964	2	0	9
10	3048	3	0	10
11	3096	4	0	11
12	3026	5	0	12
13	2950	6	0	13
14	2976	7	0	14
15	...	...	...	...
40	3042	5	0	40
41	2998	6	1	41
42	3037	7	0	42
43	3084	1	0	43
44	2928	2	0	44
45	2903	3	0	45
46	2885	4	0	46
47	2973	5	0	47

Datos: Compras de navidad

- Como podéis ver, hay una variable que es “Navidad”, otra variable que es día de la semana pero en números de 1 a 7.
- Por lo que habrá que decirle al comando que tenga en cuenta que estas variables son factores.
- ¿cómo se hace esto?

Nosotros queremos ajustar el siguiente modelo:

$$Y_t = \beta_0 + \beta_1 \text{Weekday} + \beta_2 \text{Day} + \beta_3 \text{Christmas} + ARMA(p, q)$$

model.matrix

Debemos pasar nuestras variables factor a variables “dummies”. Una variable dummie es una varible que se crea a partir de una variable nominal y que identifica sólo a una de sus caterias. Es decir, tenemos que crear diferentes variables dummie para identificar los 7 días de la semana.

```
diaSemana <- model.matrix(~as.factor(modelfitsample$Weekday))
diaSemana <- diaSemana[, -1]
colnames(diaSemana) <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat")
kable(head(diaSemana, n = 10))
```

model.matrix

```
diaSemana <- model.matrix(~as.factor(modelfitsample$Weekday))
diaSemana <- diaSemana[, -1]
colnames(diaSemana) <- c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat")
```

Como podeis ver, hemnos generado **6** variables donde antes sólo habia una. Falta el Domingo, pero el sistema toma como Domingo cuando no es ninguno de lso otros días, es decir, cuando Mon . . . Sat valen cero.

Ajuste del modelo

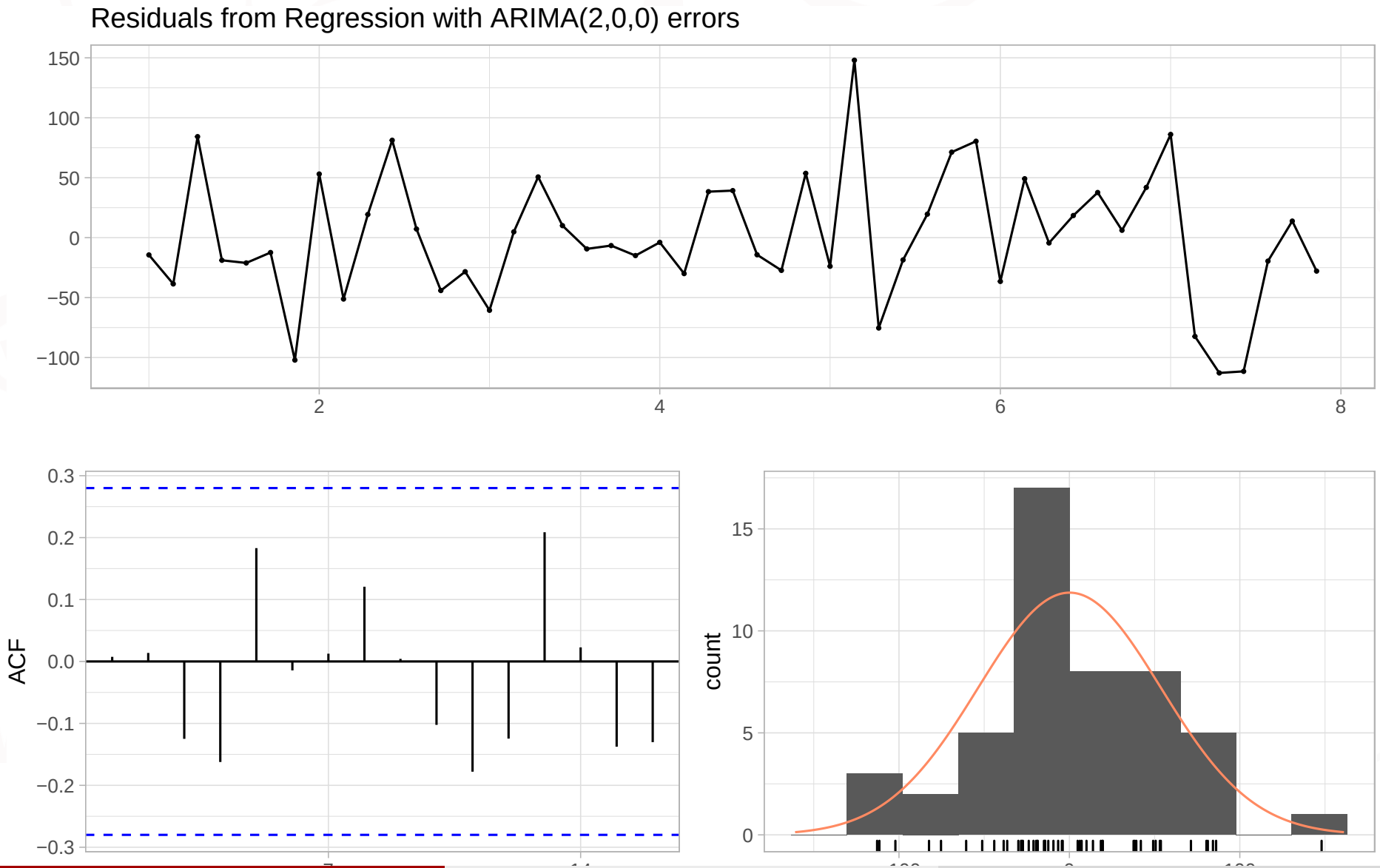
```
# Creamos las variables del modelo lineal
Variables.xreg <- cbind(Weekday=diaSemana,
                        Day=modelfitsample$Day,
                    Christmas=modelfitsample$Christmas)

# Le decimos a R cual es nuestra varia de "Time Series"
visits <- ts(modelfitsample$Customer_Visit, frequency=7)

# Ajustamos el modelo  ARIMAX (ARMA(2,0) a modo de ejemplo)
modArima <- Arima(visits, xreg=Variables.xreg, order=c(2,0,0))
```

Ajuste del modelo

Los resultados del ajuste (parámetros y error):

```
## Series: visits
## Regression with ARIMA(2,0,0) errors
##
## Coefficients:
##          ar1     ar2  intercept      Mon      Tue      Wed      Thu
##       0.0162  0.0637  3014.8611   9.8830  15.7114   7.8297   6.3939
## s.e.  0.1453  0.1485    24.0485  28.2263  27.5547  28.4301  28.4642
##           Fri      Sat      Day  Christmas
##       -20.3970  -4.5479  -0.4230    12.7212
## s.e.   29.0857  28.4466   0.5908    62.3499
##
## sigma^2 estimated as 3640:  log likelihood=-264.19
## AIC=552.39   AICc=561.06   BIC=575.09
##
## Training set error measures:
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 0.07679988 53.12831 41.33195 -0.02859856 1.373526 0.7221057
##                     ACF1
## Training set 0.0074646
```

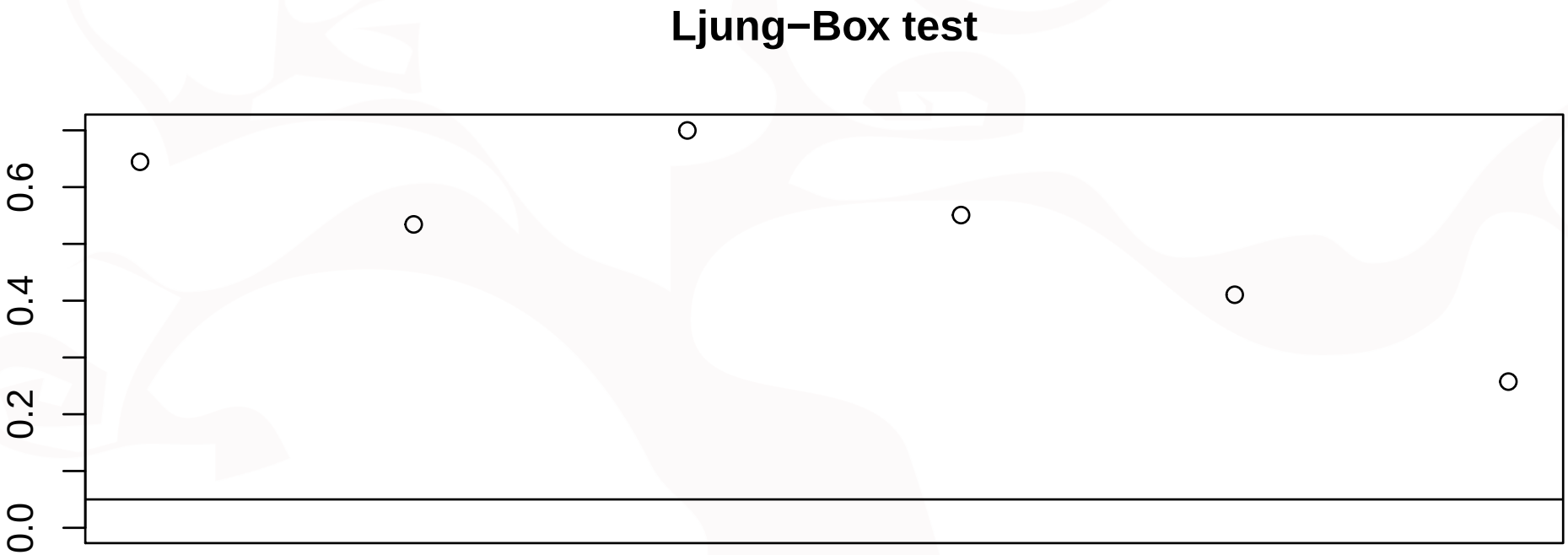
Ajuste del modelo: checkresiduals



Ajuste del modelo: checkresiduals

```
library("portes")

testresid<-LjungBox(fit)

plot(testresid[,4],
     ylim = c(0.001,max(testresid[,4])),
      xlab="", ylab="p-values",
     xaxt='n', ann=FALSE))

abline(h=0.05)
title("Ljung-Box test")
```

Ajuste del modelo: checkresiduals




```
## [1] "Ho: Los residuos son independientes"
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit)
## X-squared = 6.0465, df = 5, p-value = 0.3017
```

Predicción del modelo

El modelo es mejorable, pero esto es simplemente un ejemplo.

y como resultado, tambien podemos predecir

```
# hay que darle valores "X" para que haga la predicción
# voy a predecir un valor para el lunes sin ser Navidad del día 43
# por que tengo Mon=1 Tue=0 Wed Thu Fri Sat=0 Day=43 Xmas=0
variables.prediccion<-data.frame(1,0,0,0,0,0,43,0)
colnames(variables.prediccion)<- colnames(Variables.xreg)
prediccion<-forecast(modArima,
                     xreg=as.matrix(cbind(variables.prediccion)) )
```

Predicción del modelo

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
8	3006.449	2929.133	3083.765	2888.205	3124.693

Predicción del modelo

Y dibujamos la predicción

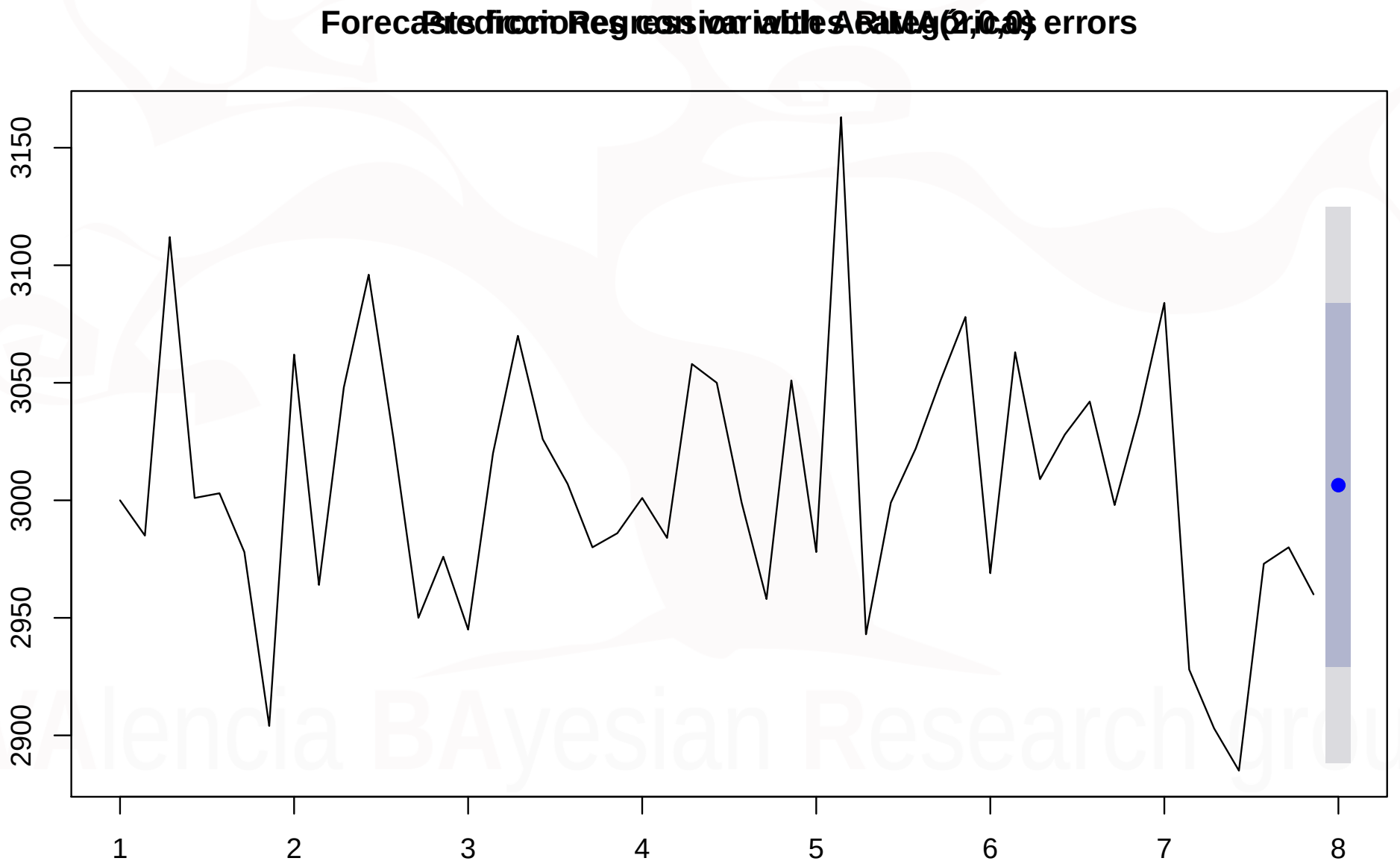
Opción 1

```
plot(prediccion)
```

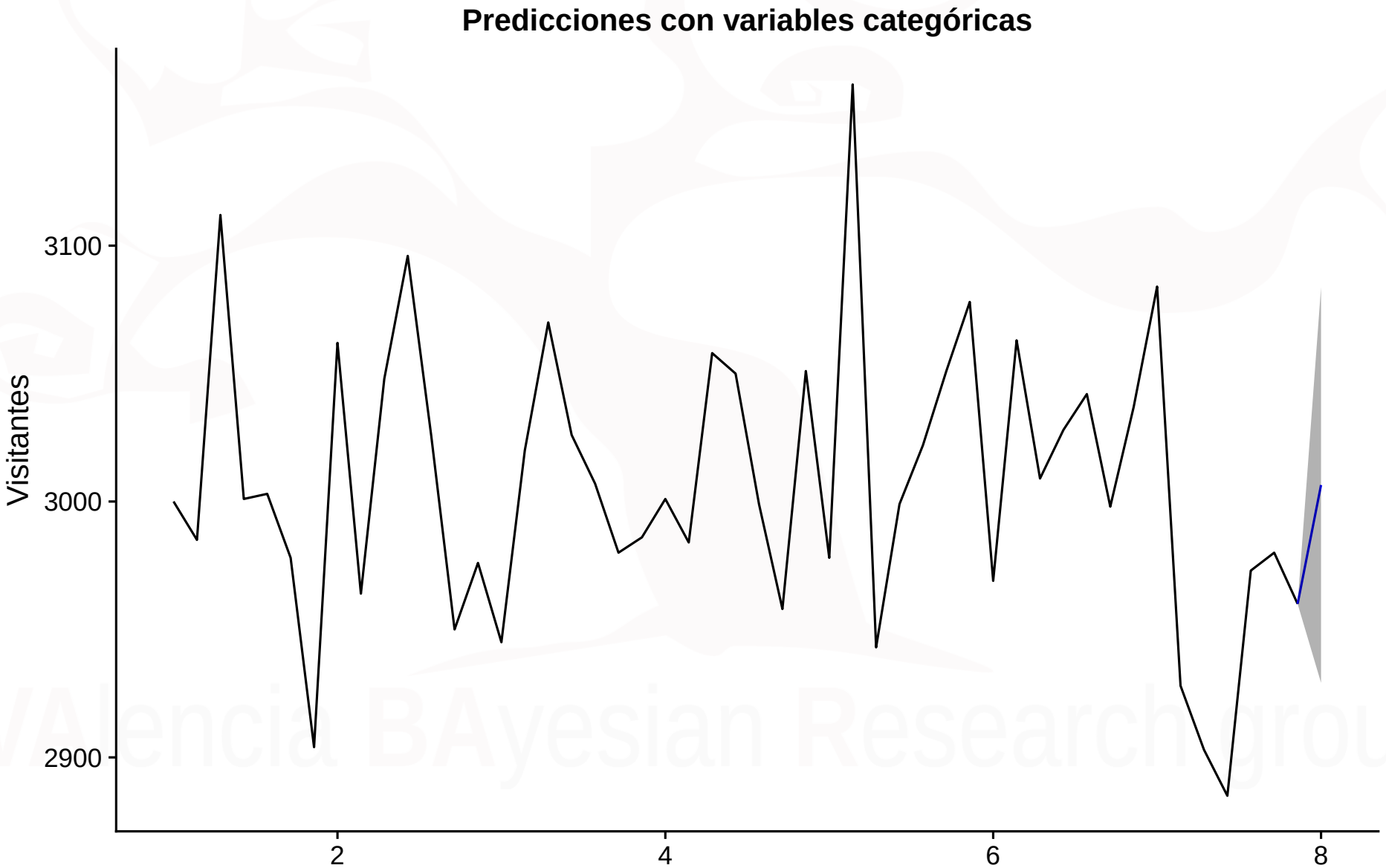
Opción 2

```
library(ggfortify)
autoplot(prediccion) + xlab("Día")+
                       ylab("Visitantes")
```

Predicción del modelo



Predicción del modelo

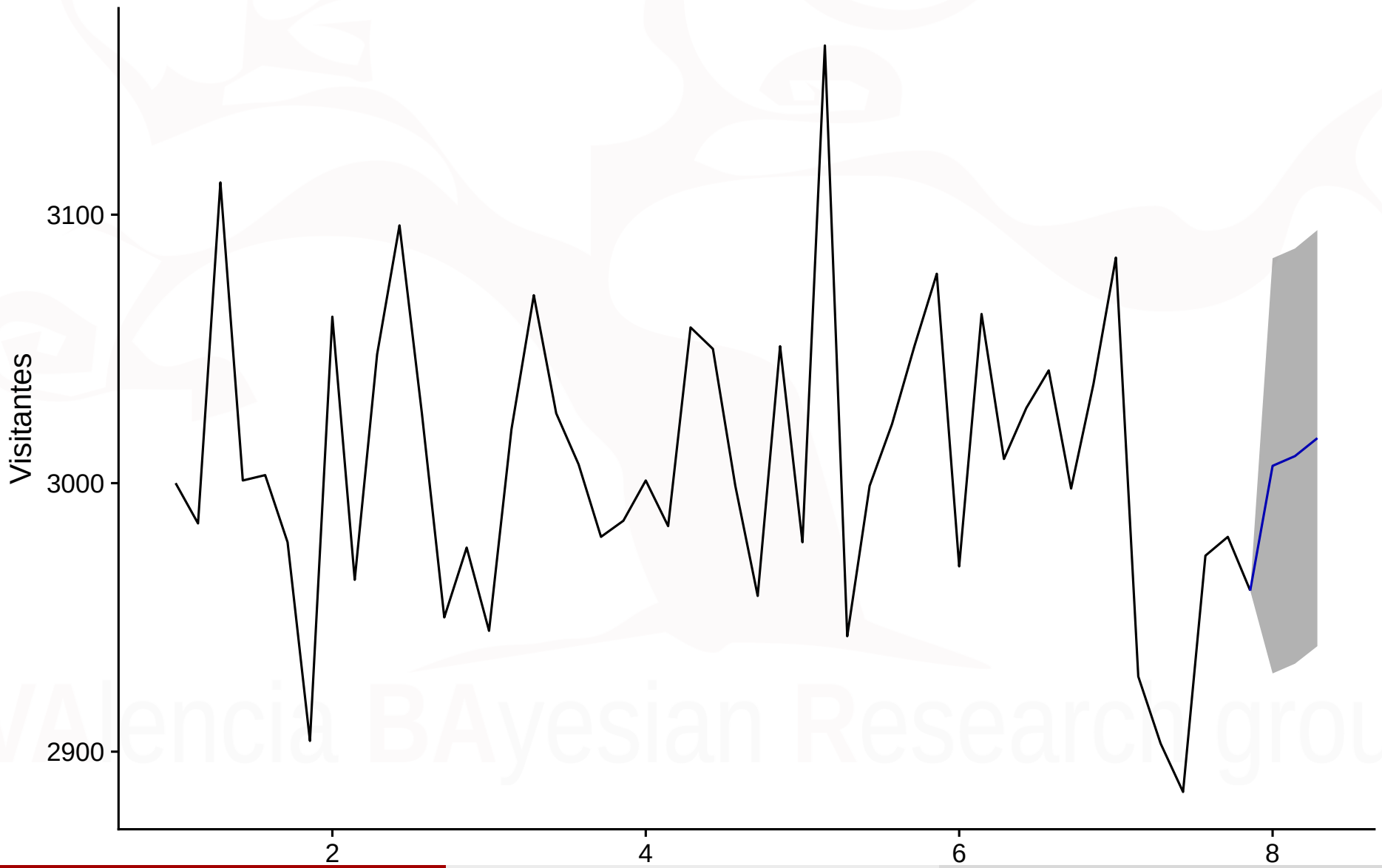


Más predicciones

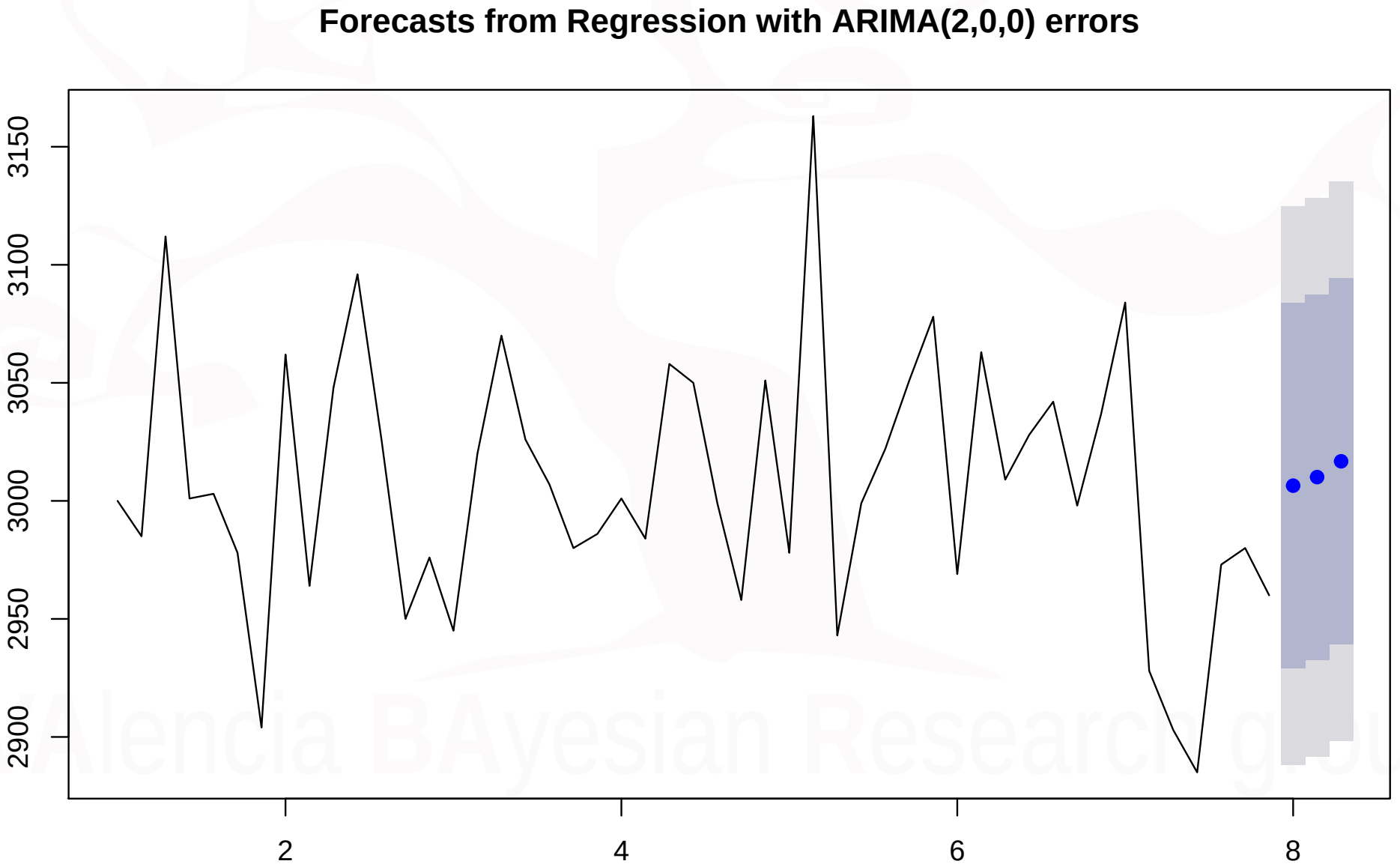
Vamos a pedir más predicciones, según valores de “X”

```
## hay que darle valores "X" para que haga la predicción

#voy a predecir tres valores lunes, martes y miercoles
# y alguno que sea Navidad (el miércoles día 44).
dia1<-c(1,0,0,0,0,0,43,0)
dia2<-c(0,1,0,0,0,0,44,0)
dia3<-c(0,0,1,0,0,0,44,1)  #Navidad
covar.pred<-as.matrix(rbind(dia1, dia2, dia3 ))
colnames(covar.pred)<- colnames(Variables.xreg)

variables.prediccion<-data.frame(rbind(dia1,dia2,dia3))

prediccion2<-forecast(modArima,
                      xreg=covar.pred)
```

Predicción del modelo



Predicción del modelo



Escenarios

- Una vez visto esto, podemos estar interesados en generar escenarios Pesimistas, neutros y Optimistas.
- Para ello lo que debemos de hacer es darle valores las **covariables** según deseemos crear esos escenarios (más altos o más bajos según el coeficiente que acompañe en el modelo).