Incumplimiento de las hipótesis básicas del modelo de regresión con R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211047

Universitat Oberta de Catalunya

Índice

Introducción

En el primer módulo hemos estudiado cómo implementar el estimador de Mínimos Cuadrados Ordinarios (MCO) para obtener una estimación de los parámetros de un modelo de regresión. El estimador MCO es la manera más simple y directa de obtener una estimación, pero para que esta sea válida es necesario que se cumplan una serie de requisitos (o restricciones) en los datos y en el modelo construido. Desafortunadamente, muy a menudo estos requisitos no se cumplen, de modo que es necesario acudir a otras técnicas para obtener una estimación fiable.

El primer capítulo de este módulo es un repaso teórico de las propiedades de la estimación de un modelo econométrico: ortogonalidad, sesgo, consistencia y eficiencia. El segundo capítulo se encarga del problema de la eficiencia de una estimación, esto es, su varianza. De esta manera, se introducen las definiciones de heteroscedasticidad y autocorrelación, fenómenos que hacen que la matriz de varianzas y covarianzas de la estimación no sea esférica. Además, con un ejemplo se estudia cómo detectar y corregir estos fenómenos con R y con R-Commander. El tercer capítulo analiza el fenómeno de errores en la muestra. La primera parte estudia la multicolinealidad, fenómeno que aparece cuando entre los regresores hay variables altamente correlacionadas entre sí, lo que dificulta la estimación y muestra resultados erróneos. La segunda parte analiza qué sucede cuando hay observaciones atípicas, esto es, muy alejadas del resto de las observaciones. Por último, el cuarto capítulo está dedicado al análisis de la permanencia estructural, es decir, si una misma estimación es válida para todos los datos de la muestra o, por el contrario, hay que dividir la muestra en varios fragmentos, ya que entre estos se detecta una relación funcional distinta.

Objetivos

1. Comprender todas las características de un modelo de regresión lineal, así como las propiedades de la estimación por mínimos cuadrados ordinarios (MCO) del mismo.

2. Entender cuál es la condición de ortogonalidad, y por qué es fundamental para el resultado de la estimación.

3. Saber diferenciar y explicar las propiedades sesgo, consistencia y eficiencia de una estimación econométrica.

4. Estudiar las propiedades de la estimación de la varianza de un modelo, esto es, la esfericidad de la matriz de varianzas y covarianzas.

5. Saber relacionar la no esfericidad de la matriz de varianzas y covarianzas con los problemas de heteroscedasticidad y autocorrelación.

6. Poder identificar la presencia de multicolinealidad entre los regresores de un modelo de regresión, además de dominar las técnicas pertinentes para solucionarlo.

7. Ser capaz de detectar la presencia de observaciones atípicas o outliers, y poder tenerlo en cuenta a la hora de efectuar la estimación econométrica.

8. Dominar las herramientas que permiten detectar una posible rotura de la permanencia estructural, así como poder efectuar estimaciones más adecuadas partiendo la muestra en diferentes partes.

1. Propiedades de la estimación del modelo

1.1. Estimador MCO y la condición de ortogonalidad

Errores y residuos

Es muy importante tener presente la diferencia entre los errores del modelo de regresión e_i y los residuos resultantes de la estimación del modelo $\hat{e}_i$.

Hay un aspecto muy importante del estimador MCO que hay que tener en cuenta. Por construcción, el estimador MCO garantiza la condición de ortogonalidad. Dicho de otra manera, una vez obtenemos los residuos de la estimación del modelo de regresión

$$\hat{e}_i = y_i - x_i'\hat{\beta},$$

siendo su expresión matricial

$$\hat{e} = Y - X\hat{\beta},$$

es imposible verificar si se cumple la condición $E(X'e) = 0$, ya que el estimador de los parámetros hace que se cumpla que:

$$X'\hat{e} = X'(Y - X\hat{\beta}) = X'Y - X'X(X'X)^{-1}X'Y = X'Y - X'Y = 0.$$

Con lo cual, el investigador deberá determinar si se cumple la condición de ortogonalidad considerando otros criterios, tema que se abordará más adelante.

1.2. Sesgo y consistencia

El estimador $\hat{\beta}$ es un estadístico, y como tal tiene una distribución. En general, esta distribución es desconocida. Si asumimos que los errores siguen una distribución normal, podemos establecer que el estimador también sigue esa distribución.

Antes de definir el sesgo y la consistencia de un estimador, resulta útil relizar la siguiente descomposición del estimador MCO:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + e) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'e \\ &= \beta + (X'X)^{-1}X'e\end{aligned}$$

Esta descomposición muestra cómo la distribución de $\hat{\beta}$ está determinada únicamente por la distrubución conjunta de (x_i, e_i).

El **sesgo** del estimador será la esperanza matemática de la diferencia entre el valor esperado del estimador y el parámetro del modelo, es decir, $E(\hat{\beta} - \beta)$. En el momento en el que se cumple $E(\hat{\beta} - \beta) = 0$, o bien $E(\hat{\beta}) = \beta$, el estimador $\hat{\beta}$ es **insesgado**. Si tomamos la expresión del estimador que viene dada por (1,29), vemos que si se cumple la condición $E(X'e) = 0$, es decir, si se cumple la condición de ortogonalidad, el estimador será insesgado:

$$E(\hat{\beta}) = \beta + E((X'X)^{-1})E(X'e) = \beta$$

El concepto de **consistencia** hace referencia a la convergencia en probabilidad del estimador con los verdaderos parámetros del modelo de regresión, a medida que el tamaño muestral n tiende a infinito. Siguiendo esta definición, diremos que el estimador $\hat{\beta}$ es consistente si se cumple que $plim_{n\to\infty}(\hat{\beta}) = \beta$, es decir, si el estimador converge en probabilidad con el verdadero parámetro del modelo.

Según la teoría asintótica, podemos entender el concepto de convergencia en probabilidad como el límite que alcanza una determinada secuencia de valores a medida que incrementa el conjunto de información.

Así pues, afirmamos que el estimador será consistente si el error es asintóticamente ortogonal a los regresores, es decir:

$$plim_{n\to\infty}\left(\frac{X'e}{n}\right) = 0$$

plim significa límite en probabilidad. Una notación también usada que indica convergecia en probabilidad de una variable con otra es $\hat{\beta} \longrightarrow_p \beta$.

En este caso, se cumplirá que:

$$plim_{n\to\infty}(\hat{\beta}) = \beta + plim_{n\to\infty}\left[\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'e}{n}\right)\right]$$

$$= \beta + plim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} plim_{n\to\infty}\left(\frac{X'e}{n}\right)$$

$$= \beta$$

Conviene recordar que, aunque un estimador sea sesgado, es decir, $E(x_i e_i) \neq 0$, es posible que asintóticamente el error sea ortogonal a los regresores, de manera que $plim_{n\to\infty}\left(\frac{X'e}{n}\right) = 0$, siendo en este caso el estimador sesgado pero consistente. Ahora bien, un estimador inconsistente **siempre** será sesgado.

1.3. Eficiencia

La eficiencia de un estimador es una propiedad que hace referencia a su varianza. Un estimador será **eficiente** si alcanza una varianza mínima entre otros posibles estimadores de los parámetros del modelo. Si retomamos el modelo de regresión lineal:

$$y_i = x_i'\beta + e_i$$

$$E(e_i|x_i) = 0$$

vemos cómo estamos imponiendo la condición de que la esperanza condicional del error es nula, siendo esta varianza condicional del modelo:

$$E(e_i^2|x_i) = \sigma_i^2$$

En el siguiente capítulo analizamos en detalle los casos particulares en los que un estimador no será eficiente, esto es, en presencia de *heteroscedasticidad* y/o autocorrelación.

1.4. Causas del sesgo y de la inconsistencia

Como se demuestra en la sección anterior, el estimador MCO garantiza la ortogonalidad de los regresores con los residuos, de manera que $E(X'\hat{e}) = 0$, por lo que es imposible saber a partir de dicha estimación si el error del modelo está correlacionado con los regresores. Dicho de otra manera, el análisis de los residuos de la regresión no contiene información sobre el sesgo y la consistencia de la estimación. En esta sección se detallan las dos situaciones en las que no se cumplen las condiciones de ortogonalidad: errores de medida y endogeneidad.

1.4.1. Errores de medida

Supongamos que disponemos del siguiente modelo de regresión lineal esférico, en el que tenemos un solo regresor (la variable x_i^*):

$$y_i = \alpha + \beta x_i^* + e_i$$

$$E(e_i|x_i^*) = 0$$

$$E(e_i^2|x_i^*) = \sigma^2$$

Si dispusiéramos de datos para las variables (y_i, x_i^*), y suponiendo que se cumplieran los dos supuestos del modelo, la estimación MCO sería (1) insesgada, (2) consistente

y (3) eficiente. Desafortunadamente, vamos a suponer que medimos el regresor con error, de manera que no observamos x_i^*, sino x_i:

$$x_i = x_i^* + v_i$$

Supongamos, además, que el error de medida v_i es una variable aleatoria, con media cero y varianza constante, no correlacionada ni con el error de la regresión ni con la auténtica variable que no podemos observar x_i^*:

$$E(v_i) = 0,$$

$$E(v_i^2) = \sigma_v^2,$$

$$E(v_i e_i) = 0,$$

$$E(v_i|x_i^*) = 0.$$

En este caso, ¿cómo afecta este error de medida en la estimación? Bien, introduzcamos el error de medida en el modelo de regresión lineal:

$$\begin{aligned} y_i &= \alpha + \beta(x_i - v_i) + e_i \\ &= \alpha + \beta x_i - \beta v_i + e_i \\ &= \alpha + \beta x_i + u_i, \\ u_i &= e_i - \beta v_i \end{aligned}$$

Introduciendo el error de medida en el modelo, vemos que el error del modelo pasa a ser $u_i = e_i - \beta v_i$. Con este error, comprobamos que la condición de ortogonalidad no se cumple:

$$\begin{aligned} E(x_i u_i) = Cov(x_i, u_i) &= Cov(x_i^* + v_i, e_i - \beta v_i) \\ &= -\beta\, Cov(v_i, v_i) = -\beta\sigma_v^2 \end{aligned}$$

Esto implica que la estimación por MCO sea sesgada e inconsistente. El estimador MCO se puede expresar de la siguiente manera:

$$\hat{\beta}_{MCO} = \frac{(1/n)\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(1/n)\sum_{i=1}^n (x_i - \bar{x})^2} = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

A continuación, analizamos la consistencia del estimador:

$$plim_{n\to\infty}\hat{\beta}_{MCO} = \beta + \frac{plim(1/n)\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{plim(1/n)(x_i - \bar{x})^2}$$

$$= \beta + \frac{Cov(x_i, u_i)}{Var(x_i)} = \beta + \frac{-\beta\sigma_v^2}{\sigma_{x^*}^2 + \sigma_v^2}$$

$$= \beta\left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2}\right)$$

Como podemos observar, en este caso el error de medida provoca un sesgo en la estimación del parámetro hacia cero, es decir, la estimación del parámetro muestra un valor inferior al del verdadero valor. Dicho sesgo crece a medida que la varianza del error σ_v^2 aumenta. Además, en el caso de que tuviéramos un modelo con varios parámetros, las estimaciones de todos ellos se verían afectadas, aun cuando el error de medida se diera en una sola variable. Cabe añadir que si hay más de un regresor medido con error, no se conocerá la dirección del sesgo.

1.4.2. Endogeneidad

Técnicamente, los conceptos de exogeneidad y predeterminación no son exactamente equivalentes, aunque nosotros usemos ambos términos de manera indistinta.

Uno de los supuestos en que nos basamos a la hora de plantear un modelo de regresión hace referencia a los regresores. Estos han de ser exógenos o predeterminados, es decir, no ha de haber ningún elemento en el modelo que los determine. Un ejemplo de endogeneidad se da en los modelos de ecuaciones simultáneas, en los que los regresores de una ecuación son generados en otras ecuaciones con una componente estocástica. Otro ejemplo lo encontramos en los modelos que consideran datos temporales cuando uno de los regresores es la variable endógena retardada, esto es:

$$y_t = \beta x_t + \gamma y_{t-1} + e_t$$

Este tipo de modelos **siempre** será sesgado, es decir, tendremos sesgo por endogeneidad. Ahora bien, dependiendo de cuál sea la estructura del error, las propiedades asintóticas del error serán unas u otras. Supongamos que el modelo es esférico, de manera que el error se caracteriza por:

$$e_t \sim iid(0, \sigma^2 I_n)$$

En este caso, si analizamos la covarianza entre regresor y error, obtenemos:

$$Cov(y_{t-1}, e_t) = Cov(\beta x_{t-1} + \gamma y_{t-2} + e_{t-1}, e_t) = Cov(e_{t-1}, e_t) = 0$$

De este modo, obtenemos consistencia en el estimador:

$$plim_{n\to\infty}(\hat{\beta}) = \beta + plim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} plim_{n\to\infty}\left(\frac{X'e}{n}\right) = \beta$$

Ahora bien, supongamos que el término de error está correlacionado, y sigue una estructura autorregresiva, con lo que el error ya no es esférico:

$$e_t = \rho\, e_{t-1} + u_t,$$

$$u_t \sim iid(0, \sigma_u^2 I_n)$$

Fijémonos en que en este caso el modelo incorpora el regresor estocástico $\rho\, e_{t-1}$. A la hora de analizar la covarianza entre regresor y error, obtenemos:

$$Cov(y_{t-1}, e_t) = Cov(\beta x_{t-1} + \gamma y_{t-2} + e_{t-1}, \rho\, e_t + u_t) = \rho\, Cov(e_{t-1}, e_{t-1}) = \rho\, \sigma^2$$

En este caso, la estimación ya no es consistente, dado que

$$plim_{n\to\infty}(\hat{\beta}) \neq \beta$$

2. Heteroscedasticidad y autocorrelación

2.1. Definición teórica

En primer lugar, definamos la matriz de varianzas y covarianzas del error del modelo de regresión:

$$MVC(e) = E(ee') = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}_{n \times n}$$

Los elementos de la diagonal son las varianzas de los errores, y fuera de la diagonal están situadas las covarianzas. Aquí nos podemos encontrar ante varias situaciones:

- **Elementos de la diagonal**. El modelo de regresión lineal es **homoscedástico** si los elementos de la diagonal son todos idénticos, esto es, si se cumple que $\sigma_i^2 = \sigma^2$. En este caso, la esperanza del cuadrado del error no varía a través de los elementos muestrales. En cambio, estaremos ante un modelo de regresión lineal **heteroscedástico** si se cumple que $\sigma_i^2 = \sigma^2(x_i)$, es decir, si σ_i^2 varía para cada elemento i.

- **Elementos fuera de la diagonal**. Si estos no son nulos, esto es, $\sigma_{ij} \neq 0,\ \forall i \neq j$, el modelo de regresión está **autocorrelacionado**, y análogamente si son nulos, el modelo no estará autocorrelacionado.

Partiendo de estas definiciones, decimos que estamos ante un **modelo de regresión lineal esférico** (también se suele denominar un modelo de regresión lineal con una matriz de varianzas y covarianzas esférica) si la matriz de varianzas y covarianzas es homoscedástica y no correlacionada, de manera que podemos expresar la matriz de varianzas y covarianzas como:

$$MVC(e) = E(ee') = \sigma^2 I_n$$

Siendo I_n la matriz identidad de dimensión $n \times n$. En este caso, la estimación del modelo por MCO es **eficiente**.

El hecho de estar ante un modelo de regresión lineal homoscedástico o heteroscedástico tiene implicaciones a la hora de valorar tanto los parámetros del modelo como la matriz de varianzas y covarianzas. Esto es, en presencia de heteroscedasticidad y/o autocorrelación, tendremos un **modelo de regresión lineal no esférico**. En este caso, la estimación del modelo por MCO no será eficiente, ya que no estaremos incorporando la estructura del error en la estimación de los parámetros. En este caso, el *Teorema de Gauss-Markov* establece que el mejor estimador lineal insesgado y de mínima varianza es el de **Mínimos Cuadrados Generalizados (MCG)**. Así, suponiendo que la matriz de varianzas y covarianzas adquiere la forma $MVC(e) = E(ee') = \Omega$, este estimador se define como:

$$\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

Frecuentemente no se conoce Ω, por lo que se ha de valorar (o bien directamente o bien imponiendo una estructura). Una vez obtenemos la estimación $\hat{\Omega}$, podemos calcular el estimador por Mínimos Cuadrados Generalizados Factibles (MCGF):

$$\hat{\beta}_{MCGF} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$$

2.2. Ejemplo práctico

En esta sección realizaremos un ejercicio práctico de análisis de heteroscedasticidad y autocorrelación con R-Commander. Para ello, analizaremos el siguiente modelo temporal de consumo con datos simulados:

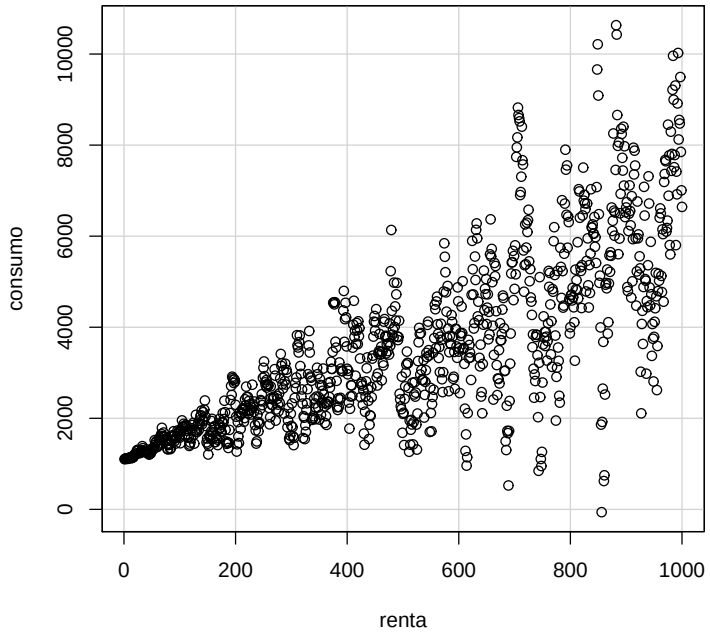
$$C_t = \beta_0 + \beta_1 R_t + e_t$$

Donde C_t corresponde al consumo y R_t es el nivel de renta. Los datos son temporales y corresponden a una economía, de manera que $t = 1, \ldots, T$.

Una vez importados los datos, un buen inicio es una representación gráfica de los datos, lo que es inmediato si solo hay un regresor. Mediante la siguiente ruta, obtenemos un diagrama de dispersión de las variables explicativa y explicada:

Gráficas / Diagrama de dispersión

Lo que resulta en el siguiente gráfico:

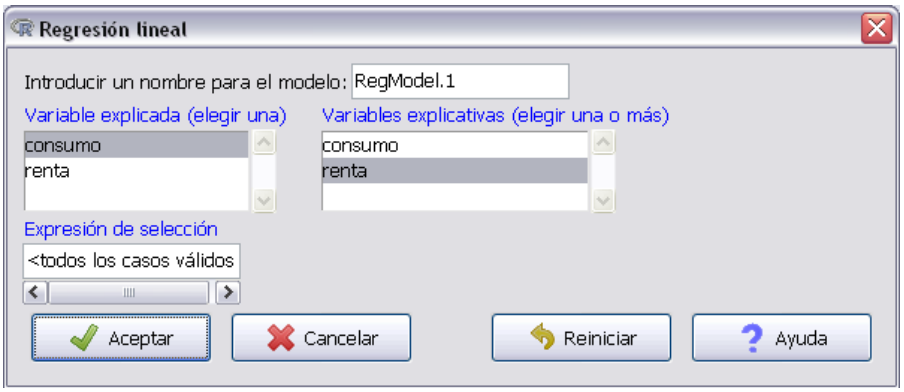


La interpretación de este gráfico es muy intuitiva. Para niveles bajos de renta, los niveles de consumo varían poco en el eje de ordenadas (y). Sin embargo, a medida que aumentan los niveles de renta, se observa una variabilidad superior de la variable explicativa. Esto es un signo de la existencia de heteroscedasticidad, cuya existencia ha de ser validadada estadísticamente mediante los contrastes correspondientes.

Para realizar la estimación del modelo con R-Commander, acudiremos a la siguiente ruta:

Estadísticos / Ajuste de modelos / Regresión lineal

Aparecerá el siguiente cuadro de diálogo, en el que introducimos la variable explicativa y la explicada:



El resultado de la estimación MCO del modelo es la siguiente:

```
> summary(RegModel.1)

Call:
lm(formula = consumo ~ renta, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max
-5475.7  -560.9    96.0   513.0  5082.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  920.372     76.855   11.97   <2e-16 ***
renta          5.250      0.133   39.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1214 on 998 degrees of freedom
Multiple R-squared:  0.6095, Adjusted R-squared:  0.6091
F-statistic:  1558 on 1 and 998 DF,  p-value: < 2.2e-16
```

Para la detección de la posible heteroscedasticidad, un test adecuado es el de Breusch-Pagan. Este test, válido cuando se dispone de muestras suficientemente grandes, presupone que es posible expresar la varianza del término de perturbación como una combinación lineal de un número determinado (p) de variables explicativas. El contraste se plantea de la siguiente manera:

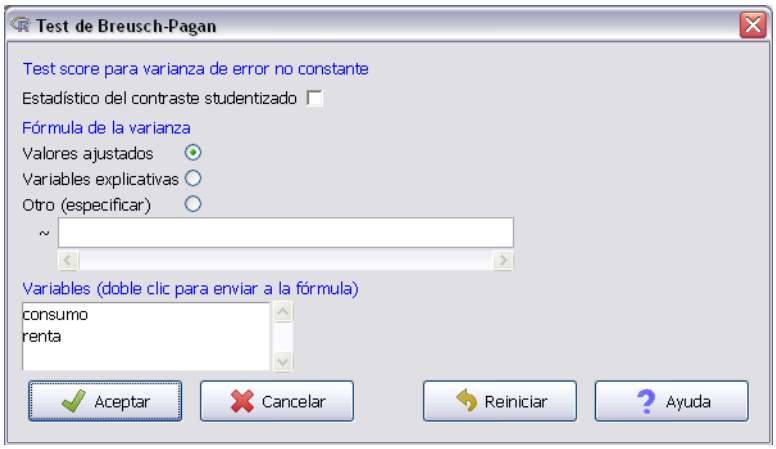
$$H_0 : \sigma_t^2 = \sigma^2$$

$$H_1 : \sigma_t^2 \neq \sigma^2$$

Es decir, bajo la hipótesis alternativa, la varianza no es constante, sino que depende de alguna variable. Con R-Commander, este test se realiza accediendo a la siguiente ruta:

Diagnósticos numéricos / Test de Breusch-Pagan para heteroscedasticidad

Aparecerá el siguiente cuadro de diálogo, en el que tendremos que introducir los valores del contraste. Es decir, tenemos la posibilidad de introducir la forma funcional de la varianza, en caso de conocerla. En nuestro caso, aceptaremos la opción por defecto, que adquiere los valores ajustados de la regresión como fórmula para la varianza:



El resultado del test nos indica que caemos en la región de rechazo de la hipótesis nula, de manera que determinamos que existe heteroscedasticidad en nuestro modelo.

```
> bptest(consumo ~ renta, varformula = ~ fitted.values(RegModel
    .1), studentize=FALSE, data=Datos)

Breusch-Pagan test

data:  consumo ~ renta
BP = 351.9272, df = 1, p-value < 2.2e-16
```

El segundo problema que hay que analizar es la posible existencia de autocorrelación en el modelo. Para esto realizaremos el contraste de Durbin-Watson. Este test permite contrastar si el término de perturbación está autocorrelacionado según un esquema AR(1), es decir, la hipótesis nula indica que si el término de perturbación es de la forma $e_t = \rho e_{t-1} + \varepsilon_t$. Específicamente, el contraste se define del siguiente modo:

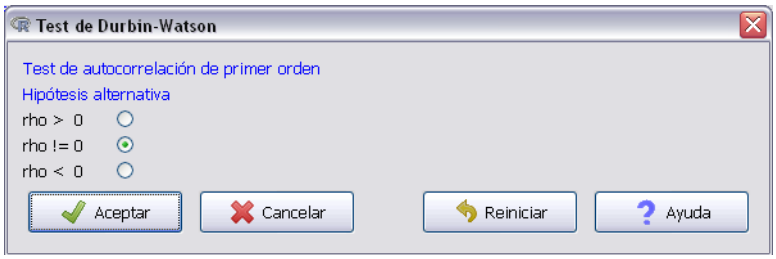
$$H_0 : e_t \sim AR(1) \qquad \text{con} \qquad \rho = 0$$

$$H_1 : e_t \sim AR(1) \qquad \text{con} \qquad \rho \gtrless 0$$

Con R-Commander, este test se realiza accediendo a la siguiente ruta:

Diagnósticos numéricos / Test de Durbin-Watson para autocorrelación

Aparecerá el siguiente cuadro de diálogo, donde tenemos que indicar la hipótesis alternativa. Si tenemos información previa de que el verdadero valor del parámetro ρ es positivo, seleccionaremos $H_1 : \rho > 0$, y lo correspondiente para un valor negativo de ρ. Si no tenemos información previa sobre este parámetro, seleccionaremos $H_1 : \rho \neq 0$:



El resultado del test nos indica claramente, para cualquier nivel de confianza, que rechazamos la hipótesis nula, es decir, existe autocorrelación en el modelo.

```
> dwtest(consumo ~ renta, alternative="two.sided", data=Datos)

Durbin-Watson test

data:  consumo ~ renta
DW = 0.4037, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is not 0
```

2.3. Estimación eficiente de la MVC

En esta sección nos encargamos de cómo realizar una estimación eficiente en presencia de autocorrelación y/o heteroscedasticidad. White (1980) argumentó que no siempre es posible conocer la estructura de los errores y valorar el modelo mediante MCG. Cuando eso sucede, en el caso de estar ante un modelo heteroscedástico, la mejor opción es valorar los parámetros del modelo mediante MCO e intentar obtener una estimación robusta de la matriz de varianzas y covarianzas de los parámetros mediante la fórmula:

$$MVC(\hat{\beta}_{MCO}) = n(X'X)^{-1}n^{-1}\sum_{i=1}^{n}\hat{e}_i^2 x_i x_i'(X'X)^{-1}$$

Este procedimiento se conoce con varios nombres en la literatura: Fórmula de White, Fórmula de Eicker-White, Fórmula de Huber, Fórmula de Huber-White o Matriz de covarianzas GMM, entre otros. Este procedimiento es problemático en muestras pequeñas.

En este capítulo veremos cómo efectuar la estimación de un modelo ante **heteroscedasticidad** y/o **autocorrelación**. Como veremos, existen dos grandes aproximaciones al respecto:

1) Estimar el modelo mediante Mínimos Cuadrados Generalizados (MCG).

2) Estimar el modelo mediante MCO y a continuación valorar eficientemente la matriz de varianzas y covarianzas.

Para ilustrarlo con un ejemplo, generaremos con R unos datos ficticios que generen un modelo heteroscedástico y autocorrelacionado. Antes de nada, cargaremos tres librerías que nos serán de ayuda:

```
> library(sandwich)
> library(lmtest)
> library(nlme)
```

Supongamos el siguiente modelo de regresión lineal:

$$y_t = \alpha + \beta x_t + u_t, \qquad t = 1, \ldots, T.$$

Simularemos los datos, de manera que los parámetros poblacionales son $\alpha = 100$ y $\beta = 5$. Además, fijamos el tamaño muestral como $T = 1000$. El modelo se construye de manera que el término de error no es esférico, ya que va a estar autocorrelacionado y va a ser heteroscedástico:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

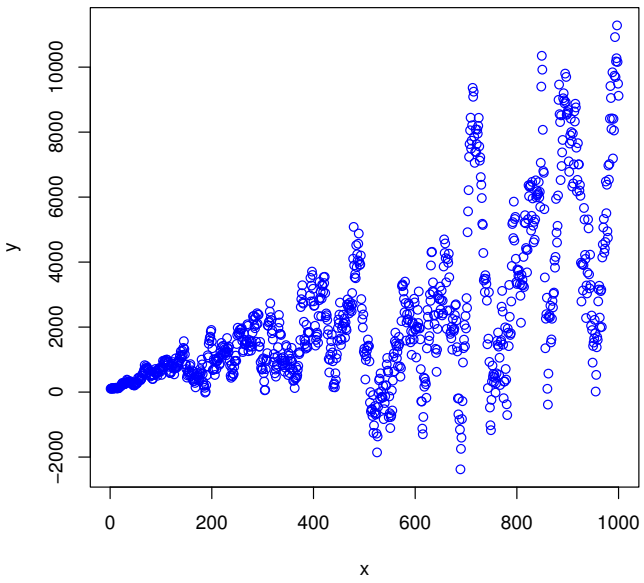
$$\varepsilon_t \sim N(0, \gamma t)$$

Fijamos los valores $\rho = 0,95$ y $\gamma = 1,1$. Con el modelo definido, lo introduccimos en R y lo representamos gráficamente:

```
> T <- 1000
> alpha <- 100
> beta <- 5
> rho <- 0.95
> gamma <- 1.1
> x <- 1:T
> y0 <- alpha + beta * x
> err <- rep(0, T)
> set.seed(12)
> err[1] <- rnorm(1, 0, 1)
> set.seed(12)
> for (i in 2:T) {
+     err[i] <- err[i - 1] * rho + rnorm(1, 0, i * gamma)
+ }
> y <- y0 + err
```

Vamos a visualizar las variables creadas para ver cómo se relacionan entre ellas. Este gráfico ya nos debe dar la impresión de que la varianza no se comporta aleatoriamente.

```
> plot(x, y, col = "blue")
```



Como vemos, el modelo por construcción no tiene un término de perturbación esférico. ¿Cuál es el problema de aplicar el estimador de mínimos cuadrados ordinarios (MCO)? Bueno, para que el estimador MCO sea eficiente (mínima varianza de la estimación), la matriz de varianzas y covarianzas de u debe ser esférica, es decir:

1) Homoscedástica: la varianza de u no varía entre los elementos de la muestra, de manera que $\sigma_t^2 = \sigma^2$ y los elementos de la diagonal de $MVC(u)$ son idénticos.

2) No autocorrelacionada: si los elementos fuera de la diagonal no son nulos ($\sigma_{ij} \neq 0,\ \forall i \neq j$), el modelo de regresión está autocorrelacionado, y viceversa.

Si *a)* y *b)* se cumplen, la matriz $MVC(u)$ será:

$$MVC(u) = E(uu') = \sigma^2 I_T$$

Siendo I_T la matriz identidad de dimensión $T \times T$.

En nuestro caso, vemos que esto no se cumple. Vamos a valorar primero el estimador MCO y ver cómo se comporta:

```
> m_mco <- lm(y ~ x)
> summary(m_mco)
```

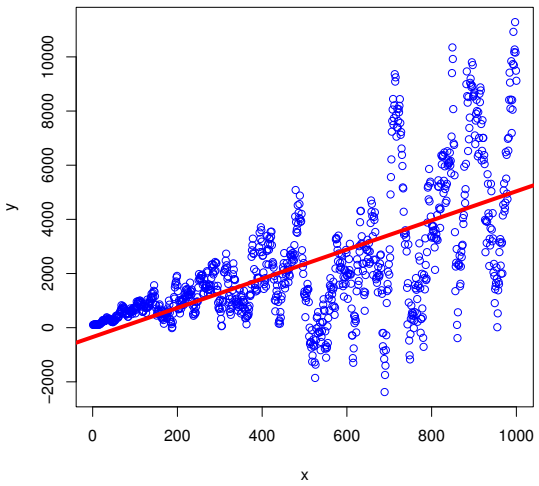
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5740.7  -964.1   157.1   724.6  6262.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -345.0854   119.8442  -2.879  0.00407 **
x              5.3828     0.2074  25.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1893 on 998 degrees of freedom
Multiple R-squared: 0.4029, Adjusted R-squared: 0.4023
F-statistic: 673.5 on 1 and 998 DF,  p-value: < 2.2e-16
```

Veamos un intervalo de confianza al 95 % para los parámetros estimados:

```
> confint(m_mco)

                 2.5 %      97.5 %
(Intercept) -580.260897 -109.909888
x              4.975785    5.789846
```

Vamos a representar visualmente la recta estimada ($\hat{\alpha}$ y $\hat{\beta}$) sobre el diagrama de dispersión de los puntos:

```
> plot(x, y, col = "blue")
> abline(lsfit(x, y), lty = 1, lwd = 4, col = "red")
```



Este estimador se construye mediante la siguiente fórmula:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

Y calcula la varianza y covarianzas de $\hat{\beta}$ así:

$$MVC(\hat{\beta}) = \hat{\sigma}_u^2(X'X)^{-1}$$

Sin embargo, hemos visto que la MVC del término de perturbación es, realmente:

$$MVC(u) = E(uu') = \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma_T^2 \end{pmatrix}_{T\times T}$$

Con lo que, en realidad, la varianza de los parámetros es:

$$MVC(\beta) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

La cuestión es, ¿cómo estimamos el modelo? Hay dos opciones. Teóricamente, si conocemos exactamente la forma de Ω, la podemos introducir directamente en el estimador por Mínimos Cuadrados Generalizados (MCG):

$$\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

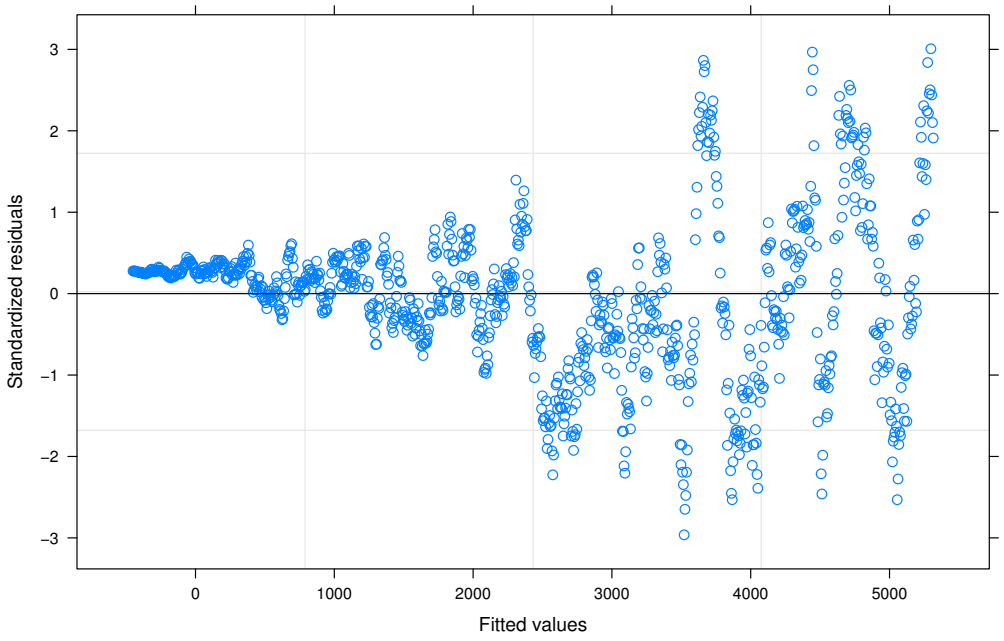
En R, vamos a estimar MCG sabiendo que $\rho = 0,95$ y $\gamma = 1,1$. Primero asumiendo solo autocorrelación:

```
> gls_1 <- gls(y ~ x, correlation = corAR1(rho))
> summary(gls_1)

Generalized least squares fit by REML
  Model: y ~ x
  Data: NULL
       AIC      BIC    logLik
  15686.43 15706.05 -7839.214

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

```
0.950404

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -455.9404  753.9217 -0.604758  0.5455
x              5.7711    1.2923  4.465694  0.0000

 Correlation:
  (Intr)
x -0.858

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.96297027 -0.53634973  0.07080453  0.37691878  3.00744997

Residual standard error: 1990.35
Degrees of freedom: 1000 total; 998 residual
```

La función `plot` aplicada al modelo estimado por MCG nos muestra el gráfico de los residuos:

```
> plot(gls_1)
```

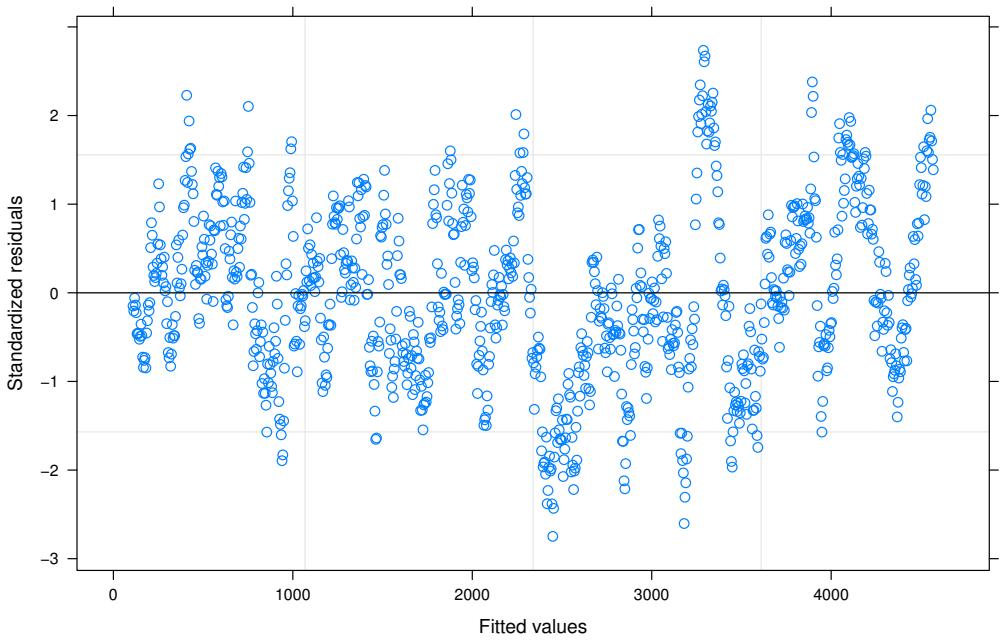


Y ahora estimamos de nuevo el modelo mediante MCG, asumiendo esta vez tanto autocorrelación como heteroscedasticidad:

```
> gls_2 <- gls(y ~ x, correlation = corAR1(rho), weights =
    varPower(gamma))
> summary(gls_2)

Generalized least squares fit by REML
  Model: y ~ x
  Data: NULL
       AIC     BIC    logLik
  14797.58 14822.1 -7393.788

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.9353005
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
   power
1.183458

Coefficients:
                Value Std.Error  t-value p-value
(Intercept) 104.94599  37.42416 2.804231  0.0051
x             4.46403   0.53408 8.358306  0.0000

 Correlation:
  (Intr)
x -0.293

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.74838335 -0.69033419 -0.05426096  0.70224709  2.73602972

Residual standard error: 0.1528423
Degrees of freedom: 1000 total; 998 residual
```

De nuevo, la función `plot` aplicada al modelo estimado por MCG nos muestra el gráfico de los residuos:

```
> plot(gls_2)
```



Es relevante recordar que White (1980) argumentó que no siempre es posible conocer la estructura de los errores y valorar el modelo mediante MCG. Cuando eso sucede, en el caso de estar ante un modelo heteroscedástico, la mejor opción es estimar los parámetros del modelo mediante MCO e intentar obtener una estimación robusta de la matriz de varianzas y covarianzas de los parámetros mediante la fórmula:

$$MVC(\hat{\beta}_{MCO}) = n(X'X)^{-1}n^{-1}\sum_{i=1}^{n}\hat{u}_i^2 x_i x_i'(X'X)^{-1}$$

En este sentido, hay muchas maneras de calcular eficientemente $\hat{\Omega}$. El programa R nos ofrece dos de ellas:

1) *HC* : *Heteroskedasticity Consistent matrix.*

2) *HAC* : *Heteroskedasticity and Autocorrelation Consistent matrix.*

Entonces, a partir de MCO, calculamos $\hat{\Omega}$ de ambas maneras y así recalculamos las varianzas (y los contrastes de significación asociados) de los coeficientes:

```
> coeftest(m_mco)

t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539  119.84419 -2.8795  0.004069 **
x              5.38282    0.20742 25.9512 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coeftest(m_mco, vcovHC(m_mco))



t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539   72.74869 -4.7435 2.406e-06 ***
x              5.38282    0.22307 24.1310 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coeftest(m_mco, vcovHAC(m_mco))

t test of coefficients:

             Estimate Std. Error t value  Pr(>t)
(Intercept) -345.08539  302.01809 -1.1426    0.2535
x              5.38282    0.75061  7.1712 1.446e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como vemos, el hecho de estimar la verdadera matriz MVC revela que las varianzas reales son en realidad mayores que las estimadas por MCO y, consecuentemente, los intervalos de confianza para $\hat{\beta}$ son también mayores.

3. Errores en la muestra

3.1. Multicolinealidad

La multicolinealidad aparece cuando dos o más variables explicativas en un modelo de regresión múltiple están altamente correlacionadas. De manera alternativa, se puede afirmar que, en presencia de multicolinealidad, una variable explicativa se puede predecir linealmente a partir de otras variables explicativas.

La multicolinealidad implica que las estimaciones de los coeficientes de la regresión múltiple pueden cambiar de forma errática ante pequeños cambios en la especificación del modelo o cambios en los datos. Además, un alto grado de multicolinealidad puede causar problemas a la hora de calcular la matriz inversa de $X'X$, necesaria para el cálculo de los coeficientes de regresión.

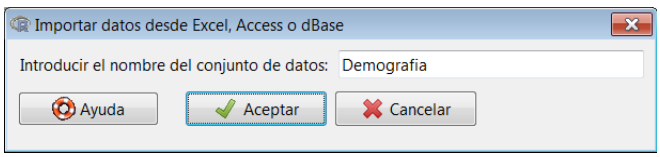
Recordemos que existen tres grados de multicolinealidad:

1) Ausencia total de multicolinealidad. Sucede cuando no existe correlación entre las variables explicativas del modelo.

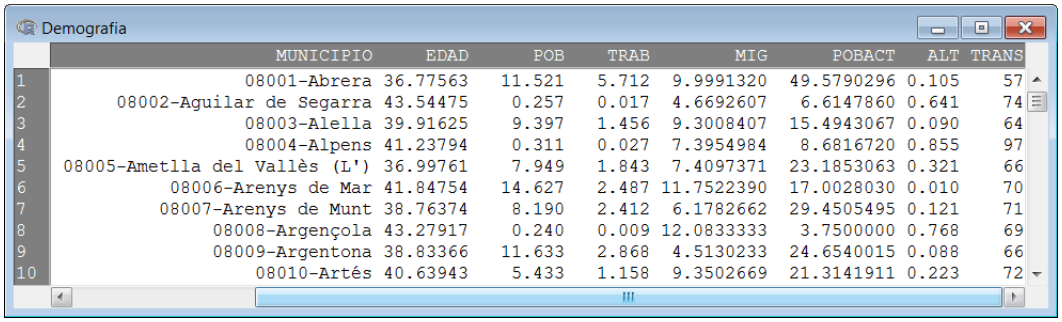
2) Presencia de un cierto grado de multicolinealidad. Existe un alto grado de correlación lineal entre algunas variables explicativas. Cuanto más elevado sea este grado de correlación (es decir, el coeficiente de correlación de Pearson se acerque a 1), mayor será el grado de multicolinealidad.

3) Presencia de multicolinealidad perfecta. Existe alguna variable explicativa que se puede obtener a partir de la combinación lineal de otras variables explicativas, lo que implica que algunas variables explicativas son linealmente dependientes entre sí. En este caso, la estimación del modelo es imposible debido a la imposibilidad de invertir la matriz $X'X$.

Veamos un ejemplo práctico, con R-Commander, de cómo analizar el problema de la multicolinealidad. Para ello consideraremos un estudio demográfico para los municipios de Cataluña en el año 2009. El primer paso será importar los datos de un archivo de Excel y crear un conjunto de datos al que daremos el nombre de *Demografia*:



Si visualizamos los datos importados, observamos que están incluidas las siguientes variables:


Demografia

	MUNICIPIO	EDAD	POB	TRAB	MIG	POBACT	ALT	TRANS
1	08001-Abrera	36.77563	11.521	5.712	9.9991320	49.5790296	0.105	57
2	08002-Aguilar de Segarra	43.54475	0.257	0.017	4.6692607	6.6147860	0.641	74
3	08003-Alella	39.91625	9.397	1.456	9.3008407	15.4943067	0.090	64
4	08004-Alpens	41.23794	0.311	0.027	7.3954984	8.6816720	0.855	97
5	08005-Ametlla del Vallès (L')	36.99761	7.949	1.843	7.4097371	23.1853063	0.321	66
6	08006-Arenys de Mar	41.84754	14.627	2.487	11.7522390	17.0028030	0.010	70
7	08007-Arenys de Munt	38.76374	8.190	2.412	6.1782662	29.4505495	0.121	71
8	08008-Argençola	43.27917	0.240	0.009	12.0833333	3.7500000	0.768	69
9	08009-Argentona	38.83366	11.633	2.868	4.5130233	24.6540015	0.088	66
10	08010-Artés	40.63943	5.433	1.158	9.3502669	21.3141911	0.223	72

La descripción de las variables es la siguiente:

MUNICIPIO: código postal y nombre del municipio.

EDAD: media de edad de la población.

POB: población total (en miles de personas).

TRAB: número de trabajadores (en miles de personas).

MIG: porcentaje de población inmigrante.

POBACT: porcentaje de población activa.

ALT: altitud del municipio (en kilómetros).

TRANS: tiempo de transporte hasta la capital más cercana.

El primer modelo de regresión considera la variable *EDAD* como variable explicada, y el resto de las variables como variables explicativas. Para valorar un modelo de regresión lineal, como sabemos, tenemos la siguiente ruta en el menú desplegable:

Estadísticos / Ajuste de modelos / Regresión lineal

Seleccionamos el nombre del modelo estimado y las variables que se deben incluir en la estimación en el siguiente cuadro de diálogo:

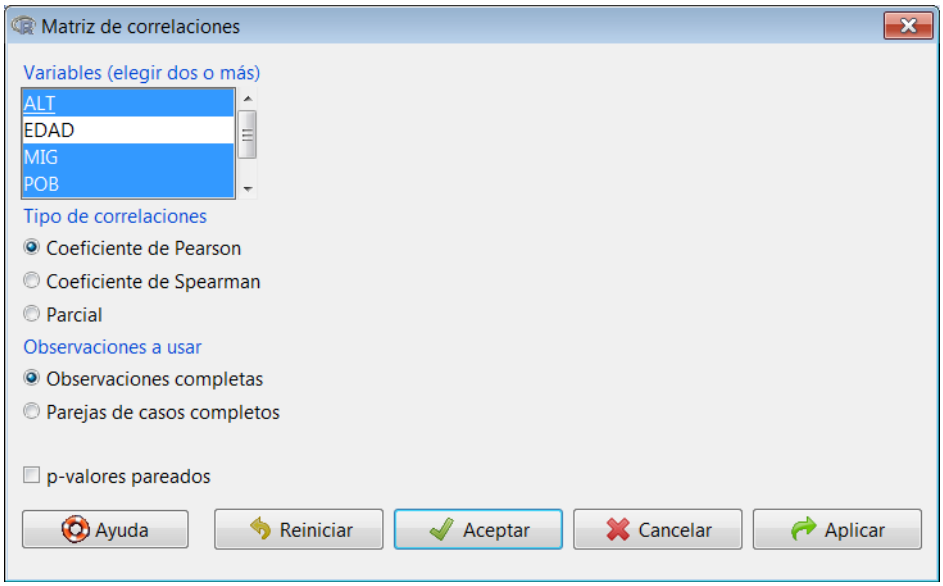
El resultado de la estimación se muestra a continuación. A simple vista, aunque el ajuste del modelo sea más bien pobre ($R^2 = 0,3$), todos los coeficientes estimados son significativos con un nivel de significación menor que 1 %, y la estimación es significativa en su conjunto, dado el resultado del test *F*.

```
> RegModel.1 <- lm(EDAD~ALT+MIG+POB+POBACT+TRAB+TRANS, data=
    Demografia)

> summary(RegModel.1)

Call:
lm(formula = EDAD ~ ALT + MIG + POB + POBACT + TRAB + TRANS,
    data = Demografia)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4061  -2.4548  -0.3131   2.3616  16.2820

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 39.634134   0.513051  77.252  < 2e-16 ***
ALT          1.407579   0.450937   3.121  0.00185 **
MIG         -0.154346   0.018279  -8.444  < 2e-16 ***
POB         -0.037418   0.012169  -3.075  0.00217 **
POBACT      -0.040815   0.005446  -7.495 1.54e-13 ***
TRAB         0.074415   0.023209   3.206  0.00139 **
TRANS        0.059654   0.006064   9.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.62 on 934 degrees of freedom
Multiple R-squared:  0.3021,  Adjusted R-squared:  0.2976
F-statistic: 67.37 on 6 and 934 DF,  p-value: < 2.2e-16
```

¿Significa esto que el resultado de la estimación es satisfactorio, y que podemos dar este resultado como válido? La verdad es que no necesariamente. Antes de realizar una estimación, resulta útil visualizar la matriz de correlaciones simple entre todas las variables. Aunque existen técnicas más avanzadas y eficientes para detectar la multicolinealidad, esta matriz siempre mostrará información útil:

Resúmenes / Matriz de correlaciones

En el cuadro de opciones resultante seleccionamos todas las variables explicativas, así como el Coeficiente de correlación de Pearson.



Esta ruta nos muestra la siguiente información:

```
> cor(Demografia[,c("ALT","MIG","POB","POBACT","TRAB","TRANS")
    ], use="complete")

          ALT    MIG    POB POBACT   TRAB  TRANS
ALT      1,00  -0,32  -0,11  -0,11  -0,07   0,43
MIG     -0,32   1,00   0,10   0,00   0,06   0,12
POB     -0,11   0,10   1,00   0,07   0,98  -0,12
POBACT  -0,11   0,00   0,07   1,00   0,08  -0,11
TRAB    -0,07   0,06   0,98   0,08   1,00  -0,08
TRANS    0,43   0,12  -0,12  -0,11  -0,08   1,00
```

Para facilitar la interpretación del resultado, se ha limitado a dos decimales cada valor de esta matriz. En realidad, el resultado muestra más decimales.

¿Qué podemos destacar de esta matriz de correlaciones? La correlación lineal entre las variables *POB* (población) y *TRAB* (trabajadores) es de *0,98*, es decir, es una correlación lineal positiva casi perfecta. Realmente, ¿es necesario incorporar en el modelo que estimar dos variables que aportan casi la misma información? Esto no solo tiene consecuencias negativas en cuanto al proceso de estimación, sino que puede llevar a estimaciones erróneas de los coeficientes.

Un procedimiento más refinado para evaluar la posible existencia de multicolinealidad entre las variables explicativas (o regresores) es el **Factor de Incremento de la Varianza (FIV)** de cada una de las variables explicativas. El FIV es un estadístico que permite determinar si la varianza de un estimador está inflada por la presencia de multicolinealidad en el modelo respecto al caso de ortogonalidad entre regresores. Esto es, si la correlación entre todos los regresores fuera igual a cero (ortogonalidad

En el cálculo del FIV no afecta cuál sea la variable explicada, ya que en su cálculo solo intervienen las variables explicativas o regresores.

perfecta), la varianza de la estimación sería óptima y el FIV de cada regresor sería igual a cero. En la práctica, cada regresor tendrá un FIV más elevado cuanto mayor sea su correlación con el resto de los regresores. En la práctica, no existe un valor umbral de los FIV a partir del cual se deba afirmar que hay problemas graves de multicolinealidad, pero se suele considerar que, para cada regresor, un $FIV > 5$ indica un grado de multicolinealidad elevado que ha de ser corregido.

A partir del modelo estimado anteriormente, con R-Commander calcularemos el FIV accediendo a la siguiente ruta:

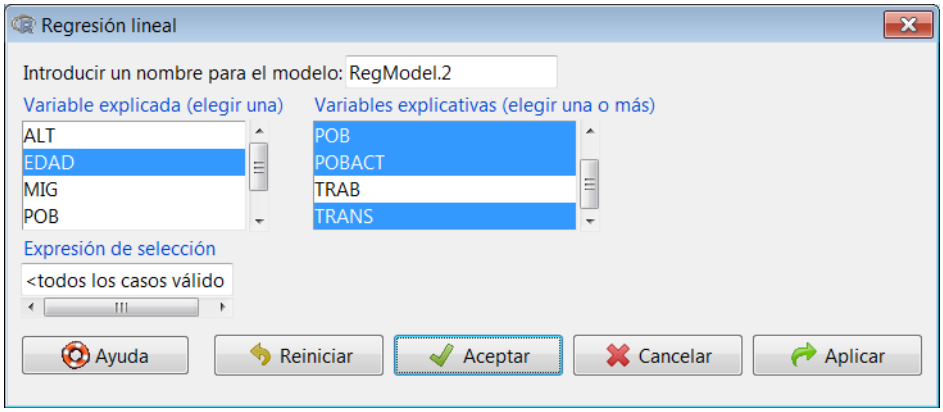
Modelos / Diagnósticos numéricos / Factores de inflación de varianza

El resultado muestra claramente cómo todas las variables tienen un FIV bajo menos dos: *POB* y *TRAB*. Para estas dos variables el valor del FIV es altísimo, con lo que una de las dos ha de ser eliminada de la especificación del modelo.

```
> vif(RegModel.1)
      ALT        MIG        POB     POBACT       TRAB      TRANS
 1.514863   1.306153  33.629262   1.029765  33.160244   1.432465
```

Ahora optaremos por retirar la variable *TRAB* de la especificación, y estimar un segundo modelo de forma análoga al caso anterior:

Estadísticos / Ajuste de modelos / Regresión lineal



El resultado del segundo modelo estimado nos muestra una contradicción respecto a la primera estimación. El coeficiente asociado a la variable *POB* ahora no es significativo, mientras que en el modelo estimado anteriormente sí lo era. ¿Qué nos indica esto? Pues que **no hay que confiar en las estimaciones de parámetros en presencia de multicolinealidad**.

```
> RegModel.2 <- lm(EDAD~ALT+MIG+POB+POBACT+TRANS, data=
    Demografia)

> summary(RegModel.2)

Call:
lm(formula = EDAD ~ ALT + MIG + POB + POBACT + TRANS, data =
    Demografia)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6718  -2.4594  -0.3481   2.4163  16.5107

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.2802715  0.5035203  78.011  < 2e-16 ***
ALT          1.4575008  0.4528992   3.218  0.00133 **
MIG         -0.1654489  0.0180364  -9.173  < 2e-16 ***
POB          0.0009936  0.0021441   0.463  0.64318
POBACT      -0.0392892  0.0054517  -7.207 1.18e-12 ***
TRANS        0.0631734  0.0059932  10.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.638 on 935 degrees of freedom
Multiple R-squared:  0.2944,  Adjusted R-squared:  0.2906
F-statistic: 78.01 on 5 and 935 DF,  p-value: < 2.2e-16
```

Para asegurarnos de que el problema de multicolinealidad está resuelto, obtendremos los VIF de los coeficientes de esta segunda estimación.

Modelos / Diagnósticos numéricos / Factores de inflación de varianza

```
> vif(RegModel.2)
     ALT      MIG      POB   POBACT    TRANS
1.513057 1.259275 1.033843 1.021903 1.385527
```

Claramente, todos los valores son menores que 5, con lo que hemos resuelto el problema de multicolinealidad.

3.2. Observaciones atípicas

Este problema surge cuando en la muestra algunas observaciones manifiestan un valor muy diferente del resto de las observaciones. Visualmente, esto se corresponde con una nube de puntos de la variable en la que un punto está muy alejado del resto de las observaciones. Dos explicaciones pueden dar respuesta a este hecho:

1) Hay errores en la recogida de la muestra, de manera que hay valores erróneos que no se corresponden con la realidad.

2) El valor recogido en la muestra de estas observaciones *outliers* se debe a particularidades de la observación, de modo que no hay ningún error en la muestra.

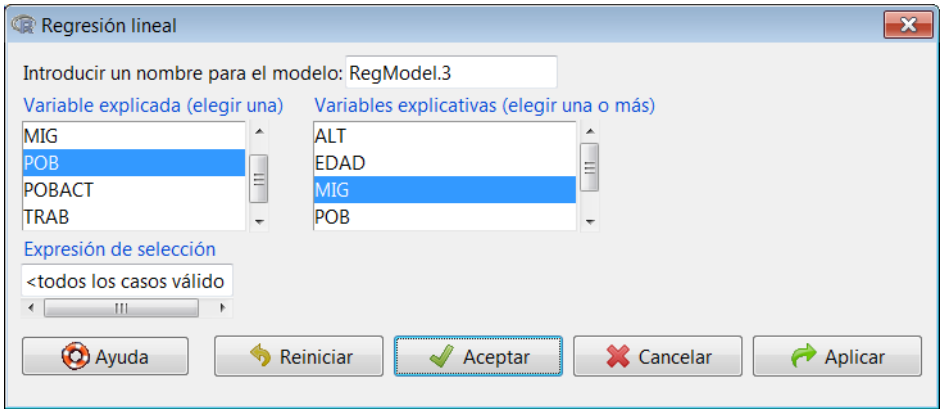
En ambos casos, la presencia de *outliers* tiene consecuencias negativas para la estimación del modelo econométrico, ya que los errores estándares de los estimadores son mayores y empeora el ajuste global del modelo (R^2 y *F de Snedecor*.

Estudiaremos este hecho a partir del conjunto de datos *Demografia*, introducido en el apartado anterior. En este caso, estimaremos un MRLS en el que el porcentaje de inmigración explica la población total de cada municipio:

$$POB_i = \beta_0 + \beta_1 MIG_i + e_i$$

Al igual que en el caso anterior, valoramos el modelo accediendo a la siguiente ruta:

Estadísticos / Ajuste de modelos / Regresión lineal

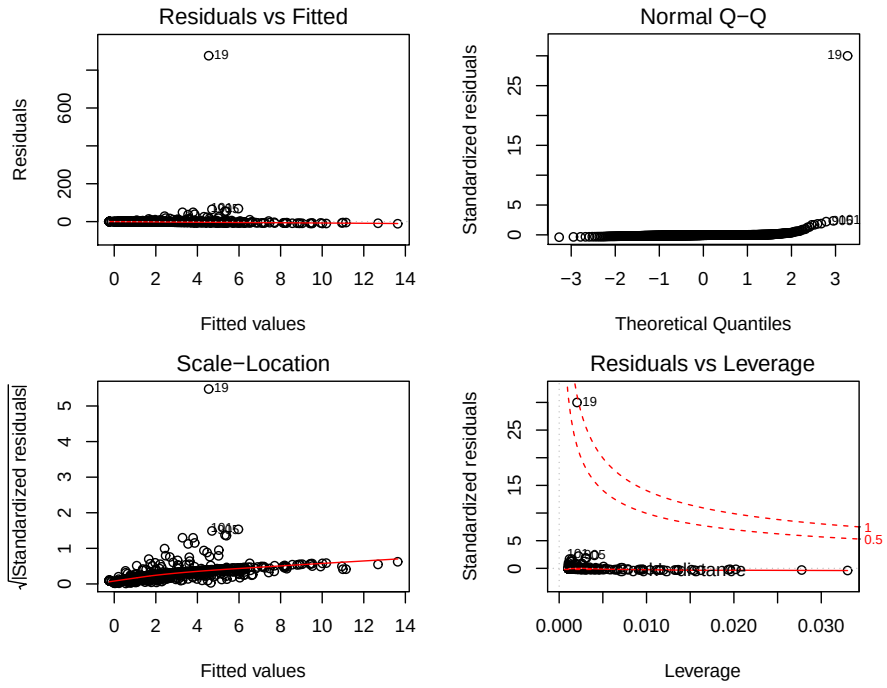


El resultado se muestra a continuación, obteniendo un efecto positivo y estadísticamente significativo del regresor sobre la variable dependiente:

```
> RegModel.3 <- lm(POB~MIG, data=Demografia)
```

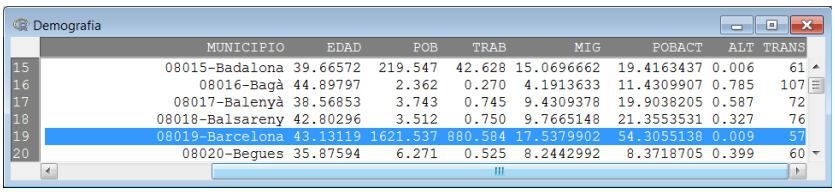
```
> summary(RegModel.3)

Call:
lm(formula = POB ~ MIG, data = Demografia)

Residuals:
    Min      1Q  Median      3Q     Max
 -31.31   -8.01   -3.91   -0.86 1607.93

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)  -0.3875     3.1626  -0.123   0.9025
MIG           0.7979     0.2474   3.226   0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.98 on 939 degrees of freedom
Multiple R-squared:  0.01096, Adjusted R-squared:  0.009906
F-statistic: 10.41 on 1 and 939 DF,  p-value: 0.0013
```

¿Es posible que exista algún *outlier* en las variables? Veamos los gráficos de diagnóstico de la estimación efectuada:



En todos los gráficos observamos que el residuo asociado a la observación 19 se aleja considerablemente del resto de los residuos. Comprobemos qué observación ocupa esa posición visualizando el conjunto de datos *Demografia*. Vemos que la observación atípica corresponde al municipio de Barcelona. Este resultado es lógico: este municipio tiene muchos más habitantes que el resto de los municipios catalanes, con lo que

la medición de esta observación no es errónea, ya que es lógico que este valor sea tan alto comparado con el resto de las observaciones.



	MUNICIPIO	EDAD	POB	TRAB	MIG	POBACT	ALT	TRANS
15	08015-Badalona	39.66572	219.547	42.628	15.0696662	19.4163437	0.006	61
16	08016-Bagà	44.89797	2.362	0.270	4.1913633	11.4309907	0.785	107
17	08017-Balenyà	38.56853	3.743	0.745	9.4309378	19.9038205	0.587	72
18	08018-Balsareny	42.80296	3.512	0.750	9.7665148	21.3553531	0.327	76
19	08019-Barcelona	43.13119	1621.537	880.584	17.5379902	54.3055138	0.009	57
20	08020-Begues	35.87594	6.271	0.525	8.2442992	8.3718705	0.399	60

¿Cómo se puede identificar la presencia de *outliers*? A partir de un modelo estimado, una opción es el test de valores atípicos de Bonferroni, el cual reporta el p-valor para los residuos estudentizados absolutos, usando la distribución t. En R-Commander, esto se hace accediendo a la siguiente ruta del menú desplegable:

Modelos / Diagnósticos numéricos / Test de valores atípicos de Bonferroni

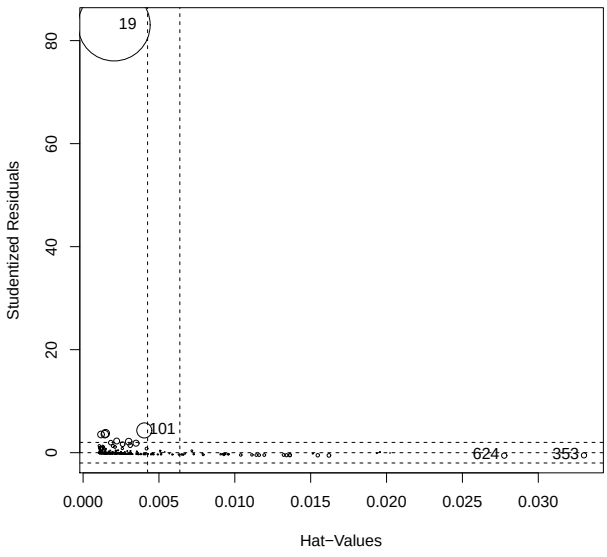
El resultado muestra dos valores atípicos, el más destacado de los cuales es la observación 19, correspondiente a Barcelona.

```
> outlierTest(RegModel.3)
     rstudent unadjusted p-value Bonferonni p
19   83.048751          0.0000e+00      0.000000
101   4.322249          1.7094e-05      0.016086
```

Alternativamente, se puede calcular el gráfico de influencias, que compara en un gráfico bidimensional los valores estimados del modelo (*hat values*) y los residuos estidentizados. Se realiza accediendo a la siguiente ruta:

Modelos / Gráficas / Gráfica de influencias

Esta acción muestra dos resultados. El primero es gráfico, en el que se ve cómo el valor de la observación 19 está claramente apartada del resto de las observaciones:



El segundo aparece en la consola. Nos muestra una lista de posibles *outliers*, mostrándose además la distancia de Cook (*CookD*). Esta medida permite detectar la extrañeza de una observación, sirviendo para detectar aquellas observaciones que tienen un efecto mayor en el ajuste que el resto, y que pueden hacer cambiar los valores estimados por los parámetros del modelo de una manera sustancial.

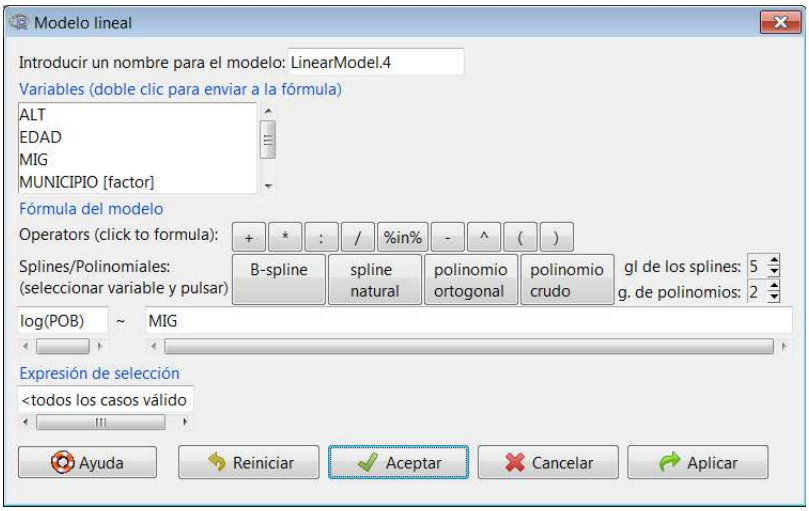
```
> influencePlot(RegModel.3, id.method="noteworthy", id.n=2)
       StudRes          Hat       CookD
19   83.0487509 0.002045796 0.92045949
101   4.3222485 0.004034997 0.19272734
353  -0.5104403 0.033011959 0.06671540
624  -0.5668982 0.027758536 0.06775765
```

Según esta medida, el principal *outlier* sigue siendo la observación 19. ¿Cuál puede ser la solución a la presencia de esta observación tan particular? Excluirla del modelo estimado podría ser una solución, pero la observación no es errónea, y obviarla significa no considerar la principal ciudad de Cataluña en un estudio sobre este territorio. No parece, pues, una solución recomendable. Una solución alternativa es cambiar la *forma funcional* de la especificación, que puede pasar por transformar alguna variable. Vamos a optar por expresar la variable dependiente en logaritmos, esto es:

$$log(POB)_i = \beta_0 + \beta_1 MIG_i + e_i$$

Se dan dos consecuencias al producirse esta transformación. La primera es que los valores de la variable *POB* se comprimen, existiendo menos distancia entre el valor 19 y el resto. Por otra parte, también cambia la interpretación de los coeficientes. Para efectuar esta estimación, hay que acceder a la ruta de un *Modelo lineal*, en cuyo cuadro de diálogo podemos especificar la relación funcional entre las variables:

Estadísticos / Ajuste de modelos / Modelo lineal



El resultado muestra una mejora significativa del ajuste del modelo y de la significación individual de los coeficientes respecto al modelo anterior.

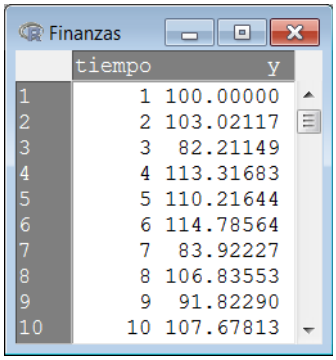
```
> LinearModel.4 <- lm(log(POB) ~ MIG, data=Demografia)

> summary(LinearModel.4)

Call:
lm(formula = log(POB) ~ MIG, data = Demografia)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3525 -1.1869 -0.2454  0.9859  6.5293

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.729731   0.087897  -8.302 3.54e-16 ***
MIG          0.090751   0.006875  13.200  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.556 on 939 degrees of freedom
Multiple R-squared:  0.1565,  Adjusted R-squared:  0.1556
F-statistic: 174.3 on 1 and 939 DF,  p-value: < 2.2e-16
```

4. Permanencia estructural

Este problema surge cuando se rompe una de las hipótesis básicas del modelo de regresión estándar, que es la hipótesis de permanencia estructural. El problema surge cuando, en una serie temporal, en un punto del tiempo cambia la relación entre la variable dependiente y uno de los regresores. Para estudiar este problema con un ejemplo sencillo, analizaremos el efecto del tiempo sobre la evolución del precio de un activo financiero ficticio, que denominaremos y. Es decir, estudiaremos el siguiente modelo:

$$y_t = \beta_0 + \beta_1 t + e_t$$

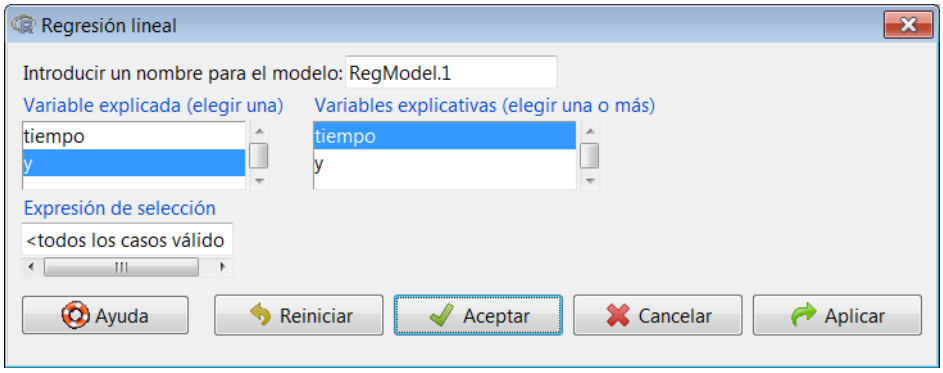
El primer paso es importar y visualizar los datos.



Aunque aquí solo se muestren 10 observaciones, el conjunto de datos contiene $T = 1000$ observaciones temporales.

El primer paso es valorar el modelo de regresión:

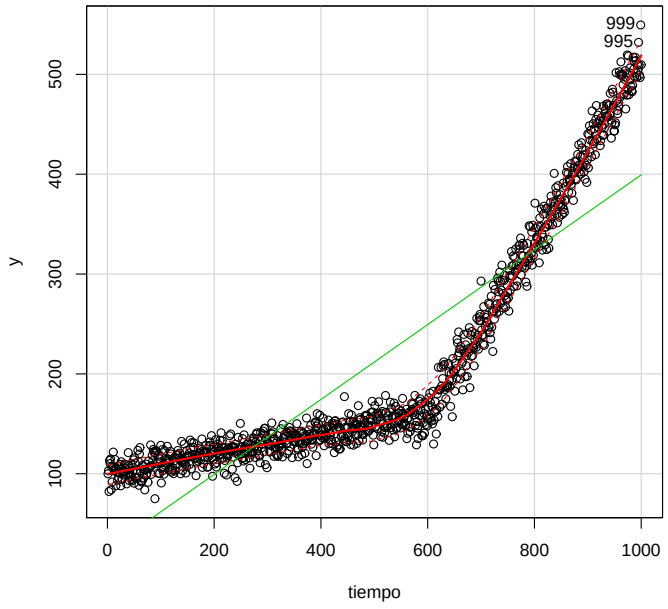
Estadísticos / Ajuste de modelos / Regresión lineal



El resultado de la estimación es el siguiente:

```
> RegModel.1 <- lm(y~tiempo, data=Finanzas)

> summary(RegModel.1)

Call:
lm(formula = y ~ tiempo, data = Finanzas)

Residuals:
     Min       1Q   Median       3Q      Max
-123.724  -46.900   -2.406   44.058  150.465

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 24.443812   3.499777   6.984 5.22e-12 ***
tiempo       0.374961   0.006057  61.903  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.29 on 998 degrees of freedom
Multiple R-squared:  0.7934,  Adjusted R-squared:  0.7932
F-statistic:  3832 on 1 and 998 DF,  p-value: < 2.2e-16
```

Como vemos, es un ajuste bastante bueno, y tanto los coeficientes valorados como el modelo estimado global son significativos estadísticamente. Esta estimación da un coeficiente $hat\beta_1 = 0,37$. ¿Hasta qué punto es esta estimación correcta? Para entender mejor el concepto de permanencia estructural, veamos en un plano cartesiano el diagrama de dispersión de las dos variables: el tiempo en el eje horizontal y el precio del activo financiero en el eje vertical. Este gráfico se obtiene acudiendo a la opción *Gráficas* del menú desplegable.



En este gráfico también aparece la recta estimada en el modelo $(24,44+0,37t)$, que es la misma para todos los puntos. Sin embargo, vemos cómo la relación funcional entre ambas variables cambia sobre el punto $t = 600$. Vemos que antes y después la pendiente cambia de manera significativa, como lo muestra la recta curva que resigue las observaciones. Así pues, parece razonable estimar dos modelos, partiendo la muestra en dos partes, con coeficientes estimados diferentes.

Estadísticamente, ¿cómo detectamos la presencia de un cambio estructural? Un test útil en este sentido es el **Test de Chow**. Este contraste consiste en estimar dos modelos separando la muestra en dos submuestras a partir de un punto de corte determinado, para después comparar las SCE de la regresión para todo el tamaño muestral con las SCE de las regresiones para cada una de las dos submuestras fijadas. Este test es un tanto arbitrario, ya que requiere que fijemos un punto de corte de antemano de manera aproximada.

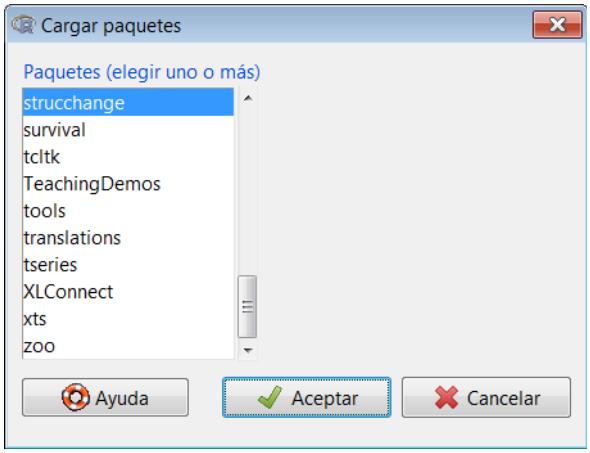
En R-Commander este contraste no está disponible en el menú, pero esto no significa que no se pueda efectuar mediante código. Para ello, hay que instalar el paquete *strucchange* en la consola:

```
> install.packages("strucchange")
```

Una vez instalada esta librería, hay que cargarla. Esto lo haremos acudiendo a la ruta del menú desplegable:

Opciones / Cargar paquetes

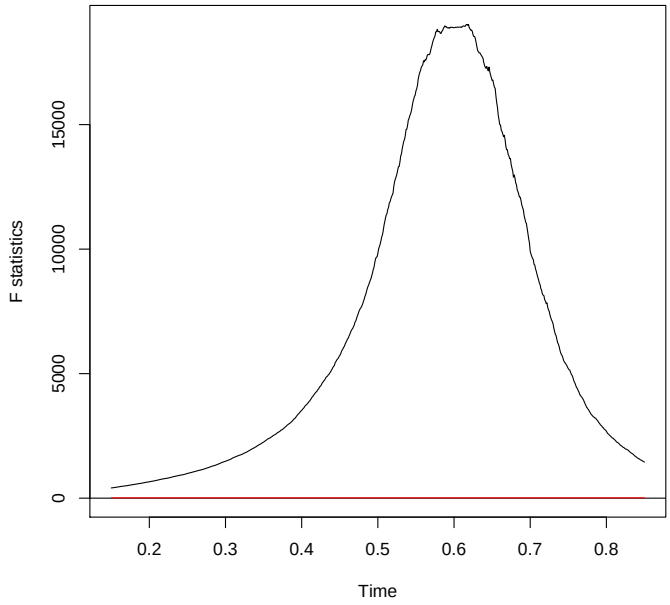
En el cuadro de diálogo que nos aparecerá, seleccionamos el paquete que acabamos de instalar.



La función de R incluida en este paquete que calcula el estadístico de Chow es `Fstats`. Un hecho positivo es que, opcionalmente, podemos introducir el período temporal en el que sospechamos que se produce el cambio estructural. Si no lo especificamos, esta función calcula el estadístico para todos los puntos de corte en la muestra. La instrucciones que debemos introducir en la ventana de instrucciones son las siguientes:

```
> Fs <- Fstats(y ~ tiempo, data = Finanzas)
> plot(Fs)
```

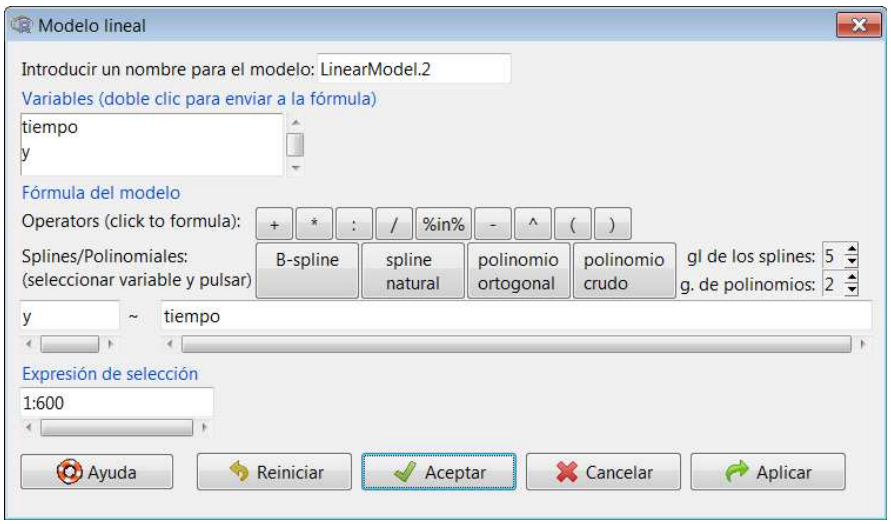
El gráfico resultante tiene la siguiente forma:



¿Qué nos dice este gráfico? Pues que el valor del estadístico F alcanza su máximo aproximadamente el en 60 % de la muestra, que coincide con el punto $t = 600$. Nuestra estrategia será estimar dos modelos, uno con la submuestra $t = 1, \ldots, 600$ y otro con la submuestra $t = 601, \ldots, 1000$. Para hacerlo, en el cuadro de diálogo del modelo lineal introduciremos, en la opción *Expresión de la selección*, la submuestra para la que queremos estimar el modelo.

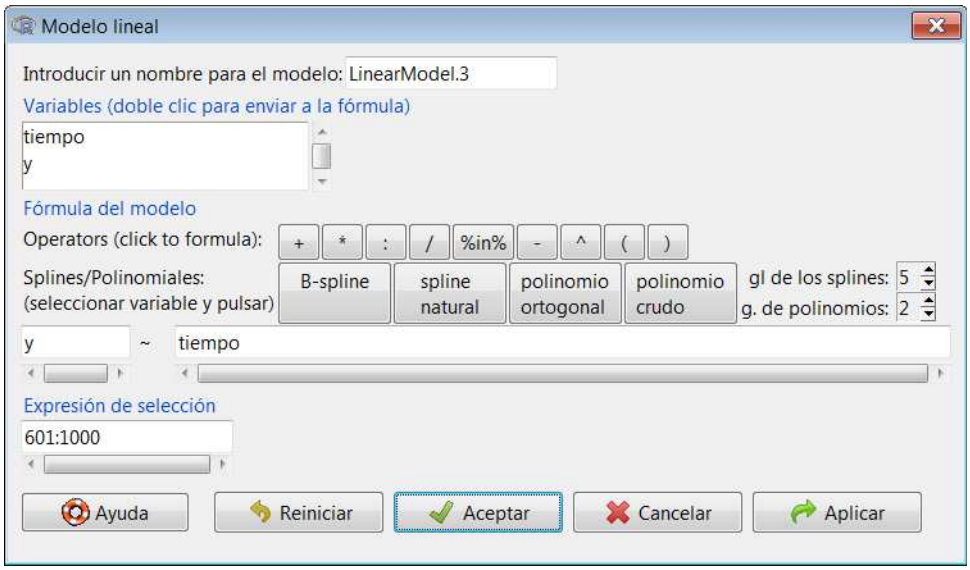
Veamos el resultado de la primera estimación para la submuestra $t = 1, \ldots, 600$.

Estadísticos / Ajuste de modelos / Modelo lineal



```
> LinearModel.2 <- lm(y ~ tiempo, data=Finanzas, subset=1:600)

> summary(LinearModel.2)

Call:
lm(formula = y ~ tiempo, data = Finanzas, subset = 1:600)

Residuals:
    Min      1Q  Median      3Q     Max
-33.897  -6.760   0.229   6.522  33.266

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.001e+02  8.268e-01  121.08   <2e-16 ***
tiempo      9.884e-02  2.384e-03   41.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.11 on 598 degrees of freedom
Multiple R-squared:  0.742, Adjusted R-squared:  0.7415
F-statistic:  1720 on 1 and 598 DF,  p-value: < 2.2e-16
```

Y ahora el resultado de la segunda estimación para la submuestra $t = 601, \ldots, 1000$.



```
> LinearModel.3 <- lm(y ~ tiempo, data=Finanzas, subset
    =601:1000)

> summary(LinearModel.3)

Call:
lm(formula = y ~ tiempo, data = Finanzas, subset = 601:1000)

Residuals:
    Min      1Q  Median      3Q     Max
-42.796 -10.810  -0.868  10.745  47.448

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -3.814e+02  5.315e+00  -71.77   <2e-16 ***
tiempo       8.957e-01  6.571e-03  136.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.18 on 398 degrees of freedom
Multiple R-squared:  0.979, Adjusted R-squared:  0.979
F-statistic: 1.858e+04 on 1 and 398 DF,  p-value: < 2.2e-16
```

De estas dos estimaciones obtenemos importantes conclusiones. La primera es que los parámetros estimados son muy diferentes, esto es, para la primera submuestra obtenemos una pendiente $\hat{\beta}_1 \simeq 0,1$; y para la segunda submuestra $\hat{\beta}_1 \simeq 0,9$. La relación entre las variables ha cambiado, pues, considerablemente en el punto $t = 600$. Además, el ajuste de los dos submodelos es mucho mejor que para el modelo global, ya que las dos rectas estimadas se ajustan mucho mejor a los dos tramos de observaciones.

Bibliografía

Artís Ortuño, M.; del Barrio Castro, T.; Clar López, M.; Guillén Estany, M.; Suriñach Caralt, J. (2011). *Econometría*. Barcelona. Material didáctico UOC.

Liviano Solís, D.; Pujol Jover, M. (2013). *Matemáticas y Estadística con R*. Barcelona. Material didáctico UOC.