Modelos econométricos avanzados con R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211048

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Este módulo está dedicado al estudio de modelos econométricos avanzados, cuya especificación y estimación adquieren cierta complejidad. Aunque el enfoque de este módulo es eminentemente práctico, y está enfocado al desarrollo de aplicaciones prácticas con R y con R-Commander, al inicio de cada capítulo se ofrece una breve explicación teórica de los modelos analizados. Sin embargo, para entender los modelos aquí expuestos, es imprescindible que el estudiante primero haya trabajado los módulos teóricos de econometría, ya que el presente manual en ningún caso sustituye a los módulos teóricos, simplemente los complementa.

El primer capítulo está dedicado a la estimación de los modelos de regresión dinámicos y multiecuacionales con R-Commander. En economía, a menudo las relaciones entre las variables no se producen únicamente en el período analizado, sino que afectan a más de un período e incluso muchas perduran en el tiempo. Así pues, cuando se trabaja con datos temporales, podemos encontrarnos con frecuencia que las relaciones dejan de ser estáticas y pasan a ser dinámicas dando lugar a los modelos de regresión dinámicos. Como mostraremos a continuación, estos modelos tienen sus propias normas de especificación y de interpretación asociada a los parámetros estimados. También veremos los distintos problemas con los que nos podremos encontrar a la hora de estimarlas y cómo utilizar la solución idónea para solventar cada uno de ellos. Finalmente, también veremos cómo muchos acontecimientos económicos se explican por variables que son exógenas y endógenas al mismo tiempo, ya que se especifican por más de una ecuación; son los modelos multiecuacionales. En este apartado mostraremos la notación empleada para la especificación de estos, su estimación y la posterior interpretación correcta de los parámetros.

En el segundo capítulo se dedica al Modelo Lineal Generalizado (GLM, en sus siglas en inglés). Este modelo es una generalización flexible del modelo de regresión lineal ordinario y permite la inclusión de variables dependientes que generan distribuciones del error distintas de una distribución normal. Esto sucede, especialmente, cuando la variable dependiente es de naturaleza cualitativa (es decir, no representa magnitudes, sino diferentes atributos o categorías), o cuando, aun siendo una variable cuantitativa, su distribución dista de seguir una ley normal. Aunque el GLM abarca una gran multitud de distribuciones, en esta sección veremos los tres modelos de regresión más comunes. Esto es, para las variables dependientes cualitativas dicotómicas analizamos los modelos de regresión logit y probit, y para las variables de recuento estudiamos el modelo de regresión de Poisson. Aunque el GLM ofrece muchas más posibilidades, estas quedan fuera del alcance de este manual.

El tercer capítulo se encarga del estudio de los modelos con datos de panel. El término *panel de datos* hace referencia a un conjunto de datos con observaciones temporales

para los propios individuos, lo que permite el seguimiento de un mismo individuo durante un período de tiempo. Debido a la amplia disponibilidad de este tipo de datos, estos son utilizados en diferentes campos, como la economía, la demografía y las finanzas, por citar algunos. Al tratarse de un conjunto de técnicas más bien complejas, el análisis aplicado se realiza exclusivamente con código R, y está precedido por una amplia explicación de la naturaleza y las características analíticas de estos modelos, además de las diferentes posibilidades de especificación y estimación existentes.

Objetivos

1. Especificar correctamente los distintos tipos de modelos de regresión dinámicos.

2. Saber calcular el multiplicador de impacto, el multiplicador retardado j períodos y el multiplicador total con R-Commander.

3. Saber calcular el retardo medio y el mediano con R-Commander así como el número de períodos que deben transcurrir para alcanzar un porcentaje determinado del cambio que se producirá en la variable endógena ante una modificación en el valor de una variable explicativa.

4. Saber calcular utilizando el método de variables instrumentales los estimadores por mínimos cuadrados en dos etapas con R-Commander.

5. Saber escribir la forma estructural, reducida y final de un modelo multiecuacional, e interpretar el significado de sus parámetros.

6. Identificar las ecuaciones de un modelo multiecuacional.

7. Ser capaz de identificar las variables dependientes que requieren de la estimación de un Modelo Lineal Generalizado.

8. Saber elegir, en cada caso, el modelo más adecuado a las características de los datos objeto de estudio.

9. Saber interpretar correctamente los coeficientes estimados en un Modelo Lineal Generalizado.

10. Poder efectuar un análisis con un panel de datos.

11. Saber, en cada caso, qué tipo de efectos incluir en el modelo: individuales, temporales o ambos.

12. Elegir correctamente la especificación correcta para cada conjunto de datos de panel.

13. Acertar con el método de estimación adecuado con un panel de datos, según la existencia de efectos fijos o aleatorios.

1. Modelos de regresión dinámicos y multiecuacionales

1.1. Modelos de regresión dinámicos

En los modelos de regresión tratamos de explicar el comportamiento de una variable Y en función de k variables X. Por tanto, siempre que se altere el valor de alguna X se modificará el valor de Y.

Cuando trabajamos con series temporales no hay que olvidar que la temporalidad estará siempre implícita o explícitamente en la especificación del modelo simplemente porque es un aspecto clave que afecta al valor que pueda tomar la variable que hemos de explicar. Si no plasmamos en el modelo el efecto temporal, el modelo estimado no será el adecuado: los residuos podrían estar autocorrelacionados indicando la omisión de variables relevantes o de una mala especificación, la bondad del ajuste no será tan alta como se podría esperar, etc. La cuestión es que en el modelo existe una relación dinámica entre las variables que no hemos contemplado.

Es habitual encontrarnos con modelos con autocorrelación residual de orden 1 que capta la relación no contemporánea que se da en los modelos con series temporales y que no hemos incluido en la especificación. Esto nos explicita las relaciones existentes en el término de perturbación (entre u_t y u_{t-1}), que nos obligaría a reespecificar el modelo incluyendo variables retardadas. Las variables retardadas no siempre han de formar parte del conjunto de variables explicativas de la especificación inicial, a veces debemos incluir como explicativa la variable endógena retardada. Por tanto, los tipos de relaciones dinámicas no siempre son los mismos, pueden surgir al cabo de cierto tiempo, pueden estar presentes únicamente en el período posterior o pueden tener un efecto indefinido pero que se va diluyendo en el tiempo.

En los modelos dinámicos nos será muy útil el operador de retardos; recordemos cómo funciona: $Y_{t-1} = LY_t$; y, en general, $Y_{t-j} = L^j Y_t$.

1.1.1. Tipos de modelos dinámicos

Los principales modelos dinámicos que podemos especificar son:

1) Modelos de retardos distribuidos (RD). En un modelo de retardos distribuidos de orden s ($RD(s)$) se explicita la dinamicidad con la introducción de variables exógenas retardadas como regresores. El orden del modelo de retardos distribuidos será el número de retardos de las variables exógenas, que puede ser finito o infinito. Por ejemplo, si una variable Y se explica por otra variable X, el modelo $RD(s)$ podría ser:

No olvidéis que un modelo $RD(2)$ siempre será estable y que sus parámetros β se interpretan como los multiplicadores (retardados o contemporáneos) de X sobre Y.

$$Y_t = \mu + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t \qquad t = 1, \ldots, T$$

que en función del operador de retardos da lugar a:

$$Y_t = \mu + \beta_0 X_t + \beta_1 L X_t + \beta_2 L^2 X_t + \ldots + \beta_s L^s X_t + u_t$$

$$= \mu + (\beta_0 + \beta_1 L + \beta_2 L^2 + \ldots + \beta_s L^s) X_t + u_t = \mu + B(L) X_t + u_t$$

Si en lugar de tener inicialmente un MRLS tuviéramos un MRLM, la especificación de un modelo $RD(s_1, s_2, s_3, \ldots, s_k)$ quedaría como:

$$Y_t = \mu + B_1(L) X_{1t} + B_2(L) X_{2t} + \ldots + B_k(L) X_{kt} + u_t$$

donde $B_j(L)$ es el polinomio s_j asociado a la variable X_{ij}.

2) Modelos autorregresivos (AR). Un modelo autorregresivo de orden r es aquel que incluye una serie de variables explicativas entre las que figuran también la misma variable endógena retardada como señal de la relación no contemporánea existente. Este modelo se define como:

En un modelo AR de orden finito una variación en X tendrá un efecto temporal indefinido sobre Y y su impacto total puede ser finito o no; mientras que en un modelo RD el impacto es finito y se distribuye en un período temporal determinado.

$$Y_t = \mu + \beta_0 X_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + u_t \qquad t = 1, \ldots, T$$

que en función del operador de retardos se transforma en:

$$Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - \ldots - \alpha_r Y_{t-r} = \mu + \beta_0 X_t + u_t$$

$$(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r) Y_t = \mu + \beta_0 X_t + u_t$$

$$A(L) Y_t = \mu + \beta_0 X_t + u_t$$

3) La hipótesis de Koyck. transforma un modelo de retardos distribuidos de orden infinito en un modelo autorregresivo de primer orden. Partimos de un modelo $RD(\infty)$:

Especificar un modelo dinámico $RD(\infty)$ o $AR(\infty)$ puede provocar problemas porque nos obliga a estimar infinitos parámetros. Por este motivo, una solución consiste en reespecificar el modelo en un AR o RD de orden finito, respectivamente.

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + u_t$$

y aplicamos la hipótesis de Koyck, que supone que el valor de los parámetros **B** disminuye de forma geométrica ($\beta_j = \beta_0 \delta^j$, $\forall\ j = 0, 1, \ldots$) donde $0 < \delta < 1$. De manera que cada vez que nos alejamos en el tiempo la influencia de X sobre Y va disminuyendo.

Si tenemos en cuenta la hipótesis anterior y la sustituimos en el modelo $RD(\infty)$ especificado, llegamos a un modelo $AR(1)$,

$$Y_t = \alpha(1-\delta) + \beta_0 X_t + \delta Y_{t-1} + v_t$$

siendo

$$v_t = u_t - \delta u_{t-1}$$

Este nuevo modelo tiene la ventaja de la reducción del número de parámetros que hay que estimar y los posibles problemas de multicolinealidad. Por contra, tenemos como explicativa la variable endógena retardada y el nuevo término de perturbación no es esférico, lo que nos condiciona el método de estimación que se debe utilizar.

4) Modelos autorregresivos y de retardos distribuidos (AD). Un modelo $AD(r, s)$ se define como:

$$A(L)Y_t = \mu + B(L)X_t + u_t$$

$$(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r)Y_t = \mu + (\beta_0 + \beta_1 L + \beta_2 L^2 + \ldots + \beta_s L^s)X_t + u_t$$

$$Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t$$

Si tuviéramos más de una variable explicativa, el modelo sería todavía más general; $X_j \ \forall \ j = 0, 1, \ldots, k$ dando lugar a $AD(r;\ s_1, s_2, ..., s_k)$. Por el contrario, modelos más simples son el $RD(s)$, que podría definirse como un $AD(0, s)$, y el $AR(r)$, definido como un $AD(r, 0)$.

5) Otros modelos dinámicos. Normalmente especificamos modelos basados en la teoría económica pero también podemos hacerlo utilizando la información histórica disponible para decidir el número de retardos asociados de cada una de las variables incluidos en la especificación. Entre los más habituales encontramos:

a) Los **modelos ARIMA(p,d,q) univariantes**:

Los modelos ARIMA(p,d,q) son modelos ARMA(p,q) que han tenido que diferenciarse d veces para convertirlos en estacionarios en media.

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p)Y_t = \mu + (\theta_0 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q)\varepsilon_t$$

$$\phi(L)Y_t = \mu + \theta(L)\varepsilon_t$$

b) Los **modelos de función de transferencia** se basan en la teoría de la cointegración, que combina la econometría clásica con el análisis de series temporales, que a su vez trata de evitar la modelización de relaciones espurias entre variables. El llamado modelo de mecanismo de corrección del error se especifica como:

La relación entre dos variables es espuria cuando no viene dada por la relación entre dichas variables sino por otras causas, como por ejemplo una tercera variable.

$$Y_t = \mu + [B(L)/A(L)]X_t + [\phi(L)/\theta(L)]\varepsilon_t$$

1.1.2. Análisis e interpretación de los modelos dinámicos

Además de la interpretación de los parámetros estimados y bondad del ajuste, en los modelos se deben tener en cuenta otros aspectos relacionados con los efectos que la temporalidad tiene en la interpretación económica de los parámetros:

1) El modelo estimado ha de ser **estable**, es decir, si se produce una variación puntual de cualquier X en un determinado momento t, la variable Y vuelve a su valor de equilibrio haciendo que el efecto total derivado de dicha variación sea finito. También se trata de un modelo estable si ante una variación permanente de X la variable Y evoluciona hacia un nuevo valor de equilibrio.

Si ante una variación puntual de X se alcanza un nuevo valor de equilibrio de Y, se trata de un modelo **inestable explosivo**, pero si no se alcanza ese nuevo valor de equilibrio el modelo es **inestable no explosivo**.

Veamos un resumen en la tabla siguiente:

Modelo	Variación de X	Valor inicial X	V. Transitorio X	V. Final X	V. Inicial Y	V. Transitorio Y	V. Final Y
Estable	Puntual	X_0	X_1	X_0	Y_0	Y_1	Y_0
Estable	Permanente	X_0	X_1	X_1	Y_0	Y_1	Y_1
Inestable expl.	Puntual	X_0	X_1	X_0	Y_0	Y_1	Y_1
Inestable no exp.	Puntual	X_0	X_1	X_0	Y_0	Y_1	$\nexists$
Inestable no exp.	Permanente	X_0	X_1	X_1	Y_0	Y_1	$\nexists$

El análisis de la estabilidad del modelo parte de la ecuación $A(L)Y_t = \mu + B(L)X_t + u_t$ y consiste en comprobar que las raíces del polinomio autorregresivo $A(L)$ ($1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_r L^r = 0$) caigan fuera del círculo de la unidad ($\forall L > 1$). Si esto se cumple, la Y recupera su valor de equilibrio ante una variación puntual de X, o alcanza un nuevo valor de equilibrio si la variación de X es permanente.

2) Se han de diferenciar los efectos contemporáneos (**multiplicador de impacto o contemporáneo**) de los no contemporáneos que tienen las X sobre la Y (**multiplicadores de retardados o una vez transcurridos j períodos, y multiplicador total**).
Vayamos por partes; dado el modelo:

Los multiplicadores son los parámetros del modelo y se definen como $\beta_j = \frac{\delta Y}{\delta X_j}$ $j = 2, \ldots, k$, es decir, nos explican el cambio en Y ante un cambio unitario de cualquier X.

$$Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_r Y_{t-r} + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_s X_{t-s} + u_t$$

definimos los conceptos de multiplicador siguientes:

a) El **multiplicador de impacto o contemporáneo** (m_0) es la variación de Y_t ante una variación unitaria de X_t, es decir, $m_0 = \dfrac{\partial Y_t}{\partial X_t} = \beta_0$

b) El **efecto multiplicativo tras haber transcurrido j períodos** (m_j) será la variación producida en Y_t ante una variación unitaria de X_{t-j}, o sea, $m_j = \dfrac{\partial Y_t}{\partial X_{t-j}}$.

Cuando $m_j \neq \beta_j$, observamos la existencia de una dependencia implícita de las variables endógenas retardadas.

c) Finalmente, el **multiplicador total (m_T)** es la suma de todos los multiplicadores contemporáneos y no contemporáneos, esto es, $m_T = \sum_{j=0}^{\infty} m_j$.

Para calcular los multiplicadores es muy útil utilizar la siguiente expresión:

$$Y_t = \mu/A(L) + [B(L)/A(L)]X_t + u_t/A(L) = \mu' + D(L)X_t + v_t$$

Cuando se trata de un modelo RD hallar los multiplicadores resulta relativamente sencillo, ya que se reducen a la identificación de las distintas β que componen el polinomio $B(L)$. En cambio, cuando existe parte AR se debe calcular $D(L) = \frac{B(L)}{A(L)}$. Además, el m_0 y el m_T serán respectivamente los valores de $D(L)$ cuando L es cero y uno, respectivamente.

donde $D(L) = \delta_0 + \delta_1 L + \delta_2 L^2 + \ldots$ ya que:

$$m_0 = \frac{\partial Y_t}{\partial X_t} = D(0) = \delta_0$$

$$m_j = \frac{\partial Y_t}{\partial X_{t-j}} = \delta_j$$

$$m_T = \sum_{j=0}^{\infty} m_j = D(1) = \delta_0 + \delta_1 + \delta_2 + \ldots$$

3) El **retardo medio** y el **retardo mediano** también nos ayudan a la interpretación de los resultados de la estimación de los modelos dinámicos.

a) El retardo medio no es más que la media ponderada de los coeficientes del polinomio $D(L)$ del modelo $Y_t = \mu' + D(L)X_t + v_t$, es decir:

$$Retardo\,medio = \frac{\sum_{j=0}^{\infty} j \cdot \delta_j}{\sum_{j=0}^{\infty} \delta_j} = \frac{D'(1)}{D(1)} = \frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$$

donde $D'(L)$, $B'(L)$ y $A'(L)$ son las derivadas de $D(L)$, $B(L)$ y $A(L)$, respecto a L y todos estos polinomios valorados en $L = 1$. De modo que un valor elevado del retardo medio implica que la contribución al comportamiento de Y de los períodos alejados en el tiempo es alta, lo que indica la concentración o dilución de los efectos de las variables exógenas.

b) El retardo mediano indica el momento en el que se alcanza el 50 % de la variación total que se produce en Y a causa de una variación en X.

Si se desea también se pueden realizar esos mismos cálculos con una hoja de cálculo como Excel.

Veamos un ejemplo de todo ello utilizando la consola de R como si fuera una calculadora. Si quisiéramos explicar el número de cervezas vendidas trimestralmente de una determinada marca (C) en función del precio medio de la cerveza (P) y el gasto trimestral en campañas de publicidad que ha realizado dicha marca de cerveza (CP) y, si fuera necesario, las ventas realizadas de esa misma marca en trimestres anteriores;

para un trimestre t determinado, podríamos especificar cualquiera de los siguientes modelos:

1) Modelo 1. Somos partidarios de la simplicidad y realmente explicamos nuestras ventas en función del precio y de los gastos que se han realizado en la campaña publicitaria del período:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + u_t \qquad t = 1, \ldots, T.$$

este modelo puede que no sea muy bueno, podríamos fácilmente detectar una autocorrelación de los residuos derivada de una mala especificación o de una omisión de variables relevantes, ya que no hemos considerado ninguna relación dinámica en nuestro modelo.

2) Modelo 2. Estamos convencidos de que los gastos en las campañas publicitarias anteriores tienen mucho impacto en las ventas de los trimestres posteriores sencillamente porque no es tan fácil que nuestros clientes se olviden de las canciones pegadizas que conscientemente hemos seleccionado para las campañas de primavera y verano; nuestro modelo pasaría a ser:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \beta_3 CP_{t-1} + \beta_4 CP_{t-2} + \beta_5 CP_{t-3} + u_t \qquad t = 1, \ldots, T.$$

3) Modelo 3. Consideramos que son las ventas de los trimestres anteriores sobre las que se basan las ventas actuales:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \alpha_1 C_{t-1} + \alpha_2 C_{t-2} + u_t \qquad t = 1, \ldots, T.$$

4) Modelo 4. Estamos de acuerdo con las dos teorías anteriores:

$$C_t = \beta_0 + \beta_1 P_t + \beta_2 CP_t + \beta_3 CP_{t-1} + \beta_4 CP_{t-2} + \beta_5 CP_{t-3} + \alpha_1 C_{t-1} + \alpha_2 C_{t-2} + u_t$$

$$t = 1, \ldots, T.$$

Una vez elegida la especificación que más se ajuste a nuestro entorno, deberíamos valorar dicho modelo utilizando la información con la que contamos. Supongamos que tenemos información trimestral de los últimos 10 años. Está claro que en este ejemplo tratamos con datos de periodicidad trimestral que supone un conjunto de 40 datos; es decir, que en nuestro caso, cuando especificamos $t = 1, \ldots, T$, nuestra $T = 40$. Supongamos ahora que las estimaciones de los modelos propuestos son:

1) Modelo 1:

$$C_t = 73{,}96 + 16{,}83 P_t + 0{,}07 CP_t$$

2) Modelo 2:

$$C_t = 59{,}52 + 20{,}51P_t + 0{,}10CP_t + 0{,}02CP_{t-1} - 0{,}05CP_{t-2} + 0{,}09CP_{t-3}$$

3) Modelo 3:

$$C_t = 127{,}62 + 22{,}44P_t + 0{,}35CP_t + 0{,}55C_{t-1} + 0{,}15C_{t-2}$$

4) Modelo 4:

$$C_t = 105{,}16 + 25{,}61P_t + 0{,}5CP_t + 0{,}35CP_{t-1} - 0{,}15CP_{t-2} + 0{,}12CP_{t-3} + 0{,}65C_{t-1} + 0{,}25C_{t-2}$$

Primero determinaremos el tipo y la estabilidad de los modelos estimados:

1) Modelo 1. Se trata de un modelo estático. No tiene sentido hablar de estabilidad.

2) Modelo 2. Es un modelo dinámico $RD(3)$. La parte de retardos distribuidos de orden 3 asociada a CP será $(0{,}10 + 0{,}02L - 0{,}05L^2 + 0{,}09L^3)$CPt. Por definición todos los modelos de retardos distribuidos son siempre estables.

3) Modelo 3. Es un modelo dinámico $AR(2)$. La parte autorregresiva de orden 2 será $(1 - 0{,}55L - 0{,}15L^2)$. Para analizar la estabilidad del modelo primero lo reescribimos:

$$(1 - 0{,}55L - 0{,}15L^2)\hat{C}_t = 127{,}62 + 22{,}44P_t + 0{,}35CP_t$$

después, calculamos las raíces del polinomio de retardos y miramos si caen o no fuera del círculo de la unidad. Efectuando los cálculos manualmente debemos escribir en la ventana de instrucciones:

```
m3 <- c(1,-0.55,-0.15)
polyroot(m3)
```

Al seleccionar lo escrito y pulsar *Ejecutar*, obtendremos en la ventana de resultados:

```
> m3 <- c(1,-0.55,-0.15)
> polyroot(m3)
[1]  1.333333+0i -5.000000-0i
```

Así, se trata de un modelo estable porque sus raíces son mayores que la unidad en valor absoluto; $L_i > |1|$

4) Modelo 4. Es un modelo dinámico $AD(2,3)$. La parte autorregresiva de orden 2 será $(1 - 0{,}65L - 0{,}25L^2)$, y la parte de retardos distribuidos (de orden 3) asociada a CP será $(0{,}5 + 0{,}35L - 0{,}15L^2 + 0{,}12L^3)$CPt. Una vez reescrito el modelo:

$$(1 - 0{,}65L - 0{,}25L^2)\hat{C}_t = 105{,}16 + 25{,}61P_t + (0{,}5 + 0{,}35L - 0{,}15L^2 + 0{,}12L^3)CP_t$$

analizamos su estabilidad del mismo modo que hemos hecho antes. En la ventana de resultados aparece:

```
> m4 <- c(1,-0.65,-0.25)
> polyroot(m4)
[1]  1.085372+0i -3.685372+0i
```

Ahora calcularemos los multiplicadores con los modelos expresados en función de sus polinomios de retardos:

1) Modelo 1. No procede.

2) Modelo 2. Es un modelo estable $RD(3)$ que se puede expresar como:

$$\hat{C}_t = 59{,}52 + 20{,}51P_t + (0{,}10 + 0{,}02L - 0{,}05L^2 + 0{,}09L^3)CP_t$$

Sabemos que en este tipo de modelos los parámetros pueden interpretarse como los multiplicadores contemporáneos o retardados de CP sobre C. Además también podemos calcular su retardo medio y su retardo mediano como sigue:

Concepto	Cálculo	Resultado
Multiplicador contemporáneo	β_2	0,10
Multiplicador retardado 1 período	β_3	0,02
Multiplicador retardado 2 períodos	β_4	−0,05
Multiplicador retardado 3 períodos	β_5	0,09
Multiplicador total	$\beta_2 + \beta_3 + \beta_4 + \beta_5$	$0{,}10 + 0{,}02 - 0{,}05 + 0{,}09 = 0{,}16$
Retardo medio	$\frac{B'(1)}{B(1)}$	$\frac{0{,}02-2*0{,}05+3*0{,}09}{0{,}16} = \frac{0{,}19}{0{,}16} = 1{,}1875$
Retardo mediano	t en alcanzar 0,08	9 meses y 18 días del mismo año

3) Modelo 3. Es un modelo estable $AR(2)$ que se puede expresar como:

$$(1 - 0{,}55L - 0{,}15L^2)\hat{C}_t = 127{,}62 + 22{,}44P_t + 0{,}35CP_t$$

Sus multiplicadores contemporáneos o retardados de CP sobre C así como sus retardo medio y mediano son:

Concepto	Cálculo	Resultado
Multiplicador contemporáneo	β_2	0,35
Multiplicador retardado 1 período	$\alpha_1 * \beta_2$	0,19
Multiplicador retardado 2 períodos	$(\alpha_1^2 + \alpha_2) * \beta_2$	0,16
Multiplicador retardado 3 períodos	$(\alpha_1^3 + 2\alpha_1\alpha_2) * \beta_2$	0,12
Multiplicador total	$\delta_0 + \delta_1 + \delta_2 + \delta_3$	0,35 + 0,19 + 0,16 + 0,12 = 0,82
Retardo medio	$\frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$	$\frac{0{,}00}{0{,}35} - \frac{-0{,}55-2*0{,}15}{1-0{,}55-0{,}15} = 0 - \frac{-0{,}85}{0{,}30} = 2{,}83$
Retardo mediano	t en alcanzar 0,41	3 meses y 18 días del próximo año

4) Modelo 4. Es un modelo estable $AD(2, 3)$ que se puede expresar como:

$$(1 - 0{,}65L - 0{,}25L^2)\hat{C}_t = 105{,}16 + 25{,}61P_t + (0{,}5 + 0{,}35L - 0{,}15L^2 + 0{,}12L^3)CP_t$$

Sus multiplicadores contemporáneos o retardados de CP sobre C así como sus retardo medio y mediano son:

Concepto	Cálculo	Resultado
Multiplicador contemporáneo	β_2	0,50
Multiplicador retardado 1 período	$\beta_3 + \alpha_1 * \beta_2$	0,675
Multiplicador retardado 2 períodos	$\beta_4 + \alpha_1 * \beta_3 + (\alpha_1^2 + \alpha_2) * \beta_2$	0,414
Multiplicador retardado 3 períodos	$\beta_5 + \alpha_1 * \beta_4 + (\alpha_1^2 + \alpha_2) * \beta_3 + (\alpha_1^3 + 2\alpha_1\alpha_2) * \beta_2$	0,558
Multiplicador total	$\delta_0 + \delta_1 + \delta_2 + \delta_3$	0,5 + 0,675 + 0,414 + 0,558 = 2,146
Retardo medio	$\frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$	$\frac{0{,}35-2*0{,}15+3*0{,}12}{0{,}5+0{,}35-0{,}15+0{,}12} - \frac{-0{,}65-2*0{,}25}{1-0{,}65-0{,}25} = \frac{0{,}41}{0{,}82} - \frac{-1{,}15}{0{,}10} = 12$
Retardo mediano	t en alcanzar 0,85	10 meses y 6 días del próximo año

En resumen:

Modelo	m_0	m_1	m_2	m_3	m_T	Retardo medio	Retardo mediano
Modelo 1	∄	∄	∄	∄	∄	∄	∄
Modelo 2	0,10	0,02	−0,05	0,09	0,16	1,1875	9 meses y 18 días del mismo año
Modelo 3	0,35	0,19	0,16	0,12	0,82	2,83	3 meses y 18 días del próximo año
Modelo 4	0,50	0,675	0,414	0,558	2,146	12	10 meses y 6 días del próximo año

1.1.3. Métodos de estimación

1) Estimación por MCO en modelos con variables exógenas retardadas. Son los modelos de retardos distribuidos $RD(s)$. Si el término de perturbación cumple todas las hipótesis básicas (tiene una matriz de varianzas y covarianzas escalar), podemos utilizar los estimadores MCO sin problemas porque serán **insesgados, eficientes y consistentes**. Sin embargo pueden aparecer algunos problemas:

a) A medida que aumenta el número de retardos de la variable exógena tendremos menos grados de libertad y disminuirá la fiabilidad de la estimación.

b) Se puede presentar un nivel de multicolinealidad elevada por utilizar como regresores la misma variable referida a diferentes momentos del tiempo.

Recordad que la multicolinealidad se trata en el módulo 2 del presente manual.

En caso de que el término de perturbación fuera no esférico esférico, deberíamos utilizar MCG para obtener estimadores eficientes.

Podéis consultar el módulo 2 del presente manual para profundizar en la estimación por MCG.

Finalmente, si nos encontráramos con la estimación de un modelo de retardos distribuidos con un número infinito de retardos no podríamos estimar por falta de datos y deberíamos transformar el modelo utilizando la hipótesis de Koyck.

2) Estimación por MCO en modelos con variables explicativas incorrelacionadas con el término de perturbación. Son modelo dinámicos en los que utilizamos como variables explicativas retardos de la propia variable endógena, y por tanto, podemos tener problemas de estimación porque tendremos regresores estocásticos. La clave está en ver si éstas están o no correlacionadas con el término de perturbación.

Para que el término de perturbación (u_t) esté incorrelacionado con los regresores su matriz de varianzas y covarianzas debe ser escalar. De este modo, todas las variables explicativas tendrán una distribución independiente del término de perturbación. Lo que ocurre es que Y_{t-1} no es independiente del término de perturbación en t ? 1 y en períodos anteriores, pero sí que es independiente en t y en períodos posteriores. Así pues, utilizando el teorema de Mann-Wald, llegamos a la conclusión de que el estimador MCO es **asintóticamente insesgado, consistente y**, si U se distribuye según una Normal, también **asintóticamente eficiente**.

3) Estimación por MCO en modelos con variables explicativas correlacionadas con el término de perturbación. En estos modelos dinámicos también se utilizan como regresores variables endógenas retardadas y están correlacionadas con el término de perturbación. Por tanto, las estimaciones por MCO son **sesgadas** (y el sesgo no tiende a cero al aumentar el tamaño de la muestra), **inconsistentes** porque no se cumple el teorema de Mann-Wald **e ineficientes** debido a la no esfericidad del término de perturbación u_t.

4) Métodos de estimación alternativos:

a) El método de estimación de *variables instrumentales (VI)* asegura la consistencia de los estimadores y se utiliza tanto en modelos dinámicos con endógenas retardadas correlacionadas con un término de perturbación autocorrelacionado como en modelos en los que hay correlación entre variables explicativas y el término de perturbación a causa del no cumplimiento del supuesto de exogeneidad de los regresores.

Recordad que el teorema de Mann-Wald asegura la consistencia de los estimadores.

En el caso que nos atañe partimos de un modelo $Y = ZB + U$, donde Z es la matriz de dimensión $T \times k$ de variables explicativas del modelo dinámico (que incluye variables endógenas retardadas). Aplicando MCO los estimadores eran inconsistentes porque existía correlación entre Z_t y u_t; por tanto $E[Z_t' u_t] \neq 0$ y no podíamos aplicar el teorema de Mann-Wald.

Lo que se propone con el método de las VI es definir una nueva matriz W_t, con tantas variables como Z_t, de manera que estas nuevas variables (instrumentos) estén lo más correlacionadas como sea posible con las variables iniciales y no lo estén con el término de perturbación del modelo para poder así aplicar el teorema de Mann-Wald y estimar por MCO.

El problema radica entonces en encontrar los instrumentos adecuados para realizar la estimación. Podemos encontrarnos ante tres posibilidades:

1) Obtener una *proxy* que nos sirva de instrumento para definir la nueva matriz W_t.
2) Utilizar como instrumento la variable que mejor explique el comportamiento de la variable endógena y así poder definir la nueva matriz W_t.
3) Obtener una estimación del instrumento a partir de una regresión auxiliar para definir la nueva matriz W_t. Cuando la variable que nos crea problemas es la endógena retardada, esta regresión auxiliar se calcula utilizando como endógena la endógena retardada y como explicativas todas las variables explicativas (excepto la propia endógena retardada) del modelo inicial retardadas un período. Si utilizamos esta última posibilidad, los estimadores obtenidos se denominan **estimadores por mínimos cuadrados en dos etapas** y tienen la particularidad de proporcionar unas estimaciones más eficientes de los parámetros iniciales que si utilizásemos otro tipo de instrumentos para la endógena retardada.

b) El método de estimación de *mínimos cuadrados no lineales (MCNL)* basado en minimizar la suma de los cuadrados de los residuos.

c) El método de estimación de *máxima verosimilitud (MV)* consiste en maximizar el logaritmo neperiano de la función de verosimilitud.

Los estimadores de estos dos últimos métodos coinciden si el término de perturbación sigue una ley de distribución Normal.

Fijaos en que el estimador de mínimos cuadrados en dos etapas está más que justificado en el modelo especificado porque incluye la variable endógena retardada un período entre sus regresores.

Veamos ahora cómo realizar con R-Commander una estimación por VI utilizando mínimos cuadrados en dos etapas. Para ello recuperaremos la esencia del ejemplo. Intentaremos explicar el número de cervezas vendidas trimestralmente de una determinada marca (C) en función de la variación del precio medio de la cerveza (P) en el trimestre actual y en el anterior; el gasto trimestral en campañas de publicidad que ha realizado en los tres últimos trimestres (CP), el porcentaje de población adulta que puede consumir cerveza (A). Además consideraremos que las ventas realizadas de esa misma marca en el trimestre anterior también es determinante para calcular las ventas actuales. Así, nuestro nuevo modelo será:

$$C_t = \mu + \beta_0 CP_t + \beta_1 CP_{t-1} + \beta_2 CP_{t-2} + \alpha_0 P_t + \alpha_1 P_{t-1} + \delta_0 A_t + \gamma_0 C_{t-1} + u_t$$

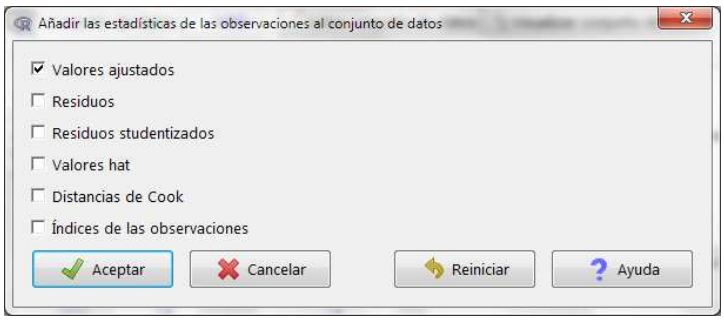
$$t = 1, \ldots, T.$$

Para estimar el modelo seguiremos los pasos siguientes:

a) Leer la base de datos denominándola MRD y ajustar el modelo por MCO excluyendo la variable C_{t-1}, denominándolo *Modelo*1. En la ventana de resultado nos aparece lo siguiente:

```
> Modelo1 <- lm(C~A+CP+CP_1+CP_2+P+P_1, data=MRD)
> summary(Modelo1)
Call:
lm(formula = C ~ A + CP + CP_1 + CP_2 + P + P_1, data = MRD)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07805 -0.03228 -0.01139  0.02653  0.11847

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) -4.81843    0.36083 -13.354 2.13e-14 ***
A            0.18891    0.29194   0.647 0.522339
CP           1.40129    0.36461   3.843 0.000564 ***
CP_1         0.72265    0.50591   1.428 0.163171
CP_2        -0.91264    0.35998  -2.535 0.016498 *
P            0.26388    0.04881   5.406 6.71e-06 ***
P_1         -0.15628    0.05605  -2.788 0.008973 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05319 on 31 degrees of freedom
Multiple R-squared:  0.9975,  Adjusted R-squared:  0.9971
F-statistic:  2090 on 6 and 31 DF,  p-value: < 2.2e-16
```

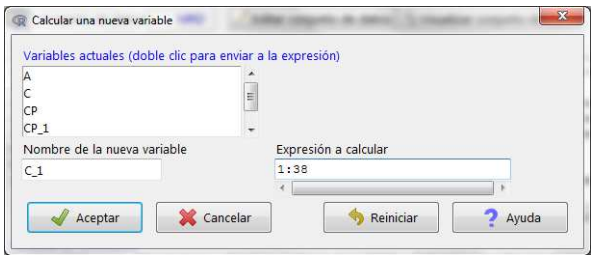
b) Guardar los valores ajustados añadiendo una nueva variable a la base de datos denominada $fitted.Modelo1$. Para ello seguiremos la ruta *Modelos / Añadir las estadísticas de las observaciones a los datos...*, y en el cuadro de diálogo que aparece marcaremos únicamente la opción de valores ajustados tal y como se muestra a continuación:



c) Construir la variable C_{t-1}. Esto lo podemos hacer directamente en la base de datos: en primer lugar, utilizamos la función `nrow` para asegurarnos del número exacto de las observaciones de nuestra base de datos, que denominaremos n.

```
n <- nrow (MRD)
```

En segundo lugar, crearemos una nueva variable llamada C_1, utilizando la opción *Datos / Modificar variables del conjunto de datos activo / Calcular una nueva variable* que vaya desde 1 hasta *n*:



Una vez creada la variable, que podía haber sido también una columna de ceros, de unos o de valores perdidos, simplemente reemplazando el 1 : 38 por 0, 1 o *NA*, respectivamente, reemplazaremos el de la variable C_1 por el de la variable C a partir de la segunda observación y dejando vacía la primera fila. Esto es muy sencillo utilizando el siguiente código:

Para más información sobre lo que hace este código podéis consultar el módulo 1 del manual *Matemáticas y Estadística con R.*

```
for (i in 1:37) {MRD$C_1[i+1] <- MRD$C[i]}
MRD$C_1[1] <- NA
```

Podemos comprobar que lo hemos hecho bien visualizando el conjunto de datos:

MRD

	C	CP	CP_1	CP_2	P	P_1	A	fitted.Modelo1	C_1
1	4.634729	7.796058	7.725330	7.654917	-0.6931472	-0.6931472	0.09139872	4.645290	NA
2	4.699257	7.850493	7.796058	7.725330	-0.6931472	-0.6931472	0.16778523	4.722850	4.634729
3	4.763786	7.881560	7.850493	7.796058	-0.6931472	-0.6931472	0.13737600	4.735427	4.699257
4	4.828314	7.967280	7.881560	7.850493	-0.3566750	-0.6931472	0.08050962	4.906362	4.763786
5	4.940471	8.024535	7.967280	7.881560	-0.3566750	-0.3566750	0.16395357	4.983363	4.828314
6	5.052629	8.130059	8.024535	7.967280	-0.3566750	-0.3566750	0.18042098	5.097487	4.940471
7	5.164786	8.231376	8.130059	8.024535	-0.5108256	-0.3566750	0.13638264	5.214468	5.052629
8	5.335131	8.313852	8.231376	8.130059	-0.5108256	-0.5108256	0.14750044	5.333144	5.164786
9	5.480639	8.368925	8.313852	8.231376	-0.3566750	-0.5108256	0.18322665	5.424878	5.335131
10	5.545177	8.439232	8.368925	8.313852	0.0953102	-0.3566750	0.11559507	5.570327	5.480639
11	5.645447	8.551595	8.439232	8.368925	-0.2231435	0.0953102	0.17827916	5.585495	5.545177
12	5.849325	8.619027	8.551595	8.439232	0.0000000	-0.2231435	0.08542789	5.788134	5.645447
13	6.075346	8.712595	8.619027	8.551595	0.6931472	0.0000000	0.18180602	6.031674	5.849325
14	6.146329	8.783243	8.712595	8.619027	0.7419373	0.6931472	0.13118231	6.031729	6.075346
15	6.240276	8.841304	8.783243	8.712595	0.9162907	0.7419373	0.15595625	6.121811	6.146329
16	6.326149	8.949625	8.841304	8.783243	1.0296194	0.9162907	0.11122376	6.245288	6.240276
17	6.393591	9.049937	8.949625	8.841304	0.9555114	1.0296194	0.14665164	6.380569	6.326149
18	6.489205	9.152605	9.049937	8.949625	0.9932518	0.9555114	0.16918275	6.523866	6.393591
19	6.622736	9.286190	9.152605	9.049937	0.9555114	0.9932518	0.19783198	6.683256	6.489205
20	6.673298	9.296885	9.286190	9.152605	1.0986123	0.9555114	0.14955891	6.735619	6.622736
21	6.744059	9.366575	9.296885	9.286190	1.5040774	1.0986123	0.08129425	6.790822	6.673298
22	6.814543	9.438113	9.366575	9.296885	1.3609766	1.5040774	0.12606960	6.838997	6.744059
23	6.892642	9.496947	9.438113	9.366575	1.2809338	1.3609766	0.08754240	6.903501	6.814543
24	6.962243	9.533872	9.496947	9.438113	1.3083328	1.2809338	0.12945782	6.960128	6.892642
25	7.006695	9.610525	9.533872	9.496947	1.3083328	1.3083328	0.08751287	7.028325	6.962243
26	7.057898	9.673068	9.610525	9.533872	1.1939225	1.3083328	0.10901158	7.111529	7.006695
27	7.109879	9.716616	9.673068	9.610525	1.0296194	1.1939225	0.16130477	7.132196	7.057898
28	7.170120	9.727585	9.716616	9.673068	1.2237755	1.0296194	0.13504321	7.193910	7.109879
29	7.229114	9.750977	9.727585	9.716616	1.5260563	1.2237755	0.12738576	7.242847	7.170120
30	7.275865	9.785998	9.750977	9.727585	1.6292405	1.5260563	0.09479042	7.272643	7.229114
31	7.331715	9.838576	9.785998	9.750977	1.6486586	1.6292405	0.13229432	7.346362	7.275865
32	7.386471	9.888425	9.838576	9.785998	1.5040774	1.6486586	0.11563319	7.377914	7.331715
33	7.483244	9.979012	9.888425	9.838576	1.3862944	1.5040774	0.17261532	7.495172	7.386471
34	7.533694	10.010547	9.979012	9.888425	1.4350845	1.3862944	0.18998090	7.593892	7.483244
35	7.591862	10.056809	10.010547	9.979012	1.4109870	1.4350845	0.11578445	7.570833	7.533694
36	7.639642	10.091957	10.056809	10.010547	1.3609766	1.4109870	0.15743518	7.623175	7.591862
37	7.707063	10.148393	10.091957	10.056809	1.2527630	1.3609766	0.15052175	7.663390	7.639642
38	7.762171	10.183881	10.148393	10.091957	1.1314021	1.2527630	0.10899437	7.698868	7.707063

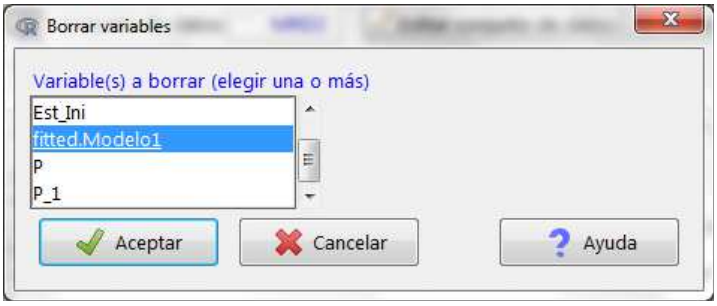
d) Construir la matriz W creando una nueva base de datos idéntica a la anterior, que llamaremos MRD2 mediante la instrucción:

```
MRD2 <- MRD
```

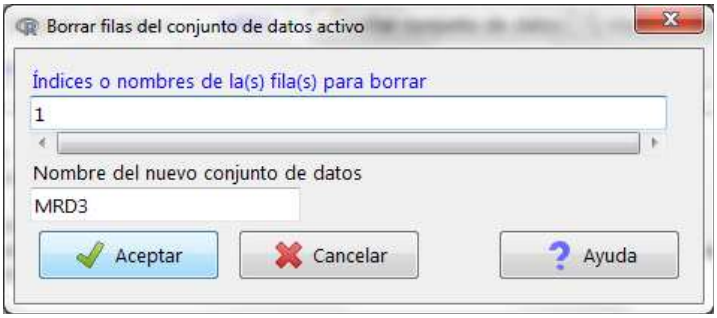
A esta nueva base de datos le insertaremos la variable de estimación del modelo retardada un período y la llamaremos Est_1ni; esto lo haremos como hemos hecho antes, cuando hemos creado C_1. Es decir:

```
MRD2$Est_Ini <- 1:38
for (i in 1:37) {MRD2$Est_Ini[i+1] <- MRD2$fitted.Modelo1[i]}
MRD2$Est_Ini[1] <- NA
```

Después eliminaremos de la base de datos la variable $fitted.Modelo1$ mediante la ruta *Datos / Modificar variables del conjunto de datos activo / Eliminar variables del conjunto de datos*, y eligiendo la variable que debemos eliminar en el cuadro de diálogo:



Luego también eliminaremos la primera observación utilizando el menú *Datos / Conjunto de datos activo / Borrar fila(s) del conjunto de datos activo*. Como vemos, aquí tenemos la posibilidad de dar un nuevo nombre a la base de datos, hagamos que sea MRD3:



En estos momentos seguiremos teniendo el mismo número de variables pero ahora tendremos una observación menos. A continuación crearemos una nueva variable llamada $UNOS$ que contenga una columna de unos utilizando de nuevo el menú *Datos / Modificar variables del conjunto de datos activo / Calcular una nueva variable*.

Finalmente, volcaremos esta nueva base datos en el espacio de trabajo para crear las matrices que necesitemos mediante la función `attach`. La matriz W contendrá las variables: $UNOS$, CP, CP_1, CP_2, P, P_1, A, Est_Ini y en el orden indicado:

```
attach (MRD3)
W <- matrix (c(UNOS, CP, CP_1, CP_2, P, P_1, A, Est_Ini),37,8)
```

Recordad que las operaciones con matrices se tratan en profundidad en el módulo 2 del manual *Matemáticas y Estadística con R*.

e) Construir la matriz Z. Partiremos de la base MRD3 que acabamos de crear pero ahora incluiremos las variables $UNOS$, CP, CP_1, CP_2, P, P_1, A, C_1:

```
Z <- matrix (c(UNOS, CP, CP_1, CP_2, P, P_1, A, C_1),37,8)
```

f) Construir la matriz Y quedándonos únicamente con la variable endógena C.

```
Y <- matrix (c(C),37,1)
```

g) Calcular la matriz $W'Z$. Primero calcularemos la matriz transpuesta con la función `t` y después multiplicamos mediante el operador $\%*\%$.

También lo podíamos haber hecho directamente mediante el código $tWZ <- t(W)\%*\%Z$.

```
tW <- t(W)
tWZ <- tW%*%Z
```

h) Calcular la matriz $(W'Z)^{-1}$. Para calcular la matriz inversa utilizaremos la función `solve`.

Mucho más corto hubiera sido haciendo $ItWZ <- solve(t(W)\%*\%Z)$.

```
ItWZ <- solve(tWZ)
```

i) Calcular la matriz $(W'Y)$.

```
tWY <- t(W)%*%Y
```

j) Y finalmente obtener la estimación de los parámetros calculando la matriz: $(W'Z)^{-1}(W'Y)$. Esto es:

De esta manera nos hubiéramos ahorrado un montón de pasos: $B <- solve(t(W)\%*\%Z, t(W)\%*\%Y)$.

```
B <- ItWZ%*%tWY
```

Con lo que la estimación por mínimos cuadrados en dos etapas de nuestro modelo dinámico es:

$$\hat{C}_t = -1{,}15 + 0{,}83CP_t - 0{,}02CP_{t-1} - 0{,}58CP_{t-2} + 0{,}11P_t - 0{,}14P_{t-1} + 0{,}12A_t + 0{,}84C_{t-1}$$

1.2. Modelos de regresión multiecuacionales

Es evidente que no todos los modelos econométricos pretenden explicar el comportamiento de una única variable endógena mediante otras variables explicativas predeterminadas (exógenas o endógenas retardadas). En ocasiones, para explicar una variable es necesario considerar otras variables endógenas como explicativas, obteniendo así más de una ecuación que estimar.

Consideraremos que una variable es endógena si no está controlada por el analista y se debe explicar por el modelo. Mientras que una variable será exógena si se puede predeterminar su valor y, por lo tanto, no se explica por el modelo.

1.2.1. Hipótesis básicas y formulación general de un modelo multiecuacional

En un principio en un modelo multiecuacional se deben cumplir las siguientes hipótesis:

1) Disponer de G variables endógenas que explicar por nuestro modelo. Esto supone especificar G ecuaciones para que el modelo sea completo. Cada variable endógena vendrá explicada por k variables predeterminadas (entre exógenas y endógenas retardadas). No es obligatorio que en cada ecuación figuren todas las variables endógenas y predeterminadas pero sí que la relación existente sea lineal. Por tanto, un modelo multiecuacional podría ser:

Recordemos que en forma matricial el conjunto de variables endógenas se representa por el vector Y y el conjunto de variables predeterminadas por la matriz Z.

$$\beta_{11}Y_{1t} + \beta_{12}Y_{2t} + \ldots + \beta_{1G}Y_{Gt} + \gamma_{11}Z_{1t} + \ldots + \gamma_{1k}Z_{kt} = u_{1t},$$

$$\beta_{21}Y_{1t} + \beta_{22}Y_{2t} + \ldots + \beta_{2G}Y_{Gt} + \gamma_{21}Z_{1t} + \ldots + \gamma_{2k}Z_{kt} = u_{2t},$$

$$\ldots \qquad \ldots \qquad \ldots \qquad \ldots \quad ,$$

$$\beta_{G1}Y_{1t} + \beta_{G2}Y_{2t} + \ldots + \beta_{GG}Y_{Gt} + \gamma_{G1}Z_{1t} + \ldots + \gamma_{Gk}Z_{kt} = u_{Gt},$$

donde $t = 1, 2, \ldots, T$ observaciones. Además es de suponer que $\beta ii = 1$.

2) Aunque los términos de perturbación de cada ecuación tendrán media nula y serán esféricos, supondremos que hay términos de perturbación referidos a diferentes ecuaciones del modelo que pueden presentar autocorrelación contemporánea.

Una vez definido un modelo multiecuacional, veamos las diferentes maneras de escribirlo:

Matricialmente, esto se escribe $E[U] = 0$ y $E[UU'] = \Sigma \otimes I$, donde $\otimes$ es el producto de Kronecker que implica que cada elemento de la primera matriz se multiplica por la segunda, y U es la matriz de dimensión $TG \times TG$ formada por todas las T observaciones de cada uno de los G términos de perturbación.

1) La **forma estructural** consiste en escribir en cada una de las G ecuaciones la combinación lineal de todas las variables endógenas y predeterminadas, e igualarlas a sus términos de perturbación.

2) La **forma reducida** es cuando expresamos cada variable endógena en función de las variables predeterminadas.

3) La **forma final** es aquella en la que las variables endógenas están en función únicamente de variables exógenas y exógenas retardadas. El punto de partida para obtener esta formulación a partir de la forma estructural requeriría utilizar el operador de retardos.

Un ejemplo de las distintas maneras de escribir un modelo multiecuacional se encuentra en el manual de la asignatura de Econometría.

1.2.2. Tipología de modelos multiecuacionales

Se distinguen cuatro tipos de modelos multiecuacionales:

1) Los **modelos multiecuacionales no relacionados** son aquellos en los que un conjunto de variables predeterminadas explican el comportamiento de G variables endógenas sin que ello suponga una relación entre ellas, es decir, que no haya una que explique la otra ni tampoco exista correlación entre los términos de perturbación de las diferentes ecuaciones.

2) Los **modelos multiecuacionales aparentemente no relacionados** son aquellos en los que también un conjunto de variables predeterminadas explican el comportamiento de G variables endógenas. No obstante, en este caso, aunque no exista relación directa entre variables endógenas (una explica la otra), sí que se pone de manifiesto una relación indirecta, ya que existe correlación entre los términos de perturbación de las diferentes ecuaciones.

3) Los **modelos multiecuacionales recursivos** son aquellos en los que un conjunto de variables predeterminadas explican el comportamiento de G variables endógenas y, además, se produce una relación de causalidad unidireccional directa entre las variables endógenas. Sin embargo, no existe correlación entre los términos de perturbación de las diferentes ecuaciones.

4) Finalmente, los **modelos multiecuacionales integrados** constituyen el caso general, en el que no se da ninguno de los casos particulares que hemos mencionado antes. Habrá relaciones directas entre las variables endógenas y, además, correlación entre los términos de perturbación de las diferentes ecuaciones.

1.2.3. El problema de la identificación

Recordad que hay que estudiar la identificación para cada ecuación individualmente. Un modelo estará no identificado si hay alguna ecuación que no lo esté. Si el modelo está identificado, por el único hecho de que haya una ecuación sobreidentificada, el modelo estará sobreidentificado. Así pues, un modelo estará exactamente identificado cuando todas las ecuaciones lo estén.

Para poder estimar un modelo multiecuacional primero debemos tratar la identificación de este, ya que puede darse el caso de que no dispongamos de suficiente información para poder resolver el sistema de ecuaciones y, por tanto, valorar los parámetros del modelo. Así, nos podemos encontrar ante tres tipos de situaciones:

1) Modelos con **ecuaciones no identificadas**, es decir, cuando no tenemos suficiente información para estimar los parámetros de la forma estructural de la ecuación. Estaríamos ante un sistema de ecuaciones incompatible. Esta ecuación tendrá que reespecificarse mediante la incorporación de nueva información en esta. Lo más habitual

es incorporar restricciones lineales sobre los parámetros de las variables endógenas o predeterminadas.

2) Modelos con **ecuaciones sobreidentificadas**, esto es, en los que hay más de una combinación de valores estimados posible de los parámetros estructurales que cumplirían la relación entre variables registrada en la ecuación. Se corresponden con sistemas de ecuaciones compatibles indeterminados.

Algunos ejemplos de restricciones son $\beta_{ij} + \beta_{ih} = 0$ o $\beta_{ij} = 0$.

3) Modelos con **ecuaciones exactamente identificadas** y que, por lo tanto, a partir de las variables incluidas en el modelo, solo podemos obtener una única estimación de los parámetros estructurales. Se corresponden con sistemas de ecuaciones compatibles determinados.

No olvidéis que los modelos multiecuacionales no relacionados, los aparentemente no relacionados y los recursivos siempre están identificados. Además, la forma reducida de un modelo también está siempre identificada.

Por las características de los diferentes modelos multiecuacionales, solo deberemos centrarnos en el estudio de la identificación de los modelos de ecuaciones simultáneas integrados. Y el problema de la identificación se centra en la forma estructural. Este problema se resuelve mediante las denominadas condiciones de rango y de orden:

1) La **condición de rango** consiste en calcular el rango de una matriz ($A\phi$), de manera que este es $G - 1$, la ecuación está identificada y no lo estará en otro caso. $A = (B|\Gamma)$ es la matriz de dimensión $G \times (G + k)$, formada por todos los parámetros de la forma estructural del modelo. ϕ es la matriz de restricciones de dimensión $(G+k)\times q$, formada por tantas filas como número de variables endógenas y predeterminadas haya, y tantas columnas (q) como restricciones presente la ecuación.

2) La **condición de orden** determina el tipo de identificación. Si el número de restricciones es $G - 1$, la ecuación está exactamente identificada y si es mayor a $G - 1$,la ecuación está sobreidentificada.

No olvidéis que la condición de orden es una condición necesaria, pero no suficiente. Podemos tener una ecuación no identificada que cumpliera dicha condición.

1.2.4. La estimación de los modelos multiecuacionales

Estos modelos se estiman por tres métodos:

1) Los **métodos directos** valoran cada ecuación por separado, sin tener en cuenta que la ecuación forma parte de un modelo multiecuacional. El método más conocido es el de MCO.

2) Los **métodos de información limitada** valoran cada ecuación por separado, pero tienen en cuenta información adicional a la ecuación estimada (registrada en el resto de las ecuaciones del modelo). Es decir, consideran si una variable explicativa es endógena o exógena, y también las variables que no están incluidas en la ecuación pero que sí están presentes en el modelo. Los métodos más habituales son el de mínimos cuadrados indirectos (MCI), el de mínimos cuadrados en dos etapas (MC2E), el de variables instrumentales (VI) y el de máxima verosimilitud de información limitada (MVIL).

3) Finalmente, los **métodos de información completa**, que estiman de manera conjunta todas las ecuaciones del modelo, por lo que tienen en cuenta toda la información del modelo. Los más habituales son el de mínimos cuadrados en tres etapas (MC3E) y el de máxima verosimilitud de información completa (MVIC).

En el cuadro siguiente aparecen las propiedades asintóticas de los estimadores MCO de los diferentes modelos multiecuacionales, siempre y cuando se cumplan las hipótesis básicas del MRLM en cada ecuación:

Modelo multiecuacional	Propiedades asintóticas de los estimadores MCO
No relacionado	Consistencia y eficiencia
Aparentemente no relacionado	Consistencia y no eficiencia
Recursivo	Consistencia y no eficiencia
Integrado	Consistencia y no eficiencia

Así, los modelos multiecuacionales no relacionados no tendrán problemas. Veamos qué ocurre con el resto de los modelos:

1) En los modelos aparentemente no relacionados, los estimadores serán ineficientes porque la estimación uniecuacional no tendrá en cuenta toda la información asociada al modelo y deberemos estimar por métodos de información completa, como el **método de estimación de Zellner**, que básicamente consiste en estimar el modelo multiecuacional completo por MCG.

2) En los modelos integrados, los estimadores serán sesgados e inconsistentes a causa de la presencia de otras variables endógenas correlacionadas con el término de perturbación como variables explicativas. Además, serán ineficientes porque no consideran toda la información disponible en el modelo. En este segundo modelo, los estimadores de información limitada serán consistentes, pero ineficientes, y los de información completa, consistentes y eficientes.

Los tres métodos de información limitada más utilizados son:

Este método resulta muy útil en ecuaciones exactamente identificadas.

1) El método de los mínimos cuadrados indirectos (MCI) consiste en estimar por MCO los parámetros Π de la forma reducida del modelo y, posteriormente, obtener las estimaciones de los parámetros de la forma estructural (ByΓ), a partir de la siguiente relación: $\Pi = -B^{-1}\Gamma \Rightarrow B\Pi + \Gamma = 0$.

Este método es útil tanto en ecuaciones sobreidentificadas como en las exactamente identificadas.

2) El método de los mínimos cuadrados en dos etapas (MC2E) consiste en primero estimar la forma reducida por MCO para obtenemer una estimación de las variables endógenas ($\hat{Y}_t$): $\hat{Y}_t = \hat{\Pi} Z_t$. Después se vuelve a valorar la ecuación estructural inicial por MCO, tras haber sustituido las variables endógenas presentes como variables explicativas (Y_t) por sus valores estimados ($\hat{Y}_t$).

3) Para acabar, el método de las variables instrumentales (VI) consiste en definir una matriz W de instrumentos y aplicar la fórmula presentada en la estimación de los modelos dinámicos. Dado que se utilizan las mismas variables como instrumentos de las variables predeterminadas de una ecuación, podemos encontrarnos con:

a) En el caso de una ecuación exactamente identificada, si como instrumentos de las variables endógenas explicativas utilizamos las variables predeterminadas omitidas en la ecuación, pero presentes en otras ecuaciones del modelo, los resultados equivaldrán a los de los MCI.

b) Por otra parte, en el caso de ecuaciones sobreidentificadas, si como instrumentos utilizamos las obtenidas en la primera etapa de los MC2E, los estimadores VI serán equivalentes a los MC2E.

1.2.5. Interpretación de los parámetros del modelo

Una vez estimado el modelo, debemos interpretar el significado de los parámetros. En este sentido, los parámetros de la forma estructural únicamente registran los efectos directos entre las variables explicativas y la variable endógena. Con el fin de registrar los efectos directos e indirectos contemporáneos (variaciones en la variable endógena ante variaciones en la variable exógena referidas todas ellas el mismo momento del tiempo), hay que calcular los parámetros de la forma reducida. Para acabar, para registrar todos los efectos contemporáneos, así como aquellos que hacen referencia a cualquier momento del tiempo, deberemos analizar los parámetros de la forma final.

2. Modelo Lineal Generalizado

2.1. Motivación

Hasta ahora, hemos asumido que la relación entre la variable dependiente y los regresores seguía la siguiente formulación matricial:

$$Y = X\beta + e$$

$$E(Y) = X\beta$$

Esta formulación puede ser limitada en el caso de que la distribución de los errores no siga una distribución normal, y esto sucede, especialmente, cuando la variable dependiente es de naturaleza cualitativa (es decir, si se refieren a atributos o categorías, y no a valores numéricos), o bien si es una variable de recuento, que son las variables cuyas observaciones son números enteros positivos.

Analíticamente, el **Modelo Lineal Generalizado** (GLM en sus siglas en inglés) es una generalización flexible del modelo de regresión lineal ordinario, que permite la inclusión de variables dependientes que generan distribuciones del error distintas de una distribución normal. El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable dependiente a través de una *función de enlace*, y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho. Esto es, la nueva formulación tiene la siguiente forma:

Acrónimos en varios idiomas

El Modelo Lineal Generalizado recibe el acrónimo **MLG** en castellano, pero es más común verlo escrito como **GLM** (Generalized Linear Model), que son sus siglas en inglés.

$$E(Y) = \mu = g^{-1}(X\beta)$$

Donde tenemos los siguientes componentes:

1) Función de distribución. Se trata de una función de distribución perteneciente a la familia exponencial, a la que pertenecen muchas de las distribuciones más comunes, como la normal, exponencial, gamma, chi-cuadrado, beta, Dirichlet, Bernoulli, categórica y Poisson. Esto permite adaptar el modelo de regresión a la distribución específica de la variable dependiente y, lógicamente, el error del modelo.

2) Media de la función de distribución (μ). Coincide con el valor esperado de Y, de manera que $E(Y) = \mu$.

3) Predictor lineal (η). Es la magnitud que incorpora la información acerca de las variables explicativas en el modelo. Así, η se expresa como una combinación lineal de los parámetros desconocidos β, de manera que $\eta = X\beta$.

4) Función de enlace (g). Esta función define la relación entre el predictor lineal (η) y la media de la función de distribución (μ), de manera que $\eta = g(\mu)$.

5) Función de la media. Se basa en invertir la función de enlace g, de manera que la media sea el predictor lineal de la función de enlace, esto es, $\mu = g^{-1}(\eta)$.

Así, el modelo de regresión lineal que hemos visto hasta ahora se puede considerar como un caso particular de GLM, en el que la función de distribución es la Normal. Esto es:

$$E(Y) = \mu = X\beta$$

$$\eta = X\beta = \mu$$

Para valorar un GLM no sirve la técnica de mínimos cuadrados ordinarios, sino que necesarias técnicas de estimación más avanzadas y complejas, como el método de máxima verosimilitud.

2.2. Modelos Logit, Probit y Poisson

Dentro de la multitud de combinaciones que ofrece la familia exponencial, en esta sección repasaremos los dos modelos más utilizados para el caso de variables dependientes cualitativas (logit y probit), y el modelo básico para analizar datos de recuento (Poisson).

Las variables cualitativas se caracterizan por que sus valores no son magnitudes, sino categorías. En el caso más simple, las variables cualitativas dicotómicas solo toman dos valores (sí o no, negro o blanco, hombre o mujer, etc.). Para estudiar estas variables, y explicar su comportamiento en función de otras variables (esto es, estimar un modelo de regresión) es muy útil codificarlas en forma de ceros y unos. Sin embargo, es fundamental tener muy claro que estos valores 0 y 1 no son magnitudes, sino que representan dos categorías diferentes.

Por una parte, el **modelo de regresión logit** se basa en la *función logit*, que es la inversa de la función logística. De esta manera, el *logit* de un número p entre 0 y 1 se define mediante la siguiente fórmula:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

Razón de probabilidad

Si p es una probabilidad, entonces $p/(1-p)$ es la correspondiente razón de probabilidad (*odds* en inglés), y el logit de la probabilidad es el logaritmo de los odds.

Esta función sirve de base para el *modelo de regresión logit*, que tiene la siguiente expresión:

$$E(Y) = \mu = g^{-1}(X\beta)$$

$$\eta = X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$$

$$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)} = \frac{1}{1+\exp(-X\beta)}$$

Modelo de regresión logit

El modelo logit fue introducido por Joseph Berkson en 1944.

Es fundamental tener en cuenta que la interpretación de los coeficientes estimados de un modelo logit o probit es probabilística, y difiere de la interpretación que se le da a un modelo de regresión lineal. En el ejemplo aplicado que sigue a continuación veremos cómo se interpretan los coeficientes.

El **modelo de regresión probit** es muy similar al modelo anterior, y la diferencia es que la distribución de referencia es la Normal, en lugar de la Logística. En estadística, la función probit es la función cuantil, es decir, la función de distribución acumulativa inversa (CDF, en sus siglas en inglés), asociado con la distribución normal estándar. De esta manera, en el modelo de regresión probit la función de enlace g adquiere la siguiente forma:

Modelo de regresión probit

La idea de probit fue publicado en 1934 por Chester Ittner Bliss en un artículo sobre la forma de tratar los datos sobre el porcentaje de una plaga eliminado por un pesticida.

$$g = \Phi^{-1}(p)$$

Siendo Φ la función de distribución acumulativa de la distribución Normal:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

El nombre *probit* surge de juntar las palabras *probability unit*.

Ambos modelos se suelen estimar usando técnicas de máxima verosimilitud.

Por último, el **Modelo de regresión de Poisson** se emplea cuando la variable dependiente es de recuento, es decir, es el recuento de ocurrencias de un evento en un espacio de tiempo determinado, y toma valores enteros y positivos. En este caso, la distribución de referencia es la distribución de Poisson, y el modelo adquiere la siguiente forma:

La distribución de Poisson

Esta distribución le debe el nombre a Siméon Denis Poisson (1781 - 1840), matemático y físico francés, que la dio a conocer en 1838 en su trabajo *Recherches sur la probabilité des jugements en matières criminelles et matière civile*.

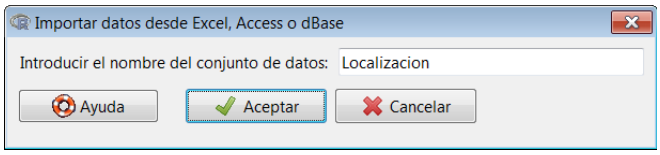
$$E(Y) = \mu = g^{-1}(X\beta)$$

$$\eta = X\beta = \ln(\mu)$$

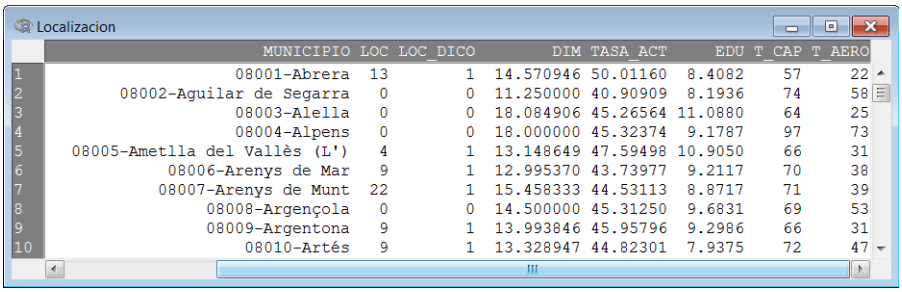
$$\mu = \exp(X\beta) = \exp(\eta)$$

2.3. Aplicación empírica

El objetivo de esta sección es presentar una aplicación de los modelos hasta ahora vistos con R-Commander. El estudio hace referencia a la localización de nuevas empresas en los municipios de Cataluña en el período 2001-2004. Como hemos venido haciendo, el primer paso es importar los datos de una hoja de cálculo y crear un conjunto de estos, al que denominaremos *Localizacion*:



Visualizando los datos, vemos las variables que componen el conjunto de estos:


Localizacion

	MUNICIPIO	LOC	LOC_DICO	DIM	TASA_ACT	EDU	T_CAP	T_AERO
1	08001-Abrera	13	1	14.570946	50.01160	8.4082	57	22
2	08002-Aguilar de Segarra	0	0	11.250000	40.90909	8.1936	74	58
3	08003-Alella	0	0	18.084906	45.26564	11.0880	64	25
4	08004-Alpens	0	0	18.000000	45.32374	9.1787	97	73
5	08005-Ametlla del Vallès (L')	4	1	13.148649	47.59498	10.9050	66	31
6	08006-Arenys de Mar	9	1	12.995370	43.73977	9.2117	70	38
7	08007-Arenys de Munt	22	1	15.458333	44.53113	8.8717	71	39
8	08008-Argençola	0	0	14.500000	45.31250	9.6831	69	53
9	08009-Argentona	9	1	13.993846	45.95796	9.2986	66	31
10	08010-Artés	9	1	13.328947	44.82301	7.9375	72	47

La descripción de las variables es la siguiente:

MUNICIPIO: nombre del municipio y su código postal.

LOC: número de empresas que se han localizado en el municipio.

LOC_DICO: variable dicotómica que toma un valor de 1 si se han localizado nuevas empresas ($LOC > 0$) y el valor 0 si no se ha localizado ninguna empresa ($LOC = 0$).

DIM: dimensión media de las empresas del municipio, expresada en número de trabajadores.

TASA_ACT: tasa de actividad del municipio, expresada como el cociente del número de trabajadores y de la población total municipal.

EDU: años medios de educación de la población municipal.

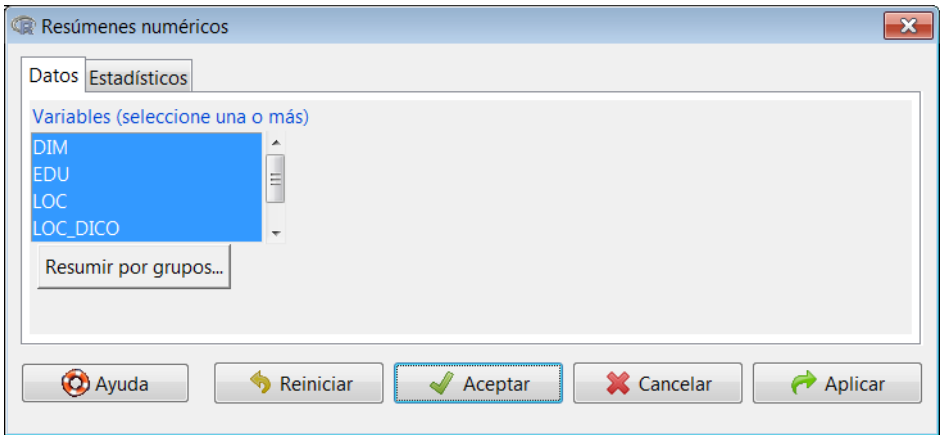
T_CAP: tiempo medio de transporte a la capital más cercana, en minutos.

T_AERO: tiempo medio de transporte al aeropuerto, en minutos.

Antes de realizar el análisis empírico, siempre es útil calcular los estadísticos básicos de las variables, lo que se realiza accediendo a la siguiente ruta del menú desplegable:

Estadísticos / Resúmenes / Resúmenes numéricos

Seleccionamos todas las variables en el cuadro de diálogo resultante:



El resultado en la consola es el siguiente:

```
> numSummary(Localizacion[,c("DIM", "EDU", "LOC", "LOC_DICO",
+ "T_AERO", "T_CAP", "TASA_ACT")], statistics=c("mean", "sd",
+ "quantiles"), quantiles=c(0,.25,.5,.75,1))

            mean      sd     0%    25%    50%     75%    100%    n
DIM        15.20    8.54   0.00  10.57  13.28   17.79  120.00  941
EDU         8.49    1.01   4.23   7.82   8.42    9.12   11.99  941
LOC         3.55   12.42   0.00   0.00   0.00    2.00  201.00  941
LOC_DICO    0.44    0.49   0.00   0.00   0.00    1.00    1.00  941
T_AERO     49.11   33.00   0.00  27.00  41.00   63.00  190.00  941
T_CAP      87.41   23.30  56.00  70.00  82.00  101.00  190.00  941
TASA_ACT   43.76    4.76  23.46  40.67  44.13   46.99   58.24  941
```

Vemos cómo el valor medio de la variable DIM es de unos 15 trabajadores por empresa, y que se han localizado empresas en el 44 % de los municipios (ya que la media de *LOC_DICO* es 0,44).

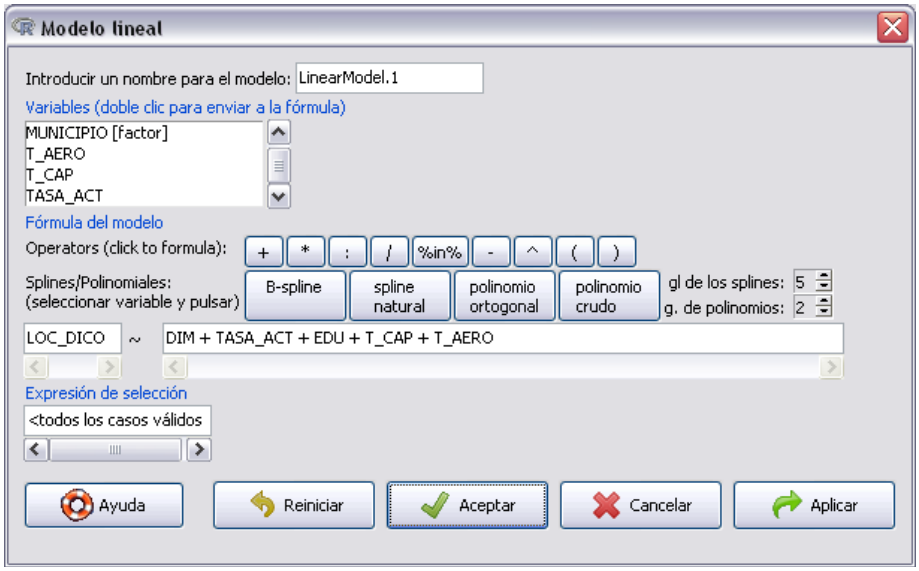
En el primer modelo que vamos a analizar, la variable dependiente será la variable dicotómica *LOC_DICO*, esto es, vamos a estudiar qué efecto tienen las variables explicativas sobre la localización de nuevas empresas en un municipio. Analíticamente, tomamos la siguiente forma funcional:

$$LOC\_DICO = f(DIM,\ TASA\_ACT,\ EDU,\ T\_CAP,\ T\_AERO)$$

El primer paso será realizar una estimación por mínimos cuadrados ordinarios (OLS en sus siglas en inglés). Para realizarla, acudimos a la siguiente ruta del menú desplegable:

Estadísticos / Ajuste de modelos / Modelo lineal

En el cuadro de diálogo resultante, introducimos la relación entre las variables, como sigue:



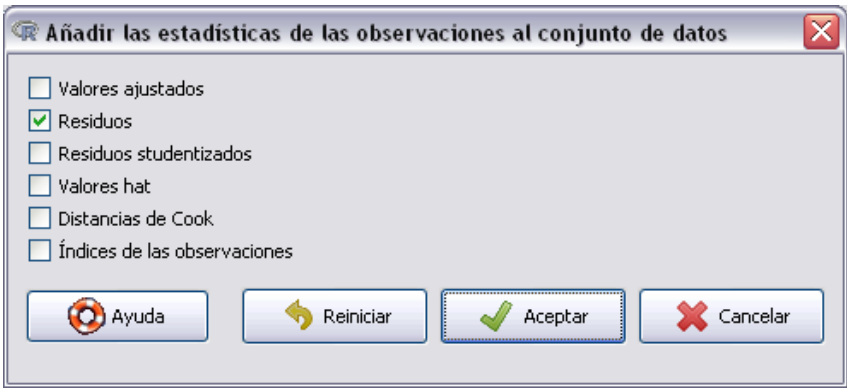
El resultado es el siguiente:

```
> LinearModel.1 <- lm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP +
+ T_AERO, data=Localizacion)

> summary(LinearModel.1)

Call:
lm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    data = Localizacion)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04317 -0.38490 -0.05569  0.38643  1.10766

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept)  0.6640403  0.1743117   3.809 0.000148 ***
DIM         -0.0119723  0.0017098  -7.002 4.81e-12 ***
TASA_ACT     0.0236518  0.0033335   7.095 2.55e-12 ***
EDU         -0.0471142  0.0156403  -3.012 0.002662 **
T_CAP       -0.0084925  0.0009044  -9.391  < 2e-16 ***
T_AERO       0.0014869  0.0006411   2.319 0.020589 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4417 on 935 degrees of freedom
Multiple R-squared:  0.216, Adjusted R-squared:  0.2118
F-statistic: 51.52 on 5 and 935 DF,  p-value: < 2.2e-16
```

Como observamos, aunque el ajuste del modelo sea más bien pobre, todos los coeficientes de las variables explicativas son significativos. Ahora bien, ¿qué validez tiene esta estimación, dada la naturaleza dicotómica de la variable dependiente? Por una parte, en una estimación por OLS las predicciones del modelo no estarán necesariamente entre cero y uno. Además, los errores se distribuyen como una distribución bimodal, y no como una Normal. Para ver esto, primero extraeremos los residuos de la estimación que acabamos de hacer, y los añadiremos a las observaciones al conjunto de datos. Acudiendo a la siguiente ruta:

Modelos / Añadir las estadísticas de las observaciones al conjunto de datos

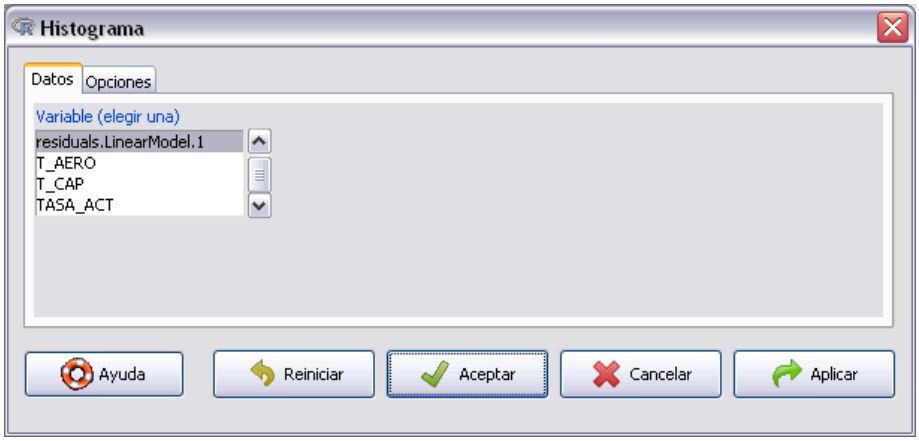
Nos aparece el cuadro de diálogo donde hemos de seleccionar qué magnitudes derivadas de la estimación deseamos almacenar como variables:



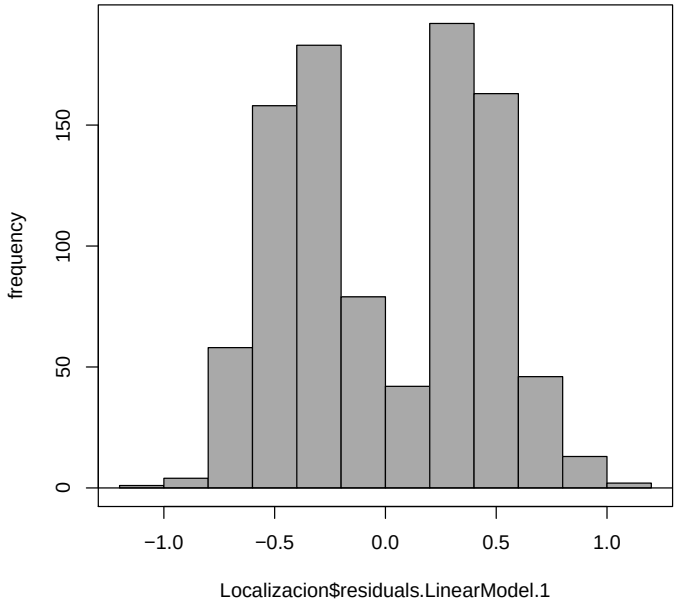
Una vez seleccionados los residuos, veremos su histograma accediendo a la siguiente ruta:

Gráficas / Histograma

Seleccionamos los residuos entre todas las variables, como sigue:



Y obtenemos el histograma de los residuos:



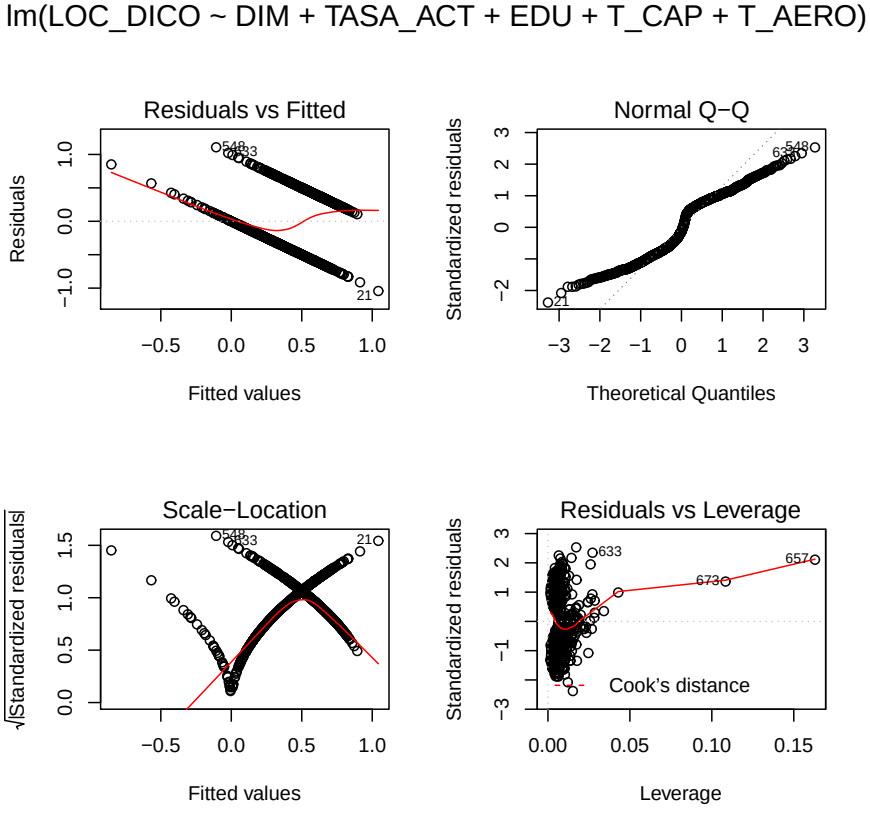
Distribución bimodal

En estadística, una distribución bimodal es una distribución de probabilidad continua con dos modas diferentes, que aparecen como dos picos distintos (máximos locales) en la función de densidad de probabilidad.

Como vemos, la distribución es bimodal, y no tiene la forma característica de una distribución Normal, lo que tiene implicaciones importantes a la hora de hacer inferencia. Podemos, alternativamente, comprobar el comportamiento particular de los residuos si obtenemos el gráfico compuesto del diagnóstico del modelo:

Modelos / Gráficas básicas de diagnóstico

Aquí podemos ver el resultado:



Sin embargo, el principal problema de la estimación MCO es que presupone que el efecto marginal de un aumento de una variable explicativa sobre la variable dependiente es el coeficiente, esto es, para una muestra de $i = 1, \ldots, n$ observaciones y k variables explicativas, el efecto marginal de un aumento del regresor x_j sobre la variable dependiente es:

$$\frac{\partial y_i}{\partial x_{ji}} = \beta_j$$

Este es un supuesto muy restrictivo, ya que la variable dependiente en un modelo binario está acotada entre 0 y 1, y se interpreta probabilísticamente. Dicho de otra manera, el incremento del efecto de un aumento de x_j sobre el incremento de y no puede ser el mismo para valores altos y bajos de x_j. Además, la relación real no es lineal, ya que es de esperar un efecto muy bajo de un cambio de x_j sobre y cuando x_j es, o bien muy bajo, o muy alto.

El *modelo de regresión logit* soluciona este problema, ya que el modelo trabaja explícitamente en términos de probabilidades. Esto es, para una muestra de $i = 1, \ldots, n$ observaciones y k variables explicativas, tenemos que:

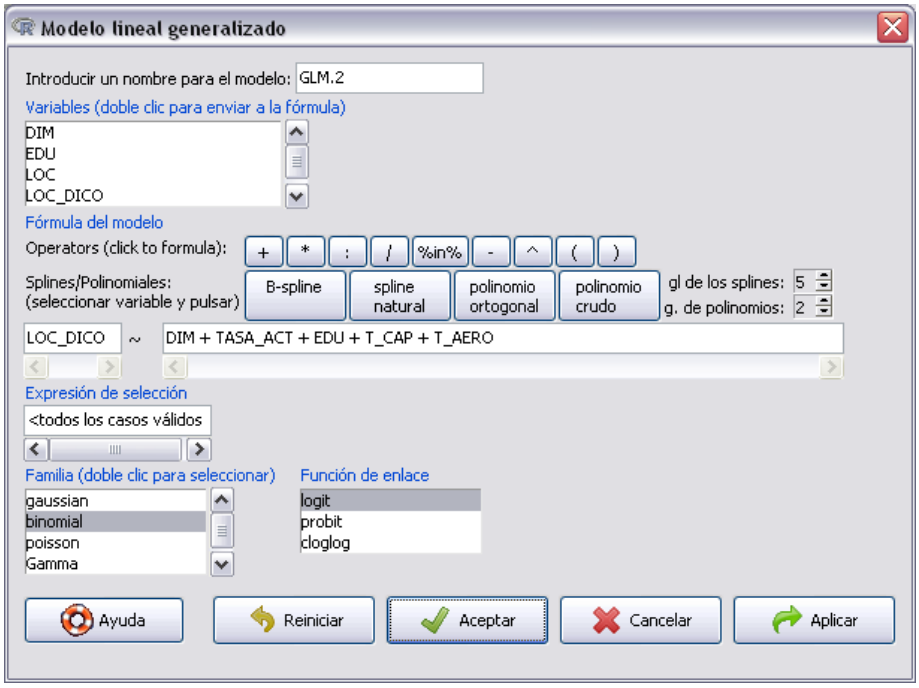
$$\text{logit}(P_i) = log\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Donde P_i es la probabilidad de ocurrencia del evento (en nuestro caso, la localización de empresas) y $\left(\frac{P_i}{1-P_i}\right)$ es la razón de probabilidad o el *odds* en su terminología inglesa, y se interpreta de la siguiente manera: si la probabilidad de que haya localizaciones empresariales es de $p = 0,75$, la probabilidad de que no haya es de $1 - p = 0,25$, de manera que la razón de probabilidad será $0,75/0,25 = 3$, esto es, la probabilidad de que haya localizaciones es de 3 a 1 a favor de que sí las haya.

Para estimar este modelo con R-Commander, acudimos a la siguiente ruta del menú desplegable:

Estadísticos / Ajuste de modelos / Modelo lineal generalizado

En el cuadro de diálogo resultante, además de introducir la variable dependiente y las variables independientes, introduciremos la familia (*binomial*) y la función de enlace (*logit*):



El resultado que obtenemos es el siguiente:

```
> GLM.2 <- glm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO
    , family=binomial(logit), data=Localizacion)

> summary(GLM.2)

Call:
glm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = binomial(logit), data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5610  -0.9145  -0.3556   0.9402   2.8054

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  1.866726   0.970452   1.924  0.05441 .
DIM         -0.095450   0.013737  -6.949 3.69e-12 ***
TASA_ACT     0.120666   0.018722   6.445 1.16e-10 ***
EDU         -0.275304   0.083802  -3.285  0.00102 **
T_CAP       -0.045395   0.005025  -9.034  < 2e-16 ***
T_AERO       0.004852   0.003555   1.365  0.17232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1294.1  on 940  degrees of freedom
Residual deviance: 1044.3  on 935  degrees of freedom
AIC: 1056.3

Number of Fisher Scoring iterations: 5
```

¿Cómo hemos de interpretar el resultado de esta estimación? A diferencia del MRLM, en un modelo binario como el logit la influencia que tienen las explicativas sobre la probabilidad de elegir la opción dada por $y_i = 1$ no depende simplemente del valor los coeficientes, sino también del valor que toman las variables explicativas, es decir, $\frac{\partial y_i}{\partial x_{ji}} = f(\beta_j, x_{ji}$. Aunque la interpretación sea compleja, podemos partir de la base de que, al incrementar una unidad la variable x_j, la probabilidad P_i pasa a ser $P_i^{'}$, de manera que:

$$\text{logit}(P_i^{'}) = \beta_j + \text{logit}(P_i)$$

Si reescribimos esta expresión, llegamos a una fórmula que nos será muy útil a la hora de interpretar el resultado:

$$\frac{\frac{P_i^{'}}{1-P_i^{'}}}{\frac{P_i}{1-P_i}} = \exp(\beta_j)$$

Esto es, $\exp(\beta_j)$ se denomina el **odds ratio**, e indica el cambio relativo que experimenta el cociente de probabilidades cuando la variable x_j aumenta una unidad. En nuestro ejemplo, el coeficiente estimado de la variable *tiempo de transporte a las capitales (T_CAP)* es de $\hat{\beta}_4 = -0,045$, esto es, $e^{-0,045} = 0,955$, lo cual indica que el aumento de 1 km en la distancia a las capitales provoca una pequeña disminución del cociente de probabilidades. La variable *años de educación (EDU)* tiene un coeficiente estimado negativo de mayor magnitud ($\hat{\beta}_3 = -0,27$), de manera que un aumento marginal de esta variable (un año más de escolarización) causa un descenso en el cociente de probabilidades de $e^{-0,27} = 0,76$, es decir, se reduce aproximadamente en una cuarta parte.

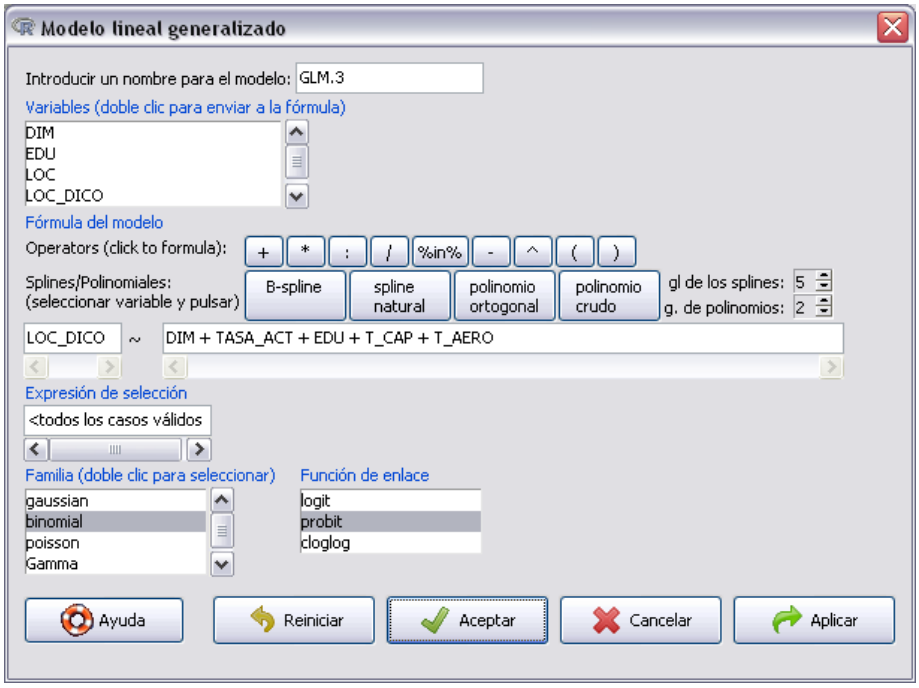
Odds ratio

No existe unanimidad sobre la traducción de este concepto al castellano, por eso aquí lo dejamos en su expresión inglesa. Algunas propuestas de traducción son razón de momios, razón relativo, razón de oportunidades, razón de productos cruzados, razón de desigualdades, razón de disparidad, razón de exceso o, simplemente, razón de odds.

El *modelo de regresión probit*, como se ha comentado antes, es muy similar al modelo anterior, y se diferencia básicamente en que el modelo probit se basa en una distribución normal acumulada. Para estimar este modelo con R-Commander, se procede de una manera muy similar al caso anterior:

Estadísticos / Ajuste de modelos / Modelo lineal generalizado

En el cuadro de diálogo resultante, además de introducir la variable dependiente y las variables independientes, introduciremos la familia (*binomial*) y la función de enlace (*probit*):



El resultado es el siguiente:

```
> GLM.3 <- glm(LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO
    ,
+ family=binomial(probit), data=Localizacion)

> summary(GLM.3)

Call:
glm(formula = LOC_DICO ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = binomial(probit), data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6358  -0.9327  -0.3466   0.9499   2.9670

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  1.030755   0.571132   1.805  0.07111 .
DIM         -0.053994   0.007728  -6.987 2.81e-12 ***
TASA_ACT     0.070029   0.010875   6.440 1.20e-10 ***
EDU         -0.154390   0.049434  -3.123  0.00179 **
T_CAP       -0.026678   0.002923  -9.128  < 2e-16 ***
T_AERO       0.003247   0.002085   1.557  0.11944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1294.1  on 940  degrees of freedom
Residual deviance: 1048.4  on 935  degrees of freedom
AIC: 1060.4

Number of Fisher Scoring iterations: 5
```

Como vemos, el resultado de la estimación, tanto los coeficientes individuales como el ajuste, es muy similar a la estimación anterior. El estadístico AIC (*Akaike Information Criterion*), muy utilizado en la estimación por máxima verosimilitud, es ligeramente menor para la estimación *logit*, de manera que nos quedaríamos con la estimación del modelo anterior.

Akaike Information Criterion (AIC)

Es una medida de la *calidad relativa* de un modelo econométrico respecto a otros, y para un conjunto específico de datos. Es decir, es una medida que sirve para comparar varios modelos, siendo un *trade-off* entre la bondad de ajuste del modelo y la complejidad de este (medida como la cantidad de parámetros que valorar).

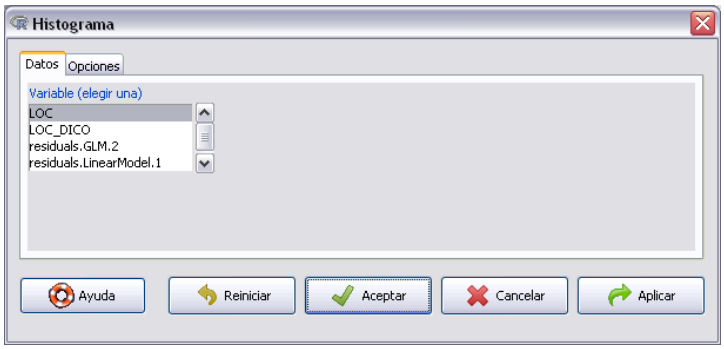
Por último, vamos a ver un ejemplo de la estimación de un *modelo de regresión de Poisson*. Para esto, vamos a estimar un modelo de regresión como el anterior, pero ahora tomando como variable dependiente el *número de empresas localizadas (LOC)*:

$$LOC = f(DIM,\ TASA\_ACT,\ EDU,\ T\_CAP,\ T\_AERO)$$

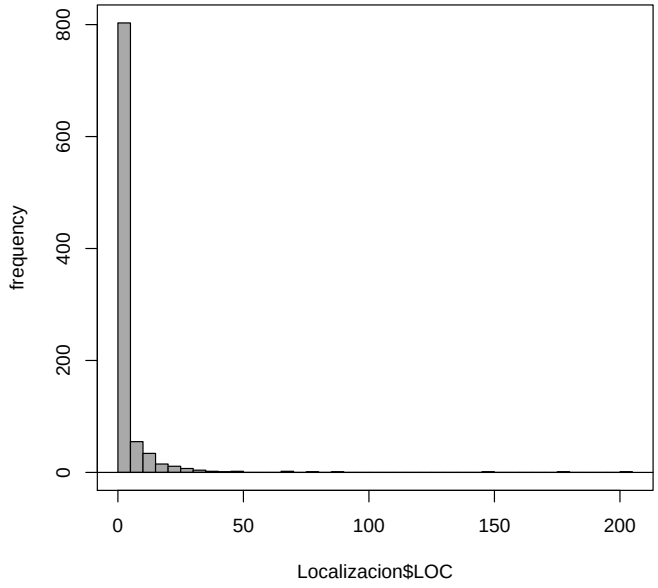
De esta manera, ahora la variable dependiente es cuantitativa, pero con una distribución peculiar. Veamos su histograma, acudiendo a la siguiente opción del menú desplegable:

Gráficas / Histograma

Seleccionamos la variable *LOC* en el cuadro de diálogo:



Obtenemos el gráfico siguiente:



Como vemos, la distribución de la variable dista mucho de ser Normal. La mayoría de las observaciones están situadas en los primeros valores $(0, 1, 2, \ldots)$, y el número de recuentos mayores de 40 es muy, muy pequeño. De hecho, la media de la variable LOC es de $3,55$. Esta distribución, pues, se puede asociar más bien a una Poisson. Este modelo tiene la siguiente formulación:

$$E(Y) = \mu = g^{-1}(X\beta)$$

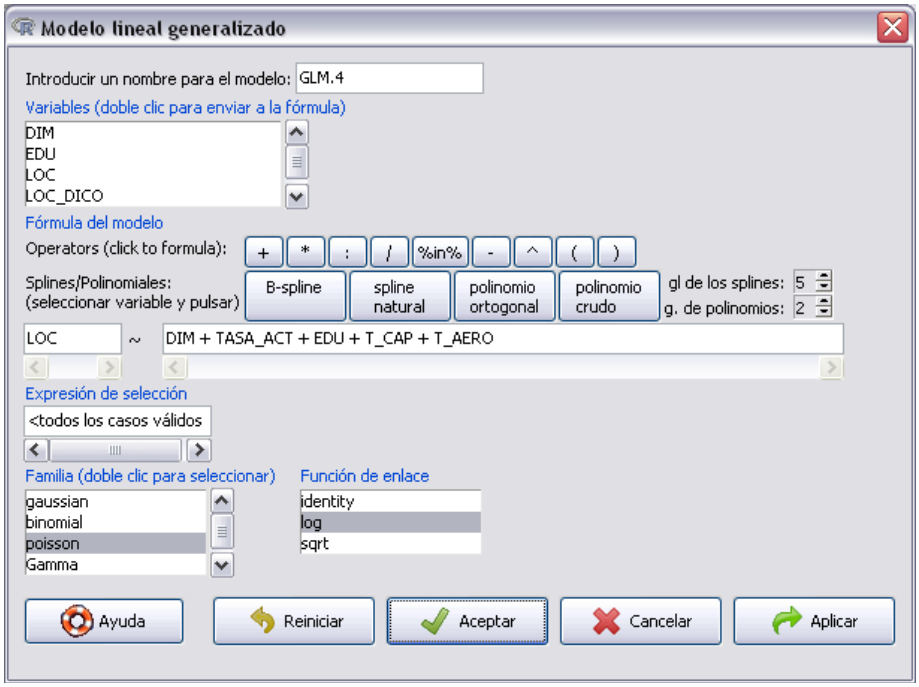
$$\eta = X\beta = \ln(\mu)$$

$$\mu = \exp(X\beta)$$

Para estimar el modelo, acudimos a la siguiente ruta del menú desplegable:

Estadísticos / Ajuste de modelos / Modelo lineal generalizado

En el cuadro de diálogo, introducimos las variables del modelo, y la familia Poisson:



El resultado es como sigue:

```
> GLM.4 <- glm(LOC ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family=poisson(log), data=Localizacion)

> summary(GLM.4)

Call:
glm(formula = LOC ~ DIM + TASA_ACT + EDU + T_CAP + T_AERO,
    family = poisson(log),
```

```
    data = Localizacion)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1733  -1.8274  -0.8423   0.0227  28.0688

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  6.549514   0.268266  24.414   <2e-16 ***
DIM         -0.102111   0.003981 -25.649   <2e-16 ***
TASA_ACT     0.061121   0.004840  12.629   <2e-16 ***
EDU         -0.175997   0.020962  -8.396   <2e-16 ***
T_CAP       -0.062467   0.001665 -37.526   <2e-16 ***
T_AERO      -0.011990   0.001207  -9.931   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11596.2  on 940  degrees of freedom
Residual deviance:  6948.9  on 935  degrees of freedom
AIC: 8292.3

Number of Fisher Scoring iterations: 7
```

La interpretación de los parámetros estimados depende de qué tipo de variable explicativa tenga asociada. Así, para el coeficiente estimado $\hat{\beta}_k$, podemos tener tres casos:

1) *Variable dummy:* la media condicional de la variable dependiente será $\exp(\hat{\beta}_k)$ veces mayor si la variable independiente x_k toma el valor 1 en lugar de 0.

2) *Variable continua:* el coeficiente asociado $\hat{\beta}_k$ es una semielasticidad, de manera que $100 \cdot \hat{\beta}_k$ es el cambio porcentual en la media de la variable dependiente cuando la variable explicativa aumenta en una unidad.

3) *Variable en logaritmos:* el coeficiente asociado $\hat{\beta}_k$ es una elasticidad, siendo el cambio porcentual en la media de la variable dependiente cuando la variable explicativa aumenta en un 1 %.

Variables dummy

Se trata de variables explicativas cualitativas dicotómicas, que adquieren el valor 0 o 1 para indicar la ausencia o la presencia de algún efecto categórico. También se conocen como variables indicador, variables de diseño o indicadores booleanos.

3. Modelos con Datos de Panel

3.1. Introducción

El objetivo de este capítulo es proporcionar una completa introducción a las técnicas de datos de panel. El término *panel de datos* hace referencia a un conjunto de datos con observaciones temporales para los mismos individuos, lo que permite al investigador seguir a un mismo individuo durante el tiempo. Debido a la amplia disponibilidad de este tipo de datos, estos son utilizados en diferentes campos, como la economía laboral, el análisis de la productividad, la demografía y las finanzas, por ejemplo.

Sobre la terminología

Mientras que en estadística y econometría se denomina **datos de panel** a un estudio con datos de estos individuos a través del tiempo, en bioestadística este conjunto de datos y técnicas reciben el nombre de **datos y estudios longitudinales**.

Las principales características del análisis de datos de panel son las siguientes:

1) *Control de la heterogeneidad individual:* en economía es fundamental el estudio de agentes económicos heterogéneos (países, personas, empresas, etc.). Los estudios solo con datos *cross-section* o con series temporales pueden dar resultados sesgados.

2) *Aumento de grados de libertad:* los datos de panel implican muestras más grandes, lo que resulta en una mayor eficiencia en la estimación.

3) *Reducción de multicolinealidad:* la unión de las dimensiones individual y temporal supone la inclusión de una gran cantidad de variabilidad y de la información. De hecho, la variación de los datos se puede descomponer en la variación entre diferentes individuos (*between*) y la variación dentro de cada individuo (*within*).

4) *Estudio de fenómenos dinámicos:* los estudios de corte transversal producen una imagen estática del tema analizado, mientras que los datos de panel pueden descubrir diferentes efectos temporales y analizar la evolución dinámica de la imagen.

5) *Identificación de efectos determinados:* hay ciertos efectos que solo pueden ser analizados mediante estudios con datos de panel.

6) *Más posibilidades de modelización:* por ejemplo, en los modelos de retardos distribuidos, al haber más variabilidad hay que imponer menos restricciones al modelo, lo que mejora el análisis.

7) *Eliminación del sesgo de agregación:* los datos que agregan a individuos, por lo general, presentan un sesgo importante.

Sin embargo, obtener datos de panel a veces comporta dificultades y limitaciones. Un problema es el diseño y la recogida de datos: falta de respuesta, *missing values*, errores de medición, etc. Otro problema es el del sesgo de selección: sesgo muestral, autoselección, pérdida de individuos, datos censurados y truncados, etc. Además, cabe destacar que las dimensiones del panel son relevantes para las propiedades asintóticas

de los estimadores. Siendo N el número de individuos y T el número de observaciones muestrales, se pueden distinguir dos casos generales:

1) $N > T$: se corresponde con datos microeconómicos o amplias muestras de individuos.

2) $T > N$: suele ser el caso de datos agregados, como regiones o países, observados durante un largo período de tiempo.

Analíticamente, una muestra de datos de panel puede ser descrita de la siguiente manera:

$$\{(y_{it}, x_{it}) : \ i = 1, \ldots, N; \ \ t = 1, \ldots, T\}.$$

Las observaciones están aparejadas (y_{it}, x_{it}), el subíndice i hace referencia a los individuos $(i = 1, \ldots, N)$ y t hace referencia al período temporal $(t = 1, \ldots, T)$. Así pues, la dimensión total de la muestra es $N \times T$. y_{it} es la variable dependiente o explicada, mientras que x_{it} es el vector de variables explicativas o regresores. Este incluye k variables: $x_{1,it}, \ldots, x_{k,it}$.

Al igual que sucede con cualquier modelo econométrico, existen dos pasos fundamentales en el análisis de un modelo con datos de panel:

1) *Especificación del modelo:* el hecho de asumir si se incluyen efectos individuales o temporales, o ambos, determinará el resultado de la estimación. El conjunto de regresores elegido, además de la forma funcional en que se incluyen (lineal, aditiva, etc.), también marcará el resultado.

2) *Estimación del modelo:* una vez especificado el modelo, hay varios métodos de estimación disponibles, cada uno con una serie de propiedades y características. El conocimiento previo del fenómeno analizado y la disponibilidad de contrastes y análisis de la varianza ayudarán a hacer una elección correcta.

Dependiendo de las hipótesis y asunciones que hagamos del modelo a estudiar, existen tres posibles especificaciones de partida (siendo u_{it} en todos los casos el término de perturbación):

1) *Modelo con efectos individuales:*

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it}.$$

2) *Modelo con efectos temporales:*

$$y_{it} = x'_{it}\beta + \eta_t + u_{it}.$$

3) *Modelo con efectos individuales y temporales:*

$$y_{it} = x'_{it}\beta + \alpha_i + \eta_t + u_{it}.$$

Un estudio descriptivo previo de las variables ayudará a realizar una elección de uno de estas tres especificaciones.

3.2. Estimación de un modelo de datos de panel

3.2.1. Mínimos Cuadrados Ordinarios (MCO)

En nomenclatura inglesa, este estimador también se conoce como *pooled OLS*. La hipótesis básica de este método de estimación es suponer que los efectos individuales son comunes entre los individuos ($\alpha_i = \alpha$). De esta manera, el modelo que estimar tiene la forma:

$$y_{it} = +\alpha + x'_{it}\beta + u_{it}.$$

Acrónimos en varios idiomas

La estimación por Mínimos Cuadrados Ordinarios recibe el acrónimo **MCO** en castellano, pero es más común verlo escrito como **OLS** (Ordinary Least Squares), que son sus siglas en inglés. En este manual usaremos ambos acrónimos indistintamente

Este método agrega las dos dimensiones i y t, sin tener en cuenta las posibles particularidades de cada individuo. Las propiedades de este estimador dependerán de la posible existencia de efectos individuales α_i correlacionados con los regresores x_i:

1) $E(x_{it}u_i) \neq 0$: en este caso, OLS será **sesgado** e **inconsistente**. La causa es que, ya que α_i no es modelizado, pasa a formar parte del término de perturbación u_i, con lo que **no** se cumple la condición de ortogonalidad $E(x_{it}u_{it}) = 0$. Esto sucederá si, por ejemplo, α_i es una variable omitida.

2) $E(x_{it}u_i) = 0$: el estimador OLS será **consistente**, pero **no será eficiente** a no ser que u_i sea esférico, es decir, homoscedástico y no correlacionado.

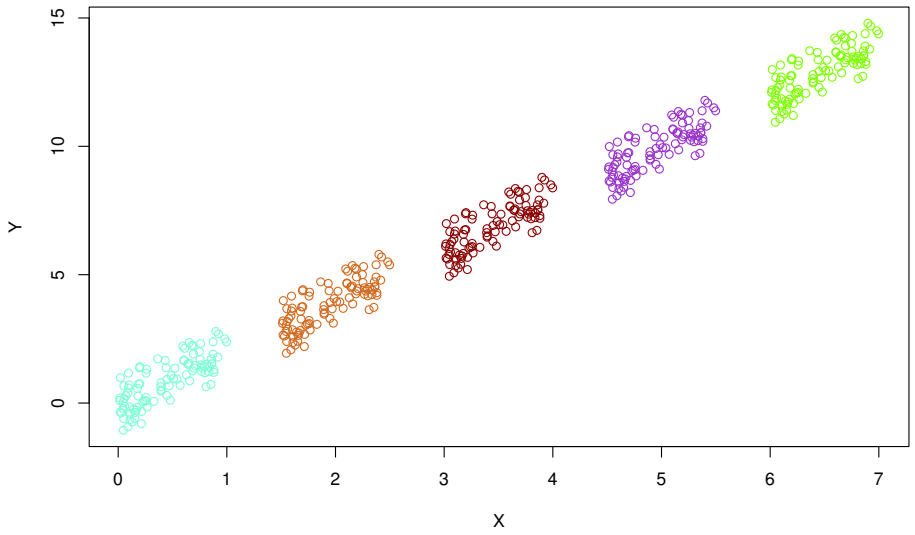
3) En general, u_i no es esférico, la estimación se puede mejorar mediante Mínimos Cuadrados Generalizados (GLS).

Veamos un ejemplo ficticio. Supongamos el siguiente modelo:

$$C_{it} = \alpha + \beta R_{it} + u_{it}$$

Donde C es el consumo y R es la renta disponible. La muestra está disponible para $N = 5$ individuos y $T = 100$ períodos temporales.

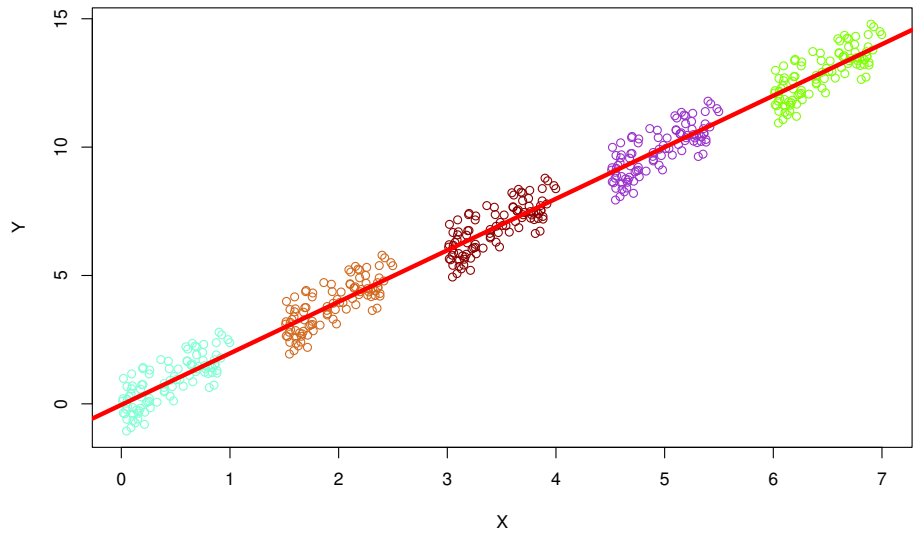
Una primera visualización de la nube de puntos permite ver cómo los cinco individuos (cada uno mostrado con un color diferente) muestran diferentes niveles de renta y consumo.



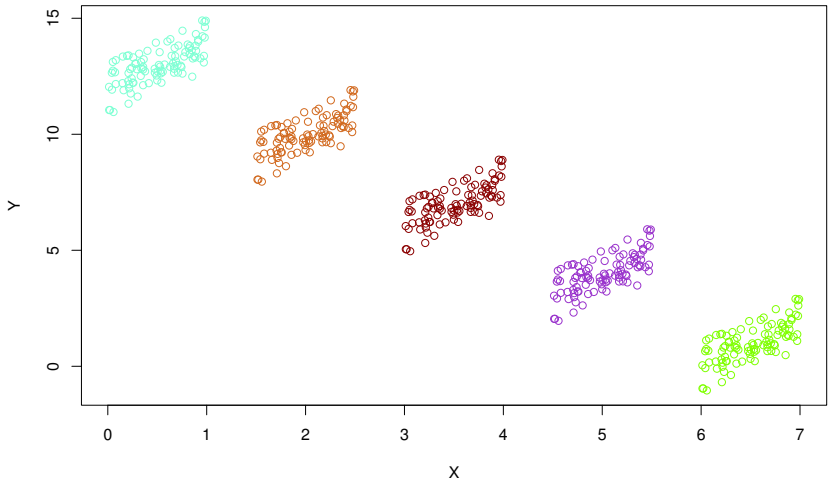
Una estimación mediante OLS permite obtener el siguiente resultado:

Variable	Coef.	Error Est.	Estad. t	Valor-p
Constante $\hat{\alpha}$	$-0,004$	$0,050$	$0,080$	$0,93$
Renta ($\hat{\beta}$)	1,996	$0,012$	$165,260$	$0,00$
Estadístico F	27312			$0,00$
R^2	$0,97$			

Como se puede comprobar, el ajuste del modelo es muy bueno. Además, el coeficiente $\hat{\beta}$ es significativo y positivo. El resultado, gráficamente, es el siguiente:



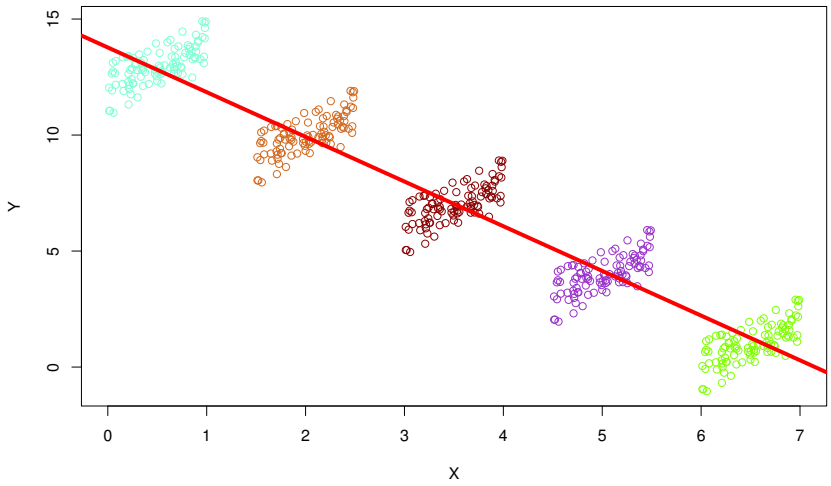
Supongamos ahora que en el anterior ejemplo hay una variable omitida: el patrimonio de cada individuo. Supongamos, además, que los individuos con un alto patrimonio generan poca renta, pero igualmente mantienen un alto nivel de consumo. El diagrama de dispersión adquiere la siguiente forma:



El resultado del estimador OLS en este nuevo escenario es el siguiente:

Variable	Coef.	Error Est.	Estad. t	Valor-p
Constante $\hat{\alpha}$	$13,77$	$0,109$	$125,6$	$0,00$
Renta ($\hat{\beta}$)	$-1,92$	$0,026$	$-72,2$	$0,00$
Estadístico F	5223			$0,00$
R^2	$0,91$			

El ajuste del modelo sigue siendo muy bueno, y los coeficientes son significativos. ¿Dónde está el problema? Para empezar, hemos de preguntarnos: ¿es lógico que la renta ejerza un efecto negativo sobre el consumo? Como vemos en el siguiente gráfico, el problema es que el estimador OLS pasa por la nube de puntos sin tener en cuenta que, para cada individuo, la pendiente individual es positiva.



3.2.2. Efectos fijos (WG/LSDV)

En inglés se denomina *within gropus (WG)* o *least squares dummy variables (LSDV) estimator*. Este método es adecuado cuando los individuos se representan a sí mismos, y es posible que difieran en características y comportamiento (regiones, países, etc.). De esta manera, diferencias entre individuos se reflejarán en diferencias en la constante. Analíticamente, se asume que los efectos individuales son un conjunto de N parámetros que estimar.

Formalmente, la especificación de efectos fijos adquiere la siguiente forma:

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}$$

$$\alpha_i = \alpha + \mu_i$$

La constante tiene dos elementos: α es la constante común y μ_i es un efecto individual que ser estimado. Es importante destacar que μ_i es un parámetro a estimar y no es parte del término de perturbación u_i. Esto provoca que se cumpla la condición de ortogonalidad $E(x_{it}u_{it}) = 0$, ya que μ_i no está recogido en u_i.

El método de estimación tiene varias alternativas.

1) Least Squares Dummy Variables (LSDV): esta transformación implica incluir una variable *dummy* para cada individuo i. Si reescribimos el modelo en notación matricial, tenemos:

$$Y_i = \iota_T \alpha_i + X_i\beta + u_i$$

Donde ι_T es un vector de unos de dimensión $T \times 1$. Si se agregan todos los individuos, se obtiene:

$$Y = D\alpha + X\beta + u$$

En que D es una matrix de *dummies* individuales y α el vector de constantes individuales. El modelo resultante es el siguiente:

$$D = \begin{pmatrix} \iota_T & \vec{0}_T & \cdots & \vec{0}_T \\ \vec{0}_T & \iota_T & \cdots & \vec{0}_T \\ \vdots & \vdots & \ddots & \vdots \\ \vec{0}_T & \vec{0}_T & \cdots & \iota_T \end{pmatrix}_{NT \times N}, \ \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} \alpha + \mu_1 \\ \alpha + \mu_2 \\ \vdots \\ \alpha + \mu_N \end{pmatrix}_{N \times 1}$$

Si a esta especificación se le aplica el estimador OLS, se obtiene el estimador LSDV. El problema del estimador LSDV es que, si N es elevado, habrá demasiadas *dummies* individuales a ser estimadas y puede haber problemas al invertir grandes matrices.

2) Within Groups (WG): esta alternativa se basa en extraer la media aritmética de todas las variables. Para la variable dependiente:

$$\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}, \qquad \text{donde} \qquad \bar{y}_{i\cdot} = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

La extracción de la media de las variables elimina los efectos individuales, ya que:

$$\bar{\alpha}_i = \alpha_i \qquad \text{de manera que} \qquad \tilde{\alpha}_i = \alpha_i - \alpha_i = 0$$

Esta transformación elimina los efectos que no varían en el tiempo. Al eliminar α_i, este ya no estará correlacionado con x_i, y esto garantiza la condición de ortogonalidad $E(x_{it}u_{it}) = 0$. Para obtener este estimador hay que aplicar algunas transformaciones. Primero, creamos una matriz de proyección de dimensión $T \times T$:

$$M_0 = I_T - \frac{1}{T}\iota\iota'$$

Donde ι es un vector de unos de dimensión $T \times 1$. Esta matriz hace que:

$$M_0 Y_i = \begin{pmatrix} \tilde{y}_{i1} \\ \vdots \\ \tilde{y}_{iT} \end{pmatrix} = \begin{pmatrix} y_{i1} - y_{i\cdot} \\ \vdots \\ y_{iT} - y_{i\cdot} \end{pmatrix}_{T\times 1}$$

Generalizamos M_0 para aplicarla a todo el modelo y extraer la media de todos los individuos:

$$M_d = I_{NT} - D(D'D)^{-1}D' = \begin{pmatrix} M_0 & 0_{T\times T} & \cdots & 0_{T\times T} \\ 0_{T\times T} & M_0 & \cdots & 0_{T\times T} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{T\times T} & 0_{T\times T} & \cdots & M_0 \end{pmatrix}_{NT\times NT}$$

Si multiplicamos M_d por D, obtenemos:

$$\begin{aligned} M_d D &= (I_{NT} - D(D'D)^{-1}D')D \\ &= I_{NT}D - D(D'D)^{-1}D'D \\ &= 0_{NT\times N}. \end{aligned}$$

Para acabar, premultiplicamos la especificación por M_d para eliminar efectos individuales:

$$M_d Y = M_d X\beta + M_d u$$

Aplicamos OLS a este modelo transformado y obtenemos el estimador WG:

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

Los efectos individuales estimados se pueden recuperar aplicando la siguiente fórmula:

$$\hat{\alpha}_i = \hat{\alpha}_0 + \hat{\mu}_i = \bar{y}_{i\cdot} - \hat{\beta}'_{WG}\bar{x}_{i\cdot}$$

El estimador de efectos fijos posee las siguientes propiedades:

1) El estimador LSDV/WG es **consistente** tanto para $E(x_{it}\mu_i) = 0$ como para $E(x_{it}\mu_i) \neq 0$, ya que la transformación elimina los efectos individuales.

2) Si se cumple que $E(x_{it}\mu_i) = 0$, el estimador LSDV/WG será **menos eficiente** a medida que $N \longrightarrow \infty$ con T fijado. Esto se debe a que estimar N constantes implica una reducción de grados de libertad.

3) Si se cumple que todos los regresores están correlacionados con μ_i, el estimador LSDV/WG será **eficiente**.

4) *Limitación:* no se pueden incluir variables que no varíen temporalmente.

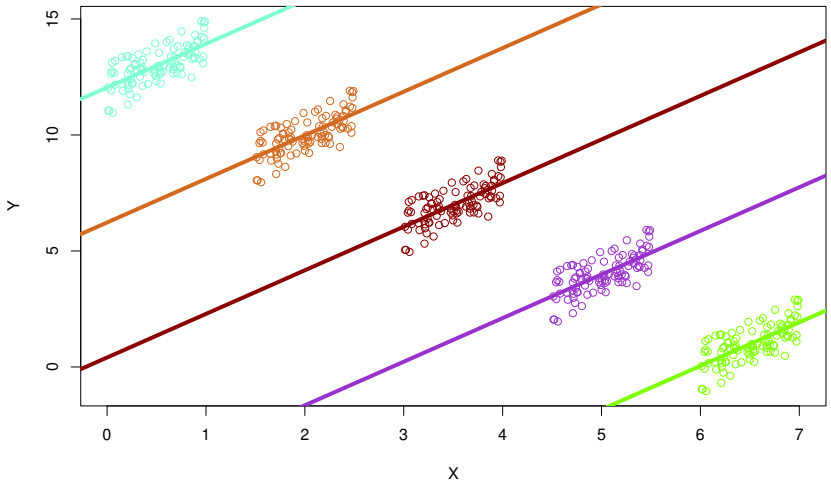
Retomamos el ejemplo anterior de la relación entre renta y consumo. Ahora, asumimos la hipótesis de efectos fijos:

$$C_{it} = \alpha + \mu_i + \beta R_{it} + u_{it}$$

En esta nueva especificación, la presencia del efecto individual μ_i puede mitigar la ausencia de la variable patrimonio. Aplicamos el estimador WG/LSDV:

Variable	Coef.	Error Est.	Estad. t	Valor-p
$\hat{\alpha}_1$	12,04	0,07	164,28	0,00
$\hat{\alpha}_2$	6,22	0,18	33,53	0,00
$\hat{\alpha}_3$	0,40	0,31	1,30	0,19
$\hat{\alpha}_4$	−5,41	0,44	−12,20	0,00
$\hat{\alpha}_5$	−11,23	0,57	−19,57	0,00
Renta ($\hat{\beta}$)	1,87	0,08	21,44	0,00
Estadístico F	459			0,00
R^2	0,48			

Vemos que, en este caso, tenemos cinco rectas diferentes, todas con la misma pendiente pero con una constante que varía de individuo a individuo.



3.2.3. Primeras diferencias (FD)

En inglés se denomina *First Differenced OLS (FD)*. Este estimador se basa en calcular primeras diferencias como alternativa para eliminar los efectos individuales. Si tomamos el modelo inicial:

$$y_{it} = \alpha + \mu_i + x'_{it}\beta + u_{it}$$

Tomando primeras diferencias se eliminan α y μ_i, ya que estos no varían con el tiempo:

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta u_{it}$$

Donde $\Delta y_{it} = y_{it} - y_{i,t-1}$, por ejemplo. Si aplicamos MCO a este modelo obtenemos el estimador $\hat{\beta}_{FD}$.

Las propiedades de este estimador son las siguientes:

1) El estimador será **consistente** si $E(\Delta x_{it}\Delta u_{it}) = 0$ (regresores exógenos).

2) Si el término u_{it} es persistente en el tiempo, Δu_{it} no tendrá correlación serial y el estimador FD será **más eficiente** que el estimador WG/LSDV.

3) En cambio, si $\tilde{u}_{it}$ no muestra correlación serial, FD será **menos eficiente** que WG/LSDV.

4) Al igual que el estimador WG/LSDV, no se pueden incluir variables que no varíen temporalmente.

3.2.4. Entre grupos (BG)

En inglés, se denomina estimador *between groups (BG)*. Este modelo incluye solo las medias individuales de las variables:

$$\bar{y}_{i\cdot} = \bar{x}_{i\cdot}'\beta + \bar{\eta}_i + \bar{v}_{i\cdot}, \qquad i = 1, \ldots, N.$$

Este modelo solo incluye la dimensión individual, y no incluye información temporal. Por ello, no se puede estudiar la tendencia dinámica de las variables. Sus principales propiedades son que este estimador BG es **consistente** si $E(x_{it}\mu_i) = 0$. Además, es un estimador **no eficiente**, ya que sólo tiene $N - k$ grados de libertad. Se suele usar como paso previo antes de efectuar otras estimaciones, para comprobar la variabilidad de las observaciones entre diferentes grupos.

3.2.5. Efectos aleatorios (RE/GLS)

En inglés se denomina *random effects generalized least squares (RE/GLS)*. Este modelo es apropiado si se extraen N individuos aleatoriamente de una población grande, y la muestra es representativa. Así, el interés no es el conjunto de características de cada individuo, sino hacer inferencia sobre las características de la población.

La especificación de partida es:

$$y_{it} = \alpha + x_{it}'\beta + u_{it}$$

$$u_{it} = \mu_i + \epsilon_i$$

El efecto individual μ_i se toma como aleatorio (debido al azar), y se modeliza como una variable (perturbación) aleatoria. El efecto μ_i es considerado como una perturbación aleatoria constante en el tiempo, homoscedástica, no autocorrelacionada y con media cero:

$$\mu_i \sim iid(0, \sigma_\mu^2)$$

$$E(\mu_i\mu_j) = 0, \quad \forall i \neq j$$

$$E(\mu_i^2) = \sigma_\mu^2$$

$$E(\mu_i) = 0.$$

Similarmente, obtenemos que:

$$\epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$$

$$E(\epsilon_{it}\epsilon_{js}) = 0, \quad \forall i \neq j,\ t \neq s$$

$$E(\epsilon_{it}^2) = \sigma_\epsilon^2.$$

$$E(\epsilon_i) = 0.$$

Para entender la constucción de este estimador, volvamos atrás al modelo de regresión lineal:

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

Si se analiza el término de perturbación u_i, se asume que tiene media condicional cero ($E(u_i|x_i) = 0$) y varianza finita. La varianza condicional del modelo será $E(u_i^2|x_i) = \sigma_i^2$, y la matriz de varianzas y covarianzas de u es:

$$MVC(u) = E(uu') = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{pmatrix}_{N\times N}$$

Recordemos que, para que el estimador MCO sea eficiente (mínima varianza de la estimación), esta matriz debe ser **esférica**, es decir:

1) Homoscedástica: la varianza de u no varía entre los elementos de la muestra, de manera que $\sigma_i^2 = \sigma^2$ 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 ambas condiciones se cumplen, la matriz $MVC(u)$ será:

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

Siendo I_N la matriz identidad de dimensión $N \times N$. En el caso de que $MVC(u)$ no sea esférica, asumimos que $MVC(u) = E(uu') = \Omega$. En este caso, MCO no incorpora la estructura del error en la estimación del modelo, y no es eficiente. El estimador eficiente será *Mínimos Cuadrados Generalizados (GLS)*:

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

Ya que Ω es desconocido, habrá que estimarlo o imponerle una estructura ($\hat{\Omega}$), y para después aplicar el estimador por Mínimos Cuadrados Generalizados Factibles (GLSF):

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

Volviendo al modelo con datos de panel, en el modelo de efectos aleatorios la matriz $MVC(u)$ no es esférica.

$$E(u_{it}^2) = \sigma_\mu^2 + \sigma_\epsilon^2$$

$$E(u_{it}u_{is}) = \sigma_\mu^2, \quad \forall t \neq s.$$

Para el individuo i, la $MVC(u)$ será:

$$\Omega_i = E(u_i u_i') = \sigma_\epsilon^2 I_T + \sigma_\mu^2 \iota'$$

Donde ι es un vector de unos $T \times 1$. La matrix Ω_i no será esférica:

$$\Omega_i = \begin{pmatrix} \sigma_\mu^2 + \sigma_\epsilon^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\epsilon^2 & \cdots & \sigma_\mu^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_\epsilon^2 \end{pmatrix}_{T\times T}$$

La matriz $MVC(u)$ para todos los individuos será entonces:

$$\Omega = E(uu') = I_N \otimes \Omega_{T\times T},$$

Con la siguiente expresión analítica:

$$\Omega = \begin{pmatrix} \Omega_1 & 0_{T\times T} & \cdots & 0_{T\times T} \\ 0_{T\times T} & \Omega_2 & \cdots & 0_{T\times T} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{T\times T} & 0_{T\times T} & \cdots & \Omega_N \end{pmatrix}_{NT\times NT}.$$

Ω es una matriz diagonal en bloques que exhibe correlación serial en el tiempo.

Para obtener una estimación eficiente del modelo, la estructura del término de perturbación se incluye en la estimación mediante Mínimos Cuadrados Generalizados (GLS):

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

Wansbeek y Kapteyn idearon un procedimiento para obtener el estimador, partiendo de la desviación de Ω^{-1} y $\Omega^{-1/2}$. Si se premultiplica el modelo:

$$Y^* = \Omega^{-1/2}Y$$

$$X^* = \Omega^{-1/2}X$$

$$\epsilon^* = \Omega^{-1/2}\epsilon,$$

Se obtiene el estimador RE/GLS aplicando mínimos cuadrados ordinarios al modelo:

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*'}X^*)^{-1}X^{*'}Y^* \\ &= (X'\Omega^{-1/2}\Omega^{-1/2}X)^{-1}X'\Omega^{-1/2}\Omega^{-1/2}Y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.\end{aligned}$$

Este prodecimiento se denomina *θ-diferenciación*, y para aplicarlo hay que definir la siguiente matriz de varianzas y covarianzas (MVC) individual:

$$\Omega_i^{-1/2} = I_T - \frac{\theta}{T}\iota\iota', \qquad \theta = 1 - \frac{\sigma_\epsilon}{\sqrt{T\sigma_\mu^2 + \sigma_\epsilon^2}},$$

Donde ι es un vector de unos $T \times 1$.

La MVC global se deriva entonces de la siguiente manera:

$$\Omega^{-1/2} = I_N \otimes \Omega_i^{-1/2}.$$

Veamos cómo la transformación afecta a las variables:

$$Y_i^* = \Omega_i^{-1/2}Y_i = \begin{pmatrix} y_{i1} - \theta\bar{y}_{i\cdot} \\ \vdots \\ y_{iT} - \theta\bar{y}_{i\cdot} \end{pmatrix}_{T\times 1}.$$

Dependiendo de los valores de las varianzas σ_μ^2 y σ_ϵ^2, el valor θ variará, lo que tiene implicaciones con respecto a la estimación. Sin embargo, los valores de estas varianzas, y por lo tanto la matriz Ω, son desconocidos. Por eso, hay que estimar $\hat{\sigma}_\mu^2$ y $\hat{\sigma}_\epsilon^2$, habiendo varios procedimientos para hacerlo. En este caso, el estimador se denomina Mínimos Cuadrados Generalizados Factibles (FGLS).

Las propiedades del estimador de efectos aleatorios son las siguientes:

1) Requiere que las variables explicativas no estén correlacionadas con los efectos individuales: $E(x_{it}\mu_i) = 0$. En este caso, el estimador RE/GLS será **consistente** y **eficiente**.

2) Si $E(x_{it}\mu_i) \neq 0$, RE/GLS es **inconsistente** a medida que $N \longrightarrow \infty$ con T fijado.

3) Si no hay efectos individuales, de manera que $\sigma_\mu^2 = 0$, entonces $\theta = 0$, y los estimadores RE/GLS y OLS coincidirán, siendo ambos **eficientes**. Esto sucede porque aplicar GLS a un modelo con MVC esférica equivale a aplicar el estimador de mínimos cuadrados OLS.

4) A medida que $T \longrightarrow \infty$, $\theta \longrightarrow 1$ y RE/GLS tenderá a coincidir con el estimador WG.

5) A diferencia del estimador de efectos fijos, este estimador permite incluir regresores que no varíen temporalmente.

3.2.6. Coeficientes variables

En inglés se denomina *variable coefficients model*. Este modelo relaja la hipótesis de que los parámetros del modelo son comunes ($\beta_{ij} = \beta$). De esta manera, el modelo de coeficientes variables aporta más flexibilidad, permitiendo estimar un coeficiente para cada individuo ($\hat{\beta}_i$) y/o para cada período ($\hat{\beta}_t$), además de la constante ($\alpha_i,\ \alpha_t$).

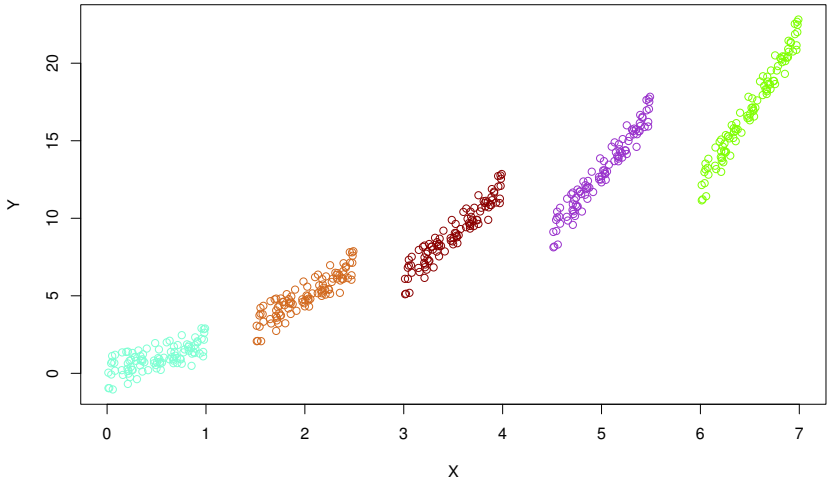
El modelo de coeficientes variables puede ser estimado con efectos fijos o efectos aleatorios:

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}.$$

Como ejemplo, retomemos el modelo de renta y consumo analizado anteriormente:

$$C_{it} = \alpha + \beta R_{it} + u_{it}$$

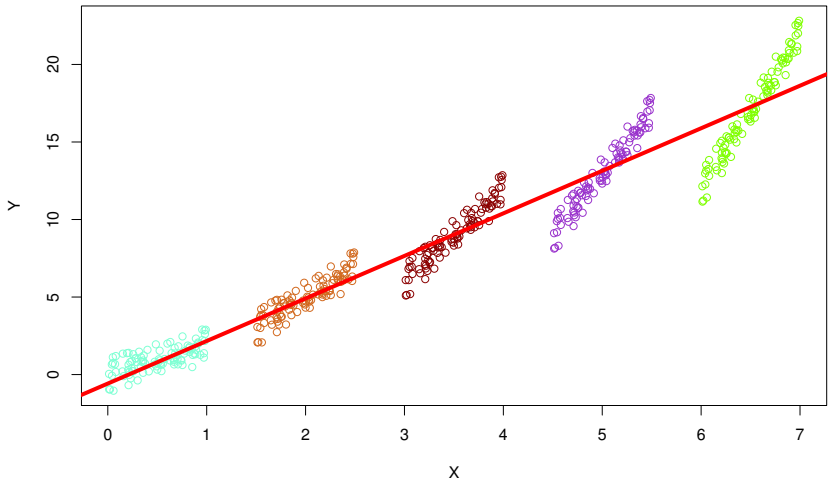
Supongamos que nos encontramos ante el siguiente diagrama de dispersión de las dos variables:



Si estimamos el modelo con *pooled OLS*, obtenemos el siguiente resultado:

Variable	Coef.	Error Est.	Estad. t	Valor-p
Constante $\hat{\alpha}$	$-0,57$	$0,11$	$-4,87$	$0,00$
Renta $\hat{\beta}$	$2,74$	$0,03$	$94,97$	$0,00$
Estadístico F	9019			$0,00$
R^2	$0,94$			

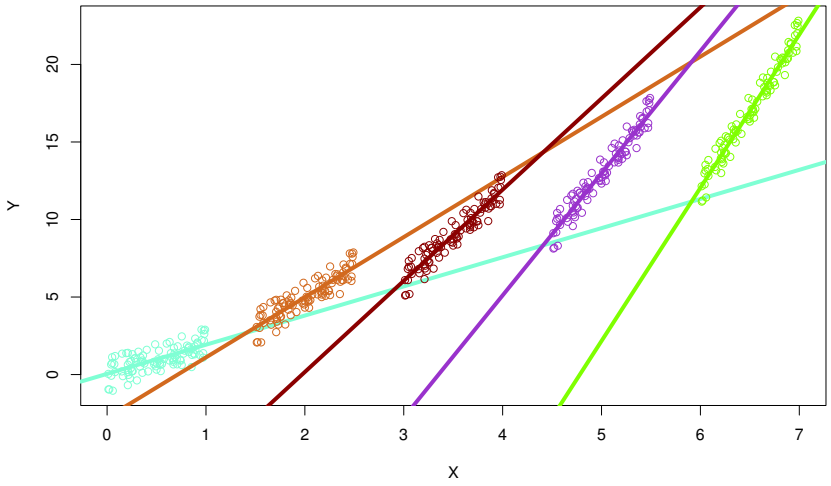
Se puede comprobar que el ajuste del modelo es bastante bueno, y los coeficientes estimados son significativos. Veamos ahora el ajuste de la recta estimada sobre las observaciones.



Se puede comprobar cómo la recta estimada no se ajusta a los diferentes individuos de la muestra. Si estimáramos un modelo de coeficientes variables con efectos fijos, obtendríamos el siguiente resultado:

i	α_i	β_i
1	0,04	1,87
2	−2,77	3,87
3	−11,59	5,87
4	−26,41	7,87
5	−47,23	9,87
R^2	0,99	

Vamos a comprobar gráficamente cómo el ajuste de la estimación a las observaciones ha mejorado con esta nueva estimación.



3.2.7. Método generalizado de los momentos (GMM):

En inglés se denomina *generalized method of moments (GMM)*. En datos de panel, es muy frecuente el análisis de ecuaciones dinámicas del tipo:

$$y_{it} = \rho y_{i,t-1} + \beta x_{it} + \mu_i + u_{it}$$

Los modelos que incorporan un retardo de la variable explicada como regresor sufren de endogeneidad. Esto es, se rompe el supuesto de exogeneidad de los regresores, y además sucede que $E(y_{i,t-1}u_{it} \neq 0)$. Esto sucede también cuando se toman primeras diferencias para eliminar los efectos individuales.

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta u_{it}$$

La solución es utilizar variables instrumentales y el estimador GMM para eliminar el sesgo causado por la endogeneidad.

3.3. Inferencia

A la hora de especificar y estimar un modelo econométrico de datos de panel, nos encontramos ante muchas posibilidades. Para acertar en la elección de la especificación y del método de estimación existen diferentes contrastes, que nos pueden ayudar a hacer una elección basada en criterios estadísticas. En esta sección se introducen solo los tres contrastes más conocidos y usados.

1) Contraste de efectos fijos: este contraste se basa en la hipótesis de que los términos constantes son todos iguales. Si esta hipótesis es cierta, no existirían los efectos individuales. La hipótesis del contraste es la siguiente:

$$H_0 : \alpha_i = \alpha, \ \ \forall i = 1, \ldots, N$$

$$H_1 : \alpha_i \neq \alpha, \ \ \forall i = 1, \ldots, N$$

El estadístico F del contraste es:

$$F = \frac{(SCR_{OLS} - SCR_{WG})/(N-1)}{SCR_{WG}/(NT-N-K)} \sim F_{(N-1),(NT-N-K)}$$

Donde SCR es la suma de los cuadrados de los residuos $\sum \hat{u}_i^2$. Si el ajuste del modelo WG mejora a OLS porque existen efectos individuales, la diferencia $SCR_{OLS} - SCR_{WG}$ será mayor y F caerá en la región crítica (rechazo de H_0).

2) Contraste de efectos aleatorios: se trata de un contraste de multiplicadores de Lagrange, y se basa en los residuos del modelo OLS. La hipótesis del contraste es la siguiente:

$$H_0 : \sigma_\mu^2 = 0 \implies corr(u_{it}u_{is}) = 0$$

$$H_1 : \sigma_\mu^2 \neq 0 \implies corr(u_{it}u_{is}) \neq 0$$

El estadístico es el siguiente:

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (\sum_{t=1}^{T} \hat{u}_{it})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2} - 1 \right]^2 \sim \chi_1^2$$

Si hay variación intra-grupos (*within*), es decir si $\sigma_\mu^2 > 0$, el estadístico LM toma un valor alto y cae en la zona de rechazo de H_0. En ese caso, existen efectos aleatorios.

3) Test de Hausman: este test se usa para decidir entre los estimadores de efectos fijos y aleatorios (WG/LSDV vs. RE/GLS). Cabe recordar que:

$E(x_{it}\mu_i) = 0$: WG y RE son consistentes, y RE es el estimador eficiente (mínima varianza).

$E(x_{it}\mu_i) \neq 0$: RE es inconsistente, y WG es consistente.

El test de Hausman se basa en que si $E(x_{it}\mu_i) = 0$, WG y RE deberían ser similares. Por eso, el test determinará si existe o no autocorrelación. La hipótesis del contraste es la siguiente:

$$H_0 : E(x_{it}\mu_i) = 0$$

$$H_1 : E(x_{it}\mu_i) \neq 0$$

El estadístico del contraste adquiere la siguiente forma:

$$\hat{q} = \hat{\beta}_{WG} - \hat{\beta}_{RE}$$

$$H = \hat{q}'[avar(\hat{q})]^{-1}\hat{q} \sim \chi^2_k$$

Donde $avar(\hat{q})$ es la varianza asintótica de $\hat{q}$. La decisión final es la siguiente:

No rechazo de H_0: estimador de efectos aleatorios es preferido, ya que $E(x_{it}\mu_i) = 0$.

Rechazo de H_0: estimador de efectos fijos (WG) es preferido, ya que $E(x_{it}\mu_i) \neq 0$.

3.4. Aplicación práctica con R

En esta sección realizamos una revisión del análisis econométrico de datos de panel con R. Para ilustrarlo, consideramos un estudio de la productividad de las manufacturas españolas para diferentes sectores y años. En este sentido, se consideran las siguientes variables:

Y : valor añadido bruto.

L : cantidad de factor trabajo.

K : stock de capital.

Estas variables están disponibles para $i = 1, \ldots, N$ sectores y para $j = 1, \ldots, T$ años. Ya que $N = 11$ sectores y $T = 17$ años (de 1980 a 1996), la dimensión total del panel es de $NT = 187$ observaciones.

El punto de partida es una especificación Cobb-Douglas de la siguiente forma:

$$Y = AL^{\beta_L}K^{\beta_K}$$

Para estimar empíricamente este modelo, se opta por una transformación logarítmica:

$$logY_{it} = \alpha + \beta_L logL_{it} + \beta_K logK_{it} + u_{it}$$

Empezaremos especificando el directorio de trabajo (donde están los datos) mediante el comando `setwd()` y cargando las librerías básicas:

```
> library(plm)
> library(car)
> library(gplots)
```

Es fundamental que las librerías han de instalarse antes de poder cargarlas. Recordemos que la instrucción para cargarlas es `install.packages()`.

La librería`plm` se denomina *Linear Models for Panel Data*, y es la librería de referencia para la estimación de este tipo de modelos. La librería `car` se denomina *Companion to Applied Regression*, y la librería `gplots` es *Various R programming tools for plotting data*. Estas dos últimas las cargamos para realizar algunos gráficos descriptivos previos multidimensionales, esto es, con datos con dos dimensiones (individuos y tiempo). Seguidamente, leemos el documento con los datos para obtener la base de datos:

```
> datos<-read.delim2("datos.txt",header=TRUE,dec=",")
```

Para una primera aproximación a los datos, haremos lo siguiente:

```
> summary(datos)

      year          sector         Y
 Min.   :1980   Min.   : 1   Min.   : 213814
 1st Qu.:1984   1st Qu.: 3   1st Qu.: 449129
 Median :1988   Median : 6   Median : 642399
 Mean   :1988   Mean   : 6   Mean   : 705955
 3rd Qu.:1992   3rd Qu.: 9   3rd Qu.: 838315
 Max.   :1996   Max.   :11   Max.   :1986698
       L                K
 Min.   : 80574   Min.   : 1492221
 1st Qu.:111936   1st Qu.: 3141607
 Median :142229   Median : 4256334
 Mean   :177993   Mean   : 5502817
 3rd Qu.:230264   3rd Qu.: 7285621
 Max.   :384167   Max.   :14903525
```

La función `head()` nos enseña la cabecera de los datos:

```
> head(datos)

  year sector      Y      L        K
1 1980      1 718333 205073 13635632
2 1981      1 687133 193424 13583555
3 1982      1 597825 173109 13290383
4 1983      1 591872 158456 13031579
5 1984      1 567590 152955 12821882
6 1985      1 545167 142229 12908390
```

El objeto creado `datos` es del tipo `"data.frame"`, es decir, una base de datos. Para efectuar el análisis de datos de panel, hemos de crear un nuevo objeto del tipo `"pdata.frame"`, que incorpora información sobre las dos dimensiones que conforman el panel, es decir, la individual y la temporal:

```
> datos.pd<-pdata.frame(datos,index=c("sector","year"))
```

Visualicemos ahora la cabecera de la nueva base de datos.

```
> head(datos.pd)

       year sector      Y      L        K
1-1980 1980      1 718333 205073 13635632
1-1981 1981      1 687133 193424 13583555
1-1982 1982      1 597825 173109 13290383
1-1983 1983      1 591872 158456 13031579
1-1984 1984      1 567590 152955 12821882
1-1985 1985      1 545167 142229 12908390
```

A continuación, introducimos las variables incluidas en la base de datos en el espacio de trabajo, para poder operar con ellas directamente.

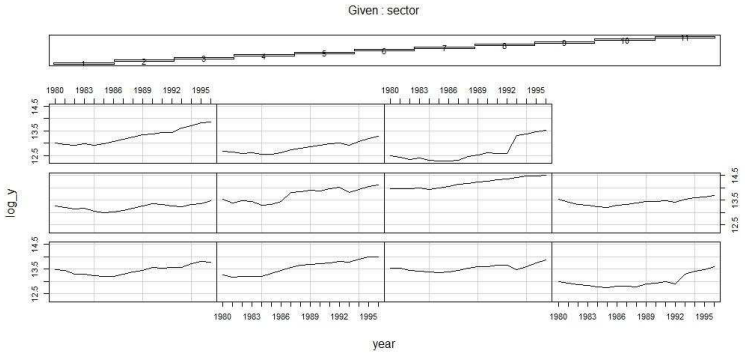
```
> attach(datos.pd)
```

Para estimar el modelo especificado, antes de nada crearemos las tres variables en logaritmos:

```
> log_y<-log(Y)
> log_l<-log(L)
> log_k<-log(K)
```

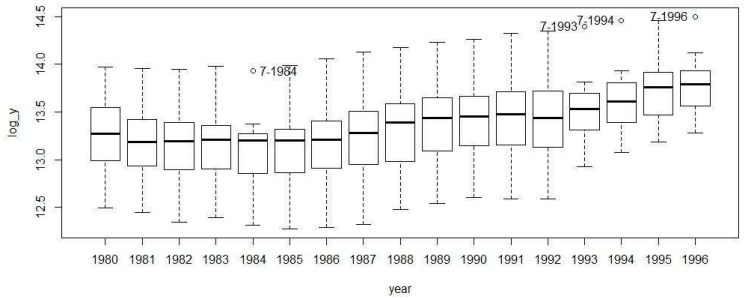
Una función que nos ayuda a entender la estructura bidimensional de los datos es `coplot()`, ya que es una herramienta gráfica muy intuitiva. Analicemos la estructura de la variable renta:

```
>coplot(log_y year|sector,type="l",data=datos.pd)
```



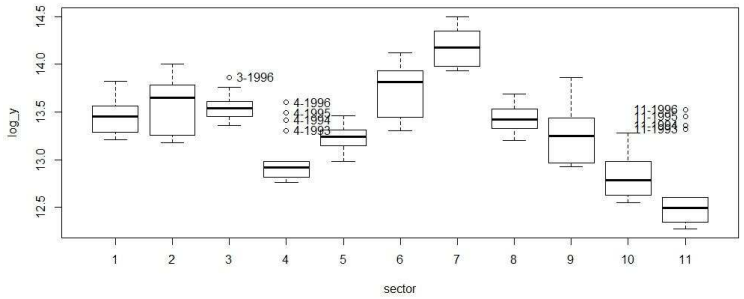
Como vemos, la evolución temporal de $log(Y)$ es positiva para todos los sectores, pero con efectos específicos en cada sector. Otra opción interesante sería ver un gráfico en el que se pueda ver, para cada año, la distribución de $log(Y)$ para los diferentes sectores. Lo haremos con la siguiente instrucción:

```
>scatterplot(log_y year|sector,data=datos.pd)
```



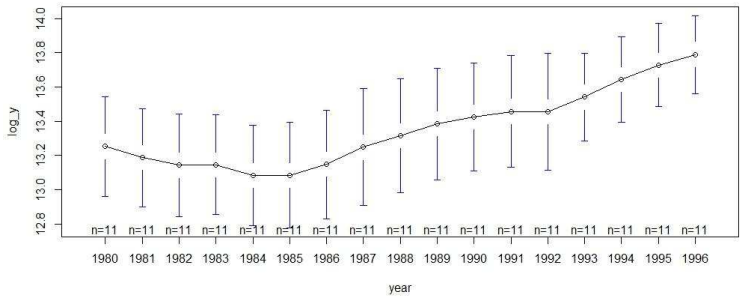
El mismo gráfico pero aplicado a diferentes sectores permite ver cómo la media temporal de $log(Y)$ es diferente para cada sector, de manera que, intuitivamente, es concebible que puedan existir efectos individuales.

```
>scatterplot(log_y sector|year,data=datos.pd)
```

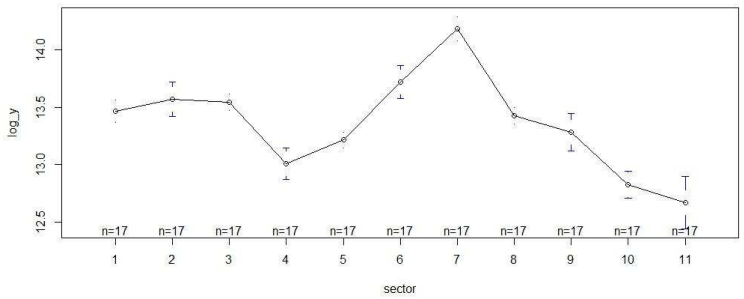


Una aproximación alternativa es usar la función `plotmeans()`, que produce resultados similares a la función anterior. Si aplicamos esta función a ambas dimensiones (sector y año), comprobamos que los efectos sectoriales son significativos.

```
>plotmeans(log_y year)
```




```
>plotmeans(log_y sector)
```



A la hora de especificar y estimar el modelo de regresión, el paquete `plm` permite trabajar con tres efectos: el individual (`effect=individual"`), el temporal (`effect="time"`) y el conjunto de ambos efectos simultáneamente (`effect="twoways"`). Si no lo especificamos, R toma por defecto efectos individuales. En lo que sigue, por simplicidad,

elaboraremos el ejemplo asumiendo solo un efecto individual (sectorial), aunque el análisis es fácilmente generalizable. Por ello, la especificación que elegimos es la siguiente:

$$logY_{it} = \alpha + \beta_L logL_{it} + \beta_K logK_{it} + \mu_i + u_{it}$$

El primer paso es introducir la especificación del modelo, es decir, la fórmula de la regresión que va a ser estimada. Lo haremos creando un objeto del tipo `"formula"`.

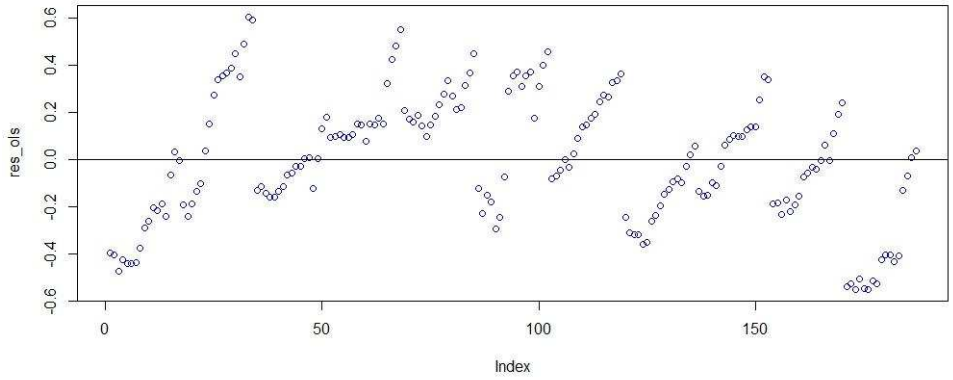
```
> eq1<-log_y~log_l+log_k
```

El primer paso será efectuar una regresión *pooled OLS*, es decir, mínimos cuadrados ordinarios ignorando los efectos individuales. El resultado es el siguiente:

```
> m_ols<-plm(eq1,data=datos.pd,model="pooling")
> summary(m_ols)

Oneway (individual) effect Pooling Model

Call:
plm(formula = eq1, data = datos.pd, model = "pooling")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.55400 -0.18300 -0.00176  0.18100  0.60600

Coefficients :
            Estimate Std. Error t-value Pr(>t)
(Intercept) 0.871465   0.618935  1.4080   0.1608
log_l       0.578346   0.051686 11.1897   <2e-16 ***
log_k       0.361428   0.038264  9.4456   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    42.488
Residual Sum of Squares: 13.223
R-Squared      :  0.68878
      Adj. R-Squared :  0.67773
F-statistic: 203.611 on 2 and 184 DF, p-value: < 2.22e-16
```

Vemos cómo el ajuste es razonablemente bueno, y los coeficientes estimados son estadísticamente significativos. Una buena herramienta de análisis del modelo es una primera inspección visual de los residuos de la regresión.

```
> res_ols<-m_ols[["residuals"]]
> plot(res_ols,col="blue")
> abline(h=0)
```



La estructura del residuo, en el gráfico, revela que el comportamiento no parece ser aleatorio, ya que los efectos individuales no han sido tenidos en cuenta y estos se manifiestan en el residuo de la regresión.

El siguiente paso es estimar el modelo mediante el estimador de efectos fijos (*within groups*).

```
> m_wg<-plm(eq1,data=datos.pd,model="within",effect="individual
    ")
```

Antes de analizar los resultados, hay que preguntarse: ¿los efectos fijos son significativos? Es decir, ¿existe una constante común para todos los individuos, o es diferente? Aquí, dos contrastes son de utilidad. El primero es un contraste del tipo F, que evalúa la hipótesis nula de que $\alpha_i = \alpha$. En R:

```
> pFtest(m_wg,m_ols)

        F test for individual effects

data:  eq1
F = 46.1892, df1 = 10, df2 = 174, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Según el contraste F, la constante es diferente para cada individuo. Una alternativa es realizar un test de Multiplicadores de Lagrange, que evalúa la propia hipótesis nula, que como vemos a continuación da el mismo resultado.

```
> plmtest(m_ols,effect="individual")

        Lagrange Multiplier Test - (Honda)

data:  eq1
normal = 21.0093, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Llegados a este punto, veamos el resultado de la estimación con efectos fijos.

```
> summary(m_wg)

Oneway (individual) effect Within Model

Call:
plm(formula = eq1, data = datos.pd, effect = "individual",
    model = "within")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.49900 -0.06520  0.00527  0.06550  0.37400

Coefficients :
      Estimate Std. Error t-value  Pr(>t)
log_l 0.757130   0.078966  9.5881 < 2.2e-16 ***
log_k 0.999751   0.074056 13.4999 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    11.582
Residual Sum of Squares: 3.6183
R-Squared      :  0.68758
      Adj. R-Squared :  0.63978
F-statistic: 191.475 on 2 and 174 DF, p-value: < 2.22e-16
```

También tenemos la opción de extraer las constantes individuales.

```
> summary(fixef(m_wg))

   Estimate Std. Error t-value  Pr(>t)
1  -12.0523     1.3056 -9.2313 < 2.2e-16 ***
2  -11.0916     1.2613 -8.7937 < 2.2e-16 ***
3  -11.2379     1.2672 -8.8682 < 2.2e-16 ***
4  -10.2121     1.1875 -8.5998 < 2.2e-16 ***
5  -10.4387     1.2092 -8.6325 < 2.2e-16 ***
6  -11.1690     1.2722 -8.7792 < 2.2e-16 ***
7  -11.4950     1.3128 -8.7559 < 2.2e-16 ***
8  -11.2453     1.2622 -8.9093 < 2.2e-16 ***
9  -10.9611     1.2394 -8.8437 < 2.2e-16 ***
10 -10.7709     1.2063 -8.9288 < 2.2e-16 ***
11 -11.0980     1.2150 -9.1341 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Una manera alternativa de eliminar los efectos fijos consiste en aplicar el estimador en primeras diferencias.

```
> m_fd<-plm(eq1,data=datos.pd,model="fd")
```

¿Qué estimador es más efectivo para eliminar los efectos individuales, *wg* o *fd*? La respuesta está en si el término de perturbación u_{it} muestra correlación serial. Para encontrar una respuesta a esta pregunta hay un test que se basa en estimar el siguiente modelo:

$$\hat{u}_{it} = \rho \hat{u}_{i,t-1} + e_{it}$$

Si $\hat{\rho} \to 0$, entonces el estimador más adecuado será efectos fijos (*wg*), mientras que si $\hat{\rho} \to 1$, entonces tendrá sentido aplicar primeras diferencias para así eliminar (además de los efectos fijos) la correlación de la perturbación ρ. El test que aplicar es el *Test de primeras diferencias de Wooldridge*, que en una de sus formas contrasta (bajo H_0) si $Cor(u_{it}u_{i,t-1}) = 0$.

```
> pwfdtest(eq1,data=datos.pd,h0="fe")

        Wooldridge's first-difference test for serial
        correlation in panels

data:  plm.model
chisq = 57.9424, p-value = 2.699e-14
alternative hypothesis: serial correlation in original errors
```

Este resultado parece indicar que $Cor(u_{it}u_{i,t-1}) \neq 0$. Esto es una indicación para usar el modelo *fd*. En el modelo de primeras diferencias el término de perturbación pasaría a ser

$$e_{it} = u_{it} - u_{i,t-1}.$$

Un segundo contraste confirmará si $Cor(e_{it}e_{i,t-1}) = 0$, lo que daría validez a la opción del modelo *fd*:

```
> pwfdtest(eq1,data=datos.pd,h0="fd")



        Wooldridge's first-difference test for serial
        correlation in panels

data:  plm.model
chisq = 0.0612, p-value = 0.8046
alternative hypothesis: serial correlation in differenced
    errors
```

Efectivamente, el estimador preferido para eliminar (*wipe out*) los efectos individuales es *fd*.

El estimador entre grupos *between groups* se aplica a las medias aritméticas de las variables, con lo que es un estimador que no aprovecha la estructura de dos dimensiones. Si hacemos una regresión tomando las medias de los sectores, obtenemos:

```
> m_bg_i<-plm(eq1,data=datos.pd,model="between",effect="
    individual")
> summary(m_bg_i)

Oneway (individual) effect Between Model

Call:
plm(formula = eq1, data = datos.pd, effect = "individual",
    model = "between")

Balanced Panel: n=11, T=17, N=187

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-0.4080 -0.1320  0.0612  0.1650  0.2220
```

```
Coefficients :
            Estimate Std. Error t-value Pr(>t)
(Intercept)  1.70898    2.31037  0.7397  0.48062
log_l        0.56305    0.19817  2.8413  0.02177 *
log_k        0.31881    0.14357  2.2206  0.05713 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1.818
Residual Sum of Squares: 0.43502
R-Squared      :  0.76072
      Adj. R-Squared :  0.55325
F-statistic: 12.7169 on 2 and 8 DF, p-value: 0.0032781
```

En cambio, tomando las medias de los años (es decir, agregando sectores):

```
> m_bg_t<-plm(eq1,data=datos.pd,model="between",effect="time")
> summary(m_bg_t)

Oneway (time) effect Between Model

Call:
plm(formula = eq1, data = datos.pd, effect = "time", model = "
    between")

Balanced Panel: n=11, T=17, N=187

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.10100 -0.04540 -0.00802  0.05150  0.10200

Coefficients :
             Estimate Std. Error t-value  Pr(>t)
(Intercept) -14.37941    2.56102 -5.6147 6.378e-05 ***
log_l         0.39246    0.21274  1.8448   0.08633 .
log_k         1.50029    0.13899 10.7942 3.596e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.80441
Residual Sum of Squares: 0.06308
R-Squared      :  0.92158
      Adj. R-Squared :  0.75895
F-statistic: 82.2647 on 2 and 14 DF, p-value: 1.8236e-08
```

Como vemos, la estimación es sensiblemente diferente, lo que nos obliga a ser muy rigurosos al elegir el modelo que estimar y al interpretar sus resultados.

El estimador que nos queda es el de efectos aleatorios (*random effects*), que será consistente siempre que los efectos individuales (que son parte de la perturbación segun el modelo *random effects*) no estén correlacionados con las variables explicativas. Primero estimamos el modelo y a continuación efectuamos el test de Hausman, que contrasta la hipótesis nula de que $E(x_{ij}\mu_i) = 0$.

```
> m_re<-plm(eq1,data=datos.pd,model="random")
> phtest(m_wg,m_re)

        Hausman Test

data:  eq1
chisq = 49.5461, df = 2, p-value = 1.743e-11
alternative hypothesis: one model is inconsistent
```

El test de Hausman parece indicar el rechazo de H_0, con lo que el modelo *random effects* no es consistente y el estimador *within effects* es consistente.

Un último modelo estático es el de coeficientes variables, que asume que $\beta_i \neq \beta$. El contraste efectuado con la función `pooltest()` nos indicará si realmente existen unos parámetros específicos para cada sector o si los parámetros son comunes.

```
> m_vc<-pvcm(eq1,data=datos.pd,model="within",effect="
    individual")
> pooltest(m_ols,m_vc)

        F statistic

data:  eq1
F = 20.3495, df1 = 30, df2 = 154, p-value < 2.2e-16
alternative hypothesis: unstability
```

El resultado del test parece indicar la existencia de parámetros individuales. Aplicando la función `summary()` a la estimación obtenemos un resumen de esta.

```
> summary(m_vc)

Oneway (individual) effect No-pooling model

Call:
pvcm(formula = eq1, data = datos.pd, effect = "individual",
    model = "within")
```

```
Balanced Panel: n=11, T=17, N=187

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.46650 -0.04995 -0.01030  0.00000  0.05380  0.38560

Coefficients:
  (Intercept)          log_l             log_k
 Min.   :-38.606   Min.   :-1.7687   Min.   :0.6171
 1st Qu.:-11.329   1st Qu.: 0.1800   1st Qu.:0.8511
 Median : -9.663   Median : 0.5613   Median :1.0105
 Mean   : -9.841   Mean   : 0.2757   Mean   :1.2908
 3rd Qu.: -5.536   3rd Qu.: 0.6973   3rd Qu.:1.4670
 Max.   : 25.708   Max.   : 1.0220   Max.   :2.7434

Total Sum of Squares: 1206.2
Residual Sum of Squares: 2.6637
Multiple R-Squared: 0.99779
```

Para obtener una estimación de todos los coeficientes individuales, haremos:

```
> summary(m_vc)[["coefficients"]]


   (Intercept)      log_l     log_k
1   -38.605905  0.5810760 2.7434385
2    -5.465967 -0.1025732 1.2826383
3    -9.663154  0.6943287 0.9480257
4   -11.426971  1.0219711 0.8702287
5    -5.605258  0.7002193 0.7183270
6    25.708400 -1.7687099 0.6170522
7    -4.221188 -0.6289064 1.6513321
8   -29.509106  0.5311384 2.3891937
9    -7.662915  0.4625070 1.0105400
10  -10.569548  0.5612735 1.1359098
11  -11.231134  0.9799564 0.8320530
```

Interpretando estos resultados se podría inferir que la contribución de los factores K y L sobre el valor añadido Y varía según el sector.

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.