UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS

ECONÓMICAS Y EMPRESARIALES

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

TRABAJO DE FIN DE GRADO

TÍTULO: *Matriculaciones de vehículos: un modelo de regresión lineal*

AUTOR: Alejandro Corcuera Ortega

TUTOR: David Casado de Lucas

CURSO ACADÉMICO: 2014/2015

CONVOCATORIA: Junio

ÍNDICE

Resumen

El objetivo principal de este trabajo es describir la dependencia del *número de matriculaciones de vehículos* en España de algunas otras variables macroeconómicas, tanto del sector de la automoción como de la economía nacional. Las variables explicativas han sido obtenidas de las bases de datos de dos organizaciones distintas: la Dirección General de Tráfico (DGT) española y Eurostat. Como herramienta estadística se ha considerado el modelo de regresión lineal, simple y múltiple. Se aplica la metodología usual de modelización para este tipo de regresión, con el objetivo de seleccionar las variables que contienen más información sobre la variable explicada. Hemos tenido en cuenta los principales problemas que pueden surgir al aplicar estos modelos. Como conclusiones principales: durante los años desde 1995 hasta 2008, el número de matriculaciones queda bien explicado mediante el producto interior bruto per cápita real. Por otro lado el uso de una variable ficticia sugiere que hay un cambio estructural a partir del año 2008 (aunque se dispone de pocos datos), y cuando se intenta ajustar un modelo lineal para los años de 1995 a 2013, la variable que mejor explica el patrón del número de matriculaciones es el índice de Gini. En este caso no se alcanza a explicar la misma variabilidad que para los datos anteriores a la crisis, lo que sugeriría buscar alguna variable nueva con más información sobre el número de matriculaciones. Hemos cuantificado la variación del número de matriculaciones en función de la variación de las variables explicativas mencionadas. En el apéndice incluimos todo el código necesario para reproducir o extender nuestros análisis.

PALABRAS CLAVE: economía, econometría, estadística, regresión lineal, homogeneidad estructural, correlación espuria, colinealidad, automoción, matriculaciones, producto interior bruto, índice de Gini.

Introducción

La industria del automóvil, una de las más importantes a nivel mundial, crea millones de empleos directos e indirectos. También genera ingresos sustanciales a muchos países, España entre ellos. Al mismo tiempo, el mercado de vehículos supone una importante fuente de beneficios para la industria de la automoción en particular y para la economía de un país en general. Para las empresas del sector, siempre interesadas en satisfacer las

demandas y necesidades de los clientes, es de gran interés la información sobre la compra y venta de automóviles.

Por tanto, explicar cómo se comportan las matriculaciones de vehículos es un asunto de interés para las organizaciones y empresas. Éste es el **objetivo econométrico** principal de nuestro trabajo, que está dirigido a describir la dependencia del *número de matriculaciones de vehículos* (variable explicada) de algunas otras variables macroeconómicas (variable explicativas): el *número de permisos de conducir emitidos*, el *número de cambios de titularidad*, el *número de bajas de vehículos*, el *censo de conductores*, el *producto interior bruto (PIB) nominal*, el *PIB per cápita nominal*, el *PIB per cápita real*, el *gasto de consumo final de los hogares*, la *población*, el *umbral de riesgo de pobreza* y el *índice de Gini*. Las variables explicativas hasta el censo de conductores, incluida, han sido obtenidas de las bases de datos de la Dirección General de Tráfico (DGT), mientras que las demás han sido obtenidas de las bases de datos de Eurostat. La disponibilidad de datos de la primera fuente hace que hayamos considerado datos desde 1995 hasta 2013. Dado que el número de variables explicativas es once y el número de datos es diecinueve, no consideramos directamente un modelo con todas las variables (además, algunas de las variables explicativas están muy correlacionadas entre sí), sino que para cada base de datos vamos a identificar las que contienen más información sobre la variable explicada y vamos a intentar combinarlas después.

Como herramienta estadística (econométrica, en este caso), utilizamos los modelos de regresión lineal simple y múltiple. Después de presentar los datos y sus características más significativas, aplicamos la metodología usual de ajuste del modelo de regresión lineal para ir seleccionando y añadiendo las variables que mejor describen el número de matriculaciones. Con una crisis económica por medio, los datos tiene más interés pero también son más difíciles de analizar, ya que registrarán hechos anómalos o situaciones más atípicas de lo normal. Durante la modelización, una vez que la definición de las variables y los datos disponibles permiten la utilización de la regresión lineal, tenemos en cuenta los principales problemas que pueden surgir al aplicar el modelo de regresión lineal: dependencia no lineal, presencia de estacionalidad, presencia de tendencia y su posible interpretación, presencia de datos atípicos, presencia de colinealidad (también denominada *multicolinealidad*) y presencia de heterocedasticidad. Como **objetivo técnico** queremos asegurarnos de que ninguno de

los problemas mencionados invalidan los resultados que obtengamos, por un lado, y buscar un modelo suficientemente bueno y sencillo, por otro lado.

Como resultado final, obtenemos que un modelo suficientemente bueno y sencillo consiste en considerar el producto interior bruto per cápita real para los años desde 1995 a 2008. A partir de ese año, la crisis económica modifica bruscamente el comportamiento de algunas variables. El propio carácter de nuestras variables, todas dependientes de la actividad económica subyacente, nos induce a pensar en un uso descriptivo del modelo de regresión lineal, no de causalidad (podría haber un incremento del producto interior bruto que tuviese poco efecto microeconómico y, por tanto, en el número de matriculaciones). También hemos considerado un modelo para todos los datos, aunque no alcanza la calidad de los anteriores. Al incluir los datos de la crisis, la variable que mejor describe el patrón de comportamiento del número de matriculaciones es el índice de Gini. En la sección de conclusiones el lector puede encontrar las variaciones cuantitativas que hemos estimado para el número de matriculaciones en función del censo de conductores, del producto interior bruto per cápita real y del índice de Gini.

Con la intención de hacer un trabajo riguroso pero no muy extenso, incluimos los resultados del modelo final únicamente, mientras que relegamos a la voluntad del lector el reproducir otros de nuestros cálculos, comprobar nuestras afirmaciones o extender nuestro trabajo en la dirección que considere oportuno. Para ello, en el apéndice incluimos todo el código necesario, empezando por la introducción de los datos (allí puede consultar la notación que hemos utilizado para nuestras variables). Hemos intentado también cumplir los **objetivos académicos y pedagógicos** que se le suponen a un trabajo de estas características.

Por último, queremos indicar que es posible aplicar otros análisis a las mismas variables y datos que vamos a considerar nosotros, por ejemplo: se puede trabajar con tasas de variación de las variables, en vez de con las mismas variables, se puede trabajar con las variables estandarizadas, se puede optar entre variables distintas muy correlacionadas entre sí, se puede aplicar una metodología distinta a la de añadir variables al modelo, etcétera.

Conjunto de datos

Presentación

De todas las variables que postulamos como posibles objetos de estudio, seleccionamos las que hemos considerado más importantes y significativas para estudiar el número de matriculaciones. Lo datos que se utilizarán en el estudio abarcan desde el año 1995 hasta el 2013, uno para cada año, es decir, son datos anuales. La primera base de datos puede considerarse «local», en el sentido de que corresponde a la información a la que podríamos acceder dentro de la DGT. La segunda base de datos sería, entonces, «global». En las referencias incluimos las direcciones en internet de las bases de datos.

- *Tiempo* (t): indica el año en que se ha medido cada dato.

Datos obtenidos de la Dirección General de Tráfico (DGT):

- *Número de matriculaciones* (Y): registra el número anual de matriculaciones de vehículos (de cualquier categoría).

- *Permisos de conducir* (X_{L1}): número de carnés de conducir emitidos.
- *Cambios de titularidad* (X_{L2}): número de cambios de titularidad.
- *Bajas* (X_{L3}): número de vehículos retirados de la circulación.
- *Censo de conductores* (X_{L4}): número de conductores activos.

Datos obtenidos de Eurostat:

- *Producto interior bruto (PIB) nominal* (X_{G1}): es un indicador del tamaño de la economía de un país, medida a precios corrientes.
- *PIB per cápita nominal* (X_{G2n}): es el PIB por habitante nominal.
- *PIB per cápita real* (X_{G2r}): es el PIB por habitante a precios fijos, con base en un año.
- *Gasto de consumo final de los hogares* (X_{G3}): registra los gastos efectuados en bienes o servicios por los hogares españoles en el territorio nacional o en el extranjero.
- *Población total* (X_{G4}): tamaño de la población del país.
- *Umbral de riesgo de pobreza* (X_{G5}): cantidad que toma como valor el 60% de la mediana de la renta nacional, a partir de la cual se define la pobreza.

- *Índice de Gini* (X_{G6}): es un número entre 0 y 100, donde 0 corresponde con la perfecta igualdad (todos los ciudadanos tendrían los mismos ingresos) y donde el valor 100 corresponde con la perfecta desigualdad (una persona tiene todos los ingresos y los demás ninguno).

Por otro lado, definimos la siguiente variable que se utilizará en la última parte del análisis:

- *F*: es una variable ficticia creada con nosotros para estudiar la homogeneidad estructural del modelo, toma el valor 0 hasta el año 2008, incluido, y 1 en adelante.

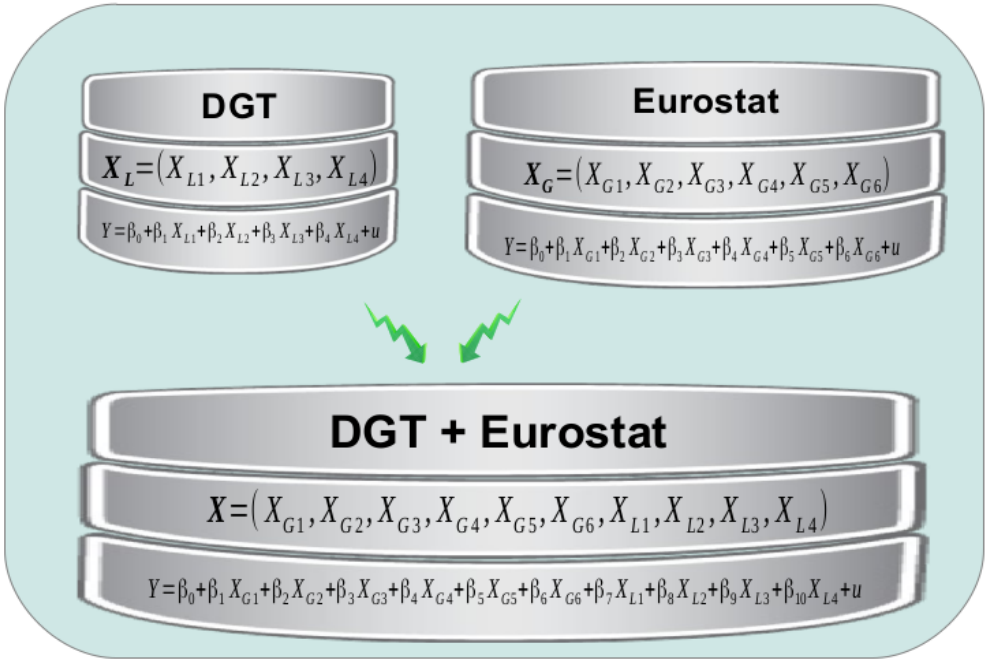


Figura 1: Representación de las variables explicativas según las fuentes, y el modelo más completo que podría construirse con ellas (si el número y la calidad de los datos lo permitiesen).

Las siguientes tablas incluyen los valores numéricos de las variables antes descritas. (Como criterio, utilizamos el punto para separar las cifras decimales. Así utilizamos el mismo criterio en el código.)

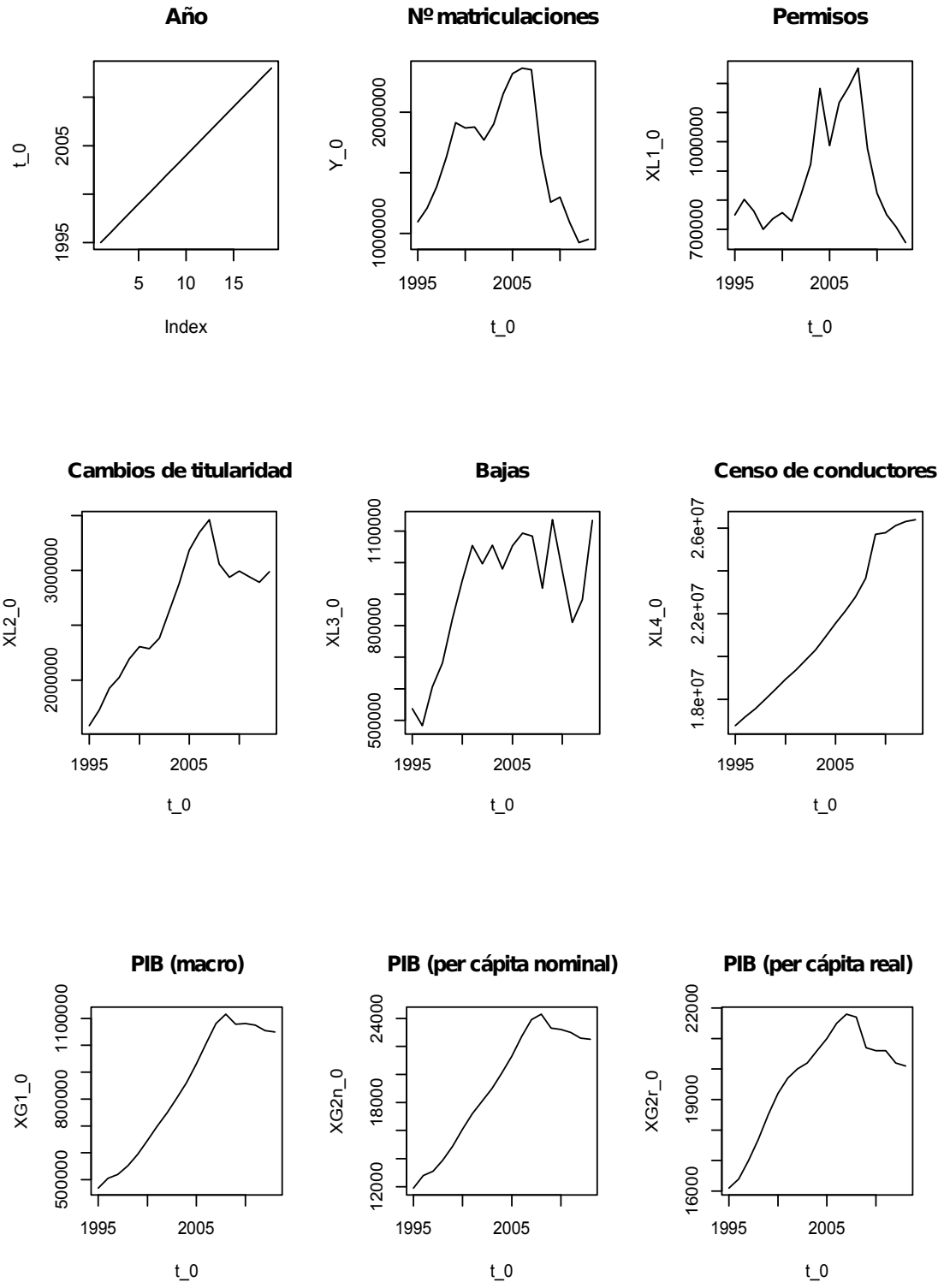
Datos DGT					
Año	**Matriculaciones**	**Permisos emitidos**	**Cambios de titularidad**	**Bajas**	**Censo de conductores**
1995	1096612	749547	1583696	536503	16761681
1996	1209197	802649	1728913	482945	17187616
1997	1385283	762586	1925559	606787	17554104
1998	1627899	700430	2023475	681643	18009374
1999	1913162	735709	2194306	822861	18459615
2000	1870262	756816	2304099	943272	18930263
2001	1875909	728665	2287072	1054181	19348667
2002	1769857	821689	2381805	996224	19823212
2003	1903801	923033	2633069	1055139	20301418
2004	2149706	1182956	2881242	979654	20919181
2005	2319590	987297	3184158	1053457	21549477
2006	2364656	1133961	3344645	1093238	22124198
2007	2350101	1186742	3460348	1083542	22777657
2008	1651013	1252354	3058363	918406	23657166
2009	1258781	977035	2938021	1135642	25713071
2010	1298809	823900	2992928	973926	25782360
2011	1091511	749810	2940634	810638	26118094
2012	924310	708631	2891722	882751	26309230
2013	949015	654924	2987708	1133504	26387882

Tabla 1: Fuente: DGT y Eurostat. Elaboración propia.

Datos Eurostat							
Año	**PIB**	**PIB per cápita nominal**	**PIB per cápita real**	**Gasto final hogares**	**Población**	**Umbral riesgo pobreza**	**Índice de Gini**
1995	468878.7	11900	16100	294834.9	39388.00	3.702	34
1996	505109.1	12800	16400	315649.7	39479.20	3.748	34
1997	519608.4	13100	17000	324558.0	39583.40	3.971	35
1998	551396.7	13900	17700	342783.5	39722.10	4.076	34
1999	594316.0	14900	18500	369697.0	39927.20	4.491	33
2000	646250.0	16100	19200	402272.0	40264.20	4.941	32
2001	699528.0	17200	19700	430871.0	40721.40	5.416	33
2002	749288.0	18100	20000	453637.0	41423.50	5.682	31
2003	803472.0	19000	20200	479081.0	42196.20	5.923	31
2004	861420.0	20100	20600	514256.0	42859.20	6.196	31.0
2005	930566.0	21300	21000	550656.0	43662.60	6.272	32.2
2006	1007974.0	22700	21500	591546.0	44360.50	6.683	31.9
2007	1080807.0	23900	21800	629402.0	45236.00	6.987	31.9
2008	1116207.0	24300	21700	646998.0	45983.20	7.577	31.9
2009	1079034.0	23300	20700	617430.0	46367.60	8.877	32.9
2010	1080913.0	23200	20600	631012.0	46562.50	8.763	33.5
2011	1075147.0	23000	20600	638036.0	46736.30	8.358	34.0
2012	1055158.0	22600	20200	635063.0	46766.40	8.321	34.2
2013	1049181.0	22500	20100	627840.0	46591.90	8.114	33.7

Tabla 2: Fuente: DGT y Eurostat. Elaboración propia.

A continuación incluimos una representación gráfica de cada una de ellas:



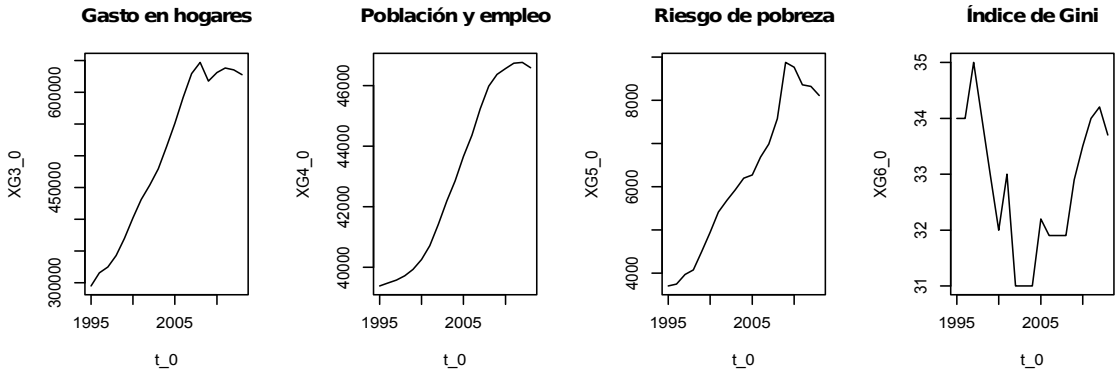


Figura 2: Fuente: DGT y Eurostat. Elaboración propia.

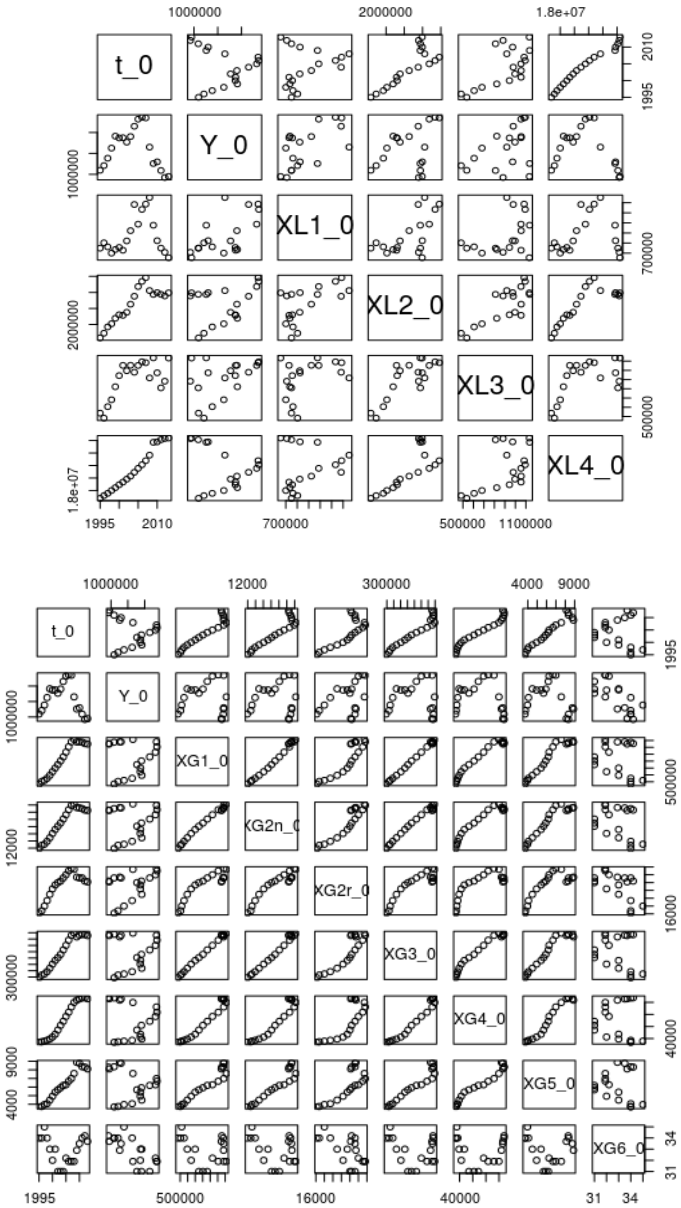


Figura 3: Fuente: DGT y Eurostat. Elaboración propia.

EJERCICIO: *Para introducir los datos y representarlos, el lector puede evaluar el código de las secciones A1 y A2 del apéndice.*

Análisis

Por motivos de extensión, y también porque puede hacerlo fácilmente el lector, no aplicamos un análisis descriptivo a los datos, sino un **análisis cualitativo** de las características de los datos que pueden afectar al análisis estadístico que vamos a aplicar.

En la representación temporal de los datos se puede apreciar que varias de las variables explicativas presentan una **tendencia** similar a la variable explicada `Y_0` (número de matriculaciones) hasta el año 2008. Por otro lado, la variable `XL1_0` (permisos de conducir) es la única que ha registrado el descenso brusco posterior a este año. Quizá la variable `XG6_0` (índice de Gini) también lo ha registrado, pero presenta un **patrón** inverso.

El gráfico de dispersión de `Y_0` (número de matriculaciones) con `XL1_0` (permisos de conducir) es el que parece indicar más claramente una **relación lineal** (directa, como es de esperar), mientras que para las demás parece haber relaciones no lineales o varios datos atípicos o, más probablemente, un cambio estructural. Sorprende el parecido de las representaciones de `Y_0` (número de matriculaciones) con `t_0` (tiempo), por un lado, y con `XL4_0` (censo de conductores), por otro lado. Respecto a las correlaciones, la variable más correlacionada linealmente (+0.62) con `Y_0` (número de matriculaciones) es `XL1_0` (permisos de conducir).

Respecto a los datos de Eurostat, el gráfico de dispersión de `Y_0` (número de matriculaciones) con `XG6_0` (índice de Gini) es el que parece indicar más claramente un patrón similar pero invertido (lo que sugiere una relación lineal inversa, como es de esperar), mientras que para las demás variables parece haber una relación no lineal o la presencia de varios datos atípicos o, más probablemente, un cambio estructural. La variable más correlacionada linealmente (–0.73) con `Y_0` (número de matriculaciones) es `XG6_0` (índice de Gini).

Aunque no conviene fiarse totalmente de los gráficos por pares de variables para juzgar la relación de la variable explicada y las variables explicativas (véase la sección 3.2.5 del libro de Montgomery y otros), aportan información útil, especialmente para la relación de unas variables explicativas con otras. En la sección de modelización tendremos en cuenta esta información previa, que al menos indica qué variables parecen contener más información sobre el número de matriculaciones. Cuando se aplica la metodología de ir añadiendo variables iterativamente en el modelo, las más correlacionadas son las primeras de deben incluirse.

Como hemos indicado ya, se aprecia un **cambio de comportamiento** a partir del año 2008, lo que puede asociarse al inicio de la crisis económica más que a otras causas, por ejemplo un posible programa estatal de ayuda a la compra de automóviles, que por otra parte ha habido de forma casi ininterrumpida. La primera observación visual de las variables sugiere que, de entre las variables de la DGT, la que muestra un patrón más proporcional o semejante al de Y (número de matriculaciones) es $XL1$ (permisos de conducir). La correlación lineal entre ambas es +0.62, y aunque consideraremos esta posibilidad al final de nuestro análisis, en principio vamos a considerar que la diferencia de comportamiento no parece ser registrada por ninguna variable con suficiente claridad como para intentar modelar todos los años a la vez con un modelo lineal (tiene interés construir el modelo como si no conociésemos esa variable). Respecto a las variables de Eurostat, casi todas muestran una **tendencia lineal** similar común a la de Y durante los años hasta el 2008. Por tratarse de datos anuales, los datos no pueden mostrar **estacionalidad**. Podrían apreciarse, si tuviésemos más datos, patrones relacionados con ciclos económicos de menor frecuencia (o mayor periodo). No se aprecian **datos atípicos** ni **erróneos** que puedan alterar las estimaciones minimocuadráticas.

Tanto el número de matriculaciones como las variables explicativas que consideramos dependen de la misma actividad económica subyacente. Esto implica que el **significado de nuestras variables** indica que nuestro objetivo es **explicar y describir** el número de matriculaciones en función de otras variables, sin intención de atribuir una relación de **causalidad**. (No se trata del tipo de relación que podría existir, por ejemplo entre número de coches fabricados y su efecto parcial sobre el número de neumáticos nuevos que se fabrican o importan.) De hecho, el modelo de regresión lineal permite estudiar la relación lineal entre las variables pero no establecer causalidad

alguna. La causalidad debe ser estudiada a partir del significado de las variables. A esto se añade que el patrón de tendencia de las variables se parece al de la variable *tiempo*, lo que en conjunto nos hace pensar en un posible problema de **correlación espuria**. Este fenómeno debe ser descrito por el significado de las variables, no por los valores numéricos de los datos (unos mismos valores podrían corresponder a variables con significados distintos). Esto puede mostrarse con el siguiente ejemplo sencillo:

$$\text{Si} \quad Y = 2t \quad \text{y} \quad X = 3t\,, \quad \text{se cumple que} \quad Y = \frac{2}{3} X.$$

Con frecuencia *t* es el tiempo. En este caso, el simple paso del tiempo podría hacernos pensar (si sólo viésemos la última relación) que hay una relación causal entre las variables. La relación causal es falsa, y lo que realmente sucede es que tanto la variable explicada como las explicativas tienen una causa común. Por tanto, la correlación espuria no es un problema técnico sino un problema de **interpretación** que consiste en asociar erróneamente relación lineal (verdadera) con causalidad (falsa). Esto no invalida el uso del modelo de regresión lineal con **carácter descriptivo**. (Si se añade tendencia a dos variables independientes, los análisis pueden mostrar la presencia de correlación, como se indica en la sección 14.3 del libro de Novales, donde se indican varias formas de estudiar variables con tendencia.)

La tendencia común que muestran algunas variables, y su propio significado, indican que tendríamos un problema de **colinealidad** si intentásemos añadirlas en la misma regresión, puesto que están aportanto la misma información sobre la variable explicada. Esto podría apreciarse tanto por los conocidos síntomas del problema de colinealidad (consúltese alguno de los textos que incluimos en las referencias) como por el que, una vez introducida una de ellas y quitado su efecto a todas las demás variables, las variables muy correlacionadas con la introducida darían lugar a unos residuos que no superarían la prueba para ser admitidas en el modelo. Si la regresión incluye un término independiente β_0, puede haber un problema de colinealidad si alguna variable explicativa es casi constante (también en la regresión simple). Por otro lado, no parece haber ningún problema de **heterocedasticidad**, lo que debe confirmarse al representar la variable explicada frente a cada variable explicativa, en los residuos de las regresiones o por otros síntomas bien conocidos de este problema (que el lector también puede consultar en la literatura).

Merece la pena hacer algunas observaciones sobre las magnitudes y las unidades de medida de las variables. Para el modelo de regresión lineal

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+u, \qquad (1)$$

La necesaria **homogeneidad dimensional** implica que:

- Las cantidades β_0 y u tienen las mismas unidades que la variable explicada Y.
- Cada cantidad β_j está media en las unidades de la variable explicada Y divididas por las unidades de la variable explicativa X_j.

Cuando los datos son temporales puede suceder que algunas unidades (que son referencia para otras) se vean afectadas por el paso del tiempo, lo que introduciría un cambio en las variables dependientes de esas unidades que no se debe al cambio de sus propios valores. Esto puede ejemplificarse analíticamente de la siguiente forma:

$$\text{Si} \quad Y=2X \quad \text{y} \quad X_t=3t, \quad \text{se cumple que} \quad Y(t)=2X_t.$$

Sin embargo, mientras que en la primera relación entre Y depende únicamente de X, en la segunda $Y(t)$ refleja los cambios propios de su dependencia de X (el coeficiente 2) más los cambios que X experimenta con el tiempo. Para que la parte de la variación de las variables económicas que se debe a la **inflación** no se interprete como cambios reales del valor de las variables, con frecuencia se trabaja con la **versión deflactada** o **a precios constantes** (tomando como referencia los valores de un determinado año) de las variables económicas. En nuestro caso, no está claro que la inflación no explique parte del comportamiento del número de matriculaciones, aunque sea de forma indirecta, por ejemplo a través del PIB per cápita nominal. Por este motivo consideramos inicialmente varias versiones de producto interior bruto y dejamos que los datos nos indiquen cuál explica mejor la variable explicativa durante el periodo que consideramos.

En resumen, en nuestros datos identificamos posibles problemas relacionados con un cambio estructural, correlación espuria y colinealidad.

Modelo de regresión lineal

Declaración

El modelo de regresión lineal queda determinado por una conjunto de hipótesis y por la siguiente **expresión matemática**

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+u, \qquad [1]$$

donde Y es la variable explicada, β_j son los parámetros, X_j son las variables explicativas y u es el término de error. Cuando sólo hay una variable explicativa (k=1), el modelo es *simple*, y si no es *múltiple*. El número de variables, y cuáles son, tiene consecuencias a la hora de interpretar los coeficientes β_j y en los casos en los que las variables X_j estén muy correlacionadas entre ellas. Las **hipótesis clásicas** de este modelo son:

[H1] Sobre X_j

[H1-1] Determinismo (no aleatoriedad): Es frecuente suponer que las variables explicativas no son aleatorias, es decir, que son deterministas. Esta suposición facilita la demostración de resultados que con frecuencia serían también válidos sin las variables fuesen aleatorias.

[H1-2] Unicidad (no colinealidad): Uno de los problemas que pueden aparecer en el modelo de regresión lineal es debido a la presencia en el modelo de variables explicativas altamente correlacionadas. Aunque sus consecuencias son graves, es un problema relativamente fácil de identificar: estudiando la relación entre las variables explicativas o los posibles síntomas en el ajuste del modelo.

[H2] Sobre Y

[H2-1] Dependencia unidireccional: Un modelo de regresión puede ser aplicado cuando las variables explicativas influyen en la variable explicada, pero no sucede lo contrario (el menos no en un alto grado). Cuando esto último sucede, habría que considerar otros modelos distintos, con ecuaciones simultáneas, por ejemplo. El significado de las variables es el principal indicador de cuáles son las relaciones de causalidad.

[H2-2] Aleatoriedad: Dado que al menos se utiliza un término de error aleatorio en el modelo, la variable explicada es también aleatoria independientemente del carácter de las variables explicativas.

[H2-3] Normalidad: Cuando el error es normalmente distribuido y las variables explicativas son deterministas, es fácil deducir (utilizando los resultados para suma de variables normales) cuál debería ser la distribución teórica de la variable explicada. Cuando las variables explicativas son estocásticas, also

similar puede deducirse fácilmente si son normalmente distribuidas, pero no si siguen otras distribuciones.

[H3] Sobre la relación funcional

[H3-1] Linealidad en los parámetros: Aunque siempre es posible redefinir los parámetros o las variables, es esta linealidad de los parámetros, no la de las variables explicativas, el motivo por el que se habla de *regresión lineal*, ya que los parámetros son las incógnitas al ajustar el modelo cuando los datos están dados. Una regresión no lineal sería, por ejemplo, $Y = \beta_0 + \beta_1^2 X + u$.

[H3-2] Homogeneidad estructural: Esta hipótesis supone que los valores de coeficientes son independientes de los valores que las variables explicativas y explicada puedan tomar, es decir, que son realmente parámetros y no variables (incluso aunque pueda tomar pocos valores). Esto implica que la misma fórmula matemática es válida para todo el rango de valores y que no hace falta considerar varias. Esta hipótesis se evalúa en la fase descriptiva de los datos (si es posible) o comparando varios modelos lineales; un método a medio camino entre los dos anteriores consiste en calcular el valor de la variable explicada para distintos rangos o estratos formados a partir de la variable explicativa «sospechosa» de tener un coeficiente no constante.

[H3-3] Linealidad en las variables: La representación de los datos suele indicar hasta qué punto es razonable suponer dependencias lineales entre las variables. Parte de la importancia del modelo lineal consiste no sólo en que suele ser una buena aproximación para muchas relaciones entre variables, sino en que con frecuencia es posible transformar las variables para seguir aplicando el modelo lineal. Esto debe tenerse en cuenta en la interpretación, en el sentido de que en $Y = \beta_0 + \beta_1 log(X) + u$ no debe olvidarse que β_1 es el coeficiente de $log(X)$, no de X.

[H4] Sobre el término de error

[H4-1] Aleatoridad: Todos los errores no aleatorios que pueden estimarse, se suponen ya corregidos en los pasos anteriores a la aplicación del modelo, por lo que los errores aleatorios son los únicos que hay que tener en cuenta en el modelo.

[H4-2] Media nula: También por lo dicho en el apartado anterior, cualquier error de media no nula se supone ya corregido antes de aplicar el modelo.

[H4-3] Homocedasticidad: También implica el que una única expresión como $Y = \beta_0 + \beta_1 X + u$ no sea válida, pero a diferencia de la homogeneidad estructural, esta hipótesis se viola cuando el término de error (no los coeficientes) no tienen una estructura constante, concretamente cuando la varianza σ^2 no es constante. En este caso de habla de heterocedasticidad, y suele ser fácil de identificar al representar los datos o mediante contrastes de hipótesis más formales.

[H4-4] Normalidad: Esta es una hipótesis usual en teoría de errores. Dado que esta distribución es bien conocida teóricamente, es fácil de evaluar su incumplimiento, por ejemplo a partir de la asimetría, el exceso de curtosis, un gráfico de cuantiles o contrastes de bondad de ajuste específicos (y por tanto más potentes) para el caso de la normalidad.

[H4-5] Autoincorrelación: El término de error cuando las variables toman ciertos valores se supone que no se ve afectado por los valores que pueda tomar cuando las variables tomas otros valores. Este hecho suele comprobarse gráficamente o aplicando contrastes de hipótesis bien conocidos.

Ajuste

El ajuste del modelo de regresión lineal, es decir, la selección de los parámetros, se basa en condiciones normales en la aplicación del **método de mínimos cuadrados**. Como su nombre indica, consiste en minimizar la suma de los cuadrados de los residuos, habitualmente denominada *suma residual*. Los residuos son las distancias entre cada valor de la variable explicada y el que proporciona la parte funcional del modelo (expresión sin el término de error). Es decir, se trata de seleccionar los valores de los coeficientes β_j (que ahora son vistos temporalmente como variables) que resuelven el problema

$$min\,\Psi(\beta_0,\beta_1,\ldots,\beta_k)=min\sum_{i=1}^{n}[Y^{(i)}-\hat{Y}^{(i)}]^2. \qquad (2)$$

El estimador de mínimos cuadrados ordinarios se comprende mejor pensando en la interpretación geométrica en el caso de la regresión lineal simple: el método elige, de entre todas las posibles rectas que cruzan la nube de puntos (datos), la recta que minimiza la suma de todas las distancias de cada punto a su proyección vertical sobre la recta.

Por otra parte, no incluimos aquí las expresiones que se obtienen como solución del anterior problema de optimización ni sus características más notables, que pueden encontrarse en los libros incluidos en las referencias. También en la literatura puede encontrarse información sobre métodos de estimación adecuados cuando hay varias medidas de la variable explicada para cada valor de las variables explicativas o cuando no se cumplen algunas hipótesis (el estimador de Newey-West se puede aplicar aunque haya heterocedasticidad o autocorrelación, por ejemplo).

Diagnosis

Para validar el ajuste de un modelo de regresión lineal, simple o múltiple, suele considerarse la información de varias medidas de distintos tipos:

Ajuste global: Las medidas más importantes son el error estándar de regresión (EER), coeficiente de determinación R^2 y la significatividad global del modelo (valor *p*). Para comparar modelos con la misma variable explicada *Y* y datos, puede utilizarse el EER, y el modelo es tanto mejor cuanto menor es esta medida. Para comparar la bondad de ajustes en los que las variables y datos pueden ser distintos, se debe utilizar el R^2 (el R^2 ajustado si hay varias variables explicativas), y el modelo es tanto mejor cuanto más cercano a uno es esta medida. Este coeficiente es adimensional e indica, en tanto por uno, la proporción de la variabilidad de *Y* que explica el modelo ajustado. En nuestro caso, podemos utilizar cualquier de estas dos medidas. (Montgomery y otros advierten, en la sección 2.8, de los problemas de utilizar el R^2 cuando se comparan las regresiones con y sin término independiente β_0.) Por último, el valor *p* será fiable cuando no haya violaciones claras de las hipótesis teóricas del modelo y cuando los datos tengan la calidad necesaria para asegurar la precisión de las estimaciones. Cuando el valor *p* es menor que 0.05 (5%), por ejemplo, se puede rechazar la hipótesis nula de que todos los coeficientes de las variables explicativas son nulos a la vez, es decir, de que el modelo es inútil para explicar la variable *Y*.

Coeficientes: Respecto a la información de las estimaciones de los coeficientes, es importante observar para cada uno la precisión y, después, la significatividad. Esta última se evalúa a partir del valor del estadístico *t* en relación con los valores críticos (que determinan la región de rechazo) o, equivalentemente, el valor *p*. El valor *p* de cada contraste individual informa del apoyo que tiene la hipótesis nula H_0: β_j=0. Como criterio usual, suele considerarse que si este valor es menor que 0.05 (5%), se rechaza la hipótesis nula y el coeficiente será significativamente distinto de 0, es decir, ese término parece útil para explicar la variable *Y*. Cuando hay una violación clara de las hipótesis (por ejemplo, heterocedasticidad, autocorrelación o colinealidad), las conclusiones sobre los coeficientes no son fiables porque no lo son las estimaciones ni los contrastes de hipótesis. (El ajuste de un modelo sin término independiente β_0 presenta algunas características propias, como indican Montgomery y otros en la sección 2.8, por lo que si no se rechaza la hipótesis H_0: β_j=0 podría ser conveniente considerar el ajuste del modelo sin este término. En la sección B1 del apéndice se indica cómo hacer esto en el lenguage R.)

Análisis de los residuos: Los residuos aportan información importante tanto sobre la información de *Y* que no ha sido explicada por la parte funcional del modelo (expresión menos el término de error) como del posible incumplimiento de alguna de las hipótesis, además de sobre la presencia de datos atípicos. En nuestro caso, hemos implementado diversos métodos formales (los profesionales no necesitan tanta

información), con la intención de aprender lo máximo posible. Entre las características que pueden tenerse en cuenta están: la media, la mediana, el coeficiente de asimetría, el exceso de curtosis, el histograma, distintos gráficos de los residuos, contrastes de aleatoriedad, contrastes de normalidad y contrastes de autocorrelación.

Aplicación

Hay distintas posibles aplicaciones de un modelo de regresión lineal, según el uso que se le esté dando al modelo. Ya hemos mencionado en un ejercicio su uso para estimar y sustraer (al considerar los residuos en adelante) la posible tendencia lineal de unos datos.

Una primera aplicación puede ser la de sustituir en el modelo los valores de las variables explicativas y considerar el valor proporcionado (*valor ajustado*) como una corrección del error del valor anterior de la variable explicada Y. Los coeficientes indican cuánto varía la variable Y, medido en sus unidades, cuando una única variable explicativa varía en una de sus unidades mientras que las demás permanecen constantes (a esta situación se la denomina *ceteris paribus*, y aunque tiene sentido matemático, no lo tiene en la práctica cuando las variables explicativas no son independientes y por tanto al variar una deberían variar también otras).

Los coeficientes también informan de los efectos, totales para las regresiones simples y directos o parciales para las múltiples, que cada variable explicativa tiene en la explicada. Dado que los coeficientes tienen unidades, la magnitud de la estimación de su coeficiente no es siempre una buena medida de la importancia de su correspondiente variable explicativa.

El modelo obtenido puede en ocasiones utilizarse para predecir valores de Y, interpolando o extrapolando valores (no muy lejanos) de las variables explicativas. En nuestro caso, y como mostramos más adelante, el modelo no debe ser utilizado con este propósito.

Para aplicaciones más avanzadas, como intervalos de confianza o contrastes de hipótesis, consúltense libros avanzados sobre regresión lineal.

Modelización

El probable **cambio estructural** que hemos mencionado en secciones anteriores nos sugiere utilizar primero los datos correspondientes a los años desde 1995 a 2008. Después utilizaremos una **variable ficticia** para evaluar si realmente se puede considerar que hay un cambio estructural en ese año. Además, intentaremos explicar el número de matriculaciones utilizando **todos los datos**.

Selección de variables

Respecto a la **selección de variables** en regresión, hoy día el ordenador permitiría considerar automáticamente **todas las posibles regresiones** (con una variable explicativa, dos, etc.) y elegir las que verifican algún criterio adecuado. Sin embargo, con frecuencia se adopta una selección por pasos, en los que **iterativamente se van introduciendo o desechando variables**. Aunque hay criterios formales para decidir cuándo incluir o excluir variables (véase, por ejemplo, la sección 9.2.2 del libro de Montgomery y otros), nosotros haremos esto basándonos en nuestra percepción de la calidad de los ajustes. Debido al escaso número de datos respecto al de variables, vamos a aplicar la primera metodología para no trabajar con pocos grados de libertad, lo que aumentaría el riesgo de sobreajuste.

Datos de 1995 a 2008

Hemos representado los datos para conocer sus características y anticipar algunos de los posibles problemas característicos de este tipo de modelos.

> EJERCICIO: *El lector puede seleccionar los datos de estos años utilizando el código de la sección A3, que, para cada par de variables, también representa un gráfico de dispersión y calcula la correlación.*

Para estos años, el gráfico de dispersión de `Y` (número de matriculaciones) con `XL2` (cambios de titularidad) es el que parece indicar una relación lineal más clara, aunque también con `XL4` (censo de conductores). Éstas son las variables más correlacionadas linealmente (+0.95 y +0.94, respectivamente) con `Y` (número de matriculaciones).

Respecto a los datos de Eurostat, el gráfico de dispersión de `Y` (número de matriculaciones) con `XG6` (índice de Gini) parece indicar una relación lineal (de pendiente negativa, como era de esperar) más clara que las demás, aunque ahora las demás variables también sugieren una relación lineal, un poco mejor para `XG2r` (producto interior bruto per cápita real). Esta última variable es la más correlacionada (positivamente, como era de esperar) linealmente con `Y` (+0.96), e incluso hay para estos años variables más correlacionadas que `XG6` (índice de Gini).

No obstante, como ya mencionamos, no conviene fiarse totalmente de los gráficos de dispersión de cada par de variables.

EJERCICIO: *El lector puede utilizar el código de la sección B3 del apéndice para evaluar la no* ***aleatoriedad*** *de las variables explicativas.*

Los modelos lineales más complejos (en términos del número de parámetros) que se pueden formar con los datos serían

$$Y=\beta_0+\beta_1 X_{L1}+\beta_2 X_{L2}+\beta_3 X_{L3}+\beta_4 X_{L4}+u, \tag{3}$$

$$Y=\beta_0+\beta_1 X_{G1}+\beta_2 X_{G2n}+\beta_3 X_{G2r}+\beta_4 X_{G3}+\beta_5 X_{G4}+\beta_6 X_{G5}+\beta_7 X_{G6}+u \tag{4}$$

y

$$\begin{aligned} Y=&\beta_0+\beta_1 X_{L1}+\beta_2 X_{L2}+\beta_3 X_{L3}+\beta_4 X_{L4} \\ &+\beta_5 X_{G1}++\beta_6 X_{G2n}+\beta_7 X_{G2r}+\beta_8 X_{G3}+\beta_9 X_{G4}+\beta_{10} X_{G5}+\beta_{11} X_{G5}+u \end{aligned} \tag{5}$$

Sin embargo, el ajuste de estos modelos indica que no son válidos para estas variables, incluso aunque en los dos últimos quitásemos algunas variables para evitar la colinealidad. Por otro lado, en el último caso el número de datos es $n = 13$ mientras que el número de variables explicativas sería $k = 11$, lo que obviamente supone un problema de falta de grados de libertad y, por tanto, de riesgo de sobreajuste.

EJERCICIO: *El lector podría intentar ajustar estos modelos que utilizan* ***todas las variables****, de cada base de datos y de la unión de ambas, evaluando*

```
regLin = lm(Y ~ XL1 + XL2 + XL3 + XL4)
regLin = lm(Y ~ XG1 + XG2n + XG2r + XG3 + XG4 + XG5 + XG6)
regLin = lm(Y ~ XL1 + XL2 + XL3 + XL4
            + XG1 + XG2n + XG2r + XG3 + XG4 + XG5 + XG6)
```

respectivamente, y el código de la sección A6 del apéndice.

Datos de la Dirección General de Tráfico (DGT)

Consideramos una posible forma de explicar el número de matriculaciones a partir de algunas variables macroeconómicas registradas junto con el número de matriculaciones en la Dirección General de Tráfico. La representación de los datos parece indicar que *XL2* y *XL4* son las que tienen más información sobre `Y` (al considerar todos los datos, la más correlacionada es `XL1_0`, el número de permisos de conducir). Es evidente que ambas comparten información entre ellas. Sin embargo, el modelo con estas dos variables no es aceptable.

EJERCICIO: *El lector puede comprobar la afirmación anterior considerando*

$$Y=\beta_0+\beta_1 X_{L2}+\beta_2 X_{L4}u \qquad \texttt{regLin = lm(Y ~ XL2 + XL4)}$$

Deducimos que el problema es la colinealidad, dado que `cor(XL2, XL4)=0.99`. En vez de considerar una de estas dos variables junto con las dos restantes, para después suprimir alguna, vamos a empezar con `XL4` y ver cuánto mejora el modelo al añadir `XL1` o `XL3` (otra opción más técnica es quitar al efecto de `XL4` a todas las variables antes de elegir la que tiene más información aún no explicada, considerando un valor mínimo de entrada para el estadístico *F* como sugiere la teoría). Elegimos `XL4` y no `XL2` porque sus correlaciones con `Y` son prácticamente iguales mientras que al estudiar todos los datos nos pareció que la primera sigue un patrón más parecido, lo que podría hacer que el modelo se extrapolase mejor a los años de la crisis. Podemos comprobar que

$$Y=\beta_0+\beta_1 X_{L4}+u \qquad \texttt{regLin = lm(Y ~ XL4)}$$

explica casi el 88% de la variabilidad del número de matriculaciones. La estimación de los coeficientes no tiene problemas de precisión y son significativamente distintos de cero. El análisis de los residuos muestra resultados suficientemente satisfactorios,

aunque su correlación con la variable explicada es `cor(regLin$residuals, Y)=0.33`. Si quitamos el efecto de `XL4` a las demás variables

```
reg1 = lm(Y ~ XL4);   reg2 = lm(XL1 ~ XL4);   reg3 = lm(XL3 ~ XL4)
Ytilde = reg1$residuals;  XL1tilde = reg2$res; XL3tilde = reg3$res
cor(cbind(Ytilde, XL1tilde, XL3tilde))
```

vemos que las nuevas versiones de `XL1` o `XL3` muestran prácticamente la misma correlación lineal **parcial** con la nueva versión de `Y`, es decir, parecen contener la misma información sobre `Y` no explicada ya por `XL4`. No calculamos los estadísticos *F* que indica la teoría para ver cuál entraría en el modelo, sino que probamos las dos regresiones

$$Y=\beta_0+\beta_1 X_{L4}+\beta_2 X_{L1}+u \quad \text{y} \quad Y=\beta_0+\beta_1 X_{L4}+\beta_2 X_{L3}+u$$

`regLin = lm(Y ~ XL4 + XL1)` y `regLin = lm(Y ~ XL4 + XL3)`

y vemos que el primer modelo se acepta mientras que para el segundo β_2 no es significativamente distinto de cero. El aumento de R^2 es sólo de aproximadamente un 2%, y la correlación de los residuos con `Y` se reduce sólo a `0.30`, por lo que no merece la pena complicar el modelo añadiendo una variable más. En resumen, a partir de las variables y datos considerados, decidimos quedarnos con el **modelo**

$$Y=\beta_0+\beta_1 X_{L4}+u \quad \text{cuya estimación es} \quad \hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X_{L4}=-2124000+0.2028\cdot X_{L4} \qquad (6)$$

Esta relación indica que cuando el censo de conductores se incrementó en cinco unidades, el número de matriculaciones lo hizo en una unidad aproximadamente.

Datos de Eurostat

Dado que hay ahora muchas variables explicativas y varias están correlacionadas entre sí, tampoco empezamos incluyendo todas las variables sino que directamente empezamos considerando la que tiene más correlación lineal total con `Y`. Vimos que era `XG2r`, por lo que empezamos considerando

$$Y=\beta_0+\beta_1 X_{G2r}+u \qquad \texttt{regLin = lm(Y ~ XG2r)}$$

Este modelo es globalmente aceptado y explica el 92% de la variabilidad de *Y*. No hay problemas de precisión en la estimación de los coeficientes, y parecen significativamente distintos de cero. El análisis de los residuos muestra resultados suficientemente satisfactorios, y la correlación de los residuos con la variable explicada es `cor(regLin$residuals, Y)=0.28`. La matriz `cor(matrizG)` indica que esta variable explicativa es la más correlacionada, pero esta diferencia es aún mayor con respecto a las otras versiones del producto interior bruto cuando se tiene en cuenta `cor(matrizG_0)`, lo que podría hacernos pensar que es la mejor variable para que el modelo se adapte a los años de crisis.

EJERCICIO: *El lector puede comprobar que quitar la tendencia implica eliminar prácticamente toda la información:*

$$\widetilde{Y}=\beta_0+\beta_1 \widetilde{X_{G2r}}+u$$

*Para suprimir la **tendencia**, las técnicas más conocidas son la de medias móviles y la de ajustar una regresión sobre el índice para quedarse con el residuo, que puede interpretarse como una transformación de la variable explicada a la que se le ha extraído el efecto de las variables explicativas que se hayan considerado (el tiempo t, en este caso).*

```
reg1 = lm(Y ~ t);      reg2 = lm(XG2r ~ t)
Ytilde = reg1$residuals;      XG2rtilde = reg2$residuals
regLin = lm(Ytilde ~ XG2rtilde)
```

Ahora quitamos el efecto de esta variable a todas las demás y tenemos en cuenta la que tiene mayor correlación parcial con la nueva versión de la variable explicada.

```
reg1 = lm(Y ~ XG2r); reg2 = lm(XG1 ~ XG2r); reg3 = lm(XG2n ~ XG2r)
reg4 = lm(XG3 ~ XG2r); reg5 = lm(XG4 ~ XG2r); reg6 = lm(XG5 ~ XG2r)
reg7 = lm(XG6 ~ XG2r);      Ytilde = reg1$residuals
XG1tilde = reg2$residuals;    XG2ntilde = reg3$residuals
XG3tilde = reg4$residuals;    XG4tilde = reg5$residuals
XG5tilde = reg6$residuals;    XG6tilde = reg7$residuals
cor(cbind(Ytilde, XG1tilde, XG2ntilde, XG3tilde,
          XG4tilde, XG5tilde, XG6tilde))
```

y vemos que, en contra de lo esperado, ahora `XG6tilde` (nueva versión del índice de Gini) no es la que más información sobre `Ytilde` tiene, sino que tiene más

`XG5tilde` (nueva versión del umbral de riesto de pobreza). Concretamente, se tiene que

`cor(Ytilde, XG5tilde)=-0.53` y `cor(Ytilde, XG6tilde)=+0.48`

así que consideramos el modelo

$$Y=\beta_0+\beta_1 X_{G2r}+\beta_2 X_{G5}+u \qquad \texttt{regLin = lm(Y ~ XG2r + XG5)}$$

con el que apenas aumenta un 2% el coeficiente R^2, añadimos complejidad y además podemos incurrir en un problema de colinealidad, dado que `cor(XG2r, XG5)=+0.98`. (Por otro lado, `cor(XG2r, XG6)=-0.82`) En resumen, a partir de las variables y datos considerados, nos quedamos con el **modelo**

$$Y=\beta_0+\beta_1 X_{G2r}+u \quad \text{cuya estimación es} \quad \hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X_{G2r}=-2165000+208.2\, X_{G2r} \qquad (7)$$

Por un incremento de un euro por habitante en el producto interior bruto real, el número de matriculaciones aumentó en 208 vehículos aproximadamente.

EJERCICIO: *Aparte de utilizando la información proporcionada por los gráficos y la matriz de correlaciones, el lector puede comprobar y estudiar el problema de* ***colinealidad*** *al considerar, por ejemplo, las regresiones*

$$X_{G1}=\beta_0+\beta_1 X_{G2n}+u \qquad Y=\beta_0+\beta_1 X_{G1}+\beta_2 X_{G2n}+u$$

$$Y=\beta_0+\beta_1 X_{Gj}+\beta_2 t+u\,, \quad j=1,2$$

`regLin = lm(XG1 ~ XG2n)` `regLin = lm(Y ~ XG1 + XG2n)` `regLin = lm(Y ~ XGj + t)`

Datos de la DGT y de Eurostat

Finalmente consideramos el modelo que podría construir un observador que tuviese acceso a todas las variables anteriores. Empezamos considerando las variables explicativas aceptadas en los modelos anteriores. Vamos a representarlas y estudiar su correlación.

```
matrizLG = cbind(Y, XL4, XG2r);  pairs(matrizLG);   cor(matrizLG)
```

dado que `cor(XL4, XG2r)=+0.98`, tenemos que seleccionar una de las dos variables para no incurrir en un problema de colinealidad en el modelo. El producto interior bruto per cápita está ligeramente más correlacionada con el número de

matriculaciones, por lo que de momento empezamos con el modelo $Y=\beta_0+\beta_1 X_{G2r}+u$.

Para intentar mejorar el modelo, y dado que ya probamos a quitarle el efecto de esta variable explicativa a las demás variables de su base de datos, hacemos ahora lo mismo con las de la base de datos de la DGT.

```
reg1 = lm(Y ~ XG2r);  reg2 = lm(XL1 ~ XG2r);  reg3 = lm(XL2 ~ XG2r)
reg4 = lm(XL3 ~ XG2r);  reg5 = lm(XL4 ~ XG2r)
Ytilde = reg1$residuals
XL1tilde = reg2$residuals;    XL2tilde = reg3$residuals
XL3tilde = reg4$residuals;   XL4tilde = reg5$residuals
cor(cbind(Ytilde, XL1tilde, XL2tilde, XL3tilde, XL4tilde))
```

Esta matriz indica que las variables `XL2` (cambios de titularidad) es la que más información nueva aportaría. Observamos

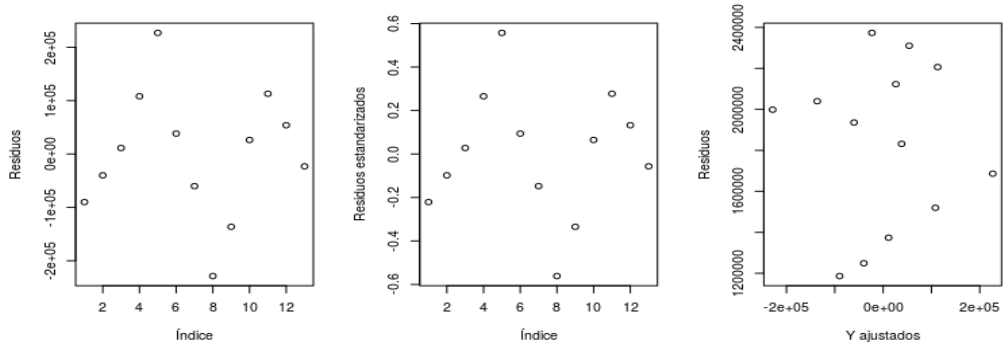
```
matriz = cbind(Y, XG2r, XL2);   pairs(matriz);   cor(matriz)
```

y como `cor(XG2r, XL2)=0.95` no probamos siquiera a añadirla. En definitiva, nos quedamos con el **modelo** $Y=\beta_0+\beta_1 X_{G2r}+u$, que ya hemos interpretado en la sección anterior. Mostramos ahora los principales resultados:

```
Residuals:
    Min      1Q  Median      3Q     Max
-228609  -60110   11290   53953  226933

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.165e+06  3.565e+05  -6.072 8.05e-05 ***
XG2r         2.082e+02  1.848e+01  11.267 2.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122800 on 11 degrees of freedom
Multiple R-squared:  0.9203,  Adjusted R-squared:  0.913
F-statistic: 126.9 on 1 and 11 DF,  p-value: 2.218e-07
```



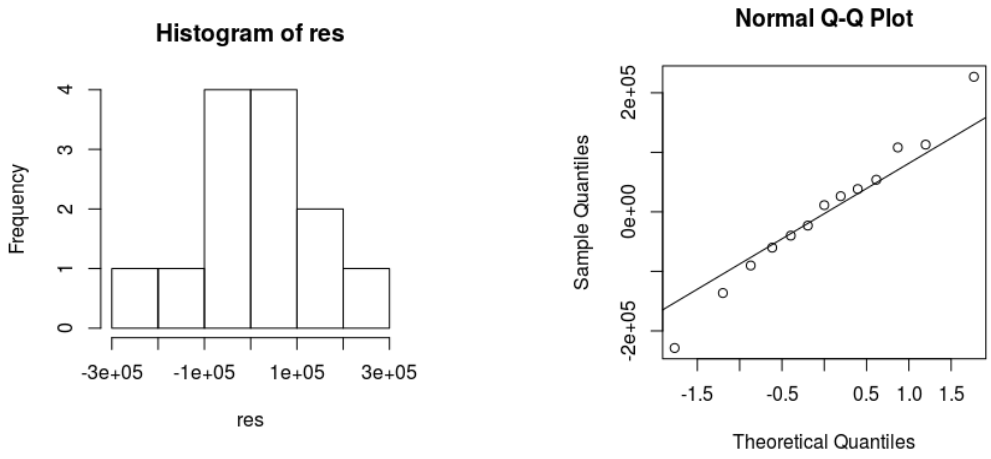


Figura 4: Gráficos para los residuos.

El lector puede aplicar los contrastes de aleatoriedad, normalidad y autocorrelación.

Datos de 1995 a 2013

El problema de modelizar todos los datos es más complicado, dado que sólo una variable parecía haber registrado claramente el mismo patrón de cambio producido en el número de matriculaciones por la crisis. Seguir los mismos pasos de antes puede conducir a resultados distintos.

El valor de las matriculaciones desde 2008 fue `Y_0[14:19]`,

```
1651013  1258781  1298809  1091511  924310  949015
```

mientras que el modelo antes calculado $\hat{Y} = -2165000 + 208.2\, X_{G2r}$ prediría, evaluando el código `-2165000+208.2*XG2r_0[14:19]`, los valores

```
2352940  2144740  2123920  2123920  2040640  2019820
```

Gráficamente, con el siguiene código representamos el número de matriculaciones real y el predicho (la regresión no muestra una línea porque los valores de la variable explicativa no están equidistribuidos a lo largo del eje de abscisas)

```
plot(Y_0, main='Modelo (en azul) de 1995-2008 aplicado a 1995-2013',
     xlab='PIB per cápita real', ylab='Matriculaciones',
     type='l', xaxt='n')
axis(1,at=c(1,13,19),
     labels=c('XG2r(1995)','XG2r(2008)','XG2r(2013)'))
lines(-2165000 + 208.2*XG2r_0, lwd=3, col='blue')
```

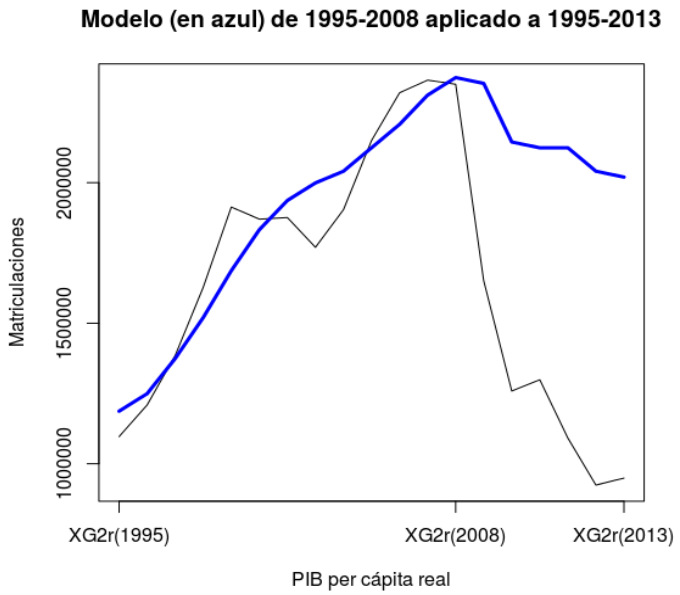


Figura 5: Número de matriculaciones reales y estimadas.

Homogeneidad estructural

Formalmente, podemos definir una variable ficticia

```
# Variable ficticia  [UNIDAD: ]
F = c(rep(0, 13), rep(1,6))
```

que, utilizada en el modelo

$$Y_0 = \beta_0 + \beta_1 X_{G2,0} + \beta_2 F + \beta_3 F \cdot X_{G2,0} + u$$ `regLin = lm(Y_0 ~ XG2r_0 + F + F*XG2r_0)`

cuando toma los valores 0 y 1 lleva respectivamente a

$$Y_0 = \beta_0 + \beta_1 X_{G2,0} + u$$ para los años de 1995 a 2008

$$Y_0 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{G2,0} + u$$ para los años de 2009 a 2013

Es decir, permite evaluar si para los últimos años parece significante una **recta de regresión con distintos parámetros** (los dos) a la obtenida anteriormente. El lector puede comprobar que los coeficientes β_2 y β_3 son significativos, es decir, que parece que a partir del año 2008 habría que considerar una recta con un término independiente y una pendiente distintas. (Nótese, sin embargo, que sólo se dispone de cinco datos posteriores a 2008, mientras que estamos estimando tres parámetros.)

Modelo con el efecto de la crisis

A la vista de la evolución temporal de todas las variables, podemos pensar en que las variables `XL1_0` (permisos de conducir) y `XG6_0` (índice de Gini) han podido registrar parte del cambio brusco de tendencia del número de matriculaciones que se produjo con la crisis. Estas son las variables que mostraron mayor correlación con `Y_0`. Estudiamos las nuevas variables

```
matrizLG_0 = cbind(Y_0, XL1_0, XG6_0);   pairs(matrizLG_0);   cor(matrizLG_0)
```

y ajustamos la regresión con

$$Y_0 = \beta_0 + \beta_1 X_{L1,0} + \beta_2 X_{G6,0} + u \qquad \texttt{regLin = lm(Y\_0 ~ XL1\_0 + XG6\_0)}$$

(si utiliza el código de la sección A6 del apéndice, ahora el lector debe escribir `Y_0` donde allí pone `Y`) lo que, dado que el coeficiente β_1 no es significativamente distinto de cero, hace pensar en la regresión simple

$$Y_0 = \beta_0 + \beta_1 X_{G6,0} + u \qquad \texttt{regLin = lm(Y\_0 ~ XG6\_0)}$$

como mejor modelo para explicar la relación lineal entre el número de matriculaciones a partir de las variables consideradas para los años desde 1995 a 2013. Este modelo explica algo más del 50% de la variabilidad del número de matriculaciones. Hay suficiente precisión al estimar los coeficientes, que parecen distintos de cero. El análisis de los residuos es satisfactorio, aunque `cor(res, Y_0)=0.68` es demasiado alta, lo que se interpreta como que los residuos todavía tienen demasiada información sobre la variable explicativa.

Para intentar mejorar el modelo, probamos a quitar el efecto de esta variable a las demás para calcular las correlaciones parciales, que nos indican cuál es la siguiente

variable que podría aportar información distinta a la de XG6_0.

```
reg1 = lm(Y_0~XG6_0); reg2 = lm(XL1_0~XG6_0); reg3 = lm(XL2_0~XG6_0)
reg4 = lm(XL3_0~XG6_0); reg5 = lm(XL4_0~XG6_0); reg6 = lm(XG1_0~XG6_0)
reg7 = lm(XG2n_0~XG6_0); reg8 = lm(XG2r_0~XG6_0); reg9 = lm(XG3_0~XG6_0)
reg10 = lm(XG4_0 ~ XG6_0);  reg11 = lm(XG5_0 ~ XG6_0)
Y_0tilde = reg1$res;  XL1_0tilde = reg2$res;  XL2_0tilde = reg3$res
XL3_0tilde = reg4$res;  XL4_0tilde = reg5$res;  XG1_0tilde = reg6$res
XG2n_0tilde = reg7$res; XG2r_0tilde = reg8$res; XG3_0tilde = reg9$res
XG4_0tilde = reg10$res;  XG5_0tilde = reg11$res
cor(cbind(Y_0tilde, XL1_0tilde, XL2_0tilde, XL3_0tilde, XL4_0tilde,
          XG1_0tilde, XG2n_0tilde, XG2r_0tilde,
          XG3_0tilde, XG4_0tilde, XG5_0tilde))
```

Esta matriz indica que las variables XL4_0 (censo de conductores) y XG5_0 (umbral de pobreza) son las que más información nueva aportarían. Observamos

```
matriz = cbind(Y_0, XL4_0, XG5_0);   pairs(matriz);   cor(matriz)
```

y observamos que cor(XL4_0, XG5_0)=0.985, por lo que sólo probamos añadir una de ellas. Elegimos la primera porque ya mostró un buen comportamiento para los años anteriores a 2009. Es decir, consideramos

$$Y_0=\beta_0+\beta_1 X_{G6,0}+\beta_2 X_{L4,0}+u$$ `regLin = lm(Y_0 ~ XG6_0 + XL4_0)`

que es un modelo globalmente significativo, explica cerca del 60% de la variabilidad de Y_0 y tal que cor(res, Y_0)=0.62, que sigue siendo alta pero se ha reducido. Sin embargo, como el coeficiente β_2 no parece distinto de cero finalmente volvemos al **modelo**

$$Y_0=\beta_0+\beta_1 X_{G6,0}+u \quad \text{cuya estimación es} \quad \hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1 X_{G6,0}=10999812-285144\, X_{G6,0} \quad (9)$$

A partir de las variables y datos disponibles, no parece que se pueda modelar mejor el número de matriculaciones, y habría que pensar en buscar alguna variable nueva que explique mejor el número de matriculaciones. Respecto a la interpretación del efecto total sobre el número de matriculaciones, se estima que, para todos los años, por cada unidad en que se incrementó el índice de Gini (en Eurostat consideran una escala de 0 a

100), se redujo el número de matriculaciones en 285144 vehículos. Mostramos los principales resultados:

```
Residuals:
    Min      1Q  Median      3Q     Max
-441440 -254623  -95715  323046  501418

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10999812    2123591   5.180 7.54e-05 ***
XG6_0        -285144      64597  -4.414 0.000379 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 336000 on 17 degrees of freedom
Multiple R-squared:  0.5341,  Adjusted R-squared:  0.5066
F-statistic: 19.48 on 1 and 17 DF,  p-value: 0.0003795
```

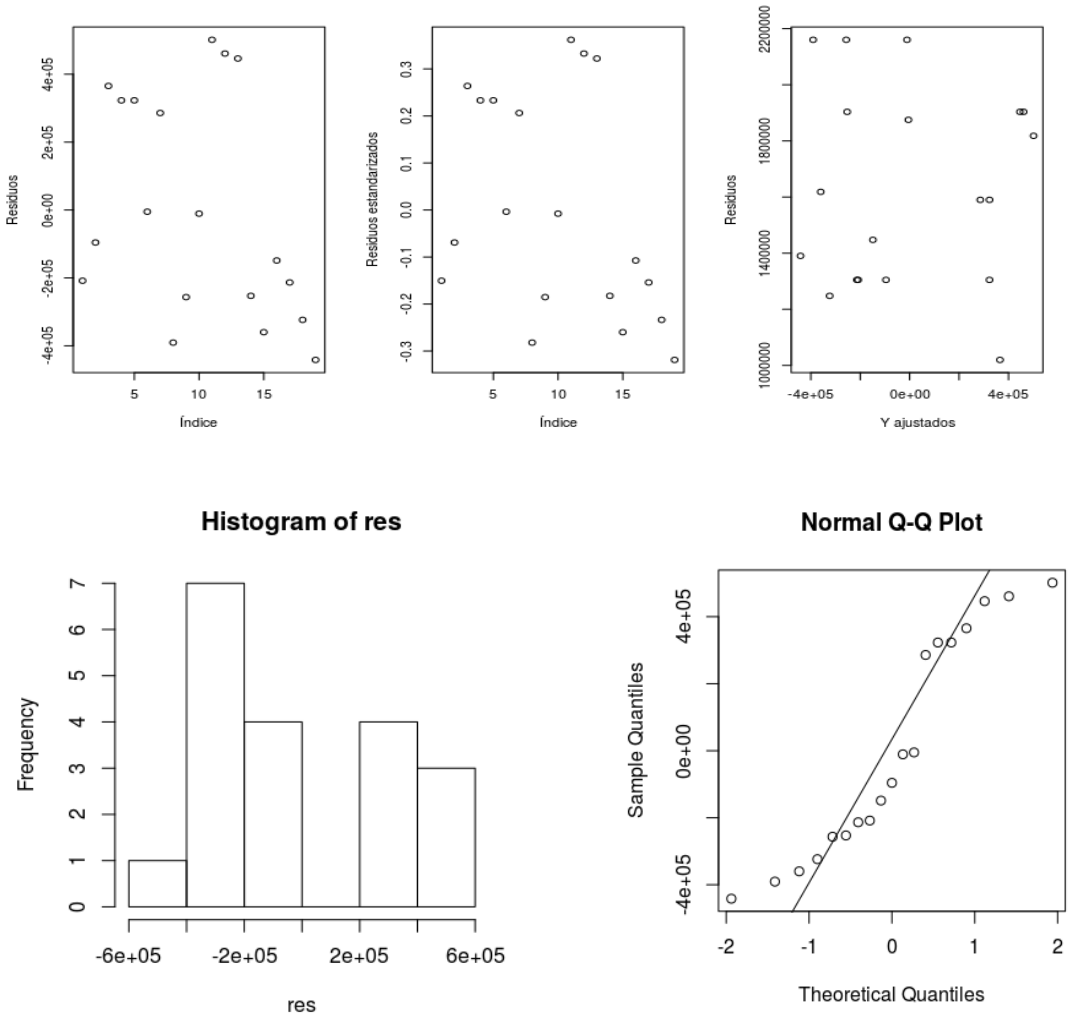


Figura 6: Gráficos para los residuos.

El lector puede aplicar los contrastes de aleatoriedad, normalidad y autocorrelación. El patrón que muestra el último gráfico corresponde, según Montgomery y otros (sección 4.2.3), a una distribución con colas pesadas (el histograma sugiere que la de la derecha lo es), aunque también aconsejan de la interpretación cautelosa de este tipo de gráfico cuando $n < 16$ (pocos datos).

Conclusiones

A partir de la información inicial, hemos considerado el posible valor que algunas variables podían tener para explicar el número de matriculaciones. También hemos comprobado los posibles problemas de nuestras variables y datos, y llegamos a la conclusión de que los únicos preocupantes serían un probable cambio estructural, la colinealidad y la correlación espuria. Hemos adaptado la metodología para evitar estos problemas y hemos encontrado modelos razonables y sencillos, sobre todo para los datos anteriores al 2008.

Para los años desde 1995 a 2008, un modelo aceptable es $Y=\beta_0+\beta_1 X_{L4}+u$, pero es algo mejor el modelo $Y=\beta_0+\beta_1 X_{G2r}+u$. Cuando se consideran todos los datos, ningún modelo puede explicar el número de matriculaciones con la calidad de los anteriores, y el mejor modelo es $Y_0=\beta_0+\beta_1 X_{G6,0}+u$, que explica algo más del 50% de la variabilidad del número de matriculaciones. Parece conveniente considerar algunas variables nuevas para construir un modelo en el que los residuos estén menos correlacionados con la variable que se desea explicar.

La interpretación de los efectos totales (las regresiones que hemos terminado aceptando son todas simples), es que

- Desde 1995 a 2008, cuando el censo de conductores se incrementó en cinco unidades, el número de matriculaciones lo hizo en una unidad aproximadamente.
- Desde 1995 a 2008, por un incremento de un euro por habitante en el producto interior bruto real, el número de matriculaciones aumentó en 208 vehículos.
- Desde 1995 a 2013, por cada unidad en que se incrementa el índice de Gini (en Eurostat consideran una escala de 0 a 100), se redujo el número de matriculaciones en 285144 vehículos.

El analista de datos, no la Estadística, es quien debe decidir la idoneidad de las variables, la representatividad de los datos y, por tanto, el valor y uso que da a cada modelo. Los años anteriores a la crisis, los de la crisis, o ambos conjuntos, podrían haber estado caracterizados por una economía singular que hace que el modelo no sea

válido para otras circunstancias. El propio carácter de nuestras variables nos indicaba que nuestro modelo tendría valor descriptivo, no de causalidad.

Como ya mencionamos en la introducción, otros análisis son posibles: utilizar las tasas de variación de las variables en vez de con las variables o trabajar con las variables estandarizadas.

Referencias

[1] Asociación Española de Fabricación de Automóviles y Camiones (ANFAC)
http://www.anfac.com/estadisticas.action

[2] Dirección General de Tráfico (DGT):
http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/
https://sedeapl.dgt.gob.es/IEST2/
http://www.dgt.es/es/la-dgt/centro-de-documentacion/

[3] Eurostat: http://ec.europa.eu/eurostat/data/database

[4] Montgomery, D.C., E.A. Peck and G.G. Vining (2001, 3rd ed). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Inc.

[5] Novales, A. (1997). *Estadística y Econometría*. McGraw-Hill

[6] Organisation Internationale des Constructeurs d'Automobiles
www.oica.net

Apéndice

Código

Sobre el lenguaje de programación R, véase http://www.r-project.org/.

A1. Código para introducir los datos

```
# Tiempo  [UNIDAD: años]
t_0 = 1995:2013

# Número de matriculaciones  [UNIDAD: ]
Y_0 = c(1096612, 1209197, 1385283, 1627899, 1913162, 1870262, 1875909, 1769857,
        1903801, 2149706, 2319590, 2364656, 2350101, 1651013, 1258781, 1298809,
        1091511, 924310, 949015)

## DGT: http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/
# Permisos de conducir emitidos  [UNIDAD: ]
XL1_0 = c(749547, 802649, 762586, 700430, 735709, 756816, 728665, 821689, 923033,
          1182956, 987297, 1133961, 1186742, 1252354, 977035, 823900, 749810, 708631,
          654924)

# Cambios de titularidad  [UNIDAD: ]
XL2_0 = c(1583696, 1728913, 1925559, 2023475, 2194306, 2304099, 2287072, 2381805,
          2633069, 2881242, 3184158, 3344645, 3460348, 3058363, 2938021, 2992928,
          2940634, 2891722, 2987708)

# Bajas de vehículos [UNIDAD: ]
XL3_0 = c(536503, 482945, 606787, 681643, 822861, 943272, 1054181, 996224, 1055139,
          979654, 1053457, 1093238, 1083542, 918406, 1135642, 973926, 810638, 882751,
          1133504)

# Censo de conductores  [UNIDAD: ]
XL4_0 = c(16761681, 17187616, 17554104, 18009374, 18459615, 18930263, 19348667,
          19823212, 20301418, 20919181, 21549477, 22124198, 22777657, 23657166,
          25713071, 25782360, 26118094, 26309230, 26387882)

## EUROSTAT: http://ec.europa.eu/eurostat/data/database
# Producto interior bruto (PIB) nominal  [UNIDAD: millones de euros, a precios actuales]
XG1_0 = c(468878.7, 505109.1, 519608.4, 551396.7, 594316.0, 646250.0, 699528.0,
          749288.0, 803472.0, 861420.0, 930566.0, 1007974.0, 1080807.0, 1116207.0,
          1079034.0, 1080913.0, 1075147.0, 1055158.0, 1049181.0)
```

```
# PIB per cápita nominal  [UNIDAD: euro por habitante, a precios actuales]
XG2n_0 = c(11900, 12800, 13100, 13900, 14900, 16100, 17200, 18100, 19000, 20100, 21300,
           22700, 23900, 24300, 23300, 23200, 23000, 22600, 22500)
# PIB per cápita real  [UNIDAD: euro por habitante]
XG2r_0 = c(16100, 16400, 17000, 17700, 18500, 19200, 19700, 20000, 20200, 20600, 21000,
           21500, 21800, 21700, 20700, 20600, 20600, 20200, 20100)

# Gasto de consumo final de los hogares  [UNIDAD: millones de euros, a precios actuales]
XG3_0 = c(294834.9, 315649.7, 324558.0, 342783.5, 369697.0, 402272.0, 430871.0,
          453637.0, 479081.0, 514256.0, 550656.0, 591546.0, 629402.0, 646998.0,
          617430.0, 631012.0, 638036.0, 635063.0, 627840.0)

# Población  [UNIDAD: miles de personas]
XG4_0 = c(39388.00, 39479.20, 39583.40, 39722.10, 39927.20, 40264.20, 40721.40,
          41423.50, 42196.20, 42859.20, 43662.60, 44360.50, 45236.00, 45983.20,
          46367.60, 46562.50, 46736.30, 46766.40, 46591.90)

# Umbral de riesgo de pobreza  [UNIDAD: euro]
XG5_0 = c(3702, 3748, 3971, 4076, 4491, 4941, 5416, 5682, 5923, 6196, 6272, 6683, 6987,
          7577, 8877, 8763, 8358, 8321, 8114)

# Índice de Gini
XG6_0 = c(34 , 34 , 35 , 34 , 33 , 32 , 33 , 31 , 31 , 31.0, 32.2, 31.9, 31.9, 31.9,
          32.9, 33.5, 34.0, 34.2, 33.7)
```

A2. Código para representar el conjunto de datos y, para cada par de variables, crear un gráfico y calcular su correlación

NOTA: Para las versiones de R para algunos sistemas operativos, la función gráfica `x11()` no abre una nueva ventana gráfica.

```
# Variable temporal y variable explicada
x11(); par(mfcol=c(1,2))
plot(t_0, main='Año', type='l')
plot(t_0, Y_0, main='N° matriculaciones', type='l')
# Variables explicativas de la DGT
x11(); par(mfcol=c(1,4))
plot(t_0, XL1_0, main='Permisos', type='l')
plot(t_0, XL2_0, main='Cambios de titularidad', type='l')
plot(t_0, XL3_0, main='Bajas', type='l')
plot(t_0, XL4_0, main='Censo de conductores', type='l')
# Variables explicativas de Eurostat
x11(); par(mfcol=c(1,7))
plot(t_0, XG1_0, main='PIB nominal', type='l')
plot(t_0, XG2n_0, main='PIB nominal per cápita', type='l')
plot(t_0, XG2r_0, main='PIB real per cápita', type='l')
plot(t_0, XG3_0, main='Gasto de consumo final de los hogares', type='l')
```

```
plot(t_0, XG4_0, main='Población y empleo', type='l')
plot(t_0, XG5_0, main='Umbral de riesgo de pobreza', type='l')
plot(t_0, XG6_0, main='Índice de Gini', type='l')

# Matriz para que las dos siguientes funciones consideren cada par de variables
matrizL_0 = cbind(t_0, Y_0, XL1_0, XL2_0, XL3_0, XL4_0)
x11();  pairs(matrizL_0)
cor(matrizL_0)

# Matriz para que las dos siguientes funciones consideren cada par de variables
matrizG_0 = cbind(t_0, Y_0, XG1_0, XG2n_0, XG2r_0, XG3_0, XG4_0, XG5_0, XG6_0)
x11();  pairs(matrizG_0)
cor(matrizG_0)
```

A3. Código para seleccionar los datos de los años 1995-2008 y, para cada par de variables, crear un gráfico y calcular su correlación

```
# Tomamos todos los datos salvo los últimos
t = t_0[1:13]
Y = Y_0[1:13]

XL1 = XL1_0[1:13]
XL2 = XL2_0[1:13]
XL3 = XL3_0[1:13]
XL4 = XL4_0[1:13]

XG1 = XG1_0[1:13]
XG2n = XG2n_0[1:13]
XG2r = XG2r_0[1:13]
XG3 = XG3_0[1:13]
XG4 = XG4_0[1:13]
XG5 = XG5_0[1:13]
XG6 = XG6_0[1:13]

# Matriz para que las dos siguientes funciones consideren cada par de variables
matrizL = cbind(t, Y, XL1, XL2, XL3, XL4)
x11();  pairs(matrizL)
cor(matrizL)

# Matriz para que las dos siguientes funciones consideren cada par de variables
matrizG = cbind(t, Y, XG1, XG2n, XG2r, XG3, XG4, XG5, XG6)
x11();  pairs(matrizG)
cor(matrizG)
```

A4. Código para instalar los paquetes adicionales

```
# Hay que evaluarlo sólo una vez
install.packages('e1071')
install.packages('tseries')
install.packages('randtests')
install.packages('lmtest')
install.packages('car')
```

A5. Código para cargar los paquetes adicionales

```
# Hay que evaluarlo en cada sesión de R
library(e1071)
library(tseries)
library(randtests)
library(lmtest)
library(car)
```

NOTA: Cuando paquetes distintos tienen alguna función con el mismo nombre, se puede llamar a la de un paquete concreto haciendo (o se puede redefinir las funciones): `nombrepaquete::nombrefuncion()`.

A6. Código para ajustar y analizar una regresión lineal

NOTA: Según las implementaciones en los distintos paquetes, para llamar a algunas funciones tenemos que introducir un vector de datos (los residuos, por ejemplo), a otras el objeto que devuelve la función `lm()` de R, y a otras la misma expresión del modelo que se introdujo en esta función `lm()`. Por otro lado, para aplicar el contraste de rachas del paquete `tseries` a los signos de los residuos, tenemos que ser nosotros quienes introduzcamos ya los factores dicotómicos (ser mayor o menor que su mediana, para una muestra de datos, o tener signo positivo o negativo, para los residuos).

```
# AJUSTE DEL MODELO
miregresion = lm(Y ~ X1 + X2)
# También se puede utilizar fácilmente este código para cada regresión
# sin más que hacer 'regLin=reg1', luego 'regLin=reg2' y así consecutivamente.
regLin = miregresion

# DIAGNOSIS: BONDAD DE AJUSTE Y COEFICIENTES
summary(regLin)

# ANÁLISIS DE LOS RESIDUOS
res = regLin$residuals  # Extraemos los residuos
resStand = res/sqrt(sum(res^2)/(length(res)-length(regLin$res)+1))  # Estandarizados
# Gráficos
```

```
x11(); par(mfcol=c(1,3))
plot(res, xlab='Índice', ylab='Residuos')   # Residuos
plot(resStand, xlab='Índice', ylab='Residuos estandarizados')
plot(res, regLin$fitted.values, xlab='Y ajustados', ylab='Residuos')
# Información sobre la variable explicada que todavía queda en los residuos
cor(res, Y)  # Correlación de los residuos e Y
# sd(res)/sd(Y)   # Ratio
# Contrastes de aleatoriedad
tseries::runs.test(as.factor(res<median(res)))  # Contraste de rachas (Wald-Wolfowitz)
randtests::runs.test(res, pvalue='exact')  # Contraste de rachas (Wald-Wolfowitz)
difference.sign.test(res)  # Contraste de diferencias sucesivas
bartels.rank.test(res)  # Contraste de rangos de Bartels
cox.stuart.test(res)  # Contraste de Cox Stuart
rank.test(res)  # Contraste de rangos (Mann-Kendall)
turning.point.test(res)  # Contraste del punto de retorno
# Media nula
mean(res)   # Media
median(res)   # Mediana
# Contrastes de heterocedasticidad
bptest(Y ~ X1 + X2)  # Contraste de Breusch-Pagan o de varianza (del error) no constante
ncvTest(regLin)  # Contraste de Breusch-Pagan o de varianza (del error) no constante
# Normalidad
hist(res)   # Histograma
skewness(res)   # Asimetría
kurtosis(res)   # Exceso de curtosis
# Contrastes de normalidad
qqnorm(res); qqline(res)   # Para crear un Q-Q plot
shapiro.test(res)
jarque.bera.test(res)   # Contraste de normalidad
# Contrastes de autocorrelación
randtests::runs.test(sign(res), pvalue='exact')  # Contraste de rachas sobre los signos
tseries::runs.test(as.factor(sign(res)))   # Contraste de rachas sobre los signos
durbinWatsonTest(res)   #  Contraste de Durbin-Watson
dwtest(regLin)   #  Contraste de Durbin-Watson
Box.test(res)   # Prueba Q de Ljung-Box
bgtest(regLin)   # Prueba LM de Breusch-Godfrey
```

Extraemos algunas líneas de código para facilitar su uso de forma independiente.

B1. Código para ajustar una regresión lineal

```
# Ajuste del modelo
miregresion = lm(Y ~ X1 + X2)
regLin = miregresion
# Gráficos
print(regLin)
plot(regLin)
# Para acceder a la información básica
```

```
summary(regLin)
# Para ver qué campos tiene este objeto y acceder a ellos
names(regLin)

# Ajuste sin término independiente (beta_0)
miregresion = lm(Y ~ 0 + X1)  # O también lm(Y ~ X1 - 1)
```

B2. Código para quitar a una variable el efecto de otras

```
# Para quitarle a la variable Z el efecto de X1 y X2, por ejemplo
miregresion = lm(Z ~ X1 + X2)
Ztilde = miregresion$residuals
```

B3. Código para estudiar la aleatoriedad

NOTA: Para aplicar el contraste de rachas del paquete `tseries` tenemos que ser nosotros quienes comparemos los datos con la media, puesto que la implementación requiere introducir un factor dicotómico.

```
# Gráficos
plot(VARIABLE)
# Contrastes de hipótesis
tseries::runs.test(as.factor(VARIABLE<median(VARIABLE)))  # Contraste de rachas (W-W)
randtests::runs.test(VARIABLE, pvalue='exact')  # Contraste de rachas (W-W)
difference.sign.test(VARIABLE)  # Contraste de diferencias sucesivas
bartels.rank.test(VARIABLE)  # Contraste de rangos de Bartels
cox.stuart.test(VARIABLE)  # Contraste de Cox Stuart
rank.test(VARIABLE)  # Contraste de rangos (Mann-Kendall)
turning.point.test(VARIABLE)  # Contraste del punto de retorno
```

B4. Código para estudiar la normalidad

```
# Medidas
mean(VARIABLE)
median(VARIABLE)
skewness(VARIABLE)
# Gráficos
hist(VARIABLE)
# Constrastes de bondad de ajuste
qqnorm(VARIABLE);  qqline(VARIABLE)  # Gráfico de cuantiles. También puede hacerse:
                                     # qqPlot(VARIABLE)  # Del paquete 'car'

shapiro.test(VARIABLE)   # Contraste de Shapiro
jarque.bera.test(VARIABLE)   # Contraste de Jarque y Bera
```

B5. Código para estudiar la autocorrelación

```
# Constrastes de ausencia de autocorrelación
randtests::runs.test(sign(RESIDUOS), pvalue='exact')  # Contraste de rachas a los signos
tseries::runs.test(as.factor(sign(RESIDUOS)))   # Contraste de rachas sobre los signos
durbinWatsonTest(RESIDUOS)   #  Contraste de Durbin-Watson
dwtest(lmOBJETO)   #  Contraste de Durbin-Watson
Box.test(RESIDUOS)   # Prueba Q de Ljung-Box
bgtest(lmOBJETO)   # Prueba LM de Breusch-Godfrey

# Ejemplo:  lmOBJETO = lm(Y ~ X1+X2);  RESIDUOS = lmOBJETO$residuals
```