Influencia del factor provincia en el análisis, la selección y la valoración del riesgo en el seguro de autos

Autor: Héctor Martín Martín

Tutores:
Adolfo Caballero Carbonell
María Pérez Martín

Máster en Ciencias Actuariales y Financieras
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense de Madrid

2015-2016

Índice

Índice de tablas

Índice de figuras

1. Introducción

El presente documento va a abordar un estudio sobre la influencia que puede tener el factor provincia en la fase de análisis, selección y valoración del riesgo de la elaboración de las normas de suscripción para el ramo de autos, y cómo puede esto afectar al importe final de las primas.

Para llevar a cabo este estudio, en primer lugar ha sido necesario construir una base de datos que va a estar formada por variables que están en su gran mayoría fuertemente relacionadas con el ramo de autos. Todos los datos utilizados están disponibles en fuentes oficiales y son del año 2014. Antes de trabajar con dicha base de datos, se ha realizado una depuración de los datos con el consiguiente tratamiento de datos atípicos, datos perdidos,...

Una vez preparada la base de datos se ha trabajado en el objetivo principal del trabajo, probar lo mucho que puede variar la prima de autos dependiendo de la provincia en la que resida el titular de la póliza de seguro. Para ello, se ha tratado de, por un lado, elaborar una clasificación de las provincias a partir de las variables recogidas y por otro, hacer una estimación de la prima de autos en función de dichas variables. Finalmente, se hace una comparación de la prima estimada con la prima del año 2014 y se resaltan las principales diferencias.

En el capítulo 2 se explica qué ha motivado este trabajo y cuál es el objetivo que se persigue con este estudio. Para ello, se hace una breve introducción al seguro dando la definición de términos básicos como son la prima y el riesgo, se introduce también el ramo de autos y se explica en qué consiste la normativa de suscripción de una entidad aseguradora.

En el capítulo 3 se trata el tema de la dificultad que existe en la actualidad para encontrar datos, y lo que es más importante, que esos datos sean de calidad. También se explica cómo ha sido la construcción de la base de datos y cuáles han sido las distintas fuentes de las que se han obtenido los datos. Todos estos datos tendrán relación con el ramo de autos. En este apartado se aborda también la depuración de la base de datos y, en concreto, se trabaja en un problema de missing. Para buscar una solución a este problema se utilizarán técnicas de análisis multivariante como regresión lineal múltiple y modelos lineales generalizados (GLM) que serán implementadas a través de SPSS y R respectivamente.

En el capítulo 4 se hace una clasificación de las provincias a partir de los datos obtenidos, y para ello se utiliza el análisis de cluster como técnica de clasificación. Se utilizan distintos tipos de ratios, así como distintas distancias y enlaces con el fin de obtener una clasificación lo más precisa posible.

En el capítulo 5 se realiza una estimación tanto del número total de primas como de la prima media de autos partiendo de los datos recogidos y utilizando como técnica la regresión lineal múltiple. Además, se compara la prima media con la prima de autos del año 2014 y se analizan las diferencias.

En el capítulo 6 se comentan cuáles son las principales conclusiones que se han extraído del trabajo.

En el capítulo 7 se plantean tres futuras líneas de investigación: ver qué ocurriría si se introduce el factor zona en este estudio, tratar de ver cómo afectarían las variables de fallecidos en cada provincia y como influye esto en el importe de la prima, y estimar la prima para cada uno de las regiones que surgen de la aplicación del análisis de cluster y valorar si las entidades aseguradoras podrían hacer grupos en función de estas regiones para posteriormente calcular la prima.

2. Motivación y objetivo

2.1. Introducción al seguro

La Ley de Contrato de Seguro [19] define el contrato de seguro como aquél por el que el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura, a indemnizar, dentro de los límites pactados el daño producido al asegurado o a satisfacer un capital, una renta u otras prestaciones convenidas. De esta definición, son interesantes para el objetivo del trabajo, el riesgo y la prima. Según Boj, Claramunt y Fortiana (2006) [2] la prima es el importe que paga el tomador del seguro a cambio de estar protegido en caso de que se produzca el riesgo que cubre la cobertura. La prima, que va a venir definida por el riesgo asegurable y por el resto de factores que conforman el coste de la empresa, es el precio del servicio más el margen explícito de beneficio y debe cumplir los principios de equidad y suficiencia de acuerdo con la naturaleza de los riesgos asumidos por el asegurador. El principio de equidad hace referencia a que la prima se ajuste al riesgo de siniestralidad de cada póliza y el principio de suficiencia se refiere a que en términos esperados las primas sean suficientes para cubrir todos los riesgos de la cartera considerada, es decir, garantizan la rentabilidad en condiciones de estabilidad a largo plazo de la entidad aseguradora. Como ya definía Guardiola (1990) [7], el riesgo podría definirse en términos generales como la posibilidad incierta de que ocurra un acontecimiento que produce una necesidad económica y cuya manifestación se previene y garantiza en la póliza y obliga al asegurador a efectuar la prestación que le corresponde. El riesgo tiene seis características fundamentales:

- Es incierto.
- Es posible.
- Es concreto, es decir, se puede analizar y valorar tanto a nivel cualitativo como cuantitativo.
- Es lícito.
- Es fortuito.
- Tiene contenido económico, ya que la realización del riesgo ha de producir una necesidad económica.

Para poder asumir la cobertura de este riesgo, el asegurador debe llevar a cabo la aplicación de una serie de técnicas que le permitan establecer la naturaleza, valoración y límites de aceptación de dicho riesgo. Estas técnicas se resumen en: selección, análisis, evaluación, compensación y distribución.

La selección de riesgos es una parte fundamental de la actividad aseguradora. En Guardiola (1990) [7] se definía como el conjunto de medidas, generalmente de carácter técnico, que adopta una entidad aseguradora para decidir cuáles son los riesgos que acepta, orientando esta decisión a aquéllos de los que se espera que, por sus propias características, no van a dar lugar necesariamente a resultados desequilibrados por encontrarse dentro del promedio en su categoría. Es decir, las entidades aseguradoras no aceptarán riesgos que por sus características se alejen de la siniestralidad propia de su ramo.

La elaboración de las normas de suscripción es el proceso que nos permite evaluar y clasificar adecuadamente el riesgo que queremos asegurar para fijar un precio justo a su cobertura. Es muy importante por tanto, el uso de unos buenos criterios de selección ya que esto podría ayudar mucho a ajustar correctamente el precio de la prima, que es el objetivo de cualquier entidad aseguradora.

Es por este motivo por el que el trabajo se centrará en hacer un estudio que pueda ayudar a la elaboración de dichas normas de suscripción. En concreto, el trabajo se enfocará en estudiar la influencia de la variable provincia en la selección y valoración de riesgos para el ramo de autos, y en consecuencia, en el importe de la prima.

2.2. Ramo de autos

El ramo de autos es aquel que tiene por objeto la prestación de indemnizaciones derivadas de accidentes producidos a consecuencia de la circulación de vehículos y es uno de los más importantes para las entidades aseguradoras.

En España, la contratación de un seguro que cubra los daños personales y materiales que se puedan ocasionar a terceras personas es obligatoria para todo propietario de un vehículo a motor [20].

El seguro obligatorio garantizará la cobertura de la responsabilidad civil en vehículos terrestres automóviles con estacionamiento habitual en España, en todo el territorio del Espacio Económico Europeo y de los Estados adheridos al Acuerdo entre las oficinas nacionales de seguros de los Estados miembros.

Los importes de la cobertura del seguro obligatorio son los siguientes:

- En los daños personales, 70 millones de euros por siniestro, cualquiera que sea el número de víctimas.
- En los daños materiales, 15 millones de euros por siniestro.

Cuando concurran daños personales y materiales, y la indemnización de estos últimos superen los 15 millones de euros por siniestro, la diferencia se indemnizará con cargo al remanente que pudiera resultar de la indemnización por los daños personales [21].

Las entidades aseguradores ofrecen las modalidades de cobertura para el seguro de autos que se pueden ver en la tabla 1:

<table>
<tr><th colspan="2">MODALIDADES</th><th colspan="2">COBERTURAS</th></tr>
<tr><td rowspan="8">Todo riesgo con/sin franquicia</td><td rowspan="6">Terceros
+ lunas
+ robo
+ incendio</td><td rowspan="5">Terceros</td><td>Resp. civil obligatoria</td></tr>
<tr><td>Resp. civil voluntaria</td></tr>
<tr><td>Asistencia en viaje</td></tr>
<tr><td>Acc. del conductor</td></tr>
<tr><td>Defensa Jurídica</td></tr>
<tr><td colspan="2">Rotura de lunas</td></tr>
<tr><td colspan="2">Robo</td></tr>
<tr><td colspan="2">Incendio</td></tr>
<tr><td colspan="3">Daños propios (con/sin franquicia)</td></tr>
</table>

Tabla 1: Modalidades de cobertura del seguro de autos
Fuente: Elaboración Propia

2.3. Normativa de suscripción

Cuando un cliente quiere contratar una póliza, las entidades aseguradoras abren un procedimiento de contratación en el que han de tomar la decisión de asumir o no la cobertura de un determinado riesgo.

Para ello, la entidad dispone de lo que se denomina política o normativa de suscripción, que son el conjunto de reglas a aplicar en el momento de la aceptación de un riesgo en base a la experiencia en dicho riesgo y que se corresponden con el siguiente esquema:

Fase I. Análisis, selección y valoración del riesgo

En esta fase, la entidad aseguradora pide al cliente que especifique las circunstancias del riesgo en base a unos parámetros prefijados. Es muy importante para la entidad aseguradora saber toda la información acerca de la zona de circulación, el vehículo asegurado y el conductor.

Son muchas las variables que influyen en la valoración del riesgo. Se pueden agrupar en:

- **Datos del conductor:** En este caso son especialmente influyentes la edad y la experiencia en la conducción, así como la zona por la que se conduce de manera habitual. En este sentido, la provincia es un elemento capital en la tarificación. También afecta si se trata de una zona urbana, periférica o rural.
- **Datos del vehículo:** El tipo, el valor y las características técnicas del vehículo serían los elementos más importantes para la tarificación.
- **Datos del seguro actual:** En este caso es influyente la información sobre su historial siniestral.

Fase II. Tarificación y proposición de seguro

Es en esta segunda fase en la que, una vez decidido que el riesgo es asegurable, se procede a calcular la prima correspondiente para esa cobertura.

Fase III. Formalización de la póliza

Una vez calculada la prima, la entidad correspondiente emite la póliza.

Este trabajo se centrará en la primera fase de la elaboración de las normas de suscripción, es decir, en la selección y valoración de los riesgos. El objetivo será ajustar mejor el valor de la prima exigida por las entidades aseguradoras en función de la provincia, que sabemos que es un factor de vital importancia en la tarificación. Para ello, se va a trabajar con datos a nivel de provincia. Es obvio que no podrá pagar la misma prima un asegurado que reside en Madrid que uno que reside en Ávila, así como un residente en Cádiz igual que un residente en Asturias, debido a las diferencias entre las distintas provincias. Ahora bien, en este estudio se pretende ir un poco más allá de la evidencia e intentar profundizar más para obtener unos resultados lo más precisos posibles.

3. Elaboración y depuración de la base de datos

3.1. Problemas en la búsqueda de datos

A partir de 1930, con el comienzo de la era de la computación se empiezan a almacenar gran cantidad de datos. A medida que se fueron desarrollando las computadoras esta cantidad fue aumentando y a día de hoy existen webs como Amazon y eBay que tienen un modelo de negocio basado en una gran base de datos.

Esto mismo ocurre en todos los sectores, incluido el sector seguros. En la actualidad, todas las entidades aseguradoras tienen una base de datos enorme y están trabajando en su explotación. Veamos un ejemplo que presenta Mayer-Schönberger y Cukier (2013) [9] que trata sobre cómo la introducción de sistemas de geolocalización en los automóviles está transformando el mundo de los seguros. Los datos ofrecen una vista pormenorizada de los tiempos, localizaciones y distancias de conducción real que permiten un precio mejor en función del riesgo. En la actualidad, las entidades aseguradoras pueden ajustar el precio del seguro del coche dependiendo de a dónde y cuándo conducen sus clientes, en lugar de basarse en su edad, sexo e historial. Esto además crea incentivos al buen comportamiento. Supone una gran transformación, ya que el seguro pasa de estar basado en el riesgo agrupado a basarse en la actuación individual.

Ahora bien, uno de los problemas que existen es que el foco en muchos casos está centrado en almacenar gran cantidad de datos y no en la calidad del dato. Desde un punto de vista estadístico, este aspecto es fundamental. Como ya hacía referencia Prieto (1980) [15] hay que cuidar mucho la calidad de la información, así como ser meticuloso en la interpretación de los resultados. Hay que tener en cuenta que el análisis de los datos tienen como objetivo ayudar en la toma de decisiones, además de en la elaboración de estrategias, políticas y tácticas de diversa índole. Para que estas decisiones sean acertadas, es importante no sólo tener mucha información sino que ésta sea de calidad, ya que si los datos de partida no son buenos los resultados tampoco lo serán con independencia de la técnica estadística que utilicemos.

Para obtener una buena base de datos es necesario tener claro cuál es el objetivo, es decir, dónde se quiere llegar con esos datos y cómo se van a tratar. De esta forma se pretende evitar acumular datos innecesarios y tener datos que no pueden ser tratados debido a su incompatibilidad.

Uno de los problemas que se presentan a día de hoy es que no se puede acceder a muchos de los datos existentes. Por ello, es bastante difícil trabajar con datos de la fecha actual.

Teniendo en cuenta el principio de calidad de datos que se ha mencionado anteriormente, la búsqueda de datos se ha basado en fuentes de confianza y por ello todas las fuentes utilizadas son fuentes oficiales como Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones (ICEA), el Instituto Nacional de Estadística (INE),... Además con el fin de que los datos sean lo más actuales posible y que a su vez se puedan recoger gran cantidad de datos, se ha optado por trabajar con datos del año 2014.

Por otro lado, se ha tratado de buscar datos muy relacionados con el ramo de autos y en concreto, datos que se puedan utilizar en la selección de riesgos para dicho ramo. Para ello, ha sido necesario pensar bien cuál era el objetivo del trabajo y cómo se quería llevar a cabo, para a partir de ahí buscar los datos más adecuados para la investigación.

3.2. Construcción de la base de datos

La base de datos está compuesta por distintas tablas de datos extraídas de documentos que están disponibles en distintas fuentes que se citan más adelante. Todas ellas tienen una identidad (ID) común, que son las provincias españolas. Se ha construido a partir de la tabla extraída del documento "Grupo 7. Accidentes y víctimas en función de la vía 2014" que está disponible en `http://www.dgt.es/es/seguridad-vial/estadisticas-eindicadores/accidentes-30dias/tablas-estadisticas/`; que es la página web de la Dirección General de Tráfico (DGT), y a partir de ahí se han ido agregando nuevos datos por provincia. Hay que recordar que todos los datos utilizados son del año 2014.

La base de datos ha quedado formada por las siguientes tablas:

- De las estadísticas e indicadores de la DGT (`http://www.dgt.es/es/seguridad-vial/estadisticas-eindicadores/`) se han obtenido tres tablas:
    - La tabla obtenida del documento *"Grupo 7. Accidentes y víctimas en función de la vía 2014"* contiene el número de víctimas, de fallecidos, de heridos hospitalizados y de heridos no hospitalizados que se han producido en cada tipo de vía distinguiendo entre los distintos tipos de vías interurbanas y urbanas que existen, desde una autopista a una calle o travesía. En total, esta tabla está formada por 28 variables. En este punto conviene hacer una aclararación. Se considerarán las siguientes definiciones sobre los accidentes de tráfico:

    - **Accidentes con víctimas:** Los que se producen, o tienen su origen en una de las vías o terrenos objeto de la legislación sobre tráfico, circulación de vehículos a motor y seguridad vial y a consecuencia de los mismos una o varias personas resultan muertas y/o heridas.
    - **Víctima mortal:** Toda persona que, como consecuencia del accidente, fallezca en el acto o dentro de los treinta días siguientes.
    - **Heridos graves:** Aquellas personas heridas en un accidente de circulación y cuyo estado precise una hospitalización superior a veinticuatro horas.
    - **Heridos leves:** Aquellas personas heridas en un accidente de circulación a los que no puede aplicarse la definición de herido grave.
  - Del documento *"Parque de Vehículos - Anuario - 2014"* se ha obtenido la tabla del parque de vehículos que está formada por los distintos tipos de vehículos como pueden ser turismos, motocicletas,... En total, esta tabla está compuesta por 7 variables.
  - La tabla de conductores extraída del documento *"Censo de Conductores - Anuario - 2014"* está formada por una sola variable que es el número de conductores.

- De la página web del INE (`http://www.ine.es/inebmenu/indice.htm`) se obtienen tres tablas:
  - Del apartado de demografía y población:
    - Una tabla de habitantes que únicamente contiene la variable número de habitantes.
  - Del apartado de mercado laboral:
    - Una tabla de actividad de la población que está compuesta por tres tasas: la tasa de actividad (cociente entre la población activa y la población de 16 años o más), la tasa de paro (cociente entre la población parada y la población activa) y la tasa de empleo (cociente entre la población ocupada y la población activa).
  - Del apartado de Industria, Energía y Construcción:
    - Una tabla de viviendas que se compone únicamente por la variable número de viviendas.

- Del catálogo y evolución de la red de carreteras del Ministerio de Fomento (`http://www.fomento.gob.es/MFOM/LANG_CASTELLANO/DIRECCIONES_GENERALES/CARRETERAS/CATYEVO_RED_CARRETERAS/`) se ha obtenido:

  - Una tabla de kilómetros de la red de carreteras que está formada por variables con la longitud de los distintos tipos de carreteras existentes, desde carreteras con calzada inferior a 5 metros hasta autopistas de peaje. En total, está compuesta por 6 variables.

- Del apartado Mercado de trabajo y pensiones en las fuentes tributarias de la Agencia Tributaria que se encuentra en (`http://www.agenciatributaria.es/AEAT.internet/datosabiertos/catalogo/hacienda/`) se ha extraído:

  - Una tabla de salarios que está compuesta por el número de asalariados y el salario medio anual.

- Del apartado catálogo de datos del Ministerio del Interior (`http://datos.gob.es/catalogo/balance-de-criminalidad-2014`) se ha obtenido:

  - Una tabla de robos que está formada por el número de robos de coches y el número de robos en domicilios, así como los correspondientes porcentajes de variación del año 2014 respecto al año 2013.

- Del apartado primas de ICEA que se encuentra en (`http://www.icea.es/es-es/informaciondelseguro/totalsector/primas/paginas/home.aspx`) se ha extraído:

  - Una tabla de primas de autos que está compuesta por dos variables que son las primas de autos de volumen de negocio y las primas de autos de nueva producción.

Finalmente, la base de datos ha quedado compuesta por 52 registros y 59 variables. En la tabla 2 se puede ver un resumen de las variables utilizadas:

Grupo	Variable	Etiqueta
Provincia	- Índice para identificar la provincia -Nombre de la provincia	ID Provincia
Accidentes según la vía	Vías Interurbanas: -Nº de víctimas en accidentes en autopista -Nº de fallecidos en accidentes en autopista -Nº de heridos hospitalizados en accidentes en autopista -Nº de heridos no hospitalizados en accidentes en autopista -Nº de víctimas en accidentes en autovía -Nº de fallecidos en accidentes en autovía -Nº de heridos hospitalizados en accidentes en autovía -Nº de heridos no hospitalizados en accidentes en autovía -Nº de víctimas en accidentes en vía convencional -Nº de fallecidos en accidentes en vía convencional -Nº de heridos hospitalizados en accidentes en vía convencional -Nº de heridos no hospitalizados en accidentes en vía convencional	 Inter_autopistas_acc_víct Inter_autopistas_fall Inter_autopistas_her_hosp Inter_autopistas_her_no_hosp Inter_autovías_acc_víct Inter_autovías_fall Inter_autovías_her_hosp Inter_autovías_her_no_hosp Inter_convencional_acc_víct Inter_convencional_fall Inter_convencional_her_hosp Inter_convencional_her_no_hosp

Accidentes según la vía	-Nº de víctimas en accidentes en otra vía interurbana -Nº de fallecidos en accidentes en otra vía interurbana -Nº de heridos hospitalizados en otra vía interurbana -Nº de heridos no hospitalizados en otra vía interurbana Vías Urbanas: -Nº de víctimas en accidentes en calle -Nº de fallecidos en accidentes en calle -Nº de heridos hospitalizados en calle -Nº de heridos no hospitalizados en calle -Nº de víctimas en accidentes en travesía -Nº de fallecidos en accidentes en travesía -Nº de heridos hospitalizados en travesía -Nº de heridos no hospitalizados en travesía -Nº de víctimas en accidentes en autovía/autopista urbana -Nº de fallecidos en accidentes en autovía/autopista urbana -Nº de heridos hospitalizados en autovía/autopista urbana -Nº de heridos no hospitalizados en autovía/autopista urbana	Inter_otro_acc_víct Inter_otro_fall Inter_otro_her_hosp Inter_otro_her_no_hosp Urb_calle_acc_víct Urb_calle_fall Urb_calle_her_hosp Urb_calle_her_no_hosp Urb_travesía_acc_víct Urb_travesía_fall Urb_travesía_her_hosp Urb_travesía_her_no_hosp Urb_autourb_acc_víct Urb_autourb_fall Urb_autourb_her_hosp Urb_autourb_her_no_hosp
Parque de Vehículos	-Nº total de vehículos -Nº de camiones y furgonetas -Nº de autobuses -Nº de turismos -Nº de motocicletas -Nº de tractores industriales -Nº de R y S -Nº de otros vehículos -Nº de conductores	Total_vehículos Camiones_furgonetas Autobuses Turismos Motocicletas Tractores_industriales R_y_S Otros_vehículos n_conductores
Red de Carreteras	-Nº total de kilómetros de la red de carreteras -Nº de kilómetros de carreteras con calzada menor de 5 metros -Nº de kilómetros de carreteras con calzada entre 5 y 7 metros -Nº de kilómetros de carreteras con calzada mayor de 7 metros -Nº de kilómetros de carreteras con doble calzada -Nº de kilómetros de autovías y autopistas libres -Nº de kilómetros de autopistas de peaje	Total_km_red_carreteras km_carr_calz_menor5 km_carr_calz_5a7 km_carr_calz_mayor7 km_carr_doble_calz km_autovías_autopistas_libres km_autopistas_peaje
Población	-Nº de habitantes -Nº de asalariados -Salario medio anual -% del cociente entre la población activa y la población de 16 o más años -% del cociente entre la población parada y la población activa -% del cociente entre la población ocupada y la población activa	n_habitantes n_asalariados salario_medio_anual Tasa_actividad Tasa_paro Tasa_empleo
Robos	-Nº de robos de vehículos en el año 2014 -% de variación del número de robos de vehículos del año 2014 respecto al año 2013 -Nº de viviendas -Nº de robos en domicilios en el año 2014 -% de variación del número de robos en domicilios del año 2014 respecto al año 2013	total_robos_vehículos_2014 porc_var_robos_vehículos n_viviendas total_robos_domicilio_2014 porc_var_robos_domicilio
Primas	-Nº total de primas del ramo de autos -Prima media del ramo de autos	Primas_Autos_Total Prima_media

Tabla 2: Variables que componen la base de datos
Fuente: Elaboración Propia

3.3. Imputación de datos perdidos

Para poder elaborar esta base de datos se ha llevado a cabo una depuración de los datos. Al obtener las tablas de distintas fuentes cada una tenía su propio formato, por lo que antes de realizar la unión se ha unificado el formato y se ha establecido en todas las tablas el mismo nombre para las provincias. Después, ha habido que tratar los datos atípicos o outliers y los datos perdidos o missing. Respecto a los outliers hay que decir que al haber utilizado datos de fuentes oficiales estaban ya muy depurados, por lo que podemos decir que no han existido problemas de outliers. En cuanto a los datos missing sí se han encontrado problemas ya que al haberse obtenido los datos de salarios de la agencia tributaria no se dispone de los datos para las provincias vascas y la Comunidad de Navarra debido a su condición especial de diputaciones forales. Se ha tratado de obtener estos datos de otros fuentes pero no ha sido posible, por lo que se ha optado por utilizar técnicas estadísticas para estimarlos. Para poder aplicar dichas técnicas estadísticas, se han buscado nuevos datos que fuesen indicadores de riqueza. De esta forma se han añadido tres nuevas tablas a la base de datos: una tabla de sociedades, una tabla del Índice de Precios de Consumo (IPC) y una tabla de depósitos.

- Todas ellas se han obtenido de la página web del INE (`http://www.ine.es/inebmenu/indice.htm`):
  - Del apartado de economía:
    - Una tabla de sociedades que está compuesta por el número de sociedades y por el capital total de éstas.
    - La tabla de depósitos contiene tres variables que son el número de depósitos vista, depósitos de ahorro y depósitos a largo plazo.
  - Del apartado nivel y condicionados de vida (IPC):
    - La tabla del IPC está formada por el propio índice así como por su variación anual.

El objetivo es aplicar técnicas estadísticas para estimar las variables número de asalariados y salario medio anual de las provincias vascas y de Navarra. Para ello, se ha optado por utilizar en primer lugar una regresión lineal múltiple.

3.3.1. Regresión Lineal Múltiple

Según Abuín (2007) [17] el modelo de regresión lineal múltiple analiza la influencia de varias variables explicativas x en el valor que toma la variable dependiente y. La

ventaja de este modelo respecto al modelo de regresión lineal simple es que al tener más variables disponemos de más información, por lo que el análisis será más preciso.

Supongamos que los valores de la variable dependiente y han sido generados por una combinación lineal de los valores de una o más variables explicativas $x_1, x_2, ..., x_n$ y un término aleatorio u:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + u$$

Se pretende saber la influencia β de cada una de las variables explicativas en la variable dependiente y.

Estos coeficientes se eligen de forma que la suma de cuadrados entre los valores observados y los pronosticados sea mínima, es decir, se minimiza la varianza residual.

Las variables explicativas elegidas deben cumplir determinados criterios:

- Deben ser numéricas
- No debe existir redundancia
- La presencia de las variables en el modelo debe estar justificada desde un punto de vista teórico.
- La relación entre variables explicativas en el modelo y casos debe de ser como mínimo de 1 a 10.
- La relación de las variables explicativas con la variable dependiente debe ser lineal o proporcional.

Además, en este modelo se han de asumir las siguientes hipótesis:

- Linealidad: los valores de la variable dependiente están generados por el siguiente modelo lineal,
$$y = \beta x + u$$
- Homocedasticidad: todas las perturbaciones tienen las misma varianza,
$$V(u_i) = \sigma^2$$
- Independencia: las perturbaciones aleatorias son independientes entre sí,
$$E(u_i u_j) = 0 \quad \forall i \neq j$$

- Normalidad: la distribución de la perturbación aleatoria tiene distribución normal,
$$u \approx N(0, \sigma^2)$$
- Las variables explicativas x_k se obtienen sin errores de medida.

Bajo estas hipótesis el teorema de Gauss-Markov establece que el método de estimación de mínimos cuadrados va a producir estimadores óptimos, en el sentido de que los parámetros estimados van a estar centrados y van a ser de mínima varianza.

En el caso particular del modelo que se va a construir las variables explicativas x_i serán 7, en concreto, el número de sociedades, el capital de estas sociedades, el IPC, la variación anual del IPC, el número de depósitos vista, el número de depósitos de ahorro y el número de depósitos a largo plazo, y las variables dependientes y_i serán el número de asalariados y el salario medio anual.

Se empezará tomando número de asalariados como variable dependiente. Al aplicar la regresión lineal múltiple para obtener el número de asalariados, SPSS proporciona la tabla 3 en la que nos indica cuál es el mejor modelo con una variable, el mejor modelo con dos variables y el mejor modelo con tres variables.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.987[a]	.975	.974	79095.264
2	.991[b]	.981	.981	68269.441
3	.994[c]	.989	.988	54255.573

a. Predictores: (Constante), n_sociedades

b. Predictores: (Constante), n_sociedades, cap_sociedades

c. Predictores: (Constante), n_sociedades, cap_sociedades, Depósitos_ahorro

Tabla 3: Modelos en la estimación del número de asalariados
Fuente: Elaboración Propia

En este caso, se obtiene un R^2 ajustado muy alto para el modelo con tres variables, por lo que en principio podríamos afirmar que el modelo formado por las variables

número de sociedades, el capital y el número de depósitos de ahorro es bueno. Ahora bien, como dicen Belsley, Kuh y Welsch (1980) [1] y Rawlings (1988) [16] hay que fijarse en el valor de la distancia de Mahalanobis y en si la distancia de Cook es mayor que $\frac{4}{n}$ donde n es el número de observaciones, para detectar si existen puntos influyentes. Para ello, SPSS proporciona la tabla 4 que contiene entre otros valores, los coeficientes de los residuos, la distancia de Mahalanobis y la distancia de Cook.

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	49947,49	2712225,25	352063,00	487483,205	48
Valor pronosticado tip.	-,620	4,842	,000	1,000	48
Error típico de valor pronosticado	8055,737	53830,813	13047,041	8756,584	48
Valor pronosticado corregido	51201,43	3281817,25	363493,16	545071,629	48
Residual	-136380,953	132314,703	,000	52495,462	48
Residuo típ.	-2,514	2,439	,000	,968	48
Residuo estud.	-2,779	2,548	-,023	1,035	48
Residuo eliminado	-578616,313	144387,750	-11430,164	102303,718	48
Residuo eliminado estud.	-3,025	2,728	-,023	1,079	48
Dist. de Mahalanobis	,057	45,288	2,937	7,862	48
Distancia de Cook	,000	27,990	,619	4,036	48
Valor de influencia centrado	,001	,964	,062	,167	48

a. Variable dependiente: n_asalariados

Tabla 4: Estadísticos sobre los residuos para el número de asalariados
Fuente: Elaboración Propia

En este caso, el rango de valores para la distancia de Mahalanobis es elevado y hay valores de la distancia de Cook superiores a $\frac{4}{48} = 0{,}083$ ya que como se puede ver su media es de 0,619.

Se elabora el gráfico 1 con el fin de detectar cuáles son las provincias influyentes, que como era de esperar resultan ser Madrid y Barcelona.

Así, en vista del gráfico se decide sacar del modelo a Madrid y Barcelona y se aplica de nuevo la regresión lineal múltiple obteniendo la tabla 5.

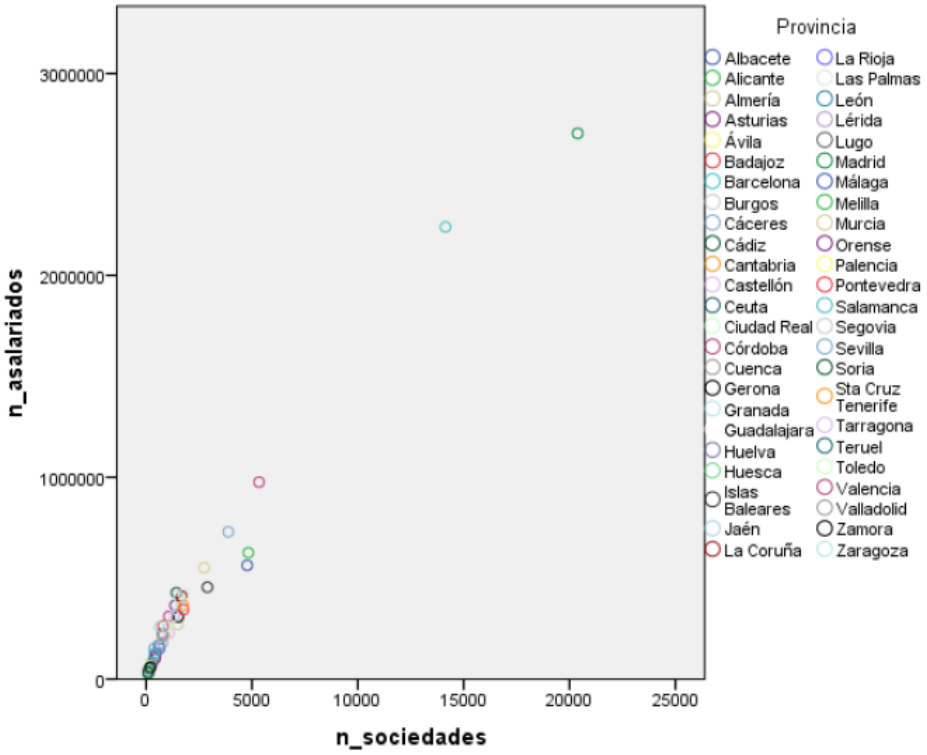

Figura 1: Número de sociedades frente a número de asalariados
Fuente: Elaboración Propia

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,942[a]	,888	,886	67977,074
2	,966[b]	,933	,929	53385,395
3	,973[c]	,946	,942	48421,898

a. Variables predictoras: (Constante), n_sociedades

b. Variables predictoras: (Constante), n_sociedades, Depósitos_ahorro

c. Variables predictoras: (Constante), n_sociedades, Depósitos_ahorro, IPC

d. Variable dependiente: n_asalariados

Tabla 5: Modelos finales en la estimación del número de asalariados
Fuente: Elaboración Propia

Se observa que el mejor modelo es el formado por las variables número de sociedades, depósitos de ahorro e IPC. Se obtiene de SPSS la tabla 6 que contiene los coeficientes estimados, los coeficientes VIF y la significatividad.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	81871.616	13827.682		5.921	.000		
	n_sociedades	147.114	7.871	.942	18.690	.000	1.000	1.000
2	(Constante)	32152.612	14323.216		2.245	.030		
	n_sociedades	100.839	10.666	.646	9.454	.000	.336	2.977
	Depósitos_ahorro	34.459	6.473	.364	5.324	.000	.336	2.977
3	(Constante)	4005876.209	1240206.880		3.230	.002		
	n_sociedades	87.646	10.515	.561	8.336	.000	.284	3.517
	Depósitos_ahorro	42.046	6.331	.444	6.642	.000	.289	3.462
	IPC	-38560.887	12034.268	-.126	-3.204	.003	.840	1.191

a. Variable dependiente: n_asalariados

Tabla 6: Coeficientes estimados para el número de asalariados
Fuente: Elaboración Propia

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	53027,13	990181,63	259918,37	195447,467	46
Valor pronosticado tip.	-1,059	3,736	,000	1,000	46
Error típico de valor pronosticado	8363,849	31663,396	13373,808	5057,928	46
Valor pronosticado corregido	54736,70	1000803,81	261714,89	199616,478	46
Residual	-106970,719	105913,945	,000	46779,998	46
Residuo típ.	-2,209	2,187	,000	,966	46
Residuo estud.	-2,525	2,364	-,017	1,039	46
Residuo eliminado	-147029,766	123875,977	-1796,525	54572,443	46
Residuo eliminado estud.	-2,709	2,508	-,016	1,075	46
Dist. de Mahalanobis	,364	18,263	2,935	3,547	46
Distancia de Cook	,000	,776	,047	,136	46
Valor de influencia centrado	,008	,406	,065	,079	46

a. Variable dependiente: n_asalariados

Tabla 7: Estadísticos finales sobre los residuos para el número de asalariados
Fuente: Elaboración Propia

Se comprueba que no existen problemas de colinealidad ya que el VIF es menor que 10, lo que nos vale para justificar la falta de problemas de colinealidad según Belsley et al. (1980) [1]. Además, se observa que todos los coeficientes son significativos. Veamos ahora si siguen existiendo puntos influyentes utilizando la tabla 7.

Se puede ver que ahora los valores de la distancia de Cook son inferiores a $\frac{4}{48} = 0{,}083$ ya que como se puede ver su media es de 0,047.

Se pasa por tanto al cálculo del número de asalariados para las provincias estudiadas aplicando

$$y = 4005876{,}209 + 87{,}646 n\_sociedades + 42{,}046 Depositos\_ahorro$$
$$+(-38560{,}887) IPC$$

De esta forma, se obtiene que el número de asalariados es 115.498 en Álava, 207.421 en Guipúzcoa, 385.823 en Vizcaya y 288.763 en Navarra. Se puede ver el resultado de la estimación en la tabla 9.

Se pasa ahora a aplicar la regresión lineal múltiple al modelo con variable dependiente el salario medio anual y se obtienen los modelos representados en la tabla 8.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.517[a]	.267	.251	2234.268
2	.621[b]	.385	.358	2068.344
3	.697[c]	.486	.451	1912.550

a. Predictores: (Constante), Depósitos_largo_plazo

b. Predictores: (Constante), Depósitos largo plazo, Var_anual

c. Predictores: (Constante), Depósitos largo plazo, Var_anual, n_sociedades

Tabla 8: Modelos en la estimación del salario medio
Fuente: Elaboración Propia

El R^2 ajustado del mejor modelo es 0.451, por lo que no es demasiado bueno el ajuste del modelo. Al tratarse de datos medios se va a optar por aplicar modelos lineales generalizados (GLM).

Provincia	Valores estimados	Valores observados
Álava	115498	0
Albacete	180757	152271
Alicante	733296	626323
Almería	280139	270807
Asturias	420746	365208
Ávila	77250	57518
Badajoz	253814	262900
Barcelona	2046623	2239578
Burgos	143688	146796
Cáceres	173916	152735
Cádiz	321904	427817
Cantabria	213373	220958
Castellón	285499	227933
Ceuta	53027	28457
Ciudad Real	215826	186798
Córdoba	248130	311065
Cuenca	124276	72888
Gerona	267136	308942
Granada	341131	336324
Guadalajara	104474	104030
Guipúzcoa	207421	0
Huelva	163793	227834
Huesca	113019	88631
Islas Baleare	430761	455110
Jaén	230393	259510
La Coruña	423678	410645
La Rioja	115203	128225
Las Palmas	314124	419869
León	198631	168732
Lérida	146511	179640
Lugo	127034	112718
Madrid	3036225	2703201
Málaga	661154	563642
Melilla	92812	27349
Murcia	468805	552349
Navarra	288763	0
Orense	135957	100335
Palencia	60481	63099
Pontevedra	348970	344678
Salamanca	137755	123768
Segovia	71780	61862
Sevilla	660334	729460
Soria	73536	35778
Sta Cruz Ten	315267	365094
Tarragona	258899	315373
Teruel	85676	51052
Toledo	273687	259776
Valencia	990184	975962
Valladolid	198049	214199
Vizcaya	385823	0
Zamora	60932	59344
Zaragoza	360466	402441

Tabla 9: Número de asalariados estimado frente a observado
Fuente: Elaboración Propia

3.3.2. Modelos Lineales Generalizados (GLM)

Los primeros trabajos donde se introduce y desarrolla el Modelo Lineal Generalizado son, respectivamente, Nelder y Wedderburn (1972) [11] y McCullagh y Nelder (1989) [12]. Para trabajar con GLM se han usado como referencia varios trabajos, entre ellos, López-González y Ruiz-Soler [8], Cañadas (2013) [3] y de Jong y Heller [5].

En un modelo lineal se asume que la variable dependiente sigue una distribución normal cuya media depende de una serie de variables explicativas y cuya varianza es constante: $y \sim N(\sum \beta_i x_i, \sigma^2)$.

En un GLM se generaliza este modelo en varias direcciones:

- La variable y sigue una distribución de probabilidad de la familia exponencial (normal, log-normal, poisson, gamma, inversa gaussiana,...).
- La esperanza de y (a la que denominamos $\mu = \mathbb{E}(y)$) ya no es directamente el predictor lineal $\eta = \sum \beta_i x_i$, sino que está relacionada con él a través de la función de enlace: $I(\mu) = \eta$. La función de enlace debe ser monótona y diferenciable.
- La varianza de y ya no es necesariamente constante, sino que es función de su media: $var(y) = \frac{\phi}{A} V(\mu)$. $V(\mu)$ se denomina función de varianza.

Los GLM son, por tanto, una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, poisson, gamma,...) y varianzas no constantes.

Según Cayuela [4] nos encontramos ante un supuesto GLM cuando la variable dependiente es:

- Un variable de conteo, ya sea expresada en términos absolutos (número de colisiones, accidentes,...) o relativos (porcentaje de heridos graves en accidentes, porcentaje de hombres...)
- Una variable binaria (vivo o muerto, hombre o mujer, joven o mayor...)

Por otro lado, según González (2014) [6] las distribuciones pertenecientes a la familia exponencial son una pieza clave en la construcción de un modelo GLM.

Las funciones de probabilidad pertenecientes a la familia exponencial son de la forma:

$$f(y) = c(y, \phi) exp\{\frac{y\theta - a(\theta)}{\phi}\}$$

donde

- θ es el parámetro canónico.
- ϕ es el parámetro de dispersión.
- La elección de $c(y, \theta)$ y $a(\phi)$ determinan la función de probabilidad de la variable respuesta, tales como la binomial, normal o gamma.

En términos de $a(\theta)$,

$$\mathbb{E}(y) = a'(\theta), \quad Var(y) = \phi a''(\theta)$$

donde $a'(\theta)$ y $a''(\theta)$ son, respectivamente, la primera y segunda derivada de $a(\theta)$ respecto a θ.

Se denominan funciones de enlace canónicas a las funciones que se aplican por defecto a cada una de las distribuciones de errores. Es recomendable comparar diferentes funciones de enlace para un mismo modelo y ver con cuál se obtiene un mejor ajuste del modelo a los datos.

Respecto a la construcción de modelos lineales generalizados, Cañuela [4] decía que es importante tener en cuenta que no existe un único modelo que sea válido. Normalmente, habrá varios modelos que puedan ajustarse al conjunto de datos tratado. Habrá que ver que modelos son adecuados y entre ellos cuál es el que explica la mayor proporción de la varianza sujeto a la restricción de que todos los parámetros del modelo deberán ser estadísticamente significativos. Esto se conoce como el modelo adecuado mínimo.

Los pasos que hay que seguir en la construcción y evaluación de un GLM según Cañuela [4] son los que se explican a continuación:

- **Exploración de los datos:** Según Tukey (1977) [18] es importante conocer bien nuestros datos. Puede ser interesante obtener gráficos que nos muestren la relación entre la variable explicada y cada una de las variables explicativas, gráficos de caja para variables categóricas, o matrices de correlación entre las variables explicativas. El objetivo de este análisis es:
  - Buscar posibles relaciones de la variable dependiente con las variables explicativas.

- Considerar la necesidad de aplicar transformaciones de las variables.
- Eliminar variables explicativas que estén altamente correlacionadas.

- **Elección de la estructura de errores y función de enlace:** Es recomendable comparar modelos con distintas funciones de enlace para ver cuál se ajusta mejor a los datos. Normalmente será a posteriori, al analizar los residuos, cuando se vea si es acertada la distribución de errores elegida.

- **Ajuste del modelo a los datos:** Debemos prestar particular atención a:
  - Los tests de significación para los estimadores del modelo.
  - La devianza, que es la varianza explicada por el modelo. Para calcularla hay que comparar la devianza D^2 del modelo nulo con la devianza residual, es decir, mide cuánto de la variabilidad de la variable dependiente es explicada por el modelo,

$$D^2 = \frac{Devianza\_modelo\_nulo - Devianza\_residual}{Devianza\_modelo\_nulo} \cdot 100$$

- **Criterios de evaluación de modelos:** Podemos utilizar la reducción de la devianza como una medida del ajuste del modelo a los datos. Los tests de significación para los parámetros del modelo son también útiles para ayudarnos a simplificar el modelo. Sin embargo, un criterio comúnmente utilizado es el Criterio de Información de Akaike (AIC), que es un índice que evalúa tanto el ajuste del modelo a los datos como la complejidad del modelo. Cuanto más pequeño es el AIC mejor es el ajuste. El AIC sirve para comparar modelos similares con distintos grados de complejidad o modelos iguales (mismas variables) pero con funciones de enlace diferentes.

- **Análisis de los residuos:** Los residuos son las diferencias entre los valores estimados por el modelo y los valores observados. Sin embargo, en muchos casos se utilizan los residuos estandarizados, que tienen que seguir una distribución normal. Conviene analizar los siguientes gráficos:
  - Histograma de los residuos.
  - Gráfico de residuos frente a valores estimados. Estos gráficos pueden indicar falta de linealidad, heterocedasticidad y valores atípicos.
  - El gráfico de normalidad (qq-plot), que permite contrastar la normalidad de la distribución de los residuos.

- **Simplificación del modelo:** El principio de parsimonia requiere que el modelo sea tan simple como sea posible. Esto significa que no debe contener parámetros de un factor que sean redundantes. La simplificación del modelo implica por tanto:
  - La eliminación de las variables explicativas que no sean significativas.
  - La agrupación de los parámetros de factores (variables categóricas) que no difieran entre sí. Esto significa que cada vez que simplificamos el modelo debemos repetir los dos puntos anteriores. La simplificación del modelo tiene que tener una lógica y no debe incrementar de manera significativa la devianza residual. Por tanto, es recomendable evitar los procedimientos automatizados (regresión stepwise, regresión backward/forward,...).

Se procede ahora a su aplicación para la estimación de los salarios medios anuales.

- **Exploración de los datos:** A primera vista observando los datos, no se puede apreciar que el salario medio anual esté asociado claramente a alguna de las variables utilizadas en el modelo.

- **Elección de las estructura de errores y función vínculo o de enlace:** Como la variable respuesta es una media se probará con dos familias de distribución de errores, que serán la inversa gaussiana y la gamma. En ambos casos se utilizará la función de enlace logarítmica por ser la que mejor suele funcionar en estos casos.

  Estas familias son muy útiles con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante.

- **Ajuste del modelo a los datos:** Se introduce el siguiente código en R:

```
mod_gaussian <- glm(salario_medio_anual ~ n_sociedades +
cap_sociedades + IPC + Var_anual + Depositos_vista +
Depositos_ahorro + Depositos_largo_plazo,
family=inverse.gaussian(link="log"),
data=subset(BBDD_riqueza,salario_medio_anual>0))

summary(mod_gaussian)
```

  Y se obtienen los siguientes resultados:

```
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.760e-03  -3.693e-04   9.287e-05   5.631e-04   1.179e-03
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.023e+01  3.042e+00   3.362  0.00171 **
## n_sociedades           -5.067e-05  2.596e-05  -1.952  0.05796 .
## cap_sociedades         -9.621e-08  1.927e-07  -0.499  0.62034
## IPC                    -2.698e-03  2.928e-02  -0.092  0.92702
## Var_anual               1.866e-01  6.382e-02   2.924  0.00566 **
## Depositos_vista        -2.080e-05  2.656e-05  -0.783  0.43819
## Depositos_ahorro       -1.440e-05  2.056e-05  -0.700  0.48783
## Depositos_largo_plazo   2.290e-05  1.390e-05   1.648  0.10728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be
## 7.742539e-07)
##
##     Null deviance: 6.3693e-05  on 47  degrees of freedom
## Residual deviance: 3.5369e-05  on 40  degrees of freedom
## AIC: 877.31
##
## Number of Fisher Scoring iterations: 4

Anova(mod_gaussian, test ="F")

## Analysis of Deviance Table (Type II tests)
##
## Response: salario_medio_anual
## Error estimate based on Pearson residuals
##
##                               SS Df      F   Pr(>F)
## n_sociedades          2.8496e-06  1 3.6806 0.062206 .
## cap_sociedades        1.7790e-07  1 0.2298 0.634251
## IPC                   6.8000e-09  1 0.0088 0.925891
## Var_anual             7.3839e-06  1 9.5372 0.003652 **
## Depositos_vista       4.7910e-07  1 0.6188 0.436139
```

```
## Depositos_ahorro      3.9540e-07  1 0.5107 0.478972
## Depositos_largo_plazo 2.1872e-06  1 2.8250 0.100599
## Residuals             3.0969e-05 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se hace lo mismo con la distribución gamma, para lo cual se introduce el siguiente código:

```
mod_gamma <- glm(salario_medio_anual ~ n_sociedades +
cap_sociedades + IPC + Var_anual + Depositos_vista +
Depositos_ahorro + Depositos_largo_plazo,
family=Gamma(link="log"),
data=subset(BBDD_riqueza, salario_medio_anual>0))

summary(mod_gamma)
```

Y se obtienen los siguientes resultados:

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.34127  -0.04848   0.01095   0.07479   0.15821
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.057e+01  2.987e+00   3.540  0.00103 **
## n_sociedades          -5.315e-05  2.630e-05  -2.021  0.05001 .
## cap_sociedades        -7.649e-08  1.849e-07  -0.414  0.68135
## IPC                   -5.916e-03  2.876e-02  -0.206  0.83805
## Var_anual              1.952e-01  6.327e-02   3.085  0.00368 **
## Depositos_vista       -2.340e-05  2.686e-05  -0.871  0.38894
## Depositos_ahorro      -1.566e-05  2.076e-05  -0.754  0.45529
## Depositos_largo_plazo  2.442e-05  1.409e-05   1.734  0.09071 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.01295672)
##
##     Null deviance: 1.07043  on 47  degrees of freedom
```

```
## Residual deviance: 0.56348  on 40  degrees of freedom
## AIC: 874.89
##
## Number of Fisher Scoring iterations: 4

Anova(mod_gamma, test = "F")

## Analysis of Deviance Table (Type II tests)
##
## Response: salario_medio_anual
## Error estimate based on Pearson residuals
##
##                            SS Df       F   Pr(>F)
## n_sociedades          0.05251  1  4.0525 0.050871 .
## cap_sociedades        0.00215  1  0.1656 0.686264
## IPC                   0.00056  1  0.0432 0.836408
## Var_anual             0.13013  1 10.0438 0.002928 **
## Depositos_vista       0.00985  1  0.7600 0.388518
## Depositos_ahorro      0.00754  1  0.5821 0.449951
## Depositos_largo_plazo 0.03963  1  3.0589 0.087967 .
## Residuals             0.51827 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Siguiendo el criterio de Akaike, como el AIC del GLM con la inversa gaussiana es 877.31 y el AIC del GLM con la gamma es 874.89, se opta por utilizar el segundo modelo.

Para este segundo modelo, se puede ver que únicamente los coeficientes de las variables número de sociedades, variación anual del IPC y depósitos a largo plazo son significativos, es decir, su $Pr(> |t|) < 0{,}1$. En vista de esto, se opta por eliminar una a una las variables explicativas que no son significativas, hasta llegar a un modelo formado únicamente por las tres variables citadas, que es el único modelo que cumple que todas sus variables son significativas.

```
mod_gamma2 <- glm(salario_medio_anual ~ n_sociedades  +
Var_anual + Depositos_largo_plazo,
family=Gamma(link="log"),
data=subset(BBDD_riqueza, salario_medio_anual>0))
```

```r
summary(mod_gamma2)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.34659  -0.04527   0.00687   0.07681   0.14402
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.982e+00  7.371e-02 135.419  < 2e-16 ***
## n_sociedades          -5.126e-05  1.795e-05  -2.855 0.006539 **
## Var_anual              2.091e-01  5.730e-02   3.649 0.000694 ***
## Depositos_largo_plazo 7.964e-06  2.182e-06   3.650 0.000692 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.0125297)
##
##     Null deviance: 1.07043  on 47  degrees of freedom
## Residual deviance: 0.60575  on 44  degrees of freedom
## AIC: 870.37
##
## Number of Fisher Scoring iterations: 4
```

```r
Anova(mod_gamma2, test = "F")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: salario_medio_anual
## Error estimate based on Pearson residuals
##
##                              SS Df       F    Pr(>F)
## n_sociedades           0.10217  1  8.1539 0.0065329 **
## Var_anual              0.17465  1 13.9392 0.0005391 ***
## Depositos_largo_plazo 0.16618  1 13.2626 0.0007095 ***
## Residuals              0.55131 44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finalmente, para el modelo formado por las variables número de sociedades, variación anual del IPC y depósitos a largo plazo se tiene que todos los coeficientes son significativos, es decir, su $Pr(>|t|) < 0{,}01$.

La varianza explicada por este modelo es un 43,41 por ciento, ya que

$$D^2 = \frac{1{,}07043 - 0{,}60575}{1{,}07043} 100 = 43{,}41$$

- **Simplificación del modelo:** Como las variables de depósitos no son todas significativas, se podría haber optado por juntarlas en una sola variable y volver a hacer el análisis. Finalmente, en función de que esto podría desvirtuar el modelo debido a que existen grandes diferencias conceptuales entre los tres tipos de depósito se ha optado por no hacer modificaciones en el modelo.

Finalmente, se calcula el salario medio anual de las provincias vascas y Navarra de la siguiente manera:

```
predict(mod_gamma, newdata=BBDD_riqueza,type="response")
```

```
##        1        2        3        4        5        6        7
## 19149.73 16383.24 15681.40 15669.00 18026.86 15558.15 16378.37
##       15       16       17       18       19       20       21
## 15337.24 17121.71 15330.70 16682.77 14877.49 17305.39 19640.76
##       22       23       24       25       26       27       28
## 15868.64 15752.15 16691.59 17138.16 17503.34 18228.12 17054.38
##       29       30       31       32       33       34       35
## 16316.29 16664.61 16837.43 24438.65 14564.51 18937.72 16693.34
##       36       37       38       39       40       41       42
## 17558.98 17643.43 17848.44 16886.33 17012.78 17076.91 13730.38
##       43       44       45       46       47       48       49
## 16785.84 15197.66 16112.35 16039.06 15488.50 16298.05 17896.00
##       50       51       52
## 16855.88 17173.18 19395.43
```

De esta forma, se obtiene que el salario medio anual para Álava es 19.150 euros, el de Guipúzcoa 19.641 euros, el de Vizcaya 16.856 euros y el de Navarra 17.559 euros. En la tabla 10 se puede ver la diferencia entre los valores estimados y los valores observados.

Provincia	Valores estimados	Valores observados
Álava	19150	
Albacete	16383	15446
Alicante	15681	15208
Almería	15669	13160
Asturias	18027	19528
Ávila	15558	15853
Badajoz	16378	13308
Barcelona	23010	21775
Burgos	18956	19578
Cáceres	17389	13991
Cádiz	15852	15418
Cantabria	16301	18980
Castellón	16497	16767
Ceuta	19197	22397
Ciudad Real	15337	15040
Córdoba	17122	13096
Cuenca	15331	14583
Gerona	16683	17627
Granada	14877	14537
Guadalajara	17305	19620
Guipúzcoa	19641	
Huelva	15869	12289
Huesca	15752	17155
Islas Baleares	16692	17247
Jaén	17138	11935
La Coruña	17503	18892
La Rioja	18228	18354
Las Palmas	17054	16247
León	16316	17715
Lérida	16665	16905
Lugo	16837	16724
Madrid	24439	24576
Málaga	14565	15128
Melilla	18938	21388
Murcia	16693	15621
Navarra	17559	
Orense	17643	16552
Palencia	17848	17754
Pontevedra	16886	17014
Salamanca	17013	17605
Segovia	17077	16818
Sevilla	13730	15580
Soria	16786	18112
Sta Cruz Tenerife	15198	15705
Tarragona	16112	18343
Teruel	16039	17114
Toledo	15489	16066
Valencia	16298	17746
Valladolid	17896	19254
Vizcaya	16856	
Zamora	17173	15623
Zaragoza	19395	19504

Tabla 10: Salarios medios estimados frente a observados
Fuente: Elaboración Propia

4. Clasificación de los datos

La base de datos estaba formada por variables numéricas, salvo la variable provincia que es de tipo cadena. A excepción de las tasas de variación del porcentaje de robo y del salario medio anual, el resto de valores estaban en valor absoluto. Ante esto, el primer paso fue estandarizar todas las variables. De esta forma, todas las variables estaban en la misma escala para poder trabajar con ellas. Ahora bien, en vista de los resultados se observó que las grandes provincias como Madrid y Barcelona se alejaban mucho del resto ya que sus valores siempre eran más altos. Ante esta situación se optó por trabajar en términos relativos. Así, como primera medida se opta por dividir todas las variables absolutas por el número de habitantes. Tras este cambio, se obtienen unos resultados más objetivos. Ahora bien, ¿son lo suficientemente precisos? Llegados a este punto se decide hacer un estudio basado en el uso de distintos tipos de ratios. Con el fin de ser lo más precisos posible, se opta por crear "estructuras" dentro de la base de datos. Así, las variables sobre los tipos de vehículos no se dividen entre el número de habitantes sino sobre el total de vehículos, de forma que se obtiene un bloque llamado parque de vehículos. Esto mismo se hace con las variables de los tipos de carreteras. Todas ellas se dividen entre el total de kilómetros de carreteras. Este tratamiento se ha realizado para todas las variables que se ha creído correspondiente. A través de estas estructuras se podrá realizar el estudio con mayor precisión. La comparación entre los resultados del estudio con los distintos tipos de ratios se verá más adelante.

Se va a utilizar el análisis de cluster como técnica de clasificación para ver cómo se agrupan las provincias.

El análisis de cluster es una técnica de clasificación, que dice si los elementos de nuestra muestra forman o no un grupo homogéneo, y en caso de que no lo formen identifican que elementos pertenecen a cada uno de los grupos existentes. O lo que es lo mismo, este método agrupa elementos en grupos homogéneos en función de las similitudes entre ellos.

Es importante destacar que en esta técnica no se parte de un conocimiento previo de los grupos y que se busca homogeneidad dentro de los grupos y heterogeneidad entre grupos.

El análisis de cluster estudia tres tipos de problemas según Peña (2002) [14]:

- **Partición de los datos:** Se dispone de datos a priori heterogéneos y se quieren dividir en un número de grupos prefijado, de forma que cada elemento pertenezca a uno y solo uno de los grupos, que todo elemento quede clasificado y que cada grupo sea internamente homogéneo.

- **Construcción de jerarquías:** El objetivo es estructurar los elementos de una muestra de forma jerárquica por su similitud. En una clasificación jerárquica los datos se ordenan por niveles, de forma que los niveles superiores contienen a los inferiores. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Sin embargo, a partir de la jerarquía construida se puede obtener también una partición de los datos en grupos.

- **Clasificación de variables:** Las variables pueden clasificarse en grupos o estructurarse en una jerarquía. Los métodos de partición utilizan la matriz de datos, pero los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. Para agrupar variables se parte de la matriz de relación entre variables. En el caso de variables continuas suele ser la matriz de correlación, y en el caso de variables discretas, se construye a partir de la distancia ji-cuadrado.

Respecto a las técnicas para encontrar cluster, se tienen los métodos jerárquicos y los métodos no jerárquicos.

Los métodos jerárquicos consisten en partir los elementos uniendo o separando cluster. En cada paso se juntan o separan dos cluster según el criterio especificado. Dentro de éstos están los métodos aglomerativos; que son los más habituales, y los métodos divisivos. Los métodos aglomerativos parten de tantos cluster como datos tiene la muestra y en cada paso se unen dos cluster según el criterio especificado hasta obtener un único cluster con todos los datos. Los métodos divisivos necesitan de más cálculos, ya que parten de un único cluster formado por todos los datos que se va dividiendo a cada paso hasta obtener tantos cluster como datos.

Los métodos no jerárquicos consisten en partir los elementos en un número prefijado de cluster siguiendo un criterio de optimización. El más utilizado es el algoritmo de k-medias introducido por MacQueen (1967) [10], ya que es fácil de programar y se obtienen unos resultados razonablemente buenos.

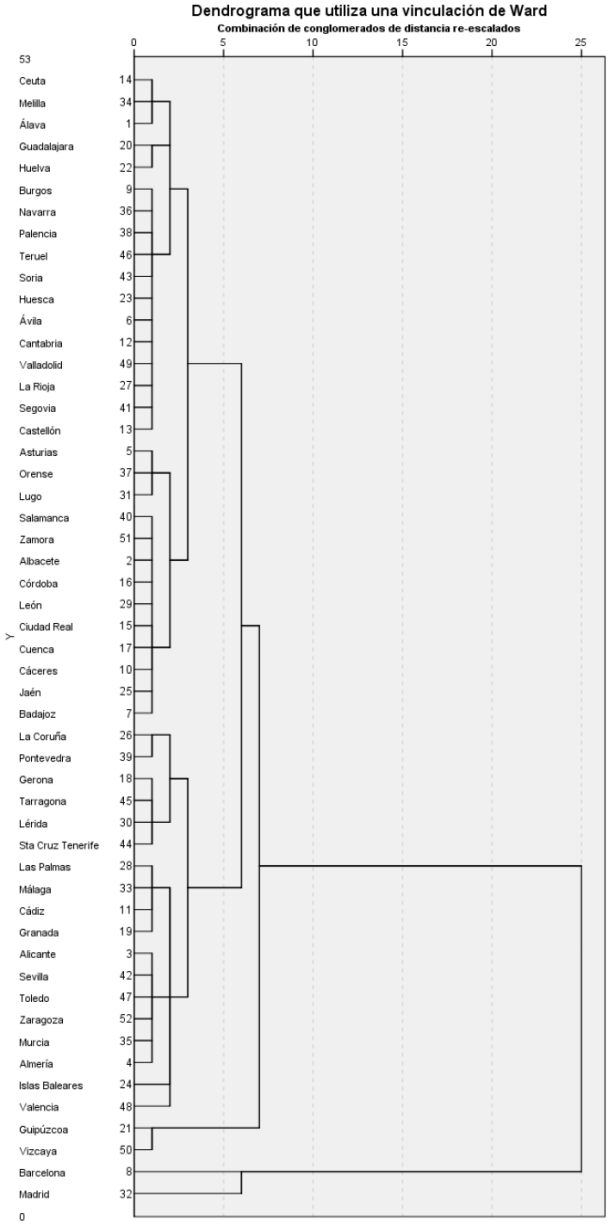

Figura 2: Dendrograma sin utilizar ratios
Fuente: Elaboración Propia

Como se ha dicho, una vez ya está completamente elaborada la base de datos se pasa a aplicar el análisis de cluster. Se opta por utilizar métodos jerárquicos, en concreto, métodos aglomerativos. Es muy importante que los datos a los que se aplica el análisis de cluster estén bien preparados. Veamos un ejemplo de mala praxis,

como sería hacer simplemente la estandarización de los datos sin haberlos observado antes. El dendrograma 2 es un buen ejemplo de ello. Se puede observar como Madrid y Barcelona se unen mucho más lejos que el resto de provincias. Para evitar esto se usan ratios. Así, a partir de la base de datos se preparan dos nuevas bases de datos. En la primera se trabajará con ratios que saldrán de dividir todas las variables por el número de habitantes, mientras que en la segunda se trabajará con "estructuras". En concreto, tendremos 6 "estructuras": accidentes según la vía, parque de vehículos, red de carreteras, población, robos y primas. Las variables de la estructura de accidentes según la vía y las de población serán divididas por el número de habitantes. Las variables pertenecientes a la estructura parque de vehículos se dividirán entre el número total de vehículos, las pertenecientes a la estructura de la red de carreteras entre el número total de kilómetros y las pertenecientes a la estructura de primas entre el número de conductores. Por último, en el caso de la estructura de robos, la variable de robo de vehículos se dividirá entre el número total de vehículos y la variable de robo en viviendas se dividirá entre el número de viviendas.

Al aplicar el análisis de cluster, se observa que no existen diferencias significativas entre trabajar con una u otra base de datos. Será por ello por lo que se opte por trabajar con sólo una de ellas.

En este punto, hay que decidir qué enlace utilizar y qué distancia aplicar. Según Pardo y Ruiz (2005) [13] los métodos de conglomeración son los procedimientos que permiten volver a calcular las distancias entre los nuevos elementos en cada etapa del proceso de fusión. Existen distintos tipos de métodos:

- **Método de vinculación por el vecino más próximo:** Selecciona los dos elementos de la matriz de distancias que se encuentran más próximos. La distancia de este nuevo conglomerado respecto al resto de elementos de la matriz será la menor de las distancias entre cada elemento del conglomerado y el resto de elementos de la matriz. De esta forma, la distancia d_{AB} entre los conglomerados A y B se calcula mediante:

  $$d_{AB} = min(d_{ij})$$

  donde d_{ij} es la distancia entre los elementos i y j; el primero perteneciente al conglomerado A y el segundo al conglomerado B.

- **Método de vinculación por el vecino más lejano:** Es opuesto al anterior. La distancia entre dos conglomerados A y B se calcula como la distancia entre sus dos elementos más lejanos

  $$d_{AB} = máx(d_{ij})$$

- **Método de vinculación inter-grupos:** La distancia entre dos conglomerados se calcula como la distancia promedio entre todos los pares de elementos de ambos conglomerados

$$d_{AB} = \frac{1}{n_A n_B} \sum_{I \in A} \sum_{j \in B} d_{ij}$$

- **Método de Ward:** Se calcula en cada conglomerado el vector de medias de todas las variables, es decir, el centroide multivariante. Seguidamente, se calculan las distancias euclídeas al cuadrado entre cada elemento y los centroides de todos los conglomerados. Y por último, se suman las distancias correspondientes a todos los elementos.

  En cada paso se unen los conglomerados que producen un menor incremento de la suma de cuadrados de las distancias intra-conglomerados. Esta suma se define como:

$$SCE = \sum_{j=1}^{k} (\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} (\sum_{i=1}^{n_j} X_{ij})^2)$$

- **Método de agrupación de centroides:** Se calcula la distancia entre dos conglomerados como la distancia entre sus vectores de medias. La distancia entre el conglomerado AB y el conglomerado o elemento C se calcula como

$$d_{(AB)C} = \frac{n_A}{n_A + n_B} d_{AC} + \frac{n_B}{n_A + n_B} d_{BC} - \frac{n_A n_B}{(n_A + n_B)^2} d_{BC}$$

- **Método de agrupación de medianas:** Los dos conglomerados que se combinan reciben idéntica ponderación en el cálculo del nuevo centroide combinado, independientemente del tamaño de cada uno de los conglomerados. Dado un conglomerado AB y un elemento C, la nueva distancia del conglomerado al elemento se calcula como

$$d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2} - \frac{d_{AB}}{4}$$

El otro aspecto clave es la medida de distancia a utilizar para cuantificar la distancia entre los elementos. Estas medidas pueden ser de similaridad o disimilaridad. Las medidas de similaridad evalúan el grado de proximidad existente entre dos elementos, mientras que las medidas de disimilaridad evalúan el grado de lejanía entre dos elementos. Existen varios tipos de medidas de distancia:

- **Distancia euclídea:** Es una medida de disimilaridad que se calcula como

$$EUCLID(X,Y) = \sqrt{\sum_i (X_i - Y_i)^2}$$

- **Distancia euclídea al cuadrado:** Es una medida de disimilaridad cuyo cálculo se hace como

$$SEUCLID(X,Y) = \sum_i (X_i - Y_i)^2$$

- **Coseno:** Es una medida de disimilaridad que puede calcularse como

$$COSINE(X,Y) = \frac{\sum_i^n X_i Y_i}{\sqrt{(\sum_i^n X_i^2)(\sum_i^n Y_i^2)}}$$

- **Correlación de Pearson:** Es una medida de disimilaridad cuyo cálculo se hace como

$$CORRELATION(X,Y) = \frac{\sum_i^n z_{x_i} z_{y_i}}{n-1}$$

  donde n es el tamaño de la muestra y z_x y z_y son las puntuaciones tipificadas del sujeto i en las variables X e Y, que son las variables entre las que se calcula la distancia.

- **Chebychev:** Es una medida de disimilaridad que se calcula como

$$CHEBYCHEV(X,Y) = máx_i|X_i - Y_i|$$

- **Bloques:** Es una medida de disimilaridad que puede calcularse como

$$BLOCK(X,Y) = \sum_i |X_i - Y_i|$$

- **Minkowsky:** Es una medida de disimilaridad que se calcula como

$$MINKOWSKY(X,Y) = (\sum_i |X_i - Y_i|^p)^{\frac{1}{p}}$$

  donde p es cualquier número entero positivo.

- **Personalizada:** Es una medida de disimilaridad cuyo cálculo se hace como

$$POWER(X,Y) = (\sum_i |X_i - Y_i|^p)^{\frac{1}{r}}$$

donde p y r son dos números enteros positivos cualquiera.

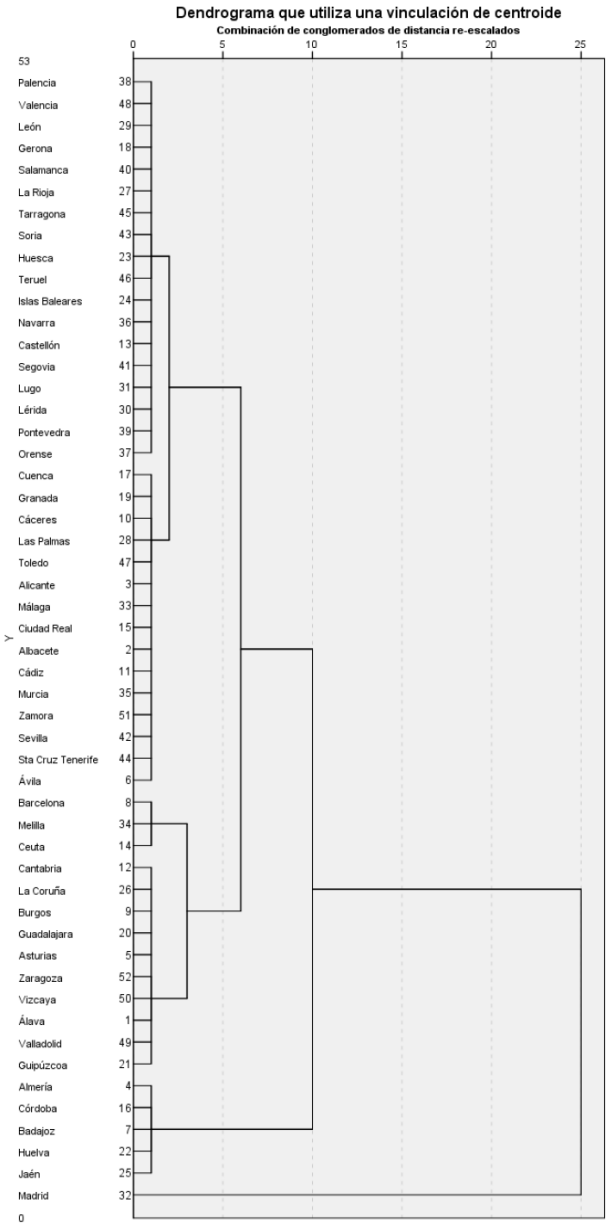

Figura 3: Dendrograma con enlace de centroides y distancia euclídea al cuadrado
Fuente: Elaboración Propia

Se empieza probando con el enlace de centroides y la distancia euclídea al cuadrado y se obtiene el dendrograma 3.

Aplicando el análisis de cluster con la distancia euclídea al cuadrado y el enlace inter-grupos, se obtiene el dendrograma 4 en SPSS.

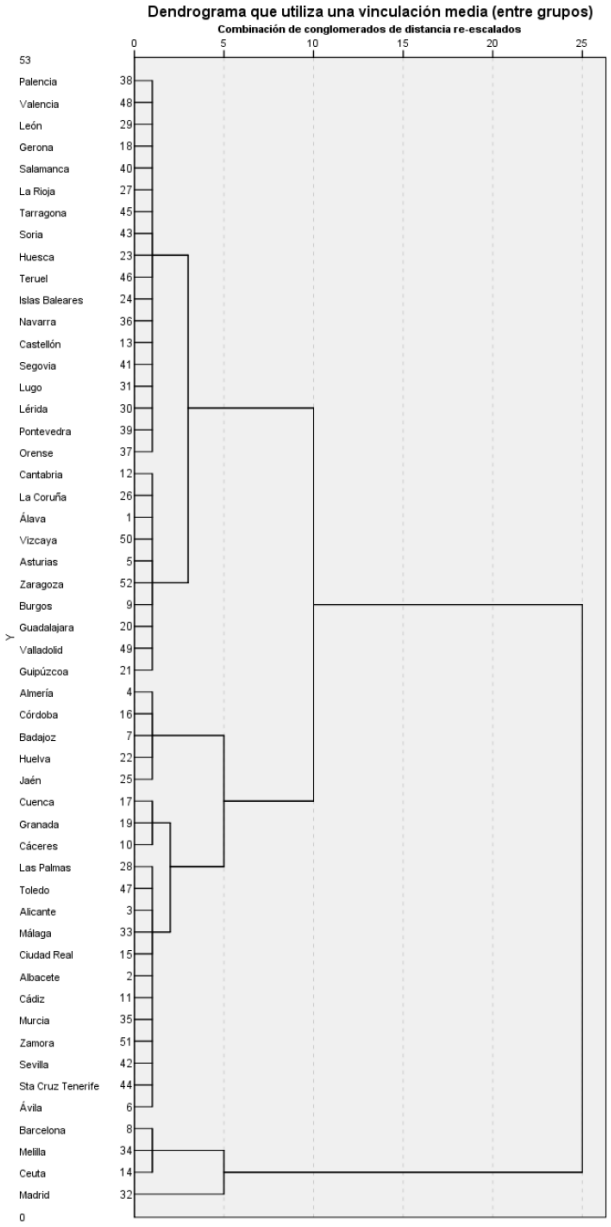

Figura 4: Dendrograma con enlace inter-grupos y distancia euclídea al cuadrado
Fuente: Elaboración Propia

Si ahora se aplica el análisis de cluster con el método de Ward y con la distancia euclídea, se obtiene el dendrograma 5 en SPSS.

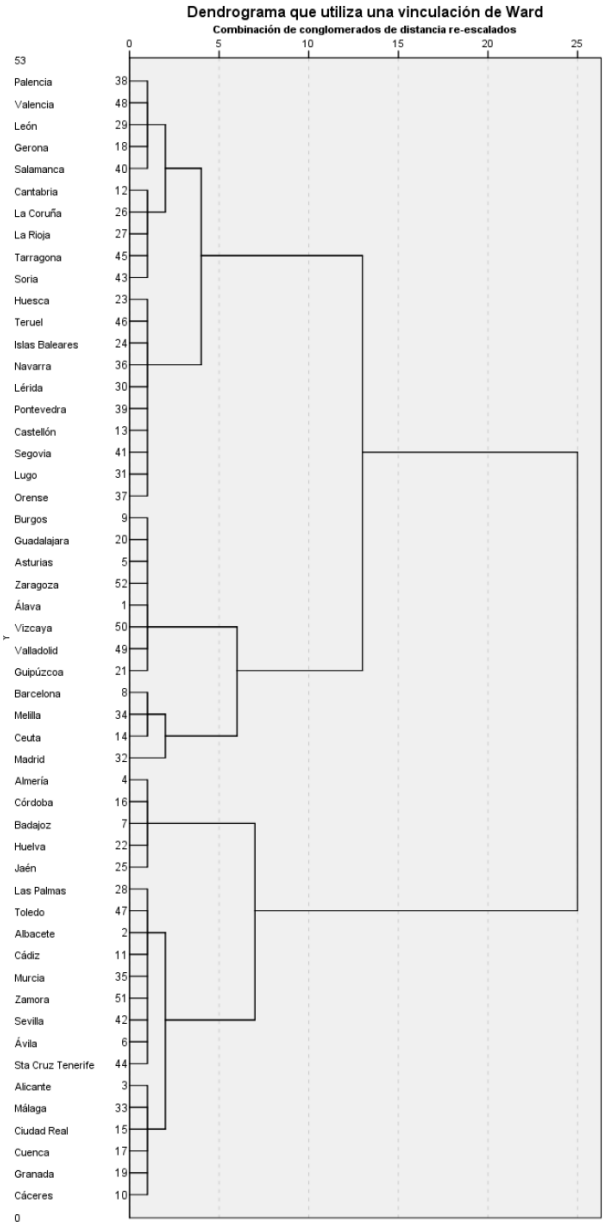

Figura 5: Dendrograma con método de Ward y distancia euclídea
Fuente: Elaboración Propia

Y si en vez de con la distancia euclídea, se aplica con la distancia euclídea al cuadrado, se obtiene en SPSS el dendrograma 6.

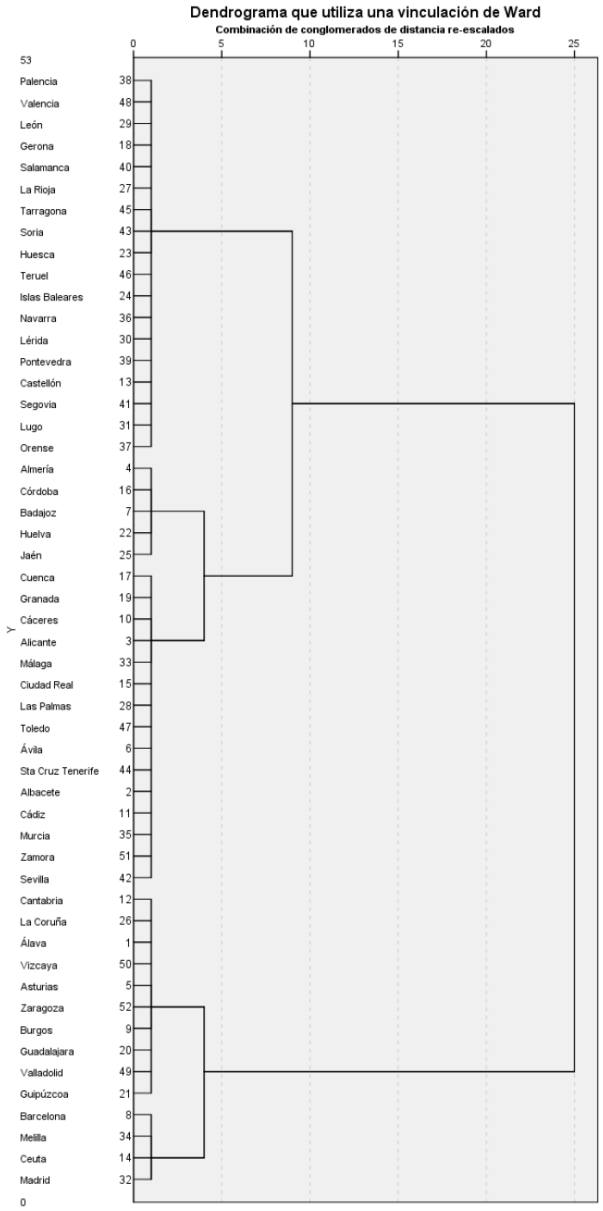

Figura 6: Dendrograma con método de Ward y distancia euclídea al cuadrado
Fuente: Elaboración Propia

Finalmente, en vista de los dendrogramas, se optará por aplicar el análisis de cluster utilizando el método de Ward y la distancia euclídea al cuadrado, ya que se obtienen 5 cluster claramente diferenciados. Además, resulta destacable que todas las provincias se unen muy pronto, es decir, los grupos son muy homogéneos entre sí.

Estos cluster son:

- **Cluster 1:** Está compuesto por Palencia, Valencia, León, Gerona, Salamanca, La Rioja, Tarragona, Soria, Huesca, Teruel, Islas Baleares, Navarra, Lérida, Pontevedra, Castellón, Segovia, Lugo y Orense. Este cluster está formado entre otras por las provincias catalanas (excepto Barcelona), las provincias de la Comunidad Valenciana (excepto Alicante), las Islas Baleares, las provincias aragonesas (excepto Zaragoza), La Rioja y Navarra. Podría decirse que está compuesto por las provincias del Noreste de España.

- **Cluster 2:** Está formado por Almería, Córdoba, Badajoz, Huelva y Jaén. Este cluster está formado por provincias del Suroeste de España con la excepción de Almería y Jaén.

- **Cluster 3:** Se compone de Cuenca, Granada, Cáceres, Alicante, Málaga, Ciudad Real, Las Palmas, Toledo, Ávila, Santa Cruz de Tenerife, Albacete, Cádiz, Murcia, Zamora y Sevilla. Este cluster está formado por las Islas Canarias y varias provincias andaluzas, extremeñas y castellano-manchegas. De esta forma, se podría decir que está compuesto por provincias del Sur de España.

- **Cluster 4:** Está compuesto por Cantabria, La Coruña, Álava, Vizcaya, Asturias, Zaragoza, Burgos, Guadalajara, Valladolid y Guipúzcoa. Podríamos decir que este cluster se compone en su mayoría por provincias del Norte de España, con la excepción de Valladolid y Guadalajara.

- **Cluster 5:** Está formado por Madrid, Barcelona, Ceuta y Melilla. Este cluster está compuesto por las dos ciudades más grandes de España y por los casos particulares Ceuta y Melilla.

En el gráfico 7 se puede ver el perfil de los cluster en función de algunas de sus variables.

- **Cluster 1:** Se observa que tiene un alto número de kilómetros de carreteras con calzada menor que 5 metros y un alto salario medio mientras que el número de robo de vehículos y de kilómetros de autopistas y autovías libres es bajo. Además tiene el menor número de camiones y furgonetas y la mayor prima media.

- **Cluster 2:** Se observan valores bajos en el número de accidentes con víctimas en calles, en el de kilómetros de carreteras con calzada menor que 5 metros y en el salario medio anual y altos en el número de camiones y furgonetas. Además tiene el el valor más alto en el número de autovías y autopistas libres y la prima media más baja.

- **Cluster 3:** Se puede ver que la mayoría de sus valores están por debajo de la media, como es el caso del número de motocicletas y de la prima media. Tiene el valor más alto en el número de camiones y furgonetas y el más bajo en el número de accidentes con víctimas en autopistas y en calles y en el salario medio anual.
- **Cluster 4:** Se observa que tiene el valor más alto en el número de accidentes con víctimas en autopistas y en calles, en el número de motocicletas y en el salario medio anual y el valor más bajo en el número de kilómetros de carretera con calzada menor que 5 metros. Además, tiene valores altos en robo de vehículos y bajos en el número de camiones y furgonetas y en la prima media.
- **Cluster 5:** Se puede ver que todos sus valores están bastante cercanos a la media. Tiene valores altos en el número de accidentes con víctimas en autopistas y en la prima media y valores más bajos en el número de accidentes con víctimas en calles, de kilómetros con autopistas y autovías libres y de robos de vehículos. Además, tiene el valor más alto en el número de kilómetros de carreteras con calzada menor que 5 metros.

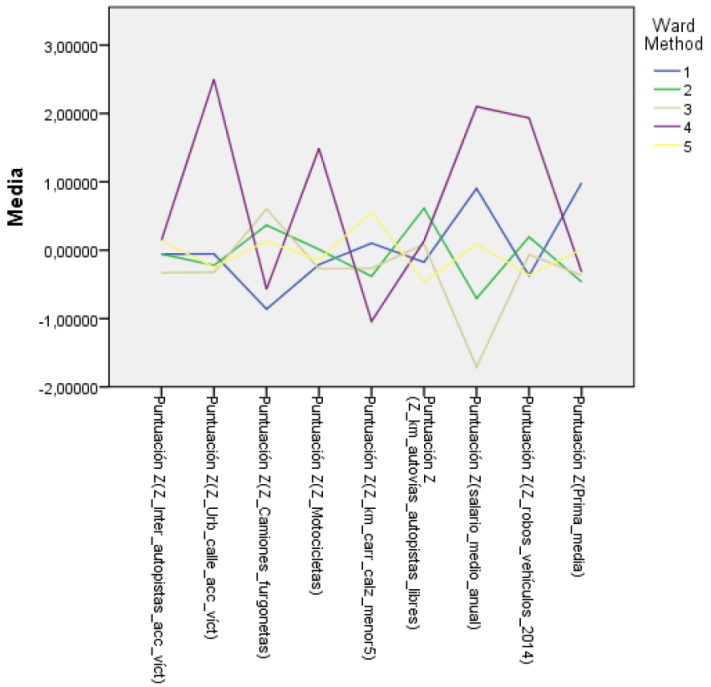

Figura 7: Gráfico del perfil de los cluster para algunas variables
Fuente: Elaboración Propia

5. Estimación de la prima del ramo de autos

Se empezará tomando el importe total de primas de autos como variable dependiente. SPSS presenta la tabla 11 con el R^2 ajustado de los mejores modelos para distinto número de variables.

Resumen del modelo[i]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,985[a]	,969	,969	59710533,702
2	,998[b]	,996	,995	22877120,942
3	,999[c]	,997	,997	17987994,819
4	,999[d]	,998	,998	15909416,712
5	,999[e]	,998	,998	14958406,006
6	,999[f]	,999	,998	13782212,565
7	,999[g]	,999	,999	12259786,224
8	1,000[h]	,999	,999	11403872,967

a. Variables predictoras: (Constante), Turismos

b. Variables predictoras: (Constante), Turismos, Urb_autourb_her_no_hosp

c. Variables predictoras: (Constante), Turismos, Urb_autourb_her_no_hosp, n_conductores

Tabla 11: Modelos en la estimación de la prima total de autos
Fuente: Elaboración Propia

Se puede ver que se obtiene un R^2 ajustado muy alto para todos los modelos, por lo que en principio los modelos son buenos. Ahora bien, hay que fijarse en el valor de la distancia de Mahalanobis y en si la distancia de Cook es mayor que $\frac{4}{n}$ donde n es el número de observaciones, para detectar si existen puntos influyentes. Esto se puede observar en la tabla 12, donde se ve que el rango de valores para la distancia de Mahalanobis es elevado y hay valores de la distancia de Cook superiores a $\frac{4}{52} = 0{,}077$ ya que como se puede ver su media es de 6,12.

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	12436255,00	2210455808,00	218996738,83	337440118,195	52
Valor pronosticado tip.	-,612	5,902	,000	1,000	52
Error típico de valor pronosticado	2280748,250	11400752,000	4353395,035	1904217,994	52
Valor pronosticado corregido	12424464,00	1602335360,00	207205272,81	269837018,803	52
Residual	-25984740,000	28053186,000	,000	10471321,903	52
Residuo típ.	-2,279	2,460	,000	,918	52
Residuo estud.	-2,886	2,549	,021	1,044	52
Residuo eliminado	-44149784,000	608453376,000	11791466,021	85464815,532	52
Residuo eliminado estud.	-3,177	2,735	,020	1,091	52
Dist. de Mahalanobis	1,059	49,991	7,846	9,430	52
Distancia de Cook	,000	316,133	6,120	43,834	52
Valor de influencia centrado	,021	,980	,154	,185	52

a. Variable dependiente: Primas_Autos_Total

Tabla 12: Estadísticos sobre los residuos para la prima total de autos
Fuente: Elaboración Propia

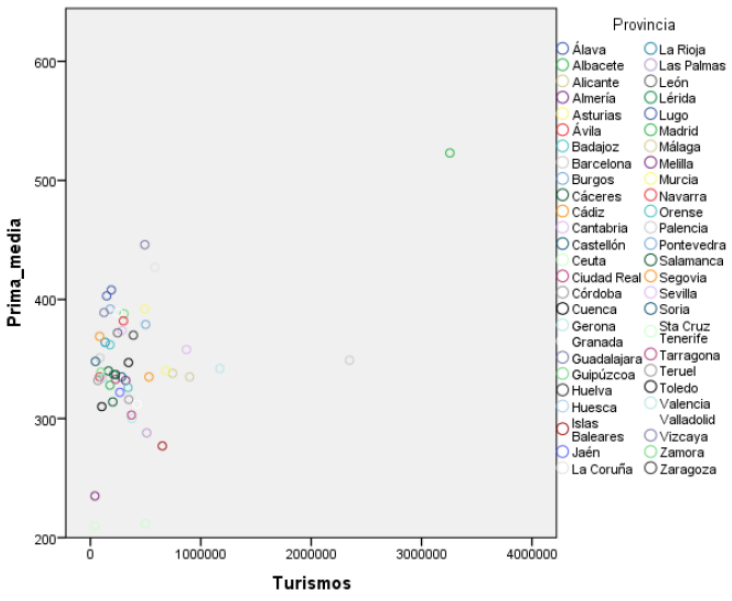

Figura 8: Número de turismos frente a prima media
Fuente: Elaboración Propia

Se crea el gráfico 8 para detectar qué provincias son influyentes.

A la vista del gráfico, se opta por sacar del modelo a Madrid y Barcelona y se vuelve a aplicar la regresión lineal múltiple obteniendo la tablas 13 y 14 donde se observa que el único modelo en el que todas las variables son significativas es el formado por las variables kilómetros de autopistas de peaje, turismos, heridos no hospitalizados en autopistas, fallecidos en autovías, heridos no hospitalizados en otros tipos de vías, kilómetros de carreteras con doble calzada, camiones y furgonetas, robos totales de vehículos en 2014, kilómetros de carreteras con calzada menor de 5 metros, número de habitantes y heridos hospitalizados en calles.

Resumen del modelo[r]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,987[a]	,974	,973	19069445,892
2	,990[b]	,981	,980	16365356,193
3	,993[c]	,985	,985	14433891,518
4	,995[d]	,990	,989	12145720,132
5	,996[e]	,992	,992	10613793,857
6	,997[f]	,994	,993	9664312,483
7	,998[g]	,995	,995	8492715,422
8	,998[h]	,996	,996	7576282,187
9	,999[i]	,997	,997	6866221,378
10	,999[j]	,997	,996	6874415,097
11	,999[k]	,997	,997	6571522,414
12	,999[l]	,998	,997	6194222,001
13	,999[m]	,998	,998	5656807,918
14	,999[n]	,998	,998	5405802,356
15	,999[o]	,999	,998	5036128,121
16	,999[p]	,999	,998	4785174,833
17	,999[q]	,999	,998	4779516,498

a. Variables predictoras: (Constante), n conductores

b. Variables predictoras: (Constante), n_conductores, km autopistas peaje

Tabla 13: Modelos finales en la estimación de la prima total de autos
Fuente: Elaboración Propia

En la tabla 14 además de observar que todos los coeficientes son significativos, se comprueba que no existen problemas de colinealidad ya que el VIF es menor que 10. A partir de ahí se estudia si siguen existiendo puntos influyentes.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados			Estadísticos de colinealidad	
		B	Error típ.	Beta	t	Sig.	Tolerancia	FIV
1	(Constante)	1536709,000	4616287,186		,333	,741		
	n_conductores	394,843	9,388	,987	42,056	,000	1,000	1,000
13	(Constante)	-4758714,184	2089419,310		-2,278	,028		
	km_autopistas_peaje	132656,666	15961,213	,066	8,311	,000	,757	1,321
	Turismos	481,422	35,172	1,017	13,688	,000	,009	114,005
	Inter_autopistas_her_no_hosp	-147605,692	14626,415	-,106	-10,092	,000	,436	2,296
	Inter_autovías_fall	-1034186,046	345620,468	-,037	-2,992	,005	,321	3,115
	Inter_otro_her_no_hosp	23593,337	4062,433	,049	5,808	,000	,679	1,473
	km_carr_doble_calz	-261351,875	44996,802	-,073	-5,808	,000	,308	3,251
	Camiones_furgonetas	-254,507	46,496	-,111	-5,474	,000	,118	8,472
	total_robos_vehículos_2014	-13805,507	3666,224	-,072	-3,766	,001	,133	7,514
	km_carr_calz_menor5	5305,020	1560,829	,027	3,399	,002	,795	1,257
	n_habitantes	58,223	17,102	,263	3,404	,002	,008	123,594
	Urb_calle_her_hosp	-75854,567	25625,612	-,038	-2,960	,005	,291	3,434

Tabla 14: Coeficientes estimados para la prima total de autos
Fuente: Elaboración Propia

En la tabla 15 se observa que ahora los valores de la distancia de Cook son sólo ligeramente superiores a $\frac{4}{52} = 0{,}077$ ya que como se puede ver su media es de 0,081.

Al aplicar este modelo se obtiene la tabla 16 en la que se puede ver la diferencia entre los valores estimados y los observados.

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	12933022,00	558284800,00	159105702,50	116041683,488	50
Valor pronosticado tip.	-1,260	3,440	,000	1,000	50
Error típico de valor pronosticado	1127559,875	4501489,500	2333327,007	985529,437	50
Valor pronosticado corregido	12728175,00	558980800,00	159254644,68	116168147,157	50
Residual	-9727792,000	11020781,000	,000	4096728,427	50
Residuo típ.	-2,035	2,306	,000	,857	50
Residuo estud.	-2,281	2,672	-,003	1,047	50
Residuo eliminado	-23054114,000	17067340,000	-148942,176	7165583,636	50
Residuo eliminado estud.	-2,431	2,942	-,003	1,079	50
Dist. de Mahalanobis	1,747	42,485	12,740	11,260	50
Distancia de Cook	,000	1,474	,081	,240	50
Valor de influencia centrado	,036	,867	,260	,230	50

a. Variable dependiente: Primas_Autos_Total

Tabla 15: Estadísticos finales sobre los residuos para la prima total de autos
Fuente: Elaboración Propia

Provincia	Valores estimados	Valores observados
Álava	71077276	81540941
Albacete	99868844	87107249
Alicante	383149069	414688576
Almería	138780243	159085148
Asturias	247027285	258838977
Ávila	44538288	41474193
Badajoz	166258563	152179156
Barcelona	1194559658	1221756642
Burgos	92065954	97261110
Cáceres	102764624	92961614
Cádiz	283682154	242649006
Cantabria	142985717	148153303
Castellón	137432643	136693741
Ceuta	20600132	12545612
Ciudad Real	125698694	112536276
Córdoba	196309681	159978323
Cuenca	62719236	50918975
Gerona	167171045	178246054
Granada	231853462	198494442
Guadalajara	49295358	68619109
Guipúzcoa	175851277	171613470
Huelva	111768627	106154370
Huesca	57245525	57218259
Islas Baleares	232870183	254068193
Jaén	152244789	137964871
La Coruña	284261389	316574736
La Rioja	77909247	71928635
Las Palmas	218657804	218164152
León	123660119	128856176
Lérida	103492715	109229739
Lugo	87173057	106800111
Madrid	1281746950	2210788652
Málaga	331079522	361288606
Melilla	21651994	14250338
Murcia	336570211	332457485
Navarra	143047708	165045359
Orense	83272402	87594123
Palencia	48170055	41847455
Pontevedra	232600505	253672049
Salamanca	98313557	76093160
Segovia	43368368	44293901
Sevilla	431461288	420529299
Soria	23904574	24651498
Sta Cruz Tenerif	206912404	157768917
Tarragona	157716526	169888813
Teruel	37408210	35576944
Toledo	143727628	170113179
Valencia	584055139	558088896
Valladolid	129078157	128415240
Vizcaya	253081020	291321843
Zamora	53570720	45854268
Zaragoza	207882139	203989235

Tabla 16: Primas totales estimadas frente a observadas
Fuente: Elaboración Propia

Se tomará ahora la prima media de autos como variable dependiente. Al aplicar la regresión lineal múltiple para obtener la prima media de autos, SPSS proporciona la tabla 17 donde se observa un R^2 ajustado muy alto para el modelo 8, por lo que en principio el modelo es bueno. Ahora bien, hay que fijarse en la tabla 18 en el valor de la distancia de Mahalanobis y en si la distancia de Cook es mayor que $\frac{4}{n}$ donde n es el número de observaciones, para detectar si existen puntos influyentes.

Resumen del modelo[i]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,621[a]	,386	,373	36,901
2	,857[b]	,735	,724	24,481
3	,881[c]	,775	,761	22,798
4	,898[d]	,807	,790	21,360
5	,916[e]	,839	,820	19,754
6	,925[f]	,856	,835	18,901
7	,934[g]	,873	,852	17,953
8	,943[h]	,888	,867	17,021

a. Variables predictoras: (Constante), Z robos vehículos 2014

b. Variables predictoras: (Constante), Z_robos_vehículos_2014, Z_Camiones_furgonetas

c. Variables predictoras: (Constante), Z robos vehículos 2014, Z_Camiones_furgonetas, Z_Otros_vehículos

Tabla 17: Modelos en la estimación de la prima media de autos
Fuente: Elaboración Propia

Se puede ver que los valores de la distancia de Cook en su mayoría son inferiores a $\frac{4}{52} = 0{,}077$ ya que como se puede ver su media es de 0,075.

En el modelo 8 no existen problemas de colinealidad como puede verse en la tabla 19 ya que el VIF es menor que 10 y además se observa que todos los coeficientes son significativos. Está formado por las variables ratio de robos totales de vehículos en 2014, ratio del número de camiones y furgonetas, ratio del número de otros vehículos, ratio de los kilómetros de autopistas y autovías libres, ratio de heridos no hospitalizados en autopistas, ratio de heridos hospitalizados en carreteras convencionales, ratio de heridos no hospitalizados en calles y tasa de paro.

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	215,86	412,08	342,14	43,920	50
Valor pronosticado tip.	-2,875	1,592	,000	1,000	50
Error típico de valor pronosticado	3,908	14,771	6,798	2,462	50
Valor pronosticado corregido	182,93	406,85	342,82	43,674	50
Residual	-37,626	33,918	,000	15,570	50
Residuo típ.	-2,211	1,993	,000	,915	50
Residuo estud.	-2,593	2,141	-,013	1,043	50
Residuo eliminado	-88,805	52,069	-,681	22,710	50
Residuo eliminado estud.	-2,801	2,244	-,018	1,073	50
Dist. de Mahalanobis	1,603	35,919	7,840	7,622	50
Distancia de Cook	,000	2,278	,075	,336	50
Valor de influencia centrado	,033	,733	,160	,156	50

a. Variable dependiente: Prima_media

Tabla 18: Estadísticos sobre los residuos para la prima media de autos
Fuente: Elaboración Propia

Por tanto, se podrán estimar los valores para la prima media a partir de los coeficientes de la tabla 19.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	373,233	7,701		48,463	,000		
	Z_robos_vehículos_2014	-48047,604	8752,268	-,621	-5,490	,000	1,000	1,000
8	(Constante)	501,601	15,477		32,410	,000		
	Z_robos_vehículos_2014	-35506,062	6210,648	-,459	-5,717	,000	,423	2,367
	Z_Camiones_furgonetas	-1017,942	122,549	-,627	-8,306	,000	,477	2,095
	Z_Otros_vehículos	4245,433	1068,837	,556	3,972	,000	,139	7,191
	Z_km_autovías_autopistas_libres	-44272,814	10935,366	-,383	-4,049	,000	,305	3,280
	Z_Inter_autopistas_her_no_hosp	-111371,157	33748,550	-,204	-3,300	,002	,716	1,397
	Z_Inter_convencional_her_hosp	-147598,437	52179,991	-,217	-2,829	,007	,464	2,154
	Z_Urb_calle_her_no_hosp	-7120,585	2551,294	-,194	-2,791	,008	,566	1,766
	Tasa_paro	-1,125	,470	-,167	-2,393	,021	,561	1,783

a. Variable dependiente: Prima_media

Tabla 19: Coeficientes estimados para la prima media de autos
Fuente: Elaboración Propia

La diferencia entre los valores estimados y los valores observados puede verse en la tabla 20 y en el mapa 9. En la tabla 20 se puede ver que, en 29 provincias la prima estimada es menor que la prima observada mientras que en 23 ocurre lo contrario.

Resulta muy destacable el caso de Madrid, donde la prima estimada es 200 euros menor que la observada. Hay que tener en cuenta que los factores de riesgo existentes son muchos y en el caso de Madrid, estos factores se multiplican. Sin tener en cuenta el caso de Madrid, el rango de la diferencia de primas está entre -34 y 38 euros. Esto demuestra que el factor de riesgo provincia es realmente relevante en algunos casos. En los casos de Badajoz, Ceuta, Córdoba, Soria, Tarragona y Zaragoza la prima estimada es bastante mayor que la prima cobrada en el año 2014, mientras que en las provincias de Barcelona, Guadalajara, La Coruña, Lugo, Madrid, Toledo y Vizcaya la prima estimada es bastante más baja que la prima cobrada.

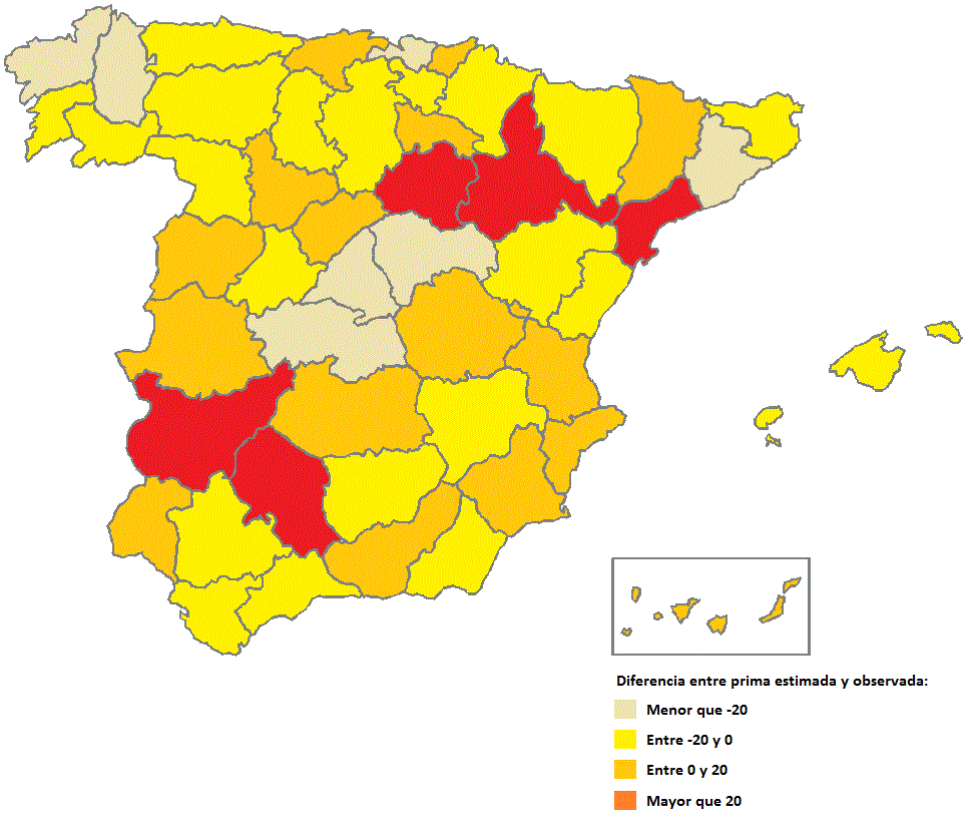

Figura 9: Mapa coloreado de la diferencia entre primas estimadas y observadas
Fuente: Elaboración Propia

Provincia	Valores estimados	Valores observados
Álava	399	403
Albacete	322	328
Alicante	337	335
Almería	313	332
Asturias	384	392
Ávila	330	335
Badajoz	364	326
Barcelona	324	349
Burgos	382	392
Cáceres	325	314
Cádiz	326	335
Cantabria	391	374
Castellón	328	335
Ceuta	232	210
Ciudad Real	338	333
Córdoba	344	316
Cuenca	311	310
Gerona	288	300
Granada	313	312
Guadalajara	367	389
Guipúzcoa	399	388
Huelva	345	337
Huesca	333	334
Islas Baleare	276	277
Jaén	305	322
La Coruña	398	427
La Rioja	373	364
Las Palmas	299	288
León	368	372
Lérida	357	337
Lugo	386	408
Madrid	323	523
Málaga	319	338
Melilla	221	235
Murcia	345	340
Navarra	378	382
Orense	354	362
Palencia	349	351
Pontevedra	372	379
Salamanca	352	340
Segovia	371	369
Sevilla	353	358
Soria	370	348
Sta Cruz Ten	216	212
Tarragona	329	303
Teruel	322	332
Toledo	324	347
Valencia	360	342
Valladolid	395	389
Vizcaya	412	446
Zamora	336	339
Zaragoza	396	370

Tabla 20: Primas medias estimadas frente a observadas
Fuente: Elaboración Propia

6. Conclusiones

La principal conclusión que se extrae de este trabajo es lo importante que es para las entidades aseguradoras elaborar una buena normativa de suscripción, ya que ello les permitirá ajustar mucho mejor el importe de la prima. En este trabajo ha quedado probado que el factor provincia es un factor de riesgo muy importante a la hora de determinar el precio de la prima, pero no sólo este factor es determinante, también lo son la edad y la experiencia del conductor, las características del vehículo, el historial siniestral,... Por este motivo, uno de los objetivos prioritarios de las entidades aseguradoras es ser capaces de seleccionar y valorar adecuadamente todos estos riesgos para ajustar el importe de la prima lo máximo posible.

Por otra parte, este trabajo sirve para demostrar lo difícil que resulta en la actualidad conseguir una buena base de datos debido a que apenas existen datos disponibles. A su vez, demuestra también que una vez elaborada y tratada la base de datos son muchas las aplicaciones que se le puede dar a los datos. En este caso, los datos se han utilizado para hacer una clasificación de las provincias y para llevar a cabo una estimación de la prima de autos, pero son muchas las aplicaciones que se les podrían haber dado. Algunas de ellas quedarán como futuras líneas de investigación.

A lo largo del trabajo se han utilizado diferentes técnicas estadísticas. Entre ellas, se han utilizado GLM; que son de vital importancia en el cálculo de la prima para las entidades aseguradoras, pero que en este caso se han utilizado para la estimación de los salarios medios anuales de las provincias vascas y Navarra, lo que sirve para demostrar que los GLM se pueden usar en otros campos más allá de la tarificación. En lugar de utilizar como herramienta SPSS, para trabajar con GLM se ha utilizado R, que al ser un software libre está empezando a ser cada vez más utilizado por las entidades aseguradoras para la tarificación de las primas. Haciendo una comparativa entre las dos herramientas, se podría decir que SPSS tiene una mejor salida de tablas y gráficos por lo que si no se requiere programar puede ser preferible a R. Ahora bien, SPSS tiene una sintaxis más complicada que la de R, por lo que cuando haya que programar siempre será preferible usar R. Además, R tiene multitud de paquetes que lo convierten en una herramienta muy potente.

Finalmente, analizando los resultados se puede ver como en provincias como Badajoz, Ceuta, Córdoba, Soria, Tarragona y Zaragoza la prima estimada es bastante superior a la prima cobrada, mientras que en Barcelona, Guadalajara, La Coruña, Lugo, Madrid, Toledo y Vizcaya la prima estimada es notablemente más baja que la prima cobrada.

7. Futuras líneas de investigación

El historial siniestral ha sido siempre uno de los factores más importantes a la hora de determinar el importe de la prima, pero no es menos cierto que existen otros factores como el probado en este trabajo, que también son muy importantes en el momento de tarificar.

En este estudio se ha visto la importancia del factor provincia en el análisis, selección y valoración del riesgo y en la posterior tarificación. Esto abre un camino a pensar en qué ocurriría si además del factor provincia se añadieran factores como la zona (urbana, periférica o rural). Para intentar facilitar la búsqueda de datos, se podría reducir a dos zonas: urbana y rural. El criterio podría ser buscar los datos de las capitales de cada provincia que serían los datos correspondientes a la zona urbana y la diferencia respecto al total de la provincia serían los datos correspondientes a la zona rural. Se trataría de crear una base de datos de 104 registros en la que cada provincia tuviese dos niveles: zona urbana y zona rural. El enfoque podría ser similar al de este trabajo y el objetivo sería ver cuanto puede influir el factor zona en el importe de la prima de autos.

En otra línea de trabajo se podría tratar de utilizar los datos sobre accidentes según la vía para, haciendo uso de técnicas estadísticas como podrían ser los GLM, tratar de ver en qué provincias afectan más las variables sobre fallecidos y ver si estas provincias se corresponden con las que tienen una prima media mayor, ya que los daños de responsabilidad civil son los que suponen un coste mayor para las entidades aseguradoras y por tanto es de esperar que éste sea uno de los motivos de aumento del importe de la prima.

Por último, otra línea de trabajo podría ser calcular la prima para cada una de las regiones que surgen de la aplicación del análisis de cluster. Para ello, simplemente habría que sumar los datos en valor en absoluto de las provincias pertenecientes a cada cluster, de forma que se tuviese una base de datos formada por 5 registros y a partir de ahí estimar la prima media. El objetivo sería ver si esta prima oscila mucho entre unas regiones y otras y valorar si las entidades aseguradoras se podrían plantear hacer grupos a la hora de tarificar en función de regiones similares a éstas.

8. Bibliografía

Referencias

[1] Belsley, D. A., Kuh, E., y Welsch, R. E. 1980. *Regression Diagnostics.* New York: John Wiley & Sons.

[2] Boj del Val, E., Claramunt Bielsa, Mª M., y Fortiana Gregori, J. 2006. *Una alternativa en la selección de los factores de riesgo a utilizar en el cálculo de primas.*

[3] Cañadas Reche, J.M. 2013. *Regresión logística. Tratamiento computacional con R.* Universidad de Granada.

[4] Cayuela L. *Modelos lineales generalizados (GLM).* Universidad de Granada.

[5] de Jong P. y Z. Heller G. Generalized Linear Models for Insurance Data. *International Series on Actuarial Science.*

[6] González Díez I. J. 2014. *Clasificación de clientes en seguros de automóviles.* Universidad Complutense de Madrid.

[7] Guardiola Lozano, A. 1990. *Manual de Introducción al Seguro.* Fundación Mapfre Estudios, Instituto Ciencias del Seguro, Ed. Mapfre, S.A. Madrid.

[8] López-González, E. y Ruiz-Soler, M. *Análisis de datos con el Modelo Lineal Generalizado. Una aplicación con R.* Universidad de Málaga.

[9] Mayer-Schönberger y V. Cukier, K. 2013. *Big data. La revolución de los datos masivos.* Turner.

[10] MacQueen J. B. 1967. *Some methods for classication and analysis of multivariate observations. Proceedings of the 5th Berkely Simposium in Mathematical Statistics and Probability, pp 281-297.* University of California Press.

[11] McCullagh, P. y Nelder, J. 1989. *Generalized Linear Models (2 ed.).* London: Chapman Hall.

[12] Nelder, J. y Wedderburn, R. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society (A), 135, pp. 370-384.* London: Chapman Hall.

[13] Pardo A. y Ruiz, M. A. 2005. *Análisis de datos con SPSS.* Mc Graw Hill.

[14] Peña, D. 2002. *Análisis de datos multivariantes.* Madrid: McGraw Hill Book co.

[15] Prieto Pérez, E. 1980. *La problemática de la estadística por ramos.* Madrid.

[16] Rawlings, J. O. 1988. *Applied Regression Analysis: A Research Tool*. Belmont, CA: Wadsworth.

[17] Rojo Abuín, J.M. 2007. *Regresión Lineal Múltiple*. Madrid.

[18] Tukey J. 1977. *E.D.A. exploratory data analysis.*

[19] *Ley 50/1980, de 8 de octubre, de Contrato de Seguro. Boletín Oficial del Estado, 17 de Octubre de 1980.*

[20] *Real Decreto Legislativo 8/2004, de 29 de octubre, por el que se aprueba el texto refundido de la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor (modificado por la Ley 21/2007, de 11 de julio) Artículo 2.*

[21] *Real Decreto 1507/2008, de 12 de septiembre, por el que aprueba el Reglamento del seguro obligatorio de responsabilidad civil en la circulación de vehículos a motor (modificado por la Ley 21/2007, de 11 de julio) Artículo 10.*