Trabajo Fin de Máster

Modelo de estimación de ingresos para clientes poco o nada vinculados con ABANCA

Francisco Manuel Fernández Agraso

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo de Fin de Máster

Título en galego: Modelo de estimación de ingresos para clientes pouco ou nada vinculados con ABANCA
Título en español: Modelo de estimación de ingresos para clientes poco o nada vinculados con ABANCA
English title: Income estimation model for little or no linked customers with ABANCA
Modalidad: Modalidad B
Autor: Francisco Manuel Fernández Agraso, Universidad de Santiago de Compostela
Director: Ricardo Cao Abad, Universidade da Coruña
Tutor: Daniel López Souto, ABANCA
Breve resumen del trabajo: Clasificación de clientes en clústeres y desarrollo de algoritmos de estimación de renta para cada uno de los clústeres identificados. La clasificación, así como el algoritmo de estimación de renta, se podrá realizar a partir de variables internas (información BBDD entidad) y/o externas
Recomendaciones: Preferiblemente desarrollos en R
Otras observaciones: La empresa desea participar en el proceso de selección

Don Ricardo Cao Abad, Catedrático de la Universidade da Coruña, y don Daniel López Souto, Técnico Especialista en el Departamento de Desarrollo de Modelos de Valoración de Riesgo de Crédito en ABANCA, informan que el Trabajo de Fin de Máster titulado

Modelo de estimación de ingresos para clientes poco o nada vinculados con ABANCA

fue realizado bajo su dirección por don Francisco Manuel Fernández Agraso para el Máster en Técnicas Estadísticas.
Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 18 de ENERO de 2018

El director: El tutor:

Ricardo Cao Abad Daniel López Souto

El autor:

Francisco Manuel Fernández Agraso

Agradecimientos

En primer lugar, me gustaría agradecer a Ricardo Cao Abad, director del presente trabajo, todos los consejos que me dio a lo largo de la realización del mismo ya que me han servido de gran ayuda tanto para el presente trabajo como para futuras investigaciones.

En segundo lugar, y aunque no menos importante, quiero agradecer a ABANCA que me dieran todas las facilidades habidas y por haber para poder trabajar en sus instalaciones siendo un componente más de su equipo de trabajo. En especial, quiero acordarme de aquellos que desde el primer día me acogieron como un compañero más, como son Daniel, Cristina, Lino, Juan, Eva y Víctor, así como al resto de compañeros de la 5ª planta de la Oficina Principal de ABANCA en Rúa Nova de A Coruña.

Índice general

x

Resumen

Resumen en español

Las entidades financieras están evolucionando de su tradicional comportamiento reactivo ante sus posibles clientes, a una conducta mucho más proactiva que consiga adaptar su cartera de productos y servicios a las necesidades particulares de cada persona.

En este trabajo se pretende estimar el ingreso mensual que pueda tener un determinado individuo, en función de la información que se tenga sobre él, con el objetivo de ofrecer el producto o servicio que más se adapte a lo que podría necesitar dicha persona. Para la estimación del ingreso mensual se desarrolla a lo largo de la memoria una serie de modelos de regresión múltiple que combinan ajustes lineales paramétricos con otros de tipo no paramétrico, así como técnicas de clusterización y segmentación de clientes.

English abstract

Financial entities are evolving from their traditional reactive behavior to their potential clients, to a much more proactive behavior that manages to adapt their portfolio of products and services to the particular needs of each client.

This paper intends to estimate the monthly income that a certain individual may have, based on the information that is held about him, in order to offer the product or service that best suits what that person might need. For the estimation of monthly income, a series of multiple regression models that combine parametric linear adjustments with others of a non-parametric type, as well as clustering and client segmentation techniques, are developed throughout the report.

Capítulo 1

Introducción

Para los bancos, y en general, para todas las entidades financieras, uno de los objetivos principales, una vez que parece ya superada la crisis económica que asoló a todo el sistema financiero en la última década, es el de captar y vincular con la entidad a la mayor cantidad de clientes posible. Una de las principales actividades bancarias es la financiación a clientes, y uno de los elementos claves en el proceso de admisión de crédito es la determinación de la capacidad de pago de los clientes. La determinación de capacidad de pago tradicionalmente es un proceso de validación soportado en la documentación que pueda aportar el solicitante de crédito, no obstante las nuevas tendencias obliga a las Entidades a ser capaces de agilizar los procesos de concesión e incluso ser capaces de dar un valor añadido al cliente el no tener que aportar tanta documentación. Para ello se hace imprescindible adoptar nuevas técnicas que permitan inferir la capacidad de pago de los clientes en base a sistemas de cálculo de ingresos estimados en función del perfil de los clientes.

El principal objetivo de este trabajo es conseguir algoritmos que permitan a una entidad financiera estimar los ingresos de un cliente, siendo las utilidades:

- Agilizar los procesos de concesión de crédito.
- Contraste de la declaración de ingresos de los clientes para detección de potencial fraude en la documentación.

La constante evolución del Big Data, principalmente en los últimos tiempos, está provocando que las entidades bancarias se encuentren ante una ingente cantidad de información acerca de sus clientes como nunca antes había podido almacenar. Dicha información, analizada profundamente y de forma correcta, puede ayudar a las entidades a conseguir una mejor segmentación de los clientes, así como a obtener modelos con una mayor capacidad de predicción que contribuyan a una mejor estimación tanto de ingresos como de riesgos de crédito.

La estimación de ingresos mensuales para clientes con nómina o pensión domiciliada se realiza en ABANCA directamente a partir de los movimientos de ingresos por estos conceptos en las cuentas corrientes de los clientes. Esta información es tratada por parte de la entidad para calcular límites de riesgo personalizados para cada cliente que luego utiliza para ofrecer de manera proactiva a través de préstamos o tarjetas preconcedidas (campañas con envío carta o SMS, ofertas en cajeros o banca electrónica, etc.). Sin embargo, para clientes que no tienen domiciliada en ABANCA su nómina o pensión, la estimación de sus ingresos resulta más compleja. El principal objetivo de este trabajo es obtener algún modelo de regresión para intentar predecir esta variable respuesta “ingresos netos mensuales por nómina/pensión” a partir del resto de información disponible en la entidad para estos clientes. La muestra proporcionada para el trabajo contiene únicamente clientes con nómina/pensión domiciliada ya que de éstos disponemos de la variable a modelizar. Se trata de clientes que en su pasado reciente han registrado movimientos en sus cuentas corrientes en concepto de pago de nómina o pensión. Los importes mensualizados de estos movimientos registrados durante algo más de un año serán los que se utilicen como variable respuesta. Puesto que no se

dispone de la misma información para todos los clientes ha sido necesario clasificar a los clientes en diferentes grupos o clústeres y obtener un modelo para cada uno de ellos.

A partir de aquí, la memoria se organiza de la siguiente forma:

- En el segundo capítulo se presentan una serie de preliminares: una sección dedicada a las modificaciones realizadas en las bases de datos utilizadas, una sección dedicada a la segmentación de los clientes en función del distinto tipo de información que se tenga sobre ellos, y una tercera sección que incluye un análisis de los ingresos mensuales, que será la variable respuesta que utilizaremos a lo largo del trabajo en los distintos ajustes, así como el proceso llevado a cabo para la obtención de dicha variable.
- En el tercer capítulo se aborda la cuestión principal, que consiste en el ajuste de los distintos modelos en aquellos segmentos en los que obtenemos resultados más o menos satisfactorios, así como un análisis exploratorio que sirva como soporte para futuras investigaciones en aquellos segmentos en los que los resultados no son los esperados inicialmente.
- Finalmente, en el capítulo 5, se presentan una serie de conclusiones finales acerca de los resultados obtenidos a lo largo del trabajo.

Capítulo 2

Preliminares

2.1. Bases de datos

La información de la que dispone ABANCA acerca de sus clientes es muy diversa y con distintos grados de calidad. Entre las bases de datos que se han utilizado para la realización del trabajo podemos encontrar clientes de los que tenemos información detallada y de alto grado de fiabilidad (Captura de Bienes y KYC), así como otros de los que apenas tenemos información o esta está almacenada de una forma que complica mucho el trabajo a la hora de intentar analizarla mediante el software estadístico R, que es el que se ha utilizado para la realización de este proyecto.

Toda la información utilizada en este trabajo ha sido proporcionada por parte de la entidad previo tratamiento de la misma para garantizar que no contuviese ningún dato que pudiese servir para identificar a los clientes. Se trata de bases de datos con identificadores de registros que permiten su uso a efectos de tratamiento estadístico pero en ningún caso es posible identificar a qué cliente corresponden los datos de informados en las distintas variables.

2.2. Segmentación de clientes

Debido a la gran heterogeneidad de clientes en cuanto a la información disponible acerca de cada uno de ellos, se plantean varios grupos, para los que será necesario desarrollar un modelo para cada uno de ellos.

Los diferentes grupos que vamos a tratar son los siguientes:

- En primer lugar, tendremos un grupo de clientes de los que tendremos información de la Captura de Bienes (CB) más reciente estableciendo como profundidad histórica máxima enero de 2014. Esta información es la más rica en cuanto a cantidad y fiabilidad de la misma ya que proviene del formulario que deben cumplimentar todos aquellos clientes (con sus correspondientes justificantes) que acudan a la entidad en busca de cualquier tipo de préstamo o crédito. Este grupo de clientes se dividirá en dos segmentos distintos:
    - El Segmento 1 será aquel en el que tenemos información tanto de la variable de ingresos fijos como la de ingresos de la unidad familiar, así como del resto de variables sociodemográficas y laborales presentes en la totalidad de los clientes.
    - El segmento 2 será aquel en el que tan solo tendremos información de los ingresos fijos, además de las mismas variables básicas del segmento anterior.

- El Segmento 3 estará formado por todos aquellos clientes de los que tenemos información de Know Your Costumer (KYC) estableciendo como profundidad histórica máxima enero de 2014 y que no están presentes en ninguno de los segmentos anteriores. El KYC es un formulario que las entidades financieras se han visto obligadas por ley a cumplimentar acerca de algunos de sus clientes con el objetivo de identificarlos y evitar mantener relaciones comerciales con personas involucradas en delitos de blanqueo de capitales, terrorismo, corrupción gubernamental o delitos relacionados con drogas, entre otros.
- El Segmento 4 lo forman aquellos clientes que no se encuentran en ninguno de los casos anteriores y de los que tan solo tendremos información sociodemográfica y laboral muy básica (sexo, edad o comunidad autónoma).
- En el Segmento 5 analizaremos a aquellos clientes que tienen contratado alguno de los productos de pasivo que ofrece ABANCA, como pueden ser depósitos (tanto a la vista como no a la vista), valores, fondos de inversiones o fondos de pensiones.

Mientras que los cuatro primeros segmentos identificados son mutuamente excluyentes entre sí, haciendo que un cliente tan solo podrá pertenecer a uno de ellos, en los dos últimos se incluyen todos aquellos clientes de los que tenemos información correspondiente de cada uno de estos segmentos. El motivo es que los resultados de los modelos que se han intentado desarrollar para ellos no son eficientes y se ha optado por llevar a cabo un análisis exploratorio de dichos segmentos, y que puedan ser de interés para futuras investigaciones acerca de los mismos.

2.3. Variable respuesta y depuración de datos

Como variable respuesta a explicar en los modelos ajustados en este trabajo se utiliza el INGRESO MEDIO. Dicha variable se obtiene como una media mensual de los ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017 por concepto de nómina o pensión.. Para que los cálculos fuesen lo más cercanos posible a la realidad, lo que se hizo fue establecer unos intervalos distintos a los de los meses naturales. De esta forma, los ingresos de un mes estarán comprendidos entre el día 16 del propio mes y el día 15 del mes siguiente para así intentar corregir aquellos casos en los que algunos de los ingresos, como pueden ser nóminas o pensiones, que se producen a principios de mes en realidad corresponden al mes anterior. Una vez calculada dicha variable obtenemos un total de 814.964 observaciones cuyo comportamiento intentaremos estudiar a lo largo del presente trabajo.

En primer lugar, nos centraremos en el análisis de la variable obtenida y que será una de las variables más importante en el ajuste de los modelos que vamos a crear. Si realizamos un primer análisis exploratorio de la variable INGRESO MEDIO a través de un resumen de los estadísticos descriptivos básicos obtenemos:

```
> summary(ingresos$INGRESO_MEDIO)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   -463.7     696.5     981.5    1180.7    1477.8   1127452.0
```

Podemos ver que: la media está en torno a los 1.180 €; el valor mínimo de los ingresos es negativo (si nos fijamos en los datos vemos que hay dos clientes con ingreso medio negativo por lo que procedemos a eliminarlos); y un 50% de los clientes tienen unos ingresos de entre 700 € y 1.500 €. También vemos que hay, al menos, un cliente con un nivel de ingresos muy alto que podría estar alterando la media significativamente.

En la figura 2.3.1 vemos un análisis gráfico que nos permita apreciar con más facilidad cómo se distribuye la variable.

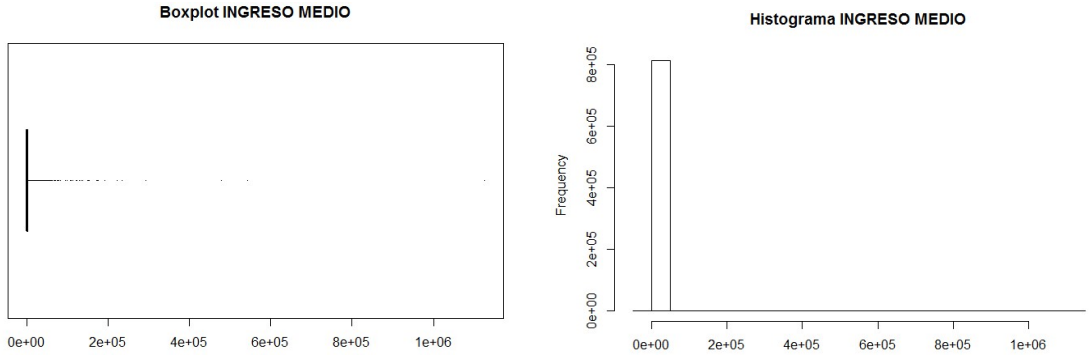


Figura 2.3.1: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma.

Observamos que hay varios clientes con unos ingresos muy altos que hacen que el histograma nos ofrezca muy poca información. Por este motivo, será necesario prescindir de estas observaciones que podemos considerar como datos atípicos.

Para elegir cuál será el valor máximo de INGRESO MEDIO a partir del cual prescindiremos de las observaciones que lo sobrepasen probamos con distintos niveles:

- Hay 13 clientes con un ingreso medio al mes superior a los 150.000 € y la distribución de los mismos se puede apreciar en la figura 2.3.2.

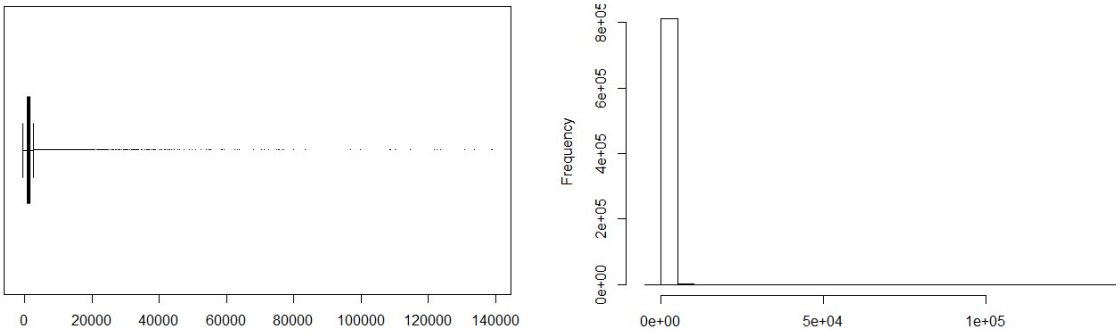


Figura 2.3.2: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma para valores menores o igual a 150.000 euros.

- Hay 24 clientes con un ingreso medio al mes superior a los 100.000 € y la distribución de los mismos se puede apreciar en la figura 2.3.3.

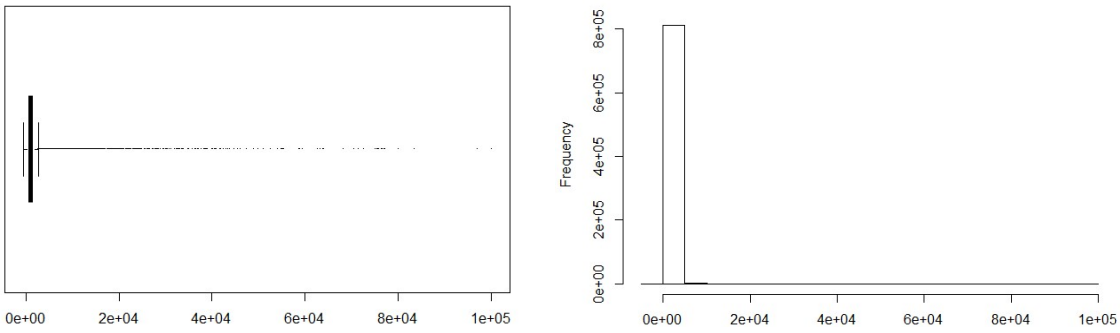


Figura 2.3.3: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma para valores menores o igual a 100.000 euros.

- Hay 56 clientes con un ingreso medio al mes superior a los 50.000 € y la distribución de los mismos se puede apreciar en la figura 2.3.4.

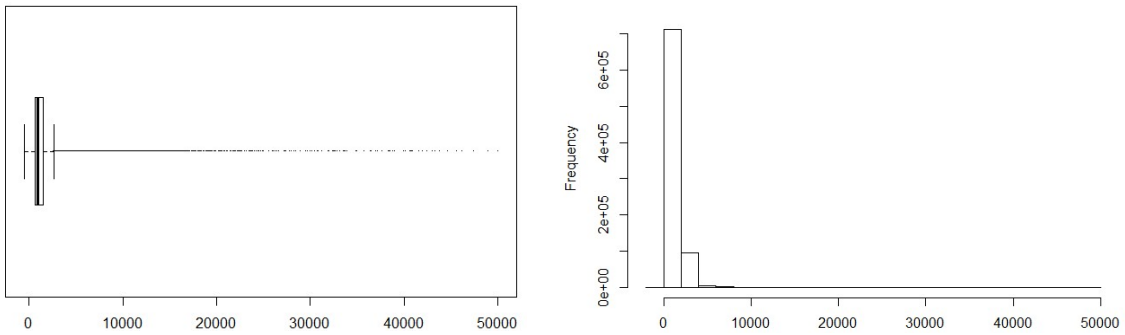


Figura 2.3.4: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma para valores menores o igual a 50.000 euros.

- Hay 630 clientes con un ingreso medio al mes superior a los 10.000 € y la distribución de los mismos se puede apreciar en la figura 2.3.5.

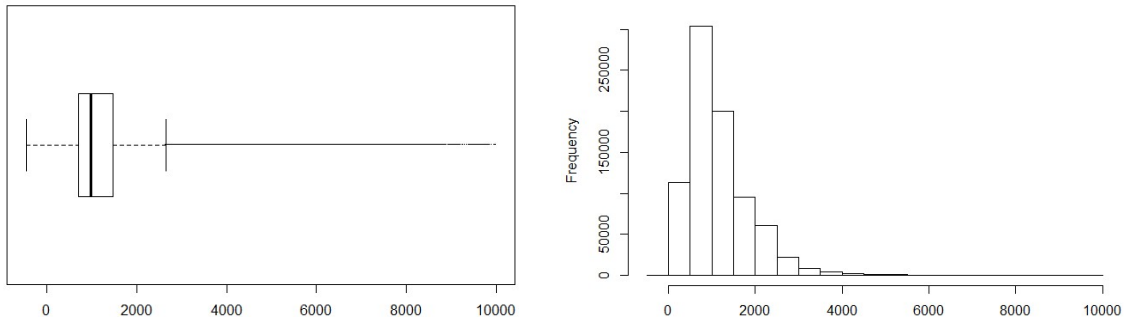


Figura 2.3.5: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma para valores menores o igual a 10.000 euros.

- Hay 3.103 clientes (tan solo un 0,38 % del total) con un ingreso medio al mes superior a los 5.000 € y la distribución de los mismos se puede apreciar en la figura 2.3.6.

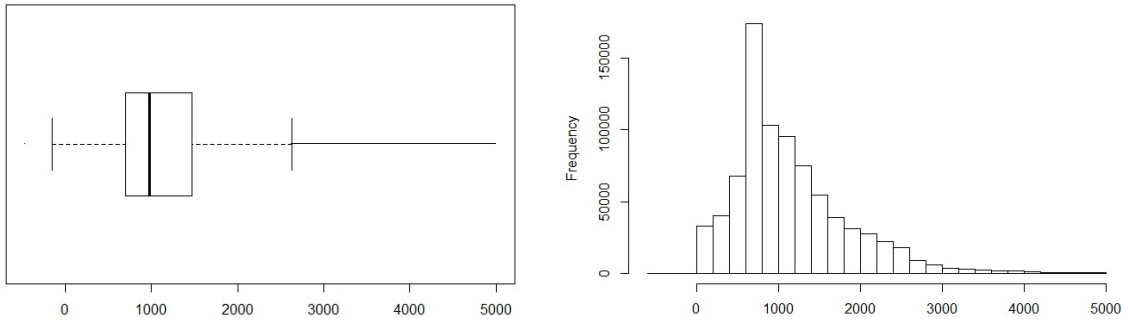


Figura 2.3.6: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO y a la derecha su correspondiente histograma para valores menores o igual a 5.000 euros.

Al analizar la variable INGRESO Medio prescindiendo de los valores más altos podemos observar que, a pesar de que cada vez es más simétrica, no sigue una distribución normal, por lo que será necesario llevar a cabo una transformación Box-Cox.

Utilizando la función powerTransform de R podemos ver que el parámetro λ más adecuado para dicha transformación es igual a 0.29 por lo que aplicaremos la fórmula de la transformación Box-Cox (3.1) con ese parámetro a la variable para que expanda los valores más bajos y contraiga los más altos, acercándose a una distribución normal.

Dicha fórmula es la siguiente:

$$\boldsymbol{Variable\ trasformada} = \frac{\boldsymbol{Variable^{\lambda} - 1}}{\boldsymbol{\lambda}} \qquad (3.1)$$

Tras llevar a cabo la transformación, si hacemos un resumen de los estadísticos descriptivos principales se obtiene:

```
> summary(ingresos$INGRESO_MEDIO_BC)
   Min.    1st Qu.     Median      Mean     3rd Qu.      Max.
 -2.541    19.573     21.982    22.148    25.187     192.878
```

Podemos ver que el 50% de los valores se encuentran entre 19.573 y 25.187, aunque se siguen apreciando valores muy altos y muy bajos (negativos) que se alejan mucho del promedio y que consideraremos como valores atípicos.

Por este motivo, una vez realizada la transformación Box-Cox nos quedaremos con aquellos clientes que no sobrepasen un nivel determinado de ingreso medio, que estableceremos en 50.000 € al mes, y que tengan un nivel de ingresos mensual de al menos 10 € (perderemos unas 684 observaciones, es decir, un 0,08 % del total).

```
> summary(ingresos1$INGRESO_MEDIO_BC)
   Min.    1st Qu.     Median      Mean     3rd Qu.      Max.
  3.279    19.573     21.986    22.159    25.189     76.075
```

De esta forma, la variable tendrá una distribución más simétrica.

Para ver la distribución definitiva de la variable de ingresos transformada representamos en la figura 2.3.7 tanto su diagrama de cajas como su histograma, donde apreciamos como se consigue disminuir la fuerte asimetría que tenían los ingresos sin transformar.

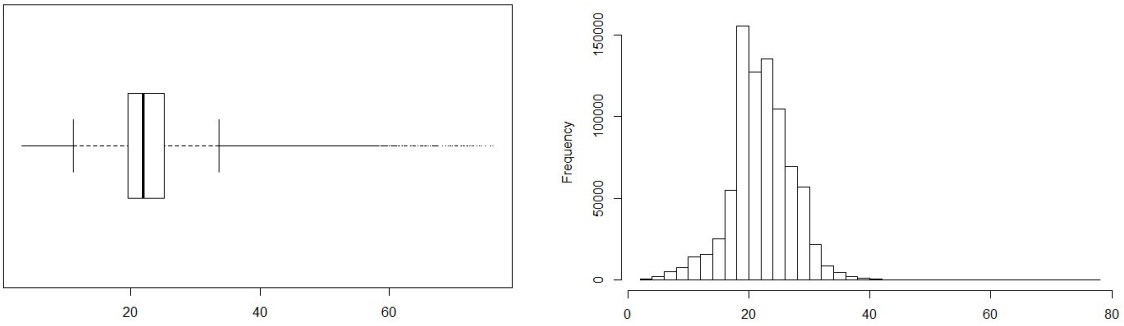


Figura 2.3.7: A la izquierda se muestra el diagrama de cajas del INGRESO MEDIO transformado, y a la derecha su correspondiente histograma, para aquellos clientes con unos ingresos mensuales de al menos 10 € y que no sobrepasen los 50.000 €.

A pesar de esta transformación, no podremos considerar que esta variable siga una distribución normal, como podemos deducir de su función de densidad estimada, que puede verse en la figura 2.3.8.

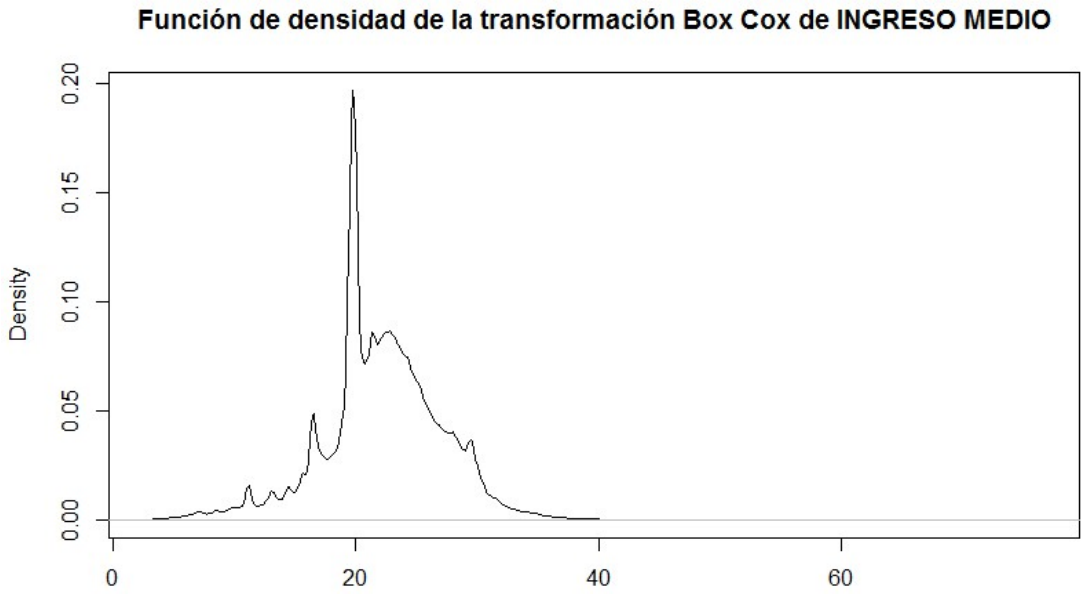


Figura 2.3.8: Función de densidad del INGRESO MEDIO transformado.

Los distintos picos en el gráfico de la función se corresponden principalmente con valores concretos de ingresos que provienen de pensiones y que se repiten en multitud de clientes.

Capítulo 3

Modelos de estimación de ingresos

En este capítulo se presentan los análisis realizados para cada uno de los segmentos que hemos establecido anteriormente. Mientras que para los tres primeros segmentos se ha logrado desarrollar un modelo que permita estimar el nivel de ingresos de forma más o menos satisfactoria, en el caso de los segmentos 5 y 6 se ha optado por realizar, simplemente, un análisis exploratorio de los datos ante la imposibilidad de obtener modelos con un grado de fiabilidad suficiente.

3.1. Segmento 1: Captura de bienes (Ingresos fijos y unidad familiar)

En un principio, para el ajuste de este modelo utilizaremos la información de aquellos clientes de los que tenemos tanto los datos de ingresos como los datos del formulario de captura de bienes. Como veremos más adelante, al introducir las variables explicativas propias de ingresos que tenemos de la captura de bienes, nos encontraremos con el hecho de que hay casos en los que estos datos de ingresos presentan valores nulos que empeoran drásticamente el ajuste (conviene recordar que trabajaremos siempre con la transformación Box-Cox de las variables de ingresos para un mejor ajuste). Por este motivo será recomendable ajustar el modelo con aquellas observaciones en las que los datos de ingresos propios de la BD de captura de bienes (tanto fijos como de la unidad familiar) sean no nulos. De esta forma pasamos de tener 141.146 observaciones a tener 132.155 observaciones, una vez que prescindimos de los autónomos, y que se quedarían en 59.091 al eliminar los clientes con valores nulos en ambas variables de las que hablamos anteriormente.

Las variables que van a formar parte del modelo son las siguientes:

- **INGRESO MEDIO:** Variable respuesta que nos dice el ingreso medio mensual de cada cliente y cuyo comportamiento intentaremos explicar a través del resto de variables (explicativas).
- **SEXO:** Variable cualitativa con dos niveles (hombre y mujer).
- **EDAD:** Variable cuantitativa y continua que nos indica la edad de cada cliente (en años).
- **AUTONOMÍA:** Variable cualitativa que nos muestra la comunidad autónoma a la que pertenece cada cliente.
- **SITUACIÓN LABORAL:** Variable cualitativa que nos indica la situación en la que se encuentra cada cliente dentro del mercado laboral. No tendremos en cuenta a los autónomos ya que sus ingresos son más difíciles de estimar ya que no se registran con concepto de nómina o pensión por lo que para la mayoría de ellos no dispondremos de variable respuesta con información fiable. Además los datos de la captura de bienes no se registran como ingresos fijos sino cómo ingresos variables.

- **SECTOR ACTIVIDAD:** Variable cualitativa que nos ofrece información acerca del sector al que pertenece el trabajo que desempeña cada cliente.
- **INGRESOS FIJOS NETOS:** Variable cuantitativa que nos indica la cantidad de ingresos fijos que percibe cada cliente al mes.
- **INGRESOS UNIDAD FAMILIAR:** Variable cuantitativa que nos indica el total de ingresos mensuales que perciben los miembros que forman parte del mismo grupo familiar. Se consideran miembros de un grupo familiar aquellos que comparten ingresos y gastos.

Prescindiremos del resto de variables presentes en la BD de Captura de Bienes ya sea porque aportan información muy similar a la que aportan algunas de las variables presentes en el modelo (por ejemplo, la variable PROFESIÓN aportaría información muy similar a la de SECTOR ACTIVIDAD) o porque la información que aportan es muy poco significativa (como es el caso de las variables UNIDAD FAMILIAR, ESTADO CIVIL o RÉGIMEN MATRIMONIAL).

Para hacernos una idea de la importancia que puede tener cada una de las variables en el ajuste del modelo será conveniente hacer una pequeña introducción individual de cada una de ellas en la que haremos un resumen de la información que nos puede aportar en relación al nivel de ingresos de cada cliente.

• INGRESO MEDIO (Variable Respuesta)

Como ya hemos visto, esta variable la obtenemos como una media mensual de los ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017.

Si hacemos un resumen de dicha variable para los clientes que pertenecen al segmento de captura de bienes:

```
> summary(z3.2$INGRESO_MEDIO2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
  3.183     19.385     21.819     21.886     24.521     59.197
  (10€)     (909€)   (1.321€)   (1.334€)   (1.924€)   (38.384€)
```

Podemos ver que el rango total de la variable es muy grande (va desde los 10 euros hasta los más de 38.000 euros), aunque gran parte de ellos se centran en torno a la media (que es de 1.334 euros). En la figura 3.1.1 podemos ver la función de densidad estimada de la transformación Box-Cox de la variable donde apreciamos mejor cómo es su distribución.

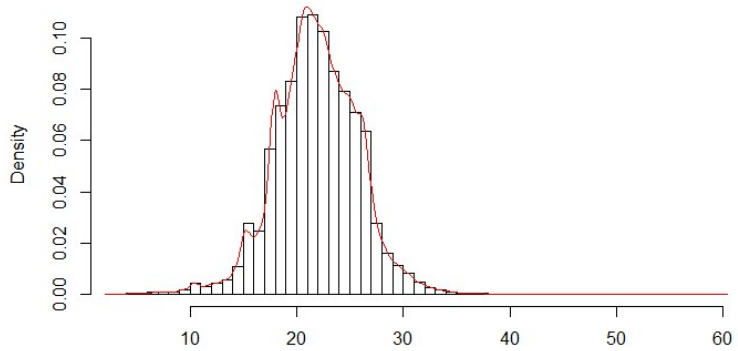


Figura 3.1.1: Histograma y estimación tipo núcleo de la función de densidad del Ingreso Medio transformado.

• SEXO

Si hacemos un resumen de la variable podemos ver que:

```
> summary(z3.2$SEXO)
Hombre  Mujer   NA's
 31966  27124      1
```

El 54,10 % de los clientes de la BD son hombres, mientras que el 45,90 % restante son mujeres (cabe destacar que hay un caso en el que no tenemos información del sexo del cliente).

Si relacionamos la variable SEXO con los ingresos podemos representar los resultados en el siguiente diagrama de cajas:

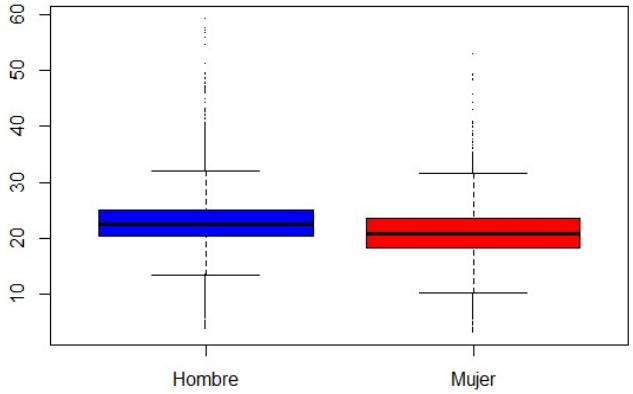


Figura 3.1.2: Diagrama de cajas del ingreso medio transformado diferenciando por sexo.

Si nos fijamos en la figura 3.1.2 lo primero que debemos destacar es la gran variabilidad que tienen los ingresos tanto en el caso de los hombres como en el de las mujeres. Lo que también podemos observar es que el nivel medio de ingresos es ligeramente superior en el caso de los hombres (con un ingreso medio de 1497,59 €) respecto de las mujeres (con una media de 1159,20 €).

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SEXO respecto del INGRESO MEDIO es de 0,04624.

- **EDAD**

Si hacemos un resumen de la variable:

```
> summary(z3.2$EDAD2)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  18.81     40.06     52.00     52.59     65.18    105.56
```

vemos que la edad de los clientes va desde los 18 años hasta los 105, siendo la media de edad de unos 52 años.

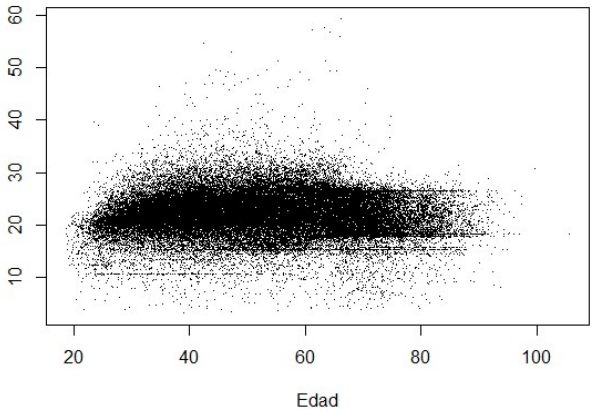


Figura 3.1.3: Diagrama de dispersión del Ingreso Medio transformado en función de la Edad.

A simple vista, en la figura 3.1.3, no se aprecia una relación lineal entre la edad y los ingresos, por este motivo es recomendable llevar a cabo una estimación polinómica o hasta una estimación no paramétrica de la relación entre ambas variables para ver de qué forma obtenemos un mejor ajuste.

En primer lugar llevaremos a cabo una estimación polinómica de dicha relación. Para ello seleccionaremos una muestra aleatoria de tamaño 10.000 de la variable edad para así poder visualizar mejor en un gráfico la estimación polinómica de grado 2 y 3 de la regresión del ingreso medio transformado respecto de la edad, junto con sus respectivos coeficientes de determinación ajustados.

Los gráficos que obtenemos para distintas muestras son los contenidos en la figura 3.1.4.

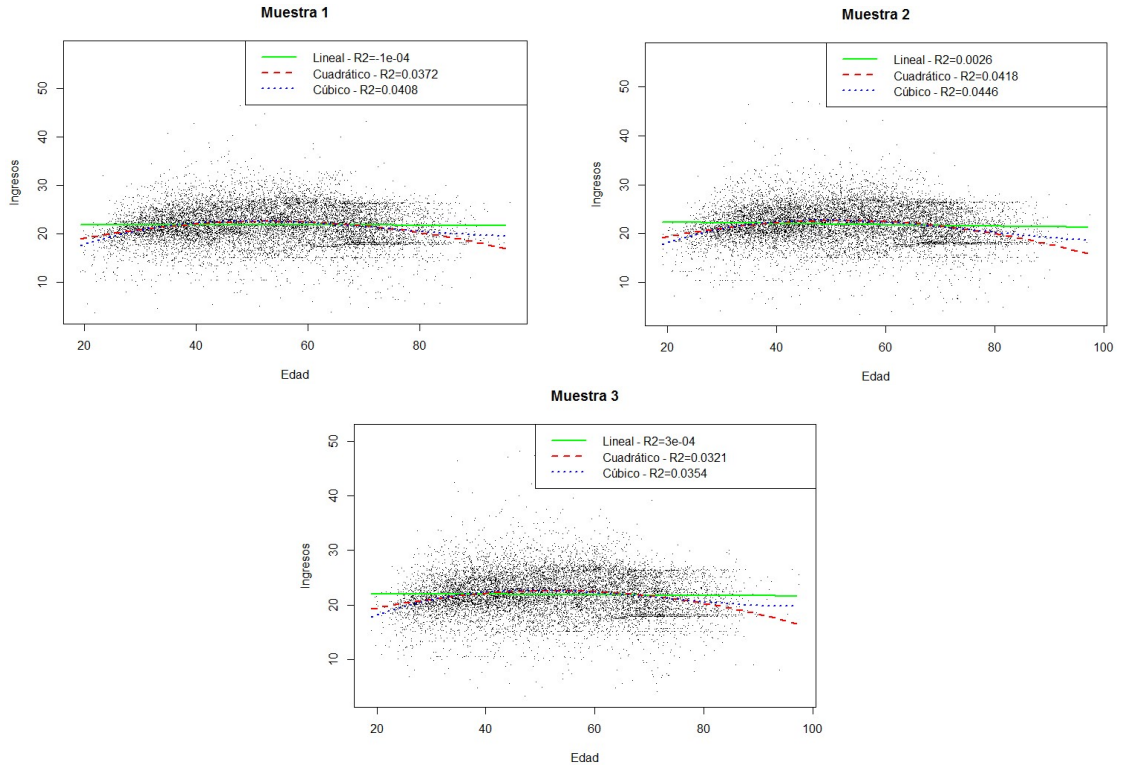
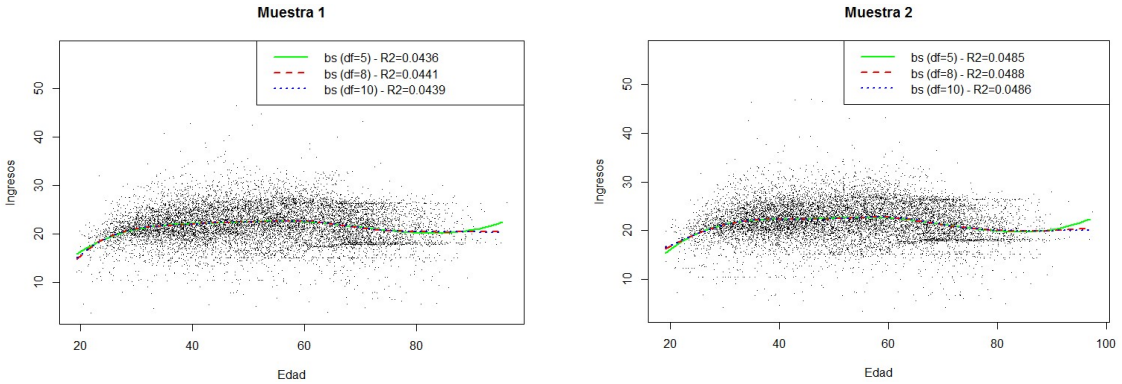


Figura 3.1.4: Representación de los ajustes polinómicos del ingreso medio transformado en función de la edad para tres muestras distintas.

Podemos ver que el coeficiente de determinación (R^2), a pesar de ser todavía muy bajo, crece significativamente cuando introducimos la estimación polinómica (tanto cuadrática como cúbica).

Para dotar de una mayor flexibilidad a la regresión podemos llevar a cabo una estimación polinómica local de la regresión mediante bases B-Splines.

Los gráficos que obtenemos para distintas muestras de esta forma son los contenidos en la figura 3.1.5.



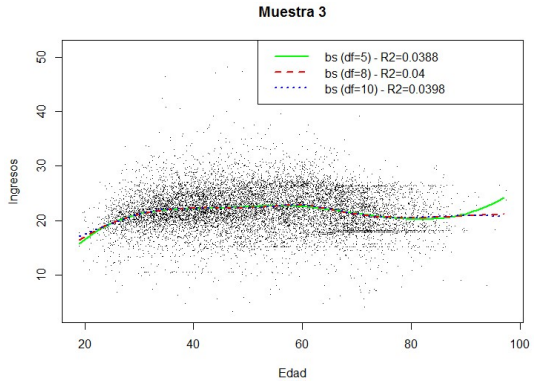


Figura 3.1.5: Representación del ajuste polinómico local del ingreso medio transformado en función de la edad mediante B-Splines con diferentes grados de libertad para tres muestras distintas.

Podemos ver que el ajuste polinómico local se parece bastante al ajuste cúbico que vimos anteriormente, y que a partir de los 8 grados de libertad el aumento del R^2 ajustado es muy poco significativo, y en algún caso puede llegar a disminuir.

Tras este análisis llegamos a la conclusión de que el mejor ajuste para la relación entre la variable EDAD y los ingresos es el que obtenemos a partir de la estimación no paramétrica mediante B-Splines con 8 grados de libertad, por lo que será la que utilizaremos en el modelo. En la figura 3.1.6 podemos observar la representación gráfica de dicho ajuste.

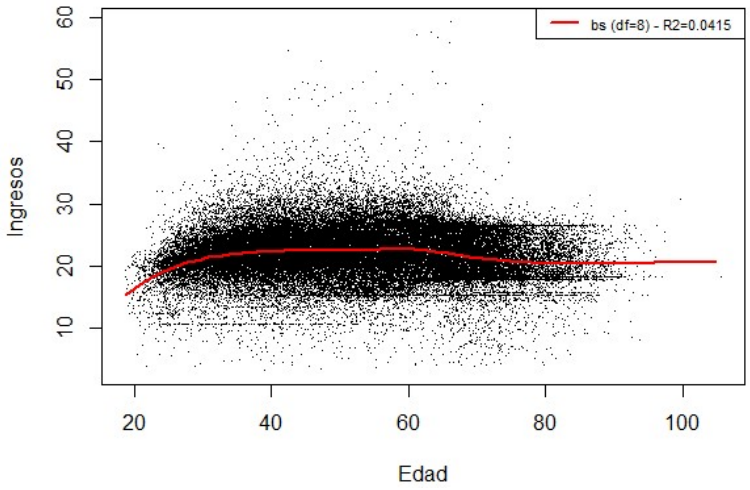


Figura 3.1.6: Diagrama de dispersión del ingreso medio transformado en función de la edad con su correspondiente ajuste polinómico local.

- **AUTONOMÍA**

Como la variable AUTONOMÍA es cualitativa con un número demasiado alto de niveles como para tratarla directamente así, es necesario realizar un análisis clúster que nos permita agruparlos en una cantidad más manejable de grupos. Para ello nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada una de las comunidades autónomas. En la figura 3.1.7 podemos observar el dendograma que nos proporciona los distintos conjuntos.

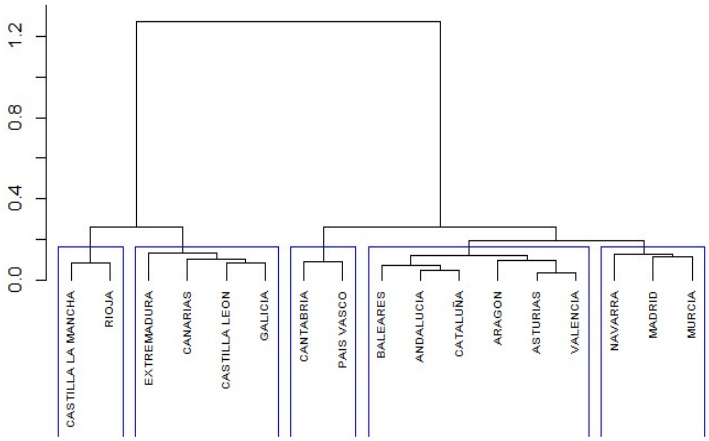


Figura 3.1.7: Dendograma que divide la variable AUTONOMÍA en 5 clústeres.

De esta forma obtenemos los siguientes conjuntos:

- **Clúster 1:** Castilla La Mancha y La Rioja (formado por 104 obs).
- **Clúster 2:** Canarias, Castilla León, Extremadura y Galicia (formado 55.427 obs).
- **Clúster 3:** Cantabria y País Vasco (formado por 220 obs).
- **Clúster 4:** Andalucía, Aragón, Asturias, Baleares, Cataluña y Valencia (formado por 2.102 obs).
- **Clúster 5:** Madrid, Murcia y Navarra (formado por 1.167 obs).
- **Clúster 6:** En este conjunto incluimos a Ceuta y Melilla (por tener datos insuficientes para poder analizarlas convenientemente) y a Otros (clientes extranjeros) aunque, en un principio, no nos centraremos en su análisis (formado por 71 obs).

En la figura 3.1.8 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable AUTONOMÍA a través de su correspondiente diagrama de cajas.

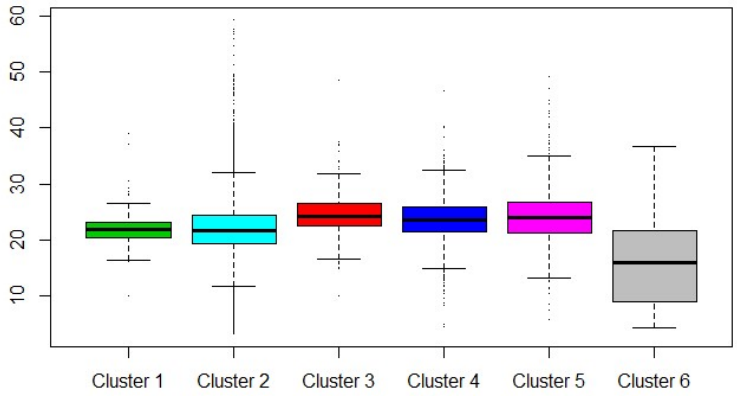


Figura 3.1.8: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

El nivel medio de ingresos en cada clúster es el siguiente:

- Para el **clúster 1** es de 1.382,36 €.
- Para el **clúster 2** es de 1.311,80 €.
- Para el **clúster 3** es de 1.910,82 €.
- Para el **clúster 4** es de 1.690,36 €.
- Para el **clúster 5** es de 1.870,62 €.
- Para el **clúster 6** es de 534,09 €.

Este último dato no es muy preciso ya que en este grupo se engloban clientes con características muy diferentes.

La distribución de los ingresos en cada uno de los distintos clústeres se puede apreciar mejor en la figura 3.1.9.

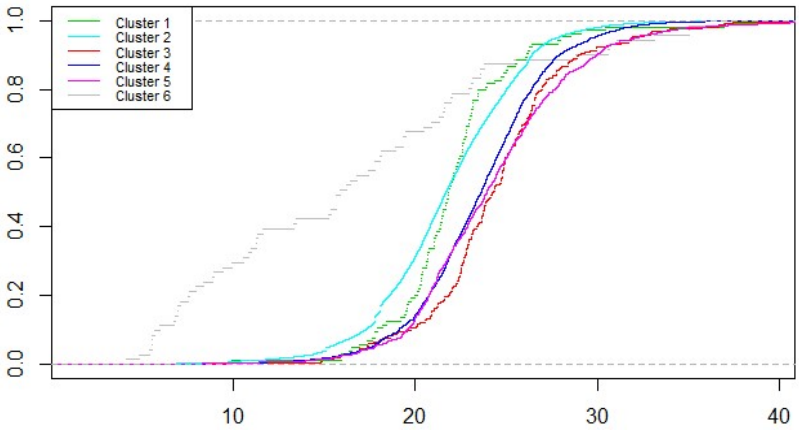


Figura 3.1.9: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable AUTONOMÍA respecto de los ingresos es de 0,01755.

- **SITUACIÓN LABORAL**

La variable SITUACION LABORAL es cualitativa y consta de 4 niveles diferentes. Si hacemos un pequeño resumen de la misma obtenemos:

```
> summary(z3.2$SITUACION_LABORAL_ID)
     Fijo     Otros   Temporal   Temporero
    30071     21542       7358        120
```

El grupo más numeroso es el de aquellos clientes con contrato fijo (un 50,89 % del total), seguido del grupo "Otros" formado por parados y pensionistas (un 36,46 % del total). En un término medio se encuentra el grupo de clientes con contrato temporal (un 12,45 % del total). Finalmente, el grupo con un número de clientes más bajo es el de los temporeros (0,20 % del total).

En la figura 3.1.10 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SITUACIÓN LABORAL a través de su correspondiente diagrama de cajas.

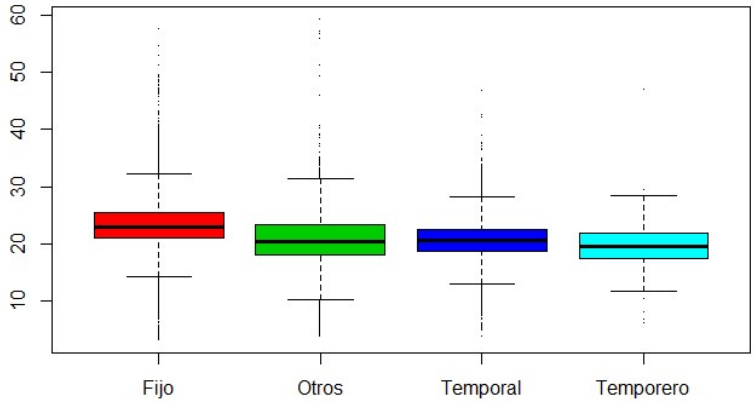


Figura 3.1.10: Diagrama de cajas del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

El nivel medio de ingresos en cada grupo de clientes es el siguiente:

- Para el grupo con contrato fijo es de 1.599,61 €.
- Para el grupo "Otros" (parados y pensionistas) es de 1.097,23 €.
- Para el grupo con contrato temporal es de 1.088,82 €.
- Para los temporeros es de 921,40 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SITUACIÓN LABORAL se puede apreciar mejor en la figura 3.1.11.

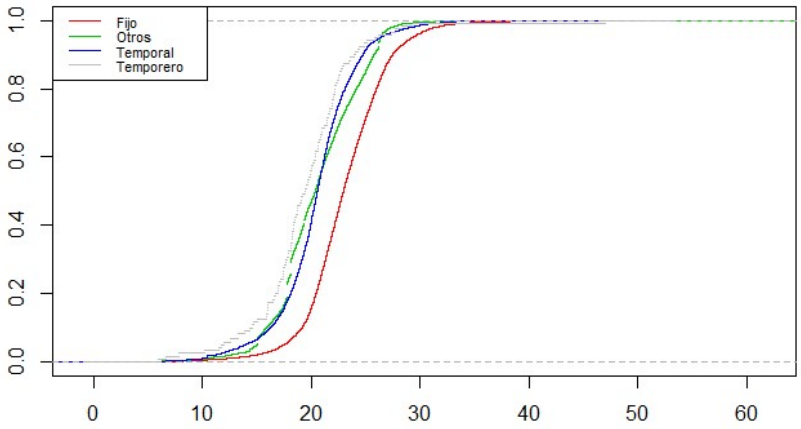


Figura 3.1.11: Comparación de la función de distribución del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SITUACIÓN LABORAL respecto de los ingresos es de 0,1025.

• SECTOR ACTIVIDAD

Al igual que ocurre con la variable AUTONOMÍA, la variable SECTOR ACTIVIDAD también es cualitativa con un número demasiado alto de niveles como para modelizarla directamente, por este motivo llevaremos a cabo un análisis clúster que nos permita quedarnos con un número más pequeño de grupos de sectores que sean lo más heterogéneos posible entre sí y en los que los sectores que forman cada grupo sean lo más semejantes posible en términos de ingresos. Para ello, y como ya hicimos anteriormente, nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada uno de los sectores de actividad. En la figura 3.1.12 podemos observar el dendograma que nos proporciona los distintos conjuntos.

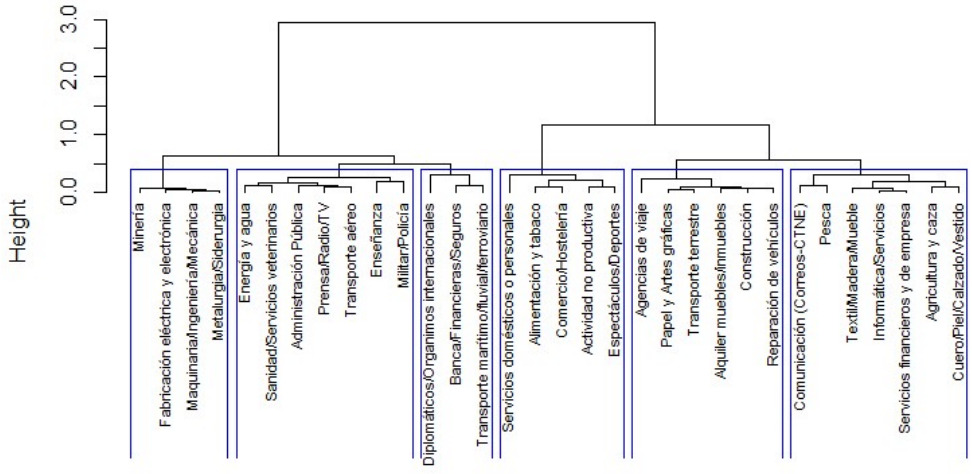


Figura 3.1.12: Dendograma que divide la variable SECTOR ACTIVIDAD en 6 clústeres.

De esta forma obtenemos los siguientes grupos:

- **Clúster 1:**
  - Metalurgia/Siderurgia
  - Maquinaria/Ingeniería/Mecánica
  - Minería
  - Fabricación eléctrica y electrónica

  4.821 obs.
- **Clúster 2:**
  - Administración Pública
  - Energía y agua
  - Enseñanza
  - Militar/Policía
  - Prensa/Radio/TV
  - Sanidad/Servicios veterinarios
  - Transporte aéreo

  12.915 obs.
- **Clúster 3:**
  - Diplomáticos/Organismos Internacionales
  - Banca/Financieras/Seguros
  - Transporte marítimo/fluvial/ferroviario

  1.388 obs.
- **Clúster 4:**
  - Actividad no productiva
  - Alimentación y tabaco
  - Comercio/Hostelería
  - Espectáculos/Deportes
  - Servicios domésticos o personales

  29.658 obs.
- **Clúster 5:**
  - Agencias de viaje
  - Alquiler muebles/Inmuebles
  - Construcción
  - Papel y artes gráficas
  - Reparación de vehículos
  - Transporte terrestre

  3.421 obs.
- **Clúster 6:**
  - Agricultura y caza
  - Comunicación (Correos/CTNE)
  - Cuero/Piel/Calzado/Vestido
  - Informática/Servicios
  - Pesca
  - Servicios financieros y de empresa
  - Textil/Madera/Mueble

  6.887 obs.

En la figura 3.1.13 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SECTOR ACTIVIDAD a través de su correspondiente diagrama de cajas.

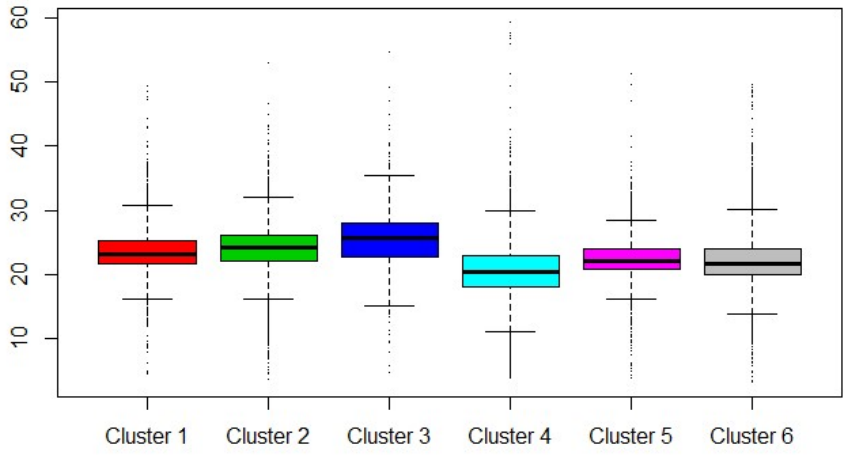


Figura 3.1.13: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

El nivel medio de ingresos en cada clúster es el siguiente:

- Para el **clúster 1** es de 1.688,05 €.
- Para el **clúster 2** es de 1.784,83 €.
- Para el **clúster 3** es de 2.168,27 €.
- Para el **clúster 4** es de 1.072,15 €.
- Para el **clúster 5** es de 1.420,70 €.
- Para el **clúster 6** es de 1.386,36 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SECTOR ACTIVIDAD se puede apreciar mejor en la figura 3.1.14.

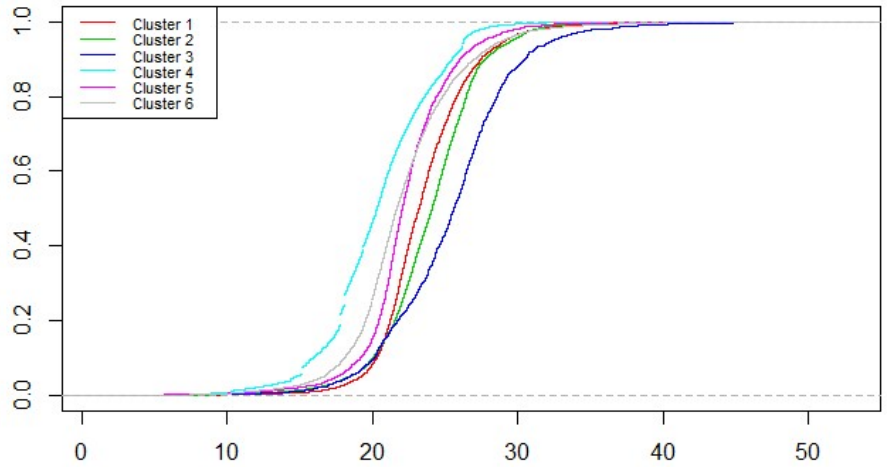


Figura 3.1.14: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SECTOR ACTIVIDAD respecto de los ingresos es de 0,1539.

• INGRESOS FIJOS

Una vez que hemos estudiado todas las variables, tanto sociodemográficas como socio-laborales, que vamos a tener en el modelo, completaremos el modelo introduciendo algunas de las variables económicas más significativas que aparecen en la captura de bienes.

En primer lugar nos centraremos en los ingresos fijos que percibe cada cliente mensualmente. A dicha variable también se le aplicará la transformación Box-Cox (con un λ de 0.275) para reducir su gran asimetría, al igual que hemos hecho con la variable ingresos que estamos utilizando como variable respuesta en el ajuste del modelo.

Si hacemos un resumen de la variable tenemos:

```
> summary(z3.2$INGRESOS_FIJOS_MES3)
    Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
    4.57       19.64       21.83       22.01       24.36       57.66
   (20€)      (947€)     (1.323€)    (1.358€)    (1.883€)    (34.995€)
```

Podemos ver que el rango total de ingresos es muy grande (va desde los 20 euros hasta los 35.000 euros), aunque gran parte de ellos se centran en torno a la media (que es de 1.358 euros). Representando la función de densidad estimada de la variable transformada apreciamos mejor cómo es su distribución.

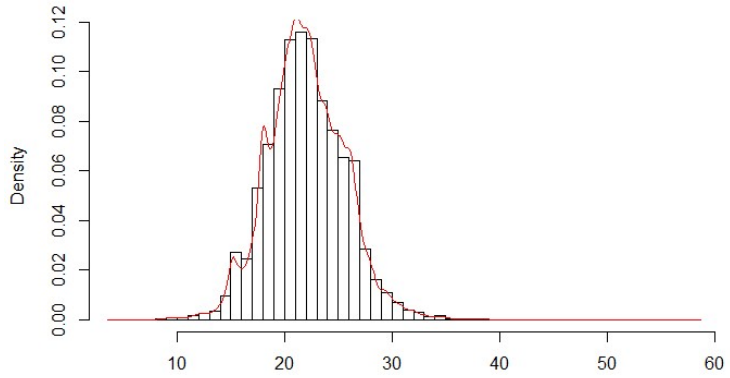


Figura 3.1.15: Histograma y estimación tipo núcleo de la función de densidad de los ingresos fijos transformados.

En la figura 3.1.15 también podemos ver algunos picos que se corresponden con ingresos fijos que se repiten en varios clientes, como pueden ser pensiones o salarios mínimos (algo que ya nos pasaba con la variable ingresos que utilizamos como variable respuesta).

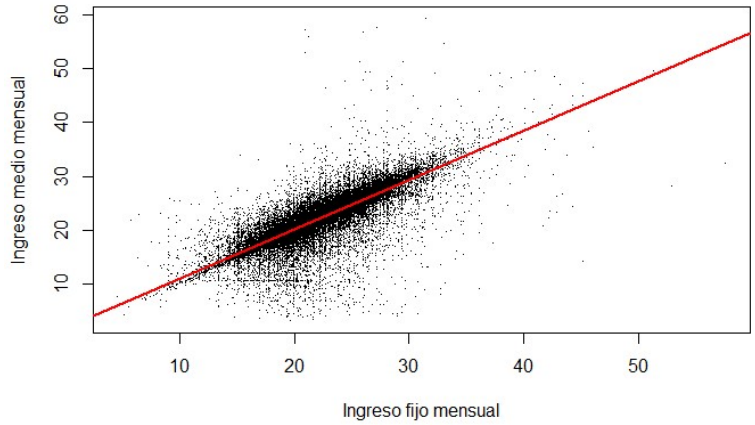


Figura 3.1.16: Diagrama de dispersión del ingreso medio transformado en función del ingreso fijo transformado con su correspondiente ajuste lineal.

En la figura 3.1.16 se aprecia una clara relación lineal entre ambas variable y se representa en rojo la línea recta que mejor se ajusta a dicha relación. Calculando el modelo de regresión lineal simple que explique los ingresos en función de los ingresos fijos obtenemos un R^2 que llega hasta el 0,6824.

• INGRESOS UNIDAD FAMILIAR

Finalmente, la última variable que introduciremos en el modelo será los ingresos de la unidad familiar que se corresponden con el total de ingresos que perciben los miembros que conforman dicha familia (padres e hijos o incapacitados) que viven en la misma vivienda. Como ya hemos hecho con las anteriores variables de ingresos, a esta también se le aplicará la transformación Box-Cox (con un λ de 0.282) para obtener un ajuste más adecuado.

Si hacemos un resumen de la variable tras aplicarle la transformación se obtiene:

```
> summary(z3.2$INGRESOS_UNIDAD_FAMILIAR2)
   Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
  3.529      20.89       23.60       23.78       26.35       44.947
  (12€)   (1.150€)    (1.700€)    (1.742€)    (2.431€)    (14.700€)
```

En esta variable también se aprecia que el rango total de valores es muy grande (desde los 12 euros a los casi 15 mil euros), a pesar de que la mayoría se encuentra en torno a la media (que es de 1.742 euros).

Esta distribución la apreciamos mejor al representar la función de densidad estimada de la transformación Box-Cox de la variable (figura 3.1.17).

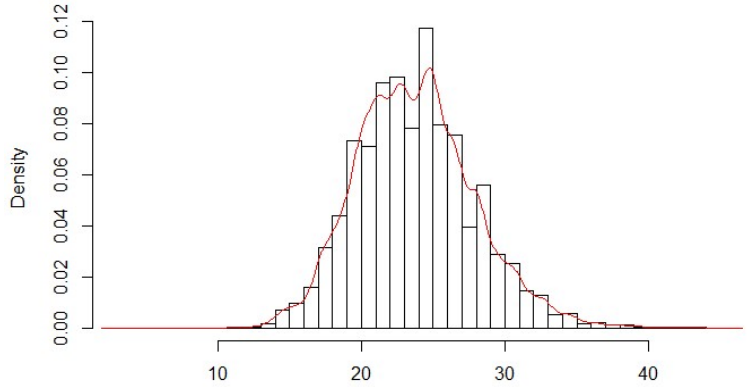


Figura 3.1.17: Histograma y estimación tipo núcleo de la función de densidad de los ingresos de la unidad familiar transformados.

Al igual que observamos en las demás variables de ingresos, en los ingresos de la unidad familiar la función de densidad estimada es una función multimodal debido a la repetición de algunos de los valores en determinados clientes, aunque en este caso es menos acusado que en las variables anteriores.

Si ponemos en relación los ingresos de la unidad familiar con los ingresos (variable respuesta) obtenemos el diagrama de dispersión de la figura 3.1.18.

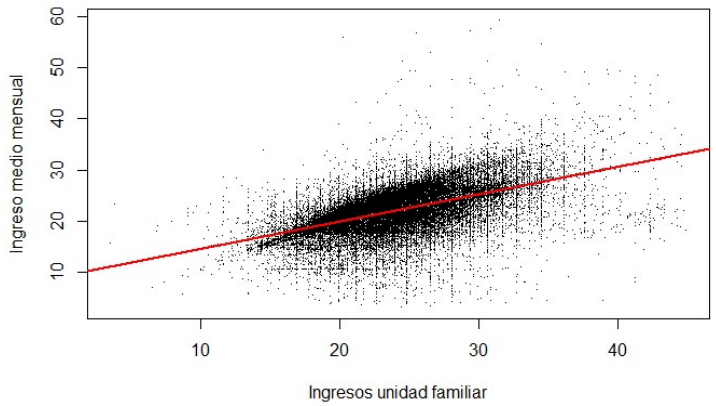


Figura 3.1.18: Diagrama de dispersión del ingreso medio transformado en función del ingreso de la unidad familiar con su ajuste lineal.

Calculando el modelo de regresión lineal simple que explique los ingresos en función de los ingresos de la unidad familiar obtenemos un R^2 de 0,3059.

Modelo de regresión

Una vez analizadas individualmente cada una de las variables que forman parte del modelo ya estamos en condiciones de realizar el ajuste que intente explicar de la mejor forma posible el comportamiento del ingreso medio transformado (variable respuesta) en función de las variables explicativas que hemos seleccionado. Entre estas últimas destacan principalmente los ingresos fijos y los ingresos de la unidad familiar por ser las dos variables que más información aportan acerca del nivel de ingresos mensuales de los clientes.

El modelo seleccionado para realizar el ajuste consiste en un modelo de regresión múltiple en el que las variables explicativas se introducen en su mayoría en forma lineal, aunque en el caso de la variable “edad”

se decide ajustarla de forma no paramétrica a través de bases B-Splines con el objetivo de obtener una mayor flexibilidad en su estimación.

De esta forma, R nos proporciona la siguiente salida con los coeficientes y los errores de las variables que forman el modelo.

```
> summary(m)
Call:
lm(formula = z3.2$INGRESO_MEDIO2 ~ z3.2$SEXO + bs(z3.2$EDAD2,
    8) + z3.2$CLUSTER_AUT + z3.2$SITUACION_LABORAL_ID + z3.2$CLUSTER_SEC_ACT +
    z3.2$INGRESOS_FIJOS_MES3 + z3.2$INGRESOS_UNIDAD_FAMILIAR2)
Residuals:
    Min      1Q  Median      3Q     Max
-26.849  -0.463   0.130   0.750  35.746
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.684331   0.337669   2.027 0.042704 *
z3.2$SEXOMujer                 -0.320484   0.020218 -15.852  < 2e-16 ***
bs(z3.2$EDAD2, 8)1              1.914279   0.357158   5.360 8.36e-08 ***
bs(z3.2$EDAD2, 8)2              1.613674   0.216808   7.443 9.99e-14 ***
bs(z3.2$EDAD2, 8)3              1.796329   0.260850   6.886 5.78e-12 ***
bs(z3.2$EDAD2, 8)4              1.705118   0.240695   7.084 1.41e-12 ***
bs(z3.2$EDAD2, 8)5              1.923029   0.251379   7.650 2.04e-14 ***
bs(z3.2$EDAD2, 8)6              1.373552   0.267514   5.134 2.84e-07 ***
bs(z3.2$EDAD2, 8)7              1.307916   0.405187   3.228 0.001248 **
bs(z3.2$EDAD2, 8)8              2.970888   0.985500   3.015 0.002574 **
z3.2$CLUSTER_AUTCluster 2       0.680747   0.221218   3.077 0.002090 **
z3.2$CLUSTER_AUTCluster 3       1.019118   0.268066   3.802 0.000144 ***
z3.2$CLUSTER_AUTCluster 4       0.727132   0.226268   3.214 0.001312 **
z3.2$CLUSTER_AUTCluster 5       0.824698   0.230486   3.578 0.000346 ***
z3.2$CLUSTER_AUTCluster 6      -4.894094   0.347011 -14.104  < 2e-16 ***
z3.2$SITUACION_LABORAL_IDOtros  0.125958   0.039105   3.221 0.001278 **
z3.2$SITUACION_LABORAL_IDTemporal -0.251734   0.031198  -8.069 7.22e-16 ***
z3.2$SITUACION_LABORAL_IDTemporero -0.182664   0.206341  -0.885 0.376023
z3.2$CLUSTER_SEC_ACTCluster 2   0.032547   0.039571   0.822 0.410801
z3.2$CLUSTER_SEC_ACTCluster 3   0.537399   0.069290   7.756 8.92e-15 ***
z3.2$CLUSTER_SEC_ACTCluster 4  -0.700371   0.042760 -16.379  < 2e-16 ***
z3.2$CLUSTER_SEC_ACTCluster 5  -0.402675   0.050540  -7.968 1.65e-15 ***
z3.2$CLUSTER_SEC_ACTCluster 6  -0.253398   0.042850  -5.914 3.37e-09 ***
z3.2$INGRESOS_FIJOS_MES3        0.836434   0.003631 230.361  < 2e-16 ***
z3.2$INGRESOS_UNIDAD_FAMILIAR2  0.038639   0.002928  13.199  < 2e-16 ***
Residual standard error: 2.251 on 59064 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6933, Adjusted R-squared:  0.6932
F-statistic:  5563 on 24 and 59064 DF,  p-value: < 2.2e-16
```

Como podemos ver en la salida que nos da R del modelo ajustado, tan solo el nivel "Temporero" de la variable SITUACIÓN LABORAL y el nivel "Cluster 2" de la variable SECTOR ACTIVIDAD no son significativos. A pesar de esto, los mantenemos en el modelo ya que proporciona un mejor ajuste. Por otro lado, también podemos ver que el coeficiente de determinación ajustado (R^2) del modelo es de 0.6932, lo que quiere decir que dicho modelo explica cerca del 70% de la variabilidad de los ingresos.

Diagnosis del modelo

Una vez ajustado el modelo de regresión, el último paso consistirá en comprobar si este verifica las hipótesis estructurales básicas de normalidad, homocedasticidad e independencia de los residuos.

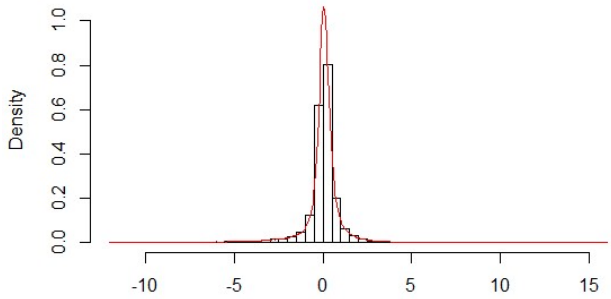


Figura 3.1.19: Histograma y estimación tipo núcleo de la función de densidad de los residuos estandarizados.

En la figura 3.1.19 vemos que la media parece estar en torno al 0, aunque también podemos observar un pico en la zona central que nos hace pensar en la falta de normalidad de los residuos.

También hacemos un gráfico de los residuos estandarizados frente a los valores ajustados que observamos en la figura 3.1.20.

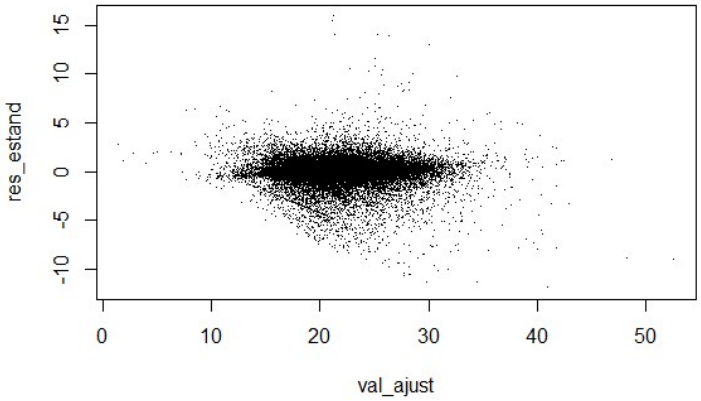


Figura 3.1.20: Diagrama de dispersión de los residuos estandarizados en función de los valores ajustados.

En la figura 3.1.20 se aprecia algo de heterocedasticidad, es decir, parece que conforme aumenta el valor ajustado de la variable respuesta va aumentando la variabilidad de la misma, hasta llegar a un punto en el que comienza a disminuir.

Por otro lado, también llevamos a cabo el test de Durbin-Watson que nos proporciona un estadístico de contraste de 1.9998, muy cercano a 2, por lo que llegamos a la conclusión de que no existe correlación lineal entre los residuos del ajuste.

Para comprobar la normalidad de los residuos estandarizados realizamos un Q-Q plot como el de la figura 3.1.21.

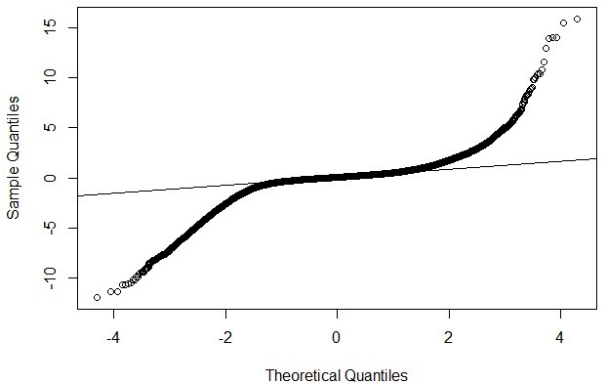


Figura 3.1.21: Q-Q plot de los residuos estandarizados

Al igual que habíamos observado en las figuras anteriores, en la figura 4.1.21 también llegamos a la conclusión de que los residuos no siguen una distribución normal ya que los puntos en las colas se encuentran muy distanciados de la diagonal que marca la normalidad.

Además, si llevamos a cabo diversos test de normalidad obtenemos los siguientes resultados:

- **Test de Lilliefors:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Anderson Darling:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Cramer von Mises:** Obtenemos un p-valor de $7.37x10^{-10}$ que nos lleva a rechazar la hipótesis de normalidad.

3.2. Segmento 2: Captura de bienes (Ingresos fijos)

Para el ajuste de este modelo utilizaremos, al igual que en el anterior, la información de aquellos clientes de los que tenemos tanto los datos de ingresos como los datos de captura de bienes. En este caso, eliminaremos la variable de ingresos de la unidad familiar respecto del modelo anterior ya que es la que más valores nulos tiene. De esta forma pasamos de tener las 141.146 observaciones, propias de la BD de Captura de Bienes, a tener 132.155 observaciones una vez que prescindimos de los autónomos, y que se quedarían en 66.905 al eliminar los clientes con valores nulos en la variable de ingresos de la unidad familiar.

Las variables que van a formar parte del modelo son las siguientes:

- **INGRESO MEDIO:** Variable respuesta que nos dice el ingreso medio mensual de cada cliente y cuyo comportamiento intentaremos explicar a través del resto de variables (explicativas).
- **SEXO:** Variable cualitativa con dos niveles (hombre y mujer).
- **EDAD:** Variable cuantitativa y continua que nos indica la edad de cada cliente (en años).
- **AUTONOMÍA:** Variable cualitativa que nos muestra la comunidad autónoma a la que pertenece cada cliente.
- **SITUACIÓN LABORAL:** Variable cualitativa que nos indica la situación en la que se encuentra cada cliente dentro del mercado laboral. No tendremos en cuenta a los autónomos ya que sus ingresos son más difíciles de estimar ya que no se registran con concepto de nómina o pensión por lo que para la mayoría de ellos no dispondremos de variable respuesta con información fiable. Además los datos de la captura de bienes no se registran como ingresos fijos sino cómo ingresos variables.
- **SECTOR ACTIVIDAD:** Variable cualitativa que nos ofrece información acerca del sector al que pertenece el trabajo que desempeña cada cliente.
- **INGRESOS FIJOS NETOS:** Variable cuantitativa que nos indica la cantidad de ingresos fijos que percibe cada cliente al mes.

Prescindiremos del resto de variables presentes en la BD de Captura de Bienes ya sea porque aportan información muy similar a la que aportan algunas de las variables presentes en el modelo (por ejemplo, la variable PROFESIÓN aportaría información muy similar a la de SECTOR ACTIVIDAD) o porque la información que aportan es muy poco significativa (como es el caso de las variables UNIDAD FAMILIAR, ESTADO CIVIL o RÉGIMEN MATRIMONIAL).

Para hacernos una idea de la importancia que puede tener cada una de las variables en el ajuste del modelo será conveniente hacer una pequeña introducción individual de cada una de ellas en la que haremos un resumen de la información que nos puede aportar en relación al nivel de ingresos de cada cliente.

- **INGRESO MEDIO (Variable Respuesta)**

Como ya hemos visto, esta variable la obtenemos como una media mensual de los ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017. A ella se le realiza una transformación Box-Cox que nos permita obtener un mejor ajuste.

Si hacemos un resumen de dicha variable para los clientes que pertenecen a este segmento de captura de bienes:

```
> summary(z3.2$INGRESO_MEDIO2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
  3.079     16.375     18.221     18.283     20.228     48.851
  (10€)     (859€)   (1.241€)   (1.256€)   (1.792€)   (50.000€)
```

Podemos ver que el rango total de la variable es muy grande (va desde los 10 euros hasta los más de 50.000 euros), aunque gran parte de ellos se centran en torno a la media (que es de 1.256 euros). Representando la función de densidad estimada de la transformación Box-Cox de la variable (figura 3.2.1) apreciamos mejor cómo es su distribución.

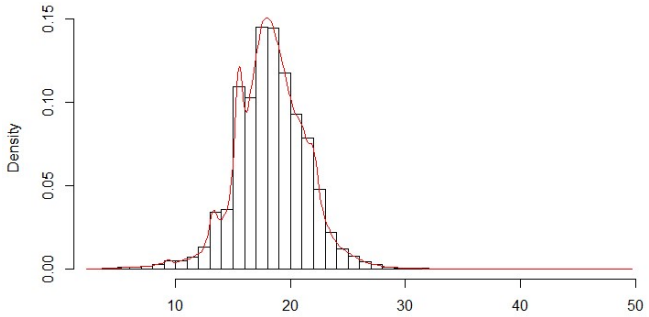


Figura 3.2.1: Histograma y estimación tipo núcleo de la función de densidad del Ingreso Medio transformado.

- **SEXO**

Si hacemos un resumen de la variable podemos ver que:

```
> summary(z3.2$SEXO)
Hombre  Mujer   NA's
 37247  29656      2
```

El 55,67% de los clientes de la BD son hombres, mientras que el 44,33 % restante son mujeres (cabe destacar que hay dos casos en el que no tenemos información del sexo del cliente).

Si relacionamos la variable SEXO con los ingresos podemos representar los resultados en el siguiente diagrama de cajas:

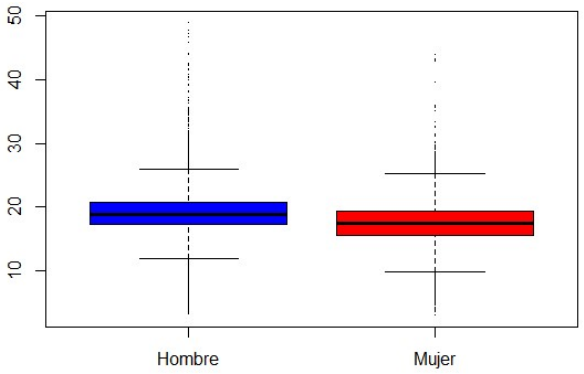


Figura 3.2.2: Diagrama de cajas del ingreso medio transformado diferenciando por sexo.

Si nos fijamos en la figura 3.2.2 lo primero que debemos destacar es la gran variabilidad que tienen los ingresos tanto en el caso de los hombres como en el de las mujeres. Lo que también podemos observar es que el nivel medio de ingresos es ligeramente superior en el caso de los hombres (con un ingreso medio de 1413,29 €) respecto de las mujeres (con una media de 1079,17 €).

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SEXO respecto de los ingresos es de 0,04885.

- **EDAD**

Si hacemos un resumen de la variable:

```
> summary(z3.2$EDAD2)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  19.03     40.05     52.71     53.15     66.04    101.89
```

Vemos que la edad de los clientes va desde los 19 años hasta los 102, siendo la media de edad de unos 53 años.

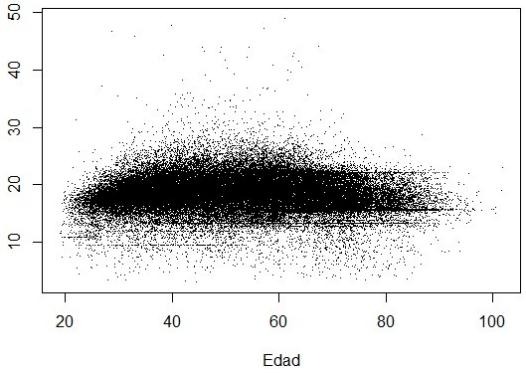
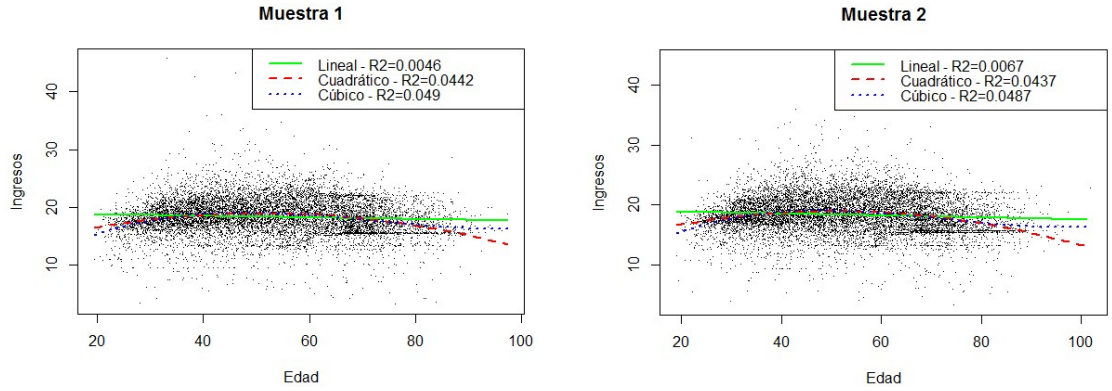


Figura 3.2.3: Diagrama de dispersión del ingreso medio transformado en función de la edad.

A simple vista, en la figura 3.2.3 no se aprecia una relación lineal entre la edad y los ingresos, por este motivo es recomendable llevar a cabo una estimación polinómica o hasta una estimación no paramétrica de la relación entre ambas variables para ver de qué forma obtenemos un mejor ajuste.

En primer lugar llevaremos a cabo una estimación polinómica de dicha relación. Para ello seleccionaremos una muestra aleatoria de tamaño 10.000 de la variable edad para así poder visualizar mejor en un gráfico la estimación polinómica de grado 2 y 3 de la regresión del ingreso medio transformado respecto de la edad, junto con sus respectivos coeficientes de correlación ajustados.

Los gráficos que obtenemos para distintas muestras son los contenidos en la figura 3.2.4.



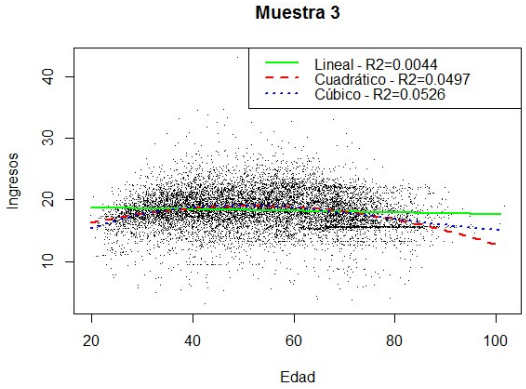


Figura 3.2.4: Representación de los ajustes polinómicos del ingreso medio en función de la edad para tres muestras distintas.

Podemos ver que el coeficiente de determinación, a pesar de ser todavía muy bajo, crece significativamente cuando introducimos la estimación polinómica (tanto cuadrática como cúbica).

Para dotar de una mayor flexibilidad a la regresión podemos llevar a cabo una estimación polinómica local de la regresión mediante bases B-Splines.

Los gráficos que obtenemos para distintas muestras de esta forma son los contenidos en la figura 3.2.5.

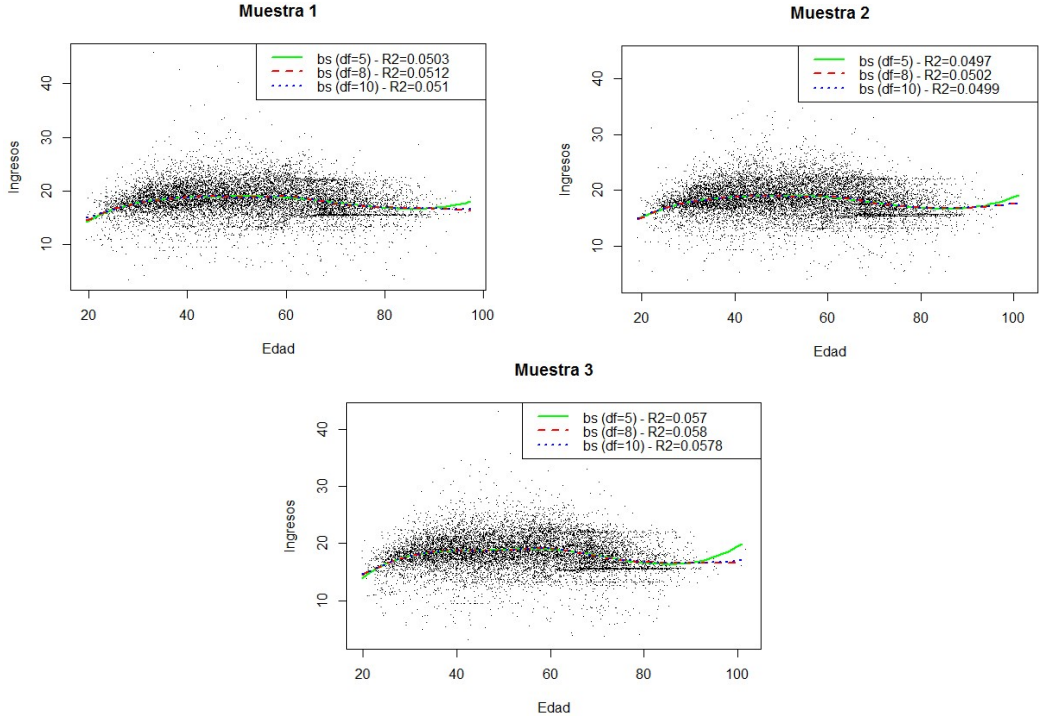


Figura 3.2.5: Representación del ajuste polinómico local del ingreso medio transformado en función de la edad mediante B-Splines con diferentes grados de libertad para tres muestras distintas.

Podemos ver que el ajuste polinómico local se parece bastante al ajuste cúbico que vimos anteriormente, y que a partir de los 8 grados de libertad el aumento del R^2 ajustado es muy poco significativo, y en algún caso puede llegar a disminuir.

Tras este análisis llegamos a la conclusión de que el mejor ajuste para la relación entre la variable EDAD y los ingresos es el que obtenemos a partir de la estimación no paramétrica mediante B-Splines con 8 grados de libertad, por lo que será la que utilizaremos en el modelo. En la figura 3.2.6 podemos observar la representación gráfica de dicho ajuste.

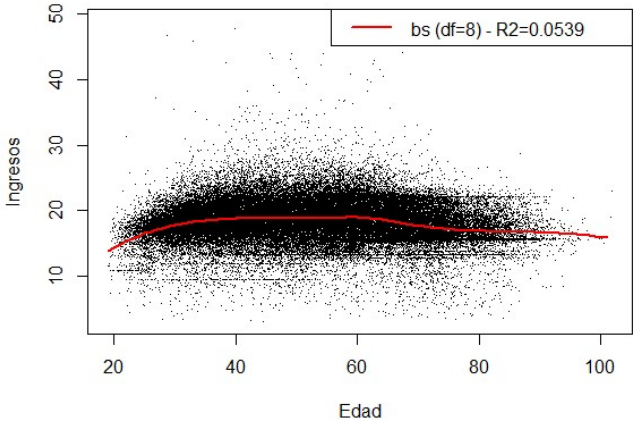


Figura 3.2.6: Diagrama de dispersión del ingreso medio transformado en función de la edad con su correspondiente ajuste polinómico local.

- **AUTONOMÍA**

Como la variable AUTONOMÍA es cualitativa con un número demasiado alto de niveles como para tratarla directamente así, es necesario realizar un análisis clúster que nos permita agruparlos en una cantidad más manejable de grupos. Para ello nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada una de las comunidades autónomas. En la figura 3.2.7 podemos observar el dendograma que nos proporciona los distintos conjuntos.

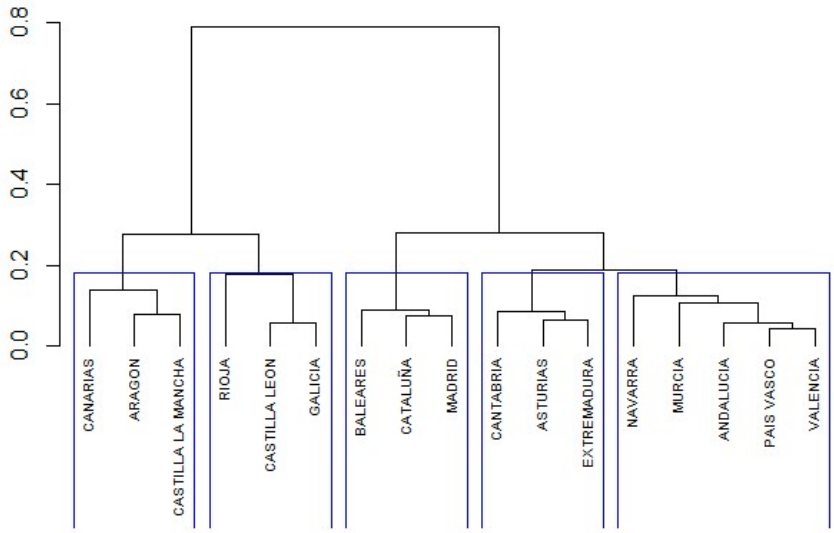


Figura 3.2.7: Dendograma que divide la variable AUTONOMÍA en 5 clústeres.

De esta forma obtenemos los siguientes conjuntos:

- **Clúster 1:** Aragón, Canarias y Castilla La Mancha (formado por 491 obs).
- **Clúster 2:** Castilla León, Galicia y La Rioja (formado 62.502 obs).
- **Clúster 3:** Baleares, Cataluña y Madrid (formado por 1.986 obs).
- **Clúster 4:** Asturias, Cantabria y Extremadura (formado por 622 obs).
- **Clúster 5:** Andalucía, Murcia, Navarra, País Vasco y Valencia (formado por 1.236 obs).
- **Clúster 6:** En este conjunto incluimos a Ceuta y Melilla (por tener datos insuficientes para poder analizarlas convenientemente) y a Otros (clientes extranjeros) aunque, en un principio, no nos centraremos en su análisis (formado por 67 obs).

En la figura 3.2.8 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable AUTONOMÍA a través de su correspondiente diagrama de cajas.

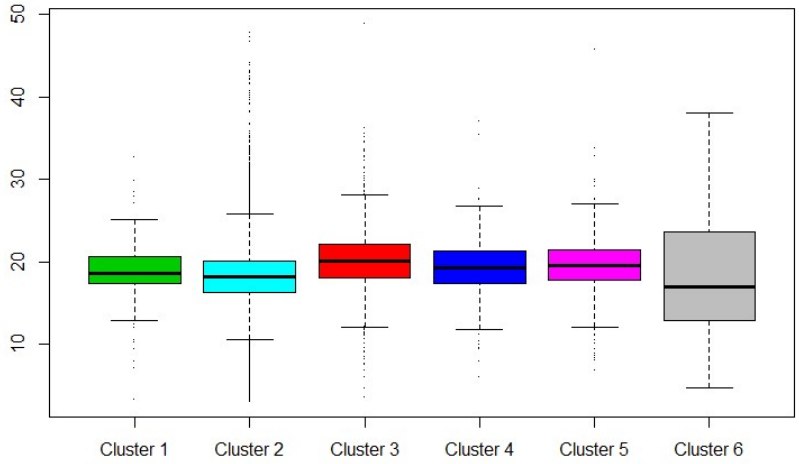


Figura 3.2.8: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

El nivel medio de ingresos en cada clúster será el siguiente:

- Para el **clúster 1** es de 1.400,23 €.
- Para el **clúster 2** es de 1.231,69 €.
- Para el **clúster 3** es de 1.783,74 €.
- Para el **clúster 4** es de 1.523,91 €.
- Para el **clúster 5** es de 1.606,09 €.
- Para el **clúster 6** es de 1.257,84 €.

Este último dato nos es muy preciso ya que en este grupo se engloban clientes con características muy diferentes.

La distribución de los ingresos en cada uno de los distintos clústeres se puede apreciar mejor en la figura 3.2.9.

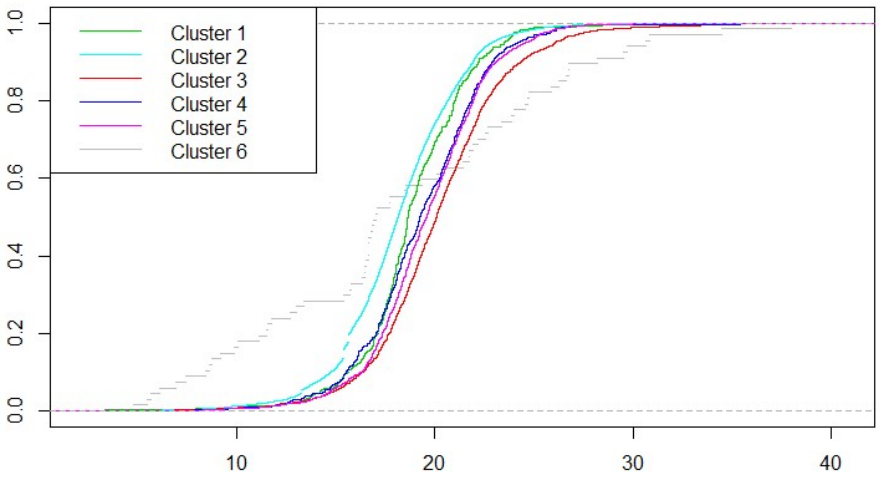


Figura 3.2.9: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable AUTONOMÍA respecto de los ingresos es de 0,01625.

- **SITUACIÓN LABORAL**

La variable SITUACION LABORAL es cualitativa y consta de 4 niveles diferentes. Si hacemos un pequeño resumen de la misma:

```
> summary(z3.2$SITUACION_LABORAL_ID)
     Fijo     Otros   Temporal   Temporero
    34016     24889       7837        163
```

El grupo más numeroso es el de aquellos clientes con contrato fijo (un 50,84 % del total), seguido del grupo "Otros" formado por parados y pensionistas (un 37,20 % del total). En un término medio se encuentra el grupo de clientes con contrato temporal (un 11,71 % del total). Finalmente, el grupo con un número de clientes más bajo es el de los temporeros (0,25 % del total).

En la figura 3.2.10 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SITUACIÓN LABORAL a través de su correspondiente diagrama de cajas.

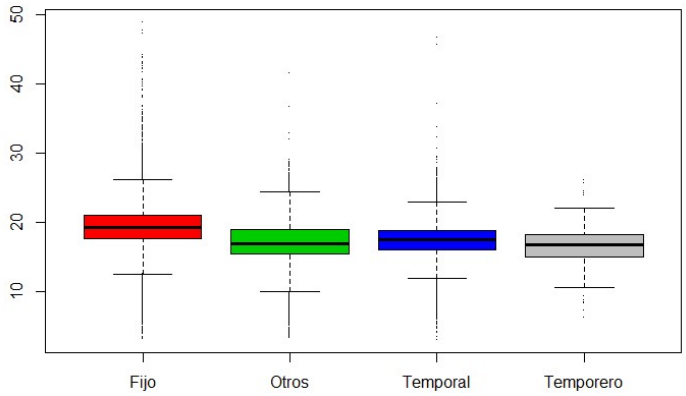


Figura 3.2.10: Diagrama de cajas del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

El nivel medio de ingresos en cada grupo de clientes es el siguiente:

- Para el grupo con contrato **fijo** es de 1.532,26 €.
- Para el grupo **"Otros"** (parados y pensionistas) es de 1.006,56 €.
- Para el grupo con contrato **temporal** es de 1.037,20 €.
- Para los **temporeros** es de 872,62 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SITUACIÓN LABORAL se puede apreciar mejor en la figura 3.2.11.

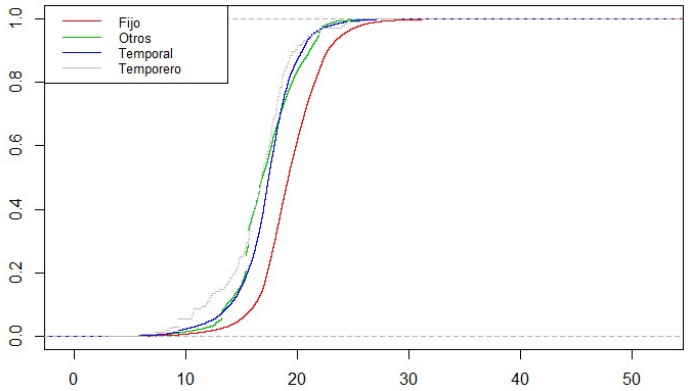


Figura 3.2.11: Comparación de la función de distribución del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SITUACIÓN LABORAL respecto de los ingresos es de 0,1181.

- **SECTOR ACTIVIDAD**

Al igual que ocurre con la variable AUTONOMÍA, la variable SECTOR ACTIVIDAD también es cualitativa con un número demasiado alto de niveles como para modelizarla directamente, por este motivo llevaremos a cabo un análisis clúster que nos permita quedarnos con un número más pequeño de grupos de sectores que sean lo más heterogéneos posible entre sí y en los que los sectores que forman cada grupo sean lo más semejantes posible en términos de ingresos. Para ello, y como ya hicimos anteriormente, nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada uno de los sectores de actividad. En la figura 3.2.12 podemos observar el dendograma que nos proporciona los distintos conjuntos.

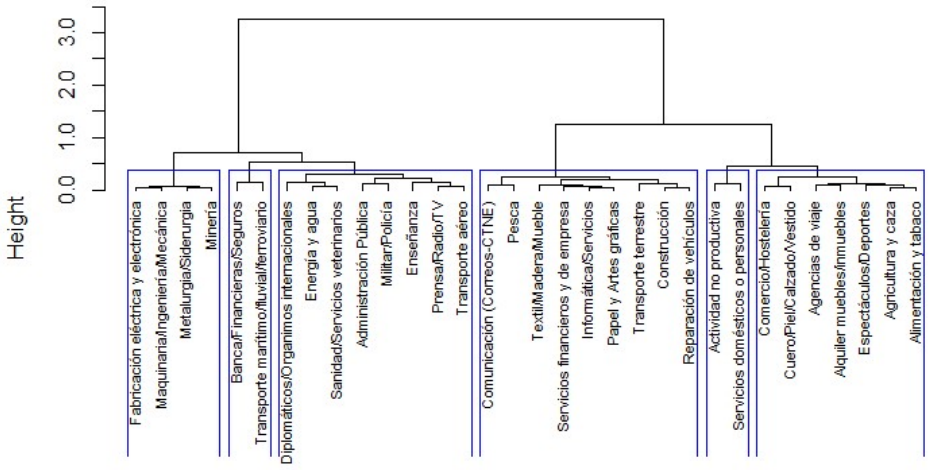


Figura 3.2.12: Dendograma que divide la variable SECTOR ACTIVIDAD en 6 clústeres.

De esta forma obtenemos los siguientes conjuntos:

- **Clúster 1:**
  - Metalurgia/Siderurgia
  - Maquinaria/Ingeniería/Mecánica
  - Minería
  - Fabricación eléctrica y electrónica

  5.837 obs.
- **Clúster 2:**
  - Banca/Financieras/Seguros
  - Transporte marítimo/fluvial/ferroviario

  1.379 obs.
- **Clúster 3:**
  - Administración Pública
  - Diplomáticos/Organismos Internacionales
  - Energía y agua
  - Enseñanza
  - Militar/Policía
  - Prensa/Radio/TV
  - Sanidad/Servicios veterinarios
  - Transporte aéreo

  12.915 obs.
- **Clúster 4:**
  - Comunicación (Correos/CTNE)
  - Construcción
  - Informática/Servicios
  - Papel y artes gráficas
  - Pesca
  - Reparación de vehículos
  - Servicios financieros y de empresa
  - Textil/Madera/Mueble
  - Transporte terrestre

  3.421 obs.

- **Clúster 5:**
  - Actividad no productiva
  - Servicios domésticos o personales

  29.658 obs.
- **Clúster 6:**
  - Agencias de viaje
  - Agricultura y caza
  - Alimentación y tabaco
  - Alquiler muebles/Inmuebles
  - Comercio/Hostelería
  - Cuero/Piel/Calzado/Vestido
  - Espectáculos/Deportes

  3.421 obs.

En la figura 3.2.13 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SECTOR ACTIVIDAD a través de su correspondiente diagrama de cajas.

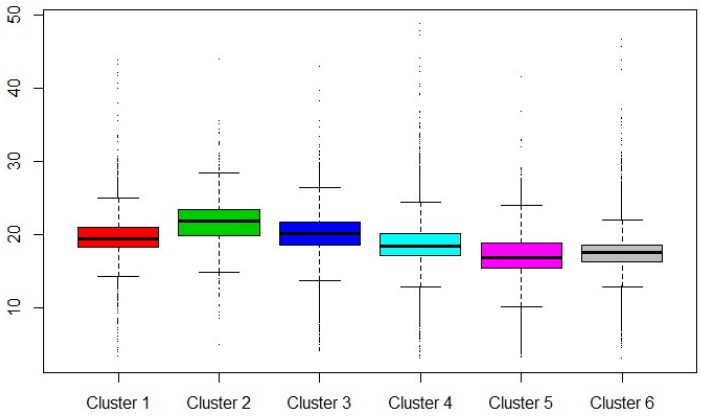


Figura 3.2.13: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

El nivel medio de ingresos en cada clúster es el siguiente:

- Para el **clúster 1** es de 1.626,42 €.
- Para el **clúster 2** es de 2.292,15 €.
- Para el **clúster 3** es de 1.731,08 €.
- Para el **clúster 4** es de 1.366,72 €.
- Para el **clúster 5** es de 988,51 €.
- Para el **clúster 6** es de 1.074,92 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SECTOR ACTIVIDAD se puede apreciar mejor en la figura 3.2.14.

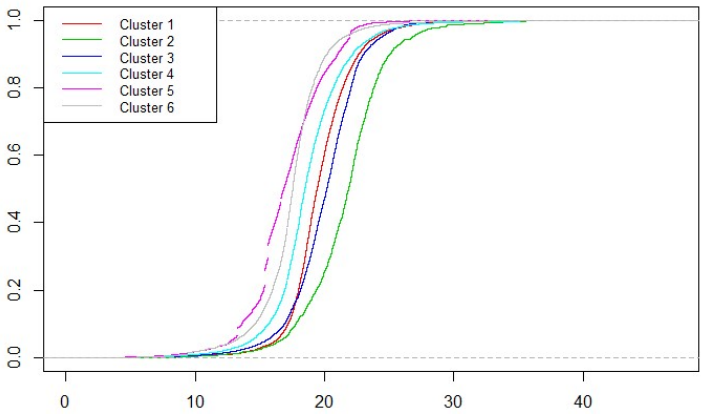


Figura 3.2.14: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SECTOR ACTIVIDAD respecto de los ingresos es de 0,1697.

- **INGRESOS FIJOS**

Una vez que hemos estudiado todas las variables, tanto sociodemográficas como socio-laborales, que vamos a tener en el modelo, completaremos el modelo introduciendo la variable de ingresos fijos que, como es lógico, será la que más información aportará de todas las variables explicativas.

A dicha variable también se le aplicará la transformación Box-Cox (con un λ de 0.269) para reducir su fuerte asimetría, al igual que hemos hecho con la variable ingresos que estamos utilizando como variable respuesta en el ajuste del modelo.

Si hacemos un resumen de la variable tenemos:

```
> summary(z3.2$INGRESOS_FIJOS_MES3)
    Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
    4.06        16.60        18.23        18.39        20.02        59.41
   (18€)       (900€)     (1.243€)     (1.282€)     (1.750€)     (109.077€)
```

Podemos ver que el rango total de ingresos es muy grande (va desde los 18 euros hasta los casi 110.000 euros), aunque gran parte de ellos se centran en torno a la media (que es de 1.282 euros). Representando la función de densidad estimada de la transformación Box-Cox de la variable apreciamos mejor cómo es su distribución.

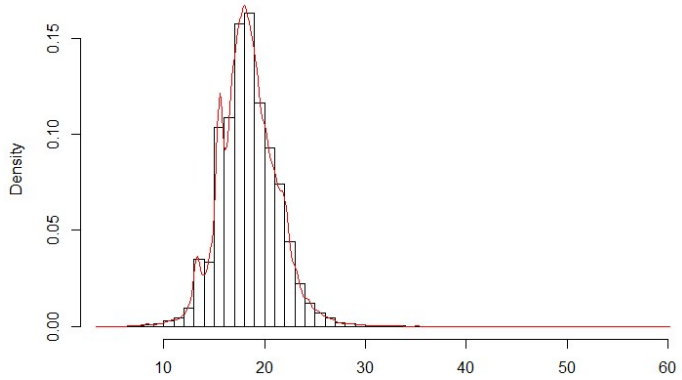


Figura 3.2.15: Histograma y estimación tipo núcleo de la función de densidad de los ingresos fijos transformados.

En la figura 3.2.15 también podemos ver algunos picos que se corresponden con ingresos fijos que se repiten en varios clientes, como pueden ser pensiones o salarios mínimos (algo que ya nos pasaba con la variable ingresos que utilizamos como variable respuesta).

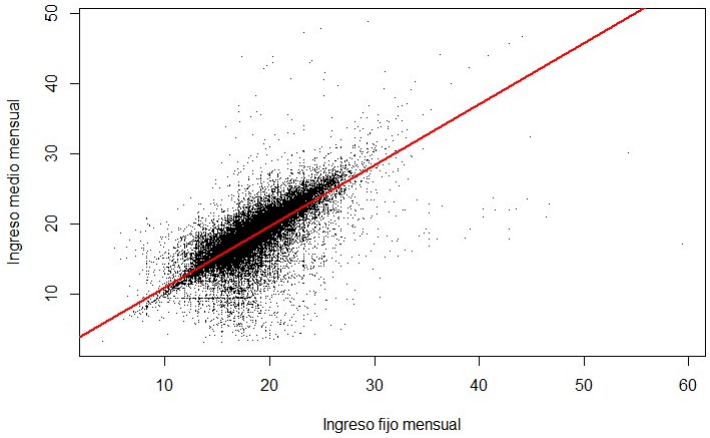


Figura 3.2.16: Diagrama de dispersión del ingreso medio transformado en función del ingreso fijo transformado con su ajuste lineal correspondiente.

En la figura 3.2.16 se aprecia una clara relación lineal entre ambas variable y se representa en rojo la línea recta que mejor se ajusta a dicha relación. Calculando el modelo de regresión lineal simple que explique los ingresos en función de los ingresos fijos obtenemos un R^2 que llega hasta el 0,6243.

Modelo de regresión

Una vez analizadas individualmente cada una de las variables que forman parte del modelo ya estamos en condiciones de realizar el ajuste que intente explicar de la mejor forma posible el comportamiento del ingreso medio transformado (variable respuesta) en función de las variables explicativas que hemos seleccionado. Entre estas últimas destaca principalmente la de ingresos fijos por ser la variable que más información aporta acerca del nivel de ingresos mensuales de los clientes.

El modelo seleccionado para realizar el ajuste consiste en un modelo de regresión múltiple en el que las variables explicativas se introducen en su mayoría en forma lineal, aunque en el caso de la variable "edad" se decide ajustarla de forma no paramétrica a través de bases B-Splines con el objetivo de obtener una mayor flexibilidad en su estimación.

De esta forma, R nos proporciona la siguiente salida con los coeficientes y los errores de las variables que forman el modelo:

```
> summary(m1)
Call:
lm(formula = z3$INGRESO_MEDIO2 ~ z3$SEXO + bs(z3$EDAD2, 8) +
    z3$CLUSTER_AUT + z3$SITUACION_LABORAL_ID + z3$CLUSTER_SEC_ACT +
    z3$INGRESOS_FIJOS_MES3)
Residuals:
    Min      1Q  Median      3Q     Max
-33.908  -0.402   0.114   0.699  26.614
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.803160   0.212378  13.199  < 2e-16 ***
z3$SEXOMujer                   -0.269587   0.016077 -16.768  < 2e-16 ***
bs(z3$EDAD2, 8)1                1.245077   0.271939   4.579 4.69e-06 ***
bs(z3$EDAD2, 8)2                1.134068   0.163567   6.933 4.15e-12 ***
bs(z3$EDAD2, 8)3                1.310818   0.198822   6.593 4.34e-11 ***
bs(z3$EDAD2, 8)4                1.166141   0.182144   6.402 1.54e-10 ***
bs(z3$EDAD2, 8)5                1.463981   0.190773   7.674 1.69e-14 ***
bs(z3$EDAD2, 8)6                0.897029   0.199561   4.495 6.97e-06 ***
bs(z3$EDAD2, 8)7                0.903311   0.262918   3.436 0.000591 ***
bs(z3$EDAD2, 8)8                1.201984   0.453896   2.648 0.008095 **
z3$CLUSTER_AUTCluster 2         0.014465   0.086836   0.167 0.867704
z3$CLUSTER_AUTCluster 3         0.199383   0.096467   2.067 0.038751 *
z3$CLUSTER_AUTCluster 4         0.100448   0.115415   0.870 0.384127
z3$CLUSTER_AUTCluster 5         0.079828   0.102000   0.783 0.433846
z3$CLUSTER_AUTCluster 6        -1.866241   0.248958  -7.496 6.65e-14 ***
z3$SITUACION_LABORAL_IDOtros    0.210500   0.045173   4.660 3.17e-06 ***
z3$SITUACION_LABORAL_IDTemporal -0.236130  0.025397  -9.297  < 2e-16 ***
z3$SITUACION_LABORAL_IDTemporero -0.477641 0.150196  -3.180 0.001473 **
z3$CLUSTER_SEC_ACTCluster 2     0.703523   0.057788  12.174  < 2e-16 ***
z3$CLUSTER_SEC_ACTCluster 3     0.131669   0.031361   4.199 2.69e-05 ***
z3$CLUSTER_SEC_ACTCluster 4    -0.227754   0.031006  -7.345 2.08e-13 ***
z3$CLUSTER_SEC_ACTCluster 5    -0.763887   0.048157 -15.862  < 2e-16 ***
z3$CLUSTER_SEC_ACTCluster 6    -0.499119   0.033555 -14.875  < 2e-16 ***
z3$INGRESOS_FIJOS_MES3          0.798950   0.003087 258.843  < 2e-16 ***

Residual standard error: 1.91 on 66873 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6364, Adjusted R-squared:  0.6363
F-statistic:  5089 on 23 and 66873 DF,  p-value: < 2.2e-16
```

Como podemos ver en la salida que nos da R del modelo la mayoría de los clusters de la variable AUTONOMÍA resultan no significativos, lo que se debe a que la diferencia entre estos clusters no es demasiado grande en relación a las diferencias que se dan con otras variables (se decide mantener dicha

variable ya que de esta forma obtenemos un mejor ajuste). Por otro lado, también podemos ver que el coeficiente de determinación ajustado (R^2) del modelo es de 0.6363, lo que quiere decir que dicho modelo explica cerca del 64% de la variabilidad de los ingresos.

Diagnosis del modelo

Una vez ajustado el modelo de regresión, el último paso consistirá en comprobar si este verifica las hipótesis estructurales básicas de normalidad, homocedasticidad e independencia de los residuos.

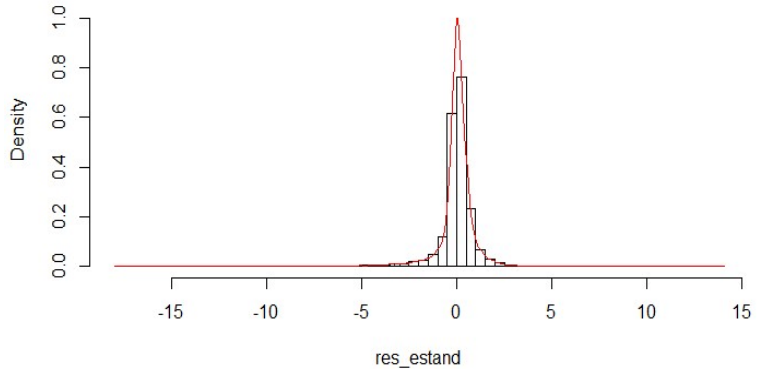


Figura 3.2.17: Histograma y estimación tipo núcleo de la función de densidad de los residuos estandarizados.

Vemos en la figura 3.2.17 que la media parece estar en torno al 0, aunque también podemos observar un pico en la zona central que nos hace pensar en la falta de normalidad de los residuos.

En la figura 3.2.18 se aprecia algo de heterocedasticidad, es decir, parece que conforme aumenta el valor ajustado de la variable respuesta va aumentando la variabilidad de la misma.

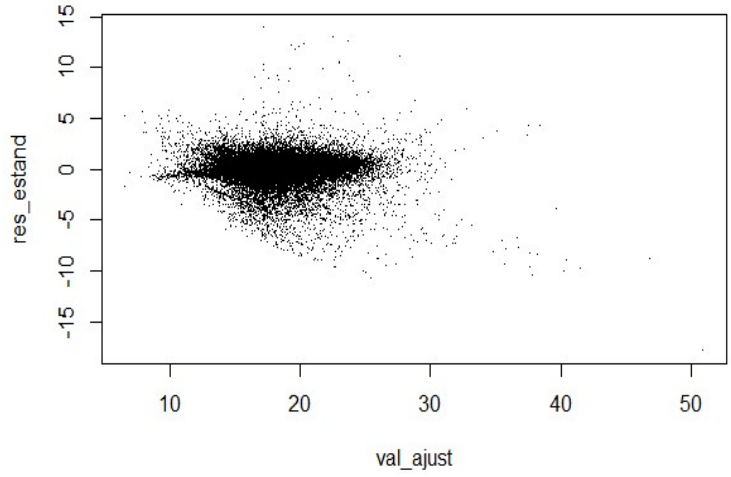


Figura 3.2.18: Diagrama de dispersión de los residuos estandarizados en función de los valores ajustados.

Para comprobar la normalidad de los residuos estandarizados realizamos un Q-Q plot como el de la figura 3.2.19.

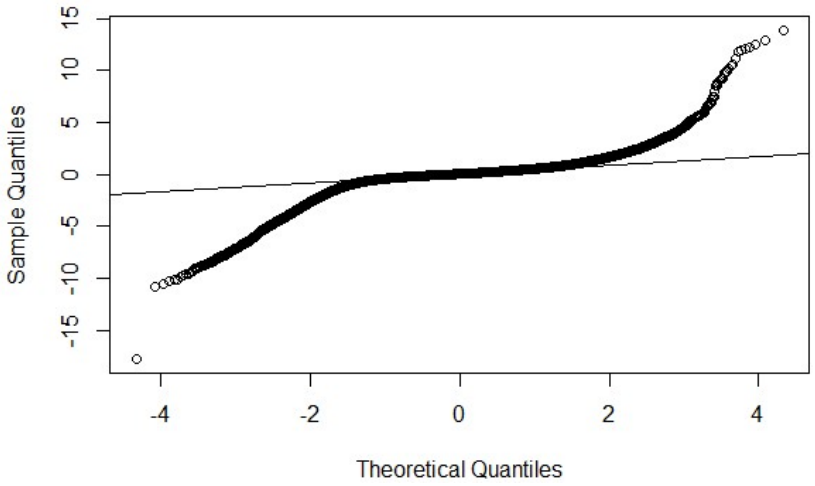


Figura 3.2.19: Q-Q plot de los residuos estandarizados

Al igual que habíamos observado anteriormente, en la figura 3.2.19 también llegamos a la conclusión de que los residuos no siguen una distribución normal ya que los puntos en las colas se encuentran muy distanciados de la diagonal que marca la normalidad.

Además, si llevamos a cabo diversos test de normalidad obtenemos los siguientes resultados:

- **Test de Lilliefors:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Anderson Darling:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Cramer von Mises**: Obtenemos un p-valor de $7.37x10^{-10}$ que nos lleva a rechazar la hipótesis de normalidad.

Por otro lado, también llevamos a cabo el test de Durbin Watson que nos proporciona un estadístico de contraste de 1.9831, muy cercano a 2, por lo que llegamos a la conclusión de que no existe correlación lineal entre los residuos del ajuste.

3.3. Segmento 3: KYC

Para el ajuste del modelo de este segmento utilizaremos la información de todos aquellos clientes de los que disponemos tanto de los datos de ingresos como los datos de KYC. Como veremos más adelante, al introducir las variables explicativas propias de ingresos que tenemos en la BD de KYC, nos encontraremos con que hay casos en los que estos datos de ingresos presentan valores nulos que empeoran drásticamente el ajuste. Por este motivo será recomendable ajustar el modelo con aquellas observaciones en las que los datos de ingresos fijos mensuales sean no nulos (como para la gran mayoría de clientes de este segmento, la información de ingresos de la unidad familiar es nula no introduciremos esta variable en el modelo). De esta forma pasamos de tener 209.872 observaciones a tener 123.805 observaciones, que se quedarían en 122.904 al prescindir de los autónomos.

Las variables que van a formar parte del modelo son las siguientes:

- **INGRESO MEDIO:** Variable respuesta que nos dice el ingreso medio mensual de cada cliente y cuyo comportamiento intentaremos explicar a través del resto de variables (explicativas).
- **SEXO:** Variable cualitativa con dos niveles (hombre y mujer).

- **EDAD:** Variable cuantitativa y continua que nos indica la edad de cada cliente (en años).
- **AUTONOMÍA:** Variable cualitativa que nos muestra la comunidad autónoma a la que pertenece cada cliente.
- **SITUACIÓN LABORAL:** Variable cualitativa que nos indica la situación en la que se encuentra cada cliente dentro del mercado laboral. No tendremos en cuenta a los autónomos ya que sus ingresos son más difíciles de estimar ya que no se registran con concepto de nómina o pensión por lo que para la mayoría de ellos no dispondremos de variable respuesta con información fiable. Además los datos de la captura de bienes no se registran como ingresos fijos sino cómo ingresos variables
- **SECTOR ACTIVIDAD:** Variable cualitativa que nos ofrece información acerca del sector al que pertenece el trabajo que desempeña cada cliente.
- **INGRESOS FIJOS NETOS:** Variable cuantitativa y continua que nos indica la cantidad de ingresos fijos que percibe cada cliente.

Como también hicimos en el segmento 1, en este caso prescindiremos del resto de variables presentes en la BD de KYC ya sea porque aportan información muy similar a la que aportan algunas de las variables presentes en el modelo (como es el caso de la variable PROFESIÓN) o porque la información que aportan es muy poco significativa (como es el caso de las variables UNIDAD FAMILIAR, ESTADO CIVIL o RÉGIMEN MATRIMONIAL).

Para hacernos una idea de la importancia que puede tener cada una de las variables en el ajuste del modelo será conveniente hacer una pequeña introducción individual de cada una de ellas en la que haremos un resumen de la información que nos puede aportar en relación al nivel de ingresos de cada cliente.

• INGRESO MEDIO (Variable Respuesta)

Como ya hemos visto, esta variable la obtenemos como una media mensual de los ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017.

Si hacemos un resumen de dicha variable para los clientes que pertenecen al segmento de KYC obtenemos:

```
> summary(z3.1$INGRESO_MEDIO2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
  3.213     18.687     20.930     21.058     23.515     66.090
  (10€)     (734€)   (1.040€)   (1.060€)   (1.496€)   (46.170€)
```

Podemos ver que el rango total de la variable es muy grande (va desde los 10 euros hasta los más de 46.000 euros), aunque gran parte de sus valores se centran en torno a la media (que es de 1.060 euros). En la figura 3.3.1 podemos ver la función de densidad estimada de la transformación Box-Cox de la variable donde apreciamos mejor cómo es su distribución.

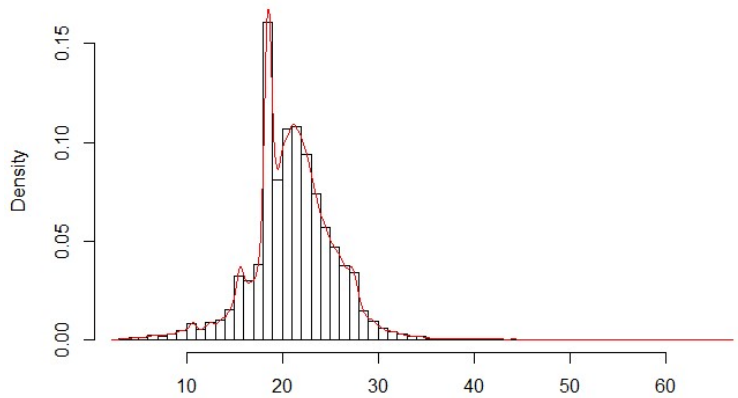


Figura 3.3.1: Histograma y estimación tipo núcleo del Ingreso Medio transformado.

- **SEXO**

Si hacemos un resumen de la variable podemos ver que:

```
> summary(z3.1$SEXO)
   Hombre    Mujer    NA's
    62913    59967      24
```

El 51,20 % de los clientes de la BD son hombres, mientras que un 48,80 % restante son mujeres (cabe destacar que hay 24 casos en los que no tenemos información del sexo del cliente).

Si relacionamos la variable SEXO con los ingresos podemos representar los resultados en el siguiente diagrama de cajas:

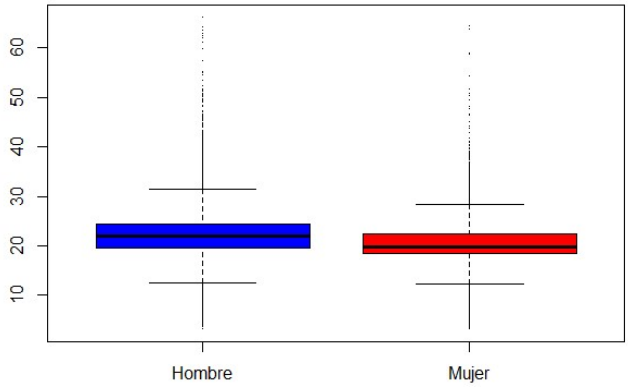


Figura 3.3.2: Diagrama de cajas del ingreso medio transformado diferenciando por sexo.

En la figura 3.3.2 destaca la gran variabilidad que tienen los ingresos tanto en el caso de los hombres como en el de las mujeres. Lo que también podemos observar es que el nivel medio de ingresos es ligeramente superior en el caso de los hombres (con un ingreso medio de 1.200,40 €) respecto de las mujeres (con una media de 924,65 €).

Si comparamos estos resultados con los que obtuvimos para el segmento 1 podemos observar que la proporción de hombres y mujeres es bastante similar, aunque el nivel medio de ingresos para ambos grupos de clientes es unos 200 – 300 € más bajo en este segmento que en el anterior.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SEXO respecto de los ingresos es de 0,04313.

- **EDAD**

Si hacemos un resumen de la variable tenemos:

```
> summary(z3.1$EDAD2)
   Min.     1st Qu.  Median    Mean   3rd Qu.      Max.
  18.42      41.23    57.02   57.57     73.08    116.95
```

Vemos que la edad de los clientes va desde los 18 años hasta casi los 117, siendo la media de edad de unos 57 años (cinco años más que en el segmento anterior).

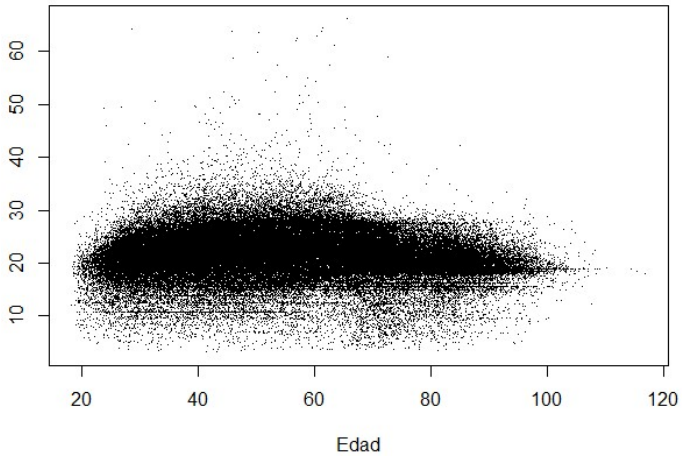


Figura 3.3.3: Diagrama de dispersión del ingreso medio transformado en función de la edad.

A simple vista, en la figura 3.3.3 no se aprecia una relación lineal entre la edad y los ingresos, por este motivo, y tal y como ya hemos hecho para el segmento 1, llevaremos a cabo una estimación polinómica o, incluso, una estimación no paramétrica de la relación entre ambas variables para ver de qué forma obtenemos un mejor ajuste.

En primer lugar llevaremos a cabo una estimación polinómica de dicha relación. Para ello seleccionaremos una muestra aleatoria de tamaño 10.000 de la variable edad para poder apreciar mejor en un gráfico la estimación polinómica de grado 2 y 3 de la regresión del ingreso medio transformado respecto de la edad, junto con sus respectivos coeficientes de correlación ajustados.

Los gráficos que obtenemos para distintas muestras son los contenidos en la figura 3.3.4.

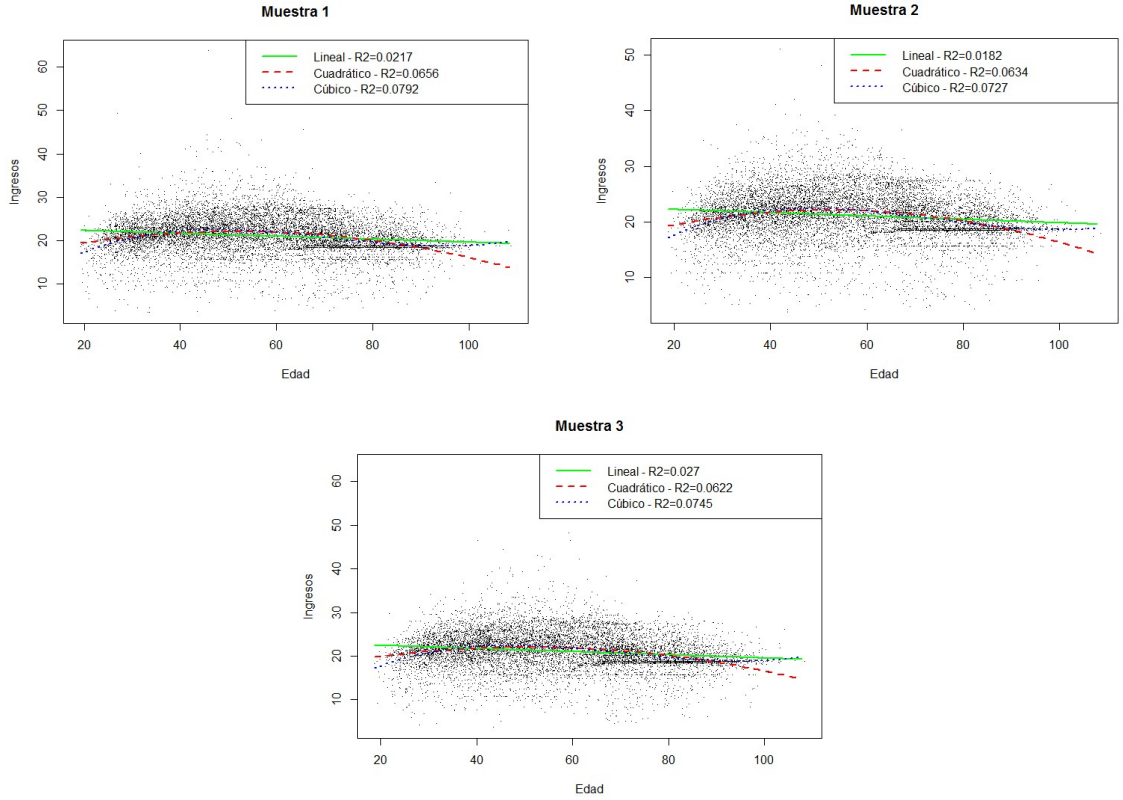


Figura 3.3.4: Representación de los ajustes polinómicos del ingreso medio en función de la edad para tres muestras distintas.

Podemos ver que el coeficiente de determinación, a pesar de ser todavía muy bajo, crece significativamente cuando introducimos la estimación polinómica (tanto cuadrática como cúbica).

Para dotar de una mayor flexibilidad a la regresión podemos llevar a cabo una estimación polinómica local de la regresión mediante bases B-Splines.

Los gráficos que obtenemos para distintas muestras de esta forma son los contenidos en la figura 3.3.5.

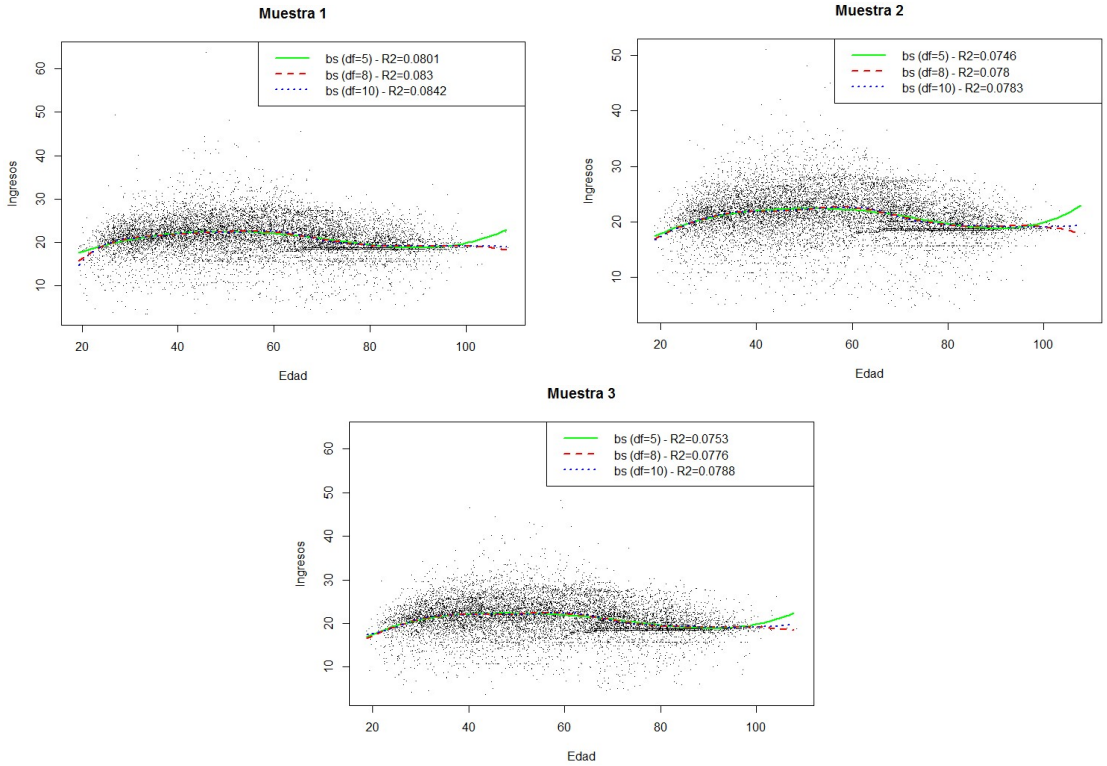


Figura 3.3.5: Representación del ajuste polinómico local del ingreso medio transformado en función de la edad mediante B-Splines con diferentes grados de libertad para tres muestras distintas.

Podemos ver que el ajuste polinómico local se parece bastante al ajuste cúbico que vimos anteriormente. También observamos que el R^2 ajustado es más alto al ajustar bases con 10 grados de libertad en todas las muestras.

Tras este análisis llegamos a la conclusión de que el mejor ajuste para la relación entre la variable EDAD y el ingreso mensual medio es el que obtenemos a partir de la estimación no paramétrica mediante B-Splines con 10 grados de libertad, por lo que será la que utilizaremos en el modelo. En la figura 3.3.6 podemos observar la representación gráfica de dicho ajuste.

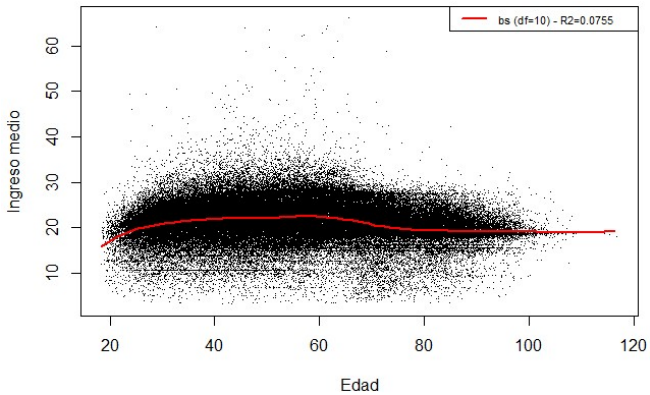


Figura 3.3.6: Diagrama de dispersión del ingreso medio transformado en función de la edad con su correspondiente ajuste polinómico local.

- **AUTONOMÍA**

Como la variable AUTONOMÍA es cualitativa con un número demasiado alto de niveles como para modelizarla directamente, es necesario realizar un análisis clúster que nos permita agruparlos en una cantidad más manejable de grupos. Para ello nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada una de las comunidades autónomas. En la figura 3.3.7 podemos observar el dendograma que nos proporciona los distintos conjuntos.

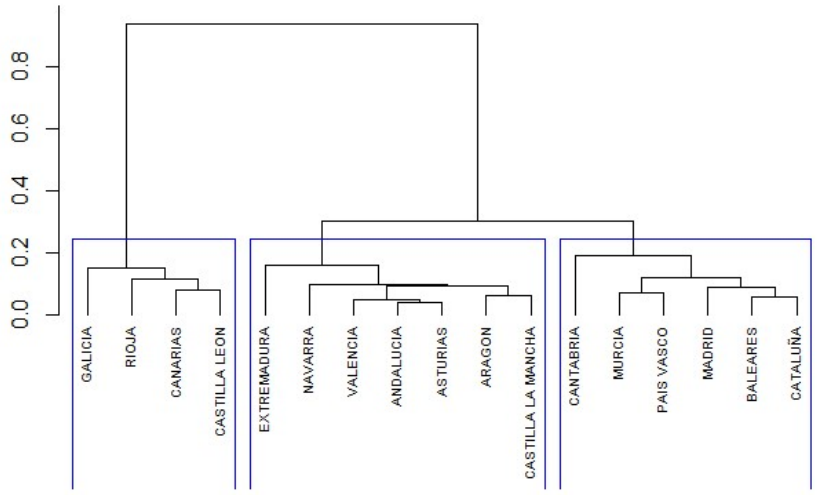


Figura 3.3.7: Dendograma que divide la variable AUTONOMÍA en 3 clústeres.

De esta forma obtenemos los siguientes grupos:

- **Clúster 1:** Canarias, Castilla León, Galicia y La Rioja (con un total de 116.216 obs).
- **Clúster 2:** Andalucía, Aragón, Asturias, Castilla La Mancha, Extremadura, Navarra y Valencia (con un total de 2.465 obs).
- **Clúster 3:** Baleares, Cantabria, Cataluña, Madrid, Murcia y el País Vasco (formado por un total de 3.698 obs).
- **Clúster 4:** En este conjunto incluimos a Ceuta y Melilla (por tener datos insuficientes para poder analizarlas convenientemente) y a Otros (clientes extranjeros) aunque, en un principio, no nos centraremos en su análisis (formado por un total de 524 obs).

También cabe destacar que hay 1 cliente para el que no tenemos información de la Comunidad Autónoma en la que reside.

En la figura 3.3.8 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable AUTONOMÍA a través de su correspondiente diagrama de cajas.

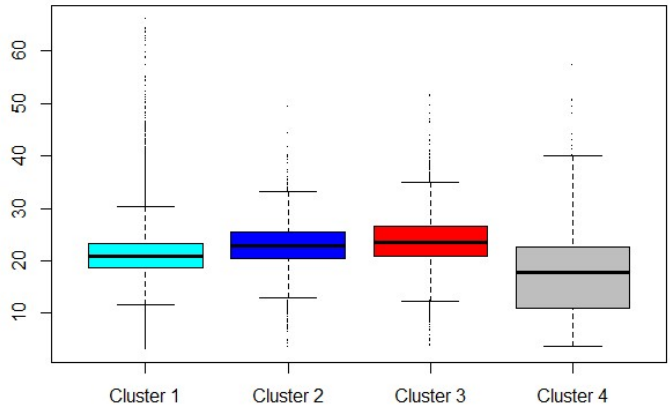


Figura 3.3.8: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

El nivel medio de ingresos en cada clúster es el siguiente:

- Para el **clúster 1** es de 1.042,84 €.
- Para el **clúster 2** es de 1.361,39 €.
- Para el **clúster 3** es de 1.511,08 €.
- Para el **clúster 4** es de 658,50 €.

Este último dato no es muy preciso ya que en este grupo se engloban clientes con características muy diferentes.

La distribución de los ingresos en cada uno de los distintos clústeres se puede apreciar mejor en la figura 3.3.9.

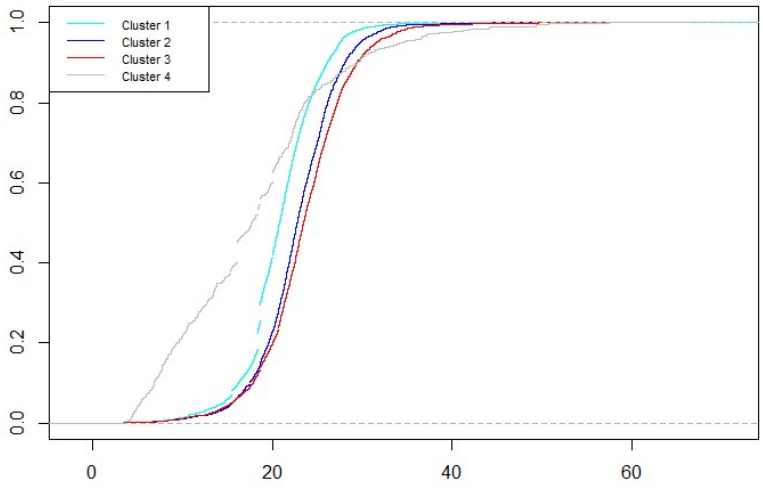


Figura 3.3.9: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable AUTONOMÍA respecto de los ingresos es de 0,01682.

• SITUACIÓN LABORAL

La variable SITUACION LABORAL es cualitativa y consta de 4 niveles diferentes. Si hacemos un pequeño resumen de la misma obtenemos:

```
> summary(z3.1$SITUACION_LABORAL_ID)
     Fijo     Otros    Temporal    Temporero
    47685     57383       17474         362
```

El grupo más numeroso es el de aquellos clientes que pertenecen al grupo "Otros" formado, principalmente, por parados y pensionistas (un 46,70 % del total), seguido del grupo de clientes con contrato fijo (un 38,80 % del total). En un término medio se encuentra el grupo de clientes con contrato temporal (un 14,22 % del total). Finalmente, el grupo con un número de clientes más bajo es el de los temporeros (0,29 % del total).

En la figura 3.3.10 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SITUACIÓN LABORAL a través de su correspondiente diagrama de cajas.

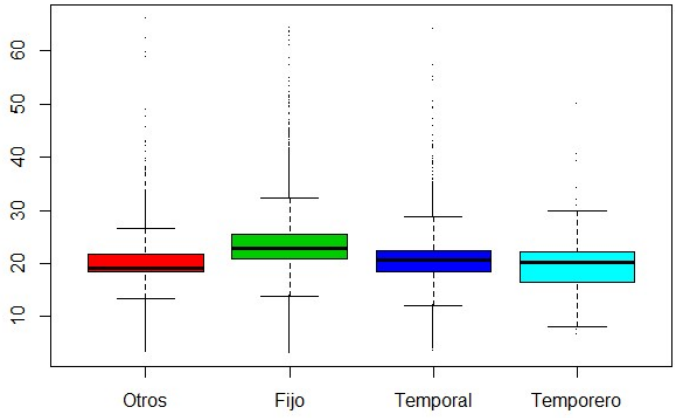


Figura 3.3.10: Diagrama de cajas del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

El nivel medio de ingresos en cada grupo de clientes es el siguiente:

- Para el grupo "**Otros**" (parados y pensionistas) es de 873,41 €.
- Para el grupo con contrato **fijo** es de 1.378,30 €.
- Para el grupo con contrato **temporal** es de 937,44 €.
- Para los **temporeros** es de 828,18 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SITUACIÓN LABORAL se puede apreciar mejor en la figura 3.3.11.

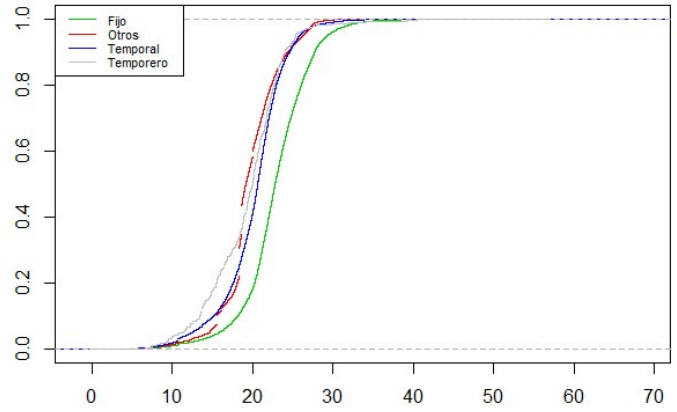


Figura 3.3.11: Comparación de la función de distribución del ingreso medio transformado diferenciando por SITUACIÓN LABORAL.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable AUTONOMÍA respecto de los ingresos es de 0,1219.

• SECTOR ACTIVIDAD

Al igual que ocurre con la variable AUTONOMÍA, la variable SECTOR ACTIVIDAD también es una variable cualitativa con un número demasiado alto de niveles como para modelizarla directamente, por este motivo llevaremos a cabo un análisis clúster que nos permita quedarnos con un número más pequeño de grupos de sectores que sean lo más heterogéneos posible entre sí y en los que los sectores que forman cada grupo sean lo más semejantes posible en términos de ingresos. Para ello, y como ya hicimos anteriormente, nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada uno de los sectores de actividad. En la figura 3.3.12 podemos observar el dendograma que nos proporciona los distintos conjuntos.

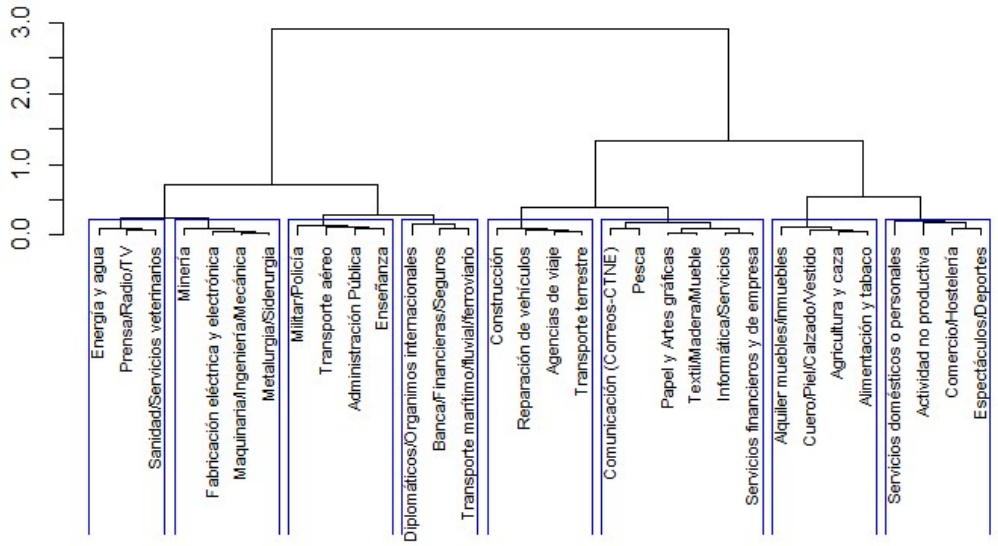


Figura 3.3.12: Dendograma que divide la variable SECTOR ACTIVIDAD en 8 clústeres.

De esta forma obtenemos los siguientes grupos:

- **Clúster 1:**
  - Energía y agua
  - Prensa/Radio/TV
  - Sanidad/Servicios Veterinarios

  7.222 obs.
- **Clúster 2:**
  - Minería
  - Maquinaria/Ingeniería/Mecánica
  - Metalurgia/Siderurgia
  - Maquinaria/Ingeniería/Mecánica

  7.876 obs.
- **Clúster 3:**
  - Militar/Policía
  - Transporte aéreo
  - Administración Pública
  - Enseñanza

  11.962 obs.
- **Clúster 4:**
  - Diplomáticos/Organismos internacionales
  - Banca/Financieras/Seguros
  - Transporte marítimo/fluvial/ferroviario

  1.638 obs.
- **Clúster 5:**
  - Construcción
  - Reparación de vehículos
  - Agencias de viaje
  - Transporte terrestre

  6.765 obs.
- **Clúster 6:**
  - Comunicación (Correos-CTNE)
  - Pesca
  - Papel y artes gráficas
  - Textil/Madera/Mueble
  - Informática/Servicios
  - Servicios financieros y de empresa

  12.054 obs.
- **Clúster 7:**
  - Alquiler muebles/Inmuebles
  - Cuero/Piel/Calzado/Vestido
  - Agricultura y caza
  - Alimentación y Tabaco

  6.161 obs.
- **Clúster 8:**
  - Servicios domésticos o personales
  - Actividad no productiva
  - Comercio/Hostelería
  - Espectáculos/Deportes

  69.134 obs.

En la figura 3.3.13 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SECTOR ACTIVIDAD a través de su correspondiente diagrama de cajas.

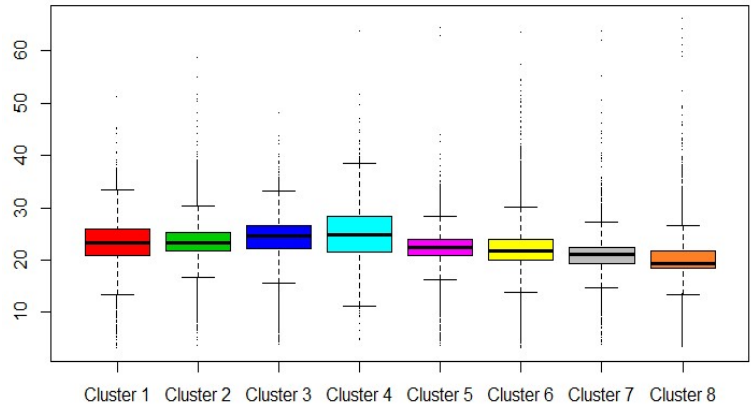


Figura 3.3.13: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

El nivel medio de ingresos en cada clúster es el siguiente:

- Para el **clúster 1** es de 1.463,20 €.
- Para el **clúster 2** es de 1.495,00 €.
- Para el **clúster 3** es de 1.578,75€.
- Para el **clúster 4** es de 1.801,82 €.
- Para el **clúster 5** es de 1.240,15 €.
- Para el **clúster 6** es de 1.214,12 €.
- Para el **clúster 7** es de 1035,15 €.
- Para el **clúster 8** es de 862,61 €.

La distribución de los ingresos en cada uno de los distintos clústeres en los que se divide la variable SECTOR ACTIVIDAD se puede apreciar mejor en la figura 3.3.14.

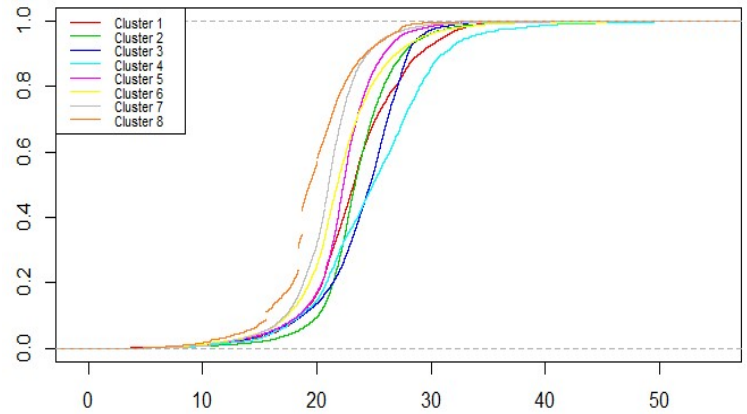


Figura 3.3.14: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de SECTOR ACTIVIDAD.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SECTOR ACTIVIDAD respecto de los ingresos es de 0,1599.

- **INGRESOS FIJOS**

Finalmente, y tras haber estudiado todas las variables sociodemográficas y socio-laborales que forman parte del modelo, acabaremos introduciendo como una variable explicativa más los ingresos fijos mensuales que podemos encontrar en la BD de KYC.

A dicha variable también se le aplicará una transformación Box-Cox (con un λ de 0.269) para reducir su fuerte asimetría, al igual que hemos hecho con la variable ingresos que estamos utilizando como variable respuesta en el ajuste del modelo.

Si hacemos un resumen de la variable tenemos:

```
> summary(z3.1$INGRESOS_FIJOS_MES2)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  2.566    6.559     6.908     6.925     7.244    12.429
  (13€)   (706€)   (1.000€)  (1.017€)  (1.400€)  (249.946€)
```

Podemos ver que el rango total de ingresos es muy grande (va desde los 13 euros hasta los casi 250.000 euros), aunque gran parte de sus valores se centran en torno a la media (que es de 1.017 euros). Representando la función de densidad estimada de la variable apreciamos mejor este hecho.

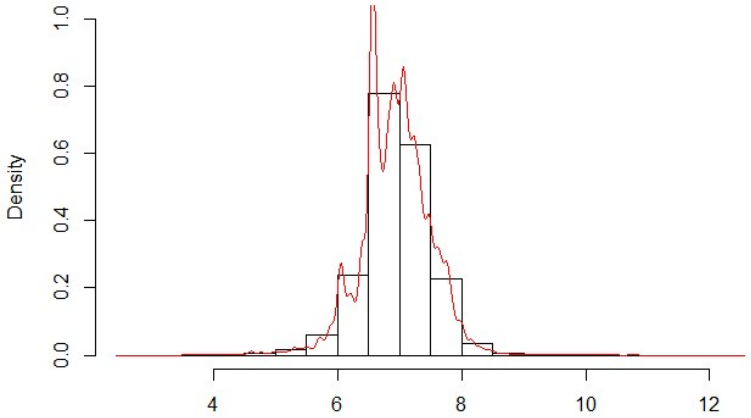


Figura 3.3.15: Histograma y estimación tipo núcleo de la función de densidad de los ingresos fijos transformados.

En la figura 3.3.15 también podemos ver algunos picos que se corresponden con ingresos que se repiten en varios clientes (algo que ya sucedía con la variable respuesta).

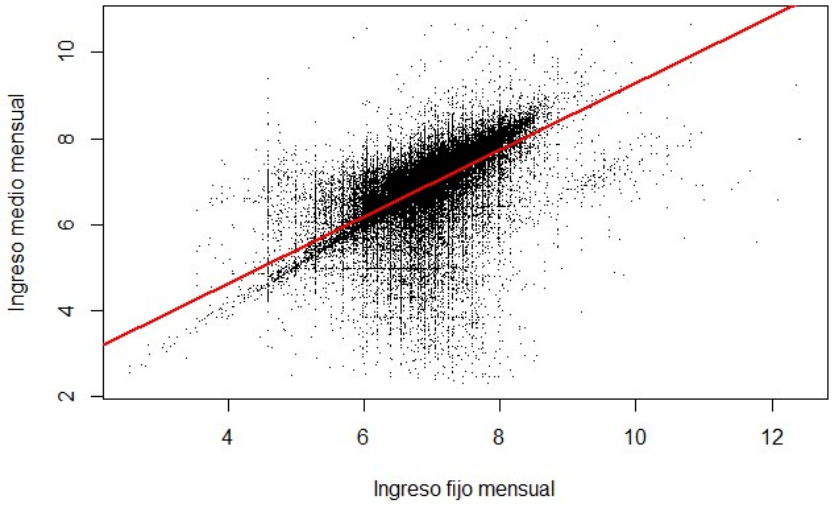


Figura 3.3.16: Diagrama de dispersión del ingreso medio transformado en función del ingreso fijo transformado con su correspondiente ajuste lineal.

En la figura 3.3.16 se aprecia una clara relación lineal entre ambas variables y se representa en rojo la línea recta que mejor se ajusta a los datos. Podemos ver que la línea no se ajusta exactamente a la diagonal, lo que se debe al efecto palanca que los datos más altos de ingresos fijos provocan sobre la recta de regresión. Calculando el modelo de regresión lineal simple que explique los ingresos en función de los ingresos fijos obtenemos un R^2 que llega hasta el 0.4229.

Modelo de regresión

Una vez analizadas individualmente cada una de las variables que forman parte del modelo ya estamos en condiciones de realizar el ajuste que intente explicar de la mejor forma posible el comportamiento del ingreso medio transformado (variable respuesta) en función de las variables explicativas que hemos seleccionado. Entre estas últimas destaca principalmente la de ingresos fijos por ser la variable que más información aporta acerca del nivel de ingresos mensuales de los clientes.

El modelo seleccionado para realizar el ajuste consiste en un modelo de regresión múltiple en el que las variables explicativas se introducen en su mayoría en forma lineal, aunque en el caso de la variable "edad" se decide ajustarla de forma no paramétrica a través de bases B-Splines con el objetivo de obtener una mayor flexibilidad en su estimación.

De esta forma, R nos proporciona la siguiente salida con los coeficientes y los errores de las variables que forman el modelo:

```
> summary(m)
Call:
lm(formula = z3.1$INGRESO_MEDIO2 ~ z3.1$SEXO + bs(z3.1$EDAD2, 10)
    + z3.1$CLUSTER_AUT + z3.1$SITUACION_LABORAL_ID + z3.1$CLUSTER_SEC_ACT
    + z3.1$INGRESOS_FIJOS_MES3)
Residuals:
    Min      1Q  Median      3Q     Max
-31.024  -0.709   0.101   1.198  44.887
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.4235927  0.2468816  21.968  < 2e-16 ***
z3.1$SEXOMujer                -0.4490605  0.0184811 -24.298  < 2e-16 ***
bs(z3.1$EDAD2, 10)1            2.3529360  0.3435810   6.848 7.51e-12 ***
bs(z3.1$EDAD2, 10)2            2.0301348  0.2089109   9.718  < 2e-16 ***
bs(z3.1$EDAD2, 10)3            2.2549289  0.2537472   8.887  < 2e-16 ***
bs(z3.1$EDAD2, 10)4            2.2834299  0.2326169   9.816  < 2e-16 ***
bs(z3.1$EDAD2, 10)5            2.4503267  0.2453861   9.986  < 2e-16 ***
bs(z3.1$EDAD2, 10)6            2.6032989  0.2389266  10.896  < 2e-16 ***
bs(z3.1$EDAD2, 10)7            1.8523637  0.2420560   7.653 1.98e-14 ***
bs(z3.1$EDAD2, 10)8            2.1724347  0.2641992   8.223  < 2e-16 ***
bs(z3.1$EDAD2, 10)9            1.7133297  0.4042107   4.239 2.25e-05 ***
bs(z3.1$EDAD2, 10)10           2.9279248  1.0532891   2.780  0.00544 **
z3.1$CLUSTER_AUTCluster 2      0.0771648  0.0608392   1.268  0.20468
z3.1$CLUSTER_AUTCluster 3      0.2185041  0.0505683   4.321 1.55e-05 ***
z3.1$CLUSTER_AUTCluster 4     -4.0314635  0.1299176 -31.031  < 2e-16 ***
z3.1$SITUACION_LABORAL_IDFijo -0.3480461  0.0403357  -8.629  < 2e-16 ***
z3.1$SITUACION_LABORAL_IDTemporal -0.7839517  0.0452763 -17.315  < 2e-16 ***
z3.1$SITUACION_LABORAL_IDTemporero -1.0272216  0.1602955  -6.408 1.48e-10 ***
z3.1$CLUSTER_SEC_ACTCluster 2 -0.0003048  0.0493753  -0.006  0.99507
z3.1$CLUSTER_SEC_ACTCluster 3  0.0866240  0.0442812   1.956  0.05044 .
z3.1$CLUSTER_SEC_ACTCluster 4  0.7345842  0.0813179   9.033  < 2e-16 ***
z3.1$CLUSTER_SEC_ACTCluster 5 -0.7290161  0.0514695 -14.164  < 2e-16 ***
z3.1$CLUSTER_SEC_ACTCluster 6 -0.4574498  0.0445550 -10.267  < 2e-16 ***
z3.1$CLUSTER_SEC_ACTCluster 7 -0.7604178  0.0519122 -14.648  < 2e-16 ***
z3.1$CLUSTER_SEC_ACTCluster 8 -1.4261641  0.0451883 -31.560  < 2e-16 ***
z3.1$INGRESOS_FIJOS_MES3       0.7043135  0.0027072 260.165  < 2e-16 ***
```

```
Residual standard error: 2.961 on 122761 degrees of freedom
  (117 observations deleted due to missingness)
Multiple R-squared:  0.5155, Adjusted R-squared:  0.5154
F-statistic:  5225 on 25 and 122761 DF,  p-value: < 2.2e-16
```

Como podemos ver en la salida que nos da R del modelo ajustado, algunas de las variables dejan de ser significativas, como es el caso del clúster 2 de la variable AUTONOMÍA y el clúster 2 de la variable SECTOR ACTIVIDAD. Este hecho se debe a que la diferencia entre los ingresos de los clientes que pertenecen a estos grupos no es demasiado importante respecto a los grupos de referencia en relación a la diferencia explicada por el conjunto de las variables del modelo (se deciden mantener dichas variables ya que de esta forma obtenemos un mejor ajuste).

Por otro lado, también podemos ver que el coeficiente de determinación ajustado (R^2) del modelo es de 0.5154, lo que quiere decir que dicho modelo explica cerca del 52% de la variabilidad de los ingresos.

Diagnosis del modelo

Una vez ajustado el modelo de regresión, el último paso consistirá en comprobar si este verifica las hipótesis estructurales básicas de normalidad, homocedasticidad e independencia de los residuos.

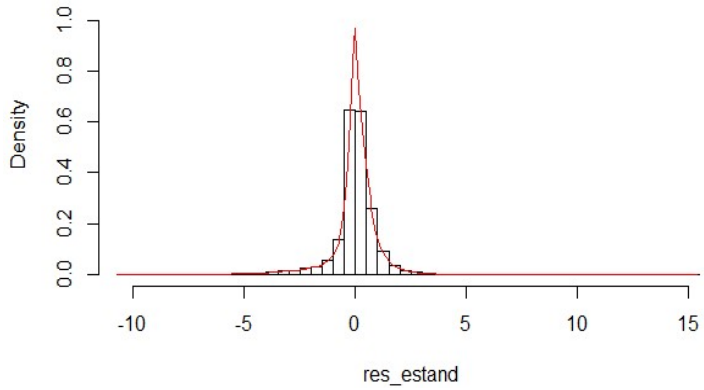


Figura 3.3.17: Histograma y estimación tipo núcleo de la función de densidad de los residuos estandarizados.

Vemos en la figura 3.3.17 que la media parece estar en torno al 0, aunque también podemos observar un apuntamiento en la zona central que nos hace pensar en la falta de normalidad de los residuos.

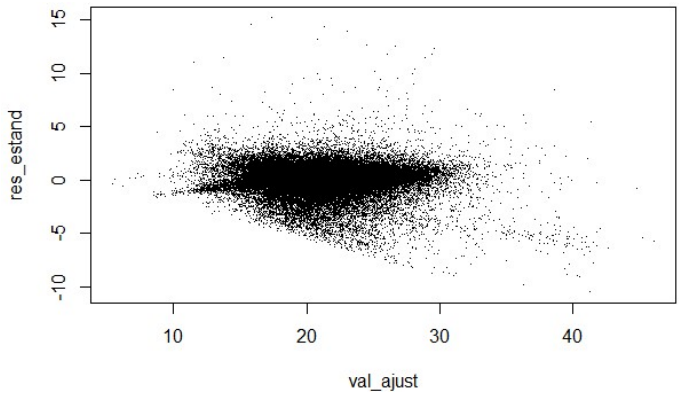


Figura 3.3.18: Diagrama de dispersión de los residuos estandarizados en función de los valores ajustados.

En la figura 3.3.18 se aprecia cierta heterocedasticidad, es decir, parece que cuanto mayor es el valor ajustado de la variable respuesta mayor es la variabilidad de los residuos.

Para comprobar la normalidad de los residuos estandarizados realizamos un Q-Q plot como el de la figura 3.3.19.

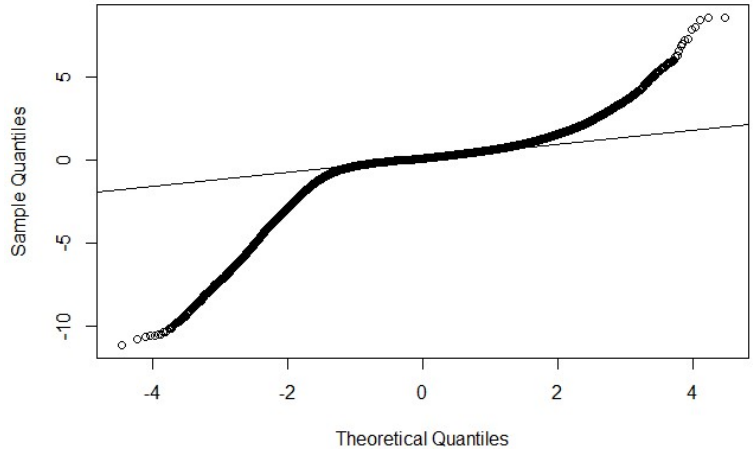


Figura 3.3.19: Q-Q plot de los residuos estandarizados

Al igual que habíamos observado en las figuras anteriores, en la figura 3.3.19 también llegamos a la conclusión de que los residuos no siguen una distribución normal ya que los puntos en las colas se encuentran muy distanciados de la diagonal que marca la normalidad.

Además, si llevamos a cabo diversos test de normalidad obtenemos los siguientes resultados:

- **Test de Lilliefors:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Anderson Darling:** Obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Cramer von Mises:** Obtenemos un p-valor de $7.37x10^{-10}$ que nos lleva a rechazar la hipótesis de normalidad.

Por otro lado, también llevamos a cabo el test de Durbin-Watson que nos proporciona un estadístico de contraste de 1.9937, muy cercano a 2, por lo que llegamos a la conclusión de que no existe correlación lineal entre los residuos del ajuste.

3.4. Segmento 4: Variables Básicas

Para ajustar el modelo de este segmento utilizaremos la información básica que nos aportan las variables que tenemos en la BD de clientes. En este segmento analizaremos a aquellos clientes que no se encuentran en ninguno de los grupos ya estudiados en los segmentos 1, 2 y 3. De esta forma trabajaremos con un total de 558.403 observaciones.

Las variables que van a formar parte del modelo son las siguientes:

- **INGRESO MEDIO**: Variable respuesta que nos dice el ingreso medio mensual de cada cliente y cuyo comportamiento intentaremos explicar a través del resto de variables (explicativas).
- **SEXO:** Variable cualitativa con dos niveles (hombre y mujer).
- **EDAD:** Variable cuantitativa y continua que nos indica la edad de cada cliente (en años).
- **AUTONOMÍA:** Variable cualitativa que nos muestra la comunidad autónoma a la que pertenece cada cliente.

- **SPID (Strategic Portfolio Identifier):** Variable cualitativa presente en el procesamiento TRIAD creado por ABANCA que nos dice el segmento con características homogéneas al que pertenece cada cliente y al que se le dará un trato diferenciado.

Prescindiremos del resto de variables presentes en la BD de clientes ya sea porque aportan información muy similar a la que aportan algunas de las variables presentes en el modelo (como es el caso de la variable PROVINCIA) o porque la información que aportan es muy poco significativa (como es el caso de las variables relacionadas con el país de residencia y la nacionalidad).

Para hacernos una idea de la importancia que puede tener cada una de las variables en el ajuste del modelo será conveniente hacer una pequeña introducción individual de cada una de ellas en la que haremos un resumen de la información que nos puede aportar en relación al nivel de ingresos de cada cliente.

• INGRESO MEDIO (Variable Respuesta)

Como ya hemos visto, esta variable la obtenemos como una media mensual de los ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017.

Si hacemos un resumen de dicha variable para los clientes que no pertenecen ni al segmento de captura de bienes ni de KYC:

```
> summary(z2$INGRESO_MEDIO2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
  3.459     22.503     26.008     26.496     30.430    106.767
  (10€)     (696€)     (914€)     (961€)   (1.400€)   (50.000€)
```

Podemos ver que el rango total de la variable es muy grande (va desde los 10 euros hasta los más de 50.000 euros), aunque gran parte de ellos se centran en torno a la media (que es de 961 euros). En la figura 3.4.1 vemos la función de densidad estimada de la transformación Box-Cox de la variable que nos sirve para apreciar mejor cómo es su distribución.

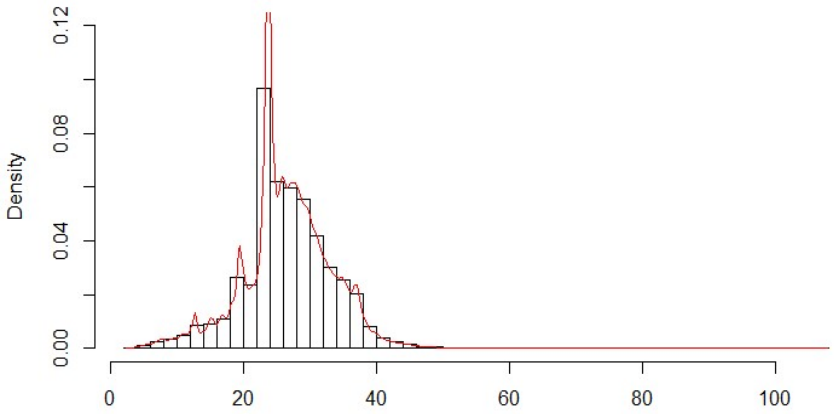


Figura 3.4.1: Histograma y estimación tipo núcleo de la función de densidad del Ingreso Medio transformado.

• SEXO

Si hacemos un resumen de la variable podemos ver que:

```
> summary(z2$SEXO)
Hombre     Mujer      NA's
257855    300484        64
```

El 46,18 % de los clientes de la BBDD son hombres, mientras que el 53,82 % restante son mujeres (cabe destacar que hay 64 casos en los que no tenemos información del sexo del cliente).

Si relacionamos la variable SEXO con los ingresos podemos representar los resultados en el siguiente diagrama de cajas:

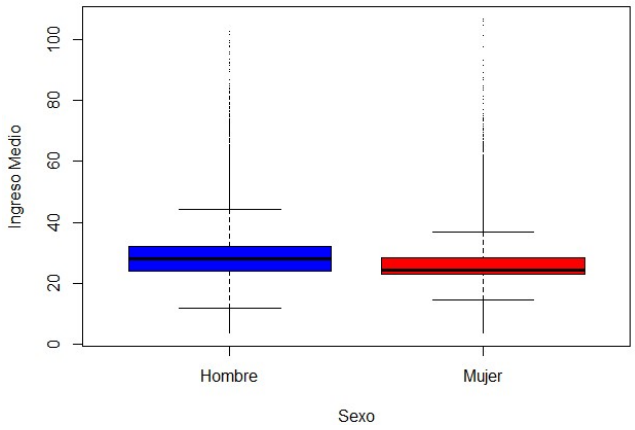


Figura 3.4.2: Diagrama de cajas del ingreso medio transformado diferenciando por sexo.

En la figura 3.4.2 destaca que, a pesar de que el rango intercuartílico no es excesivamente grande, existe una gran variabilidad en los ingresos tanto en el caso de los hombres como en el de las mujeres. Lo que también podemos observar es que el nivel medio de ingresos es ligeramente superior en el caso de los hombres (con un ingreso medio de 1.119,03 €) respecto de las mujeres (con una media de 837,58 €).

Si comparamos estos resultados con los que obtuvimos para el segmento 1 y el 2 podemos observar que la proporción de mujeres pasa a ser mayor que la de hombres, y el nivel medio de ingresos para ambos grupos de clientes es unos 300 – 350 € más bajo en este segmento que en el segmento 1 de captura de bienes, y unos 50 – 75 € más bajo que en el segmento 2 de KYC.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SEXO respecto de los ingresos es de 0,04582.

- **EDAD**

Si hacemos un resumen de la variable:

```
> summary(z2$EDAD2)
   Min.   1st Qu.   Median     Mean   3rd Qu.      Max.
  18.42     44.40    65.00    61.51    78.38    122.41
```

Vemos que la edad de los clientes va desde los 18 años hasta los 122, siendo la media de edad de unos 61 años (nueve años más que en el segmento 1 y cuatro más que en el segmento 2).

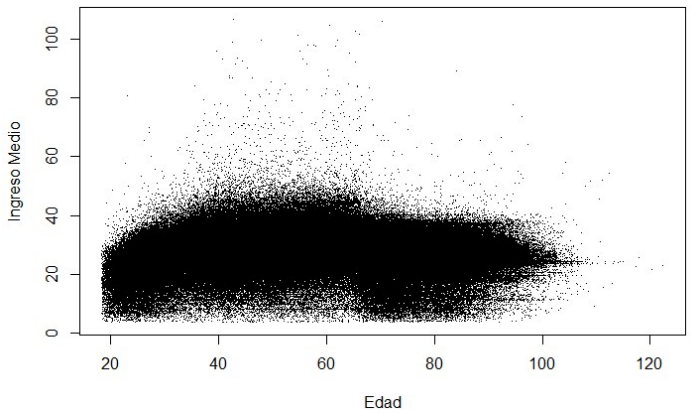


Figura 3.4.3: Diagrama de dispersión del ingreso medio transformado en función de la edad.

A simple vista, en la figura 3.4.3 no se aprecia una relación lineal entre la edad y los ingresos, por este motivo, y tal y como ya hemos hecho en los segmentos anteriores, llevaremos a cabo una estimación polinómica, o incluso una estimación no paramétrica de la relación entre ambas variables para ver de qué forma obtenemos un mejor ajuste.

En primer lugar llevaremos a cabo una estimación polinómica de dicha relación. Para ello seleccionaremos una muestra aleatoria de tamaño 10.000 de la variable edad para poder representar en un gráfico la estimación polinómica de grado 2 y 3 de la regresión del ingreso medio respecto de la edad, junto con sus respectivos coeficientes de correlación ajustados.

Los gráficos que obtenemos para distintas muestras son los contenidos en la figura 3.4.4.

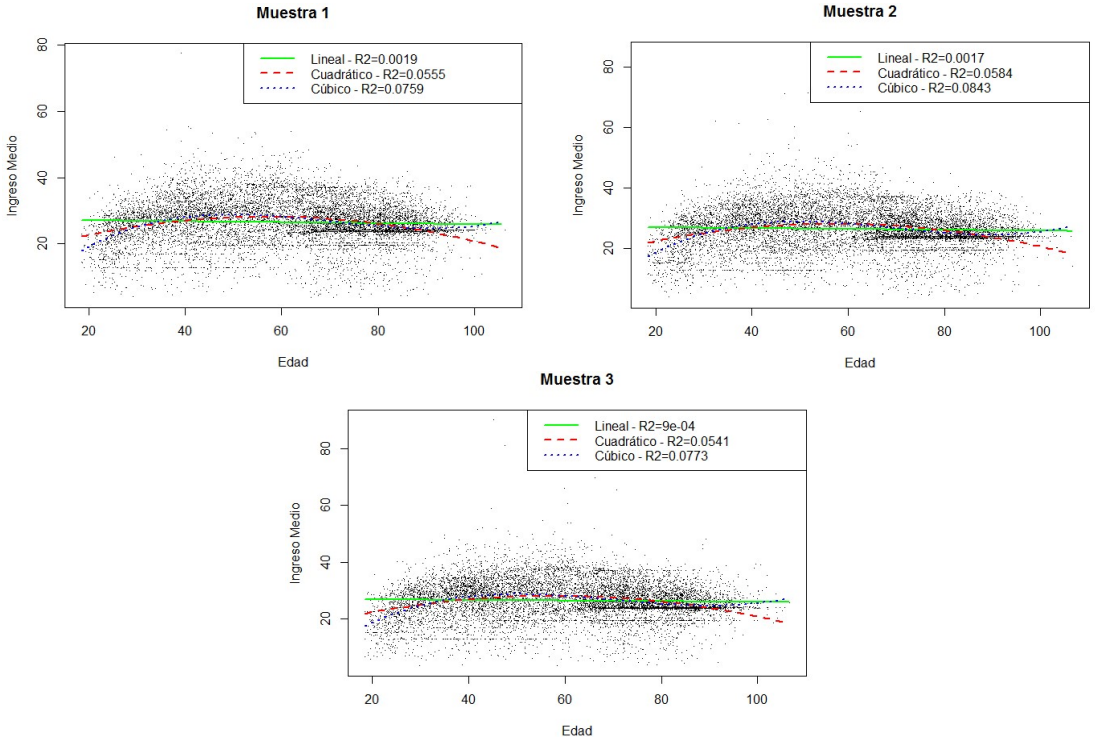
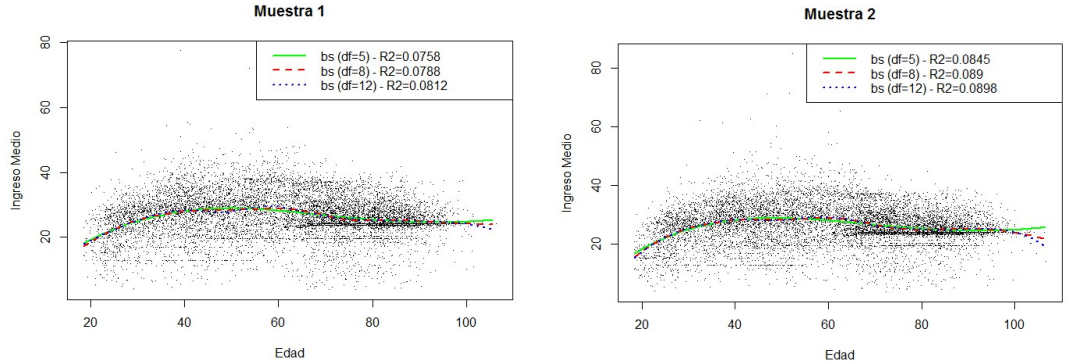


Figura 3.4.4: Representación de los ajustes polinómicos del ingreso medio transformado en función de la edad para tres muestras distintas.

Podemos ver que el coeficiente de determinación, a pesar de ser todavía muy bajo, crece significativamente cuando introducimos la estimación polinómica (tanto cuadrática como cúbica).

Para dotar de una mayor flexibilidad a la regresión podemos llevar a cabo una estimación polinómica local de la regresión mediante bases B-Splines.

Los gráficos que obtenemos para distintas muestras de esta forma son los contenidos en la figura 3.4.5



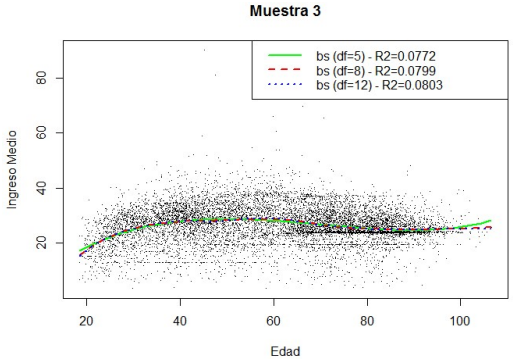


Figura 3.4.5: Representación del ajuste polinómico local del ingreso medio transformado en función de la edad mediante B-Splines con diferentes grados de libertad para tres muestras distintas.

Podemos ver que el ajuste polinómico local se parece bastante al ajuste cúbico que vimos anteriormente. También observamos que el R^2 ajustado es muy similar en todas las muestras tanto cuando ajustamos con 8 grados de libertad como cuando lo hacemos con 12.

Al realizar el ajuste con la totalidad de los datos llegamos a la conclusión de que el mejor ajuste para la relación entre la variable EDAD y el ingreso medio transformado es el que obtenemos a partir de la estimación no paramétrica mediante B-Splines con 12 grados de libertad, por lo que será la que utilizaremos en el modelo. En la figura 3.4.6 podemos observar la representación gráfica de dicho ajuste.

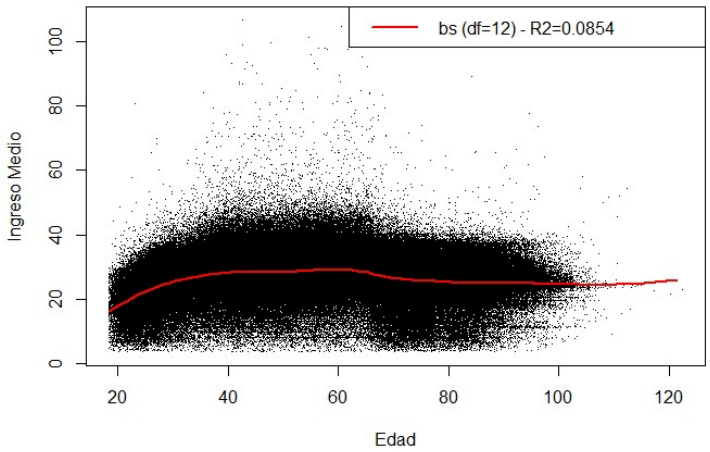


Figura 3.4.6: Diagrama de dispersión del ingreso medio transformado en función de la edad con su correspondiente ajuste polinómico local.

- **AUTONOMÍA**

Como la variable AUTONOMÍA es una variable cualitativa con un número demasiado alto de niveles como para modelizarla, es necesario realizar un análisis clúster que nos permita agruparlos en una cantidad más manejable de conjuntos. Para ello nos basaremos en la distancia de Kolmogorov-Smirnov entre las funciones de distribución de la variable ingresos de cada una de las comunidades autónomas. En la figura 3.4.7 podemos observar el dendograma que nos proporciona los distintos conjuntos.

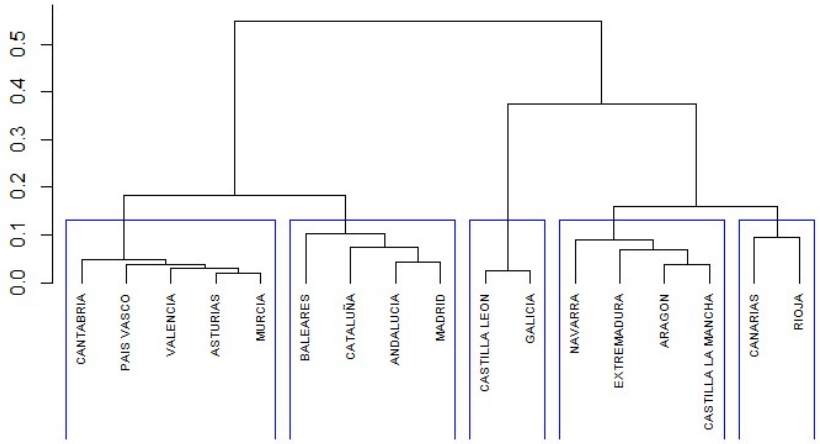


Figura 3.4.7: Dendograma que divide la variable AUTONOMÍA en 5 clústeres.

De esta forma obtenemos los siguientes conjuntos:

- **Clúster 1:** Asturias, Cantabria, Murcia, País Vasco, Valencia (formado por 9.119 obs).
- **Clúster 2:** Andalucía, Baleares, Cataluña y Madrid (formado por 13.879 obs).
- **Clúster 3:** Castilla León y Galicia (formado 529.804 obs).
- **Clúster 4:** Aragón, Castilla La Mancha, Extremadura y Navarra (formado por 2.336 obs).
- **Clúster 5:** Canarias y La Rioja (formado por 1.371 obs).
- **Clúster 6:** En este conjunto metemos a Ceuta y Melilla (por tener datos insuficientes para poder analizarlas convenientemente) y a Otros (clientes extranjeros) aunque, en un principio, no nos centraremos en su análisis (formado por un total de 1.877 obs).

También cabe destacar que hay 17 casos en los que no tenemos información de la Comunidad Autónoma en la que reside el cliente.

En la figura 3.4.8 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable AUTONOMÍA a través de su correspondiente diagrama de cajas.

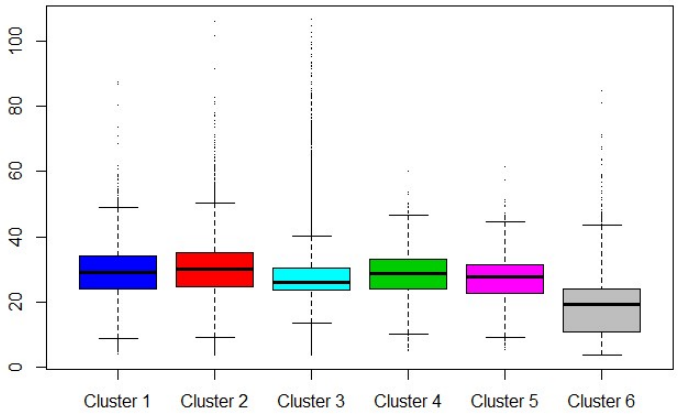


Figura 3.4.8: Diagrama de cajas del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

El nivel medio de ingresos en cada clúster será el siguiente:

- Para el **clúster 1** será de 1.189,89 €.
- Para el **clúster 2** será de 1.325,82 €.
- Para el **clúster 3** será de 950,45 €.
- Para el **clúster 4** será de 1.152,99 €.
- Para el **clúster 5** será de 1.005,96 €.

- Para el **clúster 6** será de 381,24 € (este último dato nos es muy preciso ya que en este grupo se engloban clientes con características muy diferentes).

La distribución de los ingresos en cada uno de los distintos clústeres se puede apreciar mejor en la figura 3.4.9.

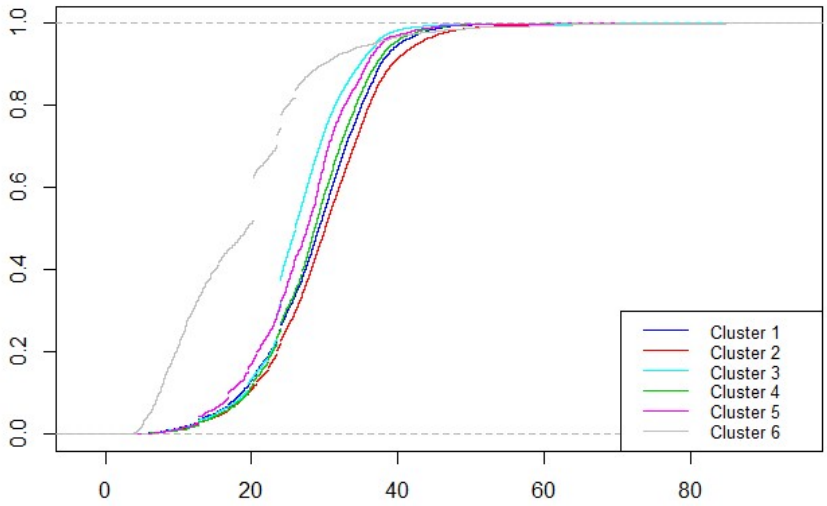


Figura 3.4.9: Comparación de la función de distribución del ingreso medio transformado diferenciando por clúster de AUTONOMÍA.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable AUTONOMÍA respecto de los ingresos es de 0,01326.

• SPID

La variable SPID (Strategic Portfolio Identifier) es una variable cualitativa creada por ABANCA para su sistema TRIAD que sirve para diferenciar a los clientes en varios grupos con características similares. Para el presente modelo solo utilizaremos clientes pertenecientes a 7 niveles de dicha variable. Si hacemos un pequeño resumen de la misma:

```
> summary(z2$SPID)
  Nómina/Pensión             Otros          Clientes sin cta. Cte.
          476834             62264                          11705
Extranjeros NO OCDE      Esp. NO resid.   Extranjero OCDE no resid.
            3786              3476                            338
```

El grupo más numeroso es el de aquellos que pertenecen al grupo de clientes que tienen su nómina o pensión en ABANCA (un 85,39 % del total), seguido del grupo de clientes perteneciente a “Otros” (un 11,15 % del total). Finalmente tendremos grupos menos numerosos como los clientes sin cuenta corriente (2,10 %), los extranjeros que no pertenecen a la OCDE (0,68 %), los españoles no residentes (0,62 %) y los extranjeros no residentes de países pertenecientes a la OCDE (0,06 %).

En la figura 3.4.10 podemos ver la distribución de los ingresos en cada uno de los clústeres en los que se divide la variable SPID a través de su correspondiente diagrama de cajas.

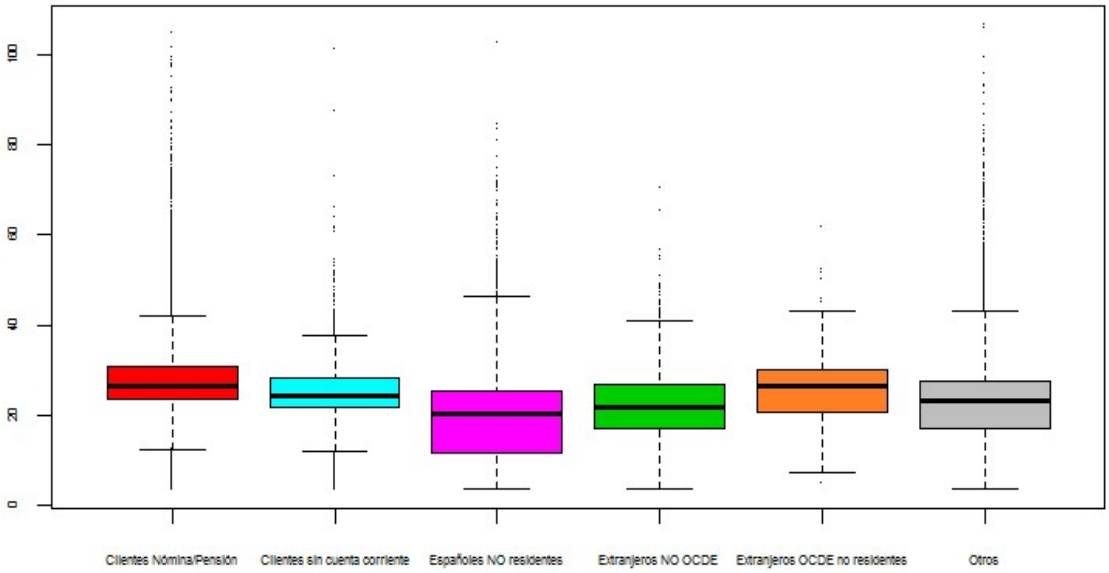


Figura 3.4.10: Diagrama de cajas del ingreso medio transformado diferenciando por SPID.

El nivel medio de ingresos en cada grupo de clientes será el siguiente:

- Para el grupo de **clientes con nómina o pensión** en ABANCA será de 1.027,32 €.
- Para el grupo de **clientes sin cuenta corriente** será de 766,41 €.
- Para los **españoles no residentes** será de 482,04 €.
- Para los **extranjeros que no pertenecen a la OCDE** será de 561,45 €.
- Para los **extranjeros no residentes y pertenecientes a la OCDE** será de 869,26 €.
- Para el grupo "**Otros**" será de 615,08 €.

La distribución de los ingresos en cada uno de los distintos clústeres se puede apreciar mejor en la figura 3.4.11.

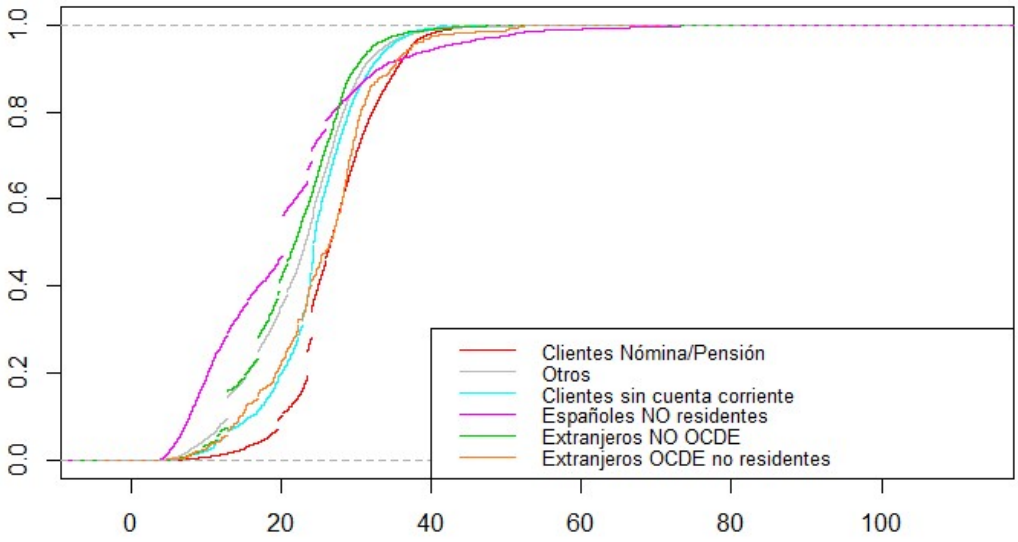


Figura 3.4.11: Comparación de la función de distribución del ingreso medio transformado diferenciando por SPID.

Si llevamos a cabo un ANOVA de una vía, el R^2 de la variable SPID respecto de los ingresos es de 0,06136.

Modelo de regresión

Una vez analizadas individualmente cada una de las variables que forman parte del modelo ya estamos en condiciones de realizar el ajuste que intente explicar de la mejor forma posible el comportamiento del ingreso medio transformado (variable respuesta) en función de las variables explicativas que hemos seleccionado.

El modelo seleccionado para realizar el ajuste consiste en un modelo de regresión múltiple en el que las variables explicativas se introducen en su mayoría en forma lineal, aunque en el caso de la variable "edad" se decide ajustarla de forma no paramétrica a través de bases B-Splines con el objetivo de obtener una mayor flexibilidad en su estimación.

De esta forma, R nos proporciona la siguiente salida con los coeficientes y los errores de las variables que forman el modelo.

```
> summary(m5)
Call:
lm(formula = z2$INGRESO_MEDIO2 ~ z2$SEXO + bs(z2$EDAD2, 12) +
    z2$CLUSTER_AUT + z2$SPID)
Residuals:
    Min      1Q  Median      3Q     Max
-28.904  -3.229  -0.250   3.485  83.826
Coefficients:
                                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)                        21.09030    0.16247  129.808  < 2e-16 ***
z2$SEXOMujer                       -2.69651    0.01619 -166.578  < 2e-16 ***
bs(z2$EDAD2, 12)1                   5.63895    0.24180   23.320  < 2e-16 ***
bs(z2$EDAD2, 12)2                   9.66783    0.13984   69.135  < 2e-16 ***
bs(z2$EDAD2, 12)3                  11.14122    0.17229   64.664  < 2e-16 ***
bs(z2$EDAD2, 12)4                  10.54083    0.15268   69.040  < 2e-16 ***
bs(z2$EDAD2, 12)5                  11.52769    0.16583   69.515  < 2e-16 ***
bs(z2$EDAD2, 12)6                  11.04597    0.15889   69.518  < 2e-16 ***
bs(z2$EDAD2, 12)7                   8.06021    0.15878   50.762  < 2e-16 ***
bs(z2$EDAD2, 12)8                   7.82644    0.15794   49.554  < 2e-16 ***
bs(z2$EDAD2, 12)9                   7.16929    0.15703   45.656  < 2e-16 ***
bs(z2$EDAD2, 12)10                  8.32822    0.21948   37.945  < 2e-16 ***
bs(z2$EDAD2, 12)11                  8.45039    0.60620   13.940  < 2e-16 ***
bs(z2$EDAD2, 12)12                 10.18060    2.24376    4.537 5.70e-06 ***
z2$CLUSTER_AUTCluster 2             1.05334    0.08069   13.054  < 2e-16 ***
z2$CLUSTER_AUTCluster 3            -1.59465    0.06355  -25.092  < 2e-16 ***
z2$CLUSTER_AUTCluster 4            -0.29464    0.13879   -2.123  0.03377 *
z2$CLUSTER_AUTCluster 5            -1.40652    0.17337   -8.113 4.96e-16 ***
z2$CLUSTER_AUTCluster 6            -5.40838    0.18406  -29.383  < 2e-16 ***
z2$SPIDClientes sin cuenta corriente -2.48167  0.05642  -43.986  < 2e-16 ***
z2$SPIDEspañoles NO residentes     -5.00450    0.12602  -39.711  < 2e-16 ***
z2$SPIDExtranjeros NO OCDE         -6.17080    0.09840  -62.711  < 2e-16 ***
z2$SPIDExtranjeros OCDE no residentes -0.98858 0.33186   -2.979  0.00289 **
z2$SPIDOtros                       -4.72621    0.02713 -174.190  < 2e-16 ***

Residual standard error: 5.985 on 558298 degrees of freedom
  (81 observations deleted due to missingness)
Multiple R-squared:  0.1863, Adjusted R-squared:  0.1862
F-statistic:  5556 on 23 and 558298 DF,  p-value: < 2.2e-16
```

Como podemos ver en la salida que nos da R del modelo ajustado, todos los coeficientes son significativos con un 95% de nivel de confianza.

Por otro lado, también podemos ver que el coeficiente de determinación ajustado (R^2) del modelo es de 0.1862, lo que quiere decir que dicho modelo explica casi un 19 % de la variabilidad de los ingresos, lo

que es bastante lógico ya que solo disponemos de variables sociodemográficas que no dan demasiada información acerca de los ingresos que puedan tener los clientes.

Diagnosis del modelo

Una vez ajustado el modelo de regresión, el último paso consistirá en comprobar si este verifica las hipótesis estructurales básicas de normalidad, homocedasticidad e independencia de los residuos.

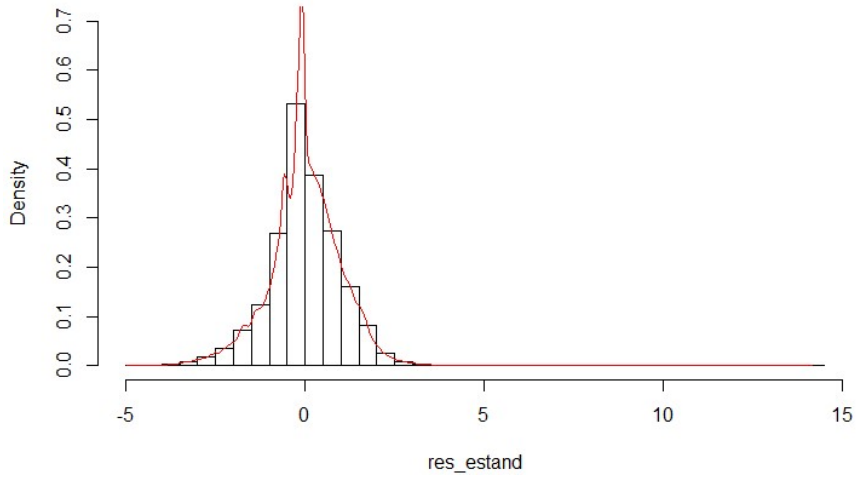


Figura 3.4.12: Histograma y estimación tipo núcleo de la función de densidad de los residuos estandarizados.

En la figura 3.4.12 vemos que la media parece estar en torno al 0, aunque también podemos observar un pico en la zona central que nos hace pensar en la falta de normalidad de los residuos.

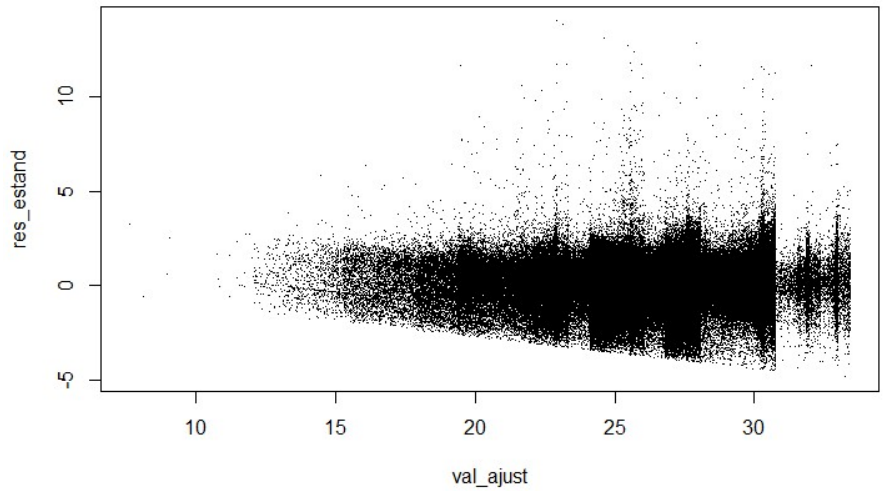


Figura 3.4.13: Diagrama de dispersión de los residuos estandarizados en función de los valores ajustados.

En la figura 4.4.13 se aprecia una ligera heterocedasticidad, es decir, parece que cuanto mayor es el valor ajustado de la variable respuesta mayor es la variabilidad de la misma.

Por otro lado, se aprecia un ligero patrón en los residuos que podrían ser un síntoma de falta de linealidad. Si llevamos a cabo el test de Durbin Watson, este nos proporciona un estadístico de contraste de 1.928, muy cercano a 2, por lo que llegamos a la conclusión de que no existe correlación lineal entre los residuos del ajuste.

Si llevamos a cabo el Test de Breusch-Pagan obtenemos un p-valor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis nula de homocedasticidad.

Para comprobar la normalidad de los residuos estandarizados realizamos un Q-Q plot como el de la figura 3.4.14.

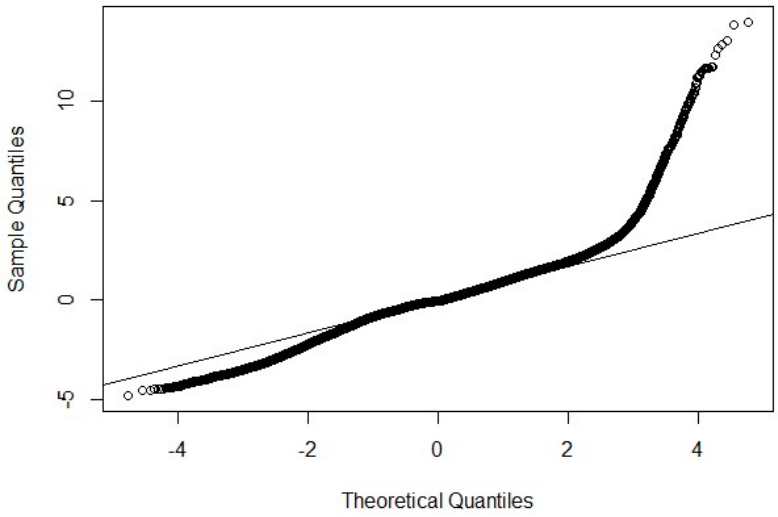


Figura 4.4.14: Q-Q plot de los residuos estandarizados

Al igual que habíamos observado en las figuras anteriores, en la figura 4.4.14 también llegamos a la conclusión de que los residuos no siguen una distribución normal ya que los puntos en las colas, principalmente los correspondientes a la cola derecha, se encuentran muy distanciados de la diagonal que marca la normalidad. Este hecho se debe tanto a al apuntamiento en torno a la media como a la asimetría que tiene su distribución.

Además, si llevamos a cabo diversos test de normalidad obtenemos los siguientes resultados:

- **Test de Lilliefors:** Obtenemos un pvalor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Anderson Darling:** Obtenemos un pvalor de $2x10^{-16}$ que nos lleva a rechazar la hipótesis de normalidad.
- **Test de Cramer von Mises:** Obtenemos un pvalor de $7.37x10^{-10}$ que nos lleva a rechazar la hipótesis de normalidad.

3.5. Modelo general

Una vez que ya hemos ajustado el modelo correspondiente a cada uno de los segmentos analizados hasta ahora, estamos en disposición de llevar a cabo un análisis del modelo general resultado de la combinación de los 4 segmentos.

A pesar de que al llevar a cabo la diagnosis de cada uno de los ajustes hemos visto que no se cumplen la mayor parte de las hipótesis de los modelos de regresión, se decide hacer el modelo general con estos ajustes. El motivo principal es que las diferentes transformaciones de las variables presentes en ellos no proporcionaron ajustes que si cumplieran dichas hipótesis y además los coeficientes de determinación ajustados resultaban incluso menos interesantes.

Haciendo una pequeña tabla veremos el número de clientes que pertenecen a cada uno de los segmentos:

	Número de clientes	**Porcentaje sobre el total**	**Coeficiente de determinación del modelo (R^2)**
Segmento 1 – Capt. Bienes	59.091	7,32 %	0,6932
Segmento 2 – Capt. Bienes (2)	66.905	8,29 %	0,6363
Segmento 3 – KYC	122.904	15,22 %	0,5154
Segmento 4 – Var. Básicas	558.403	69,17 %	0,1862
TOTAL	**807.303**	**100,00 %**	**0,3840**

Tabla 3.5.1: Proporción de clientes pertenecientes a cada segmento con el R^2 de su correspondiente modelo

Como se puede apreciar en la tabla, el segmento más numeroso es aquel del que solo disponemos de variables sociodemográficas y socio-laborales, y cuyo modelo es el que tiene un coeficiente de determinación más pequeño.

Por otro lado, podemos ver que los modelos ajustados para los segmentos de captura de bienes y de KYC tienen coeficientes de determinación significativamente más altos.

Por estos motivos, es más recomendable analizar el modelo formado solo por los tres primeros segmentos, de los que poseemos una información más completa. De esta forma, el coeficiente de determinación de este modelo alcanzará el 0.6275, lo que quiere decir que el modelo consigue explicar cerca del 63 % de la variabilidad del ingreso medio transformado mensual de los clientes de la entidad pertenecientes a dichos segmentos.

En la figura 3.5.1 comparamos la distribución de los ingresos predichos para cada una de las observaciones con las que estamos trabajando en cada uno de estos segmentos.

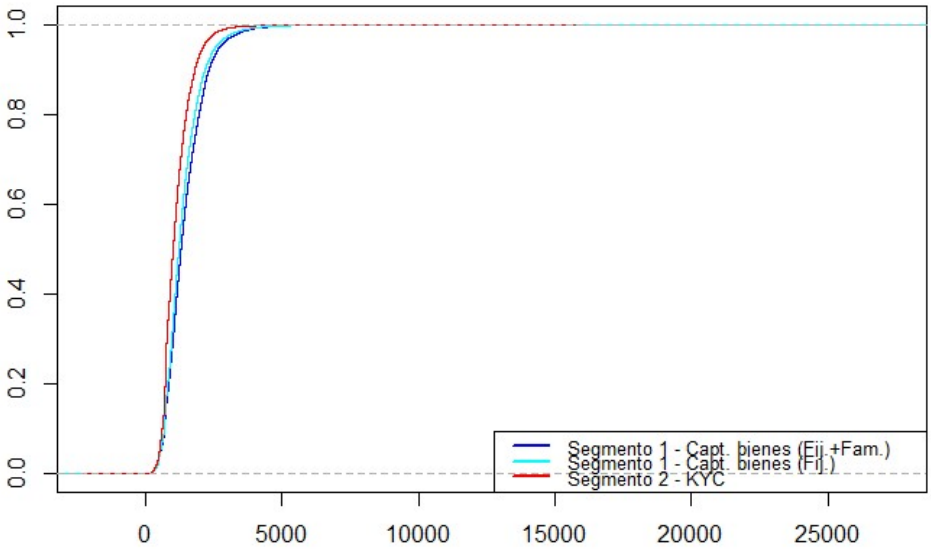


Figura 3.5.1: Comparación de la función de distribución de los ingresos predichos por segmento

En la figura 3.5.1 no se puede apreciar muy bien la diferencia entre los distintos segmentos ya que hay valores altos de ingresos. Por este motivo, es aconsejable representar las distribuciones de aquellos casos en los que el valor de ingresos no sobrepase los 7.000 euros, que como vemos corresponde a la mayor parte de los casos (figura 3.5.2).

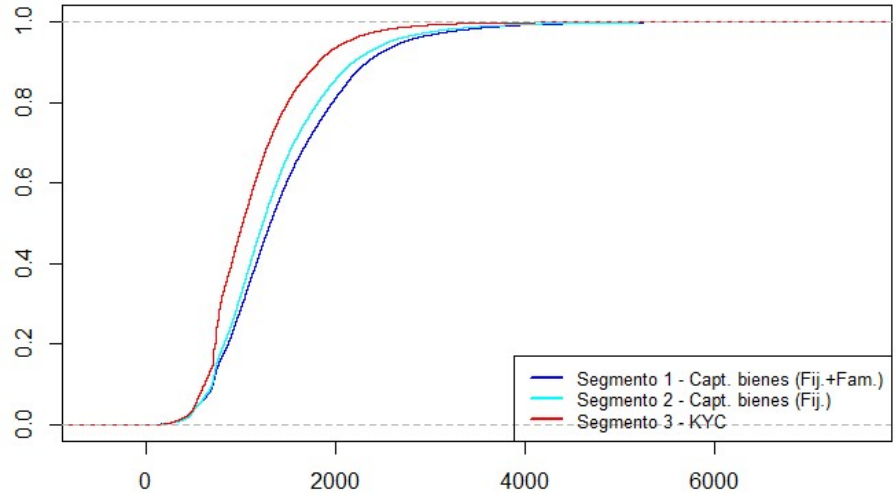


Figura 3.5.2: Comparación de la función de distribución de los ingresos predichos por segmento para aquellos casos en los que el valor del ingreso no supere los 7.000 euros.

En la figura 3.5.2 podemos ver que el segmento 1 de captura de bienes, en el que se encuentran los clientes tanto con información de ingresos fijos como de los ingresos de la unidad familiar, es el que proporciona un nivel de ingresos más alto, seguido muy de cerca por el segmento 2, en el que se encuentran los clientes con captura de bienes de los que tan solo tenemos información de ingresos fijos. Por otro lado, vemos que el nivel medio de ingresos en el segmento 3 es algo más bajo que en los otros dos segmentos.

El nivel medio de ingresos que proporciona el modelo para cada segmento es el siguiente:

- Segmento 1 – Captura de bienes (Fijos + Familiares) ---- **1.334,24 €**
- Segmento 2 – Captura de bienes (Fijos) ---- **1.256,15 €**
- Segmento 3 – KYC ---- **1.059,72 €**

Para comprender mejor cómo es la distribución del ingreso medio en cada uno de estos segmentos representamos sus correspondientes funciones de densidad estimadas en la figura 3.5.3.

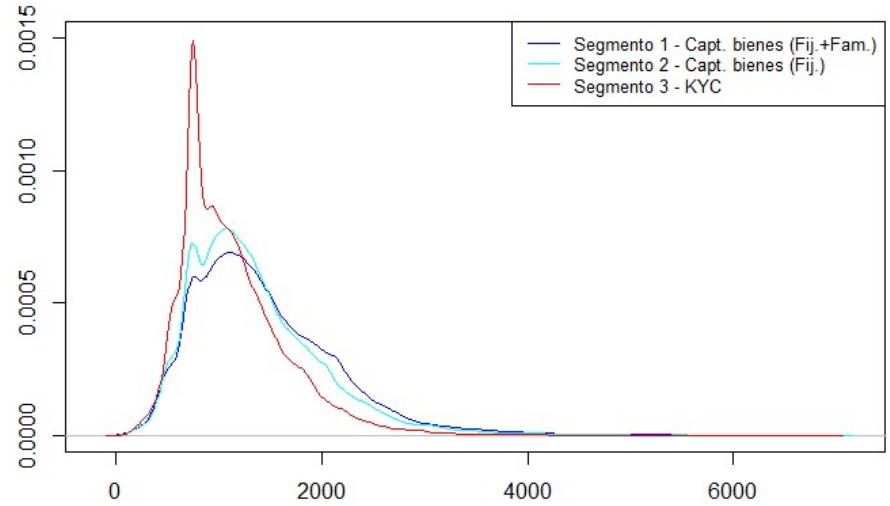


Figura 3.5.3: Comparación de la función de densidad estimada de los ingresos predichos por segmento para aquellos casos en los que el valor del ingreso no supere los 7.000 euros.

En la figura 3.5.3 podemos ver que la función de densidad estimada de los ingresos medios en los dos segmentos correspondientes a la captura de bienes son muy similares, mientras que en el segmento 3, correspondiente a clientes con KYC, se observa un pico muy importante en torno a los 800 euros que es el motivo principal por el que el nivel medio de ingresos es más bajo que en los otros dos segmentos.

3.6. Segmento 5: Productos de pasivo

En este segmento, al no obtener resultados satisfactorios para la creación de un modelo que nos permita obtener una estimación suficientemente fiable de los ingresos a través de los saldos de cada uno de los principales productos de pasivo que ABANCA ofrece a sus clientes, nos limitaremos a llevar a cabo un análisis exploratorio individualizado de cada uno de estos productos. La finalidad de estos análisis es que puedan servir para futuras investigaciones acerca de este tipo de productos.

Los productos que se van a analizar son:

- Depósitos a la vista
- Depósitos no a la vista
- Valores
- Fondos de inversión
- Fondos de pensiones

Para cada uno de estos productos realizaremos un análisis exploratorio individualizado y comprobaremos la relación que puedan tener los saldos medios de cada uno de ellos (calculando la media de los saldos medios de los últimos 12 meses que se presentan al final de cada trimestre) con la variable de ingreso medio mensual que hemos obtenido de los datos de ingresos en cuenta corriente que cada cliente ha recibido durante todo el año 2016 y los dos primeros meses del año 2017.

- Depósitos a la vista

Lo primero que demos tener en cuenta es que este tipo de producto es el que posee la mayor parte de los clientes de la entidad. En la base de datos con la que estamos trabajando, de los 800.620 clientes de los que tenemos información de ingresos, un total de 798.372 tienen al menos un depósito de este tipo (99,72 % del total).

Si hacemos un resumen de los saldos en depósitos a la vista obtenemos:

```
> summary(x1$SALDO_MEDIO_ANUAL_Vista)
   Min.    1st Qu.   Median     Mean     3rd Qu.      Max.
-369475      1022      5083     14156     16768     3189998
```

Podemos ver que la media es mucho mayor que la mediana debido a la existencia de algunos saldos muy grandes que llegan hasta los más de 3.000.000 millones de euros y que hacen crecer la media de una forma exagerada. Este efecto se puede observar mejor si hacemos un diagrama de cajas como el de la figura 3.6.1.

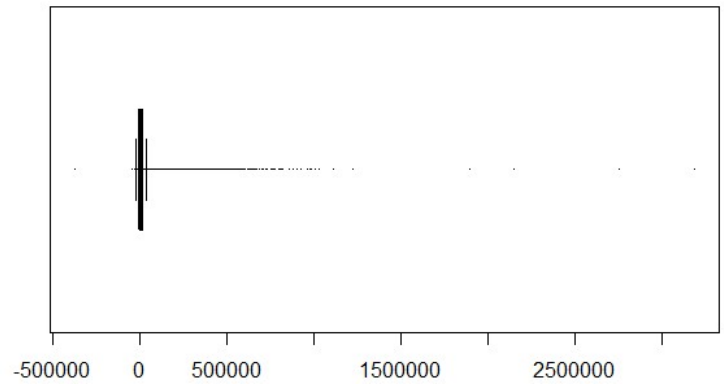


Figura 3.6.1: Diagrama de cajas del saldo medio en Depósitos a la vista.

En la figura 3.6.1 se puede apreciar mejor que la mayor parte de los saldos no superan el millón de euros, y que existen valores muy grandes que están afectando al cálculo de la media (al menos cuatro de ellos están por encima de 1.800.000 euros).

Por otro lado, también podemos observar la presencia de un cliente con un descubierto de cerca de 400.000 euros que podríamos considerar como dato atípico debido a que está muy distante del resto de los datos.

Si eliminamos estas observaciones que consideramos como atípicas, la distribución de la variable de saldo medio en depósitos a la vista quedaría de la siguiente forma:

```
> summary(x1.1$SALDO_MEDIO_ANUAL_Vista)
   Min.    1st Qu.    Median      Mean    3rd Qu.      Max.
  -4554       1022       5083     14137     16768     991285
```

Podemos observar que la media de la variable desciende un poco, aunque sigue siendo mucho más grande que la mediana.

Al volver a hacer el diagrama de cajas (figura 3.6.2) seguimos observando la fuerte asimetría que presentan los saldos medios en los depósitos a la vista.

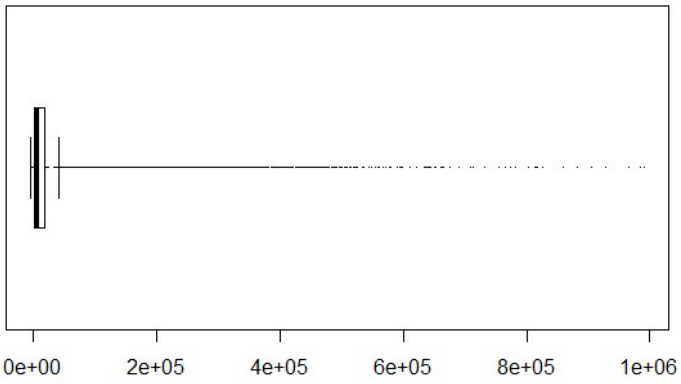


Figura 3.6.2: Diagrama de cajas del saldo medio en Depósitos a la vista positivos y menores de 1.000.000 euros.

Debido a dicha variabilidad será conveniente llevar a cabo, al igual que para la variable ingreso medio, una transformación Box-Cox que nos permitan reducir la misma y transformarla en una variable con una distribución más cercana a la normal.

Al transformar la variable "Saldo medio en Depósitos a la vista" a través del método Box-Cox y con su lambda correspondiente (0.315) obtenemos la siguiente distribución de la variable:

```
> summary(x1.1$SALDO_MEDIO_ANUAL_Vista2)
   Min.     1st Qu.     Median       Mean       3rd Qu.       Max.
 -3.726      13.602      19.682      19.707      25.457      57.461
   (0€)    (1.081€)    (5.218€)    (5.248€)   (16.966€)   (991.270€)
```

Podemos ver como los valores más altos que teníamos de la variable se han contraído de una forma bastante considerable, lo que produce que la media se acerque mucho a la mediana de los datos.

También podemos ver gráficamente cómo será la distribución de la variable transformada en la figura 3.6.3.

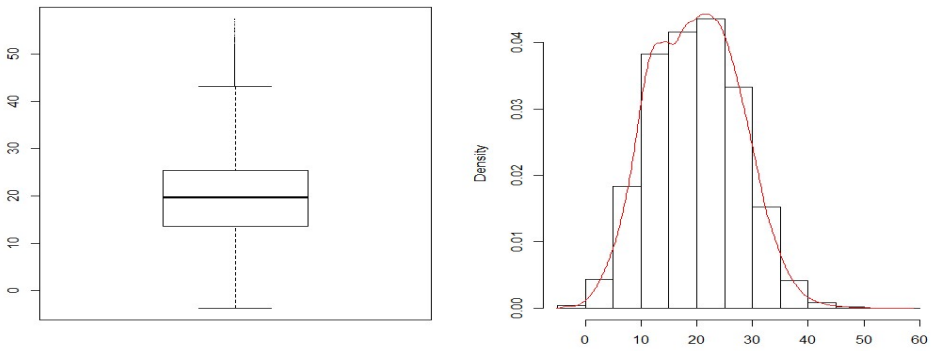


Figura 3.6.3: Diagrama de cajas e histograma con la estimación tipo núcleo de la función de densidad de la variable saldo medio en Depósitos a la vista transformada.

Una vez que hemos llevado a cabo la transformación, tanto de esta variable como de la variable ingreso medio (que ya hemos transformado anteriormente), estamos en disposición de ver la relación entre el ingreso mensual medio y el saldo medio en depósitos a la vista. Para ello, hacemos un diagrama de dispersión que ponga en relación ambas variables (figura 3.6.4).

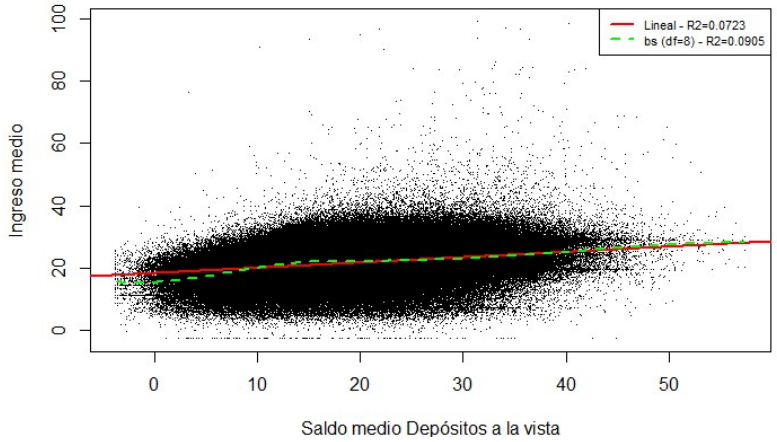


Figura 3.6.4: Diagrama de dispersión del ingreso medio transformado en función del saldo medio en depósitos a la vista transformado, y su correspondiente ajuste lineal y mediante B-Splines.

Podemos ver que existe una ligera relación positiva entre ambas variables, es decir, existe una tendencia que parece indicar que cuanto mayor es el ingreso medio transformado de un cliente mayor será el saldo medio que tendrá en su depósito a la vista. En el gráfico se representa tanto el ajuste lineal como un ajuste no paramétrico, cuyos coeficientes de determinación ajustados no llegan ni al 10 % debido a la alta dispersión que presenta la nube de puntos.

- Depósitos no a la vista

A diferencia de los depósitos que hemos analizado anteriormente, los depósitos no a la vista no tienen tanto éxito a pesar de ofrecer mayores rentabilidades. Mientras que casi todos los clientes de los que tenemos información de ingresos en cuenta corriente tienen un depósito a la vista, tan solo 167.856 clientes poseen un depósito no a la vista (un 20,97 % del total).

Si hacemos un resumen de los saldos en depósitos no a la vista obtenemos:

```
> summary(x2$SALDO_MEDIO_ANUAL_NoVista)
   Min.   1st Qu.    Median      Mean    3rd Qu.      Max.
     -1      9392     28409     50593     63000   5155324
```

Al igual que ocurría con los saldos de los depósitos a la vista, la media vuelva a ser mucho más grande que la mediana debido a que existen valores muy altos que la hacen aumentar considerablemente. Lo podemos apreciar mejor si hacemos un diagrama de cajas de la variable como el de la figura 3.6.5.

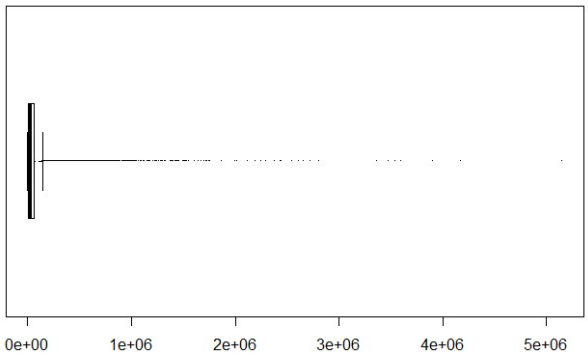


Figura 3.6.5: Diagrama de cajas del saldo medio en Depósitos No a la vista.

En el gráfico podemos ver que la mayor parte de los saldos no superan los 2.000.000 €, y los casos en los que se supera esta cantidad los valores están cada vez más dispersos.

Por este motivo, podríamos considerar estas observaciones como atípicas y será conveniente eliminarlas de nuestro análisis. Al hacerlo, la distribución de los saldos en depósitos no a la vista quedaría así:

```
> summary(x2.1$SALDO_MEDIO_ANUAL_NoVista)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
       -1      9384     28391     50119     63000   1872499
```

Vemos que la media ha disminuido cerca de 500 €, pero aún sigue siendo muy alta respecto a la mediana.

En la figura 3.6.6 podemos observar la fuerte asimetría que siguen presentando los saldos de los depósitos no a la vista.

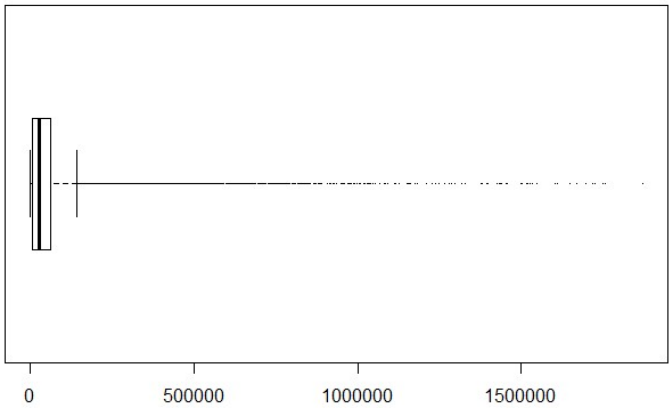


Figura 3.6.6: Diagrama de cajas del saldo medio en Depósitos No a la vista positivos y menores de 2.000.000 euros.

Para intentar reducir dicha variabilidad será conveniente llevar a cabo una transformación Box-Cox de la variable (con un λ de 0.304) que nos permita normalizar un poco la distribución de la misma.

La distribución de los saldos de depósitos no a la vista tras la transformación Box-Cox correspondiente la podemos intuir al hacer un resumen de la nueva variable:

```
> summary(x2.1$SALDO_MEDIO_ANUAL_NoVista2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -2.962     41.110     56.185     54.540     70.327     181.374
   (0€)  (10.000€)  (29.151€)  (26.304€)  (63.730€)  (1.872.500€)
```

La transformación de la variable hace que se contraigan los valores más altos de manera que la media se acerque a la mediana de la distribución.

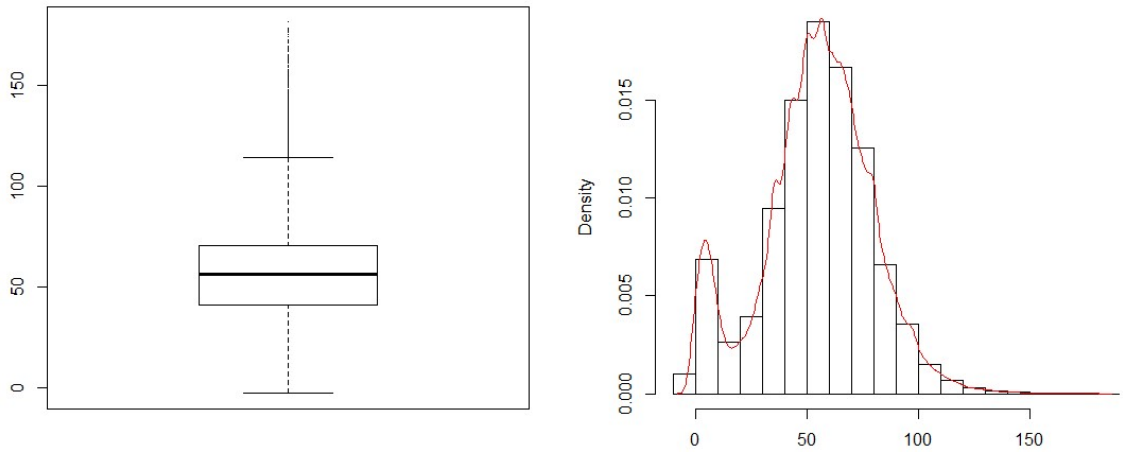


Figura 3.6.7: Diagrama de cajas e histograma con la estimación tipo núcleo de la función de densidad de la variable saldo medio en Depósitos a la vista transformada.

En el histograma de la variable apreciamos un pico muy considerable en torno al cero que se debe a que hay un importante número de clientes que tienen un saldo de no más de 10 euros en este tipo de depósito.

Una vez que tenemos transformadas tanto la variable de saldos en depósitos no a la vista como la variable de ingreso medio transformado podemos hacer un diagrama de dispersión que ponga en relación ambas variables para ver si existe algún tipo de correlación entre ambas.

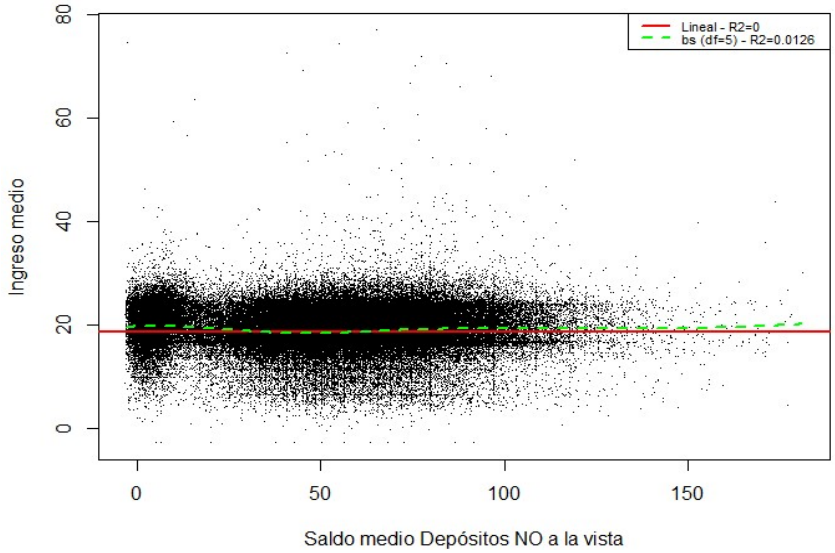


Figura 3.6.8: Diagrama de dispersión del ingreso medio transformado en función del saldo medio en depósitos No a la vista transformado, y su correspondiente ajuste lineal y mediante B-Splines.

En la figura 3.6.8 no se aprecia ningún tipo de relación entre ambas variables. De hecho, al intentar ajustar dicha relación tanto lineal como paramétricamente, obtenemos un coeficiente de determinación prácticamente nulo.

- Valores

La cuenta de valores es un producto que permite al cliente gestionar todas sus acciones de compra-venta de valores negociables en bolsa.

Hay que destacar que tan solo 30.514 clientes (de los que tenemos información de ingresos en cuenta corriente) poseen una cuenta de este tipo (solo un 3,81 % del total), lo que hace que sea el producto con menos clientes de los que estamos analizando.

Si hacemos un resumen de los saldos en valores:

```
> summary(x3$SALDO_MEDIO_ANUAL_Valores)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   -1839       617      2563     14385     10066  10227038
```

Podemos ver, como ya ocurría en los casos anteriores, que la media es mucho mayor que la mediana, y en este caso es incluso más grande que el tercer cuartil. Este hecho nos lleva a pensar que el efecto que los valores excepcionalmente grandes tienen sobre la media es aún más importante que en los depósitos. Si representamos el diagrama de cajas de la variable lo podemos apreciar mejor (figura 3.6.9).

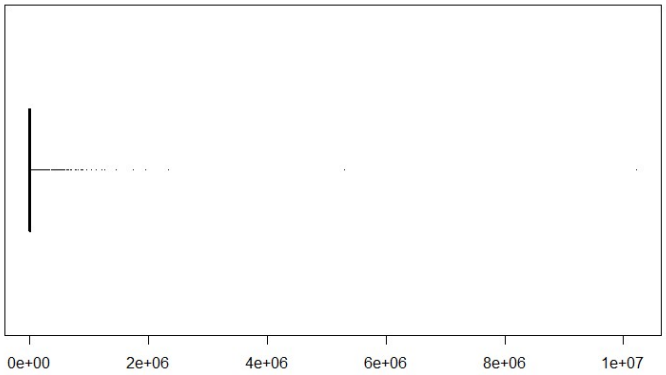


Figura 3.6.9: Diagrama de cajas del saldo medio en Valores.

En la figura 3.6.9 podemos ver que existe al menos una observación en torno a los 10.000.000 € que está tremendamente alejada del resto de observaciones, y que junto con otra cuyo valor está cercano a los 5.000.000 € están afectando mucho al cálculo de la media. Al eliminar estas observaciones, los principales estadísticos que resumen la distribución de la variable quedarían de la siguiente forma:

```
> summary(x3.1$SALDO_MEDIO_ANUAL_Valores)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   -1839       617      2562     13541     10057   2343164
```

Podemos ver como la media ha disminuido significativamente, aunque sigue siendo mucho más alta que la mediana.

Si hacemos un nuevo diagrama de cajas tras eliminar estas dos observaciones obtenemos el de la figura 3.6.10.

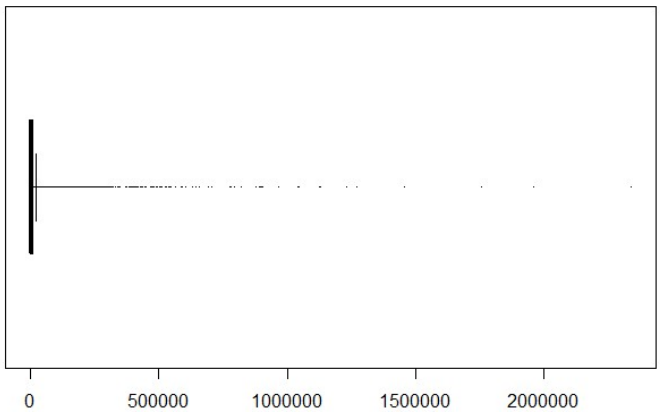


Figura 3.6.10: Diagrama de cajas del saldo medio en Valores positivos y menores de 2.500.000 euros.

En la figura 3.6.10 podemos ver que siguen existiendo observaciones que están muy alejadas del resto por lo que consideraremos como datos atípicos todas aquellas observaciones cuyo valor sobrepase los 300.000 €. De esta forma, los estadísticos que resumen la distribución de la variable quedarían así:

```
> summary(x3.1$SALDO_MEDIO_ANUAL_Valores)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   -1839      617     2540    11619     9887   297784
```

Podemos ver que la media vuelve a disminuir, no obstante sigue por encima del tercer cuartil de la distribución.

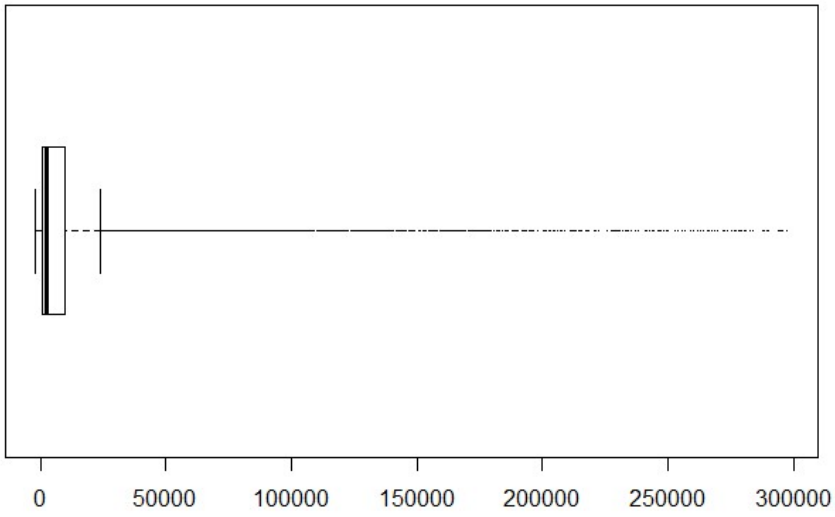


Figura 3.6.11: Diagrama de cajas del saldo medio en Valores positivos y menores de 300.000 euros.

Como vemos en la figura 3.6.11, se aprecia una fuerte asimetría en los saldos de las cuentas de Valores. Para intentar reducirla realizamos una transformación Box-Cox de la variable que nos permita contraer los valores más altos y que normalice un poco la distribución de la misma.

Si hacemos un resumen de la nueva variable podemos ver que la distribución de los saldos de las cuentas de valores tras la transformación Box-Cox correspondiente es:

```
> summary(x3.1$SALDO_MEDIO_ANUAL_Valores2)
    Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
  -2.494    8.620   11.263   11.168   14.124   22.984
   (0€)    (617€)  (2.544€) (2.425€) (9.891€) (297.795€)
```

Podemos ver que debido a la transformación de la variable, la media resulta casi del mismo valor que la mediana de la distribución.

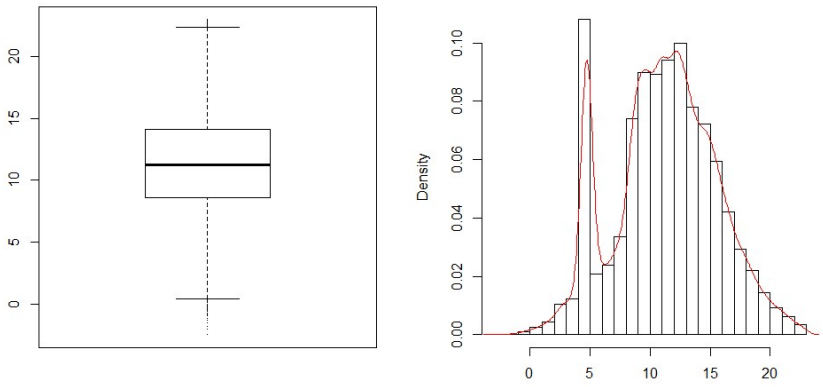


Figura 3.6.12: Diagrama de cajas e histograma con la estimación tipo núcleo de la función de densidad de la variable saldo medio en valores transformada.

En el histograma de la figura 3.6.12 se puede apreciar un gran pico en torno a los 50 euros que coincide con un importe del saldo medio de valores que se repite en cerca de 2.500 clientes (un 8,2 % del total) y que corresponde con algún tipo de producto estándar que la entidad ofrece a sus clientes.

Una vez transformadas tanto la variable de saldos en cuentas de valores como la variable de ingreso medio podemos hacer un diagrama de dispersión que ponga en relación ambas variables para ver si existe algún tipo de correlación entre ambas.

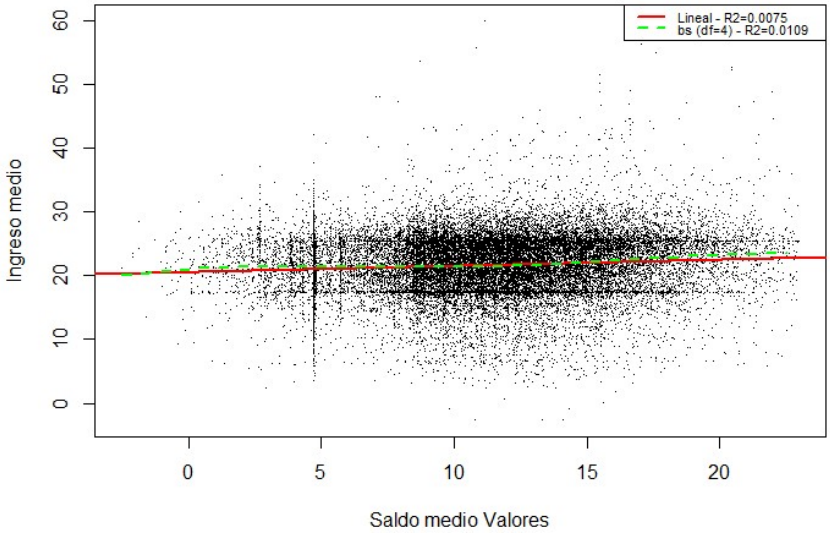


Figura 3.6.13: Diagrama de dispersión del ingreso medio transformado en función del saldo medio en Valores transformado, y su correspondiente ajuste lineal y mediante B-Splines.

En la figura 3.6.13 no se aprecia ningún tipo de relación entre ambas variables. De hecho, al intentar ajustar dicha relación tanto lineal como paramétricamente, obtenemos un coeficiente de determinación prácticamente nulo.

- Fondos de inversión

El fondo de inversión es un producto en el que el cliente adquiere una participación en un determinado paquete de activos financieros, que es gestionado por la entidad emisora del fondo, a cambio de una rentabilidad esperada determinada.

Al igual que ocurre con las cuentas de valores, los fondos de inversión son productos que poseen un número poco considerable de clientes de la entidad. En este caso, solo 35.151 clientes (de los que tenemos información de ingresos en cuenta corriente) poseen un producto de este tipo (solo un 4,39 % del total).

Si hacemos un resumen de los saldos en fondos de inversión:

```
> summary(x4$SALDO_MEDIO_ANUAL_F_Inversion)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
       0      5192     16462     39987     42057   8458328
```

Como viene siendo habitual en los saldos de todos los productos que estamos analizando, la media vuelve a ser mucho más grande que la mediana debido a la presencia de observaciones con un saldo medio extraordinariamente alto. Este hecho se aprecia mejor en la figura 3.6.14.

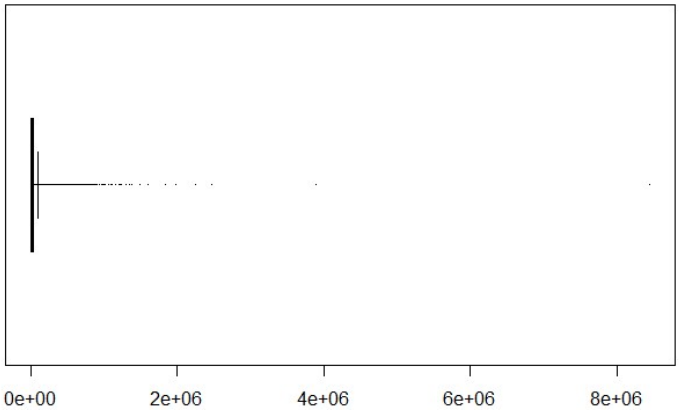


Figura 3.6.14: Diagrama de cajas del saldo medio en Fondos de inversión.

En la figura 3.6.14 observamos al menos dos observaciones, una por importe de más de 8.000.000 € y otra de cerca de 4.000.000 € que están afectando mucho al cálculo de la media. Si eliminamos estas observaciones que están excepcionalmente alejadas de las demás, los principales estadísticos que resumen la distribución de la variable quedarían así:

```
> summary(x4.1$SALDO_MEDIO_ANUAL_F_Inversion)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
       0      5192     16448     39637     42055   2484969
```

Eliminando dichas observaciones vemos que la media desciende sensiblemente, aunque sigue siendo muy alta respecto de la mediana. Por este motivo, representamos nuevamente el diagrama de cajas para establecer el límite a partir del cual podemos considerar a algunas de las observaciones como datos atípicos.

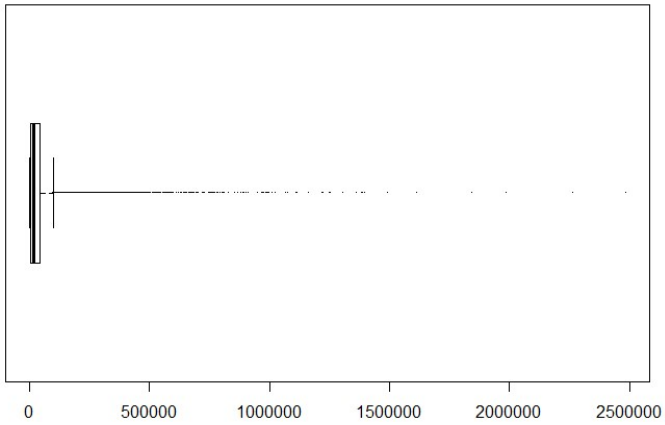


Figura 3.6.15: Diagrama de cajas del saldo medio en Fondos de inversión positivos y menores de 2.500.000 euros.

En la figura 3.6.15 seguimos observando un grupo de observaciones que siguen estando muy alejadas del resto, por lo que podemos establecer que a partir de los 800.000 € dichos saldos pueden considerarse como datos atípicos. Eliminándolos de la muestra, la variable quedaría de la siguiente forma:

```
> summary(x4.1$SALDO_MEDIO_ANUAL_F_Inversion)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
       0      5183     16381     37920     41815    794993
```

En esta ocasión vemos que la media sufre una disminución más considerable que la hace acercarse un poco más a la mediana de la distribución.

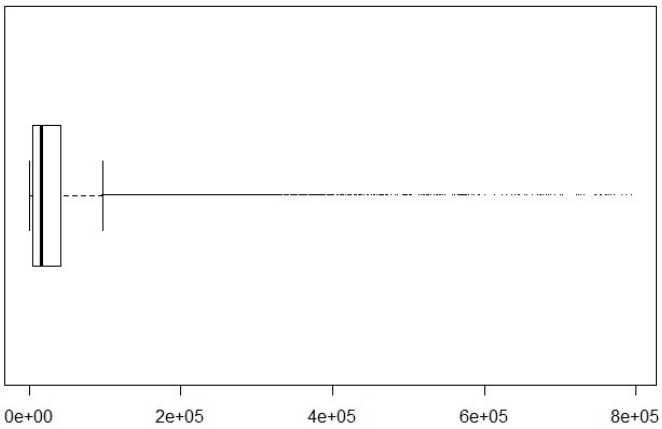


Figura 3.6.16: Diagrama de cajas del saldo medio en Fondos de inversión positivos y menores de 800.000 euros.

Una vez eliminados todos aquellos saldos que consideramos como datos atípicos, en la figura 3.6.16 podemos observar que los saldos en Fondos de inversión, al igual que ocurre con los saldos de los demás productos de pasivo, presentan una fuerte asimetría. Por este motivo se le aplicará una transformación Box-Cox (con un λ de 0.319) que nos permita obtener una variable con una distribución que se aproxime a la de una normal.

Una vez que llevamos a cabo la transformación Box-Cox correspondiente, el resumen con los estadísticos principales que nos describen la distribución de la nueva variable es:

```
> summary(x4.1$SALDO_MEDIO_ANUAL_F_Inversion2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -2.044    16.532      20.677     20.634     24.582     40.800
   (0€)   (5.183€)   (16.383€)  (16.203€)  (41.814€)  (795.000€)
```

Por lo que se observa en el resumen, la transformación ha conseguido centrar la distribución de los saldos en Fondos de inversión haciendo que la media sea muy parecida a la mediana de la distribución.

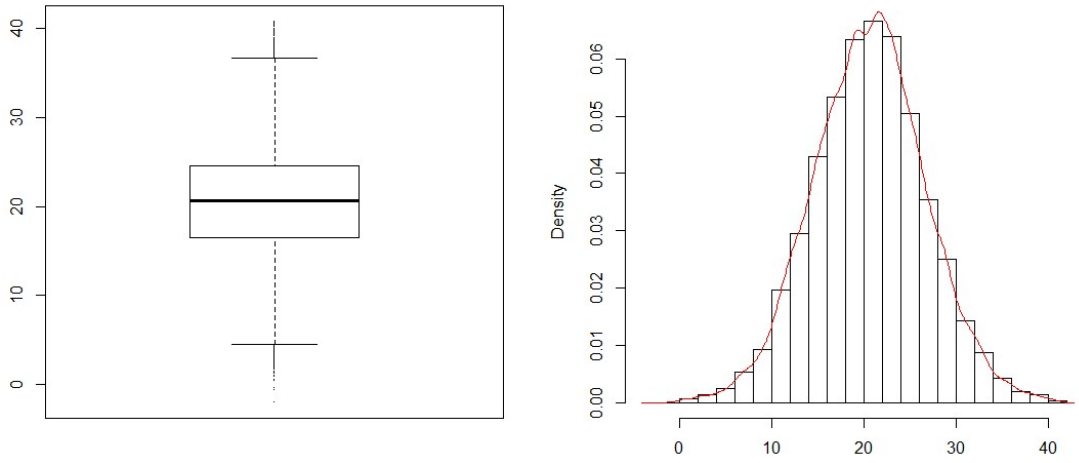


Figura 3.6.17: Diagrama de cajas e histograma con la estimación tipo núcleo de la función de densidad de la variable saldo medio en fondos de inversión transformada.

Una vez transformadas tanto la variable de saldos en Fondos de inversión como la variable de ingreso medio podemos hacer un diagrama de dispersión que ponga en relación ambas variables para ver si existe algún tipo de correlación entre ambas.

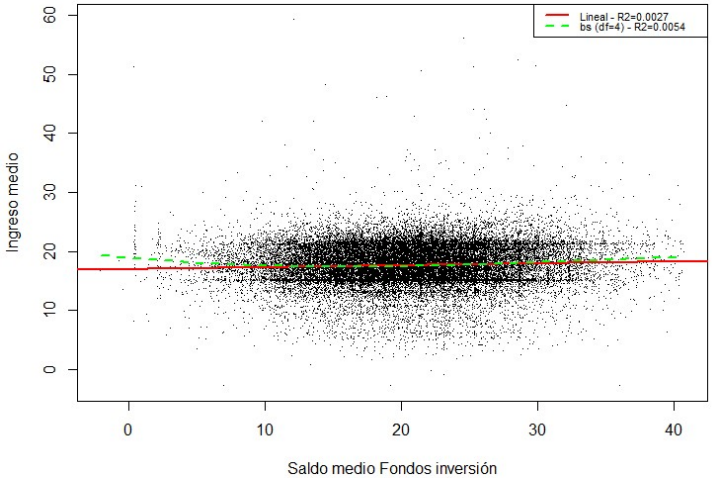


Figura 3.6.18: Diagrama de dispersión del ingreso medio transformado en función del saldo medio en Fondos de inversión transformado, y su correspondiente ajuste lineal y mediante B-Splines.

En la figura 3.6.18 no se aprecia ningún tipo de relación entre ambas variables. De hecho, al intentar ajustar dicha relación tanto lineal como paramétricamente, obtenemos un coeficiente de determinación prácticamente nulo.

- Fondos de pensiones

El fondo de pensiones es un instrumento financiero mediante el cual se canalizan los distintos flujos monetarios generados por un plan de pensiones, que a su vez es un instrumento de ahorro-inversión a largo plazo cuyo fin principal es el de complementar la pensión pública cubierta por la Seguridad Social.

Este tipo de producto cuenta con algo más de aceptación que los valores y los fondos de inversión dentro de la clientela de la entidad aunque se encuentra muy alejado del número de clientes que tienen contratado algún depósito, ya sean a la vista o no a la vista. En el caso de los fondos de pensiones, hay 54.707 clientes (de los que tenemos información de ingresos en cuenta corriente) que poseen un producto de este tipo (solo un 6,83 % del total).

Si hacemos un resumen de los saldos en fondos de pensiones:

```
> summary(x5$SALDO_MEDIO_ANUAL_F_Pension)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
      0       590      3896      8575     10126    524098
```

Como en los casos anteriores, podemos apreciar que la media de la variable es mucho más grande que la mediana de su distribución. Hay que destacar que la diferencia no es tan acusada como en casos anteriores ya que las observaciones más alejadas no tienen un importe tan elevado sino que el saldo máximo es de algo más de 500.000 €. En este caso, y fijándonos en la figura 3.6.19, estableceremos como datos atípicos todas aquellas observaciones en la que el saldo en fondos de pensiones sobrepasen los 200.000 €.

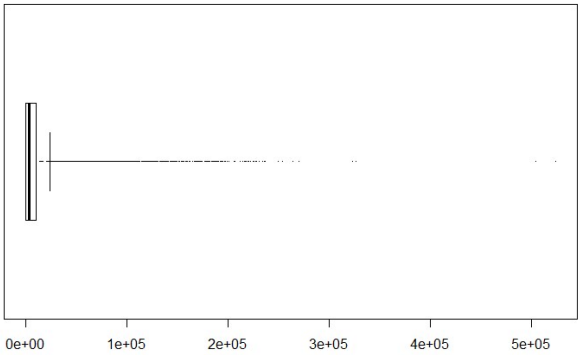


Figura 3.6.19: Diagrama de cajas del saldo medio en Fondos de pensiones.

Si hacemos un resumen de la variable tras eliminar las observaciones que consideramos como datos atípicos, los estadísticos principales que nos describen a la variable quedarían de la siguiente forma:

```
> summary(x5.1$SALDO_MEDIO_ANUAL_F_Pension)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       0      588     3892     8453    10113   199648
```

Podemos ver que la media no sufre una disminución demasiado importante, aunque al representar el nuevo diagrama de cajas ya no apreciamos un salto en los importes de los saldos como sucedía anteriormente.

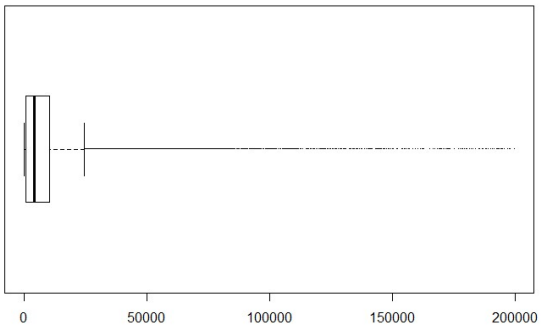


Figura 3.6.20: Diagrama de cajas del saldo medio en Fondos de pensiones positivos y menores de 200.000 euros.

Una vez llegados a este punto, y a tenor de la fuerte asimetría que se aprecia en los saldos, será conveniente llevar a cabo la correspondiente transformación Box-Cox (con un λ de 0.323) que nos permita contraer los valores más altos, así como obtener una distribución de la variable que se asemeje a la de una normal.

Si hacemos un resumen de la nueva variable podemos ver que la distribución de los saldos en fondos de pensiones tras la transformación es:

```
> summary(x5.1$SALDO_MEDIO_ANUAL_F_Pension2)
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
 -3.271     11.012     17.044     16.118     20.852     37.223
   (0€)     (589€)   (3.892€)   (3.014€)  (10.114€)  (199.650€)
```

Como podemos ver, la media de la variable transformada se acerca mucho a la mediana de la distribución, llegando incluso a tener un importe sensiblemente menor que esta.

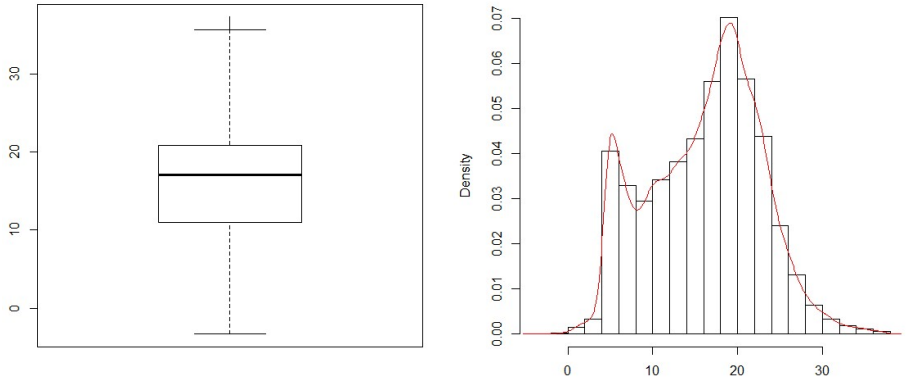


Figura 3.6.21: Diagrama de cajas e histograma con la estimación tipo núcleo de la función de densidad de la variable saldo medio en fondos de pensión transformada.

En el histograma de la figura 3.6.21 se puede apreciar un pico importante a partir de los 30 euros que coincide con la aportación mínima necesaria para contratar un plan de pensiones con la entidad.

Una vez transformadas tanto la variable de saldos en Fondos de pensiones como la variable de ingreso medio podemos hacer un diagrama de dispersión que ponga en relación ambas variables para ver si existe algún tipo de correlación entre ambas.

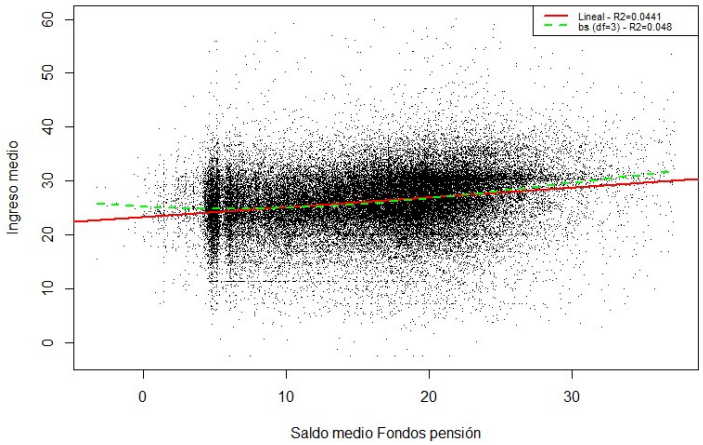


Figura 3.6.22: Diagrama de dispersión del ingreso medio transformado en función del saldo medio en Fondos de pensión transformado, y su correspondiente ajuste lineal y mediante B-Splines.

Podemos considerar que existe una muy ligera relación positiva entre ambas variables, es decir, existe una tendencia que parece indicar que cuanto mayor es el saldo medio que tenga un cliente en su fondo de pensiones mayor será su ingreso medio. En la figura 3.6.22 se representa tanto el ajuste lineal como un ajuste no paramétrico, cuyos coeficientes de determinación ajustados no llegan ni al 5 % debido a la alta dispersión que presenta la nube de puntos.

Capítulo 4

Conclusiones

Las principales conclusiones que podemos sacar del trabajo realizado son:

- Los coeficientes de determinación obtenidos en los diferentes modelos ajustados no son lo suficientemente altos para que estos modelos puedan ser utilizados para una predicción suficientemente fiable de los ingresos de un cliente potencial.
    - El modelo para el segmento 1 correspondiente a los clientes con ingresos positivos tanto de ingresos fijos como de ingresos de la unidad familiar en la captura de bienes tiene un coeficiente de determinación del 69’32 %.
    - El modelo para el segmento 2 correspondiente a los clientes con ingresos positivos tan solo de ingresos fijos en la captura de bienes tiene un coeficiente de determinación del 63’63 %.
    - El modelo para el segmento 3 correspondiente a los clientes con KYC y que no se encuentran en ninguno de los segmentos anteriores tiene un coeficiente de determinación del 51’54 %.
    - El modelo para el segmento 4 correspondiente a los clientes con tan solo información sociodemográfica básica tiene un coeficiente de determinación del 18’62 %.
- Se confirma que, si bien tiene bastante fiabilidad usar el dato de ingresos fijos netos tecleado en una captura de bienes relativamente reciente, resulta significativamente menos fiable usar este mismo dato del KYC. No obstante, el coeficiente de determinación de la variable ingresos fijos netos mensuales de la captura de bienes resulta inferior a lo que la entidad esperaba por lo que el uso de esta información en la estimación de los ingresos de los clientes menos vinculados deberá hacerse con prudencia.
- Como cabía esperar, si se utiliza únicamente información sociodemográfica básica no es posible hacer una estimación de ingresos mínimamente fiable. Con estas variables el coeficiente de determinación que se obtiene es muy reducido, es decir, la variabilidad explicada de la variable objetivo es muy poca.

Referencias

[1] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59–88.

[2] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.

[3] Christauskas, C., Stunguriene, S. (2007). Research on Difficulties of Financial Decision Making Under Uncertainty Conditions. Transformations in Business & Economics(2), 98-113.

[4] Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447–1465.

[5] Henley, W. E., Hand, D. J. (1996) A K-Nearest-Neighbour Classifier For Assessing Consumer Credit Risk, Statistician , 45(1), 77-95.

[6] Hllig, K., Hrner, J. (2013). Approximation And Modelling With B-Splines. Society for Industrial and Applied Mathematics Philadelphia.

[7] Ince, H., Aktan, B. (2009). A Comparison Of Data Mining Techniques For Credit Scoring In Banking: A Managerial Perspective. Journal of Business Economics and Management(10), 233-240.

[8] Sakia, R. M., (1992). The Box-Cox Transformation Technique: A Review. Journal of the Royal Statistical Society, 41(2), 169-178

[9] Vojtek, M., Kocenda, E., (2006). Credit Scoring Methods. Czech Journal of Economics and Finance (56), 152-167.

[10] Wei, G., Yun-Zhong, C., & Ming-shu, C. (2014). A new dynamic credit scoring model based on the objective cluster analysis. In Practical applications of intelligent systems. In Advances in Intelligent Systems and Computing. Springer Berlin Heidelberg, 279, 579–589.

[11] Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. European Journal of Operational Research, 183(3), 1521–1536.