Técnicas de estadística multivariante para el estudio de la variabilidad metabólica asociada a la presencia de los distintos alelos del gen APOE

Trabajo de Fin de Máster

Máster en Modelización e Investigación Matemática, Estadística y Computación

Autor:

David Mompel Lancina

Directores:

José Miguel Arbonés Mainar

Beatriz Lacruz Casaucau

Ana Pérez Palomares

Universidad de Zaragoza

Facultad de Ciencias

Noviembre 2017

Abstract

The APOE gene is a polymorphic gene that has three major alleles: APOE2, APOE3, and APOE4. The interest of their study is that the presence of one allele or another has physiological consequences, for example, the presence of APOE2 is associated with a decrease in the risk of Alzheimer's and cardiovascular disease, while the presence of APOE4 increases the risk of suffering from these diseases in addition to metabolic syndrome. The APOE3 variant is the most frequent allele and corresponds to 65-70 % of the Spanish population, while APOE2 and APOE4 correspond to 10-15 % each.

Metabolites are the final products of all processes that occur in cells and it is estimated that there are more than 2,000 different ones that can be synthesized endogenously. The present quantities of a certain metabolite in the body reflect the adaptation of the organism to a certain stimulus (illness, feeding, genetic variation, etc.). The different metabolites interact with each other, they are organized in the form of networks and they can be modulated by the presence (or absence) of different genes or alleles.

In this work we study the relationship among APOE gene and several metabolites belonging to the sets of fatty acids, amino acids, biochemestry analysis and carnitines. For this purpose we apply several multivariate statistical techniques as principal component analysis, multinomial logistic regression and classification trees. We have four data sets, one for each kind of metabolite, with different number of cases and different number of variables. Since the number of cases is very small in relation to the total number of variables, we tackle the problem set by set. After this analysis we obtain the set of variables (metabolites) which better explain the APOE gene and they are used to obtain a global predictive model. The final conclusion is that, even though any metabolite has enough predictive power, there are several variables which appear as very important in several models.

Key words: APOE gene, metabolites, principal component analysis, multinomial logistic regression, classification trees.

Resumen

El gen APOE es un gen polimórfico que tiene tres alelos principales: APOE2, APOE3 y APOE4. El interés de su estudio radica en que la presencia de un alelo u otro tiene consecuencias fisiológicas profundas, por ejemplo, la presencia de APOE2 se asocia a una disminución del riesgo de Alzheimer y enfermedad cardiovascular, mientras que la presencia de APOE4 aumenta el riesgo de padecer estas enfermedades además de síndrome metabólico. La variante APOE3 es mayoritaria y corresponde a un 65-70 % de la población española, mientras que APOE2 y APOE4 corresponden a un 10-15 % cada uno.

Los metabolitos son los productos finales de todos los procesos que se producen en las células y se estima que hay más de 2.000 diferentes que pueden ser sintetizados de forma endógena. Las cantidades presentes de un determinado metabolito reflejan la adaptación del organismo ante un determinado estímulo (enfermedad, alimentación, variación genética, etc). Los diferentes metabolitos interaccionan entre sí, están organizados en forma de redes y pueden ser modulados por la presencia (o ausencia) de distintos genes o alelos.

En este trabajo estudiaremos la relación entre el gen APOE y los distintos metabolitos pertenecientes a los grupos de ácidos grasos, aminoácidos, análisis bioquímico y carnitinas. Para ello, utilizaremos diferentes técnicas de estadística multivariante como análisis de componentes principales, regresión logística multinomial y árboles de clasificación. Disponemos de cuatro conjuntos de datos, cada uno correspondiente a cada tipo de metabolito con diferente número de observaciones y variables. Ya que el número de observaciones es muy pequeño en relación con el número total de variables, abordaremos el problema conjunto a conjunto. Tras este análisis hemos obtenido un conjunto de variables que mejor explican el gen APOE y estas han sido usadas para obtener un modelo global predictivo. La conclusión final es que, a pesar de que ningún metabolito tiene suficiente poder predictivo, hay varias variables que aparecen como muy importantes en varios modelos.

Palabras clave: gen APOE, metabolitos, análisis de componentes principales, regresión logística multinomial, árboles de clasificación.

Índice general

Capítulo 1

Descripción del problema y de los datos disponibles

1.1. Introducción

El gen APOE es un gen polimórfico que tiene tres alelos principales: APOE2, APOE3 y APOE4. El interés de su estudio radica en que la presencia de un alelo u otro tiene consecuencias fisiológicas profundas, por ejemplo, la presencia de APOE2 se asocia a una disminución del riesgo de Alzheimer y enfermedad cardiovascular, mientras que la presencia de APOE4 aumenta el riesgo de padecer estas enfermedades además de síndrome metabólico. La variante APOE3 es mayoritaria y corresponde a un 65-70 % de la población española, mientras que APOE2 y APOE4 corresponden a un 10-15 % cada uno.

Los metabolitos son los productos finales de todos los procesos que se producen en las células y se estima que hay más de 2.000 diferentes que pueden ser sintetizados de forma endógena. Las cantidades presentes de un determinado metabolito reflejan la adaptación del organismo ante un determinado estímulo (enfermedad, alimentación, variación genética, etc). Los diferentes metabolitos interaccionan entre sí, están organizados en forma de redes y pueden ser modulados por la presencia (o ausencia) de distintos genes o alelos.

El objetivo de este Trabajo de Fin de Máster consiste en la construcción de modelos que permitan determinar qué metabolitos están asociados con cada una de las tres variantes del gen APOE. Su determinación nos permitiría conocer los cambios metabólicos que se asocian a la presencia de cada uno de los alelos del gen APOE.

Los datos para el estudio provienen de 242 pacientes intervenidos en los servicios de cirugía de los hospitales Miguel Servet y Royo Villanova (Zaragoza) [1]. Los metabolitos están organizados en cuatro conjuntos de datos que recogen el análisis de ácidos grasos (con 47 variables), aminoácidos (con 26 variables), análisis bioquímico (con 22 variables) y carnitinas (con 31 variables). Un quinto archivo contiene los alelos del gen APOE de cada paciente. Solo se dispone de los datos completos para el conjunto correspondiente al análisis bioquímico. El tamaño de los conjuntos de datos es respectivamente de 231, 158, 242, 238 y 238, observaciones.

En este estudio se aplican técnicas estadísticas para determinar cuáles son los metabolitos (va-

riables) que tienen mayor poder predictivo, se construyen modelos de clasificación supervisada con diferentes métodos como la regresión logística multinomial y los árboles de clasificación. Finalmente, se valida y compara la eficacia de los mismos. Para el análisis estadístico se ha utilizado R [2] y SAS Enterprise Guide.

El trabajo está organizado como sigue. En este primer capítulo se describen los conjuntos de datos que contienen la información sobre cada uno de los tipos de metabolitos antes mencionados, así como del que contiene los alelos del gen APOE. El capítulo 2 contiene la descripción de las principales técnicas estadística que se han utilizado para determinar las variables que tienen mayor poder predictivo: análisis de componentes principales, regresión logística multinomial y árboles de clasificación. En el capítulo 3 se presentan los resultados del análisis exploratorio de cada uno de los conjuntos de datos según los alelos del gen APOE. Para ello se han aplicado por una parte, herramientas de estadística descriptiva calculando medidas de posición y dispersión y realizando representaciones gráficas como los diagramas de caja y la estimación de funciones de densidad. También se han aplicado herramientas de inferencia estadística como el test de Kruskal-Wallis, que permite detectar diferencias significativas en el comportamiento de las variables, y dentro del conjunto de ácidos grasos, se ha realizado un estudio de correlaciones sugerido por la naturaleza de las variables que contiene. Además, dado que el número de casos no es suficiente para abordar la construcción de un modelo con todas las variables disponibles, se ha optado por intentar encontrar dentro de cada conjunto de metabolitos aquellos que tienen más capacidad para explicar la variabilidad de los datos mediante un análisis de componentes principales, así como aquellos que tienen mayor capacidad para predecir el alelo del gen APOE de un individuo mediante la construcción de modelos de regresión logística multinomial y árboles de clasificación incluyendo y sin incluir las componentes principales que mayor variabilidad explican. En el capítulo 4 se han construido los modelos finales con las variables seleccionadas de cada conjunto de datos. Este capítulo termina con las conclusiones finales sobre cuáles son los metabolitos que están más asociados con cada una de las tres variantes de APOE. Se incluyen 6 anexos que incluyen tablas y gráficas que no han sido incluidos dentro del texto para facilitar la lectura del mismo.

1.2. Conjuntos de datos

Para el análisis de datos, disponemos de 5 conjuntos de datos, que son los siguientes: ácidos grasos, aminoácidos, bioquímicas, carnitinas y APOE. Todos los conjuntos tienen dos variables en común, que son el identificador y el número de muestra, que son los identificadores de cada caso. En alguno de los conjuntos, hay una variable identificador más (*numero.analisis*), pero, al no estar en todos los conjuntos, ignoraremos esta variable de los conjuntos.

La documentación utilizada en esta sección puede consultarse en [3, 4, 5].

1.2.1. Ácidos Grasos

Los ácidos grasos son biomoléculas lípidas formadas por una larga cadena hidrocarbonada de tipo alifático, es decir, lineal ($-CH_2-CH_2-CH_2-$); con un número par de átomos de carbono, el último de los cuales forma un grupo carboxilo ($-COOH$), también llamado grupo ácido. Los ácidos grasos son los principales constituyentes de ciertos lípidos, como las grasas.

El conjunto de datos de Ácidos Grasos tiene 231 casos y 47 variables que hacen referencia a los ácidos grasos que o bien circulan libres por el plasma sanguíneo o bien unidos a la membrana de lo glóbulos rojos.

Los ácidos grasos que trataremos en el conjunto de datos son los siguientes:

Ácido láurico (X120), Ácido mirístico (X140), Ácido palmítico (X160), Ácido almitoleico (X161), Ácido esteárico (X180), Ácido oleico (X181), Ácido linoleico (X182), Ácido γ-linoleico (G183), Ácido α-linoleico (A183), Ácido octadecatetraenoico (X184), Ácido araquídico (X200), Ácido eicosamonoenoico (X201), Ácido eicosadienoico (X202), Ácido eicosatrienoico (X203), Ácido araquidónico (X204), Ácido eicosapentaenoico (X205), Ácido behénico (X220), Ácido erúcico (X221), Ácido docosatetraenoico (X224), Ácido docosapentaenoico (X225), Ácido docosahexaenoico (X226), Ácido lignocérico (X240), Ácido nervoico (X241) y Ácido hexacosanoico (X260).

Y las variables que trataremos serán las siguientes, donde la terminación "H" indica que va unido a la membrana del glóbulo rojo y la terminación "P" que circula libre en el plasma:

v120H, v140H, v160H, v161H, v180H, v181H, v182H, g183H, a183H, v184H, v200H, v201H, v202H, v203H, v204H, v205H, v220H, v221H, v224H, v225H, v240H, v226H, v241H, v260H, v140P, v160P, v161P, v180P, v181P, v182P, g183P, a183P, v184P, v200P, v201P, v202P, v203P, v204P, v205P, v220P, v221P, v224P, v225P, v240P, v226P, v241P, v260P.

Todas las variables de este conjunto son variables cuantitativas y se miden en $\%$.

1.2.2. Aminoácidos

Los aminoácidos son compuestos orgánicos, de baja masa molecular, que se caracteriza por poseer un grupo carboxilo $-COOH$ y un grupo amino $-NH_2$. Son compuestos sólidos, cristalinos, solubles en agua, con un punto de fusión elevado y con actividad óptica.

Se pueden clasificar según su obtención en esenciales y no esenciales. Los esenciales no se sintetizan por el organismo y deben de ser ingeridos en la dieta, y los no esenciales son sintetizados por el organismo.

Según la ubicación del grupo amino se pueden clasificar como alfa-, beta- o gamma-aminoácidos.

El conjunto de Aminoácidos tiene 158 casos y 26 variables, y las variables que trataremos son las siguientes:

Fosfoserina, Taurina, Ácido Aspártico, Treonina, Serina, Asparragina, Ácido Glutámico, Glutamina, Glicina, Alanina, Citrulina, Ácido α-Aminobutírico, Valina, Cistina, Metionina, Isoleucina, Leucina, Tirosina, Fenil-Alanina, Ornitina, Lisina, X1-Metil-Histidina, Histidina, Triptófano, Arginina, Prolina.

Todas estas variables son variables cuantitativas y se miden en $\mu mol/L$ *sangre*.

1.2.3. Bioquímicas

Las variables bioquímicas son las variables de ciertos compuestos que se recogen en los análisis de sangre.

El conjunto de Bioquímicas tiene 242 casos y 22 variables. Las variables que trataremos son las siguientes:

Proteína C reactiva (*PCRU*), Glucosa en suero (*GLU*), Triglicéridos en suero (*TRIG*), Colesterol en suero (*CHOL*), Colesterol HDL en suero (*cHDL*), Colesterol LDL en suero (*LDL*), *GGT, GOT, GPT*, Insulina (*INS*), Leptina (*LEPT*), Apolipoproteína A1 en suero (*APOA*), Apolipoproteína B en suero (*APOB*), Lipoproteína en suero (*LPA*), Beta Hidroxibutirato (*BHID*), *NEFA*, Hemoglobina glicosilada (*HEMG*), Hemoglobina glicosilada mmol/mol (*HEMGm*), Selenio en suero (*SE*), *BGP*, Telopéptidos C-terminal del colágeno Tipo I = Beta-CrossLaps = CTX-I en suero (*CTx*) y Vitamina D en suero (*VitD*).

Todas las variables son cuantitativas.

Las variables $PCRU$, GLU, $TRIG$, $CHOL$, $cHDL$, LDL, $APOA$, $APOB$, LPA y $BHID$ se miden en mg/dL; las variables GGT, GOT y GPT se miden en U/L; la variable INS se mide en $\mu Ul/mL$; la variable $LEPT$ se mide en ng/mL; la variable $HEMG$ se mide en $\%$; la variable $HEMGm$ se mide en $mmol/mol$; la variable SE se mide en $\mu g/mL$; la variable CTx se mide en pg/mL; la variable $VitD$ se mide en $nmol/L$.

1.2.4. Carnitinas

La carnitina es una amina cuaternaria sintetizada por el hígado, los riñones y el cerebro a partir de dos aminoácidos esenciales, la lisina y la metionina. Es la responsable del transporte de ácidos grasos al interior de las mitocondrias, más concretamente del grupo acilo de éstos. El proceso por el que se produce el transporte es el siguiente:

La enzima carnitina palmitoiltransferasa I (CPTI) escinde el coenzima A (CoA) de la molécula acilo-CoA y une el grupo acilo a la carnitina formando acilcarnitina, que puede atravesar la membrana de la mitocondria. El CoA resultante se puede unir a otro ácido graso y así formar acilo-CoA. La carnitina, una vez que ha dejado el grupo acilo dentro de la mitocondria, sale al exterior donde se repite el proceso.

Las variables que trataremos en el conjunto de Carnitinas son diferentes tipos de acilcarnitinas. El conjunto tiene 238 casos y 31 variables. Las variables son las siguientes:

C0n, C2n, C3n, C3DCn, C4n, C4DCn, C5n, C51n, C5DCn, C6n, C6DCn, C8n, C81n, C10n, C101n, C102n, C12n, C121n, C14n, C14On, C141n, C142n, C16n, C161n, C16On, X161On, C18n, C18On, C181n, C181On, C182n.

Todas las variables son cuantitativas y se miden en $\mu mol/L$ *sangre*.

1.2.5. APOE

La apolipoproteína E (APOE) es una molécula de la familia de apoproteínas que se encuentra en los quilomicrones y lipoproteínas de densidad intermedia (IDLs) que es esencial para el normal catabolismo de proteínas ricas en triglicéridos.

El gen está localizado en el cromosoma 19 en el mismo clúster con las Apolipoproteínas C1 y C2.

El gen consiste en 4 exones y 3 intrones, con una total de 3597 pares de bases.

El conjunto de APOE tiene 238 casos y 4 variables, que hacen referencia al fenotipo y genotipo del gen APOE de cada caso.

Las variables son las siguientes:

snp112, snp158, apoE y apoE2.

La variable *apoE* muestra el genotipo del gen APOE de cada caso, a saber: *E2E3, E2E4, E3E3* y *E3E4*. La variable *apoE2* muestra el fenotipo del gen y toma los siguientes valores: *E2, E3* y *E4*.

1.3. Combinación de los conjuntos de datos

A la hora de combinar todos los conjuntos, se ha seleccionado un conjunto de datos en el que estuvieran todos los pacientes a estudiar. Este conjunto es el conjunto de Bioquímicas. Sobre este conjunto se han ido añadiendo las variables de los demás conjuntos.

Como se indicó en la Sección 1.2, se dispone de dos identificadores de caso que aparecen en todos los conjuntos de datos: el identificador (*ID*) y número de muestra (*muestra*). Sin embargo, en el proceso de depuración se han detectado algunos errores de transcripción y se ha producido a utilizar únicamente el número de muestra para unir los conjuntos de datos adecuadamente. Los errores detectados son la duplicidad de los pares $(ID,\ muestra)$. Las parejas afectadas son las siguientes:

muestra:	63	$\rightarrow$	53913	muestra:	238	$\rightarrow$	262430
			539113				626430
muestra:	117	$\rightarrow$	265971	muestra:	211	$\rightarrow$	1136578
			264971				389045
muestra:	221	$\rightarrow$	40515	muestra:	213	$\rightarrow$	1136578
			404515				389045

1.4. Otras observaciones a los conjuntos

La variable $X161On$ del conjunto de *Carnitinas* se renombra a $C161On$ ya que es una carnitina, y todas de este conjunto empiezan por C.

La variable *ID* de todos los conjuntos se elimina por la duplicidad vista en la Sección 1.3. También se elimina la variable *numero.analisis* por no estar presente en todos los conjuntos.

Capítulo 2

Metodología

La estadística multivariante [6] es un área de la estadística que se dedica al análisis simultáneo de varias variables. La aplicación práctica del análisis multivariante implica la utilización de técnicas de análisis univariante y multivariante para entender las relaciones entre las variables y el problema a tratar.

Además, la estadística multivariante utiliza distribuciones de probabilidad multivariantes para ver cómo pueden utilizarse para representar las distribuciones de los datos y cómo pueden ser usados para la inferencia.

En este trabajo abarcaremos las técnicas de Análisis de Componentes Principales, Regresión Logística y Árboles de Clasificación que se explican a continuación.

2.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA) es una técnica estadística que consiste en la reducción de la dimensionalidad del conjunto de datos, es decir, si es posible describir con precisión la información aportada por p variables a través de un pequeño subconjunto $r < p$ de ellas. Con ello, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. Así, el PCA tiene como objetivo que si dadas n observaciones de p variables, se analiza si utilizando un número menor de variables, combinación lineal de las originales, se puede representar la información adecuadamente.

Supongamos que se dispone de una matriz $\mathbf{X}$ de dimensión $n \times p$ de rango p, donde las columnas representan las variables y las filas las observaciones. Supongamos que se han estandarizado los datos, es decir, que a cada variable se le ha restado su media y se ha dividido por su desviación típica. De esta manera, $\mathbf{X}$ tiene media cero y su matriz de correlaciones es $\mathbf{S} = \frac{1}{n}\mathbf{X}^t\mathbf{X}$.

A continuación veremos el cálculo de las componentes.

2.1.1. Cálculo de las componentes

La primera componente principal se define como la combinación de las variables originales de varianza máxima. Los valores de esta componente se representan por un vector z_1 dado por $z_1 = \mathbf{X}a_1$. Como $\mathbf{X}$ es una matriz de media nula, z_1, al ser suma de variables de media cero, también será de media cero, y la varianza será: $\frac{1}{n}z_1^t z_1 = \frac{1}{n}a_1^t \mathbf{X}^t\mathbf{X}a_1 = a_1^t \mathbf{S}a_1$ donde $\mathbf{S}$ es la matriz de correlaciones de $\mathbf{X}$.

Se puede maximizar la varianza aumentando el módulo del vector a_1. Para que la maximización de la varianza tenga solución, tenemos que imponer que $a_1^t a_1 = 1$. Introduciendo esta restricción mediante el multiplicador de Lagrange:

$$M = a_1^t \mathbf{S}a_1 - \lambda(a_1^t a_1 - 1).$$

Derivando respecto a_1 e igualando a cero tenemos

$$\frac{\partial M}{\partial a_1} = 2\mathbf{S}a_1 - 2\lambda a_1 = 0$$

con solución

$$\mathbf{S}a_1 = \lambda a_1.$$

Por tanto, la solución a_1 es el vector propio de $\mathbf{S}$ y λ es el valor propio asociado. Para ver qué valor propio es λ, multiplicamos por la izquierda por a_1^t tenemos

$$a_1^t \mathbf{S}a_1 = \lambda a_1^t a_1 = \lambda.$$

Por tanto, λ es la varianza de z_1, que, como queremos maximizar, será el mayor valor propio de $\mathbf{S}$.

La segunda componente se calcula estableciendo como función objetivo la suma de las varianzas $z_1 = \mathbf{X}a_1$ y $z_2 = \mathbf{X}a_2$, con a_1 y a_2 vectores ortonormales. El objetivo es obtener la máxima varianza, y para ello, utilizando multiplicadores de Lagrange, tenemos la siguiente función objetivo:

$$\phi = a_1^t \mathbf{S}a_1 + a_2^t \mathbf{S}a_2 - \lambda_1(a_1^t a_1 - 1) - \lambda_2(a_2^t a_2 - 1).$$

Derivando respecto a_1 y a_2 e igualando a cero:

$$\frac{\partial \phi}{\partial a_1} = 2\mathbf{S}a_1 - 2\lambda_1 a_1 = 0$$
$$\frac{\partial \phi}{\partial a_2} = 2\mathbf{S}a_2 - 2\lambda_2 a_2 = 0$$

con soluciones:

$$\mathbf{S}a_1 = \lambda_1 a_1$$
$$\mathbf{S}a_2 = \lambda_1 a_2.$$

Con lo que tenemos que a_1 y a_2 son vectores propios asociados a los valores propios λ_1 y λ_2, y, en el máximo, la función objetivo es $\phi = \lambda_1 + \lambda_2$. Por tanto, los valores propios que maximizan la función son los dos valores propios mayores de $\mathbf{S}$.

Podemos generalizar el cálculo hasta p componentes principales por ser p el rango de $\mathbf{X}$ y de $\mathbf{S}$. Los valores propios que se obtendrán serán $\lambda_1, \dots \lambda_p$, es decir, los valores propios de $\mathbf{S}$, que se calculan mediante:

$$|\mathbf{S} - \lambda \mathbf{I}| = 0$$

con vectores propios asociados:

$$(\mathbf{S} - \lambda_i \mathbf{I})\, a_i = 0, \;\; i = 1, \dots, p.$$

Los valores λ_i son reales y positivos, al ser $\mathbf{S}$ simétrica y definida positiva.

Si llamamos $\mathbf{Z}$ a la matriz de dimensión $n \times p$ referente a los valores de las nuevas componentes, esta matriz está relacionada con la original $\mathbf{X}$ mediante $\mathbf{Z} = \mathbf{XA}$, con $\mathbf{A}$ una matriz ortogonal $p \times p$ que cumple $\mathbf{A}^t\mathbf{A} = \mathbf{I}_{p\times p}$. Por tanto, calcular las componentes principales es equivalente a aplicar una transformación ortogonal $\mathbf{A}$ a las variables de $\mathbf{X}$ para obtener unas nuevas variables de $\mathbf{Z}$ que sean incorreladas entre sí.

2.1.2. Selección del número de componentes

En busca del objetivo de esta técnica, que era reducir la dimensionalidad del problema, no se toman todas las componentes calculadas, sino sólo una parte de ellas. Para saber elegir cuántas se toman, hay diferentes criterios para su selección:

1. Realizar un gráfico de los valores propios ordenados. Se seleccionan componentes hasta que en el gráfico se observe que los valores propios empiezan a ser muy similares. Es decir, se busca una variación brusca en el gráfico, a partir de la cual los valores propios son aproximadamente iguales. El objetivo es excluir las componentes asociadas a valores propios pequeños y del mismo tamaño.

2. Seleccionar componentes hasta cubrir un cierto porcentaje de la varianza. Esta regla se debe aplicar con cierto cuidado ya que una componente puede explicar una gran cantidad de la varianza, pero las restantes recojan la información necesaria para apreciar la variabilidad del conjunto de datos.

3. Seleccionar aquellas componentes que superen una cota, que suele ser la varianza media. En nuestro caso la cota tomada será 1.

2.1.3. Librerías de *R* y documentación utilizada

Los paquetes utilizados para realizar el Análisis de Componentes Principales son *RcmdrMisc* [7], *nortest* [8], *car* [9] y *psych* [10].

La referencia utilizada en esta sección ha sido el documento [11].

2.2. Modelo de Regresión Logística

Los métodos de regresión son herramientas muy útiles a la hora de estudiar la relación entre una variable y una o más variables explicativas.

En algunos casos, la variable respuesta es discreta, pudiendo tomar dos o más valores. Por ello, la regresión logística es una herramienta a utilizar en estas situaciones.

2.2.1. Introducción al modelo de Regresión Logística

En un modelo de regresión logística binario, la variable respuesta es una variable dicotómica o binaria. Supongamos que nuestro objetivo es identificar si los individuos pertenecen o no a una población. La variable respuesta toma los valores y_i:

$$y_i = \begin{cases} 1, & \text{si el individuo } i \text{ pertenece a la población} \\ 0, & \text{en otro caso} \end{cases}$$

Supongamos también el caso más sencillo en el que sólo disponemos de una variable explicativa.

El primer modelo que podemos plantear es

$$y = \beta_0 + \beta_1^t x + u$$

con u variable normal de media cero y varianza 1. Dado el valor de x_i, obtendríamos

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

con u_i variables normales, lo cual no es un modelo adecuado.

Por ello, no se utiliza el modelo lineal, sino una transformación de este, que aporte ciertas propiedades deseadas. Para transformar el modelo, se suelen utilizar funciones con *forma de S* o sigmoides. Se pueden ver ejemplos de estas funciones en la Figura 2.1.

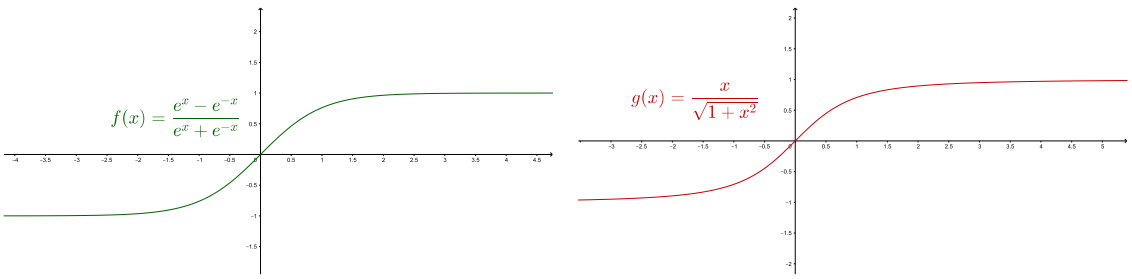

Figura 2.1: Ejemplos de funciones sigmoides.

En el modelo de regresión logística, utilizaremos la función logística

$$f(x) = \frac{1}{1+e^{-x}}$$

que tiene la gráfica que se puede ver en la Figura 2.2.

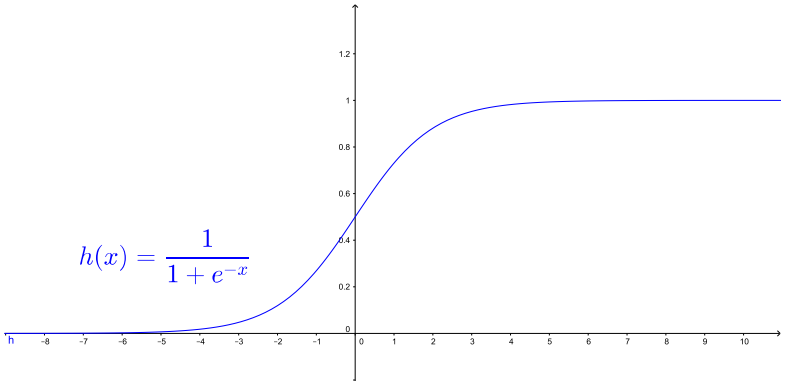

Figura 2.2: Gráfica de la función logística

Para simplificar la notación, utilizaremos $\pi(x) = E[y|x]$. Siguiendo esta notación y la transformación logística, nuestro modelo a calcular es

$$\pi(x) = \frac{1}{1+e^{-\beta_0-\beta_1^t x}}.$$

Esta función tiene diversas propiedades, entre ellas, que si tomamos logaritmos de la función entre la unidad menos la función, obtenemos el *logit* de la función

$$g(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1^t x$$

que como se ve, se asemeja al modelo planteado anteriormente.

Estimación de los parámetros

Para estimar los coeficientes no se puede utilizar el método de mínimos cuadrados ya que la variable respuesta es dicotómica. El método utilizado es el de máxima verosimilitud. Para aplicar este método, primero debemos construir la función de verosimilitud, y los valores que maximicen esta función serán los estimadores máximo verosímiles. El valor $\pi(x)$ nos proporciona una probabilidad condicionada de que $y = 1$ dado el valor de x ($P(y = 1|x)$). El valor $1 - \pi(x)$ da la probabilidad de que y tome el valor 0. Entonces, cuando $y_i = 1$, el valor aportado a la función de verosimilitud será $\pi(x_i)$. Por el contrario, cuando $y_i = 0$, el valor que aporta es $1 - \pi(x_i)$. Para el par (y_i, x_i), la contribución a la función de verosimilitud es

$$\pi(x_i)_i^y (1-\pi(x_i))^{1-y_i}.$$

Como las observaciones son independientes, la función de verosimilitud es el producto de las contribuciones de cada par, por lo que la función de verosimilitud es la siguiente

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)_i^y (1-\pi(x_i))^{1-y_i}.$$

Tomando el logaritmo de la función anterior, es más fácil estimar los parámetros β_0 y β_1. El logaritmo de la función es

$$L(\beta) = \sum_{i=1}^{n} y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i)).$$

Derivando esta expresión respecto β_0 y β_1 e igualando a cero obtenemos las ecuaciones verosímiles

$$\sum_{i=1}^{n} y_i - \pi(x_i) = 0$$

y

$$\sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0.$$

Estas ecuaciones se resuelven por métodos numéricos, como el algoritmo de *Newton-Raphson*. Una vez calculados $\hat{\beta}_0$ y $\hat{\beta}_1$, tenemos la siguiente propiedad

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$$

es decir, que la suma de los valores observados de y es igual que la suma de los valores esperados $\hat{\pi}(x)$.

Significación del modelo

Una vez estimados los coeficientes, tenemos que evaluar la significación de las variables en el modelo. Una aproximación para probar la significación de un coeficiente del modelo se basa en la relación entre lo que podemos explicar con el modelo utilizando la variable o no. Este hecho se comprueba comparando los valores observados con los estimados con ambos modelos. En el modelo de regresión logística, esta comparación se basa en el logaritmo de la función de verosimilitud. Para ello, se utilizan las predicciones con un modelo saturado, un modelo con tantos parámetros como observaciones.

La comparación de los valores observados y estimados utilizando la función de verosimilitud es

$$D = -2\log\left(\frac{\text{verosimilitud del modelo}}{\text{verosimilitud del modelo saturado}}\right)$$

que se llama desviación global.

Como la variable de salida sólo toma los valores 0 y 1, el valor de $\log(\text{verosimilitud del modelo saturado})$ es 0, por lo que la expresión anterior toma el valor

$$D = -2\sum_{i=1}^{n}(y_i \log \hat{\pi}(x_i) + (1-y_i)\log(1-\hat{\pi}(x_i)))$$

y si el modelo es correcto, sigue una distribución χ^2 con $n-1$ grados de libertad.

Otro estadístico utilizado es el test de Wald, que compara el estimador $\hat{\beta}_1$ con el estimador del error estándar. Bajo la hipótesis $H_0 : \beta_1 = 0$, el siguiente ratio se distribuye como una variable normal estándar

$$W = \frac{\hat{\beta}_1}{\hat{SE}\left(\hat{\beta}_1\right)}.$$

2.2.2. Modelo de Regresión Logística Múltiple

Consideramos ahora un conjunto de p variables independientes denotadas como $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicionada se denota $P(y = 1|\mathbf{x}) = \pi(\mathbf{x})$. El *logit* del modelo múltiple es

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

y el modelo de regresión queda

$$\pi(\mathbf{x}) = \frac{1}{1 + e^{-g(\mathbf{x})}}.$$

Estimación de los parámetros

Para obtener los parámetros $\beta^t = (\beta_0, \beta_1, \ldots, \beta_n)$ utilizaremos el mismo método que en el modelo univariante, el método de máxima verosimilitud. Utilizaremos la misma función de verosimilitud que en el caso univariante, pero tomaremos $\pi(\mathbf{x})$ como acabamos de definir. Ahora hay $p+1$ ecuaciones para determinar $p+1$ coeficientes. Las ecuaciones son las siguientes:

$$\sum_{i=1}^{n} y_i - \pi(\mathbf{x}_i) = 0$$

y

$$\sum_{i=1}^{n} x_{ij}(y_i - \pi(\mathbf{x}_i)) = 0$$

para $j = 1, 2, \ldots, p$, que utilizaremos para calcular los coeficientes con métodos numéricos.

Supongamos que los estimadores calculados son $\hat{\beta}$ y las probabilidades condicionadas son $\pi(\mathbf{x}_i)$. Para estimar las varianzas y covarianzas de los estimadores se utilizan las derivadas segundas del logaritmo de la función de verosimilitud. Las derivadas segundas son:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i)$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i)$$

para $j, l = 1, 2, \ldots, n$, donde $\pi_i = \pi(\mathbf{x}_i)$.

Consideramos $\mathbf{I}(\beta)$ la matriz $(p+1) \times (p+1)$ de los términos de las ecuaciones anteriores. Las varianzas y covarianzas se obtienen de la inversa de esta matriz, $Var(\beta) = \mathbf{I}^{-1}(\beta)$. No siempre es posible escribir una expresión para esta matriz, pero utilizaremos la notación $Var(\beta_j)$ para los elementos diagonales, que es la varianza de $\hat{\beta}_j$, y $Cov(\beta_j, \beta_l)$ para los elementos no diagonales, que es la covarianza de $\hat{\beta}_j$ y $\hat{\beta}_l$. Los estimadores de las varianzas y covarianzas se denotan por $\hat{Var}(\hat{\beta})$ y se obtienen evaluando $Var(\beta)$ en $\hat{\beta}$. Utilizaremos $\hat{Var}(\hat{\beta})$ y $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_l)$, $j, l = 1, 2, \ldots, p$ para denotar los valores de esta matriz. A partir de estos valores podemos calcular los errores estandarizados de los estimadores que calcularemos así

$$\hat{SE}(\hat{\beta}_j) = \left(\hat{Var}(\hat{\beta}_j)\right)^{\frac{1}{2}}$$

para $j = 0, 1, \ldots, p$.

Una formulación de la matriz de información es $\hat{\mathbf{I}}(\hat{\beta}) = \mathbf{X}^t\mathbf{V}\mathbf{X}$, donde $\mathbf{X}$ es la matriz

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

de dimensión $n \times (p+1)$ y la matriz $\mathbf{V}$ es

$$\mathbf{V} = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}.$$

Significación del modelo

De manera análoga a la regresión con una variable, podemos calcular el test de Wald para cada variable es

$$W_j = \frac{\hat{\beta}_j}{\hat{SE}\left(\hat{\beta}_j\right)}$$

para $j = 1,2,\ldots,p$ sigue una distribución normal estándar.

El test de Wald para varias variables simultáneamente se obtiene del siguiente cálculo matricial

$$W = \hat{\beta}^t \left(\hat{Var}\left(\hat{\beta}\right)\right)^{-1} \hat{\beta} = \hat{\beta}^t \left(\mathbf{X}^t\mathbf{V}\mathbf{X}\right) \hat{\beta}$$

que se distribuye como una χ^2 de $(p+1)$ grados de libertad bajo la hipótesis de que todos los coeficientes son cero.

Interpretación de los coeficientes

Para interpretar los coeficientes se utiliza el concepto de *odds*-ratio. Este ratio mide el efecto de una variable por unidad de cambio. Supongamos que tenemos el caso $\mathbf{x}_i = (x_{i1},\ldots,x_{ij},\ldots,x_{ip})$ y denotemos como $\mathbf{x}_i + 1$ a $(x_{i1},\ldots,x_{ij}+1,\ldots,x_{ip})$. El *odds*-ratio de la variable j es el siguiente

$$OR_j = \frac{\frac{P(Y=1|\mathbf{x}_i+1)}{(1-P(Y=1|\mathbf{x}_i+1))}}{\frac{P(Y=1|\mathbf{x}_i)}{(1-P(Y=1|\mathbf{x}_i))}}.$$

Este ratio toma valores en $(0,\infty)$. Una medida cercana a 1 indica poca relación entre la variable explicativa y la variable respuesta. Si toma valores mayores que uno, se ve fácilmente, pero si es menor que uno, basta con invertir el número para ver mejor la relación.

Por ejemplo, supongamos que el *OR* de una variable es 1,57. Esto significa que aumentando en una unidad la variable, hay un 57 % más de probabilidad de que $Y = 1$.

2.2.3. Modelo de Regresión Logística Multinomial

Hemos visto los modelos cuando la variable respuesta tomaba valores 0 y 1. Supongamos ahora que la variable respuesta toma más de dos valores. Supongamos que toma 3 valores para simplificar la notación, y supongamos que estos valores son 0, 1 y 2. Para el modelo, necesitaremos 2 funciones logísticas, y decidir qué categoría utilizaremos para comparar. En este caso, utilizaremos el valor 0 como referencia.

Las variables las representaremos con un vector de coordenadas $\mathbf{x}$ de $p+1$ coordenadas fijando la primera coordenada a 1, $\mathbf{x}^t = (1, x_1, x_2, \ldots, x_p)$. Las funciones son las siguientes:

$$g_1(\mathbf{x}) = \log\left(\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p = \mathbf{x}^t\beta_1$$

y

$$g_2(\mathbf{x}) = \log\left(\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p = \mathbf{x}^t\beta_2.$$

Las fórmulas para las probabilidades condicionadas son

$$P(Y=0|\mathbf{x}) = \frac{1}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}},$$

$$P(Y=1|\mathbf{x}) = \frac{e^{g_1(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}}$$

y

$$P(Y=2|\mathbf{x}) = \frac{e^{g_2(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}+e^{g_2(\mathbf{x})}}.$$

Tomando la dinámica del modelo, denotaremos como $\pi_j(\mathbf{x}) = P(Y=j|\mathbf{x}),\ j=0,1,2$. Cada probabilidad es una función de $2(p+1)$ parámetros $\beta^t = (\beta_1^t, \beta_2^t)$.

Estimación de los parámetros

Para construir la función de verosimilitud, se construyen tres variables binarias $Y_j,\ j=0,1,2$ donde $Y_j = 1$ si $Y = j$ e $Y_k = 0,\ k \neq j$. Con la codificación vista, se tiene que $\sum_{j=0}^{2} Y_j = 1$. Usando la notación, para una muestra de n observaciones, la función de verosimilitud es

$$l(\beta) = \prod_{i=1}^{n} \pi_0(\mathbf{x}_i)^{y_{0i}} \pi_1(\mathbf{x}_i)^{y_{1i}} \pi_2(\mathbf{x}_i)^{y_{2i}}.$$

Tomando el logaritmo y el hecho de que $\sum_{j=0}^{2} y_{ji} = 1,\ i = 1,2,\ldots,n$, el logaritmo de la función de verosimilitud es

$$L(\beta) = \sum_{i=1}^{n} y_{1i}g_1(\mathbf{x}_i) + y_{2i}g_2(\mathbf{x}_i) - \log\left(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}\right).$$

Para hallar los parámetros, hay que calcular las derivadas parciales del logit de la función respecto a los $2(p+1)$ parámetros. Tomando como $\pi_{ji} = \pi_j(\mathbf{x}_i)$ se tiene que las derivadas parciales son de la forma

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - \pi_{ji})$$

para $j = 1,2$ y $k = 0,1,2,\ldots,p$, con $x_{0i} = 1$ para cada caso.

Igualando a cero y resolviendo para β, hallamos el valor de los estimadores $\hat{\beta}$ con métodos numéricos.

La matriz de información y el estimador de la matriz de covarianzas se calculan con las segundas derivadas de la función logit. Las fórmulas son las siguientes

$$\frac{\delta^2 L(\beta)}{\partial\beta_{jk}\partial\beta_{jl}} = -\sum_{i=1}^{n} x_{li}x_{ki}\pi_{ji}(1-\pi_{ji})$$

y

$$\frac{\partial^2 L(\beta)}{\partial\beta_{jk}\partial\beta_{hl}} = \sum_{i=1}^{n} x_{hl}x_{ki}\pi_{ji}\pi hl$$

para $j,h = 1,2$ y $k,l = 0,1,2,\ldots,p$. La matriz de información $\mathbf{I}(\hat{\beta})$ es la matriz con el valor de las ecuaciones anteriores evaluada en $\hat{\beta}$. El estimador de la matriz de covarianzas es la inversa de la matriz de información, $\hat{Var}(\hat{\beta}) = \mathbf{I}(\hat{\beta})^{-1}$.

Otra manera de calcular la matriz de información es similar al modelo logístico binario. Sea $\mathbf{X}$ la matriz $n \times (p+1)$ con los valores de la covarianza para cada caso. Sea $\mathbf{V}_j$ la matriz diagonal $n \times n$ con los valores $\hat{\pi_{ji}}(1-\hat{\pi_{ji}})$ para $j = 1,2$ e $i = 1,2,\ldots,n$ y sea $\mathbf{V}_3$ la matriz diagonal $n \times n$ con los elementos $\hat{\pi_{1i}}\hat{\pi_{2i}}$, el estimador de la amtriz de información es

$$\mathbf{I}(\hat{\beta}) = \begin{pmatrix} \hat{\mathbf{I}}(\hat{\beta})_{11} & \hat{\mathbf{I}}(\hat{\beta})_{12} \\ \hat{\mathbf{I}}(\hat{\beta})_{21} & \hat{\mathbf{I}}(\hat{\beta})_{22} \end{pmatrix}$$

donde

$$\hat{\mathbf{I}}(\hat{\beta})_{11} = \left(\mathbf{X}^t\mathbf{V}_1\mathbf{X}\right)$$
$$\hat{\mathbf{I}}(\hat{\beta})_{22} = \left(\mathbf{X}^t\mathbf{V}_2\mathbf{X}\right)$$

y

$$\hat{\mathbf{I}}(\hat{\beta})_{12} = \hat{\mathbf{I}}(\hat{\beta})_{21} = -\left(\mathbf{X}^t\mathbf{V}_3\mathbf{X}\right).$$

Significación del modelo

De manera análoga a la regresión logística utilizando una variable, se puede calcular el test de Wald, con la única diferencia de que por cada variable, habrá dos coeficientes, y con ello, dos test de Wald. El test es el siguiente

$$W_{kj} = \frac{\hat{\beta}_{kj}}{\hat{SE}\left(\hat{\beta}_{kj}\right)},$$

para $k = 1,2$ y $j = 1,2,\ldots,p$, que sigue una distribución normal estándar.

Interpretación de los coeficientes

De igual manera que antes, podemos calcular los *odds*-ratios de las variables. Ahora la variable respuesta tiene más de dos categorías, y por ello hay que realizar dos *odds*-ratios, cada uno comparando la categoría de referencia con cada categoría. Los *odds*-ratios para las variable j para la categoría

$Y = 1$ es

$$OR_{j1} = \frac{\frac{P(Y=1|\mathbf{x}_i+1)}{P(Y=0|\mathbf{x}_i+1)}}{\frac{P(Y=1|\mathbf{x}_i)}{P(Y=0|\mathbf{x}_i)}}$$

y para la categoría $Y = 2$ es

$$OR_{j2} = \frac{\frac{P(Y=2|\mathbf{x}_i+1)}{P(Y=0|\mathbf{x}_i+1)}}{\frac{P(Y=2|\mathbf{x}_i)}{P(Y=0|\mathbf{x}_i)}}$$

que tienen la misma interpretación que antes.

2.2.4. Librerías de *R* y documentación utilizada

Los librerías utilizadas para realizar las Regresiones Logísticas son *RcmdrMisc* [7], *nnet* y *MASS* [12].

Las referencias utilizadas para la elaboración de esta sección han sido los documentos [11, 13, 14].

2.3. Árboles de Clasificación

Los árboles de clasificación son uno de los dos principales tipos de aprendizaje basado en árboles de decisión. Un árbol de decisión es un modelo de predicción que se basa en nodos y reglas. Los nodos son subconjuntos del conjunto de elementos de estudio y las reglas son reglas lógicas binarias que dividen cada nodo en dos subnodos disjuntos hasta llegar a nodos terminales que no se pueden dividir. El objetivo del método es clasificar los individuos de estudio en J clases.

La construcción del método se basa en los siguientes elementos:

- Un conjunto de preguntas binarias del tipo ¿$a \in \mathbf{A}$?, $\mathbf{A} \subset \chi$, con el que formularemos los *splits*, que son las reglas que se utilizarán para dividir el árbol.
- Una función de medida de la bondad del *split*, $\Phi(s,t)$ que se puede evaluar para cualquier *split* s y nodo t.
- El criterio para decidir cuando se para o se sigue dividiendo el árbol.
- La regla de asignación de cada nodo terminal a una clase.

Para construir los árboles, supongamos que tenemos un conjunto de tamaño n con M variables, y supongamos que cada caso tiene una estructura de vector $\mathbf{X} = (x_1, x_2, \ldots, x_M)$. Supondremos que la variable respuesta tiene J clases diferentes.

2.3.1. Conjunto de preguntas estándar

Para poder realizar *splits*, al conjunto de datos, se le formula una serie de preguntas. Cada *split* que se forma depende únicamente del valor de una variable que responde a una pregunta. Si la variable x_m es numérica, las preguntas son del tipo ¿$x_m \leq c$?, donde $c \in (-\infty, \infty)$. Por el contrario, si la variable

es categórica, tomando valores $\{c_1, c_2, \ldots, c_N\}$, las preguntas son del tipo ¿$x_m \in S$?, donde S es un subconjunto de todos los valores $\{c_1, c_2, \ldots, c_N\}$ que puede tomar.

De esta manera se genera el conjunto $\mathscr{S}$ de todos los *splits* posibles que dividan un nodo t en dos subnodos t_L y t_R para poder empezar a generar el árbol.

2.3.2. Reglas de división y parada del algoritmo

El primer paso para empezar a dividir el árbol es establecer una función de bondad del *split*. Para ello, es necesario definir antes qué es una función de impureza sobre un nodo.

Una función de impureza es una función ϕ definida sobre el conjunto de todas las $J-$tuplas, $(p_1, p_2, \ldots, p_J)$ que cumplen que $0 \leq p_j \leq 1$ y que $\sum_{j=1}^{J} p_j = 1$, que representan las proporciones de cada clase en el nodo. La función ϕ tiene las siguientes propiedades:

1. ϕ tiene máximo solamente en $\left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right)$.
2. ϕ tiene mínimo solamente en $(1,0,\ldots,0)$, $(0,1,\ldots,0)$, …, $(0,0,\ldots,1)$.
3. ϕ es una función simétrica respecto $p_1, \ldots, p_J$.

Una vez definida una función de impureza ϕ, podemos definir una función de medida de la impureza $i(t)$ del nodo t como $i(t) = \phi(p(1|t), p(2|t), \ldots, p(J|t))$, donde $p(j|t)$ es la proporción de la clase j en el nodo t.

Si un *split* s del nodo t envía la proporción p_L al nodo t_L y la proporción p_R al nodo t_R, entonces podemos definir el decrecimiento de la impureza como $\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$. Se define la bondad del *split* s en el nodo t como $\Phi(s,t) = \Delta(s,t)$.

Supongamos que hemos realizado varios *splits*, el conjunto de los *splits* y el orden en que se han realizado determinan el árbol T. Denotaremos los nodos terminales de este árbol como $\mathbb{T}$. Definimos la impureza del nodo t del árbol T como $I(t) = i(t)p(t)$, donde $i(t)$ es la impureza definida anteriormente y $p(t)$ la proporción del nodo t respecto el árbol T, y definimos la impureza del árbol $I(T)$ como

$$I(T) = \sum_{t \in \mathbb{T}} I(t) = \sum_{t \in \mathbb{T}} i(t)p(t).$$

Es fácil ver que seleccionar los *splits* que maximicen $\Delta i(s,t)$ es equivalente a seleccionar los *splits* que minimicen la impureza $I(T)$. Tomando un nodo $t \in \mathbb{T}$ y usando un *split* s, separamos el nodo t en t_L y t_R. El nuevo árbol T' tiene una impureza $I(T') = \sum_{t' \in \mathbb{T} - \{t\}} I(t') + I(t_L) + I(t_R)$.

El decrecimiento de la impureza es $I(T) - I(T') = I(t) - I(t_L) - I(t_R)$, que depende únicamente del nodo t y el *split* s. Entonces, el decrecimiento de la impureza del árbol por *splits* en t es equivalente a maximizar la expresión $\Delta I(s,t) = I(t) - I(t_L) - I(t_R)$.

Definiendo, $p_L = \frac{p(t_L)}{p(t)}$ y $p_R = \frac{p(t_R)}{p(t)}$, se tiene que $p_L + p_R = 1$ y la expresión anterior de $\Delta I(s,t)$ se puede reescribir como

$$\Delta I(s,t) = (i(t) - p_L i(t_L) - p_R i(t_R))p(t) = \Delta i(s,t)p(t).$$

Como se puede ver, $\Delta I(s,t)$ y $\Delta i(s,t)$ difieren sólo en el factor $p(t)$. Por tanto, el mismo *split* $s*$ maximiza ambas expresiones, por lo que la selección del *split* idóneo que minimice la impureza de un nodo puede enfocarse como la minimización de la impureza del árbol.

Para establecer el criterio de parada, se fija un umbral β, y un nodo t se declara terminal si

$$\max_{s\in\mathscr{S}} \Delta I(s,t) < \beta.$$

2.3.3. Asignación de cada clase

Supongamos que tenemos un árbol T con nodos terminales $\mathbb{T}$.

Una regla de asignación es una regla que asigna una clase $j \in \{1,2,\ldots,J\}$ a cada nodo terminal $t \in \mathbb{T}$. La clase asignada al nodo t se designa como $j(t)$.

Para cualquier regla de asignación $j(t)$ en el nodo $t \in \mathbb{T}$, tenemos que $\sum_{j\neq j(t)} p(j|t)$ es una estimación de la probabilidad de no acertar la predicción en el nodo t. El objetivo es obtener una regla que minimice esta estimación.

Consideremos $j^*(t)$ la regla de asignación que minimiza la estimación anterior. Esta regla está definida como: si $p(j|t) = \max_i p(i|t)$, entonces $j^*(t) = j$. Si el máximo se obtiene en varias clases, se asigna el valor de $j^*(t)$ arbitrariamente.

Utilizando esta regla, tenemos que la estimación de la probabilidad de no acertar en la asignación, $r(t)$, está dada por $r(t) = 1 - \max_j p(j|t)$.

Si denotamos por $R(t) = r(t)p(t)$, podemos estimar la probabilidad de no acertar en la predicción $R^*(T)$ del árbol T como $R^*(T) = \sum_{t\in\mathbb{T}} R(t)$.

2.3.4. Valores *missing*

En los conjuntos de datos, usualmente nos enfrentamos a datos ausentes. En estos casos, se podrían eliminar del conjunto, pero se podría perder mucha información del conjunto de datos, por lo que se mantienen y se procede a manejar esta información de otra manera.

Primero, se dividie el nodo en dos subnodos y se calculan las impurezas de ambos nodos sin tener en cuenta los datos ausentes. Una vez calculados, se ajustan las proporciones p_L y p_R para que puedan sumar 1 ambas proporciones a la hora de calcular y maximizar $\Delta I(s,t)$.

Una vez se ha elegido el *split*, queda elegir qué se hace con las variables con datos ausentes que no han podido ser seleccionados en algún nodo.

Con todos los casos con datos ausentes, se toma el resto de variables y se utilizan para realizar el *split*. Se ordenan conforme a la predictividad del *split* que realizan y se utiliza la que más predictividad tiene. Si la variable con más predicitividad no está informada, se utilizaría la siguiente. A estos *splits* se les llama *surrogate splits*.

2.3.5. Podado del árbol

Una vez que hemos construido un árbol, en ocasiones suele ser o muy largo o muy complejo, entonces toca decidir cuánto del modelo queremos mantener. De esta manera tomaremos un subconjunto de nodos del árbol que formarán un subárbol.

Sea T un árbol no trivial calculado. Dado un número real $\alpha \in [0,\infty)$, sea $R_\alpha(t) = R(t) + \alpha$, con $t \in T_0$. Dado un subárbol T' de T, se define $R_\alpha(T) = R(T) + \alpha|\mathbb{T}|$ como el coste del árbol, y se define como T_α el subárbol del modelo completo con mínimo coste.

$R(T)$ se interpreta como el coste de añadir otra variable al modelo, y a α se le llama complejidad. Obviamente, T_0 es el modelo completo y T_∞ es el modelo con un nodo y ningún *split*. Se tienen los siguientes resultados:

1. Si T_1 y T_2 son subárboles de T con $R_\alpha(T_1) = R_\alpha(T_2)$, entonces o bien $T_1 < T_2$ o bien $T_2 < T_1$.
2. Si $\alpha > \beta$, entonces $T_\alpha = T_\beta$ o $T_\alpha < T_\beta$.

Utilizando el primer resultado, podemos definir como T_α como el subárbol más pequeño que minimiza $R_\alpha(T)$. Como cualquier árbol anidado basado en T tiene como mucho $|T|$ nodos, y utilizando el segundo resultado, podemos agrupar todos los valores de α en $m \leq |T|$ intervalos $I_1 = [0,\alpha_1]$, $I_2 = (\alpha_1,\alpha_2], \ldots, I_m = (\alpha_m,\infty]$, donde cada I_i comparte el mismo subárbol minimizado.

Para tomar el valor α óptimo, se utiliza el algoritmo de validación cruzada que consta de los siguientes pasos:

1. Se calculan los intervalos I_1, I_2 $\ldots$, I_m.

   Se fija $\beta_1 = 0$
   $\beta_2 = \sqrt{\alpha_1\alpha_2}$
   $\beta_3 = \sqrt{\alpha_2\alpha_3}$
   $\vdots$
   $\beta_{m-1} = \sqrt{\alpha_{m-2}\alpha_{m-1}}$
   $\beta_m = \infty$.

2. Se divide el conjunto en s grupos $G_1,\ G_2,\ldots,\ G_s$ equitativamente y se procede de la siguiente manera para cada grupo:
   - Se calcula el modelo para todo el conjunto de datos sin los datos del grupo G_i y se determinan los subárboles $T_{\beta_1},\ T_{\beta_2},\ldots,T_{\beta_m}$.
   - Se calcula la predicción para cada elemento de G_i bajo los subárboles T_{β_j}, con $1 \leq j \leq m$.
   - Se calcula el error de cada elemento de G_i.
3. Se suma el error cometido en cada G_i para cada β_j. Para cada β se toma el conjunto con menos error y se calcula el subárbol T_β para el conjunto de datos completo, con lo que se tiene un valor de complejidad β y el mejor subárbol podado para este valor.

En la práctica, un valor de $s = 10$ es suficiente para el algoritmo y la eficiencia de este.

2.3.6. Librerías de *R* y documentación utilizada

Las referencias utilizadas en esta sección han sido los documentos [11, 15, 16, 17, 18].

En la práctica, las librerías utilizadas de *R* para realizar Árboles de Clasificación son *rattle* [19] y *rpart* [20].

Para utilizar esta función, se introducen una serie de argumentos de control del árbol. Hay tres tipos de argumentos, por: control de crecimiento, control de salida y control del algoritmo interno. Los argumentos que se utilizan para controlar el crecimiento del árbol son *minbucket*, *minsplit*, *maxdepth* y *cp*. Los argumentos que controlan la salida son *maxcompete* y *maxsurrogate*. Los argumentos de control del algoritmo interno son *usesurrogate*, *xval* y *surrogatestyle*. El significado de cada argumento se encuentra en la siguiente lista:

- minbucket: número mínimo de observaciones en un nodo terminal.
- minsplit: el número mínimo de observaciones que debe existir en un nodo de cara a intentar un *split*.
- maxdepth: máximo de profundidad del árbol final.
- maxcompete: el número de *splits* competidores retenidos en la salida.
- cp: parámetro de complejidad.
- usesurrogate: cómo se usan los *surrogate splits*. Con un valor de 0, los registros con un valor ausente no se sigue utilizando; con un valor de 1, se utilizarán *surrogate splits*, y si las variables implicadas tienen algún dato faltante, se elimina; y con un valor de 2, se utilizarán *surrogate splits*, y si un registro tiene todos los valores de las variables de los *surrogate splits* están ausentes, se envían a la mayoría.
- maxsurrogate: el número de *surrogate splits* retenidos en la salida.
- xval: número de validaciones cruzadas.
- surrogatestyle: cómo se selecciona el mejor *surrogate split*. Si se evalúa como 0, el programa usa el total de clasificaciones correctas para variables *surrogate* potenciales. Si se evalúa como 1, se utiliza el porcentaje exacto, calculado con los valores no ausentes de las variables *surrogate*.

Una vez desarrollado el algoritmo, *R* proporciona una salida que consiste en primer lugar en el error del nodo principal, las variables utilizadas en los *splits* del árbol generado y una tabla donde se indican los distintos árboles obtenidos, cada uno más ramificado que el anterior. La tabla proporciona el parámetro de complejidad utilizado, el número de *splits*, el error relativo del árbol y los errores medios y desviación estándar de los árboles utilizados a la hora de realizar los diferentes árboles en la validación cruzada. El objetivo es obtener un árbol con un error medio lo más bajo posible.

Capítulo 3

Análisis por conjuntos de variables

3.1. Introducción

En este capítulo vamos a realizar un análisis preliminar de los datos. Se aplicarán técnicas de estadística descriptiva y las técnicas mencionadas en la metodología para poder seleccionar variables que tengan una mayor influencia en el APOE.

Primero se va a realizar un primer análisis exploratorio de los datos, utilizando técnicas de estadística descriptiva. Después se realizará un análisis de componentes principales a las variables de los conjuntos de datos para ver posibles agrupaciones de variables. Finalmente se construirán modelos de regresión multinomial y árboles de clasificación para determinar las variables más influyentes en la determinación del fenotipo del gen APOE.

3.2. Análisis exploratorio

En este primer análisis exploratorio, se va a realizar un análisis descriptivo numérico y gráfico de las variables de los conjuntos para poder ver si alguna de ellas puede ser una candidata para separar correctamente la variable respuesta. Tras este análisis, se realizará el test de Kruskal-Wallis para poder confirmar si estas variables obtenidas en el primer análisis pueden separar bien por grupos la variable respuesta. Finalmente se calculan las correlaciones para el conjunto de datos de Ácidos grasos con el fin de ver si alguna de las parejas de variables está correlada.

3.2.1. Análisis descriptivo numérico y gráfico

En esta sección vamos a realizar un primer análisis descriptivo numérico y gráfico. El objetivo del análisis es concluir qué variables pueden clasificar a priori de manera correcta la variable respuesta, *apoE2*, que es el fenotipo del gen APOE. Para realizar el análisis, primero se obtendrá la distribución de la variable *apoE2* y luego se realizarán los análisis descriptivos y gráficos del resto de variables.

La distribución de la variable puede verse en la Tabla 3.1. De esta se destaca que hay un grupo que predomina sobre los demás, el grupo *E3*, puesto que el 67,77 % de los registros tienen este valor, distribución que se asemeja a la dada por la población española.

Valor	Número	Porcentaje
E2	28	11,57 %
E3	164	67,77 %
E4	35	14,46 %
NA's	15	6,20 %

Tabla 3.1: Distribución de la variable *apoE2*.

El análisis numérico consistirá en calcular el porcentaje de casos ausentes, media, mediana y desviación típica de cada variable por grupos de la variable *apoE2*. También se mirarán los cuartiles de cada variable y algunos percentiles representativos. Las tablas con los estadísticos descriptivos pueden consultarse en el Anexo A.

Del primer análisis numérico, comparando los percentiles de las variables, las variables que parecen distinguir de algún modo los grupos de la variable respuesta son las siguientes: del conjunto de Ácidos grasos las variables *g183H*, *v120H*, *v204P* y *v226H*; del conjunto Aminoácidos las variables *AcAlfaAminobutirico*, *Alanina* y *Prolina*; del conjunto de Bioquímicas las variables *APOB*, *CHOL*, *CTx*, *LDL*, *LPA* y *PCRU*; y del conjunto de Carnitinas las variables *C4DCn* y *C18n*.

Tras este primer análisis descriptivo de las variables, se puede concluir que no se pueden apreciar muchas diferencias por grupos en la mayoría de las variables. También se puede observar que hay un gran porcentaje de datos faltantes en el conjunto de Aminoácidos, ya que todas las variables tienen un porcentaje superior al 33 %.

Respecto al análisis gráfico de las variables, se han obtenido los diagramas de caja y estimado las funciones de densidad de todas las variables por grupo de *apoE2*, con el objetivo de ver si alguna variable presenta diferencias muy claras respecto a la variable respuesta. Para la estimación de las densidades se ha fijado el parámetro de suavizado eligiendo entre los valores que proporciona el método de Silverman [21] y seleccionado tras la inspección uno a uno de todos los gráficos. El método de Silverman obtiene el valor de ajuste de la siguiente manera. Se toma el mínimo entre la desviación típica y el rango intercuartílico dividido entre 1,349 y se multiplica por $1{,}06 \times n^{-1/5}$. En el método práctico, se analizó una a una las variables y se anotó el valor de ajuste que parecía correcto. El valor de ajuste final que se ha tomado ha sido el máximo de estos dos valores.

Los diagramas de caja y funciones de densidad pueden consultarse en el Anexo B.

Comparando los grupos de las variables, la distribución de ambas tanto en los diagramas de caja como en las estimaciones de densidad, pueden apreciarse diferencias en algunas en cuanto a la variable *apoE2*. Tras estas observaciones, las variables que a priori presentan distribuciones diferentes son las siguientes: del conjunto de Ácidos grasos las variables *v201P*, *v224P*, *v160H*, *g183H*, *v201H*, *v221H* y *v226H*; del conjunto de Aminoácidos las variables *Fosfoserina*, *Alanina*, *Valina*, *Cistina*, *Arginina* y *Prolina*; del conjunto de Bioquímicas las variables *CHOL*, *LDL* y *APOB*; y del conjunto de Carnitinas las variables *C3n*, *C18On* y *C182n*.

De este primer análisis se puede concluir que no hay mucha evidencia de que haya variables que por sí solas puedan clasificar correctamente la variable *apoE2*, ya que ninguna de las herramientas utilizadas revela que existan grandes diferencias en el comportamiento según los grupos definidos por dicha variable.

El resumen de las variables que en principio presentan algunas diferencias según la variable *apoE2* puede verse en la Tabla 3.2 donde se aprecia que algunas diferencias detectadas en el análisis numérico no son detectadas en el gráfico.

Conjunto de datos	Análisis numérico	Análisis gráfico
Ácidos grasos	*v120H*, *v204P*	-
	-	*v201P*, *v224P*, *v160H*, *v201H*, *v221H*
	g183H, *v226H*	g183H, *v226H*
Aminoácidos	*AcAlfaAminobutirico*	-
	-	*Fosfoserina*, *Valina*, *Cistina*, *Arginina*
	Alanina, *Prolina*	*Alanina*, *Prolina*
Bioquímicas	*CTx*, *LPA*, *PCRU*	-
	APOB, *CHOL*, *LDL*	*APOB*, *CHOL*, *LDL*
Carnitinas	*C4DCn*, *C18n*	-
	-	*C3n*, *C12n*, *C14n*

Tabla 3.2: Resumen de los resultados obtenidos en el análisis descriptivo.

3.2.2. Test de Kruskal-Wallis

En esta sección, se va a realizar el test de Kruskal-Wallis para todas las variables de los conjuntos, con el fin de confirmar qué variables separan correctamente los grupos de la variable *apoE2*.

El test de Kruskal-Wallis [22] es un método no parámetrico que se utiliza para determinar si varias muestras proceden de la misma distribución. El test es usado para comparar dos o más muestras independientes del mismo o diferente tamaño. Es el equivalente no paramétrico del test ANOVA. El test, al ser no paramétrico, no asume una distribución normal de los datos.

Las hipótesis del test son las siguientes:

H_0: Las medianas de los grupos son iguales.
H_1: Al menos la mediana de un grupo es diferente que las del resto de la población.

Los resultados del test para los conjuntos de Ácidos grasos, Aminoácidos, Bioquímicas y Carnitinas pueden verse respectivamente en las Tablas 3.3, 3.4, 3.5 y 3.6.

Las variables que pasan el test con un nivel de significación de $\alpha = 0{,}05$ son *v201P*, *v224P*, *g183H*, *v201H* y *v260H* del conjunto de Ácidos grasos; del conjunto de Aminoácidos la variable *Alanina*; las variables *APOB*, *CHOL*, *CTx* y *VitD* del conjunto de Bioquímicas; y las variables *C6n* y *C16On* del conjunto de Carnitinas.

Con los resultados obtenidos en el test, y comparando con los resultados vistos en los análisis descriptivos anteriores, se puede confirmar que las variables *v201P*, *v224P*, *g183H*, *v201H*, *Alanina*, *APOB*, *CHOL* y *CTx* tienen un comportamiento algo diferente según los grupos de la variable *apoE2*, ya que también hemos obtenido estas variables en el análisis descriptivo numérico o gráfico, o incluso en ambos análisis, por lo que habrá que tener especial consideración con ellas.

La comparación de los análisis descriptivos y los resultados del test puede ser consultada en la Tabla 3.7.

Variable	***p*-valor**	**Variable**	***p*-valor**
v120P	–	v120H	0.814791
v140P	0.703973	v140H	0.864251
v160P	0.170677	v160H	0.261506
v161P	0.467986	v161H	0.504649
v180P	0.761026	v180H	0.649898
v181P	0.697544	v181H	0.749776
v182P	0.162495	v182H	0.391211
g183P	0.494311	**g183H**	**0.000543**
a183P	0.059522	a183H	0.699045
v184P	0.257150	v184H	0.161171
v200P	0.614300	v200H	0.488016
v201P	**0.049426**	**v201H**	**0.049816**
v202P	0.174756	v202H	0.555406
v203P	0.277309	v203H	0.535634
v204P	0.271565	v204H	0.602514
v205P	0.654558	v205H	0.326922
v220P	0.562291	v220H	0.740213
v221P	0.904376	v221H	0.057408
v224P	**0.035224**	v224H	0.628998
v225P	0.389523	v225H	0.101893
v240P	0.670263	v240H	0.344139
v226P	0.244068	v226H	0.067595
v241P	0.713611	v241H	0.776477
v260P	0.852232	**v260H**	**0.038689**

Tabla 3.3: Resultados del test de Kruskal-Wallis para las variables del conjunto de Ácidos grasos. En negrita se resaltan los p-valores menores que 0,05.

Variable	*p*-valor	Variable	*p*-valor
Fosfoserina	0.206401	Cistina	0.165002
Taurina	0.819239	Metionina	0.576901
AcAspartico	0.947797	Isoleucina	0.656071
Treonina	0.897992	Leucina	0.276179
Serina	0.918512	Tirosina	0.971120
Asparragina	0.971408	FenilAlanina	0.369918
AcGlutamico	0.550806	Ornitina	0.864951
Glutamina	0.743152	Lisina	0.261980
Glicina	0.211860	X1MetilHistidina	0.780686
Alanina	**0.033018**	Histidina	0.419969
Citrulina	0.348460	Triptofano	0.563330
AcAlfaAminobutirico	0.123647	Arginina	0.329902
Valina	0.096233	Prolina	0.164005

Tabla 3.4: Resultados del test de Kruskal-Wallis para las variables del conjunto de Aminoácidos. En negrita se resaltan los *p*-valores menores que 0,05.

Variable	*p*-valor	Variable	*p*-valor	Variable	*p*-valor
PCRU	0.067141	GPT	0.168154	NEFA	0.860839
GLU	0.064972	INS	0.593915	HEMG	0.176337
TRIG	0.588003	LEPT	0.554863	HEMGm	0.190281
CHOL	**0.020317**	APOA	0.983308	SE	0.274113
cHDL	0.968944	**APOB**	**0.004786**	BGP	0.943245
LDL	0.069963	LPA	0.252937	**CTx**	**0.032750**
GGT	0.884129	BHID	0.979577	**VitD**	**0.038101**
GOT	0.539416				

Tabla 3.5: Resultados del test de Kruskal-Wallis para las variables del conjunto de Bioquímicas. En negrita se resaltan los *p*-valores menores que 0,05.

Variable	*p*-valor	Variable	*p*-valor	Variable	*p*-valor
C0n	0.950051	C8n	0.680352	C142n	0.429966
C2n	0.954504	C81n	0.335594	C16n	0.483813
C3n	0.690698	C10n	0.644037	C161n	0.810456
C3DCn	0.722320	C101n	0.631454	**C16On**	**0.019167**
C4n	0.212597	C102n	0.699658	C161On	0.148338
C4DCn	0.108912	C12n	0.453545	C18n	0.226941
C5n	0.799595	C121n	0.993969	C18On	0.282013
C51n	0.542221	C14n	0.198426	C181n	0.412894
C5DCn	0.868069	C14On	0.624174	C181On	0.656420
C6n	**0.018045**	C141n	0.678428	C182n	0.264194
C6DCn	0.624098				

Tabla 3.6: Resultados del test de Kruskal-Wallis para las variables del conjunto de Carnitinas. En negrita se resaltan los *p*-valores menores que 0,05.

Si tomamos como nivel de significación un valor algo mayor, un $\alpha = 0{,}10$, se puede comprobar que pasan el test de Kruskal-Wallis las variables *a183P*, *v221H*, *v226H*, *Valina*, *PCRU*, *GLU* y *LDL*. Estas variables, aunque no hayan resultado significativas para $\alpha = 0{,}05$, han estado cerca de hacerlo, por lo que puede que aparezcan en los modelos junto a otras variables que complementen la pequeña deficiencia que les ha hecho no pasar el test.

3.2.3. Correlaciones en Ácidos grasos

En el conjunto de Ácidos grasos, las variables van por parejas, ya que hacen referencia al mismo ácido graso, con la única diferencia que varía de dónde se ha tomado el dato. En las variables acabadas en *H*, el dato se ha tomado cuando el ácido graso va unido a la membrana del glóbulo rojo, mientras que las acabadas en *P* se han tomado del plasma sanguíneo. Por tanto, tiene sentido ver si estas parejas de variables están correladas entre sí, ya que a la hora de crear un modelo final, es preferible que entren variables que no estén correladas, para que puedan aportar mas riqueza al modelo.

Para comprobar si las variables tienen un coeficiente de correlación de Pearson que sea significativo, realizamos el siguiente test de significancia [23, 24] a cada uno de los coeficientes obtenidos:

H_0: El coeficiente de correlación es cero ($\rho_{xy} = 0$).
H_1: El coeficiente de correlación es distinto de cero ($\rho_{xy} = 0$).

Los resultados de los coeficientes de correlación por parejas y del test se pueden ver en Tabla 3.8. Ahí se puede ver que las variables *v140, v161, v180, v183, a183* y *v221* no están correladas. Del resto, las variables con una correlación, en valor absoluto, es superior a 0,5 son *v181, v200, v203, v204, v205, v220, v224, v225, v226* y *v241*. Las variables con valores inferiores a 0,5 son *v160, v182, v184, v201, v202, v240* y *v260*.

Conjunto de datos	Análisis numérico	Análisis gráfico	Test Kruskal-Wallis
Ácidos grasos	*v120H*, *v204P*	-	-
	-	*v201P*, *v224P*, *v160H*, *v201H*, *v221H*	*v201P*, *v224P*, *v201H*
	g183H, *v226H*	*g183H*, *v226H*	*g183H*
	-	-	*v260H*
Aminoácidos	*AcAlfaAminobutirico*	-	-
	-	*Fosfoserina*, *Valina*, *Cistina*, *Arginina*	-
	Alanina, *Prolina*	*Alanina*, *Prolina*	*Alanina*
Bioquímicas	*CTx*, *LPA*, *PCRU*	-	*CTx*
	APOB, *CHOL*, *LDL*	*APOB*, *CHOL*, *LDL*	*APOB*, *CHOL*
	-	-	*VitD*
Carnitinas	*C4DCn*, *C18n*	-	-
	-	*C3n*, *C12n*, *C14n*	-
	-	-	*C6n*, *C16On*

Tabla 3.7: Resumen de los resultados obtenidos en el análisis descriptivo y test de Kruskal-Wallis.

3.3. Análisis de Componentes Principales

En esta sección se va a realizar un análisis de componentes principales de cada conjunto.

Para realizar cada análisis, se procede de la siguiente manera. Se calcula la matriz de correlaciones, y a partir de ella se realizará el análisis y se calcularán las componentes.

Para determinar el número de componentes a calcular, se calculará el número de valores propios de la matriz de correlaciones. El número de factores a calcular se determinará con el número de valores propios mayores que 1 y su gráfico de sedimentación, donde se ve cómo disminuye la variabilidad explicada con cada componentes.

Cuando se ha elegido el número de componentes, se realiza el análisis de componentes principales, y se evalúa la variabilidad explicada por cada componente, la distribución de estos, y se representan los dos o tres que más variabilidad explican.

3.3.1. Ácidos grasos

Se va a realizar en este apartado el análisis de componentes principales con el conjunto de datos de Ácidos grasos. El conjunto tiene 47 variables con las que se calcularán las componentes.

Primero hay que determinar el número de componentes a calcular, y para ello utilizaremos los valores propios y el gráfico de sedimentación. Hay 15 valores propios mayores que uno, un número muy grande, y, si nos apoyamos en el gráfico de sedimentación, que se puede ver en la Figura 3.1, se puede comprobar que a partir del factor 10-11 apenas se explica varianza.

Si realizamos un primer análisis de componentes principales con número de factores desde 7 hasta 15 para ver la evolución de la varianza explicada. Esta evolución puede verse en la Tabla 3.9.

Variable	Correlación	Variable	Correlación	Variable	Correlación
v120	–	a183	0,12649674	v220	0,58292625 ∗∗
v140	0,01734069	v184	0,25808501 ∗∗	v221	0,13972357
v160	0,32663467 ∗∗	v200	0,62504895 ∗∗	v224	0,55132596 ∗∗
v161	-0,05714245	v201	0,43973284 ∗∗	v225	0,65460311 ∗∗
v180	0,12324818	v202	0,41904137 ∗∗	v240	0,34811635 ∗∗
v181	0,99978624 ∗∗	v203	0,59166598 ∗∗	v226	0,71476992 ∗∗
v182	0,16166086∗	v204	0,65538246 ∗∗	v241	0,50895565 ∗∗
g183	-0,07413649	v205	0,70356402 ∗∗	v260	0,26013543 ∗∗

Tabla 3.8: Valores de las correlaciones por variable del conjunto de Ácidos. Valores del test de correlación: $(**)$: p-valor $< 0{,}01$, $(*)$: p-valor $< 0{,}05$.

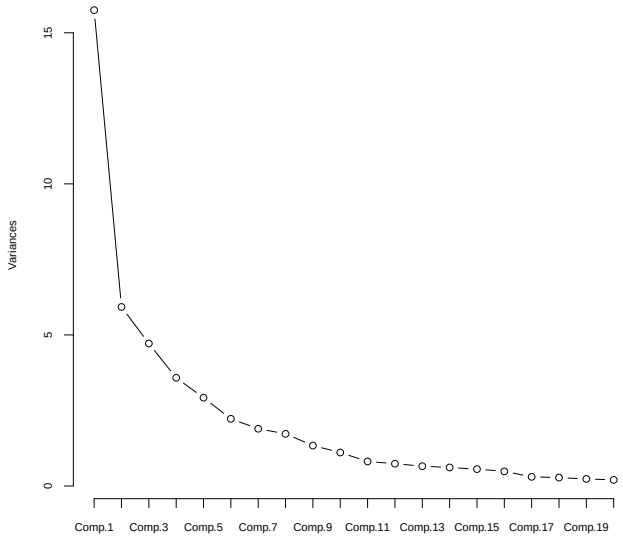

Figura 3.1: Gráfico de sedimentación del conjunto de Ácidos grasos.

A la vista de los resultados obtenidos, un número correcto de factores a calcular es 9, ya que se explica un 62,8 % de varianza y se utiliza un número no muy grande de factores.

Una vez realizado el análisis, se obtienen los 9 factores, y obtenemos que explican un 62,8 % de la variabilidad del conjunto. Los primeros tres factores explican un 10,4 %, un 10,4 % y un 7,9 % de la variabilidad cada uno.

De estos tres factores, el primero se compone de 6 variables, las parejas de variables *v205P* y *v205H*; *v225P* y *v225H*; y *v226P* y *v226H*, que hemos visto que están muy correladas entre sí. El segundo se compone de 9 variables, de las cuales 4 variables han sido obtenidas en análisis anteriores como buenas explicadoras de la variabilidad de la variable *apoE2*. Las cuatro variables son *v260H*, *v224P*, *v201P* y *v201H*. Cabe destacar que la variable *v184H* no se ha visto implicada en ningún factor.

Factores	Varianza explicada
7	55,1 %
8	59,2 %
9	62,8 %
10	66,0 %
11	68,9 %
12	71,6 %
13	74,0 %
14	76,4 %
15	78,5 %

Tabla 3.9: Evolución de varianza explicada por factores.

La representación gráfica de todos los factores con sus respectivas variables puede verse en la Figura 3.2, y en detalle puede verse en la Tabla 3.10.

Si representamos las tres componentes más influyentes en el plano, dos a dos, podemos ver que están dispersas sobre el plano, pero no se puede apreciar que ningún grupo de *apoE2* se aísle en alguna zona del plano. Estas representaciones pueden verse en el Anexo C.1.

3.3.2. Aminoácidos

Con el conjunto de Aminoácidos se va a realizar un análisis de componentes principales que nos permita ver cómo se pueden agrupar las 26 variables del conjunto de manera que no se pierda apenas información.

Se calcula la matriz de correlaciones, y a partir de ella, se calcula el número de factores. La matriz tiene dos valores propios mayores que 1, pero, al ser un número pequeño, vamos a tomar 4, ya que en el gráfico de sedimentación (Figura 3.3), a partir del cuarto factor, apenas hay variación en la varianza explicada añadiendo más factores.

Una vez realizado el análisis, los cuatro factores explican un 84,6 % de la varianza. El primer factor explica un 27,8 % de la varianza, el segundo un 13,5 %, el tercero explica un 17,1 % y por último el cuarto un 26,2 %. La primera componente se compone de 9 variables, siendo una de ellas la variable *Arginina*, que ya había sido obtenida anteriormente. El segundo factor se compone de 9 variables, entre ellas, las variables *Alanina*, *Prolina* o *Cistina*, variables obtenidas anteriormente como candidatas a un modelo final. El tercer factor se compone de 5 variables, entre ellas, las variables *AcAlfaAminobutirico*, *Fosfoserina* y *Valina*.

Las componentes puede verse en la Figura 3.4, y con más detalle en la Tabla 3.11.

Las representaciones dos a dos de las componentes que más varianza explican puede verse en el Anexo C.2. La distribución de las componentes sobre el plano no proporciona información de si se separan correctamente los grupos de *apoE2*, ya que las observaciones se encuentran dispersas y no agrupadas.

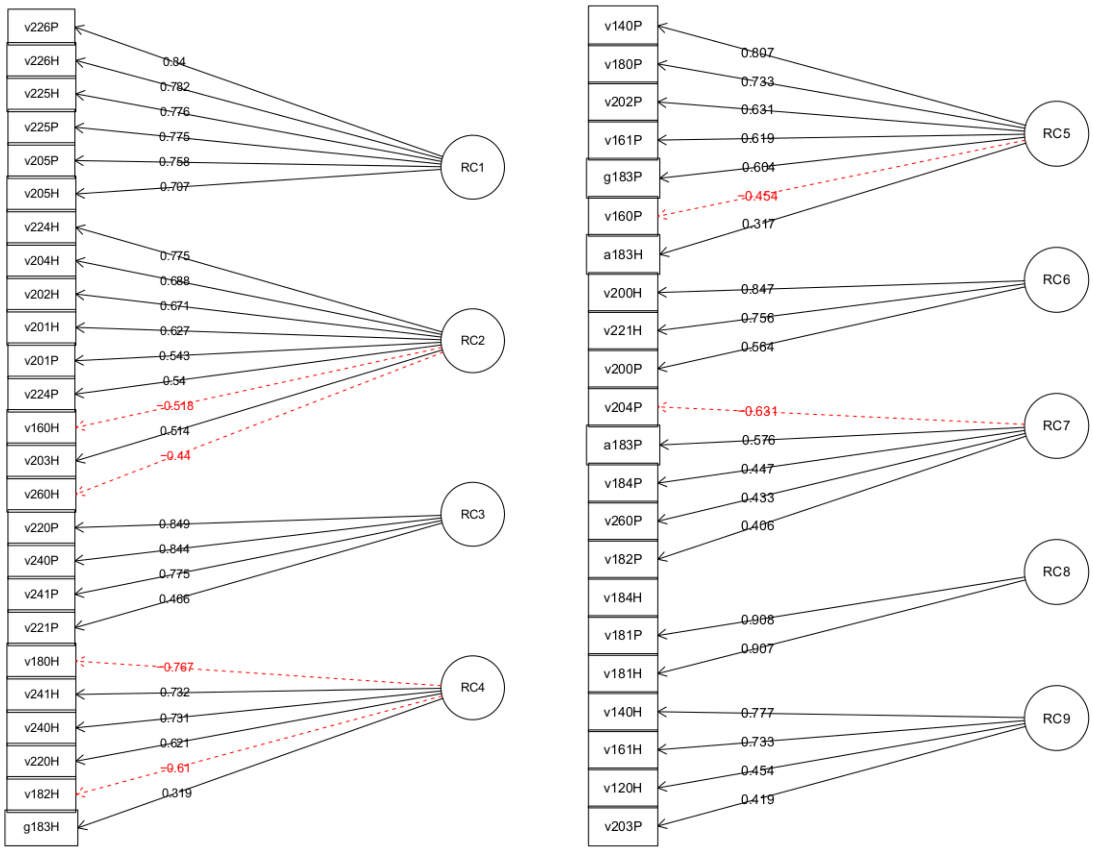

Figura 3.2: Representación de los factores del conjunto de Ácidos grasos con sus respectivas variables.

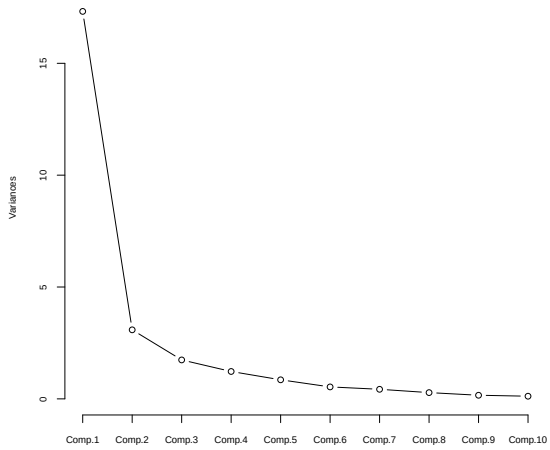

Figura 3.3: Gráfico de sedimentación del conjunto de Aminoácidos

Factor	Variables implicadas	Varianza explicada	Varianza acumulada
RC1	*v205P*, *v225P*,*v226P*, *v205H*, *v225H*,*v226H*	10,5 %	10,5 %
RC2	*v260H*, *v203H*, *v160H*, *v224P*, *v201P*, *v201H*, *v202H*, *v204H*, *v224H*	10,3 %	20,8 %
RC3	*v220P*, *v240P*, *v241P*, *v221P*	7,9 %	28,7 %
RC4	*g183H*, *v182H*, *v220H*, *v240H*, *v241H*, *v180H*	6,8 %	35,5 %
RC5	*v140P*, *v180P*, *v202P*, *v161P*, *g183P*, *v160P*, *a183H*	6,7 %	42,2 %
RC6	*v200P*, *v200H*, *v221H*	6,1 %	48,3 %
RC7	*v204P*, *a183P*, *v184P*, *v260P*, *v182P*	5,2 %	53,5 %
RC8	*v181P*, *v181H*	4,8 %	58,3 %
RC9	*v140H*, *v161H*, *v120H*, *v203P*	4,5 %	62,8 %

Tabla 3.10: Detalle de los factores del conjunto de Ácidos Grasos.

Factor	Variables implicadas	Varianza explicada	Varianza acumulada
RC1	*Arginina*, *Citrulina*, *Lisina*, *Asparragina*, *Histidina*, *Serina*, *Glutamina*, *Ornitina*, *Isolecuona*, *FenilAlanina*	27,8 %	27,8 %
RC2	*Triptofano*, *Prolina*, *Alanina*, *Tirosina*, *Treonina*, *Metionina*, *Glicina*, *Cistina*	26,2 %	54,0 %
RC3	*AcAlfaAminobutirico*, *Fosfoserina*, *X1MetilHistidina*, *Leucina*, *Valina*	17,1 %	71,1 %
RC4	*AcGlutamico*, *AcAspartico*, *Taurina*	13,5 %	84,6 %

Tabla 3.11: Detalle de los factores del conjunto de Aminoácidos.

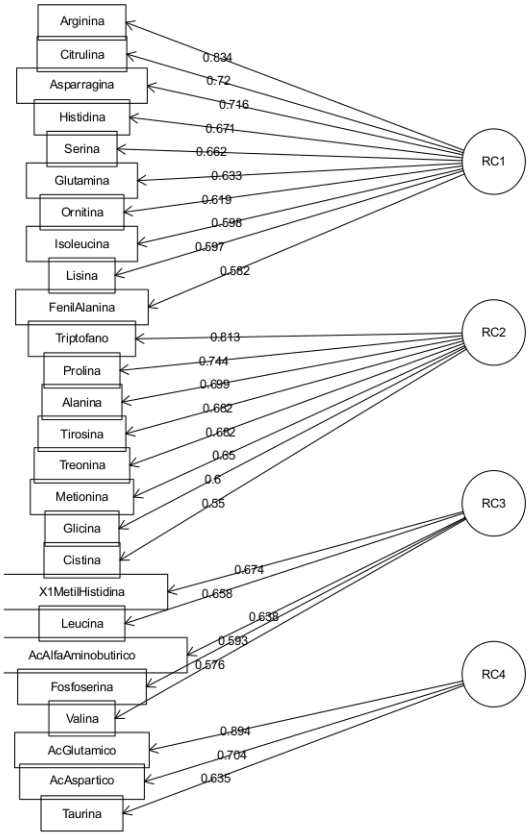

Figura 3.4: Representación de los factores del conjunto de Aminoácidos.

3.3.3. Bioquímicas

El conjunto de Bioquímicas está formado por 22 variables, de las cuales se va a intentar reducir el número mediante un análisis de componentes principales.

Primero se decide el número de componentes a calcular utilizando el gráfico de sedimentación (Figura 3.5) y el número de valores propios de la matriz de correlaciones mayores que 1. El número de valores propios mayores que 1 es 8, y junto al gráfico de sedimentación se corrobora que 8 es el número de factores a calcular.

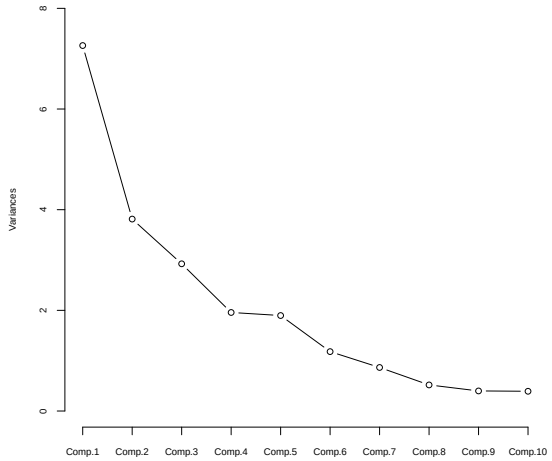

Figura 3.5: Gráfico de sedimentación del conjunto de Bioquímicas.

Con las componentes principales calculadas, se obtiene que se explica un 76,4 % de la varianza. Las componentes principales se pueden ver en la Figura 3.6 y con más detalle en la Tabla 3.12. La primera componentes explica un 14,3 % de la variabilidad del conjunto, y se compone de las variables *HEMG*, *HEMGm*, *GLU* y *INS*, que son variables que por separado no habían sido seleccionadas anteriormente, pero juntas puede que aporten información a la variable *apoE2*. La segunda componente está formada por tres variables vistas en los análisis descriptivos anteriores, las variables *APOB*, *LDL* y *CHOL*. Las tercera y cuarta componentes son componentes formadas por variables que no habían aparecido, como la primera componente. Están formadas por las variables *GOT*, *GPT* y *GGT*, y *cHDL*, *APOA* y *TRIG*, respectivamente.

Si realizamos la representación gráfica de las componentes dos a dos sobre el plano, se puede ver que ninguna de las parejas se distribuye de forma agrupada, sino que están distribuidas dispersamente y sin ningún tipo de orden. Estas representaciones pueden verse en el Anexo C.3.

3.3.4. Carnitinas

El conjunto de Carnitinas está formado por 31 variables, y para intentar reducir este número tan elevado se va a realizar un análisis de componentes principales.

El número de factores a calcular se calcula utilizando el número de valores propios mayores que 1 de la matriz de correlaciones y el gráfico de sedimentación (Figura 3.7). Una vez vistos estos resultados, se van a calcular 6 componentes.

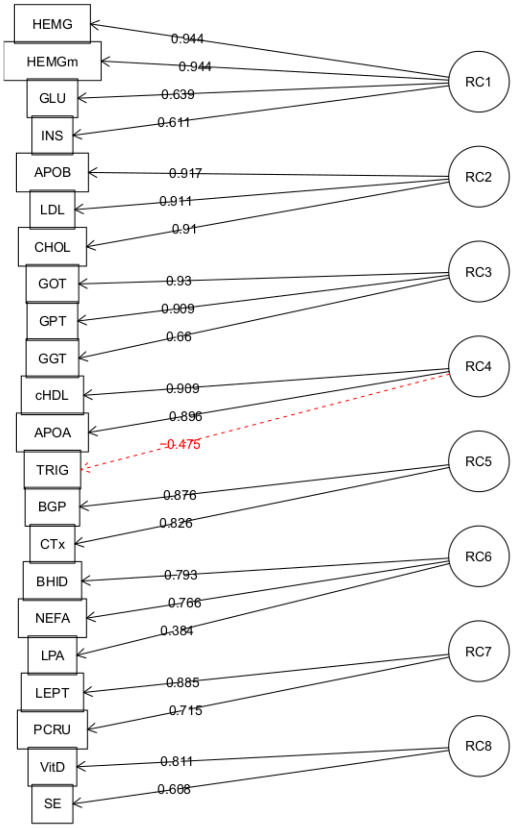

Figura 3.6: Representación de los factores del conjunto de Bioquímicas.

Factor	Variables implicadas	Varianza explicada	Varianza acumulada
RC1	*HEMG*, *HEMGm*, *GLU*, *INS*	14,3 %	14,3 %
RC2	*APOB*,*LDL*,*CHOL*	13,4 %	27,7 %
RC3	*GOT*,*GPT*,*GGT*	10,2 %	37,9 %
RC4	*cHDL*,*APOA*,*TRIG*	10,2 %	48,1 %
RC5	*BGP*,*CTx*	8,1 %	56,2 %
RC6	*BHID*,*NEFA*,*LPA*	7,6 %	63,8 %
RC7	*LEPT*,*PCRU*	6,8 %	70,6 %
RC8	*VitD*,*SE*	5,8 %	76,4 %

Tabla 3.12: Detalle de los factores del conjunto de Bioquímicas.

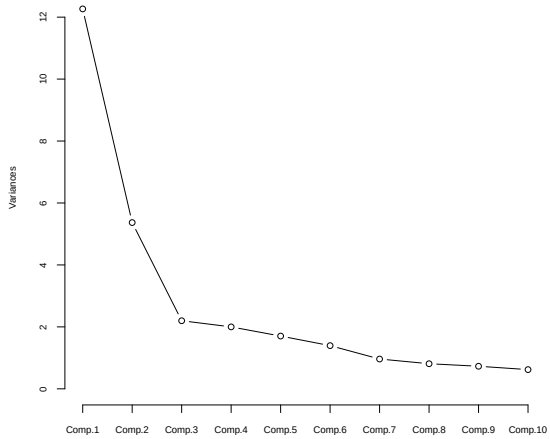

Figura 3.7: Gráfico de sedimentación del conjunto de Carnitinas.

Una vez calculadas las 6 componentes, con ellas, se explica un 65,2 % de la varianza total. Las componentes pueden verse en la Figura 3.8 y la Tabla 3.13.

La primera componente explica un 18,8 % de la varianza y está formada por 9 variables, de las cuales únicamente la variable *C14n* ha sido obtenida en resultados anteriores. Con un 15,9 % de varianza explicada está la segunda componentes, formada por 7 variables que en principio por separado no parecían explicar la variabilidad de la variable *apoE2*. La tercera componentes está formada por 5 variables, y entre ellas las variables *C3n* y *C4DCn*, explicando un 10,1 % de la varianza.

Factor	**Variables implicadas**	**Varianza explicada**	**Varianza acumulada**
RC1	*C16n*, *C161On*, *C14n*, *C181n*, *C18n*, *C161n*, *C182n*, *C0n*, *C2n*	18,8 %	18,8 %
RC2	*C121n*, *C141n*, *C12n*,*C142n* *C101n*, *C10n*, *C8n*	15,9 %	34,7 %
RC3	*C4n*, *C5n*, *C51n*, *C3n*, *C4DCn*	10,1 %	44,8 %
RC4	*C14On*, *C16On*, *C181On*, *C3DCn*, *C18On*, *C5DCn*	7,9 %	52,7 %
RC5	*C81n*, *C102n*, *C6DCn*	6,7 %	59,4 %
RC6	*C6n*	5,8 %	65,2 %

Tabla 3.13: Detalle de los factores del conjunto de Carnitinas.

La representación de las componentes calculadas dos a dos, puede verse en el Anexo C.4, en el cual no puede observarse una distribución por grupos de las componentes.

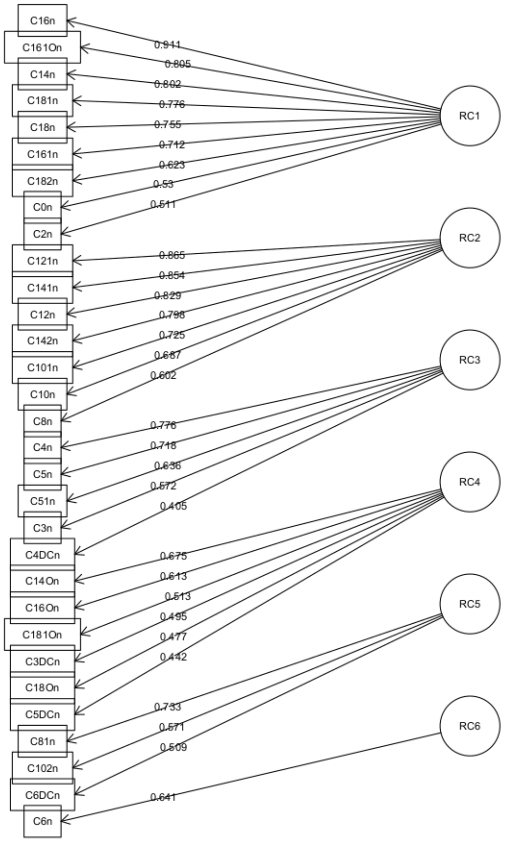

Figura 3.8: Representación de los factores del conjunto de Carnitinas.

3.4. Regresión Logística Multinomial

En esta sección se va a realizar una regresión logística multinomial para cada conjunto de variables con el fin de obtener las variables que mejor separan por grupos el *apoE2*. De esta manera, se puede reducir el número de variables a la hora de elaborar el modelo final.

En primer lugar se realiza una regresión logística hacia delante con todas las variables, para descartar las variables menos influyentes y mantener las más predictivas. Una vez seleccionadas las variables en la regresión anterior, se añadirán las variables que se han seleccionado en los análisis anteriores y se realizará una regresión paso a paso *backward/forward* con el fin de descartar las variables menos influyentes.

3.4.1. Ácidos grasos

En este apartado se va a realizar una regresión logística con el conjunto de Ácidos grasos. El conjunto de Ácidos grasos tiene 47 variables, por lo que el objetivo va a ser eliminar variables para poder quedarnos con las que realmente separan bien la variable *apoE2*.

Se realiza el primer modelo y la primera regresión hacia delante y se obtienen los resultados que podemos ver en el Anexo D.1.1.

En esta regresión, las variables finalistas han sido *g183H*, *v182P*, *v226H*, *v221H*, *v184H* y *a183P*, de las cuales, las variables *g183H*, *v226H* y *v221H* las habíamos obtenido en análisis anteriores.

Con las variables finales del modelo hacia delante y con las variables de los análisis anteriores, se va a realizar una regresión paso a paso. Las variables que se utilizarán forman la siguiente lista de 13 variables: *g183H*, *v182P*, *v226H*, *v221H*, *v184H*, *a183P*, *v120H*, *v204P*, *v201P*, *v224P*, *v160H*, *v201H* y *v260H*.

Una vez realizada la regresión paso a paso final, se obtienen las mismas variables que en el primer modelo paso a paso. Si miramos la significación de cada variable, obtenemos que solo las variables *v221H* y *v184H* obtienen un p-valor inferior a 0,05. Los resultados de la regresión se pueden ver en el Anexo D.1.2.

Para ver la eficacia del modelo, se calcula la matriz de confusión, con la que se puede comparar las observaciones reales con las predichas por el modelo, obteniendo la Tabla 3.14. El modelo tras los resultados obtenidos falla en un 25,59 % de los casos, pero clasifica mal las observaciones de los grupos no mayoritarios.

3.4.2. Aminoácidos

En este apartado se va a realizar una regresión logística con el conjunto de Aminoácidos. El conjunto tiene 24 variables, de manera que se intentará reducir para tener las que separan correctamente la variable respuesta *apoE2*.

Se realiza el primer modelo hacia delante y se obtienen los resultados que podemos ver en el Anexo D.2.1.

En esta regresión, las variables finalistas han sido *Cistina*, *Arginina* y *Glicina*, de las cuales, *Cistina* y *Arginina* son variables que habíamos obtenido en resultados anteriores como candidatas a buenas variables para separar el *apoE2*.

Con las variables finales del modelo anterior y con las variables de los análisis anteriores, se va a realizar una regresión paso a paso. Las variables que se utilizarán forman la siguiente lista de 8 variables: *Cistina*, *Arginina*, *Glicina*, *AcAlfaAminobutirico*, *Fosfoserina*, *Valina*, *Alanina* y *Prolina*.

Una vez realizada la regresión paso a paso final, se obtienen las mismas variables que en el primer modelo paso a paso y la variable *Alanina*, variable obtenida en los análisis anteriores. Si miramos la significación de cada variable, obtenemos que todas las variables obtienen un p-valor inferior a 0,05. Los resultados de la regresión se pueden ver en el Anexo D.2.2.

Para ver si el modelo predice correctamente, se calcula la matriz de confusión, con la que se puede comparar las observaciones reales con las predicciones del modelo, obteniendo la Tabla 3.14. El modelo obtenido falla en un 25,83 % de los casos. El modelo solo clasifica 5 observaciones como distintas del grupo *E3*, y de estas, se clasifican correctamente 2 de los grupos *E2* y *E4* .

3.4.3. Bioquímicas

En este apartado se va a realizar una regresión logística con el conjunto de Bioquímicas. El conjunto tiene 22 variables, las cuales se intentará reducir a un número menor para tener las que separan correctamente la variable respuesta *apoE2*.

Se realiza la primera regresión hacia delante y se obtienen los resultados que podemos ver en el Anexo D.3.1.

En esta regresión, las variables finalistas han sido *APOB*, *LPA* y *VitD*, todas ellas variables que habíamos obtenido en resultados anteriores como candidatas a buenas variables para separar el *apoE2*.

Como todas las variables obtenidas ya las habíamos obtenido en los análisis anteriores, se va a realizar una regresión paso a paso. Dichas variables son las 7 siguientes: *CTx*, *LPA*, *PCRU*, *APOB*, *CHOL*, *LDL* y *VitD*.

Una vez realizada la regresión paso a paso final, se obtienen las mismas variables que en el primer modelo y la variable *PCRU*, variable obtenida en los análisis anteriores. Si miramos la significación de cada variable, obtenemos que sólo las variables *APOB*, y *PCRU* obtienen un p-valor inferior a 0,05. Los resultados de la regresión se pueden ver en el Anexo D.3.2.

Para ver si el modelo predice correctamente, se calcula la matriz de confusión, que se puede ver en la Tabla 3.14, con la que se comparan las observaciones reales con las predicciones del modelo. El modelo final obtenido falla en un 28,04 % de los casos y además no hace ninguna clasificación como grupo *E4*.

3.4.4. Carnitinas

En este apartado se va a realizar una regresión logística con el conjunto de Carnitinas. El conjunto tiene 31 variables, el objetivo es reducir este número de variables para poder quedarnos con las más

influyentes de cara a encontrar las variables que separen correctamente la variable respuesta *apoE2*.

Se realiza el primer modelo y la primera regresión paso a paso hacia delante y se obtienen los resultados que podemos ver en el Anexo D.4.1.

En esta regresión, las variables finalistas han sido *C16On*, *C6n*, *C14n*, *C182n*, *C4DCn*, *C5n* y *C18On*. En resultados anteriores, de estas variables, habíamos obtenido que las variables *C16On*, *C6n*, *C14n* y *C4DCn* eran candidatas a buenas variables para separar la variable *apoE2*.

Con las variables obtenidas en esta primera regresión y las obtenidas en los análisis descriptivos, se va a realizar una regresión paso a paso. Las variables para hacer la regresión son las 10 siguientes: *C18n*, *C3n*, *C12n*, *C16On*, *C6n*, *C14n*, *C182n*, *C4DCn*, *C5n* y *C18On*.

Una vez realizada la regresión paso a paso final, se obtienen los mismos resultados que en el primer modelo paso a paso. Si miramos la significación de cada variable, obtenemos que sólo las variables *C14n*, y *C182n* obtienen un *p*-valor inferior a 0,05. Los resultados de la regresión se pueden ver en el Anexo D.4.2.

Para ver si el modelo predice correctamente, se calcula la matriz de confusión, que puede verse en la Tabla 3.14. El modelo final obtenido falla en un 27,35 % de los casos y con una pésima clasificación de los grupos minoritarios.

3.4.5. Conclusiones

En general, no se han obtenido buenos modelos que permitan la clasificación de los grupos de la variable respuesta, como puede verse en la Tabla 3.14. Todos coinciden en que apenas hay clasificaciones de los grupos *E2* y *E4*.

(a)		Predicción		
		E2	E3	E4
apoE2	E2	2	19	1
	E3	1	118	3
	E4	1	18	5

(b)		Predicción		
		E2	E3	E4
apoE2	E2	2	17	0
	E3	0	108	0
	E4	1	21	2

(c)		Predicción		
		E2	E3	E4
apoE2	E2	2	23	0
	E3	3	152	0
	E4	0	34	0

(d)		Predicción		
		E2	E3	E4
apoE2	E2	1	23	3
	E3	1	156	4
	E4	0	30	5

Tabla 3.14: Matrices de confusión de las regresiones finales paso a paso de los conjuntos de Ácidos grasos (a), Aminoácidos (b), Bioquímicas (c) y Carnitinas (d).

Del conjunto de Ácidos grasos, se ha obtenido en la regresión tres variables que ya estaban siendo presentes en los resultados de los análisis realizados anteriormente. Dichas variables son *v221H*, *g183H* y *v226H*, por lo que probablemente estarán en el modelo final que se calculará. También han aparecido en los resultados de la regresión las variables *v182P*, *v184H* y *a183P*, y es la primera vez que aparecen en los resultados de un análisis.

Del conjunto de Aminoácidos, se ha conseguido una regresión con cuatro variables finalistas que son *Alanina*, *Cistina*, *Arginina* y *Glicina* De estas variables, la variable *Glicina* es la primera vez que aparece tras algún análisis. La variable *Alanina* ha aparecido en todos los análisis, por lo que puede tener importancia de cara a calcular un modelo final.

Del conjunto de Bioquímicas se ha conseguido una regresión de cuatro variables que ya habían aparecido en análisis anteriores. Estas variables son *LPA*, *PCRU*, *APOB* y*VitD*.

Por último, del conjunto de Carnitinas se ha obtenido una regresión logística de siete variables, de las cuales cuatro ya habían aparecido en análisis anteriores, y las otras tres restantes han aparecido por primera vez como posibles candidatas que separen correctamente el *apoE*2. Las variables ya vistas antes son *C4DCn*,*C14n*, *C6n* y *C16On*, y las nuevas variables son *C182n*, *C5n* y *C18On*.

El resumen de las variables obtenidas a lo largo de los análisis realizados puede verse en la Tabla 3.15.

Conjunto de datos	Análisis descriptivo	Test de Kruskal-Wallis	Regresión logística
Ácidos grasos	*v120H*, *v204P*, *v201P*, *v224P*, *v160H*, *v201H*, *v221H*, *g183H*, *v226H*	*v201P* *v224P*, *v201H*, *g183H*	*v221H*, *g183H*, *v226H*
	-	-	*v182P*, *v184H*, *a183P*
Aminoácidos	*AcAlfaAminobutirico*, *Fosfoserina*, *Valina*, *Cistina*, *Arginina*, *Alanina*, *Prolina*	*Alanina*	*Cistina*, *Arginina*, *Alanina*
	-	-	*Glicina*
Bioquímicas	*CTx*, *LPA*, *PCRU*, *APOB*, *CHOL*, *LDL*	*CTx*, *APOB*, *CHOL*, *VitD*	*LPA*, *PCRU*, *APOB*, *VitD*
Carnitinas	*C4DCn*, *C18n*, *C3n*, *C12n*, *C14n*	*C6n*, *C16On*	*C4DCn*, *C14n*, *C6n*, *C16On*
	-	-	*C182n*, *C5n*, *C18On*

Tabla 3.15: Resumen de los resultados obtenidos tras la regresión logística.

3.5. Regresión Logística Multinomial utilizando las componentes principales

En esta sección se va a realizar una regresión logística multinomial para cada conjunto, como en la sección anterior, pero añadiendo a los modelos obtenidos anteriormente las componentes principales que se han calculado en la Sección 3.3. De esta manera podremos ver si alguna de las variables es sustituida por alguna componente principal.

3.5.1. Ácidos grasos

Cuando se ha realizado la regresión logística en el apartado anterior, el modelo logístico de Ácidos grasos se componía de las variables *v221H*, *g183H*, *v226H*, *v182P*, *v184H* y *a183P*.

En el análisis de componentes principales realizado anteriormente se han obtenido 9 factores que incluiremos con las variables anteriores en la regresión logística que vamos a realizar.

Una vez realiza la regresión cuyos resultados pueden verse el Anexo E.1, se comprueba que el modelo obtenido es el mismo que sin incluir las componentes, por lo que las componentes no han aportado nada en este conjunto.

3.5.2. Aminoácidos

Con el conjunto de Aminoácidos se ha obtenido antes una regresión de 4 variables. Añadiendo las 4 variables de las componentes principales, tenemos la siguiente lista de variables para hacer regresiones: *Cistina*, *Arginina*, *Alanina*, *Glicina*, *RC1*, *RC2*, *RC3* y *RC4*.

Una vez realizada la regresión logística con las variables anteriores, se obtiene que la variable *Alanina* se cae de la regresión y en su lugar entran las componentes *RC3* y *RC4*. Los resultados del modelo pueden verse en el Anexo E.2.

El test de significación de las variables es superado por todas las variables excepto por *Cistina*.

Si calculamos la matriz de confusión en la Tabla 3.16, vemos que no se predice como *E4* ningún caso, y que se tiene un error de predicción del 28,24 %. En este caso, a diferencia que en el modelo sin las componentes, no se ha clasificado ninguna observación como *E4*, lo que no favorece al modelo.

		Predicción		
		E2	E3	E4
apoE2	E2	4	15	0
	E3	2	90	0
	E4	0	20	0

Tabla 3.16: Matriz de confusión de la regresión paso a paso del conjunto de Aminoácidos y las componentes principales.

3.5.3. Bioquímicas

El conjunto de Bioquímicas en la regresión logística había obtenido un modelo final formado por las variables *LPA*, *PCRU*, *APOB* y *VitD*. Ahora vamos a calcular la regresión logística incluyendo las componentes principales, que son 8, de manera que realizaremos la regresión logística paso a paso con 12 variables.

Una vez realizada la regresión, cuyos resultados se pueden ver en el Anexo E.3, se tiene una ninguna de las componentes principales es incluida en el modelo.

3.5.4. Carnitinas

Con el conjunto de Carnitinas se ha obtenido anteriormente una regresión logística de 7 variables, que son *C16On*, *C6n*, *C14n*, *C182n*, *C4DCn*, *C5n* y *C18On*. En el análisis de componentes principales, se han obtenido 6 componentes para este conjunto, por lo que a la hora de hacer una regresión se utilizarán 13 variables.

Una vez calculado el modelo, se obtiene el mismo modelo que anteriormente, con las 7 variables originales, por lo que las componentes principales no han aportado nada. Dichos resultados pueden comprobarse en el Anexo E.4.

3.5.5. Conclusiones

Las conclusiones obtenidas tras realizar una regresión logística añadiendo las componentes principales son que las componentes no han aportado nada de información en comparación con las variables finalistas del modelo de regresión logística.

3.6. Árboles de Clasificación

En esta sección vamos a realizar árboles de clasificación. Para ello, utilizaremos la librería *rpart* de *R* [20].

El objetivo del análisis es obtener un árbol con un error medio menor que uno, y en caso contrario, con el menor error posible. Una vez obtenido el árbol, se calcularán las variables más importantes del árbol y se compararán con resultados anteriores.

3.6.1. Ácidos grasos

El conjunto de Ácidos tiene 47 variables originales, y aplicando esta técnica intentaremos obtener las variables más importantes del conjunto de datos y así poder reducir el número de variables de cara a un modelo finalista.

El árbol óptimo puede verse en la Figura 3.9. Este tiene un error en el nodo raíz del 27,49 % y se compone de las variables *g183H* y *v205P*.

En este árbol, las 10 variables más importantes por orden son *v205P*, *v201H*, *g183H*, *v160P*, *v161P*, *v226P*, *v200H*, *g183P*, *v200P* y *v140P*. De todas estas variables, las variables *v201H* y *g183H* ya nos habían aparecido en los análisis anteriores, por tanto podemos considerarlas de gran importancia para un modelo.

Si calculamos la matriz de confusión obtenemos la Tabla 3.17, en la que podemos ver que el árbol falla en un 25,11 % de los casos a la hora de hacer una nueva predicción y no se clasifica ninguna variable como *E2*, y la gran mayoría de los casos son clasificados como *E3*, que es el grupo mayoritario, Debido a esto no es un buen modelo de clasificación y le faltaría el apoyo de alguna variable extra.

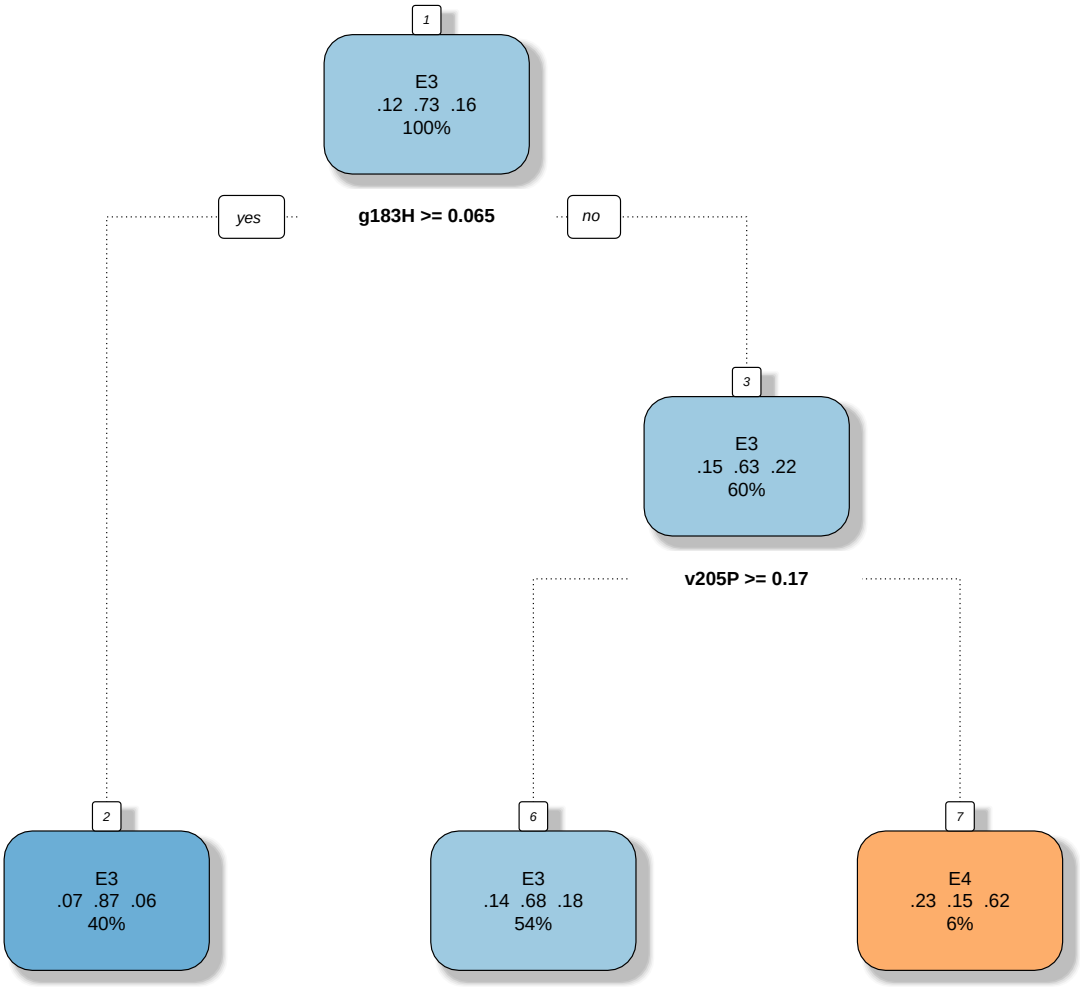

Figura 3.9: Árbol óptimo del conjunto Ácidos grasos.

Con este árbol no se han obtenido buenos resultados, pero sí se han obtenido variables que hasta ahora no habían aparecido, como es la variable *v205P* que es utilizada por el árbol para realizar un *split*. También aparecido como variables importantes las variables *v160P* y *v161P*.

3.6.2. Aminoácidos

Se va a realizar un árbol de clasificación con el conjunto de Aminoácidos. El conjunto tiene 24 variables, de las cuales el objetivo es obtener las más importantes de cara a un modelo final.

No se ha obtenido un árbol con un error medio inferior a 1, pero el árbol con menor error es el que se puede ver en la Figura 3.10. El árbol está formado por las variables *Cistina*, *Alanina*, *Valina*, *Isoleucina* y *Serina*.

Las variables más importantes son *Cistina*, *Serina*, *Isoleucina*, *Alanina*, *Valina*, *Leucina*, *Tirosina*, *Treonina* y *Glutamita*, que, siguiendo ese orden, son las cinco variables que forman el árbol y cuatro variables nuevas que hasta ahora no habían aparecido. De las variables que forman el árbol, ya habíamos obtenido en análisis anteriores las variables *Cistina*, *Alanina* y *Valina*, destacando las dos primeras que han aparecido en la regresión logística.

Si calculamos la matriz de confusión obtenemos la Tabla 3.17, donde vemos que el árbol falla en un 23,78 % de los casos a la hora de hacer una nueva predicción. Si nos fijamos en detalle, se puede ver que todas las observaciones salvo 9 han sido clasificadas como grupo *E3*, por lo que no es un árbol que inspire mucha confianza.

3.6.3. Bioquímicas

El conjunto de Bioquímicas tiene 22 variables originales, y utilizando esta técnica, llegaremos a las variables más importantes del conjunto que mejor separen el *apoE2*.

En este conjunto no ha podido calcularse un árbol óptimo con un error medio inferior que 1, pero sin embargo se ha calculado el mejor árbol posible. El árbol puede verse en la Figura 3.11. El árbol obtenido tiene un error en el nodo raíz del 27,75 %.

Las variables que forman el árbol son *APOB*, *VitD*, *BHID* y *CHOL*. De estas variables, en la regresión logística se habían utilizado las variables *APOB* y *VitD*, lo que denota su importancia.

Las variables más importantes del árbol son, en orden de importancia, *APOB*, *CHOL*, *VitD*, *BGP*, *BHID*, *GOT* y *LPA*. Como podemos observar, la variable *BGP* es más importante que una variable que ha formado parte del árbol, por lo que habrá que tenerla en cuenta.

Si calculamos la matriz de confusión obtenemos la Tabla 3.17. En ella vemos que el árbol no predice correctamente el 24,23 % de nuevas observaciones. La gran mayoría de las predicciones se hacen al grupo *E3*, el grupo mayoritario, por lo que este árbol no clasifica correctamente los grupos y necesita alguna variable de apoyo más para completar esta debilidad.

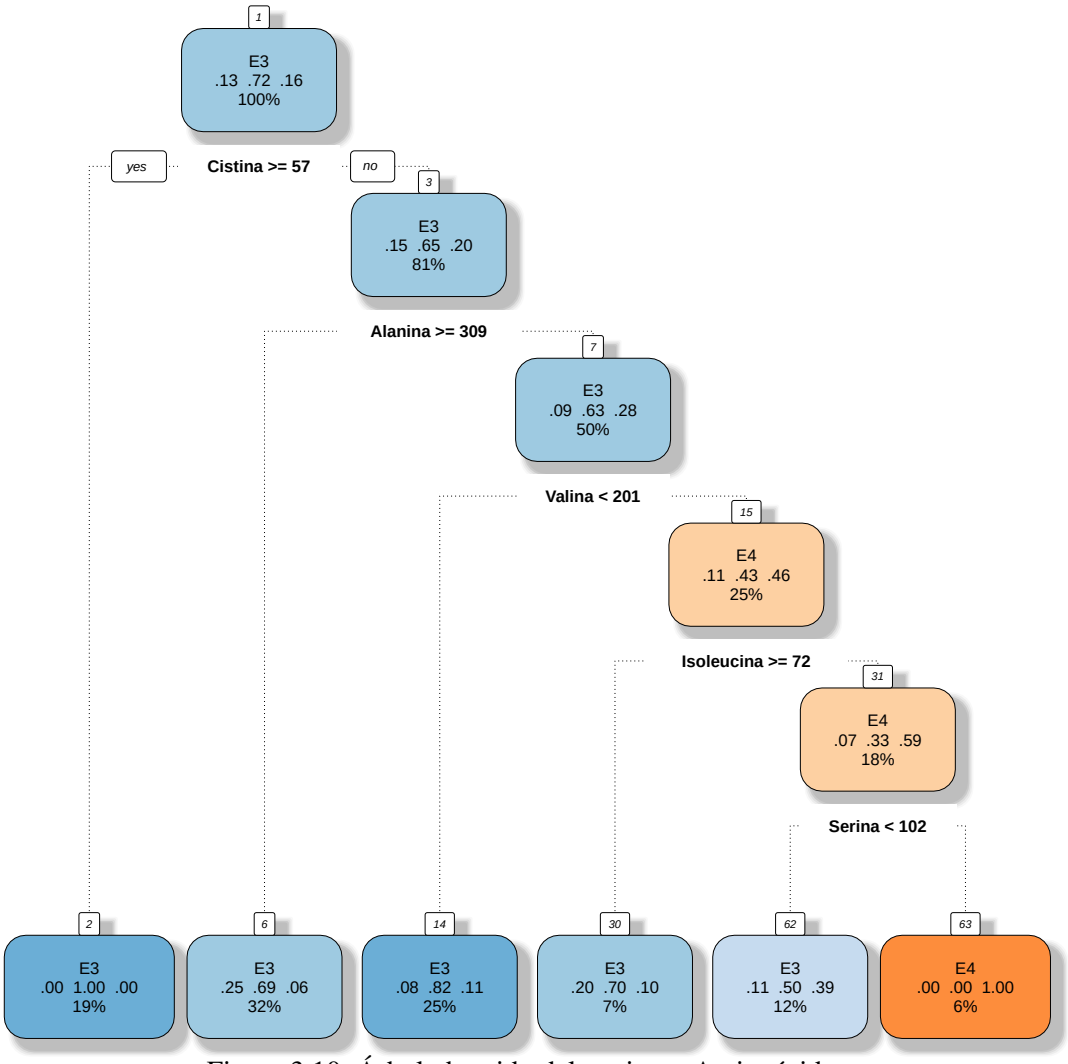

Figura 3.10: Árbol obtenido del conjunto Aminoácidos.

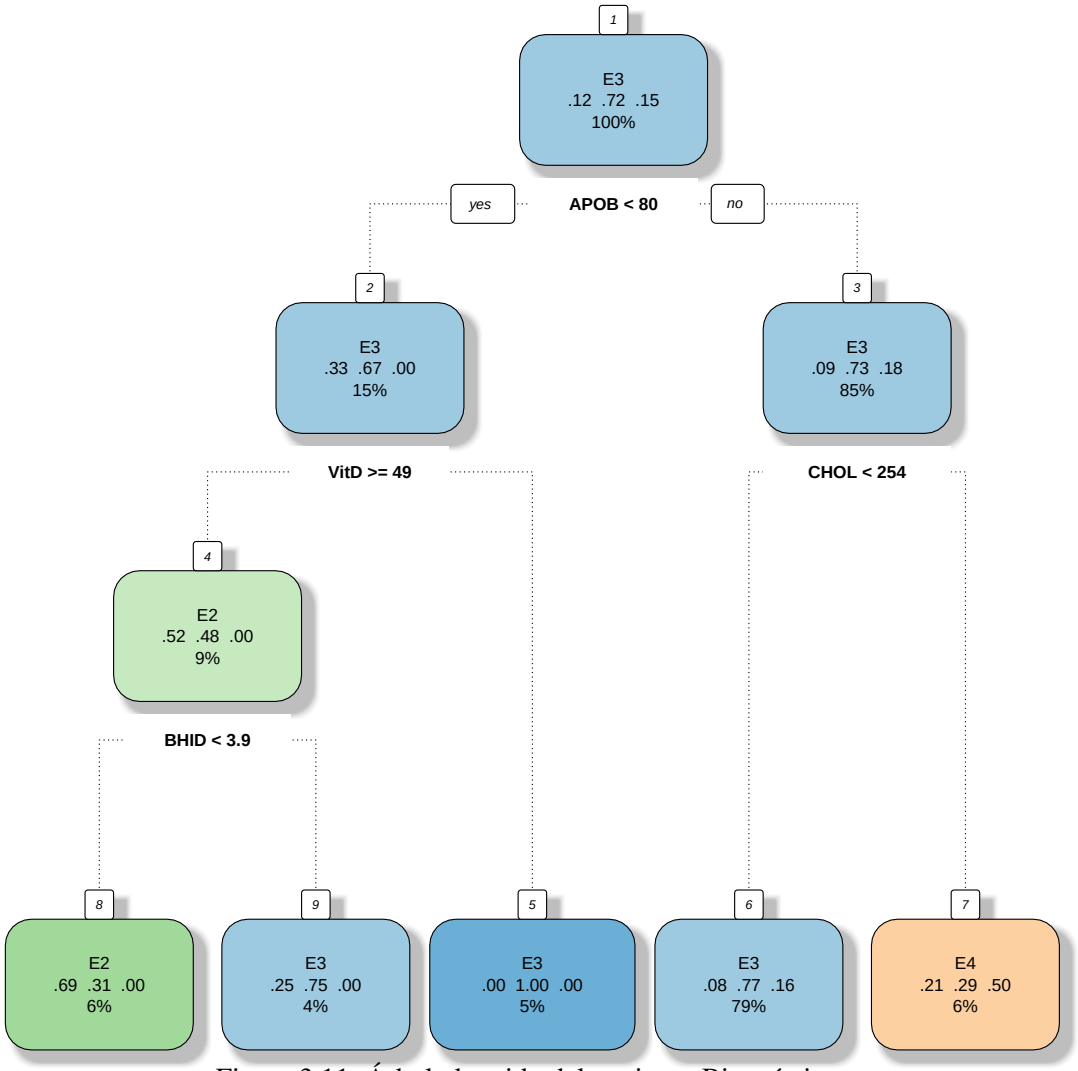

Figura 3.11: Árbol obtenido del conjunto Bioquímicas.

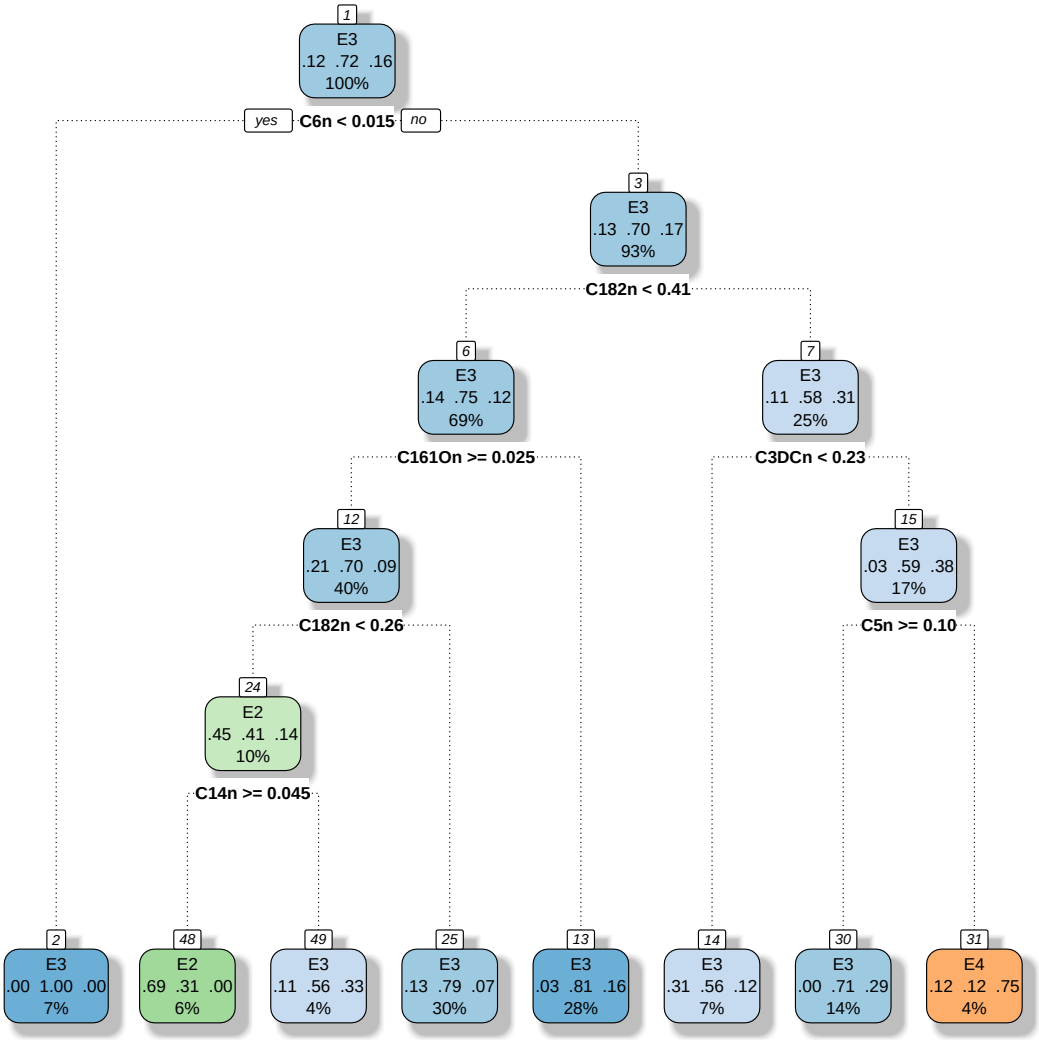

Figura 3.12: Árbol óptimo del conjunto Carnitinas.

3.6.4. Carnitinas

Se va a calcular un árbol de clasificación para el conjunto de Carnitinas. El conjunto en su origen tiene 31 variables, de las cuales seleccionaremos las más influyentes de cara a obtener el modelo final.

El árbol óptimo calculado se puede ver en la Figura 3.12. El nodo raíz tiene un error de 27,80 %, y las variables que lo forman son las variables *C6n*, *C182n*, *C161On*, *C3DCn*, *C5n* y *C14n*. Las variables más importantes en el árbol son las anteriores y las variables *C2n* y *C181n*. Estas tres variables tienen más importancia en el árbol que alguna de las variables que han entrado en este.

De todas estas variables mencionadas, en análisis anteriores no se habían obtenido hasta ahora las variables *C2n*, *C161On*, *C181n*, *C3DCn* y *C121n*, con lo que son variables que posiblemente haya que tener en cuenta.

Si calculamos la matriz de confusión obtenemos la Tabla 3.17, donde vemos que el árbol falla en un 23,35 % de los casos a la hora de hacer una nueva predicción. Como viene pasando hasta ahora, el árbol tiene el mismo comportamiento que los anteriores, clasifica la mayor parte de las observaciones

como el grupo mayoritario e ignora las de los grupos minoritarios.

3.6.5. Conclusiones

En general, no se han obtenido buenos árboles de regresión logística. El gran problema de los árboles ha sido la clasificación incorrecta de las observaciones de los grupos minoritarios, los grupos *E2* y *E4*. La predicción de los modelos se puede ver en sus matrices de confusión de los árboles, que se pueden ver en la Tabla 3.17.

(a)		Predicción		
		E2	E3	E4
apoE2	E2	0	25	3
	E3	0	162	2
	E4	0	27	8

(b)		Predicción		
		E2	E3	E4
apoE2	E2	0	28	0
	E3	0	164	0
	E4	0	26	9

(c)		Predicción		
		E2	E3	E4
apoE2	E2	9	16	3
	E3	4	156	4
	E4	0	28	7

(d)		Predicción		
		E2	E3	E4
apoE2	E2	9	18	1
	E3	2	159	3
	E4	0	29	6

Tabla 3.17: Matrices de confusión de los árboles obtenidos con los conjuntos de Ácidos grasos (a), Aminoácidos (b), Bioquímicas (c) y Carnitinas (d).

Del conjunto de Ácidos grasos se ha obtenido un árbol óptimo de dos variables, *g183H* y *v205P*. La variable *g183H* siempre ha ido apareciendo a lo largo de los análisis realizados, por lo que se puede ver la gran importancia de esta en el conjunto. Como variables importantes en el árbol, han aparecido las variables *g183H*, *v205P*, *v160P* y *v161P*.

Del conjunto de Aminoácidos, el árbol obtenido está formado por cinco variables, que son *Valina*, *Cistina*, *Alanina*, *Isoleucina* y *Serina*, de las cuales *Cistina* y *Alanina* habían entrado en el modelo de regresión logística.

Del conjunto de Bioquímicas se ha obtenido un árbol formado por cuatro variables, que son *APOB*, *CHOL*, *VitD* y *BHID*, y de estas, *APOB* y *VitD* ya habían entrado en el modelo de regresión logística, lo que indica que habrá que tenerlas en cuenta. Como variables importantes, han aparecido las variables del árbol, pero también la variable *BGP*, que tiene más importancia que una de las variables que ha entrado en el árbol.

Del conjunto de Carnitinas se ha obtenido un árbol formado por seis variables, que son *C6n*, *C182n*, *C161On*, *C3DCn*, *C5n* y *C14n*, y cuatro de ellas también han aparecido en el modelo de regresión logística. Estas cuatro variables *C6n*, *C182n*, *C5n* y *C14n* habrá que tenerlas en cuenta. Además, las variables *C2n* y *C181n* han tenido una importancia en el árbol mayor que alguna variable que ha entrado a formar parte del árbol.

El resumen detallado de las variables tras los análisis se puede ver en la Tabla 3.18.

Conjunto de datos	Primeros análisis	Regresión logística	Árboles de clasificación
Ácidos grasos	*v120H*, *v204P*, *v201P*, *v224P*, *v160H*, *v201H*, *v221H*, *g183H*, *v226H*	*v221H*, *g183H*, *v226H*	*g183H*
	-	*v182P*, *v184H*, *a183P*	-
	-	-	*v205P*
Aminoácidos	*AcAlfaAminobutirico*, *Fosfoserina*, *Valina*, *Cistina*, *Arginina*, *Alanina*, *Prolina*	*Cistina*, *Arginina*, *Alanina*	*Valina*, *Cistina*, *Alanina*
	-	*Glicina*	-
	-	-	*Isoleucina*, *Serina*
Bioquímicas	*CTx*, *LPA*, *PCRU*, *APOB*, *CHOL*, *LDL*, *VitD*	*LPA*, *PCRU*, *APOB*, *VitD*	*APOB*, *CHOL*, *VitD*
	-	-	*BHID*
Carnitinas	*C4DCn*, *C18n*, *C3n*, *C12n*, *C14n*, *C6n*, *C16On*	*C4DCn*, *C14n*, *C6n*, *C16On*	*C14n*, *C6n*
	-	*C182n*, *C5n*, *C18On*	*C182n*, *C5n*
	-	-	*C161On*, *C3DCn*

Tabla 3.18: Resumen de los resultados obtenidos tras los árboles de clasificación.

3.7. Árboles de Clasificación utilizando las componentes principales

De manera similar que antes, en esta sección se van a calcular árboles de clasificación para cada conjunto de datos, con la única diferencia de que se añaden las variables calculadas en el análisis de componentes principales, que se puede ver en la Sección 3.3.

3.7.1. Ácidos grasos

En el conjunto de Ácidos grasos incluyendo las componentes, se ha obtenido el mismo árbol que en el apartado anterior, con la única diferencia que ha entrado una componente como variable importante, la componente *RC3*, formada por las variables *v220P*, *v240P*, *v241P* y *v221P*.

3.7.2. Aminoácidos

En el conjunto de Aminoácidos, se ha intentado calcular un árbol de clasificación óptimo incluyendo las componentes principales calculadas, pero no se ha conseguido, por lo que se ha calculado el mejor de los árboles obtenidos. El árbol obtenido ha sido el mismo que en el apartado anterior, en el cual tampoco han entrado las componentes como variables importantes.

3.7.3. Bioquímicas

Con el conjunto de Bioquímicas se ha intentado calcular un árbol de clasificación incluyendo las componentes principales. No se ha conseguido un árbol óptimo, y el árbol de menor error medio, es el mismo que en el apartado anterior. Sin embargo, la componente principal *RC1* ha entrado a la lista de variables importantes a la hora de creación del árbol. Esta componente está formada por las variables *HEMG*, *HEMGm*, *GLU* y *INS*.

3.7.4. Carnitinas

De la misma manera que antes, se calcula un árbol de clasificación para el conjunto de Carnitinas incluyendo las componentes principales. El árbol óptimo obtenido es el mismo que antes, sin utilizar las componentes, salvo que entran a la lista de variables importantes las componentes *RC1* y *RC4*, formadas respectivamente por las variables *C16n*, *C161On*, *C14n*, *C181n*, *C18n*, *C161n*, *C182n*, *C0n* y *C2n*; y las variables *C14On*, *C16On*, *C181On*, *C3DCn*, *C18On* y *C5DCn*.

3.8. Conclusiones

Después de todas las técnicas realizadas, se han ido obteniendo las siguientes conclusiones sobre las variables de los conjuntos de datos y su relación con la variable respuesta *apoE2*. Estas conclusiones se recogen a continuación organizadas por conjunto.

3.8.1. Ácidos grasos

El conjunto de datos de Ácidos grasos en principio tenia 47 variables. De estas 47 variables, se ha visto en análisis descriptivos tanto numéricos como gráficos que podían separar correctamente los grupos de *apoE2* las variables *v120H*, *v204P*, *v201P*, *v224P*, *v160H*, *v201H*, *v221H*, *g183H* y *v226H*.

Con el test de Kruskal-Wallis se ha visto que de estas variables, las variables *v201P*, *v224P*, *v201H* y *g183H*, pasaban el test, lo que indicaba con evidencia estadística que provienen de grupos distintos.

Se ha realizado un análisis de componentes principales y se han calculado 9 componentes del conjunto, pero se ha visto que al representar gráficamente las que más varianza explicaban no separaban los grupos de *apoE2* correctamente.

Después se ha realizado una regresión logística para todas las variables, y se ha visto que las variables que han entrado en el modelo han sido *v221H*, *g183H*, *v226H*, *v182P*,*v184H* y *a183P*.

Con el modelo de regresión calculado se han intentado añadir las componentes principales calculadas anteriormente a la regresión, pero no han entrado en el modelo final y el modelo de regresión logística ha quedado de la misma manera.

Se ha utilizado árboles de clasificación y con esta técnica se ha encontrado un árbol óptimo de dos variables, *v205P* y *g183H*. Cabe destacar, que en la creación del árbol, las variables *v160P* y *v161P* han sido las dos variables más importantes del árbol, por detrás de las variables que lo forman.

Una vez calculado el árbol óptimo, se ha intentado introducir las componentes calculadas en el análisis anterior, pero el árbol obtenido ha sido el mismo. Sin embargo, ha la componente *RC3* como variable importante, pero no con un papel importante. Esta componente está formada por las variables *v220P, v240P, v241P* y *v221P*.

Con todo esto se puede concluir que la variable más importante del conjunto de cara a separar por grupos de *apoE2* es la variable *g183H* y posiblemente entre en el modelo final con todos los conjuntos. También han destacado en menor medida variables como *v205P*, *v221H* o *v226H*.

Las variables que se utilizarán para calcular el modelo final del conjunto de Ácidos grasos son las variables *v221H*, *g183H*, *v226H*, *v182P*, *v184H*, *a183P*, *v205P*, *v160P* y *v161P*.

3.8.2. Aminoácidos

El conjunto de datos de Aminoácidos en principio tenía 26 variables. De estas variables, en análisis descriptivos tanto numéricos como gráficos se ha visto que podían separar correctamente los grupos de *apoE2* las variables *AcAlfaAminobutirico*, *Fosfoserina*, *Valina*, *Cistina*, *Arginina*, *Alanina* y *Prolina*.

Con el test de Kruskal-Wallis se ha obtenido que la única variable que lo ha pasado ha sido la variable *Alanina*.

Se ha realizado un análisis de componentes principales y se han calculado 4 componentes del conjunto, pero no lograban separar los grupos de *apoE2* correctamente.

Con la regresión logística que se ha calculado, han entrado al modelo las variables *Cistina*, *Alanina*, *Arginina* y *Glicina*. Cuando se añadían a la regresión las componentes principales, las componentes no entraban ni aportaban nada al modelo.

Cuando se han calculado árboles de regresión, no se ha obtenido ningún árbol óptimo que tuviera un error medio inferior a 1, pero el árbol que se ha calculado con menor error estaba formado por las variables *Cistina*, *Alanina*, *Valina*, *Isoleucina* y *Serina*, que también eran las variables más importantes de este. Una vez calculado el árbol, se ha intentado añadir las componentes principales y ni han entrado al modelo y ni han entrado como variables importantes del árbol.

A la vista de los resultados, se puede ver que la variable *Alanina* es la variable más importante del conjunto, puesto que aparece en todos los análisis realizados. También destacan las variables *Valina* y *Cistina*.

Las variables que se utilizarán para calcular el modelo final del conjunto de Aminoácidos son las variables *Valina*, *Cistina*, *Alanina*, *Isoleucina*, *Serina*, *Glicina*, *Arginina*, *RC3* y *RC4*.

3.8.3. Bioquímicas

El conjunto de datos de Bioquímicas tiene 22 variables. De estas variables, se ha visto en análisis descriptivos tanto numéricos como gráficos que podían separar correctamente los grupos de *apoE2* las variables *CTx*, *LPA*, *PCRU*, *APOB*, *CHOL* y *LDL*.

Con el test de Kruskal-Wallis se ha visto que de todas las variables, las variables *CTx*, *APOB*, *CHOL* y *VitD* pasaban el test lo que indica que provienen de grupos distintos.

Se ha realizado un análisis de componentes principales y se han calculado 8 componentes del conjunto, pero una vez que se representaban en el plano las más explicativas, no se separaban claramente los grupos de *apoE2* de ninguna manera.

Después se ha realizado una regresión logística para todas las variables, y, una vez realizada, se ha visto que las variables que han entrado en el modelo han sido *LPA*, *PCRU*, *APOB* y *VitD*. Estas 4 variables, ya las habíamos obtenido en los análisis anteriores, por lo que indica que pueden ser buenas candidatas al modelo final con todos los conjuntos. Cuando se intentaba realizar la regresión logística con las componentes principales, no se ha obtenido ningún resultado ya que no ha entrado ninguna componente en el modelo final.

Se ha utilizado árboles de clasificación y con esta técnica no se ha encontrado un árbol óptimo con un error medio menor que 1, pero se ha calculado el árbol con menor error, y en este árbol se han utilizado las variables *APOB*, *CHOL*, *VitD* y *BHID*. Cuando se han calculado las variables más importantes para la formación del árbol, se han obtenido, además de las que forman el árbol, las siguientes variables *BGP*, *GOT*, *LPA* y *CTx*. Cuando se han calculado árboles de clasificación incluyendo las componentes principales, no se ha obtenido ningún resultado con estas.

Con todos estos resultados se puede concluir que las variables *APOB*, *CHOL* y *VitD* son las más importantes del conjunto de datos, ya que han aparecido prácticamente en todas las técnicas que se han utilizado para identificar variables que separen por grupos el *apoE2*. También destacarían las variables *LPA* y *CTx* que han aparecido en un primer análisis y han sido marcadas como importantes en la creación del árbol, y la variable *BHID* que aparece en el árbol final.

Las variables que se utilizarán para calcular el modelo final del conjunto de Bioquímicas son las variables *APOB*, *CHOL*, *VitD*, *BHID*, *LPA*, *PCRU* y *BGP*, *GOT*.

3.8.4. Carnitinas

El conjunto de datos de Carnitinas tiene 31 variables. De estas, se ha visto en análisis descriptivos tanto numéricos como gráficos que podían separar correctamente los grupos de *apoE2* las variables *C4DCn*, *C18n*, *C3n*, *C12n* y *C14n*.

Con el test de Kruskal-Wallis se ha obtenido que las variables *C6n* y *C16On* pasaban el test y antes no habían sido marcadas como posibles para separar el *apoE2*.

Se ha realizado un análisis de componentes principales y se han calculado 6 componentes del conjunto, pero no se ha visto evidencias de que separasen los grupos de *apoE2* correctamente.

Después se ha realizado una regresión logística para todas las variables, y se ha visto que las más variables que han entrado en el modelo han sido *C4DCn*, *C14n*, *C6n*, *C16On*, *C182n*, *C5n* y *C18On*, de las cuales las cuatro primeras ya las habíamos detectado en los análisis anteriores. Intentando añadir las componentes principales no se ha obtenido éxito, ninguna de ellas ha entrado al modelo logístico final.

Utilizando árboles de clasificación se ha encontrado un árbol óptimo de seis variables, *C14n*, *C6n*, *C182n*, *C5n*, *C161On* y *C3DCn*. A la hora de ver la importancia de las variables, además de las que han entrado al modelo, han sido seleccionadas las siguientes variables *C16n*, *C2n* y *C181n*.

Una vez calculado el árbol optimo, se ha intentado introducir las componentes calculadas en el análisis anterior, pero el árbol obtenido ha sido el mismo, salvo que como variables importantes han sido marcadas las componentes *RC1* y *RC6*, formadas respectivamente por las variables *C16n*, *C161On*, *C14n*, *C181n*, *C18n*, *C161n*, *C182n*, *C0n* y *C2n*; y las variables *C14On*, *C16On*, *C181On*, *C3DCn*, *C18On* y *C5DCn*. Estas dos componentes han obtenido una importancia superior a la de alguna variable que había entrado al modelo.

Con todos estos análisis se puede concluir que las variable más importantes del conjunto son *C16On* y *C14n*, ya que han aparecido en algún análisis anterior y han entrado en ambos modelos propuestos. También destacarían las variables *C182n* y *C5n*, que a pesar de no aparecer en los primeros análisis, han entrado también en ambos modelos.

Las variables que se utilizarán para calcular el modelo final del conjunto de Carnitinas son las variables *C14n*, *C6n*, *C16On*, *C182n*, *C5n*, *C18On*, *C161On*, *C3DCn*, *C4DCn*, *C2n*, *C181n*, *RC1* y *RC4*.

Capítulo 4

Análisis global: Cálculo del modelo final

En este capítulo trataremos de crear un modelo final con todos los conjuntos de datos utilizados anteriormente, pero utilizando únicamente las variables que hemos visto que eran más influyentes de cada conjunto. De esta manera se evitan cálculos innecesarios para variables que, como se ha visto nada, no aportan nada a la clasificación por grupos de la variable *apoE2*.

Las variables que se han obtenido como finalistas de cada conjunto son las siguientes:

- Ácidos grasos: *v221H*, *g183H*, *v226H*, *v182P*, *v184H*, *a183P*, *v205P*, *v160P* y *v161P*.
- Aminoácidos: *Valina*, *Cistina*, *Alanina*, *Isoleucina*, *Serina*, *Glicina*, *Arginina*, *AmiRC3* y *AmiRC4*.
- Bioquímicas: *APOB*, *CHOL*, *VitD*, *BHID*, *LPA*, *PCRU*, *BGP* y *GOT*.
- Carnitinas: *C14n*, *C6n*, *C16On*, *C182n*, *C5n*, *C18On*, *C161On*, *C3DCn*, *C4DCn*, *C2n*, *C181n*, *CarRC1* y *CarRC4*.

De cara a evitar confusiones con los nombres de las componentes principales, se han renombrado con las tres primeras letras del nombre de su conjunto, así, por ejemplo, la componente *RC1* del conjunto de Ácidos grasos pasaría a llamarse *AciRC1*.

Con estas 39 variables finalistas de cada conjunto, se va a proceder a crear un modelo de regresión logística y un árbol de clasificación para obtener las variables más influyentes y predictivas de todos los conjuntos en común.

4.1. Regresión Logística Multinomial

Se va a realizar una regresión logística con todas las variables indicadas anteriormente. El problema de esta técnica es el tratamiento de los valores *missing*. El algoritmo no trata los datos faltantes, por ello hay que eliminar las observaciones con alguna variable sin informar.

El conjunto de Aminoácidos tiene un porcentaje muy elevado de datos faltantes, y, por eso, se va a realizar dos modelos de regresión logística, el primero utilizando las variables del conjunto de Aminoácidos, y el segundo sin utilizar este conjunto.

De esta manera, en un modelo utilizaremos todas las variables de los conjuntos, tendremos más variables pero se perderán observaciones con algún dato faltante, y en el otro modelo no utilizaremos el conjunto Aminoácidos, perdemos variables pero tendremos más observaciones con las que modelizar.

Para el modelo con el conjunto de Aminoácidos, se realizará una regresión logística hacia delante para ver que variables entran en el modelo.

Para el modelo sin el conjunto de Aminoácidos, se generarán dos regresiones logísticas paso a paso, la primera será hacia delante para ver que variables entran de primeras al modelo, y la segunda será *backward/forward* para ver con qué variables decide quedarse el modelo.

4.1.1. Modelo Logístico completo

Con las 39 variables, se construye el modelo y se obtienen los resultados que se pueden ver en el Anexo F.1.

En este modelo entran 6 variables que son *g183H, v182P* y *v226H* del conjunto de Ácidos grasos, *APOB* y *VitD* del conjunto de Bioquímicas, y *C16On* del conjunto de Carnitinas. No entra ninguna variable del conjunto de Aminoácidos en el modelo.

Las variables que pasan el test de Wald de significación son las variables *g183H* y *v226H* del conjunto de Ácidos grasos, *APOB* y *VitD* del conjunto de Bioquímicas y *C16On* del conjunto de Carnitinas.

Si calculamos el poder de predicción del modelo a través de su matriz de confusión, en la Tabla 4.1, podemos comprobar que el modelo falla en un 24,07 % de los casos. En la predicción se puede comprobar que el modelo no tiende a clasificar todas las observaciones como *E3* como lo hacían los modelos anteriores. También se ve que, aunque se ha perdido un poco de poder de predicción del grupo *E3*, se ha ganado en los otros dos grupos minoritarios. Este hecho marca la diferencia con los modelos anteriores, ya que antes las observaciones de los grupos *E2* y *E4* se clasificaban la gran mayoría como *E3*.

		Predicción		
		E2	E3	E4
apoE2	E2	4	13	4
	E3	4	110	4
	E4	2	12	9

Tabla 4.1: Matriz de confusión del modelo de regresión logística con todos los conjuntos.

4.1.2. Modelo Logístico sin el conjunto de Aminoácidos

En esta sección, se procederá como la sección anterior y se realizarán dos regresiones paso a paso. En esta sección se eliminarán las variables del conjunto de Aminoácidos y perderemos riqueza de variables pero ganaremos observaciones.

Ahora, con la eliminación de las variables del conjunto de Aminoácidos se tienen 30 variables con las que modelizar. Los resultados de ambas regresiones están en los Anexos F.2.1 y F.2.2 correspondientes a los modelos hacia delante y *backward/forward* respectivamente.

En el primer modelo, el modelo hacia delante, entran 9 variables, que son las variables *g183H*, *v182P*, *v221H*, *v226H* y *v184H* del conjunto de Ácidos grasos, las variables *VitD* y *CHOL* del conjunto de Bioquímicas, y las variables *C16On* y *C6n* del conjunto de Carnitinas.

De este modelo generado, las variables que pasan el test de Wald, y por tanto son significativas, las variables *g183H, v221H*, y *v184H* del conjunto de Ácidos grasos, las variables *CHOL* y *VitD* del conjunto de Bioquímicas y las variables *C16On* y *C6n* del conjunto de Carnitinas.

Por otra parte, en el modelo *backward/forward*, se obtienen 17 variables de las 30 posibles. Estas variables son las variables *v221H*, *g183H*, *v226H*, *v182P*, *v184H* y *v161P* del conjunto de Ácidos grasos, las variables *CHOL*, *VitD* y *BHID* del conjunto de Bioquímicas y las variables *C14n*, *C6n*, *C16On*, *C18On*, *C2n*, *CarRC1* y *CarRC4* del conjunto de Carnitinas.

En este segundo modelo generado, las variables significativas, según el test de Wald son *v221H*, *g183H*, *v226H* y *v184H* del conjunto de Ácidos grasos, *CHOL* y *VitD* del conjunto de Bioquímicas, y *C14n*, *C6n*, *C16On*, *C5n, C18On, C2n*, *CarRC1* y *CarRC4*.

Las matrices de confusión de ambos modelo se pueden ver en la Tabla 4.2. En ellas se ve que el primer modelo falla al predecir en un 15,82 % de los casos, y el segundo modelo solo en un 13,92 %, lo que indica que el segundo modelo es mejor. En ambos modelos obtenidos, se puede ver que se empiezan a clasificar de mejor manera los grupos menos representativos, y se puede ver que el segundo mejor es mejor, ya que obtiene un menor error de predicción y una mejor clasificación de los grupos con menos representación.

(a)		**Predicción**		
		E2	E3	E4
apoE2	E2	7	12	2
	E3	1	115	1
	E4	0	9	11

(b)		**Predicción**		
		E2	E3	E4
apoE2	E2	8	11	2
	E3	1	114	
	E4	1	5	14

Tabla 4.2: Matriz de confusión de los modelos de regresión logística sin el conjunto de Aminoácidos. La matriz (a) corresponde al modelo hacia delante y la matriz (b) al modelo *backward/forward*.

4.2. Árbol de clasificación

Con todas las variables seleccionadas de los conjuntos, se procede a realizar un árbol de clasificación. El objetivo es encontrar un árbol con un error medio menor que 1, pero si no se pudiese encontrar, se procedería a buscar algún otro árbol con el menor error medio posible.

Realizando la técnica, se ha obtenido un árbol óptimo con 11 *splits* y puede verse en la Figura 4.1.

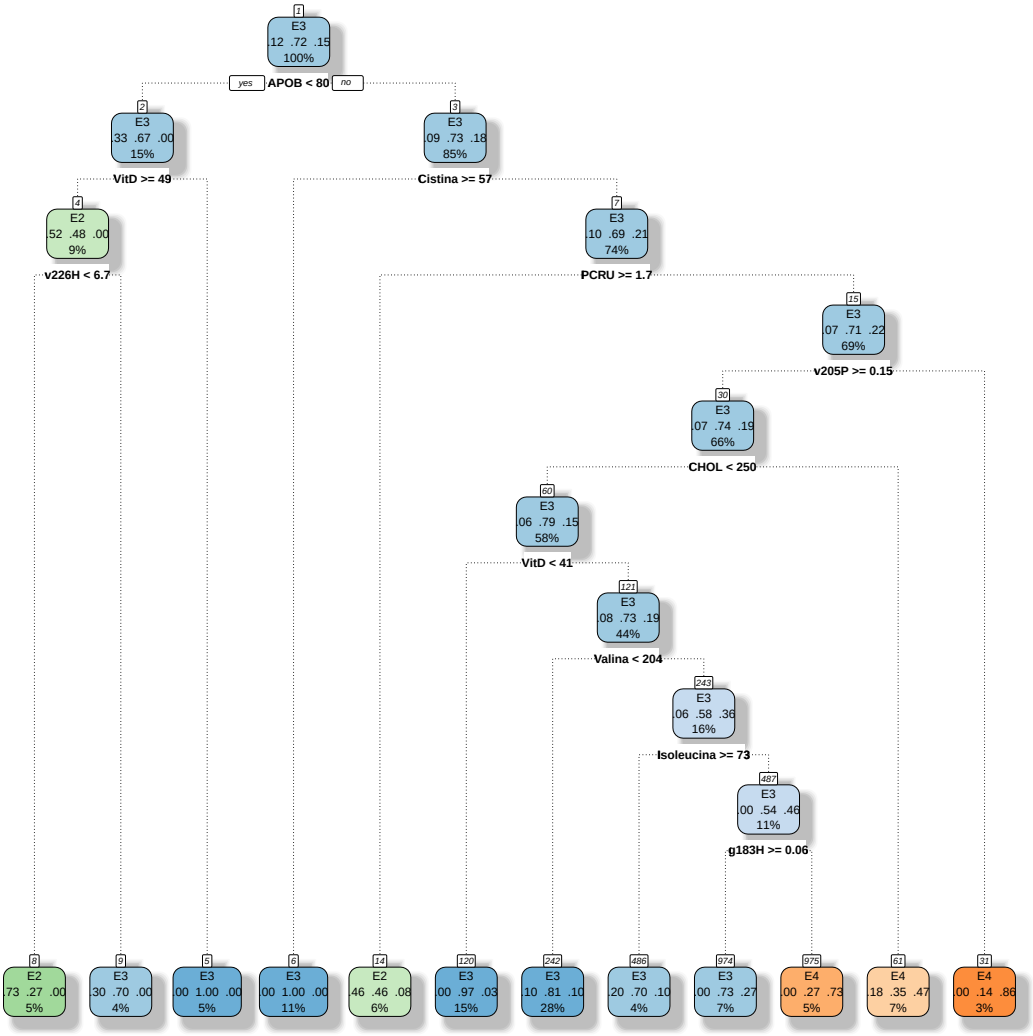

Figura 4.1: Árbol de clasificación óptimo obtenido con todos los conjuntos.

El árbol está formado por las variables *g183H* , *v205P* y *v226H* del conjunto de Ácidos grasos, las variables *Cistina*, *Isoleucina* y *Valina* del conjunto de Aminoácidos, y las variables *APOB*, *CHOL*, *PCRU* y *VitD* del conjunto de Bioquímicas. Del conjunto de Carnitinas no ha resultado ninguna variable.

A la hora de crear los *splits*, las 15 variables más importantes del árbol son *g183H* , *v205P*, *v226H* y *v221H* del conjunto de Ácidos grasos, las variables *Cistina*, *Isoleucina* y *Valina* del conjunto de Aminoácidos, las variables *APOB*, *CHOL*, *PCRU* y *VitD* del conjunto de Bioquímicas, y las variables *C181n, C2n, C3DCn* y *C5n* del conjunto de Carnitinas.

Como puede verse, no ha entrado ninguna variable del conjunto de Carnitinas al árbol óptimo resultante, pero sí que han entrado como de las variables más importantes a la hora de crear los *subrrogate splits*. Las variables que han entrado para este fin, pero no han entrado en el árbol son *v221H, C181n, C3DCn, C2n* y *C5n*.

El árbol tiene un poder predictivo alto y predice de manera asumible los grupos de *apoE2*, como se puede ver en la matriz de confusión en la Tabla 4.3. El árbol erra en un 20,70 % de los casos, un error menor que el del nodo raíz, que es un 27,75 %. En este árbol se ven evidencias de una buena clasificación de los conjuntos, ya que en los árboles anteriores se clasificaba la mayoría como el grupo mayoritario, el grupo *E3*, y el porcentaje de correcta predicción global era elevado. En este árbol no solo se consigue un poder predictivo total elevado, sino que se consiguen un porcentaje mayor o igual al 50 % de acierto en todos los grupos, algo que hasta ahora únicamente se conseguía en el grupo *E3*.

		Predicción		
		E2	E3	E4
apoE2	E2	14	11	3
	E3	9	142	13
	E4	1	10	24

Tabla 4.3: Matriz de confusión del árbol de clasificación óptimo obtenido.

4.3. Conclusiones

Como se ha comentado en la introducción, el objetivo del presente trabajo ha sido doble: por un lado establecer qué variables de las observadas influyen en el fenotipo APOE y por otro predecir con dichas variables, si es posible, el tipo de APOE que presenta un individuo. A lo largo del trabajo se han aplicado distintas técnicas para abordar los objetivos propuestos. En primer lugar, se ha realizado un análisis individual de cada variable, es decir, se han aplicado las técnicas habituales para analizar la influencia individual de cada variable. Con este primer análisis se ha comprobado que individualmente la influencia de las variables sobre el fenotipo APOE no era muy evidente.

Para el análisis conjunto, debido al elevado número de variables del que disponíamos (con no demasiados individuos) se ha realizado la selección de las variables en dos fases: la primera dentro de cada conjunto de variables y la segunda a partir de todas la variables seleccionadas en la primera fase. Además de la dificultad del alto número de variables frente al bajo tamaño de la muestra, se añade

el problema de los datos faltantes en algunos conjuntos de variables. Las técnicas aplicadas en la selección de variables influyentes en el fenotipo APOE, resuelve el problema de los datos faltantes de dos maneras distintas: en la regresión logística se eliminan y en los árboles de clasificación se puede trabajar con datos faltantes sin necesidad de ser eliminados.

Aplicadas estas técnicas en cada conjunto de datos y posteriormente en el análisis global, hemos obtenido los siguientes resultados.

Utilizando modelos de regresión logística se han obtenido tres modelos. En el primero se ha utilizado el conjunto de Aminoácidos, que estaba caracterizado por tener mucho dato sin informar e impedía el correcto funcionamiento al algoritmo. En los dos últimos se han eliminado las variables de este conjunto con el fin de que algoritmo pudiera trabajar mejor.

Con el modelo utilizando las variables del conjunto de Aminoácidos se ha obtenido un error de predicción del 24 %. El modelo obtenido estaba formado por seis variables. que son las variables *g183H*, *v182P*, *v226H*, *APOB*, *VitD* y *C16On*. Es decir, en el modelo utilizando el conjunto de Aminoácidos, no entra ninguna variable de Aminoácidos.

Con los modelos realizados sin el conjunto de Aminoácidos se han obtenido dos modelos de nueve y quince variables con errores de predicción del 16 % y del 14 %. Las variables que coinciden en ambos modelos son *g183H*, *v182P*, *v184H*, *v221H* y *v226H* del conjunto de Ácidos grasos, las variables *CHOL* y *VitD* del conjunto de Bioquímicas, y las variables *C16On* y *C6n* del conjunto de Carnitinas. De estas variables que coinciden, en ambos modelos son significativas las variables *v184H, g183H*, *v221H, CHOL, VitD*, *C16On* y *C6n*.

La conclusión que se puede sacar de los cuatro modelos es que la mayor parte de las variables que se han seleccionado para estos últimos modelos son importantes, pero cabe destacar el comportamiento de las variables *g183H, VitD*, *C16On* y *C6n* que han aparecido en todos los modelos que se han generado y han sido variables significativas. También se destaca que las variables de los conjuntos de Ácidos grasos y Carnitinas son las más importantes. También hay que destacar el papel de variables como *v226H, v221H, v182P, APOB, CHOL*, *C182n* y *C5n*, que aunque no hayan resultado ser significativas en alguno de los modelos han ido apareciendo en la mayoría de los que se han generado desde el principio.

Respecto a los árboles de clasificación, se ha obtenido un árbol de once *splits* y doce nodos terminales. En este modelo no se utiliza ninguna variable del conjunto de Carnitinas, pero sin embargo, sí que se hace un gran uso de las variables del conjunto de Aminoácidos. En este árbol han participado variables han estado presentes en los distintos árboles que se han generado. Estas variables tan presentes en todos los árboles son *g183H*, *Valina*, *Cistina*, *APOB*, *VitD* y *CHOL*.

El modelo generado por el árbol de clasificación utilizando once variables, y entre ellas, estas variables seis variables comentadas, tiene una poder predictivo alto, con un 21 % de fallo. Sin embargo, lo bueno de este modelo es que se ha conseguido un modelo donde no solo no se clasifican la gran mayoría de observaciones al grupo mayoritario, sino que se consigue evitar el impedimento de los valores faltantes con los *surrogate splits* y se obtiene un modelo que predice cada grupo correctamente por separado con una tasa de acierto del 50 % o más.

En resumen, de ambos modelos se pueden destacar como variables influyentes en el gen APOE las variables *g183H* (ácido *gamma*-linoleico en glóbulos rojos), *v184H* (ácido octadecatetraenoico en

glóbulos rojos), *v221H* (ácido erúcico en glóbulos rojos), *v226H* (ácido docosahexaenoico en glóbulos rojos) y *v182P* (ácido linoleico en plasma) del conjunto de Ácidos grasos, las variables *Valina* y *Cistina* del conjunto de Aminoácidos, las variables *APOB* (apolipoproteína B en suero), *CHOL* (colesterol en suero) y *VitD* (vitamina D en suero) del conjunto de Bioquímicas y las variables *C16On* (acilcarnitina C16-OH en sangre) y *C6n* (acilcarnitina C6 en sangre) del conjunto de Carnitinas.

Bibliografía

[1] TORRES-PEREZ et al. The FAT expandability (FATe) Project: Biomarkers to determine the limit of expansion and the complications of obesity. *Cardiovasc Diabetol* 2015 14: 40.

[2] R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[3] JIMENO, A. (2009). *Biología 2, Bachillerato* (primera edición). Madrid: Santillana.

[4] Carnitina. (2017, 22 de enero). *Wikipedia, La enciclopedia libre*. Recuperado de https://es.wikipedia.org/w/index.php?title=Carnitina&oldid=96371175.

[5] Apolipoprotein E. (2017, 4 de mayo). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Apolipoprotein_E&oldid=778723012.

[6] Multivariate statistics. (2017, 3 de agosto). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Multivariate_statistics&oldid=793719135.

[7] FOX, J. (2016). *RcmdrMisc: R Commander Miscellaneous Functions. R package version 1.0-5*. URL: https://CRAN.R-project.org/package=RcmdrMisc.

[8] GROSS, J., LIGGES, U. (2015). *nortest: Test for Normality. R package version 1.0-4*. URL: https://CRAN.R-project.org/package=nortest.

[9] FOX, J., WEISBERG, S. (2011). *An R Companion to Applied Regression, Segunda Edición*. Thousand Oaks (CA): Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[10] REVELLE, W. (2017). *psych: Procedures for Personality and Psycological Research. R package version 1.7.5*. Northwestern University, Evanston, Illinois, USA. URL: https://CRAN.R-project.org/package=psych.

[11] PEÑA, D. (2010). *Análisis de datos multivaiantes*. Madrid [etc.]: McGraw-Hill.

[12] VENABLES, W. N., RIPLEY, B. D. (2002). *Modern Applied Statistics with S, Cuarta Edición*. Nueva York: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[13] HOSMER, D., LEMESHOW, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, Inc.

[14] Logistic regression. (2017, 24 de noviembre). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=811862357.

[15] BREIMAN, L., FRIEDMAN, J., OLSHEN, R., STONE, C. (1984). *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group.

[16] Decision tree learning. (2017, 13 de junio). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=785380783.

[17] Árbol de decisión. (2017, 21 de mayo). *Wikipedia, la enciclopedia libre*. Recuperado de https://es.wikipedia.org/w/index.php?title=%C3%81rbol_de_decisi%C3%B3n&oldid=99282090.

[18] THERNEAU, T.M., ATKITSON, E.J. (12 marzo 2017). *An Introduction to Recursive Partitioning Using the RPART Routine*. Mayo Foundation. http://www.mayo.edu/research/documents/rpartminipdf/doc-10027257.

[19] GRAHAM, J. W. (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)*. Springer.

[20] THERNEAU, T., ATKINSON, B., RIPLEY, B. (2017). *rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11*.

[21] MIÑARRO, A. (Enero 1998). *Estimación no paramétrica de la función de densidad*. Universidad de Barcelona. http://www.ub.edu/stat/personal/minarro/documents/Nonpar.pdf.

[22] Kruskal-Wallis one-way analysis of variance. (2017, 6 de abril). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Kruskal%E2%80%93Wallis_one-way_analysis_of_variance&oldid=774108236.

[23] Pearson correlation coefficient. (2017, 18 de noviembre). *Wikipedia, The Free Encyclopedia*. Recuperado de https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=810884273.

[24] HARRELL JR, F.E. con contribuciones de DUPONT, C. et al. (2017). *Hmisc: Harrell Miscellaneous. R package version 4.0-3*. URL: https://CRAN.R-project.org/package=Hmisc.

Anexos

Anexo A

Tablas de distribución de las variables

Acontinuación se recogen las distribuciones de las variables, donde se recogen porcentaje de datos faltantes, media, mediana, desviación típica y percentiles. Estos datos se agrupan por grupo de la variable *apoE2*. Los correspondientes al grupo *E2* se recogen en el Anexo A.1, los correspondientes al *E3* en el Anexo A.2 y los correpsondientes al grupo *E4* en el Anexo A.3.

A.1. Grupo Apoe E2

Las distribuciones de las variables de los casos con alelo *E2* son:

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v140P	10,7143 %	0,4120	0,2484	0,1200	0,1200	0,1400	0,1500	0,2700	0,3500	0,4900	0,5800	1,1000	1,1400	1,1400
v160P	10,7143 %	19,5716	2,1458	15,9400	15,9400	16,1900	17,0200	17,7900	19,9700	20,8500	21,7800	22,8300	25,0600	25,0600
v161P	10,7143 %	2,0784	1,2908	1,1800	1,1800	1,1900	1,2600	1,5200	1,7200	2,0900	3,1300	3,6100	7,6200	7,6200
v180P	10,7143 %	7,8572	1,0310	5,4200	5,4200	6,3100	6,6800	7,2000	7,8200	8,6500	9,0500	9,3900	9,7600	9,7600
v181P	10,7143 %	26,1176	4,1874	19,2500	19,2500	20,8900	21,2300	22,7900	26,0800	28,7100	30,9900	32,9800	37,0000	37,0000
v182P	10,7143 %	24,0496	4,3147	12,8100	12,8100	17,7900	20,0000	21,7700	23,3000	27,0800	29,6400	30,6300	31,6600	31,6600
g183P	10,7143 %	0,3104	0,1452	0,1000	0,1000	0,1400	0,1700	0,2300	0,2900	0,4000	0,4600	0,5100	0,7800	0,7800
a183P	21,4286 %	0,2073	0,0846	0,1000	0,1000	0,1200	0,1200	0,1300	0,1950	0,2400	0,3200	0,3300	0,4200	0,4200
v184P	21,4286 %	0,0623	0,0349	0,0100	0,0100	0,0100	0,0200	0,0400	0,0600	0,0900	0,0900	0,1200	0,1500	0,1500
v200P	10,7143 %	0,2936	0,1229	0,1100	0,1100	0,1500	0,2100	0,2300	0,2500	0,3300	0,4600	0,6000	0,6200	0,6200
v201P	10,7143 %	0,2620	0,0710	0,1100	0,1100	0,1600	0,1700	0,2200	0,2700	0,3300	0,3500	0,3700	0,3800	0,3800
v202P	10,7143 %	0,3468	0,1044	0,0800	0,0800	0,2100	0,2500	0,3000	0,3400	0,4000	0,4400	0,4800	0,6500	0,6500
v203P	10,7143 %	2,0672	0,5797	1,0200	1,0200	1,4300	1,4400	1,6600	1,9700	2,4300	3,0100	3,1100	3,2100	3,2100
v204P	10,7143 %	9,6272	2,3170	5,8400	5,8400	5,9400	6,5700	8,2800	9,5600	11,4500	13,6100	13,6600	14,1300	14,1300
v205P	10,7143 %	0,4600	0,2596	0,1200	0,1200	0,1500	0,1700	0,2400	0,4600	0,5300	0,8700	1,0000	1,1700	1,1700
v220P	10,7143 %	0,6424	0,2629	0,3000	0,3000	0,3200	0,3400	0,4600	0,5700	0,9200	1,0100	1,1300	1,1800	1,1800
v221P	10,7143 %	0,1536	0,1070	0,0400	0,0400	0,0500	0,0500	0,0900	0,1200	0,1800	0,3200	0,4100	0,4800	0,4800
v224P	10,7143 %	0,3168	0,1050	0,1100	0,1100	0,1300	0,1800	0,2500	0,3100	0,3900	0,4500	0,4700	0,5300	0,5300
v225P	21,4286 %	0,4523	0,1455	0,2500	0,2500	0,2700	0,2900	0,3200	0,4550	0,5300	0,6700	0,7300	0,7500	0,7500
v240P	10,7143 %	0,5620	0,2350	0,1900	0,1900	0,2200	0,3000	0,4100	0,5000	0,6200	0,9400	0,9500	1,0400	1,0400

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v226P	10,7143 %	2,7764	1,0208	1,6000	1,6000	1,6100	1,6300	1,8400	2,8400	3,5200	3,7500	3,9500	5,8400	5,8400
v241P	10,7143 %	1,1584	0,5745	0,4400	0,4400	0,4800	0,5700	0,7600	0,9400	1,5200	1,9500	2,2800	2,6600	2,6600
v260P	10,7143 %	0,0652	0,0340	0,0100	0,0100	0,0100	0,0200	0,0500	0,0600	0,0900	0,1000	0,1300	0,1500	0,1500
v120H	21,4286 %	0,0245	0,0250	0,0100	0,0100	0,0100	0,0100	0,0100	0,0150	0,0400	0,0400	0,0500	0,1200	0,1200
v140H	21,4286 %	0,2414	0,3499	0,0200	0,0200	0,0700	0,0800	0,1200	0,1700	0,2300	0,3600	0,3600	1,7600	1,7600
v160H	21,4286 %	16,9527	2,4404	12,6700	12,6700	12,8100	13,7400	15,4400	16,7700	19,0100	19,7400	19,9400	22,2200	22,2200
v161H	21,4286 %	0,4377	0,1954	0,2200	0,2200	0,2500	0,2600	0,3000	0,3650	0,5400	0,7200	0,7700	0,9900	0,9900
v180H	21,4286 %	15,6586	1,2977	13,3600	13,3600	13,7700	14,2200	14,6500	15,7100	16,7700	17,3000	17,5800	18,1900	18,1900
v181H	21,4286 %	14,4532	2,2772	11,2400	11,2400	11,5100	12,0000	12,8800	14,4500	15,6100	16,5000	19,3500	20,0600	20,0600
v182H	21,4286 %	8,7914	1,8868	4,6500	4,6500	6,4300	6,9100	7,6400	8,5700	10,1600	11,0600	11,9700	12,1900	12,1900
g183H	21,4286 %	0,0568	0,0383	0,0100	0,0100	0,0200	0,0200	0,0300	0,0500	0,0700	0,1000	0,1100	0,1800	0,1800
a183H	21,4286 %	0,0509	0,0254	0,0100	0,0100	0,0100	0,0200	0,0300	0,0500	0,0700	0,0900	0,0900	0,0900	0,0900
v184H	21,4286 %	0,1145	0,0842	0,0100	0,0100	0,0200	0,0400	0,0700	0,0850	0,1500	0,2100	0,2400	0,3800	0,3800
v200H	21,4286 %	0,4218	0,1057	0,2200	0,2200	0,2200	0,3000	0,3600	0,4200	0,4900	0,5400	0,5900	0,6100	0,6100
v201H	21,4286 %	0,3709	0,0974	0,2400	0,2400	0,2500	0,2600	0,2800	0,3550	0,4600	0,4700	0,4900	0,5800	0,5800
v202H	21,4286 %	0,3986	0,1130	0,1300	0,1300	0,2700	0,2900	0,3100	0,4000	0,4800	0,5400	0,5600	0,5800	0,5800
v203H	21,4286 %	2,0045	0,4197	1,4100	1,4100	1,4200	1,4900	1,6800	1,9800	2,1600	2,6800	2,7100	2,8600	2,8600
v204H	21,4286 %	17,4964	2,1694	13,7600	13,7600	14,6000	15,0800	15,7500	17,4350	18,6500	20,6300	21,0100	21,7400	21,7400
v205H	21,4286 %	0,4986	0,2114	0,2500	0,2500	0,2600	0,2700	0,3200	0,4750	0,5400	0,8200	0,8400	1,0100	1,0100
v220H	21,4286 %	1,2582	0,3487	0,7100	0,7100	0,7800	0,8400	1,0600	1,2650	1,4900	1,6500	1,9100	2,0000	2,0000
v221H	21,4286 %	0,0873	0,0392	0,0300	0,0300	0,0300	0,0400	0,0600	0,0800	0,1200	0,1300	0,1500	0,1700	0,1700
v224H	21,4286 %	4,2114	1,1022	2,3900	2,3900	2,6200	3,0300	3,1600	4,1950	4,9700	5,6200	5,8400	6,2500	6,2500
v225H	21,4286 %	2,0055	0,4316	1,3000	1,3000	1,4300	1,5700	1,7200	1,9400	2,1700	2,5500	2,8500	3,0800	3,0800
v240H	21,4286 %	4,1282	2,0086	1,6100	1,6100	1,7100	1,7500	2,1800	4,0000	5,2200	6,5400	7,6200	8,5300	8,5300
v226H	21,4286 %	5,7477	1,7412	3,6100	3,6100	3,6800	3,7800	4,2100	5,4200	6,6800	7,9200	9,3600	9,4500	9,4500
v241H	21,4286 %	4,1359	1,5679	1,2900	1,2900	1,5500	2,1400	2,9400	4,2700	5,1300	6,0600	6,1500	6,9900	6,9900
v260H	21,4286 %	0,3205	0,2384	0,0200	0,0200	0,0300	0,0400	0,1300	0,2550	0,4600	0,7000	0,7300	0,7900	0,7900
Fosfoserina	32,1429 %	22,0146	8,3531	8,4410	8,4410	8,4410	13,6470	17,270	20,5960	24,8310	30,3980	49,1690	49,1690	49,1690

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
Taurina	32,1429 %	50,6308	19,2691	24,9000	24,9000	24,9000	35,2130	41,6340	48,1580	52,9060	66,3650	121,1560	121,1560	121,1560
AcAspartico	32,1429 %	3,9947	3,6858	1,5020	1,5020	1,5020	1,6950	1,7960	3,1020	4,2690	7,8670	17,8080	17,8080	17,8080
Treonina	32,1429 %	106,9616	16,6382	77,7160	77,7160	77,7160	80,8250	93,1850	110,1300	118,5610	130,5210	133,8000	133,8000	133,8000
Serina	32,1429 %	100,7371	24,1355	68,2300	68,2300	68,2300	68,6790	82,1520	96,4090	118,3640	136,4880	150,5120	150,5120	150,5120
Asparragina	32,1429 %	80,7604	14,1839	55,4470	55,4470	55,4470	55,9660	69,9280	79,2620	95,1710	98,3160	101,9420	101,9420	101,9420
AcGlutamico	32,1429 %	51,4079	28,3305	5,5740	5,5740	5,5740	22,0030	35,7100	43,1700	60,0890	91,3960	135,6980	135,6980	135,6980
Glutamina	32,1429 %	530,3586	71,3619	386,5190	386,5190	386,5190	442,5490	478,5510	521,3050	566,5380	630,3970	685,0090	685,0090	685,0090
Glicina	32,1429 %	186,1209	59,5002	125,8200	125,8200	125,8200	130,1120	141,1560	157,4160	233,5190	286,3500	342,9260	342,9260	342,9260
Alanina	32,1429 %	332,3042	76,7386	209,1830	209,1830	209,1830	219,4030	284,3350	333,5550	366,0650	429,8250	546,9390	546,9390	546,9390
Citrulina	32,1429 %	25,4389	9,8955	11,9240	11,9240	11,9240	14,2900	15,4000	25,5450	32,5070	43,6370	45,0730	45,0730	45,0730
AcAlfaAminobutirico	32,1429 %	15,6105	7,6131	4,7360	4,7360	4,7360	6,0930	10,4530	15,0320	18,2150	24,2230	39,5850	39,5850	39,5850
Valina	32,1429 %	190,4123	32,5095	138,2480	138,2480	138,2480	140,3810	171,5290	178,8880	220,8560	243,9100	245,9560	245,9560	245,9560
Cistina	32,1429 %	41,1319	6,1281	25,7750	25,7750	25,7750	32,4900	39,4220	41,8480	43,4270	48,5150	54,9190	54,9190	54,9190
Metionina	32,1429 %	17,5422	3,7569	10,1230	10,1230	10,1230	12,0170	15,2260	17,6550	20,5600	22,3000	26,8110	26,8110	26,8110
Isoleucina	32,1429 %	57,7009	15,5012	30,4110	30,4110	30,4110	41,1180	44,7500	55,4470	71,8080	82,9690	85,6000	85,6000	85,6000
Leucina	32,1429 %	106,7452	28,7207	57,6210	57,6210	57,6210	69,6250	89,6500	99,9040	129,1570	157,0920	158,4210	158,4210	158,4210
Tirosina	32,1429 %	55,0612	14,8115	36,9260	36,9260	36,9260	42,5060	45,6440	52,3350	61,5380	73,7080	103,0550	103,0550	103,0550
FenilAlanina	32,1429 %	43,9916	6,7258	33,0680	33,0680	33,0680	36,0290	38,7930	45,0000	48,1750	56,5770	57,5280	57,5280	57,5280
Ornitina	32,1429 %	92,5248	37,1061	46,8750	46,8750	46,8750	57,2120	67,2590	88,2240	105,7820	159,2140	201,3740	201,3740	201,3740
Lisina	32,1429 %	184,2178	42,4856	140,2000	140,2000	140,2000	141,9200	148,6070	179,7120	198,8410	270,0520	292,6820	292,6820	292,6820
X1MetilHistidina	32,1429 %	10,2429	5,8056	0,3620	0,3620	0,3620	2,1390	4,4640	10,7320	14,6170	19,0350	19,8850	19,8850	19,8850
Histidina	32,1429 %	67,0919	9,5672	55,3010	55,3010	55,3010	56,1980	60,6210	64,2880	71,3760	86,7590	91,7480	91,7480	91,7480
Triptofano	32,1429 %	34,9526	13,7586	11,4190	11,4190	11,4190	18,8560	25,9130	34,9170	45,6530	56,9890	59,0850	59,0850	59,0850
Arginina	32,1429 %	43,3823	35,2550	6,0760	6,0760	6,0760	6,2930	18,2830	34,8980	60,7370	75,0930	159,0060	159,0060	159,0060
Prolina	32,1429 %	192,6406	68,3348	105,7300	105,7300	105,7300	116,2060	151,4500	170,9210	232,4790	307,4840	359,7320	359,7320	359,7320
PCRU	7,1429 %	0,8738	0,8130	0,0800	0,0800	0,1100	0,1300	0,3300	0,5150	1,2000	1,9200	2,7600	3,0100	3,0100
GLU	0,0000 %	107,7143	35,5412	60,0000	60,0000	65,0000	75,0000	86,5000	102,5000	112,5000	151,0000	186,0000	229,0000	229,0000
TRIG	14,2857 %	123,5542	45,1596	73,8000	73,8000	74,0000	76,0000	89,7500	109,0000	157,0000	183,0000	209,0000	240,0000	240,0000

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
CHOL	0,0000 %	181,8214	45,2762	111,0000	111,0000	115,0000	124,0000	143,5000	177,0000	210,5000	255,0000	261,0000	283,0000	283,0000
cHDL	21,4286 %	49,0909	9,4512	29,0000	29,0000	32,0000	38,0000	44,0000	50,0000	54,0000	60,0000	61,0000	70,0000	70,0000
LDL	25,0000 %	109,9524	46,0472	37,0000	37,0000	46,0000	64,0000	78,0000	98,0000	134,0000	172,0000	177,0000	202,0000	202,0000
GGT	7,1429 %	29,6154	22,2316	9,0000	9,0000	10,0000	11,0000	17,0000	23,0000	28,0000	69,0000	78,0000	100,0000	100,0000
GOT	7,1429 %	31,9615	12,7984	17,0000	17,0000	18,0000	19,0000	22,0000	28,5000	39,0000	52,0000	59,0000	64,0000	64,0000
GPT	3,5714 %	35,3704	19,4049	9,0000	9,0000	12,0000	15,0000	19,0000	34,0000	50,0000	60,0000	65,0000	86,0000	86,0000
INS	3,5714 %	12,2981	12,6968	2,0000	2,0000	2,0000	3,3700	4,6000	9,2000	15,6000	28,2000	28,5000	64,4000	64,4000
LEPT	14,2857 %	22,2538	21,3643	2,1900	2,1900	2,5300	4,3500	8,5300	14,2000	30,7000	38,9000	62,3000	93,6000	93,6000
APOA	3,5714 %	138,8815	23,6220	96,4000	96,4000	97,4000	102,0000	121,0000	136,0000	157,0000	169,0000	173,0000	190,0000	190,0000
APOB	3,5714 %	91,7667	27,4306	50,5000	50,5000	53,8000	60,9000	72,5000	87,1000	110,0000	127,0000	144,0000	155,0000	155,0000
LPA	3,5714 %	26,1656	29,4043	1,9600	1,9600	2,0000	2,0000	2,5900	10,5000	45,1000	76,5000	83,1000	110,0000	110,0000
BHID	3,5714 %	5,7996	7,8972	0,3200	0,3200	0,4100	0,4300	1,8800	3,3400	6,2500	11,9000	16,3100	40,5000	40,5000
NEFA	3,5714 %	1,0596	0,5131	0,3900	0,3900	0,4100	0,4900	0,7000	0,9700	1,4300	1,5900	1,6700	2,8000	2,8000
HEMG	10,7143 %	5,9520	0,9752	4,8000	4,8000	5,0000	5,0000	5,4000	5,7000	6,2000	7,4000	7,9000	8,9000	8,9000
HEMGm	10,7143 %	41,6400	10,6922	29,0000	29,0000	31,0000	31,0000	36,0000	39,0000	44,0000	57,0000	63,0000	74,0000	74,0000
SE	17,8571 %	93,7957	8,4942	73,8000	73,8000	78,8000	82,4000	87,5000	94,7000	99,9000	102,1000	104,2000	108,0000	108,0000
BGP	10,7143 %	9,0120	5,8222	2,0000	2,0000	2,5000	2,6000	5,7000	7,6000	10,4000	16,5000	19,7000	26,7000	26,7000
CTx	7,1429 %	436,5500	192,9219	153,0000	153,0000	173,2000	173,2000	339,9000	409,6500	521,8000	673,2000	851,5000	941,0000	941,0000
VitD	7,1429 %	69,4577	31,2738	22,6000	22,6000	26,6000	35,7000	50,8000	63,0000	84,6000	104,4000	134,5000	159,5000	159,5000
C0n	3,5714 %	28,5689	8,5322	14,1900	14,1900	14,3200	16,9900	22,6300	26,9300	33,1200	38,8000	42,2200	52,3400	52,3400
C2n	3,5714 %	16,1589	5,6062	7,6700	7,6700	8,3600	9,2000	10,8500	16,6700	20,5100	23,9000	25,8500	26,0000	26,0000
C3n	3,5714 %	1,4133	0,5818	0,7500	0,7500	0,7900	0,8700	0,9800	1,1900	1,6100	2,3600	2,4800	2,9100	2,9100
C3DCn	3,5714 %	0,2804	0,2024	0,1000	0,1000	0,1100	0,1200	0,1700	0,2100	0,3100	0,4300	0,8400	0,9700	0,9700
C4n	3,5714 %	0,1733	0,0646	0,0700	0,0700	0,1100	0,1200	0,1300	0,1500	0,2300	0,2800	0,2800	0,3200	0,3200
C4DCn	3,5714 %	1,2337	0,3799	0,6000	0,6000	0,6900	0,7500	0,9600	1,1600	1,6100	1,7200	1,8200	1,9100	1,9100
C5n	3,5714 %	0,1115	0,0552	0,0600	0,0600	0,0700	0,0700	0,0800	0,1000	0,1300	0,1500	0,1600	0,3500	0,3500
C51n	3,5714 %	0,0081	0,0062	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0200	0,0200	0,0200	0,0200
C5DCn	3,5714 %	0,0874	0,0336	0,0200	0,0200	0,0200	0,0300	0,0700	0,0900	0,1100	0,1200	0,1300	0,1700	0,1700
C6n	3,5714 %	0,0556	0,0314	0,0200	0,0200	0,0200	0,0300	0,0300	0,0500	0,0600	0,1000	0,1200	0,1600	0,1600

varName	PctMissing	Media	s	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
C6DCn	3,5714 %	0,0563	0,0322	0,0000	0,0000	0,0000	0,0100	0,0400	0,0500	0,0800	0,1000	0,1100	0,1300	0,1300
C8n	3,5714 %	0,0856	0,0371	0,0400	0,0400	0,0400	0,0400	0,0600	0,0800	0,1200	0,1300	0,1400	0,1800	0,1800
C81n	3,5714 %	0,0522	0,0278	0,0200	0,0200	0,0200	0,0200	0,0300	0,0500	0,0600	0,0900	0,0900	0,1500	0,1500
C10n	3,5714 %	0,0922	0,0473	0,0300	0,0300	0,0400	0,0400	0,0500	0,0900	0,1200	0,1300	0,1900	0,2400	0,2400
C101n	3,5714 %	0,0626	0,0297	0,0300	0,0300	0,0300	0,0300	0,0400	0,0600	0,0800	0,1000	0,1200	0,1500	0,1500
C102n	3,5714 %	0,0074	0,0059	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200
C12n	3,5714 %	0,0415	0,0207	0,0200	0,0200	0,0200	0,0200	0,0300	0,0300	0,0500	0,0700	0,1000	0,1000	0,1000
C121n	3,5714 %	0,0507	0,0277	0,0200	0,0200	0,0200	0,0300	0,0300	0,0400	0,0600	0,0900	0,0900	0,1500	0,1500
C14n	3,5714 %	0,0696	0,0301	0,0400	0,0400	0,0400	0,0500	0,0500	0,0600	0,0800	0,1200	0,1300	0,1700	0,1700
C14On	3,5714 %	0,0078	0,0042	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100
C141n	3,5714 %	0,0659	0,0360	0,0300	0,0300	0,0400	0,0400	0,0500	0,0500	0,0700	0,1000	0,1200	0,2100	0,2100
C142n	3,5714 %	0,0263	0,0150	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0500	0,0500	0,0800	0,0800
C16n	3,5714 %	0,9037	0,4849	0,4200	0,4200	0,4800	0,5200	0,6300	0,7600	0,9300	1,7700	1,9100	2,5400	2,5400
C161n	3,5714 %	0,0893	0,0536	0,0400	0,0400	0,0400	0,0400	0,0500	0,0700	0,1100	0,2000	0,2100	0,2200	0,2200
C16On	3,5714 %	0,0093	0,0047	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200
C161On	3,5714 %	0,0337	0,0115	0,0200	0,0200	0,0200	0,0200	0,0300	0,0300	0,0400	0,0500	0,0600	0,0700	0,0700
C18n	3,5714 %	0,4644	0,1561	0,2300	0,2300	0,2500	0,2700	0,3500	0,4700	0,5400	0,6500	0,8000	0,8700	0,8700
C18On	3,5714 %	0,0056	0,0051	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100
C181n	3,5714 %	1,4756	0,6340	0,7400	0,7400	0,7600	0,8600	1,0400	1,3600	1,6700	2,4900	2,9900	3,2000	3,2000
C181On	3,5714 %	0,0204	0,0076	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200	0,0300	0,0300	0,0400	0,0400
C182n	3,5714 %	0,3293	0,1316	0,1500	0,1500	0,1800	0,2000	0,2400	0,2900	0,3900	0,5500	0,5800	0,6500	0,6500

A.2. Grupo Apoe E3

Las distribuciones de las variables de los casos con alelo *E3* son:

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v140P	7,3171 %	0,5642	1,7704	0,0800	0,0800	0,1500	0,1900	0,2500	0,3700	0,5250	0,7300	0,9000	1,7700	22,0400
v160P	7,3171 %	19,2422	2,6665	4,3000	13,7700	15,4800	16,1900	17,7950	19,3700	20,7600	22,2600	23,5800	24,6000	27,7700
v161P	7,3171 %	1,9297	0,8832	0,6400	0,9500	1,0900	1,1800	1,3950	1,7200	2,2850	2,8000	3,2600	5,8200	8,3700
v180P	7,3171 %	7,9803	1,8972	0,4700	1,7000	6,4500	6,6600	7,2100	7,8050	8,5650	9,4700	9,7900	12,2700	24,2300
v181P	7,3171 %	45,8270	247,6884	2,2200	16,8000	19,0200	20,6300	23,6150	25,6200	28,3450	30,7200	32,3900	38,5300	3079,0000
v182P	7,3171 %	24,6863	4,7164	16,5600	16,6500	17,4200	18,6400	21,4150	23,8250	27,8150	31,4600	33,3100	39,1400	39,4500
g183P	7,3171 %	0,3369	0,1636	0,0700	0,0700	0,1300	0,1600	0,2200	0,3100	0,4200	0,5600	0,6500	0,9000	0,9000
a183P	8,5366 %	0,1927	0,0810	0,0600	0,0700	0,1000	0,1000	0,1300	0,1800	0,2400	0,2900	0,3600	0,4400	0,4500
v184P	8,5366 %	0,0579	0,0438	0,0100	0,0100	0,0100	0,0100	0,0200	0,0500	0,0900	0,1200	0,1300	0,2000	0,2100
v200P	7,3171 %	0,2841	0,1023	0,0200	0,1300	0,1400	0,1600	0,2100	0,2700	0,3400	0,4400	0,4700	0,5400	0,5600
v201P	7,3171 %	0,2523	0,0751	0,1000	0,1100	0,1500	0,1600	0,2000	0,2500	0,2900	0,3500	0,3900	0,4600	0,5800
v202P	7,3171 %	0,3556	0,1166	0,1400	0,1600	0,2000	0,2200	0,2700	0,3300	0,4400	0,4900	0,5800	0,6600	0,8400
v203P	7,3171 %	2,0588	0,6530	0,9400	0,9700	1,1300	1,3500	1,5850	1,9650	2,4200	2,8900	3,2100	4,1600	4,8100
v204P	7,3171 %	9,1066	2,3323	4,6300	4,9100	5,6800	6,3400	7,2350	8,8600	10,8300	12,1000	12,7300	14,9000	16,3900
v205P	7,3171 %	0,4703	0,3547	0,1100	0,1100	0,1700	0,1900	0,2450	0,3750	0,5500	0,8500	1,2700	1,9400	2,3000
v220P	7,3171 %	0,6863	0,2495	0,2000	0,2800	0,3300	0,3800	0,4800	0,6800	0,8450	1,0200	1,1400	1,2400	1,6400
v221P	7,3171 %	0,1541	0,0991	0,0200	0,0300	0,0500	0,0600	0,0800	0,1350	0,2000	0,2800	0,3400	0,4700	0,5900
v224P	7,3171 %	0,2975	0,1170	0,1200	0,1200	0,1500	0,1700	0,2100	0,2700	0,3700	0,4600	0,5000	0,6600	0,7500
v225P	8,5366 %	0,4424	0,1701	0,1200	0,1200	0,2100	0,2500	0,3200	0,4100	0,5700	0,6650	0,7400	0,9100	1,1100
v240P	7,3171 %	0,6016	0,2352	0,1800	0,2300	0,2700	0,3300	0,4100	0,5900	0,7300	0,9600	1,0200	1,3500	1,4100

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v226P	7,3171 %	2,9543	1,1407	0,6200	0,9000	1,2400	1,4700	2,1950	2,8150	3,7100	4,5400	4,9200	6,4900	6,8200
v241P	7,3171 %	1,2141	0,5449	0,3000	0,3300	0,4500	0,6200	0,7900	1,1250	1,6150	1,9200	2,3100	2,6900	2,9900
v260P	7,3171 %	0,0711	0,0605	0,0100	0,0100	0,0100	0,0100	0,0300	0,0500	0,1000	0,1500	0,2000	0,2900	0,3500
v120H	25,0000 %	0,0282	0,0414	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100	0,0300	0,0600	0,1100	0,2100	0,3000
v140H	25,0000 %	0,3124	1,1664	0,0300	0,0400	0,0700	0,0800	0,1200	0,1600	0,2500	0,3200	0,4400	1,7400	12,9400
v160H	25,0000 %	15,9580	3,7540	0,3400	1,4400	12,4800	13,0300	14,0500	15,9000	18,2600	20,0000	21,0200	22,6900	23,0200
v161H	25,0000 %	1,0436	2,9972	0,1300	0,1400	0,2000	0,2600	0,3300	0,4100	0,5100	0,7200	1,1200	16,4800	17,7700
v180H	25,0000 %	15,4931	1,2417	12,6800	13,0300	13,7000	14,0000	14,5900	15,3300	16,2400	17,2300	17,5300	18,7300	19,0000
v181H	25,0000 %	26,7821	140,6641	1,2500	7,2200	12,0900	12,3600	13,3900	14,3000	15,2000	15,8600	16,7000	17,9500	1574,0000
v182H	25,0000 %	8,7670	1,9690	0,1500	1,2200	6,1800	6,5700	7,6100	8,7200	9,9400	11,1400	11,8400	13,1300	13,2600
g183H	25,0000 %	0,1511	0,7676	0,0200	0,0200	0,0200	0,0300	0,0500	0,0700	0,1000	0,1500	0,2000	0,4900	8,5700
a183H	25,0000 %	0,0657	0,0593	0,0100	0,0100	0,0200	0,0200	0,0400	0,0500	0,0700	0,1000	0,1800	0,3400	0,4400
v184H	25,0000 %	0,0854	0,0613	0,0100	0,0100	0,0100	0,0100	0,0500	0,0700	0,1100	0,1500	0,2100	0,3100	0,3100
v200H	25,0000 %	0,4552	0,1677	0,2200	0,2300	0,2700	0,2900	0,3400	0,4200	0,5100	0,6900	0,7600	1,0600	1,2200
v201H	25,0000 %	0,3947	0,0933	0,1800	0,2200	0,2600	0,2700	0,3300	0,3900	0,4500	0,5100	0,5500	0,6200	0,7600
v202H	25,0000 %	0,4323	0,1088	0,2000	0,2400	0,2700	0,3000	0,3500	0,4300	0,5000	0,5600	0,6100	0,7000	0,8700
v203H	25,0000 %	2,0165	0,4709	0,9400	1,1800	1,3400	1,4500	1,7100	1,9500	2,3300	2,7200	2,8500	3,1500	3,2300
v204H	25,0000 %	17,6460	2,3960	12,3800	12,6000	14,1400	14,6700	15,6600	17,5600	19,9400	20,8200	21,3300	21,8900	22,2100
v205H	25,0000 %	0,5194	0,3167	0,1800	0,1900	0,2400	0,2700	0,3400	0,4500	0,6100	0,8000	1,0300	1,5400	2,7200
v220H	25,0000 %	1,3105	0,4527	0,5300	0,6700	0,7500	0,8700	0,9800	1,2400	1,5200	1,8300	2,1800	2,7000	3,5200
v221H	25,0000 %	0,1555	0,2610	0,0300	0,0400	0,0500	0,0500	0,0700	0,1000	0,1600	0,2500	0,3600	0,5400	2,8200
v224H	25,0000 %	4,4146	1,1292	1,5500	2,4200	2,8700	3,0300	3,5400	4,2400	5,2200	5,9500	6,3100	7,1600	7,5900
v225H	25,0000 %	2,0396	0,5055	0,6300	1,1400	1,2700	1,4200	1,7000	2,0400	2,3500	2,5800	2,8400	3,4200	3,6600
v240H	25,0000 %	3,8251	1,6714	0,0900	0,2500	1,8400	2,0600	2,6200	3,5500	5,0000	5,8400	6,5000	9,0300	9,5700
v226H	25,0000 %	6,2291	1,8840	2,6400	2,7700	3,5400	3,9200	4,6800	6,0200	7,5000	8,8200	9,6000	10,4900	11,9300
v241H	25,0000 %	4,1891	1,7101	1,6400	1,7500	1,9300	2,1800	2,8400	3,7800	5,5500	6,6300	6,9900	8,3300	9,1000
v260H	25,0000 %	0,2592	0,2214	0,0100	0,0100	0,0400	0,0600	0,1100	0,2000	0,3600	0,5200	0,6300	1,0100	1,5100
Fosfoserina	34,1463 %	24,3649	15,1715	9,8140	9,9490	10,5830	12,7400	15,1130	18,9110	26,3290	48,2940	59,6310	72,2820	91,9080

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
Taurina	34,1463 %	55,3337	28,5588	28,5560	28,7280	31,4410	33,6910	38,9870	48,3805	56,4270	96,8370	119,1040	174,8850	179,4910
AcAspartico	39,0244 %	3,6499	2,4229	1,0010	1,1540	1,5715	1,8035	2,4035	3,2015	3,9485	6,0570	7,1510	15,8755	21,1660
Treonina	34,1463 %	135,1467	121,7483	47,1840	55,6730	73,2490	79,4960	95,4180	106,3270	129,6190	165,9920	207,8120	794,0370	840,3290
Serina	34,1463 %	111,9325	71,7583	50,1040	56,2420	64,2710	69,8680	82,1705	96,2925	111,8355	145,6610	171,8100	467,1920	479,7150
Asparragina	34,7561 %	92,7413	59,9525	40,3550	42,3070	48,5970	58,1190	67,1800	80,9790	93,3530	116,1100	156,3510	400,0190	431,5420
AcGlutamico	35,3659 %	50,1353	30,8671	1,2810	7,8970	15,3660	17,7850	26,1050	42,0190	61,6110	98,1290	110,8170	144,8960	146,3260
Glutamina	34,1463 %	611,7889	431,9028	330,4150	350,1940	408,8960	440,4500	472,5845	516,1395	568,5240	650,8610	826,8670	2755,5660	2804,7070
Glicina	34,1463 %	203,3991	151,3397	99,5190	106,1960	125,4630	132,3810	145,1190	164,8010	193,2070	264,7150	381,3290	953,0590	1017,7350
Alanina	34,1463 %	336,4710	193,8370	154,9770	163,0970	195,9720	222,0130	253,4315	289,6155	344,6140	420,2680	588,2820	1201,4440	1601,4130
Citrulina	34,1463 %	33,2195	35,6562	6,0750	9,8420	13,6440	15,9500	19,4605	25,9365	33,1710	39,1460	50,0060	202,2690	273,2510
AcAlfaAminobutirico	34,1463 %	19,8236	10,3586	3,0840	7,4790	9,1290	10,1400	12,7590	17,7085	23,5105	32,3260	41,1890	61,0290	66,9700
Valina	34,1463 %	223,0392	112,0421	109,4460	119,2710	131,5950	153,7240	174,9300	198,6425	232,0955	276,6680	306,1170	749,2330	803,5620
Cistina	34,1463 %	55,5390	43,6369	14,0780	23,4830	28,0950	30,9260	36,3965	44,6855	57,8990	72,6130	87,7330	258,3640	263,0760
Metionina	35,3659 %	19,8516	14,5020	7,1430	8,2400	10,2560	11,6570	14,4490	16,5840	19,7490	23,9360	28,7490	86,9250	92,9260
Isoleucina	34,1463 %	65,6342	33,8192	28,5780	34,0380	39,0820	43,9200	48,9135	58,5370	71,1185	85,6280	91,7200	202,5550	255,9560
Leucina	34,1463 %	118,1210	47,3248	49,2510	55,7950	69,0380	78,5070	89,6065	111,4060	131,9385	156,8670	202,0570	281,6060	368,9220
Tirosina	34,1463 %	59,0868	32,5087	26,3110	27,6750	32,3640	37,9480	45,5435	50,5450	60,3140	78,8240	104,9380	189,7460	232,7970
FenilAlanina	34,1463 %	52,4776	29,7803	29,4150	30,4510	34,8510	35,9980	41,0870	46,2960	51,3500	59,8780	73,6170	195,0490	212,8100
Ornitina	34,1463 %	94,7226	62,1339	28,8710	38,3600	48,4790	55,8560	68,8960	80,9670	101,5550	117,7590	171,8400	350,3930	507,7000
Lisina	34,1463 %	197,2023	121,3204	84,1370	97,7880	117,1070	136,0470	152,4845	170,2725	194,0310	224,9810	303,0800	762,3110	862,2980
X1MetilHistidina	35,9756 %	12,9081	12,2988	1,1910	2,1110	3,3030	3,4890	5,6000	9,3760	17,1860	22,0770	33,9610	56,3650	93,1120
Histidina	34,1463 %	75,2831	54,3520	33,5240	40,5640	46,8300	50,2320	56,1280	64,1365	71,5380	87,2080	95,6090	321,5690	378,7800
Triptofano	34,7561 %	39,3720	25,2950	12,0680	13,6280	16,1100	18,1100	24,3050	33,1570	47,1030	62,0320	84,3900	124,3510	195,3560
Arginina	34,1463 %	41,3259	41,4374	3,3680	3,3910	10,2100	14,6620	21,0350	30,6445	49,6010	61,0140	73,1170	247,6200	288,0400
Prolina	34,1463 %	195,3498	142,4168	74,8610	76,6420	106,3530	115,9240	134,7095	167,3245	198,1950	265,8450	447,8780	853,9230	1176,0870
PCRU	3,6585 %	0,6594	0,7467	0,0200	0,0300	0,0600	0,0900	0,1800	0,4150	0,9500	1,5200	1,7600	3,2000	6,3100
GLU	0,0000 %	106,5366	35,1822	64,0000	65,0000	74,0000	76,0000	83,0000	98,0000	115,0000	152,0000	175,0000	216,0000	326,0000
TRIG	13,4146 %	135,7739	192,1739	41,7000	56,0000	61,0000	70,0000	83,2000	100,7500	136,0000	189,0000	227,0000	553,0000	2257,0000

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
CHOL	0,6098 %	194,7853	36,6703	110,0000	121,0000	128,0000	151,0000	167,0000	194,0000	221,0000	240,0000	250,0000	299,0000	313,0000
cHDL	21,3415 %	49,0465	11,5182	29,0000	31,0000	32,0000	34,0000	40,0000	46,0000	56,0000	64,0000	69,0000	79,0000	86,0000
LDL	25,6098 %	120,4098	33,5061	46,0000	53,0000	65,0000	76,0000	98,0000	120,5000	143,0000	163,0000	173,0000	205,0000	219,0000
GGT	1,2195 %	31,0679	28,4581	8,0000	8,0000	12,0000	13,0000	17,0000	23,0000	36,0000	52,0000	68,0000	208,0000	249,0000
GOT	4,2683 %	29,6051	13,3377	10,0000	11,0000	15,0000	16,0000	20,0000	26,0000	36,0000	51,0000	57,0000	70,0000	82,0000
GPT	1,2195 %	28,2222	16,5127	4,0000	9,0000	10,0000	12,0000	16,0000	23,5000	35,0000	49,0000	59,0000	81,0000	109,0000
INS	1,8293 %	13,7137	23,5695	1,0000	1,0000	1,5000	2,0000	3,5300	8,1500	15,0000	26,2000	38,0000	142,6000	231,0000
LEPT	7,3171 %	26,0879	27,8121	1,5600	1,5600	2,4200	3,3300	8,4200	18,5000	34,8000	53,4000	75,9000	103,5000	251,0000
APOA	2,4390 %	140,9925	28,8664	83,0000	86,2000	100,5000	107,0000	121,0000	138,5000	154,0000	180,0000	193,5000	237,0000	244,0000
APOB	2,4390 %	104,6138	24,9113	42,9000	46,3000	63,8500	75,0500	87,9000	102,5000	121,0000	133,0000	150,0000	170,0000	198,0000
LPA	2,4390 %	34,7119	41,4947	1,9600	2,0000	2,0000	2,0100	7,0050	22,6500	45,4000	81,2000	113,0000	261,0000	275,0000
BHID	2,4390 %	4,8436	4,3510	0,1500	0,2000	0,2550	1,1250	2,0100	3,8200	6,2150	9,2350	13,1400	24,5000	25,1000
NEFA	2,4390 %	0,9717	0,3384	0,1000	0,3100	0,4750	0,6100	0,7300	0,9600	1,1400	1,3500	1,5450	2,0200	2,7000
HEMG	4,2683 %	5,9924	1,2261	4,4000	4,7000	4,9000	5,1000	5,3000	5,6000	6,2000	7,1000	8,5000	11,2000	12,6000
HEMGm	4,2683 %	41,9682	13,4027	25,0000	28,0000	30,0000	32,0000	34,0000	38,0000	44,0000	54,0000	69,0000	99,0000	114,0000
SE	15,8537 %	92,2877	16,2478	35,0000	41,7000	63,0000	74,0000	83,2000	92,6500	102,0000	112,0000	122,3000	127,0000	128,4000
BGP	6,7073 %	9,7229	10,8481	0,5000	0,9000	1,6000	2,6000	4,3000	7,2000	12,0000	17,3000	20,9000	46,2000	114,0000
CTx	7,9268 %	349,9146	205,3615	102,7000	105,0000	133,2000	161,5000	220,9000	308,6000	435,9000	555,8000	652,0000	1245,0000	1695,0000
VitD	3,6585 %	55,0481	22,7545	16,8000	16,8000	25,3000	32,0000	40,1000	52,1000	64,8000	82,9000	98,6000	125,4000	185,3000
C0n	1,8293 %	29,4335	9,5564	9,6200	10,0400	15,8700	18,6800	22,8300	28,0700	34,9500	41,1400	47,2000	59,8100	63,5100
C2n	1,8293 %	16,8004	6,2254	4,7500	5,1700	8,3400	10,1400	12,8000	15,4100	19,7600	24,7300	28,4600	39,1900	41,9500
C3n	1,8293 %	1,5341	0,6962	0,4200	0,4600	0,6400	0,7600	1,0300	1,4000	1,9500	2,4900	2,6700	3,6200	4,1100
C3DCn	1,8293 %	0,2963	0,2046	0,0700	0,0700	0,1100	0,1200	0,1700	0,2300	0,3400	0,5500	0,6800	1,2300	1,3300
C4n	1,8293 %	0,1981	0,1080	0,0400	0,0600	0,0900	0,1000	0,1300	0,1700	0,2500	0,3200	0,4100	0,7200	0,7500
C4DCn	1,8293 %	1,1516	0,4840	0,3900	0,4000	0,6000	0,6800	0,8300	1,0500	1,3800	1,7500	2,1100	2,8400	3,2100
C5n	1,8293 %	0,1158	0,0524	0,0200	0,0300	0,0500	0,0600	0,0800	0,1100	0,1400	0,1800	0,2100	0,3200	0,3600
C51n	1,8293 %	0,0096	0,0068	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0400
C5DCn	1,8293 %	0,0887	0,0390	0,0000	0,0000	0,0400	0,0400	0,0600	0,0900	0,1100	0,1400	0,1600	0,1900	0,2300
C6n	1,8293 %	0,0492	0,0308	0,0000	0,0000	0,0100	0,0200	0,0300	0,0500	0,0600	0,0800	0,1000	0,1600	0,2000

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
C6DCn	1,8293 %	0,0663	0,0488	0,0000	0,0000	0,0200	0,0300	0,0400	0,0600	0,0800	0,1100	0,1400	0,2700	0,3900
C8n	1,8293 %	0,0894	0,0421	0,0200	0,0300	0,0400	0,0500	0,0600	0,0800	0,1100	0,1400	0,1700	0,2300	0,2500
C81n	1,8293 %	0,0461	0,0267	0,0000	0,0100	0,0200	0,0200	0,0300	0,0400	0,0600	0,0700	0,0900	0,1400	0,2500
C10n	1,8293 %	0,0986	0,0497	0,0300	0,0300	0,0400	0,0500	0,0700	0,0900	0,1200	0,1500	0,1900	0,2900	0,3100
C101n	1,8293 %	0,0648	0,0289	0,0200	0,0200	0,0300	0,0300	0,0500	0,0600	0,0800	0,1000	0,1100	0,1400	0,2500
C102n	1,8293 %	0,0066	0,0059	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0200	0,0300
C12n	1,8293 %	0,0376	0,0189	0,0100	0,0100	0,0200	0,0200	0,0200	0,0300	0,0400	0,0600	0,0800	0,1000	0,1300
C121n	1,8293 %	0,0481	0,0225	0,0100	0,0100	0,0200	0,0200	0,0300	0,0400	0,0600	0,0800	0,0800	0,1200	0,1600
C14n	1,8293 %	0,0592	0,0230	0,0200	0,0200	0,0300	0,0300	0,0400	0,0500	0,0700	0,0900	0,1100	0,1200	0,1500
C14On	1,8293 %	0,0087	0,0046	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200
C141n	1,8293 %	0,0663	0,0337	0,0200	0,0200	0,0300	0,0300	0,0400	0,0600	0,0800	0,1000	0,1200	0,1900	0,2100
C142n	1,8293 %	0,0255	0,0136	0,0000	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0400	0,0500	0,0800	0,0900
C16n	1,8293 %	0,7831	0,2944	0,3200	0,3300	0,4100	0,4600	0,5700	0,7000	0,9300	1,1900	1,3700	1,6800	1,7600
C161n	1,8293 %	0,0745	0,0334	0,0200	0,0300	0,0300	0,0400	0,0500	0,0700	0,0900	0,1200	0,1300	0,2100	0,2400
C16On	1,8293 %	0,0106	0,0045	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200	0,0200
C161On	1,8293 %	0,0299	0,0104	0,0100	0,0100	0,0200	0,0200	0,0200	0,0300	0,0400	0,0400	0,0500	0,0700	0,0700
C18n	1,8293 %	0,4432	0,1688	0,2000	0,2000	0,2400	0,2700	0,3200	0,4100	0,5300	0,6800	0,7900	0,9600	1,0400
C18On	1,8293 %	0,0050	0,0050	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100
C181n	1,8293 %	1,4380	0,5435	0,5400	0,6200	0,7400	0,8200	1,1100	1,3300	1,7200	2,0200	2,5000	3,4900	3,7500
C181On	1,8293 %	0,0193	0,0083	0,0100	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0400	0,0400	0,0500
C182n	1,8293 %	0,3516	0,1309	0,1400	0,1400	0,1700	0,2100	0,2700	0,3300	0,4000	0,5300	0,6300	0,8200	0,8500

A.3. Grupo Apoe E2

Las distribuciones de las variables de los casos con alelo *E4* son:

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v140P	5,7143 %	0,4503	0,2364	0,1700	0,1700	0,2000	0,2400	0,2700	0,3700	0,5500	0,7000	1,0200	1,1300	1,1300
v160P	5,7143 %	20,0391	1,8925	14,7300	14,7300	17,6400	17,9000	18,9700	20,1800	21,0900	21,9500	23,1300	25,2500	25,2500
v161P	5,7143 %	2,0024	0,7169	1,1200	1,1200	1,2300	1,2500	1,6500	1,8800	2,2200	2,6400	2,7700	5,1000	5,1000
v180P	5,7143 %	7,7676	0,9202	6,4000	6,4000	6,4800	6,5900	7,1300	7,7500	8,2300	8,8100	10,0800	10,1500	10,1500
v181P	5,7143 %	25,4191	3,5302	18,0500	18,0500	19,6100	21,8500	22,7200	25,0500	27,5800	30,1500	31,3700	34,8300	34,8300
v182P	5,7143 %	31,5945	35,4067	9,2700	9,2700	18,7500	19,7900	23,7900	26,2800	28,6000	29,6700	34,7700	227,1200	227,1200
g183P	5,7143 %	0,2930	0,1144	0,0900	0,0900	0,1100	0,1600	0,2200	0,2900	0,3700	0,4400	0,4900	0,5900	0,5900
a183P	5,7143 %	0,1624	0,0728	0,0600	0,0600	0,0800	0,1000	0,1100	0,1500	0,1900	0,2400	0,3000	0,4100	0,4100
v184P	5,7143 %	0,0476	0,0367	0,0100	0,0100	0,0100	0,0100	0,0100	0,0400	0,0700	0,1000	0,1100	0,1400	0,1400
v200P	5,7143 %	0,2648	0,0890	0,1100	0,1100	0,1300	0,1800	0,2000	0,2500	0,3100	0,3800	0,4400	0,4800	0,4800
v201P	5,7143 %	0,2242	0,0573	0,1300	0,1300	0,1500	0,1600	0,1800	0,2200	0,2500	0,3000	0,3500	0,3600	0,3600
v202P	5,7143 %	0,3170	0,1020	0,1700	0,1700	0,1700	0,2000	0,2600	0,3100	0,3400	0,4500	0,5600	0,6100	0,6100
v203P	5,7143 %	1,9233	0,7133	0,9300	0,9300	0,9700	1,0600	1,4400	1,7700	2,1700	3,1600	3,4700	3,6800	3,6800
v204P	5,7143 %	8,7182	2,3732	5,0000	5,0000	5,1900	6,1600	7,0200	8,1000	10,3300	12,0300	13,0200	14,8600	14,8600
v205P	5,7143 %	0,4524	0,3366	0,1100	0,1100	0,1300	0,1300	0,1800	0,3400	0,5900	1,0000	1,1700	1,2100	1,2100
v220P	5,7143 %	0,6897	0,2022	0,2300	0,2300	0,4200	0,4400	0,5400	0,6900	0,8100	0,9300	1,0100	1,2000	1,2000
v221P	5,7143 %	0,1345	0,0585	0,0500	0,0500	0,0500	0,0600	0,0800	0,1300	0,1800	0,2100	0,2300	0,2900	0,2900
v224P	5,7143 %	0,2621	0,1304	0,1000	0,1000	0,1300	0,1400	0,1700	0,2300	0,2800	0,3900	0,5900	0,6300	0,6300
v225P	5,7143 %	0,4009	0,1446	0,1800	0,1800	0,1800	0,2200	0,2900	0,3600	0,5100	0,5800	0,6600	0,7300	0,7300
v240P	5,7143 %	0,5948	0,2199	0,2100	0,2100	0,2800	0,3400	0,4500	0,5800	0,7300	0,8000	0,9900	1,2100	1,2100

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
v226P	5,7143 %	2,5909	0,8516	1,0100	1,0100	1,1300	1,4700	2,1400	2,4300	3,2400	3,7400	4,1500	4,3000	4,3000
v241P	5,7143 %	1,2036	0,4061	0,6100	0,6100	0,6800	0,7100	0,8500	1,1600	1,4900	1,8100	1,9700	2,0500	2,0500
v260P	5,7143 %	0,0664	0,0455	0,0100	0,0100	0,0100	0,0100	0,0300	0,0600	0,0900	0,1300	0,1400	0,2000	0,2000
v120H	31,4286 %	0,0500	0,1207	0,0100	0,0100	0,0100	0,0100	0,0100	0,0100	0,0350	0,1000	0,1100	0,6000	0,6000
v140H	31,4286 %	0,2425	0,3117	0,0600	0,0600	0,0600	0,0900	0,1350	0,1700	0,2250	0,3500	0,3700	1,6600	1,6600
v160H	31,4286 %	17,1329	2,6389	12,0500	12,0500	13,0100	13,9100	15,4950	16,6250	18,8100	21,1600	22,3300	22,3600	22,3600
v161H	31,4286 %	1,0196	2,8209	0,2600	0,2600	0,2700	0,2800	0,3700	0,4300	0,5150	0,6100	0,8300	14,2500	14,2500
v180H	31,4286 %	15,2238	1,1990	12,3400	12,3400	13,6800	13,7500	14,2350	15,3850	16,1900	16,5300	16,6800	17,2700	17,2700
v181H	31,4286 %	13,8233	1,9074	8,2900	8,2900	11,2300	11,6500	12,9350	13,8600	15,3750	15,9300	16,2100	16,9200	16,9200
v182H	31,4286 %	9,2475	2,8038	0,0700	0,0700	6,0300	6,7100	8,0200	9,3650	11,4750	12,2100	12,4000	13,8600	13,8600
g183H	31,4286 %	0,0492	0,0280	0,0200	0,0200	0,0200	0,0200	0,0300	0,0400	0,0550	0,1000	0,1100	0,1100	0,1100
a183H	31,4286 %	0,0571	0,0418	0,0100	0,0100	0,0100	0,0200	0,0250	0,0450	0,0750	0,1000	0,1500	0,1800	0,1800
v184H	31,4286 %	0,0813	0,0580	0,0100	0,0100	0,0100	0,0100	0,0500	0,0600	0,1050	0,1600	0,1800	0,2600	0,2600
v200H	31,4286 %	0,4196	0,1485	0,2200	0,2200	0,2700	0,2700	0,3250	0,3750	0,4850	0,6500	0,6800	0,7600	0,7600
v201H	31,4286 %	0,3442	0,1124	0,1800	0,1800	0,2000	0,2100	0,2550	0,3450	0,4000	0,5400	0,5600	0,5800	0,5800
v202H	31,4286 %	0,4183	0,0768	0,2700	0,2700	0,3100	0,3100	0,3550	0,4300	0,4900	0,5200	0,5200	0,5300	0,5300
v203H	31,4286 %	1,9158	0,5086	1,3200	1,3200	1,3300	1,3800	1,5050	1,8750	2,2000	2,3400	2,3500	3,6400	3,6400
v204H	31,4286 %	17,1792	2,1212	14,2300	14,2300	14,9500	14,9800	15,4200	16,8400	18,4950	20,0600	20,3200	22,5400	22,5400
v205H	31,4286 %	0,4604	0,2649	0,1800	0,1800	0,2200	0,2300	0,2650	0,3650	0,5650	0,8800	1,0300	1,0600	1,0600
v220H	31,4286 %	1,3904	0,4899	0,7400	0,7400	0,7500	0,9200	1,0950	1,2300	1,6500	2,3400	2,4100	2,4500	2,4500
v221H	31,4286 %	0,1213	0,1038	0,0300	0,0300	0,0400	0,0400	0,0500	0,0850	0,1350	0,3000	0,3800	0,4000	0,4000
v224H	31,4286 %	4,3200	1,3983	2,5400	2,5400	2,6500	2,6700	3,2100	4,0850	5,3850	5,8800	7,3100	7,4200	7,4200
v225H	31,4286 %	1,8250	0,4410	1,1700	1,1700	1,3300	1,3800	1,4550	1,7800	2,1350	2,4400	2,5700	2,8700	2,8700
v240H	31,4286 %	4,3775	1,6633	1,9800	1,9800	2,0100	2,5100	3,1700	4,1200	5,2800	6,4900	6,9900	8,8300	8,8300
v226H	31,4286 %	5,3546	1,6202	3,1600	3,1600	3,3800	3,5500	4,1750	5,1850	6,2250	7,2200	7,8700	9,9100	9,9100
v241H	31,4286 %	4,3817	1,5510	2,1200	2,1200	2,2900	2,5900	3,3750	4,0850	5,3250	6,5600	7,3900	8,1500	8,1500
v260H	31,4286 %	0,4621	0,6453	0,0200	0,0200	0,0600	0,0600	0,2100	0,3650	0,5150	0,5700	0,6600	3,3700	3,3700
Fosfoserina	31,4286 %	27,7058	16,8470	8,4640	8,4640	12,0330	13,5920	17,9245	23,6240	29,1395	54,3140	72,0750	75,5760	75,5760

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
Taurina	31,4286 %	52,6070	14,7405	34,6520	34,6520	36,5200	36,5380	44,0830	48,6645	58,2355	79,9450	84,0660	84,7160	84,7160
AcAspartico	37,1429 %	3,2704	1,2529	0,5860	0,5860	1,5010	1,7500	2,2940	3,2275	4,1450	4,8910	5,0610	5,3390	5,3390
Treonina	31,4286 %	113,1271	29,4556	54,7970	54,7970	73,2220	81,9060	96,3070	106,5370	137,5160	158,6120	167,2060	169,2120	169,2120
Serina	31,4286 %	99,7768	21,5411	63,8730	63,8730	69,5810	71,4780	85,1205	99,2855	106,4620	126,5730	140,8120	151,6360	151,6360
Asparragina	31,4286 %	82,6941	20,1113	43,9260	43,9260	54,4000	55,3990	68,2780	85,9925	92,6935	117,8230	118,0660	120,1470	120,1470
AcGlutamico	34,2857 %	42,9397	23,3574	10,6500	10,6500	13,8800	23,1730	28,2000	36,2600	51,2150	66,7920	75,6180	121,8240	121,8240
Glutamina	31,4286 %	531,6473	62,7632	415,8800	415,8800	429,0620	462,8950	494,5275	526,9390	563,8590	579,8890	636,5880	704,0590	704,0590
Glicina	31,4286 %	155,9747	28,4064	106,3870	106,3870	116,4420	117,0850	135,6160	155,0505	172,0740	195,1030	206,5950	209,9430	209,9430
Alanina	31,4286 %	283,9312	75,9630	212,5190	212,5190	215,7480	220,8330	232,0865	270,7345	291,3630	427,5810	443,7120	516,6830	516,6830
Citrulina	31,4286 %	24,1843	8,1218	7,6510	7,6510	15,1690	16,5930	18,0755	22,9370	29,3700	35,2460	37,0650	41,7370	41,7370
AcAlfaAminobutirico	31,4286 %	18,0865	9,5217	7,7880	7,7880	8,5010	9,4880	11,1245	14,9455	24,7630	31,6280	37,0910	42,0160	42,0160
Valina	31,4286 %	215,9759	36,8955	128,5210	128,5210	163,1660	173,6840	197,2695	216,8940	232,5390	255,3130	287,2340	290,9370	290,9370
Cistina	31,4286 %	41,2263	11,4087	1,5290	1,5290	30,8040	34,1830	35,9635	41,2235	46,6650	56,2360	56,6810	57,2490	57,2490
Metionina	34,2857 %	16,4098	2,6724	11,8670	11,8670	12,8640	13,5010	14,8950	16,2270	17,4300	19,8360	19,8620	24,5960	24,5960
Isoleucina	31,4286 %	60,7988	13,5416	39,2000	39,2000	41,0190	46,2290	53,8930	59,5910	66,2095	72,1080	83,0550	103,0770	103,0770
Leucina	31,4286 %	118,4611	21,5005	73,8160	73,8160	86,4100	89,7020	104,9305	119,1230	130,6275	155,4000	155,4680	161,0920	161,0920
Tirosina	31,4286 %	52,0187	7,1558	39,3340	39,3340	39,8220	41,8260	46,9820	52,7915	57,2250	62,7890	63,0060	63,0470	63,0470
FenilAlanina	31,4286 %	46,1541	6,3320	32,2570	32,2570	35,6030	38,7860	41,8945	44,9890	50,6240	54,5940	55,7950	57,1430	57,1430
Ornitina	31,4286 %	83,6983	24,5982	48,1160	48,1160	48,8470	55,2290	72,2635	79,1480	92,9640	116,9290	124,5970	155,9690	155,9690
Lisina	31,4286 %	164,1039	24,5375	111,8750	111,8750	133,4360	142,4770	148,7540	160,0190	181,4120	185,4540	211,7500	219,3140	219,3140
X1MetilHistidina	34,2857 %	10,6406	7,3412	2,2860	2,2860	2,3570	2,6030	4,0510	8,7210	15,6740	19,9620	20,5000	31,3970	31,3970
Histidina	31,4286 %	62,7966	9,6711	46,7890	46,7890	49,6120	54,2860	56,9660	59,6960	69,2840	78,5880	79,4120	84,7210	84,7210
Triptofano	31,4286 %	33,0196	14,8906	16,4740	16,4740	17,6320	22,4670	24,3545	29,7295	33,8750	51,5090	58,0970	85,0260	85,0260
Arginina	31,4286 %	40,8370	17,8077	4,8230	4,8230	10,3960	11,4900	30,4365	45,0645	55,8895	59,4160	66,4030	68,6250	68,6250
Prolina	31,4286 %	156,2250	51,2847	58,4020	58,4020	103,3680	107,6360	127,9125	146,8600	185,8975	202,7240	209,9980	325,3050	325,3050
PCRU	0,0000 %	0,5103	0,5386	0,0300	0,0300	0,0400	0,0600	0,1200	0,2200	0,8600	1,3500	1,5700	2,0400	2,0400
GLU	8,5714 %	92,7500	19,7288	66,0000	66,0000	71,0000	72,0000	79,0000	87,5000	102,0000	112,0000	137,0000	162,0000	162,0000
TRIG	14,2857 %	123,9600	70,0421	45,0000	45,0000	55,4000	59,5000	70,0000	102,5000	156,0000	228,0000	274,0000	345,0000	345,0000

varName	PctMissing	Media	*s*	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
CHOL	8,5714 %	213,2188	45,6747	123,0000	123,0000	137,0000	159,0000	175,0000	210,0000	251,5000	277,0000	297,0000	302,0000	302,0000
cHDL	22,8571 %	49,0370	11,1130	33,0000	33,0000	34,0000	36,0000	39,0000	47,0000	58,0000	61,0000	68,0000	75,0000	75,0000
LDL	25,7143 %	135,2692	39,7548	57,0000	57,0000	67,0000	80,0000	119,0000	133,5000	167,0000	188,0000	191,0000	214,0000	214,0000
GGT	0,0000 %	38,4571	39,9811	9,0000	9,0000	10,0000	13,0000	16,0000	23,0000	44,0000	60,0000	145,0000	181,0000	181,0000
GOT	5,7143 %	33,7879	33,0953	13,0000	13,0000	14,0000	16,0000	21,0000	28,0000	34,0000	46,0000	55,0000	210,0000	210,0000
GPT	0,0000 %	38,5429	61,5216	8,0000	8,0000	11,0000	12,0000	16,0000	23,0000	38,0000	69,0000	80,0000	377,0000	377,0000
INS	0,0000 %	11,7054	14,5144	1,0000	1,0000	2,0000	2,0000	3,5600	6,6100	12,5000	31,0000	48,5000	69,5000	69,5000
LEPT	8,5714 %	26,1438	30,2311	1,5600	1,5600	1,5600	1,8900	2,9500	19,2000	28,0500	75,4000	100,7000	108,6000	108,6000
APOA	0,0000 %	139,0829	26,7232	90,5000	90,5000	93,4000	107,0000	118,0000	138,0000	156,0000	178,0000	188,0000	194,0000	194,0000
APOB	0,0000 %	116,3886	29,7865	81,1000	81,1000	82,8000	84,4000	91,6000	113,0000	130,0000	167,0000	189,0000	200,0000	200,0000
LPA	0,0000 %	31,2046	50,9063	2,0000	2,0000	2,0000	2,6000	6,5300	10,3000	34,3000	109,0000	134,0000	254,0000	254,0000
BHID	0,0000 %	4,5434	3,2044	0,2500	0,2500	0,3500	1,1700	2,0200	4,1100	6,3000	8,8200	11,3000	14,6000	14,6000
NEFA	0,0000 %	1,0591	0,4979	0,5100	0,5100	0,5400	0,6400	0,7600	0,8600	1,2800	1,5500	2,1100	3,1000	3,1000
HEMG	0,0000 %	5,6371	0,6924	5,0000	5,0000	5,1000	5,1000	5,2000	5,4000	5,8000	6,1000	7,8000	8,3000	8,3000
HEMGm	0,0000 %	38,0286	7,6369	31,0000	31,0000	32,0000	32,0000	33,0000	36,0000	40,0000	43,0000	62,0000	67,0000	67,0000
SE	20,0000 %	88,9750	15,7484	55,4000	55,4000	55,4000	73,9000	79,8500	89,6000	95,8500	111,0000	112,0000	130,0000	130,0000
BGP	5,7143 %	9,4455	9,0618	1,2000	1,2000	1,4000	2,1000	3,7000	7,2000	13,8000	14,9000	16,4000	51,5000	51,5000
CTx	8,5714 %	387,8531	222,8235	106,5000	106,5000	136,6000	171,9000	246,6000	319,5000	514,6000	614,5000	898,5000	1103,0000	1103,0000
VitD	2,8571 %	55,4235	16,1337	23,4000	23,4000	24,8000	36,8000	45,4000	53,7000	65,4000	74,9000	84,5000	92,2000	92,2000
C0n	0,0000 %	29,2309	9,4035	17,2000	17,2000	18,0000	21,3400	23,2200	27,4900	32,9500	37,4800	39,7500	70,7500	70,7500
C2n	0,0000 %	16,9286	6,3636	6,5100	6,5100	6,5500	11,0900	12,7300	15,6600	19,2500	25,5400	33,5500	35,9900	35,9900
C3n	0,0000 %	1,5137	0,6688	0,5100	0,5100	0,7100	0,7800	0,9100	1,4600	1,9100	2,4200	2,9600	3,3000	3,3000
C3DCn	0,0000 %	0,3149	0,2269	0,0800	0,0800	0,0800	0,1000	0,1800	0,2600	0,4000	0,5700	0,7900	1,2100	1,2100
C4n	0,0000 %	0,2126	0,1015	0,0000	0,0000	0,0700	0,0900	0,1400	0,2000	0,2700	0,3500	0,4300	0,4400	0,4400
C4DCn	0,0000 %	1,3020	0,5576	0,4500	0,4500	0,5700	0,7900	0,9900	1,2000	1,5200	2,2100	2,5300	2,8900	2,8900
C5n	0,0000 %	0,1117	0,0436	0,0400	0,0400	0,0500	0,0600	0,0800	0,1000	0,1500	0,1800	0,1900	0,2000	0,2000
C51n	0,0000 %	0,0097	0,0057	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200	0,0200
C5DCn	0,0000 %	0,0871	0,0514	0,0000	0,0000	0,0000	0,0300	0,0600	0,0800	0,1100	0,1400	0,1600	0,2800	0,2800
C6n	0,0000 %	0,0657	0,0356	0,0200	0,0200	0,0200	0,0300	0,0400	0,0600	0,0800	0,1200	0,1300	0,1900	0,1900

varName	PctMissing	Media	s	Mínimo	p1	p5	p10	p25	Mediana	p75	p90	p95	p99	Máximo
C6DCn	0,0000 %	0,0629	0,0250	0,0200	0,0200	0,0200	0,0300	0,0500	0,0600	0,0800	0,0900	0,1100	0,1200	0,1200
C8n	0,0000 %	0,0914	0,0342	0,0400	0,0400	0,0500	0,0600	0,0700	0,0800	0,1100	0,1300	0,1600	0,1900	0,1900
C81n	0,0000 %	0,0423	0,0152	0,0200	0,0200	0,0200	0,0300	0,0300	0,0400	0,0500	0,0700	0,0700	0,0800	0,0800
C10n	0,0000 %	0,0994	0,0389	0,0400	0,0400	0,0500	0,0600	0,0700	0,0900	0,1200	0,1700	0,1700	0,1800	0,1800
C101n	0,0000 %	0,0680	0,0289	0,0300	0,0300	0,0400	0,0400	0,0400	0,0600	0,0800	0,1100	0,1200	0,1600	0,1600
C102n	0,0000 %	0,0063	0,0065	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200
C12n	0,0000 %	0,0383	0,0129	0,0200	0,0200	0,0200	0,0200	0,0300	0,0400	0,0500	0,0600	0,0600	0,0600	0,0600
C121n	0,0000 %	0,0460	0,0156	0,0200	0,0200	0,0200	0,0300	0,0300	0,0500	0,0600	0,0700	0,0700	0,0700	0,0700
C14n	0,0000 %	0,0609	0,0217	0,0200	0,0200	0,0300	0,0400	0,0400	0,0600	0,0800	0,0900	0,1000	0,1000	0,1000
C14On	0,0000 %	0,0083	0,0051	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0200
C141n	0,0000 %	0,0671	0,0260	0,0300	0,0300	0,0300	0,0400	0,0500	0,0600	0,0800	0,1000	0,1300	0,1400	0,1400
C142n	0,0000 %	0,0274	0,0115	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0300	0,0400	0,0500	0,0600	0,0600
C16n	0,0000 %	0,8231	0,2976	0,2700	0,2700	0,3800	0,4400	0,5900	0,8300	1,0200	1,1700	1,4400	1,5300	1,5300
C161n	0,0000 %	0,0740	0,0329	0,0200	0,0200	0,0300	0,0400	0,0500	0,0700	0,0900	0,1100	0,1400	0,1800	0,1800
C16On	0,0000 %	0,0083	0,0045	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200
C161On	0,0000 %	0,0323	0,0117	0,0100	0,0100	0,0200	0,0200	0,0200	0,0300	0,0400	0,0500	0,0500	0,0600	0,0600
C18n	0,0000 %	0,5031	0,2232	0,1600	0,1600	0,2100	0,2500	0,3700	0,4400	0,6200	0,8300	0,9000	1,2000	1,2000
C18On	0,0000 %	0,0037	0,0049	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100
C181n	0,0000 %	1,5126	0,4885	0,4200	0,4200	0,6000	0,9700	1,2000	1,5100	1,8200	2,2900	2,3100	2,3900	2,3900
C181On	0,0000 %	0,0194	0,0080	0,0100	0,0100	0,0100	0,0100	0,0100	0,0200	0,0200	0,0300	0,0400	0,0400	0,0400
C182n	0,0000 %	0,3783	0,1464	0,1400	0,1400	0,1700	0,2000	0,2500	0,3600	0,4700	0,5900	0,6200	0,8000	0,8000

Anexo B

Diagramas de caja y funciones de densidad

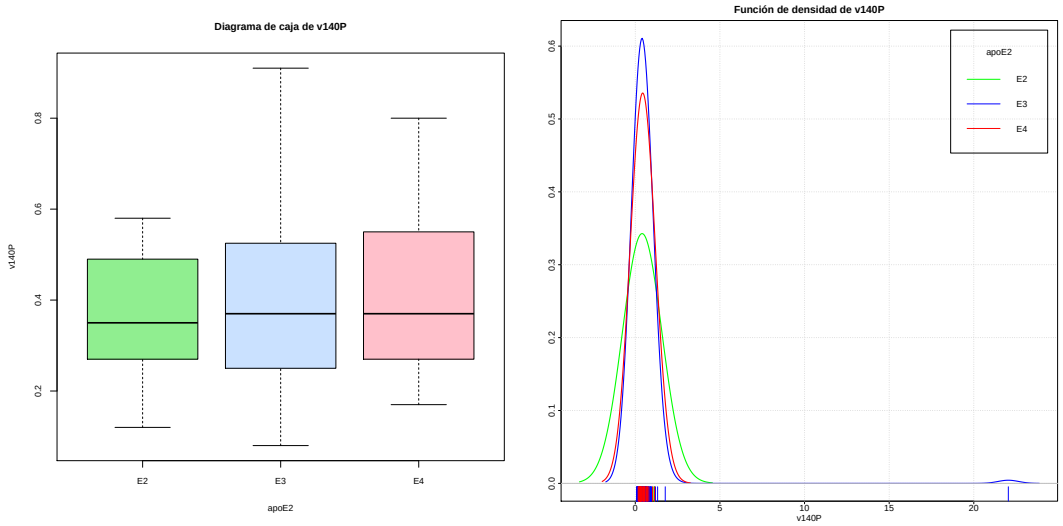

Figura B.1: Diagrama de caja y función de densidad de la variable $v140P$.

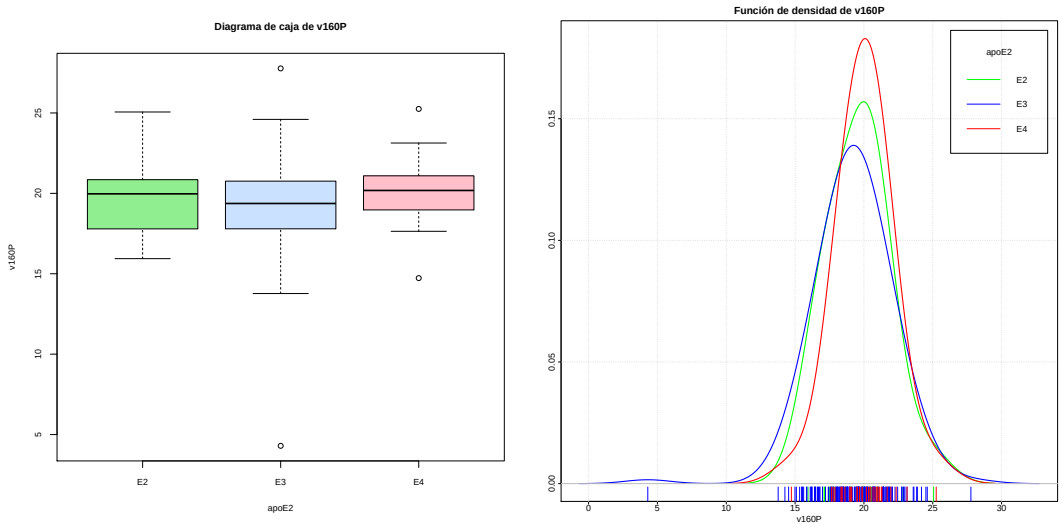

Figura B.2: Diagrama de caja y función de densidad de la variable $v160P$.

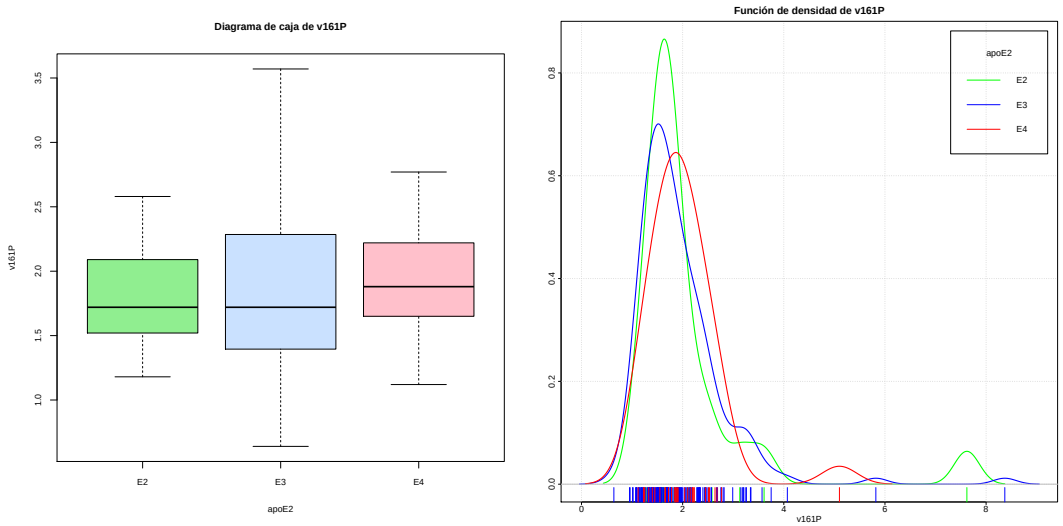

Figura B.3: Diagrama de caja y función de densidad de la variable $v161P$.

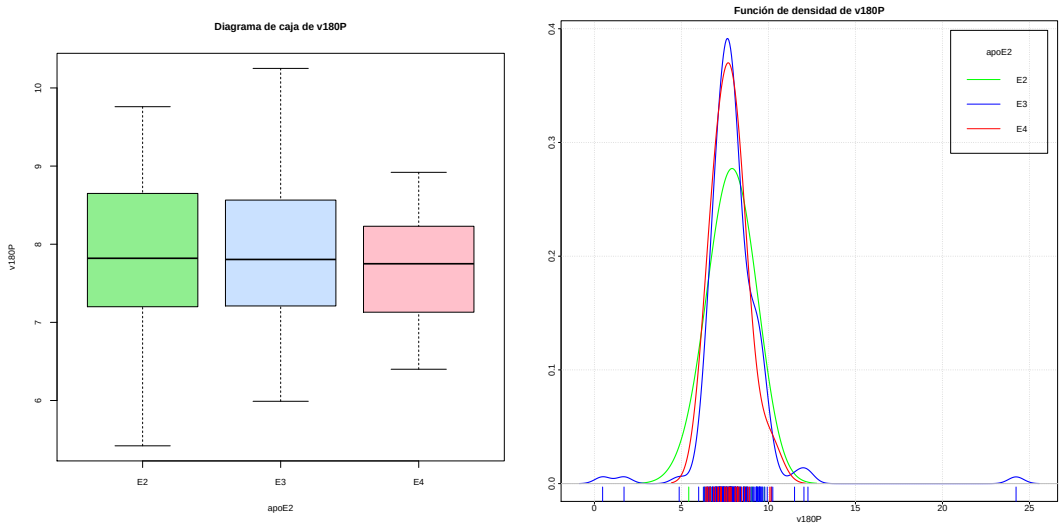

Figura B.4: Diagrama de caja y función de densidad de la variable $v180P$.

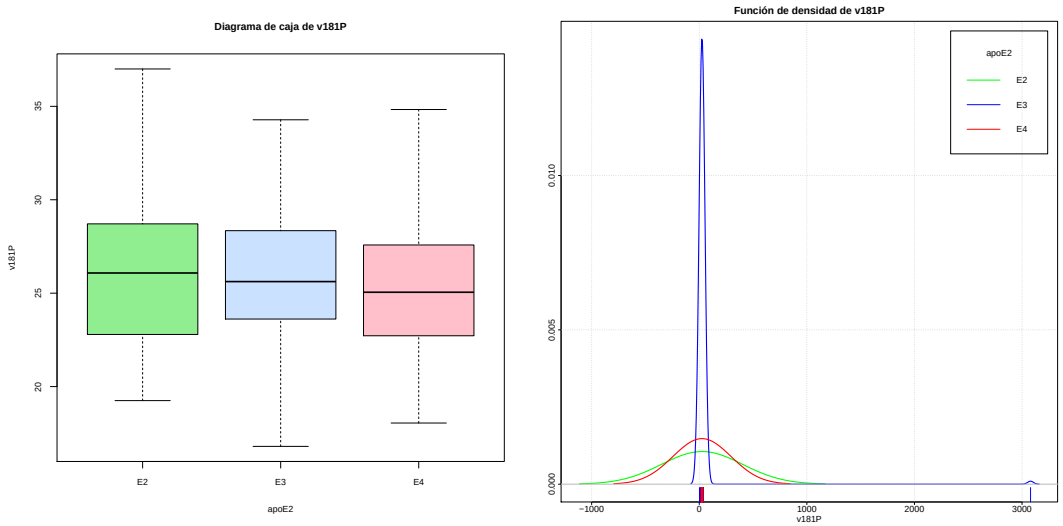

Figura B.5: Diagrama de caja y función de densidad de la variable $v181P$.

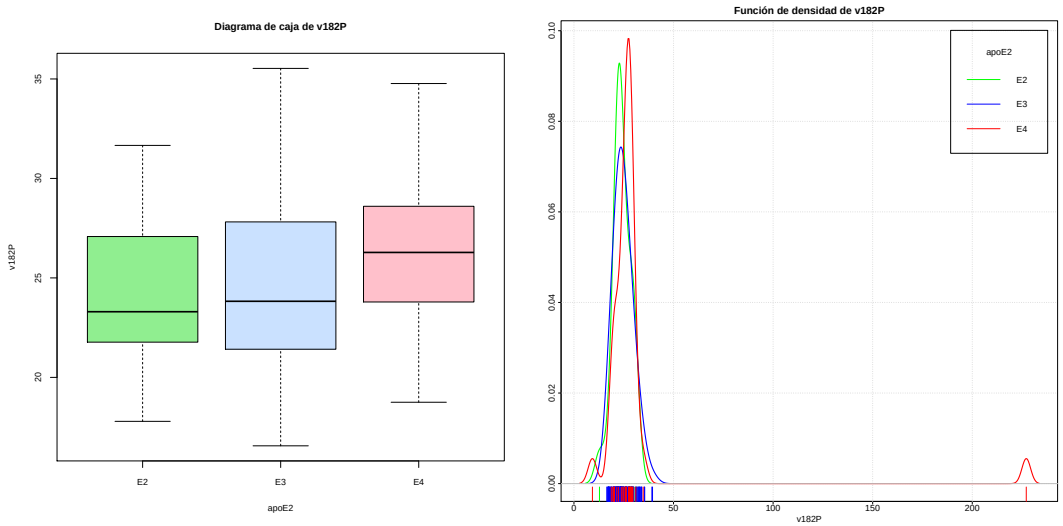

Figura B.6: Diagrama de caja y función de densidad de la variable $v182P$.

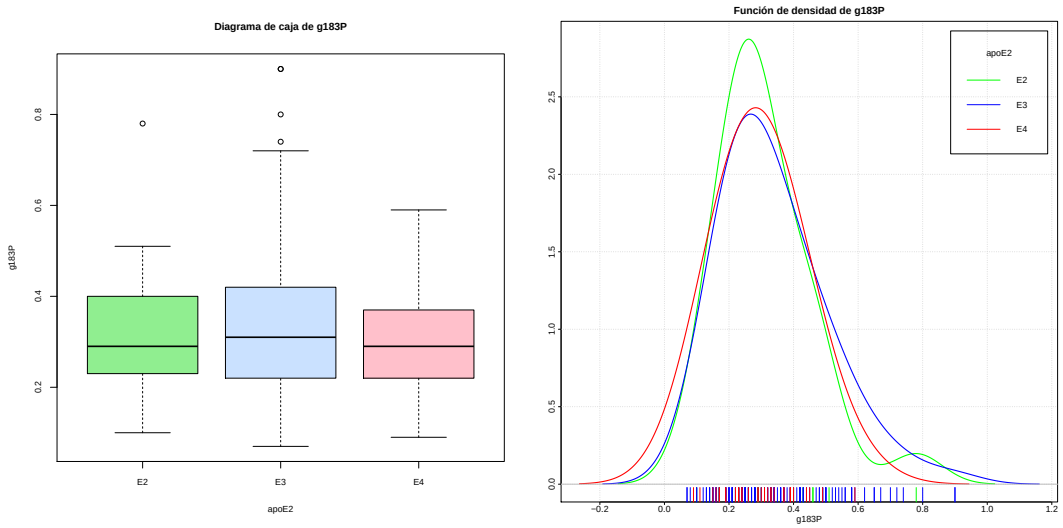

Figura B.7: Diagrama de caja y función de densidad de la variable $g183P$.

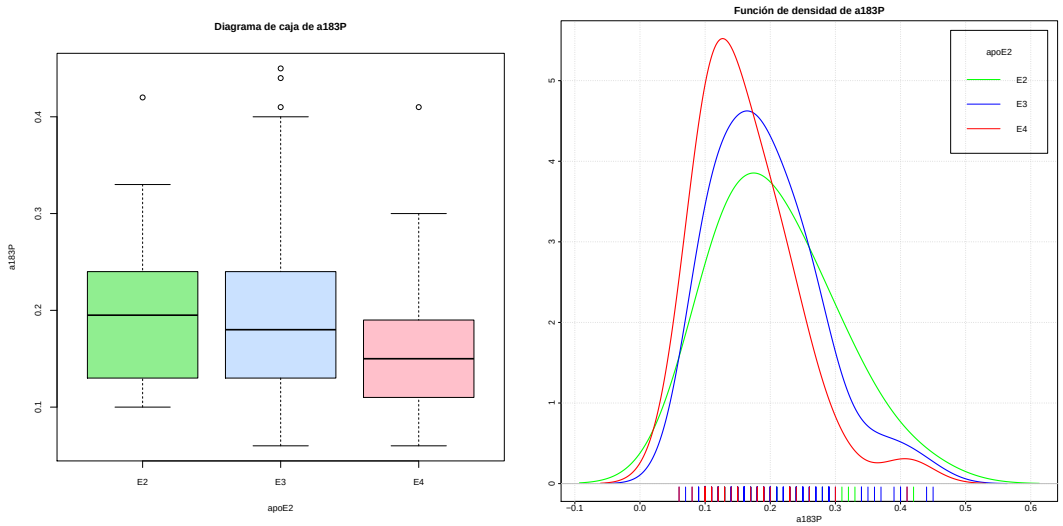

Figura B.8: Diagrama de caja y función de densidad de la variable $a183P$.

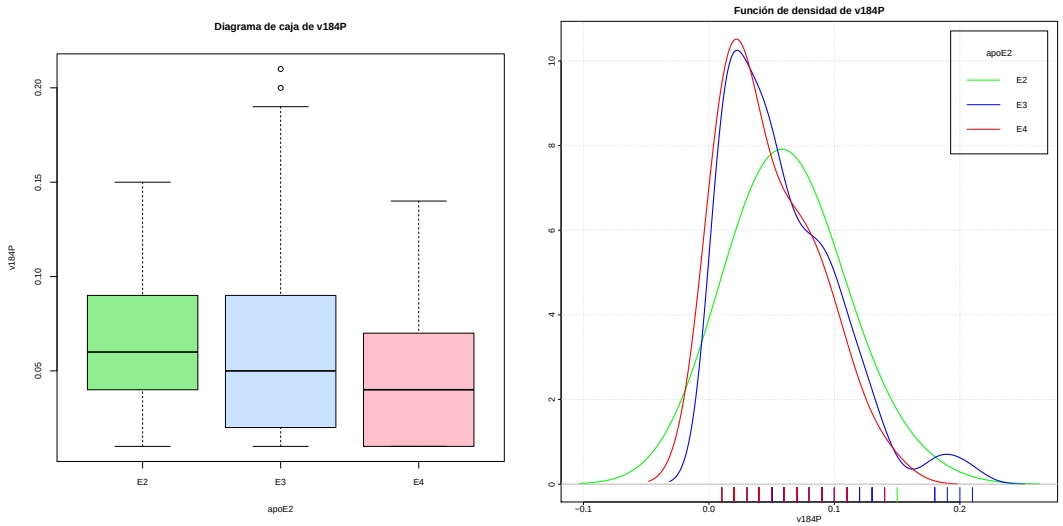

Figura B.9: Diagrama de caja y función de densidad de la variable $v184P$.

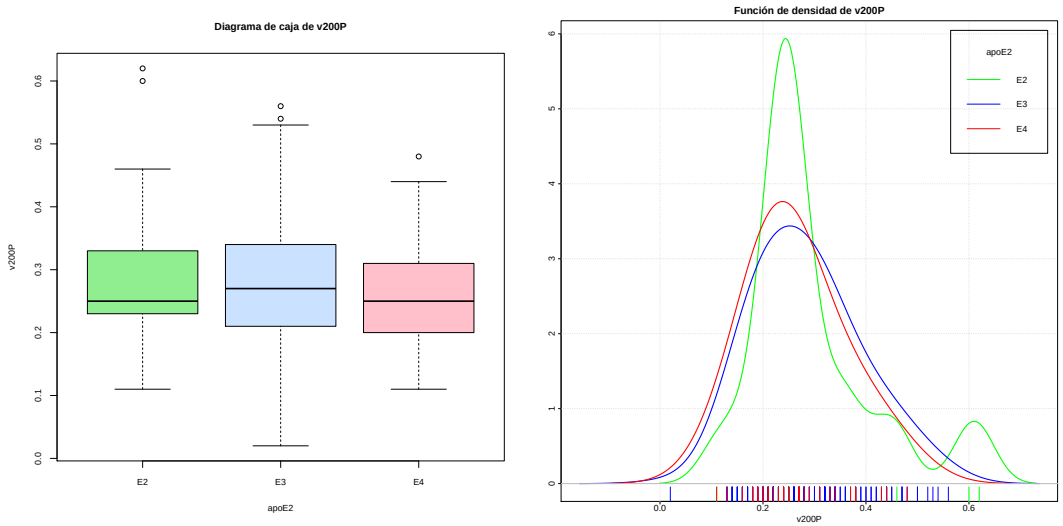

Figura B.10: Diagrama de caja y función de densidad de la variable *v*200*P*.

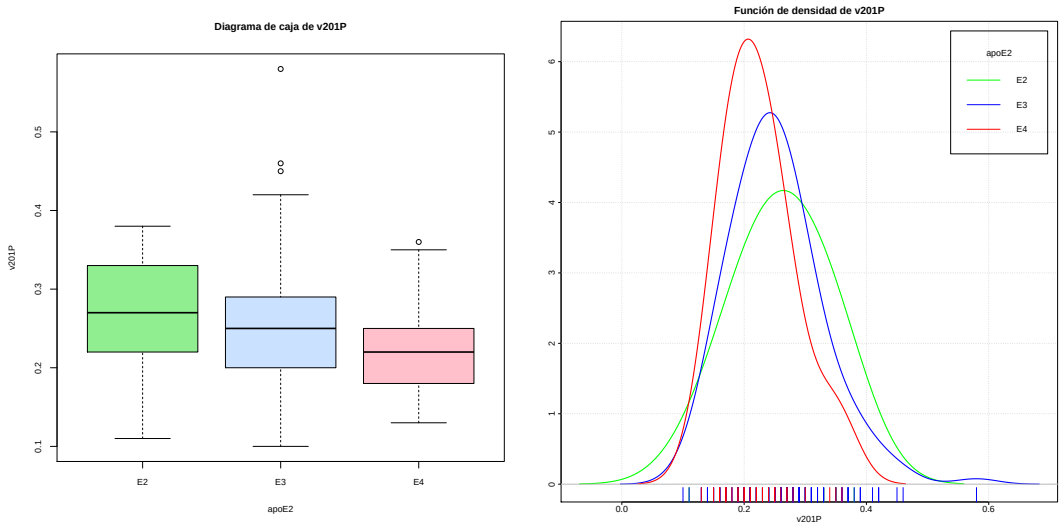

Figura B.11: Diagrama de caja y función de densidad de la variable *v*201*P*.

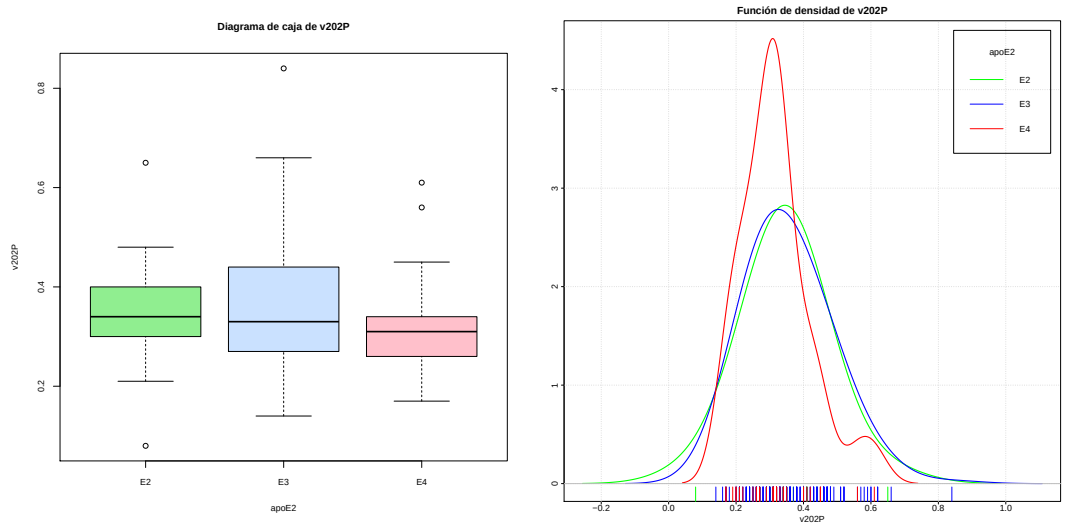

Figura B.12: Diagrama de caja y función de densidad de la variable *v*202*P*.

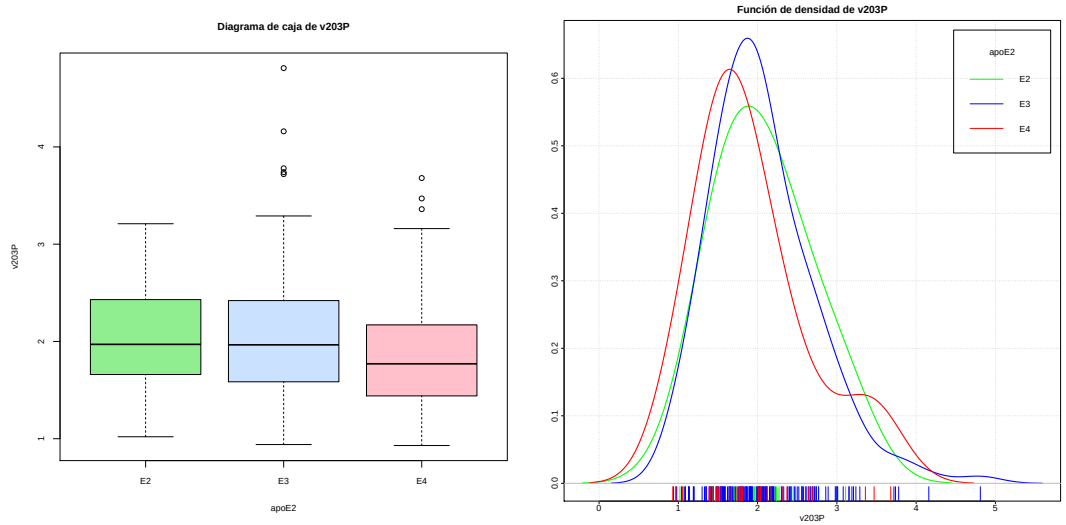

Figura B.13: Diagrama de caja y función de densidad de la variable *v*203*P*.

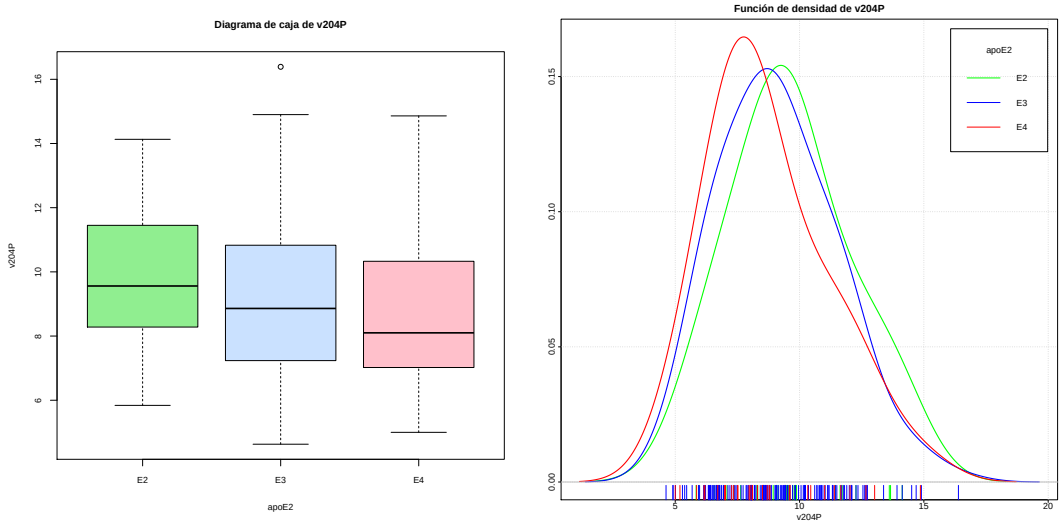

Figura B.14: Diagrama de caja y función de densidad de la variable *v*204*P*.

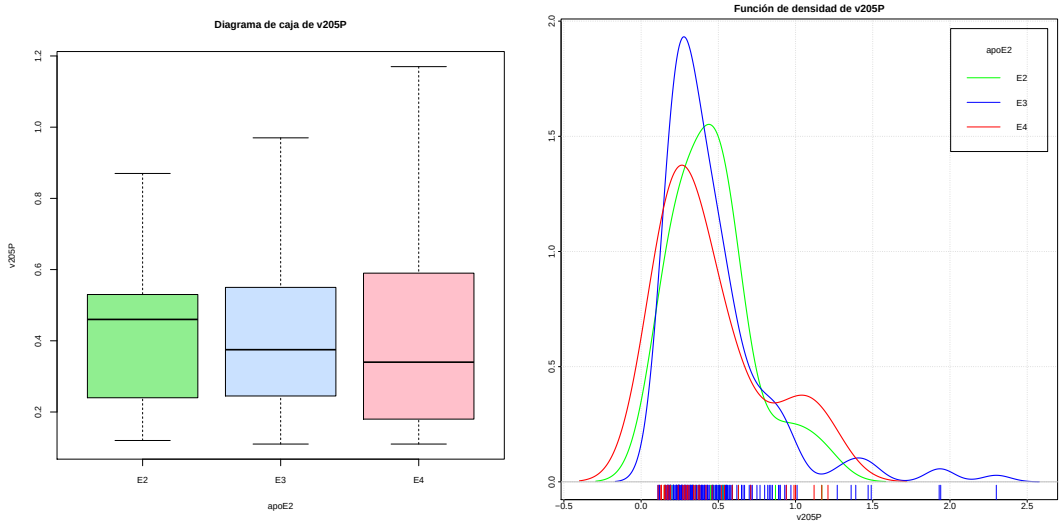

Figura B.15: Diagrama de caja y función de densidad de la variable *v*205*P*.

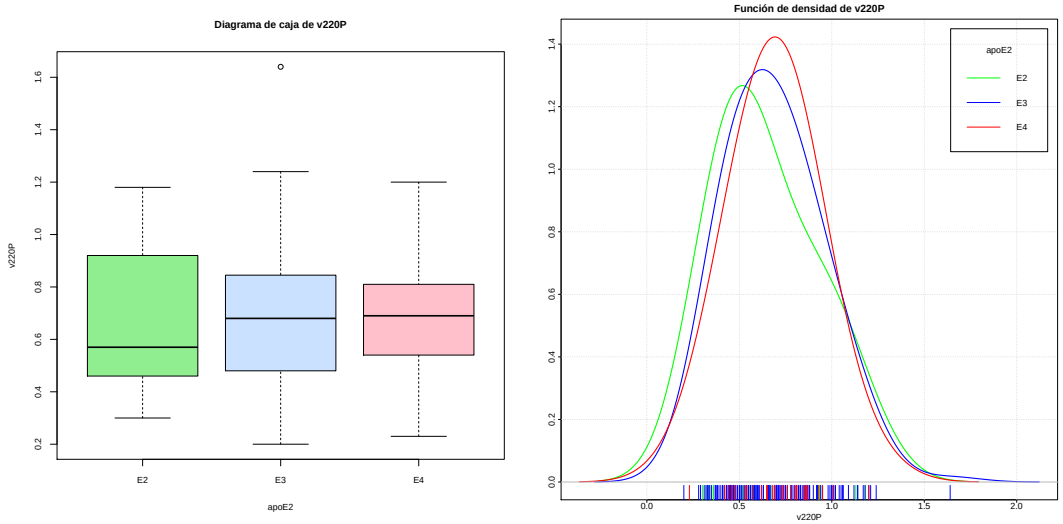

Figura B.16: Diagrama de caja y función de densidad de la variable *v*220*P*.

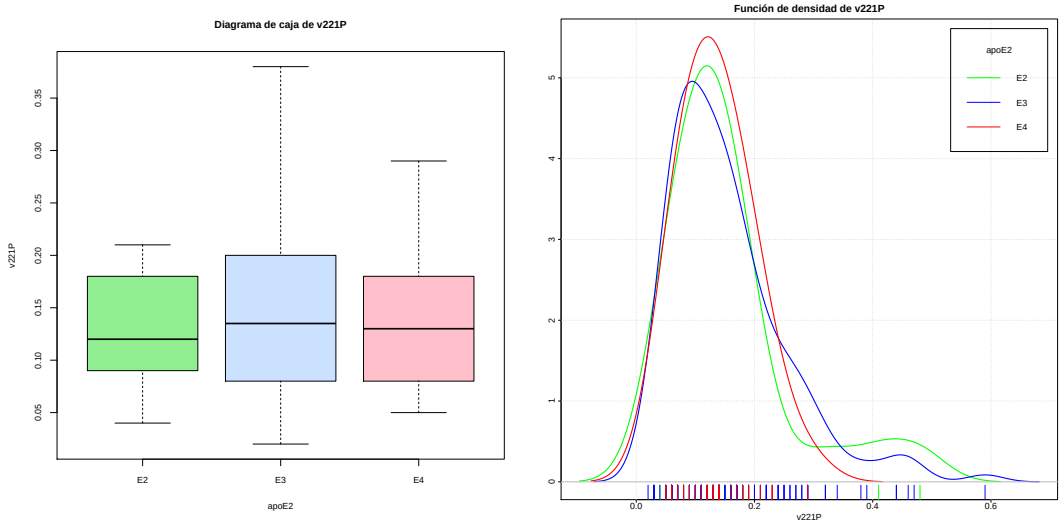

Figura B.17: Diagrama de caja y función de densidad de la variable *v*221*P*.

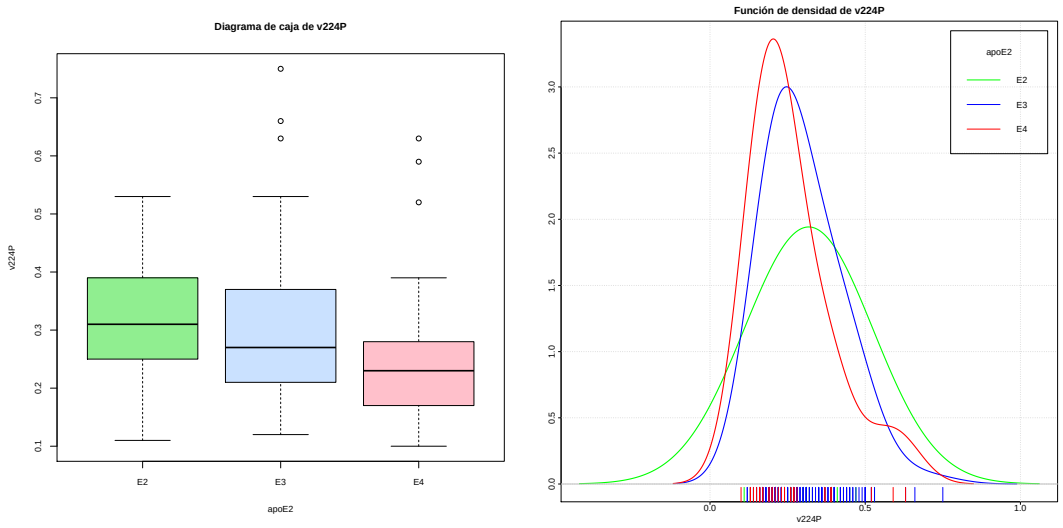

Figura B.18: Diagrama de caja y función de densidad de la variable *v*224*P*.

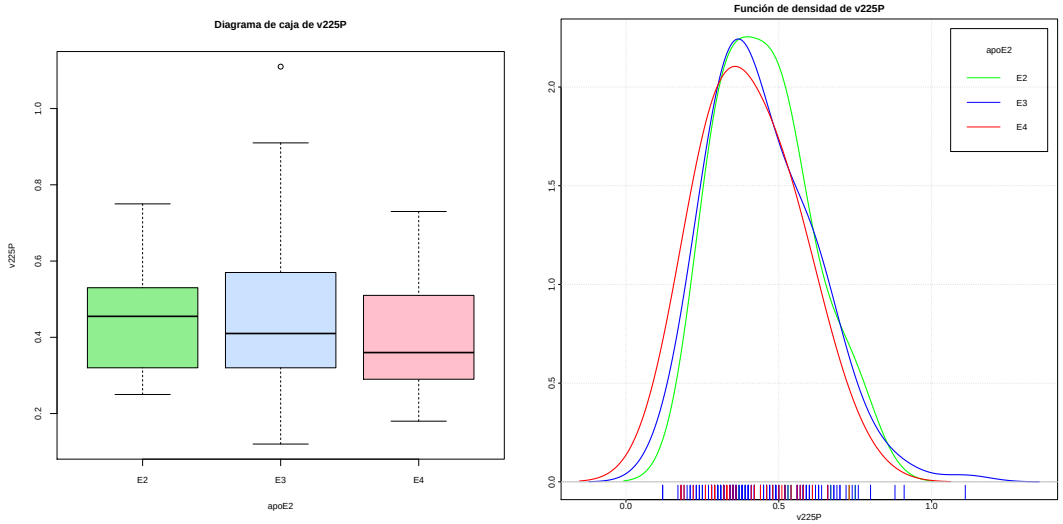

Figura B.19: Diagrama de caja y función de densidad de la variable *v*225*P*.

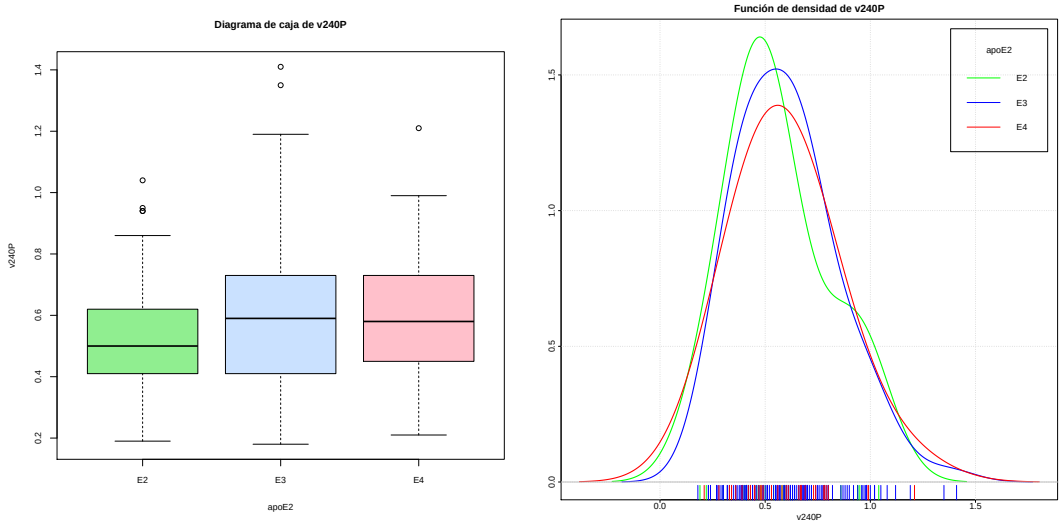

Figura B.20: Diagrama de caja y función de densidad de la variable *v*240*P*.

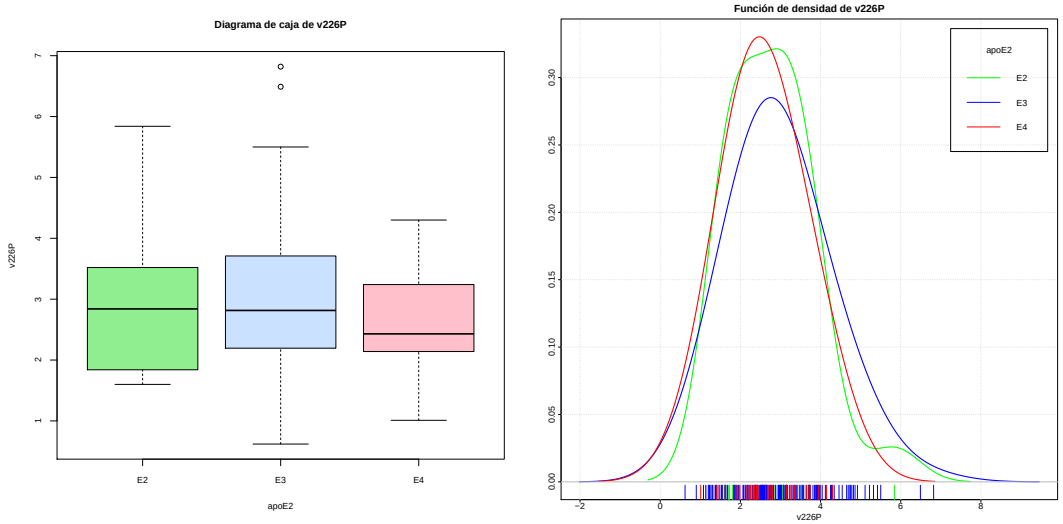

Figura B.21: Diagrama de caja y función de densidad de la variable *v*226*P*.

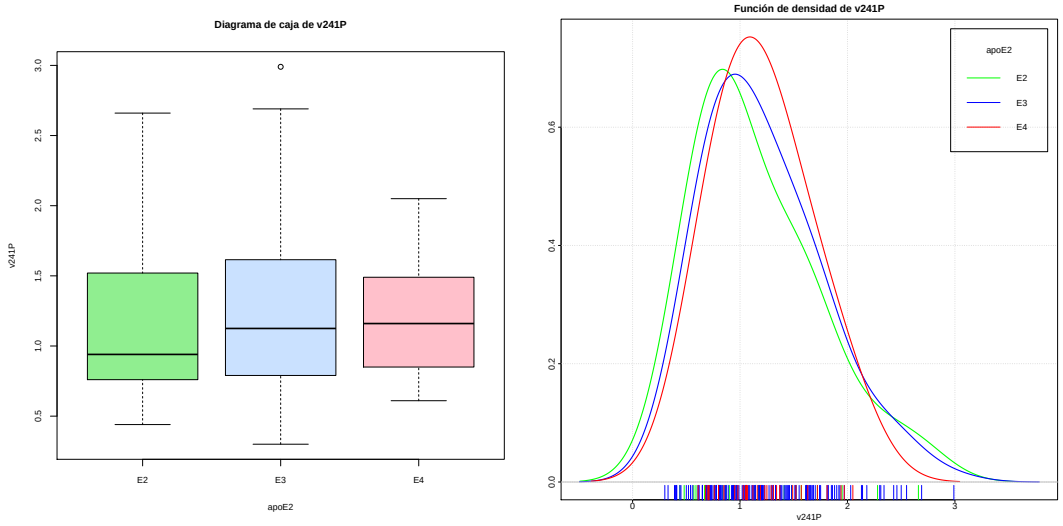

Figura B.22: Diagrama de caja y función de densidad de la variable *v*241*P*.

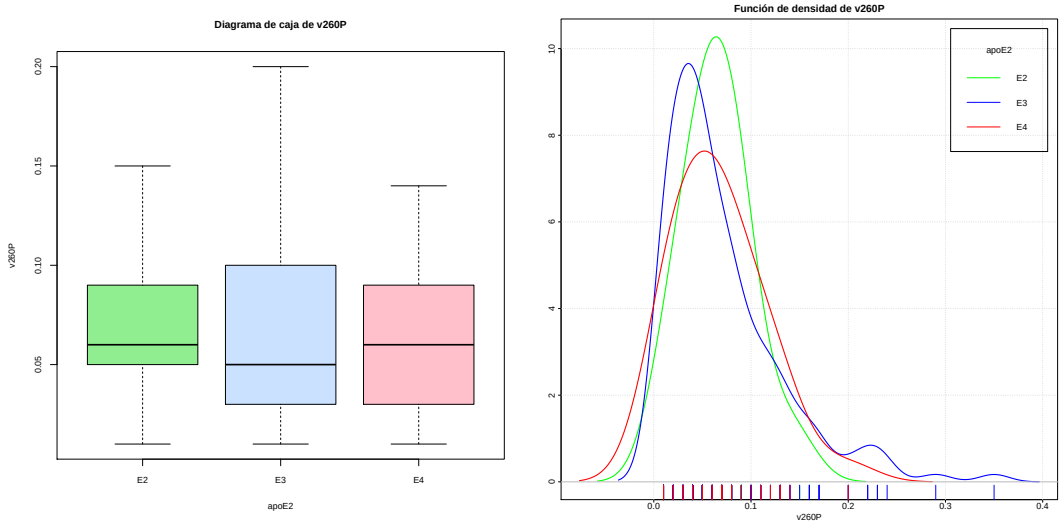

Figura B.23: Diagrama de caja y función de densidad de la variable *v*260*P*.

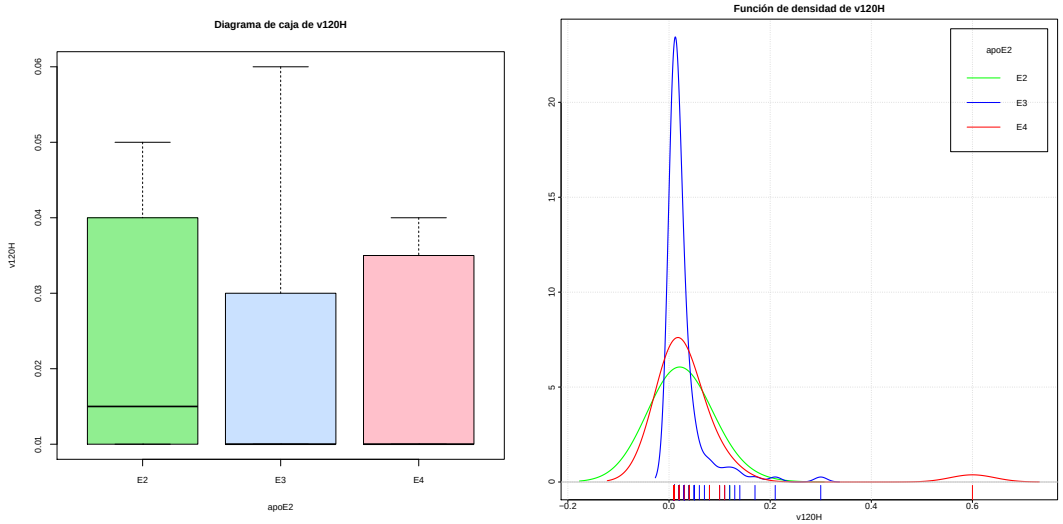

Figura B.24: Diagrama de caja y función de densidad de la variable $v120H$.

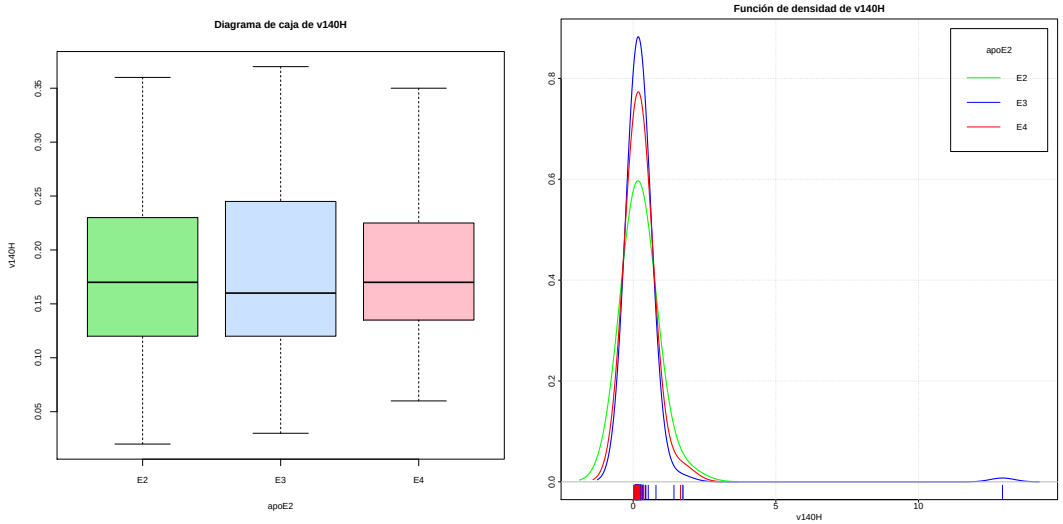

Figura B.25: Diagrama de caja y función de densidad de la variable $v140H$.

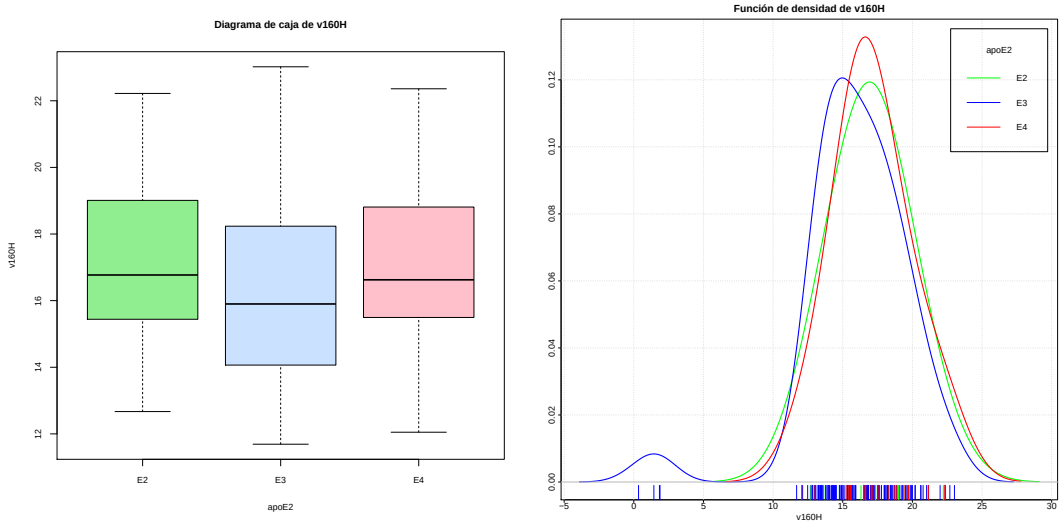

Figura B.26: Diagrama de caja y función de densidad de la variable $v160H$.

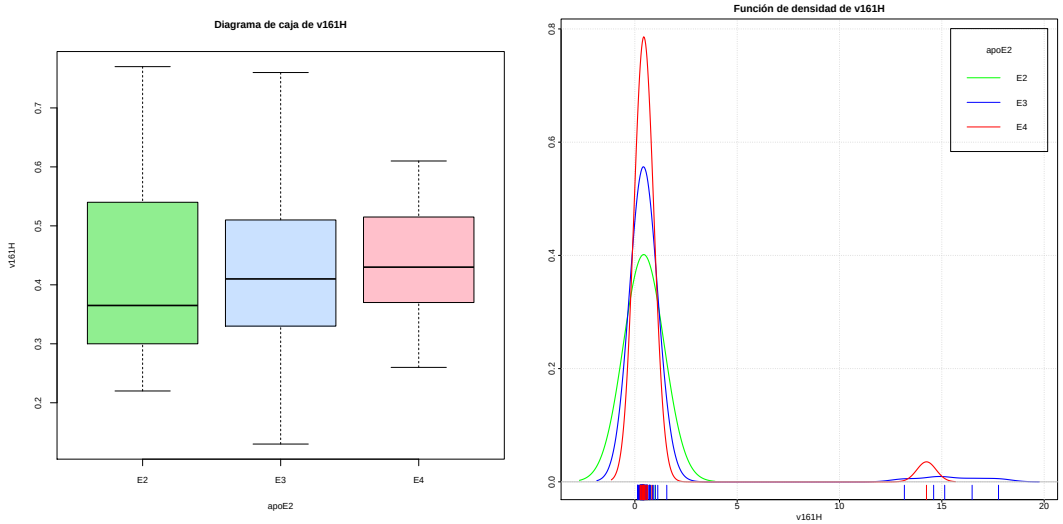

Figura B.27: Diagrama de caja y función de densidad de la variable $v161H$.

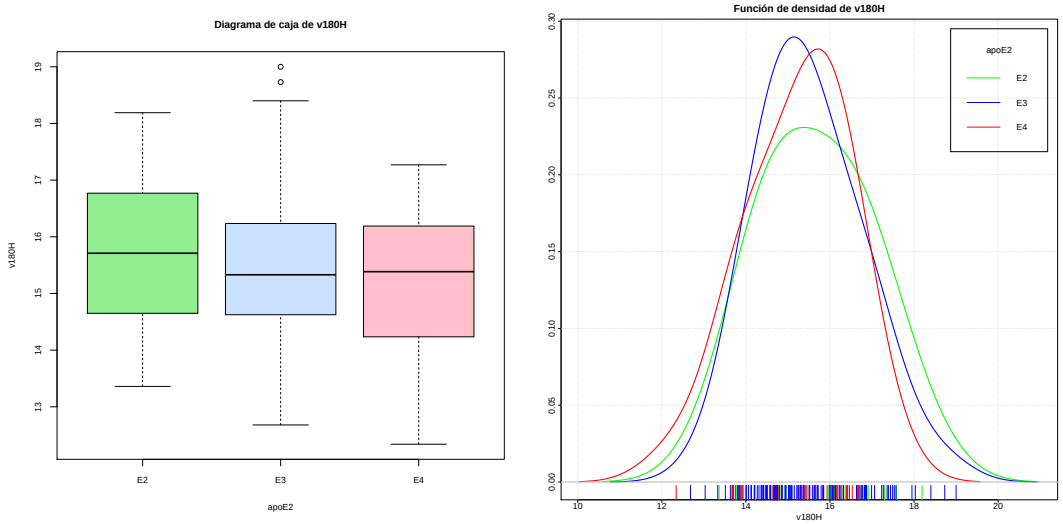

Figura B.28: Diagrama de caja y función de densidad de la variable $v180H$.

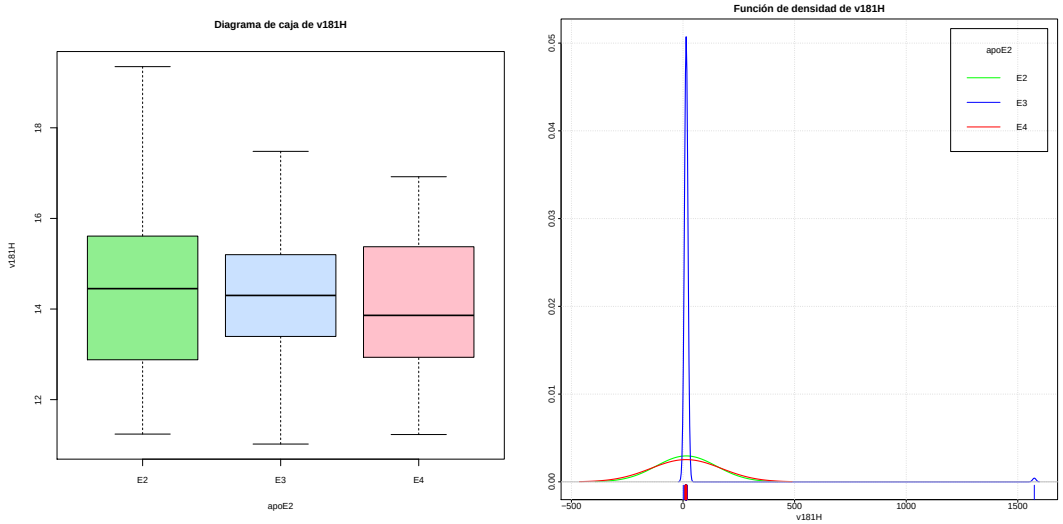

Figura B.29: Diagrama de caja y función de densidad de la variable $v181H$.

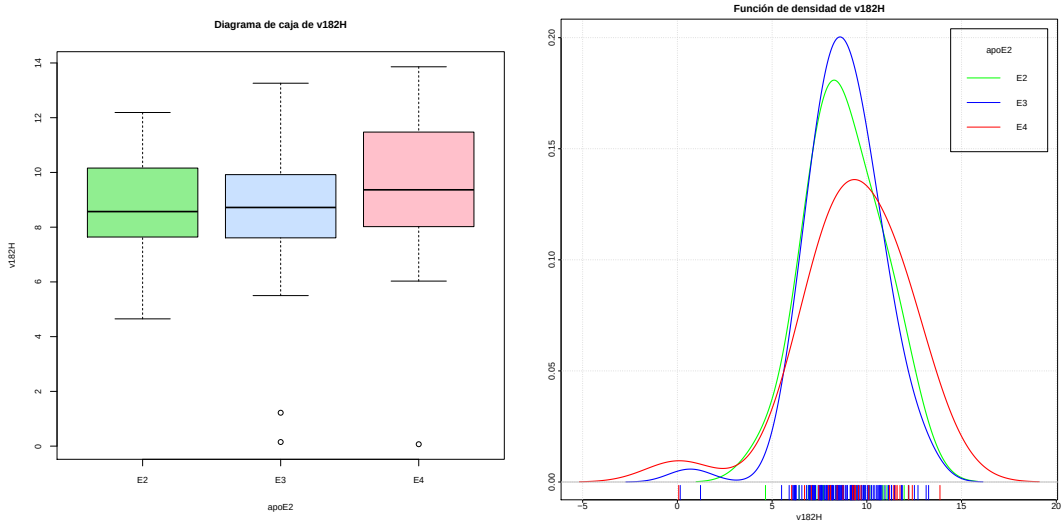

Figura B.30: Diagrama de caja y función de densidad de la variable $v182H$.

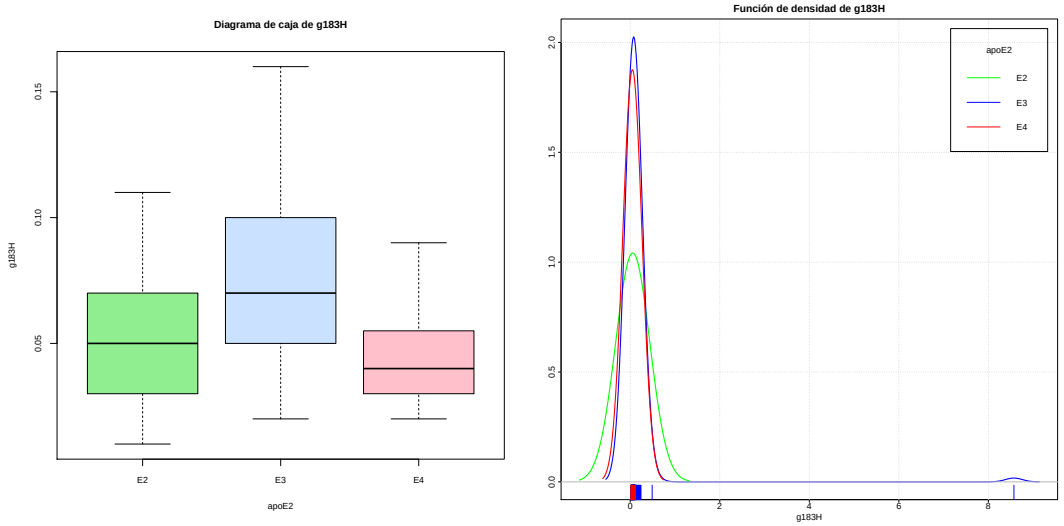

Figura B.31: Diagrama de caja y función de densidad de la variable $g183H$.

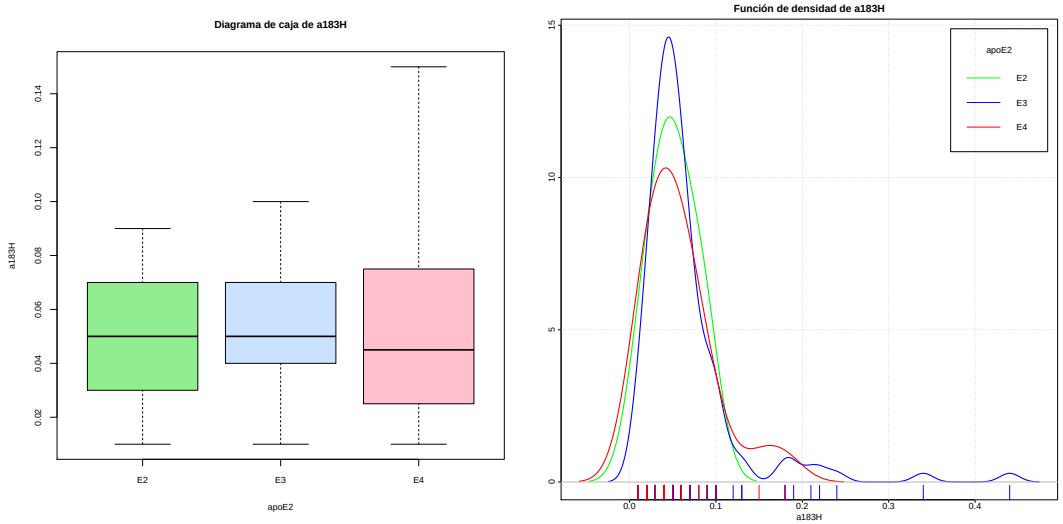

Figura B.32: Diagrama de caja y función de densidad de la variable $a183H$.

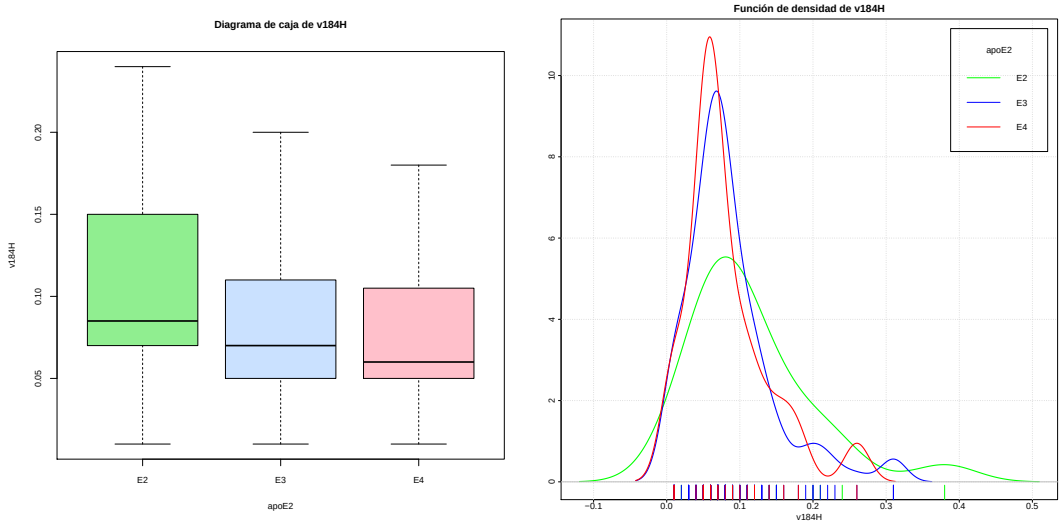

Figura B.33: Diagrama de caja y función de densidad de la variable $v184H$.

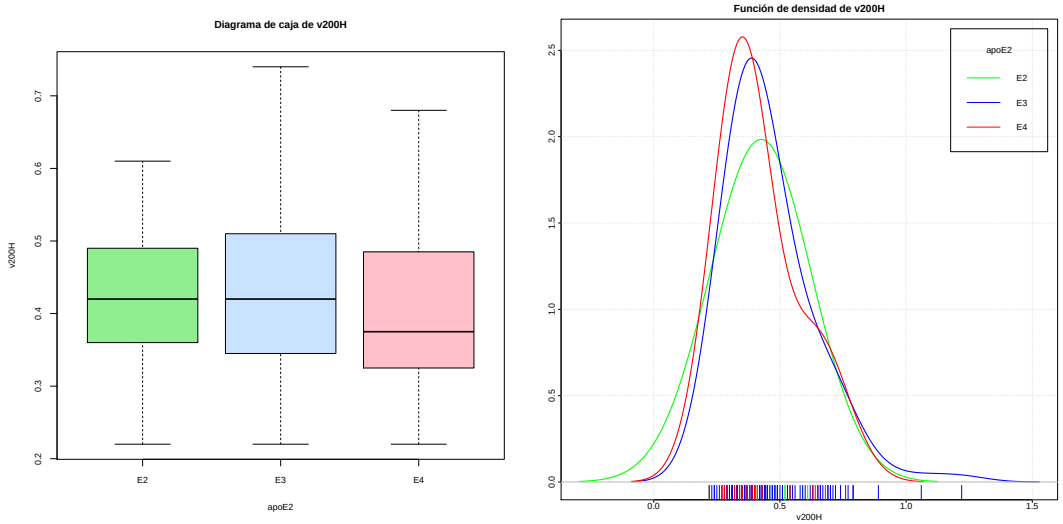

Figura B.34: Diagrama de caja y función de densidad de la variable $v200H$.

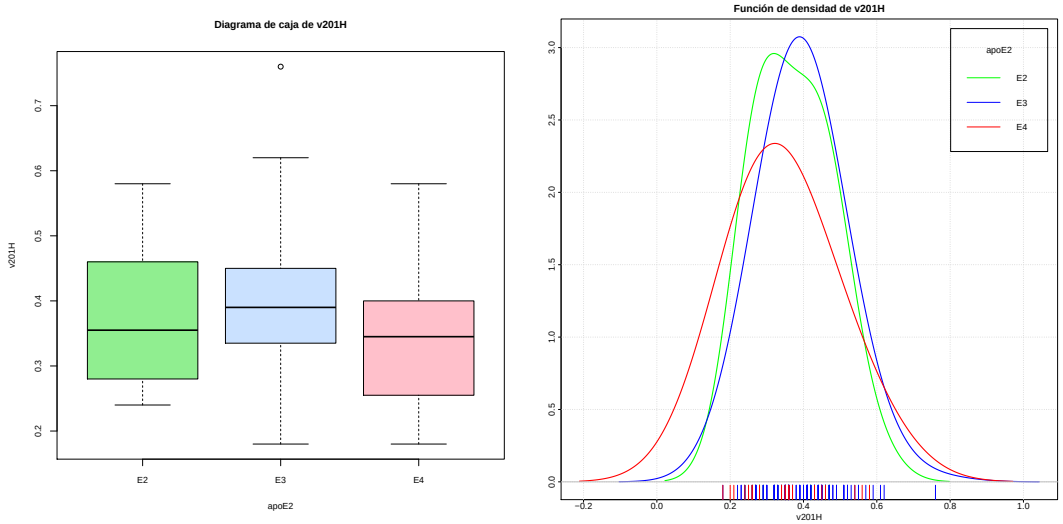

Figura B.35: Diagrama de caja y función de densidad de la variable $v201H$.

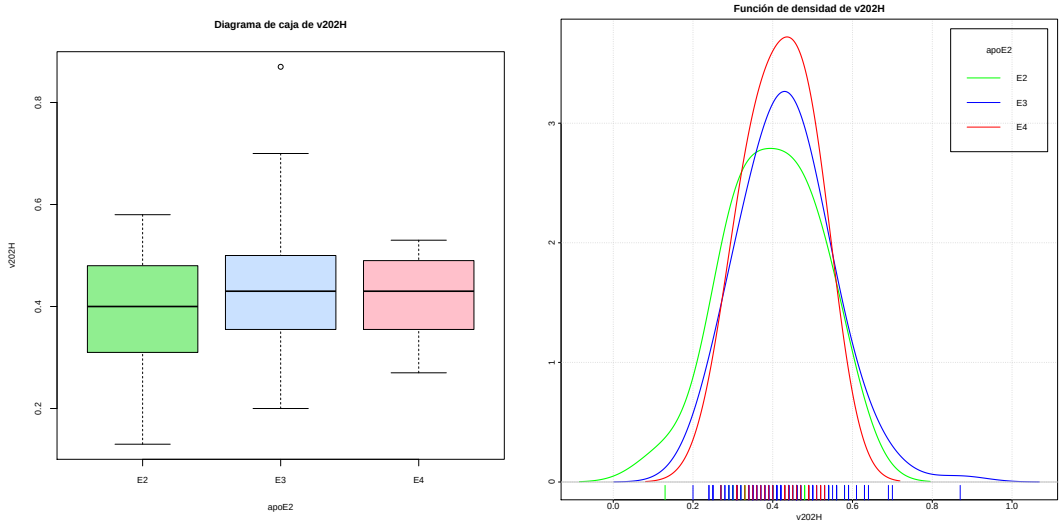

Figura B.36: Diagrama de caja y función de densidad de la variable $v202H$.

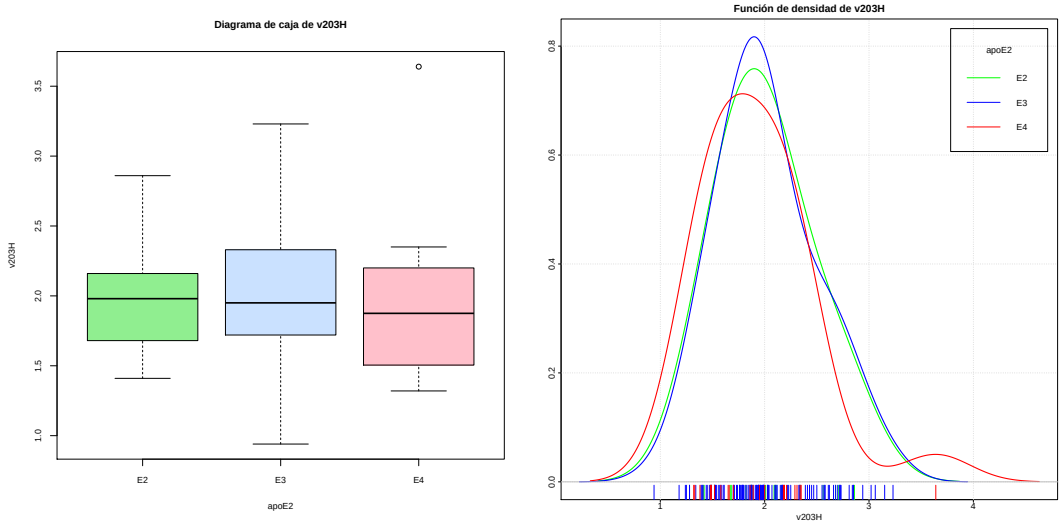

Figura B.37: Diagrama de caja y función de densidad de la variable $v203H$.

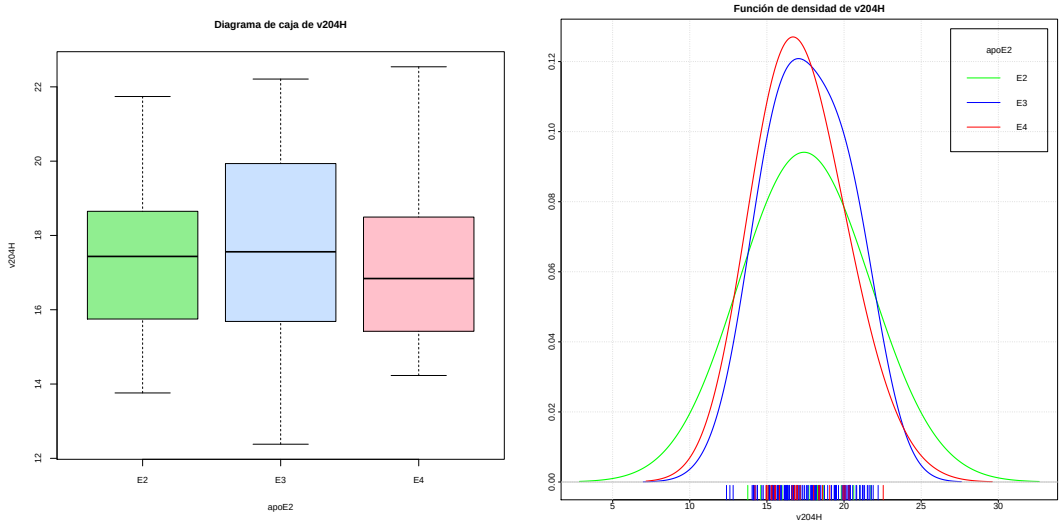

Figura B.38: Diagrama de caja y función de densidad de la variable $v204H$.

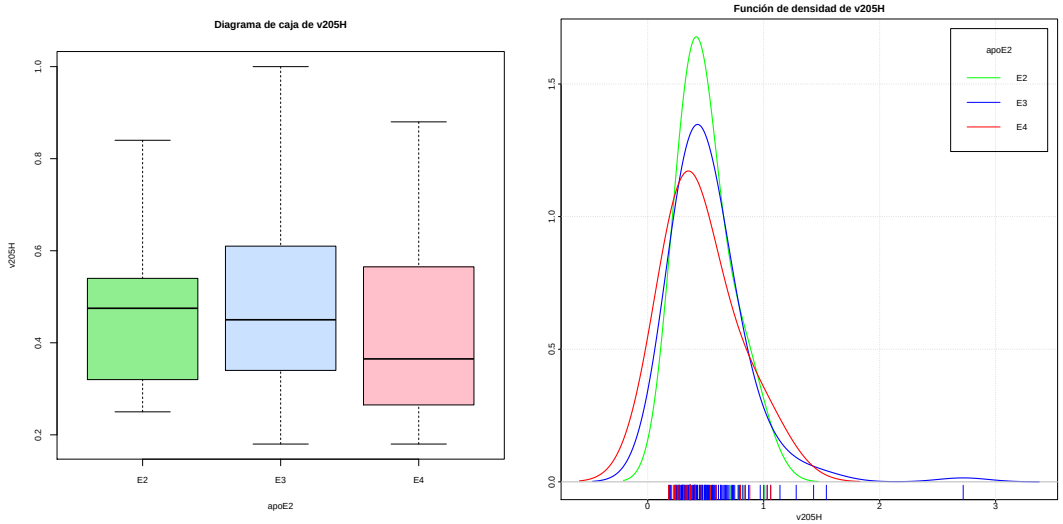

Figura B.39: Diagrama de caja y función de densidad de la variable $v205H$.

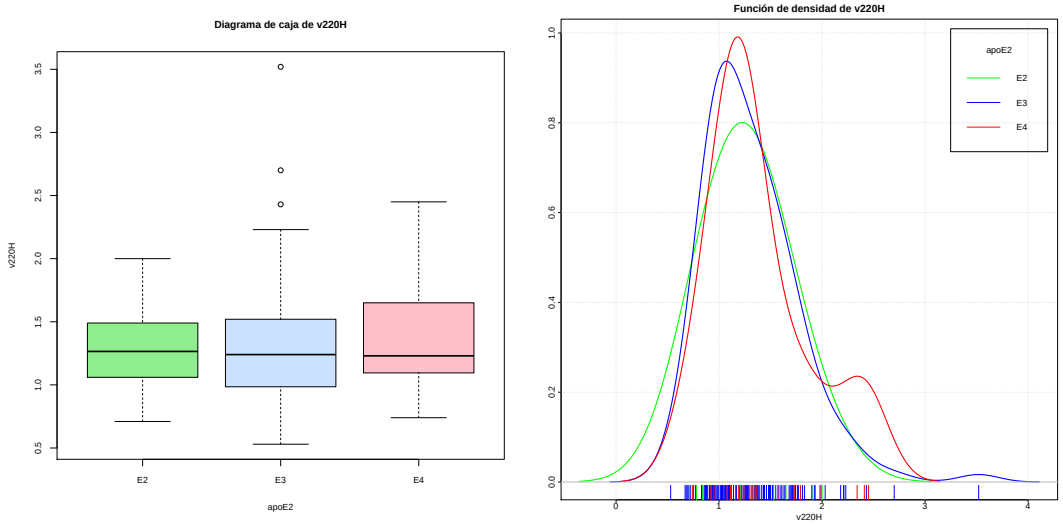

Figura B.40: Diagrama de caja y función de densidad de la variable $v220H$.

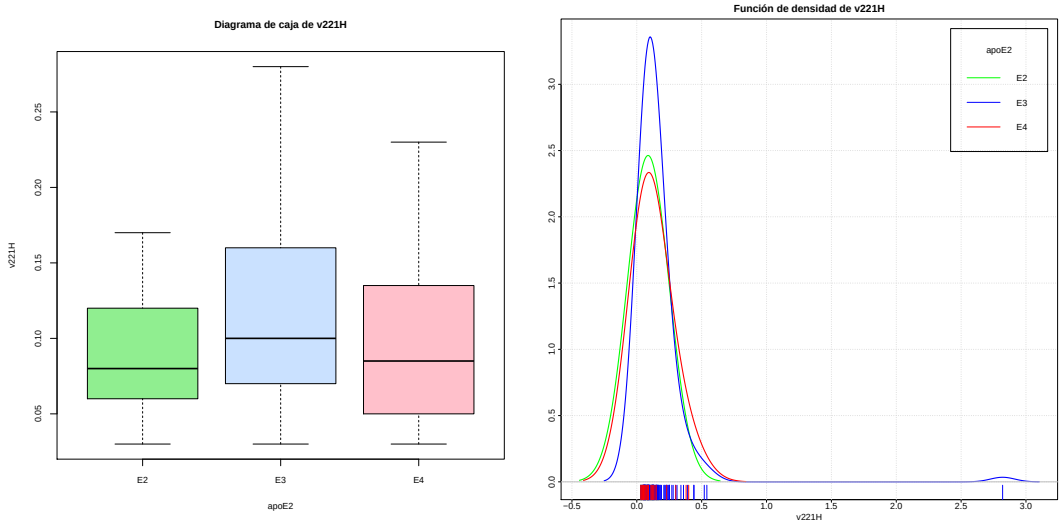

Figura B.41: Diagrama de caja y función de densidad de la variable $v221H$.

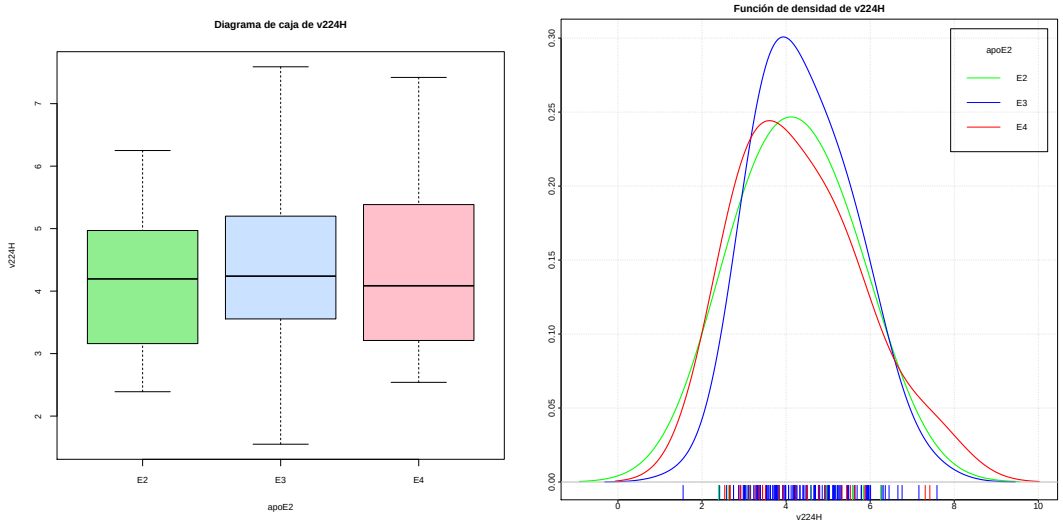

Figura B.42: Diagrama de caja y función de densidad de la variable $v224H$.

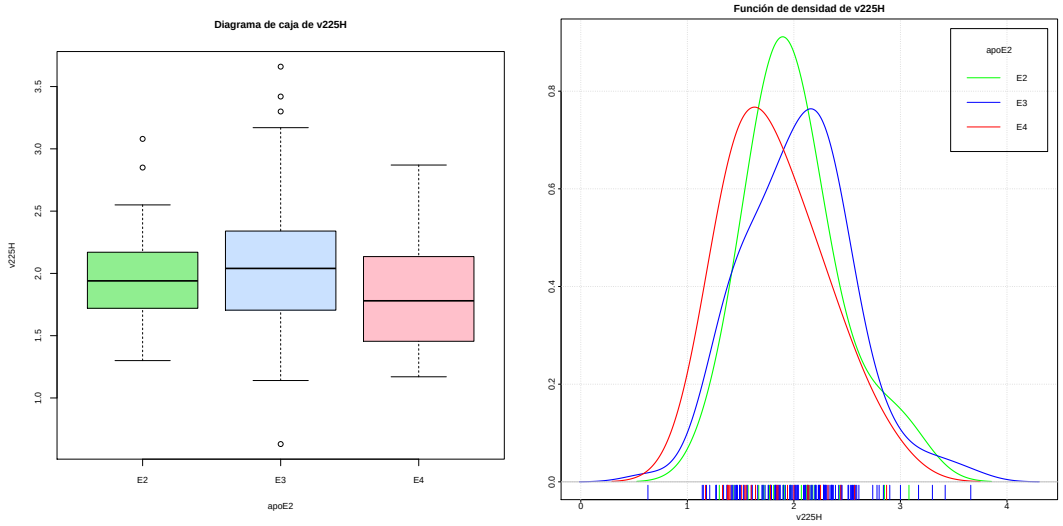

Figura B.43: Diagrama de caja y función de densidad de la variable $v225H$.

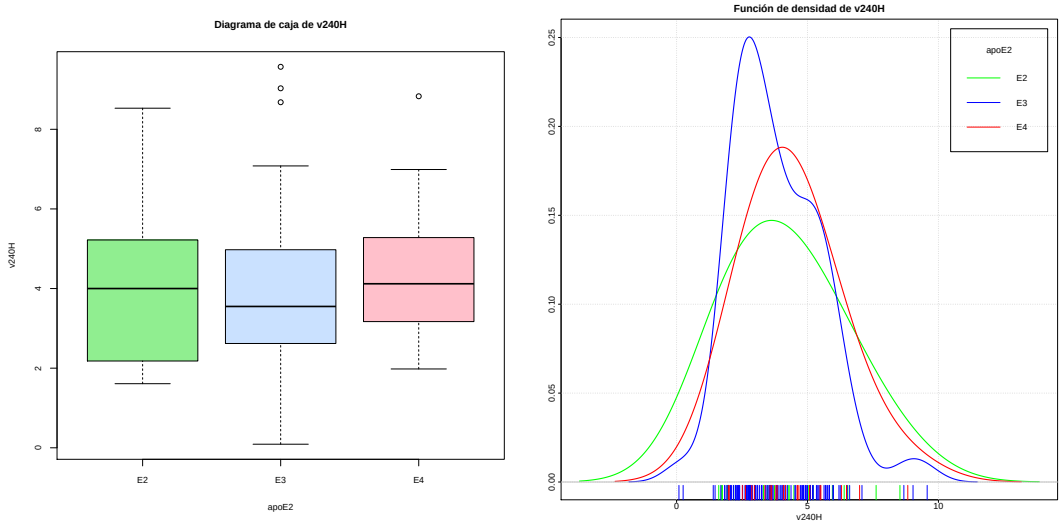

Figura B.44: Diagrama de caja y función de densidad de la variable $v240H$.

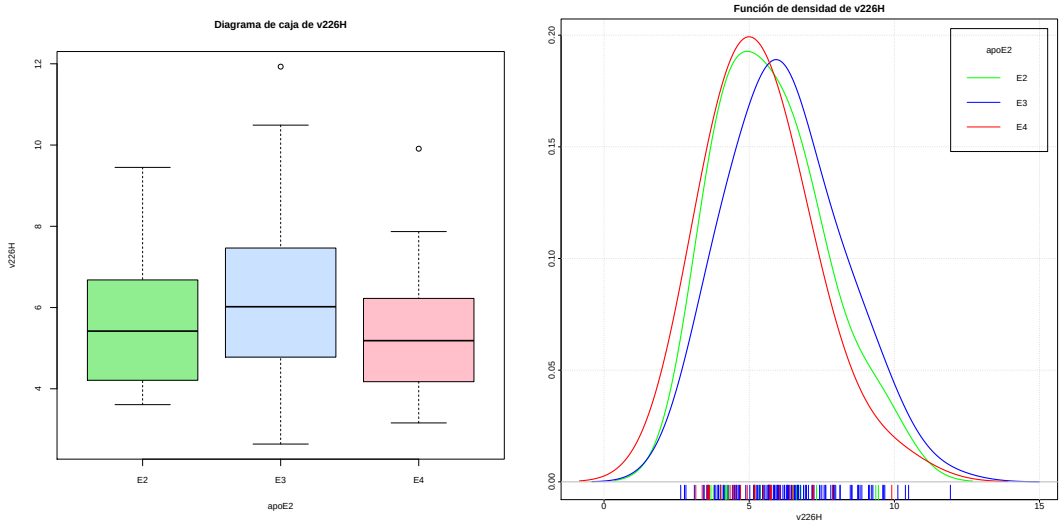

Figura B.45: Diagrama de caja y función de densidad de la variable $v226H$.

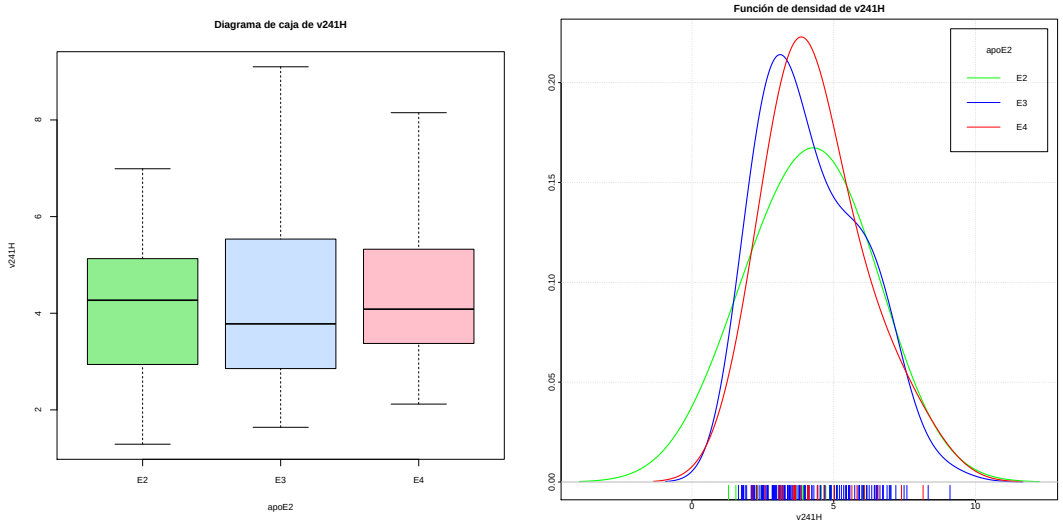

Figura B.46: Diagrama de caja y función de densidad de la variable $v241H$.

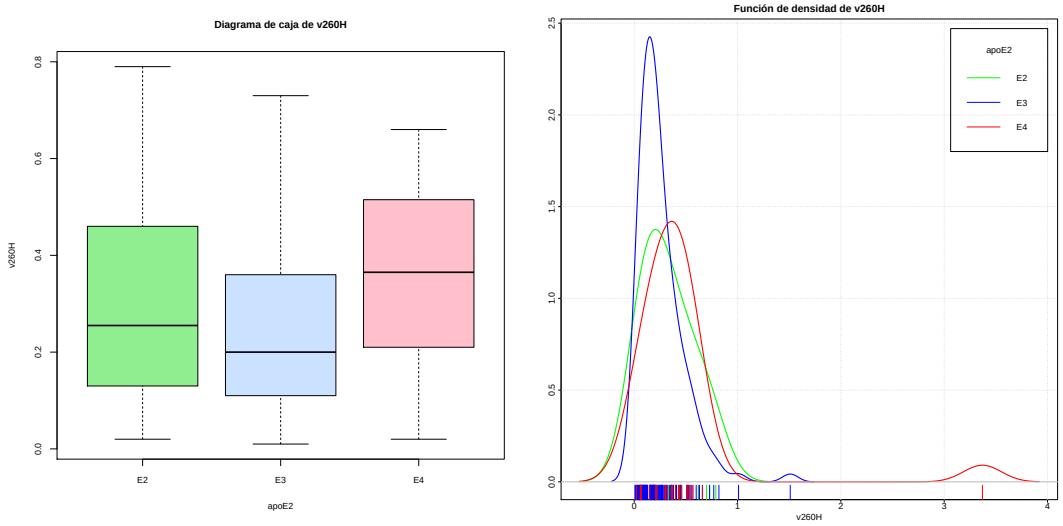

Figura B.47: Diagrama de caja y función de densidad de la variable $v260H$.

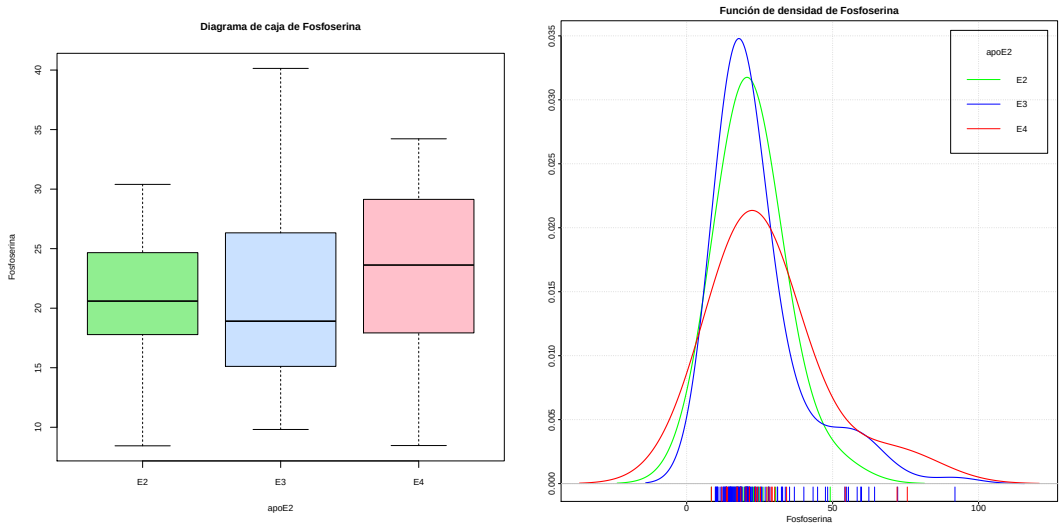

Figura B.48: Diagrama de caja y función de densidad de la variable *Fosfoserina*.

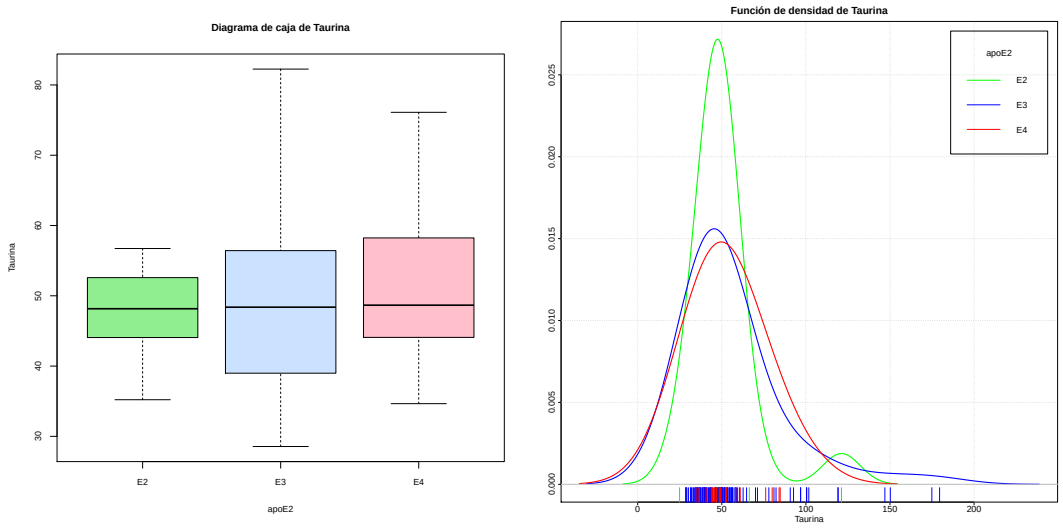

Figura B.49: Diagrama de caja y función de densidad de la variable *Taurina*.

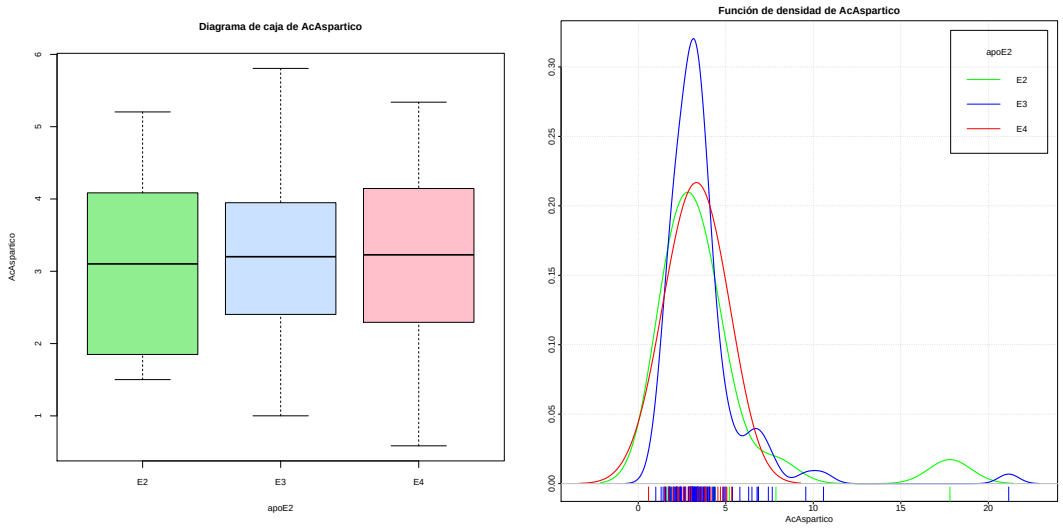

Figura B.50: Diagrama de caja y función de densidad de la variable *AcAspartico*.

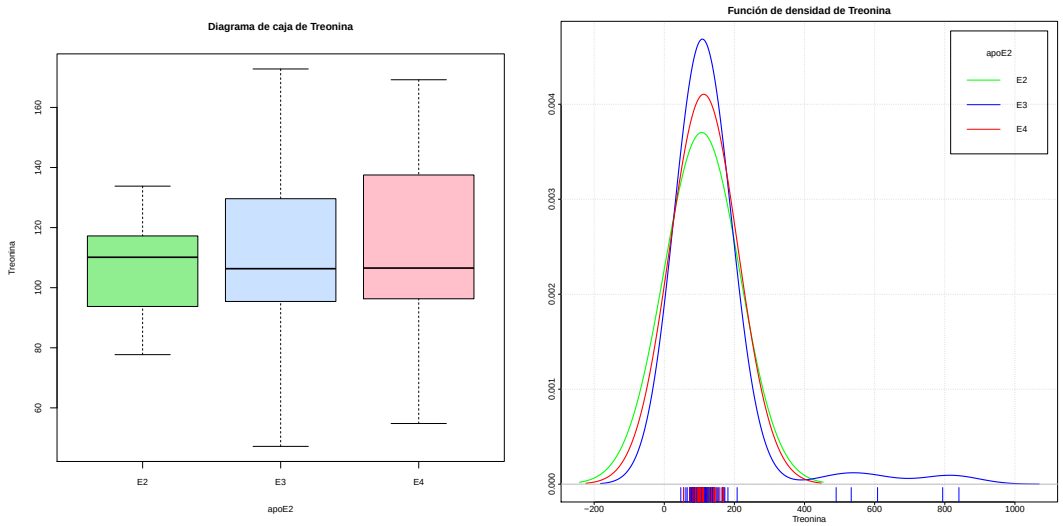

Figura B.51: Diagrama de caja y función de densidad de la variable *Treonina*.

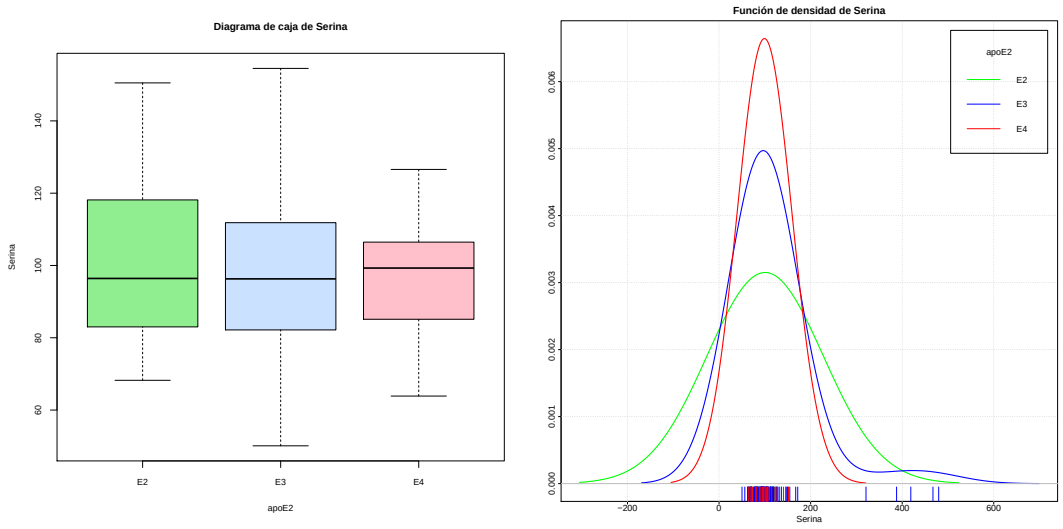

Figura B.52: Diagrama de caja y función de densidad de la variable *Serina*.

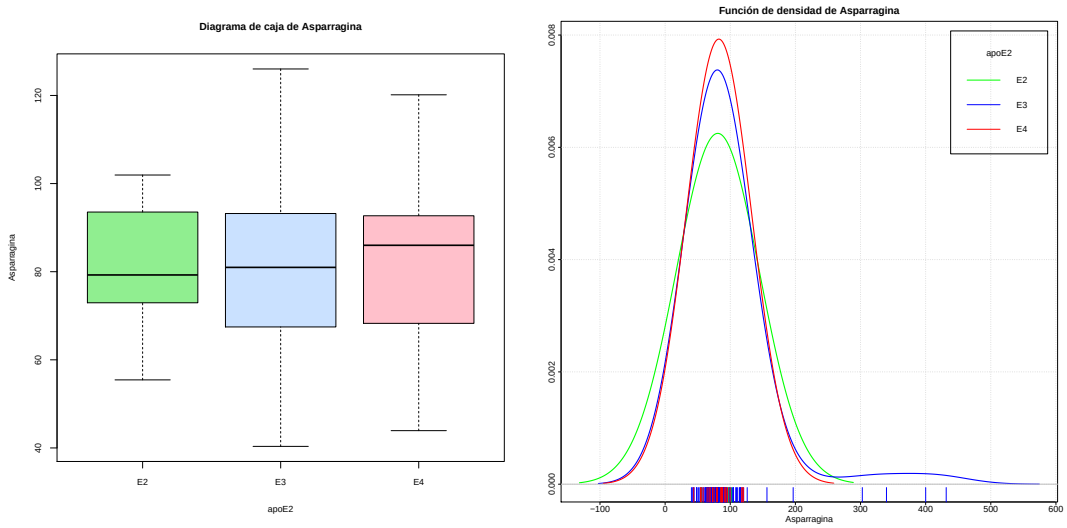

Figura B.53: Diagrama de caja y función de densidad de la variable *Asparragina*.

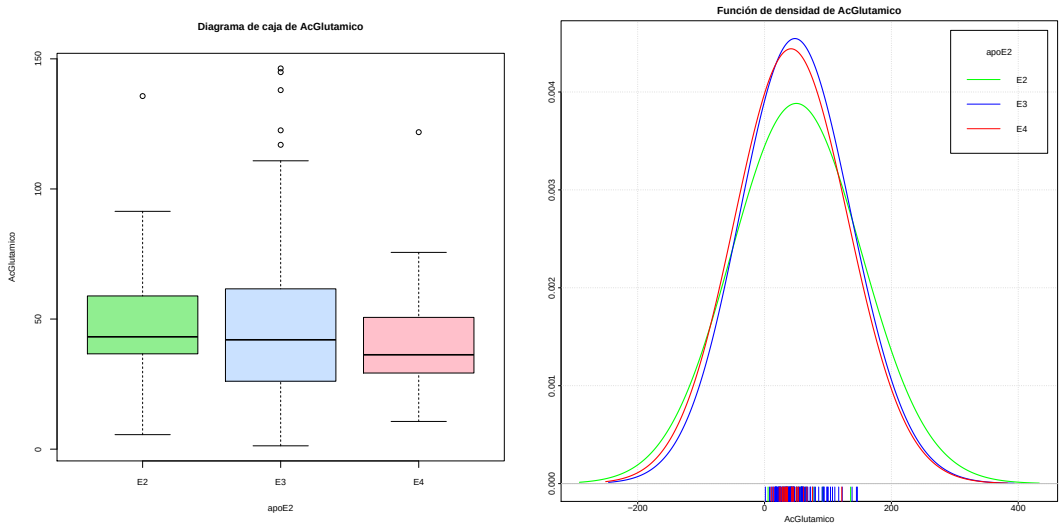

Figura B.54: Diagrama de caja y función de densidad de la variable *AcGlutamico*.

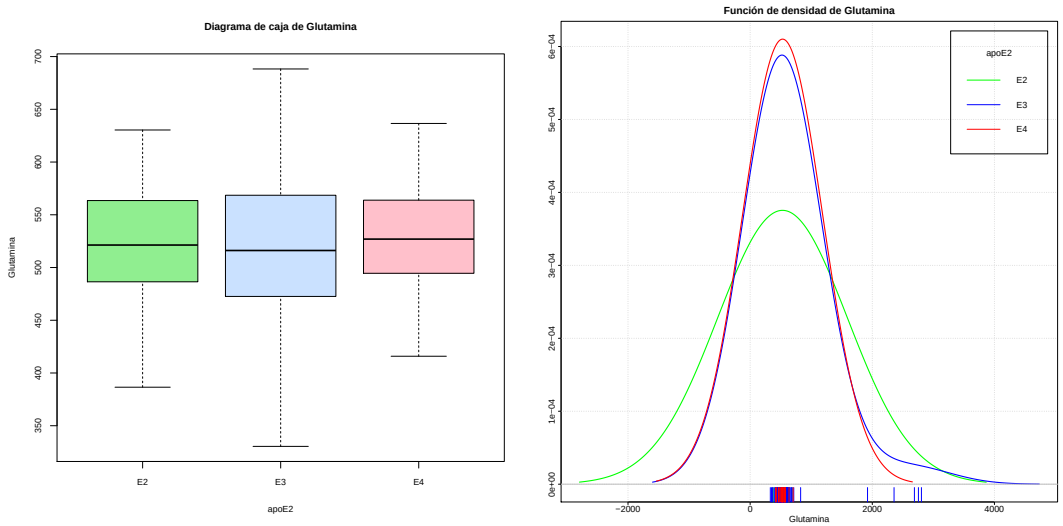

Figura B.55: Diagrama de caja y función de densidad de la variable *Glutamina*.

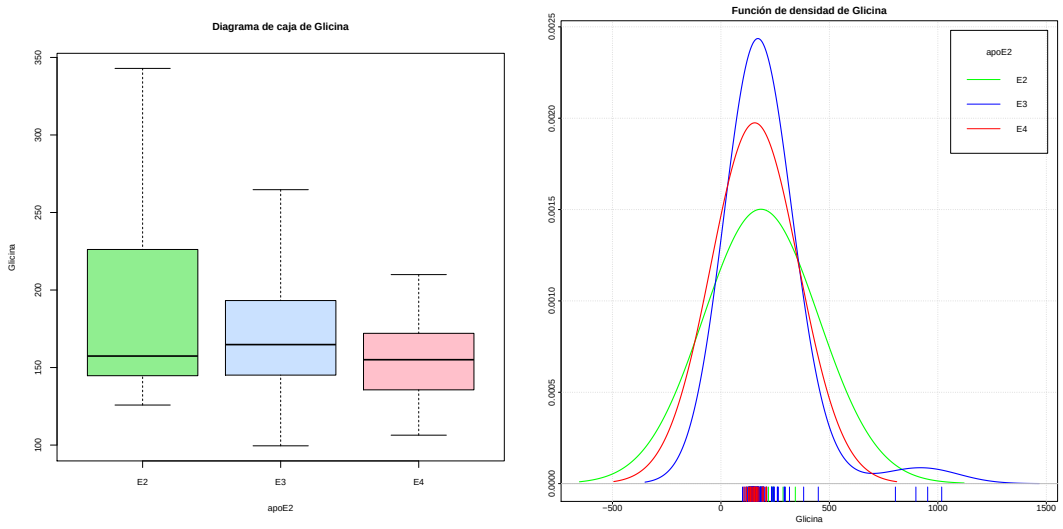

Figura B.56: Diagrama de caja y función de densidad de la variable *Glicina*.

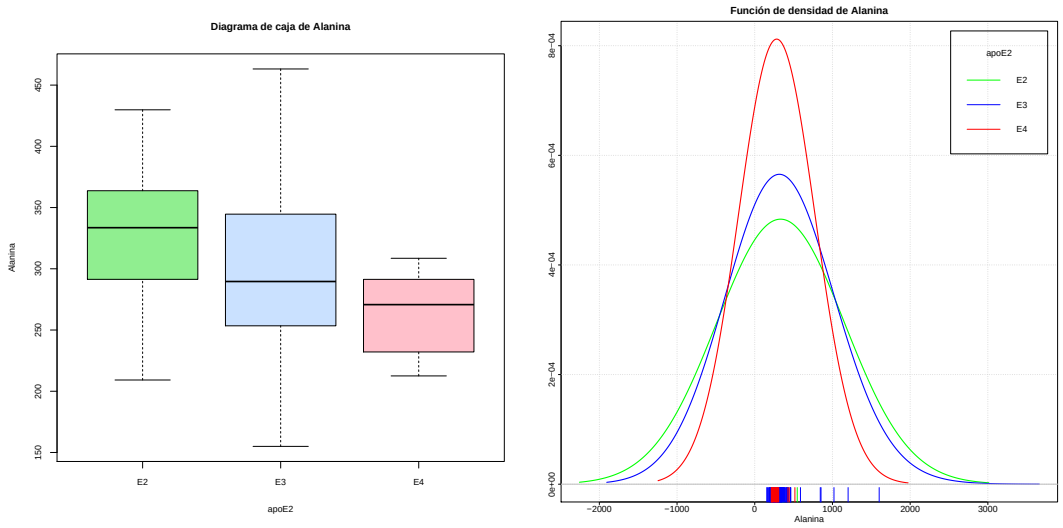

Figura B.57: Diagrama de caja y función de densidad de la variable *Alanina*.

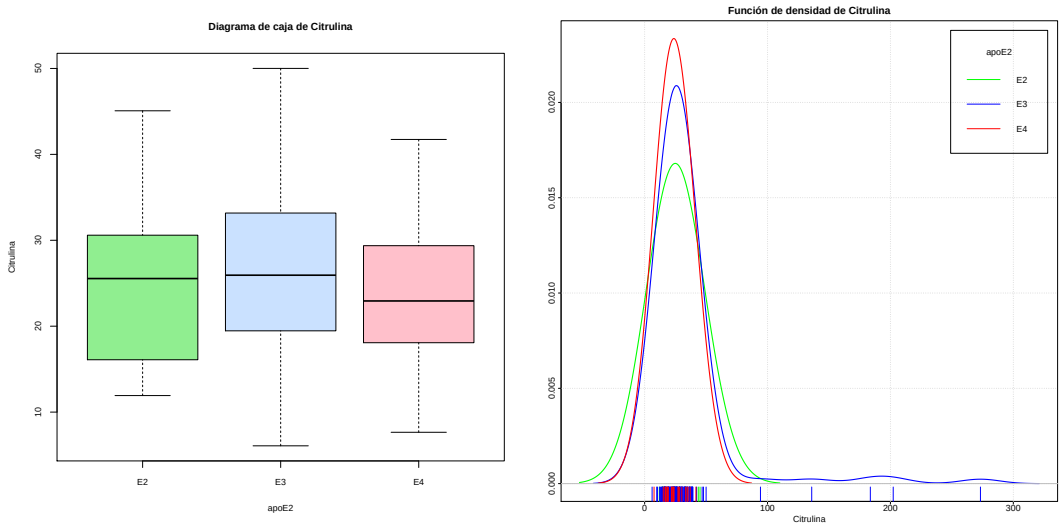

Figura B.58: Diagrama de caja y función de densidad de la variable *Citrulina*.

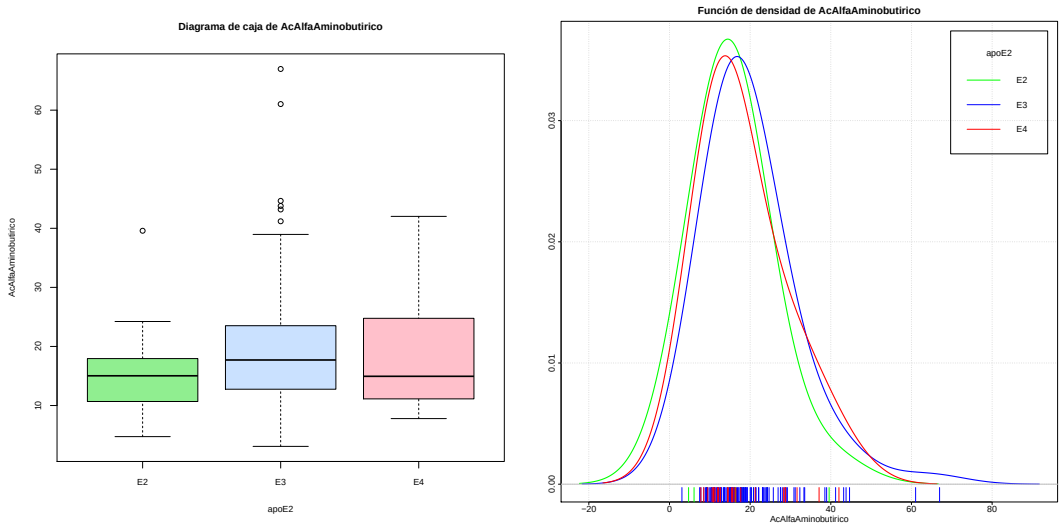

Figura B.59: Diagrama de caja y función de densidad de la variable *AcAlfaAminobutirico*.

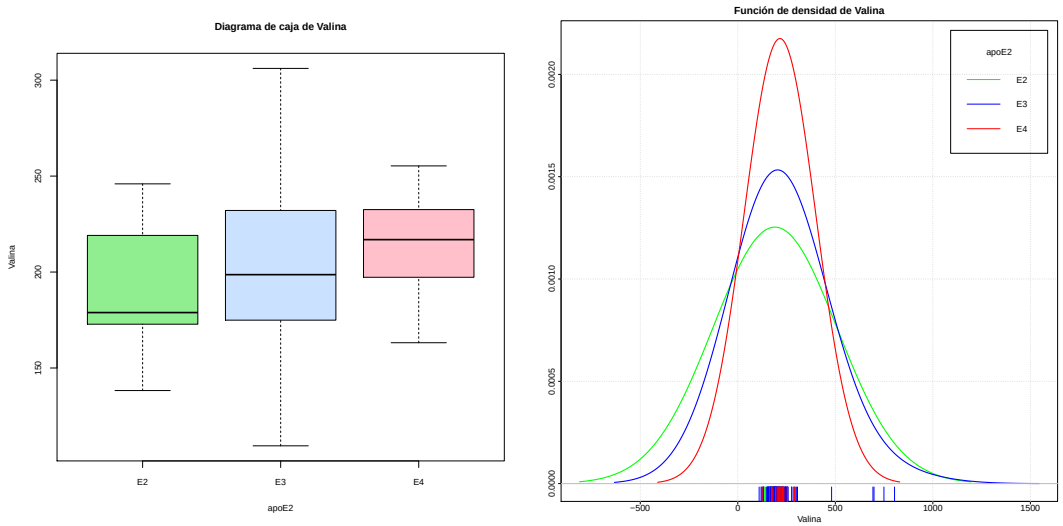

Figura B.60: Diagrama de caja y función de densidad de la variable *Valina*.

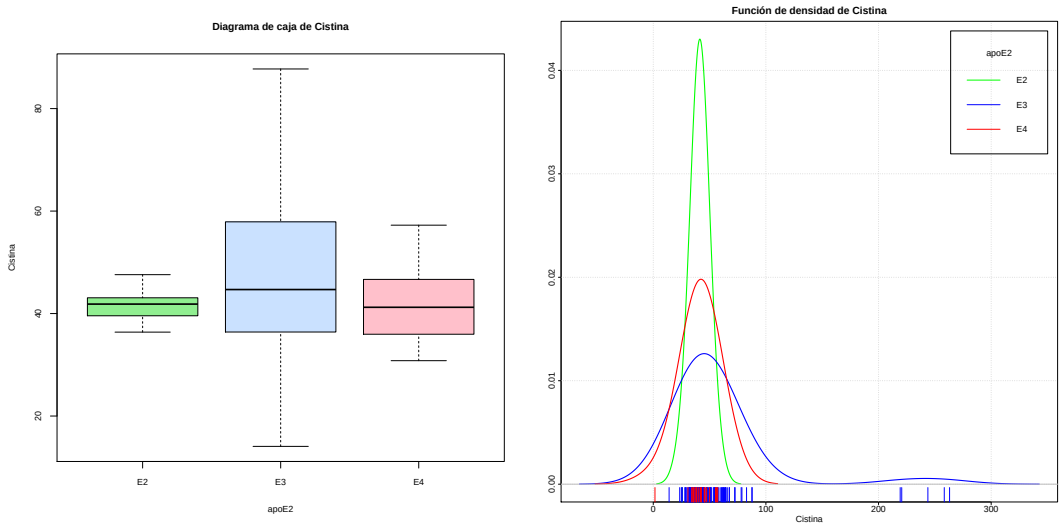

Figura B.61: Diagrama de caja y función de densidad de la variable *Cistina*.

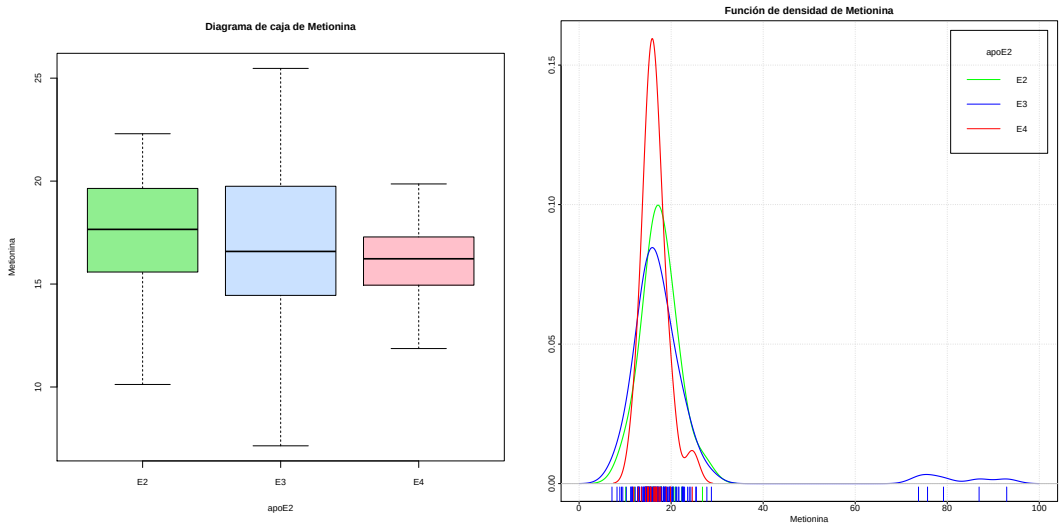

Figura B.62: Diagrama de caja y función de densidad de la variable *Metionina*.

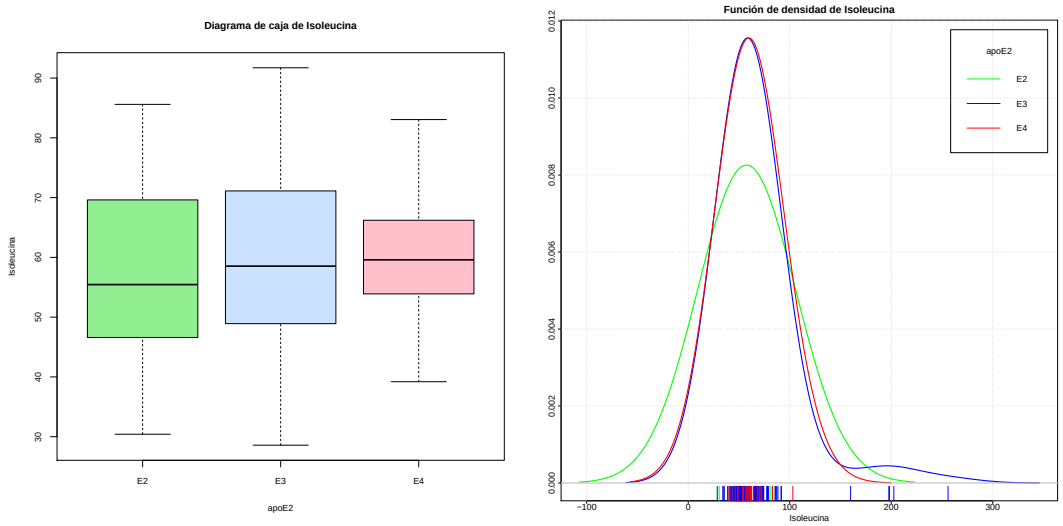

Figura B.63: Diagrama de caja y función de densidad de la variable *Isoleucina*.

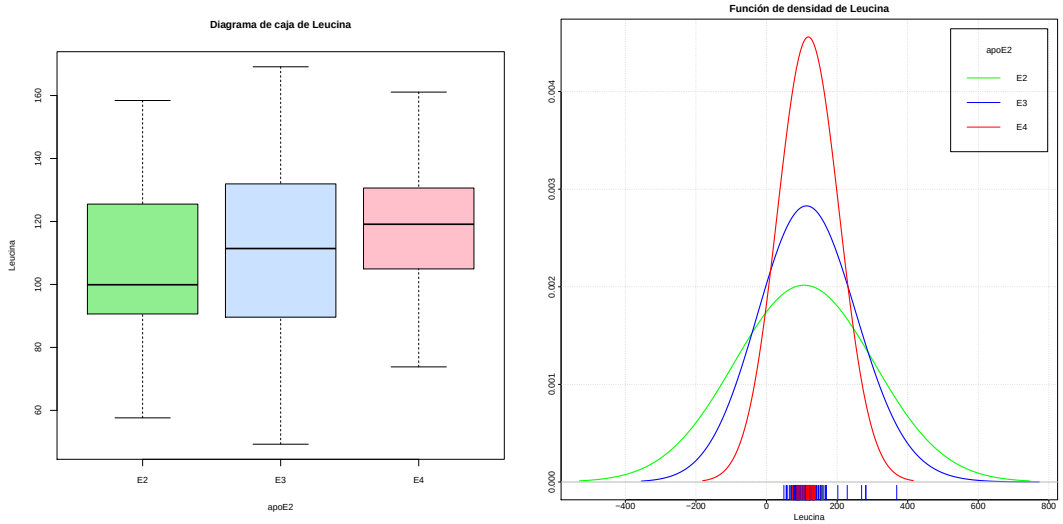

Figura B.64: Diagrama de caja y función de densidad de la variable *Leucina*.

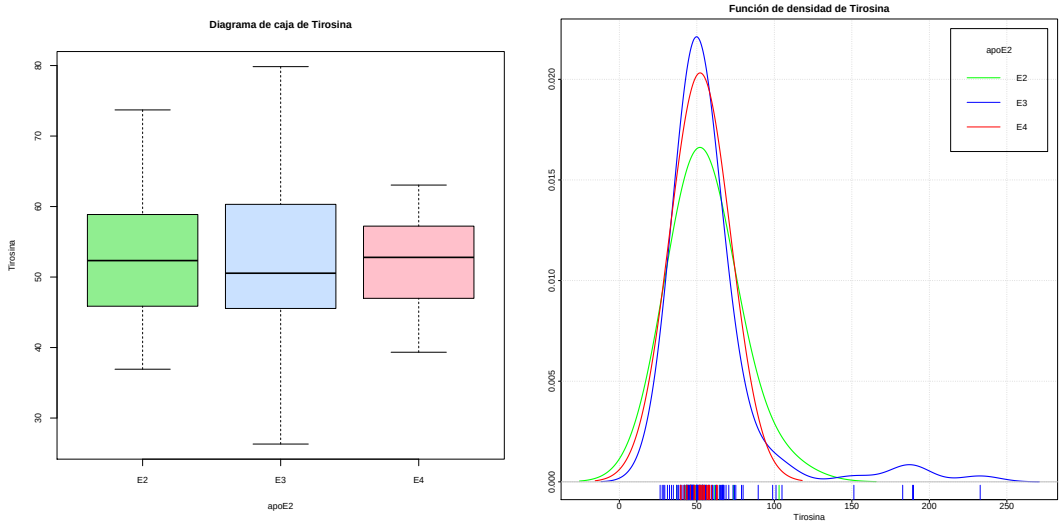

Figura B.65: Diagrama de caja y función de densidad de la variable *Tirosina*.

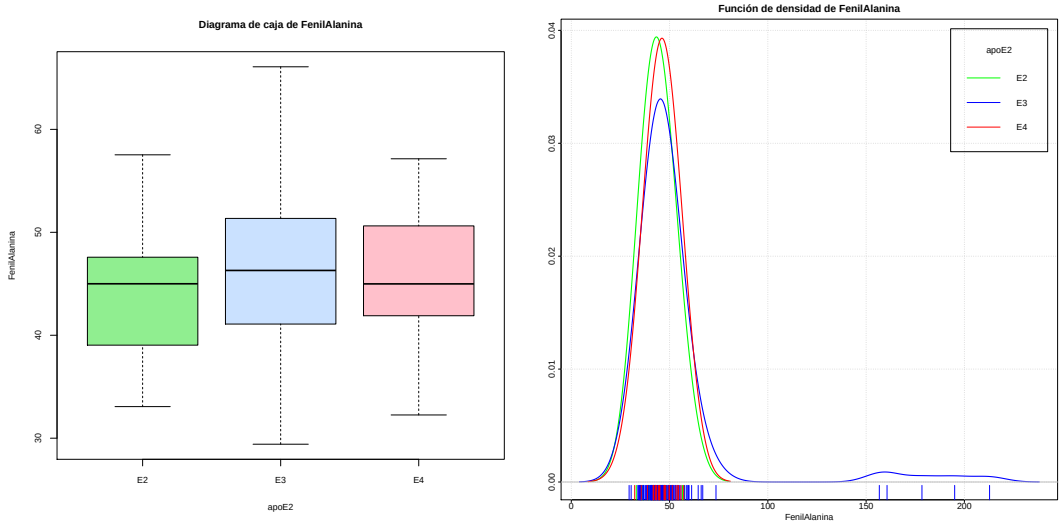

Figura B.66: Diagrama de caja y función de densidad de la variable *FenilAlanina*.

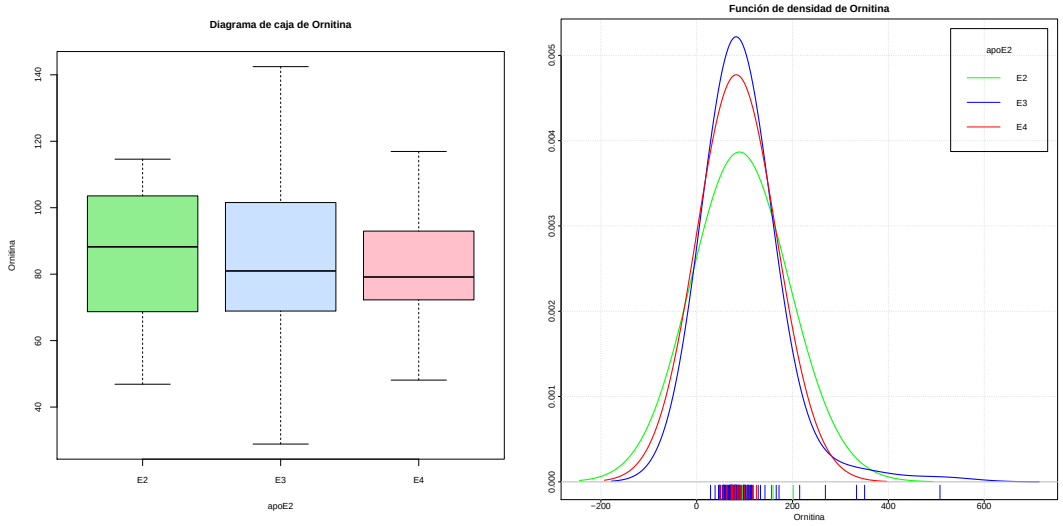

Figura B.67: Diagrama de caja y función de densidad de la variable *Ornitina*.

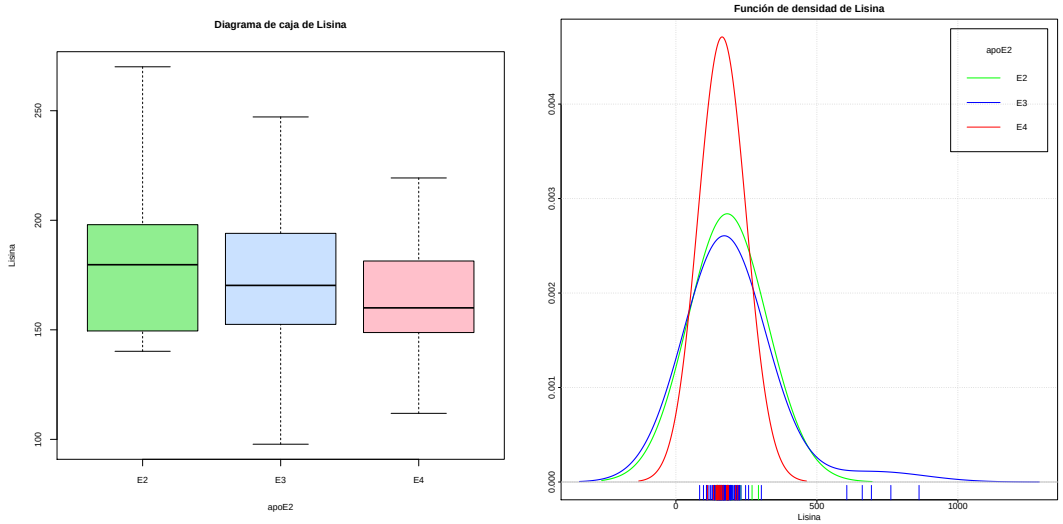

Figura B.68: Diagrama de caja y función de densidad de la variable *Lisina*.

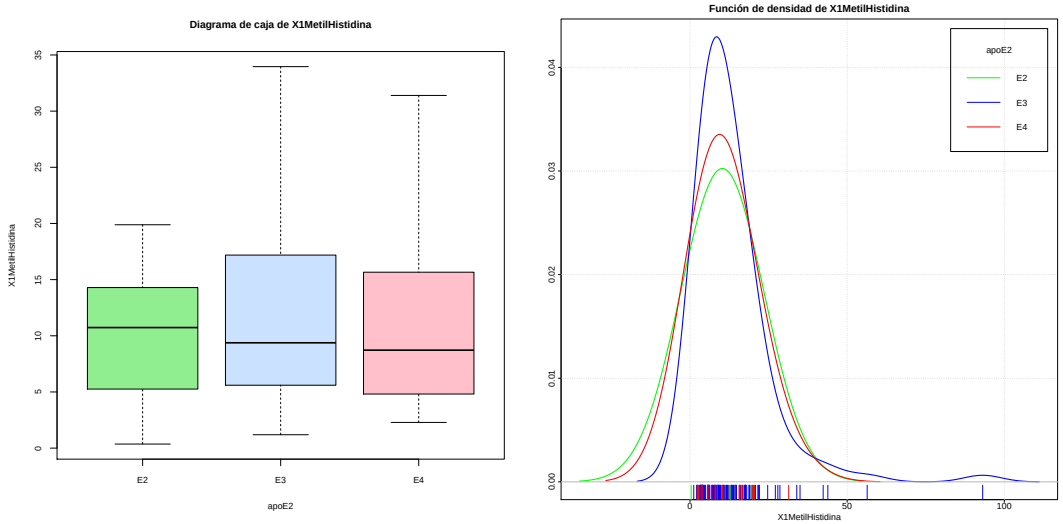

Figura B.69: Diagrama de caja y función de densidad de la variable $X1MetilHistidina$.

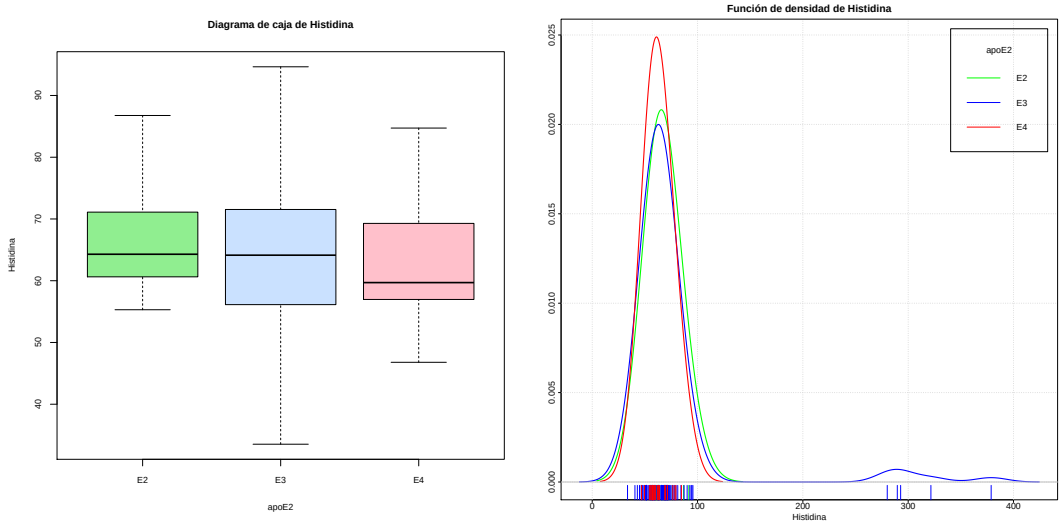

Figura B.70: Diagrama de caja y función de densidad de la variable *Histidina*.

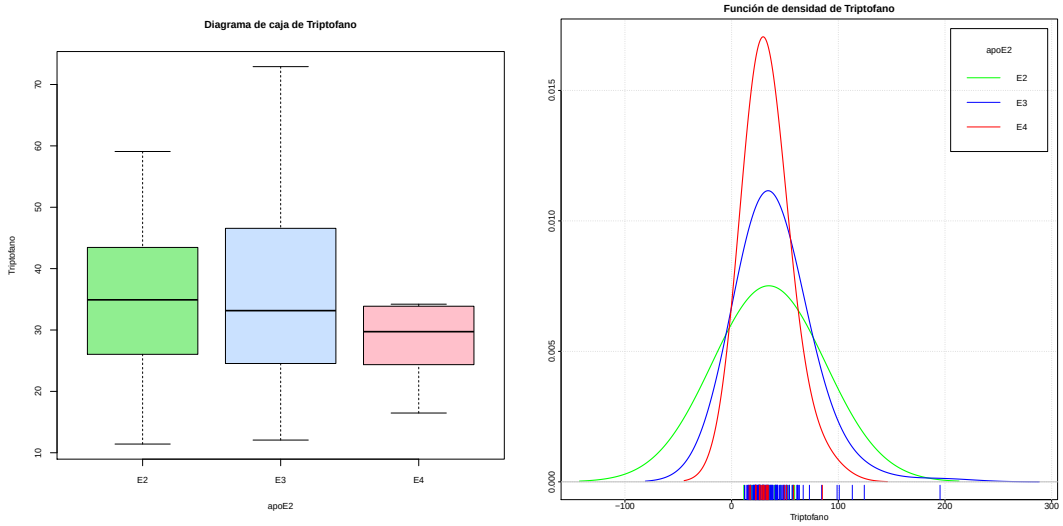

Figura B.71: Diagrama de caja y función de densidad de la variable *Triptofano*.

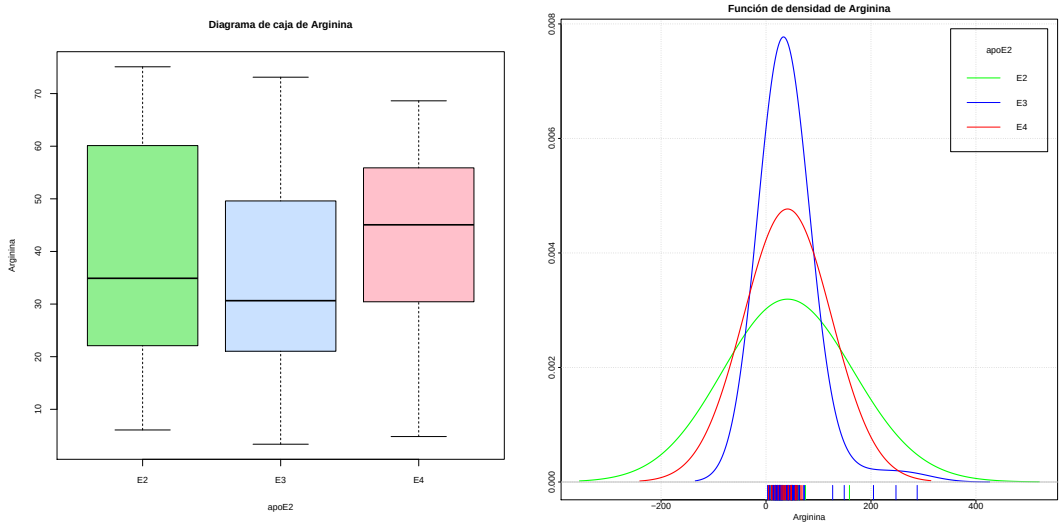

Figura B.72: Diagrama de caja y función de densidad de la variable *Arginina*.

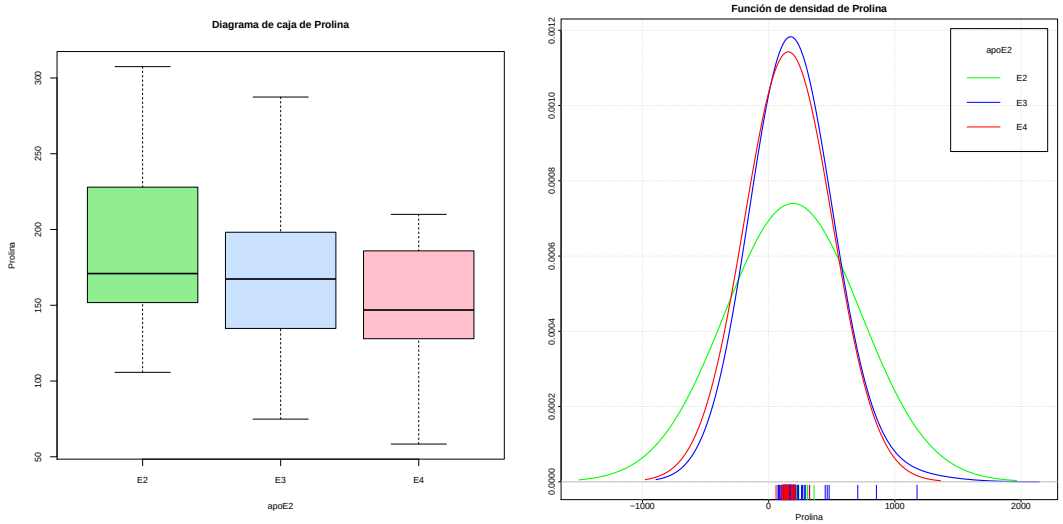

Figura B.73: Diagrama de caja y función de densidad de la variable *Prolina*.

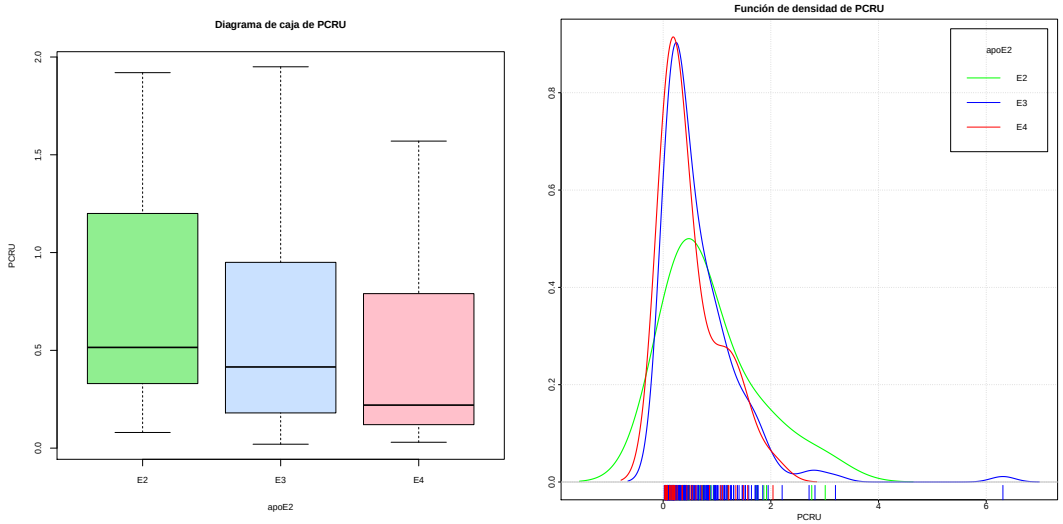

Figura B.74: Diagrama de caja y función de densidad de la variable *PCRU*.

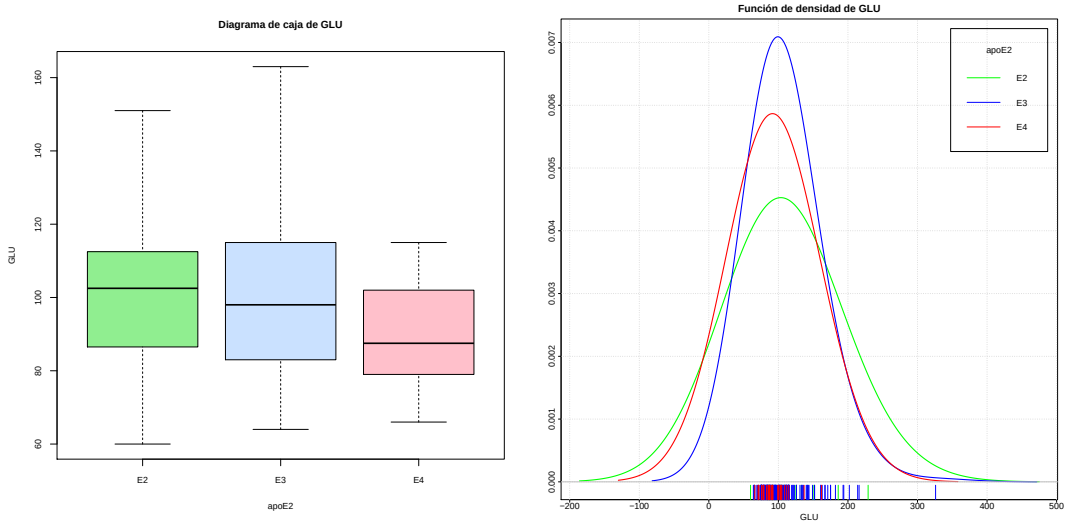

Figura B.75: Diagrama de caja y función de densidad de la variable *GLU*.

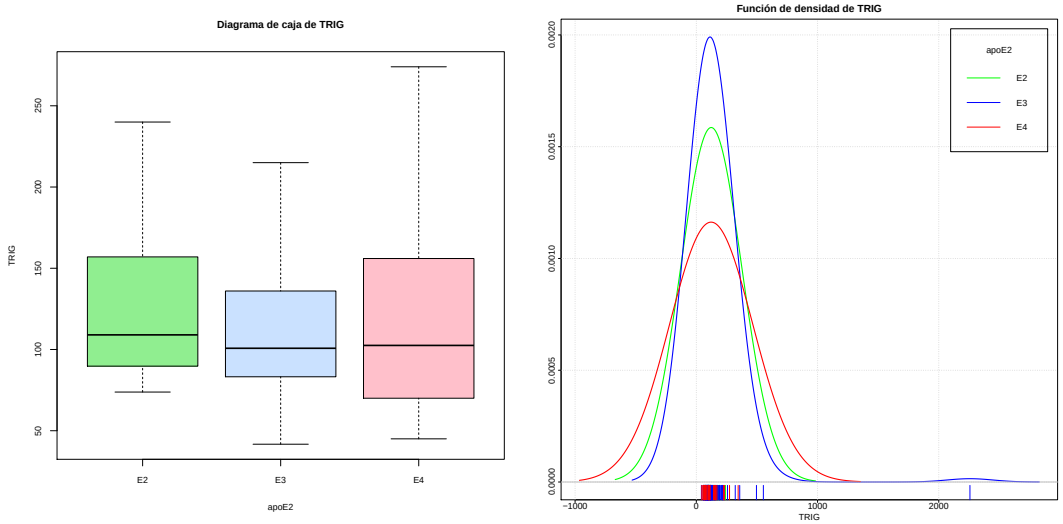

Figura B.76: Diagrama de caja y función de densidad de la variable *TRIG*.

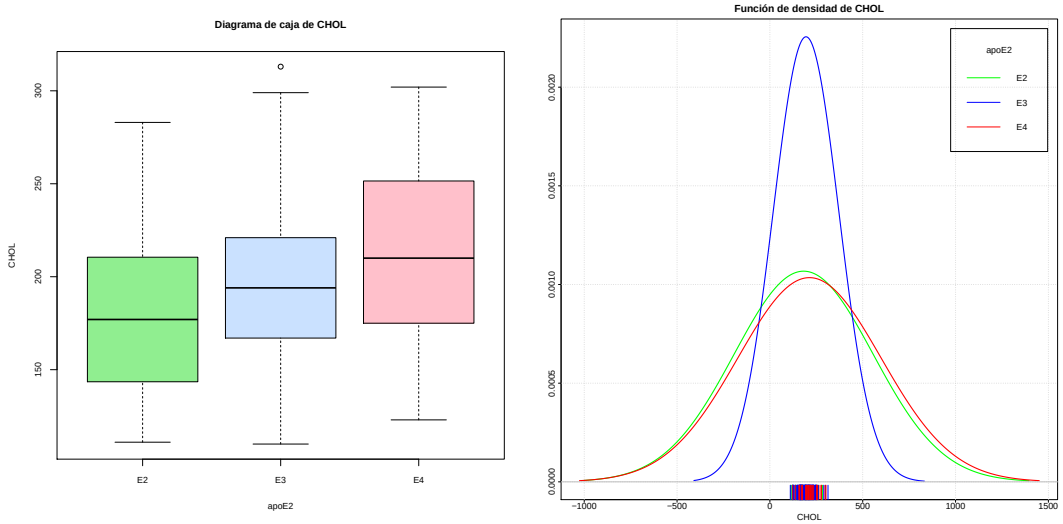

Figura B.77: Diagrama de caja y función de densidad de la variable *CHOL*.

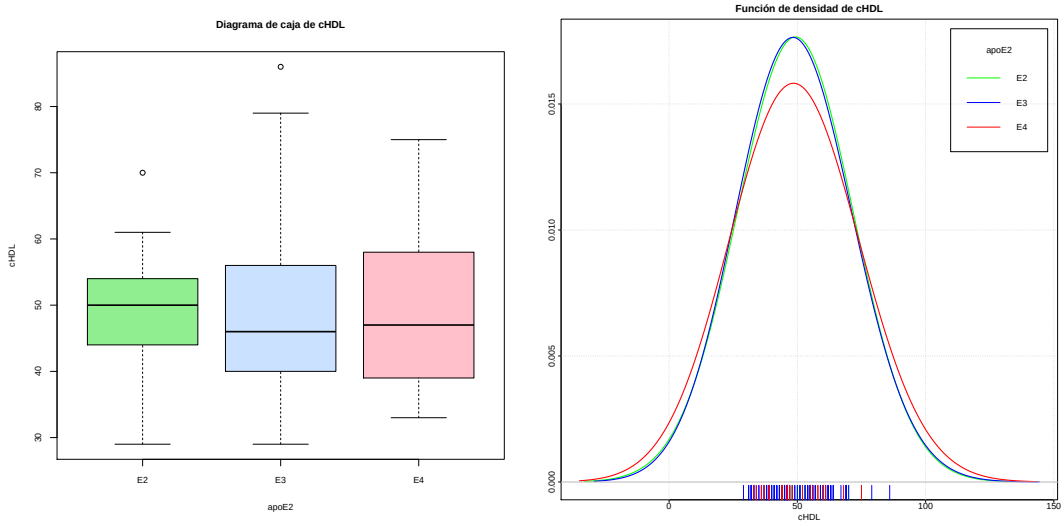

Figura B.78: Diagrama de caja y función de densidad de la variable *cHDL*.

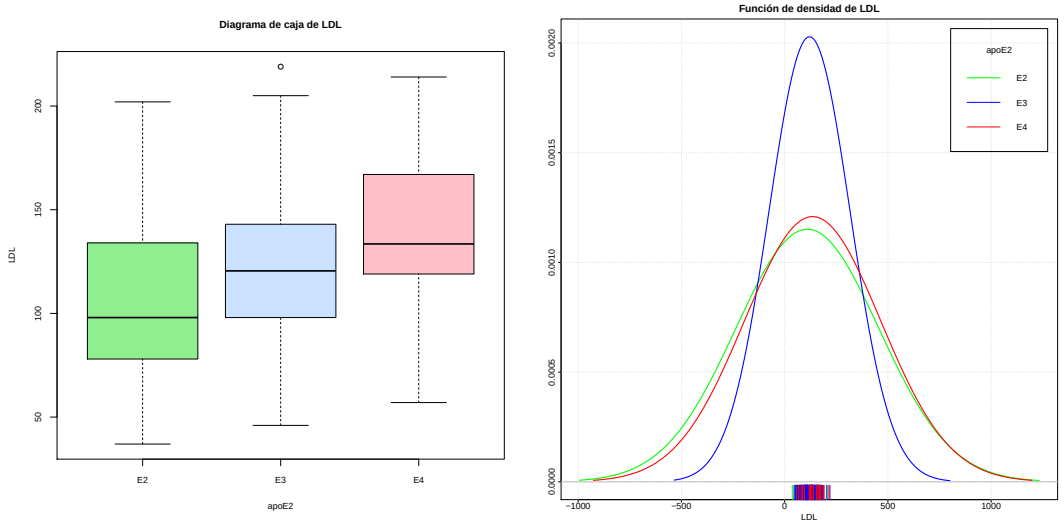

Figura B.79: Diagrama de caja y función de densidad de la variable *LDL*.

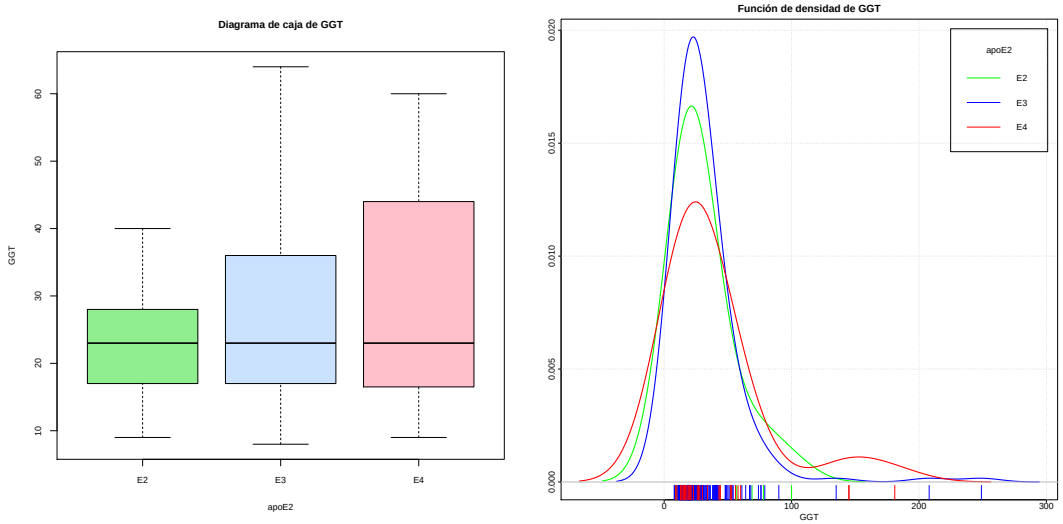

Figura B.80: Diagrama de caja y función de densidad de la variable *GGT*.

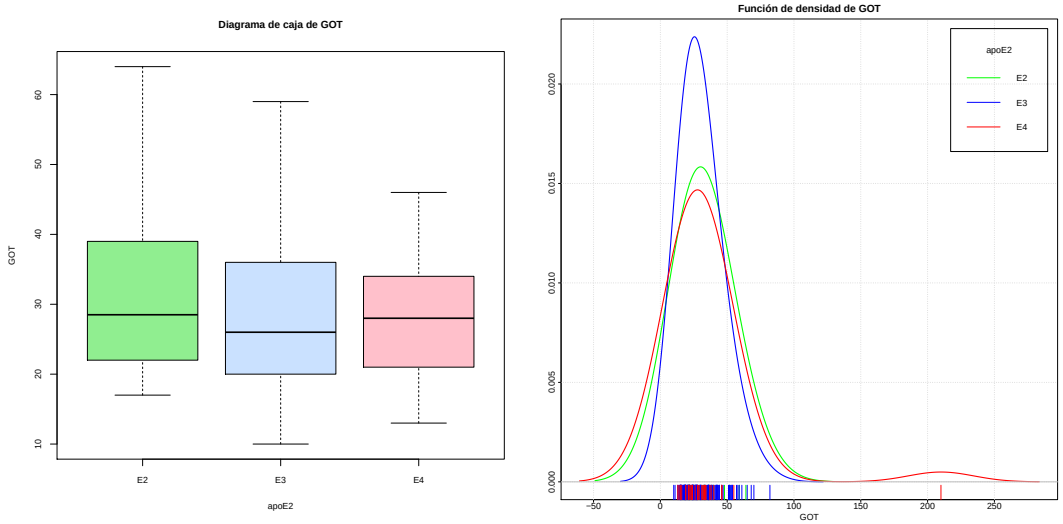

Figura B.81: Diagrama de caja y función de densidad de la variable *GOT*.

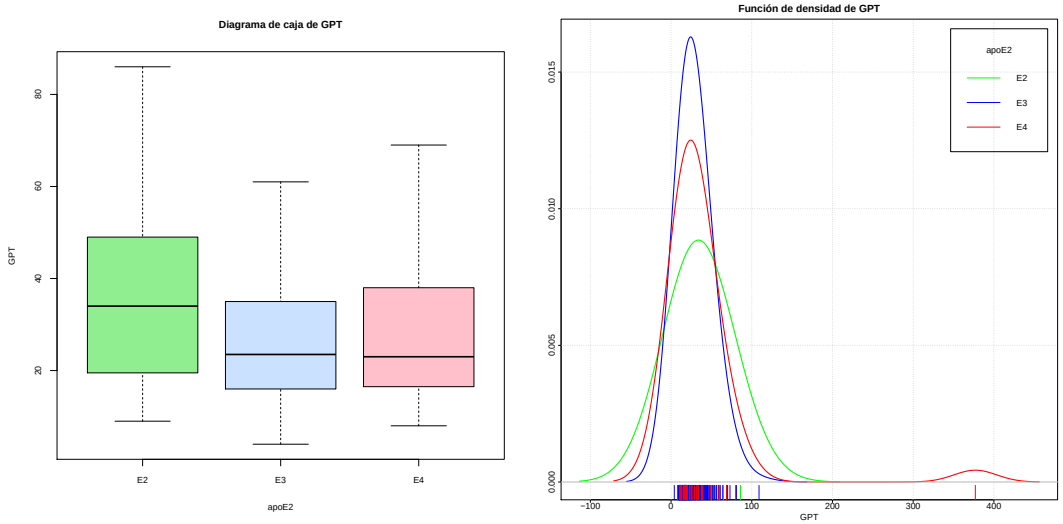

Figura B.82: Diagrama de caja y función de densidad de la variable *GPT*.

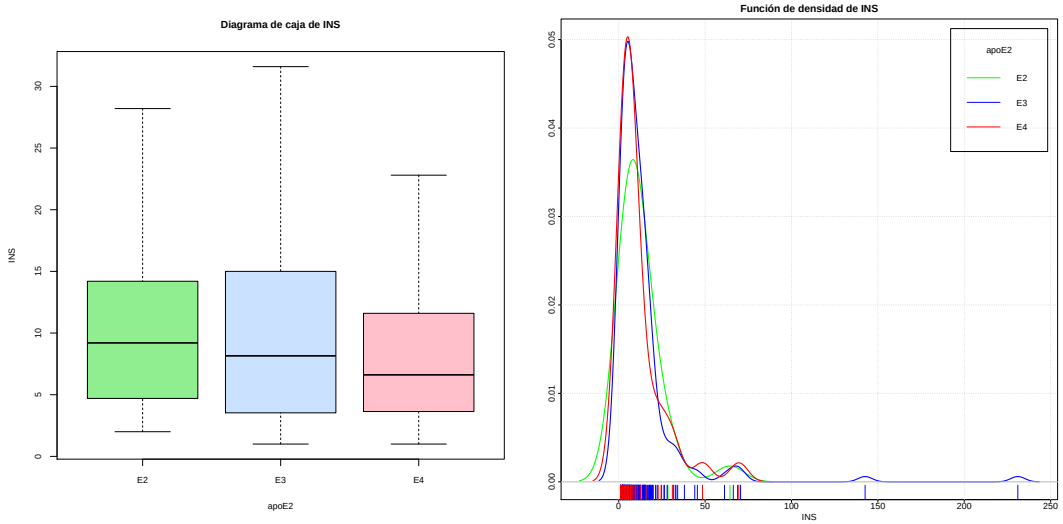

Figura B.83: Diagrama de caja y función de densidad de la variable *INS*.

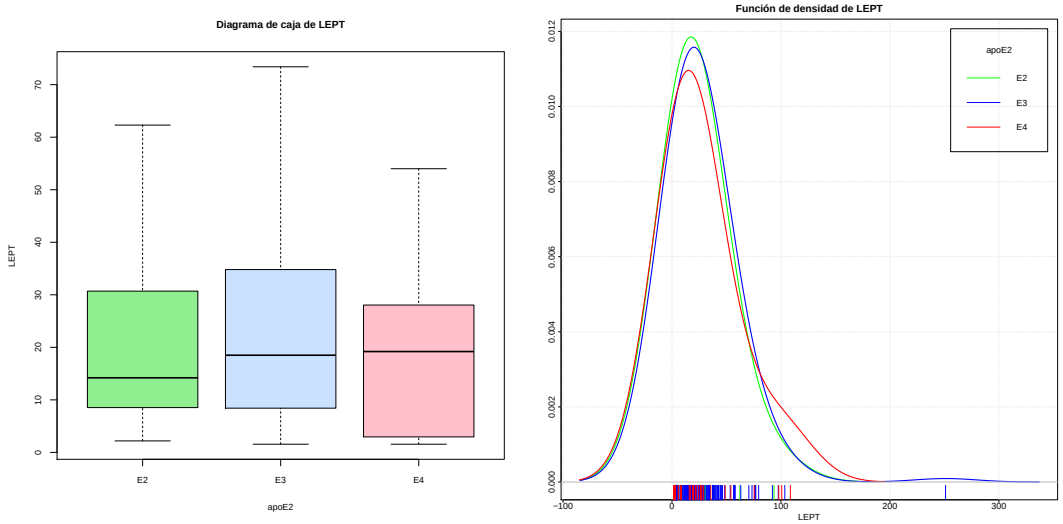

Figura B.84: Diagrama de caja y función de densidad de la variable *LEPT*.

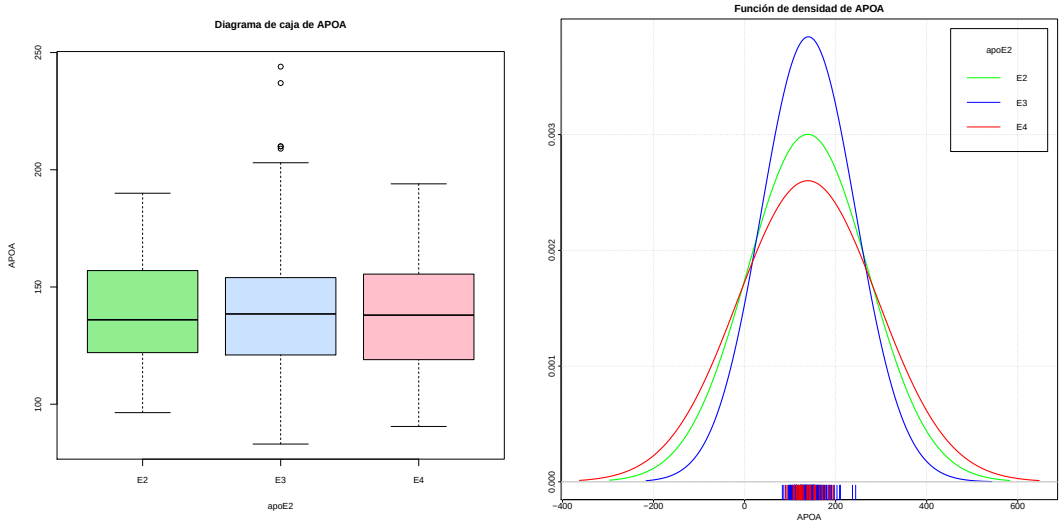

Figura B.85: Diagrama de caja y función de densidad de la variable *APOA*.

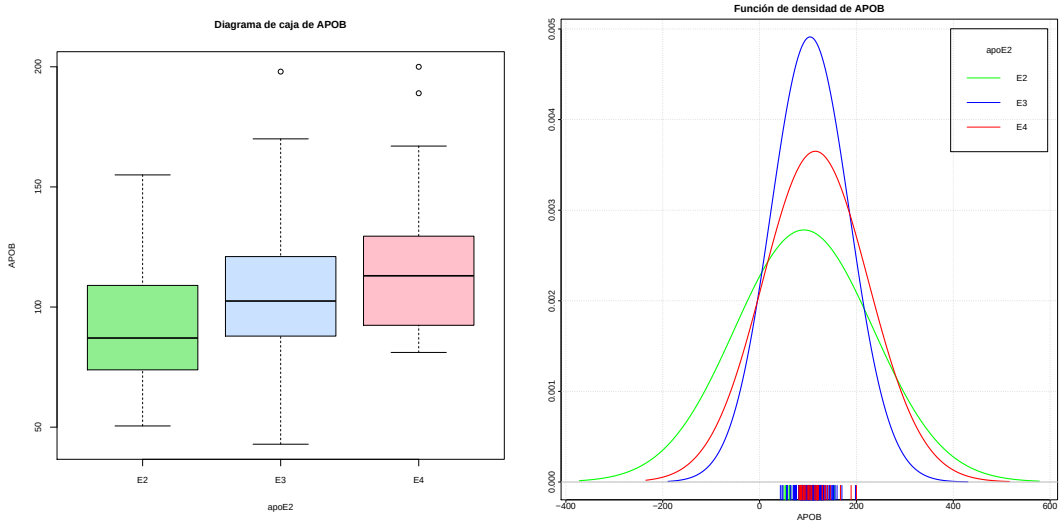

Figura B.86: Diagrama de caja y función de densidad de la variable *APOB*.

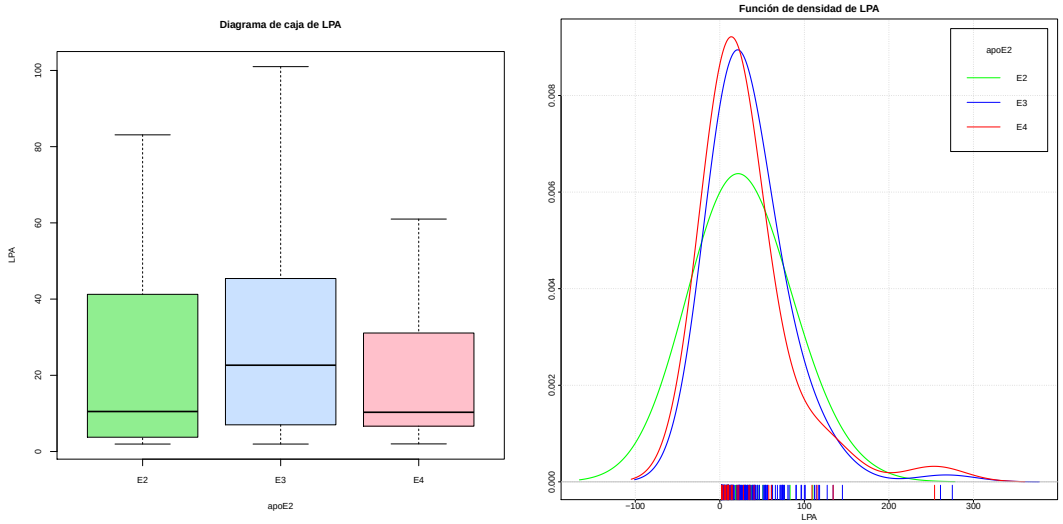

Figura B.87: Diagrama de caja y función de densidad de la variable *LPA*.

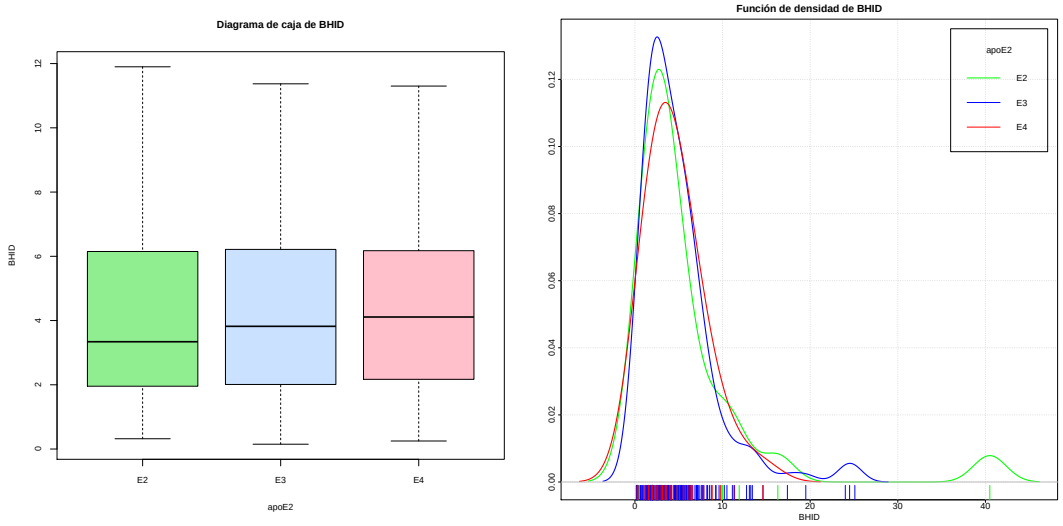

Figura B.88: Diagrama de caja y función de densidad de la variable *BHID*.

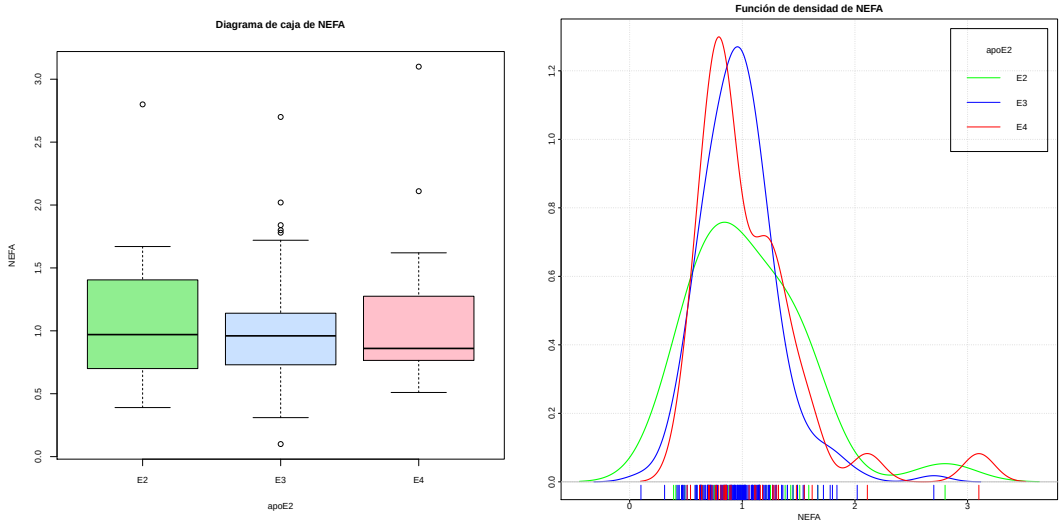

Figura B.89: Diagrama de caja y función de densidad de la variable *NEFA*.

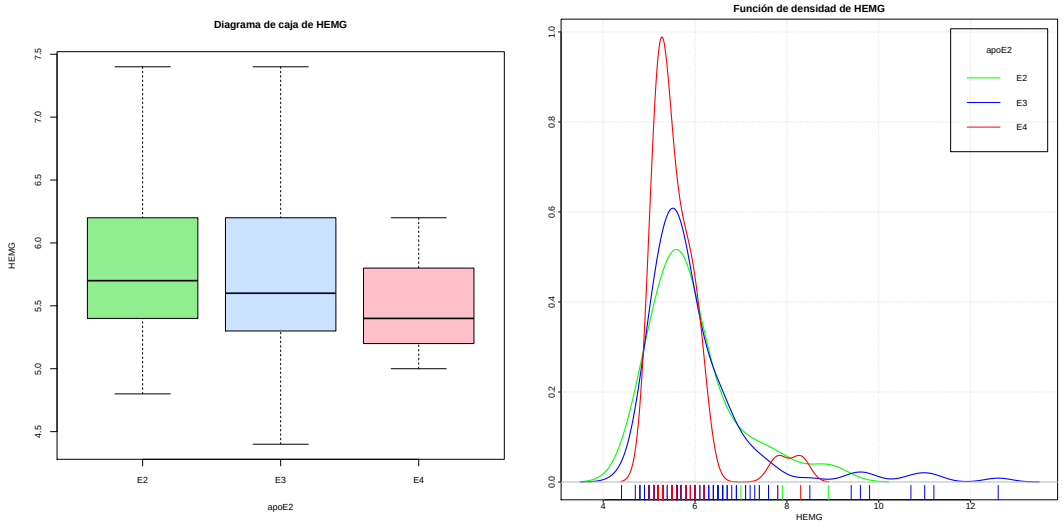

Figura B.90: Diagrama de caja y función de densidad de la variable *HEMG*.

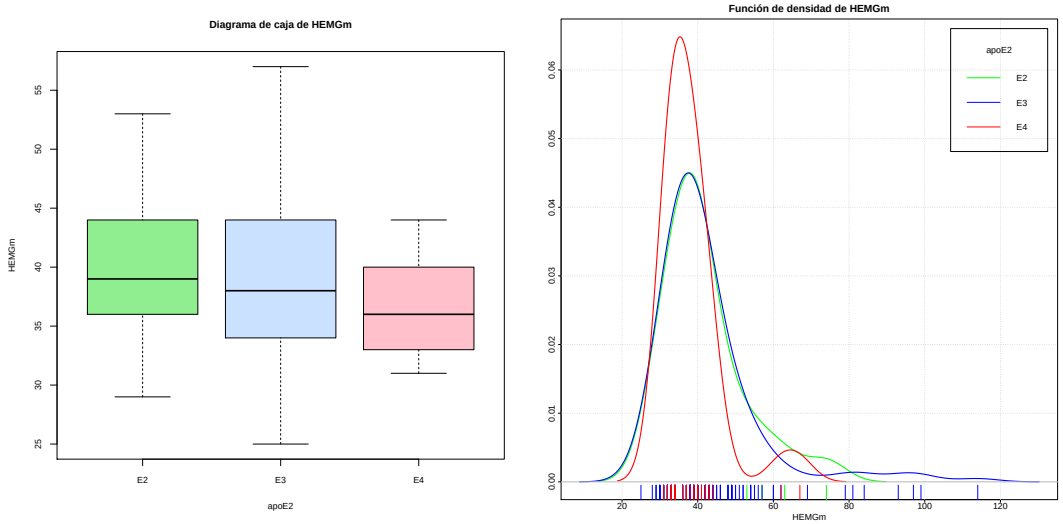

Figura B.91: Diagrama de caja y función de densidad de la variable *HEMGm*.

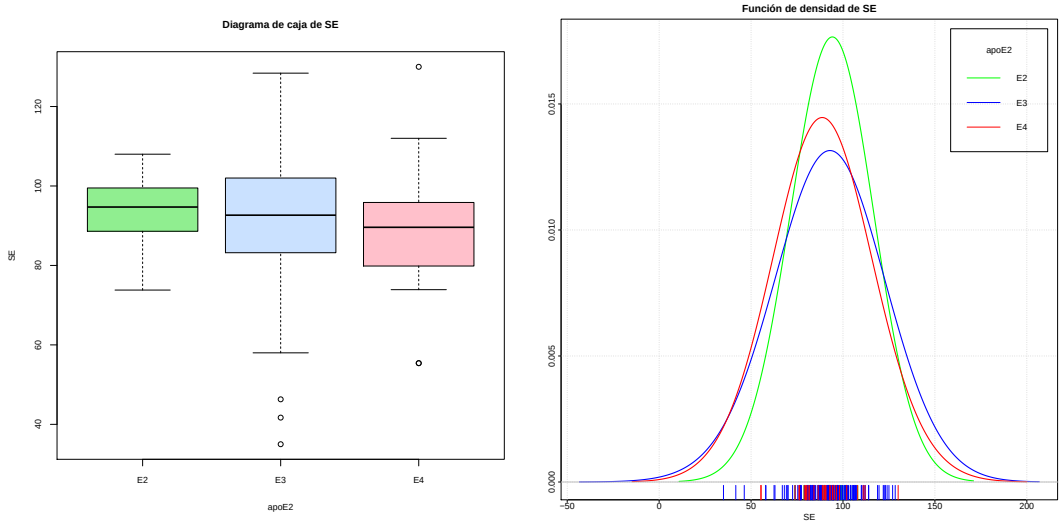

Figura B.92: Diagrama de caja y función de densidad de la variable *SE*.

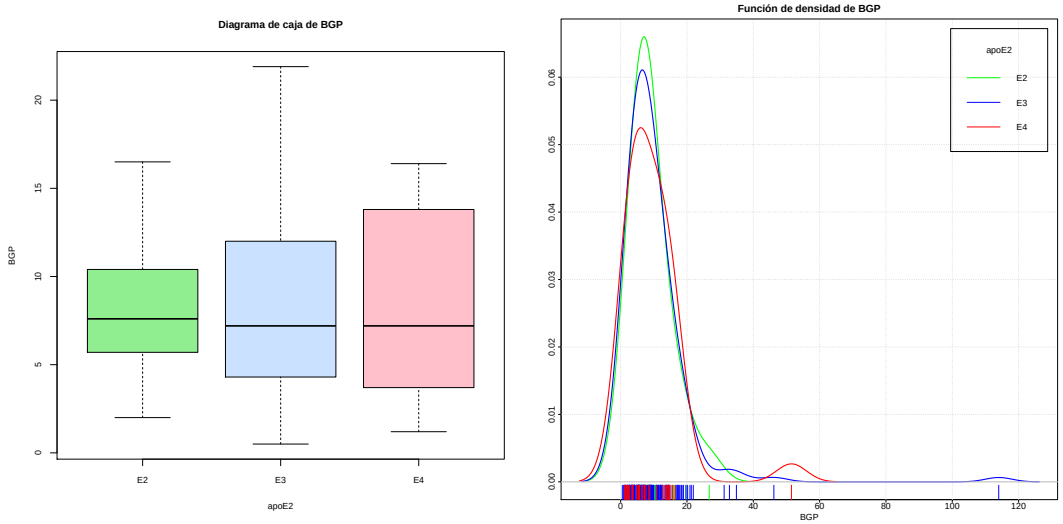

Figura B.93: Diagrama de caja y función de densidad de la variable *BGP*.

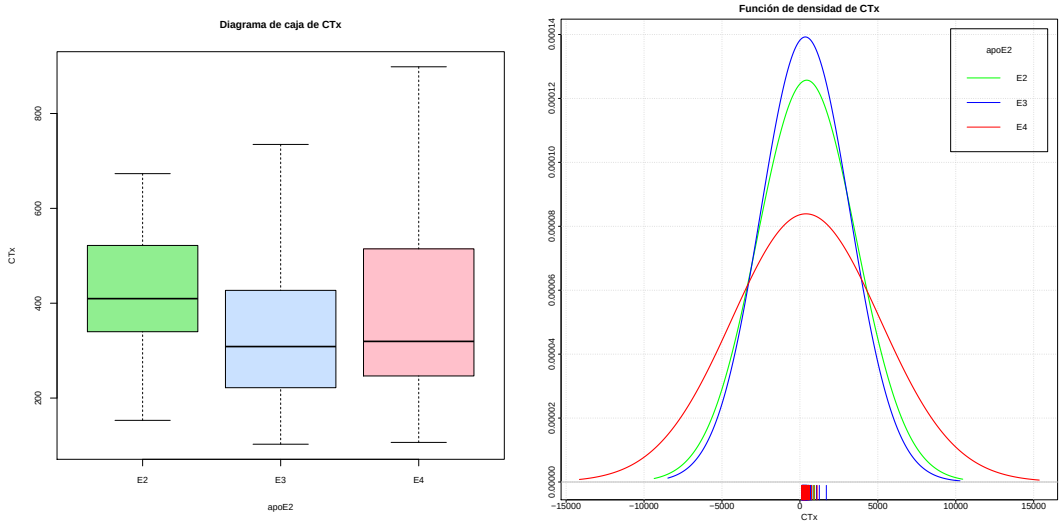

Figura B.94: Diagrama de caja y función de densidad de la variable CTx.

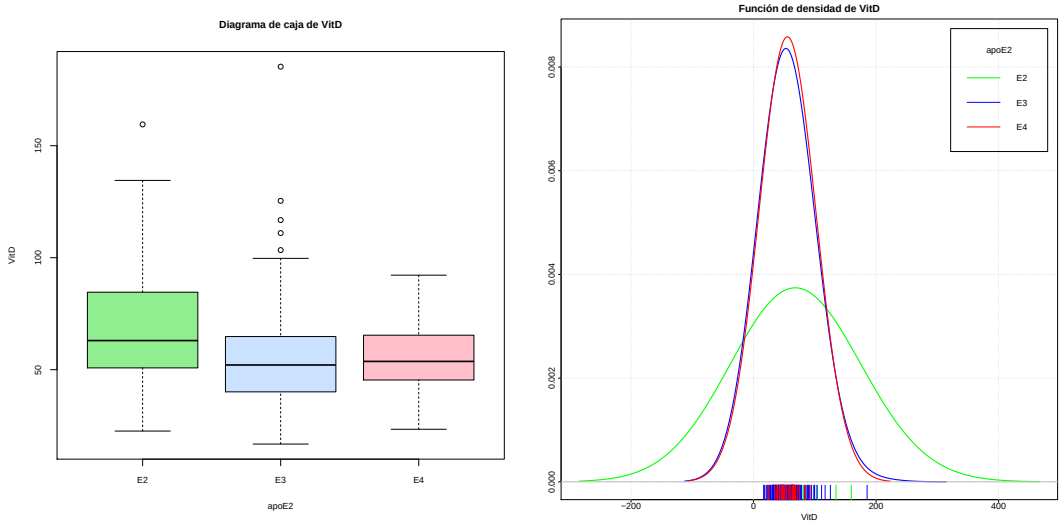

Figura B.95: Diagrama de caja y función de densidad de la variable $VitD$.

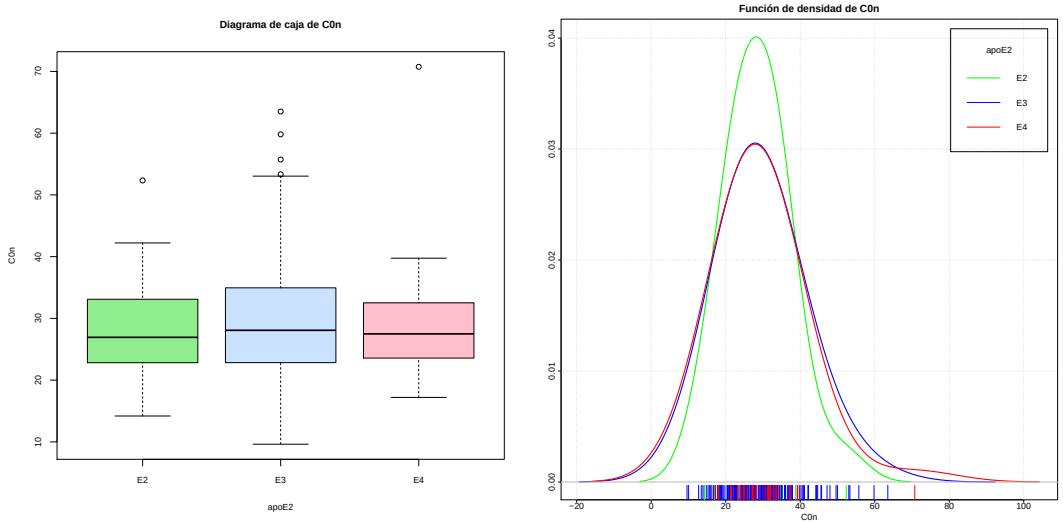

Figura B.96: Diagrama de caja y función de densidad de la variable *C*0*n*.

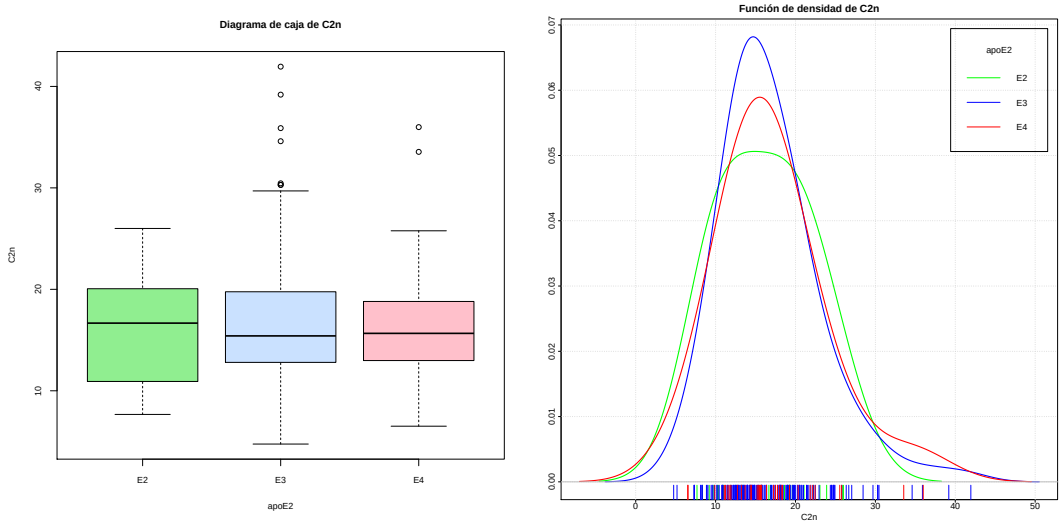

Figura B.97: Diagrama de caja y función de densidad de la variable *C*2*n*.

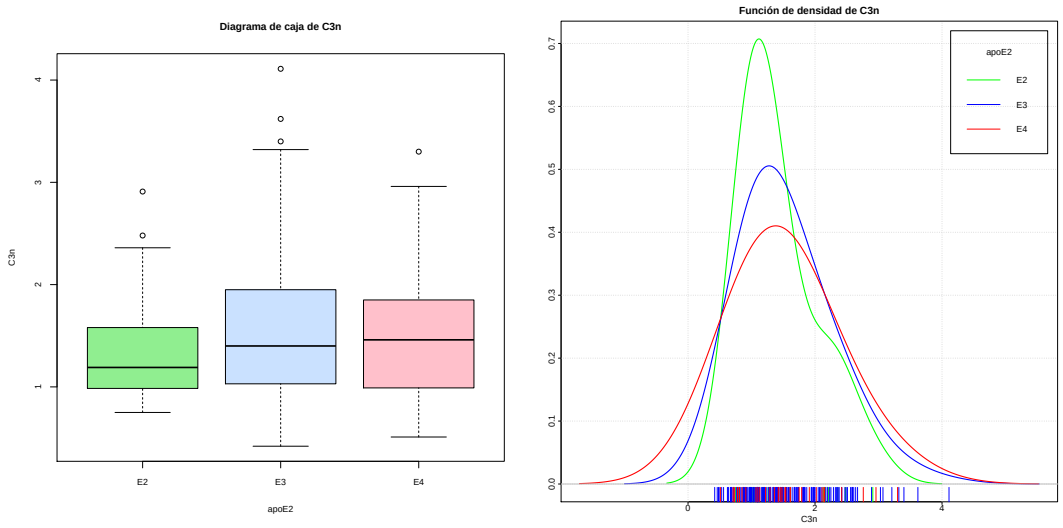

Figura B.98: Diagrama de caja y función de densidad de la variable *C*3*n*.

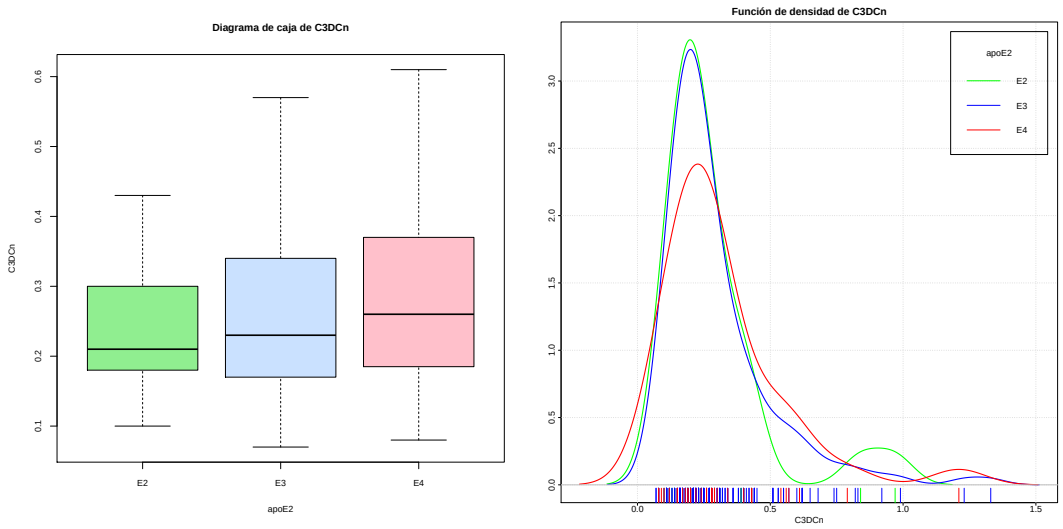

Figura B.99: Diagrama de caja y función de densidad de la variable *C*3*DCn*.

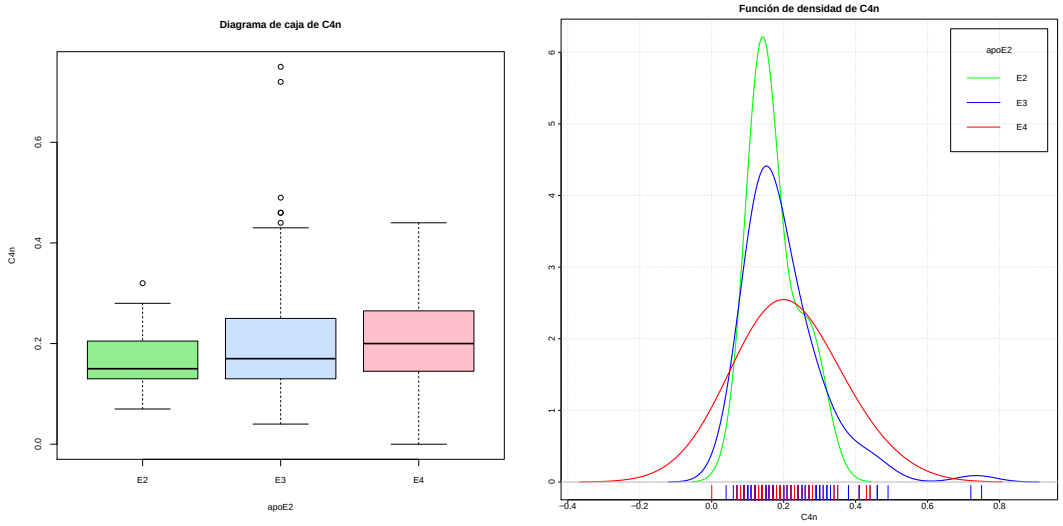

Figura B.100: Diagrama de caja y función de densidad de la variable *C4n*.

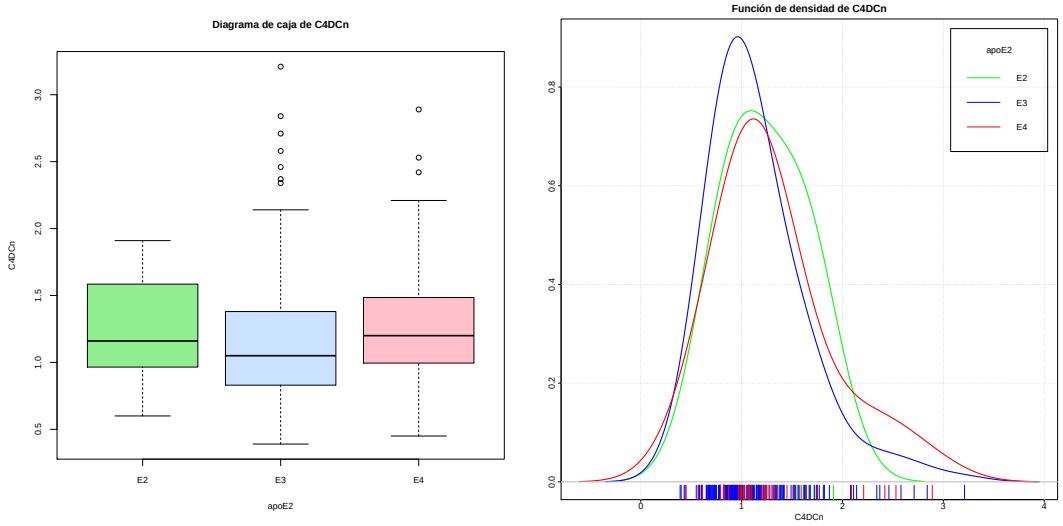

Figura B.101: Diagrama de caja y función de densidad de la variable *C4DCn*.

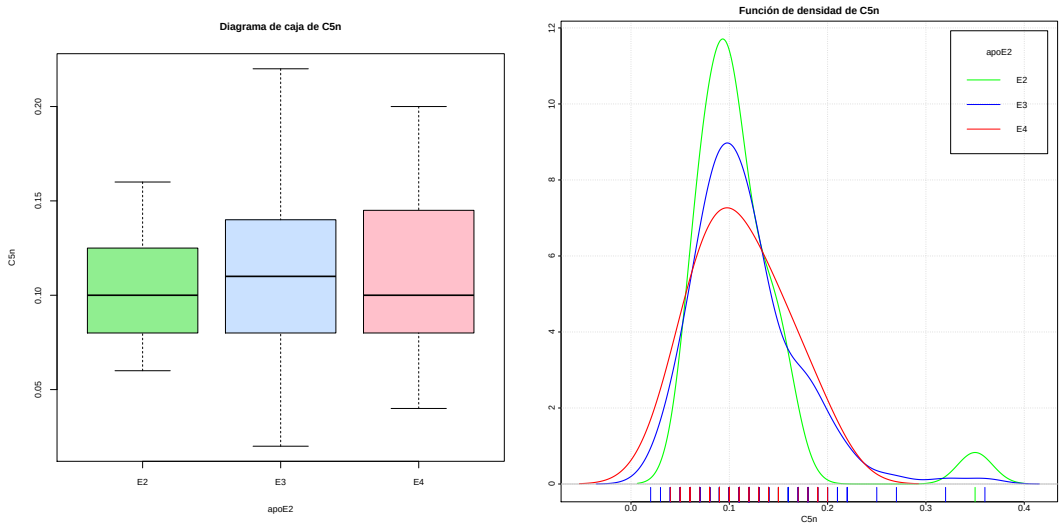

Figura B.102: Diagrama de caja y función de densidad de la variable *C*5*n*.

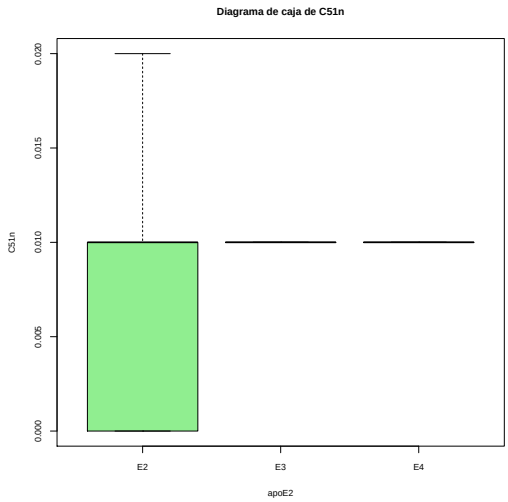

Figura B.103: Diagrama de caja y función de densidad de la variable *C*51*n*.

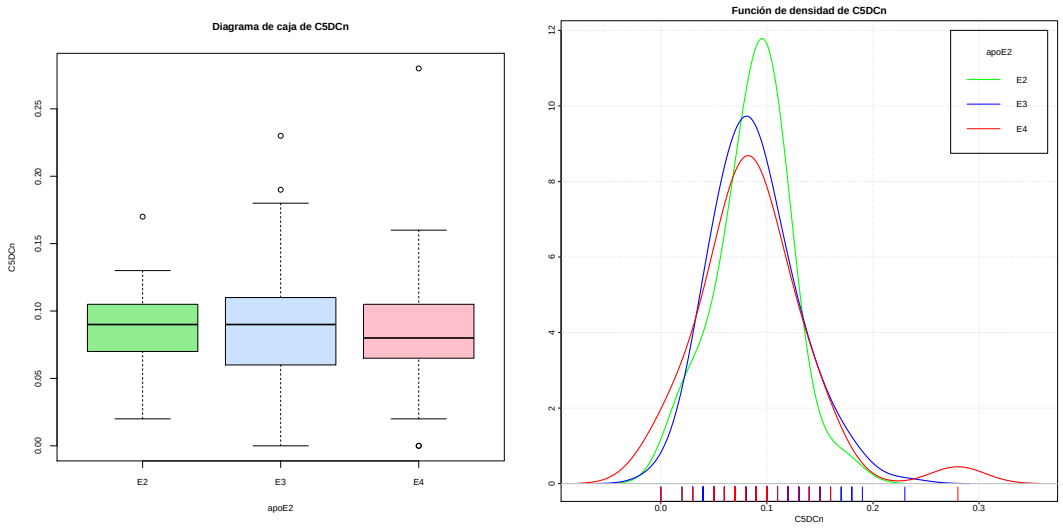

Figura B.104: Diagrama de caja y función de densidad de la variable *C5DCn*.

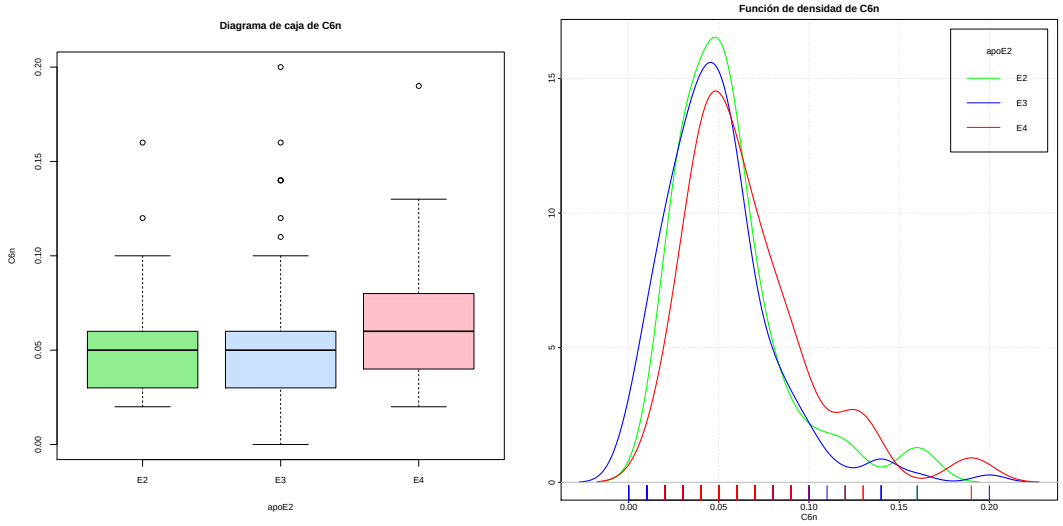

Figura B.105: Diagrama de caja y función de densidad de la variable *C6n*.

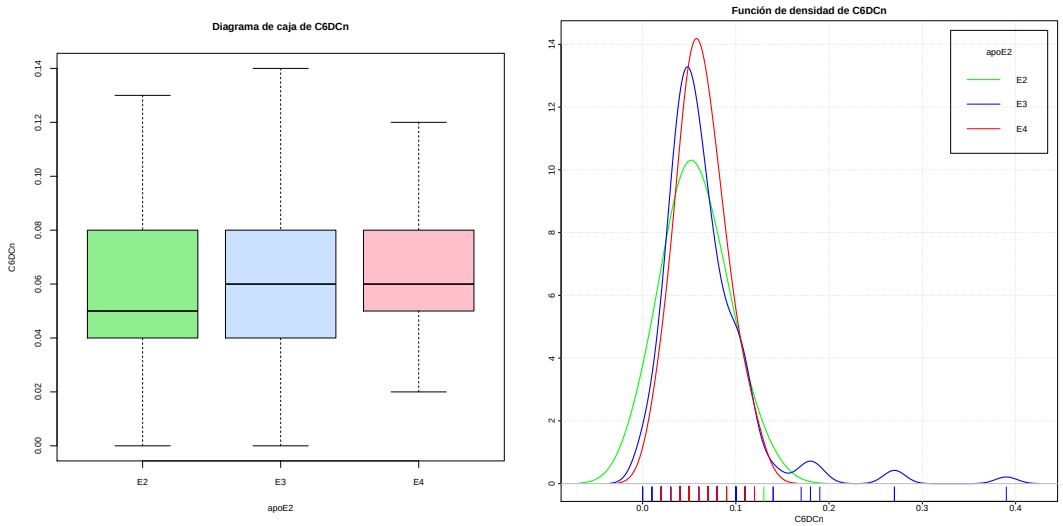

Figura B.106: Diagrama de caja y función de densidad de la variable *C6DCn*.

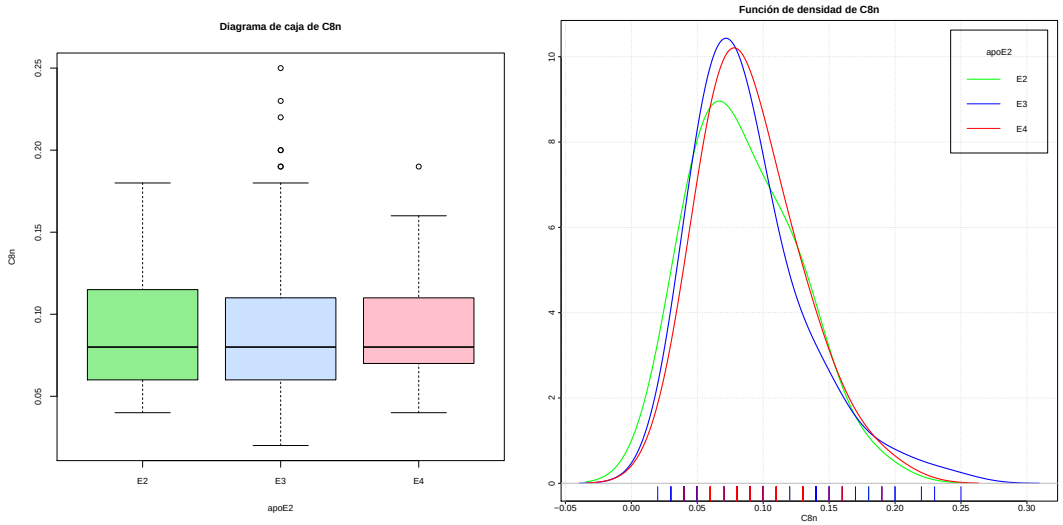

Figura B.107: Diagrama de caja y función de densidad de la variable *C8n*.

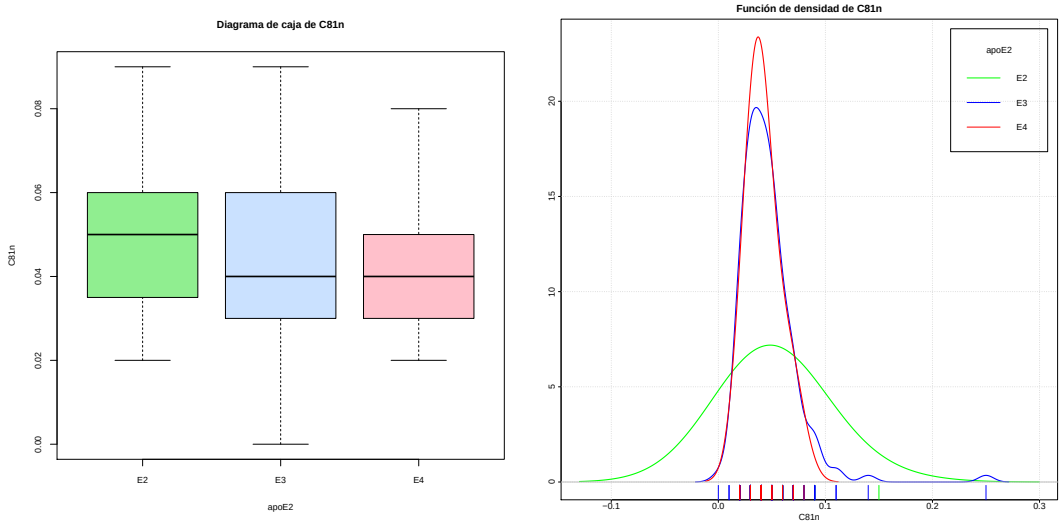

Figura B.108: Diagrama de caja y función de densidad de la variable $C81n$.

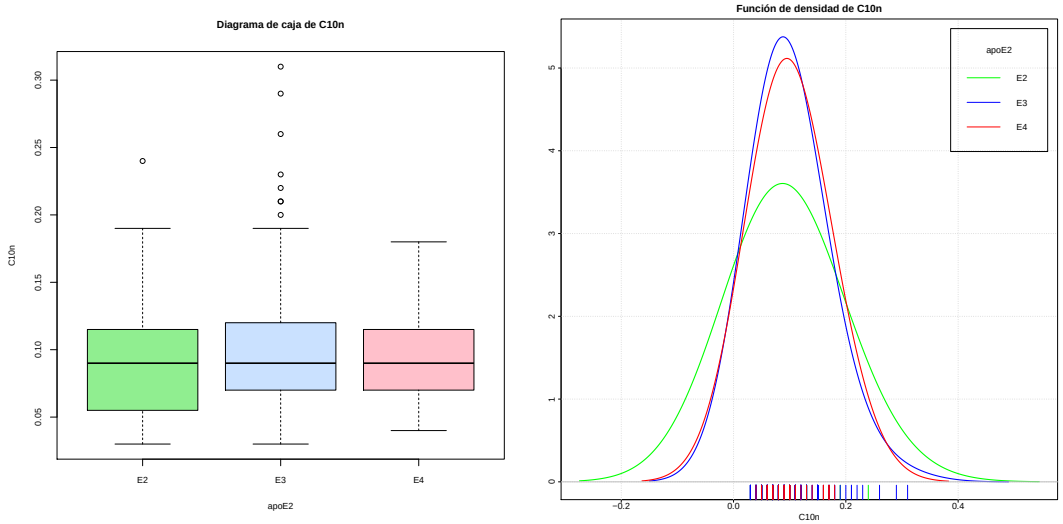

Figura B.109: Diagrama de caja y función de densidad de la variable $C10n$.

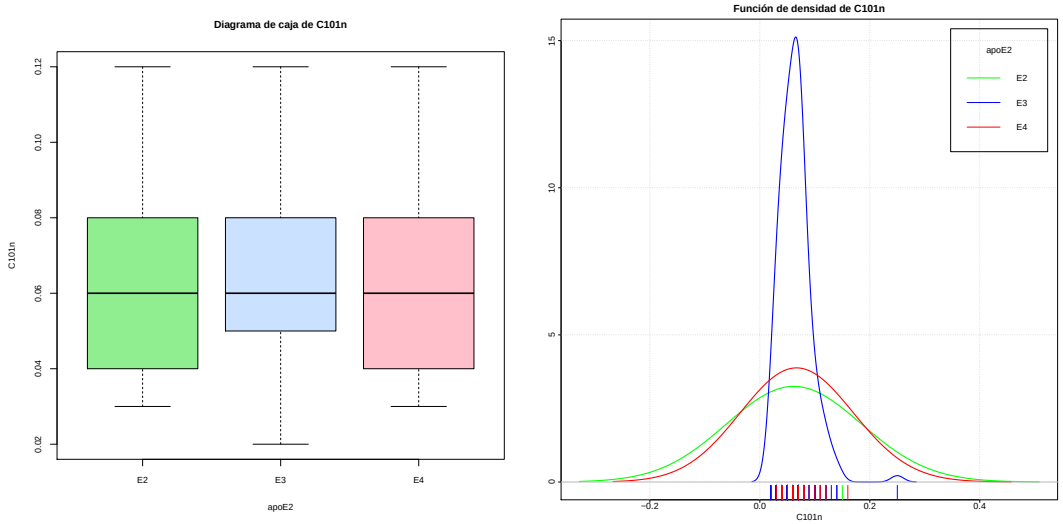

Figura B.110: Diagrama de caja y función de densidad de la variable $C101n$.

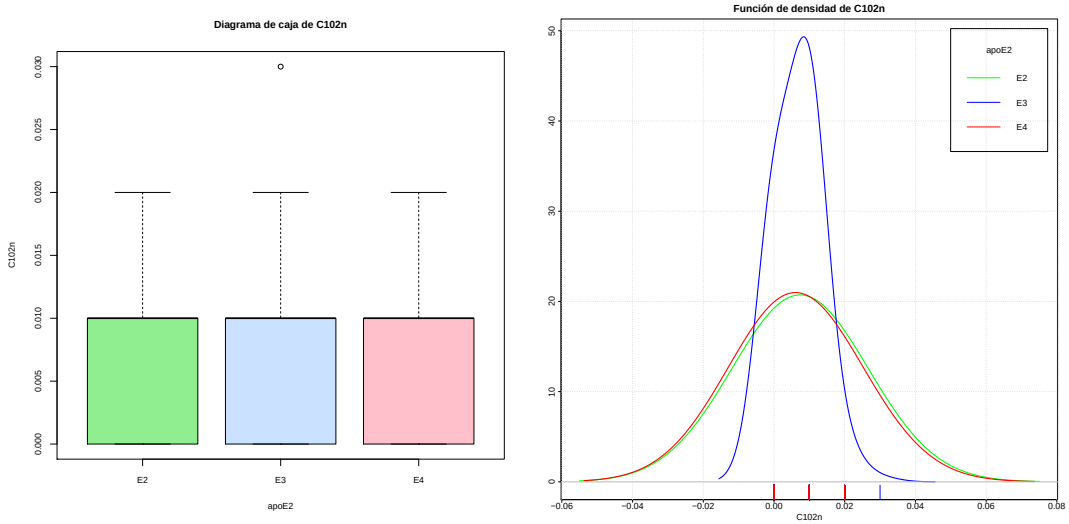

Figura B.111: Diagrama de caja y función de densidad de la variable $C102n$.

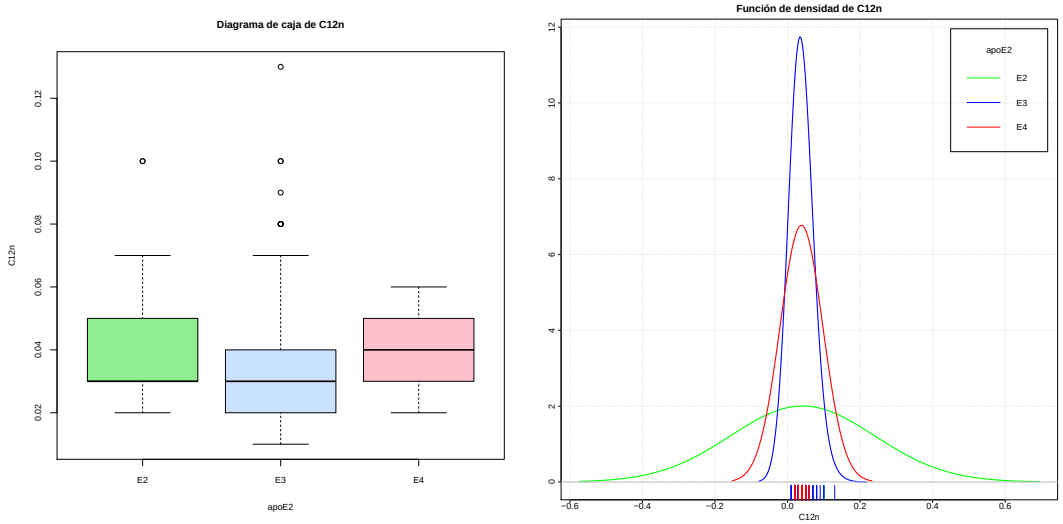

Figura B.112: Diagrama de caja y función de densidad de la variable $C12n$.

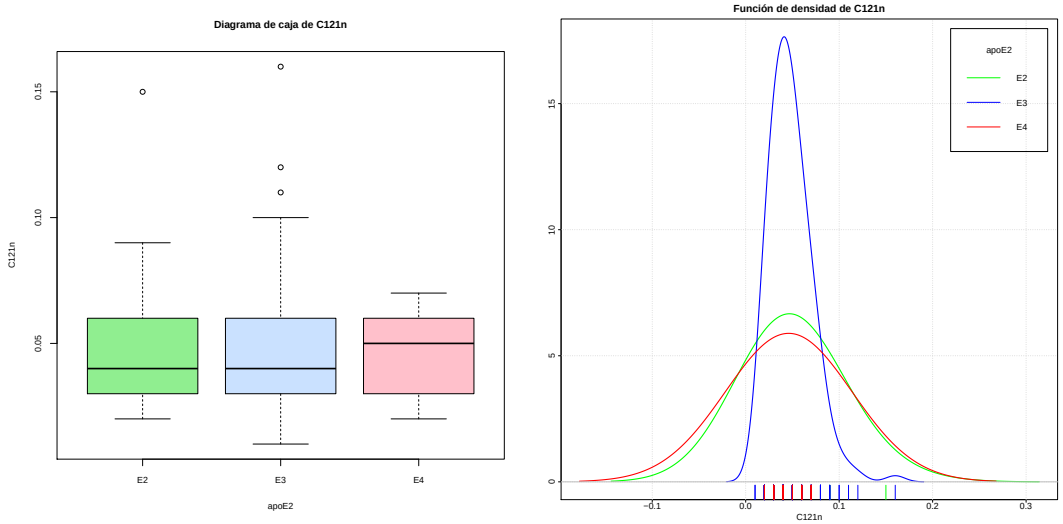

Figura B.113: Diagrama de caja y función de densidad de la variable $C121n$.

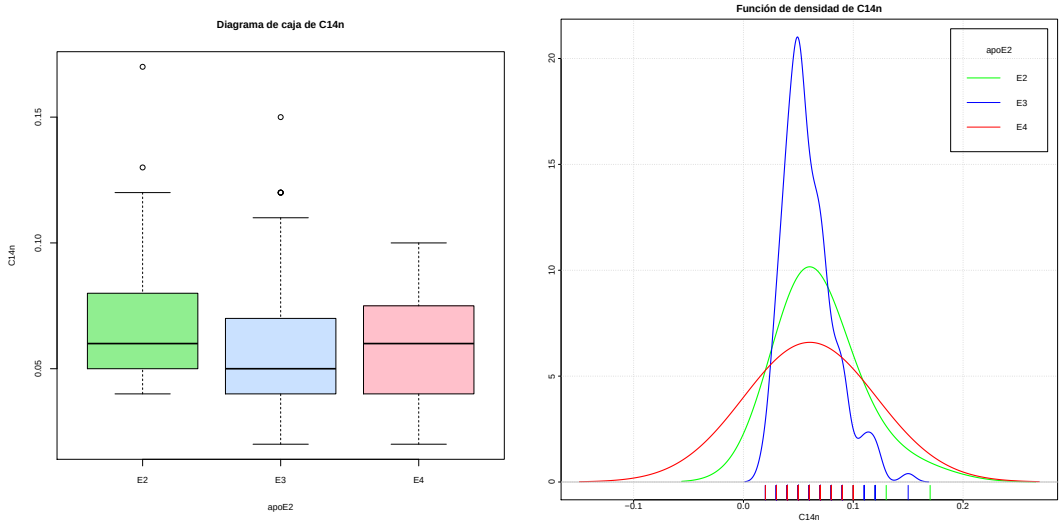

Figura B.114: Diagrama de caja y función de densidad de la variable $C14n$.

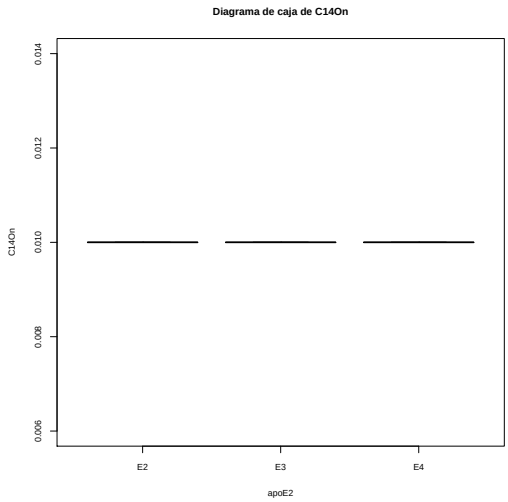

Figura B.115: Diagrama de caja y función de densidad de la variable $C14On$.

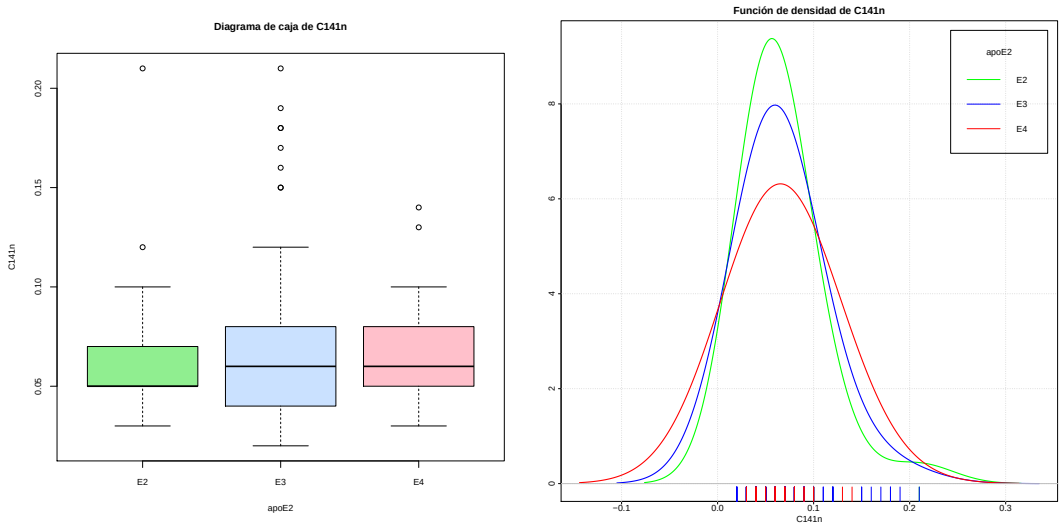

Figura B.116: Diagrama de caja y función de densidad de la variable $C141n$.

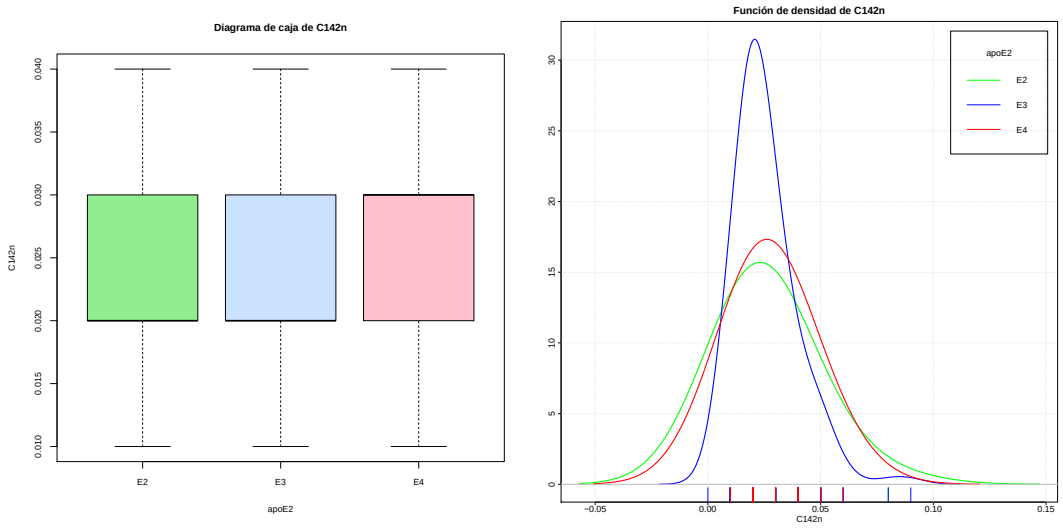

Figura B.117: Diagrama de caja y función de densidad de la variable $C142n$.

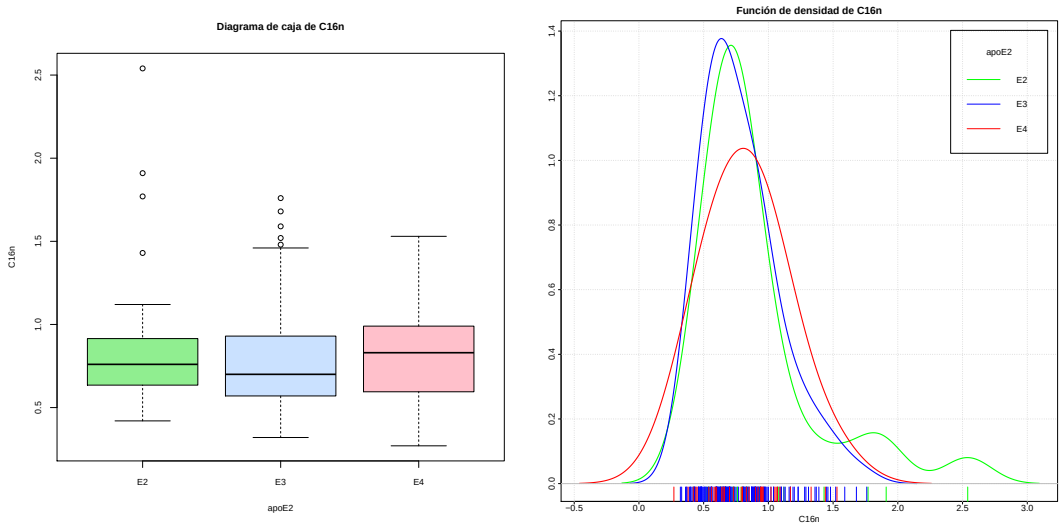

Figura B.118: Diagrama de caja y función de densidad de la variable $C16n$.

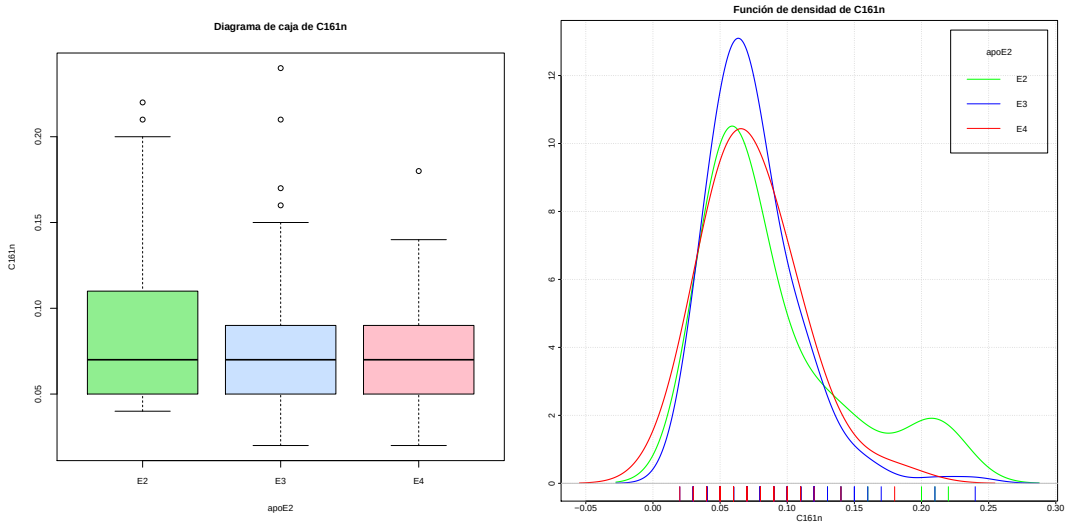

Figura B.119: Diagrama de caja y función de densidad de la variable $C161n$.

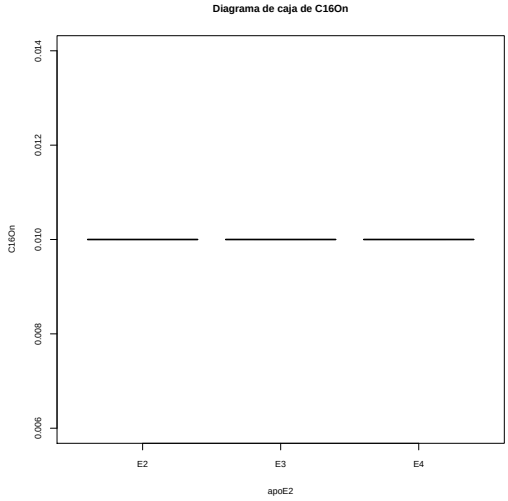

Figura B.120: Diagrama de caja y función de densidad de la variable *C*16*On*.

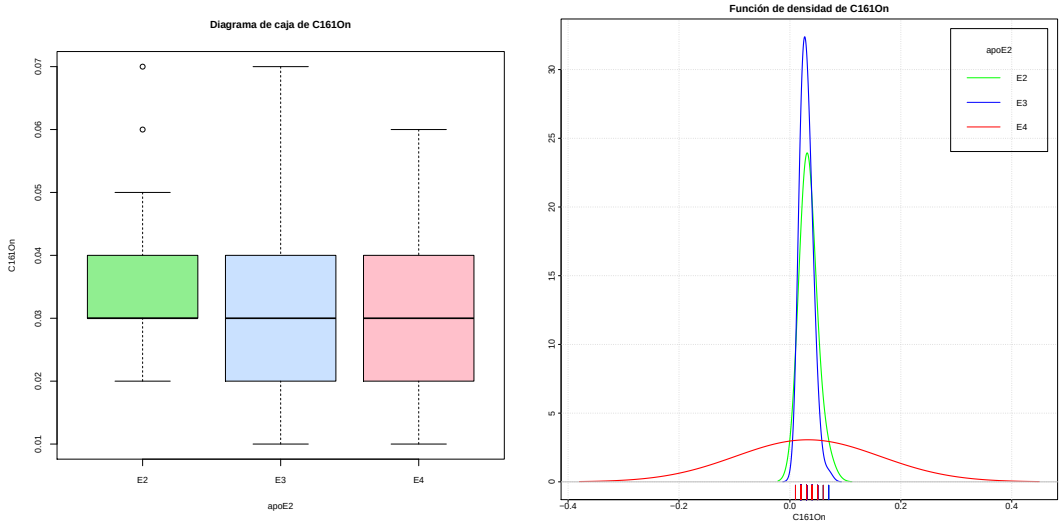

Figura B.121: Diagrama de caja y función de densidad de la variable *C*161*On*.

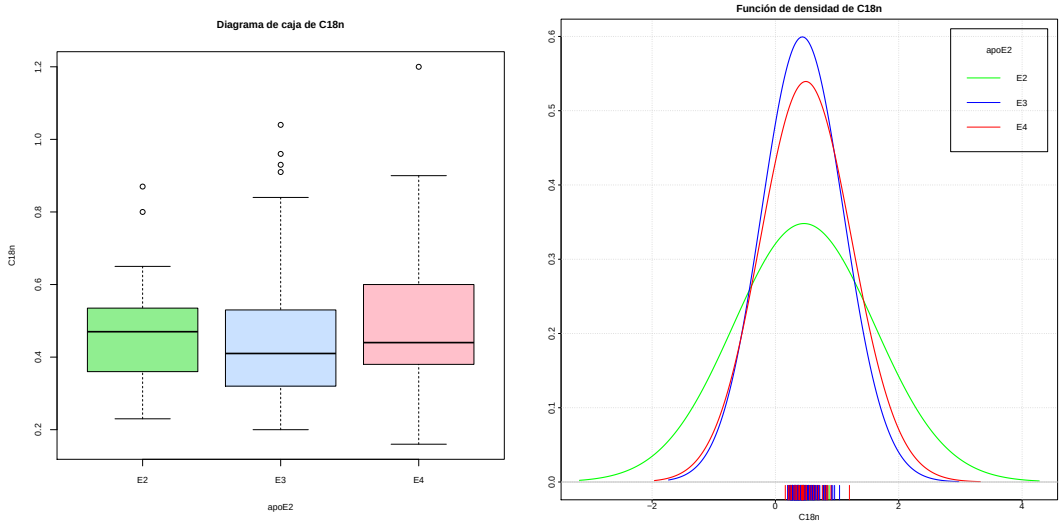

Figura B.122: Diagrama de caja y función de densidad de la variable $C18n$.

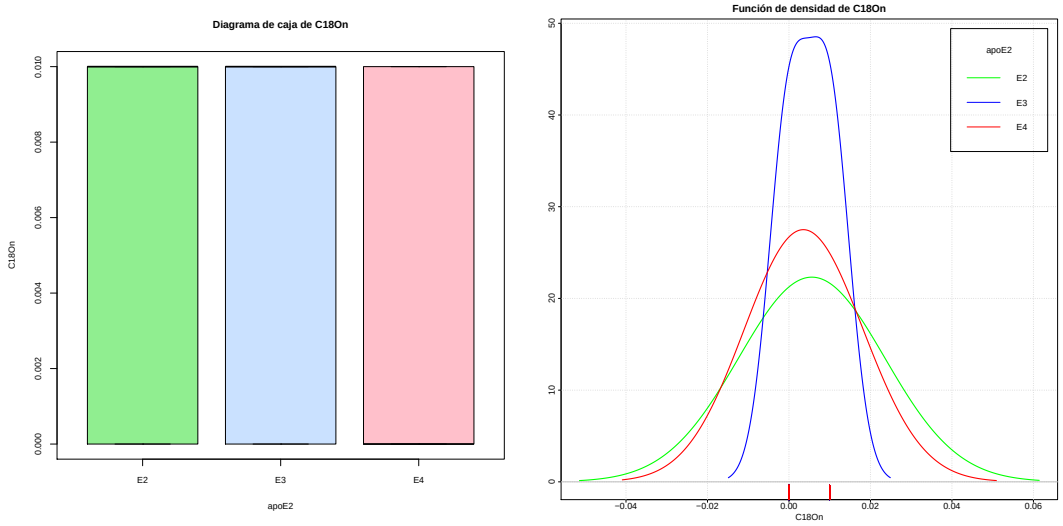

Figura B.123: Diagrama de caja y función de densidad de la variable $C18On$.

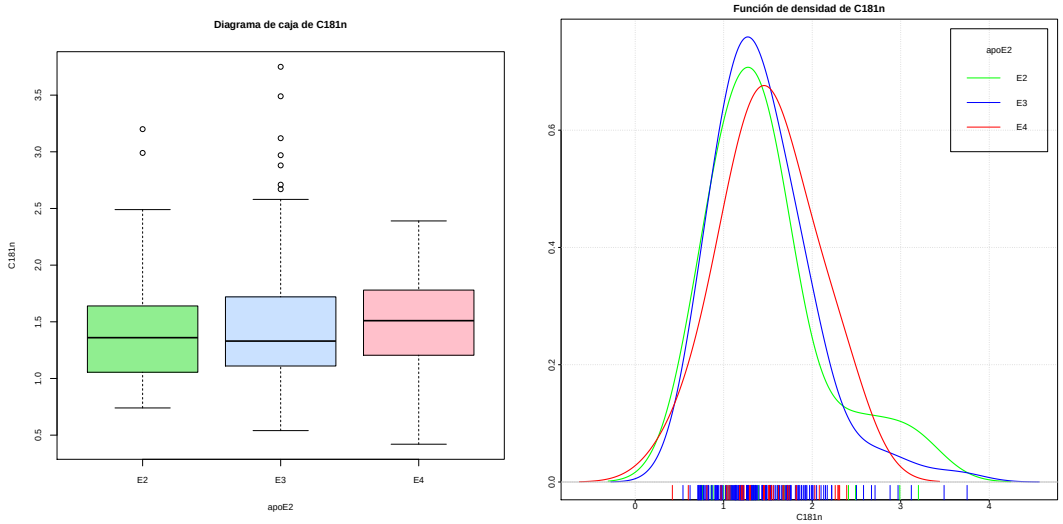

Figura B.124: Diagrama de caja y función de densidad de la variable $C181n$.

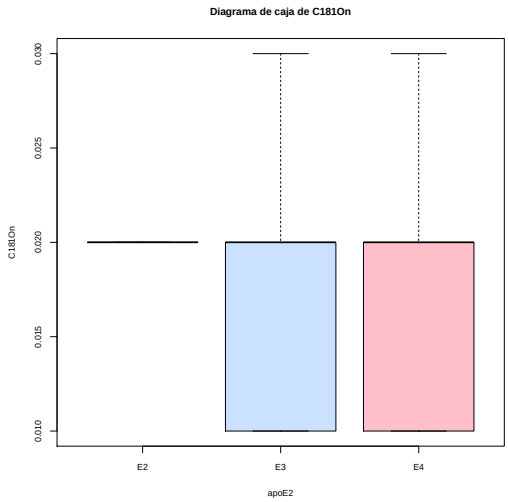

Figura B.125: Diagrama de caja y función de densidad de la variable $C181On$.

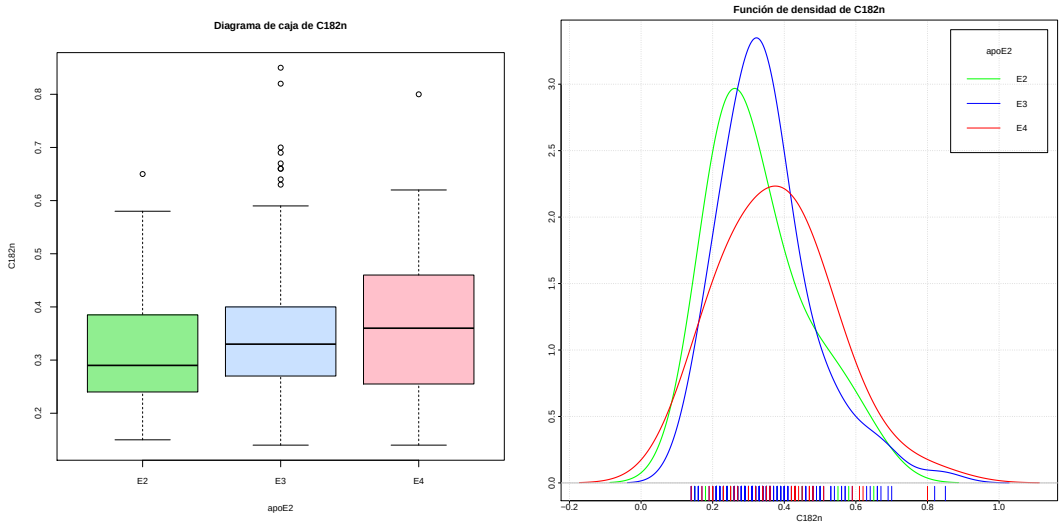

Figura B.126: Diagrama de caja y función de densidad de la variable $C182n$.

Anexo C

Representación de los factores

En este Anexo se presentan los resultados de la representación en el plano de las componentes más influyentes de cada conjunto, que se han calculado como combinación lineal de las variables más representativas que aparecen en el gráfico de la representación de todas las componentes.

C.1. Ácidos grasos

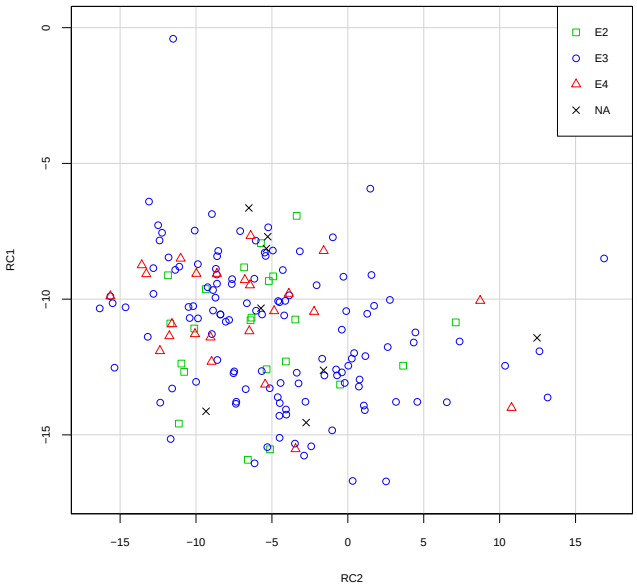

Figura C.1: Representación en el plano de los factores 1 y 2 del conjunto de Ácidos grasos.

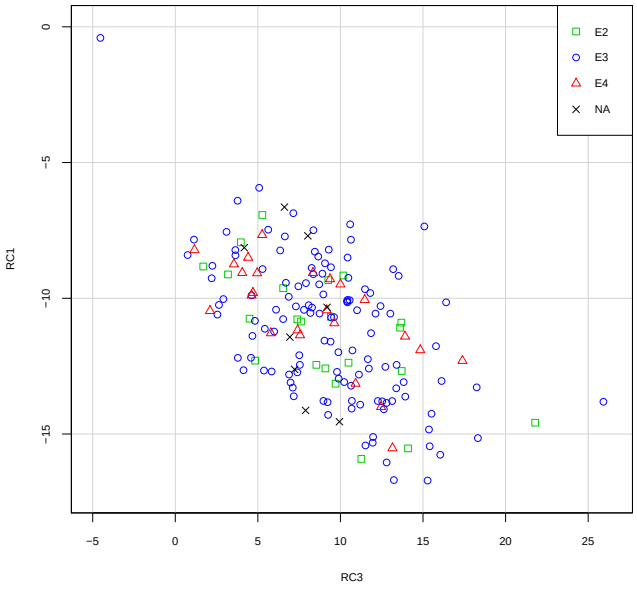

Figura C.2: Representación en el plano de los factores 1 y 3 del conjunto de Ácidos grasos.

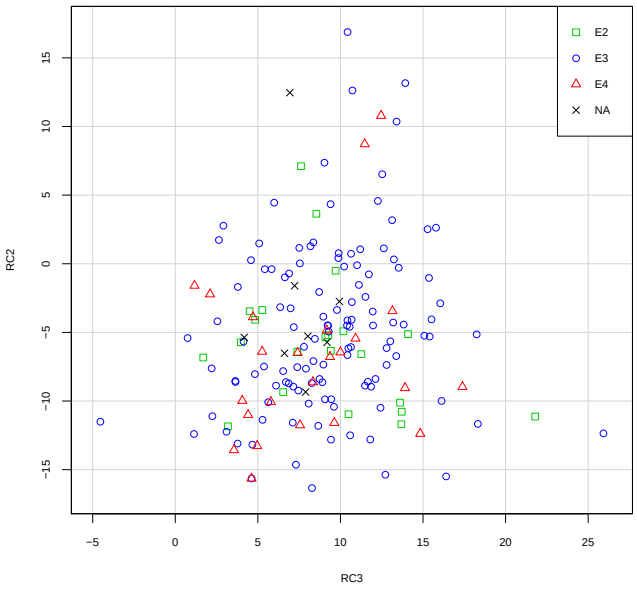

Figura C.3: Representación en el plano de los factores 2 y 3 del conjunto de Ácidos grasos.

C.2. Aminoácidos

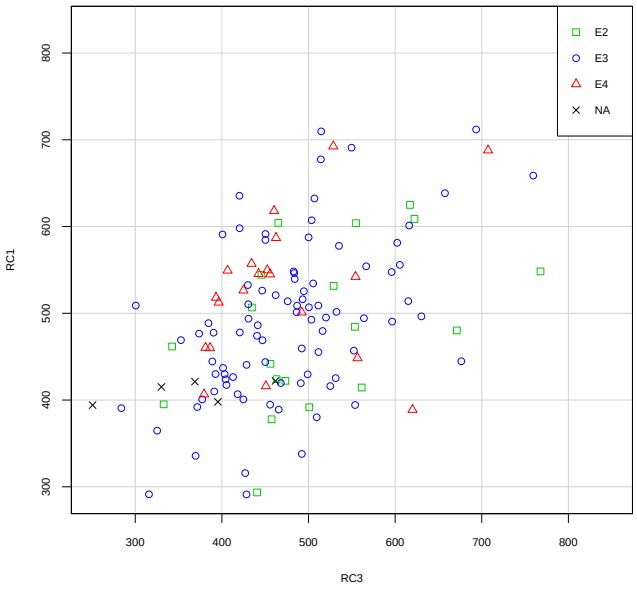

Figura C.4: Representación en el plano de los factores 1 y 2 del conjunto de Aminoácidos.

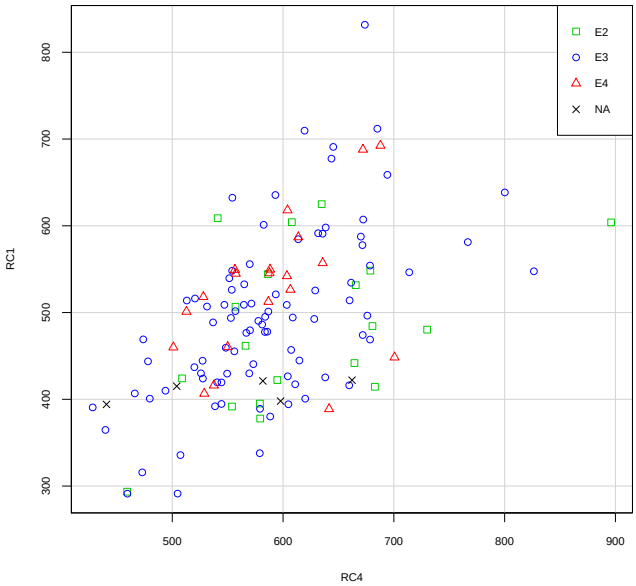

Figura C.5: Representación en el plano de los factores 1 y 3 del conjunto de Aminoácidos.

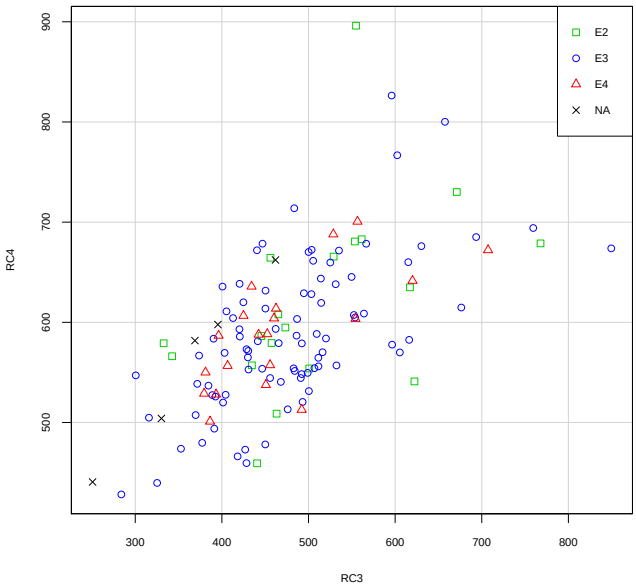

Figura C.6: Representación en el plano de los factores 2 y 3 del conjunto de Aminoácidos.

C.3. Bioquímicas

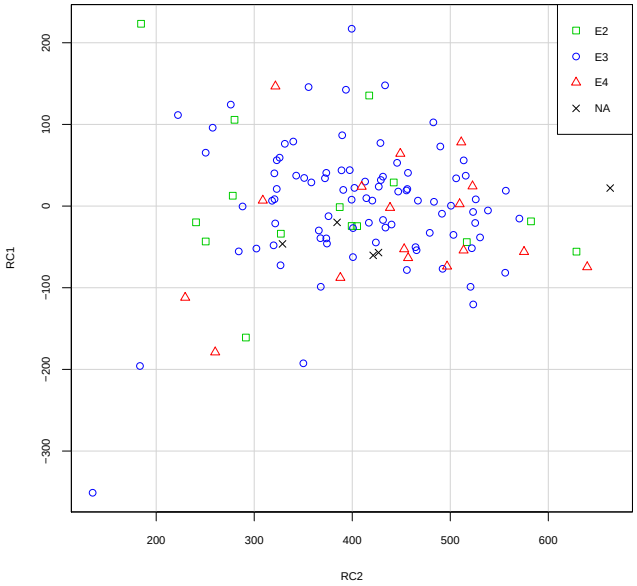

Figura C.7: Representación en el plano de los factores 1 y 2 del conjunto de Bioquímicas.

Figura C.8: Representación en el plano de los factores 1 y 3 del conjunto de Bioquímicas.

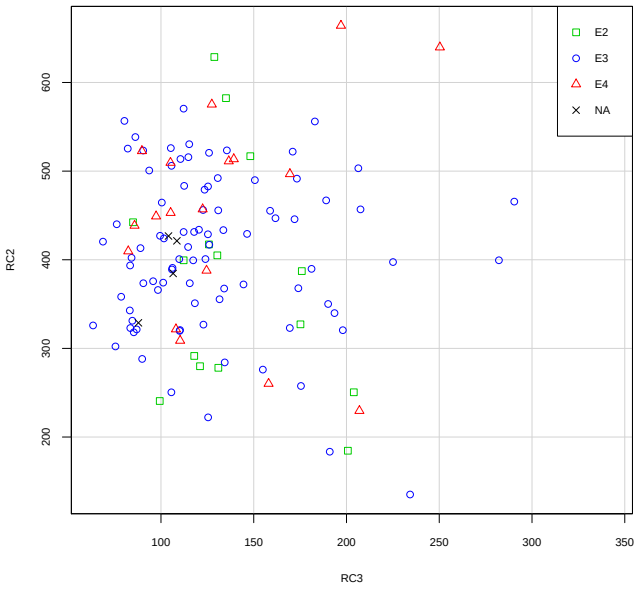

Figura C.9: Representación en el plano de los factores 2 y 3 del conjunto de Bioquímicas.

C.4. Carnitinas

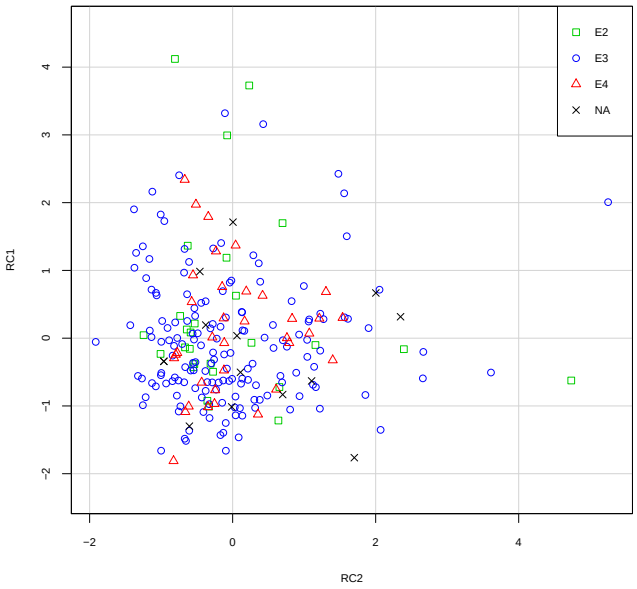

Figura C.10: Representación en el plano de los factores 1 y 2 del conjunto de Carnitinas.

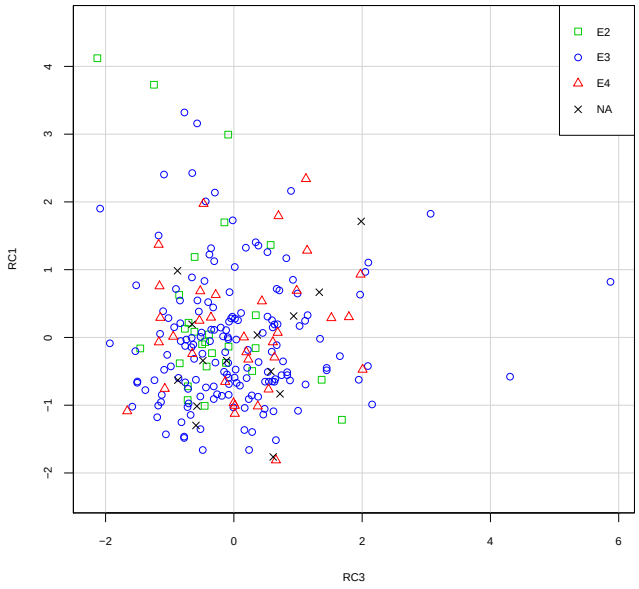

Figura C.11: Representación en el plano de los factores 1 y 3 del conjunto de Carnitinas.

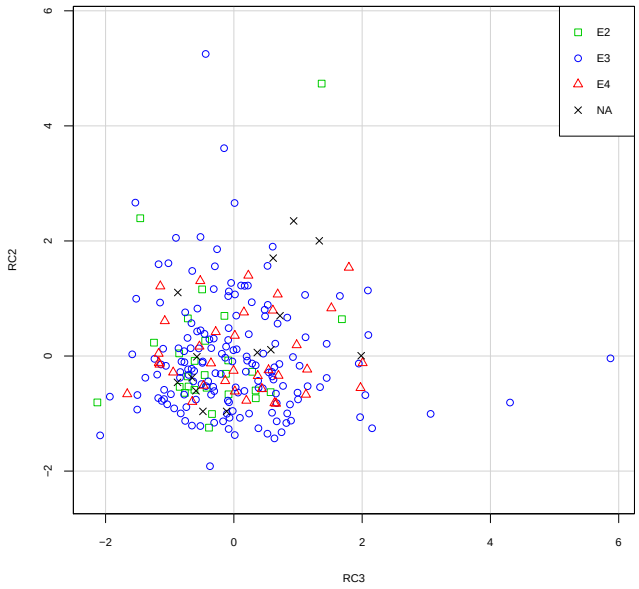

Figura C.12: Representación en el plano de los factores 2 y 3 del conjunto de Carnitinas.

Anexo D

Resultados de las regresiones multinomiales por conjuntos

D.1. Ácidos grasos

D.1.1. Primera regresión hacia delante

```
> regAcidosRL2<-multinom(apoE2~., data=AcidosRL2,maxit=10000)
> regAcidosPasoRL2<-stepwise(regAcidosRL2,direction="forward",
+ criterion="AIC",maxit=1000)
...
        Df    AIC
        Df    AIC
<none>     245.26
+ v240P  2 246.25
+ v120H  2 246.29
+ v220P  2 246.37
+ v161H  2 246.38
+ v200P  2 246.70
+ v200H  2 246.78
+ v140P  2 246.84
+ a183H  2 246.90
+ v241P  2 247.20
+ v260H  2 247.21
+ v182H  2 247.28
+ v221P  2 247.47
+ v204P  2 247.52
+ v161P  2 247.64
+ v181H  2 247.81
+ v202H  2 247.89
+ v201P  2 247.91
```

```
+ v260P  2 248.32
+ v201H  2 248.38
+ v160H  2 248.38
+ v184P  2 248.56
+ v203P  2 248.66
+ v202P  2 248.78
+ v180H  2 248.83
+ v160P  2 248.84
+ g183P  2 248.90
+ v181P  2 248.91
+ v225H  2 249.01
+ v240H  2 249.02
+ v203H  2 249.03
+ v220H  2 249.06
+ v140H  2 249.09
+ v224H  2 249.09
+ v226P  2 249.10
+ v204H  2 249.10
+ v225P  2 249.11
+ v205P  2 249.13
+ v180P  2 249.13
+ v224P  2 249.16
+ v205H  2 249.23
+ v241H  2 249.25
> regAcidosPasoRL2
Call:
multinom(formula = apoE2 ~ g183H + v182P + v226H + v221H + v184H +
    a183P, data = AcidosRL2, maxit = 10000)

Coefficients:
   (Intercept)      g183H      v182P      v226H    v221H     v184H     a183P
E3  -0.3970592   9.289896 0.01936891  0.1722238 10.87125 -7.810451 -2.130366
E4   0.3069252 -15.446728 0.09621001 -0.1360441 11.13595 -7.637845 -8.110303

Residual Deviance: 217.2601
AIC: 245.2601
```

D.1.2. Regresión paso a paso final

```
> regAcidosRL2<-multinom(apoE2~., data=AcidosPasoFinal,maxit=10000)
> regAcidosPasoFinalBF<-stepwise(regAcidosRL2,direction="backward/forward",
+ criterion="AIC",maxit=1000)
...
        Df    AIC
```

```
<none>     245.26
- a183P  2 245.60
- v226H  2 246.01
+ v120H  2 246.29
- v184H  2 246.46
+ v260H  2 247.21
+ v204P  2 247.52
+ v201P  2 247.91
+ v201H  2 248.38
+ v160H  2 248.38
- v221H  2 248.81
+ v224P  2 249.16
- v182P  2 251.26
- g183H  2 254.47
> regAcidosPasoFinalBF
Call:
multinom(formula = apoE2 ~ g183H + v182P + v226H + v221H + v184H +
    a183P, data = AcidosPasoFinal, maxit = 10000)

Coefficients:
   (Intercept)      g183H      v182P      v226H    v221H     v184H     a183P
E3  -0.3970592   9.289896 0.01936891  0.1722238 10.87125 -7.810451 -2.130366
E4   0.3069252 -15.446728 0.09621001 -0.1360441 11.13595 -7.637845 -8.110303

Residual Deviance: 217.2601
AIC: 245.2601
> z<-summary(regAcidosPasoFinalBF)$coefficients /
+ summary(regAcidosPasoFinalBF)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)     g183H     v182P     v226H      v221H      v184H      a183P
E3   0.8077351 0.1508610 0.7358388 0.2466592 0.03361932 0.02622515 0.48559537
E4   0.8843683 0.1484103 0.1833090 0.4975732 0.03334565 0.11598545 0.05727444
> predictAcidosPasoFinal<-predict(regAcidosPasoFinalBF,
+ newdata=Acidos,type="class")
> tblAcidosPasoRL2<-table(Acidos$apoE2,predictAcidosPasoFinal)
> tblAcidosPasoRL2
    predictAcidosPasoFinal
      E2  E3  E4
  E2   2  19   1
  E3   1 118   3
  E4   1  18   5
> 1-sum(diag(tblAcidosPasoRL2))/sum(tblAcidosPasoRL2)
[1] 0.2559524
```

D.2. Aminoácidos

D.2.1. Primera regresión hacia delante

```
> regAminoRL2<-multinom(apoE2~., data=AminoR2L,maxit=1000)
> regAminoPasoRL2<-stepwise(regAminoRL2,direction="forward",
+ criterion="AIC",maxit=1000)
...
                      Df    AIC
<none>                   215.38
+ Serina              2 215.49
+ Triptofano          2 216.13
+ Valina              2 216.20
+ Alanina             2 216.41
+ AcGlutamico         2 216.62
+ Treonina            2 217.06
+ Prolina             2 217.09
+ Ornitina            2 217.15
+ Lisina              2 217.17
+ Asparragina         2 217.26
+ AcAspartico         2 217.31
+ Taurina             2 217.49
+ AcAlfaAminobutirico  2 217.69
+ FenilAlanina        2 217.73
+ Glutamina           2 217.75
+ Leucina             2 217.96
+ Fosfoserina         2 218.01
+ Metionina           2 218.04
+ Histidina           2 218.22
+ Citrulina           2 218.33
+ Tirosina            2 218.61
+ Isoleucina          2 218.99
+ X1MetilHistidina    2 219.05
> regAminoPasoRL2
Call:
multinom(formula = apoE2 ~ Cistina + Arginina + Glicina, data = AminoRL2,
    maxit = 1000)

Coefficients:
   (Intercept)    Cistina     Arginina      Glicina
E3    1.254807 0.04700427 -0.022018348 -0.004968717
E4    1.852846 0.02535922 -0.003610715 -0.016186397

Residual Deviance: 199.3762
```

```
AIC: 215.3762
```

D.2.2. Regresión paso a paso final

```
> regAminoRL2<-multinom(apoE2~., data=AminoPasoFinal,maxit=10000)
> regAminoPasoFinalBF<-stepwise(regAminoRL2,direction="backward/forward",
+ criterion="AIC",maxit=1000)
...
                      Df    AIC
<none>                   230.27
+ Prolina              2 231.37
+ AcAlfaAminobutirico  2 232.22
+ Glicina              2 232.31
+ Fosfoserina          2 232.77
- Arginina             2 234.42
- Alanina              2 234.94
- Valina               2 235.61
- Cistina              2 238.81
> regAminoPasoFinalBF
Call:
multinom(formula = apoE2 ~ Cistina + Arginina + Valina + Alanina,
    data = AminoPasoFinal, maxit = 10000)

Coefficients:
   (Intercept)      Cistina    Arginina     Valina      Alanina
E3   0.9319050  0.044270227 -0.03442584 0.01371560 -0.008057817
E4  -0.4900941 -0.006109293 -0.01809216 0.02712575 -0.012337855

Residual Deviance: 210.2702
AIC: 230.2702
> z<-summary(regAminoPasoFinalBF)$coefficients /
+ summary(regAminoPasoFinalBF)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)    Cistina    Arginina      Valina     Alanina
E3   0.3523873 0.04071741 0.009023253 0.069794870 0.011238959
E4   0.7486978 0.82053761 0.223519975 0.003887659 0.006020541
> predictAminoPasoFinalBF<-predict(regAminoPasoFinalBF,
+ newdata=Amino,type="class")
> tblAminoPasoRL2<-table(Amino$apoE2,predictAminoPasoFinalBF)
> tblAminoPasoRL2
    predictAminoPasoFinalBF
      E2  E3  E4
  E2   2  17   0
```

```
  E3   0 108   0
  E4   1  21   2
> 1-sum(diag(tblAminoPasoRL2))/sum(tblAminoPasoRL2)
[1] 0.2582781
```

D.3. Bioquímicas

D.3.1. Primera regresión hacia delante

```
> regBioquimicaRL2<-multinom(apoE2~., data=BioquimicaRL2,maxit=1000)
> regAminoPasoRL2<-stepwise(regAminoRL2,direction="forward",
+ criterion="AIC",maxit=1000)
...
        Df    AIC
<none>     186.71
+ PCRU   2 187.76
+ TRIG   2 187.85
+ BGP    2 188.31
+ GLU    2 188.74
+ INS    2 188.98
+ NEFA   2 189.10
+ APOA   2 189.45
+ GGT    2 189.46
+ CTx    2 189.57
+ GPT    2 189.76
+ CHOL   2 189.84
+ LDL    2 189.91
+ GOT    2 190.05
+ BHID   2 190.19
+ SE     2 190.22
+ cHDL   2 190.25
+ LEPT   2 190.45
+ HEMGm  2 190.56
+ HEMG   2 190.59
> regBioquimicaPasoRL2
Call:
multinom(formula = apoE2 ~ APOB + LPA + VitD, data = BioquimicaRL2,
    maxit = 1000)

Coefficients:
   (Intercept)       APOB          LPA        VitD
E3    1.955848 0.01363187  0.009304792 -0.03068557
E4   -2.877366 0.04003707 -0.023768237 -0.01020046
```

```
Residual Deviance: 170.7066
AIC: 186.7066
```

D.3.2. Regresión paso a paso final

```
> regBioquimicaRL2<-multinom(apoE2~., data=BioquimicaPasoFinal,maxit=10000)
> regBioquimicaPasoFinalBF<-stepwise(regBioquimicaRL2,direction="backward/forward",
+ criterion="AIC",maxit=1000)
...
       Df    AIC
<none>    242.20
- VitD  2 242.31
+ CHOL  2 243.04
- LPA   2 243.81
- PCRU  2 243.94
+ LDL   2 244.75
+ CTx   2 245.53
- APOB  2 246.13
> regBioquimicaPasoFinalBF
Call:
multinom(formula = apoE2 ~ LPA + PCRU + APOB + VitD, data = BioquimicaPasoFinal,
    maxit = 10000)

Coefficients:
   (Intercept)          LPA       PCRU       APOB        VitD
E3    1.222844  0.009862354 -0.5133741 0.01900567 -0.02145933
E4   -1.713475 -0.008976054 -1.3614688 0.03554224 -0.01278834

Residual Deviance: 222.1969
AIC: 242.1969
> z<-summary(regBioquimicaPasoFinalBF)$coefficients /
+ summary(regBioquimicaPasoFinalBF)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)       LPA       PCRU        APOB       VitD
E3   0.3559097 0.3262987 0.14964440 0.085103523 0.05090442
E4   0.3103718 0.5103248 0.02755199 0.007746989 0.38218204
> predictBioquimicaPasoFinal<-predict(regBioquimicaPasoFinalBF,
+ newdata=Bioquimica,type="class")
> tblBioquimicaPasoRL2<-table(Bioquimica$apoE2,predictBioquimicaPasoFinal)
> tblBioquimicaPasoRL2
    predictBioquimicaPasoFinal
      E2  E3  E4
```

```
  E2   2  23   0
  E3   3 152   0
  E4   0  34   0
> 1-sum(diag(tblBioquimicaPasoRL2))/sum(tblBioquimicaPasoRL2)
[1] 0.2803738
```

D.4. Carnitinas

D.4.1. Primera regresión hacia delante

```
> regCarnitinasRL2<-multinom(apoE2~., data=CarnitinasRL2,maxit=1000)
> regCarnitinasPasoRL2<-stepwise(regCarnitinasRL2,direction="forward",
+ criterion="AIC",maxit=1000)
...
         Df    AIC
<none>      339.64
+ C4n     2 340.32
+ C6DCn   2 340.54
+ C81n    2 340.78
+ C18n    2 341.11
+ C51n    2 341.65
+ C161On  2 341.79
+ C16n    2 341.99
+ C161n   2 342.35
+ C10n    2 342.48
+ C181On  2 342.51
+ C142n   2 342.74
+ C3n     2 342.75
+ C102n   2 342.75
+ C3DCn   2 342.83
+ C8n     2 342.87
+ C121n   2 343.02
+ C2n     2 343.13
+ C14On   2 343.25
+ C0n     2 343.32
+ C141n   2 343.42
+ C181n   2 343.57
+ C12n    2 343.61
+ C101n   2 343.62
+ C5DCn   2 343.77
> regCarnitinasPasoRL2
Call:
multinom(formula = apoE2 ~ C16On + C6n + C14n + C182n + C4DCn +
```

```
    C5n + C18On, data = CarnitinasRL2, maxit = 1000)

Coefficients:
   (Intercept)     C16On        C6n      C14n    C182n       C4DCn       C5n
E3 2.140794111  87.85514 -12.170863 -33.39329 5.625742 -0.99454661 8.4411653
E4 0.007638532 -42.48992   7.274048 -26.94467 6.445974  0.09691212 0.2091391
        C18On
E3  -22.89289
E4 -101.67111

Residual Deviance: 307.6399
AIC: 339.6399
```

D.4.2. Regresión paso a paso final

```
> regCarnitinasRL2<-multinom(apoE2~., data=CarnitinasPasoFinal,maxit=10000)
> regCarnitinasPasoFinalBF<-stepwise(regCarnitinasRL2,direction="backward/forward",
+ criterion="AIC",maxit=1000)
...
        Df    AIC
<none>     339.64
- C18On  2 339.88
- C5n    2 340.42
+ C18n   2 341.11
- C182n  2 342.06
+ C3n    2 342.75
- C4DCn  2 343.12
+ C12n   2 343.61
- C16On  2 345.16
- C14n   2 346.07
- C6n    2 347.49
> regCarnitinasPasoFinalBF
Call:
multinom(formula = apoE2 ~ C16On + C6n + C14n + C182n + C4DCn +
    C5n + C18On, data = CarnitinasPasoFinal, maxit = 10000)

Coefficients:
   (Intercept)     C16On        C6n      C14n    C182n       C4DCn       C5n
E3 2.140794111  87.85514 -12.170863 -33.39329 5.625742 -0.99454661 8.4411653
E4 0.007638532 -42.48992   7.274048 -26.94467 6.445974  0.09691212 0.2091391
        C18On
E3  -22.89289
E4 -101.67111
```

```
Residual Deviance: 307.6399
AIC: 339.6399
> z<-summary(regCarnitinasPasoFinalBF)$coefficients /
+ summary(regCarnitinasPasoFinalBF)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)      C16On        C6n       C14n      C182n      C4DCn       C5n
E3  0.01695419 0.08807298 0.08123658 0.00147682 0.02909786 0.06500064 0.1159868
E4  0.99439211 0.49127139 0.34017052 0.04164252 0.02633515 0.87878756 0.9756682
        C18On
E3 0.62520988
E4 0.07990282
> predictCarnitinasPasoFinal<-predict(regCarnitinasPasoFinalBF,
+ newdata=Carnitinas,type="class")
> tblCarnitinasPasoRL2<-table(Carnitinas$apoE2,predictCarnitinasPasoFinal)
> tblCarnitinasPasoRL2
    predictCarnitinasPasoFinal
      E2  E3  E4
  E2   1  23   3
  E3   1 156   4
  E4   0  30   5
> 1-sum(diag(tblCarnitinasPasoRL2))/sum(tblCarnitinasPasoRL2)
[1] 0.2735426
```

Anexo E

Resultados de las regresiones logísticas con las componentes principales

E.1. Ácidos grasos

```
> regAcidos2<-multinom(apoE2~., data=AcidosTodo2[,-c(1)],maxit=10000)
> regAcidosPaso2<-stepwise(regAcidos2,direction="forward/backward",
+ criterion="AIC",maxit=1000)
...
        Df    AIC
<none>     245.26
- a183P  2 245.60
- v226H  2 246.01
- v184H  2 246.46
+ RC8    2 248.64
- v221H  2 248.81
+ RC4    2 248.91
+ RC6    2 248.97
+ RC7    2 249.03
+ RC3    2 249.04
+ RC2    2 249.17
+ RC1    2 249.23
+ RC5    2 249.24
- v182P  2 251.26
- g183H  2 254.47
> regAcidosPaso2
Call:
multinom(formula = apoE2 ~ g183H + v182P + v226H + v221H + v184H +
    a183P, data = AcidosTodo2[, -c(1)], maxit = 10000)

Coefficients:
```

```
   (Intercept)      g183H      v182P      v226H    v221H     v184H     a183P
E3  -0.3970592   9.289896 0.01936891  0.1722238 10.87125 -7.810451 -2.130366
E4   0.3069252 -15.446728 0.09621001 -0.1360441 11.13595 -7.637845 -8.110303

Residual Deviance: 217.2601
AIC: 245.2601
> z<-summary(regAcidosPaso2)$coefficients /
+ summary(regAcidosPaso2)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)     g183H     v182P     v226H      v221H      v184H      a183P
E3   0.8077351 0.1508610 0.7358388 0.2466592 0.03361932 0.02622515 0.48559537
E4   0.8843683 0.1484103 0.1833090 0.4975732 0.03334565 0.11598545 0.05727444
> predictAcidosPaso2<-predict(regAcidosPaso2,
+ newdata=AcidosTodo2,type="class")
> tblAcidosPaso2<-table(AcidosTodo2$apoE2,predictAcidosPaso2)
> tblAcidosPaso2
    predictAcidosPaso2
      E2  E3  E4
  E2   2  19   1
  E3   1 118   3
  E4   1  18   5
> 1-sum(diag(tblAcidosPaso2))/sum(tblAcidosPaso2)
[1] 0.2559524
```

E.2. Aminoácidos

```
> regAmino2<-multinom(apoE2~., data=AminoTodo2[,-c(1)],maxit=10000)
> regAminoPaso2<-stepwise(regAmino2,direction="backward/forward",
+ criterion="AIC",maxit=1000)
...
           Df    AIC
<none>        213.90
- Cistina   2 214.38
- Glicina   2 215.60
+ Alanina   2 215.71
+ RC2       2 217.13
+ RC1       2 217.70
- Arginina  2 217.88
- RC3       2 218.44
- RC4       2 218.71
> regAminoPaso2
Call:
```

```
multinom(formula = apoE2 ~ RC2 + RC3 + Cistina + Arginina + Glicina,
    data = AminoTodo2[, -c(1)], maxit = 10000)

Coefficients:
   (Intercept)         RC4        RC3     Cistina    Arginina      Glicina
E3   0.8910085 -0.02069257 0.01784414 0.039567775 -0.04366851 -0.005957521
E4   0.5512375 -0.03132739 0.02842308 0.004140921 -0.03813593 -0.019174296

Residual Deviance: 189.8996
AIC: 213.8996
> z<-summary(regAminoPaso2)$coefficients /
+ summary(regAminoPaso2)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)        RC4         RC3   Cistina   Arginina   Glicina
E3  0.24442943 0.01731644 0.022571299 0.1098861 0.01027232 0.2489629
E4  0.06827112 0.01018054 0.007363939 0.8964372 0.08847876 0.0275499
> predictAminoPaso2<-predict(regAminoPaso2,
+ newdata=AminoTodo2,type="class")
> tblAminoPaso2<-table(AminoTodo2$apoE2,predictAminoPaso2)
> tblAminoPaso2
    predictAminoPaso2
     E2 E3 E4
  E2  4 15  0
  E3  2 90  0
  E4  0 20  0
> 1-sum(diag(tblAminoPaso2))/sum(tblAminoPaso2)
[1] 0.2824427
```

E.3. Bioquímicas

```
> regBioquimica2<-multinom(apoE2~., data=BioquimicaTodo2[,-c(1)],maxit=10000)
> regBioquimicaPaso2<-stepwise(regBioquimica2,direction="forward/backward",
+ criterion="AIC",maxit=1000)
...
       Df    AIC
<none>    186.71
+ PCRU  2 187.76
- VitD  2 188.63
+ RC6   2 188.89
+ RC3   2 189.71
+ RC8   2 189.72
+ RC5   2 189.72
```

```
+ RC4   2 189.82
+ RC7   2 189.90
+ RC2   2 189.96
+ RC1   2 190.09
- LPA   2 190.22
- APOB  2 191.48
> regBioquimicaPaso2
Call:
multinom(formula = apoE2 ~ LPA + APOB + VitD, data = BioquimicaTodo2[,
    -c(1)], maxit = 10000)

Coefficients:
   (Intercept)          LPA       APOB        VitD
E3    1.955848  0.009304792 0.01363187 -0.03068557
E4   -2.877366 -0.023768237 0.04003707 -0.01020046

Residual Deviance: 170.7066
AIC: 186.7066
> z<-summary(regBioquimicaPaso2)$coefficients /
+ summary(regBioquimicaPaso2)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)       LPA        APOB       VitD
E3   0.1963410 0.4412493 0.273274737 0.02772626
E4   0.1523303 0.1974993 0.009124525 0.58613281
> predictBioquimicaPaso2<-predict(regBioquimicaPaso2,
+ newdata=BioquimicaTodo2,type="class")
> tblBioquimicaPaso2<-table(BioquimicaTodo2$apoE2,predictBioquimicaPaso2)
> tblBioquimicaPaso2
    predictBioquimicaPaso2
     E2 E3 E4
  E2  1 14  0
  E3  1 86  1
  E4  0 18  1
> 1-sum(diag(tblBioquimicaPaso2))/sum(tblBioquimicaPaso2)
[1] 0.2786885
```

E.4. Carnitinas

```
> regCarnitinas2<-multinom(apoE2~., data=CarnitinasTodo2[,-c(1)],maxit=10000)
> regCarnitinasPaso2<-stepwise(regCarnitinas2,direction="forward/backward",
+ criterion="AIC",maxit=1000)
...
```

```
        Df    AIC
<none>     339.64
- C18On  2 339.88
- C5n    2 340.42
- C182n  2 342.06
+ RC3    2 342.25
+ RC5    2 342.45
+ RC4    2 342.54
+ RC2    2 342.70
+ RC1    2 342.78
+ RC6    2 342.97
- C4DCn  2 343.12
- C16On  2 345.16
- C14n   2 346.07
- C6n    2 347.49
> regCarnitinasPaso2
Call:
multinom(formula = apoE2 ~ C16On + C6n + C14n + C182n + C4DCn +
    C5n + C18On, data = CarnitinasTodo2[, -c(1)], maxit = 10000)

Coefficients:
   (Intercept)     C16On        C6n      C14n    C182n       C4DCn       C5n
E3 2.140794111  87.85514 -12.170863 -33.39329 5.625742 -0.99454661 8.4411653
E4 0.007638532 -42.48992   7.274048 -26.94467 6.445974  0.09691212 0.2091391
        C18On
E3  -22.89289
E4 -101.67111

Residual Deviance: 307.6399
AIC: 339.6399
> z<-summary(regCarnitinasPaso2)$coefficients /
+ summary(regCarnitinasPaso2)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)      C16On        C6n       C14n      C182n      C4DCn       C5n
E3  0.01695419 0.08807298 0.08123658 0.00147682 0.02909786 0.06500064 0.1159868
E4  0.99439211 0.49127139 0.34017052 0.04164252 0.02633515 0.87878756 0.9756682
        C18On
E3 0.62520988
E4 0.07990282
> predictCarnitinasPaso2<-predict(regCarnitinasPaso2,
+ newdata=CarnitinasTodo2,type="class")
> tblCarnitinasPaso2<-table(CarnitinasTodo2$apoE2,predictCarnitinasPaso2)
> tblCarnitinasPaso2
```

```
    predictCarnitinasPaso2
      E2  E3  E4
  E2   1  23   3
  E3   1 156   4
  E4   0  30   5
> 1-sum(diag(tblCarnitinasPaso2))/sum(tblCarnitinasPaso2)
[1] 0.2735426
```

Anexo F

Regresión logística final

F.1. Regresión logística con todos los conjuntos

```
> regDatosTodo2<-multinom(apoE2~., data=DatosTodo2[,-c(1)],maxit=1000000)
> regDatosTodoPaso2F<-stepwise(regDatosTodo2,direction="forward",
+ criterion="AIC",maxit=1000000)
...
             Df    AIC
<none>          126.11
+ v161P       2 126.52
+ CHOL        2 126.68
+ C6n         2 126.93
+ Arginina    2 127.06
+ v221H       2 127.42
+ C18On       2 127.52
+ Glicina     2 127.75
+ v184H       2 127.78
+ Valina      2 127.85
+ Cistina     2 127.89
+ PCRU        2 128.65
+ CarRC1      2 128.77
+ C3DCn       2 128.86
+ CarRC4      2 128.91
+ C182n       2 128.92
+ Isoleucina  2 128.93
+ Serina      2 129.02
+ BGP         2 129.07
+ BHID        2 129.09
+ GOT         2 129.09
+ v160P       2 129.14
+ LPA         2 129.15
```

```
+ C2n         2 129.20
+ AmiRC4      2 129.29
+ C181n       2 129.30
+ AmiRC3      2 129.35
+ C161On      2 129.37
+ C5n         2 129.50
+ Alanina     2 129.63
+ v205P       2 129.66
+ C14n        2 129.67
+ C4DCn       2 129.67
+ a183P       2 129.68
> regDatosTodoPaso2F
Call:
multinom(formula = apoE2 ~ g183H + v182P + VitD + C16On + v226H +
    APOB, data = DatosTodo2[, -c(1)], maxit = 1e+06)

Coefficients:
   (Intercept)     g183H       v182P        VitD     C16On       v226H
E3   -1.395861  20.51458 -0.06710545 -0.03983976  101.2927  0.45913865
E4   -6.149151 -66.74725  0.04926358  0.05176575 -100.1321 -0.08174769
         APOB
E3 0.01805661
E4 0.05026537

Residual Deviance: 98.10634
AIC: 126.1063
> z<-summary(regDatosTodoPaso2F)$coefficients /
+ summary(regDatosTodoPaso2F)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)      g183H     v182P       VitD C16On      v226H       APOB
E3  0.62869597 0.05182382 0.3098412 0.03262005     0 0.03716098 0.20050050
E4  0.09756295 0.00000000 0.2896714 0.05079231     0 0.77085561 0.01118721
> predictDatosTodoF<-predict(regDatosTodoPaso2F,
+ newdata=DatosTodo[,-c(1)],type="class")
> tblDatosTodoF<-table(DatosTodo$apoE2,predictDatosTodoF)
> tblDatosTodoF
    predictDatosTodoF
      E2  E3  E4
  E2   4  13   4
  E3   4 110   4
  E4   2  12   9
> 1-sum(diag(tblDatosTodoF))/sum(tblDatosTodoF)
[1] 0.2407407
```

F.2. Regresión Logística Multinomial sin el conjunto de Aminoácidos

F.2.1. Regresión paso a paso hacia delante

```
> regDatosNoAmino2<-multinom(apoE2~., data=DatosNoAmino2[,-c(1)],maxit=1000000)
> regDatosNoAminoPaso2F<-stepwise(regDatosNoAmino2,direction="forward",
+ criterion="AIC",maxit=1000000)
...
         Df    AIC
<none>      176.43
+ C182n   2 176.74
+ CarRC4  2 177.01
+ CarRC1  2 177.06
+ C4DCn   2 177.20
+ LPA     2 178.25
+ C18On   2 178.28
+ v161P   2 178.37
+ C5n     2 178.90
+ APOB    2 178.91
+ v160P   2 179.06
+ C2n     2 179.10
+ C3DCn   2 179.11
+ PCRU    2 179.19
+ C181n   2 179.57
+ a183P   2 179.68
+ BGP     2 179.69
+ GOT     2 180.23
+ C14n    2 180.27
+ v205P   2 180.34
+ BHID    2 180.40
+ C161On  2 180.43
+ C16n    2 180.57
> regDatosNoAminoPaso2F
Call:
multinom(formula = apoE2 ~ g183H + CHOL + v182P + C16On + v221H +
    VitD + v226H + v184H + C6n, data = DatosNoAmino2[, -c(1)],
    maxit = 1e+06)

Coefficients:
   (Intercept)      g183H         CHOL       v182P      C16On    v221H
E3   0.6079842   6.552362 0.0003950524 -0.02349397   96.84998 15.88974
E4  -4.8996233 -29.858987 0.0313121079  0.03607710 -116.30974 14.88476
          VitD      v226H     v184H      C6n
E3 -0.03203272  0.3526138 -9.867776 -9.10528
```

```
E4 -0.02059431 -0.2412772  1.473006 15.87229

Residual Deviance: 136.4255
AIC: 176.4255
> z<-summary(regDatosNoAminoPaso2F)$coefficients /
+ summary(regDatosNoAminoPaso2F)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept)     g183H        CHOL     v182P C16On        v221H      VitD
E3   0.7947500 0.2838958 0.963571751 0.5285147     0 0.000000e+00 0.0144647
E4   0.1177232 0.0000000 0.007521845 0.1422981     0 2.398082e-14 0.2095416
        v226H      v184H          C6n
E3 0.06987574 0.01136962 1.427547e-01
E4 0.38102293 0.78682253 2.222456e-05
> predictDatosNoAminoF<-predict(regDatosNoAminoPaso2F,
+ newdata=DatosNoAminoRL[,-c(1)],type="class")
> tblDatosNoAminoF<-table(DatosNoAminoRL$apoE2,predictDatosNoAminoF)
> tblDatosNoAminoF
    predictDatosNoAminoF
      E2  E3  E4
  E2   7  12   2
  E3   1 115   1
  E4   0   9  11
> 1-sum(diag(tblDatosNoAminoF))/sum(tblDatosNoAminoF)
[1] 0.1582278
```

F.2.2. Regresión paso a paso *backward/forward*

```
> regDatosNoAmino2<-multinom(apoE2~., data=DatosNoAmino2[,-c(1)],maxit=1000000)
> regDatosNoAminoPaso2BF<-stepwise(regDatosNoAmino2,direction="backward/forward",
+ criterion="AIC",maxit=1000000)
...
         Df    AIC
<none>      180.81
- v161P   2 181.03
- C18On   2 181.72
- C6n     2 181.96
+ C182n   2 182.35
+ LPA     2 183.51
+ C3DCn   2 183.51
+ v160P   2 183.57
+ C181n   2 183.68
- BHID    2 183.68
+ GOT     2 183.88
```

```
- C5n     2 184.13
- v184H   2 184.45
+ APOB    2 184.63
+ v205P   2 184.63
+ a183P   2 184.96
+ C4DCn   2 184.97
+ BGP     2 185.10
- C14n    2 185.33
+ PCRU    2 185.85
+ C161On  2 185.99
- C2n     2 186.09
- VitD    2 186.29
- CarRC4  2 188.15
- CarRC1  2 188.73
- g183H   2 189.01
- v221H   2 190.12
- C16On   2 190.41
- v226H   2 190.52
- CHOL    2 195.16
- v182P   2 197.81
> regDatosNoAminoPaso2BF
Call:
multinom(formula = apoE2 ~ v221H + g183H + v226H + v182P + v184H +
    v161P + CHOL + VitD + BHID + C14n + C6n + C16On + C5n + C18On +
    C2n + CarRC1 + CarRC4, data = DatosNoAmino2[, -c(1)], maxit = 1e+06)

Coefficients:
   (Intercept)    v221H     g183H      v226H       v182P     v184H
E3   -4.053223 19.65652  10.19533  0.5489301 -0.03191435 -9.426097
E4  -13.433832 20.70141 -33.41528 -0.5826304  0.06257534  7.909129
         v161P        CHOL        VitD       BHID       C14n       C6n
E3  0.80853854 0.005931983 -0.04405290  0.1245057 -49.554926 -16.13138
E4 -0.07091041 0.057595892 -0.01017413 -0.4080101   4.840686  11.32179
       C16On       C5n      C18On        C2n    CarRC1    CarRC4
E3  144.1482  12.30525   26.05722  0.4833955  1.034569 -3.234218
E4 -188.9523 -16.31227 -183.43931 -2.1486477 -3.375825 12.801955

Residual Deviance: 108.8057
AIC: 180.8057
> z<-summary(regDatosNoAminoPaso2BF)$coefficients /
+ summary(regDatosNoAminoPaso2BF)$standard.errors
> p<-(1-pnorm(abs(z), 0, 1)) * 2
> p
   (Intercept) v221H     g183H      v226H      v182P      v184H     v161P
```

```
E3  0.29259027     0 0.1320466 0.03016465 0.59852566 0.02469569 0.0597181
E4  0.01533372     0 0.0000000 0.09704418 0.07633066 0.25295063 0.9181955
           CHOL        VitD       BHID C14n          C6n C16On          C5n
E3 5.645996e-01 0.002345844 0.36890098    0 5.420109e-13     0 4.715802e-02
E4 7.783142e-05 0.620965108 0.06261222    0 0.000000e+00     0 7.549517e-15
   C18On        C2n     CarRC1    CarRC4
E3     0 0.40920093 0.25135802 0.3125107
E4     0 0.02931494 0.02225701 0.0169666
> predictDatosNoAminoBF<-predict(regDatosNoAminoPaso2BF,
+ newdata=DatosNoAminoRL[,-c(1)],type="class")
> tblDatosNoAminoBF<-table(DatosNoAminoRL$apoE2,predictDatosNoAminoBF)
> tblDatosNoAminoBF
    predictDatosNoAminoBF
      E2  E3  E4
  E2   8  11   2
  E3   1 114   2
  E4   1   5  14
> 1-sum(diag(tblDatosNoAminoBF))/sum(tblDatosNoAminoBF)
[1] 0.1392405
```