

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**ANÁLISIS DE SUBGRUPOS EN
ENsayos CLÍNICOS**

TESIS

QUE PARA OBTENER EL TÍTULO DE

MATEMÁTICAS APLICADAS

PRESENTA

REGINA CEBALLOS MONDRAGÓN

ASESOR: DR. LUIS ENRIQUE NIETO BARAJAS

REVISOR: DR. ERNESTO JUVENAL BARRIOS ZAMUDIO

CIUDAD DE MÉXICO

2019

«Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada “**Análisis de subgrupos en ensayos clínicos**”, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación».

FECHA

REGINA CEBALLOS MONDRAGÓN

«This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.»

«Esta publicación está basada en investigación utilizando información obtenida de www.projectdatasphere.org, la cual es mantenida por Project Data Sphere, LLC. Ni Project Data Sphere, LLC ni el (los) dueño(s) de cualquier información del sitio web han contribuido, aprobado o son de alguna forma responsables de los contenidos de esta publicación.»

Índice general

Introducción	1
Descripción del problema	1
Objetivos	2
Estructura	3
1. Enfoques de los análisis de subgrupos	5
1.1. Exploratorios	10
1.2. Predefinidos	12
1.3. Otras consideraciones	13
2. Preliminares	15
2.1. Fundamentos Estadísticos	15
2.1.1. Aleatorización	17
2.1.2. Urna de Pòlya	17
2.2. Estadística Bayesiana	20
2.2.1. Estimación puntual	22
2.2.2. Conjuntos creíbles	23
2.2.3. Pruebas de hipótesis	23
2.2.4. Distribución de Jeffreys	24
2.2.5. Distribuciones <i>g-prior</i>	27

2.2.6. Mezclas de <i>g-priors</i>	28
2.3. Simulación Monte Carlo de Cadenas de Markov	31
2.3.1. Muestreo de Gibbs	31
2.4. Aprendizaje de Máquina	32
2.4.1. Árboles de regresión y clasificación CART	33
2.4.2. Bosques aleatorios	34
2.4.3. Validación cruzada	35
2.4.4. <i>Bootstrap</i>	36
3. Métodos de análisis de subgrupos exploratorios	37
3.1. Método <i>Virtual Twins</i>	37
3.1.1. Medida de desempeño de la región <i>A</i>	42
3.2. Método de árboles de interacción	46
4. Métodos de análisis de subgrupos predefinidos	55
4.1. Selección Bayesiana de modelo basado en la urna de Pòlya	55
4.1.1. Definición de posibles modelos	58
4.1.2. Distribución del modelo <i>M</i> basándose en Urnas de Pòlya	59
4.1.3. Distribuciones iniciales de los parámetros desconocidos bajo un modelo	63
4.1.4. Probabilidades finales	68
4.1.5. Ampliación a múltiples covariables	69
4.1.6. Decisión de Bayes	71
4.2. Método de Dixon y Simon	75
4.2.1. Definición del modelo	75
4.2.2. Probabilidades iniciales y final	79
4.2.3. Simplificaciones	82
4.2.4. Inferencia	89
4.2.5. Modificación con <i>JAGS</i>	90

5. Ejemplos ilustrativos con datos simulados	93
5.1. <i>Virtual Twins</i>	101
5.2. Árboles de Interacción	112
5.3. Método de Dixon y Simon	121
5.4. Modificación con <i>JAGS</i>	132
5.5. Comparación entre métodos	146
6. Ejemplos ilustrativos con ensayos clínicos	149
6.1. Cáncer en el área de la cabeza y cuello	150
6.1.1. Exploración	154
6.1.2. <i>Virtual Twins</i>	158
6.1.3. Método de árboles de interacción	170
6.1.4. Conclusión de la base de datos de cáncer de cabeza y cuello	177
6.2. Leucemia	178
6.2.1. Definición de subgrupos	180
6.2.2. Exploración	181
6.2.3. Método de Dixon y Simon	183
6.2.4. Modificación con <i>JAGS</i>	187
6.2.5. Conclusión de la base de datos de leucemia	192
7. Conclusiones	195
Referencias	203

Introducción

Descripción del problema

En la actualidad, las decisiones que toman los doctores sobre el cuidado de un paciente están fuertemente determinadas por los tratamientos y medicamentos disponibles en el mercado. Las regulaciones para que un medicamento se pueda comerciar aseguran que, en promedio, los pacientes tengan una mejora significativa gracias al tratamiento. Sin embargo, muchas veces un tratamiento que para la mayoría de las personas tiene un efecto positivo, puede tener un efecto negativo para una minoría. Este fenómeno es común, por ejemplo, en pacientes con diversos tipos de cáncer, problemas cardiovasculares o enfermedades mentales. En algunos casos, prolongar el uso de dichos tratamientos puede resultar en el deterioro del bienestar de los pacientes (Rothwell 2005).

El análisis de subgrupos surge de la búsqueda de tratamientos que respondan de mejor manera a las necesidades específicas de cada paciente enfermo. Se define como las prácticas estadísticas que buscan determinar la heterogeneidad de efectos de un tratamiento para un grupo particular

de pacientes (R. Wang et al. 2007). Es decir, busca encontrar si existen subgrupos determinados por características físicas, demográficas, genéticas o del historial clínico para los cuales un tratamiento tiene efectos diferentes. Al identificar los subgrupos, no sólo se tiene un mayor conocimiento sobre los posibles riesgos y beneficios de un tratamiento, sino que se puede dar un tratamiento más personalizado y de recuperación más rápida al paciente.

Por lo tanto, es relevante estudiar y evaluar los métodos estadísticos que se han propuesto para hacer análisis de subgrupos. Un mayor conocimiento acerca de su funcionamiento inevitablemente resultará en ideas para mejorarlos, lo que a su vez ayudará a mejorar la identificación, valoración y confirmación de efectos en subgrupos.

Objetivos

El objetivo principal de este trabajo es describir y ejemplificar el funcionamiento de distintos métodos para realizar análisis de subgrupos. En específico, se comparan métodos que utilizan subgrupos predefinidos y subgrupos encontrados después de aplicar el método. Para evaluar los diferentes métodos se usan bases de datos simuladas y de ensayos clínicos reales. Para los métodos de subgrupos predefinidos se utilizan subgrupos para los cuales previamente se hayan encontrado efectos heterogéneos, con el motivo de hacer un estudio confirmatorio, no sólo exploratorio.

Algunos problemas que se pueden esperar son dificultades computacionales para aplicar los métodos y falta de información para definir distribuciones iniciales informativas en un enfoque Bayesiano. Este enfoque se utiliza en algunos de los métodos para contrastar con el

INTRODUCCIÓN

enfoque Frecuentista también se utiliza en la tesis. Se hablará más de los diferentes enfoques estadísticos en el siguiente capítulo.

Un objetivo adicional del trabajo es la aplicación de los conocimientos adquiridos en la carrera de Matemáticas Aplicadas. En particular, se utilizan conceptos de los cursos de Estadística Matemática, Estadística Bayesiana y Aprendizaje de Máquina. La falta de entendimiento de conceptos y herramientas matemáticas que no son básicas, genera dudas acerca de la importancia del estudio de las Matemáticas. En este trabajo se muestra una aplicación de la Estadística relevante en la actualidad, con repercusiones directas en la salud humana. Se busca ampliar el panorama del lector acerca de las múltiples aplicaciones de la Estadística y de las Matemáticas.

Estructura

El trabajo se desarrolla de la siguiente forma. En el Capítulo 1, se da una definición detallada del análisis de subgrupos, donde se exponen los principales enfoques que se han investigado y propuesto sobre este tema. En particular, se describen las diferencias entre análisis para encontrar subgrupos y análisis para confirmar la existencia de subgrupos. A estas prácticas se les refiere como subgrupos exploratorios y subgrupos predefinidos, respectivamente. También se exponen problemas que surgen al hacer análisis de subgrupos y las soluciones propuestas por diferentes autores. Se busca plasmar de manera más detallada los objetivos y aplicaciones del trabajo a través de este capítulo.

El Capítulo 2 es una introducción a la teoría que facilita la

INTRODUCCIÓN

compresión del trabajo. Una sección del capítulo se dedica a explicar conceptos de la Estadística, haciendo énfasis en exponer los fundamentos de la Estadística Bayesiana. Por otro lado, se describen brevemente los algoritmos de Aprendizaje de Máquina en los que se basan algunos de los métodos para analizar subgrupos.

Los métodos que se utilizan en esta tesis se describen exhaustivamente en los Capítulos 3 y 4, donde cada capítulo trata métodos para subgrupos exploratorios y predefinidos, respectivamente. Los criterios para la exposición fueron que los métodos revisados tengan aplicaciones adicionales a los ensayos clínicos, y que tengan diferencias importantes con el resto de los métodos. Dicho esto, se exponen ventajas y desventajas de cada método, así como ejemplos de aplicaciones. Se termina por enlistar otros métodos de análisis de subgrupos para complementar los métodos mencionados.

Algunos de los métodos expuestos en los Capítulos 3 y 4 se aplican a datos simulados en el Capítulo 5 y a datos reales en el Capítulo 6. En cada capítulo, se da una breve descripción de las bases de datos a las cuales se le aplican los métodos. Posteriormente, se exponen las covariables seleccionadas para predefinir o identificar subgrupos. Asimismo, se hace un análisis exploratorio para conocer más sobre la base de datos y se explican correcciones o modificaciones de variables.

Por último, se exponen las conclusiones del proyecto en el Capítulo 7. En este capítulo se incluyen sugerencias de usos de los métodos diferentes a la medicina y para finalizar se dan ejemplos relevantes de aplicaciones a futuro.

Capítulo 1

Enfoques de los análisis de subgrupos

Ante la necesidad de evaluar la calidad, efectos y seguridad de nuevas medicinas o tratamientos médicos han surgido diferentes métodos, experimentos y normas a lo largo de la historia. Hoy en día, la herramienta más utilizada y confiable para asegurar el uso seguro y útil de un tratamiento es el ensayo clínico (Spiegelhalter, Abrams y Myles 2004). Los ensayos clínicos son experimentos en los se administra de manera controlada uno o más tratamientos a un grupo de pacientes con características determinadas y se supervisan los resultados. En general, los ensayos clínicos suelen ser doblemente ciegos para asegurar aleatoriedad, lo que significa que los pacientes y los investigadores no saben qué tratamiento se está administrando.

Una práctica común dentro de los ensayos clínicos es el análisis de subgrupos que, como se definió en la introducción, busca determinar la

heterogeneidad de un tratamiento para un grupo de pacientes. Se distingue entre dos tipos de heterogeneidad: cualitativa y cuantitativa (Yusuf, Wittes et al. 1991). Cuando un tratamiento tiene el mismo efecto en toda la población, pero es de diferente magnitud en subgrupos se dice que presenta heterogeneidad cuantitativa. Por otro lado, se puede encontrar que un tratamiento tiene efectos positivos para la mayor parte de la población, pero dañinos para el resto. A esta diferencia en el sentido del efecto del tratamiento (positiva y negativa) se le llama heterogeneidad cualitativa.

El estudio de ambos tipos de heterogeneidad es de suma importancia. El autor Rothwell (2005) propone una analogía con el sistema judicial para reflejar la gravedad de tratar a todos los pacientes con la misma medicina. Si todos los sospechosos con altas probabilidades de asesinato se metieran a la cárcel, no se dejaría libre a nadie culpable y habría menos criminales en libertad. Sin embargo, condenar a alguien inocente a cadena perpetua es inadmisible. En cambio, sin importar las diferencias en el riesgo de un tratamiento, en la práctica se trata a todos los pacientes por igual.

El riesgo de utilizar un tratamiento no sólo se genera cuando para una minoría de la población el tratamiento puede ser dañino, sino también porque al escoger un tratamiento no efectivo sobre otro mejor, se evita que los pacientes se curen. En enfermedades crónicas resulta en un deterioro en la calidad de vida y en otras enfermedades puede resultar en la muerte. A pesar de ser una situación difícil de resolver, es menester que los doctores y pacientes sean conscientes de los beneficios y riesgos que un tratamiento puede tener.

Los análisis de subgrupos tienen aplicaciones que no sólo benefician a los pacientes. Se puede, por ejemplo, buscar demostrar significancia

estadística del tratamiento en un subgrupo cuando no la hay para la población general, lo cual podría ser un motivo para continuar un ensayo clínico que de otra forma se cancelaría (Song y Chi 2007). También pueden influir en la forma en que se comercializa un nuevo tratamiento en el mercado. En estos ejemplos, las compañías farmacéuticas se benefician de la posible producción de más medicamentos y, por lo tanto, más ventas.

Aunado a los avances en la biología molecular, así como en la minería de datos, la aplicación máxima del análisis de subgrupos podría ser la completa personalización del cuidado médico. A pesar de los inmensurables beneficios que conllevaría, hay gran desacuerdo dentro de la comunidad científica sobre la validez y regulación de los análisis de subgrupos. Los argumentos en contra surgen de que muchas veces los análisis son mal informados y engañosos, cuando existen, por ejemplo, conflictos de interés (Pocock et al. 2002).

Un ejemplo muy difundido del daño que pueden causar los análisis de subgrupos es un estudio en el que se mostró que la aspirina no prevenía infartos en mujeres, pero sí en hombres (Rothwell 2005 citando a The Canadian Cooperative Study Group 1978). Gracias a este estudio, se trató inefficientemente a muchas mujeres, al no prescribirles aspirina. Sin embargo, estudios posteriores llegaron a la conclusión de que era una observación falsa. Incluso se realizó un estudio con técnicas similares para mostrar diferentes efectos en subgrupos determinados por el signo zodiacal de los pacientes, el cual claramente no define la fisiología de un paciente ni su respuesta a un medicamento (Rothwell 2005 citando a Yusuf, Collins y Peto 1984).

Los resultados de estos análisis se deben posiblemente a uno de los

problemas principales a los que se enfrentan los análisis de subgrupos: el problema de la multiplicidad (Rothwell 2005). Este ocurre cuando se realizan múltiples pruebas con la misma muestra, las cuales hacen que la probabilidad del Error Tipo I, o rechazar la hipótesis nula cuando es verdadera, crezca rápidamente. Esto se da ya que entre más pruebas se hagan sobre la misma muestra, más probabilidad hay de que alguno de los resultados no sea correcto por mera casualidad. Por lo que hay más probabilidad de que se rechace la hipótesis nula de manera errónea. En el contexto del análisis de subgrupos, hay más probabilidad de encontrar efectos con significancia estadística por mera casualidad, a pesar de que no exista el efecto.

En adición a lo anterior, existe el problema de que la mayoría de los ensayos clínicos están diseñados para tener la potencia suficiente para encontrar sólo efectos generales (Rothwell 2005). Potencia en este contexto significa la probabilidad de encontrar un efecto en el tratamiento cuando el efecto existe (rechazar la hipótesis nula cuando es falsa). La potencia depende del nivel de significancia de las pruebas, del tamaño de la muestra y de la magnitud del efecto. Que un ensayo clínico esté diseñado para encontrar un efecto significa que tiene el tamaño de muestra necesario para encontrar un efecto de una magnitud prevista cierto nivel de confianza. Por lo que si se busca encontrar efectos en subgrupos es necesario aumentar el tamaño de muestra. El no prever la existencia de efectos en subgrupos resulta en pérdida de información que se puede extraer de ensayos clínicos.

Por otro lado, se ha encontrado que muchos de los artículos que reportan efectos en subgrupos no son rigurosos y tienen conclusiones exageradas (Pocock et al. 2002; Assmann et al. 2000). Un ejemplo popular es el estudio de la aspirina que se mencionó anteriormente. Esto

disminuye la credibilidad en los estudios y de los métodos para encontrar efectos en subgrupos. A pesar de los múltiples problemas, es claro que la solución a los análisis de subgrupos no es dejar de hacerlos. Es por esto que ha habido gran esfuerzo de parte de la comunidad médica y estadística por establecer ciertas pautas para realizar e identificar estudios creíbles y con validez estadística (R. Wang et al. 2007; Rothwell 2005; Sun et al. 2014; Yusuf, Wittes et al. 1991; Lagakos 2006). Incluso, en el año 1996 se creó la declaración de Estándares Consolidados para Reportar Ensayos, CONSORT por sus siglas en inglés (*CONsolidated Standars Of Reporting Trials*) (*The CONSORT Website* 2019), para mejorar la calidad de las publicaciones sobre ensayos clínicos publicada en el artículo de Begg et al. (1996). Actualmente se utiliza la versión del Schulz et al. (2010).

Entre las sugerencias para mejorar la calidad de los estudios de subgrupos está utilizar pruebas de interacción formales. Estas se pueden hacer por ejemplo con una regresión lineal de la respuesta al tratamiento, donde uno de los términos es una covariante de interés multiplicada por el tratamiento. Los análisis con mayor validez son aquellos que comparan subgrupos complementarios. Además, no se deben utilizar los valores de p (p -values) de cada subgrupo para reportar significancia ya que estos tienen tasas altas de falsos positivos (Rothwell 2005; Lagakos 2006; Yusuf, Wittes et al. 1991; R. Wang et al. 2007). El aumento en falsos positivos se da porque para obtener los valores de p se hacen múltiples comparaciones sobre la misma muestra. Similarmente al aumento del Error Tipo I, al hacer muchas comparaciones es probable que al menos uno de los valores de p obtenidos sea menor que 0.05 por mera casualidad, por lo que aumenta el número de falsos positivos. 0.05 es un número utilizado comúnmente para denotar significancia.

La mayoría de las sugerencias se basan en diferenciar entre subgrupos predefinidos y subgrupos exploratorios. Utilizando las definiciones de Yusuf (1991), en este trabajo nos referiremos a subgrupos predefinidos como un grupo de pacientes caracterizado por parámetros iniciales comunes, y a subgrupos exploratorios o post-hoc como un grupo de pacientes caracterizados por una variable medida después de la aleatorización y posiblemente afectada por el tratamiento.

Hay un consenso general de que la validez de los subgrupos incrementa si los subgrupos a evaluar se definieron previamente a realizar el estudio. Esto se debe principalmente a que, si los subgrupos son predefinidos, la aleatorización del tratamiento puede ser estratificada. Es decir, se puede diseñar la muestra para que cada subgrupo tenga un número balanceado de pacientes tomando el tratamiento y el control (Assmann et al. 2000). Para subgrupos exploratorios, es más probable que las diferencias en el tratamiento se den por casualidad y no porque en realidad haya un efecto. Es por esto que sólo sirven para crear hipótesis, no para probarlas (Rothwell 2005; Lagakos 2006; Yusuf, Wittes et al. 1991; R. Wang et al. 2007). A continuación, se dan sugerencias y precauciones para análisis con cada tipo de subgrupo.

1.1. Exploratorios

Algunos autores argumentan que el uso de subgrupos exploratorios es necesario ya que es la única forma de identificar subgrupos no intuitivos y proponen varias formas de mejorar la validez estadística. Por ejemplo, Su et al. (2009) sugieren dividir la muestra sobre la cual se hace el análisis en tres conjuntos: un conjunto \mathcal{L}_1 de aprendizaje, \mathcal{L}_2 de validación y \mathcal{L}_3

de prueba. A los primeros dos conjuntos se le aplican los métodos y se encuentran los subgrupos. Luego, el conjunto de prueba se utiliza para confirmar el buen funcionamiento de los subgrupos.

De igual forma, otros autores sugieren que el sesgo se elimina al predefinir la metodología, en vez de predefinir los subgrupos (J. C. Foster, Taylor y Ruberg 2011). Se argumenta que siempre va a haber subjetividad presente en el análisis de subgrupos, incluso cuando son predefinidos. El escoger los subgrupos depende de la decisión de los científicos y se pueden evitar identificar o reportar subgrupos intencionalmente (Su et al. 2009). También, se pueden hacer muchas pruebas y reportar sólo aquellas en las que se encuentra un resultado atractivo o aparentemente novedoso.

En general, los métodos para encontrar subgrupos utilizan algoritmos de Aprendizaje de Máquina. Una gran parte de métodos utilizan árboles aleatorios basándose en el algoritmo CART (Breiman et al. 1984; Therneau y Atkinson 1997; Ciampi, Negassa y Lou 1995). Se basa en dividir a la muestra de pacientes en particiones de forma recursiva. Las diferencias entre los métodos que utilizan CART son principalmente en los criterios de división de los nodos y los criterios de paro. Tienen como ventaja que los árboles que resultan de los métodos son fáciles de interpretar. Además, la mayoría tienen paquetes en softwares públicos como **R** (R Core Team 2018) o **python** (Python Core Team 2015), por lo que su implementación es fácil. Otros métodos se basan en algoritmos como regresión Ridge o LASSO y máquinas de soporte vectorial, pero es probable que no se utilicen tan frecuentemente porque sus resultados son más difíciles de interpretar (Xu et al. 2015; Chipman, George y McCulloch 2010; Imai y Ratkovic 2013).

1.2. Predefinidos

Los análisis con subgrupos predefinidos tienen mayor aceptación dentro de la comunidad científica principalmente por dos razones: se pueden diseñar los ensayos clínicos de tal forma que haya una muestra balanceada y se puede controlar el crecimiento del Error Tipo I a causa de hacer múltiples pruebas. El método más conocido para hacer ajustes de multiplicidad es el método Bonferroni, donde se determina el número k de pruebas a realizar y cada prueba se hace a un nivel de significancia menor que α/k . Con esto, la suma de las probabilidades de cometer un Error Tipo I se mantiene por debajo del nivel deseado α (Proschan y Waclawiw 2000).

La mayoría de los métodos de análisis de subgrupos predefinidos se basan en la idea del método Bonferroni de controlar el Error Tipo I. Este se puede controlar por ejemplo a través de las pruebas de hipótesis o eligiendo parámetros dentro de los métodos por simulación. Dichos métodos son atractivos ya que hay una gran aceptación y uso, tanto de investigadores como del público en general, de la teoría que sustenta las pruebas de hipótesis Frecuentistas.

Además de variaciones a las pruebas de hipótesis, otros métodos incluyen regresiones lineales, logísticas y análisis a través del modelo de riesgos proporcionales de Cox como el método de Simon (2002).

A pesar de que hay un consenso de que los subgrupos predefinidos tienen mayor rigor científico que los exploratorios, es menester asegurar su validez estadística. Dentro de las sugerencias para aumentar la credibilidad de los análisis, se incluye utilizar métodos para obtener muestras aleatorias confiables y simples. Los subgrupos se deben estratificar a través de covariables conocidas. Asimismo, es conveniente

mantener el número de subgrupos limitado a pesar de los ajustes de multiplicidad (Assmann et al. 2000; R. Wang et al. 2007; Yusuf, Wittes et al. 1991).

Por otro lado, se sugiere que la selección de los subgrupos se haga basada en observaciones biológicas o patológicas para justificar la heterogeneidad. A pesar de los grandes avances en las técnicas matemáticas, estadísticas y computacionales, es importante que el manejo de los subgrupos esté respaldado por intuición clínica (Feinstein 1998). Particularmente porque tienen consecuencias directas en la salud humana.

1.3. Otras consideraciones

Independientemente del análisis que se lleve a cabo, es imperativo informar si los subgrupos son predefinidos o exploratorios, el número de análisis que se hicieron en total y la técnica que se utilizó para obtener una muestra aleatoria. Si el objetivo principal del estudio no es el análisis de subgrupos se debe especificar, para no poner demasiado énfasis en sus resultados, ya que esto disminuye la credibilidad de los estudios. Se sugiere también tomar en cuenta que los efectos estén siempre en un rango de probabilidad continua (Sun et al. 2014). Es decir, el tomar un efecto como total o como nulo es una simplificación que es preferible evitar (Assmann et al. 2000).

Finalmente, se debe tener precaución cuando un análisis de subgrupos se utiliza para tomar decisiones sobre el tratamiento de un paciente o para plantear políticas públicas. Los análisis deben ser leídos siempre de forma escéptica, considerando la probabilidad de falsos positivos. La mejor forma

de comprobar si hay un efecto en subgrupos es que dicho efecto se pueda replicar en otros estudios, con otras muestras y utilizando otras técnicas (Rothwell 2005). Por lo tanto, es importante realizar la mayor cantidad de análisis de subgrupos para mejorar los métodos y así la credibilidad de los efectos en subgrupos.

Capítulo 2

Preliminares

En este capítulo se presenta una breve introducción a los fundamentos de la Estadística, haciendo énfasis en la Estadística Bayesiana, y a los fundamentos del Aprendizaje de Máquina con el propósito de facilitar la comprensión de métodos descritos en los siguientes capítulos.

2.1. Fundamentos Estadísticos

Una de las formas de probar la validez de una hipótesis es a través de experimentos científicos que se evalúan con técnicas estadísticas. Dentro de la Estadística hay dos enfoques para hacer inferencia: la Estadística Frecuentista y la Estadística Bayesiana. Durante muchos años, la Estadística Frecuentista fue más utilizada por los científicos y el público general para hacer inferencia. Sin embargo, a finales del siglo XX surgieron avances en la Estadística Bayesiana y en la computación que

lograron resolver problemas que impedían que se pudiera poner en práctica con facilidad (Lynch 2007). Hoy en día se utilizan ambos enfoques dependiendo del tema de las hipótesis a evaluar, así como de la preferencia del investigador que realiza los análisis.

La principal diferencia entre los dos enfoques de la Estadística es que el enfoque Frecuentista se basa en asociar la probabilidad a la frecuencia a largo plazo, mientras que el enfoque Bayesiano interpreta la probabilidad como una medida subjetiva de incertidumbre (van de Schoot et al. 2014 citando a de Finetti 1974). Suponiendo que se puede hacer un experimento infinitas veces, la frecuencia (el número de veces relativo al total) que ocurre cada evento es equivalente a la probabilidad en el primer enfoque. De ahí el nombre de la Estadística Frecuentista.

En cambio, en la Estadística Bayesiana la probabilidad de que ocurra un evento depende de la información que tiene cada persona antes de empezar el experimento. Una vez que el experimento ha tenido lugar, se adquiere más información y la probabilidad se actualiza. Entonces, la probabilidad depende de la incertidumbre que le asigna el investigador a la ocurrencia de un evento antes de realizar un experimento y la forma en que actualiza la incertidumbre con la nueva información. El enfoque se llama Bayesiano ya que la forma de actualizarla es a través del teorema de Bayes (van de Schoot et al. 2014).

En este trabajo se utilizan métodos para hacer análisis de subgrupos basados en ambos enfoques. Sin embargo, la Estadística Frecuentista y sus conceptos son ampliamente conocidos, por lo que no se describen en esta sección. Por lo tanto, en los siguientes párrafos se comienza por describir técnicas comunes a los dos enfoques, seguidas por los fundamentos de la Estadística Bayesiana, así como conceptos particulares de ella.

2.1.1. Aleatorización

Cuando se lleva a cabo un ensayo clínico, se reclutan a personas que cumplen ciertos criterios. Ejemplos de criterios son que los pacientes tengan la enfermedad que se busca tratar, que estén en un rango de edad y que puedan recibir el tratamiento en el hospital donde se realiza el ensayo clínico. A estos pacientes, se les asigna un grupo: de control o de tratamiento. El grupo de control puede recibir diferentes tratamientos. Ejemplos de tratamientos son un placebo, una menor dosis o una medicina anteriormente probada y se utilizan dependiendo de los objetivos del ensayo clínico. Al proceso de asignar un grupo a cada paciente se le llama aleatorización, por hacerse de forma aleatoria. Esto es, ni los pacientes ni los doctores escogen el grupo asignado.

La aleatorización asegura que los grupos sean comparables y provee una base para las densidades con las que se harán análisis estadísticos (Spiegelhalter, Abrams y Myles 2004). Por ejemplo, si se busca comparar el efecto de un tratamiento en subgrupos definidos por el sexo, la aleatorización se puede hacer de tal forma que haya el mismo número de mujeres que de hombres en los grupos de control y de tratamiento.

2.1.2. Urna de Pòlya

La distribución de Urna de Pòlya es una distribución no paramétrica que consiste en tomar pelotas de una urna con varios colores (Mahmoud 2009). La variable aleatoria X_n es el color observado de tomar una pelota de la urna en el paso n . Para el paso $n+1$, se agrega una pelota del mismo color que X_n . Se presenta un efecto en el que “los ricos se vuelven más ricos”. Cada vez que se toma una pelota de un color, aumenta la probabilidad de

que vuelve a salir ese color en pasos futuros.

La figura 2.1 representa la acción de tomar una pelota azul en el paso n de una urna que contiene una pelota de color azul, una de color naranja y una de color verde. La variable aleatoria X_n toma el valor azul. Para el siguiente paso $n + 1$, se regresa la pelota tomada en el paso n y se agrega una pelota azul adicional, que cambiará la probabilidad de tomar una pelota de cada color en el paso $n + 1$.

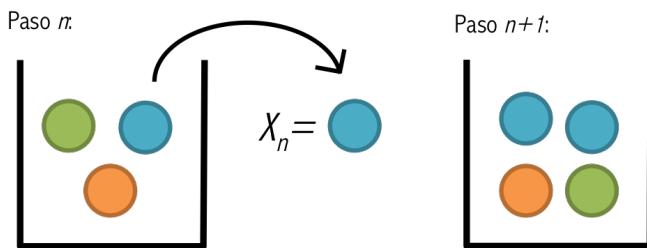


Figura 2.1: Urna de Pòlya: se ejemplifica la acción de tomar una pelota de color azul en el paso n , regresarla a la urna y agregar una pelota adicional de color azul para el paso $n + 1$.

Urna de Pòlya enriquecida por cero

La Urna de Pòlya enriquecida por cero es una versión de la Urna de Pòlya utilizada por Laud et al. (2013) donde se agrega la probabilidad de que una observación sea cero. En este caso, no siempre se toma una pelota de la urna. La versión de la urna funciona de la siguiente forma:

En dado paso n , la urna tiene $m \leq n$ pelotas de $k \leq m$ colores diferentes. Puede ocurrir una de las tres siguientes opciones, cada una con diferentes probabilidades:

1. No tomar ninguna pelota. $X_n = 0$. Este evento ocurre con probabilidad p .
2. Tomar una pelota de la urna. A X_n se le asigna el valor del color de la pelota que se tomó. Se regresa la pelota a la urna y se agrega una pelota adicional del mismo color. Este evento ocurre con probabilidad $(1 - p)(\frac{n}{\varphi + n})$.
3. Agregar una nueva pelota a la urna. No se toma ninguna pelota y a X_n se le asigna el valor de un nuevo color. Se agrega la pelota nueva a la urna, ahora habrán $k + 1$ colores en la urna. Este evento ocurre con probabilidad $(1 - p)(\frac{\varphi}{\varphi + n})$.

Los valores de p y φ se establecen dependiendo del objetivo del estudio. A mayor p , mayor es la probabilidad de que no se tome una pelota. A mayor φ , mayor es la probabilidad de que se agreguen pelotas de colores nuevos a la urna en vez de tomar una ya existente. Al parámetro φ se le llama parámetro de masa total.

Las tres opciones se ilustran en la figura 2.2:

Paso n :

Con probabilidad p , $X_n=0$

Con probabilidad $(1-p)$ se saca una bola de la urna

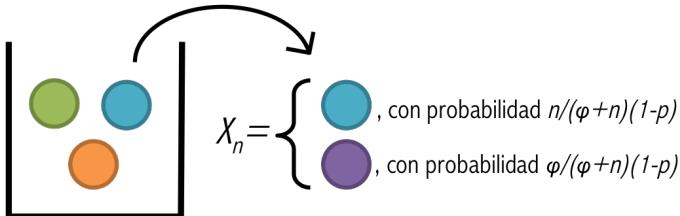


Figura 2.2: Urna de Pòlya enriquecida por cero: se ejemplifican las diferentes opciones, con sus respectivas probabilidades en el paso n : no tomar una pelota, tomar una pelota de un color existente, agregar una pelota de un color nuevo.

2.2. Estadística Bayesiana

La Estadística Bayesiana busca representar incertidumbre inicial sobre un evento. El evento (o serie de eventos) se puede describir a través de un modelo, el cual puede ser definido por parámetros. Los parámetros a su vez se pueden definir a través de una distribución de probabilidad. Las muestras aleatorias son datos que sirven para actualizar la distribución de probabilidad inicial, buscando eliminar la incertidumbre. A través de la actualización se obtiene una distribución final, con la cual se puede hacer inferencia sobre los parámetros del modelo con menos incertidumbre (Lynch 2007).

La Estadística Bayesiana recibe su nombre ya que utiliza el teorema de Bayes para hacer inferencia sobre parámetros de interés. Recordando el

teorema de Bayes:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \quad P(B) > 0.$$

Esto es, la probabilidad de que ocurra el evento B dado que ocurrió el evento A es igual a la probabilidad de que A dado B por la probabilidad de A entre la probabilidad de B .

Este teorema aplicado a densidades de probabilidad da el siguiente resultado:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)},$$

donde θ es el parámetro de interés sobre el cual se hace inferencia y $X = (X_1, X_2, \dots, X_n)$ es la información que proviene de una muestra. En otras palabras, la probabilidad del parámetro θ dado que se obtuvo la muestra X es igual la probabilidad de obtener la muestra X dado θ por la probabilidad marginal de θ entre la probabilidad marginal de la muestra. Notar que $f(X|\theta)$ se puede expresar como la verosimilitud de la muestra y $f(\theta)$ es una distribución que se establece de manera subjetiva y representa la incertidumbre inicial o a priori. Entonces, la probabilidad final, también llamada posterior, de θ bajo X es:

$$\text{Posterior} \propto \text{Verosimilitud} \times \text{Inicial},$$

donde el símbolo \propto se utiliza para denotar proporcionalidad.

Dicho esto, es importante mencionar que los fundamentos teóricos de la Estadística Bayesiana provienen de la teoría de la decisión, que como su nombre lo dice, estudia el problema de tomar decisiones. En particular, se toman decisiones sobre el valor del parámetro desconocido θ , que pertenece al espacio paramétrico Θ . Una decisión se denota por \mathbf{d}

y todas las decisiones conforman el espacio \mathcal{D} (la decisión también se puede llamar acción y se denota por \mathbf{a} que pertenece al espacio de acciones \mathcal{A}) (Berger 1980).

Una forma de tomar una decisión es asignando una función de pérdida (o utilidad) al espacio de decisiones \mathcal{D} . Esta es una función que asigna una pérdida por tomar la decisión $\mathbf{d}_i \in \mathcal{D}$ cuando el valor real del parámetro es $\theta_j \in \Theta$. Se supone que las funciones de pérdida $L(\mathbf{d}, \theta)$ están definidas para toda $(\mathbf{d}, \theta) \in \mathcal{D} \times \Theta$ y que satisfacen $L(\mathbf{d}, \theta) \geq -K > -\infty$, $K \in \mathbb{R}^+$.

Se define la pérdida esperada para la función $L(\mathbf{d}, \theta)$ como

$$\rho(\mathbf{d}) = E[L(\mathbf{d}, \theta)] = \int_{\Theta} L(\mathbf{d}, \theta)p(\theta)d\theta.$$

Entonces, a la decisión $\mathbf{d}^* \in \mathcal{D}$ que minimiza $\rho(\mathbf{d})$ se le conoce como la decisión de Bayes (Berger 1980).

Los problemas de inferencia se pueden resolver como problemas de decisión en un contexto Bayesiano. A continuación, se exponen algunos problemas básicos de inferencia resueltos desde un enfoque Bayesiano.

2.2.1. Estimación puntual

La medida que comúnmente se utiliza para estimar un parámetro θ a través de la distribución posterior $f(\theta|X)$ es la media posterior. Esto se debe a que utilizando la función de pérdida cuadrática $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, el valor que minimiza la esperanza de esta función es $E[\theta|X]$. Otras medidas para estimar al parámetro θ son la moda posterior (el valor $\hat{\theta}$ que maximiza $f(\theta|X)$) y la mediana posterior (el valor $\hat{\theta}$ para el cual $P(\theta \leq \hat{\theta}|X) = 0.5$).

2.2.2. Conjuntos creíbles

A diferencia de la Estadística Frecuentista, en la Estadística Bayesiana la estimación por intervalos no se hace a través intervalos de confianza. En el enfoque Bayesiano se utiliza el concepto de conjuntos creíbles, los cuales se definen de la siguiente forma (Berger 1980):

Un conjunto $100(1 - \alpha)\%$ creíble para θ es un subconjunto $C \subset \Theta$ que cumple $1 - \alpha \leq P(\theta \leq C|x)$, $\alpha \in \mathbb{R} \cap (0, 1)$. Dado que $1 - \alpha$ determina la probabilidad que se acumula en C , se le conoce como nivel de credibilidad.

En particular a los conjuntos $C^* \subset \Theta$ que cumplen $C^* = \{\theta \in \Theta : p(\theta|x) \geq k(\alpha)\}$, donde $k(\alpha)$ es la mayor constante tal que $P(\theta \leq C|x) \geq 1 - \alpha$, se les llama conjuntos creíbles de máxima densidad posterior. Para estos, C^* es la menor región de θ donde se acumula al menos una probabilidad de $100(1 - \alpha)\%$.

2.2.3. Pruebas de hipótesis

Supongamos que se busca comparar las hipótesis $H_0 : \theta \in \theta_0$ y $H_1 : \theta \in \theta_1$. Entonces, se rechaza H_0

$$\begin{aligned} &\iff E[L(\theta_1, \theta|X)] > E[L(\theta_0, \theta|X)] \\ &\iff p(\theta_0|X)L(\theta_1, \theta_0|X) + p(\theta_1|X)L(\theta_1, \theta_1|X) \\ &\quad > p(\theta_0|X)L(\theta_0, \theta_0|X) + p(\theta_1|X)L(\theta_0, \theta_1|X) \\ &\iff p(X|\theta_0)p(\theta_0)L(\theta_1, \theta_0|X) + p(X|\theta_1)p(\theta_1)L(\theta_1, \theta_1|X) \\ &\quad > p(X|\theta_0)p(\theta_0)L(\theta_0, \theta_0|X) + p(X|\theta_1)p(\theta_1)L(\theta_0, \theta_1|X) \\ &\iff p(X|\theta_1)p(\theta_1)(L(\theta_1, \theta_1|X) - L(\theta_0, \theta_1|X)) \\ &\quad > p(X|\theta_0)p(\theta_0)(L(\theta_0, \theta_0|X) - L(\theta_1, \theta_0|X)) \end{aligned}$$

$$\begin{aligned} &\iff \left(\frac{L(\theta_1, \theta_1|X) - L(\theta_0, \theta_1|X)}{L(\theta_0, \theta_0|X) - L(\theta_1, \theta_0|X)} \right) \frac{p(\theta_1)}{p(\theta_0)} > \frac{p(X|\theta_0)}{p(X|\theta_1)} \\ &\iff k > \frac{p(X|\theta_0)}{p(X|\theta_1)}. \end{aligned}$$

Por lo tanto, el contraste de hipótesis en el enfoque Bayesiano es una comparación del coeficiente de verosimilitud $\frac{p(X|\theta_0)}{p(X|\theta_1)}$.

2.2.4. Distribución de Jeffreys

La distribución inicial de Jeffreys es una distribución inicial no informativa que es invariante con respecto a transformaciones monótonas (*The Jeffreys Prior* 2018 citando a Jeffreys 1946).

Las distribuciones iniciales no informativas se utilizan para representar la falta de conocimientos del investigador sobre las condiciones iniciales del experimento. Esto es, se busca que de inicio los parámetros tengan una probabilidad similar o que no afecte a las distribuciones finales. En un principio, se propuso utilizar la distribución uniforme como la distribución mínimo informativa de referencia. Sin embargo, si se le aplica una transformación monótona a una variable uniforme, la distribución resultante no es uniforme. No conserva las mismas propiedades (*The Jeffreys Prior* 2018).

Como solución al problema, Jeffreys propuso la siguiente distribución inicial para el modelo $X \sim f(x|\theta)$:

$$p^J(\theta) \propto \sqrt{I^F(\theta)},$$

donde $I^F(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right] = -E \left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]$ es la

información de Fisher. El símbolo \sim se utiliza para denotar que una variable aleatoria sigue cierta distribución de probabilidad.

Supongamos que se tiene el modelo $X \sim f(x|\theta)$ y la reparametrización $X \sim g(x|h(\theta))$ con la transformación monótona $h(\theta) = \eta$. Entonces el principio de invarianza con respecto a transformaciones monótonas se refiere a que se obtiene el mismo resultado al aplicarle la distribución de Jeffreys a la transformación que al aplicarla al modelo original y deducir la transformación de este (*The Jeffreys Prior* 2018). Este proceso se representa en la figura 2.3:

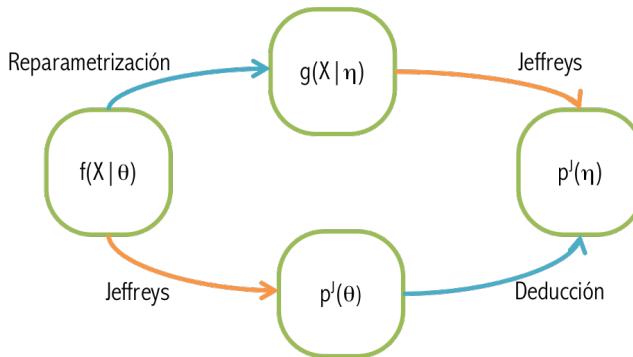


Figura 2.3: Invarianza con respecto a transformaciones monótonas: la figura representa que aplicar procesos diferentes a la misma densidad lleva a los mismos resultados. El primer proceso es una reparametrización, de la cual se obtiene la distribución de Jeffreys. El segundo proceso consta en primero aplicar la distribución de Jeffreys y luego deducir la transformación. Con ambos se llega al mismo resultado.

Si se aplica el principio de Jeffreys a la reparametrización de $f(x)$, $g(x|h(\theta))$, se obtiene $p^J(h(\theta)) = p^J(\eta) \propto \sqrt{IF(\eta)}$, el cual también se puede deducir al obtener $p_\eta(\eta)$ como función de θ . En este caso, $p_\eta(\eta)$ es la

densidad de probabilidad que evalúa η bajo el espacio de η y $p_\theta(h^{-1}(\eta))$ es la densidad de probabilidad que evalúa $h^{-1}(\eta)$ bajo el espacio de θ :

$$\begin{aligned} p_\eta(\eta) &= p_\theta(h^{-1}(\eta)) \left| \frac{\partial}{\partial \eta} h^{-1}(\eta) \right| \\ &= p(\theta) \left| \frac{\partial \theta}{\partial \eta} \right| \\ &\propto \sqrt{I^F(\theta)} \left| \frac{\partial \theta}{\partial \eta} \right| \\ &= \sqrt{I^F(\theta) \left(\frac{\partial \theta}{\partial \eta} \right)^2} \\ &= \sqrt{E \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right] \left(\frac{\partial \theta}{\partial \eta} \right)^2} \\ &= \sqrt{\int \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 f(x|\theta) dx \left(\frac{\partial \theta}{\partial \eta} \right)^2} \\ &= \sqrt{\int \left(\frac{\partial}{\partial \theta} \log g(x|h(\theta)) \right)^2 g(x|h(\theta)) dx \left(\frac{\partial \theta}{\partial \eta} \right)^2} \\ &= \sqrt{\int \left(\frac{\partial \log g(x|\eta)}{\partial \eta} \frac{\partial \eta}{\partial \theta} \right)^2 g(x|\eta) dx \left(\frac{\partial \theta}{\partial \eta} \right)^2} \text{ (regla de la cadena)} \\ &= \sqrt{\int \left(\frac{\partial \log g(x|\eta)}{\partial \eta} \right)^2 \left(\frac{\partial \eta}{\partial \theta} \right)^2 \left(\frac{\partial \theta}{\partial \eta} \right)^2 g(x|\eta) dx} \\ &= \sqrt{\int \left(\frac{\partial \log g(x|\eta)}{\partial \eta} \right)^2 g(x|\eta) dx} \\ &= \sqrt{E \left[\left(\frac{\partial}{\partial \eta} \log g(x|\eta) \right)^2 \right]} \\ &= \sqrt{I^F(\eta)} \\ &\propto p^J(\eta) \end{aligned}$$

Por lo tanto, $p^J(\eta)$ se puede deducir de $p_\theta(\theta)$ si $h(\theta) = \eta$ es una transformación monótona.

2.2.5. Distribuciones *g-prior*

Previo exponer las distribuciones *g-prior*, es conveniente establecer la notación de algunos conceptos. Se utiliza $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$, para denotar la distribución de probabilidad normal con media μ y varianza σ^2 , donde \mathbb{R}^+ es el conjunto de números reales positivos. Para el caso de la distribución normal multivariada, se utiliza $N(\mu, \Sigma)$, con vector de medias μ y matriz de covarianzas Σ y se especifica la dimensión del vector y la matriz. Por ejemplo, si $\mu \in \mathbb{R}^n$ y $\Sigma \in \mathbb{R}^{n \times n}$, la variable que siga la distribución $N(\mu, \Sigma)$ será un vector de dimensión n .

Las distribuciones *g-prior* fueron propuestas por Zellner (1986) y sirven para establecer las distribuciones iniciales para los coeficientes de un modelo de regresión lineal múltiple normal. Considerar el modelo:

$$\begin{aligned}y &= X\beta + \varepsilon, \\y &\in \mathbb{R}^n, X \in \mathbb{R}^{n \times k}, \text{ rango}(X) = k, \beta \in \mathbb{R}^k, \\\varepsilon &\sim N(\mathbf{0}_n, \frac{1}{\phi} I_n),\end{aligned}$$

donde I_n es la matriz identidad de tamaño n , $\mathbf{0}_n$ es el vector de tamaño n cuyos elementos son 0 (de aquí en adelante se utilizará esa notación), y y X son conocidos y β y ϕ son desconocidos. Entonces, la distribución inicial para β y ϕ es (Liang et al. 2008):

$$\begin{aligned}p(\phi) &\propto \frac{1}{\phi}, \\\beta | \phi &\sim N(\beta_a, \frac{g}{\phi} (X^T X)^{-1}),\end{aligned}$$

donde $\beta_a = E[\beta|D_0] = \hat{\beta}_0 = (X^T X)^{-1} X^T y_0$, para la muestra imaginaria y_0 que se supone se genera por el modelo

$$\begin{aligned} D_0 : y_0 &= X\beta + \varepsilon_0, \\ \varepsilon_0 &\sim N(\mathbf{0}_n, \sigma_0^2 I_n). \end{aligned}$$

La muestra es imaginaria porque realmente nunca se genera, sólo se utiliza como fundamento para obtener las distribuciones de β y ϕ . En realidad, se asigna un valor inicial a $\hat{\beta}_0$ la media a priori con base en los conocimientos del investigador. La matriz de covarianzas inicial es un múltiplo escalar de la matriz de covarianzas del estimador de máxima verosimilitud (Li y Clyde 2018). La escala proviene de que como la muestra es imaginaria, se puede tomar el parámetro $\sigma_0^2 = g/\phi$, con $g \in (0, \infty)$. De aquí viene el nombre de las distribuciones *g-prior*. El parámetro g ayuda a controlar el encogimiento hacia la media inicial y hay muchos métodos para seleccionar el valor de g (Li y Clyde 2018).

2.2.6. Mezclas de *g-priors*

Basándose en la idea de Zellner (1986), ha habido varias adaptaciones de las distribuciones *g-prior*. Por ejemplo, para la selección de modelos Bayesiana, Liang et al. (2008) proponen comparar los modelos \mathcal{M}_γ :

$$\begin{aligned} y &\sim N(\mu, \frac{1}{\phi} I_n), \\ \mathcal{M}_\gamma : \mu &= \alpha \mathbf{1}_n + X_\gamma \beta_\gamma, \\ y \in \mathbb{R}^n, \alpha \in \mathbb{R}^n, X &\in \mathbb{R}^{n \times k}, \text{ rango}(X) = k, \beta \in \mathbb{R}^{k \times 1}, \end{aligned}$$

donde $\mathbf{1}_n$ es el vector de tamaño n cuyos elementos son 1, y y X son conocidas y α , β y ϕ son desconocidas. Se utiliza α para continuar con la notación de Zellner, pero cabe mencionar que no es el α de los intervalos

de confianza o los intervalos creíbles. Para todos los modelos α es común, entonces las hipótesis a comparar son:

$$H_0 : \beta_\gamma = \mathbf{0}_k, \text{ contra}$$

$$H_1 : \beta_\gamma \in \mathbb{R}^k.$$

Con algunos supuestos, Liang et al. (2008) proponen utilizar las siguientes distribuciones iniciales para los parámetros desconocidos bajo el modelo \mathcal{M}_γ :

$$p(\alpha, \phi) = p(\alpha)p(\phi), \quad (2.1)$$

$$p(\alpha) \propto 1, \quad (2.2)$$

$$p(\phi) \propto \frac{1}{\phi}, \quad (2.3)$$

$$\beta_\gamma | \phi, \mathcal{M}_\gamma \sim N(\mathbf{0}_n, \frac{g}{\phi}(X^T X)^{-1}). \quad (2.4)$$

El parámetro g nuevamente aparece en la matriz de covarianzas de la distribución inicial de β . Al asignar una distribución inicial al parámetro g en (2.4), se obtiene una mezcla de g -*priors*. Las mezclas de g -*priors* son atractivas porque resuelven varios problemas que resultan de asignar un g fija, tienen propiedades teóricas atractivas y son más eficientes que la g -*prior* original (Li y Clyde 2018).

Hyper-gs

Una familia de mezclas de g -*priors* propuesta por Liang et al. (2008) es aquella con la siguiente probabilidad:

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}, \quad (2.5)$$

$$g > 0, \quad (2.6)$$

$$a > 2, (\text{si } a \leq 2, \text{ la inicial } p(g) \text{ es impropia}). \quad (2.7)$$

La familia anterior se basa en la familia de *hyper-gs* también propuesta por Liang et al. (2008). En las *hyper-gs*, se le asigna una distribución Beta($1, a/2 - 1$) al parámetro $g/(1 + g)$. Para esta familia si $a \in \mathbb{Z}$, la distribución de $g/(1 + g)$ es muy similar a la ecuación (2.5):

$$\begin{aligned} \text{Si } \frac{g}{1+g} &\sim \text{Beta}\left(1, \frac{a}{2} - 1\right) \\ \implies p\left(\frac{g}{1+g}\right) &= \frac{\Gamma(\frac{a}{2})}{\Gamma(1)\Gamma(\frac{a}{2}-1)} \left(\frac{g}{1+g}\right)^{1-1} \left(1 - \frac{g}{1+g}\right)^{a/2-1-1} \\ &= \frac{\left(\frac{a}{2}-1\right)!}{\left(\frac{a}{2}-2\right)!} \left(\frac{1+g-g}{1+g}\right)^{a/2-2} \\ &= \left(\frac{a}{2}-1\right) \left(\frac{1}{1+g}\right)^{a/2-2} \\ &= \frac{a-2}{2} (1+g)^{-a/2+2}. \end{aligned}$$

Algunas de las propiedades de la familia de *hyper-gs* son las siguientes (Li y Clyde 2018):

- La distribución final de g dado un modelo es cerrada.
- Si $a = 4$, entonces $\frac{g}{1+g}$ es uniforme.
- Si $a = 4$, entonces la distribución de g corresponde a la inicial de Jeffreys para g .

2.3. Simulación Monte Carlo de Cadenas de Markov

Para algunos modelos, es difícil calcular las integrales que definen la distribución posterior. Una solución a esto son los métodos de muestreo, que simulan una muestra de la distribución de interés a la cual se le aplican fórmulas discretas para resumir la información de los parámetros y aproximar las integrales (Lynch 2007).

Uno de los métodos más conocidos de muestreo son los métodos de Monte Carlo de Cadenas de Markov (MCMC). En estos métodos, se simula un proceso de manera aleatoria (Monte Carlo) y se obtiene una muestra de la distribución final dado la muestra anterior. La repetición de este proceso produce una cadena de Markov de muestras de la distribución posterior (Lynch 2007).

2.3.1. Muestreo de Gibbs

El muestreo de Gibbs es un método de Monte Carlo de Cadenas de Markov y es uno de los métodos de muestreo más utilizados en análisis Bayesiano. El muestreo de Gibbs sigue el siguiente algoritmo (Lynch 2007):

- Asignar valores iniciales a los parámetros de los cuáles se quiere obtener una muestra $\theta^j = (\theta_1^j, \dots, \theta_m^j)$, con $j = 0$.
- Para $j = 1, \dots, n$:
 - Obtener una muestra de $(\theta_1^j | \theta_2^{j-1}, \theta_3^{j-1}, \dots, \theta_m^{j-1})$.

- Obtener una muestra de $(\theta_2^j | \theta_1^{j-1}, \theta_3^{j-1}, \dots, \theta_m^{j-1})$.
- ⋮
- Obtener una muestra de $(\theta_m^j | \theta_1^{j-1}, \theta_2^{j-1}, \dots, \theta_{m-1}^{j-1})$.
- $j = j + 1$

Entonces, en cada iteración se obtiene una muestra aleatoria de una distribución que depende de la muestra anterior. De este modo, las distribuciones se van actualizando con los valores que se obtienen en la muestra cada iteración. En el muestreo de Gibbs, a un nuevo valor de un parámetro se le llama un valor actualizado.

Algunos de los softwares que utilizan el muestreo de Gibbs para hacer inferencia Bayesiana son *BUGS* (*OpenBUGS* 2018) y *JAGS* (Plummer 2013). En estos métodos, el usuario describe las relaciones entre los parámetros de interés a través de un modelo. En particular, en esta tesis se utilizará *JAGS* por su buen funcionamiento en varios sistemas operativos y su código abierto.

2.4. Aprendizaje de Máquina

El Aprendizaje de Máquina se refiere a las técnicas utilizadas para extraer información y patrones de un conjunto de datos (Hastie, Tibshirani y Friedman 2009). Una máquina aprende de una experiencia si mejora su desempeño con respecto a cierta tarea al procesar la información de la experiencia (Mitchell 1997). En general, hay dos tipos de aprendizajes que puede realizar una máquina: supervisado o no supervisado.

El aprendizaje supervisado busca predecir un valor de salida al asignar una función a una serie de parámetros iniciales. Si el valor que regresa la función es numérico se le conoce al algoritmo como de regresión y si es categórico se le conoce como de clasificación. Se le llama supervisado porque los datos con los que entrena la máquina tienen valores de salida conocidos. Entonces, se puede evaluar el desempeño de la función al comparar los valores de salida predichos con los reales y se pueden modificar parámetros de la función para mejorar las predicciones.

En cambio, con el aprendizaje no supervisado no se conocen los valores de salida para los datos, y por lo tanto no se puede saber si la función realiza predicciones acertadas. Su objetivo principal entonces no es hacer predicciones, sino describir atributos de los datos y agruparlos.

Las bases de datos de ensayos clínicos normalmente están compuestas por valores de entrada (covariables), el tratamiento que se administra y un valor de salida que puede ser el tiempo de supervivencia o una variable indicadora que describe la eficacia del tratamiento para una serie de pacientes. Por lo tanto, los métodos que se utilizan en esta tesis son de aprendizaje supervisado.

2.4.1. Árboles de regresión y clasificación CART

Los árboles CART por sus siglas en inglés (*Classification And Regression Trees*) fueron creados por Breiman et al. (1984) y dividen un conjunto de datos en particiones y le asignan a cada partición una función. A grandes rasgos el algoritmo que sigue CART es el siguiente:

1. Crecer un árbol: se establece una regla para la división de los nodos.

Cada nodo se divide hasta cumplir un criterio. Se obtiene un árbol inicial grande T_0 .

2. Podar el árbol: se refiere a eliminar las ramas de un árbol. Típicamente se poda el árbol T_0 de abajo hacia arriba progresivamente. Esto es, se van eliminando ramas iterativamente hasta llegar al árbol nulo o raíz. Cada iteración produce un subárbol de T_0 , por lo que se obtiene una sucesión de subárboles.
3. Escoger el árbol de mejor tamaño: se establece una regla para escoger un árbol de la sucesión de subárboles de T_0 .

Una vez que se obtiene el árbol, se le asigna una función a cada nodo terminal. Si la función regresa valores continuos, el árbol es de regresión y si asigna valores discretos es de clasificación.

2.4.2. Bosques aleatorios

Otro algoritmo creado por Breiman son los bosques aleatorios. Los bosques aleatorios consisten de un ensamble o conjunto de árboles de decisión que se definen de la siguiente forma (Breiman 2001):

Un árbol de decisión es un clasificador (una regresión) que consiste de una colección de árboles de clasificación (regresión) $h(X, \Theta_k)$, $k = 1, \dots$, donde Θ_k es una serie de vectores aleatorios $(\theta_{k1}, \dots, \theta_{kn_k})$ independiente e idénticamente distribuidos y cada árbol emite un voto para elegir la clase (valor) más popular para los valores de entrada X .

En este caso, un vector aleatorio Θ_k es una muestra aleatoria con reemplazo T_k del conjunto de observaciones T formado por el vector de

respuestas y para las covariables X . Para crecer cada árbol $h(X, \Theta_k)$, $k = 1, \dots$, se utiliza la muestra T_k y una serie de covariables $X_k \in X$ elegidas al azar. Los árboles se crecen siguiendo el algoritmo de CART (Breiman et al. 1984) pero sólo se utilizan las covariables X_k (no todas) para determinar las divisiones del árbol. En cada nodo, la división se hace con la covariable de esa muestra que maximice la regla de división de los nodos. El árbol se crece hasta que haya una observación en cada nodo (los nodos no se puedan dividir más) y no se poda.

Al final, se obtiene un conjunto de árboles para los cuales se clasifican los valores de entrada X . En bosques de clasificación se toma la moda de los valores obtenidos al clasificar X con los árboles aleatorios y en bosques de regresión se utiliza la media de las regresiones para X .

2.4.3. Validación cruzada

La validación cruzada es una técnica de Aprendizaje de Máquina que se utiliza para eliminar sesgo en muestras a las cuales se les hace análisis. Consiste en dividir la muestra en diferentes subconjuntos y utilizar uno como conjunto de prueba y los demás como conjuntos de entrenamiento de manera iterativa para el algoritmo deseado. Al terminar las iteraciones, todos los subconjuntos habrán sido utilizados como de prueba una vez y como de entrenamiento $w - 1$ veces si la muestra se dividió en w subconjuntos (Hastie, Tibshirani y Friedman 2009).

2.4.4. *Bootstrap*

Consiste en tomar múltiples muestras con reemplazo de los datos con el propósito de obtener muestras independientes. Por ejemplo, si la muestra a evaluar es de tamaño N , se pueden producir un número m de muestras nuevas de tamaño N a partir de la original, logrando mejorar los análisis para algunos métodos de Aprendizaje de Máquina (Hastie, Tibshirani y Friedman 2009).

Capítulo 3

Métodos de análisis de subgrupos exploratorios

3.1. Método *Virtual Twins*

El método *Virtual Twins* de Foster et al. (2011) busca definir un subgrupo a través de una región A en el espacio de covariables donde el tratamiento es mejor substancialmente que para el promedio o para un umbral. La región A depende de un número pequeño de variables y es estimada por el método. Es atractivo porque sus resultados son fáciles de entender y es eficiente computacionalmente. Tiene un enfoque Frecuentista y sirve para evaluar tratamientos con mejora general no lo suficientemente grande para ser adoptados y tratamientos con aparente mejora general, pero mejora substancial para un subgrupo.

Los análisis que realiza *Virtual Twins* para estimar la región A son un

bosque aleatorio para estimar la diferencia general en el efecto del tratamiento Z y un árbol aleatorio para encontrar un número pequeño de covariables X que estén fuertemente asociadas a Z .

Supongamos que un paciente puede sólo tener un resultado positivo o negativo a tratamiento. Si la variable de salida no es binaria, se dicotomiza de tal forma que $Y_i \in \{0, 1\}$ para cada paciente $i = 1, 2, \dots, n$. Además, para cada paciente se tiene un vector de covariables X_i y una variable indicadora T_i que señala si el paciente pertenece al grupo control ($T_i = 0$) o si está recibiendo el tratamiento ($T_i = 1$). Si el paciente i pertenece al grupo control, se observa $Y_i|T_i = 0, X_i$ y si pertenece al grupo tratado, se observa $Y_i|T_i = 1, X_i$. Notar que sólo se puede observar uno de los resultados. El método se basa en crear un individuo gemelo para estimar el resultado que falta, de aquí es de donde recibe su nombre.

El algoritmo que aplica el método es el siguiente:

1. Ajustar un bosque aleatorio a los datos. El bosque utiliza los valores de $(Y_i, T_i, X_i, X_i I(T_i = 0), X_i I(T_i = 1))$ como datos de aprendizaje y crea una predicción de caja negra donde entran nuevos valores $(T_j, X_j, X_j I(T_j = 0), X_j I(T_j = 1))$ y sale un estimador de $P(Y_j = 1)$. La función $I(E)$ es la función indicadora, la cual toma el valor 1 si se cumple el evento E y 0 si no se cumple el evento E . Foster et al. mencionan que agregar los términos $X_i I(T_i = 0)$ y $X_i I(T_i = 1)$ al bosque aleatorio mejora el desempeño del método. A pesar de que no se da una explicación concreta de la razón de la mejora, se expone una posible explicación.

Empecemos analizando los valores que toman $X_i I(T_i = 0)$ y

$X_i I(T_i = 1)$:

$$X_i I(T_i = 0) = \begin{cases} X_i, & T_i = 0 \\ 0, & T_i = 1 \end{cases}$$

$$X_i I(T_i = 1) = \begin{cases} 0, & T_i = 0 \\ X_i, & T_i = 1 \end{cases}$$

Si se hace un estudio con los individuos $\{1, 3\}$ en el grupo tratamiento y $\{2\}$ en el grupo control, por ejemplo, entonces se obtienen las siguientes matrices:

$$\left[\begin{array}{c|c|c} X_1 & 0 & X_3 \end{array} \right] \text{ y } \left[\begin{array}{c|c|c} 0 & X_2 & 0 \end{array} \right],$$

donde X_i son los vectores de covariables de los pacientes.

Por lo tanto, se obtiene una matriz con únicamente las covariables de los pacientes en el grupo de tratamiento y una matriz con únicamente las covariables de los pacientes del grupo control. Es probable que, al utilizar estas matrices en el algoritmo, el efecto de cada grupo se amplifique y sean más fáciles de diferenciar.

El primer paso de este algoritmo crea la caja negra que se utilizará para estimar diferentes probabilidades. La figura 3.1 representa una caja negra en la que entran las covariables, los tratamientos y la multiplicación de las covariables con las indicadoras y se obtiene la estimación de la probabilidad de obtener una respuesta favorable para cada individuo. Notar que los resultados verdaderos Y_i para los pacientes ya los tenemos, pero lo que se busca son las probabilidades $P(Y_i = 1)$ dependiendo de los valores de las covariables.

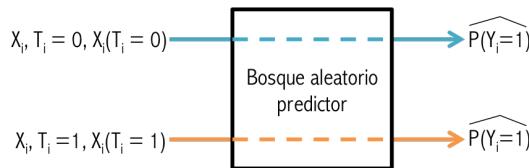


Figura 3.1: Bosque aleatorio para predecir probabilidades.

2. Para cada paciente con tratamiento, estimar $\hat{P}_{1i} = P(Y_i = 1|T_i = 1, X_i)$ y para cada paciente con control, estimar $\hat{P}_{0i} = P(Y_i = 1|T_i = 0, X_i)$ con el bosque aleatorio.
3. Invertir los tratamientos de los pacientes. Si el paciente está en el grupo control se tomará como tratado y viceversa. Es decir, si el tratamiento verdadero para el sujeto i es el tratamiento k , se toma momentáneamente $T_i = 1 - k$, donde $k \in \{0, 1\}$.
4. Usar el bosque aleatorio para predecir el efecto de los pacientes con el tipo de tratamiento invertido y estimar la probabilidad $\hat{P}_{(1-k)i}$, $k \in \{0, 1\}$ que falta para cada individuo. A esta se le llama probabilidad gemela. Los últimos dos pasos se muestran ilustrados en la figura 3.2:

Intercambiar T_i a los pacientes:



Figura 3.2: Bosques aleatorio original con probabilidades gemelas.

5. Definir $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$, el estimador del efecto del tratamiento para cada paciente.

6. Obtener covariables de X fuertemente asociadas a Z que definan una región \hat{A} . Esto se puede ser con dos métodos:

- Árbol de regresión VT(R): Ajustar un árbol de regresión para predecir Z como respuesta de X . Si las \hat{Z}_i predicciones de Z_i son mayores que un valor de umbral c , el paciente i está en \hat{A} . Es decir, si $Z_i > c \implies X_i \in \hat{A}$. Entonces, $\hat{A} = \emptyset$ si $\forall i = 1, \dots, n, Z_i \leq c$
- Árbol de clasificación VT(C): Crear la variable

$$Z_i^* = \begin{cases} 1, & Z_i > c \\ 0, & Z_i \leq c. \end{cases}$$

Ajustar un árbol de clasificación a Z^* como respuesta a X . Clasificar individuos que pertenezcan a \hat{A} . Entonces, $\hat{A} = \emptyset$ si el árbol de clasificación no tiene ramas (es sólo la raíz).

El tamaño de \hat{A} dependerá del valor que tome c . Foster et al. sugieren tomar $c = \delta + 0.1$ o $c = \delta + 0.05$, donde δ es un estimador del efecto del tratamiento $P(Y_i = 1|T_i = 1) - P(Y_i = 1|T_i = 0)$. La selección de c altera la región \hat{A} ya que para que un individuo esté en \hat{A} , necesita haber suficiente evidencia de que tuvo una respuesta al tratamiento sustancialmente mejor que el promedio. Por lo tanto, c logra que sólo se seleccionen pacientes para los cuales su estimador del efecto sea mayor estrictamente que el estimador del efecto general.

No es clara la razón para escoger los valores 0.1 y 0.05 para agregar a δ al tomar el valor de c . Una explicación puede ser que los valores 0.1 y 0.05 se utilizan para representar que si un paciente está en el grupo de tratamiento al menos debería tener 10% o 5% más probabilidad de obtener un resultado positivo que si está en el grupo control. Por

ejemplo, tal vez si una enfermedad normalmente dura 20 días, con un tratamiento durará como mucho 19 días. Tal vez, se utiliza el 5 % y el 10 % por ser valores comunes de efectividad de tratamientos, sobre los cuales la comunidad médica toma decisiones. No se encontraron estudios que hablaran sobre estos valores, por lo que esta explicación es un mero supuesto. Lo que se sabe es que entre mayor es c , más difícil es que pertenezca una observación a la región \hat{A} .

Una alternativa del método es dividir los datos de acuerdo al tratamiento y ajustar un bosque a cada tratamiento (ilustrado en la figura 3.3). Con ayuda de los dos bosques se predicen \hat{P}_{0i} y \hat{P}_{1i} para todos los pacientes. Asimismo, se pueden utilizar el método original y el alternativo y crear el estimador con una combinación lineal de las probabilidades predichas. Por último, se puede definir Z_i en escala logit dependiendo del contexto de donde provienen los datos, donde $Z_i = \text{logit}(\hat{P}_{1i}) - \text{logit}(\hat{P}_{0i})$ y $\text{logit}(p) = \log \frac{p}{1-p}$. El uso de modificaciones queda al criterio del investigador. La figura 3.3 muestra la alternativa del método:

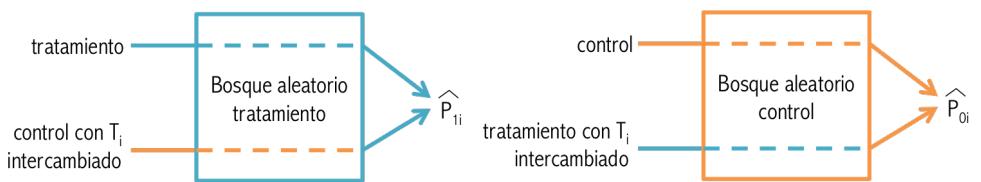


Figura 3.3: *Virtual Twins* alternativo.

3.1.1. Medida de desempeño de la región A

Para evaluar el funcionamiento de *Virtual Twins*, se crea la medida

$$Q(A) = (P(Y = 1|T = 1, X \in A) - P(Y = 1|T = 0, X \in A))$$

$$- (P(Y = 1|T = 1) - P(Y = 1|T = 0)),$$

que es la medida del tratamiento en A contra el tratamiento en general. Entre más grande sea $Q(\hat{A})$, mejor es el tratamiento en \hat{A} en comparación con el tratamiento en general. Se estima $Q(\hat{A})$ y no $Q(A)$ ya que la primera es la medida verdadera del efecto en la región estimada A . En general, $A \neq \hat{A}$ y como $Q(A)$ es la medida sobre la región desconocida A , no la estimamos. Dicho esto, hay varias formas de estimar $Q(\hat{A})$:

- 1. Resubstitución:** estimar las cantidades y utilizarlas como substituto de las cantidades verdaderas en la ecuación. Los mismos datos que se usan para construir la región \hat{A} son los que la evalúan. Los métodos para calcular \hat{A} pueden sobre ajustar los datos, entonces $\hat{Q}(\hat{A})$ puede ser sesgado. El sesgo se da en particular cuando la base de datos consta de pocas observaciones con muchas covariables.

El estimador de resubstitución tiene la forma:

$$\hat{Q}(\hat{A}) = \left[\frac{1}{|\hat{A}_1|} \sum_{\substack{X_i \in \hat{A} \\ T_i=1}} \hat{P}_{1i} - \frac{1}{|\hat{A}_0|} \sum_{\substack{X_i \in \hat{A} \\ T_i=0}} \hat{P}_{0i} \right] - \left[\frac{1}{n_1} \sum_{T_i=1} \hat{P}_{1i} - \frac{1}{n_0} \sum_{T_i=0} \hat{P}_{0i} \right],$$

donde $|\hat{A}_j| = \#X_i \in \hat{A}$ tal que $T_j, j = 0, 1$.

- 2. Simular datos nuevos:** De aplicar *Virtual Twins* se tiene \hat{P}_{1i} y \hat{P}_{0i} . Se simulan observaciones de

$$\begin{aligned} Y_i^* &\sim \text{Bernoulli}(\hat{P}_{1i}) \text{ para } T_i = 1 \text{ y} \\ Y_i^* &\sim \text{Bernoulli}(\hat{P}_{0i}) \text{ para } T_i = 0. \end{aligned}$$

De Y^* , se estiman las cantidades de $Q(\hat{A})$. Como los nuevos datos son independientes de los pasados, $\hat{Q}(\hat{A})$ tendrá menos sesgo aunque es imposible eliminarlo por el sobreajuste presente en \hat{P}_{1i} y \hat{P}_{0i} .

3. Validación cruzada para estimar \hat{P}_{ki} :

- a) Dividir los datos en w subconjuntos del mismo tamaño. Entre más subconjuntos dividan a los datos, menor es el sesgo de las muestras, pero se incrementa el esfuerzo computacional. Dependiendo del tamaño de los datos, si w es muy pequeño o muy grande, se reduce el número posible de diferentes conjuntos. Por ejemplo, si $w = 1$ sólo hay un posible subconjunto, el que contiene a todos los datos. Si w es igual al número de datos, cada subconjunto tendrá a sólo un elemento en él, resultando en conjuntos no funcionales. Se suele tomar $w = 10$ porque los conjuntos resultantes suelen ser funcionales y diferentes si se repite el experimento.
- b) Aplicar el bosque aleatorio a $w - 1$ de w conjuntos de los datos.
- c) Calcular \hat{P}_{1i} y \hat{P}_{0i} para el conjunto de datos que sobra.
- d) Repetir pasos b) y c) alternando todos los conjuntos. Se obtienen \hat{P}_{1i} y $\hat{P}_{0i} \forall i$.
- e) Obtener $\hat{Q}(\hat{A})$ con el método 1 o 2.

4. Validación cruzada completa:

- a) Dividir los datos en w subconjuntos del mismo tamaño.
- b) Definir \hat{A}_k con $w - 1$ de w conjuntos de los datos.
- c) Guardar observaciones del conjunto k que sobra que están en \hat{A}_k .
- d) Repetir los pasos b,c) alternando todos los conjuntos.
- e) $\hat{A} = \bigcup_k \hat{A}_k$.
- f) Con las observaciones guardadas, estimar $P(Y = 1|T = 1, X \in \hat{A})$ y $P(Y = 1|T = 0, X \in \hat{A})$.

g) Obtener $\hat{Q}(\hat{A})$ con el método 1.

5. Bootstrap:

Sea F la distribución real de los datos desconocida \hat{F} la distribución de los datos *bootstrap*, A la región real desconocida, \hat{A}_F la región estimada de datos observados, $\hat{A}_{\hat{F}}$ la región estimada de datos *bootstrap*. Entonces, $Q(\cdot)$ depende de F , $\hat{Q}_F(\cdot)$ depende de los datos observados, y $\hat{Q}_{\hat{F}}(\cdot)$ depende de los datos *bootstrap*. Se busca $Q(\hat{A}_F) = [Q(\hat{A}_F) - Q(A)] + Q(A)$. Sea $S = Q(\hat{A}_F) - Q(A)$ y $R = Q(A)$. Se aproxima S con $\hat{Q}_F(\hat{A}_F) - (\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}}) - \hat{Q}_F(\hat{A}_F))$ y R con $\hat{Q}_F(\hat{A}_{\hat{F}}) - \hat{Q}_F(\hat{A}_F)$. Se obtiene $\hat{Q}(\hat{A}_F) = \hat{Q}_F(\hat{A}_F) + \hat{Q}_F(\hat{A}_{\hat{F}}) - \hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$ y para calcularlo con *bootstrap* se sigue el siguiente procedimiento:

- a) Calcular $\hat{Q}_F(\hat{A}_F)$.
- b) Obtener una muestra de los datos con reemplazo tamaño n .
- c) Calcular $\hat{Q}_F(\hat{A}_{\hat{F}})$ y $\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$ con esta muestra.
- d) Repetir los pasos a) y b) un número u de veces. Se escoge u de forma que el esfuerzo computacional sea razonable pero que el número de muestras sea suficiente para disminuir el sesgo.
- e) Promediar $\hat{Q}_F(\hat{A}_{\hat{F}})$ y $\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$ de cada muestra.
- f) Calcular $\hat{Q}(\hat{A})$.

6. Simular datos + bootstrap:

- a) Obtener muestra de datos con reemplazo.
- b) Calcular \hat{P}_{1i} y \hat{P}_{0i} de la nueva muestra.
- c) Simular Y^* y calcular $\hat{Q}_F(\hat{A}_{\hat{F}})$ y $\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$.
- d) Repetir los pasos a)-c) un número u de veces.

- e) Promediar $\hat{Q}_F(\hat{A}_{\hat{F}})$ y $\hat{Q}_{\hat{F}}(\hat{A}_{\hat{F}})$ de cada muestra.
- f) Calcular $\hat{Q}(\hat{A})$.

En general, *Virtual Twins* tiene como ventaja que es fácil de interpretar, explica las interacciones con pocas variables y es eficiente computacionalmente. Hay métodos tradicionales que sólo pueden encontrar interacciones incluidas en el modelo, con el orden de los términos pequeño y no son viables si el número de covariables es grande, problemas que *Virtual Twins* soluciona según sus creadores (J. C. Foster, Taylor y Ruberg 2011). No obstante, es importante tomar en cuenta las dificultades de las cuales se habló en el Capítulo 1 para encontrar subgrupos.

3.2. Método de árboles de interacción

Una aplicación de árboles aleatorios es el método de árboles de interacción de Su et al. (2009). En él se escogen los criterios de división y selección de árboles de forma que se encuentren efectos en subgrupos. El enfoque que se le da al método es de tipo Frecuentista y se busca que la identificación sea guiada por datos y de forma recursiva. Consta básicamente de 3 pasos que son crecer un árbol inicial, podar el árbol y fusionar nodos terminales similares. Resulta en un árbol binario para el cual cada división está determinada por una covariable importante. Los nodos terminales corresponden a los individuos asignados a cada subgrupo por el método. A continuación, se da una descripción detallada de cada paso:

1. Crecer un árbol inicial T_0 : Se empieza con un nodo (raíz) y se divide

el nodo en 2 hijos, a los cuales se vuelve a dividir y así sucesivamente. Para cada nodo, la división está determinada de la siguiente forma:

- Sea s una división del árbol y t_L, t_R los nodos a la izquierda y derecha que resultan de la división. Para cada nodo $j \in \{L, R\}$ se tiene $\mu_{ij}, \bar{y}_{ij}, s_{ij}^2, n_{ij}$, $i = 0, 1$, la media poblacional, la media muestral, la varianza muestral y el tamaño de la muestra para individuos control ($i = 0$) e individuos con tratamiento ($i = 1$). La heterogeneidad de efectos entre t_L y t_R está dada por la diferencia entre los efectos que tiene el tratamiento para cada nodo: $\mu_{1L} - \mu_{0L}$ y $\mu_{1R} - \mu_{0R}$. Como las medias son desconocidas, se propone utilizar la siguiente estadística:

$$t(s) = \frac{(\bar{y}_{1L} - \bar{y}_{0L}) - (\bar{y}_{1R} - \bar{y}_{0R})}{\hat{\sigma} \sqrt{\frac{1}{n_{1L}} + \frac{1}{n_{1R}} + \frac{1}{n_{0L}} + \frac{1}{n_{0R}}}}.$$

Y para facilitar cálculos, se eleva al cuadrado y se obtiene $G(s) = (t(s))^2$. Entonces, se busca la división s^* , tal que:

$$\begin{aligned} G(s^*) &= \max_s G(s), \\ G(s) &= (t(s))^2 \rightsquigarrow \mathcal{X}_{(1)}^2, \\ \hat{\sigma}^2 &= \sum w_{ij} s_{ij}^2, \\ w_{ij} &= \frac{n_{ij} - 1}{\sum(n_{uv} - 1)}, \end{aligned}$$

donde el símbolo \rightsquigarrow se utiliza para denotar que $G(s)$ converge a una distribución ji-cuadrada con 1 grado de libertad ($\mathcal{X}_{(1)}^2$).

El crecimiento del árbol tiene como propósito encontrar aquellas ramas que tengan mayor diferencia en efectos.

Para cada nodo se busca la división óptima hasta llegar a una condición de paro. Si se llega a la condición de paro, se busca la

división para el siguiente nodo. Una vez que todos los nodos hayan llegado a la condición de paro, se dejan de buscar divisiones y se sigue con el siguiente paso. Las condiciones de paro pueden ser las siguientes:

- Todas las covariables en un nodo tienen el mismo valor para diferentes observaciones.
 - El número de observaciones en un nodo es igual a un mínimo establecido anteriormente.
 - La profundidad del árbol en este nodo es igual a un máximo establecido anteriormente. Esto es, del nodo terminal a la raíz hay un número determinado de nodos igual a un máximo.
2. Podar el árbol: Como en CART (Breiman et al. 1984), se crea una sucesión de árboles anidados (cada árbol es un subárbol del anterior) y se escoge el de mejor tamaño. En cada árbol se busca podar la rama más débil, que es aquella que tiene la menor interacción con el tratamiento por nodo en el árbol. A mayor interacción con el tratamiento, mayor es la diferencia en el efecto con respecto a otras ramas. Por lo tanto, una rama con poca interacción es ineficiente. Uno de los pasos se ilustra en la figura 3.4 y el algoritmo es el siguiente:

- Empezar con T_0 , el árbol que se obtuvo en el primer paso y que será el más grande de todos los árboles.
- Mientras $T_i \neq T_M$, donde T_M es árbol raíz (aquel para el cual todas las observaciones se agrupan en la raíz ya que no tiene divisiones):
 - Buscar aquel nodo $h^* \in T_i$ tal que:

$$g(h^*) = \min_h g(h),$$

$$g(h) = \frac{G(T_h)}{|T_h - \tilde{T}_h|},$$

$$G(T_h) = \sum_{i \in T_h - \tilde{T}_h} G(i)$$

$$= \sum_{i \in T_h - \tilde{T}_h} (t(i))^2.$$

Donde $|\cdot|$ denota el número de nodos que tiene un árbol T y \tilde{T} es el conjunto de nodos terminales del árbol T . Por lo tanto, $T - \tilde{T}$ es el árbol conformado por los nodos internos de T . De este modo, $g(h)$ es la medida de interacción del tratamiento en el árbol T_h promediada por nodo. Esta medida está compuesta por $G(T_h)$ que es la medida de interacción general de T_h , definida en el paso (1) del algoritmo y $|T_h - \tilde{T}_h|$, que son el número de nodos internos del árbol T_h .

- Podar rama h^* .
- El árbol que queda es T_{i+1} .
- $i = i + 1$

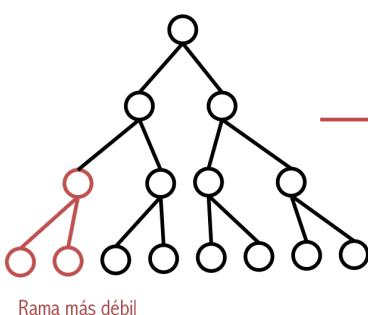
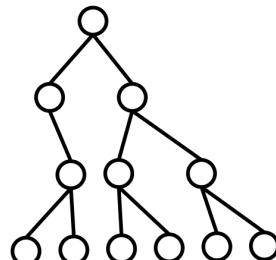
Árbol inicial T_0 :Árbol anidado T_i :

Figura 3.4: Árbol anidado después de podar la rama más débil.

- Se tiene la sucesión de subárboles anidada $T_M \prec \dots \prec T_0$. Un ejemplo de una sucesión es la figura 3.5:

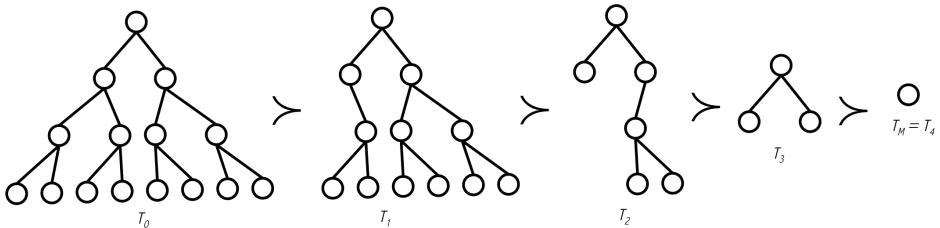


Figura 3.5: Sucesión de árboles anidados.

- Escoger árbol T_m^* tal que:

$$G_\lambda(T_m^*) = \max_{T_m} G_\lambda(T_m),$$

$G_\lambda(T_m) = G(T_m) - \lambda|T_m - \tilde{T}_m|$, medida de interacción-complejidad,
 $|\cdot|$, número de nodos de un árbol,
 \tilde{T}_m , árboles terminales de T_m ,
 $\lambda \geq 0$, parámetro de complejidad.

La función $G_\lambda(T_m)$ tiene dos componentes: $G(T_m)$ mide la interacción en el árbol y $|T_m - \tilde{T}_m|$ que a través del número de nodos mide la complejidad. Se busca que la mayor interacción pero se penaliza a los árboles más complejos. Para escoger λ , Su et al. hicieron simulaciones y concluyeron que el que tiene mejor funcionamiento es $\lambda = \ln(n)$. Este corresponde a la penalización del criterio de información Bayesiano (BIC) (Su et al. 2009 citando a Schwarz 1978).

3. Fusionar nodos terminales con efectos parecidos para facilitar interpretación del árbol: hay veces en las que un tratamiento tiene

efectos similares en diferentes ramas del árbol por diferentes razones. Por ejemplo, para un subgrupo se tiene un efecto nulo y para otro una combinación de un efecto positivo con un efecto negativo. Entonces, se puede determinar un mínimo en la diferencia de los efectos entre nodos. Si los nodos no cumplen con la condición, se toman a los individuos como parte del mismo subgrupo.

Árbol final:

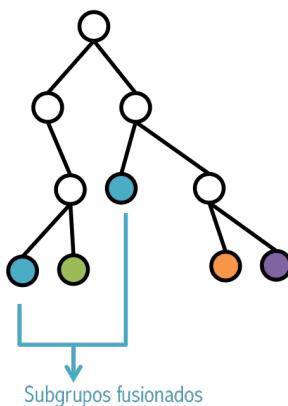


Figura 3.6: Árbol final después de fusionar nodos terminales similares.

Como se comparan los nodos terminales, puede ser que en algunos casos se fusionen nodos de diferentes niveles (véase la figura 3.6). Sólo se evalúa que los nodos tengan efectos parecidos, no la razón por la que los tienen. Entonces, no se toma en cuenta la parte del árbol en la que están los nodos, sólo los efectos que presentan.

Adicionalmente, es posible que un objetivo sea determinar las covariables con mayor importancia (influencia) para determinar la estructura del árbol final. A estas variables se les llama “modificadoras de

efecto”. Para encontrarlas, se crea un bosque aleatorio compuesto por varios árboles de interacción de la siguiente forma:

- Sea \mathcal{L} el conjunto de datos a analizar.
- Sea V_j el costo de excluir a la covariable X_j , $1 \leq j \leq J$. Se empieza con $V_j = 0$.
- Para $b = 1, \dots, B$:
 - a) Generar muestra con reemplazo (*bootstrap*) \mathcal{L}_b .
 - b) Crecer árbol inicial T_b con el algoritmo de los árboles de interacción sin podar: Para cada nodo, seleccionar m_0 covariables al azar. Determinar la división óptima s^* de forma que sólo dependa de una de las m_0 covariables seleccionadas.
 - c) Clasificar el conjunto de datos $\mathcal{L} - \mathcal{L}_b$ con el árbol inicial T_b y calcular la estadística $G(T_b)$.
 - d) Para $j = 1, \dots, J$:
 - a) Permutar los valores de X_j en los datos $\mathcal{L} - \mathcal{L}_b$ aleatoriamente.
 - b) Aplicar el árbol a $\mathcal{L} - \mathcal{L}_b$ con la permutación y calcular $G_j(T_b)$.
 - c) $V_j = V_j + \frac{G(T_b) - G_j(T_b)}{G(T_b)}$.
 - d) $j = j + 1$
 - e) $b = b + 1$
- $V_j = V_j/B$
- Seleccionar un número determinado de covariables con mayor V_j .

El funcionamiento del algoritmo se basa en que si una variable es importante, afecta mucho a una interacción alta $G(T_b)$. Al hacer una permutación, $G_j(T_b)$ será pequeña (los valores de la covariable dejan de afectar al resultado porque son al azar). Por otro lado, si la interacción alta $G(T_b)$ se debe a covariables que no son j , $G_j(T_b)$ no tendrá muchos cambios independientemente de la permutación y se penalizará restando constantemente. El uso del *bootstrap* da como resultado muestras independientes, por lo que se pueden utilizar los mismos datos y hacer múltiples cálculos sin un aumento desproporcionado del sesgo. Utilizar muestras con *bootstrap* asegura que el resultado de V_j sea válido y no sólo determinado por una muestra no representativa.

El método es atractivo porque puede encontrar interacciones cualitativas, así como cuantitativas. Es aplicable a cualquier base de datos que tenga covariables, una variable de respuesta y una covariable binaria. En los ensayos clínicos la covariable binaria es el tratamiento, pero no es necesario que lo sea. Por ejemplo, Su et al. utilizan el género como covariable binaria para analizar diferencias en el salario de trabajadores. Además, el árbol es fácil de interpretar gracias al uso cotidiano por público general de gráficas en la vida diaria. Por lo tanto, diferentes aplicaciones del método se pueden dar a conocer en áreas de conocimiento que no utilicen estadística o matemáticas, gracias a que no hay necesidad de adentrarse en el funcionamiento del método para comprender sus resultados.

Capítulo 4

Métodos de análisis de subgrupos predefinidos

4.1. Selección Bayesiana de modelo basado en la urna de Pòlya

Laud, Sivaganesan y Müller (2013; 2011; 2010) presentan el análisis de subgrupos como un problema de selección de modelos. Es decir, hay varios modelos y se usa una selección de modelos Bayesiana para elegir el óptimo. El modelo óptimo indicará si hay heterogeneidad de efectos en subgrupos y la cantidad de efectos diferentes. Los subgrupos son predefinidos a través de covariables importantes.

Como los subgrupos son predefinidos, se pueden conocer todos los modelos posibles para dichos subgrupos. Supongamos que se busca probar la diferencia en efectos en subgrupos definidos por el género del

paciente. Entonces, los pacientes se dividirán en dos subgrupos y los posibles modelos que evalúa este método son los siguientes:

1. Los pacientes de ambos sexos presentan el mismo efecto y es diferente de cero. No hay efectos en subgrupos.
2. El sexo femenino presenta un efecto diferente a cero y el masculino no presenta efecto. Hay efectos en subgrupos.
3. El sexo femenino no presenta efectos y el masculino presenta un efecto diferente a cero. Hay efectos en subgrupos.
4. Ambos sexos presentan efectos diferentes a cero, pero de distinta magnitud. Hay efectos en subgrupos.
5. Ninguno de los sexos presentan efectos. El tratamiento tiene un efecto nulo y no hay efectos en subgrupos.

Normalmente, cuando se realizan ensayos clínicos para probar la efectividad de un tratamiento se comparan los modelos general y nulo (1. y 5. de la lista anterior). Como se ha visto en capítulos anteriores, una práctica común es comparar estos modelos a través de pruebas de hipótesis. Sin embargo, todos los modelos enlistados anteriormente tienen una probabilidad, por más pequeña que sea, diferente a cero. La selección de modelos Bayesiana compara todos los modelos posibles de acuerdo a que tan probables son y selecciona uno de acuerdo a una regla de decisión.

Desde un enfoque Bayesiano, la probabilidad de cada modelo está determinada por dos elementos: el conocimiento subjetivo inicial del investigador sobre los modelos y los resultados del ensayo clínico.

Inicialmente, se supone que todos los modelos tienen una probabilidad similar. Por lo tanto, se utilizan distribuciones no informativas iniciales para representar la falta de conocimiento subjetiva de cada modelo. Posteriormente, se utilizan las respuestas que tienen los pacientes al tratamiento como información para actualizar las probabilidades de los modelos.

En notación matemática, se busca calcular y comparar la probabilidad final de cada modelo a través de la fórmula:

$$\begin{aligned} P(M|y) &= \frac{P(y|M)P(M)}{P(y)} \\ &= \frac{P(y|M)P(M)}{\sum_{M_i \in \mathcal{M}} P(y|M_i)P(M_i)}, \end{aligned}$$

donde y son las respuestas de los pacientes al tratamiento que se suponen continuas y M es uno de los modelos pertenecientes al espacio de modelos \mathcal{M} .

Entonces, los elementos necesarios para tomar una decisión de Bayes son los siguientes:

- La probabilidad inicial de cada modelo $P(M)$.
- La verosimilitud o probabilidad de la respuesta de los individuos bajo un modelo $P(y|M)$.
- La función de utilidad $u(\mathbf{d}, M, y)$ que se utiliza para tomar la decisión de Bayes \mathbf{d}^* . Recordar que la decisión de Bayes es aquella que maximiza la utilidad esperada de cada decisión \mathbf{d} . Se busca elegir un modelo que tenga evidencia fuerte a su favor de ser el real y poca evidencia en contra, no el modelo que maximiza la probabilidad final. Esto con el fin de controlar la probabilidad del

Error Tipo I y otros errores que se calculan a través de simulaciones. Tal control se puede dar a través de diferentes funciones de utilidad.

El método es relevante ya que resuelve el problema de multiplicidad controlando el Error Tipo I y gracias a que tiene una justificación teórica Bayesiana, el método es replicable para distintos experimentos. A continuación, se detalla cada parte del método, empezando por generalizar la definición de los posibles modelos.

4.1.1. Definición de posibles modelos

Para simplificar la explicación del método, se supone que los subgrupos predefinidos $\{1, \dots, S\}$ dependen de una sola covariante X_i . Si hay S subgrupos, entonces puede haber a lo máximo S efectos del tratamiento diferentes. Si se tiene el máximo número de efectos, estos pueden ser $\{0, 1, 2, \dots, S - 1\}$ si un subgrupo tiene el efecto nulo o $\{1, 2, \dots, S\}$ si todos los subgrupos tienen efectos distintivos y diferentes a cero. El número total de efectos distintivos y diferentes a cero se denota por K , con $K \leq S - 1$. Los modelos se enlistan utilizando un vector $\gamma = (\gamma_1, \dots, \gamma_S)$, donde $\gamma_s = 0$ si no hay efecto en el subgrupo y $\gamma_s = k$, $0 < k \leq S$ en otro caso. Los subgrupos que tienen el mismo efecto reciben

el mismo número, asignados en orden de apariencia. Es decir, para $r > s$

$$\gamma_r = \begin{cases} 0, & \text{si el tratamiento no tiene efecto para el subgrupo } r. \\ \gamma_s, & \text{si el tratamiento tiene el mismo efecto } (\neq 0) \text{ para} \\ & \text{el subgrupo } r \text{ que para el } s. \\ k \neq \gamma_s, & \text{si el tratamiento tiene un efecto diferente a cero} \\ & \text{distintivo de todos efectos anteriores.} \end{cases}$$

Por lo tanto, $K = \max\{\gamma_s | s \leq S\}$. Por ejemplo, el modelo $\gamma = (1, 0, 2, 1)$ es aquel en el que el primer y cuarto subgrupo tienen el mismo efecto, el segundo subgrupo no tiene efecto y el tercer subgrupo tiene un efecto diferente a los demás. Hay $K = 2$ efectos distintivos diferentes a cero para $S = 4$ subgrupos.

4.1.2. Distribución del modelo M basándose en Urnas de Pòlya

Para encontrar la probabilidad de un modelo M , es decir $P(M)$, Laud et al. (2013) modifican la Urna de Pòlya para hacerla enriquecida por cero (explicada en el Capítulo 2). Para los diferentes modelos, se toman los colores de las pelotas equivalentes a los efectos de subgrupos. En particular, se reparametriza $q = \frac{1}{1+\varphi}$ y se asignan distribuciones tal que $p \sim \text{Beta}(\varphi_1, \eta_1)$ y $q \sim \text{Beta}(\varphi_2, \eta_2)$, con $\varphi_1, \eta_1, \varphi_2, \eta_2$ conocidos.

Antes de calcular probabilidades se define:

- $K_s = \max\{\gamma_{s'} \neq 0 | s' \leq s\}$, el número de efectos distintivos diferentes de cero en los primeros s subgrupos. ($K_S = K$)

- $N_{sk} = \#\{s' | \gamma_{s'} = k, s' \leq s\}$, el número de subgrupos con efectos de tipo $k \neq 0$ en los primeros s subgrupos para $1 \leq k \leq K_s$.
- $L_s = \#\{s' | \gamma_{s'} \neq 0, s' \leq s\}$, el número de subgrupos con efectos diferentes a cero en los primeros s subgrupos. ($\sum_k N_{sk} = L_s$)
- $N_k = N_{Sk} = \#\{s | \gamma_s = k, 1 \leq s \leq S\}$, el número de subgrupos del tipo $k \in \{0, 1, \dots, K\}$ totales. ($\sum N_k = S$)

Para el modelo $\gamma = (1, 0, 1, 2)$, se muestran los valores de K_s, N_{sk}, L_s y N_k para distintos valores de s y k en el cuadro 4.1:

s	1	2	3	$S = 4$
K_s	1	1	1	2
N_{s1}	1	1	2	2
N_{s2}	-	-	-	1
L_s	1	1	2	3
	N_0	1		
	N_1	2		
	N_2	1		

Cuadro 4.1: Diferentes valores para el modelo $\gamma = (1, 0, 1, 2)$.

Como distribución inicial de M se asigna $P(M|p, q) = c(p, q)P(N_0, \dots, N_K|p, q)$, donde $c(p, q)$ es una constante de normalización. Para obtener $P(N_0, \dots, N_K|p, q)$, es necesario definir las siguientes probabilidades:

$$P(\gamma_{s+1} = 0 | \gamma_1, \dots, \gamma_s) = p, \quad (4.1)$$

$$P(\gamma_{s+1} = K_s + 1 | \gamma_1, \dots, \gamma_s) = (1 - p) \frac{1 - q}{1 - q + L_s q}, \quad K_s \geq 0, \quad (4.2)$$

$$P(\gamma_{s+1} = k | \gamma_1, \dots, \gamma_s) = (1 - p) \frac{N_{sk} q}{1 - q + L_s q}, \quad k = 1, \dots, K_s \geq 1. \quad (4.3)$$

El desarrollo de las probabilidades es el siguiente:

- La ecuación (4.1) está basada en que como la urna es enriquecida por cero, en cada paso hay una probabilidad p de no sacar una bola o sacar cero como el efecto del subgrupo $s + 1$.
- La ecuación (4.2) es el equivalente a tomar una bola de un color nuevo en la urna de Pòlya. Como en algunos pasos sale cero, no en todos los pasos se agrega una pelota a la urna. Entonces, se agrega $(1 - p)$ a ecuación. También, se puede tomar una bola de un color nuevo o uno que ya estaba en la urna y como la probabilidad de tomar una bola de color nuevo es $\frac{\varphi}{\varphi + L_s}$, se agrega a la ecuación. En el contexto de la urna, N_{sk} es el número de pelotas de color k en el paso s y L_s es número de pelotas en la urna en el paso s . Entonces se tiene:

$$\begin{aligned} P(\gamma_{s+1} = K_{s+1} | \gamma_1, \dots, \gamma_s) &= P(\gamma_{s+1} \neq 0) \frac{\varphi}{\varphi + L_s} \\ &= (1 - p) \frac{\frac{1-q}{q}}{\frac{1-q}{q} + L_s} \\ &= (1 - p) \frac{1 - q}{1 - q + L_s q}. \end{aligned}$$

Cabe mencionar que en este modelo se toma la primera bola diferente a cero como una nueva. Es decir, la urna comienza vacía. Si sale un efecto cero con probabilidad p , la urna continúa vacía. Si sale una bola diferente a cero con probabilidad $1 - p$, esta será la primera bola de la urna.

- La ecuación (4.3) es el equivalente a tomar una bola de un color que ya está en la urna. Por lo tanto, en el paso $s + 1$, k puede tomar valores

desde 1 hasta el efecto máximo K_s . En este caso, la probabilidad de tomar una bola de un color que ya está en la urna es $\frac{L_s}{\varphi + L_s}$. De este modo,

$$\begin{aligned} P(\gamma_{s+1} = k | \gamma_1, \dots, \gamma_s) &= P(\gamma_{s+1} \neq 0) \left(\frac{N_{sk}}{L_s} \right) \left(\frac{L_s}{\varphi + L_s} \right) \\ &= (1-p) \frac{N_{sk}}{\frac{1-q}{q} + L_s} \\ &= (1-p) \frac{N_{sk}q}{1 - q + L_sq}. \end{aligned}$$

Con esto es posible calcular $P(N_0, \dots, N_K | p, q)$, tomando en cuenta que:

- N_0 depende únicamente de p .
- Para el resto de los colores $k = 1, \dots, K$, es necesario que no haya salido el cero, que se haya tomado un nuevo color para poder introducir el color k a la urna y que se haya tomado ese color $N_k - 1$ veces más. Estos eventos tienen probabilidad

$$\begin{aligned} &\underbrace{(1-p) \frac{\varphi}{\varphi + 1}}_{\text{bola color nuevo } k} \cdot \underbrace{(1-p) \frac{1}{\varphi + L_{k,2}}}_{\text{bola color } k \text{ habiendo 1 en la urna}} \cdot \underbrace{(1-p) \frac{2}{\varphi + L_{k,3}}}_{\text{bola color } k \text{ habiendo 2 en la urna}} \cdots \cdot \underbrace{(1-p) \frac{N_k - 1}{\varphi + L_{k,N_k}}}_{\text{bola color } k \text{ habiendo } N_k - 1 \text{ en la urna}} \\ &= \frac{(1-p)^{N_k} \cdot \varphi \cdot (1)(2)\dots(N_k - 1)}{(\varphi + 1)(\varphi + L_{k,2})(\varphi + L_{k,3})\dots(\varphi + L_{k,N_k})} \\ &= \frac{\varphi(1-p)^{N_k}(N_k - 1)!}{\prod_{i=1}^{N_k} (\varphi + L_{k,i})}. \end{aligned}$$

Donde $L_{k,i}$ es el número de pelotas que había en la urna al tomar la i -ésima pelota del color k . Se toma $L_{k,1} = 1 \forall k$.

- Tomando en cuenta las probabilidades anteriores se obtiene:

$$\begin{aligned}
 P(N_0, \dots, N_K | p, q) &= [p^{N_0}] \left[\frac{\varphi(1-p)^{N_1}(N_1-1)!}{\prod_{i=1}^{N_1}(\varphi+L_{1,i})} \right] \dots \left[\frac{\varphi(1-p)^{N_K}(N_K-1)!}{\prod_{i=1}^{N_K}(\varphi+L_{K,i})} \right] \\
 &= \frac{p^{N_0}\varphi^K(1-p)^{\sum_{k=1}^K N_k} \prod_{k=1}^K (N_k-1)!}{\varphi(\varphi+1)(\varphi+2)\dots(\varphi+L_S-1)} \\
 &= p^{N_0}(1-p)^{S-N_0} \frac{\varphi^K \prod_{i=1}^K [N_i-1]!}{\prod_{i=1}^{S-N_0} (\varphi+i-1)}.
 \end{aligned}$$

La constante de normalización $c(p, q)$ cuenta el número de distintos modelos para todas las k y (N_0, \dots, N_K) posibles. Sirve para que la probabilidad del modelo sea propia. Su fórmula es:

$$c(p, q) = \frac{1}{\sum_{k=0}^{S-1} \sum_{N_i} \binom{S}{N_0} P(N_0, \dots, N_K | p, q)},$$

para todos los posibles $N_0 \geq 0, N_j > 0, j \neq 0$ que cumplan $\sum N_i = S$.

Con esto, se tienen todos los elementos para calcular la distribución inicial de M ,

$$P(M) = \int P(M|p, q)P(p, q) dp dq, \text{ donde} \quad (4.4)$$

$$P(M|p, q) = c(p, q)P(N_0, \dots, N_K | p, q). \quad (4.5)$$

4.1.3. Distribuciones iniciales de los parámetros desconocidos bajo un modelo

Esta sección está dedicada a definir los elementos necesarios para calcular la verosimilitud o la probabilidad de obtener ciertas respuestas a tratamiento bajo un modelo $P(y|M)$. Para cada modelo M se tienen

$J_0 + J_1$ observaciones donde las respuestas de los pacientes al tratamiento son y_{1j} , $j = 1, \dots, J_1$ y las respuestas de los pacientes control son y_{0j} , $j = 1, \dots, J_0$. Se supone que $y_{1j} \sim N(\mu_{1s}, \sigma^2)$ y $y_{0j} \sim N(\mu_{0s}, \sigma^2)$, donde $x_{ij} = s$ si el individuo j pertenece al subgrupo s predefinido por la covariable i . Esto hace que las medias sean iguales para los pacientes dentro de cada subgrupo. En el siguiente cuadro 4.2 se muestran las distribuciones para cada paciente:

Control	Tratamiento
$y_{01} \sim N(\mu_{0,x_{01}}, \sigma^2)$	$y_{11} \sim N(\mu_{1,x_{11}}, \sigma^2)$
$y_{02} \sim N(\mu_{0,x_{02}}, \sigma^2)$	$y_{12} \sim N(\mu_{1,x_{12}}, \sigma^2)$
\vdots	\vdots
$y_{0J_0} \sim N(\mu_{0,x_{0J_0}}, \sigma^2)$	$y_{1J_1} \sim N(\mu_{1,x_{1J_1}}, \sigma^2)$

Cuadro 4.2: Tabla de distribuciones de cada paciente.

Si los individuos del grupo de control con $j = \{2, 5, 7\}$ y los individuos del tratamiento $j = \{3, 5, 8\}$ pertenecen al subgrupo $s = 2$ por ejemplo, entonces seguirán una distribución $N(\mu_{02}, \sigma^2)$ y $N(\mu_{12}, \sigma^2)$, respectivamente.

Por lo tanto, la verosimilitud es

$$\begin{aligned} P(y|M) &= \prod P(y_{ij}|M) \\ &= \left[\prod_{j=1}^{J_0} P(y_{0j}|M) \right] \left[\prod_{j=1}^{J_1} P(y_{1j}|M) \right]. \end{aligned}$$

Para calcular la verosimilitud, es necesario conocer la media y la varianza de cada subgrupo. Como la varianza es común para todos los

subgrupos no es necesario realizar muchos cálculos para definirla. En cambio, cada subgrupo tiene una media diferente para los pacientes control y los pacientes con tratamiento. Para simplificar, se define $\delta_s = \mu_{1s} - \mu_{0s}$, el efecto para cada subgrupo $s = 1, \dots, S$. Este vector se utiliza porque no es importante conocer la magnitud del efecto control o placebo (esta se puede suponer nula). Sólo interesa probar si al aplicar el tratamiento se obtienen efectos significantes diferentes. Entonces el cuadro 4.2 se puede expresar de una forma diferente, mostrada en el cuadro 4.3:

Control	Tratamiento
$y_{01} \sim N(\mu_{0,x_{01}}, \sigma^2)$	$y_{11} \sim N(\mu_{0,x_{11}} + \delta_{x_{11}}, \sigma^2)$
$y_{02} \sim N(\mu_{0,x_{02}}, \sigma^2)$	$y_{12} \sim N(\mu_{0,x_{12}} + \delta_{x_{12}}, \sigma^2)$
\vdots	\vdots
$y_{0J_0} \sim N(\mu_{0,x_{0J_0}}, \sigma^2)$	$y_{1J_1} \sim N(\mu_{0,x_{1J_1}} + \delta_{x_{1J_1}}, \sigma^2)$

Cuadro 4.3: Tabla de distribuciones utilizando δ .

De esta forma, se obtiene

$$\begin{aligned}y_{0j} &\sim N(\mu_{0s}, \sigma^2) \\y_{1j} &\sim N(\mu_{0s} + \delta_s, \sigma^2),\end{aligned}$$

donde $x_{ij} = s$ si el individuo j pertenece al subgrupo s predefinido por la covariante i . Por lo tanto, todos los individuos de un subgrupo tendrán la misma distribución si están en el grupo control y la misma distribución si están en el grupo de tratamiento. Con esto, se puede definir $y_0 \sim N_S(\mu_0, \sigma^2 I_S)$ la distribución normal multivariada que define los subgrupos del grupo control y $y_1 \sim N_S(\mu_0 + \delta, \sigma^2 I_S)$ que define los subgrupos del grupo con tratamiento. En las distribuciones,

$\mu_0 = (\mu_{01}, \dots, \mu_{0S})$ es el vector de medias control y $\delta = (\delta_1, \dots, \delta_S)$ es el vector de diferencias en el efecto para cada subgroupo. Con las δ_s se forma el vector de K efectos diferentes a cero $\delta^* = (\delta_1^*, \dots, \delta_K^*)$.

Por lo tanto, los parámetros desconocidos bajo un modelo son δ , μ_0 y σ^2 (δ^* se puede conocer a través de δ). Esto que quiere decir que, si conocemos estos parámetros, podemos conocer cualquier modelo M y calcular $P(y|M)$. Por lo tanto, se establecen distribuciones iniciales para los parámetros.

Laud, Sivaganesan y Müller sugieren asignar distribuciones no informativas a todos los parámetros para que los modelos tengan probabilidades similares y que la selección del modelo óptimo sea impulsada por los datos. En particular, se basan en la comparación de modelos de Liang et al. de mezcla de distribuciones *g-prior* (Liang et al. 2008 basados en Zellner 1986), las cuales se describen en el Capítulo 2, para asignar las distribuciones iniciales. El modelo puede ser visto de la siguiente forma:

$$y = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim N(M, \sigma^2 I_{2S}),$$

$$M = \begin{bmatrix} \mu_0 \\ \mu_0 \end{bmatrix} + I_{2S} \begin{bmatrix} \mathbf{0}_S \\ \delta \end{bmatrix},$$

donde se toma n como $2S$, M como μ , σ^2 como $\frac{1}{\phi}$, $(\mu_0, \mu_0)^T$ en vez de $\alpha\mathbf{1}_n$, I_{2S} como X_γ y $(\mathbf{0}_S, \delta)^T$ como β_γ de la notación del modelo original (Liang et al. 2008). La única diferencia con el modelo original de Liang et al. es que, en este modelo, se utiliza $(\mu_0, \mu_0)^T$ en vez de $\alpha\mathbf{1}_n$ y μ_0 no es constante para todos los datos. Esto es, los datos no están centrados como $\alpha\mathbf{1}_n$ está centrado en α . Esto se puede arreglar con una reparametrización de los datos y continuar como lo hacen Liang et al.

Siguiendo las propuestas de Liang et al., las probabilidades iniciales se definen de la siguiente forma:

$$\delta^*|g, \mu_0, \sigma^2 \sim N_K(\mathbf{0}_K, g\sigma^2 I_K), \text{ que proviene de} \quad (4.6)$$

$$(\mathbf{0}_S, \delta)^T | g, \mu_0, \sigma^2 \sim N_{2S}(\mathbf{0}_{2S}, g\sigma^2 I_{2S}), \text{ donde} \quad (4.7)$$

$$P(g, \mu_0, \sigma^2) = P(g)P(\mu_0, \sigma^2), \quad (4.8)$$

$$P(\mu_0, \sigma^2) = P(\mu_0)P(\sigma^2), \quad (4.9)$$

$$P(\mu_0) \propto 1, \quad \mu_0 \in \mathbb{R}^S, \quad (4.10)$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2}, \quad \sigma^2 > 0, \quad (4.11)$$

$$P(g) = \frac{1}{(1+g)^2}, \quad g > 0. \quad (4.12)$$

El modelo anterior es una mezcla de *g-priors* (Liang et al. 2008; Li y Clyde 2018) ya que se utiliza una distribución inicial para determinar el parámetro g . En particular, se utiliza la distribución parecida a las *hyper-g* definida por Liang et al. (2008), expuesta en el Capítulo 2. Recordando la ecuación (2.5), esta distribución es

$$p(g) = \frac{a-2}{2}(1+g)^{-a/2}, g > 0$$

utilizando $a = 4$.

Una vez definidas las distribuciones iniciales de los parámetros desconocidos de cada modelo se puede calcular $P(y|M)$. Esta se calcula de la siguiente forma:

$$P(y|M) = \int P(y|M, g)P(g) dg, \text{ donde} \quad (4.13)$$

$$P(y|M, g) = \int P(y|\mu_0, \delta^*, \sigma^2)P(\delta^*|g, \mu_0, \sigma^2)P(\mu_0, \sigma^2) d\mu_0 d\delta^* d\sigma^2 \quad (4.14)$$

La selección de distribuciones iniciales se hace con motivo de que la integral $P(y|M, g)$ tenga una expresión de forma cerrada. Es decir, que se pueda calcular de forma analítica y, por lo tanto, no se requiera de mayor esfuerzo computacional.

4.1.4. Probabilidades finales

Recordando que el objetivo final del método es seleccionar el modelo M que maximice una función de utilidad esperada, se busca calcular la probabilidad final o posterior de cada modelo $P(M, y)$. Recordando las ecuaciones (4.4)-(4.14), se obtiene la probabilidad final de cada modelo:

$$P(M|y) = \frac{P(y|M)P(M)}{\sum_{M' \in \mathcal{M}} P(y|M')P(M')}, \text{ donde}$$

$$P(y|M) = \int P(y|M, g)P(g) dg,$$

$$P(M) = \int P(M|p, q)P(p, q) dp dq.$$

Los componentes de las ecuaciones pasadas se habían definido anteriormente. La probabilidad $P(y|M)$ está compuesta por

$$P(y|M, g) = \int P(y|\mu_0, \delta^*, \sigma^2)P(\delta^*|g, \mu_0, \sigma^2)P(\mu_0, \sigma^2) d\mu_0 d\delta^* d\sigma^2 \text{ y}$$

$$P(g) = \frac{1}{(1+g)^2}.$$

Por otro lado, $P(M)$ se compone por:

$$P(M|p, q) = c(p, q)P(N_0, \dots, N_K|p, q), \text{ donde}$$

$$c^{-1}(p, q) = \sum_{k=0}^{S-1} \sum_{N_i} \binom{S}{N_0} P(N_0, \dots, N_K|p, q),$$

$$P(N_0, \dots, N_K | p, q) = p^{N_0} (1-p)^{S-N_0} \frac{\varphi^K \prod_{i=1}^K [N_i - 1]!}{\prod_{i=1}^{S-N_0} (\varphi + i - 1)}, \text{ y}$$

$P(p, q) = P(p)P(q)$, donde $p \sim \text{Beta}(\varphi_1, \eta_1)$ y $p \sim \text{Beta}(\varphi_2, \eta_2)$.

4.1.5. Ampliación a múltiples covariables

Hasta el momento se ha supuesto que los subgrupos están definidos por una sola covariable X_i . Sin embargo, el modelo se puede ampliar de tal forma que cada covariable defina a cierto número de subgrupos. Por ejemplo, para un grupo de pacientes se pueden seleccionar como covariables de interés el sexo y la edad. Los subgrupos determinados por el sexo son femenino y masculino y los subgrupos determinados por la edad se definen como personas menores de 45 años y personas mayores de 45 años. Entonces, los posibles efectos que se encuentran en este método son los siguientes:

- Hay un efecto general del tratamiento pero no hay efectos en subgrupos para ninguna de las dos covariables.
- Hay efectos en subgrupos para la covariable de sexo, pero no hay efectos en subgrupos para la covariable de edad. Esto es, hay diferencias del tratamiento entre los hombres y las mujeres, pero el tratamiento no cambia dependiendo de la edad.
- Hay efectos en subgrupos para la covariable de edad, pero no hay efectos en subgrupos para la covariable de sexo. Es decir, el tratamiento tiene diferente magnitud dependiendo de la edad pero no del sexo del paciente.
- Hay efectos en subgrupos para las dos covariables. Hay diferencias en tratamiento entre hombres y mujeres, pero también entre personas

menores de 45 y mayores de 45 años. Por lo tanto, para cada uno de los siguientes subgrupos el tratamiento causa un efecto diferente: mujeres menores de 45, mujeres mayores de 45, hombres menores de 45 y hombres mayores de 45.

- El tratamiento no tiene efecto para ningún paciente.

Estos escenarios toman en cuenta que, si hay efectos en subgrupos, puede ser que algunos de los subgrupos no tengan efecto mientras que los otros sí lo tengan o que todos los subgrupos tengan efecto, pero en diferente magnitud. Se procede a definir cómo el método incluye los subgrupos determinados por distintas covariables.

Sean X_1, \dots, X_I las variables que definirán los subgrupos. Cada covariable X_i se divide en S_i subgrupos y se define un espacio de modelos \mathcal{M}_i y H_i . Los modelos de \mathcal{M}_i se etiquetan con M_{ih} donde i corresponde a la i -ésima covariable y h corresponde al modelo, $0 \leq h \leq H_i$. Esto es, $\mathcal{M}_i = \{M_{ih}\}$, donde $0 \leq i \leq I, 0 \leq h \leq H_i$.

Para todos los espacios de modelos \mathcal{M}_i se reserva el índice $h = 0$ para el modelo nulo, aquel donde todos los subgrupos tienen efecto cero: $M_{i0} = (0, 0, \dots, 0)$, $\forall i$. Para facilitar la notación, el modelo nulo es M_{00} . De igual forma, para todos los espacios se reserva el índice $h = H_i$ para el modelo de efecto general, aquel donde todos los subgrupos tienen el mismo efecto diferente a cero: $M_{iH_i} = (1, 1, \dots, 1)$, $\forall i$. El modelo general es M_{01} y $\mathcal{M}_0 = M_{00} \cup M_{01}$.

Sea $\Gamma_i = \mathcal{M}_i \setminus (M_{i0} \cup M_{iH_i})$ los modelos con efectos en subgrupos para la covariable i , $\mathcal{M} = \mathcal{M}_0 \cup \Gamma_1 \cup \dots \cup \Gamma_I = \mathcal{M}_0 \cup \Gamma$ el conjunto de posibles modelos a elegir. Para poder elegir un modelo, necesitamos especificar un modelo de probabilidad $P(M)$ sobre el espacio \mathcal{M} . Denotamos $P_i(M)$ y

$\bar{P}_i(M) = P_i(M|y)$ la probabilidad inicial y la probabilidad final del modelo M bajo el espacio \mathcal{M}_i , respectivamente (son las que se definieron anteriormente). Entonces, construimos

$$P(M) = \begin{cases} \pi_0, & M = M_{00} \\ \pi_1, & M = M_{01} \\ \frac{P_i(M)}{I}, & M = M_{i\gamma}, \gamma \in \Gamma \end{cases}$$

Si para alguna covariable se tiene que $S_i > 2$, se utiliza

$$P(M) = \begin{cases} \pi_0, & M = M_{00} \\ \pi_1, & M = M_{01} \\ P_i(M) \frac{1-\pi_0-\pi_1}{1-\pi_{i0}-\pi_{iH_i}}, & M = M_{i\gamma}, \gamma \in \Gamma, \end{cases}$$

donde $\pi_{i0} = P_i(M_{00})$ y $\pi_{iH_i} = P_i(M_{01})$ bajo \mathcal{M}_i .

4.1.6. Decisión de Bayes

Sea \mathbf{d} la decisión de elegir un modelo. \mathbf{d} puede ser: a) M_{00} , b) M_{01} o c) uno o más modelos de subgrupos. Cada modelo de subgrupos reporta la heterogeneidad en el efecto del tratamiento para una covariable. La diferencia en el efecto no es simultánea. Esto es, no se toman en cuenta las interacciones entre covariables, sólo con el tratamiento. Elegir un modelo con efectos en subgrupos es equivalente a reportar las covariables donde hay efectos y los subgrupos correspondientes a los efectos. Entonces, se define $A_I = \{i_1, \dots, i_m\} \subset \{1, \dots, I\}$ índices de las covariables con efectos en subgrupos y $\Gamma_I = \{\gamma_i, i \in A_I\}$ el modelo escogido para cada covariable. Si $A = (A_I, \Gamma_I)$ y $n_A = |A_I|$, entonces el espacio de decisiones es $\mathcal{D} = \{M_{00}, M_{01}, A\}$.

Se define la función de utilidad como:

$$u(\mathbf{d}, M, y) = \begin{cases} u_0 I(M = M_{00}), & \mathbf{d} = M_{00} \\ u_1 I(M = M_{01}), & \mathbf{d} = M_{01} \\ u_2 I(M \in A) - (n_A - 1), & \mathbf{d} = A \end{cases}$$

$u(\mathbf{d}, M, y)$ es una función de utilidad escogida arbitrariamente que premia por elegir el modelo correcto y penaliza por cada modelo de efectos en subgrupos. La decisión de Bayes es aquella que maximiza la utilidad esperada. Para denotar la utilidad esperada definimos:

- $\gamma_i^* = \arg \max_{\gamma \in \Gamma} \{P(M_{i\gamma}|y)\}.$
- $A_I^* = \left\{ i : P(M_{i\gamma_i^*}|y) > \frac{1}{u_2} \max\{u_0 \bar{P}(M_{00}), u_1 \bar{P}(M_{01})\} \right\}.$
- $\Gamma_I^* = \{\gamma_i^*, i \in A_I^*\}.$

Si $\mathbf{d}^* = A$, entonces es $\mathbf{d}^* = A^* = (A_I^*, \Gamma_I^*)$. Esto es, sólo vamos a elegir los modelos de subgrupos que dan la máxima probabilidad final para cada covariable y sólo los vamos a reportar si su probabilidad es mayor que la probabilidad del modelo nulo y del modelo general multiplicadas por una constante.

La fórmula de la utilidad esperada es:

$$U(\mathbf{d}, M, y) = \begin{cases} u_0 P(M_{00}|y), & \mathbf{d} = M_{00} \\ u_1 P(M_{01}|y), & \mathbf{d} = M_{01} \\ u_2 \sum_{i \in A_I^*} P(M_{i\gamma_i^*}|y) - (n_{A^*} - 1), & \mathbf{d} = A^*. \end{cases}$$

Para facilitar la notación, $M_i^* = M_{i\gamma_I^*}$, $M^* = \arg \max \bar{P}(M_I^*)$ e i^* el índice i que denota M^* . El desarrollo para encontrar la decisión de Bayes es el siguiente:

- Se elige M_{01} si

$$\begin{aligned}
 & u_1 P(M_{01}|y) > u_0 P(M_{00}|y) \text{ y} \\
 & u_1 P(M_{01}|y) > u_2 \sum_{i \in A_I^*} P(M_{i\gamma_I^*}|y) - (n_{A^*} - 1) \\
 \iff & \frac{\bar{P}(M_{01})}{\bar{P}(M_{00})} > \frac{u_0}{u_1} \text{ y} \\
 & \frac{\bar{P}(M_{01})}{\bar{P}(M^*)} > \frac{u_2}{u_1} + \frac{u_2}{u_1 \bar{P}(M^*)} \left(\sum_{i \in A_I^* \setminus i^*} \bar{P}(M_i^*) - (n_{A^*} - 1) \right).
 \end{aligned}$$

- Se elige A^* si

$$\begin{aligned}
 & u_2 \sum_{i \in A_I^*} P(M_{i\gamma_I^*}|y) - (n_{A^*} - 1) > u_0 P(M_{00}|y) \text{ y} \\
 & u_2 \sum_{i \in A_I^*} P(M_{i\gamma_I^*}|y) - (n_{A^*} - 1) > u_1 P(M_{01}|y) \\
 \iff & \frac{u_2}{u_0} + \frac{u_2}{u_0 \bar{P}(M^*)} \left(\sum_{i \in A_I^* \setminus i^*} \bar{P}(M_i^*) - (n_{A^*} - 1) \right) > \frac{\bar{P}(M_{00})}{\bar{P}(M^*)} \text{ y} \\
 & \frac{u_2}{u_1} + \frac{u_2}{u_1 \bar{P}(M^*)} \left(\sum_{i \in A_I^* \setminus i^*} \bar{P}(M_i^*) - (n_{A^*} - 1) \right) > \frac{\bar{P}(M_{01})}{\bar{P}(M^*)}.
 \end{aligned}$$

- Se elige M_{00} en otro caso.

Con algunas simplificaciones, se elige M_{01} si

$$\frac{\bar{P}(M_{01})}{\bar{P}(M_{00})} > \frac{u_0}{u_1} \text{ y } \frac{\bar{P}(M_{01})}{\bar{P}(M^*)} > \frac{u_2}{u_1}.$$

Se elige A^* si

$$\frac{u_2}{u_0} > \frac{\bar{P}(M_{00})}{\bar{P}(M^*)} \text{ y } \frac{u_2}{u_1} > \frac{\bar{P}(M_{01})}{\bar{P}(M^*)}.$$

Se elige M_{00} en otro caso. Por lo tanto, la decisión de Bayes es:

$$\begin{aligned} \mathbf{d}^* &= \begin{cases} M_1, & \frac{\bar{P}(M_{01})}{\bar{P}(M_{00})} > \frac{u_0}{u_1} \text{ y } \frac{\bar{P}(M_{01})}{\bar{P}(M^*)} > \frac{u_2}{u_1} \\ A^*, & \frac{u_2}{u_0} > \frac{\bar{P}(M_{00})}{\bar{P}(M^*)} \text{ y } \frac{u_2}{u_1} > \frac{\bar{P}(M_{01})}{\bar{P}(M^*)} \\ M_0, & \text{e.o.c.} \end{cases} \\ &= \begin{cases} M_1, & \frac{\bar{P}(M_{01})}{\bar{P}(M_{00})} > \frac{u_0}{u_1} \text{ y } \frac{\bar{P}(M^*)}{\bar{P}(M_{01})} < \frac{u_1}{u_2} \\ A^*, & \frac{\bar{P}(M^*)}{\bar{P}(M_{00})} > \frac{u_0}{u_2} \text{ y } \frac{\bar{P}(M^*)}{\bar{P}(M_{01})} > \frac{u_1}{u_2} \\ M_0, & \text{e.o.c.} \end{cases} \end{aligned}$$

Tomando en cuenta que $\bar{P}(M^*) > \bar{P}(M_i^*) \forall i \implies$ si $\bar{P}(M^*) < x$, entonces $\bar{P}(M_i^*) < x$ y si $\bar{P}(M_i^*) > x$ para alguna i , entonces $\bar{P}(M^*) > x$, la decisión de Bayes en términos de M_i^* es:

$$\mathbf{d}^* = \begin{cases} M_1, & \frac{\bar{P}(M_1)}{\bar{P}(M_0)} > t_0 \text{ y } \frac{\bar{P}(M_i^*)}{\bar{P}(M_1)} < t_1, \forall i \\ A^*, & \text{p.a. } i \text{ se cumple } \frac{\bar{P}(M_i^*)}{\bar{P}(M_0)} > t_0 t_1 \text{ y } \frac{\bar{P}(M_i^*)}{\bar{P}(M_1)} > t_1 \\ M_0, & \text{e.o.c.} \end{cases}$$

$$\text{con } t_0 = u_0/u_1, \quad t_1 = u_1/u_2 \text{ y } t_0 t_1 = u_0/u_2.$$

En resumen, el método elige alguno de los siguientes efectos:

- Efecto general: si es mejor proporcionalmente que el modelo de efecto nulo y no hay modelos de subgrupos mejores que el de efecto total.
- Efectos en subgrupos: si son el mejor modelo dentro de su espacio y son mejores proporcionalmente que el modelo de efecto total y nulo.

- Efecto nulo: si ningún otro modelo es satisfactorio.

El método tiene como ventaja que es una aplicación directa de la Estadística Bayesiana, por lo que tiene los mismos fundamentos y han sido evaluados y respaldados a lo largo de muchos años, por diferentes científicos. Tiene como desventaja que utiliza modelos semi paramétricos. Son difíciles de programar y aplicar a datos simulados o reales. Es por esto que en esta tesis este método no se aplica a ninguna base de datos.

4.2. Método de Dixon y Simon

El método de Dixon y Simon (1991) es un método Bayesiano en el que se estiman los efectos del tratamiento para diferentes subgrupos a través de una regresión lineal. La inferencia de los subgrupos se hace a través una comparación entre los coeficientes de regresión. Se supone que el tratamiento y las covariables son dicotómicas. Los subgrupos se establecen antes de realizar el análisis y se definen sólo dos subgrupos por cada covariable.

4.2.1. Definición del modelo

Primero, considerar el modelo parcial con una sola covariable x_i que define a dos subgrupos (un subgrupo es cuando $x_i = 0$ y el otro cuando $x_i = 1$) para una observación i :

$$E(y_i) = \mu + \tau t_i + \beta x_i + \gamma t_i x_i,$$

donde μ es la respuesta de control,

- τ es la respuesta al tratamiento,
 β es la respuesta del subgrupo al control, y
 γ es la respuesta del subgrupo al tratamiento.

Se estiman los valores de μ, τ, β y γ . Para cada valor de la covariable x_i se puede estimar una respuesta y_i . Los valores que puede tomar $E[y_i]$ dependen de x_i y t_i y se resumen en el siguiente cuadro 4.4:

	$x_i = 1$	$x_i = 0$
$t_i = 1$	$\mu + \tau + \beta + \gamma$	$\mu + \tau$
$t_i = 0$	$\mu + \beta$	μ
Efecto de t.	$\tau + \gamma$	τ

Cuadro 4.4: Respuestas de subgrupos para una covariable.

Entonces, la diferencia de que un individuo i tenga el tratamiento $t_i = 1$ o $t_i = 0$ se da por los valores que se encuentran en la fila de efecto de t. También, se puede observar en el cuadro 4.4 que lo que marca la diferencia entre un subgrupo y otro es γ . Por lo tanto, si $\gamma \neq 0$ hay diferencia en subgrupos.

La misma idea se puede extender a más subgrupos definidos por un número m de covariables, considerando el siguiente modelo parcial para una observación i :

$$E(y_i) = \mu + \tau t_i + \sum_{j=1}^m (\beta_j x_{ij} + \gamma_j t_i x_{ij}),$$

donde β_j es la respuesta del subgrupo definido por la covariable j al control, y γ_j es la respuesta del subgrupo j al tratamiento.

El modelo es parcial ya que no se especifica la distribución de Y , sólo su esperanza. Normalmente, $Y \sim N(E[Y], V(Y))$ pero en el caso general la distribución de las y_i puede ser desconocida.

Al igual que en el método de Laud et. al. (2013), los subgrupos dependen de una covariable a la vez. Esto es, no se toman interacciones entre covariables, sólo la interacción con el tratamiento. Los valores que $E[y_i]$ puede tomar se resumen en el cuadro 4.5 (notar que como t_i y x_{ij} son dicotómicas, hay un número finito de valores de $E[y_i]$):

	$x_{ik} = 1$	$x_{ik} = 0$
$t_i = 1$	$\mu + \tau + \beta_k + \gamma_k + \sum_{j \neq k} x_{ij}(\beta_j + \gamma_j)$	$\mu + \tau + \sum_{j \neq k} x_{ij}(\beta_j + \gamma_j)$
$t_i = 0$	$\mu + \beta_k + \sum_{j \neq k} x_{ij}\beta_j$	$\mu + \sum_{j \neq k} x_{ij}\beta_j$
Efecto de t.	$\tau + \gamma_k + \sum_{j \neq k} x_{ij}\gamma_j$	$\tau + \sum_{j \neq k} x_{ij}\gamma_j$

Cuadro 4.5: Respuestas de subgrupos con múltiples covariables.

Como en este modelo hay más covariables, la inferencia se tiene que hacer sobre todas las γ_j al mismo tiempo. Por ejemplo, para ver si el tratamiento tiene efectos heterogéneos para el subgrupo definido por la covariable k , se necesita comparar $\tau + \gamma_k + \sum_{j \neq k} x_{ij}\gamma_j$ con $\tau + \sum_{j \neq k} x_{ij}\gamma_j$. Pero para $j \neq k$, x_{ij} no es necesariamente igual a 1 cuando $x_{ik} = 1$ y no es igual a 0 cuando $x_{ik} = 0$. Por lo tanto, hay más términos a comparar que en el modelo de una covariable y la interacción de γ_k con el tratamiento no es significativa a menos que $\gamma_j = 0$ para $j \neq k$. La solución que le dan Dixon y Simon a este problema, es crear la combinación lineal $\eta = \ell^T \theta$, con $\theta = (\mu, \tau, \beta_1, \dots, \beta_m, \gamma_1, \dots, \gamma_m)$, donde ℓ es un vector de elementos 0 y 1. Entonces, se utilizará θ para obtener las

distribuciones y se obtendrá un estimador de η con enfoque Bayesiano. El valor de ℓ es determinado por el investigador dependiendo del análisis que busca realizar. Se sugiere establecer ℓ de forma que se estimen las modas para los efectos del tratamiento vistos en el cuadro 4.5.

Además de dar los estimadores para los efectos del tratamiento para cada subgrupo, se puede buscar hacer un promedio ponderado de los efectos del tratamiento para cada variable que determina un subgrupo. Por ejemplo, si se busca analizar el efecto del tratamiento cuando $x_{i1} = 1$ y hay tres covariables, se calculan los siguientes estimadores (donde w_{uvt} es la proporción de pacientes tal que $x_{i1} = u$, $x_{i2} = v$ y $x_{i3} = t$, $u, v, t = \{0, 1\}$):

$$\begin{aligned} E[y_i|x_{i1} = 1, t_i = 1] &= w_{111}(\mu + \tau + \beta_1 + \gamma_1 + \beta_2 + \gamma_2 + \beta_3 + \gamma_3) + \\ &\quad w_{110}(\mu + \tau + \beta_1 + \gamma_1 + \beta_2 + \gamma_2) + \\ &\quad w_{101}(\mu + \tau + \beta_1 + \gamma_1 + \beta_3 + \gamma_3) + \\ &\quad w_{100}(\mu + \tau + \beta_1 + \gamma_1) \\ &= \mu + \tau + \beta_1 + \gamma_1 + (\beta_2 + \gamma_2)(w_{111} + w_{110}) + \\ &\quad \beta_3(w_{111} + w_{101}). \end{aligned}$$

$$\begin{aligned} E[y_i|x_{i1} = 1, t_i = 0] &= w_{111}(\mu + \beta_1 + \beta_2 + \beta_3) + w_{110}(\mu + \beta_1 + \beta_2) + \\ &\quad w_{101}(\mu + \beta_1 + \beta_3) + w_{100}(\mu + \beta_1) \\ &= \mu + \beta_1 + \beta_2(w_{111} + w_{110}) + \beta_3(w_{111} + w_{101}). \end{aligned}$$

$$\begin{aligned} \text{Efecto de } t \text{ cuando } x_1 = 1 &= w_{111}(\tau + \gamma_1 + \gamma_2 + \gamma_3) + w_{110}(\tau + \gamma_1 + \gamma_2) + \\ &\quad w_{101}(\tau + \gamma_1 + \gamma_3) + w_{100}(\tau + \gamma_1) \\ &= \tau + \gamma_1 + (\beta_2 + \gamma_2)(w_{111} + w_{110}) + \\ &\quad \beta_3(w_{111} + w_{101}). \end{aligned}$$

Estos estimadores se utilizan más adelante.

4.2.2. Probabilidades iniciales y final

Los pasos que toman Dixon y Simon para encontrar la distribución final de $\underline{\theta}$ son los siguientes:

1. Realizar un análisis inicial para estimar $\underline{\theta}$, tomando en cuenta que $\hat{\underline{\theta}}|\underline{\theta} \sim N(\underline{\theta}, C)$. $\hat{\underline{\theta}} = T(\underline{y})$ es una estadística suficiente en el sentido Bayesiano para \underline{y} (por eso se puede utilizar el modelo parcial) y C es una matriz positiva definida conocida. La estadística $T(\underline{y})$ es suficiente porque $P(\underline{\theta}|\underline{y})$ depende de \underline{y} sólo a través de $T(\underline{y})$. Esto es, se puede calcular $P(\underline{\theta}|\underline{y})$ utilizando $P(\underline{\theta}|T(\underline{y}))$.
2. Definir distribuciones iniciales necesarias para obtener $\underline{\theta}$.

Inicialmente:

- $\mu \sim N(\mu_0, \sigma_\mu^2)$,
- $\tau \sim N(0, \sigma_\tau^2)$,
- $\beta_j \sim N(0, \sigma_{\beta_j}^2)$, $j = 1, \dots, m$,
- $\gamma|\xi^2 \sim N(\underline{0}, \xi^2 I)$ (normal multivariada),
- $p(\xi^2) \propto [\max(\xi^2, \varepsilon)]^{-1}$.

La media inicial de los parámetros τ , β_j y γ es cero porque se busca ser neutral en cuanto la magnitud y sentido de los efectos de subgrupos. Además, posteriormente se tomará en cuenta que $\sigma_\mu^2, \sigma_\tau^2, \sigma_{\beta_j}^2 \rightarrow \infty$, $j = 1, \dots, m$ para tener distribuciones iniciales no informativas. Al aumentar la varianza, la distribución tiende a hacerse plana lo que da probabilidades similares a todos los valores.

La distribución inicial de $\underline{\gamma}$ está hecha de tal forma que haya intercambiabilidad entre efectos de subgrupos. Esto es, la distribución de $\underline{\gamma}$ no cambia si se permutan las γ_j (Lindley y Smith 1972; Spiegelhalter, Abrams y Myles 2004). Por lo tanto, inicialmente no se prefiere un modelo con efectos en un subgrupo que en los demás subgrupos. Es una distribución jerárquica en el sentido que depende de la distribución de ξ^2 .

La distribución inicial de ξ^2 está basada en la distribución de Jeffreys $(\xi^2)^{-1}$. Es de segundo nivel ya que la distribución inicial de $\underline{\gamma}$ está condicionada sobre ξ^2 . La modificación que se hace a $(\xi^2)^{-1}$ es con motivo de obtener una densidad propia final para $\underline{\theta}$. Se toma $\xi^2 \leq M$, con M muy grande.

$$\begin{aligned} \text{Si } f(\xi^2) &= \frac{c}{\xi^2}, \xi^2 \in (0, M), \text{ donde } c \text{ no depende de } \xi^2, \\ \implies \int_0^M f(\xi^2) d\xi^2 &= \int_0^M \frac{c}{\xi^2} d\xi^2 \\ &= c \ln |\xi^2| \Big|_{\xi^2=0}^{\xi^2=M} \\ &= c \ln \xi^2 \Big|_{\xi^2=0}^{\xi^2=M} \text{ porque } \xi^2 \text{ siempre es positivo} \end{aligned}$$

Como $\lim_{\xi^2 \rightarrow 0} c \ln \xi^2 = -\infty$,

$$\begin{aligned} \implies \int_0^M f(\xi^2) d\xi^2 &\text{ no converge} \\ \implies f(\xi^2) &\text{ no es propia.} \end{aligned}$$

$$\text{Pero si } f(\xi^2) = \frac{c}{\max(\xi^2, \varepsilon)}$$

$$\begin{aligned}
 \implies \int_0^M f(\xi^2) d\xi^2 &= \int_0^\varepsilon f(\xi^2) d\xi^2 + \int_\varepsilon^M f(\xi^2) d\xi^2 \\
 &= \int_0^\varepsilon \frac{c}{\varepsilon} d\xi^2 + \int_\varepsilon^M \frac{c}{\xi^2} d\xi^2 \\
 &= \frac{c}{\varepsilon} \int_0^\varepsilon 1 d\xi^2 + c \int_\varepsilon^M \frac{1}{\xi^2} d\xi^2 \\
 &= \frac{c}{\varepsilon} (\xi^2) \Big|_{\xi^2=0}^{\xi^2=\varepsilon} + c (\ln |\xi^2|) \Big|_{\xi^2=\varepsilon}^{\xi^2=M} \\
 &= \frac{c}{\varepsilon} \varepsilon + c \ln |M| - c \ln |\varepsilon| \\
 &= c(1 + \ln M - \ln \varepsilon).
 \end{aligned}$$

La cual es propia con $c = 1/(1 + \ln M - \ln \varepsilon)$. En particular, Dixon y Simon toman $\varepsilon = 0.005$.

3. Obtener distribución final de $\underline{\theta}$ con las el Teorema de Bayes:

$$\begin{aligned}
 f(\underline{\theta}|y) &= f(\underline{\theta}|\hat{\theta}) \text{ porque } \hat{\theta}(y) \text{ es estadística suficiente de } \underline{\theta} \\
 &= \frac{f(\hat{\theta}, \underline{\theta})}{f(\hat{\theta})} \\
 &= \frac{\int f(\underline{\theta}, \hat{\theta}|\xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}.
 \end{aligned}$$

Como $f(\underline{\theta}, \hat{\theta}|\xi^2) = f(\underline{\theta}|\hat{\theta}, \xi^2) f(\hat{\theta}|\xi^2)$,

$$\implies f(\underline{\theta}|\hat{\theta}) = \frac{\int f(\underline{\theta}|\hat{\theta}, \xi^2) f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}.$$

En otras palabras,

$$q(\underline{\theta}|\hat{\theta}) = \frac{\int q(\underline{\theta}|\hat{\theta}, \xi^2) g(\hat{\theta}|\xi^2) p(\xi^2) d\xi^2}{\int g(\hat{\theta}|\xi^2) p(\xi^2) d\xi^2},$$

con p densidad inicial, g verosimilitud y q densidad final.

La distribución final $f(\underline{\theta}|\hat{\theta}) = \frac{\int f(\underline{\theta}|\hat{\theta}, \xi^2) f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}$ es propia a pesar del

uso de impropias iniciales.

4.2.3. Simplificaciones

Para calcular la distribución final, Dixon y Simon realizan varias simplificaciones. Entre ellas está utilizar los resultados matriciales de Lindley y Smith (1972) para calcular $f(\hat{\theta}|\xi^2)$ y $f(\theta|\hat{\theta}, \xi^2)$. Además, se sustituye $f(\hat{\theta}|\xi^2)$ por $f(\hat{\gamma}|\xi^2)$. El uso de las simplificaciones disminuye los cálculos computacionales. Cabe mencionar que el método fue creado en 1991, por lo que era importante reducir el trabajo computacional para asegurar su funcionamiento. En la actualidad no es necesario hacer tantas simplificaciones pero se explicarán para ratificar el entendimiento del método.

De Lindley y Smith (1972) tenemos el siguiente teorema:

Teorema 4.2.1. *Suponer $y|\theta_1 \sim N(A_1\theta_1, C_1)$, normal multivariada donde θ_1 tiene p_1 parámetros, C_1 es positiva semidefinida. Por otro lado, suponer $\theta_1|\theta_2 \sim N(A_2\theta_2, C_2)$, normal multivariada donde θ_2 tiene p_2 hiperparámetros, C_2 es positiva semidefinida. Entonces, $y|\theta_2 \sim N(A_1A_2\theta_2, C_1 + A_1C_2A_2^T)$, $\theta_1|y, \theta_2 \sim N(Bb, B)$, con $B^{-1} = A_1^TC_1^{-1}A_1 + C_2^{-1}$ y $b = A_1C_1^{-1}y + C_2^{-1}A_2\theta_2$.*

Se tiene que $\underline{\theta} = (\mu, \tau, \beta_1, \dots, \beta_m, \gamma_1, \dots, \gamma_m)$, pero a su vez estos son parámetros desconocidos que dependen de distribuciones normales. Entonces, condicionalmente:

$$\underline{\theta}|(\mu, \tau, \beta_1, \dots, \beta_m, \gamma_1, \dots, \gamma_m)^T, \xi^2 \sim N(\mu^*, \sigma^{*2}),$$

$$\text{donde } \mu^* = \begin{bmatrix} \mu_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ y } \sigma^{*2} = \begin{bmatrix} \sigma_\mu^2 & 0 & \dots & \dots & \dots & 0 \\ 0 & \sigma_\tau^2 & 0 & \dots & \dots & 0 \\ 0 & 0 & \sigma_{\beta_1}^2 & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_{\beta_2}^2 & 0 & \dots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \sigma_{\beta_m}^2 & 0 \\ 0 & \dots & \dots & 0 & 0 & I\xi^2 \end{bmatrix}.$$

Se toma

$$y = \hat{\theta}|\xi^2, \quad \theta_1 = \underline{\theta}|\xi^2$$

$$A_1 = I, \quad A_2 = I,$$

$$C_1 = C, \quad C_2 = \sigma^{*2},$$

$$\theta_2 = \mu^*.$$

El Teorema 4.2.1 implica que

$$\begin{aligned} \hat{\theta}|\xi^2 &\sim N(\mu^*, C + \sigma^{*2}), \\ y \underline{\theta}|\hat{\theta}, \xi^2 &\sim N(Bb, B), \\ \text{con } B \text{ tal que } B^{-1} &= C^{-1} + (\sigma^{*2})^{-1}, \\ y b &= C^{-1}\hat{\theta} + (\sigma^{*2})^{-1}\mu^*. \end{aligned}$$

Como C es semipositiva definida tiene inversa y como σ^{*2} es una matriz diagonal, $(\sigma^{*2})^{-1} = \text{diag} \left[\frac{1}{(\sigma^{*2})_{ii}} \right]$. Por lo tanto,

$$\begin{aligned} B^{-1} &= C^{-1} + \text{diag} \left[\frac{1}{(\sigma^{*2})_{ii}} \right] \\ b &= C^{-1}\hat{\theta} + \left[\frac{\mu_0}{\sigma_{\mu_0}^2}, 0, 0, \dots, 0 \right]^T. \end{aligned}$$

Para $\hat{\theta}|\xi^2$, se analiza que pasa cuando $\sigma_\mu^2, \sigma_\tau^2, \sigma_{\beta_j}^2 \rightarrow \infty$, $j = 1, \dots, m$, empezando por σ_μ^2 . Recordemos que la función de densidad de $\hat{\theta}|\xi^2$ por ser una normal multivariada es

$$f(\hat{\theta}|\xi^2) = \det(2\pi(C + \sigma^{*2}))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\hat{\theta} - \mu^*)^T(C + \sigma^{*2})^{-1}(\hat{\theta} - \mu^*)\right\}.$$

En esta ecuación, σ_μ^2 solo está presente en la matriz de covarianzas $H = C + \sigma^{*2}$, por lo que nos enfocamos en ver el comportamiento de H cuando $\sigma_\mu^2 \rightarrow \infty$.

$$H = C + \sigma^{*2}$$

$$= \begin{bmatrix} c_{11} + \sigma_\mu^2 & c_{21} & c_{31} & \dots & c_{k1} \\ c_{21} & c_{22} + \sigma_\tau^2 & c_{32} & \dots & c_{k2} \\ c_{31} & c_{32} & c_{33} + \sigma_{\beta_1}^2 & \dots & c_{k3} \\ \vdots & \vdots & \ddots & & \vdots \\ c_{k1} & c_{k2} & \dots & & c_{kk} + \xi^2 \end{bmatrix}, k = 2m + 2$$

$$= \begin{bmatrix} c_{11} + \sigma_\mu^2 & h^T \\ h & H_1 \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} + \sigma_\mu^2 & \mathbf{0}_k^T \\ h & I_{k-1} \end{bmatrix} \begin{bmatrix} 1 & \frac{h^T}{c_{11} + \sigma_\mu^2} \\ \mathbf{0}_k^T & H_1 - \frac{hh^T}{c_{11} + \sigma_\mu^2} \end{bmatrix}, \text{ donde}$$

$$H_1 = \begin{bmatrix} c_{22} + \sigma_\tau^2 & c_{32} & \dots & c_{k2} \\ c_{32} & c_{33} + \sigma_{\beta_1}^2 & & c_{k3} \\ \vdots & \ddots & & \vdots \\ c_{k2} & \dots & & c_{kk} + \xi^2 \end{bmatrix},$$

$$h = \begin{bmatrix} c_{21} \\ c_{31} \\ \vdots \\ c_{k1} \end{bmatrix}.$$

Se utiliza $|\cdot|$ para denotar el determinante y se descompone la matriz de covarianzas en bloques para facilitar el cálculo del determinante y la inversa de H , ya que son los elementos presentes en la densidad de probabilidad de $\hat{\theta}|\xi^2$. Se aísle σ_μ^2 para evaluar H cuando σ_μ^2 tiende a infinito.

Notar que ninguno de los elementos del vector h y de la matriz H_1 contiene a μ . Incluso la matriz H_1 se puede expresar $H_1 = C_R + \sigma_R^{*2}$ donde C_R es una matriz positiva definida que muestra las interacciones entre los parámetros $\tau, \beta_j, j = 1, \dots, m$ y γ (pero no μ). Por otro lado, σ_R^{*2} es una matriz con las varianzas de los parámetros anteriores en la diagonal y el resto de los elementos son cero. H_1 es la matriz de covarianzas reducida para los parámetros quitando μ . Por propiedades de las matrices se obtiene:

$$\begin{aligned} |H| &= |c_{11} + \sigma_\mu^2| |I_{k-1}| \cdot |1| \left| H_1 - \frac{hh^T}{c_{11} + \sigma_\mu^2} \right|, \\ &= (c_{11} + \sigma_\mu^2) \left| H_1 - \frac{hh^T}{c_{11} + \sigma_\mu^2} \right| \\ \text{y } H^{-1} &= \left[\begin{array}{c|c} L^{-1} & -L^{-1}h^T H_1^{-1} \\ \hline -H_1^{-1}hL^{-1} & H_1^{-1} + H_1^{-1}hL^{-1}h^T H_1^{-1} \end{array} \right], \\ \text{con } L^{-1} &= \frac{1}{c_{11} + \sigma_\mu^2 - h^T H_1^{-1}h} \in \mathbb{R}. \end{aligned}$$

Por lo tanto, la densidad de $\hat{\theta}|\xi^2$ es:

$$f(\hat{\theta}|\xi^2) = \left((2\pi)^k (c_{11} + \sigma_\mu^2) \left| H_1 - \frac{hh^T}{c_{11} + \sigma_\mu^2} \right| \right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\hat{\theta} - \mu^*)^T H^{-1}(\hat{\theta} - \mu^*) \right\}$$

Aunque lo que interesa es el límite de la densidad de $\hat{\theta}|\xi^2$ se sustituye $f(\hat{\theta}|\xi^2)$ en la distribución final de $\underline{\theta}|\hat{\theta}$. De este modo, se puede eliminar el término $c_{11} + \sigma_\mu^2$ de los numeradores y cuando se calcule el límite, más términos tenderán a cero. Cabe mencionar que $c_{11} + \sigma_\mu^2$ se puede sacar de la integral como una constante para luego cancelarlo porque no dependen de ξ^2 .

$$\begin{aligned} & \frac{\int f(\underline{\theta}|\hat{\theta}, \xi^2) f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2} = \\ &= \frac{\int |2\pi H|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int |2\pi H|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\xi^2) d\xi^2} \\ &= \frac{\int ((2\pi)^k (c_{11} + \sigma_\mu^2) |H|)^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int ((2\pi)^k (c_{11} + \sigma_\mu^2) |H|)^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\xi^2) d\xi^2} \\ &= \frac{((2\pi)^k (c_{11} + \sigma_\mu^2))^{-\frac{1}{2}} \int |H|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{((2\pi)^k (c_{11} + \sigma_\mu^2))^{-\frac{1}{2}} \int |H|^{-\frac{1}{2}} e^{-\frac{1}{2}(\hat{\theta}-\mu^*)^T H^{-1}(\hat{\theta}-\mu^*)} f(\xi^2) d\xi^2} \\ &= \frac{\int |H|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\hat{\theta} - \mu^*)^T H^{-1}(\hat{\theta} - \mu^*) \right\} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int |H|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\hat{\theta} - \mu^*)^T H^{-1}(\hat{\theta} - \mu^*) \right\} f(\xi^2) d\xi^2}. \end{aligned} \tag{4.15}$$

Se procede a calcular los siguientes límites:

$$\begin{aligned} & \lim_{\sigma_\mu^2 \rightarrow \infty} |H| \\ & \lim_{\sigma_\mu^2 \rightarrow \infty} (\hat{\theta} - \mu^*)^T H^{-1}(\hat{\theta} - \mu^*) \end{aligned}$$

Es fácil ver que

$$\lim_{\sigma_\mu^2 \rightarrow \infty} |H| = \lim_{\sigma_\mu^2 \rightarrow \infty} \left| H_1 - \frac{hh^T}{c_{11} + \sigma_\mu^2} \right| = |H_1|,$$

pero para el otro límite conviene tomar en cuenta que

$$\begin{aligned}\hat{\underline{\theta}} - \mu^* &= \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \\ \vdots \\ \hat{\theta}_k \end{bmatrix} - \begin{bmatrix} \mu_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \hat{\theta}_1 - \mu_0 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \\ \vdots \\ \hat{\theta}_k \end{bmatrix} \\ &= \begin{bmatrix} \hat{\theta}_1 - \mu_0 \\ \vdots \\ \hat{\theta}_R \end{bmatrix}. \\ \hat{\underline{\theta}}_R &\sim N(0, \sigma_R^{*2}),\end{aligned}$$

donde $\hat{\underline{\theta}}_R$ es el vector de parámetros estimados reducido que no estima a μ . Entonces,

$$\begin{aligned}
 & (\hat{\underline{\theta}} - \mu^*)^T H^{-1} (\hat{\underline{\theta}} - \mu^*) = \\
 &= \left[\begin{array}{c|c} \hat{\theta}_1 - \mu_0 & \hat{\underline{\theta}}_R^T \end{array} \right] \left[\begin{array}{c|c} L^{-1} & -L^{-1} h^T H_1^{-1} \\ \hline -H_1^{-1} h L^{-1} & H_1^{-1} + H_1^{-1} h L^{-1} h^T H_1^{-1} \end{array} \right] \left[\begin{array}{c} \hat{\theta}_1 - \mu_0 \\ \hat{\underline{\theta}}_R \end{array} \right] \\
 &= \left[\begin{array}{c} (\hat{\theta}_1 - \mu_0) L^{-1} - \hat{\underline{\theta}}_R^T H_1^{-1} h L^{-1} \\ \hline -(\hat{\theta}_1 - \mu_0) L^{-1} h^T H_1^{-1} + \hat{\underline{\theta}}_R^T (H_1^{-1} + H_1^{-1} h L^{-1} h^T H_1^{-1}) \end{array} \right]^T \left[\begin{array}{c} \hat{\theta}_1 - \mu_0 \\ \hat{\underline{\theta}}_R \end{array} \right] \\
 &= (\hat{\theta}_1 - \mu_0)^2 L^{-1} - (\hat{\theta}_1 - \mu_0) \hat{\underline{\theta}}_R^T H_1^{-1} h L^{-1} \\
 &\quad - (\hat{\theta}_1 - \mu_0) L^{-1} h^T H_1^{-1} \hat{\underline{\theta}}_R + \hat{\underline{\theta}}_R^T (H_1^{-1} + H_1^{-1} h L^{-1} h^T H_1^{-1}) \hat{\underline{\theta}}_R
 \end{aligned}$$

Para calcular el límite, tomamos en cuenta que

$$\lim_{\sigma_\mu^2 \rightarrow \infty} L^{-1} = \lim_{\sigma_\mu^2 \rightarrow \infty} \frac{1}{c_{11} + \sigma_\mu^2 - h^T H_1^{-1} h} = 0.$$

Entonces, todos los términos de la suma que tengan L^{-1} tienden a cero cuando σ_μ^2 tiende a infinito. Por lo tanto,

$$\lim_{\sigma_\mu^2 \rightarrow \infty} (\hat{\underline{\theta}} - \mu^*)^T H^{-1} (\hat{\underline{\theta}} - \mu^*) = \hat{\underline{\theta}}_R^T H_1^{-1} \hat{\underline{\theta}}_R.$$

Se procede a calcular el límite de la distribución posterior de $\underline{\theta}|\hat{\theta}$ utilizando el resultado de la ecuación (4.15):

$$\begin{aligned}
 & \lim_{\sigma_\mu^2 \rightarrow \infty} \frac{\int f(\underline{\theta}|\hat{\theta}, \xi^2) f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2} = \\
 &= \lim_{\sigma_\mu^2 \rightarrow \infty} \frac{\int |H|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\hat{\theta} - \mu^*)^T H^{-1} (\hat{\theta} - \mu^*)\} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int |H|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\hat{\theta} - \mu^*)^T H^{-1} (\hat{\theta} - \mu^*)\} f(\xi^2) d\xi^2} \\
 &= \frac{\int |H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\underline{\theta}}_R^T H_1^{-1} \hat{\underline{\theta}}_R\} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int |H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\underline{\theta}}_R^T H_1^{-1} \hat{\underline{\theta}}_R\} f(\xi^2) d\xi^2}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\int ((2\pi)^{k-1})^{-\frac{1}{2}} |H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\theta}_R^T H^{-1} \hat{\theta}_R\} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int ((2\pi)^{k-1})^{-\frac{1}{2}} |H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\theta}_R^T H^{-1} \hat{\theta}_R\} f(\xi^2) d\xi^2} \\
 &= \frac{\int |2\pi H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\theta}_R^T H^{-1} \hat{\theta}_R\} f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int |2\pi H_1|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\hat{\theta}_R^T H^{-1} \hat{\theta}_R\} f(\xi^2) d\xi^2} \\
 &= \frac{\int f(\hat{\theta}_R|\xi^2) f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}_R|\xi^2) f(\xi^2) d\xi^2}.
 \end{aligned}$$

Entonces, se observa en el límite resultante que la densidad final deja de depender de μ y los parámetros μ_0, σ_μ^2 si se hace tender σ_μ^2 a infinito. Análogamente, se puede ver que si se hace tender a $\sigma_\tau^2, \sigma_{\beta_j}^2, j = 1, \dots, m$ a infinito

$$\lim_{\sigma_\mu^2 \rightarrow \infty} \frac{\int f(\hat{\theta}|\xi^2) f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\theta}|\xi^2) f(\xi^2) d\xi^2} = \frac{\int f(\hat{\gamma}|\xi^2) f(\underline{\theta}|\hat{\theta}, \xi^2) f(\xi^2) d\xi^2}{\int f(\hat{\gamma}|\xi^2) f(\xi^2) d\xi^2},$$

donde $\hat{\gamma} \sim N(0, I\xi^2)$. Es decir, el declarar distribuciones iniciales no informativas resulta en que no se influyan en la distribución final de $\underline{\theta}|\hat{\theta}$.

4.2.4. Inferencia

Una vez obtenida la distribución final simplificada se hace inferencia sobre $\underline{\theta}$. Los resultados se reportan a través de la moda de cada valor de $\underline{\theta}$ y el intervalo de mayor densidad posterior al 95 %, aunque no son los únicos resúmenes posibles.

Se debe mencionar que si hay subgrupos de interés *a priori* se deben considerar como no intercambiables. Esto elimina la posibilidad de utilizar los resultados de Lindley y Smith, y se requieren más cálculos computacionales. En este caso se utilizan distribuciones normales como

iniciales, pero no son las únicas posibles. De hecho, si hay alta correlación entre covariables se sugiere establecer distribuciones iniciales informativas, en las que se tome en cuenta la correlación.

Asimismo, una desventaja del método es que sólo se pueden predefinir dos subgrupos por covariable y las covariables se tienen que transformar a binarias si son categóricas o continuas, lo que puede dificultar la definición de los subgrupos.

Por otro lado, gracias a que se basa en una regresión lineal, el funcionamiento del método se puede intuir si se tiene conocimiento previo de Estadística y es posible modificar el método. Aunado con el hecho de que fue publicado hace más de 20 años, ha dado como resultado que se ha utilizado como base para distintos métodos. Por ejemplo, se ha adaptado a una regresión logística o incluso a un modelo de riesgos proporcionales. Ejemplos de modificaciones publicadas son los métodos de Simon (2002) y de Jones et al. (2011).

4.2.5. Modificación con *JAGS*

En el modelo anterior, la matriz C y el vector $\hat{\theta}$ son conocidos, o se han obtenido anteriormente de forma analítica. Alternativamente, se puede definir el modelo completo y no sólo la esperanza y hacer inferencia sobre él. Por ejemplo, suponer

$$y_i \sim N(\mu_i, \frac{1}{\rho}), \text{ donde} \quad (4.16)$$

$$\mu_i = E[y_i] = \mu + \tau t_i + \sum_{j=1}^m (\beta_j x_{ij} + \gamma_j t_i x_{ij}),$$

con ρ un parámetro desconocido, y

m el número de covariables que determinan a los subgrupos.

En este caso, se obtiene la distribución final de la siguiente forma:

$$f(\underline{\theta}|\underline{y}) = \frac{f(\underline{y}|\underline{\theta})f(\underline{\theta})}{f(\underline{y})},$$

donde la verosimilitud es simplemente la conjunta $f(\underline{y}|\underline{\theta}, \phi) = \prod_{i=1}^n f(y_i|\mu_i, \phi) = \prod_{i=1}^n \sqrt{\frac{\phi}{2\pi}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\phi}\right\}$. Estableciendo una distribución inicial para $\underline{\theta}$ y ϕ , se puede hacer inferencia Bayesiana sobre $\underline{\theta}$ dado \underline{y} .

Para las distribuciones iniciales de $\underline{\theta}$, se pueden elegir las definidas por Dixon y Simon. De igual forma, para ϕ se puede asignar una distribución no informativa, por ejemplo $f(\phi) \propto \frac{1}{\phi}$, $\phi > 0$.

Una de las formas de hacer inferencia sobre el modelo es a través paquetes computacionales modernos como *JAGS* (Plummer 2013), explicado en la sección 2.3.1. No obstante, las distribuciones iniciales no informativas impropias no se pueden utilizar en *JAGS*. Por lo tanto, se utilizan aproximaciones, en particular se utilizan para $\underline{\theta}$ son

$$\begin{aligned} \mu &\sim N(0, 0.0001), \\ \tau &\sim N(0, 0.0001), \\ \beta_j &\sim N(0, 0.0001), j \in \{1, \dots, m\}, \\ \gamma_j &\sim N(0, v), j \in \{1, \dots, m\}, \\ v &= \frac{1}{\xi^2}, \\ \xi^2 &= \exp\{\rho\}, \\ \rho &\sim \text{Unif}(\log(1e-10), \log(1e-10)). \end{aligned} \tag{4.17}$$

Las varianzas de los parámetros son cercanas a cero para simular iniciales no informativas. En el caso de v , se busca aproximar la inicial de Jeffreys

para la varianza. Por otro lado, para ϕ se define

$$\phi \sim \text{Gamma}(0.001, 0.001),$$

la cual aproxima la distribución $f(\phi) \propto \frac{1}{\phi}$.

En capítulos siguientes se aplicará este modelo, además de los vistos anteriormente a diferentes bases de datos.

Capítulo 5

Ejemplos ilustrativos con datos simulados

Este capítulo está dedicado a la aplicación de varios métodos expuestos en los Capítulos 3 y 4: *Virtual Twins*, el método de árboles de interacción, el método de Dixon y Simon y su modificación de *JAGS* a una serie de datos simulados. La base de datos se simula de tal forma que la respuesta al tratamiento es diferente que para el promedio para subgrupos conocidos. Para los métodos exploratorios (los primeros dos), se busca analizar si los métodos encuentran el subgrupo para el cual hay heterogeneidad en el tratamiento. En cambio, para los métodos confirmatorios se evalúa su capacidad para estimar las diferencias en los efectos del tratamiento. No se aplica el método de selección Bayesiana de un modelo basado en la urna de Pòlya de Laud et al. (2013) porque al ser un método que utiliza distribuciones no paramétricas, su implementación se dificulta.

Para la simulación de los datos y la aplicación de los métodos se

utilizó como software R (R Core Team 2018) a través de RStudio (RStudio Team 2016). El código para crear los datos simulados se obtuvo de *Biopharmaceutical Network* (Lipkovich 2018), una organización que promueve el estudio y uso de software para análisis en ensayos clínicos. En esta página, autores de diferentes métodos de análisis de subgrupos, entre ellos los autores del método de árboles de interacción, Su et al. (2009), comparten los códigos de sus métodos para la utilización de cualquier usuario de forma gratuita.

Se obtuvo una muestra aleatoria con 1000 observaciones del modelo adquirido en *Biopharmaceutical Network* (Lipkovich 2018) del código de simulación de Su et al. (2009). El modelo es muy parecido al expuesto por Dixon y Simon (1991) y se describe a continuación:

$$\begin{aligned} y_i = & \mu + \tau t_i + \beta_1 I(x_{1i} \leq c_1) + \beta_2 I(x_{2i} \leq c_2) + \\ & \gamma_1 t_i I(x_{1i} \leq c_1) t_i + \gamma_2 I(x_{2i} \leq c_2) t_i + e, \end{aligned} \quad (5.1)$$

donde

$$\begin{aligned} e &\sim N(0, \sigma^2), \\ t_i &\sim \text{Bernoulli}\left(\frac{1}{2}\right), \\ x_{ji} &\sim \frac{\text{Uniforme discreta}\{1, U\}}{U}, j = 1, 2, 3, 4 \\ U &\geq 2, U \in \mathbb{N}. \end{aligned}$$

La respuesta al tratamiento para el paciente i es y_i . Cada paciente se asigna aleatoriamente al grupo de control o al grupo del tratamiento a través de la variable aleatoria t_i . Un paciente tiene la misma probabilidad de estar en el grupo control que en el tratamiento. Si $t_i = 0$, el paciente está en el grupo control y si $t_i = 1$ está en el grupo con tratamiento.

Por otro lado, se crean las covariables x_{ji} con $j \in \{1, 2, 3, 4\}$. Estas covariables provienen de una distribución uniforme discreta que toma valores en el rango de enteros de 1 a U . La razón para hacer una muestra aleatoria discreta es que, en ensayos clínicos, la mayoría de las variables son de tipo categórico (por ejemplo, edad, sexo, raza, etapa del cáncer, etcétera). Por lo tanto, hay individuos que tienen valores en las covariables iguales. Si se utilizara una distribución uniforme continua, sería improbable obtener una muestra donde algunos de los valores en las covariables fueran iguales. Después obtener la muestra aleatoria de la variable uniforme discreta, esta se divide entre U con el motivo de que $x_{ji} \in (\frac{1}{U}, 1]$. Las covariables se utilizarán para determinar los subgrupos con efectos heterogéneos.

Para determinar si un paciente está en el subgrupo j , se utilizan los valores de umbral c_j . Los valores de umbral dividen a los pacientes en 2 subgrupos dependiendo del valor que toma una covariable. Es decir, si $x_{ji} \leq c_j$ entonces el tratamiento será diferente para el paciente i . En esta simulación sólo se aparecen los valores c_1 y c_2 en el modelo. Por lo tanto, hay 4 subgrupos determinados por las covariables x_1 y x_2 .

El resto de los parámetros que determinan el modelo son los siguientes: μ la media de control que determina el efecto que tienen todos los pacientes independientemente de obtener el tratamiento o no, τ la respuesta al tratamiento de los pacientes, β_j la respuesta de control del subgrupo determinado por x_j y γ_j la respuesta al tratamiento de los pacientes del subgrupo determinado por x_j .

En este caso, se utilizaron los siguientes valores de los parámetros para simular los datos:

$$n = 10000,$$

$$\begin{array}{ll} \mu = 2, & \tau = 4, \\ \beta_1 = 2, & \beta_2 = 0, \\ \gamma_1 = 4, & \gamma_2 = -2, \\ c_1 = 0.3, & c_2 = 0.3, \\ \sigma^2 = 0, & U = 50. \end{array}$$

Esto quiere decir que el control tiene un efecto positivo, es decir que se supone respuesta positiva inicial para todos los individuos que participaron en el ensayo clínico. También, el tratamiento tiene un efecto positivo en la respuesta de los individuos. Por otro lado, los pacientes que estén en el subgrupo donde $x_1 \leq c_1$ tendrán un efecto positivo control mayor a los demás y también el tratamiento tendrá un efecto positivo mayor que a los demás pacientes. Este subgrupo deberá estar compuesto por el 30% de los pacientes. El otro subgrupo, del mismo tamaño, contiene a los pacientes cuya covariable $x_2 \leq c_2$. En este subgrupo el tratamiento tiene un efecto negativo. Es decir, a este subgrupo el tratamiento le hace daño. Las covariables x_3 y x_4 no tienen efecto sobre la respuesta. Esto es, $\beta_3 = \beta_4 = \gamma_3 = \gamma_4 = 0$.

Haciendo un análisis exploratorio a la muestra aleatoria, se obtienen los siguientes resúmenes, mostrados en el cuadro 5.1:

Covariable	Resumen o Clase	Observaciones
Grupo	Control	5057
	Tratamiento	4943
x_1	$x_1 \leq c_1$	3045
	$x_1 > c_1$	6955
	Mínimo	0.0200
	Máximo	1
	Media	0.5063
x_2	$x_2 \leq c_2$	3012
	$x_2 > c_2$	6988
	Mínimo	0.0200
	Máximo	1
	Media	0.5106
x_3	Mínimo	0.0200
	Máximo	1
	Media	0.5028
x_4	Mínimo	0.0200
	Máximo	1
	Media	0.5098
Respuesta	Mínimo	-1.315
	Máximo	14.740
	Media	4.913

Cuadro 5.1: Análisis exploratorio para datos simulados.

Los individuos asignados al grupo control y de tratamiento fueron balanceados, casi el 50 % de los pacientes se asignaron a cada grupo.

Por otro lado, en promedio, cada valor de cada covariante x_{ij} se obtuvo 200 veces en la muestra. Las covariantes x_{ij} pueden tomar $U = 50$ valores diferentes. Todos los valores posibles salieron en la muestra al menos una vez. Como era de esperarse, las muestras de x_{ij} fueron muy parecidas. Esto porque las covariantes siguen la misma distribución.

En cuanto a la variable de respuesta, se tiene que su media es de 4.913. Hubo respuestas negativas a pesar de que de inicio todas las pacientes tienen una respuesta $\mu = 2$. Los valores negativos se pueden obtener si hay un e negativo y si el paciente pertenece subgrupo determinado por x_2 .

En la figura 5.1 se muestra la distribución de los pacientes de acuerdo al tratamiento y a los subgrupos.

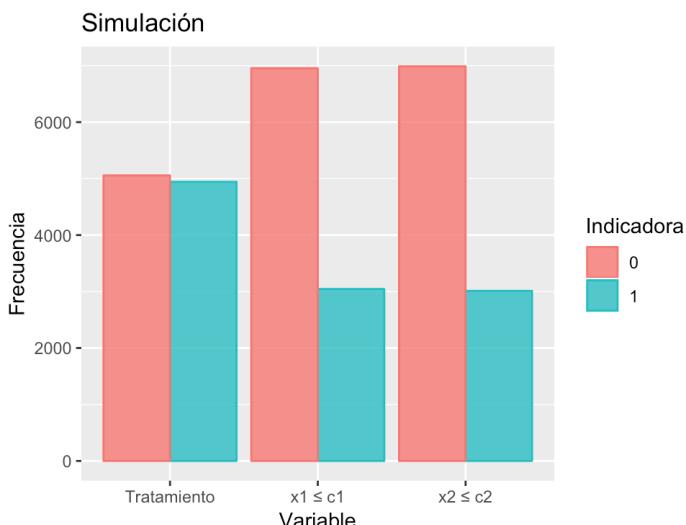


Figura 5.1: Frecuencia de pacientes con respecto a las variables.

Con la figura anterior 5.1 se puede visualizar que la probabilidad de

que un individuo estuviera en el grupo de control era la misma que en el tratamiento. Asimismo, se puede ver en la gráfica como los valores de umbral c_1 y c_2 dividen a la muestra.

Otro análisis interesante es la distribución que tiene la variable de respuesta con respecto al tratamiento y a las covariables que definen los subgrupos. En la figura 5.2 se muestra la frecuencia de los valores de respuesta. El color rojo representa a los pacientes en el grupo control y el azul en el tratamiento.

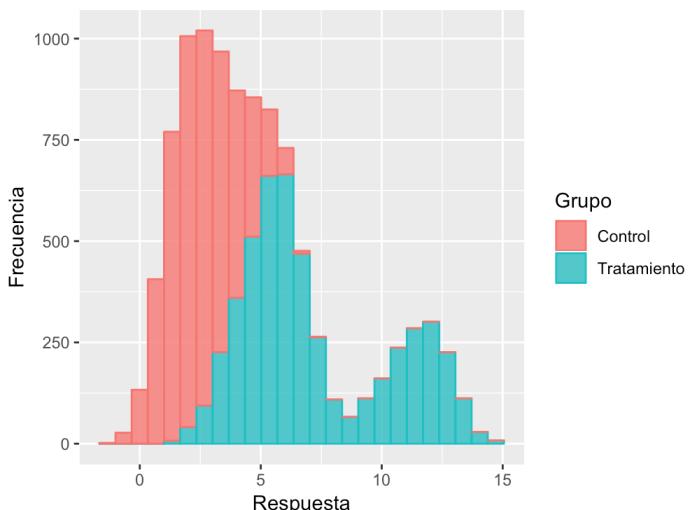


Figura 5.2: Respuesta del grupo control y tratamiento.

Se puede ver en la figura 5.2 que en general los individuos del grupo control obtuvieron respuestas menores que los del tratamiento. Las observaciones de color rojo se encuentran en el lado izquierdo de la gráfica mientras que las azules al derecho. Esto era de esperarse por la forma en que se definió el modelo.

La figura 5.3 representa a la frecuencia de las respuestas con respecto a los subgrupos definidos por la covariable x_1 . Es claro que los individuos cuya covariable $x_1 \leq c_1$ tienen en general resultados mayores.

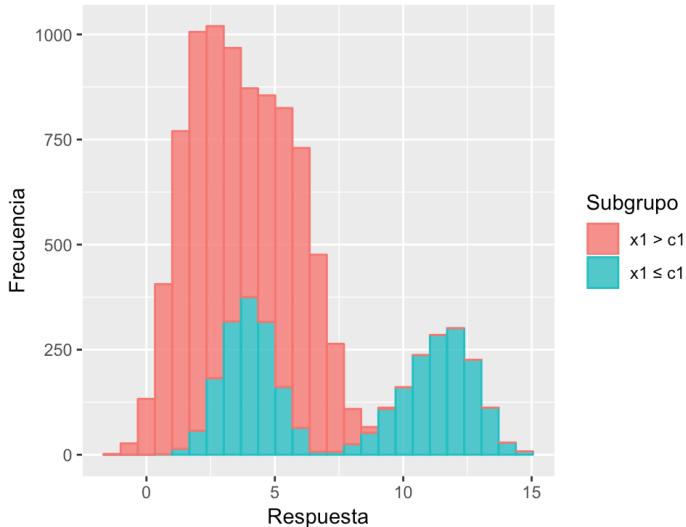


Figura 5.3: Respuesta de los subgrupos determinados por x_1 .

Por último, se grafica la respuesta de los subgrupos determinados por x_2 . En general, tienen resultados menores como era de esperarse. En la figura 5.4 se aprecia que los individuos con $x_2 \leq c_2$ se concentran al lado izquierdo de la gráfica. Es decir, tienen menos probabilidades de obtener un resultado positivo.

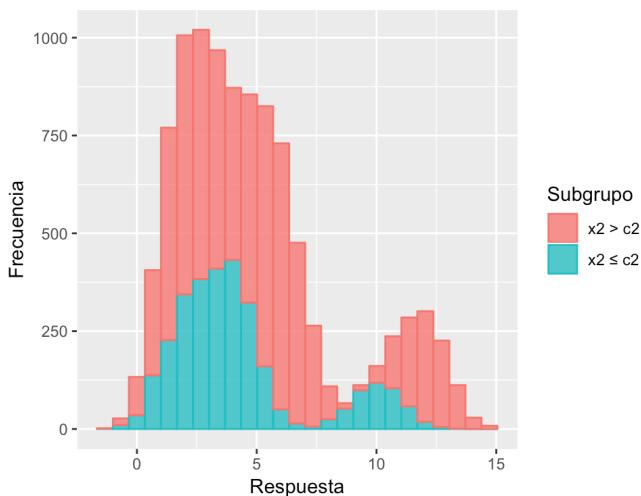


Figura 5.4: Respuesta de los subgrupos determinados por x_2 .

5.1. *Virtual Twins*

Para la aplicación del método *Virtual Twins* se utilizó el paquete de R, **aVirtualTwins** (Vieille y J. Foster 2018). Para que el método funcione, la variable de respuesta tiene que ser dicotómica. Por lo tanto, se asigna el valor 1 si la respuesta es mayor o igual a la media y 0 si es menor. Para cada bosque aleatorio (original y gemelo) se utilizaron 500 árboles. Se hizo la prueba un número mayor de árboles, pero se notó que los resultados eran similares. Los resultados de las predicciones se muestran en el cuadro 5.2:

	Media de $E[y] = \hat{P}_1$	Media de $E[y] = \hat{P}_0$
$T_i = 0$	0.788	0.042
$T_i = 1$	0.777	0.039

Cuadro 5.2: Predicciones de probabilidades.

Las predicciones de \hat{P}_1 siempre son mayores que las de \hat{P}_0 lo que muestra que independientemente de los valores de las covariables, siempre se predice un mejor resultado si se supone el tratamiento. Además, para los individuos del grupo tratado se estima mayor esperanza de una respuesta positiva que para el grupo control. Por otro lado, al calcular el estimador del efecto del tratamiento para cada paciente $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$, se obtiene que la media de Z_i es 0.7421. Es decir, el tratamiento tiene en general un efecto positivo, bastante grande si se toma en cuenta que la respuesta de toma valores de 0 y 1.

Asimismo, se calcula δ , el estimador del efecto del tratamiento $P(Y_i = 0|T_i = 0) - P(Y_i = 1|T_i = 0)$ para utilizarse dentro del valor umbral en los árboles. Se tiene que $P(Y_i = 0|T_i = 0) - P(Y_i = 1|T_i = 0) = \frac{P(Y_i=0, T_i=0)}{P(T_i=0)} - \frac{P(Y_i=1, T_i=0)}{P(T_i=0)}$ y este se estima con:

$$\begin{aligned}\delta &= \frac{\#(X_i|Y_i = 0, T_i = 0)}{\#(X_i|T_i = 0)} - \frac{\#(X_i|Y_i = 1, T_i = 0)}{\#(X_i|T_i = 0)} \\ &= \frac{\sum_{i=0}^n I(Y_i = 0, T_i = 0)}{\sum_{i=0}^n I(T_i = 0)} - \frac{\sum_{i=0}^n I(Y_i = 1, T_i = 0)}{\sum_{i=0}^n I(T_i = 0)} \\ &= \frac{\sum_{i=1}^n Y_i \cdot T_i}{\sum_{i=0}^n T_i} - \frac{\sum_{i=0}^n Y_i \cdot (1 - T_i)}{\sum_{i=0}^n (1 - T_i)}, \\ &\quad (\text{porque } Y_i \text{ y } T_i \text{ son binarias})\end{aligned}$$

Para los datos, $\delta = 0.71619$. Lo que quiere decir que utilizar el tratamiento aumenta en 0.71 la probabilidad de obtener un resultado favorable que

utilizando el control.

Con propósito de comparación, se aplica a la base de datos el método de *Virtual Twins* con la variación del árbol de regresión y clasificación. Cada árbol tiene como criterios de paro una máxima profundidad de 15 nodos donde cada nodo debe tener como mínimo 20 observaciones, para replicar los criterios establecidos por Foster et al. Asimismo, se utilizan dos valores para c , el valor de umbral: $c = \delta + 0.1$ y $c = \delta + 0.05$. Entre mayor es el valor de umbral, menor será el tamaño de la región A que describe los subgrupos con heterogeneidad del tratamiento.

***Virtual Twins* con árbol de regresión**

Con $c = \delta + 0.1$

El método divide los datos simulados en el árbol mostrado en la figura 5.5:

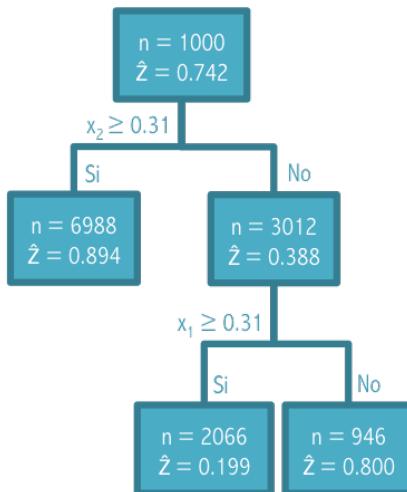


Figura 5.5: Árbol de regresión con $c = \delta + 0.1$.

En cada nodo se muestra el número de observaciones y la media efecto del tratamiento \hat{Z} para las variables de ese nodo. En primer nodo del árbol calcula el efecto del tratamiento para todas las variables. En este nodo, el efecto es casi 0.7421, o la media general de Z_i . Que sea un valor positivo indica que pertenecer al tratamiento aumenta el valor de la respuesta.

El método identifica correctamente que los subgrupos dependen de las covariables x_1 y x_2 y que no dependen de x_3 y x_4 . Asimismo, identifica correctamente los valores de umbral $c_1 = c_2 = 0.3$.

La división principal que encuentra el método es si el valor de la covariable $x_2 \geq 0.31$. Les asigna una mayor respuesta a los datos cuando $x_2 \geq 0.31$, observación correcta si se toma en cuenta la forma en que se definió el modelo. La otra división que se encuentra es $x_1 \geq 0.31$ para la cual se predice un menor valor de respuesta. Esta predicción también

corresponde al modelo.

A continuación, se muestra el subgrupo donde hay mayor heterogeneidad en el tratamiento encontrado por el método. Este subgrupo forma la región \hat{A} . En el cuadro 5.3 se muestra el porcentaje de respuestas favorables en el subgrupo para individuos en el grupo de tratamiento y de control.

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
$x_2 \geq 0.31$	6988	90.7 %	5.1 %

Cuadro 5.3: Probabilidad para el subgrupo con $c = \delta + 0.1$.

La diferencia entre el número de pacientes con resultados favorables en el grupo con tratamiento es del 90.7% mientras que para pacientes en el grupo control es del 5.1%. La diferencia es bastante grande, lo que indica que si un paciente está tiene un valor para la covariable $x_2 \geq 0.31$ y recibe el tratamiento es mucho más probable que obtenga un resultado favorable que si está en el grupo control.

Con $c = \delta + 0.05$

Relajando el valor de umbral c , se obtiene una región \hat{A} con más observaciones. Para $c = \delta + 0.05$, el árbol de regresión es el mismo que para $c = \delta + 0.1$ (véase la figura 5.5). Una explicación de porqué se obtiene el mismo árbol es que el umbral del 5% es suficiente para discriminar las diferencias en los resultados gracias al tratamiento a pesar de ser un umbral más relajado. Es quizás una razón de Foster et al. para escoger los valores 10% y 5% para definir los umbrales del método.

La diferencia entre los dos valores de umbral se refleja de manera más explícita en los subgrupos encontrados que forman la región \hat{A} , mostrados en el cuadro 5.4:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
$x_2 \geq 0.31$	6988	90.7 %	5.1 %
$x_2 < 0.31$ y $x_1 < 0.31$	946	100 %	19.3 %

Cuadro 5.4: Probabilidades para subgrupos con $c = \delta + 0.05$.

En ese caso, hay un mayor número de individuos en la región \hat{A} . Además, también aparece el subgrupo determinado por x_1 . Parece que el aumento en el efecto del tratamiento en el subgrupo $x_1 \leq 0.3$ es disminuido por el efecto negativo del subgrupo $x_2 \leq 0.3$. A pesar de que los árboles encontraron las divisiones de manera correcta, no se encontró el subgrupo con mayor efecto positivo del tratamiento: los individuos para los cuales $x_1 \leq 0.3$ y $x_2 > 0.3$. Esto se debe probablemente a que el método no encuentra suficiente heterogeneidad entre individuos que tienen $x_2 > 0.3$ y $x_1 \leq 0.3$ e individuos que tienen $x_2 > 0.3$ y $x_1 > 0.3$. Posiblemente se deba a que no hay una gran diferencia entre número de pacientes que obtuvieron un resultado favorable en el tratamiento y en el control para los dos subgrupos.

Finalmente, se evalúa el desempeño de las diferentes regiones \hat{A} obtenidas para cada valor de c . Para estimar $Q(\hat{A})$ se utilizó la fórmula

revisada en el Capítulo 3:

$$\hat{Q}(\hat{A}) = \left[\frac{1}{|\hat{A}_1|} \sum_{\substack{X_i \in \hat{A} \\ T_i=0}} \hat{P}_{1i} - \frac{1}{|\hat{A}_0|} \sum_{\substack{X_i \in \hat{A} \\ T_i=0}} \hat{P}_{0i} \right] - \left[\frac{1}{n_1} \sum_{T_i=0} \hat{P}_{1i} - \frac{1}{n_0} \sum_{T_i=0} \hat{P}_{0i} \right].$$

Las características para cada región se pueden resumir en el cuadro 5.5:

\hat{A}	$ \hat{A} $	$\hat{Q}(\hat{A})$
$c = \delta + 0.1$	6988	0.1465
$c = \delta + 0.05$	7934	0.1357

Cuadro 5.5: Estadísticas para la región \hat{A} .

Se puede ver que la región de $c = \delta + 0.05$ contiene más observaciones que la región obtenida con $c = \delta + 0.1$. Sin embargo, la medida de desempeño para esta región es menor. Por lo tanto, se puede suponer que las observaciones adicionales de la región de $c = \delta + 0.05$ no aportan información relevante a la región.

***Virtual Twins* con árbol de clasificación**

A diferencia de los árboles de regresión, los árboles de clasificación predicen un valor de respuesta 0 o 1 para observación si se utiliza el tratamiento, no la probabilidad de obtener un resultado favorable. A continuación, se muestran los resultados para cada valor de c .

Con $c = \delta + 0.1$

El árbol de clasificación que se encontró utilizando el método de con $c = \delta + 0.1$ se ilustra en la siguiente figura 5.6:

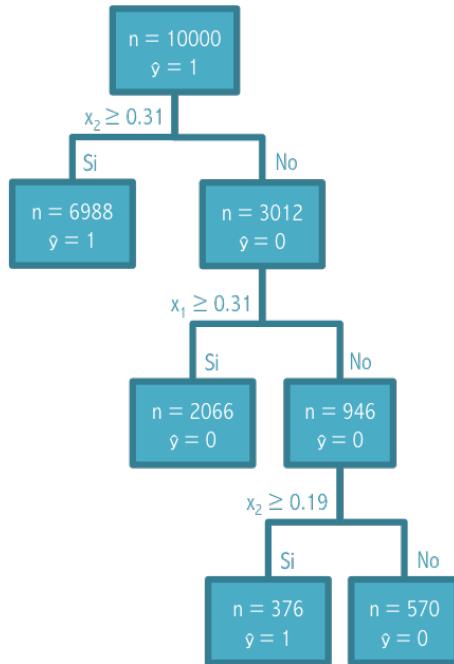


Figura 5.6: Árbol de clasificación con $c = \delta + 0.1$.

Para las divisiones que están definidas por x_2 , en ambos casos si x_2 es mayor, se predice una respuesta favorable. Sin embargo, de simplemente interpretar el árbol no es claro cómo influye la división de x_1 en las predicciones. En los subgrupos encontrados que forman la región \hat{A} , los cuales se muestran en el cuadro 5.6, se puede hacer una interpretación más precisa:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
$0.19 \leq x_2 < 0.31$ y $x_1 < 0.31$	376	100 %	17.3 %
$x_2 \geq 0.31$	6988	90.7 %	5.1 %

Cuadro 5.6: Probabilidades para subgrupos con $c = \delta + 0.1$.

Se identifica correctamente el efecto que tiene x_1 en el tratamiento, ya que uno de los subgrupos que pertenecen a \hat{A} está determinado por $x_1 < 0.31$. Sin embargo, para x_2 se identifica una división donde no hay efectos de subgrupos. De cierta manera el árbol sobre ajusta la clasificación. Es decir, se adapta a los datos demasiado y toma en cuenta patrones no relevantes que han surgido por casualidad. Entonces, da una idea de precisión que es falsa.

Con $c = \delta + 0.05$

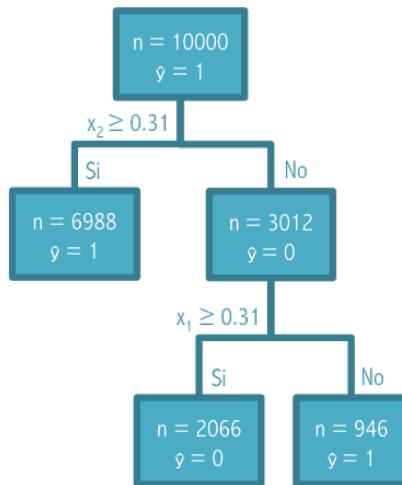


Figura 5.7: Árbol de clasificación con $c = \delta + 0.05$.

Con un valor umbral menor, se encuentra un árbol con menos divisiones. Este árbol, ilustrado en la figura 5.7, está determinado por las mismas divisiones que los árboles obtenidos en el método de regresión (véase la figura 5.5).

A pesar de que el árbol que se obtiene cuando se utiliza $c = \delta + 0.1$ tiene más divisiones que el que se obtiene cuando $c = \delta + 0.05$, el tamaño de la región \hat{A} es mayor cuando $c = \delta + 0.05$. Es posible que esto se deba a que el método sólo selecciona observaciones de los nodos terminales para formar \hat{A} . Por lo tanto, con un valor umbral más relajado se puede preferir no hacer divisiones en algunos nodos para obtener una región \hat{A} con más observaciones.

Las predicciones que se hacen sobre el valor del resultado concuerdan con el modelo original y se muestran en el cuadro siguiente 5.7:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
$x_2 < 0.31$ y $x_1 < 0.31$	946	100 %	19.3 %
$x_2 \geq 0.31$	6988	90.7 %	5.1 %

Cuadro 5.7: Probabilidades para subgrupos con $c = \delta + 0.05$.

El desempeño de los árboles de clasificación para los dos valores de umbral se resume en el cuadro 5.8:

\hat{A}	$ \hat{A} $	$\hat{Q}(\hat{A})$
$c = \delta + 0.1$	7364	0.1427
$c = \delta + 0.05$	7934	0.1357

Cuadro 5.8: Estadísticas para la región \hat{A} .

Al igual que con los árboles de regresión, la región con mejor desempeño es la que utiliza $c = \delta + 0.1$. Esto se traduce en que las observaciones adicionales en la región con $c = \delta + 0.05$ disminuyen la diferencia entre el efecto que tiene el tratamiento en la región \hat{A} con el tratamiento en general. De otra forma, las observaciones adicionales tienen un efecto más parecido al promedio que el resto de las observaciones en la región \hat{A} . Esta es probablemente otra razón por la cual Foster et al. hayan escogido $\delta + 0.1$ y $\delta + 0.05$ como valores de umbral. El umbral $\delta + 0.05$ es suficiente para obtener un árbol funcional pero no para obtener una región \hat{A} con desempeño óptimo. Mientras que el umbral $\delta + 0.1$ no encuentra los árboles óptimos.

El desempeño de los métodos de clasificación es muy parecido a los de regresión. Una explicación para este fenómeno es que en todos los casos los árboles fueron muy parecidos y para todos los métodos, el subgrupo $x_2 \geq 0.31$ pertenecía a la región \hat{A} .

A pesar de que no se encontró al subgrupo cuando $x_2 > 0.3$ y $x_1 \leq 0.3$, todas las variaciones del método encontraron correctamente las variables que determinaban a los subgrupos con efectos diferentes del tratamiento y los valores c_1 y c_2 . Además, en todas se determinó correctamente el sentido del efecto en el subgrupo. Por último, ninguna variación encontró equivocadamente algún subgrupo determinado por las covariables x_3 y x_4 . Por lo tanto, se puede concluir que el método tiene un buen funcionamiento en un ambiente controlado, pero su funcionamiento no es óptimo. Será interesante probar su desempeño en la exploración de datos reales.

5.2. Árboles de Interacción

Como fue mencionado anteriormente, los autores del método de árboles de interacción (Su et al. 2009) compartieron el código de su método en la página de *Biopharmaceutical Network* (Lipkovich 2018). Este código se utilizó y adaptó para que funcionara con las bases de datos de interés.

Para aplicar el método de árboles de interacción, se comenzó siguiendo la sugerencia de Su et al. (2009) de dividir el conjunto de datos en un conjunto de aprendizaje y uno de prueba. Para el primer conjunto, se tomó una muestra aleatoria con el 80% de las observaciones de la base de datos. El resto de las observaciones conformaron el conjunto de prueba. Recordando que el método consta de tres pasos, se muestran a continuación

los resultados de cada paso:

1. Árbol inicial T_0 : Se utiliza la estadística $t(s)$ descrita en el Capítulo 3 para crear las divisiones de cada nodo y los siguientes criterios de paro (siguiendo las sugerencias de Su et al.):

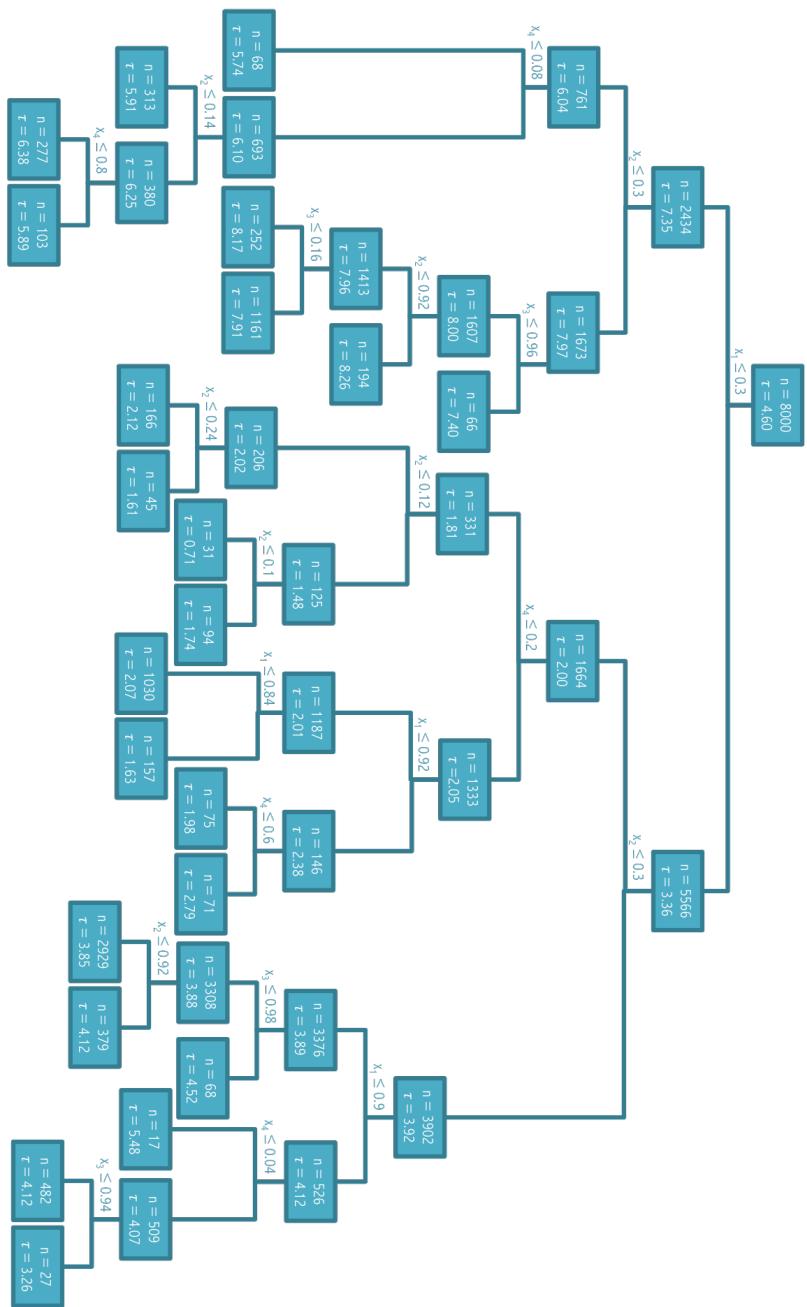
- a) Mínimo número de observaciones: 20
- b) Máxima profundidad del árbol: 5

El método divide a los datos simulados en el árbol inicial T_0 , mostrado en la figura 5.8.

2. Podar el árbol:

- a) Se obtiene la sucesión de árboles podados, ilustrada en la figura 5.9. El código de colores que determina cada árbol de la sucesión se muestra en la figura 5.10.

Las figuras 5.8, 5.9 y 5.10 se muestran a continuación:

Figura 5.8: Árbol inicial T_0

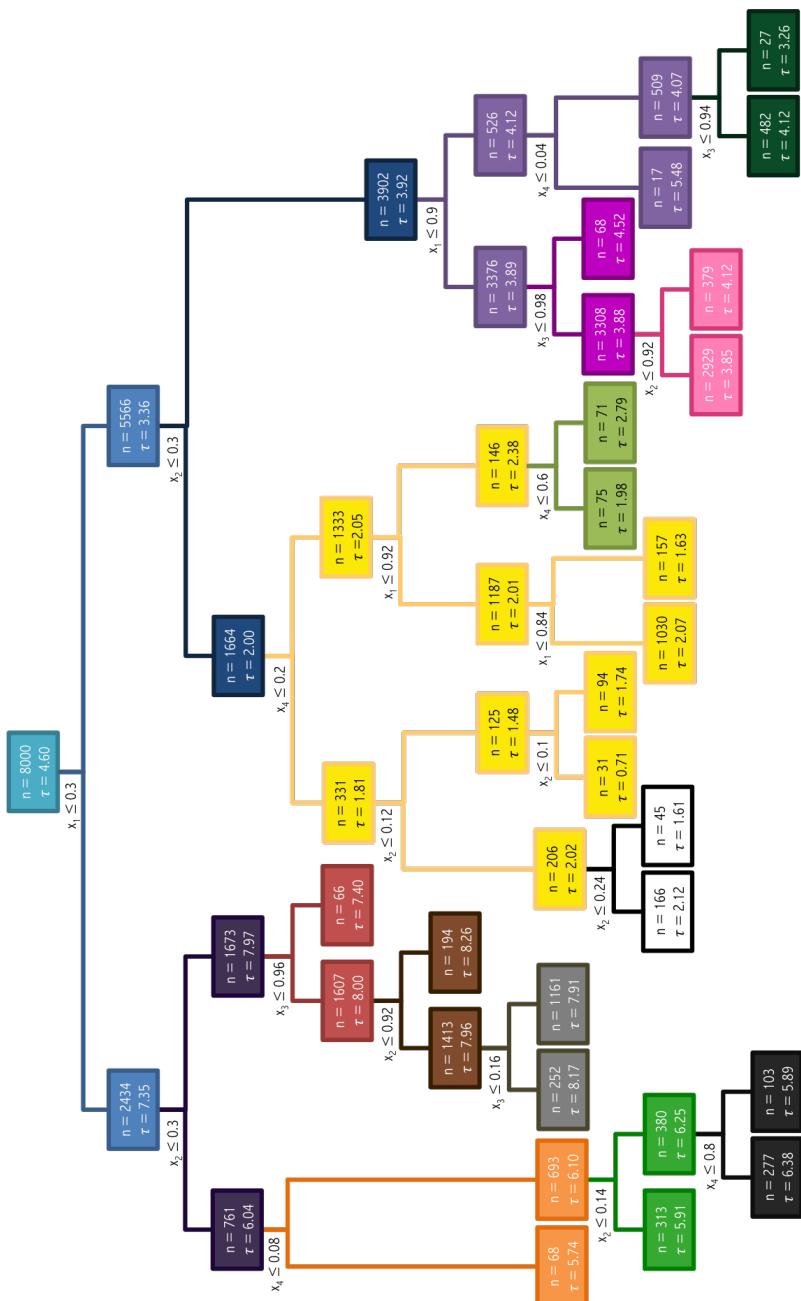


Figura 5.9: Sucesión de árboles anidados. Cada árbol anidado está representado por un conjunto diferente de colores que se describen en la siguiente figura 5-10.

Árboles:

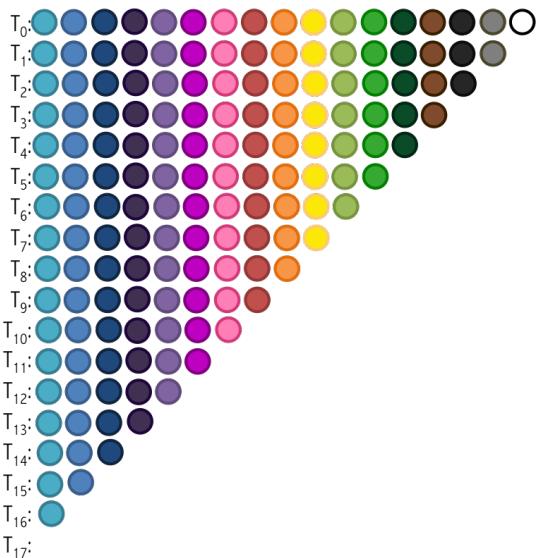


Figura 5.10: Cada color representa la rama más débil elegida en una de las iteraciones (concepto descrito en el Capítulo 3) para el árbol mostrado en la figura 5.9. Se empieza con el árbol completo y se termina con la raíz. En cada iteración se elimina una rama, la rama más débil.

- b) Para cada árbol T_m de la sucesión se calcula la estadística $G_\lambda(T_m)$. Para efectos de comparación se toma $\lambda \in \{2, 3, 4, \ln(n)\}$. Los resultados se presentan en el cuadro 5.9:

Subárbol	$G_2(T_m)$	$G_3(T_m)$	$G_4(T_m)$	$G_{\ln(n)}(T_m)$
T_0	1545.24	1524.24	1503.24	1427.63
T_1	1547.19	1527.19	1507.19	1435.18
T_2	1546.68	1527.68	1508.68	1440.26
T_3	1548.53	1530.53	1512.53	1447.71
T_4	1549.53	1532.53	1515.53	1454.32
T_5	1551.01	1535.01	1519.01	1461.40
T_6	1551.28	1536.28	1521.28	1467.27
T_7	1553.28	1539.28	1525.28	1474.87
T_8	1558.29	1549.29	1540.29	1507.88
T_9	1556.73	1548.73	1540.73	1511.93
T_{10}	1558.71	1551.71	1544.71	1519.50
T_{11}	1560.62	1554.62	1548.62	1527.02
T_{12}	1561.26	1556.26	1551.26	1533.26
T_{13}	1564.70	1561.70	1558.70	1547.90
T_{14}	1447.74	1445.74	1443.74	1436.54
T_{15}	1198.17	1197.17	1196.17	1192.57
T_{16}	0	0	0	0
Máximo	T_{13}	T_{13}	T_{13}	T_{13}

Cuadro 5.9: Estadística $G_\lambda(T_m)$ para cada árbol de sucesión anidada.

Para todos los valores de λ , el máximo de $G_\lambda(T_m)$ se obtiene con el subárbol T_{13} . Inicialmente, el número de nodos adicionales aumenta la interacción en promedio. Después llega a un máximo cuando $m = 13$ y disminuye. A partir de $m = 13$, el número de nodos adicionales a la raíz no aumenta lo suficiente la interacción en el árbol. Es decir, el tratamiento no es lo suficientemente heterogéneo en las divisiones a partir de

T_{13} como para agregar nodos. Al ser el máximo para todos los valores de λ , se toma el árbol T_{13} como el óptimo:

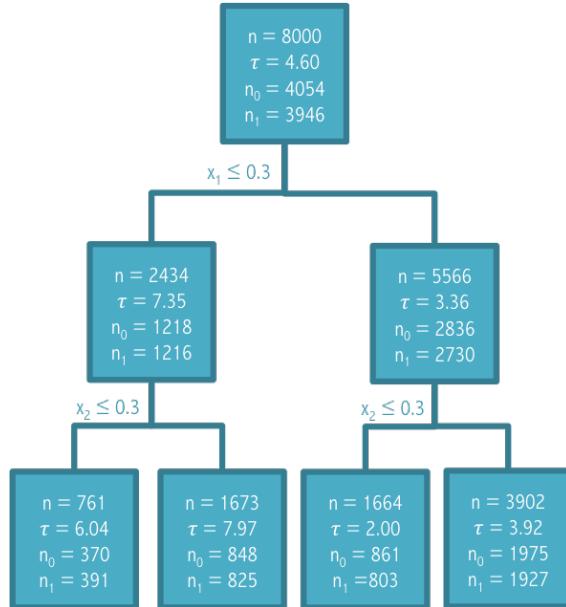


Figura 5.11: Árbol final.

A pesar de que en un inicio se encontró un árbol con muchas divisiones y muchos niveles, el método regresó un árbol cuyas divisiones están determinadas sólo por las covariables que realmente tienen diferencia en el tratamiento: x_1 y x_2 (véase la figura 5.11). De igual forma, se determinaron correctamente los valores de umbral c_1 y c_2 en todos los casos. No se omite ninguno de los subgrupos con efectos diferentes, como es el caso del subgrupo donde $x_1 \leq 0.3$ y $x_2 > 0.3$ en el método de Virtual Twins.

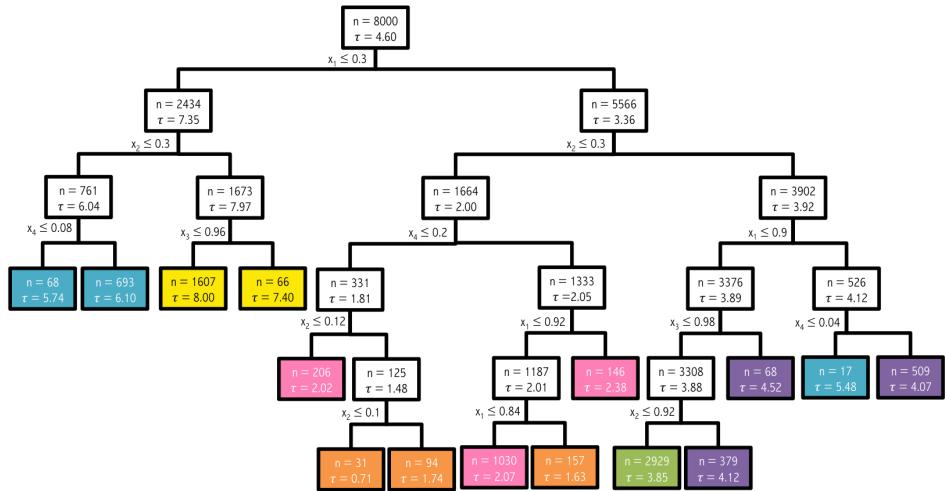
Por otra parte, los valores predichos para el efecto tratamiento comparado con el control en cada subgrupo son cercanos a las

medias reales. El siguiente cuadro 5.10 refleja ese resultado:

Subgrupo	$E[y t = 1] - E[y t = 0]$	Predicción
$x_1 \leq 0.3, x_2 \leq 0.3$	6	6.04
$x_1 \leq 0.3, x_2 > 0.3$	8	7.97
$x_1 > 0.3, x_2 \leq 0.3$	2	2.00
$x_1 > 0.3, x_2 > 0.3$	4	3.92

Cuadro 5.10: Esperanza y predicción para subgrupos encontrados.

3. Fusionar nodos: Se puede buscar fusionar los nodos terminales para obtener un árbol con un número determinado de nodos. En este caso no es necesario por el tamaño del árbol y su fácil comprensión. Sin embargo, con el propósito de exemplificar este paso, se le aplica la fusión al árbol T_7 , el cual tiene 15 nodos terminales en 3 distintos niveles de profundidad del árbol. Se reducen los 15 subgrupos de los nodos terminales a 6 subgrupos (este número es escogido por el investigar de acuerdo a los objetivos del análisis). El árbol resultante se muestra en la figura 5.12:


 Figura 5.12: Fusión a los nodos terminales del árbol T_7 .

En la figura 5.12, los nodos fusionados que forman a los subgrupos están representados por 6 colores diferentes. Se puede ver que el algoritmo de fusión no distingue entre la profundidad del árbol ni en las ramas de las cuales provienen los nodos para hacer una fusión. Se basa sólo en la diferencia entre la predicción del efecto del tratamiento en cada nodo. En particular, para este ejemplo los nodos fusionados con la máxima diferencia se dan en el subgrupo naranja, con 1.03 unidades en los nodos con más heterogeneidad dentro del subgrupo (en este subgrupo se fusionan 3 nodos). Por otro lado, en el subgrupo color verde sólo hay un nodo, lo que quiere decir que este nodo no se fusionó con ninguno otro.

Regresando al árbol obtenido en el paso 2., para el árbol final se puede calcular la importancia de las covariables según la estadística V_j como análisis adicional. El cuadro 5.11 muestra los resultados de dicha estadística:

Covariable	V_j
x_1	307.32
x_2	96.92
x_3	37.51
x_4	32.70

Cuadro 5.11: Estadística V_j para cada covariable x_j .

A diferencia del método *Virtual Twins*, se distingue a la covariable x_1 como la de mayor importancia: es la que más influye en la determinación de la estructura del árbol. Puede ser una razón por la que el método de árboles de interacción encuentra correctamente todos los subgrupos mientras que en *Virtual Twins* no se incluye uno.

Por otro lado, las covariables x_3 y x_4 muestran valores positivos para V_j . Sin embargo, si se comparan contra x_1 , se puede concluir que las covariables no son relevantes para determinar la estructura del árbol.

En conclusión, el método encontró correctamente todos los subgrupos y los valores que los determinan. De igual forma, la dirección y la magnitud del efecto del tratamiento en los subgrupos fue correcta. Entonces, se puede concluir que su desempeño fue superior a *Virtual Twins* para este ejercicio y que su funcionamiento es óptimo en un entorno controlado.

5.3. Método de Dixon y Simon

Para probar los métodos que confirman la existencia de subgrupos que se han predefinido antes del análisis se simula una nueva muestra con

500 observaciones que provienen de modelo (5.1). Esto con el propósito de que la muestra sea independiente de aquella para la cual se hizo la exploración de subgrupos. En análisis de subgrupos para ensayos clínicos, idealmente se generan hipótesis de la existencia de subgrupos con efectos del tratamiento heterogéneos a través de un análisis exploratorio. Sólo se confirma la hipótesis si se realizan análisis con subgrupos predefinidos en otros ensayos clínicos diferentes al que generó la hipótesis (Rothwell 2005).

Para aplicar el método de Dixon y Simon (1991), se utiliza el paquete de R, **DSBayes** (Varadhan y Yao 2014). Si los estimadores $\hat{\theta}$ y C no se proporcionan, como es el caso en este ejemplo, el paquete aplica una regresión lineal a los datos y extrae los coeficientes la regresión para estimar $\hat{\theta}$. La matriz C la estima con la función **vcov** que regresa una matriz de covarianzas para la misma regresión lineal (R Core Team 2018).

Los subgrupos a evaluar son aquellos definidos por las covariables x_1 y x_2 y los valores c_1 y c_2 . Para cada observación se crean las variables indicadoras I_1 e I_2 que determinan la pertenencia de un individuo a un subgrupo, donde $I_1 = I(x_1 \leq 0.3)$ e $I_2 = I(x_2 \leq 0.3)$. Estos fueron los subgrupos encontrados por el método de árboles de interacción.

Siguiendo la sugerencia de Dixon y Simon, no se calculan las medias para los coeficientes de la regresión, sino que los resultados que se presentan son la moda y los intervalos creíbles para cada subgrupo. Los subgrupos a evaluar están determinados por los valores mostrados en el cuadro 5.12:

Subgrupo	Tratamiento	Control	Diferencia
$x_1 \leq .3$ y $x_2 \leq .3$	$\mu + \tau + \beta_1 + \beta_2 + \gamma_1 + \gamma_2$	$\mu + \beta_1 + \beta_2$	$\tau + \gamma_1 + \gamma_2$
$x_1 \leq .3$ y $x_2 > .3$	$\mu + \tau + \beta_1 + \gamma_1$	$\mu + \beta_1$	$\tau + \gamma_1$
$x_1 > .3$ y $x_2 \leq .3$	$\mu + \tau + \beta_2 + \gamma_2$	$\mu + \beta_2$	$\tau + \gamma_2$
$x_1 > .3$ y $x_2 > .3$	$\mu + \tau$	μ	τ

Cuadro 5.12: Posibles respuestas para cada subgroupo.

Los valores que determinan si hay heterogeneidad en los subgrupos son aquellos que se encuentran en la columna de diferencia del cuadro anterior. Si estos valores son muy diferentes entre sí, sugiere la existencia de efectos en subgrupos. Los resultados para cada subgroupo se exponen en el cuadro 5.13:

Subgrupo	Grupo	Moda	Int. creíble	$E[y]$
$x_1 \leq .3$ y $x_2 \leq .3$	Tratamiento	9.837	(9.529, 10.143)	10
	Control	3.898	(3.587, 4.210)	4
	Diferencia	5.938	(5.502, 6.373)	6
$x_1 \leq .3$ y $x_2 > .3$	Tratamiento	11.948	(11.716, 12.177)	12
	Control	3.992	(3.748, 4.239)	4
	Diferencia	7.954	(7.611, 8.295)	8
$x_1 > .3$ y $x_2 \leq .3$	Tratamiento	3.838	(3.604, 4.070)	4
	Control	1.916	(1.664, 2.170)	2
	Diferencia	1.918	(1.571, 2.269)	2
$x_1 > .3$ y $x_2 > .3$	Tratamiento	5.949	(5.780, 6.116)	6
	Control	2.011	(1.847, 2.178)	2
	Diferencia	3.934	(3.700, 4.171)	4

Cuadro 5.13: Modas obtenidas para cada subgrupo.

Se puede ver en el cuadro 5.13 que los valores estimados a través de la moda son muy parecidos a las esperanzas para cada subgrupo. Todos están a menos de 0.2 unidades de las esperanzas. Esto quiere decir que, si los subgrupos están definidos anteriormente, el método los encuentra de manera óptima. Además, los intervalos creíbles son relativamente pequeños, que se puede traducir en poca incertidumbre sobre los estimadores obtenidos (véase la figura 5.13).

Moda e intervalos de confianza

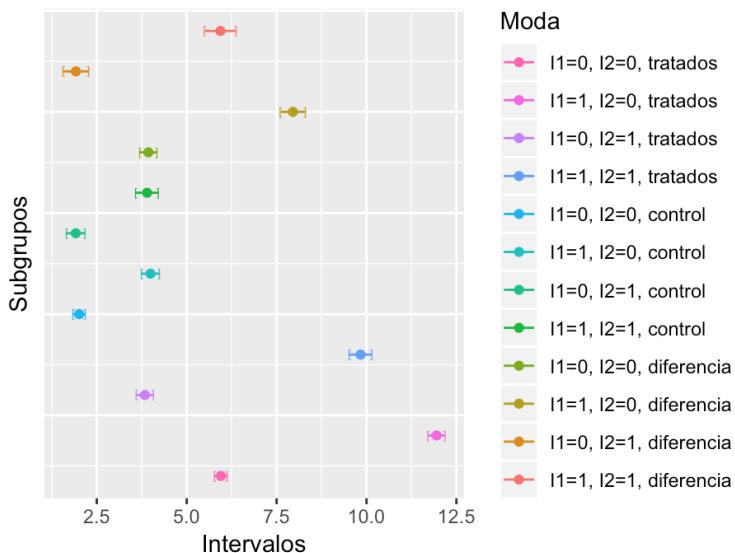


Figura 5.13: Intervalos creíbles con 95 % de probabilidad y la moda de los subgrupos.

Dixon y Simon no calculan estimadores para ninguno de los parámetros, ya que el valor de cada subgrupo se estima de manera individual. Esto es, no se estiman los valores de los parámetros y luego se suman dependiendo del subgrupo. Es una forma de tomar en cuenta la interacción entre los parámetros y no verlos individualmente. Es por esto que en el cuadro 5.13 la diferencia entre los grupos con tratamiento y con control no es el resultado de la resta. Sin embargo, en esta tesis se calculan las modas de los parámetros para complementar el análisis y se muestran en la tabla 5.14 y en la figura 5.14.

Coefficiente	Moda	Intervalo crefble	Valor real
μ	2.011	(1.847, 2.178)	2
τ	3.934	(3.700, 4.171)	4
β_1	1.981	(1.706, 2.257)	2
β_2	-0.095	(-0.375, 0.185)	0
γ_1	4.019	(3.629, 4.406)	4
γ_2	-2.016	(-2.408, -1.624)	-2

Cuadro 5.14: Modas obtenidas para los parámetros. En todos los casos, la moda es muy cercana al valor real del parámetro.

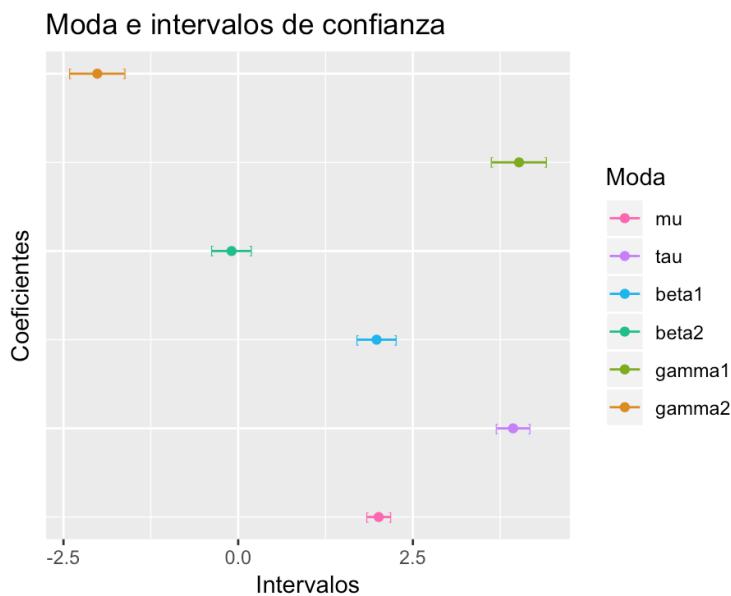


Figura 5.14: Intervalos crefables con 95 % de probabilidad y la moda de los parámetros. En todos los casos, los intervalos son de poca longitud, sugiriendo precisión en las estimaciones.

Nuevamente, el resultado de los estimadores es muy cercano al valor real. Es claro que el método logra estimar los efectos en subgrupos y los valores de los parámetros satisfactoriamente si los subgrupos son predefinidos correctamente.

Un análisis adicional que se puede realizar es probar el funcionamiento del método si los subgrupos no son conocidos con totalidad. Esto asemeja con mayor cercanía un escenario real. Entonces, se definen los siguientes subgrupos determinados por el valor de las covariables: $x_1 \leq 0.5$, $x_2 \leq 0.7$, $x_3 \leq 0.1$ y $x_4 \leq 0.5$. Los resultados se muestran en el cuadro 5.15. Los subgrupos están definidos por las indicadoras I_j con $j = 1, 2, 3, 4$, donde $I_1 = I(x_1 \leq 0.5)$, $I_2 = I(x_2 \leq 0.7)$, $I_3 = I(x_3 \leq 0.1)$ e $I_4 = I(x_4 \leq 0.5)$.

Subgrupo		G	Moda	$E[y]$		Subgrupo	G	Moda	$E[y]$
$I_1 = 0$	$I_3 = 0$	T	5.99	6	$I_1 = 0$	$I_3 = 0$	T	5.28	5.14
	$I_2 = 0$	$I_4 = 0$	C	2.01		$I_4 = 0$	C	1.97	4
		D		3.99		D		3.30	1.14
	$I_3 = 0$	T	5.80	6		$I_3 = 0$	T	5.09	5.14
		$I_4 = 1$	C	1.92		$I_4 = 1$	C	1.89	4
		D		3.88		D		3.19	1.14
	$I_3 = 1$	T	6.10	6		$I_3 = 1$	T	5.39	5.14
		$I_4 = 0$	C	1.82		$I_4 = 0$	C	1.78	4
		D		4.27		D		3.61	1.14
$I_1 = 1$	$I_2 = 0$	$I_3 = 1$	T	5.91	$I_1 = 1$	$I_3 = 1$	T	5.20	5.14
		$I_4 = 1$	C	1.74		$I_4 = 1$	C	1.70	4
		D		4.17		D		3.50	1.14
	$I_2 = 1$	$I_3 = 0$	T	9.30		$I_3 = 0$	T	8.58	8.74
		$I_4 = 0$	C	3.25		$I_4 = 0$	C	3.21	5.2
		D		6.04		D		5.37	3.54
	$I_3 = 0$	T	9.11	9.6		$I_3 = 0$	T	8.39	8.74
		$I_4 = 1$	C	3.17		$I_4 = 1$	C	3.13	5.2
		D		5.93		D		5.26	3.54
	$I_3 = 1$	T	9.40	9.6		$I_3 = 1$	T	8.68	8.74
		$I_4 = 0$	C	3.06		$I_4 = 0$	C	3.03	5.2
		D		6.32		D		5.64	3.54
	$I_4 = 1$	$I_3 = 1$	T	9.21		$I_3 = 1$	T	8.49	8.74
		$I_4 = 1$	C	2.98		$I_4 = 1$	C	2.94	5.2
		D		6.21		D		5.53	3.54

Cuadro 5.15: Modas obtenidas para los subgrupos determinados por cuatro covariables. Se utiliza la columna titulada por G para denotar uno de los dos grupos: Tratamiento (T) o Control (C) y la diferencia (D) que hay entre la moda o esperanza entre ellos.

Para cada subgrupo definido por las variables I_j , se estima la moda del subgrupo en el tratamiento y en el control. Además, se estima la moda de la diferencia entre individuos control y tratamiento. En cada columna $E[y]$ se muestran los valores esperados para estos subgrupos.

Para las observaciones que la variable $I_1 = 1$ e $I_2 = 1$, todas las observaciones cumplen $x_1 \leq 0.5$ y $x_2 \leq 0.7$. Como las covariables siguen una distribución uniforme discreta, se puede estimar que aproximadamente el 60% de las observaciones que cumplen $x_1 \leq 0.5$ cumplen también $x_1 \leq 0.3$ y el 42.8% de las observaciones que cumplen $x_2 \leq 0.7$ cumplen también $x_2 \leq 0.3$. Entonces, para este subgrupo se estima la $E[y]$ para los de la siguiente forma: $\hat{E}[y] = \mu + \tau + 0.6\beta_1 + 0.6\gamma_1 + 0.428\beta_2 + 0.428\gamma_2$. El cálculo es análogo para el resto de los subgrupos.

Se puede ver en el cuadro 5.15 que los valores de la moda están determinados por las variables I_1 e I_2 y no influyen tanto las variables I_3 e I_4 , lo que significa que el método toma en cuenta las variables importantes de forma correcta. Para el subgrupo donde $I_1 = 0$ e $I_2 = 0$, todos los estimadores son muy cercanos a la $E[y]$. En los otros subgrupos, las modas del grupo del tratamiento son cercanas a las $E[y]$ pero las del control y las de la diferencia no lo son. Una explicación para esto es que para $I_1 = 0$ e $I_2 = 0$, todas las observaciones pertenecen al subgrupo cuando $x_1 > 0.3$ y $x_2 > 0.3$. Por lo tanto, no se tiene que hacer un cálculo especial para $E[y]$. Para el resto de subgrupos determinados por I_1 e I_2 , la respuesta viene de diferentes modelos (la esperanza de los individuos dentro de un subgrupo depende de diferentes coeficientes). Un fenómeno peculiar de las modas es que para las observaciones donde no se cumple $I_1 = 0$ e $I_2 = 0$, la moda del grupo de control es cercana a la esperanza de la diferencia y viceversa.

Nuevamente se calculan las modas de los parámetros y se muestran en el cuadro 5.16 y la figura 5.15:

Coeficiente	Moda	Intervalo creíble	Valor real
μ	2.011	(1.387, 2.640)	2
τ	3.990	(3.135, 4.827)	4
β_1	1.239	(0.701, 1.778)	2
β_2	-0.040	(-0.593, 0.519)	0
β_3	-0.184	(-1.007, 0.630)	0
β_4	-0.083	(-0.595, 0.431)	0
γ_1	2.069	(1.300, 2.824)	4
γ_2	-0.670	(-1.446, 0.079)	-2
γ_3	0.279	(-0.820, 1.426)	0
γ_4	-0.108	(-0.831, 0.614)	0

Cuadro 5.16: Modas e intervalos creíbles para los parámetros que determinan la $E[y]$ para los pacientes simulados.

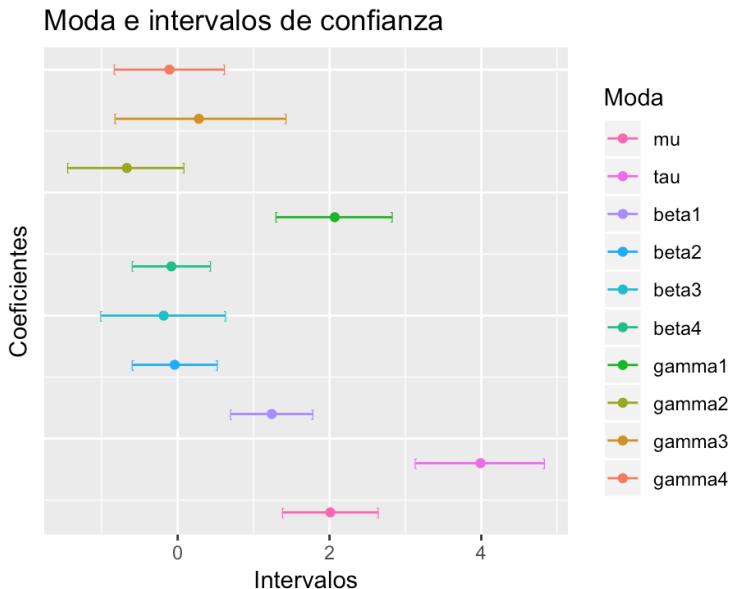


Figura 5.15: Intervalos creíbles con 95 % de probabilidad y moda de los parámetros que determinan la $E[y]$ para los pacientes simulados.

Aunque la mayoría de los parámetros son cercanos al valor real, los parámetros β_1 , γ_1 y γ_2 no están contenidos en los intervalos creíbles. Esto es preocupante porque son los parámetros que determinan los efectos del tratamiento en los subgrupos de interés. Aunado con los resultados de las modas de efectos en los subgrupos, se concluye que a pesar de que el método sirve como referencia, si los subgrupos no están bien definidos antes de hacer el análisis se puede llegar a resultados incorrectos.

5.4. Modificación con *JAGS*

Como se expuso en el Capítulo 4, se puede hacer un análisis alternativo al método de Dixon y Simon (1991) utilizando el paquete computacional *JAGS* (Plummer 2013). Empezaremos a evaluar el caso cuando se utilizan los subgrupos definidos por x_1 y x_2 . En el modelo expuesto en la sección 4.2.5, $m = 2$ (el número de covariables que definen a los subgrupos).

Se crea el modelo con 10,000 simulaciones y 3 cadenas de Markov. Después, se actualiza el modelo con 10,000 simulaciones más. Finalmente, se hace una muestra de 10,000 valores de cada cadena. Se escogen estos números ya que se considera que 30,000 es un número de iteraciones suficiente para llegar a una cadena estable. Se utilizan 3 cadenas para comparar los diferentes resultados obtenidos al iniciar desde puntos diferentes.

En la figura 5.16 se muestran los valores obtenidos de la muestra. A estas gráficas se les llama traza. Para cada parámetro, en el eje x se grafica el número de iteración y en el eje y se grafica la muestra obtenida para el parámetro en esa iteración. No se grafican las primeras 10,000 iteraciones porque se considera que tienen mayor variación al ser el inicio de las cadenas. El máximo y mínimo del eje y es el rango en el que está contenido un parámetro. Se busca que este rango vaya decreciendo conforme se aumenten las iteraciones.

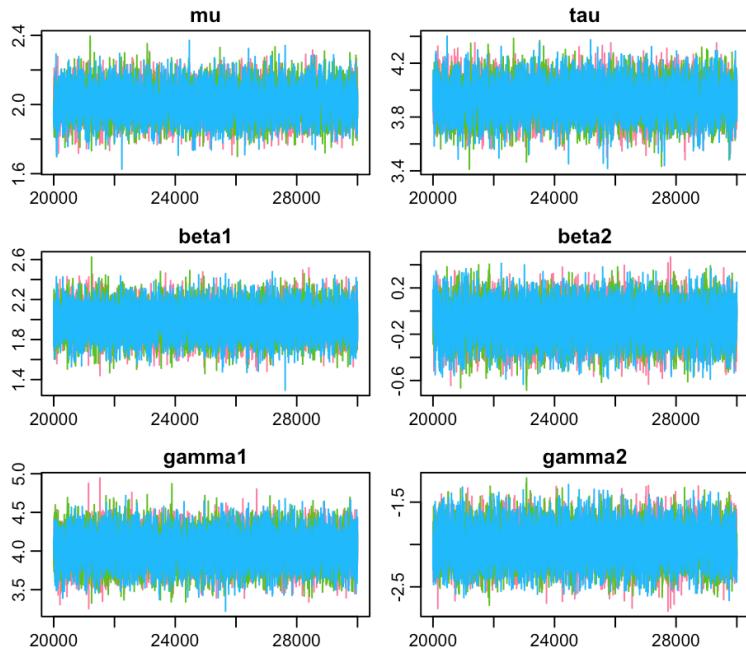


Figura 5.16: Trazas para los parámetros de cada cadena.

De las gráficas en la figura 5.16 podemos ver que las muestras para cada parámetro están concentradas en un rango relativamente pequeño. Calculando la frecuencia de los valores obtenidos para cada parámetro, se obtiene la densidad de los parámetros de cada cadena y se ilustran en la figura 5.17. En el eje x se grafica el valor de cada parámetro y en el eje y se grafica la frecuencia relativa de cada valor obtenido para un parámetro.

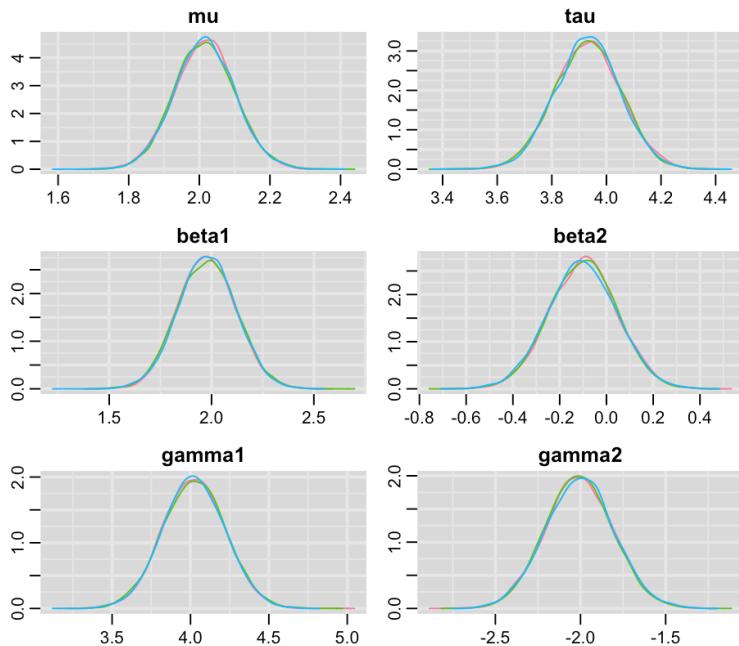


Figura 5.17: Densidades para los parámetros de cada cadena.

El que las todas las cadenas tengan una forma similar significa que independientemente de los valores especificados al inicio, las muestras de los parámetros son parecidas. En la figura 5.18 se muestra la densidad y la moda de los parámetros utilizando las tres cadenas:

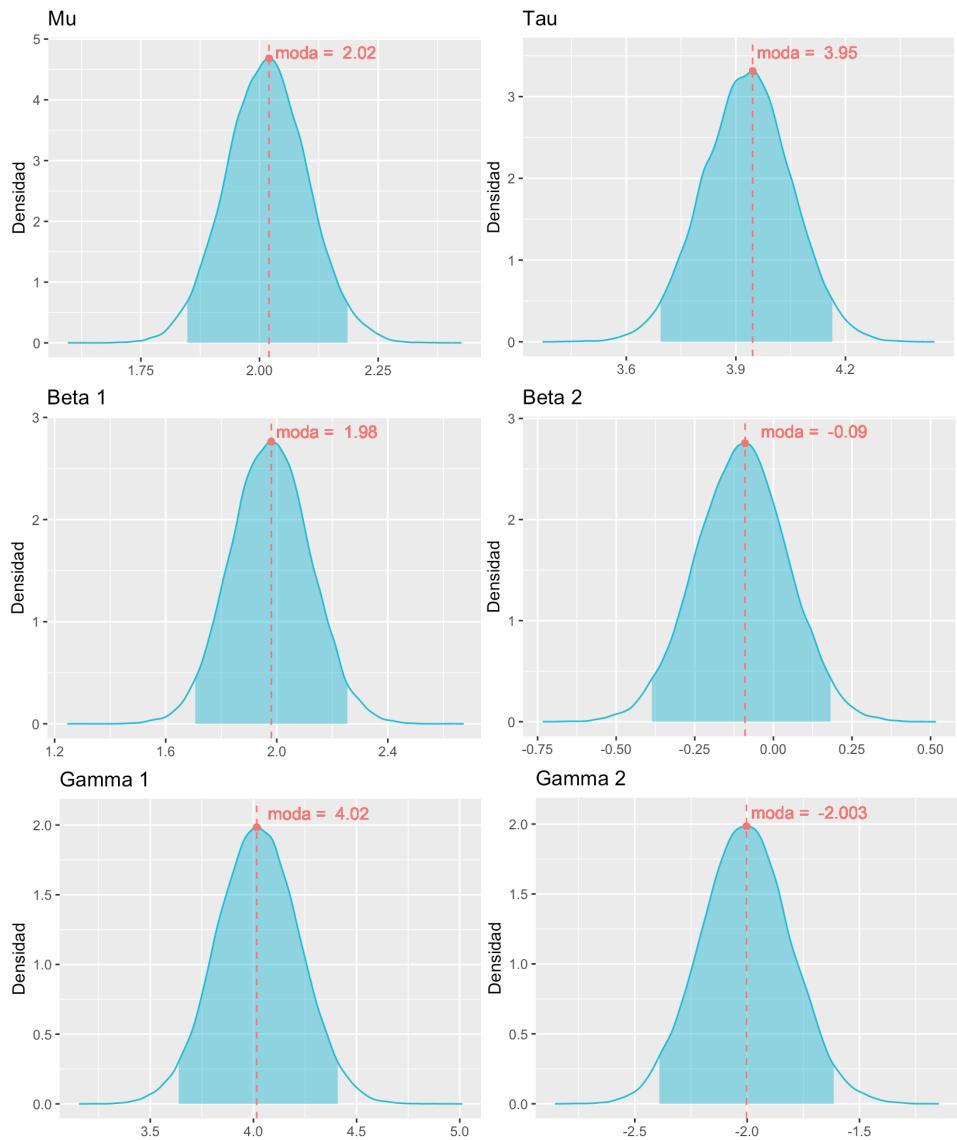


Figura 5.18: Densidades de los parámetros. Las densidades son unimodales, lo que refleja que hay un valor que se obtiene con mayor frecuencia en las muestras.

Las medias y modas que se estiman para cada coeficiente se resumen en el siguiente cuadro 5.17:

Coeficiente	Media	Moda	Cuantil del 2.5 %	Cuantil del 97.5 %
μ	2.015	2.019	1.847	2.185
τ	3.930	3.945	3.692	4.165
β_1	1.977	1.980	1.704	2.255
β_2	-0.100	-0.090	-0.387	0.182
γ_1	4.022	4.016	3.636	4.411
γ_2	-2.006	-2.002	-2.393	-1.614

Cuadro 5.17: Estimadores para cada coeficiente. Las medias y moda son muy parecidas, lo que significa que el valor para cada parámetro que se obtiene con mayor frecuencia es cercano a la media.

Para todos los parámetros, las medias y las modas son parecidas y son cercanas al valor real del coeficiente en el modelo original (véanse los cuadros 5.14 y 5.17). Además, todos los valores reales están contenidos en los intervalos creíbles, como se muestra en la figura 5.19.

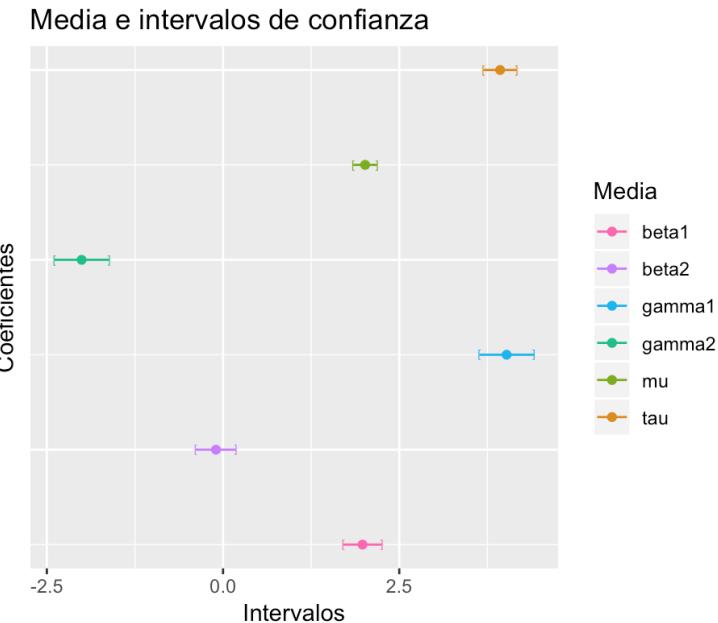


Figura 5.19: Intervalos creíbles con 95 % de probabilidad y la media. El haber obtenido densidades unimodales, donde la moda es parecida a la media y está contenida por intervalos creíbles pequeños sugiere que el modelo describe cercanamente a los valores reales de los parámetros.

Para comparar con el método aplicado con `DSBayes`, se muestran los resultados para cada subgrupo en el cuadro 5.18:

Subgrupo	Grupo	Estimador	$E[y]$
$x_1 \leq .3$ y	Tratamiento	9.839	10
$x_2 \leq .3$	Control	3.892	4
	Diferencia	5.946	6
$x_1 \leq .3$ y	Tratamiento	11.946	12
$x_2 > .3$	Control	3.993	4
	Diferencia	7.953	8
$x_1 > .3$ y	Tratamiento	3.838	4
$x_2 \leq .3$	Control	1.914	2
	Diferencia	1.923	2
$x_1 > .3$ y	Tratamiento	5.945	6
$x_2 > .3$	Control	2.015	2
	Diferencia	3.930	4

Cuadro 5.18: Estimador para cada subgrupo.

Al igual que utilizando el paquete `DSBayes`, los valores estimados para cada subgrupo son muy cercanos a su esperanza, lo que refleja un buen funcionamiento del método en un entorno controlado.

Por último, se puede realizar un análisis adicional para checar la convergencia de los parámetros: el diagnóstico Gelman-Rubin. En Gelman-Rubin, un factor de 1 significa que la varianza entre diferentes cadenas es igual. Entre mayor es el valor Gelman-Rubin, significa que hay mayor diferencia entre las cadenas. Se busca que los valores tiendan a 1. En general, se usa la regla por conveniencia de que si se tiene un valor de Gelman-Rubin menor que 1.1, hay convergencia entre cadenas (*Gelman–Rubin convergence diagnostic using multiple chains* 2019).

A continuación, se muestran los valores de Gelman-Rubin para cada parámetro en una gráfica. Para cada parámetro, el eje x es el número de iteración y el eje y es el factor que se obtiene con Gelman-Rubin. En la figura 5.20, la línea verde indica el número 1.1 y se puede ver claramente que todos los parámetros están por debajo de la línea, incluyendo su estimador. Además, se puede apreciar que conforme aumenta el número de iteraciones, el factor disminuye. Por lo tanto, se puede concluir que los parámetros convergen.

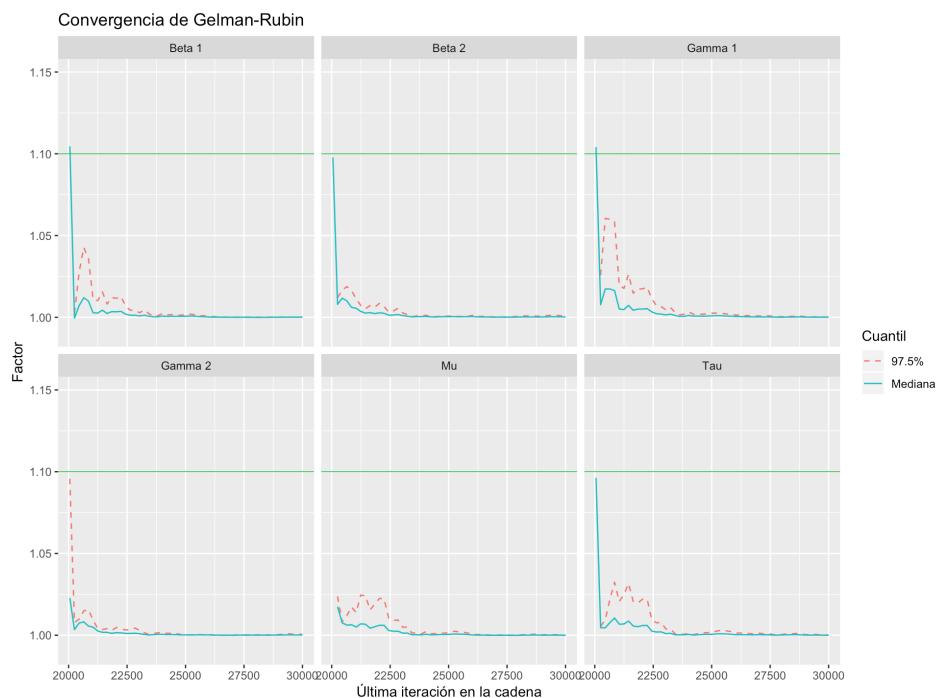


Figura 5.20: Convergencia de Gelman-Rubin.

Con estos análisis, se puede concluir que el método tiene un funcionamiento favorable en entornos controlados si los subgrupos están

predefinidos correctamente.

Para seguir evaluando al método, se realiza el mismo análisis, incluyendo las covariables que no tienen subgrupos con efectos heterogéneos en el tratamiento. En este caso, $m = 4$ ya que se agregan 2 covariables y se utilizan 3 cadenas de Markov. Nuevamente se utilizan 10,000 simulaciones para crear el modelo, para actualizarlo y para obtener una muestra.

Las trazas obtenidas en esta ocasión son las siguientes:

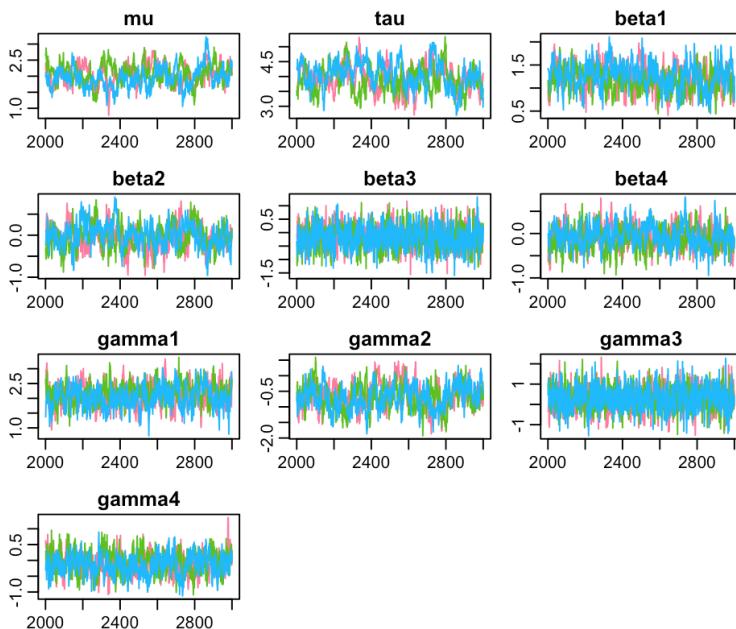


Figura 5.21: Trazas para los parámetros de cada cadena.

En este caso, las muestras de los parámetros no están tan concentradas como en la variación pasada (comparar la figura 5.21 con la figura 5.16).

Esto se puede ver especialmente en la traza de μ , τ y β_2 . Asimismo, si se grafican las densidades de los parámetros, se observa mayor diferencia entre las cadenas que en el ejemplo anterior (comparar la figura 5.22 con la figura 5.17). Algunas de las cadenas producen densidades que no tienen forma de normales, por ejemplo, la densidad de β_2 que es bimodal. Esto no necesariamente es algo negativo, pero puede ser que los parámetros tarden más iteraciones en converger.

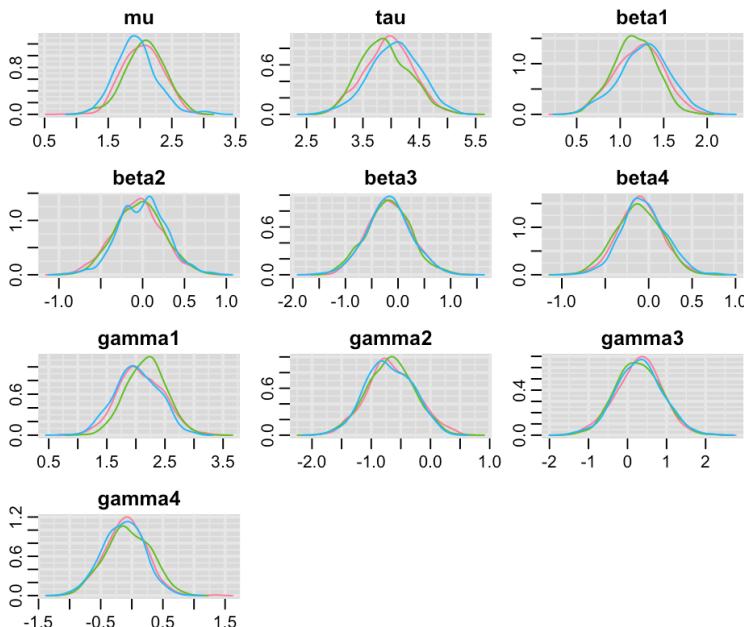


Figura 5.22: Densidades para los parámetros de cada cadena. Muestra las frecuencias relativas obtenidas en las múltiples iteraciones de las 3 cadenas. Las cadenas presentan diferencias notables y algunas no son unimodales.

En el cuadro 5.19 y la figura 5.23 se muestran las modas y las medias que el método produce para los parámetros, con sus intervalos creíbles al

95 %:

Coeficiente	Media	Moda	Cuantil del 2.5 %	Cuantil del 97.5 %
μ	2.018	3.928	1.396	2.645
τ	3.968	3.928	3.116	4.821
β_1	1.218	1.293	0.664	1.745
β_2	-0.032	0.027	-0.605	0.551
β_3	-0.175	-0.211	-1.005	0.690
β_4	-0.091	-0.131	-0.598	0.411
γ_1	2.092	1.985	1.358	2.820
γ_2	-0.672	-0.752	-1.452	0.107
γ_3	0.293	0.366	-0.868	1.441
γ_4	-0.100	-0.095	-0.771	0.560

Cuadro 5.19: Estimadores para coeficientes.

Para algunos parámetros, la media difiere de la moda, lo que lleva a cuestionar que tan confiables son los resultados del método.

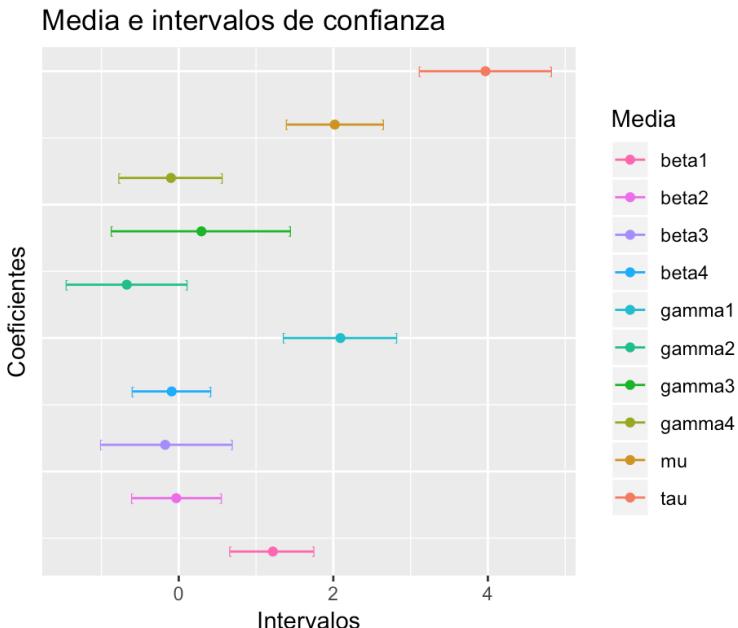


Figura 5.23: Intervalos creíbles con 95 % de probabilidad y la media.

Las medias de los parámetros son muy parecidas a las modas obtenidas en el método de Dixon y Simon sin modificaciones. Para continuar con la comparación entre métodos, se exponen las estimaciones del tratamiento en cada subgrupo en el cuadro 5.20:

Subgrupo		G	Moda	$E[y]$	Subgrupo		G	Moda	$E[y]$
$I_1 = 0$	$I_3 = 0$	T	5.98	6	$I_1 = 0$	$I_3 = 0$	T	5.28	5.14
	$I_2 = 0$	$I_4 = 0$	C	2.01		$I_4 = 0$	C	1.98	4
		D		3.96		D		3.29	1.14
	$I_3 = 0$	T	5.79	6		$I_3 = 0$	T	5.09	5.14
		$I_4 = 1$	C	1.92		$I_4 = 1$	C	1.89	4
		D		3.86		D		3.19	1.14
	$I_3 = 1$	T	6.10	6		$I_3 = 1$	T	5.40	5.14
		$I_4 = 0$	C	1.84		$I_4 = 0$	C	1.81	4
		D		4.26		D		3.58	1.14
$I_1 = 1$	$I_3 = 1$	T	5.91	6	$I_1 = 1$	$I_3 = 1$	T	5.20	5.14
	$I_2 = 0$	$I_4 = 1$	C	1.75		$I_4 = 1$	C	1.71	4
		D		4.16		D		3.48	1.14
		$I_3 = 0$	T	9.29		$I_3 = 0$	T	8.59	8.74
		$I_4 = 0$	C	3.23		$I_4 = 0$	C	3.20	5.2
		D		6.06		D		5.38	3.54
	$I_3 = 0$	T	9.10	9.6		$I_3 = 0$	T	8.40	8.74
		$I_4 = 1$	C	3.14		$I_4 = 1$	C	3.11	5.2
		D		5.96		D		5.28	3.54
$I_1 = 1$	$I_3 = 1$	T	9.41	9.6	$I_1 = 1$	$I_3 = 1$	T	8.71	8.74
	$I_2 = 0$	$I_4 = 0$	C	3.06		$I_4 = 0$	C	3.02	5.2
		D		6.35		D		5.68	3.54
		$I_3 = 1$	T	9.22		$I_3 = 1$	T	8.51	8.74
	$I_2 = 1$	$I_4 = 1$	C	2.97		$I_4 = 1$	C	2.93	5.2
		D		6.25		D		5.58	3.54

Cuadro 5.20: Modas y esperanzas para los subgrupos determinados por cuatro covariables. Se utiliza la columna titulada por G para denotar uno de los dos grupos: Tratamiento (T) o Control (C) y la diferencia (D) que hay entre la moda o esperanza entre ellos.

Al igual que con los estimadores de los coeficientes, los estimadores para los efectos en los subgrupos son casi los mismos que utilizando el método original de Dixon y Simon (1991). Por lo tanto, las simplificaciones y resúmenes que hacen Dixon y Simon no alteran los resultados de las estimaciones. Por otro lado, se concluye que ninguno de los dos métodos es apto para hacer análisis de subgrupos si no son conocidos con anterioridad y los subgrupos están predefinidos correctamente. Esto es una carencia de los métodos, ya que en escenarios reales es difícil que los subgrupos se puedan definir con tanta precisión como se hizo en el entorno controlado.

Por último, se calculan los valores de convergencia Gelman-Rubin y se muestran en la figura 5.24. A pesar de que los valores se mueven en un rango más grande que en las muestras utilizando sólo las covariables x_1 y x_2 , todos se encuentran debajo de 1.1 por lo que se puede suponer convergencia.



Figura 5.24: Convergencia de Gelman-Rubin.

5.5. Comparación entre métodos

La principal diferencia entre los métodos exploratorios fue que el método *Virtual Twins* no encontró uno de los subgrupos con efectos heterogéneos mientras que el método de árboles de interacción encontró todos los subgrupos. En cuanto a las variables que definían los subgrupos x_1 y x_2 y los valores c_1 y c_2 , ambos métodos los lograron identificar correctamente. Ninguno de los métodos produjo un árbol que tomara en cuenta a las covariables x_3 o x_4 para definir su estructura.

Entre los métodos confirmatorios, aparentemente no hubo una diferencia importante. Ambos métodos funcionaron satisfactoriamente en el caso en que se predefinían las covariables que determinaban los subgrupos correctamente. Sin embargo, su desempeño no fue favorable si los subgrupos no eran conocidos perfectamente, un gran déficit en los métodos.

Capítulo 6

Ejemplos ilustrativos con ensayos clínicos

Una vez que se ha probado el desempeño de los métodos en un entorno simulado, se aplican los métodos a dos distintas bases de datos de ensayos clínicos reales. Los datos se obtuvieron a través de la plataforma de *Project Data Sphere, LLC* (2018), la cual es una iniciativa para fomentar la investigación en ensayos clínicos de cáncer. El objetivo de *Project Data Sphere, LLC*, es promover la innovación a través de una plataforma que integre las investigaciones de la comunidad científica con las de la industria farmacéutica. Dentro de la plataforma, laboratorios como Bayer Healthcare Pharmaceuticals Inc, Eli Lilly, The National Cancer Institute, Pfizer y otros, comparten bases de datos de ensayos clínicos de pacientes con cáncer. Cualquier investigador puede acceder a estas bases de datos si justifica sus estudios a través de una aplicación en línea. Se busca que los resultados de las investigaciones se publiquen en

artículos científicos abiertos al público para fomentar la difusión de información acerca de diferentes tipos de cáncer.

La primera base de datos que se utiliza es un ensayo clínico para probar el funcionamiento de un nuevo tratamiento para pacientes con cáncer de la cabeza y el cuello. A esta base de datos se le aplican los métodos para explorar subgrupos. Se busca comparar las diferencias entre árboles resultantes de los dos métodos para los mismos datos.

La otra base de datos es un ensayo clínico de un tratamiento para pacientes con leucemia. Con base en diferentes estudios de otros ensayos clínicos, se definen subgrupos a evaluar y se le aplica el método de Dixon y Simon con su modificación con *JAGS*. Los resultados de los análisis se muestran en las siguientes secciones.

6.1. Cáncer en el área de la cabeza y cuello

El estudio a analizar contiene la información obtenida de un ensayo clínico cuyos resultados y documentación se publicaron en un artículo de Vermorken et al. (2013). El estudio compara el efecto de quimioterapia acompañada de un anticuerpo monoclonal llamado Panitumumab contra el efecto de quimioterapia sola. Se aplica a pacientes con cáncer en la cabeza o en el cuello. La quimioterapia estaba compuesta por los agentes cisplatino y 5-fluorouracilo.

Un anticuerpo monoclonal es aquel que se genera de un solo clon de linfocitos B (*What are monoclonal antibodies?* 2018). Para generar el anticuerpo, primero se extraen linfocitos de la médula ósea. Se separan los linfocitos de tipo B y estos se fusionan con células cancerígenas

(mielomas) para que los linfocitos sobrevivan fuera del cuerpo y se puedan cultivar. Posteriormente, se seleccionan y diluyen los híbridos de tal forma que se obtengan híbridos con una sola célula de tipo B que sean capaces de crear anticuerpos. Cada célula se clona múltiples veces y como resultado se obtienen anticuerpos idénticos, a los cuales se les llama monoclonales. Al utilizar anticuerpos monoclonales en el ensayo clínico, se eliminan algunas posibles causas de heterogeneidad en el tratamiento.

El principal criterio de elegibilidad para que un paciente fuera admitido al ensayo clínico era que tuviera cáncer recurrente y/o metastásico en la cabeza y cuello. Esto es que, tras la eliminación de los tumores, el cáncer haya vuelto a surgir en la misma zona o que se haya propagado a diferentes órganos.

Además, el cáncer debía manifestarse a través de carcinomas (tumores malignos) en células planas o escamosas del epitelio. El epitelio es un tejido que se presenta en múltiples órganos del cuerpo, como por ejemplo en la capa externa de la piel, la epidermis. Las células planas o escamosas del epitelio se presentan en la cabeza y cuello en la boca, esófago, laringe y faringe. Se llaman planas porque vistas lateralmente, las células tienen poca altura o escamosas porque vistas desde arriba, las células tienen forma de hexágono.

Algunos otros criterios para la inclusión en el ensayo clínico eran que los pacientes fueran mayores de edad (18 años), que la enfermedad se hubiera determinado incurable con cirugía o radioterapia y que obtuvieran un nivel de 0 o 1 en la escala ECOG*.

La escala ECOG de estatus de desempeño mide el impacto que tiene

*Eastern Cooperative Oncology Group.

una enfermedad en las habilidades diarias de un paciente (*ECOG Performance Status* 2018). Los niveles 0 y 1 se refieren a estados donde la enfermedad no tiene mucho impacto en el paciente y puede realizar sus tareas cotidianas de una forma casi normal. Los niveles 2 a 4 se refieren a estados en los que el paciente necesita distintos grados de ayuda para sobrevivir y el nivel 5 refleja la muerte del paciente.

El estudio se realizó con una muestra de 657 individuos, los cuales se dividieron en un grupo de control y un grupo de tratamiento con 330 y 327 individuos, respectivamente. Los grupos se seleccionaron a través de aleatorización estratificada por tratamientos anteriores, el lugar del tumor principal y el nivel en la escala ECOG. La base de datos obtenida de *Project Data Sphere*, contenía información de 520 sujetos, de los cuales cada grupo (control y tratamiento) tenía asignados a 260 individuos. El estudio es de fase III, ya que compara un tratamiento nuevo (quimioterapia con Panitumumab) con el tratamiento estándar (quimioterapia) para el cáncer de cabeza y cuello.

El objetivo principal del estudio era evaluar si el nuevo tratamiento mejora la supervivencia general de los pacientes. Es decir, aumenta el tiempo en el que tardan en fallecer los pacientes. Otros objetivos eran evaluar si aumentaba el tiempo sin que avanzara la enfermedad, la respuesta general al tratamiento, el tiempo que tardaba un paciente en responder al tratamiento e investigar si había subgrupos determinados por marcadores biológicos o genéticos.

El estudio se condujo en múltiples centros de investigación alrededor del mundo. Se administraron máximo 6 ciclos del tratamiento cada 21 días y se hicieron evaluaciones de los tumores cada 6 semanas. Si los pacientes presentaban un avance en la enfermedad, eran intolerantes al tratamiento

o se morían, se dejaba de administrar el tratamiento. Independientemente de si los pacientes recibían todos los ciclos, se buscaba hacer una revisión mensual a la conclusión de la administración del tratamiento, así como esporádicos a largo plazo. Al menos 325 pacientes en cada grupo recibieron una dosis de quimioterapia o el anticuerpo, y se incluyeron en el estudio. Sin embargo, todos los pacientes no acabaron los ciclos de quimioterapia y casi todos (324) no acabaron los ciclos del anticuerpo.

En ensayo clínico evaluado por Vermorken et al. (2013) se encontró una diferencia en la mediana de la supervivencia general. Sin embargo, los intervalos de confianza se empalmaban y el valor-p no fue lo suficientemente pequeño como para decir que la diferencia entre tratamientos fue significativa. De igual forma, se encontró que había un aumento significativo en el tiempo sin que avanzara el cáncer y una disminución significativa en la recurrencia del cáncer.

Los análisis de subgrupos en el artículo mostraron ciertas diferencias, pero no se deben tomar en cuenta porque no siguen las estipulaciones de varias guías para la publicación de análisis de subgrupos (R. Wang et al. 2007; Rothwell 2005; Yusuf, Wittes et al. 1991; Lagakos 2006). Por ejemplo, no se especifica en el artículo la cantidad de pruebas realizadas, los métodos que se utilizaron para realizar los análisis, si los subgrupos fueron predefinidos o encontrados después del análisis ni si se hicieron pruebas de interacción. A pesar de las carencias, Vermorken et al. (2013) recalcan que los resultados de los análisis de subgrupos no se deben sobreinterpretar, algo positivo en la publicación de subgrupos. También explican que los subgrupos están formados por un número pequeño de individuos. Esto surge de que el estudio fue diseñado para encontrar principalmente el efecto general del tratamiento.

6.1.1. Exploración

Los archivos obtenidos de *Project Data Sphere* fueron varias bases de datos, en las que se medían diferentes criterios para cada individuo. La preparación de los datos siguió el siguiente proceso:

1. Seleccionar covariables de interés: Dentro de más de 100 covariables se escogieron para evaluar la edad, el sexo, la raza del paciente y la frecuencia con que el paciente fuma tabaco o bebe alcohol. Estas variables se seleccionaron por su simplicidad y fácil interpretación.
2. Crear resúmenes de algunas covariables: Al recopilar la historia médica de los pacientes, se les preguntó si nunca, alguna vez o actualmente fumaban tabaco o tomaban alcohol. En la base de datos, se incluye solamente la frecuencia, pero no la substancia que el paciente utilizó. Es decir, se puede saber si el paciente alguna vez fumó tabaco o tomó alcohol, pero no cuál de las dos substancias utilizó. Por lo tanto, se tomó la frecuencia más alta para las dos substancias y se resumió en una covariable que indica la frecuencia de uso para ambas. Por ejemplo, si un paciente nunca fumó tabaco, pero actualmente toma alcohol, la covariable que indica si fumó o tomó tendrá el valor correspondiente a que actualmente utiliza alguna de las substancias.
3. Escoger una variable de respuesta: En el artículo de Vermorken et al. (2013) se encontró que no hubo una mejora significativa en la supervivencia general para pacientes con tratamiento contra pacientes del grupo control, pero hubo un aumento significativo en el tiempo sin que avanzara el cáncer. Por lo tanto, se utilizó un resumen del número de días sin que progresara la enfermedad o el

paciente muriera como variable de respuesta. Como en el método *Virtual Twins* se necesita que la variable de respuesta sea binaria, se creó una variable indicadora de si el número de días estaba por debajo de la media o no y se utilizó como la variable de respuesta.

4. Crear indicadoras auxiliares para las variables categóricas: Para cada clase de las variables categóricas se crearon nuevas variables auxiliares que indican si un individuo pertenece a una clase. Por ejemplo, para la covariante fuma o toma, si un paciente no ha tomado nunca, la covariante de nunca fuma o toma tendrá el valor 1, la de anteriormente el valor de 0 y la de actualmente el valor 0.
5. Unir las covariables en una única base de datos: Al encontrarse las covariables en distintas bases de datos, fue necesario fusionarlas para crear una base con toda la información de cada paciente.
6. Eliminar observaciones con información incompleta: Si el valor de una o más de las covariables estaba vacío o era “NA” se eliminaba la observación.

Después de la preparación de los datos se obtuvo la base de datos que se utilizó para hacer los análisis y para cual los valores de sus covariables se pueden resumir en el cuadro 6.1:

Covariable	Resumen o Clase	Observaciones
Grupo	Control	248
	Tratamiento	251
Edad	Mínimo	30
	1er cuartil	53
	Mediana	58
	3er cuartil	63
	Máximo	82
	Media	57.8
Sexo	Femenino	64
	Masculino	435
Raza	Blanca o caucásica	453
	Asiática	42
	Negra o Afroamericana	1
	Hispana o Latina	1
	Otras	2
Fuma o toma	Nunca	39
	Anteriormente	217
	Actualmente	243

Cuadro 6.1: Análisis exploratorio para base de datos de cáncer en cabeza y cuello.

Se grafican las características de las observaciones para ver la composición de la base de datos y se muestran en la figura 6.1:

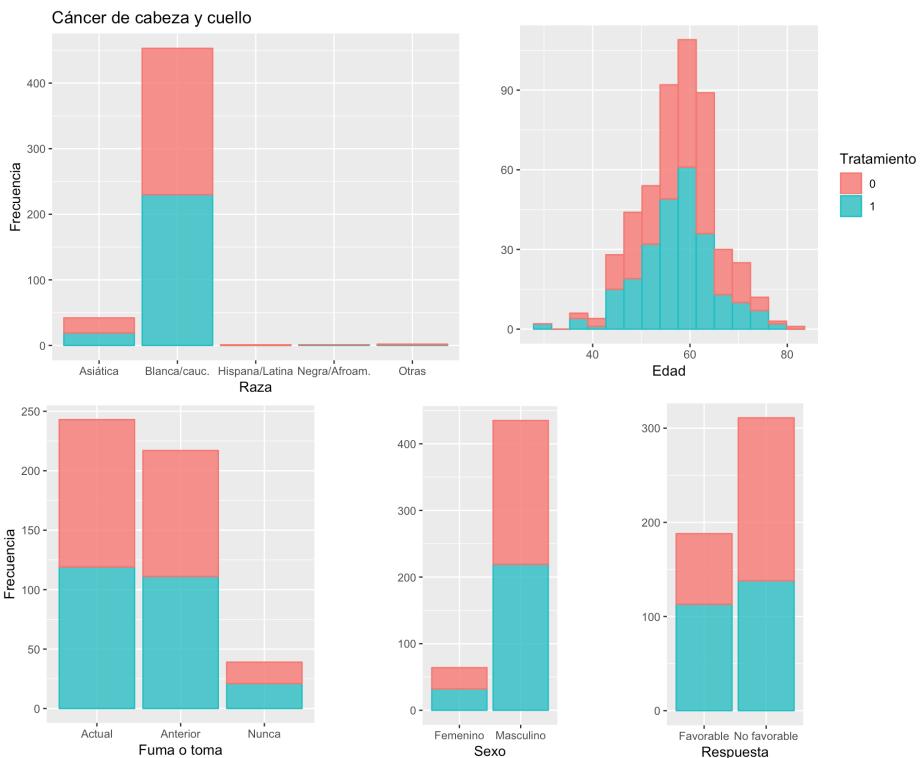


Figura 6.1: Frecuencia de pacientes con respecto a variables.

Se puede ver que para la mayoría de las categorías el número de personas en el grupo de tratamiento y el control está balanceado. Sin embargo, el número de personas en cada categoría no está balanceado para las variables. Esto es particularmente notable en la covariable de raza. Se debe tener cuidado con estas variables ya que pueden dar lugar a resultados no confiables.

6.1.2. *Virtual Twins*

Una vez preparada base de datos, se le aplica el método *Virtual Twins*. El proceso de aplicación del método y los resultados expuestos son los mismos que se utilizaron en la base de datos simulados. Lo primero que se hizo para aplicar el método fue ajustar los bosques aleatorios para los tratamientos originales y los tratamientos invertidos (gemelo). Cada bosque se crea con 500 árboles. En el siguiente cuadro, se resumen los resultados de las predicciones de los estimadores:

	Media de \hat{P}_1	Media de \hat{P}_0
$T_i = 1$	0.311	0.117
$T_i = 0$	0.101	0.293

Las predicciones del árbol con tratamiento original son mayores que para el árbol gemelo. Esto es, se predice una mayor probabilidad de un resultado favorable para el tratamiento real del paciente que para el tratamiento invertido. También, se calcula Z_i para cada individuo. En el caso de los datos, la media de Z_i es 0.20234, lo que significa que hay en promedio una diferencia de 0.202 entre el resultado de pacientes con tratamiento y con control.

Por otro lado, también se calcula δ , que para los datos es 0.1477798. Quiere decir que utilizar el tratamiento aumenta en 0.148 la probabilidad de obtener un resultado favorable.

Las variaciones que se aplican de *Virtual Twins* son nuevamente el árbol de regresión y el árbol de clasificación, cada uno con los valores de umbral $c = \delta + 0.1$ y $c = \delta + 0.05$.

***Virtual Twins* con árbol de regresión**

Con $c = \delta + 0.1$

Para los datos de cáncer de cabeza y cuello se obtuvo el siguiente árbol de regresión:

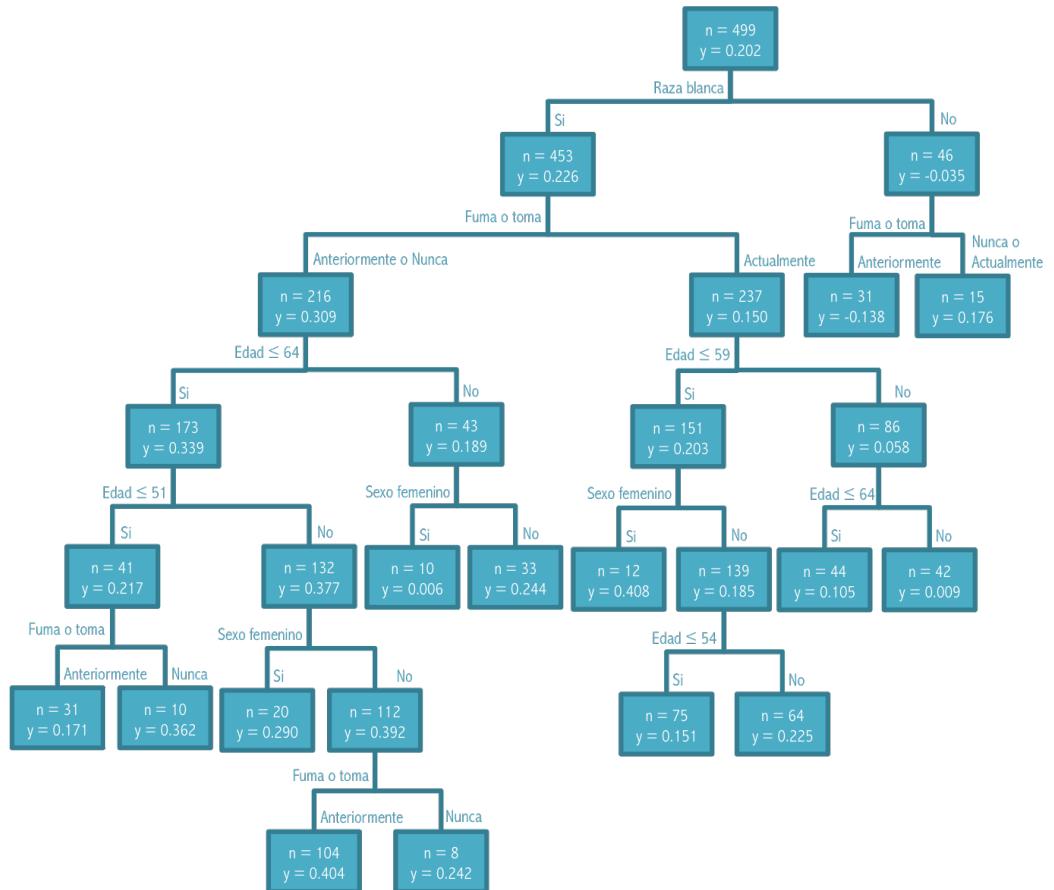


Figura 6.2: Árbol de regresión con $c = \delta + 0.1$

Las divisiones en el árbol dependen de todas las covariables. Para cada una se hace un análisis:

- Raza: La primera división del árbol separa a las observaciones en individuos de raza blanca y de otras razas. Se puede ver que la respuesta predicha para individuos de raza blanca es mayor que para el de otras razas. En algunos estudios se ha encontrado que pacientes de raza blanca reciben diferentes tratamientos que otras razas (Bach et al. 1999). A pesar de que en el estudio clínico el tratamiento es el mismo para todas las razas, puede ser que un mejor cuidado de los pacientes de raza blanca sea la causa de una mayor probabilidad de un resultado favorable. Una causa más probable de la diferencia de efectos es que el número de individuos de raza blanca es mayor que de otras razas.
- Fuma o toma: En general, se prefiere que el paciente fumara o tomara anteriormente o nunca lo hiciera a que actualmente fume o tome. No hay mucha diferencia entre pacientes que fumaban o tomaban anteriormente y pacientes que nunca lo hicieron, ya que en algunas divisiones una categoría resulta favorable y en otras perjudica.
- Edad: No se puede concluir que la edad influye en el resultado del tratamiento ya no hay un comportamiento constante en las divisiones del árbol que dependen de la edad.
- Sexo: Al igual que con la edad, el comportamiento de las divisiones determinadas por el sexo del paciente no es constante, por lo que no se puede concluir que afecte el funcionamiento del tratamiento.

A continuación, se muestra el subgrupo donde hay mayor heterogeneidad en el tratamiento encontrado por el método. En el cuadro

6.2 se muestra el porcentaje de respuestas favorables en el subgrupo para individuos en el grupo de tratamiento y de control.

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
Raza blanca, fuma/toma actual, edad ≤ 59 y sexo masculino.	12	71.4 %	20 %
Raza blanca, fuma/toma nunca y edad ≤ 51 .	10	50 %	0
Raza blanca, fuma/toma anterior, fuma/toma nunca, $52 \leq$ edad ≤ 64 y sexo masculino.	20	60 %	50 %
Raza blanca, fuma/toma anterior, $52 \leq$ edad ≤ 64 y sexo femenino.	104	53.6 %	22.9 %

Cuadro 6.2: Probabilidades para subgrupos con $c = \delta + 0.1$

Con $c = \delta + 0.05$

El árbol de regresión del método con $c = \delta + 0.05$ es el mismo que para $c = \delta + 0.1$ (véase figura 6.2). Sin embargo, la región \hat{A} tiene una configuración diferente, mostrada en el cuadro 6.3:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
Raza blanca, fuma/toma actual, $55 \leq \text{edad} \leq 59$ y sexo masculino.	64	54.3 %	24.1 %
Raza blanca, fuma/toma actual, $\text{edad} \leq 59$ y sexo femenino.	12	71.4 %	20 %
Raza blanca, fuma/toma anterior, fuma/toma nunca, $\text{edad} \geq 65$ y sexo masculino.	33	35.3 %	25 %
Raza blanca, fuma/toma nunca y $\text{edad} \leq 50$.	10	50 %	0
Raza blanca, fuma/toma anterior, fuma/toma nunca, $52 \leq \text{edad} \leq 64$ y sexo femenino.	20	60 %	50 %
Raza blanca, fuma/toma nunca, $52 \leq \text{edad} \leq 64$ y sexo masculino.	8	40 %	33.3 %
Raza blanca, fuma/toma actual, $52 \leq \text{edad} \leq 64$ y sexo masculino.	104	53.6 %	22.9 %

Cuadro 6.3: Probabilidades para subgrupos con $c = \delta + 0.05$.

Para $c = \delta + 0.05$, hay más subgrupos y más individuos que pertenecen

a \hat{A} . No obstante, con ninguno de los valores c se encuentra una regla simple para describir a la región \hat{A} . La pertenencia de una observación a \hat{A} depende de una serie de covariables que tienen valores diferentes. Es decir, los efectos heterogéneos del tratamiento dependen de diferentes interacciones entre covariables. Se puede entonces concluir que ninguna covariable presenta efectos heterogéneos en subgrupos por sí sola.

Otra diferencia entre los resultados que se obtienen al cambiar los valores de c se da en la medida de desempeño $\hat{Q}(\hat{A})$:

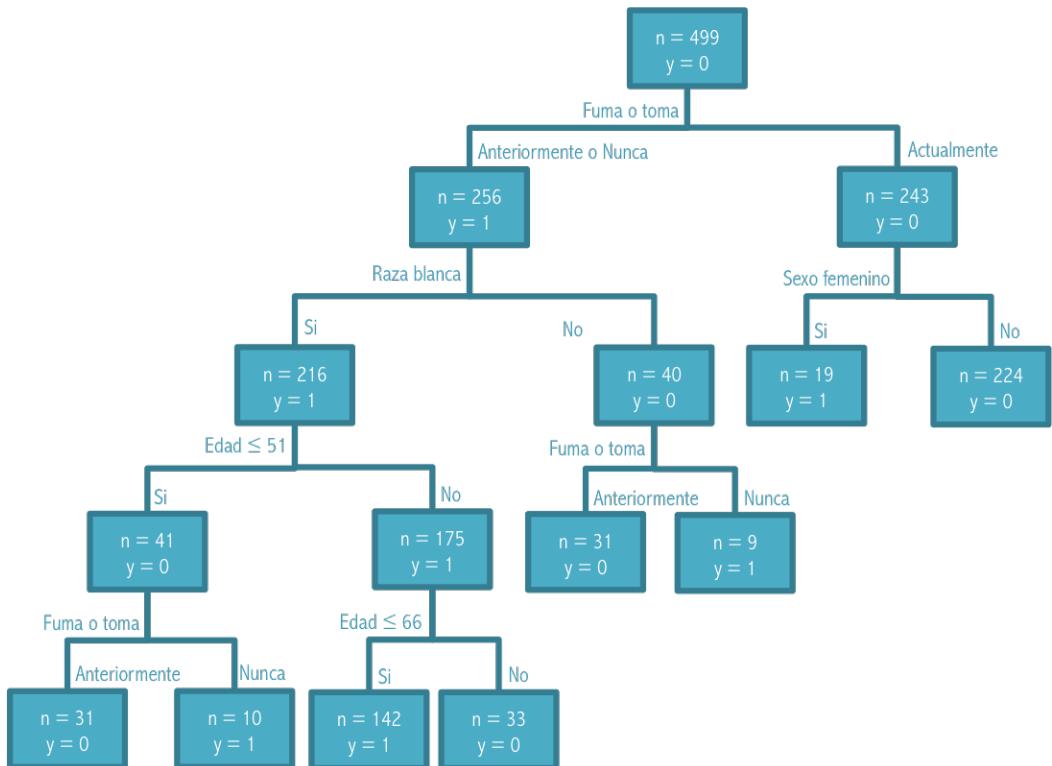
\hat{A}	$ \hat{A} $	$\hat{Q}(\hat{A})$
$c = \delta + 0.1$	146	0.1785338
$c = \delta + 0.05$	251	0.1177343

La medida de desempeño para la región \hat{A} obtenida con $c = \delta + 0.05$ es menor que con $c = \delta + 0.1$ y como en el ejemplo de simulación se puede suponer que las observaciones adicionales no aportan información relevante cuando se relaja el umbral a $c = \delta + 0.05$.

***Virtual Twins* con árbol de clasificación**

Con $c = \delta + 0.1$

El árbol de clasificación que se encuentra para $c = \delta + 0.1$ tiene menos divisiones que los árboles de regresión, como se muestra en la figura 6.3:


 Figura 6.3: Árbol de clasificación con $c = \delta + 0.1$.

El análisis de las divisiones por covariable es el siguiente:

- Fuma o toma: parece que es la variable más importante para la definición del subgrupo ya que aparece en la primera división del árbol. En todos los casos, el que un paciente fume o tome anterior o actualmente resulta en una predicción de respuesta no favorable. Por otro lado, si un paciente nunca tomó o fumó hace que la predicción de su respuesta sea favorable.

- Raza: Al igual que en el árbol de regresión, el que un paciente sea de raza blanca mejora la respuesta al tratamiento.
- Sexo: Hay sólo una división en el árbol que depende del sexo. Esta división se encuentra en la rama donde los pacientes fuman actualmente. Hay estudios que muestran que, en promedio, las mujeres fuman menos que los hombres, lo cual puede ser una explicación de una mejor respuesta predicha para mujeres (Dorak y Karpuzoglu 2012).
- Edad: Hay dos divisiones en el árbol que dependen de la edad. En una de ellas, una menor edad significa una mejor respuesta al tratamiento, mientras que, en la otra significa una peor respuesta. Por lo tanto, no se pueden hacer conclusiones sobre el efecto que tiene la edad sobre la respuesta.

Con base en estas variables se crean los subgrupos que conforman la región \hat{A} , los cuales se muestran en el cuadro 6.4:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
Fuma/toma actual y $56 \leq \text{edad} \leq 58.$	41	52 %	12.5 %
Fuma/toma nunca, raza blanca y edad $\leq 50.$	7	50 %	0
Fuma/toma nunca, raza blanca, $51 \leq \text{edad} \leq 62$ y sexo femenino.	10	75 %	33.3 %
Fuma/toma nunca, fuma/toma anterior, raza blanca, $51 \leq \text{edad} \leq 72$ y sexo masculino.	142	50 %	25 %

Cuadro 6.4: Probabilidades para subgrupos con $c = \delta + 0.1.$

Los subgrupos que pertenecen a \hat{A} no se pueden describir con reglas simples, lo cual, al igual que con el árbol de regresión sugiere que la heterogeneidad del tratamiento no depende de una covariable.

Con $c = \delta + 0.05$

El árbol de clasificación obtenido con $c = \delta + 0.05$ se grafica en la figura 6.4:

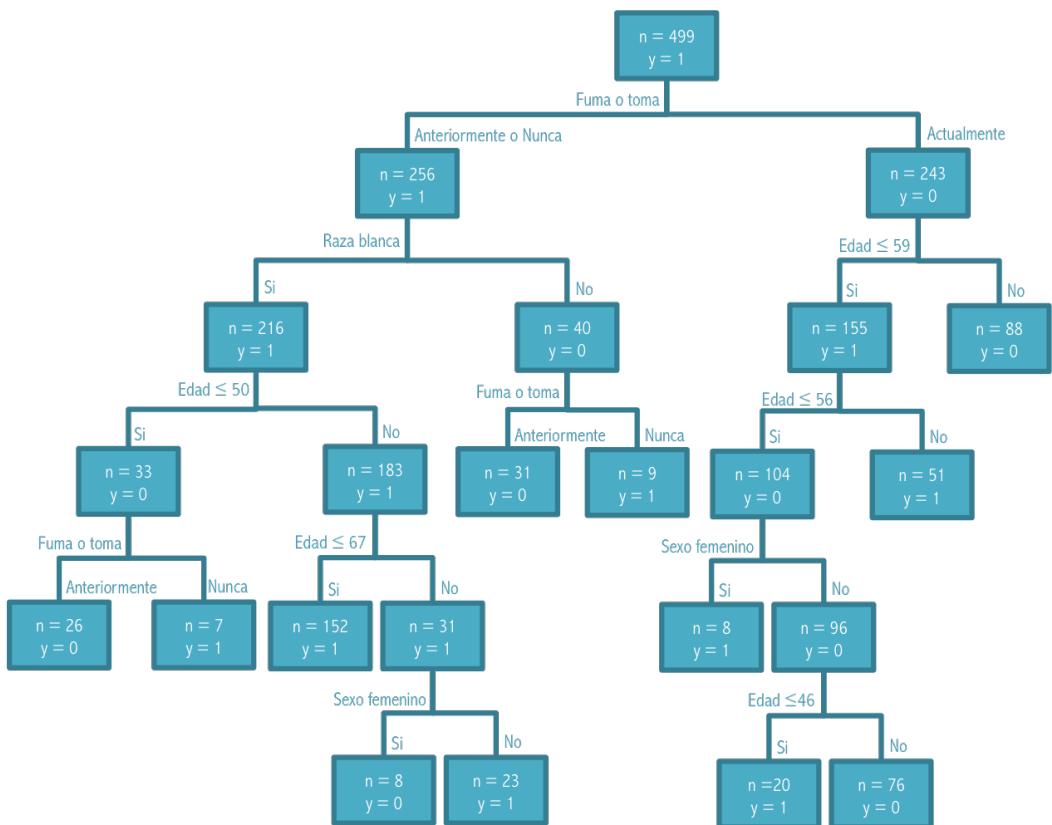


Figura 6.4: Árbol de clasificación con $c = \delta + 0.05$.

Para este árbol, sólo una de las covariables tiene un comportamiento constante: raza. Sólo una división está determinada por la raza y esta separa a los pacientes en aquellos de raza blanca y las demás razas. Al igual que en los otros árboles, si un paciente es de raza blanca se predice una respuesta favorable.

Como las demás variables no tienen un comportamiento constante (el resultado favorable no depende del valor que toman en las divisiones) es

difícil interpretar el árbol. Además, los subgrupos encontrados para esta variación no se pueden describir con una regla simple, por lo que se puede inferir que la heterogeneidad en efectos depende de interacciones de las covariables. Esto se muestra en el cuadro 6.5:

Subgrupo	Tamaño	$Y = 1$ con tratamiento	$Y = 1$ con control
Fuma/toma actual y $45 \leq \text{edad} \leq 47.$	15	50 %	28.6 %
Fuma/toma actual y $55 \leq \text{edad} \leq 59.$	71	55 %	22.6 %
Fuma/toma nunca y raza no blanca.	9	40 %	0
Fuma/toma nunca, raza blanca y edad $\leq 50.$	7	50 %	0
Fuma/toma anterior, Fuma/toma nunca, raza blanca, $51 \leq \text{edad} \leq 64$ y sexo femenino.	23	54.5 %	41.7 %
Fuma/toma anterior, Fuma/toma nunca, raza blanca, edad ≥ 51 y sexo masculino.	150	48.8 %	24.3 %

Cuadro 6.5: Probabilidades para subgrupos con $c = \delta + 0.05.$

Las características de la región \hat{A} de los árboles de clasificación se resumen el cuadro 6.6:

\hat{A}	$ \hat{A} $	$\hat{Q}(\hat{A})$
$c = \delta + 0.1$	200	0.1475226
$c = \delta + 0.05$	275	0.1108037

Cuadro 6.6: Región \hat{A} obtenida con los árboles de clasificación.

En todos los análisis hechos en esta tesis se obtuvo una región con mejor desempeño cuando $c = \delta + 0.1$, por lo que se entiende que sólo se deben agregar observaciones a la región si el efecto hay mucha heterogeneidad en tratamiento, si no la penalización es grande. En este caso, los árboles de regresión tienen un mejor desempeño que los árboles de clasificación.

En ninguna variación del método se encontró una región $\hat{A} = \emptyset$, lo que quiere decir que hay heterogeneidad en el tratamiento. No obstante, para todas las variaciones del método la región \hat{A} encontrada dependía de muchas interacciones de las covariables y no se encontraron similitudes en los subgrupos de los diferentes métodos. Esto hace pensar que la heterogeneidad, si es que existe, depende de las interacciones complicadas entre covariables. Por lo tanto, se puede concluir que no hay efectos en subgrupos para las covariables seleccionadas.

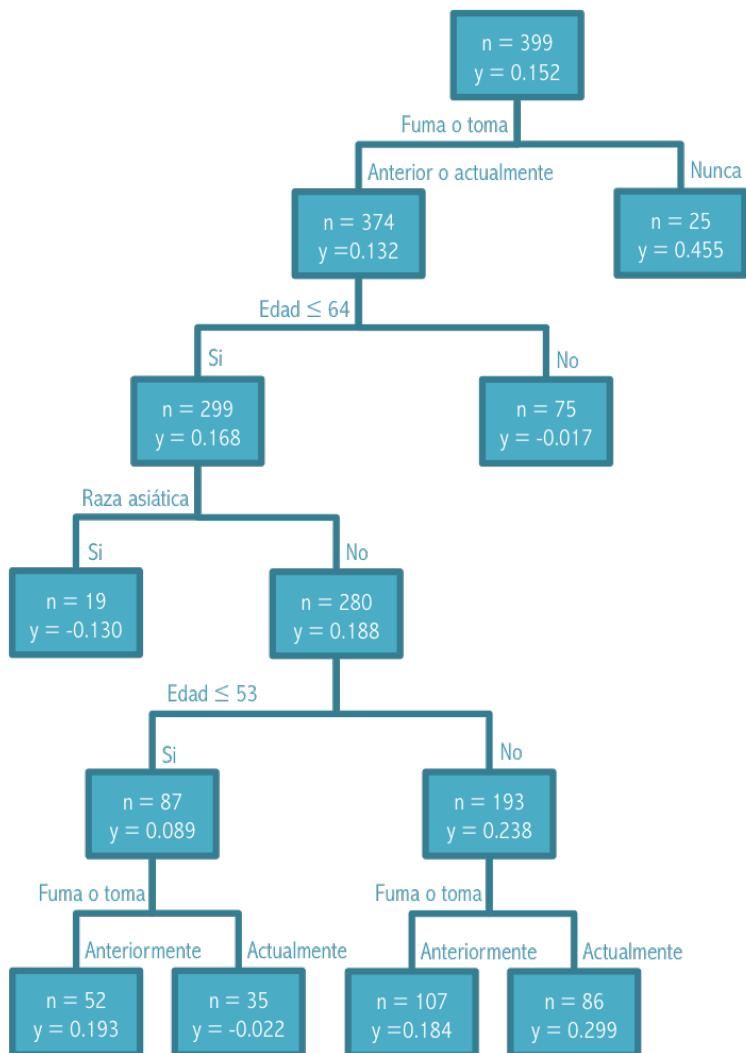
Por otro lado, la medida de desempeño de la región fue positiva en todos los casos, lo que significa que hay una diferencia en el efecto del tratamiento. Un análisis posterior podría incluir a aplicación del método las observaciones descritas con otras covariables. Sin embargo, el objetivo de los análisis no debe ser encontrar subgrupos, sino conocer la base de datos. Realizar un número indefinido de análisis hasta encontrar subgrupos es conflictivo y engañoso. Por lo que la conclusión del análisis es que el método *Virtual Twins* no encontró subgrupos para la base de datos.

6.1.3. Método de árboles de interacción

Con motivo de comparación, se aplica el método de árboles de interacción a la base de datos de cáncer de cabeza y cuello. La base de datos del ensayo clínico se divide en un conjunto de aprendizaje y uno de prueba. Como en la base de datos simulados, el conjunto de aprendizaje se conforma con 80 % de las observaciones.

A continuación, se muestran los resultados de cada paso que realiza el método de árboles de interacción:

1. El árbol inicial T_0 obtenido se muestra en la figura 6.5:

Figura 6.5: Árbol inicial T_0 .

Los criterios de paro para obtener el árbol fueron que cada nodo tuviera como mínimo 20 observaciones y que el árbol tuviera una profundidad de máximo 15 nodos. Se utilizan estos criterios porque

son los predeterminados para el método definido por Su et al. en su código para R.

2. Podar el árbol:

- a) La sucesión anidada de árboles podados se representa en la figura 6.6, donde cada color representa la rama más débil elegida en una de las iteraciones:

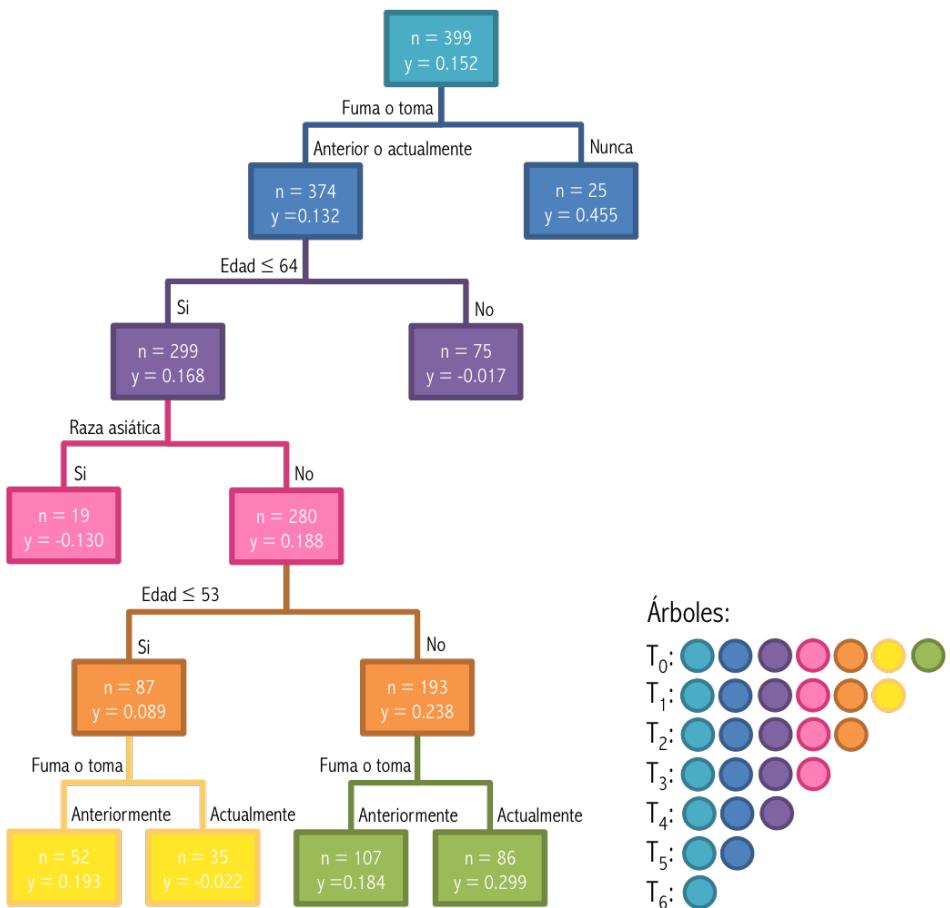


Figura 6.6: Sucesión de árboles anidados.

- a) El cálculo de la estadística $G_\lambda(T_m)$, con $\lambda \in \{2, 3, 4, \ln(n)\}$, para cada árbol de la sucesión anidada se presenta en el cuadro 6.7:

Subárbol	$G_2(T_m)$	$G_3(T_m)$	$G_4(T_m)$	$G_{\ln(n)}(T_m)$
T_0	-6.14	-12.14	-18.14	-21.77
T_1	-4.37	-9.37	-14.37	-17.40
T_2	-2.72	-6.73	-10.73	-13.15
T_3	-0.95	-3.95	-6.95	-8.77
T_4	-0.99	-1.99	-4.99	-6.20
T_5	0.93	-0.06	-1.06	-1.67
T_6	0	0	0	0
Máximo	T_5	T_6	T_6	T_6

Cuadro 6.7: Estadística $G_\lambda(T_m)$ para la sucesión anidada de árboles con diferentes valores de λ .

Para todos los árboles, exceptuando el nulo y T_5 con $\lambda = 2$, $G_\lambda(T_m)$ es negativo. Esto quiere decir que, en general, el número de nodos adicionales a la raíz no aumenta lo suficiente la interacción en el árbol. Dicho de otra forma, el tratamiento no es lo suficientemente heterogéneo en alguna de las divisiones como para agregar nodos. Como Su et al. (2009) mencionan que el mejor criterio para escoger el árbol óptimo es $\lambda = \ln(n)$, se utiliza $G_{\ln(n)}(T_m)$ como criterio de selección. Por lo tanto, se toma el árbol nulo como el óptimo. Entonces, de acuerdo a este método no hay diferencias en efectos de subgrupos significativas para la base de datos analizada.

$$\begin{aligned} n &= 399 \\ y &= 0.152 \\ n_0 &= 196 \\ n_1 &= 203 \end{aligned}$$

Figura 6.7: Árbol final

3. El paso en el que se fusionan nodos similares no se realiza ya que el árbol final es aquél que sólo tiene la raíz.

A pesar de que el árbol final que se encontró fue el árbol nulo, cabe evaluar árboles anteriores como complemento al análisis. En el árbol inicial, la división más importante está determinada por la frecuencia con la que fuma un paciente. Las divisiones en el árbol reflejan que, si un paciente nunca ha fumado o tomado, el efecto del tratamiento y es mayor que para cualquier otro subgrupo. Esta observación concuerda con la observación médica que fumar tabaco en cigarro es una causa del cáncer. Parece que no hay mucha diferencia entre los pacientes que fuman o toman alcohol actualmente y aquellos que lo hacían anteriormente ya que en una división favorece a unos pacientes y otra los perjudica.

En las divisiones que corresponden a la edad, no se encuentra un comportamiento constante. Es decir, dependiendo del subgrupo una menor edad puede sugerir una mejor o una peor respuesta al tratamiento. Esto se debe probablemente a que la muestra no está balanceada respecto a la edad. Por otro lado, ninguna de las divisiones depende del sexo del paciente. Esto nos hace pensar que es una covariante irrelevante para la definición de subgrupos.

Por último, se calcula la importancia de las covariables a través de V_j en el cuadro 6.8:

Covariable	V_j	
Edad	45.73	
Sexo	21.38	
Fuma o toma	9.11	
	Anteriormente	23.92
	Actualmente	24.68
Raza	Blanca o caucásica	10.02
	Asiática	8.84
	Negra o Afroamericana	0
	Hispana o Latina	0
	Otras razas	0

Cuadro 6.8: Estadística de importancia V_j para las diferentes covariables.

La covariable de edad es la segunda covariable con mayor importancia según el algoritmo, a pesar de que esta importancia no se encontró un comportamiento constante en las divisiones del árbol. La importancia de la edad tiene fundamentos biológicos ya que, para una persona joven el cuerpo humano tiene mayor habilidad para sanar y su esperanza de vida es mayor que el de una persona mayor.

Una de las variables con menor importancia V_j es la raza, incluso para algunas categorías su valor es cero. Esto se debe probablemente a que mayoría de las observaciones la raza es blanca o caucásica. Las variables que obtuvieron cero importancia tenían dos o menos observaciones del total, lo cual dificulta el análisis. Por lo tanto, si la muestra no está balanceada no se

puede concluir que la covariable no tenga importancia en el funcionamiento del tratamiento.

De la misma forma, es probable que la indicadora auxiliar de que una persona nunca fumó o tomó tenga menos importancia que aquellas con otras frecuencias porque tiene un número de observaciones menor.

Por último, el sexo tiene una importancia relativamente alta a pesar de que la muestra está compuesta en su mayoría por hombres. Su importancia tiene fundamentos biológicos ya que hay muchas diferencias en el funcionamiento del cuerpo humano para los dos sexos. Ejemplos de diferencias son la composición hormonal y características físicas. Además, se han publicado estudios que sugieren que el sexo es un factor de heterogeneidad en tratamientos de cáncer (Kim, Lim y Moon 2018). Esto puede significar que el sexo influye en el funcionamiento del tratamiento.

Es importante recalcar que el método no encontró subgrupos. Esto habla de un buen diseño del método ya que no se busca arrojar subgrupos sólo para dar un análisis más interesante. Si se encuentran subgrupos con este método, significa que tienen una diferencia significativa en el tratamiento.

6.1.4. Conclusión de la base de datos de cáncer de cabeza y cuello

Una diferencia entre los métodos fueron las variables que definían los árboles y los tamaños de los árboles. A pesar de que los criterios de paro para ambos métodos eran similares, en *Virtual Twins* se encontraron árboles más grandes.

Por otro lado, el primero encontró subgrupos en todas sus variaciones, mientras que el segundo no encontró efectos en subgrupos. Los subgrupos encontrados en el primer método no eran descritos por reglas simples, lo que hace pensar que se dieron por mera probabilidad. Por lo tanto, se puede concluir que para esta base de datos no hay efectos en subgrupos y se reafirma la observación de que el método de árboles de interacción tiene un desempeño mejor que *Virtual Twins*.

6.2. Leucemia

La base de datos contiene la información obtenida de un ensayo clínico aplicado a pacientes con leucemia mieloide aguda. Sus resultados y documentación se publicaron en un artículo de Kolitz et al. (2010).

La leucemia mieloide aguda es un tipo de cáncer que afecta a los glóbulos blancos (leucocitos) que se encuentran en la médula ósea. La proliferación de leucocitos cancerígenos resulta en la disminución de glóbulos rojos, plaquetas y leucocitos saludables. Los pacientes con este tipo de leucemia suelen complicaciones derivadas de una disminución en su salud.

En el ensayo clínico se buscó comparar los efectos de aplicar quimioterapia solamente o aplicar quimioterapia acompañada por una serie de tratamientos. Los tratamientos se dividieron en tres etapas sucesivas:

1. Modulación MDR a través del tratamiento PSC-833 (también llamado Valspodar).

2. Terapia citogenética de intensificación con riesgo adaptado.
3. Inmunoterapia utilizando RIL-2

El objetivo principal del estudio era determinar si el tratamiento PSC-833 aumentaba el tiempo que los pacientes permanecían en remisión y el tiempo que los pacientes estaban vivos después del tratamiento. El resto de los objetivos tenían que ver con las otras etapas de tratamientos.

En esta tesis se analizarán los resultados de la primera etapa, ya que es la única etapa descrita en la base de datos obtenida de *Project Data Sphere*. Por lo tanto, se buscará analizar los efectos en subgrupos causados por quimioterapia acompañada por Valspodar, tomando como grupo de control la quimioterapia sola.

El tratamiento Valspodar es una droga que hace que las células cancerígenas tengan más sensibilidad a los efectos de la quimioterapia. Esto lo hace al inhibir p-glicoproteína en las células cancerígenas. A su vez, esto hace que la quimioterapia permanezca dentro de las células cancerígenas durante más tiempo. El uso de quimio sensibilizadores como PSC-833 buscan ser una alternativa a aumentar la dosis de la quimioterapia en algunos pacientes, ya que el aumento puede resultar en la intoxicación del paciente.

Los principales criterios para que un paciente fuera aceptado en el ensayo clínico eran que no hubieran recibido tratamientos contra la leucemia anteriormente, salvo algunas excepciones en tratamientos, y que tuvieran menos de 60 años. El tratamiento se administró en un régimen de 14 días. Al terminar, se hizo una biopsia de la médula ósea de los pacientes para evaluar la presencia de células cancerígenas. Si un paciente presentaba leucemia, era sometido a un segundo régimen de tratamiento

y se realizaba nuevamente una biopsia de la médula ósea. Se evaluaba el tiempo que los pacientes permanecían en remisión (no presentaban células cancerígenas) y el tiempo que permanecían vivos. Para las siguientes etapas, se volvían a asignar a los pacientes a grupos de tratamiento y control de manera aleatoria.

6.2.1. Definición de subgrupos

Como se ha establecido en capítulos anteriores, el funcionamiento de los métodos para confirmar subgrupos depende de la definir los mismos antes de realizar el análisis. Se buscó que anteriormente se hubieran encontrado efectos de subgrupos para tratamientos contra el cáncer en otros estudios. Este criterio se cumplió para subgrupos definidos por el sexo del paciente y son los subgrupos que se utilizarán para esta base de datos.

Las siguientes son diferencias determinadas por el sexo en pacientes con cáncer (Kim, Lim y Moon 2018; Dorak y Karpuzoglu 2012; J. Wang y Huang 2007):

- Los hombres presentan más neoplasias hematológicas, así como la mayoría de los otros tipos de cáncer. Las excepciones son cáncer en la tiroides, vesícula biliar y el ano, donde las mujeres son más susceptibles a tenerlos.
- En la infancia (menos de 12 años), los niños son generalmente más susceptibles a tener cáncer que las niñas.
- Las mujeres presentan más y mayores efectos secundarios a causa de la quimioterapia por toxicidad: diarrea, alopecia, estomatitis, entre otras.

6.2.2. Exploración

La preparación de la base de datos siguió un proceso similar al descrito en la sección de métodos exploratorios, pero cabe explicar la variable de respuesta. Se selecciona como variable de respuesta la indicadora de si el paciente entró en completa remisión después del tratamiento. En el artículo de Kolitz et al. (2010) se encontró que, tanto para el grupo de control como para el grupo con el tratamiento, el porcentaje de pacientes que entraron a remisión completa era del 75 %. Sin embargo, si se compara con otras variables, se descubre que el estudio tampoco encontró diferencias significativas entre el tratamiento y el control para otras posibles respuestas, como por ejemplo la supervivencia. Por otro lado, una ventaja de utilizar la indicadora de remisión como variable de respuesta es que es binaria. Por lo tanto, no se necesita modificar para poder utilizarla con el método de Dixon y Simon. A pesar de que no hay una diferencia entre los grupos, en los análisis de subgrupos se puede encontrar una parte de la población para la cual el efecto en el grupo control y de tratamiento sea distinto.

La base de datos obtenida en *Project Data Sphere* contiene una muestra de 302 pacientes de los cuales la mitad estaban en el grupo de control y la mitad en el grupo con tratamiento. Para cada observación había 14 covariables y la mayoría eran de tipo categórico. Tras la preparación se obtuvo una base de datos con los resúmenes mostrados en el cuadro 6.9 y en la figura 6.8:

Covariable	Clase	Observaciones
Grupo	Control	151
	Tratamiento	151
Sexo	Femenino	142
	Masculino	160

Cuadro 6.9: Características de las observaciones de acuerdo al tratamiento y sexo.

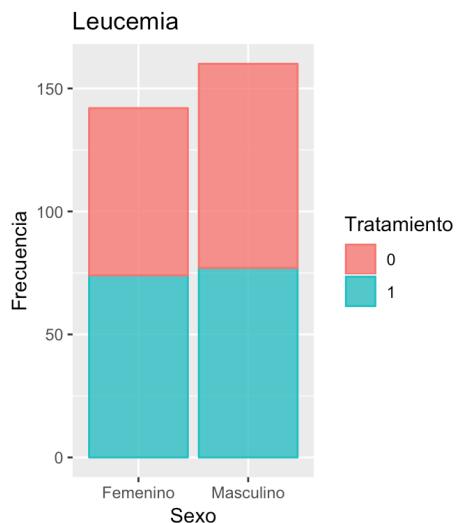


Figura 6.8: Histograma de observaciones de la base de datos de leucemia.

El ensayo clínico no fue diseñado para hacer análisis de subgrupos, pero la covariable de sexo está balanceada, ya que hay casi el mismo número de observaciones para el sexo femenino que para el sexo masculino. Como los subgrupos a evaluar están determinados por el sexo, este balance es importante para obtener resultados confiables.

6.2.3. Método de Dixon y Simon

Como se hizo en el capítulo de simulación, para aplicar el método de Dixon y Simon a la base de datos se utiliza el paquete **DSBayes** y se calculan la moda y los intervalos creíbles para cada subgrupo. Sin embargo, como en este caso sólo hay una covariable que define a los subgrupos, hay menos estimadores que calcular, mostrados en el cuadro 6.10:

Subgrupo	Tratamiento	Control	Diferencia
Mujer	$\mu + \tau + \beta + \gamma$	$\mu + \beta$	$\tau + \gamma$
Hombre	$\mu + \tau$	μ	τ

Cuadro 6.10: Estimadores a calcular para los dos subgrupos.

Los resultados de las estimaciones con sus intervalos creíbles se presentan en el cuadro 6.11 y la figura 6.9:

Subgrupo	Estimador	Moda	Intervalo creíble
Mujer	Tratamiento	1.288	(0.735, 1.852)
	Control	0.913	(0.341, 1.374)
	Diferencia	0.397	(-0.338, 1.193)
Hombre	Tratamiento	0.916	(0.422, 1.421)
	Control	1.282	(0.808, 1.948)
	Diferencia	-0.402	(-1.201, 0.323)

Cuadro 6.11: Moda e intervalo creíble para estimadores de los subgrupos.

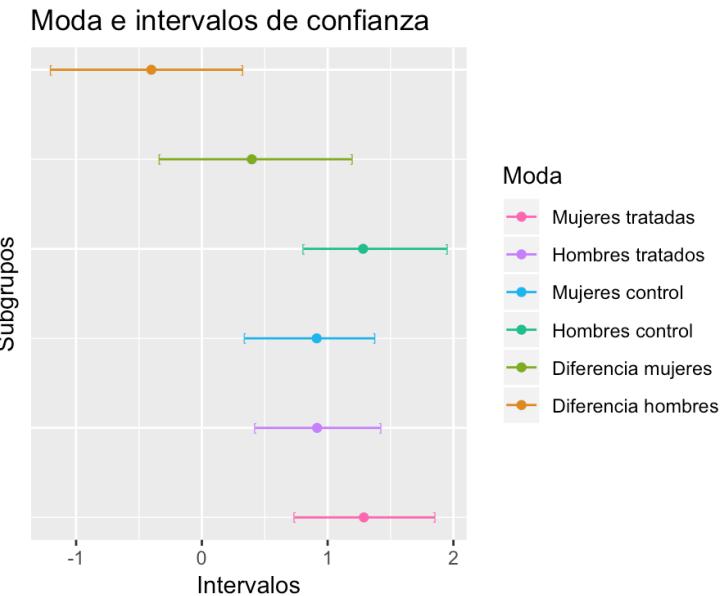


Figura 6.9: Intervalos creíbles con 95 % de probabilidad y la moda.

Para los dos subgrupos determinados por el sexo, tanto el estimador para el grupo control como para el grupo con tratamiento es positivo. Entonces, los dos tratamientos ayudan a que los pacientes entren en remisión completa.

La diferencia entre el estimador para el grupo con tratamiento y el grupo con control de las mujeres es positivo. Esto quiere decir que, para las mujeres el tratamiento aumenta las probabilidades de entrar en remisión completa. Por otro lado, hay una diferencia negativa en el subgrupo de los hombres. Es decir, utilizar el tratamiento disminuye las probabilidades de que un hombre entre en remisión.

Esta observación sugiere la existencia de heterogeneidad en subgrupos

determinados por el sexo. Sin embargo, los intervalos creíbles de las diferencias para cada subgrupo se empalman, por lo que no se puede confirmar su existencia (véase figura 6.9). Para lograr la confirmación se debe realizar análisis de subgrupos para un ensayo clínico del mismo tratamiento, pero con otras observaciones.

Como análisis adicional, se calculan las modas y los intervalos creíbles para los parámetros. Se exponen en el cuadro 6.12 y la figura 6.10:

Coeficiente	Moda	Intervalo creíble
μ	1.282	(0.808, 1.948)
τ	-0.402	(-1.201, 0.323)
β	-0.001	(-1.324, 0.117)
γ	0.741	(-0.741, 1.980)

Cuadro 6.12: Modas obtenidas para los parámetros.

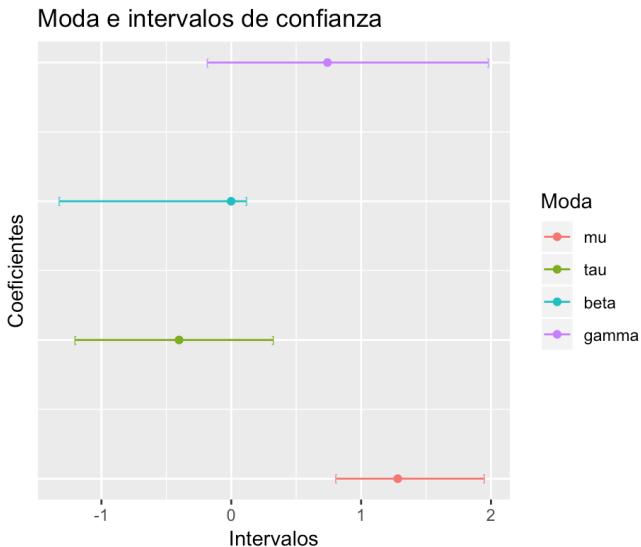


Figura 6.10: Intervalos creíbles del 95 % de probabilidad y la moda para los parámetros.

Analizando la moda encontrada para los parámetros, parece que la quimioterapia acompañada por los tratamientos no tiene un efecto extra al de la quimioterapia sola. Esto porque el valor estimado es cercano a cero y el intervalo creíble contiene al cero, una observación que complementa los resultados del estudio de Kolitz et. al (2010). Para la quimioterapia sola no hay efectos en subgrupos. Sin embargo, para el coeficiente que denota los efectos de subgrupos de la quimioterapia acompañada por los tratamientos, se encontró una moda igual a 0.741, sugiriendo efectos en subgrupos. El intervalo creíble contiene al cero por lo que, aunado con el análisis de las modas de la diferencia entre subgrupos, hace necesario realizar más análisis con diferentes ensayos clínicos para confirmar la existencia de efectos en subgrupos.

6.2.4. Modificación con *JAGS*

A pesar de que no se encontró gran diferencia entre los resultados del método de Dixon y Simon original y su modificación con *JAGS* en los datos simulados, se aplica la modificación con *JAGS* a la base de datos de leucemia como análisis adicional. Se utiliza un modelo creado con 10,000 simulaciones y 3 cadenas de Markov que se actualiza con 10,000 simulaciones. Se obtiene una muestra de 10,000 valores de cada cadena, para los cuales se grafica la traza y la densidad en la figura 6.11 y la figura 6.12

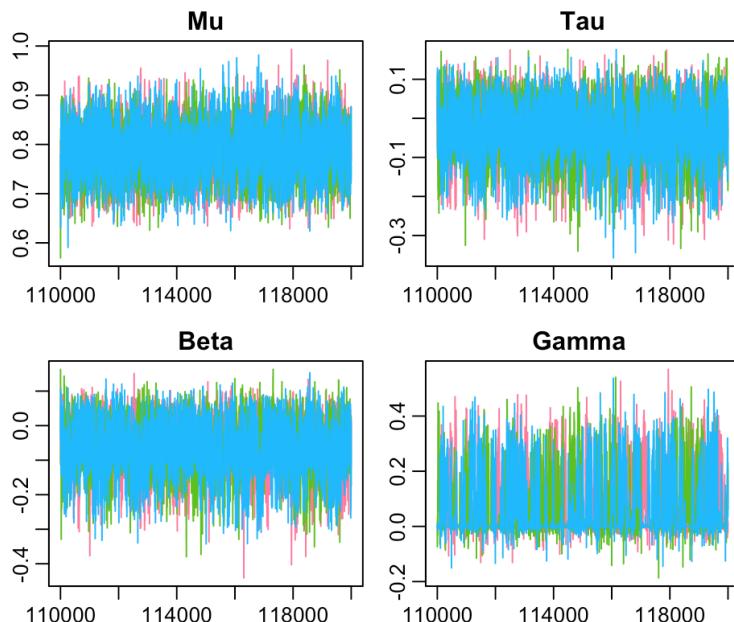


Figura 6.11: Trazas de los parámetros de cada cadena.

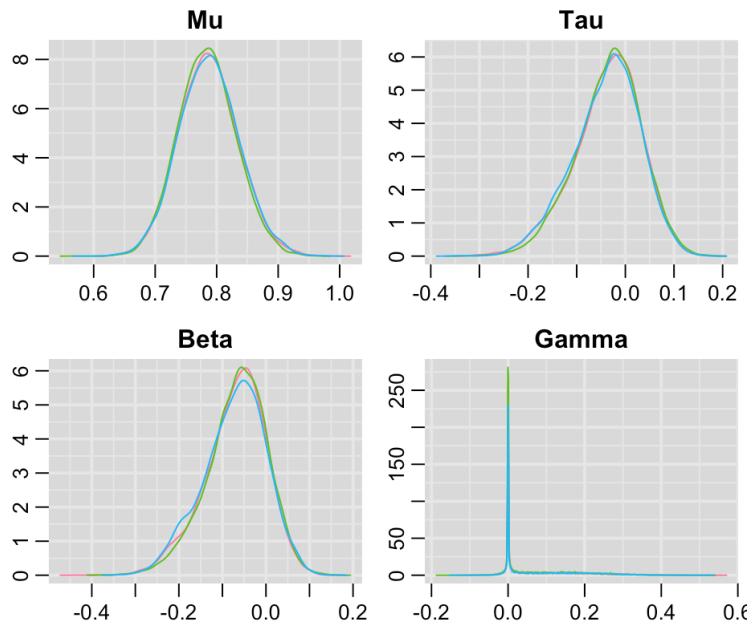


Figura 6.12: Densidades para los parámetros de cada cadena.

Todas las cadenas tienen forma similar, lo que sugiere que la distribución de los parámetros converge. En este caso, los resultados de aplicar el modelo con *JAGS* son diferentes a los del método de Dixon y Simon original. En el cuadro 6.13 y la figura 6.13 se muestran las medias y modas de cada parámetro, así como sus intervalos creíbles utilizando información de las tres cadenas:

Coeficiente	Media	Moda	Intervalo creíble
μ	0.786	0.788	(0.808, 1.948)
τ	-0.037	-0.022	(-1.201, 0.323)
β	-0.069	-0.528	(-1.324, 0.117)
γ	0.068	0.0007	(-0.741, 1.980)

Cuadro 6.13: Modas obtenidas para los parámetros.

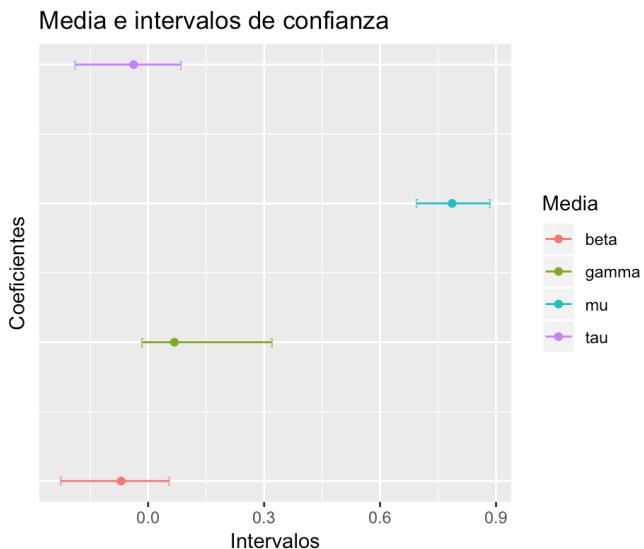


Figura 6.13: Intervalos creíbles con 95 % de probabilidad y la media.

Con este método, el coeficiente del efecto de la quimioterapia sola es del 0.786, menor que con **DSBayes**. El resto de los coeficientes tienen un estimador más cercano al cero. Lo que sugiere que el tratamiento no tiene efecto y tampoco hay efectos en subgrupos.

Notar que para la variable γ , la media y la moda no son parecidas, la moda es 0.0007 y la media 0.068. Esto se da porque, aunque la densidad es unimodal, tiene una cola larga a la derecha. Se puede apreciar mejor cuando se utiliza la información de las tres cadenas para graficar la densidad en la figura 6.14:

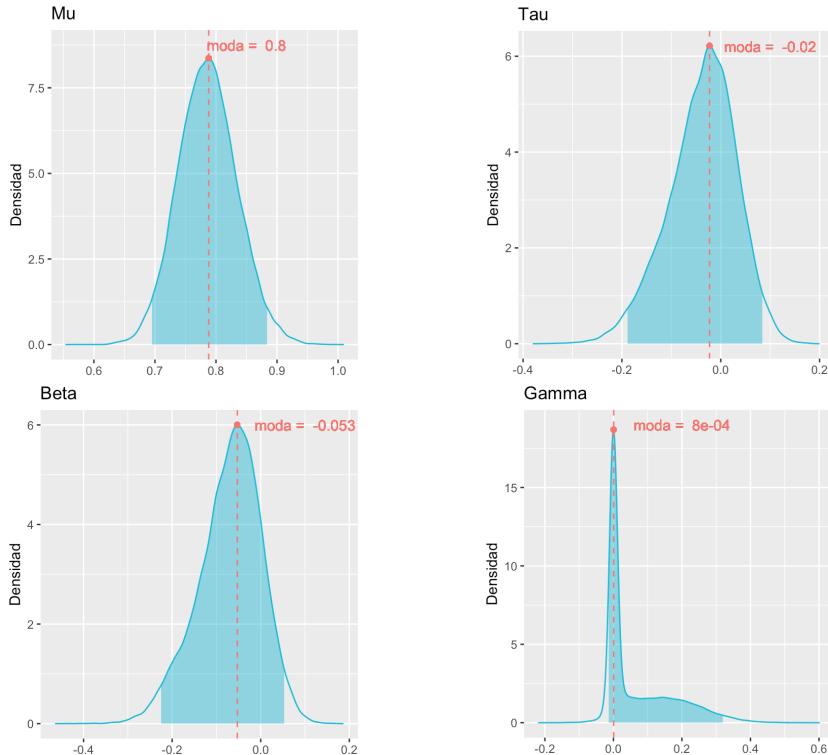


Figura 6.14: Densidades de los parámetros.

Para cada subgrupo, se calcula la media del tratamiento en el cuadro 6.14:

Subgrupo	Estimador	Media
Mujer	Tratamiento	0.747
	Control	0.716
	Diferencia	0.030
Hombre	Tratamiento	0.748
	Control	0.786
	Diferencia	-0.037

Cuadro 6.14: Media de los subgrupos.

En este caso, también se estima una diferencia entre el tratamiento y el control positiva para las mujeres y negativa para los hombres. No obstante, estos valores son más pequeños y cercanos a cero.

Para poder concluir que no hay efectos de subgrupos para este tratamiento, es importante analizar la convergencia de los estimadores ya que un estimador que no converge no es confiable. Por lo tanto, se evalúan los valores de Gelman-Rubin en la figura 6.15:

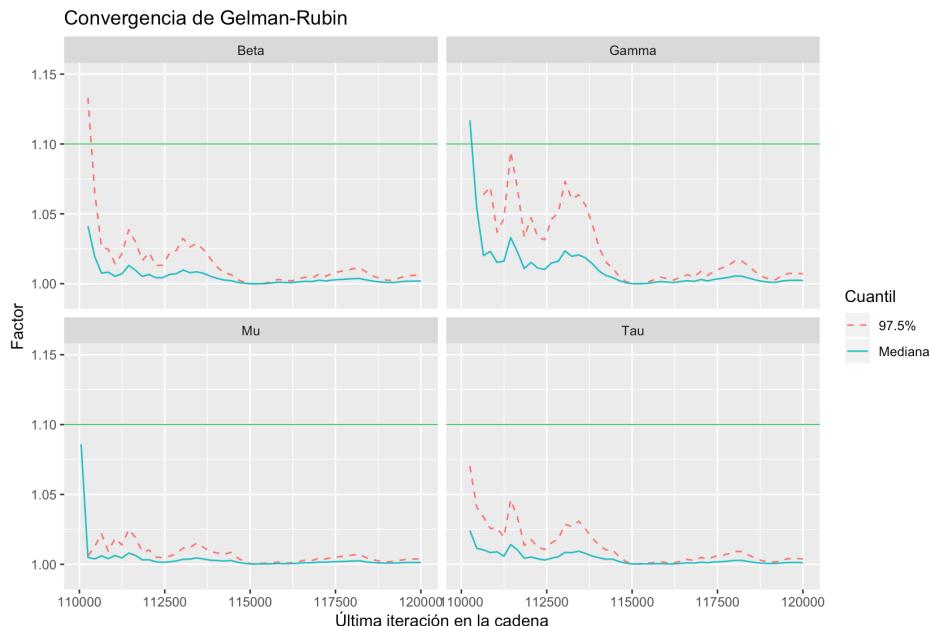


Figura 6.15: Convergencia de Gelman-Rubin.

Es claro que para todos los parámetros hay convergencia. Entonces, se concluye que los estimadores de los coeficientes son confiables. La modificación con *JAGS* no encuentra efectos en subgrupos.

6.2.5. Conclusión de la base de datos de leucemia

Inicialmente, el análisis utilizando el método de Dixon y Simon sugería la existencia de efectos en subgrupos para el tratamiento a pesar de que el tratamiento no tuviera un efecto en general. Tras hacer el análisis utilizando la modificación con *JAGS*, se encontró que no había efectos en subgrupos. A causa de la diferencia de resultados entre métodos, es difícil

concluir que haya efectos en subgrupos. Puede ser que no se hayan encontrado los efectos en subgrupos a causa de que el ensayo no fue diseñado para hacer análisis de subgrupos. Sin embargo, se toman en cuenta las sugerencias de no exagerar los resultados de los subgrupos y se concluye que no hay efectos de subgrupos para este tratamiento.

Capítulo 7

Conclusiones

Se comenzó por investigar varios métodos para hacer análisis de subgrupos con dos enfoques: cuando los subgrupos eran desconocidos y cuando existía una hipótesis de la existencia de subgrupos que no estaba confirmada. Una vez explicados los métodos se aplicaron diferentes bases de datos.

La primera aplicación de los métodos se hizo analizando una base de datos simulados, que se dividía en cuatro subgrupos conocidos con efectos heterogéneos en el tratamiento. Para el primer enfoque, se llevaron a cabo los métodos como si no se conocieran los subgrupos. Los dos métodos encontraron subgrupos, pero sólo uno encontró todos los subgrupos.

En la segunda aplicación, se comenzó por evaluar, a través de datos simulados, el escenario de predefinir los subgrupos correctos eran los predefinidos. Las estimaciones fueron muy cercanas a los parámetros reales y se encontraron los efectos correctos para cada subgrupo. Después, se evaluó el escenario de predefinir los subgrupos de manera incorrecta y

se agregar subgrupos que no tenían efectos heterogéneos al análisis. Algunos estimadores fueron cercanos a los parámetros reales a pesar haber definido los subgrupos incorrectamente. Como esto no sucedió con todos los estimadores, se concluye que, si los subgrupos no están bien definidos en un inicio, se pueden obtener resultados incorrectos.

Posteriormente, se aplicaron los métodos de exploración de subgrupos a un ensayo clínico de pacientes con cáncer de cabeza y cuello. Para esta base de datos, las variaciones del método *Virtual Twins* encontraron subgrupos en los que el tratamiento tenía un efecto mayor. No obstante, los subgrupos sólo se podían describir a través de interacciones entre covariables. Por otro lado, el método de árboles de interacción no encontró ningún subgrupo. Por lo tanto, se concluyó que no hay efectos en subgrupos para el tratamiento analizado.

Finalmente, se aplicó el método de confirmación de Dixon y Simon original y con modificaciones de *JAGS* a un ensayo clínico para probar la efectividad de un tratamiento contra leucemia. Se predefinieron los subgrupos de interés como el subgrupo de pacientes de sexo femenino y de sexo masculino. El método original y el método con modificaciones de *JAGS* encontraron que los subgrupos no tienen diferencias en el tratamiento.

En ninguno de los dos ensayos clínicos se encontraron efectos en subgrupos. Esto responde a dos observaciones. Primero, uno de los objetivos de la estadística es evitar cometer errores tipo I y II. Por lo tanto, un método bien diseñado no encontrará efectos en subgrupos a menos de que haya mucha evidencia de su existencia. Puede ser que esto haya ocurrido con las bases de datos analizadas. Segundo, los ensayos clínicos no fueron diseñados para tener la potencia para encontrar efectos

en subgrupos, sólo para encontrar el efecto del tratamiento en general. Al no tener más información, se concluyó que no hay efectos de subgrupos para los tratamientos.

A pesar de que un ensayo clínico en el que se encuentran efectos en subgrupos es más interesante, es de suma importancia que no se publiquen efectos de subgrupos si no hay evidencia fuerte a favor de que existen y que no se aumente el número de análisis con tal de encontrar subgrupos. Un análisis con resultados de este tipo puede causar fuertes repercusiones en la salud de pacientes. Por ejemplo, el caso visto en el Capítulo 2 en el que no se utilizó durante muchos años la aspirina como tratamiento para prevenir infartos en mujeres a causa de un estudio engañoso (Sun et al. 2014).

Por otro lado, es imperativo que los datos de más ensayos clínicos sean públicos. Esto con el motivo de promover la investigación de tratamientos médicos, obtener la mayor información posible de los ensayos clínicos y promover la innovación médica. De aquí la importancia de apoyar a organizaciones como *The Project Data Sphere* (2018) y dar incentivos para que laboratorios médicos compartan los resultados de sus ensayos clínicos.

A pesar de que la plataforma *The Project Data Sphere* es de suma utilidad, se debe mencionar que muchos de los ensayos clínicos encontrados en la plataforma no eran útiles para hacer análisis. Por ejemplo, se encontraron algunas bases de datos en las que solo se publicó un brazo del ensayo. Es decir, sólo se publicaron los resultados para los pacientes en el grupo control o para los pacientes en el grupo tratamiento, lo cual hace imposible su comparación. Además, algunos de los ensayos publicados evaluaban tratamientos que no tienen efecto en los

pacientes, como es el caso de un ensayo analizado en esta tesis. Esto dificulta el análisis de subgrupos.

Se debe incentivar a las compañías farmacéuticas a compartir sus resultados de una forma activa, por ejemplo, monetaria o legal, si se busca aumentar la participación en la publicación de los resultados de ensayos clínicos. De no ser así, se continuará teniendo un acceso limitado a información necesaria para mejorar los métodos para hacer análisis de subgrupos.

Los análisis de subgrupos no sólo tienen beneficios para la medicina. Una aplicación de los métodos expuestos en esta tesis diferente a ensayos clínicos es, por ejemplo, explicar las diferencias entre salarios de hombres y mujeres a causa del sector en el que trabajan o el número de hijos que tienen (Su et al. 2009). También se puede realizar la evaluación de una política pública que se haya llevado a cabo en diferentes países o ciudades. No obstante, hay una gran motivación para aplicar los métodos a ensayos clínicos por su impacto directo en la salud de las personas y porque la que la evaluación de los tratamientos es un requisito para las farmacéuticas para llevarlos a venta.

El siguiente paso para continuar estudiando el análisis de subgrupos es evaluar el funcionamiento de más métodos. En esta tesis sólo se realizó análisis de subgrupos predefinidos a través del método de Dixon y Simon. Al aplicar Dixon y Simon se observó que, si los subgrupos no estaban perfectamente definidos de acuerdo a los efectos verdaderos, los resultados encontrados eran incorrectos. Este es déficit del método ya que en escenarios reales es altamente improbable que se conozcan con exactitud los efectos de un tratamiento para un subgrupo. Es necesario evaluar otros métodos para conocer el potencial del análisis de subgrupos.

Si todos los métodos se encuentran con el mismo problema que Dixon y Simon, el análisis de subgrupos continuará siendo cuestionado y no se podrá utilizar para tomar decisiones sobre los tratamientos de pacientes.

Asimismo, se deben crear criterios para comparar diferentes métodos. Algunos métodos proponían criterios que evaluaban su desempeño. Por ejemplo, las variables de importancia V_j del método de árboles de interacción. Sin embargo, los criterios encontrados sólo se utilizaban para evaluar un método. Esto hace que la comparación entre ellos no sea objetiva. Se sugiere basarse en la estadística $Q(\hat{A})$ de *Virtual Twins* para analizar varios métodos. Antes de poder utilizar $Q(\hat{A})$, se debe comprobar que funcione de manera adecuada en la mayoría, sino es que todos los métodos de análisis de subgrupos. Una vez que se haya encontrado un criterio de evaluación común, se pueden valorar muchos métodos y llegar a un consenso sobre cuáles métodos usar en ámbitos formales.

Por otro lado, también se deben realizar las mejoras necesarias para que los métodos exploratorios sean aceptados por la comunidad científica como confirmatorios. En los análisis realizados en capítulos anteriores, se encontró que los métodos exploratorios tenían un buen funcionamiento, incluso mejor que el método confirmatorio de Dixon y Simon. Sin embargo, enfrentan varios problemas que se deben solucionar antes de adaptarlos.

Uno de los problemas es la falta de estandarización de los métodos. Se debe buscar llegar a resultados similares independientemente de los parámetros iniciales que se utilicen. En general, los métodos exploratorios utilizan herramientas del Aprendizaje de Máquina que, aunque son muy poderosas, pueden dar resultados engañosos. Cambiar algunos de los parámetros iniciales puede dar resultados diferentes, lo que disminuye la formalidad de los resultados publicados. Independientemente de la

existencia de efectos en subgrupos, un investigador puede ser capaz de encontrarlos si se escogen los parámetros iniciales incorrectos.

A pesar de que ya ha habido esfuerzos mejorar la forma en que se reportan ensayos clínicos, estos no se enfocan lo suficiente en análisis de subgrupos (*The CONSORT Website* 2019). Se suele sugerir evitar hacer análisis de subgrupos exploratorios. Esta sugerencia limita la investigación y posible mejora de los métodos exploratorios. Una alternativa sería llegar a consensos por parte de la comunidad científica de cómo utilizarlos y aprovechar el enorme potencial que tienen para encontrar efectos en subgrupos. Se puede además optar por métodos que estén sustentados por fundamentos de la Estadística y Medicina y no sólo de Aprendizaje de Máquina.

El tener que definir inicialmente los subgrupos para que puedan ser considerados como verdaderos no es una solución viable al análisis de subgrupos. Por un lado, el costo de tener que realizar primero un análisis exploratorio y luego uno confirmatorio en bases de datos independientes es alto y muchas veces prescindible. Asimismo, se limita a los investigadores a estudiar efectos conocidos y no buscar efectos nuevos. Adoptar de manera controlada el análisis de subgrupos puede significar una mayor personalización del cuidado médico y un aumento de bienestar en pacientes con tratamientos.

Por último, un análisis a considerar a futuro es la comparación entre métodos con enfoque Bayesiano y enfoque Frecuentista. Se deben tomar en cuenta las características que describen las bases de datos de ensayos clínicos y cómo cada enfoque las toma en cuenta para hacer análisis. Es un análisis complicado. A la fecha no se ha concluido cuál de los enfoques Estadísticos es mejor en un ámbito general, mucho menos en el análisis de

subgrupos. Hay opiniones divididas entre la comunidad científica acerca de los dos enfoques. En el ámbito de los análisis de subgrupos se debe considerar la posibilidad de que uno de los enfoques funcione mejor. De ser así, priorizar los métodos del enfoque que funciona mejor puede significar en un grado mayor de aceptación de los análisis de subgrupos.

Son claros los beneficios de los análisis de subgrupos, pero no se podrán explotar hasta que los ensayos clínicos se diseñen de tal forma que tengan la potencia para encontrar efectos en subgrupos y los métodos se evalúen de forma profunda y constante. A pesar de que en esta tesis no se encontraron subgrupos en ensayos reales, se comprobó la utilidad y la efectividad de los métodos para encontrar y confirmar la existencia de subgrupos a través de simulaciones. Es un primer paso para demostrar el inmenso potencial del análisis de subgrupos.

Referencias

- Assmann, Susan F., Stuart J. Pocock, Laura E. Enos y Linda E. Kasten (2000). «Subgroup analysis and other (mis)uses of baseline data in clinical trials». En: *The Lancet* 355 (9209), págs. 1064-1069. DOI: [10.1016/S0140-6736\(00\)02039-0](https://doi.org/10.1016/S0140-6736(00)02039-0).
- Bach, Peter B., Laura D. Cramer, Joan L. Warren y Colin B. Begg (1999). «Racial Differences in the Treatment of Early-Stage Lung Cancer». En: *The New England Journal of Medicine* 341.16, págs. 1198-1205. DOI: [10.1056/NEJM199910143411606](https://doi.org/10.1056/NEJM199910143411606).
- Begg, Colin et al. (1996). «Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement». En: *Journal of the American Statistical Association* 276 (8), págs. 637-639. DOI: [10.1001/jama.1996.03540080059030](https://doi.org/10.1001/jama.1996.03540080059030).
- Berger, James O. (1980). *Statistical Decision Theory and Bayesian Analysis*. 2.^a ed. Springer. ISBN: 0-387-96098-8.
- Breiman, Leo (2001). «Random Forests». En: *Machine Learning* 45 (1), págs. 5-32.
- Breiman, Leo, Jerome H. Friedman, Richard A. Olshen y Charles J. Stone (1984). *Classification and regression trees*. Chapman & Hall/CRC Press, págs. 202-265. ISBN: 0-412-04841-8.

REFERENCIAS

- Chipman, Hugh A., Edward I. George y Robert E. McCulloch (2010). «BART: Bayesian additive regression trees». En: *The Annals of Applied Statistics* 4.1, págs. 266-298. DOI: 10.1214/09-AOAS285.
- Ciampi, Antonio, Abdissa Negassa y Zihyi Lou (1995). «Tree-structured prediction for censored survival data and the Cox model». En: *Journal of Clinical Epidemiology* 48.5, págs. 675-689.
- De Finetti, Bruno (1974). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons Ltd. ISBN: 978-1119286370.
- Dixon, Dennis O. y Richard Simon (1991). «Bayesian subset analysis». En: *Biometrics* 47 (3), págs. 871-881. DOI: 10.2307/2532645.
- Dorak, M.Tevfik y Ebru Karpuzoglu (2012). «Gender differences in cancer susceptibility: an inadequately addressed issue». En: *Frontiers in Genetics* 3, págs. 1-11. DOI: 10.3389/fgene.2012.00268.
- ECOG Performance Status* (2018). ECOG-ACRIN Cancer Research Group. URL: <https://ecog-acrin.org/resources/ecog-performance-status> (visitado 08-2018).
- Feinstein, Alvan R. (1998). «The problem of cogent subgroups: A clinicostatistical tragedy». En: *Journal of Clinical Epidemiology* 51.4, págs. 297-299. DOI: 10.1016/s0895-4356(98)00004-3.
- Foster, Jared C., Jeremy M.G. Taylor y Stephen Ruberg (2011). «Subgroup identification from randomized clinical trial data». En: *Statistics in Medicine* 30.24, págs. 2867-2880. DOI: 10.1002/sim.4322.
- Gelman-Rubin convergence diagnostic using multiple chains* (2019). StataCorp LLC. URL: <https://blog.stata.com/2016/05/26/gelman-rubin-convergence-diagnostic-using-multiple-chains/> (visitado 01-2019).

- Hastie, Trevor, Robert Tibshirani y Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer. ISBN: 978-1-4200-5983-0.
- Imai, Kosuke y Marc Ratkovic (2013). «Estimating treatment effect heterogeneity in randomized program evaluation». En: *The Annals of Applied Statistics* 7.1, págs. 443-470. DOI: 10.1214/12-AOAS593.
- Jeffreys, Harold (1946). «An invariant form for the prior probability in estimation problems». En: *Proceedings of the Royal Society of London* 186, págs. 453-461. DOI: 10.1098/rspa.1946.0056.
- Jones, Hayley E., David I. Ohlssen, Beat Neuenschwander, Amy Racine y Michael Branson (2011). «Bayesian models for subgroup analysis in clinical trials». En: *Clinical Trials* 8 (2), págs. 129-143. DOI: 10.1177/1740774510396933.
- Kim, Hae-In, Hyesol Lim y Aree Moon (2018). «Sex Differences in Cancer: Epidemiology, Genetics and Therapy». En: *Biomolecules and Therapeutics* 26.4, págs. 335-342. DOI: 10.4062/biomolther.2018.103.
- Kolitz, Jonathan E. et al. (2010). «P-glycoprotein inhibition using valsphodar (PSC-833) does not improve outcomes for patients younger than age 60 years with newly diagnosed acute myeloid leukemia: Cancer and Leukemia Group B study 19808». En: *Blood Journal* 116.9, págs. 1413-1421. DOI: 10.1182/blood-2009-07-229492.
- Lagakos, Stephen W. (2006). «The challenge of subgroup analyses - reporting without distorting». En: *The New England Journal of Medicine* 354 (16), págs. 1667-1669. DOI: 10.1056/NEJMp068070.
- Laud, Purushottam W., Siva Sivaganesan y Peter Müller (2013). «Subgroup analysis». En: *Bayesian Theory and Applications*. Ed. por Paul Damien, Petros Dellaportas, Nicholas G. Polson y David A. Stephens. Oxford University Press, págs. 576-592. ISBN:

- 978-0-19-969560-7. DOI: 10.1093/acprof:oso/9780199695607.001.0001.
- Li, Yingbo y Merlise A. Clyde (2018). «Mixtures of *g*-Priors in Generalized Linear Models». En: *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2018.1469992.
- Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde y Jim O. Berger (2008). «Mixtures of *g* Priors for Bayesian Variable Selection». En: *Journal of the American Statistical Association* 103 (481), págs. 410-423. DOI: 10.1198/016214507000001337.
- Lindley, D. V. y A. F. M. Smith (1972). «Bayes estimates for the linear model». En: *Journal of the Royal Statistical Society* 34.1, págs. 1-41.
- Lipkovich, Ilya (2018). *Subgroup analysis software*. Biopharmaceutical Network. URL: <http://biopharmnet.com/subgroup-analysis-software/> (visitado 09-2018).
- Lynch, Scott M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer. ISBN: 978-0-387-71264-2.
- Mahmoud, Hosam M. (2009). *Pólya Urn Models*. Chapman & Hall/CRC Press, págs. 45-65. ISBN: 978-1-4200-5983-0.
- Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill, págs. 1-15. ISBN: 0070428077.
- Müller, Peter, Siva Sivaganesan y Purushottam W. Laud (2010). «A Bayes rule for subgroup reporting». En: *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. Ed. por Ming-Hui Chen, Dipak K. Dey, Peter Müller, Dongchu Sun y Keying Ye. Springer, págs. 277-284. ISBN: 978-1-4419-6944-6.
- OpenBUGS* (2018). The BUGS project. URL: <http://openbugs.net> (visitado 12-2018).

- Plummer, Martyn (2013). *JAGS Version 3.4.0 user manual*. [Online]. URL: http://www.stats.ox.ac.uk/~nicholls/MScMCMC15/jags_user_manual.pdf.
- Pocock, Stuart J., Susan E. Assmann, Laura E. Enos y Linda E. Kasten (2002). «Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems». En: *Statistics in Medicine* 21 (19), págs. 2917-2930. DOI: 10.1002/sim.1296.
- Project Data Sphere (2018). CEO Life Sciences Consortium. URL: <https://www.projectdatasphere.org> (visitado 09-2018).
- Proschan, Michael A. y Myron A. Waclawiw (2000). «Practical guidelines for multiplicity adjustment in clinical trials». En: *Controlled Clinical Trials* 21, págs. 527-539.
- Python Core Team (2015). *Python: A dynamic, open source programming language*. URL: <https://www.python.org/>.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rothwell, Peter M. (2005). «Subgroup analysis in randomised controlled trials: importance, indications, and interpretation». En: *The Lancet* 365 (9454), págs. 176-186. DOI: 10.1016/S0140-6736(05)17709-5.
- RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA. URL: <http://www.rstudio.com/>.
- Schulz, Kenneth F., Douglas G. Altman, David Moher y CONSORTGroup (2010). «CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials». En: *Annals of Internal Medicine* 152 (11), págs. 826-732. DOI: 10.7326/0003-4819-152-11-201006010-00232.

- Schwarz, Gideon (1978). «Estimating the dimention of a model». En: *The Annals of Statistics* 6.2, págs. 161-164. DOI: 10.1214/aos/1176344136.
- Simon, Richard (2002). «Bayesian subset analysis: applications to studying treatment-by-gender interactions». En: *Statistics in Medicine* 21 (19), págs. 2909-2916. DOI: 10.1002/sim.1295.
- Sivaganesan, Siva, Purushottam W. Laud y Peter Müller (2011). «A bayesian subgroup analysis with a zero-enriched Polya urn scheme». En: *Statistics in Medicine* 30 (4), págs. 312-323. DOI: 10.1002/sim.4108.
- Song, Yang y George Y.H. Chi (2007). «A method for testing a prespecified subgroup in clinical trials». En: *Statistics in Medicine* 26 (19), págs. 3535-3549. DOI: 10.1002/sim.2825.
- Spiegelhalter, David J., Keith R. Abrams y Jonathan P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons Ltd. ISBN: 0-471-49975-7.
- Su, Xiaogang, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson y Bogong Li (2009). «Subgroup analysis via recursive partitioning». En: *UC Davis Graduate School of Management Research Paper* 12.9, págs. 1-20. DOI: 10.2139/ssrn.1341380.
- Sun, Xin, John P.A. Ioannidis, Thomas Agoritsas, Ana C. Alba y Gordon Guyatt (2014). «How to use a subgroup analysis: users' guide to the medical literature». En: *Journal of the American Statistical Association* 311.4, págs. 405-411. DOI: 10.1001/jama.2013.285063.
- The Canadian Coooperative Study Group (1978). «A randomized trial of aspirin and sulfinpyrazone in threatened stroke». En: *The New England Journal of Medicine* 299.2, págs. 53-59. DOI: 10.1056/NEJM197807132990201.
- The CONSORT Website* (2019). The CONSORT Group. URL: <http://www.consort-statement.org> (visitado 01-2019).

REFERENCIAS

- The Jeffreys Prior (2018). Duke University. URL: <https://www2.stat.duke.edu/courses/Fall11/sta114/jeffreys.pdf> (visitado 11-2018).
- Therneau, Terry M. y Elizabeth J. Atkinson (1997). *An introduction to recursive partitioning using the RPART routines*. Inf. téc.
- Van de Schoot, Rens, David Kaplan, Jaap Denissen, Jens B. Asendorpf, Franz J. Neyer y Marcel A.G. van Aken (2014). «A gentle introduction to Bayesian analysis: Applications to developmental research». En: *Child Development* 85.3, págs. 1-19. DOI: 10.1111/cdev.12169.
- Varadhan, Ravi y Wenliang Yao (2014). *DSBayes: Bayesian subgroup analysis in clinical trials*. R package version 1.1. URL: <https://CRAN.R-project.org/package=DSBayes>.
- Vermorken, Jan B et al. (2013). «Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SPECTRUM): an open-label phase 3 randomised trial». En: *The Lancet* 14 (8), págs. 697-710. DOI: 10.1016/S1470-2045(13)70181-5.
- Vieille, Francois y Jared Foster (2018). *aVirtualTwins: Adaptation of Virtual Twins Method from Jared Foster*. R package version 1.0.1. URL: <https://CRAN.R-project.org/package=aVirtualTwins>.
- Wang, Jeffrey y Ying Huang (2007). «Pharmacogenomics of Sex Difference in Chemotherapeutic Toxicity». En: *Current Drug Discovery Technologies* 4, págs. 56-68.
- Wang, Rui, Stephen W. Lagakos, James H. Ware, David J. Hunter y Jeffrey M. Drazen (2007). «Statistics in medicine - reporting of subgroup analyses in clinical trials». En: *The New England Journal of Medicine* 357.21, págs. 2189-2194. DOI: 10.1056/NEJMsr077003.

REFERENCIAS

- What are monoclonal antibodies?* (2018). Genscript Co. URL: <https://www.genscript.com/how-to-make-monoclonal-antibodies.html> (visitado 11-2018).
- Xu, Yaoyao, Menggang Yu, Ying-Qi Zhao, Quefeng Li, Sijian Wang y Jun Shao1 (2015). «Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects». En: *Biometrics* 71, págs. 645-653. DOI: 10.1111/biom.12322.
- Yusuf, Salim, Rory Collins y Richard Peto (1984). «Why do we need some large, simple randomized trials?» En: *Statistics in Medicine* 3.4, págs. 409-420. DOI: 10.1002/sim.4780030421.
- Yusuf, Salim, Janet Wittes, Jeffrey Probstfield y Herman A. Tyrolier (1991). «Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials». En: *Journal of the American Statistical Association* 266.1, págs. 93-98. DOI: 10.1001/jama.1991.03470010097038.
- Zellner, Arnold (1986). «Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti». En: ed. por Prem K. Goel y Arnold Zellner. Elsevier, págs. 233-243. ISBN: 0-444-87712-6.