UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Cotas generales para el problema del incumplimiento parcial del tratamiento

Lucio José Pantazis

Directora: Mariela Sued

Fecha de Presentación: 27 de Diciembre de 2012

Resumen

Consideremos un ensayo clínico en el que el tratamiento fue asignado de forma completamente aleatoria, pero la población de interés no necesariamente cumple con las asignaciones de tratamiento. En este contexto, ante la imposibilidad de identificar el efecto medio del tratamiento **recibido**, se buscan cotas para el mismo.

Para este problema asumimos un Modelo Causal Funcional que retrata la dinámica de las variables del sistema.

Usando la distribución empírica de las variables observadas (tratamiento asignado, tratamiento recibido y respuesta) y valiéndonos del modelo asumido, formulamos un problema de programación lineal a partir del cual logramos encontrar estimadores para dichas cotas.

Nuestro aporte consiste de encontrar cotas para parámetros causales muy generales (no sólo el efecto medio del tratamiento) en el caso en que la variable respuesta toma finitos valores, generalizando trabajos anteriores.

Agradecimientos

La extensión de esta sección puede resultar excesiva, pero tengo como regla estar agradecido por las cosas buenas que me pasan y la buena gente con la que me he encontrado. Además, siempre sostuve que uno no es nada sin la gente que lo acompaña. Por suerte, en estos años, tengo una infinidad de cosas por las que estar agradecido y la fortuna de haber conocido muchas personas.

A Mariela, mi directora. Podría escribir un sinfín de adjetivos para intentar describir a esta inmensidad de persona y siempre quedarme cortísimo. Por lo que seré conciso: Gracias Mariela, por ser vos, única e irrepetible. Gracias por bancar mis delirios, por el apoyo, la motivación y por la constante sonrisa, entre millones de otras cosas.

Gracias a la inmensa ayuda de Julieta Molina, por transformar un camino pedregoso de subida en una bajada sin asperezas. Sin ella y su arduo trabajo a esta altura recién estaría entendiendo el objetivo de esta tesis.

A mis viejos, por bancarme y facilitar mis estudios, acontecimiento por el que estoy eternamente agradecido, aunque no se los diga lo suficiente. Pero sobre todo, por ser mis modelos de persona, su bondad e infinita tolerancia han echado raices en mi personalidad y puedo decir con sumo orgullo que me parezco a ellos en varios aspectos.

A mis hermanos, por bancarse el mal humor previo a numerosos exámenes. A mis abuelas, por ser una fuente inagotable de comida, especialmente a «La Noni» que a lo largo de la carrera me ha simplificado el viaje tomándome como huésped. A mis primas, por ser una fuente inagotable de abrazos. A mis tíos. Gracias por cada domingo compartido. A mi abuelo, por ser un ejemplo a la determinación y el esfuerzo, cualidades que necesité mucho cuando nos dejó...

A los chicos de «Asaduli»: Santi, Pablito, Fer, Pachu, Manu, Barret, Guido y Mati. Gracias por tantas risas, canciones, charlas, consejos y las necesarias salidas. Gracias por siempre cultivar la curiosidad, la inquietud por saber y conocer. La «escasa información del entorno» sería muchísimo menor sin ustedes. Sobre todo, gracias por estar, a pesar de las diferencias.

A las chicas del Acosta: Belu, Cele, Lau, Marian, Lu, Andre y Euge. Me han bancado montones y el inmenso cariño que demuestran día a día levántadome de muchos bajones.

A Agustín, por siempre ser compañero de distensiones, por ayudar a desconectarme cuando era necesario y sobre todo por las estufas (Flema no existe). Además, con Ale

(Cuña), Roberto (CadPo) y Adrián (A.C.T.) hacen revivir al adolescente/niño que llevo dentro, me han hecho divertir montonazos. También a Noe, Gise, Paulita, Maru (la Rubia), Marie (la Morocha), Lucas Corach, Lu Salva, Iru, Hernán, Santi, Andrés, Alejandra, Mariela (Okus Pokus) y Estefi.

A grandes amigos como Manuel Benjamín (yo no me olvido...), Pablo Escobar, Herman, Mateo, Fede Martínez, Maxi, Marianito y Julián que si bien la carrera ha hecho que nuestros rumbos difieran, han sido excelentes compañeros de cursada. Además, a Vero, Paulette, Florence, Yani, Ema (yo no me olvido...), Pablo Colombo, Romi, Lucha, Carla, Ceci, Xime (yo no me olvido...), Mer, Rosario (yo no me olvido...), Belén, Flor Slezak, Carito, Rocío, Marie y Carla Baroncini (yo no me olvido...).

A Laura Cacheiro, cuya excelente tesis me sirvió de guía para adentrarme en el mundo de la inferencia causal que me era tan ajeno en un principio.

Hay tres personas que he tenido la fortuna de tenerlos en dos circunstancias diferentes: como docentes y como amigos. Al Dr. Lucas Bali, Ignacio Ojea (Expreso Liniers) y Santiago Saglietti. Estas tres personas, sus enormes capacidades y su innato interés por pensar problemas nuevos, han aportado muchísimo a mi formación, me han ayudado montones cuando no tenían por qué hacerlo, además de ser modelos a seguir como miembros de la academia. Me han asistido para pasar el mayor obstáculo de la tesis, por eso, mil gracias.

A la gente del Instituto de Cálculo: Paula, Marina, Stella, Agustín y Alejandra. Han sido un sostén para el último tramo de la carrera, pero sobre todo, les agradezco por hacerme sentir parte.

A algunos docentes que han logrado un gran interés de mi parte en sus clases, dejándome infinitas enseñanzas para mis estudios y mis capacidades como docente: Ana Bianco, Pablo Groisman, Daniela Rodríguez, María Eugenia Szretter, Sandra Martínes, Daniel Galicer, Victoria Paternostro, Román Villafañe, Matías Graña, Daniel Perrucci, Carlos Gustavo López Pombo, Mariano Moscato, Susana Puddu y Juan Pablo Pinasco.

A algunos compañeros de trabajo, con los que me he llevado de maravillas: Federico Quallbrun, Sebastián Freyre, Julián Haddad, Facundo Poggi y Marcela Fabio, además de los ya mencionados en otros párrafos.

A los alumnos. «El día que dejamos de enseñar, es el día que dejamos de aprender».

A Graciela, por proveerme de líquido elemento (es decir, mate).

A los chicos de «Dale Chechu», por la gloria que tuvo, tiene y tendrá este equipo.

Felicitaciones, todos ustedes han contribuido para hacer de mí un licenciado.

Índice general

Introducción

En el ámbito de la inferencia causal, muchas veces se busca determinar la efectividad de un tratamiento. Para lograr este objetivo, hay que determinar una magnitud que refleje un «nivel de efectividad». Cuando tratamos con ensayos clínicos, usualmente se considera el «Efecto medio del tratamiento» o «ATE».

En esta tesis abordamos el problema en el que se asigna de manera completamente aleatoria tratamiento a un grupo de pacientes pero no todos los pacientes cumplen con dicha asignación. Ante este inconveniente, encontramos cotas óptimas para el ATE, teniendo en cuenta que el cumplimiento de los pacientes no fue perfecto.

Las cotas obtenidas generalizan los trabajos de Pearl [1], que enfrenta este problema en un caso en el que la variable respuesta es binaria, nosotros extenderemos estos resultados al caso en el que la variable respuesta toma finitos valores.

Hemos elaborado esta tesis procurando que resulte autocontenida. Para ello, hemos consultado variada bibliografía. El Capítulo 1 fue motivado por los trabajos de Miguel Hernán y J.M. Robins [7] y [8]. El libro de Pearl [13] junto con las notas preparadas por la Profesora Andrea Rotnitzky para el Congreso Monteiro [16] han sido fundamentales para la elaboración de los capítulos 2 y 3. Para los problemas relacionados con la programación hemos utilizado el libro [4], disponible en internet. Por último, queremos mencionar la Tesis de Licenciatura de Laura Cacheiro [2], que fue una importante fuente de consulta.

La tesis se organiza de la siguiente manera:

En el Capítulo 1, introducimos varias nociones de la inferencia causal, fundamentalmente definiendo las llamadas «variables contrafactuales» y remarcando su diferencia con las «variables factuales». Además, hacemos mención del «parámetro causal» y como buscar cotas para el mismo, que será de suma importancia a la hora de determinar efectos causales. En este capítulo abundan nuevos conceptos, que no demandan de mucha matemática. Hemos optado por un estilo un tanto informal esperando tornar su lectura lo más amena posible.

En el Capítulo 2, detallamos una relación entre grafos y distribuciones de probabilidad. Esta relación nos permitirá asegurar independencias o independencias condicionales entre las variables de un modelo.

En el Capítulo 3, damos un nuevo enfoque a los modelos causales, presentando los

llamados Modelos Causales Funcionales. En el Capítulo 4, atacamos el Problema del incumplimiento parcial del tratamiento, presentando antes un ejemplo de mayor simpleza para acostumbrar al lector a las construcciones que serán hechas. Luego, detallaremos el problema ya mencionado en el caso resuelto por Pearl, para luego generalizar estos trabajos.

Algunas cuestiones de notación:

- Nos será de mucha utilidad ver a las funciones de probabilidad puntual como vectores. Para esto, dado $n \in \mathbb{N}$, denotamos por $\mathcal{S}_n$ al siguiente conjunto:

$$\mathcal{S}_n = \{(x_1, x_2, ..., x_n) \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \wedge x_i \geq 0 \ \forall \ 1 \leq i \leq n\}$$

  Este conjunto es llamado un simplex en $\mathbb{R}^n$.

  Teniendo esto en cuenta, la función de probabilidad puntual de todo vector aleatorio que toma n valores posibles, se corresponde con un vector de $\mathbb{R}^n$, con la particularidad de que sus coordenadas son no negativas y además la suma de las mismas da como resultado 1. Es decir, la función de probabilidad puntual de tal vector puede ser considerada en $\mathcal{S}_n$. El lector sabrá diferenciar cada caso.

- En aras de simplificar algunas construcciones, a veces no haremos distinción entre un vector aleatorio y un conjunto de variables aleatorias. Por lo que un vector aleatorio podrá aparecer de las siguientes maneras: $\widetilde{X} = (X_1, X_2, ..., X_n)$ o $\widetilde{X} = \{X_1, X_2, ..., X_n\}$.

- Usaremos la notación $X \coprod Y$ para simbolizar que X e Y son independientes. Mientras que $X \coprod Y|Z$ significará que X e Y son independientes **condicionales a** Z.

- Por último, como trabajaremos con variables discretas, en varios pasos asumiremos que podemos condicionar a ciertos eventos, suponiendo que tienen probabilidad positiva para no incurrir en casos patológicos.

Capítulo 1

Breve introducción a la Causalidad

1.1. Efectos causales

1.1.1. Efectos causales individuales

Quien no se haya preguntado alguna vez «¿Qué hubiera pasado si las circunstancias hubieran sido distintas?» que tire la primera piedra... Para introducir la noción de efecto causal individual, pensemos en el siguiente ejemplo. Un ser querido nuestro llamado Juan ha fallecido pasado un mes de que le hayan realizado un transplante de corazón. Ante la impotencia y falta de control sobre lo sucedido, uno empieza a buscar explicaciones y se pregunta si esa persona hubiera podido continuar su vida en el caso de no haber sido transplantado. La imposibilidad de volver el tiempo atrás hace que esta sea una pregunta que no se puede responder (haciendo crecer aún más la impotencia). O sea, no podemos observar ambos escenarios, lo único que podemos tomar como dato es lo que sucedió.

¿Qué pasaría si pudiéramos ver ambas situaciones? Uno podría ir a una tarotista (o cualquier persona que posea una bola de cristal) dispuesto a contestar esa pregunta. Se pueden dar dos situaciones:

- Por un lado, la bola de cristal puede decirnos que efectivamente Juan seguiría vivo de no ser por el transplante. En un arrebato de ira, uno va a amenazar de muerte al cirujano, de incendios premeditados al hospital y demás actos de vandalismo a todo transeúnte que se imponga en el camino. Esta ira se debe a que cualquier persona entiende que el transplante ocasionó la muerte de nuestro querido Juan. En otros términos, concluimos que **el transplante tuvo un efecto causal** sobre el fallecimiento de Juan.

- Por otro lado, la bola puede decirnos que Juan hubiera fallecido igual en caso de no haber sido transplantado. Acto seguido compramos resignados un nuevo paquete de pañuelos descartables porque nos vuelve a abrumar la falta de explicaciones.

> Nuevamente, esa sensación de resignación se debe a que ya nadie consideraría «echarle la culpa al transplante». Dicho de otra forma, **no se considera que el transplante haya tenido un efecto causal** sobre la muerte de Juan.

O sea, **si pudiéramos** observar el estado de Juan en ambas situaciones (es decir, transplantado y sin transplante), podríamos determinar si (a nivel individual) el transplante tuvo un **efecto causal** sobre el fallecimiento de Juan.

Efectos causales individuales no pueden ser abordados, pues resulta un obstáculo insuperable el hecho de que hayan datos faltantes (ver el trabajo de Paul Holland [9]). Sin embargo, hay muchas personas con la misma patología que Juan. Algunos han sido transplantados y, como Juan, han fallecido, mientras que otros transplantados han tenido mejor suerte. La misma ambivalencia de situaciones (morir o no morir, esa es la cuestión) se puede presentar en gente que no fue transplantada. Considerando como población de interés al grupo de pacientes que sufren la misma patología que Juan, la propuesta estadística procura dar respuesta al siguiente interrogante: ¿Cómo podemos definir efectos causales sobre nuestra población de interés?

1.1.2. Efectos causales poblacionales

Los directivos del hospital, cansados de las amenazas hacia sus integrantes, empiezan a controlar con mayor énfasis los transplantes de corazón, para personas con determinada enfermedad o patología (la población de interés), con el objetivo de determinar si efectivamente hay algún problema relacionado al transplante que ocasione el fallecimiento de los pacientes. En el caso de que esto sea cierto, los futuros pacientes que padezcan esa enfermedad no serán sometidos al transplante. En el caso contrario, el transplante seguirá siendo considerada la mejor opción ante la patología estudiada.

Con los datos recogidos, se necesita una propuesta estadística para sacar conclusiones al respecto.

La propuesta de la inferencia causal es comparar hipotéticos escenarios donde:

1. Todos los pacientes son transplantados.

2. Ningún paciente es transplantado.

Esta conceptualización permite formular la siguiente pregunta: ¿Es mejor para el hospital tomar como política realizar el transplante a pacientes con la mencionada patología o es preferible no transplantar a ninguno?

Esta pregunta lleva consigo muchos otros interrogantes: ¿Cómo comparamos estas distintas circunstancias, pudiendo observar sólo una de ellas en cada individuo? ¿Cómo inferimos desde lo observado en cada individuo a efectos causales poblacionales? ¿Habrá otros factores (edad, peso, debilidades inmunológicas, etc.) que puedan aportar mayor información y no estén siendo consideradas? ¿Cómo medir que una política determinada sea «mejor» que otra?

Estas son algunas de las dificultades matemáticas y/o filosóficas que se presentan ante cualquier modelo causal. Siempre surge una numerosa cantidad de interrogantes que complican el modelado a la hora de determinar efectos causales, ya que hay muchos supuestos y variables que uno puede tener en cuenta o no. Como los obstáculos que aparecen son varios, el modelo que uno plantea para estos problemas es **crucial**.

Hemos presentado en este ejemplo los efectos de un transplante de corazón sobre la tasa de mortalidad de una cierta población, pero más generalmente, podríamos estar evaluando los efectos de dos posibles «acciones» (transplantar - no transplantar) sobre una «respuesta física» de una cierta población. De ahora en adelante, salvo aviso previo, se usará *tratamiento* (operar) y *control o no tratado* (no operar) para denominar las dos acciones posibles cuyos efectos queremos comparar.

Cuando encaramos desde un punto de vista estadístico preguntas sobre posibles efectos de dos acciones en cierta variable respuesta, debemos encontrar parámetros que permitan cuantificar los efectos causales, para responder a las preguntas planteadas. La expresión que logre el cometido será llamado «parámetro causal».

En el ejemplo, un posible parámetro causal sería la diferencia entre las siguientes dos magnitudes:

1. La proporción de pacientes que se mantendrían con vida en el caso hipotético de que todos fueran transplantados

2. La proporción de pacientes que se mantendrían con vida en el caso hipotético de que ninguno fuera transplantado.

Recordemos que estas magnitudes dependen de variables que no son completamente observadas, pero comparar ambas proporciones sería una buena forma de cuantificar un efecto causal del tratamiento sobre la población.

1.2. Variables factuales y contrafactuales

Ahora sí, vamos a la matemática. Supongamos que tenemos una población A (que supondremos finita) y P una probabilidad **uniforme** definida sobre A. En esta población queremos determinar los efectos de dos posibles acciones (tratamiento y control) sobre cierta variable respuesta.

Los individuos de la población serán denotados como a y a cada $a \in A$ se le asignará una de las posibles acciones: tratamiento o control. Además, en cada individuo observamos cierta variable respuesta. En el presente trabajo resulta conveniente pensar en las variables respuesta siendo el resultado de cierto experimento físico realizado en cada individuo de la población, bajo ciertas circunstancias. Podemos pensar en un análisis clínico o incluso, como ocurre en el ejemplo que estamos abordando, determinar si un individuo sobrevive o no. Las condiciones en la que se realice el experimento darán

origen a distintas variables, denotadas todas ellas mediante perturbaciones de una letra en común (generalmente Y, Y_{x_0}, Y_{x_1}) .

Consideremos una variable $X : A \to \{0, 1\}$, donde $X(a)$ representa la acción que fue asignada al individuo a, mientras que la variable $Y : A \to \{0, 1\}$ denotará la variable respuesta de interés (momentáneamente la supondremos binaria). Para $a \in A$, tendremos:

- $X(a) = x_1$ en el caso de que la persona a fue tratada.
- $X(a) = x_0$ en el caso de que la persona a no fue tratada (control).
- $Y(a) = 1$ si el individuo a responde «positivamente».
- $Y(a) = 0$ si el individuo a responde «negativamente».

En el ejemplo de los directivos del hospital, ellos considerarán como «tratamiento» el transplante de corazón y la «respuesta negativa» será (claramente) que el individuo pase a mejor vida pasado un mes de la decisión entre realizar el transplante o no.

Ejemplo 1.2.1 *Supongamos que la población A consta de 9 personas, y luego de hacer ciertas observaciones tenemos los siguientes valores para X e Y, dados en el Cuadro 1.1:*

a	$X(a)$	$Y(a)$
1	x_0	0
2	x_0	0
3	x_0	1
4	x_1	0
5	x_1	0
6	x_1	1
7	x_1	1
8	x_1	1
9	x_1	1

Cuadro 1.1: Valores de acción y la respuesta para los individuos de la población A

Introduciremos ahora a las variables contrafactuales (también llamadas «respuestas potenciales»), un concepto propio de la inferencia causal. En teoría, antes de determinar qué acción recibirá, cada individuo a podría ser expuesto tanto al tratamiento como al control, pero una vez realizada la asignación, sólo podemos ver la respuesta pertinente (dada por $Y(a)$), en presencia de una asignación fija dada por $X(a)$.

Supongamos que tenemos una variable Y_{x_1} definida sobre A que responde de antemano cual va a ser la respuesta de cada individuo si toda la población fuera tratada, y otra variable Y_{x_0}, también definida sobre A, que detalla la respuesta de cada individuo si toda la población estuviera bajo la acción control. Como mencionamos al comienzo de la sección, estas variables representan la variable «respuesta» en nuevas condiciones. Si el experimento fuera una película, cada variable contrafactual proyectaría un final alternativo a la película que hemos filmado.

Cuando la respuesta es binaria, indicando respuestas favorables y desfavorables, para cada $a \in A$ tenemos:

- $Y_{x_1}(a) = 1$ si la persona a responde bien cuando toda la población es tratada.
- $Y_{x_1}(a) = 0$ si la persona a responde de forma negativa al tratamiento cuando toda la población es tratada.
- $Y_{x_0}(a) = 1$ si la persona a responde bien cuando toda se aplica el control en toda la población.
- $Y_{x_0}(a) = 0$ si la persona a responde de forma negativa cuando toda se aplica el control en toda la población.
- Además de estas variables contrafactuales, seguimos bajo la presencia de las siguientes variables **factuales**:
- $\left\{ \begin{array}{ll} X(a) = x_0 & \text{si a la persona } a \text{ se le asigna control.} \\ X(a) = x_1 & \text{si a la persona } a \text{ se le asigna tratamiento.} \end{array} \right\}$
- $\left\{ \begin{array}{ll} Y(a) = 0 & \text{si se } \textbf{observa} \text{ una respuesta positiva en la persona } a. \\ Y(a) = 1 & \text{si se } \textbf{observa} \text{ una respuesta negativa en la persona } a. \end{array} \right\}$

Hemos dicho que la variable Y no tiene porqué ser binaria, por lo que extenderemos la definición de variables contrafactuales sin recurrir a valores numéricos, simplemente nos referiremos a la «acción» y la «respuesta». Así, tenemos una definición más amplia de variables contrafactuales.

Definición 1.2.2 *Sean las variables X, Y, que denotan la acción asignada a cada individuo y la respuesta en él observada respectivamente, éstas se denominan variables factuales.*

En cambio, denotamos por Y_{x_i} a la variable respuesta que observaríamos en el caso en que toda la población sea expuesta a la acción x_i, para $i = 0, 1$. Y_{x_1} y Y_{x_0} se llaman variables contrafactuales (o respuestas potenciales), dado que representan a la variable respuesta en escenarios ***hipotéticos****.*

Variables factuales y contrafactuales se relacionan mediante el **supuesto** de consistencia, que será asumido de ahora en adelante a lo largo del presente trabajo. El mismo establece que existe una igualdad entre $Y(a)$ y $Y_{x_0}(a)$, para todo a que cumple $X(a) = x_0$, del mismo modo $Y(a) = Y_{x_1}(a)$, $\forall a : X(a) = x_1$. Esta hipótesis considera que la respuesta observada en una persona tratada, coincide con la que observaríamos en **la misma** persona si todas las personas fueran tratadas, y lo mismo ocurre con el control. Matemáticamente, la consistencia se puede expresar en la siguiente ecuación:

$$Y(a) = Y_{x_1}(a) \cdot I_{\{x_1\}}(X(a)) + Y_{x_0}(a) \cdot I_{\{x_0\}}(X(a)) \tag{1.1}$$

Es decir, para los individuos tratados, la respuesta $Y(a)$ coincide con la respectiva variable contrafactual $Y_{x_1}(a)$, y para los individuos bajo control, $Y(a) = Y_{x_0}(a)$.

Nos valdremos ahora del Ejemplo 1.2.1 para retratar **una posible configuración** para las variables contrafactuales.

Ejemplo 1.2.3 *Para la población A del Ejemplo 1.2.1, los valores para X e Y coinciden con los que ya fueron observados, pero incluimos **potenciales** respuestas, es decir, valores que pueden tomar Y_{x_0} e Y_{x_1} (aunque no es necesario que tomen **estos** valores):*

a	$X(a)$	$Y(a)$	$Y_{x_0}(a)$	$Y_{x_1}(a)$
1	x_0	0	0	1
2	x_0	0	0	0
3	x_0	1	1	1
4	x_1	0	0	0
5	x_1	0	1	0
6	x_1	1	0	1
7	x_1	1	0	1
8	x_1	1	1	1
9	x_1	1	1	1

Cuadro 1.2: Valores del tratamiento y la respuesta para los individuos de A y las «potenciales respuestas»

Por ejemplo, el primer individuo (o sea, $a = 1$) no ha sido tratado ($X(1) = x_0$) y hemos observado una respuesta negativa ($Y(1) = 0$). Pero la variable contrafactual Y_{x_1} nos dice que si hubiera sido tratado el primer individuo, se hubiera observado una respuesta positiva ($Y_{x_1}(1) = 1$), por lo que el tratamiento hubiera sido beneficioso para el primer individuo. Para $a = 5$ hubiera pasado lo contrario, el tratamiento fue perjudicial pues $X(5) = x_1$ (fue tratado), observamos $Y(5) = 0$ (respuesta negativa), pero $Y_{x_0}(5) = 1$ (en el caso de ser sometido a control, tendría respuesta positiva). Con esto último, hemos visto que si bien la hipótesis de consistencia permite igualar $Y(a)$ con $Y_{x_1}(a)$ si $X(a) = x_1$, no tiene por qué darse la igualdad $Y(a) = Y_{x_1}(a)$ si $X(a) = x_0$.

Notar que en la construcción de las variables contrafactuales, ya no empleamos tiempos verbales en algún pretérito o un condicional. Ponemos, por ejemplo, «la persona **responde** bien», esto es porque las variables contrafactuales «llevan consigo la verdad», sin lugar a cuestionamientos (como cuando en estadística paramétrica a veces se toma el «parámetro real», ya sea dado por Dios, la naturaleza, Maradona, Einstein o algún tipo de fuerza mayor). Estas variables vendrían a representar la «bola de cristal» mencionada en la primer sección del capítulo, dando respuestas en escenarios no necesariamente coincidentes con la realidad.

Su denominación de contrafactuales, viene de la imposibilidad de observar estas variables en su totalidad, pues en cada individuo podemos observar sólo una de las posibles situaciones. Pero bajo ciertas condiciones, podemos sacar conclusiones sobre la distribución de estas variables o ciertos **funcionales** asociados a la distribución de las variables contrafactuales.

1.3. Clasificación de variables

Cabe hacer una distinción a esta altura. En algún momento de la primer sección nos preguntamos si estábamos teniendo en cuenta todos los factores que puedan influir en la respuesta del paciente. Por ejemplo, la edad de una persona claramente puede influir en la respuesta a un tratamiento, para considerarla como un posible factor, sólo basta preguntarle a los pacientes la fecha de nacimiento. Sin embargo, como no se lo consideró, no se observó la edad (como variable) de cada paciente. Pero **podrían haberlo hecho**, aportando así mayor información, permitiendo quizás mejores conclusiones.

Si de «aportar información» hablamos, las variables contrafactuales, si pudiéramos tenerlas como dato, nos darían muchísima información sobre efectos del tratamiento, pero como dijimos, las variables contrafactuales le hacen honor al nombre porque **no pueden** ser observadas en su totalidad.

Observación 1.3.1 *Decimos que las variables contrafactuales no son observadas «en su totalidad» porque bajo la hipótesis de consistencia, las observamos parcialmente. Es decir observamos:* $Y_{x_0}(a) = Y(a), \forall a : X(a) = x_0$ *y* $Y_{x_1}(a) = Y(a), \forall a : X(a) = x_1$*. Lo que no es observado* ***ni puede serlo*** *es* $Y_{x_1}(a)$ *si* $X(a) = x_0$ *ni* $Y_{x_0}(a)$ *si* $X(a) = x_1$*.*

Esto lleva a que consideremos la siguiente clasificación:

- **Variables observadas:** Son aquellas variables que los investigadores se deciden a tomar como datos, porque piensan que pueden ser suficientes para sacar conclusiones.
- **Variables factuales:** Que una variable no haya sido observada, no implica que no pueda pasar serlo. Las variables que podamos incluir en nuestro modelo como variables explicativas serán llamadas variables factuales.

- **Variables contrafactuales:** Son variables que no podemos observar en su totalidad, ya que observarla totalmente no está a nuestro alcance o genera un absurdo con la continuidad del espacio-tiempo.

Observación 1.3.2 *Notar que las variables observadas son variables factuales, pero no necesariamente se da la recíproca.*

Hechas estas salvedades, nos disponemos a continuar.

1.4. Efecto medio del tratamiento - Parámetros causales

¿Cuál será la forma de determinar si un tratamiento fue efectivo o no? Necesitaremos una magnitud que nos dé información sobre la influencia del tratamiento sobre la respuesta de interés.

En el ejemplo anterior, recordemos que los directivos del hospital querían ver qué política aplicar: si transplantar a todos los pacientes con esa enfermedad o no transplantar a ninguno, eligiendo el que disminuya la proporción de fallecidos en cada caso.

Consideremos los siguientes conjuntos:

- $A_1 = \{a \in A : \text{ el individuo } a \text{ vive pasado un mes del transplante}\}$
- $A_0 = \{a \in A : \text{ el individuo } a \text{ no ha sido transplantado y vive pasado un mes}\}$

La definición de las variables contrafactuales nos permitirá expresar matemáticamente estas proporciones. Tomemos los siguientes subconjuntos de A:

$$A_1 = \{a \in A : Y_{x_1}(a) = 1\} \tag{1.2}$$
$$A_0 = \{a \in A : Y_{x_0}(a) = 1\} \tag{1.3}$$

La comparación entre dichas proporciones resulta:

$$\frac{\#A_1}{\#A} - \frac{\#A_0}{\#A}$$

Como la probabilidad en A es uniforme y usando las ecuaciones (1.2) y (1.3), se obtiene la siguiente expresión:

$$P(Y_{x_1} = 1) - P(Y_{x_0} = 1) \tag{1.4}$$

Si esta expresión es positiva, al hospital le convendrá tomar la política de transplantar a los que padezcan la enfermedad cardíaca, pues se obtendrá mayor proporción

de supervivientes. En caso de que sea negativo, para disminuir actos de violencia y juicios de mala praxis, convendrá que no se realicen transplantes a los que sufren dicha enfermedad.

Podemos concluir que esta expresión nos da información sobre el efecto causal del tratamiento sobre la respuesta, sin embargo, no es la única forma de lograr este objetivo.

El «parámetro causal» o «parámetro causal de interés» será la expresión fijada por el investigador para medir efectos causales de dos posibles acciones sobre cierta variable respuesta, que en la gran mayoría de los casos tendrá como protagonista a las variables contrafactuales. En general, el parámetro causal es un funcional que se aplica a la distribución de éstas últimas.

A modo de ejemplo, hemos planteado a Y tomando valores en $\{0, 1\}$, pero en principio puede tomar cualquier cantidad de valores (incluso no tienen porqué ser discretas, aunque para introducir estas cuestiones, pensaremos a las variables del modelo como variables aleatorias discretas de rango finito). En este caso, donde Y es una variable dicotómica, tenemos que la ecuación (1.4) se transforma en:

$$ATE = E[Y_{x_1}] - E[Y_{x_0}] \tag{1.5}$$

Esta magnitud es llamada ATE: «Average treatment effect» («Efecto medio del tratamiento» en español) y es un parámetro causal muy utilizado en las diversas áreas de aplicación de la inferencia causal, pudiendo ser definido para **cualquier** tipo de variable respuesta. A lo largo del presente trabajo consideraremos el efecto medio del tratamiento $E[Y_{x_1} - Y_{x_0}]$ como parámetro causal de interés.

1.5. Identificabilidad

Por lo general, el parámetro causal de interés depende de la distribución de variables contrafactuales. Por ejemplo, acabamos de definir el efecto medio del tratamiento, dado por $ATE = E[Y_{x_1} - Y_{x_0}]$. Sin embargo, las variables contrafactuales no son observadas en todos los individuos de la población y el investigador dispone de una muestra de ciertas variables factuales (variables observadas). Resta determinar si esta información resulta suficiente para determinar (y luego estimar) el parámetro causal de interés. Para ello, necesitaremos estipular ciertos supuestos en el modelo que, por lo general, involucran variables factuales y contrafactuales. Tenemos entonces la siguiente definición:

Definición 1.5.1 *Si el parámetro causal queda determinado (bajo el modelo asumido) a partir de la distribución de las variables observadas (llamémosla p), decimos que el parámetro causal está* ***identificado****. Esto es equivalente a pedir que el parámetro causal sea un funcional de p. Es decir, tenemos identificabilidad cuando existe T_{caus} tal que el parámetro causal (llamémoslo PC) se pueda escribir como $PC = T_{caus}(p)$.*

Observación 1.5.2 *Notar que la definición depende fuertemente de cuales sean las variables observadas y el modelo asumido. No tener identificabilidad sería tener dos parámetros causales distintos «compatibles» con la misma distribución de variables observadas.*

Como mencionamos, para que valga la condición de identificabilidad, el modelo requerido involucra propiedades conjuntas entre variables observadas y contrafactuales. La imposibilidad de observar estas últimas impide verificar las condiciones de identificabilidad asumidas mediante procedimientos estadísticos. Es decir, dada una muestra de datos, no existe un «test» para garantizar identificabilidad. Por eso las hipótesis sobre el modelo, que son de tanta importancia, deben ser asumidas en función del conocimiento específico del que se disponga sobre el problema que se trata.

Estudiar si el parámetro causal está identificado es la primer inquietud que surge a la hora de trabajar con causalidad, por lo que se plantean varias propuestas para abordar esta cuestión:

1. Una forma de encarar esto es determinar qué supuestos sobre las variables (observadas y contrafactuales) son necesarios para poder identificar el parámetro causal de interés. Queda en manos del experto determinar si los supuestos necesarios para identificar resultan apropiados para el problema a resolver.

2. Cuando las variables observadas resulten insuficientes para poder identificar el parámetro causal de interés, podemos procurar medir nuevas variables (factuales) que permitan asumir las suposiciones necesarias para garantizar identificabilidad.

3. Por último, ante la falta de identificabilidad, se pueden buscar cotas para el parámetro causal que permitan esbozar una idea de posibles valores para el mismo.

Veremos como funciona cada una de estas propuestas en el ejemplo que hemos estado considerando.

Siguiendo el enfoque propuesto en el punto 1, un posible modelo consiste en suponer que el mecanismo de asignación del tratamiento es **completamente aleatorio**. Entendemos por mecanismo completamente aleatorio a cualquiera que permita establecer que tratados y no tratados (controles) conforman grupos intercambiables, con mismas características estadísticas a los efectos del problema en cuestión. Por ejemplo, si se extrajera al azar una cantidad **fija** de nombres de una bolsa y a los pacientes cuyo nombre fue extraído se les asigna tratamiento, estaríamos en presencia de un mecanismo de asignación de tratamiento completamente aleatorio. En tal caso, la asignación del tratamiento resulta independiente de toda variable relacionada con el experimento. En particular, con las variables contrafactuales, por lo que tenemos:

$$Y_{x_i} \coprod X, \forall i \in \{0, 1\}. \tag{1.6}$$

Con esta suposición (junto con $P(X = x_1) > 0, P(X = x_0) > 0$) se ve simplemente lo siguiente:

$$E[Y|X = x_1] \quad = \quad \underbrace{E[Y_{x_1}|X = x_1]}_{consistencia} \quad = \quad \underbrace{E[Y_{x_1}]}_{\text{por (1.6)}} \tag{1.7}$$

Entonces, $E[Y_{x_1}]$ se puede obtener a partir de la distribución de las variables observadas (X, Y). De manera similar, se puede obtener $E[Y_{x_0}]$ como $E[Y|X = x_0]$. Como ambas esperanzas se obtienen a partir de la distribución de los datos observados, nuestro parámetro causal está identificado a partir de la distribución de (X, Y), bajo el supuesto de aleatorización completa.

Observación 1.5.3 *La igualdad dada en (1.7) casi siempre lleva a una gran confusión para los que se introducen en el mundo de la causalidad. Pues se tiende a confundir la distribución de la variable contrafactual Y_{x_i} con la distribución de $Y|X = x_i$, para $i \in \{0, 1\}$. La diferencia entre ambas está explicada en el trabajo de Robins [7]. Cuando la asignación del tratamiento es completamente al azar, las distribuciones coinciden.*

Puede ocurrir que la aleatorización completa resulte un supuesto demasiado exigente que no se condiga con la realidad. En tal caso, veremos ahora un ejemplo donde se hace uso de la propuesta hecha en el punto 2, agregando variables observadas al problema y nuevos supuestos para el modelo, que permiten garantizar la identificabilidad. Consideremos el caso en el que el tratamiento no haya sido asignado de forma completamente aleatoria, sino que se ha priorizado a las personas cuya condición de salud inicial fue considerada crítica. Incorporamos la variable **factual** C donde $C(a) = 1$ si el individuo a está en condición inicial crítica y $C(a) = 0$ en el caso contrario. Observar que C pasa a ser una variable factual observada.

Supongamos ahora que entre las personas de situación crítica ($C = 1$) han sido elegidos los tratados de forma completamente aleatoria y entre el resto de la población también, pero en cada uno de estos grupos (críticos y no críticos) tratados y controlados se eligen con mecanismos distintos. Es decir, para cada nivel de la variables C el tratamiento se asigna mediante aleatorización completa. En tales circunstancias, decimos que el tratamiento se asignó por «aleatorización condicional». Un ejemplo para retratar esto sería suponer que el tratamiento se asignó al 50 por ciento de los pacientes en situación crítica y al 25 por ciento del resto de la población, es decir:

$$P(X = x_1|C = 1) = \frac{1}{2}$$
$$P(X = x_1|C = 0) = \frac{1}{4}$$

Ahora, **fijado el valor de** C, tenemos aleatorización completa, por lo que la asignación del tratamiento resulta nuevamente independiente de cualquier otra variable correspondiente al experimento. En particular:

$$Y_{x_i} \coprod X|C, \forall i \in \{0,1\}. \tag{1.8}$$

Veremos que en este nuevo escenario también se logra identificabilidad del ATE conociendo la distribución de las variables (C, X, Y). Hallemos la expresión para $E[Y_{x_i}], i \in \{0,1\}$ en términos de la distribución de las variables observadas:

$$E[Y_{x_i}] \;=\; E[E[Y_{x_i}|C]] \;= \underbrace{E[E[Y_{x_i}|X = x_i, C]]}_{\text{por (1.8)}} \;=\; \underbrace{E[E[Y|X = x_i, C]]}_{consistencia}$$

Conclusión: Antes no teníamos identificabilidad pero agregando una variable al modelo y asumiendo el supuesto de aleatorización condicional, conseguimos identificabilidad.

Para hablar del punto 3 referente a cotas, supongamos que se han extraviado los datos correspondientes a la condición inicial del individuo. Veremos más adelante que observando tan sólo (X, Y), sin imponer modelo alguno, el efecto medio del tratamiento no queda identificado mediante la distribución del par (X, Y). Ante esta contingencia, acotaremos el valor que pueda tomar el ATE, a partir de la información dada por la distribución de (X, Y).

Consideremos entonces conocidos los valores de $p_{(X,Y)}(i, j), \forall i, j \in \{0,1\}$, dados por la siguiente tabla:

$X \setminus Y$	0	1
x_0	p_1	p_2
x_1	p_3	p_4

Sabemos entonces que $\sum_{i=1}^{4} p_i = 1$ y $p_i \geq 0, \forall 1 \leq i \leq 4$, por ser las probabilidades puntuales de un vector aleatorio discreto.

Recordemos que Y queda determinada por la acción asignada (X) y ambas variables contrafactuales $Y_{x_i}, \forall i \in \{0,1\}$, mediante la hipótesis de consistencia dada en (1.1). Si bien no sabemos la distribución de Y_{x_i}, para $i \in \{0,1\}$, sabemos que toma valores finitos, al igual que X, por lo que el vector (X, Y_{x_1}, Y_{x_0}) también lo hace. Si bien no son conocidos, consideremos los valores para las probabilidades puntuales de este vector:

$X \setminus (Y_{x_1}, Y_{x_0})$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
x_0	r_1	r_2	r_3	r_4
x_1	r_5	r_6	r_7	r_8

De nuevo, $\sum_{i=1}^{8} r_i = 1$ y $r_i \geq 0, \forall 1 \leq i \leq 8$.

Este vector que parece sacado de la galera, busca relacionar lo que sabemos de Y a las variables contrafactuales, sin deshacernos de la información dada por X, aprovechando que Y está determinada por Y_{x_1}, Y_{x_0}, X. Por ejemplo, observemos lo siguiente:

$$\begin{aligned} p_1 &= P(X = x_0, Y = 0) \\ &= \underbrace{P(X = x_0, Y_{x_0} = 0)}_{consistencia} \\ &= \sum_{i=0}^{1} P(X = x_0, Y_{x_1} = i, Y_{x_0} = 0) \\ &= r_1 + r_3 \end{aligned}$$

O sea, podemos determinar las probabilidades puntuales de (X, Y) como función de los r_i de forma análoga. Resultando en las siguientes ecuaciones:

$$p_1 = P(X = x_0, Y = 0) = r_1 + r_3 \tag{1.9}$$
$$p_2 = P(X = x_0, Y = 1) = r_2 + r_4 \tag{1.10}$$
$$p_3 = P(X = x_1, Y = 0) = r_5 + r_6 \tag{1.11}$$
$$p_4 = P(X = x_1, Y = 1) = r_7 + r_8 \tag{1.12}$$

Observación 1.5.4 *Notemos que como $r_i \geq 0, \forall i$, tenemos las siguientes desigualdades:*

$$r_3 = p_1 - r_1 \leq p_1 \tag{1.13}$$
$$r_2 = p_2 - r_4 \leq p_2 \tag{1.14}$$
$$r_6 = p_3 - r_5 \leq p_3 \tag{1.15}$$
$$r_7 = p_4 - r_8 \leq p_4 \tag{1.16}$$

Recordemos el ATE del caso binario, dado por

$$P(Y_{x_1} = 1) - P(Y_{x_0} = 1).$$

Tenemos una nueva ventaja, es que podemos expresar la distribución de las variables contrafactuales en función de estas incógnitas. Calculemos cada una de estas probabilidades en función de los r_i.

$$P(Y_{x_1} = 1) = \underbrace{\sum_{i,j=0}^{1} P(X = x_i, Y_{x_1} = 1, Y_{x_0} = j)}_{\text{Probabilidad total}} = r_3 + r_7 + r_4 + r_8$$

$$P(Y_{x_0} = 1) = \underbrace{\sum_{i,j=0}^{1} P(X = x_i, Y_{x_1} = j, Y_{x_0} = 1)}_{\text{Probabilidad total}} = r_2 + r_6 + r_4 + r_8$$

De aquí que nuestro parámetro causal de interés lo podemos escribir de la siguiente forma:

$$ATE(r) = r_3 + r_7 + r_4 + r_8 - (r_2 + r_6 + r_4 + r_8) = r_3 + r_7 - r_2 - r_6 \tag{1.17}$$

Mostraremos ahora dos posibles distribuciones $(r, \widetilde{r} \in \mathcal{S}_8)$ para el vector (X, Y_{x_0}, Y_{x_1}), que inducen la misma distribución en el par (X, Y) teniendo asociados diferentes valores para el efecto medio del tratamiento. Detallamos los valores para $r = (r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8)$ y $\widetilde{r} = (\widetilde{r}_1, \widetilde{r}_2, \widetilde{r}_3, \widetilde{r}_4, \widetilde{r}_5, \widetilde{r}_6, \widetilde{r}_7, \widetilde{r}_8)$ en los Cuadros 1.3 y 1.4 respectivamente:

$X \setminus (Y_{x_1}, Y_{x_0})$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
x_0	$5/40$	$5/40$	$5/40$	$5/40$
x_1	$5/40$	$5/40$	$5/40$	$5/40$

Cuadro 1.3: Valores del vector r

$X \setminus (Y_{x_1}, Y_{x_0})$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
x_0	$5/40$	$5/40$	$5/40$	$5/40$
x_1	$8/40$	$2/40$	$9/40$	$1/40$

Cuadro 1.4: Valores del vector $\widetilde{r}$

Verifiquemos primero que ambas inducen la misma probabilidad puntual para (X, Y), usando (1.9), (1.10), (1.11) y (1.12):

- $\begin{cases} p_1 = r_1 + r_3 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \\ p_1 = \widetilde{r}_1 + \widetilde{r}_3 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \end{cases}$
- $\begin{cases} p_2 = r_2 + r_4 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \\ p_2 = \widetilde{r}_2 + \widetilde{r}_4 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \end{cases}$
- $\begin{cases} p_3 = r_5 + r_6 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \\ p_3 = \widetilde{r}_5 + \widetilde{r}_6 = \dfrac{8}{40} + \dfrac{2}{40} = \dfrac{1}{4} \end{cases}$

- $$\begin{cases} p_4 = r_7 + r_8 = \dfrac{5}{40} + \dfrac{5}{40} = \dfrac{1}{4} \\ \\ p_4 = \widetilde{r}_7 + \widetilde{r}_8 = \dfrac{9}{40} + \dfrac{1}{40} = \dfrac{1}{4} \end{cases}$$

Sin embargo, usando la ecuación (1.17) vemos que:

$$ATE(r) = \frac{5}{40} + \frac{5}{40} - \frac{5}{40} - \frac{5}{40} = 0 \neq \frac{7}{40} = \frac{5}{40} + \frac{9}{40} - \frac{5}{40} - \frac{2}{40} = ATE(\widetilde{r})$$

Por lo que el parámetro causal no es identificable a partir de (X, Y).

Ante la falta de indentificabilidad, buscaremos mínimo y máximo para la expresión dada por (1.17). Si queremos que las cotas halladas sean precisas, debemos valernos de toda la información de la cual disponemos.

Buscamos un máximo y un mínimo para la expresión dada por (1.17) en el conjunto de distribuciones r compatibles con la distribución de los datos observados. Es decir, dado p, r será un vector del subconjunto de $\mathcal{S}_8$ que cumple las ecuaciones (1.9), (1.10), (1.11) y (1.12). Notemos que, gracias a la positividad de cada r_i y usando las desigualdades dadas por (1.13), (1.14), (1.15) y (1.16):

$$r_3 + r_7 - r_2 - r_6 \leq r_3 + r_7 \leq p_1 + p_4$$

$$r_3 + r_7 - r_2 - r_6 \geq -(r_2 + r_6) \geq -(p_2 + p_3)$$

Como conclusión tenemos que

$$-(p_2 + p_3) \leq P(Y_{x_1} = 1) - P(Y_{x_0} = 1) \leq p_1 + p_4. \tag{1.18}$$

Observación 1.5.5 *Recordemos que el ATE (dado por (1.4)), en el caso de ser positivo, refleja que el tratamiento genera mayor proporción de respuestas positivas que el control. De la misma forma, en el caso en que el ATE sea negativo, el control es el que deja mayor proporción de respuestas positivas. Si la cota inferior es positiva o la cota inferior es negativa, podemos deducir el signo del ATE y concluir que acción es mas ventajosa.*

Observemos las desigualdades dadas en (1.18) y consideremos los siguientes casos:

1. $\underline{p_2 = p_3 = 0:}$
   Esto nos diría que $-(p_2 + p_3) = 0 \leq P(Y_{x_1} = 1) - P(Y_{x_0} = 1)$, *dando así un ATE que no puede ser negativo. En este caso, los investigadores se decidirán por el tratamiento.*

2. $\underline{p_1 = p_4 = 0:}$
   Esto nos diría que $p_1 + p_4 = 0 \geq P(Y_{x_1} = 1) - P(Y_{x_0} = 1)$, *dando así un ATE que no puede ser positivo. En este caso, los investigadores se decidirán por el control.*

En el último capítulo, daremos un enfoque más amplio a este problema.

1.6. Cotas

Recordemos que una vez planteado un modelo, no todas las variables tienen por qué ser observadas, por lo que tendremos un conjunto S de variables observadas cuya distribución es p. Esta distribución pertenecerá a un conjunto $\mathcal{F}$, que estará constituida por las posibles distribuciones para S. Consideremos un parámetro causal de interés (llamado PC) que no es identificado a partir de nuestras variables observadas. En el caso de que exista, sea R a un conjunto de variables con distribución r perteneciente a cierto modelo $\mathcal{R}$ que cumpla las siguientes condiciones:

- Existe un operador T_{caus} tal que $T_{caus}(r) = PC$
- Existe un operador T_{mg} tal que $T_{mg}(r) = p$

Es decir, R será un conjunto de variables que, cumpliendo un modelo $\mathcal{R}$, determine el PC al igual que la distribución de las variables observadas dada por p. Notemos que estos operadores no son necesariamente inyectivos, podría haber muchos $r \in \mathcal{R}$ que cumplan $T_{caus}(r) = PC$ o $T_{mg}(r) = p$.

Observación 1.6.1 *Usamos el término «operadores» porque se aplica a las posibles r* ***funciones de distribución*** *de R.*

En el ejemplo anterior (ver el la Sección 1.5), tenemos que:

- $R = (X, Y_{x_0}, Y_{x_1}) \sim r \in \mathcal{R} = \{r = (r_1, \cdots, r_8) : r_i \geq 0, \sum_{i=1}^{8} r_i = 1\}$.
- $T_{caus}(r) = r_3 + r_7 - r_2 - r_6$
- $T_{mg}(r) = (r_1 + r_3, r_2 + r_4, r_5 + r_6, r_7 + r_8)$.

Para cada p posible distribución, queremos $m(p), M(p)$ cotas para el parámetro causal (mínimo y máximo respectivamente), sabiendo que el mismo es compatible con p.

Formalizaremos un poco lo mencionado anteriormente:

Definición 1.6.2 *Consideremos $\widetilde{X}$ un conjunto de variables, S un conjunto de variables observadas con distribución $p \in \mathcal{F}$ y PC un parámetro causal de interés. Sea R un conjunto de variables con distribución r perteneciente al modelo $\mathcal{R}$ de forma que $\exists T_{caus}, T_{mg}$ tal que $T_{caus}(r) = PC$ y $T_{mg}(r) = p$. Bajo estas condiciones, decimos que $m, M : \mathcal{F} \to \mathbb{R}$ son cotas para el PC si*

$$\forall p \in \mathcal{F}, m(p) \leq T_{caus}(r) \leq M(p), \forall r \in \mathcal{R} : T_{mg}(r) = p$$

Definición 1.6.3 *Bajo las mismas hipótesis que la definición anterior, decimos que* m, M *son cotas finas si* $\forall p \in \mathcal{F}, \exists r_1, r_2 \in \mathcal{R}$ *tal que:*

$$T_{caus}(r_1) = m(p), T_{mg}(r_1) = p$$

$$T_{caus}(r_2) = M(p), T_{mg}(r_2) = p$$

Notemos que si podemos asegurar que el mínimo y/o máximo se alcanza, tenemos las siguientes cotas finas

$$m(p) = \min_{r:T_{mg}(r)=p} T_{caus}(r)$$

$$M(p) = \max_{r:T_{mg}(r)=p} T_{caus}(r)$$

Por lo que, en el fondo, buscar cotas finas será resolver los siguientes problemas de optimización:

$$\left\{ \begin{array}{c} \min T_{caus}(r) \\ \\ T_{mg}(r) = p \\ \\ r \in \mathcal{R} \end{array} \right\} \quad y \quad \left\{ \begin{array}{c} \max T_{caus}(r) \\ \\ T_{mg}(r) = p \\ \\ r \in \mathcal{R} \end{array} \right\} \tag{1.19}$$

En el Capítulo 4 veremos que para el problema que abordaremos, existirá una matriz $A \in \mathbb{R}^{m \times n}$ y un vector $c \in \mathbb{R}^n$ (determinaremos también n y m) para los cuales encontrar cotas finas para el parámetro causal resultará equivalente a resolver los siguientes problemas de optimización:

$$\left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \sum_{i=1}^{n} r_i = 1 \\ \widetilde{r} \geq 0 \end{array} \right\} \quad y \quad \left\{ \begin{array}{c} \max c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = p \\ \sum_{i=1}^{n} \widetilde{r}_i = 1 \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{1.20}$$

Es por ello que terminaremos el presente capítulo mencionando algunos resultados concernientes a la programacion lineal.

1.7. Programación lineal

Hemos hecho uso del Libro de Castillo et. al. [4] para nuestra recolección de propiedades relativas a la programación lineal. La formalización de estos asuntos requiere adentrarse en un mar de numerosas definiciones que no aportan al objetivo de este trabajo, sin embargo necesitaremos usar varias de ellas como herramientas. Trataremos de

mencionar lo indispensable, pues en la bibliografía estas cuestiones están completamente detalladas.

Consideremos el siguiente problema de optimización:

$$\left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \sum_{i=1}^{n} \widetilde{r}_i = 1 \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{1.21}$$

En este problema tenemos que optimizar una función, conscientes de que las variables cumplen ciertas restricciones. Estas restricciones nos determinan un conjunto llamado **región de factibilidad** o **conjunto de soluciones factibles**, que consta del subconjunto (en este caso, subconjunto de $\mathbb{R}^n$) que cumple las restricciones. Por ejemplo, en (1.21) el conjunto de soluciones factibles es

$$S = \{\widetilde{r} \in \mathbb{R}^n : A \cdot \widetilde{r} = \widetilde{p} \wedge \sum_{i=1}^{n} \widetilde{r}_i = 1 \wedge \widetilde{r} \geq 0\}.$$

Cuando el funcional a optimizar y las restricciones de las soluciones factibles son lineales en las variables (y finitas) se dice que tenemos un problema de programación lineal (para simplificar, lo llamaremos a veces PPL). Las restricciones pueden estar dadas por sistemas de ecuaciones e inecuaciones. En el ejemplo, denotando por $(A)_{i,*}$ la fila i de la matriz A, a $(\widetilde{p})_i$ la i-ésima coordenada del vector $\widetilde{p}$ y U_n el vector n-dimensional de unos, el problema dado por (1.21) consta de las siguientes restricciones:

$$(A)_{i,*} \cdot \widetilde{r} = (\widetilde{p})_i, \quad \forall 1 \leq i \leq m$$

$$\vec{U_n} \cdot \widetilde{r} = 1$$

$$\widetilde{r}_j \geq 0, \quad \forall 1 \leq j \leq n$$

Notar que una ecuación de la forma $a \cdot \widetilde{r} = cte$ determina un hiperplano en $\mathbb{R}^n$. El mismo hiperplano determina dos semiespacios $\{a \cdot \widetilde{r} \leq cte\}$ y $\{a \cdot \widetilde{r} \geq cte\}$. Por lo que en un problema de programación lineal, el conjunto de soluciones factibles es una intersección **finita** de hiperplanos y semiespacios, que se define como un **poliedro convexo**. Definiremos el concepto de conjunto convexo a continuación:

Definición 1.7.1 *Un conjunto C se dice convexo si $\forall x, y \in C$, se cumple*

$$\lambda.x + (1 - \lambda).y \in C, \forall \lambda \in [0, 1].$$

Es decir, dados dos elementos del conjunto, el segmento que los une también está incluido en el conjunto.

Ya caracterizado el conjunto de soluciones factibles, dentro de ellas, tendremos un mínimo en $\widetilde{r}_0$ si $c \cdot \widetilde{r}_0 \leq c \cdot \widetilde{r}$, $\forall \widetilde{r}$ solución factible. Si se busca minimizar la función, se denominará **solución óptima** a un $\widetilde{r}_0 \in \mathbb{R}^n$ que realice este mínimo.

Del mismo modo, si se quiere maximizar la función, $\widetilde{r}_0 \in \mathbb{R}^n$ será una **solución óptima** si $c \cdot \widetilde{r}_0 \geq c \cdot \widetilde{r}$, $\forall \widetilde{r}$ solución factible.

Cabe hacer la distinción entre solución óptima y óptimo, si $\widetilde{r}_0$ es una solución óptima llamaremos **óptimo** a $c \cdot \widetilde{r}_0$. Notar que el óptimo es un número real, mientras que una solución óptima es un vector en $\mathbb{R}^n$. Además, el óptimo (en caso de existir) es único, mientras que la solución óptima puede no serlo.

Observación 1.7.2 *Notemos que la única diferencia entre (1.20) y (1.21) es que no aparece el problema de maximización. Con las definiciones hechas podemos explicar el motivo de esta omisión. Consideremos el siguiente problema:*

$$\left\{ \begin{array}{c} min \ -c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \sum_{i=1}^{n} \widetilde{r}_i = 1 \\ \widetilde{r} \geq 0 \end{array} \right\}$$

Llamemos $\widetilde{r}_0$ a la solución óptima de este problema y $-c \cdot \widetilde{r}_0$ su respectivo óptimo. Como $\widetilde{r}_0$ es solución óptima, $-c \cdot \widetilde{r}_0 \leq -c \cdot \widetilde{r}$, $\forall \widetilde{r}$ solución factible. Multiplicando por -1 a ambos lados, $c \cdot \widetilde{r}_0 \geq c \cdot \widetilde{r}$, $\forall \widetilde{r}$ solución factible. Entonces, por definición, $\widetilde{r}_0$ resulta solución óptima del problema de maximización dado en (1.20).

Esta equivalencia nos permite ver cualquier PPL de maximización como uno de minimización. Haremos uso de esta propiedad para evitar mencionar ambos problemas y tratar sólo con el de minimización cuando sea necesario.

Una observación interesante es la que daremos a continuación:

Observación 1.7.3 *Si $\widetilde{r}_1$ y $\widetilde{r}_2$ son soluciones óptimas de un PPL con óptimo $m = c \cdot \widetilde{r}_1 = c \cdot \widetilde{r}_2$, entonces $\lambda.\widetilde{r}_1 + (1-\lambda).\widetilde{r}_2$ es solución óptima para todo $\lambda \in [0,1]$, pues*

$$\begin{aligned} c \cdot (\lambda \cdot \widetilde{r}_1 + (1-\lambda) \cdot \widetilde{r}_2) &= \lambda \cdot c \cdot \widetilde{r}_1 + (1-\lambda) \cdot c \cdot \widetilde{r}_2 \\ &= \lambda \cdot m + (1-\lambda) \cdot m \\ &= m \end{aligned}$$

En definitiva, cualquier punto del segmento comprendido entre dos soluciones óptimas es solución óptima.

Notar que usamos fuertemente que la región de factibilidad es convexa y que la función a optimizar es lineal. Por lo que esta es una característica que puede no darse si el problema de optimización no es de programación lineal.

Daremos la definición de punto extremo de un poliedro convexo, que luego se verá que caracterizan las posibles soluciones óptimas:

Definición 1.7.4 *Dado C un conjunto convexo, $z \in C$ se dice punto extremo si dados $x, y \in C$ y $\lambda \in (0,1)$ tal que $\lambda.x + (1-\lambda).y = z$, necesariamente $x = y = z$. Es decir, un punto extremo no pertenece a ningún segmento comprendido entre puntos distintos de C, a menos que sea justamente un extremo del mismo.*

Un resultado importantísimo, por algo se llama como se llama, es el siguiente:

Teorema 1.7.5 *(Propiedad fundamental de la Programación Lineal) Si un Problema de Programación Lineal tiene **óptimo**, existe un punto extremo que es **solución óptima** del problema.*

Se puede demostrar que los puntos extremos son finitos, por lo que, si se sabe que un PPL tiene óptimo, la solución óptima estará entre los finitos puntos extremos. Tendremos que hacer uso de unas definiciones más para esbozar la demostración:

Definición 1.7.6 *Dado el PPL:*

$$\left\{ \begin{array}{c} min\ c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = b \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{1.22}$$

*Supongamos que la matriz $A \in \mathbb{R}^{m \times n}$ tiene rango $m \leq n$. Sea $B \in \mathbb{R}^{m \times m}$ una submatriz de A, ésta se denomina **matriz básica** si B tiene rango m. Además, se llama **matriz básica y factible** si $B^{-1} \cdot b \geq 0$.*

Definición 1.7.7 *Para cada matriz básica y factible $B = \left(\begin{array}{c|c|c|c} A_{*,j_1} & A_{*,j_2} & \cdots & A_{*,j_m} \end{array} \right)$ (donde $A_{*,j}$ es la j-ésima columna de A y $j_1 < j_2 < \cdots < j_m$) podemos construir de forma única el vector $\widetilde{r}_B \in \mathbb{R}^n$ tal que*

$$(\widetilde{r}_B)_i = \left\{ \begin{array}{cc} (B^{-1}.b)_l & si\ \exists 1 \leq l \leq m\ tal\ que\ i = j_l \\ 0 & si\ \ \nexists 1 \leq l \leq m : i = j_l \end{array} \right\}$$

*Notemos que r_B es una solución factible del problema ($A.\widetilde{r}_B = b$ y $\widetilde{r}_B \geq 0$). Al estar asociada a una matriz básica y factible, diremos que $\widetilde{r}_B$ es una **solución básica y factible** del problema.*

Se puede ver que todos los puntos extremos son además soluciones básicas y factibles. Se da también la recíproca. Como las posibles matrices básicas y factibles son finitas (son submatrices de A), también lo son las soluciones básicas y factibles, por lo que también son finitos los puntos extremos. La finitud de los mismos será usado en el último capítulo.

Hay algoritmos que resuelven estos problemas, el más conocido es el algoritmo Simplex, que resuelve este problema en tiempo polinomial para el caso promedio. Aunque hay otros procedimientos que lo resuelven en tiempo polinomial, el Simplex le hace honor al nombre (en comparación, se entiende).

Capítulo 2

Grafos y Probabilidades

Supongamos que tenemos $\widetilde{X} = (X_1, X_2, ..., X_n)$ un vector aleatorio discreto y $\vec{x} = (x_1, x_2, ..., x_n)$ una realización del vector $\widetilde{X}$. Siempre que tenga sentido condicionar, sabemos que, por regla multiplicativa,

$$P(\widetilde{X} = \vec{x}) = \prod_{j=1}^{n} P(X_j = x_j | X_i = x_i, \forall i < j) \tag{2.1}$$

Pero quizás condicionar a **todas** las variables X_i con $i < j$ sea innecesario. Por ejemplo, es razón de serio escepticismo pensar que tomar café una hora antes del partido de Racing influirá sobre el resultado del partido, por lo que condicionar o no al consumo de la tan popular infusión no modificará las probabilidades de que el amado club de Avellaneda logre otros tres puntos.

Entonces uno se pone a pensar si puede disminuir la cantidad de variables a las que condiciona X_j de forma de obtener las mismas probabilidades **condicionales**. Es decir, ¿existirán $S_1, S_2, ...S_n$, con $S_j \subseteq \{1, 2, ..., j-1\}$ tal que

$$P(X_j = x_j | X_i = x_i, \forall i < j) = P(X_j = x_j | X_i = x_i, \forall i \in S_j)?$$

Dependiendo de la naturaleza de las variables en cuestión, podemos plantear modelos en los que no sólo esto es cierto, sino que podemos identificar dichos S_j. Para ello, una herramienta muy usada es asumir un modelo gráfico, que facilitará el estudio de las relaciones existentes entre las coordenadas del vector $\widetilde{X}$.

2.1. Grafos

Un grafo es un par ordenado de conjuntos $G = (V, E)$ con V un conjunto finito (llamado conjunto de **nodos** o **vértices**) y E es un subconjunto de $V \times V$ (llamado conjunto de **aristas** o **ramas**) que marcan una conexión entre nodos (de ahí que deviene

en un subconjunto de $V \times V$). Si $(u, v) \in E$ diremos que u y v son adyacentes. De aquí en adelante, utilizaremos indistintamente los términos nodos y vértices para referirnos a elementos de V, al igual que serán usados tanto ramas como aristas para elementos de E.

2.1.1. Grafos dirigidos y no dirigidos

Un grafo **no dirigido** es uno que no distingue entre las aristas (u, v) y (v, u), simplemente dos nodos son adyacentes si hay una «relación simétrica» entre ambos (por ejemplo, si los nodos representaran personas, ser hermanos, ser compatriotas o ser compañeros de truco daría una adyacencia en un grafo no dirigido).

Pero a la hora de modelar, no siempre tiene sentido suponer una relación necesariamente ambivalente (no es lo mismo ser padre de alguien que ser hijo de alguien, no es lo mismo ganar que perder y, a todos nos ha pasado, la atracción física por otra persona no siempre es recíproca). Por lo que a veces tiene sentido pensar que las aristas (u, v) y (v, u) puedan ser distintas. Notaremos una arista dirigida (u, v) como $u \to v$. En este caso tenemos un grafo dirigido y tenemos una noción de incidencia de un nodo sobre otro. Observemos que esta noción de incidencia difiere de la de adyacencia. Para retratar esta observación, u puede ser adyacente a v, pero u puede no incidir en v (es el caso $v \to u$).

Los grafos son una herramienta matemática muy útil a la hora de plantear modelos que establecen relaciones entre objetos, pondremos como ejemplo grafos genealógicos, representados en las figuras 2.1 y 2.2:

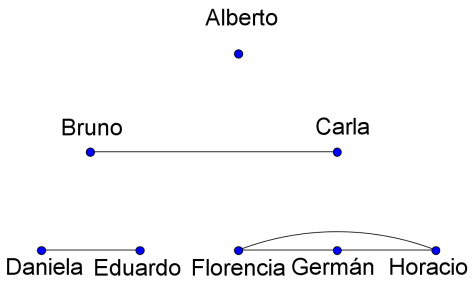


Figura 2.1: Grafo no dirigido donde los nodos son personas que componen una familia, se agrega una arista entre dos personas si ambos son hermanos.

Para nuestras aplicaciones, no permitiremos la existencia (por dictatorial que suene) de ambas aristas dirigidas entre dos nodos, es decir, si $u \to v$, entonces $v \not\to u$. Nos

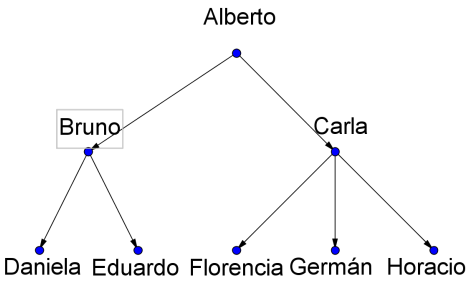


Figura 2.2: Grafo dirigido donde los nodos son personas que componen la misma familia, pero se agrega una arista desde progenitores hasta sus respectivos hijos.

referiremos siempre a grafos dirigidos, asumiendo implícitamente que se cumple tal propiedad.

Uno de los ejemplos dados para simbolizar la necesidad de poder pensar aristas dirigidas fue la de la relación padre-hijo. Esto no fue un mero hecho del azar, ya que es muy usada esta analogía en el ámbito de grafos dirigidos. Dado un grafo dirigido G y dos nodos $u, v \in V$, decimos que u es padre de v si $u \to v$, dejando espacio a la obvia definición que v es hijo de u si $u \to v$ (notar que para la definición no tendría sentido biológico la simultaneidad de $u \to v$ y $v \to u$). Dado $v \in V$, denotamos el conjunto de padres de v como

$$pa_G(v) = \{u \in V : u \to v\}$$

y extendemos esta definición a subconjuntos de V de la siguiente forma:

$$pa_G(W) = \bigcup_{v \in W} pa_G(v) \text{ para } W \subseteq V.$$

2.1.2. Caminos y caminos dirigidos en un grafo

Imaginemos que debemos transitar la ciudad de Buenos Aires, teniendo una esquina de la cual partimos y otra a la cual queremos llegar.

Estando a pie, para llegar de una esquina a otra esquina «vecina» (es decir, no hay necesidad de pasar por otra esquina para llegar a ella) debemos recorrer una calle. Nuestra condición de peatones nos permite obviar el sentido del tránsito al que están sujetos los automovilistas para llegar de una esquina a otra. Entonces, para caminar de un punto de partida hasta uno de llegada, lo único que tendré que determinar son una cantidad de esquinas «vecinas» cada una con la anterior para lograr mi objetivo.

Pero una vez que nos hacemos de un vehículo de 4 ruedas, no podemos omitir el sentido que tienen las calles a la hora de transitarlas (al menos, sin tener consecuencias

desastrosas). Entonces, aunque dos esquinas sean «vecinas», no siempre podemos llegar de una a otra sin pasar por otra esquina.

Si en este ejemplo pensamos las esquinas (en realidad, las 4 esquinas de una bocacalle) como nodos y las calles que las unen como aristas, podemos establecer un grafo (enorme por cierto) que represente gráficamente nuestra ciudad. Podemos pensar un camino (no necesariamente único) de la esquina A hasta la esquina B como el conjunto de aristas empleadas para cumplir el recorrido. Otra forma de determinarlo es considerando las esquinas por las cuales se pasó y el orden en el que fueron visitadas. Con este ejemplo, damos una idea de caminos (posibles caminos de un peatón) y caminos dirigidos (camino de un vehículo, respetando los sentidos de las calles) en un grafo G.

Definición 2.1.1 *Un **camino** en un grafo (dirigido o no dirigido) G es una sucesión finita de **aristas** $(e_1, e_2, ..., e_r) \in E^r$, determinado por una tira (también finita) de nodos **adyacentes** $(v_0, v_1, ..., v_r) \in V^{r+1}$ tal que $e_i = (v_{i-1}, v_i)$ o $e_i = (v_i, v_{i-1}), \forall 1 \leq i \leq r$. En este caso, decimos que hay un camino entre v_0 y v_r.*

Notar que recorrer un camino en un grafo según el orden dado por los vértices que lo determinan no tiene porqué coincidir con el sentido de cada arista (se permite ir «a contramano»).

Diferente es la definición de un **camino dirigido** en un grafo dirigido G. En esta definición sí se respeta la dirección de las aristas.

Definición 2.1.2 *Dado G un **grafo dirigido**, un camino dirigido en G es una sucesión finita de aristas $(e_1, e_2, ..., e_r) \in E^r$ de forma que si $e_i = (v_i, v_{i+1})$ y $e_{i+1} = (w_{i+1}, w_{i+2})$, necesariamente $\forall 1 \leq i \leq r-1$ se cumple $v_{i+1} = w_{i+1}$. En este caso, tenemos un camino dirigido de v_1 a v_{r+1}.*

Deducimos que podemos darnos el lujo de pensar a cada $e_i = (v_i, v_{i+1})$ sin necesidad de hacer distinción entre los posibles nodos de cada arista. A partir de estas aclaraciones, podemos decir que en un camino dirigido, donde $e_i = (v_i, v_{i+1})$, v_i es **incidente** en v_{i+1} para todo $1 \leq i \leq r$. A grandes rasgos, siguiendo el orden dado por el camino, cada arista «comienza» donde «termina» la arista anterior.

En el campo de la causalidad se usan como modelos gráficos los llamados grafos acíclicos dirigidos (DAG: directed acyclic graph). Para eso, definiremos un grafo acíclico:

Definición 2.1.3 *Sea G un grafo dirigido, decimos que G es acíclico si $\forall v \in V$ no existe un camino dirigido de v a v.*

Ejemplo 2.1.4 *Observemos el siguiente grafo*

*En este grafo tenemos $V = \{v_1, v_2, v_3, v_4\}$ y $E = \{(v_4, v_1); (v_3, v_1); (v_3, v_2); (v_1, v_2)\}$. Analicemos los **caminos dirigidos** posibles:*

- $\{(v_4, v_1); (v_1, v_2)\}$ *camino dirigido de v_4 a v_2*

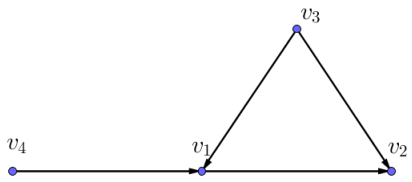


Figura 2.3: Un grafo dirigido que usaremos mucho en este trabajo.

- $\{(v_3, v_1); (v_1, v_2)\}$ *camino dirigido de* v_3 *a* v_2
- $\{(v_3, v_2)\}$ *camino dirigido de* v_3 *a* v_2
- $\{(v_3, v_1)\}$ *camino dirigido de* v_3 *a* v_1
- $\{(v_1, v_2)\}$ *camino dirigido de* v_1 *a* v_2
- $\{(v_4, v_1)\}$ *camino dirigido de* v_4 *a* v_1

Agregando los ***caminos*** *posibles:*

- $\{(v_4, v_1); (v_3, v_1)\}$ *camino de* v_4 *a* v_3
- $\{(v_4, v_1); (v_3, v_1); (v_3, v_2)\}$ *camino de* v_4 *a* v_2
- $\{(v_1, v_2); (v_1, v_4)\}$ *camino de* v_2 *a* v_4
- $\{(v_4, v_1); (v_1, v_2); (v_3, v_2)\}$ *camino de* v_4 *a* v_3
- $\{(v_4, v_1)\}$ *camino de* v_1 *a* v_4
- $\{(v_3, v_1)\}$ *camino de* v_1 *a* v_3
- $\{(v_3, v_2)\}$ *camino de* v_2 *a* v_3
- $\{(v_2, v_1)\}$ *camino de* v_2 *a* v_1

Dado G un grafo dirigido y $v, w \in V$ diremos que v es ancestro de w, si existe un camino dirigido de v a w. Ante esta situación diremos que w es descendiente de v.

Denotamos el conjunto de ancestros de w como

$$an_G(w) = \{v \in V : \exists \text{ un camino dirigido de } v \text{ a } w\}.$$

De manera similar, el conjunto de descendientes de v como

$$de_G(v) = \{w \in V : \exists \text{ un camino dirigido de } v \text{ a } w\}.$$

Al igual que antes, para subconjuntos $W \subseteq V$, $an_G(W) = \bigcup_{v \in W} an_G(v)$ y $de_G(W) = \bigcup_{v \in W} de_G(v)$.

Se puede demostrar que en un DAG, se pueden enumerar los vértices $V = \{v_1, v_2, ..., v_n\}$ de forma que $\forall 1 \leq i \leq n$, $an_G(v_i) \subseteq \{v_1, ...v_{i-1}\}$. A esto se le llama numeración topológica de los vértices. Siguiendo la numeración, cada v_j no tiene caminos dirigidos a ningún v_i con $i < j$. Una observación es que $pa_G(v_i) \subseteq an_G(v_i) \subseteq \{v_1, ...v_{i-1}\}$.

Ejemplo 2.1.5 *En la Figura 2.3 tenemos que el nodo v_3 es padre de v_2 por lo que no se cumple $pa_G(v_2) \subseteq \{v_i : 1 \leq i < 2\} = \{v_1\}$, pero daremos una nueva numeración de los vértices de este grafo para que se cumpla esta condición. Aprovechando que el grafo es acíclico, se elige algún vértice del cual sólo «salgan flechas» (sino exisitiera tal nodo, habría un ciclo) y se elige como primer nodo. Se eliminan las flechas que salen de él y nuevamente, tiene que haber un vértice del cual «salgan flechas», se elige este como segundo y se repite el procedimiento.*

Tenemos así, la nueva numeración de los vértices dando un grafo isomorfo al anterior, representado en la figura 2.1.5:

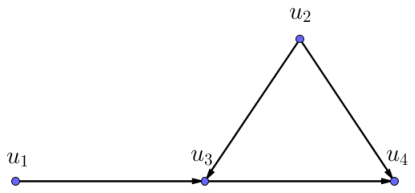


Figura 2.4: Un grafo isomorfo al de la Figura 2.3, pero numerado topológicamente.

Todas estas definiciones nos serán de utilidad en las siguientes secciones.

2.2. Grafos compatibles con una probabilidad P

Uno podría preguntarse (con razón) qué relación guardan todas estas definiciones con lo mencionado al principio del capítulo. Usaremos grafos dirigidos para identificar las variables que son determinantes en la distribución condicional de las restantes.

Supongamos que tenemos un grafo dirigido acíclico numerado topológicamente con $V = \{v_1, v_2, ..., v_n\}$ y $X = (X_{v_1}, X_{v_2}, ..., X_{v_n})$ (manteniendo el orden de los nodos de G) un vector aleatorio n-dimensional con distribución conjunta P.

Observación 2.2.1 *Notar que cada variable X_{v_i} se corresponde unívocamente con un nodo v_i del grafo G. Muchos autores, aprovechando esta correspondencia, no hacen la distinción de notación entre las variables de un vector y los nodos de un grafo. Sin embargo, al ser distintos como objetos matemáticos nos parece pertinente diferenciar la notación, porque no hacerlo pueda generar cierta confusión entre las herramientas gráficas y las herramientas probabilísticas que uno usa.*

Definición 2.2.2 *Sea $G = (V, E)$ un grafo acíclico dirigido y sea $X = (X_{v_1}, X_{v_2}, ...X_{v_n})$ un vector aleatorio con distribución P. Diremos que la probabilidad P (o el vector X) es compatible con el grafo G si:*

$$X_{v_j}|\{X_{v_i}, \forall i < j\} \sim X_{v_j}|\{X_{v_i} : v_i \in pa_G(v_j)\}, \forall j. \tag{2.2}$$

Observación 2.2.3 *Notar que si G es un DAG cuyos nodos están numerados topológicamente, dada la numeración de las variables, tenemos que $\forall j, \{X_{v_i} : v_i \in pa_G(v_j)\} \subseteq \{X_{v_i} : \forall i < j\}$.*

Es decir, un vector X resulta compatible con un DAG G cuando, para cada variable X_{v_j}, la distribución de X_{v_j} dadas todas las variables «anteriores» (el subconjunto $\{X_{v_i} : i < j\}$) coincide con la distribución condicional utilizando solo las variables asociadas a los padres del nodo v_j en el grafo G, desechando variables innecesarias.

Recordemos la regla multiplicativa, dada por (2.1). Si sabemos que la distribución P es compatible con el grafo G, por (2.2), tenemos la siguiente igualdad:

$$P(X_{v_1} = x_1, X_{v_2} = x_2, ..., X_{v_n} = x_n) = \prod_{j=1}^{n} P(X_j = x_j | X_i = x_i, \forall i < j) \tag{2.3}$$

$$= \prod_{j=1}^{n} P(X_{v_j} = x_j | X_{v_i} = x_i : v_i \in pa_G(v_j)). \tag{2.4}$$

Esta es llamada la **factorización Markov de** P.

Hemos visto aquí una implicación, pero se puede ver que vale la recíproca, dada en el siguiente lema:

Lema 2.2.4 *Sea $G = (V, E)$ un DAG con $\#V = n$, $\widetilde{X} = (X_1, X_2, ..., X_n)$ un vector aleatorio con distribución P. Además, sea $\vec{x} = (x_1, x_2, ..., x_n)$ un posible valor para $\widetilde{X}$. Entonces:*

$$P \text{ es compatible con } G \Leftrightarrow P(\widetilde{X} = \vec{x}) = \prod_{j=1}^{n} P(X_{v_j} = x_j | X_{v_i} = x_i : v_i \in pa_G(v_j))$$

En los trabajos de Pearl, se hace uso de esta equivalencia, ya que se define directamente a un grafo compatible con P como un grafo que cumple la Factorización Markov dada en (2.4).

Pasaremos a unos ejemplos:

Ejemplo 2.2.5 *Como todos sabemos, el aprendizaje es un proceso de construcción. Uno empieza por lo básico para luego abordar lo más complejo. Por ejemplo, primero uno aprende a sumar, luego a restar, después a multiplicar, dividir y a realizar potencias (en ese orden).*

Supongamos que una carrera tiene 5 materias completamente correlativas (lo que nos da una noción de «primera» materia), donde cada materia usa herramientas de todas las «anteriores». Entonces, si pudiéramos cuantificar un «nivel de aprendizaje» de cada materia, esta magnitud se verá influida por lo que se aprendió en las materias anteriores. Es decir, si $X_i =$«nivel de aprendizaje de la materia i» $(1 \leq i \leq 5)$, el vector $\widetilde{X} = (X_1, X_2, X_3, X_4, X_5)$ tendrá distribución compatible con el siguiente grafo:

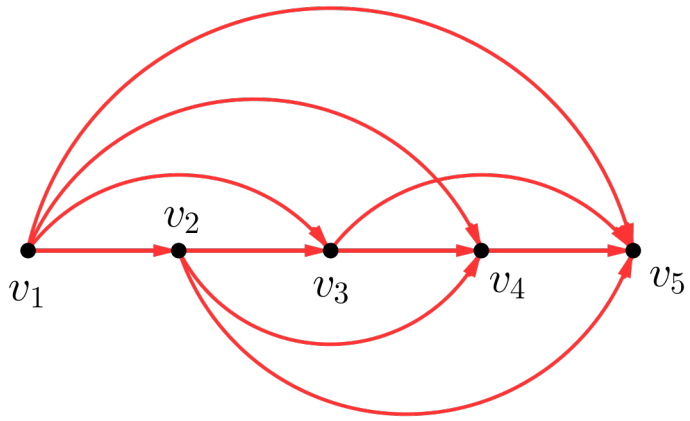


Figura 2.5: Un grafo dirigido compatible con la distribución de $\widetilde{X}$

Ejemplo 2.2.6 *Un jugador tira 5 veces una moneda equilibrada. En cada tirada, el jugador gana \$1 si sale cara y pierde \$1 si sale ceca. Consideremos las siguientes variables aleatorias:*

- $X_i = \left\{ \begin{array}{ll} 1 & \textit{si la i-ésima tirada salió cara} \\ & \\ -1 & \textit{si la i-ésima tirada salió ceca} \end{array} \right\} \ (1 \leq i \leq 5)$
- $S_n = \sum_{i=1}^{n} X_i, \ 1 \leq i \leq 5$

Notemos que la variable S_n representa la ganancia parcial del jugador luego de la tirada n, que dependerá exclusivamente de la ganancia hasta la tirada anterior, pues si $X_n = 1 \Rightarrow S_n = S_{n-1} + 1$ y si $X_n = -1 \Rightarrow S_n = S_{n-1} - 1$. Por lo que el vector $\widetilde{S} = (S_1, S_2, S_3, S_4, S_5)$ tendrá distribución compatible con el siguiente grafo:

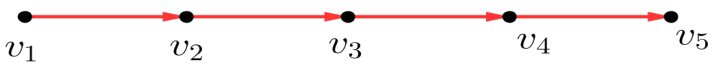


Figura 2.6: Un grafo dirigido compatible con la distribución de $\widetilde{S}$

Notemos además que la distribución del vector $\widetilde{S}$ también es compatible con el grafo dado en la Figura 2.5, porque agrega aristas al grafo de la Figura 2.6, sin quitar las aristas «importantes».

Observación 2.2.7 *Todo vector $X = (X_{v_1}, X_{v_2}, ...X_{v_n})$ siempre tiene una distribución compatible con el siguiente grafo dirigido: $G_{comp} = (V, E_{comp})$, con $V = \{v_1, v_2, ..., v_n\}$ y $E_{comp} = \bigcup_{j=1}^{n} \{(v_i, v_j) : 1 \leq i < j\}$, pues el conjunto $pa_{G_{comp}}(v_j)$ es exactamente $\{v_i : 1 \leq i < j\}$, cumpliendo trivialmente la condición de compatibilidad. De aquí podemos deducir también que el grafo también puede contener aristas superfluas, por lo que no siempre se queda con las variables «influyentes», sino que desecha variables innecesarias a la hora de condicionar.*

Se puede ver que a partir de un DAG G (con n vértices), se puede construir un vector aleatorio n-dimensional con una distribución P compatible con G. Recíprocamente, a partir de un vector aleatorio con distribución P, se puede construir un DAG

G compatible con P, demostrado en la Tesis de Licenciatura de Laura Cacheiro [2]. La íntima relación entre grafos y distribuciones de probabilidad a veces es nucleada en un par (G_P, P) (es llamado una **red bayesiana**) donde G_P y P son compatibles.

2.3. Criterios para determinar independencia e independencia condicional entre variables

En la primer sección hemos visto la noción de camino (no necesariamente dirigido) que todavía no ha sido utilizada. En esta sección haremos uso de la mencionada definición.

Usaremos los posibles caminos en un grafo G para detectar independencias (posiblemente condicionales) entre variables de cualquier vector aleatorio con probabilidad P compatible con G.

2.3.1. D-separación en grafos

Definición 2.3.1 *Sea $G = (V, E)$ un grafo acíclico dirigido y p un camino (no necesariamente dirigido):*

- *Cadena: diremos que el camino p tiene una cadena con centro v_j si incluye la siguiente estructura:*
$$v_i \to v_j \to v_k$$
- *Tenedor: diremos que el camino p tiene un tenedor con centro en v_s si incluye la siguiente estructura:*
$$v_r \leftarrow v_s \to v_t$$
- *Colisionador: diremos que el camino p tiene un colisionador con centro en v_g si incluye la siguiente estructura:*
$$v_e \to v_g \leftarrow v_f$$

Observación 2.3.2 *Entre tres nodos «consecutivos» de un camino (siguiendo el orden dado por los vértices del mismo) estas son las tres estructuras que se pueden encontrar, por lo que analizar estas configuraciones nos permitirán sacar conclusiones sobre cualquier camino (que contenga al menos 2 aristas) en G.*

Observación 2.3.3 *En estas definiciones queda muy claro que el camino p no es necesariamente dirigido.*

Ahora veremos cuáles son las condiciones para que un conjunto de vértices «bloquee» un camino, que será importante a la hora de detectar independencias condicionales.

Definición 2.3.4 *Dados un DAG $G = (V, E)$, $Z = \{z_1, z_2, ..., z_k\} \subset V$ y un camino p, decimos que p está bloqueado por Z si se verifica alguna de las siguientes condiciones:*

- *p tiene una cadena o tenedor con centro en Z, es decir, $\exists z_j \in Z$ tal que:*

$$v_i \to z_j \to v_k \text{ o } v_i \leftarrow z_j \to v_k$$

- *p tiene un colisionador tal que ni él ni sus descendientes pertenecen a Z:*

$$v_e \to v_g \leftarrow v_f, \text{ con } v_g \notin Z \text{ y } de_G(v_g) \cap Z = \emptyset$$

Observación 2.3.5 *Notemos que en el caso $Z = \emptyset$, la noción de bloqueo es tan simple como ver que p tiene un colisionador, pues las otras condiciones o bien no se cumplen o no es necesario verificarlas.*

Ahora extenderemos la definición a subconjuntos de V y los posibles caminos entre sus nodos.

Definición 2.3.6 *(d-separación de conjuntos) Dados un DAG $G = (V, E)$, Z, T y W subconjuntos disjuntos (dos a dos) de V, decimos que Z d-separa a los conjuntos T y W, si Z bloquea **todos** los caminos p que unen vértices de W con los de T. La notación utilizada suele ser $(T \coprod W|Z)_G$ para simbolizar que T y W están d-separados por Z en el grafo G.*

Observación 2.3.7 *Nuevamente en el caso $Z = \emptyset$, para verificar d-separación entre W y T, basta ver que **todo** camino entre nodos de W y T tienen un colisionador. La notación para este caso es $(T \coprod W)_G$.*

Ejemplo 2.3.8 *En el ejemplo de la Figura 2.4, veremos que $(\{u_1\} \coprod \{u_2\})_G$ y $(\{u_1\} \coprod \{u_4\}|\{u_2, u_3\})_G$, aprovechando que los caminos a analizar son pocos.*

Para ver que $(\{u_1\} \coprod \{u_2\})_G$, tenemos que ver que los caminos de u_1 a u_2 tienen un colisionador. Tenemos las siguientes posibilidades:

- *$u_1 \to u_3 \leftarrow u_2$ tiene un colisionador en u_3.*

- *$u_1 \to u_3 \to u_4 \leftarrow u_2$ tiene un colisionador en u_4*

Como todos los caminos entre u_1 y u_2 están bloqueados, $\{u_1\}$ y $\{u_2\}$ están d-separados en G.

Para ver $(\{u_1\} \coprod \{u_4\}|\{u_2, u_3\})_G$, analizamos los siguientes caminos de u_1 a u_4, viendo si están bloqueados por $Z = \{u_2, u_3\}$.

- *$u_1 \to u_3 \to u_4$ tiene una cadena con centro en $u_3 \in Z$*

- *$u_1 \to u_3 \leftarrow u_2 \to u_4$ tiene un tenedor con centro en $u_2 \in Z$*

Entonces, $\{u_1\}$ y $\{u_4\}$ están d-separados por $\{u_2, u_3\}$ en G.

2.3.2. Independencias condicionales de vectores a partir de d-separación en un grafo

Hasta ahora vimos definiciones estrictamente relacionadas con grafos, pues la d-separación es un criterio gráfico que, valga la redundancia, se utiliza en grafos. Recordar que la definición de d-separación depende fuertemente de G y no se vale de ninguna variable aleatoria ni probabilidad. Veamos cómo se traducen estas propiedades en un marco probabilśitico. Como dijimos antes, la profunda relación entre ellas lleva a muchas confusiones, por eso nos parece importante mencionar repetitivamente en cuál de los ámbitos nos encontramos.

Supongamos que ahora tenemos un vector aleatorio $X = (X_v : v \in V)$ con distribución P compatible con un grafo dirigido $G = (V, E)$. Dados T, W y Z subconjuntos disjuntos de V, denotemos con X_T, X_W, X_Z a los subvectores de X dados por $X_T = (X_t, t \in T)$, $X_W = (X_w, w \in W)$, $X_Z = (X_z, z \in Z)$. La independencia condicional entre X_T y X_V dado X_Z será denotada por $(X_T \coprod X_W | X_Z)_P$. El siguiente resultado (disponible en [20] o [10]) permite establecer condiciones de independencia condicional estudiando conjuntos d-separados en el grafo. De esta forma, modelos probabilísticos pueden ser propuestos mediante DAG's.

Teorema 2.3.9 *Sea $G = (V, E)$ un DAG, con W, T y Z subconjuntos disjuntos de V. Sea $X = (X_{v_1}, X_{v_2}, ..., X_{v_n})$ un vector aleatorio. con distribución P compatible con G. Entonces*

$$(T \coprod W | Z)_G \Leftrightarrow (X_T \coprod X_W | X_Z)_P \text{ para toda } P \text{ compatible con } G.$$

Sea G el grafo de la Figura 4, sea $\widetilde{X} = (Z, U, X, Y)$ un vector aleatorio cuya distribución es compatible con G. Tendríamos que la variable Z está representada por u_1, U por u_2, X por u_3, mientras que Y está representada por u_4. Usando lo visto en el Ejemplo 2.3.8, el Teorema 2.3.9 nos permite asegurar lo siguiente:

$$(Z \coprod U)_P \tag{2.5}$$

$$(Z \coprod Y | X, U)_P$$

En resumen, los grafos nos dirán mucho sobre nuestras variables, la relación entre ellas y como detectar independencia.

Capítulo 3

Modelos Causales Funcionales

En el capítulo anterior hemos visto que a partir de un grafo podemos deducir qué variables de nuestro modelo podemos considerar relacionadas. Pero nunca se dijo nada de cómo se relacionan. Es decir, cuales son los **valores** que tomará una variable (por ejemplo, X_{v_j}) a partir de los valores que toman las variables representativas (el vector $\{X_{v_i} : v_i \in pa_G(v_j)\}$).

Para esto usaremos lo que se llaman **Modelos Causales Funcionales**. Aquel que quiera profundizar puede ver el Libro de Pearl [13]. Estos modelos asumen que cada una de las variables resulta ser una función determinística de otras variables combinadas con cierta perturbación. Es decir, cada variable factual X_j puede ser escrita como $f_j(PA_j, U_j)$, donde $PA_j \subset \{X_1, \cdots, X_{j-1}\}$, usando como **herramienta** las llamadas *ecuaciones estructurales*. A partir del conjunto de funciones f_j construiremos un grafo de forma tal que, **cuando las perturbaciones U_j son independientes**, la distribución del vector $(X_1, \cdots, X_n)$ resulta compatible con dicho grafo, permitiendo usar las herramientas del capítulo anterior. Por otra parte, este enfoque permite construir las variables contrafactuales modificando algunas de las funciones f_j. Esta propuesta causal será utilizada en el próximo capítulo, a la hora de encontrar cotas para ciertos parámetros causales.

3.1. Sistemas de ecuaciones estructurales

Sea $X = \{X_1, X_2, ..., X_n\}$ un conjunto de variables aleatorias y supongamos que cada variable X_j está determinada por:

- un conjunto de variables $PA_j \subseteq X \setminus \{X_j\}$
- una variable **aleatoria** U_j llamada perturbación (factores externos que puedan influir en el valor de X_j)

Bajo estas condiciones, tenemos que para todo j hay una función **determinística** f_j tal que

$$f_j(PA_j, U_j) = X_j.$$

A esta ecuación se la llama ecuación estructural, que da una forma de determinar el valor de X_j, usando las variables que influyen en ella. Notar que estando fijos los valores del vector PA_j y la variable U_j, tenemos un **valor fijo** para X_j, por eso enfatizamos que la función es determinística, ya que la aleatoriedad del sistema proviene sólo de las perturbaciones U_j. Al conjunto de estas ecuaciones (hay una para cada j) se le llama sistema de ecuaciones estructurales.

Definición 3.1.1 *Un modelo M que asume la existencia de ecuaciones estructurales para cada coordenada de un vector aleatorio $X = (X_1, X_2, ..., X_n)$ es llamado un modelo de ecuaciones estructurales para X (denotado SEM «Structural equation model»). El modelo M, sólo consta de las funciones f_j, las perturbaciones U_j e hipótesis conjuntas sobre las mismas.*

Observación 3.1.2 *Notar que las funciones f_j no tienen por qué ser conocidas. Del mismo modo, las variables X_j pueden no ser observadas, simplemente es un modelo que plantea la existencia de dichas funciones.*

Que las f_j no sean conocidas, no implica que uno no pueda asumir que dependan de ciertos parámetros. Por ejemplo, a veces se asume un modelo donde f_j es lineal en las variables que lo determinan, es decir:

$$f_j(PA_j, U_j) = \sum_{X_i \in PA_j} b_i.X_i + U_j$$

Los coeficientes lineales b_i son parámetros (pueden ser conocidos o no) que determinan la función f_j. Entonces, bajo esta suposición en la estructura de las funciones f_j, se tiene que éstas dependen de **finitos** parámetros.

Pero no siempre podemos asumir suposiciones que involucren la «forma» que tienen las funciones. Dando lugar a la siguiente definición:

Definición 3.1.3 *Cuando un SEM sólo asume la existencia de las funciones f_j, o no puede asumir que queden determinados a partir de finitos parámetros, se dice que tenemos un NPSEM («Non-parametric structural equations model»)*

Daremos a continuación unos ejemplos:

Ejemplo 3.1.4 *Un ejemplo simple es el de una fila de n dominós alineados en posición vertical (las fichas tienen altura h). Supongamos que U_j es la distancia entre la j-ésima ficha y la anterior ($2 \leq j \leq n$). Tomamos $X_j = 1$ si la j-ésima ficha se cae y $X_j = 0$*

en el caso contrario. Queremos ver cuales son las fichas que se caen si se impulsa la primer ficha dejándola a merced de la gravedad. Una vez que cae una ficha, la siguiente caerá si la distancia entre ellos es menor que h, en caso contrario no habrá impacto. Podemos decir también que si una ficha no cae, la siguiente no caerá pues no habrá quien la derrumbe.

Representamos gráficamente la situación en la figura 3.1:

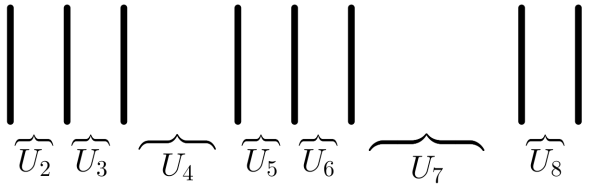


Figura 3.1: El gráfico consta de 8 dominós alineados, con distintas separaciones entre ellos

Planteamos entonces que el vector $(X_1, X_2, ..., X_n)$ *está sujeto al siguiente modelo de ecuaciones estructurales:*

- $f_1(U_1) \equiv 1$
- $f_j(X_{j-1}, U_j) = \left\{ \begin{array}{ll} 1 & si\ X_{j-1} = 1 \wedge U_j \leq h \\ 0 & caso\ contrario \end{array} \right\}$ *para todo* $2 \leq j \leq n$
- $\{U_1, U_2, ..., U_n\}$ *independientes*

En este caso las funciones no dependen de ningún parámetro desconocido. Al conocer el sistema que determina nuestras variables, podemos afirmar que no tenemos un NPSEM.

Este ejemplo intenta retratar que las ecuaciones estructurales representan los «mecanismos» por los cuales se generan los valores de nuestras variables, pues cada ecuación estructural detalla de qué modo los valores de ciertas variables «generan» los valores

de otra. En el ejemplo, fijados los valores de U_j, una vez que cae un dominó, la caída o no del resto está **determinada**.

Además, este ejemplo da una noción de **orden cronológico** entre las variables, es decir, todo el proceso está comandado por la caída o no del **primer** dominó, ya que si éste no cae, no caerán los **siguientes**. Esto no siempre se cumple, pero es de suma importancia cuando se pueda asumir un cierto orden.

Por último, da una idea de por qué «las perturbaciones» reciben ese nombre, pues representa toda la variabilidad que no puede ser explicada con las variables consideradas y pueda influir en la evolución del proceso. En este caso, una vez caído un dominó, **lo único** que puede impedir que el siguiente caiga es la distancia entre ellos.

Ejemplo 3.1.5 *Consideremos el ejemplo de ensayos clínicos presentado en el Capítulo 1. Recordemos que X, Y son variables binarias, donde X representa la acción que se aplica a cada individuo (tratamiento o control), mientras que Y denota la respuesta observada en el paciente. Podríamos incluir una variable W que detalla los factores que pueden influir tanto en la asignación del tratamiento como en la respuesta del paciente.*

Bajo estas suposiciones, podemos considerar un modelo en el que existen funciones f_X, f_Y, f_W y perturbaciones U_X, U_Y, U_W (donde U_X, U_Y, U_W son independientes) tal que:

- $W = f_W(U_W)$ *(esta función genera los valores de W)*
- $X = f_X(W, U_X)$ *(los valores de X están influidos por W y la perturbación correspondiente)*
- $Y = f_Y(W, X, U_Y)$ *(los valores de Y están influidos por W, X y la perturbación correspondiente)*

determinada

En este caso, como sólo asumimos la existencia de estas funciones, sin ninguna hipótesis sobre su estructura, tenemos un NPSEM.

Observación 3.1.6 *Notar que, siguiendo el orden de las ecuaciones dadas en el Ejemplo 3.1.5, si U_W, U_X y U_Y están fijos, las variables X, Y y W también pasan a tener valores fijos.*

El razonamiento que acabamos de hacer es un cabal ejemplo de la manera en la que queremos apropiarnos de los modelos estructurales a la hora de trabajar con causalidad. El presente abordaje establece que cada variable resulta ser una función determinística de otras, combinadas con una perturbación aleatoria. De esta manera, conocer la naturaleza del sistema con el que se trabaja permite plasmar qué variables se encuentran relacionadas, estableciendo cuales son la necesarias a la hora de determinar el **valor** de

una de ellas. Es decir, por más que no podamos precisar cuál es la forma de f_j podremos establecer de qué variables esta función depende.

Tanto en este capítulo como el anterior hemos hablado de «influencia» de una variable sobre otra. En la próxima sección relacionaremos ambos capítulos, pues obtendremos grafos a partir de un SEM.

3.2. Diagramas causales

Supongamos que tenemos un vector $X = (X_1, X_2, ..., X_n)$ cumpliendo un sistema de ecuaciones estructurales:

$$f_j(PA_j, U_j) = X_j, \forall 1 \leq j \leq n.$$

Teniendo en cuenta este modelo se construye un grafo $G = (V, E)$ donde:

- $V = \{v_1, v_2, ..., v_n\}$ y cada v_j representa a la variable X_j.
- Se agrega una arista dirigida (v_i, v_j) si la variable X_i está en PA_j (es una de las variables determinantes)
- Se agrega una arista bidirigida punteada (v_j, v_k) si las variables U_j y U_k no son independientes.

Definición 3.2.1 *Al grafo construido se le llama un diagrama causal de un SEM. Si el mismo resulta ser un DAG, se dice que tenemos un DAG causal.*

Observación 3.2.2 *El hecho de que el diagrama causal sea un DAG nos da la siguiente información sobre el modelo:*

- *Las perturbaciones $\{U_1, U_2, ..., U_n\}$ son independientes, pues en caso contrario habría un camino dirigido de un nodo a sí mismo (por la arista bidirigida), contradiciendo que sea acíclico.*
- *Como todo DAG se puede numerar de forma que $pa_G(v_j) \subseteq \{v_i : 1 \leq i < j\}$ (numeración topológica), podemos reordenar nuestras variables de forma que podamos considerar un orden cronológico entre las mismas. Es decir, podemos asumir que observamos una variable primero, luego otra, después la siguiente y así sucesivamente hasta que se generen los valores de todas las variables, siempre respetando el orden, a diferencia del problema del huevo y la gallina.*
- *Siguiendo el orden cronológico, podemos decir que fijados los valores de U_j, los valores de las variables X_j están **determinados**.*

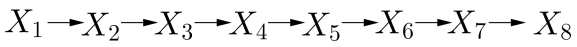


Figura 3.2: Diagrama causal correspondiente al modelo del Ejemplo 3.1.4

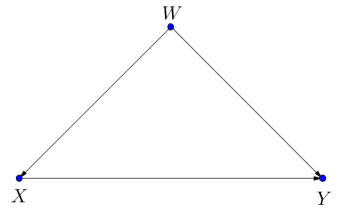


Figura 3.3: Diagrama causal correspondiente al modelo del Ejemplo 3.1.5

De los Ejemplos 3.1.4 y 3.1.5 podemos deducir los diagramas causales representados en las Figuras 3.2 y 3.3 respectivamente:

Notemos que en las Figuras 3.2 y 3.3, pese a las advertencias hechas en las secciones anteriores, estamos utilizando las mismas letras para denotar variables aleatorias y nodos del grafo. En adelante, habiendo explicado exhaustivamente la diferencia entre ambos, incurriremos en este abuso, con la siguiente salvedad: las letras utilizadas en los nodos del grafo estarán asociadas a mediciones, resultados asociados a ciertos experimentos, representados en las **variables aleatorias**. Pero cuando aparecen en el grafo son **nodos**.

Resumiendo lo dicho hasta el momento, enfaticemos el siguiente hecho:

Observación 3.2.3 *Un modelo de ecuaciones estructurales no paramétricas con per-*

turbaciones $\{U_1, U_2, ..., U_n\}$ *independientes, tiene asociado un DAG causal:*

$$M = \left\{ \begin{array}{l} \{U_1, U_2, ..., U_n\} \text{ independientes} \\ \{f_i : v_i \in V\} \text{ compatible con } G \end{array} \right\} \Leftrightarrow G = (V, E)$$

Cuando decimos que las funciones f_i *son «compatibles con* G*», nos referimos a que* G *es el diagrama causal correspondiente a* M*.*

El siguiente resultado muestra que, ante un vector aleatorio $\widetilde{X}$ satisfaciendo un sistema de ecuaciones estructurales M cuyo diagrama causal resulte un DAG G, podemos aplicar los métodos gráficos presentados en el Capítulo 2 para estudiar independencias o independencias condicionales entre sus coordenadas.

Lema 3.2.4 *Sea* $\widetilde{X} = (X_1, X_2, ...X_n)$ *un vector de variables aleatorias satisfaciendo un sistema de ecuaciones estructurales cuyo diagrama causal resulta un grafo acíclico dirigido* G*. Tenemos entonces que la distribución de* $\widetilde{X}$ *resulta compatible con* G*.*

Demostración 3.2.5 *Sea* $\vec{x} = (x_1, x_2, ..., x_n)$ *un valor para* $\widetilde{X}$ *y, para cada* X_j $(1 \leq j \leq n)$ *tenemos* $PA_j \subset \{X_1, X_2, ..., X_{j-1}\}$ *de forma que* $X_j = f_j(PA_j, U_j)$*. Denotemos por*
$pa_j = (x_i : i \in PA_j)$ *el valor para* PA_j *cuando* $\widetilde{X} = \vec{x}$*. Además sabemos que* $\{U_1, U_2, ..., U_n\}$ *son independientes entre sí y* $U_j \coprod X_i$*,* $\forall i < j$*, en particular,* $U_j \coprod PA_j$*.*

Para demostrar el resultado, por la equivalencia dada en el Lema 2.2.4, basta ver que la distribución de X *verifica la factorización de Markov detallada en (2.4):*

$$\begin{aligned} P\left(\bigcap_{j=1}^{n} X_j = x_j\right) &= P\left(\bigcap_{j=1}^{n} f_j(PA_j, U_j) = x_j\right) \\ &= P\left(\bigcap_{j=1}^{n} f_j(pa_j, U_j) = x_j\right) \\ &= \prod_{j=1}^{n} P(f_j(pa_j, U_j) = x_j) \\ &= \prod_{j=1}^{n} P(f_j(pa_j, U_j) = x_j | PA_j = pa_j) \\ &= \prod_{j=1}^{n} P(X_j = x_j | PA_j = pa_j) \end{aligned}$$

En adelante, trabajaremos con modelos de ecuaciones estructurales asociados a DAG's causales, por lo que se asumirá que las perturbaciones U_j son independientes y que toda variable que pueda afectar a dos de las presentes, forma parte del sistema.

Trabajo interdisciplinario

En muchas ocasiones nos hemos referido al «investigador», pero nunca aclaramos mucho al respecto. El investigador es un experto de otra disciplina que nos plantea un problema para resolver, pues necesita ayuda en el desarrollo matemático del mismo. Muchas veces, el diálogo entre ambos profesionales se ve afectado por la diferencia entre sus formaciones, ya sea por los términos que usan (en el momento que un matemático dice «Markov», el resto de la sala hace silencio y esboza expresiones faciales semejantes a la repugnancia), formas de pensar el problema y otras inconsistencias.

Pero notemos que, ya sea plantear ecuaciones estructurales o el diagrama causal correspondiente, los modelos que hemos detallado resultan de absoluta simpleza, pues cualquier persona que disponga de un lápiz y sepa dibujar flechas (conscientes de que la flecha representa que una variable «causa» a la otra) o que esté familiarizado la definición de una función puede entender de qué se trata y además, participar en variaciones del modelo en cuestión, ya que su conocimiento sobre el problema nos será de mucha utilidad para concretar un modelo de mayor precisión.

3.3. Modelos intervenidos y variables contrafactuales

Daremos un ejemplo (legalmente incorrecto, por eso pedimos a los abogados que se abstengan de comentarios) que introduce la noción de modelos intervenidos.

Supongamos que se lleva a cabo una elección bipartidista (o un ballotage si se quiere, es para simplificar) entre el partido A y el partido B (claro que la existencia de un partido con una sola sigla resulta algo cuasi hilarante y de una extrema pereza de parte de los mismos partidos, pero a modo de ejemplo sirve). En una cierta mesa de votación, llega una cantidad X de votantes decididos (o no, como sabemos, casi nunca es algo seguro) a emitir sufragio. De esas personas, hay una cantidad Y que vota por el partido A (o la cantidad de votos a favor de A que terminan en la urna), claramente afectada por la cantidad de votantes (al menos, tendrá que ser menor o igual). Por último, a la hora del escrutinio, el presidente de mesa cuenta una cantidad Z de votos para el partido A, como la autoridad de mesa ha sido abandonado por sus compañeros de mesa, luego de más de 10 horas de obligaciones (como ser humano) puede equivocarse al realizar la suma. Como también tiene deseos imperiosos de volver a su hogar, ni se fija si la cantidad de votos que contó coincide con la cantidad de votantes.

Para este caso tenemos un modelo en el que

$$\begin{aligned} X &= f_X(U_X) \\ Y &= f_Y(X, U_Y) \\ Z &= f_Z(Y, U_Z) \end{aligned}$$

son las ecuaciones estructurales y el DAG asociado es:

$$X \to Y \to Z$$

Tenemos entonces que el modelo M al que están sujetas las variables **factuales** está compuesto por las funciones $\{f_X, f_Y, f_Z\}$ junto con las perturbaciones $\{U_X, U_Y, U_Z\}$.

Consideremos un escenario hipotético (contrafactual) en el que, aprovechando el receso del presidente de mesa para visitar el sanitario, un enviado por el partido A vacía la urna y la llena con a votos a favor del partido A.

Es decir, se produce una intervención mediante la cual se establece que la urna contiene a votos en favor del candidato A. Si bien esto no modifica la cantidad de votantes que llega a la mesa ni la forma en la que el presidente de mesa cuenta los votos, determina que en la urna van a encontrarse a votos en favor del candidato A. Utilizaremos X_a, Y_a y Z_a para denotar a las variables contrafactuales correspondientes a este nuevo escenario, construídas a partir del siguiente sistema de ecuaciones:

$$\begin{aligned} X_a &= f_X(U_X) \\ Y_a &= a \\ Z_a &= f_Z(Y_a, U_Z) \end{aligned}$$

Estas nuevas ecuaciones estructurales se representan gráficamente en el siguiente DAG:

$$X \quad Y \quad \to \quad Z$$

Observación 3.3.1 *Podemos sacar las siguientes conclusiones:*

- *Cambiar el contenido de la urna no influye en la cantidad de votantes, por eso es que la función que determina la cantidad de votantes en este nuevo escenario sigue siendo f_X. En particular, tenemos que $X_a = X$.*
- *Modificar el contenido se la urna no cambia la manera en que el presidente de mesa procede a la hora de contar los votos, él (con todos los errores que pueda tener), en definitiva, cuenta. Por eso no cambia la función f_Z, pero sí cambia la variable que recibe dicha función.*
- *Claramente, cambia la función que determina cuantos votos hay en favor del candidato A en la urna, razón por la cual la nueva función que utilizamos para determinar la variable asociada al nodo Y esta dada por la constante a.*

Podemos pensar entonces que el modelo intervenido M_a que da origen a las variables X_a, Y_a, Z_a, queda determinado por las funciones $\{f_X^a, f_Y^a, f_Z^a\}$ junto con las perturbaciones $\{U_X, U_Y, U_Z\}$, siendo

$$f_X^a = f_X \ , f_Y^a = a \ , f_Z^a = f_Z.$$

Luego de este ejemplo, pasaremos a la definición general de modelo intervenido. La idea es dejar que todo evolucione según la dinámica factual, salvo en lo que respecta a los nodos donde intervenimos. Las funciones asociadas a tales nodos estarán dadas por el valor de la constante con la que queremos intervenir. Este nuevo sistema de ecuaciones dará origen a las variables contrafactuales asociadas a la intervención fijada.

Definición 3.3.2 *Dado el siguiente SEM para un vector aleatorio* $X = (X_1, X_2, ..., X_n)$*:*

$$M = \left\{ \begin{array}{c} \{U_1, U_2, ..., U_n\} \text{ independientes} \\ \{f_i : v_i \in V\} \text{ compatible con } G \end{array} \right\}$$

Donde G *es el diagrama causal asociado a* M*.*

Además, sea A *un subconjunto de variables* ($A = \{X_{i_l} / 1 \leq l \leq k\} \subseteq X$) *y* $\overline{a} = (a_{i_1}, a_{i_2}, ...a_{i_k})$ *un posible valor para el vector* A*. Definimos un nuevo modelo (llamado modelo intervenido)* $M_{\overline{a}}$*, donde:*

- *Las perturbaciones* $\{U_1, U_2, ...U_n\}$ *coinciden con las de* M
- *Las nuevas funciones (serán llamadas* $\{f_i^{\overline{a}} : 1 \leq i \leq n\}$ *se modifican de la siguiente forma:*
  - $f_j^{\overline{a}}(PA_j, U_j) := f_j(PA_j, U_j)$ *si* $X_j \notin \{X_{i_1}, X_{i_2}, ..., X_{i_k}\}$
  - $f_j^{\overline{a}}(PA_j, U_j) := a_{i_l}$ *si* $\exists l$ *tal que* $j = i_l$

Observación 3.3.3 *Como dijimos, cada NPSEM tiene su respectivo DAG, entonces si* M *es un modelo para* X *y* $M_{\overline{a}}$ *es el modelo intervenido en las variables* $A \subseteq X$*, tenemos las siguientes correspondencias entre grafos y sistemas de ecuaciones estructurales:*

$$M \Leftrightarrow G = (V, E) \text{ y } M_{\overline{a}} \Leftrightarrow G_{\overline{A}} = (V, E_{\overline{A}})$$

donde $E_{\overline{A}}$ *es el conjunto resultante de remover todas las flechas que llegan a nodos de* A*. Esto es intuitivo porque una vez que forzamos a las variables de* A *a tomar ciertos valores, ya tenemos que las variables de* A *son determinísticas, sin depender de otras variables, por lo que no tendrá sentido establecer relaciones entre las mismas.*

Es decir, tenemos:

$$M_{\overline{a}} = \left\{ \begin{array}{c} \{U_1, U_2, ..., U_n\} \text{ independientes} \\ \{f_i^{\overline{a}} : v_i \in V\} \text{ compatible con } G_{\overline{A}} \end{array} \right\}$$

Estas ideas de modelos intervenidos nos ayudarán a estudiar el comportamiento de algo que parece inalcanzable como las variables contrafactuales.

Recordemos el ejemplo del primer capítulo, donde teníamos la variable aleatoria Y que detallaba la respuesta de un individuo a la acción dada por X. A partir de

estas variables factuales, construimos la variable contrafactual Y_{x_1}, que representaba la potencial respuesta de un individuo en el caso de que éste fuera tratado, aunque no necesariamente haya sido sometido a esa acción. Pero hemos mencionado la gran cantidad de inconvenientes que presentan estas variables contrafactuales pues, ante todo, no son observadas en su totalidad. Sin embargo, asumiendo un Modelo Causal Funcional, si tenemos que $Y = f_Y(X, U_Y)$ donde X es el tratamiento asignado por los investigadores, el «mecanismo» por el cual se genera Y a partir de X no hubiera cambiado si se hubiera tratado a toda la población, **pues el origen de la aleatoriedad dada por U_Y no se ve afectada**. Es decir, $Y_{x_1} = f_Y(x_1, U_Y)$ pues al fin y al cabo estoy forzando a toda la población a recibir tratamiento x_1 sin alterar U_Y. Del mismo modo, tenemos que la variable contrafactual Y_{x_0} responde a la ecuación estructural dada por $Y_{x_0} = f_Y(x_0, U_Y)$.

Notemos que de estas últimas expresiones para las variables contrafactuales **se deduce** la hipótesis de consistencia, por lo que no parece justificado el énfasis que se le dio en el Capítulo 1. Por esto nos parece importante remarcar la razón.

Sin decirlo, en el primer capítulo presentamos el llamado **Modelo Contrafactual de Rubin** (ver el trabajo de Rubin [17]), que asume la existencia de las variables factuales, las variables contrafactuales e hipótesis conjuntas entre ellas. Por ejemplo, en el Capítulo 1 hemos considerado las variables X, Y y C como variables factuales, Y_{x_0}, Y_{x_1} como variables contrafactuales. Luego, hemos planteados distintas hipótesis para los distintos problemas, como la hipótesis de consistencia, la de aleatorización completa y aleatorización condicional.

En cambio, ahora la hipótesis de consistencia (por ejemplo) se deduce porque hemos asumido un **Modelo Causal Funcional**, que a diferencia del Modelo Contrafactual de Rubin asume la existencia de un sistema de ecuaciones estructurales, en los cuales **las perturbaciones son independientes**. Esta suposición es más fuerte que la de consistencia y es por eso que en un Modelo Causal Funcional la consistencia es una propiedad y no una hipótesis.

Estos son dos posibles formas muy distintas de abordar un modelo causal, nosotros asumiremos un modelo causal funcional para los problemas del próximo capítulo.

Capítulo 4

Cotas generales para problemas de incumplimiento parcial del tratamiento

En este capítulo, para cada problema que será considerado, asumiremos Modelos Causales Funcionales. Deduciremos a partir de ellos los diagramas causales pertinentes y construiremos las variables contrafactuales con la metodología propuesta en el capítulo anterior.

Como mencionamos en el Capítulo 1, cuando el parámetro causal de interés no es identificable a partir de la distribución de las variables observadas podemos buscar cotas para el mismo. Balke y Pearl [1] consideran este problema en estudios experimentales donde el tratamiento es asignado de manera completamente aleatoria pero el cumplimiento por parte de los pacientes del mismo no es perfecto. En tal caso, el efecto real del tratamiento **asignado** puede diferir de la diferencia correspondiente a las medias de las respuestas entre los tratamientos **recibidos**. En este contexto, los autores encuentran cotas óptimas para el parámetro causal de interés, asumiendo que la variable respuesta es binaria. Los valores de estas cotas pueden asegurar la positividad o negatividad del ATE. Los autores asumen un modelo causal funcional, para construir en ese contexto las variables contrafactuales mediante el modelo intervenido. El objetivo de esta tesis es extender los resultados obtenidos por Balke et. al. para el caso en que la variable respuesta es discreta de rango finito. También se obtienen cotas para otros posibles funcionales asociados a la distribución de variables contrafactuales.

Si bien no obtenemos una fórmula explícita para las cotas (como sí lo hacen Balke y Pearl) mostramos que éstas quedan determinadas por la solución de un problema de programación lineal. Hemos implementado en lenguage R un algoritmo que, tomando como parámetro la distribución de las variables observadas, resuelve el problema de optimización, devolviendo el valor de las cotas. Además, Balke y Pearl hallan cotas para el ATE, mientras que nosotros encontramos cotas finas para otros parámetros

causales.

Desde el punto de vista estadístico, la distribución de las variables observadas es estimada empíricamente a partir de una muestra y las cotas se estiman resolviendo el problema de programación lineal asociado a la distribución empírica. Luego usamos propiedades de programación lineal para garantizar la consistencia de este procedimiento estadístico.

Para comodidad del usuario, está a su disposición la función que calcula las cotas partiendo de un conjunto de datos.

El Capítulo se organiza de la siguiente manera: en la primer sección, hemos decidido adaptar la metodología implementada por Balke et. al a un *toy example*, donde las cotas pueden determinarse mediante aritmética elemental. Hemos seguido este camino entendiendo que la simpleza de este primer problema permite enfatizar en los conceptos propuestos para encontrar cotas en problemas más complejos. Cabe mencionar que este mismo caso ha sido tratado en la Sección 1.5 del Capítulo 1.

En la segunda sección, siguiendo el trabajo de Balke et. al. , detallamos el problema de incumplimiento parcial del tratamiento, en el caso de una variable respuesta binaria. Presentamos el problema de programación lineal que devuelve cotas óptimas para el parámetro causal y en este contexto se encuentran fórmulas explícitas para dichas cotas.

En la tercer sección, presentamos el principal aporte original de este trabajo, extendiendo los resultados de Balke et.al. al caso en el que la variable respuesta toma una cantidad finita de valores.

Para finalizar, en la última sección damos una descripción del estimador, demostramos su consistencia y dedicamos varias subsecciones a las simulaciones que fueron realizadas.

4.1. Casos simples para introducir el problema

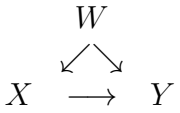
Recordemos nuevamente el estudio experimental en el que se asignaba tratamiento (dada por la variable X) a un grupo de pacientes y se observaba la respuesta de cada uno (representada por la variable Y). En el Capítulo 1 (ver la Sección 1.5) hemos demostrado que el efecto medio de tratamiento (dado por $ATE = E[Y_{x_1} - Y_{x_0}]$) queda identificado por la distribución de las variables (X, Y) bajo el supuesto de aleatorización completa del tratamiento. Asumiendo un modelo funcional causal, tenemos el siguiente DAG causal:

$$X \to Y$$

¿Qué pasa si esta hipótesis es muy fuerte? Cuando la aleatorización *completa* no puede ser asumida, pueden existir factores influyendo sobre los resultados de X e Y que no estamos teniendo en cuenta. En este caso, tendríamos que considerar la existencia de una variable W (que puede ser un vector, puede ser de naturaleza continua o discreta,

etc.) que tiene en cuenta **todos** los factores que influyen sobre el tratamiento X y la respuesta Y.

Notar que esta situación es idéntica a la dada en el Ejemplo 3.1.5, dando lugar al mismo Modelo Causal Funcional, que tiene el siguiente diagrama causal:



Si la variable W es observada, estamos en un caso similar al de la primera sección, en el cual medíamos una nueva variable C para lograr identificabilidad (ver la Sección 1.5). Sólo que ahora en vez de considerar la variable C, estamos considerando W. La construcción de variables contrafactuales bajo el modelo causal funcional permite demostrar que Y_{x_i} resulta condicionalmente independiente de X dado W y por consiguiente, como demostramos en la Sección 1.5, el parámetro causal es identificable.

Pero claro que la vida no es tan linda y la hipótesis de observar W más que fuerte es herculeana...

Al no observar W ya no tenemos identificabilidad (lo hemos visto en la Sección 1.5). Ante este percance, procuraremos dar cotas óptimas para el parámetro causal.

Asumiendo que nuestras variables satisfacen un sistema de ecuaciones estructurales, al igual que en el Ejemplo 3.1.5, tenemos que existen funciones f_X,f_Y y respectivas perturbaciones U_X, U_Y tal que

$$X = f_X(W, U_X)$$
$$Y = f_Y(X, W, U_Y)$$

Como estamos incluyendo en W a **todos** los factores que influyen sobre X e Y, ya habiendo perdido esperanzas de identificar nuestro parámetro causal, no será de mayor utilidad contar las perturbaciones U_X, U_Y. Por lo que uno puede incluir en W a dichas perturbaciones, resultando así:

$$X = f_X(W)$$
$$Y = f_Y(X, W)$$

simplificando un poco las ecuaciones anteriores.

Lo ingenioso de los trabajos de Pearl [1] es que en vez de tener en cuenta toda la información proveída por W, se fija para cada w cuál de las posibles funciones $\{g \,/\, g : \{x_0, x_1\} \to \{0, 1\}\}$ coincide con $f_Y(\cdot, w)$. Una vez conocida dicha función, conociendo X podemos determinar el valor de Y, para cada w valor de W.

Además, con lo visto de ecuaciones estructurales y modelos intervenidos, también (siempre fijo $W = w$) podremos determinar las variables contrafactuales $Y_{x_i}, i \in \{0, 1\}$.

Con estos fines creamos $R_Y = R_Y(W)$ tal que la distribución de (X, R_Y) determina no sólo la distribución de las variables observadas (X, Y), además permitirá obtener la

distribución de las variables contrafactuales Y_{x_0}, Y_{x_1}. Con estas distribuciones podemos calcular nuestro parámetro causal. Pese a que (X, R_Y) no es enteramente observado, vamos a decir que resulta suficiente para determinar el parámetro causal.

Pasemos ahora a la construcción de R_Y. Sea w una realización de W, y fijemos $W = w$. Sabemos que $f_Y(\cdot, w)$ es una función que tiene como dominio el conjunto $\{x_0; x_1\}$ y respectivo codominio $\{0; 1\}$. Es decir:

$$f_Y(\cdot, w) : \{x_0; x_1\} \longmapsto \{0; 1\}$$

Concluimos que tenemos 4 posibilidades para $f_Y(\cdot, w)$:

$$\begin{array}{l} g_0(x_0) = 0 \ ; \ g_0(x_1) = 0 \\ g_1(x_0) = 0 \ ; \ g_1(x_1) = 1 \\ g_2(x_0) = 1 \ ; \ g_2(x_1) = 0 \\ g_3(x_0) = 1 \ ; \ g_3(x_1) = 1 \end{array}$$

Entonces tenemos que para cada w, $\exists R_Y = R_Y(w) \in \{0; 1; 2; 3\}$ tal que

$$f_Y(\cdot, w) = g_{R_Y}(\cdot)$$

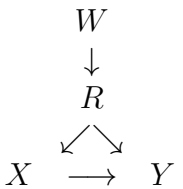
Componiendo con la variable W, tenemos $R_Y = R_Y(W)$ que cumple

$$f_Y(x_0, W) = g_{R_Y(W)}(x_0) = g_{R_Y}(x_0)$$

$$f_Y(x_1, W) = g_{R_Y(W)}(x_1) = g_{R_Y}(x_1)$$

Observación 4.1.1 *Notar que, a fines de simplificar la notación, ocasionalmente omitimos enfatizar que R_Y es una función de W. Además querremos independizarnos lo más posible de la variable W. También notemos que R_Y es discreta, tomando valores en $\{0, 1, 2, 3\}$*

Tomando $R = (X, R_Y)$, consideraremos el vector (W, R, X, Y), asociado al siguiente DAG:



Estamos considerando también nuevas ecuaciones estructurales $f_X^*(R) = X$ y $f_Y^*(R, X) = g_{R_Y}(X)$, mientras que $f_R^*(W) = (f_X(W), R_Y(W))$. La evolución de las variables bajo este nuevo modelo coincide con el proceso original. Con esto nos referimos

	x_0	x_1
g_0	0	0
g_1	0	1
g_2	1	0
g_3	1	1

Cuadro 4.1: Posibles funciones de $\{x_0, x_1\}$ a $\{0, 1\}$

a que mismos valores de W en ambos modelos gráficos tendrán como resultado los mismos valores para X e Y, al igual que Y_{x_i}, $i \in \{0, 1\}$.

En el Cuadro 4.1 exhibimos una tabla que facilitará el proceso de identificar cuales son los valores que puede toma R_Y a partir de los valores de X e Y.

Por ejemplo, si sabemos que $X = x_0$ y $Y = 0$ tendremos que los valores para R_Y pueden ser 0 o 1, pues g_0, g_1 son las únicas funciones que mandan x_0 a 0. Si sabemos que $X = x_1$ y $Y = 1$, $R_Y \in \{1, 3\}$ por análogas razones.

Esto servirá para igualar probabilidades que involucran a X e Y con probabilidades que involucran a X e R_Y.

Por ejemplo, usando los ejemplos presentados, tenemos que

$$P(X = x_0, Y = 0) = P(X = x_0, R_Y = 0 \cup X = x_0, R_Y = 1).$$

Usando que la unión es claramente disjunta (R_Y toma distintos valores) podemos aplicar la aditividad para decir que

$$P(X = x_0, Y = 0) = P(X = x_0, R_Y = 0) + P(X = x_0, R_Y = 1).$$

En la mayoría de las demostraciones usaremos que las uniones son disjuntas para poder usar la aditividad.

Denotemos por $r_{ij} = P_R(x_i, j) = P(X = x_i, R_Y = j)$ la probabilidad puntual de R, con $i \in \{0; 1\}$ y $j \in \{0; 1; 2, 3\}$. La ventaja de esta «factorización» del grafo es que tanto la distribución de (X, Y) como la de las variables contrafactuales (Y_{x_0}, Y_{x_1}) se pueden obtener a partir de la probabilidad puntual de R (que ahora resulta un vector discreto con 8 posibles valores). Ya no nos es de importancia la naturaleza de W. Demostraremos esto en el siguiente lema.

Lema 4.1.2 *Ambas distribuciones de las variables contrafactuales y la distribución conjunta de (X, Y) puede obtenerse a partir de la distribución de R. Más aún, se consiguen fórmulas explícitas para cada distribución como función de las r_{ij} con $i \in \{0; 1\}$ y $j \in \{0; 1; 2, 3\}$*

Demostración 4.1.3 *Veremos que se cumple para algunos posibles valores, ya que esta sección es para introducir el problema, más adelante nos interesará las fórmulas explícitas para todas las probabilidades puntuales.*

Por ejemplo,

$$\begin{aligned}
p_{(X,Y)}(x_0,0) &= P(X=x_0, Y=0)\\
&= P(X=x_0, f_Y^*(X,R_Y)=0)\\
&= P(X=x_0, f_Y^*(x_0,R_Y)=0)\\
&= P(X=x_0, g_{R_Y}(x_0)=0)\\
&= P\left(\bigcup_{0\le j\le 3: g_j(x_0)=0} \{X=x_0, R_Y=j\}\right)\\
&= P(\{X=x_0,R_Y=0\}\bigcup\{X=x_0,R_Y=1\})\\
&= P(X=x_0,R_Y=0)+P(X=x_0,R_Y=1)\\
&= r_{00}+r_{01}
\end{aligned}$$

De forma similar se determinan el resto de las probabilidades puntuales de (X,Y). *Recordando que* $Y_{x_0} = f_Y^*(x_0,W) = g_{R_Y(W)}(x_0)$, *tenemos que*

$$Y_{x_0}=1 \Leftrightarrow g_{R_Y}(x_0)=1 \Leftrightarrow R_Y\in\{2,3\}$$

Entonces,

$$P(Y_{x_0}=1)=P(R_Y\in\{2,3\})=P(R_Y=2)+P(R_Y=3)$$

El término de la izquierda (refiriéndonos a $P(R_Y=2)$*) se puede expresar, por probabilidad total, como*

$$P(R_Y=2)=P(X=x_0,R_Y=2)+P(X=x_1,R_Y=2)=r_{02}+r_{12}$$

De la misma forma tenemos que

$$P(R_Y=3)=P(X=x_0,R_Y=3)+P(X=x_1,R_Y=3)=r_{03}+r_{13}.$$

Resultando así: $P(Y_{x_0}=1)=r_{02}+r_{12}+r_{03}+r_{13}=T_0(r)$
Análogamente: $P(Y_{x_1}=1)=r_{01}+r_{11}+r_{03}+r_{13}=T_1(r)$

Observación 4.1.4 *Llamando* $r=(r_{00},r_{01},r_{02},r_{03},r_{10},r_{11},r_{12},r_{13})\in\mathcal{S}_8$, *y* $p=p_{(X,Y)}$ *(visto como vector de* $\mathcal{S}_4$*), tenemos que* $\exists T_{mg}:\mathcal{S}_8\to\mathcal{S}_4$ *y* $\exists T_{ATE}:\mathcal{S}_8\to\mathbb{R}$ *(ambos* ***operadores lineales****) tal que:*

$$T_{mg}(r)=p$$

$$T_{ATE}(r)=T_1(r)-T_0(r)=P(Y_{x_1}=1)-P(Y_{x_0}=1)=ATE$$

Notemos que existe una correspondencia biunívoca entre el vector $R = (X, R_Y)$ y el vector (X, Y_{x_0}, Y_{x_1}) . El lector puede comprobar que, escencialmente, hemos reescrito los resultados obtenidos en la Sección 1.5 con mínimos cambios en la notación (hemos cambiado a $r_i,\ 1 \leq i \leq 8$ por $r_{ij},\ 1 \leq i \leq 2,\ 1 \leq j \leq 4$), donde probamos la falta de identificabilidad y encontramos cotas óptimas para el parámetro causal de interés, utlizando ahora la variable R_Y. Esta idea será de suma utilidad en los próximos ejemplos.

4.2. Caso binario del problema de incumplimiento parcial del tratamiento

Supongamos que tenemos un tratamiento cuya «bondad» queremos analizar. Entonces tendremos una población A a la que le asignaremos tratamiento y control de forma completamente aleatoria. Por diferentes motivos (ya sea un carácter olvidadizo, desconfianza en el respectivo profesional, la clásica «mi mamá y/o novia no me deja», etc.) los individuos de la población pueden no seguir la indicación médica recibida (ya sea tratamiento o control). En este caso, medir el efecto sobre la respuesta (positiva o negativa) del tratamiento **asignado** no será un método empleado por alguien en su sano juicio. Convendrá entonces analizar el efecto del tratamiento que realmente **recibió** la población A.

Ahora tendremos X, Y, Z variables observadas. X y Z serán binarias y, para seguir introduciendo el problema que resolveremos y su respectiva resolución, consideraremos Y también binaria (después lo extenderemos para finitos valores de Y). Donde $Z(a)$ es la acción **asignada** a la persona a, $X(a)$ es la acción que **recibió** la persona a y por último, $Y(a)$ será la respuesta observada en el paciente. Todo esto bajo la presencia de U, que representa todos los factores externos que pueden influir sobre X e Y. Notar que como el tratamiento Z es asignado de forma **completamente** aleatoria, podemos asumir que U no tendrá influencia sobre Z. Tendremos los siguientes valores para las variables:

- $Z(a) = z_1$ si se le asigna tratamiento a la persona a.
- $Z(a) = z_0$ si se le asigna control a la persona a.
- $X(a) = x_1$ si la persona a recibe tratamiento.
- $X(a) = x_0$ si la persona a no recibe tratamiento.
- $Y(a) = 1$ si se observa respuesta positiva en la persona a.
- $Y(a) = 0$ si se observa respuesta negativa en la persona a.

El siguiente DAG representa la situación que estamos considerando:

$$\begin{array}{ccccc} & & U & & \\ & & \swarrow\ \searrow & & \\ Z & \rightarrow & X & \rightarrow & Y \end{array} \tag{4.1}$$

Notar que coincide con el DAG dado en la Figura 2.4.

Ya que queremos evaluar el efecto medio del tratamiento **recibido**, el parámetro causal de interés será:

$$ATE = E(Y_{x_1}) - E(Y_{x_0})$$

que como hemos visto, al ser Y binaria, coincide con

$$P(Y_{x_1} = 1) - P(Y_{x_0} = 1)$$

Nuevamente, nuestro parámetro causal no es identificable a partir de las variables observadas (la presencia de U arruina la identificabilidad). En esta sección reproducimos los resultados obtenidos por Balke y Pearl [1], quienes proponen un algoritmo para encontrar cotas óptimas para el ATE a partir de la distribución de (Z, X, Y).

A partir del DAG planteado, nuestro modelo asume la existencia de dos funciones f_X, f_Y tal que:

$$\begin{array}{l} X = f_X(Z, U) \\ Y = f_Y(X, U) \end{array} \tag{4.2}$$

De nuevo hemos incluido en U las perturbaciones U_X, U_Y, para simplificar la notación. Además, conscientes de que el DAG coincide con el de la Figura 2.4, hemos visto independencias a partir de caminos en dicho grafo en el Ejemplo 2.3.8, deduciendo que $Z \coprod U$ (ver (2.5)).

Veremos que entonces la distribución de Z y U es suficiente para determinar la distribución de las variables observadas y las variables contrafactuales. Al igual que en la sección anterior, la complejidad de U puede ser simplificada tomando una «versión discreta» teniendo en cuenta sólo las influencias de U sobre los valores de X e Y.

Construiremos entonces $\widetilde{R} = (R_X, R_Y) = (R_X(U), R_Y(U))$ de manera similar a la sección anterior. Fijando $U = u$ sabemos que $f_X(\cdot, u)$ es alguna función que parte de $\{z_0, z_1\}$ y llega a $\{x_0, x_1\}$. Caractericemos las 4 posibles funciones $\{h_i, 0 \leq i \leq 3\}$ en el Cuadro 4.2:

Para cada u, tomamos $R_X(u)$ tal que $f_X(\cdot, u) = h_{R_X(u)}(\cdot)$. Componiendo, $f_X(Z, U) = h_{R_X(U)}(Z) = h_{R_X}(Z)$, con $R_X \in \{0, 1, 2, 3\}$.

Al igual que en la sección anterior tenemos $R_Y = R_Y(U)$ tal que $f_Y(X, U) = g_{R_Y(U)}(X) = g_{R_Y}(X)$, con $R_Y \in \{0, 1, 2, 3\}$.

Combinando éstas últimas expresiones con las ecuaciones vistas en (4.2), tenemos que:

$$\begin{array}{l} X = h_{R_X}(Z) \\ Y = g_{R_Y}(X) \end{array} \tag{4.3}$$

	z_0	z_1
h_0	x_0	x_0
h_1	x_0	x_1
h_2	x_1	x_0
h_3	x_1	x_1

Cuadro 4.2: Posibles funciones de $\{z_0, z_1\}$ a $\{x_0, x_1\}$

Por lo que X e Y son funciones **determinísticas** de $R = (Z, R_X, R_Y)$.

Tomemos $r_{ijk} = P(Z = z_i, R_X = j, R_Y = k)$ la probabilidad puntual (representado en un vector de $\mathcal{S}_{32}$) del vector $R = (Z, R_X, R_Y)$. Por la independencia de Z y $\widetilde{R}$ (pues $\widetilde{R}$ es función de U y, por (2.5), $Z \coprod U$), tenemos que

$$r_{ijk} = P(Z = z_i)P(R_X = j, R_Y = k) = r_i^Z.\widetilde{r}_{jk}$$

Es decir, la probabilidad puntual de R pertenece al modelo

$$\mathcal{R} = \{r \in \mathcal{S}_{32} \,/\, r_{ijk} = P(Z = z_i)P(R_X = j, R_Y = k) = r_i^Z\widetilde{r}_{jk}, r^Z \in \mathcal{S}_2, \widetilde{r} \in \mathcal{S}_{16}\}.$$

Haciendo $\widehat{r} = (r^Z, \widetilde{r})$, con $r^Z \in \mathcal{S}_2, \widetilde{r} \in \mathcal{S}_{16}$ tenemos parametrizado el modelo $\mathcal{R}$. Veremos ahora que tanto el parámetro causal de interés como la distribución p de las variables observadas (Z, X, Y) pueden ser determinados por las coordenadas del vector $\widehat{r}$.

Lema 4.2.1 *La distribución de (Z, X, Y) y la de las variables contrafactuales Y_{x_0},Y_{x_1} son funciones de $\widehat{r} = (r^Z, \widetilde{r})$. Más aún, se deducen fórmulas explícitas para las probabilidades puntuales de Y_{x_0},Y_{x_1} y (Z, X, Y).*

Demostración 4.2.2 *Lo veremos para la siguiente la probabilidad puntual de (Z, X, Y), después daremos las fórmulas explícitas para las restantes probabilidades:*

$$
\begin{aligned}
&p_{(Z,X,Y)}(z_0,x_0,0) = \\
&= P(Z=z_0, X=x_0, Y=0) \\
&= P(Z=z_0, f_X(Z,U)=x_0, Y=0) \\
&= P(Z=z_0, f_X(z_0,U)=x_0, Y=0) \\
&= P\left(\bigcup_{0\le j\le 3: h_j(z_0)=x_0} \{Z=z_0, R_X=j, Y=0\}\right) \\
&= P(Z=z_0, R_X=0, Y=0) + P(Z=z_0, R_X=1, Y=0) \\
&= P(Z=z_0, R_X=0, f_Y(X,U)=0) + P(Z=z_0, R_X=1, f_Y(X,U)=0) \\
&= P(Z=z_0, R_X=0, f_Y(x_0,U)=0) + P(Z=z_0, R_X=1, f_Y(x_0,U)=0)
\end{aligned}
$$

$$
\begin{aligned}
&= P\left(\bigcup_{0\le k\le 3: g_k(x_0)=0} \{Z=z_0, R_X=0, R_Y=k\}\right) \\
&+ P\left(\bigcup_{0\le k\le 3: g_k(x_0)=0} \{Z=z_0, R_X=1, R_Y=k\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= P(Z=z_0, R_X=0, R_Y=0) + P(Z=z_0, R_X=0, R_Y=1) \\
&+ P(Z=z_0, R_X=1, R_Y=0) + P(Z=z_0, R_X=1, R_Y=1)
\end{aligned}
$$

$$
\begin{aligned}
&= P(Z=z_0)[P(R_X=0, R_Y=0) + P(R_X=0, R_Y=1)] \\
&+ P(Z=z_0)[P(R_X=1, R_Y=0) + P(R_X=1, R_Y=1)]
\end{aligned}
$$

$$
= r_0^Z(\widetilde{r_{00}} + \widetilde{r_{01}} + \widetilde{r_{10}} + \widetilde{r_{11}})
$$

Análogamente deducimos las siguientes expresiones:

$$
\begin{aligned}
P(Z=z_0, X=x_0, Y=1) &= r_0^Z(\widetilde{r_{02}} + \widetilde{r_{03}} + \widetilde{r_{12}} + \widetilde{r_{13}}) \\
P(Z=z_0, X=x_1, Y=0) &= r_0^Z(\widetilde{r_{20}} + \widetilde{r_{22}} + \widetilde{r_{30}} + \widetilde{r_{32}}) \\
P(Z=z_0, X=x_1, Y=1) &= r_0^Z(\widetilde{r_{21}} + \widetilde{r_{23}} + \widetilde{r_{31}} + \widetilde{r_{33}}) \\
P(Z=z_1, X=x_0, Y=0) &= r_1^Z(\widetilde{r_{00}} + \widetilde{r_{01}} + \widetilde{r_{20}} + \widetilde{r_{21}}) \\
P(Z=z_1, X=x_0, Y=1) &= r_1^Z(\widetilde{r_{02}} + \widetilde{r_{03}} + \widetilde{r_{22}} + \widetilde{r_{23}})
\end{aligned}
$$

$$P(Z = z_1, X = x_1, Y = 0) = r_1^Z(\widetilde{r_{10}} + \widetilde{r_{12}} + \widetilde{r_{30}} + \widetilde{r_{32}})$$

$$P(Z = z_1, X = x_1, Y = 1) = r_1^Z(\widetilde{r_{11}} + \widetilde{r_{13}} + \widetilde{r_{31}} + \widetilde{r_{33}})$$

Para Y_{x_0}, recordando que $Y_{x_0} = f_Y(x_0, U) = g_{R_Y}(x_0)$:

$$\begin{aligned} P(Y_{x_0} = 1) &= P(g_{R_Y}(x_0) = 1) \\ &= P\left(\bigcup_{0 \le k \le 3 : g_k(x_0) = 1} \{R_Y = k\}\right) \\ &= P(R_Y = 2) + P(R_Y = 3) \\ &= \sum_{i=0}^{4} P(R_Y = 2, R_X = i) + \sum_{i=0}^{4} P(R_Y = 3, R_X = i) \\ &= \widetilde{r_{02}} + \widetilde{r_{12}} + \widetilde{r_{22}} + \widetilde{r_{32}} + \widetilde{r_{03}} + \widetilde{r_{13}} + \widetilde{r_{23}} + \widetilde{r_{33}} \end{aligned}$$

De manera similar, obtenemos:

$$P(Y_{x_1} = 1) = \widetilde{r_{01}} + \widetilde{r_{11}} + \widetilde{r_{21}} + \widetilde{r_{31}} + \widetilde{r_{03}} + \widetilde{r_{13}} + \widetilde{r_{23}} + \widetilde{r_{33}}$$

Como Y_{x_i} es binaria $\forall i \in \{0, 1\}$, basta conseguir $P(Y_{x_i} = 1)$ para conseguir su distribución ya que $P(Y_{x_i} = 0) = 1 - P(Y_{x_i} = 1)$.

Corolario 4.2.3 *Podemos determinar el parámetro causal de interés como función de $\widetilde{r}$ de la siguiente forma:*

$$L_{ATE}(\widetilde{r}) = \widetilde{r_{01}} + \widetilde{r_{11}} + \widetilde{r_{21}} + \widetilde{r_{31}} - [\widetilde{r_{02}} + \widetilde{r_{12}} + \widetilde{r_{22}} + \widetilde{r_{32}}] \tag{4.4}$$

Observación 4.2.4 *Notemos que en realidad las probabilidades de nuestras variables contrafactuales son **funciones lineales** de $\widetilde{r}$ (al igual que nuestro parámetro causal), no así las probabilidades puntuales del vector (Z, X, Y). La falta de linealidad se debe al factor r_i^Z, pero nos podemos deshacer de r_i^Z reemplazándolo por $P(Z = z_i)$. Equivalentemente, podemos pasarlo dividiendo, obteniendo fórmulas para las probabilidades condicionales de (X, Y) dado Z en función de $\widetilde{r}$.*

En adelante, utilizaremos la siguiente notación:

$$\begin{aligned} p_{ij,0} &= P(X = x_i, Y = j | Z = z_0) \quad 0 \le i, j \le 1 \\ p_{ij,1} &= P(X = x_i, Y = j | Z = z_1) \quad 0 \le i, j \le 1 \end{aligned}$$

Por la observación 4.2.4, $p_{ij,l}$ resultan lineales en $\widetilde{r}$, porque se simpilifican los r_i^Z, dando los siguientes resultados:

$$p_{00,0} = P(X = x_0, Y = 0|Z = z_0) = \frac{P(X = x_0, Y = 0, Z = z_0)}{P(Z = z_0)} = \widetilde{r_{00}} + \widetilde{r_{01}} + \widetilde{r_{10}} + \widetilde{r_{11}}$$

$$p_{01,0} = P(X = x_0, Y = 1|Z = z_0) = \frac{P(X = x_0, Y = 1, Z = z_0)}{P(Z = z_0)} = \widetilde{r_{02}} + \widetilde{r_{03}} + \widetilde{r_{12}} + \widetilde{r_{13}}$$

$$p_{10,0} = P(X = x_1, Y = 0|Z = z_0) = \frac{P(X = x_1, Y = 0, Z = z_0)}{P(Z = z_0)} = \widetilde{r_{20}} + \widetilde{r_{22}} + \widetilde{r_{30}} + \widetilde{r_{32}}$$

$$p_{11,0} = P(X = x_1, Y = 1|Z = z_0) = \frac{P(X = x_1, Y = 1, Z = z_0)}{P(Z = z_0)} = \widetilde{r_{21}} + \widetilde{r_{23}} + \widetilde{r_{31}} + \widetilde{r_{33}}$$

$$p_{00,1} = P(X = x_0, Y = 0|Z = z_1) = \frac{P(X = x_0, Y = 0, Z = z_1)}{P(Z = z_1)} = \widetilde{r_{00}} + \widetilde{r_{01}} + \widetilde{r_{20}} + \widetilde{r_{21}}$$

$$p_{01,1} = P(X = x_0, Y = 1|Z = z_1) = \frac{P(X = x_0, Y = 1, Z = z_1)}{P(Z = z_1)} = \widetilde{r_{02}} + \widetilde{r_{03}} + \widetilde{r_{22}} + \widetilde{r_{23}}$$

$$p_{10,0} = P(X = x_1, Y = 0|Z = z_1) = \frac{P(X = x_1, Y = 0, Z = z_1)}{P(Z = z_1)} = \widetilde{r_{10}} + \widetilde{r_{12}} + \widetilde{r_{30}} + \widetilde{r_{32}}$$

$$p_{11,1} = P(X = x_1, Y = 1|Z = z_1) = \frac{P(X = x_1, Y = 1, Z = z_1)}{P(Z = z_1)} = \widetilde{r_{11}} + \widetilde{r_{13}} + \widetilde{r_{31}} + \widetilde{r_{33}} \tag{4.5}$$

Entonces, tomando

$$\widetilde{p} = L(p) = (p_{00,0}; p_{01,0}; p_{10,0}; p_{11,0}; p_{00,1}; p_{01,1}; p_{10,1}; p_{11,1}) \in \mathcal{S}_4 \times \mathcal{S}_4, \tag{4.6}$$

tenemos que $\exists L_o : \mathcal{S}_{16} \to \mathcal{S}_4 \times \mathcal{S}_4$ tal que $L_o(\widetilde{r}) = \widetilde{p}$ y L_o es una transformación lineal. Notemos que $\widetilde{p} \in \mathcal{S}_4 \times \mathcal{S}_4$ porque $\sum_{i,j=0}^{1} \widetilde{p}_{ij,0} = \sum_{i,j=0}^{1} \widetilde{p}_{ij,1} = 1$.

Luego de varios prolegómenos, recordemos que estamos buscando cotas finas para el parámetro causal (que puede ser obtenido por $L_{ATE}(\widetilde{r})$), sabiendo que $L_o(\widetilde{r}) = \widetilde{p}$. O sea, buscamos resolver los siguientes problemas:

$$\left\{ \begin{array}{c} \min\ L_{ATE}(\widetilde{r}) \\ \\ L_o(\widetilde{r}) = \widetilde{p} \\ \sum_{i,j=0}^{4} \widetilde{r_{ij}} = 1 \\ \widetilde{r} \geq 0 \end{array} \right\} \text{ y } \left\{ \begin{array}{c} \max\ L_{ATE}(\widetilde{r}) \\ \\ L_o(\widetilde{r}) = \widetilde{p} \\ \sum_{i,j=0}^{4} \widetilde{r_{ij}} = 1 \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{4.7}$$

Escribamos a $\widetilde{r}$ de la siguiente forma:

$$\widetilde{r} = (\widetilde{r}_{00}, \widetilde{r}_{01}, \widetilde{r}_{02}, \widetilde{r}_{03}, \widetilde{r}_{10}, \widetilde{r}_{11}, \widetilde{r}_{12}, \widetilde{r}_{13}, \widetilde{r}_{20}, \widetilde{r}_{21}, \widetilde{r}_{22}, \widetilde{r}_{23}, \widetilde{r}_{30}, \widetilde{r}_{31}, \widetilde{r}_{32}, \widetilde{r}_{33}) \in \mathcal{S}_{16}.$$

Usando las expresiones obtenidas para $L_{ATE}(\widetilde{r})$ y $L_o(\widetilde{r})$ en (4.4) y (4.5) respectivamente, deducimos los siguientes problemas de programación lineal:

$$\left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\} \text{ y } \left\{ \begin{array}{c} \max c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{4.8}$$

donde $c \in \mathbb{R}^{16}$ y la matriz $A \in \mathbb{R}^{8\times 16}$ están dados por :

$$c = (0, 1, -1, 0, 0, 1, -1, 0, 0, 1, -1, 0, 0, 1, -1, 0) \tag{4.9}$$

$$A = \left(\begin{array}{cccc|cccc|cccc|cccc} 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array} \right) \tag{4.10}$$

Observación 4.2.5 *La matriz A se puede pensar como una matriz de 16 bloques de $\mathbb{R}^{2\times 4}$:*

$$A = \left(\begin{array}{c|c|c|c} B_1 & B_1 & 0 & 0 \\ \hline 0 & 0 & B_2 & B_2 \\ \hline B_1 & 0 & B_1 & 0 \\ \hline 0 & B_2 & 0 & B_2 \end{array} \right)$$

Donde B_1, B_2 son de la forma:

$$B_1 = \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array} \right) = \left(\begin{array}{cc} \vec{U_2} & \vec{0_2} \\ \vec{0_2} & \vec{U_2} \end{array} \right)$$

$$B_2 = \left(\begin{array}{c|c} I_2 & I_2 \end{array} \right)$$

Denotando $\vec{U_k}$ el vector k-dimensional de unos, $\vec{0_k}$ el vector k-dimensional de ceros, mientras que I_k es la identidad de $\mathbb{R}^{k\times k}$. Otra observación es que hemos omitido la restricción $\sum_{i=0}^{16} \widetilde{r}_i = 1$, que no ha sido omitida en la primer sección (ver (1.20)), esto es porque termina siendo una restricción superflua, ya que multiplicar matricialmente

las primeras 4 filas de A con $\widetilde{r}$ y luego sumar las coordenadas resultantes termina siendo $\sum_{i=0}^{16} \widetilde{r}_i$. Por otro lado, las restricciones dicen que tienen que sumar lo mismo las primeras 4 coordenadas de p, pero al ser éstas las probabilidades de $(X,Y)|Z=z_0$, suman 1.

Observación 4.2.6 *El vector c en realidad puede ser visto como*

$$c = (\check{c}, \check{c}, \check{c}, \check{c}),$$

donde

$$\check{c} = (0,1,-1,0) = (0,1,0,1) + (0,0,-1,-1)$$

Notemos que 1 y -1 son los coeficientes que multiplican a $P(Y_{x_1}=1)$ y $P(Y_{x_0}=1)$ respectivamente en nuestro parámetro causal (ATE).

Estas últimas apreciaciones nos serán de mucha utilidad a la hora de generalizar el problema.

Resumiendo, hemos demostrado el siguiente resultado (obtenido por Balke et. at.)

Lema 4.2.7 *Sea (Z,X,Y) con función de probabilidad puntual p satisfaciendo un modelo causal funcional asociado al DAG (4.1). Sea $L(p)=\widetilde{p}$ el vector de probabilidades condicionales detallado en (4.6).*

Tenemos entonces que las cótas óptimas para $E(Y_{x_1})-E(Y_{x_0})$ están dadas por $m(p)$ y $M(p)$ definidos por

$$m(p) = \left\{ \begin{array}{c} min\ c\cdot\widetilde{r} \\ \\ A\cdot\widetilde{r}=\widetilde{p} \\ \widetilde{r}\geq 0 \end{array} \right\} \quad y \quad M(p) = \left\{ \begin{array}{c} max\ c\cdot\widetilde{r} \\ \\ A\cdot\widetilde{r}=\widetilde{p} \\ \widetilde{r}\geq 0 \end{array} \right\}$$

done c y A están definidos en (4.9) y (4.10), respectivamente.

Observación 4.2.8 *Notemos que las cotas dependen de p mediante $\widetilde{p}$. Es decir, dados $p_1, p_2 \in \mathcal{S}_8$ tal que $L(p_1)=L(p_2)=\widetilde{p}\in\mathcal{S}_4\times\mathcal{S}_4$, se cumple que $m(p_1)=m(p_2)$ y $M(p_1)=M(p_2)$. Es por ello que, abusando de la notación, escribimos indistintamente $m(p)$ o $m(\widetilde{p})$ y $M(p)$ o $M(\widetilde{p})$, a pesar de que $p\in\mathcal{S}_8$ y $\widetilde{p}\in\mathcal{S}_4\times\mathcal{S}_4$.*

Observación 4.2.9 *En el caso de que $m(p)$ sea estrictamente positivo, tendremos como consecuencia un ATE positivo y por lo tanto, evidencia a favor del tratamiento.*

En el caso de que $M(p)$ sea estrictamente negativo, tendremos como consecuencia un ATE negativo y por lo tanto, evidencia a favor del control.

4.3. Generalizando el problema a respuestas discretas

En este caso, tomaremos finitos valores posibles de respuesta. Por ejemplo, podría tener sentido tener en cuenta distintos niveles de mejoría del paciente con valores del 0 al 10. En términos de variables, permitiremos a Y tomar valores en $\{y_0, y_1, ..., y_{k-1}\}$, pero manteniendo X, Z binarias. Además, seguimos considerando a la variable U, y que las variables responden al DAG causal dado en (4.1).

Mantenemos X y Z binarias, pues al fin y al cabo, siempre estamos comparando dos posibles tratamientos. Decimos posibles tratamientos pues no siempre podemos estar comparando tratamiento vs. placebo. Podríamos comparar distintas dosis de un tratamiento (por lo que X y Z podrían tomar más valores). Pero al fin y al cabo, para medir el efecto causal, estaremos comparando dos valores de Z y, por consiguiente, de X.

Esto nos da otra fórmula para el parámetro causal:

$$ATE = E[Y_{x_1}] - E[Y_{x_0}] = \sum_{j=0}^{k-1} y_j \cdot [P(Y_{x_1} = y_j) - P(Y_{x_0} = y_j)] \tag{4.11}$$

El hecho de tener más valores posibles para Y no cambia la falta de identificabilidad del parámetro causal (recordar que la definición de identificabilidad depende de cuales son las variables observadas), por lo que buscaremos cotas para este caso general, con construcciones muy parecidas al del caso binario.

De nuevo, tomaremos el vector discreto $\widetilde{R} = (R_X, R_Y)$ cuya disribución será suficiente para determinar la distribución de las variables contrafactuales (en consecuencia, determinará el parámetro causal) y junto con la distribución de Z determinarán la distribución las variables observadas (Z, X, Y). Ahora R_X será igual que antes (los valores de Z y X no cambiaron, por lo que la función que los relaciona sigue siendo una $h_j : \{z_0, z_1\} \to \{x_0, x_1\}$ con $0 \leq j \leq 3$).

La variación surge en la construcción de R_Y, pues ahora $Y = f_Y(\cdot, U)$ es una función que parte de $\{x_0, x_1\}$ y llega a $\{y_0, y_1, ..., y_{k-1}\}$ (hay k^2 posibilidades). Por eso, determinaremos R_Y mediante la tabla dada en el Cuadro 4.3:

Es decir, si $s = lk + r$ con $0 \leq l \leq k-1, 0 \leq r \leq k-1$, tendremos que $g_s(x_0) = y_l, g_s(x_1) = y_r$. De la misma forma que antes, definiremos R_Y de forma que $f_Y(X, U) = g_{R_Y}(X)$.

Siguiendo este festival de dèjá vús, tomaremos las probabilidades puntuales de $\widetilde{R}$ (ya vimos que tomar las puntuales de Z no nos ayudarán a la hora de resolver el problema de programación lineal) en el vector $\widetilde{r}_{ij} = P(R_X = i, R_Y = j)$ ahora nuestro $\widetilde{r}$ pertenecerá a $\mathbb{R}^{4k^2}$.

Recordando que $Y_{x_1} = g_{R_Y}(x_1)$ y $Y_{x_0} = g_{R_Y}(x_0)$, concluimos que

g $\setminus$ X	x_0	x_1
g_0	y_0	y_0
g_1	y_0	y_1
$\vdots$	$\vdots$	$\vdots$
g_{k-1}	y_0	y_{k-1}
$\vdots$	$\vdots$	$\vdots$
g_{lk}	y_l	y_0
g_{lk+1}	y_l	y_1
$\vdots$	$\vdots$	$\vdots$
g_{lk+k-1}	y_l	y_{k-1}
$\vdots$	$\vdots$	$\vdots$
$g_{(k-1)k}$	y_{k-1}	y_0
g_{k^2-k+1}	y_{k-1}	y_1
$\vdots$	$\vdots$	$\vdots$
g_{k^2-1}	y_{k-1}	y_{k-1}

Cuadro 4.3: Posibles funciones de $\{x_0, x_1\}$ a $\{y_0, y_1, ..., y_{k-1}\}$

$$\begin{aligned} P(Y_{x_1} = y_j) &= P(g_{R_Y}(x_1) = y_j) \\ &= P(R_Y \in \{lk + j : 0 \leq l \leq k-1\}) \\ &= \sum_{l=0}^{k-1} P(R_Y = lk + j) \\ &= \sum_{l=0}^{k-1} \sum_{i=0}^{3} P(R_X = i, R_Y = lk + j) \\ &= \sum_{i=0}^{3} \sum_{l=0}^{k-1} \widetilde{r}_{i,lk+j} \end{aligned}$$

y

$$\begin{aligned} P(Y_{x_0} = y_j) &= P(g_{R_Y}(x_0) = y_j) \\ &= P(R_Y \in \{jk + r : 0 \leq r \leq k-1\}) \\ &= \sum_{r=0}^{k-1} P(R_Y = jk + r) \\ &= \sum_{r=0}^{k-1} \sum_{i=0}^{3} P(R_X = i, R_Y = jk + r) \\ &= \sum_{i=0}^{3} \sum_{r=0}^{k-1} \widetilde{r}_{i,jk+r} \end{aligned}$$

Concluyendo que el ATE se puede expresar mediante la siguiente función lineal de $\widetilde{r}$:

$$L_{ATE}(\widetilde{r}) = \sum_{j=0}^{k-1} y_j \cdot [P(Y_{x_1} = y_j) - P(Y_{x_0} = y_j)] = \sum_{j=0}^{k-1} y_j \cdot \left[\sum_{i=0}^{3} \sum_{l=0}^{k-1} \widetilde{r}_{i,lk+j} - \sum_{i=0}^{3} \sum_{r=0}^{k-1} \widetilde{r}_{i,jk+r} \right]$$

Intercambiando sumatorias (son finitas), tenemos:

$$L_{ATE}(\widetilde{r}) = \sum_{i=0}^{3} \sum_{j=0}^{k-1} y_j \left[\sum_{l=0}^{k-1} \widetilde{r}_{i,lk+j} - \sum_{r=0}^{k-1} \widetilde{r}_{i,jk+r} \right]$$

También deducimos las siguientes fórmulas para las probabilidades condicionales $p_{(X,Y)|Z}$, fijando $0 \leq j \leq k-1$

$$p_{0j,0} = P(X = x_0, Y = y_j | Z = z_0) = P(R_X \in \{0,1\}, R_Y \in \{jk + r : 0 \leq r \leq k-1\})$$

$$p_{0j,0} = \sum_{r=0}^{k-1} \widetilde{r}_{0,jk+r} + \widetilde{r}_{1,jk+r}$$

$$p_{1j,0} = P(X = x_1, Y = y_j | Z = z_0) = P(R_X \in \{2,3\}, R_Y \in \{lk + j : 0 \leq l \leq k-1\})$$

$$p_{1j,0} = \sum_{l=0}^{k-1} \widetilde{r}_{2,lk+j} + \widetilde{r}_{3,lk+j}$$

$$p_{0j,1} = P(X = x_0, Y = y_j | Z = z_1) = P(R_X \in \{0,2\}, R_Y \in \{jk + r : 0 \leq r \leq k-1\})$$

$$p_{0j,1} = \sum_{r=0}^{k-1} \widetilde{r}_{0,jk+r} + \widetilde{r}_{2,jk+r}$$

$$p_{1j,1} = P(X = x_1, Y = y_j | Z = z_1) = P(R_X \in \{1,3\}, R_Y \in \{lk + j : 0 \leq l \leq k-1\})$$

$$p_{1j,1} = \sum_{l=0}^{k-1} \widetilde{r}_{1,lk+j} + \widetilde{r}_{3,lk+j}$$

Organizaremos los vectores $\widetilde{r} \in \mathcal{S}_{4k^2}$ y $\widetilde{p} \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$ de la siguiente forma:

$$\widetilde{r} = (\widetilde{r}_{0,0}; \widetilde{r}_{0,1}; ...; \widetilde{r}_{0,k^2-1}; \widetilde{r}_{1,0}; \widetilde{r}_{1,1}; ...; \widetilde{r}_{1,k^2-1}; \widetilde{r}_{2,0}; \widetilde{r}_{2,1}; ...; \widetilde{r}_{2,k^2-1}; \widetilde{r}_{3,0}; \widetilde{r}_{3,1}; ...; \widetilde{r}_{3,k^2-1})$$

$$\widetilde{p} = L(p) = \begin{pmatrix} p_{00,0} \\ p_{01,0} \\ \vdots \\ p_{0(k-1),0} \\ p_{10,0} \\ p_{11,0} \\ \vdots \\ p_{1(k-1),0} \\ p_{00,1} \\ p_{01,1} \\ \vdots \\ p_{0(k-1),1} \\ p_{10,1} \\ p_{11,1} \\ \vdots \\ p_{1(k-1),1} \end{pmatrix} \tag{4.12}$$

A partir de estas expresiones podemos deducir una transformación lineal $L_o : \mathcal{S}_{4k^2} \to \mathcal{S}_{2k} \times \mathcal{S}_{2k}$ tal que $L_o(\widetilde{r}) = \widetilde{p}$. Luego, tenemos que las cotas para el parámetro causal se encuentran resolviendo los problemas

$$\left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\} \quad y \quad \left\{ \begin{array}{c} \max c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{4.13}$$

donde, la organización de las coordenadas de $\widetilde{r}$ y $\widetilde{p}$, dan las siguientes estructuras para la matriz de restricciones A (la veremos como bloques de $\mathbb{R}^{k \times k^2}$) y el vector c, similares a las obtenidas en el caso binario:

$$A = \left(\begin{array}{c|c|c|c} B_1 & B_1 & 0 & 0 \\ \hline 0 & 0 & B_2 & B_2 \\ \hline B_1 & 0 & B_1 & 0 \\ \hline 0 & B_2 & 0 & B_2 \end{array} \right) \tag{4.14}$$

Donde B_1, B_2 son de la forma:

$$B_1 = \begin{pmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$B_1 = \begin{pmatrix} \vec{U_k} & \vec{0_k} & \cdots & \vec{0_k} \\ \vec{0_k} & \vec{U_k} & \cdots & \vec{0_k} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{0_k} & \vec{0_k} & \cdots & \vec{U_k} \end{pmatrix}$$

$$B_2 = \begin{pmatrix} I_k \,\big|\, \cdots \,\big|\, I_k \end{pmatrix}$$

Recordemos que denotamos por $\vec{U_k}$ al vector k-dimensional de unos, $\vec{0_k}$ el vector k-dimensional de ceros e I_k la matriz identidad de $\mathbb{R}^{k\times k}$

Notar que estos bloques son de $k \times k^2$, por lo que la matriz A resulta de $4k \times 4k^2$.

Por otro lado, tenemos que

$$c = (\widetilde{c}, \widetilde{c}, \widetilde{c}, \widetilde{c}) \tag{4.15}$$

con $\widetilde{c}$ dado por

$$(y_0 - y_0; y_1 - y_0; ...; y_{k-1} - y_0; ...; y_0 - y_l; y_1 - y_l; ...;$$
$$y_{k-1} - y_l; ...; y_0 - y_{k-1}; y_1 - y_{k-1}; ...; y_{k-1} - y_{k-1}).$$

Observación 4.3.1 *Este vector puede ser visto como* $\sum_{i=0}^{k-1} y_i \cdot [\vec{v_i} - \vec{w_i}]$, *donde* $\vec{v_i} = (e_{\vec{i+1}}|e_{\vec{i+1}}|\cdots|e_{\vec{i+1}})$ *se obtiene concatenando* k *veces* $e_{\vec{i+1}}$, *el vector canónico de* $\mathbb{R}^k$ *mientras que* $\vec{w_i} = (\vec{0_k}|\cdots|\vec{0_k}|\vec{U_k}|\vec{0_k}|\cdots|\vec{0_k})$ *es un vector en* $\mathbb{R}^{k^2}$ *con unos en entre las coordenadas* $ik+1$ *y la* $ik+k = k(i+1)$, *y ceros en las restantes.*

Observación 4.3.2 *Notemos que el vector* $\widetilde{r}$ *está en* $\mathcal{S}_{4k^2}$, *como sus coordenadas suman 1 y son mayores o iguales a 0, podemos decir que* $\widetilde{r} \in [0,1]^{4k^2}$ *(que es compacto), por lo que el funcional lineal (en particular, continuo) es acotado en* $[0,1]^{4k^2}$, *por lo que ahí tendremos realización del máximo y el mínimo.*

El siguiente resultado, generaliza el Lemma 4.2.7.

Teorema 4.3.3 *Sea* (Z, X, Y) *con función de probabilidad puntual* p *satisfaciendo un modelo causal funcional asociado al DAG (4.1). Sea* $L(p) = \widetilde{p}$ *el vector de probabilidades condicionales dado en (4.12).*

Tenemos entonces que las cotas óptimas para $E(Y_{x_1}) - E(Y_{x_0})$ *están dadas por* $m(p)$ *y* $M(p)$ *definidos por*

$$m(p) = \left\{ \begin{array}{c} min\ c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\} \quad y \quad M(p) = \left\{ \begin{array}{c} max\ c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p} \\ \widetilde{r} \geq 0 \end{array} \right\}$$

done c y A están definidos en (4.15) y (4.14), respectivamente.

Mas generalmente, bajo estas mismas condiciones la cotas óptimas para $E[t(Y_{x_1})] - E[t(Y_{x_0})]$ están dadas por la solución de los mismos problemas, reemplazando con c por $c_t = (\widetilde{c}_t, \widetilde{c}_t, \widetilde{c}_t, \widetilde{c}_t)$, dado por $\widetilde{c}_t = \sum_{i=0}^{k-1} t(y_i) \cdot [\vec{v}_i - \vec{w}_i] \in \mathbb{R}^{k^2}$, donde $\vec{v}_i, \vec{w}_i$ son los dados en 4.3.1.

Nuevamente, las cotas dependen de p mediante $\widetilde{p}$. Es por ello que, abusando de la notación, escribimos indistintamente $m(p)$ o $m(\widetilde{p})$ y $M(p)$ o $M(\widetilde{p})$.

4.4. Estadística

El Teorema 4.3.3 nos permite caracterizar las cotas óptimas para el efecto medio del tratamiento conociendo la distribución p del vector (Z, X, Y). Cuando hacemos estadística, la distribución p es desconocida pero disponemos de una muestra (Z_i, X_i, Y_i) del vector (Z, X, Y) y asumiremos conocido el rango de la variable Y. En tal caso, podemos estimar p con $\widehat{p}_n$, la funcion de frecuencia relativa asociada a la muestra. La ley de los grandes números garantiza que $\widehat{p}_n$ converge a p, lo que sugiere estimar las cotas con $m(\widehat{p}_n)$, $M(\widehat{p}_n)$. La consistencia de esta propuesta depende de la continuidad de las funciones $m(\cdot) : \mathcal{S}_{4k} \to \mathbb{R}$ y $M(\cdot) : \mathcal{S}_{4k} \to \mathbb{R}$ en el punto p.

Dada una muestra (Z_i, X_i, Y_i), con $1 \leq i \leq n$, las siguientes frecuencias relativas aproximan a las probabilidades puntuales de la variable Z y las probabilidades condicionales de (X, Y) dado Z:

$$P_n(Z = 1) = \frac{\sum_{i=1}^{n} I_{\{z_1\}}(Z(i))}{n} \tag{4.16}$$

$$P_n(Z = z_0) = \frac{\sum_{i=1}^{n} I_{\{z_0\}}(Z(i))}{n} = 1 - P_n(Z = 1) \tag{4.17}$$

$$p^n_{ij,0} = \frac{\frac{\sum_{i=1}^{n} I_{\{(x_i,y_j,z_0)\}}((X(i), Y(i), Z(i)))}{n}}{P_n(Z = z_0)}$$

Entonces, usando la ecuación (4.17) simplificando los n,

$$p^n_{ij,0} = \frac{\sum_{i=1}^{n} I_{\{(x_i,y_j,z_0)\}}((X(i), Y(i), Z(i)))}{\sum_{i=1}^{n} I_{\{z_0\}}(Z(i))}, \forall i \in \{0, 1\}, \forall 1 \leq j \leq k-1$$

De manera similar, usando (4.16)

$$p^n_{ij,1} = \frac{\sum_{i=1}^n I_{\{(x_i,y_j,z_1)\}}((X(i),Y(i),Z(i)))}{\sum_{i=1}^n I_{\{z_1\}}(Z(i))}, \forall i \in \{0,1\}, \forall 1 \leq j \leq k-1$$

Formando con estas últimas (de la misma forma que en (4.12)) expresiones un vector $\widetilde{p}_n \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$ resolvemos los problemas de programación lineal detallados en (4.13), pero cambiando el vector de restricciones por $\widetilde{p}_n$, dando así estimadores para las cotas del parámetro causal:

$$m(\widehat{p}_n) = m(\widetilde{p}_n) = \left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p}_n \\ \widetilde{r} \geq 0 \end{array} \right\} \quad y \quad M(\widehat{p}_n) = M(\widetilde{p}_n) = \left\{ \begin{array}{c} \max c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p}_n \\ \widetilde{r} \geq 0 \end{array} \right\} \tag{4.18}$$

Recordemos que por la Observación 4.3.2 podemos garantizar la obtención de los óptimos del funcional $m(\widetilde{p}_n)$ y $M(\widetilde{p}_n)$, para cada poliedro de soluciones factibles determinado por $\widetilde{p}_n$.

4.4.1. Consistencia

Recién vimos que $\widehat{p}_n \xrightarrow{c.s.} p$, pero para cada $\widetilde{p}_n$ resolvemos los problemas de optimización dados por (4.18). ¿Podemos inferir que entonces $m(\widehat{p}_n) \to m(p)$ y $M(\widehat{p}_n) \to M(p)$?

Para demostrar eso, antes tendremos que hablar largo y tendido sobre algunas cuestiones de programación lineal.

Sea una matriz $A \in \mathbb{R}^{m \times n}$, $b, b_n \in \mathbb{R}^m$ y $c, r \in \mathbb{R}^n$, donde r es un vector de variables.

Definición 4.4.1 *Dado el siguiente problema de programación lineal, llamado problema primal, dado por:*

$$(P) \left\{ \begin{array}{c} min\ c \cdot r \\ \\ A \cdot r = b \\ r \geq 0 \end{array} \right\} \tag{4.19}$$

A partir de (P) se formula el llamado ***problema dual****, de la siguiente manera:*

$$(D) \left\{ \begin{array}{c} max\ b^t \cdot y \\ \\ A^t \cdot y \leq c^t \\ y \geq 0 \end{array} \right\}$$

Observación 4.4.2 *Notar que para que las multiplicaciones matriciales tengan sentido, el vector de variables y tiene que tener igual cantidad de variables que restricciones del problema primal. En el libro de Castillo et. al. [4], se detalla una interpretación de la correspondencia entre variables duales y restricciones del problema primal.*

Un resultado más que importante de dualidad, se da en el siguiente teorema:

Teorema 4.4.3 *Dados los problemas*

$$(P)\left\{\begin{array}{c} min\ c\cdot r \\ \\ A\cdot r = b \\ r \geq 0 \end{array}\right\}\ (D)\left\{\begin{array}{c} max\ b^t\cdot y \\ \\ A^t\cdot y \leq c^t \\ y \geq 0 \end{array}\right\}$$

Si (P) tiene una ***solución óptima*** r_0*, entonces (D) tiene una* ***solución óptima*** y_0*, donde además se cumple la igualdad entre los* ***óptimos****, es decir:* $c\cdot r_0 = b^t\cdot y_0$*, donde* $c\cdot r_0$ *es* ***óptimo*** *de (P) y* $b^t\cdot y_0$ *es* ***óptimo*** *de (D).*

Observación 4.4.4 *Notar que en el teorema anterior se hace hincapié en la diferencia entre* ***solución óptima*** *y* ***óptimo****.*

El siguiente resultado nos garantizará la consistencia del estimador.

Teorema 4.4.5 *Sean los siguientes problemas:*

$$(P)\left\{\begin{array}{c} min\ c\cdot r \\ \\ A\cdot r = b \\ r \geq 0 \end{array}\right\}\quad (P_n)\left\{\begin{array}{c} min\ c\cdot r \\ \\ A\cdot r = b_n \\ r \geq 0 \end{array}\right\} \tag{4.20}$$

de forma existen $c\cdot r_0^n$ *y* $c\cdot r_0$ *los óptimos para los problemas* (P_n) *y (P), respectivamente. Entonces, si* $b_n \to b$*, tenemos que*

$$c\cdot r_0^n \to c\cdot r_0. \tag{4.21}$$

Demostración 4.4.6 *Para demostrar este resultado tenemos el problema de que los conjuntos de soluciones factibles de cada* (P_n) *son distintos. ¿En qué nos simplifica el problema el potencial uso de este teorema de dualidad? Los respectivos duales de estos problemas están dados por:*

$$(D)\left\{\begin{array}{c} max\ b^t\cdot y \\ \\ A^t\cdot y \leq c^t \\ y \geq 0 \end{array}\right\}\ (D_n)\left\{\begin{array}{c} max\ b_n^t\cdot y \\ \\ A^t\cdot y \leq c^t \\ y \geq 0 \end{array}\right\}$$

Notar que la región de factibilidad de (D) y el de (D_n) coinciden para cualquier $n \in \mathbb{N}$ y por consiguiente, coinciden las regiones de factibilidad de (D_n) y (D_m) $\forall n \neq m$, simplificando el problema que generaba la diferencia.

Por el teorema de dualidad, el hecho de que (P) y (P_n) alcancen su óptimo, implica que (D) y (D_n) lo hacen. Tomemos como $b_n^t \cdot y_0^n$ al óptimo de (D_n) para cada n, mientras que llamaremos $b^t \cdot y_0$ al óptimo de (D). Además, por el teorema de dualidad, $b^t \cdot y_0 = c \cdot r_0$ y $b_n^t \cdot y_0^n = c \cdot r_0^n, \forall n \in \mathbb{N}$. De aquí concluimos que verificar la convergencia dada por (4.4.5), se traduce en verificar que se cumple la siguiente convergencia:

$$b_n^t \cdot y_0^n \to b^t \cdot y_0 \tag{4.22}$$

Demostraremos a continuación esto último. Supondremos sin pérdida de generalidad que cada y_0^n e y_0 son puntos extremos del ***mismo*** *poliedro de soluciones $S = \{y \in \mathbb{R}^m : A^t \cdot y \leq c^t, y \geq 0\}$. Si no fueran puntos extremos, hay algún punto extremo que también alcanza dicho óptimo. Como estos puntos extremos son finitos, sabemos que $\exists M > 0$ tal que $||y_0^n|| \leq M, \forall n \in \mathbb{N}$, donde $||\cdot||$ denota la norma de un vector de $\mathbb{R}^m$. Podemos pedirle a M que también cumpla $||y_0|| \leq M$.*

Fijemos $\varepsilon > 0$, debemos ver que $\exists n_0 = n_0(\varepsilon)$ tal que $|b_n^t \cdot y_0^n - b^t \cdot y_0| < \varepsilon, \forall n \geq n_0$. Veamos esto:

Por empezar, tenemos que $b_n \longrightarrow b$, por lo que $||b_n - b|| \longrightarrow 0 \Rightarrow \exists n_0 = n_0(\varepsilon)$ tal que $||b_n - b|| < \dfrac{\varepsilon}{3M}$.

$$\begin{aligned} |b_n^t \cdot y_0^n - b^t \cdot y_0| &= |b_n^t \cdot y_0^n - b_n^t \cdot y_0 + b_n^t \cdot y_0 - b^t \cdot y_0| \\ &\leq |b_n^t \cdot y_0^n - b_n^t \cdot y_0| + |b_n^t \cdot y_0 - b^t \cdot y_0| \end{aligned}$$

Analicemos estos dos términos por separado.
Para el segundo término:

$$\begin{aligned} |b_n^t \cdot y_0 - b^t \cdot y_0| &= |(b_n^t - b^t) \cdot y_0| \\ &\leq ||b_n^t - b^t|| \cdot ||y_0|| \text{ (por Cauchy-Schwarz)} \\ &< \frac{\varepsilon}{3}, \text{ si } n \geq n_0 \end{aligned}$$

Para el primer término, observemos lo siguiente:

1. *y_0^n es solución óptima de (D_n), por lo que $b_n^t \cdot y_0^n \geq b_n^t \cdot y_0$*

2. *y_0 es solución óptima de (D), por lo que $b^t \cdot y_0 \geq b^t \cdot y_0^n$*

$$\begin{aligned}
|b_n^t \cdot y_0^n - b_n^t \cdot y_0| &= b_n^t \cdot y_0^n - b_n^t \cdot y_0 \text{ (por 1.)} \\
&= b_n^t \cdot y_0^n - b^t \cdot y_0^n + b^t \cdot y_0^n - b_n^t \cdot y_0 \\
&\leq b_n^t \cdot y_0^n - b^t \cdot y_0^n + b^t \cdot y_0 - b_n^t \cdot y_0 \text{ (por 2.)} \\
&= (b_n^t - b^t) \cdot y_0^n + (b^t - b_n^t) \cdot y_0 \\
&\leq ||b_n^t - b^t|| \cdot ||y_0^n|| + ||b^t - b_n^t|| \cdot ||y_0|| \text{ (por C-S)} \\
&\leq ||b_n^t - b^t|| \cdot M + ||b^t - b_n^t|| \cdot M \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3}, \forall n \geq n_0
\end{aligned}$$

Entonces, si $n \geq n_0$, conseguimos:

$$|b_n^t \cdot y_0 - b^t \cdot y_0| < \varepsilon$$

De esto deducimos que:

$$|b_n^t \cdot y_0 - b^t \cdot y_0| = |c \cdot r_0^n - c \cdot r_0| \longrightarrow 0$$

Como hemos visto en la Observación 4.2.6, si tomamos $b \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$ y la matriz A definida en (4.14), por la estructura de la matriz A, la restricción $A \cdot r = b$ garantiza que el vector r está en el compacto $\mathcal{S}_{4k^2}$. Por la Observación 4.3.2 podemos asegurar existencia del óptimo para el problema dado en (4.19), **para todo** $b \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$. Además, si $b_n \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$, $\forall n \in \mathbb{N}$ y $b \in \mathcal{S}_{2k} \times \mathcal{S}_{2k}$ tal que $b_n \to b$, estamos en las condiciones del Teorema 4.4.5, por lo que podemos asegurar la convergencia de los mínimos de los problemas dados en (4.13).

La Observación 1.7.2 nos permite también asegurar la convergencia de los máximos de los problemas detallados en (4.13).

Terminamos la sección enunciando el resultado de consistencia para las cotas propuestas.

Teorema 4.4.7 *Sea $\widehat{p}_n$ la función de frecuencias relativas asociada a la muestra (Z_i, X_i, Y_i), para $1 \leq i \leq n$, del vector (Z, X, Y) con función de probabilidad puntual p. Entonces,*

$$m(\widehat{p}_n) \xrightarrow{c.s.} m(p)$$
$$M(\widehat{p}_n) \xrightarrow{c.s.} M(p)$$

Demostración 4.4.8 *Por la ley fuerte de los grandes números, sabemos que $\widehat{p}_n \xrightarrow{c.s.} p$. Como las funciones continuas preservan la convergencia casi segura, $\widetilde{p}_n \xrightarrow{c.s.} \widetilde{p}$. Luego, como los óptimos de los problemas dados en 4.13 se alcanzan, el Teorema 4.4.5 permite garantizar que:*

$$m(\widehat{p}_n) = m(\widetilde{p}_n) \xrightarrow{c.s.} m(\widetilde{p}) = m(p)$$
$$M(\widehat{p}_n) = M(\widetilde{p}_n) \xrightarrow{c.s.} M(\widetilde{p}) = M(p)$$

4.4.2. Métodos para generar datos

Para generar datos, vimos que era suficiente saber las distribuciones de la variable Z y de $R = (R_X, R_Y)$, ya que la distribución de (X, Y, Z) se puede determinar a partir de ellas.

Veamos como se determina cada observación de la muestra $(Z(i), X(i), Y(i))$.

1. **Para obtener $Z(i)$:**

   - Como Z es una variable que toma valores en $\{0, 1\}$, nos basta saber cuánto es $P(Z = 1)$ (ya que $P(Z = z_0) = 1 - P(Z = 1)$).
   - Dado $p_Z = P(Z = 1)$, en cada iteración $1 \leq i \leq n$ generamos un número aleatorio U_i entre 0 y 1 ($U_i \sim \mathcal{U}([0, 1])$).
   - Si $U_i \leq p_Z$, le otorgamos el valor 1 a $Z(i)$, en caso contrario se tendrá $Z(i) = 0$. Notar que

   $$P(Z(i) = 1) = P(U_i \leq p_Z) = p_Z = P(Z = 1)$$

2. **Para determinar $X(i), Y(i)$:**

   - En cada iteración generamos otro número aleatorio $\widetilde{U}_i \sim \mathcal{U}([0, 1])$, independiente de U_i.
   - Dado $p_R = (\widetilde{r}_{00}, \widetilde{r}_{01}, ..., \widetilde{r}_{0,k-1}, ..., \widetilde{r}_{40}, \widetilde{r}_{41}, ..., \widetilde{r}_{4,k-1}) \in \mathcal{S}_{4k^2}$, construimos $F_R \in \mathbb{R}^{4k^2}$, acumulando los valores de p_R, es decir: $F_R(j) = \sum_{l=1}^{j} p_R(l) \ \ \forall 1 \leq j \leq 4k^2$.
   - Notar que $F_R(4k^2) = \sum_{l=1}^{4k^2} p_R(l) = 1$ pues $p_R \in \mathcal{S}_{4k^2}$. Podemos construir una partición del intervalo $[0, 1]$ de la siguiente manera:
     $[0, 1] = \cup_{j=1}^{4k^2} I_j$, donde $I_1 = [0, F_R(1)]$ e $I_j = (F_R(j-1), F_R(j)], \forall 2 \leq j \leq 4k^2$.
   - Sea u el valor que toma $\widetilde{U}_i$, entonces $\exists! 1 \leq j_0 \leq 4k^2$ tal que $u \in I_{j_0}$. Como $1 \leq j_0 \leq 4k^2$, $\exists! 0 \leq s \leq 3, 1 \leq t \leq k^2$ que cumple $j_0 = s.k^2 + t$. Una vez encontrados s, t, asignamos a $R_X(i)$ el valor s, y $R_Y(i) = t - 1$ (recordemos que los valores de R_Y pertenecen a $\{0, 1, ..., k^2 - 1\}$ y $1 \leq t \leq k^2$).
   - Con estos datos, sabemos (por construcción de R_X, R_Y) que $X(i) = h_{R_X(i)}(Z(i)) = h_s(Z(i))$, y una vez obtenido $X(i)$, tenemos que $Y(i) = g_{R_Y(i)}(X(i)) = g_{t-1}(X(i))$.

Así es como generamos una muestra de tamaño n para (X, Y, Z).

4.4.3. El algoritmo

Hemos programado en lenguaje R un programa que a partir de un vector p que representa la distribución de (X, Y, Z) y un vector constituido por un rango de valores para Y, construye el vector $\widetilde{p}$ y resuelve los problemas dados en (4.13), devolviendo un vector $(m(p), M(p))$ con las cotas óptimas para el ATE.

El algoritmo toma como parámetro un vector p que debe estar organizado del siguiente modo:

$$p = \begin{pmatrix} p_{000} \\ p_{010} \\ \vdots \\ p_{0(k-1)0} \\ p_{100} \\ p_{110} \\ \vdots \\ p_{1(k-1)0} \\ p_{001} \\ p_{011} \\ \vdots \\ p_{0(k-1)1} \\ p_{101} \\ p_{111} \\ \vdots \\ p_{1(k-1)1} \end{pmatrix}$$

donde $p_{ijl} = P(X = x_i, Y = y_j, Z = z_l)$ con $0 \leq i, l \leq 1 \ \wedge \ 0 \leq j \leq k-1$.

En el caso de tener una muestra, programamos otra función que dada una matriz $A \in \mathbb{R}^{n\times 3}$ (donde la primer columna corresponde a las observaciones de X, la segunda a las de Y y la tercera a las de Z) estima p con $\widehat{p}_n$ y (tomando también el rango de Y) usa la función anterior para encontrar los estimadores $m(\widehat{p}_n)$ y $M(\widehat{p}_n)$.

Además, tenemos versiones similares a estos últimos programas para casos en los que el parámetro causal no es el ATE, sino de la siguiente forma:

$$\sum_{j=0}^{k-1} a_{j1} \cdot P(Y_{x_1} = y_j) + a_{j0} \cdot P(Y_{x_0} = y_j) \tag{4.23}$$

El algoritmo requiere como parámetros los vectores $(a_{j1} : 1 \leq j \leq k-1)$ y $(a_{j0} : 1 \leq j \leq k-1)$. Los hemos presentado en programas separados porque tomando $a_{j1} = y_j, a_{j0} = -y_j$ y reemplazando en (4.23), obtenemos el ATE.

En el caso general de que el parámetro causal esté dado por $E[t(Y_{x_1}) - t(Y_{x_0})]$, habrá que pasar como parámetro los vectores cuyas coordenadas son $a_{j1} = t(y_j), a_{j0} = -t(y_j)$,

con $0 \leq j \leq k-1$.

Si se dispone del vector p, para encontrar las cotas para el parámetro causal dado en (4.23) ya no es necesario pasar como parámetro el rango de Y. Sin embargo, si se dispone de una muestra, requerimos del rango para construir la distribución empírica $\widehat{p}_n$, y, a partir de él conseguir $m(\widehat{p}_n)$ y $M(\widehat{p}_n)$.

4.4.4. Cotas teóricas

Para el caso en que Y es binaria, Balke y Pearl [1] han encontrado expresiones teóricas para las cotas del ATE, dependiendo de ciertas condiciones que deben cumplir las probabilidades condicionales $p_{ij,0}$, $p_{ij,1}, i \in \{0,1\}, j \in \{0,...,k-1\}$. Recordemos que

$$p_{ij,1} = P(X = x_i, Y = y_j | Z = z_1)$$

$$p_{ij,0} = P(X = x_i, Y = y_j | Z = z_0).$$

Estas cotas teóricas están detalladas en los Cuadros 4.4 y 4.5

Generando muestras como detallamos en la sección anterior, para distintos valores de p_Z y p_R (en el caso binario, con $k = 2$) y comprobamos que las cotas obtenidas mediante nuestro algoritmo coinciden con las cotas teóricas. En la mayoría de los casos la diferencia entre las cotas obtenidas por el algoritmo y las cotas teóricas daba como resultado **exacto** 0, es decir, sin errores numéricos. Cuando hubo errores numéricos, no superaban el orden de 10^{-16}.

Llamaremos $M(p)$ al máximo teórico para el ATE, $m(p)$ el mínimo teórico y $\widehat{M(p)}, \widehat{m(p)}$ los óptimos obtenidos por el algoritmo. Además, llamamos $\widehat{r} = (r^Z, \widetilde{r})$ al vector dado en el Lema 4.2.1. Con estas notaciones, explayaremos a continuación resultados de dicha simulación en la siguiente lista:

-
  - $r^Z = (0{,}6; 0{,}4)$
  - $\widetilde{r} = (0; 0{,}0667; 0{,}0444, 0{,}0778; 0{,}1222; 0{,}2222; 0; 0; 0; 0{,}0778; 0{,}1333; 0; 0{,}0778; 0; 0{,}1; 0{,}0778)$
  - $L_{ATE}(\widetilde{r}) = 0{,}089$
  -

$m(p)$	$\widehat{m(p)}$	$m(p) - \widehat{m(p)}$	$M(p)$	$\widehat{M(p)}$	$M(p) - \widehat{M(p)}$
$-0{,}2889$	$-0{,}2889$	$5{,}55111512312 \cdot 10^{-17}$	$0{,}4334$	$0{,}4334$	$-5{,}55111512312 \cdot 10^{-17}$

-
  - $r^Z = (0{,}3; 0{,}7)$
  - $\widetilde{r} = (0; 0{,}4194; 0; 0{,}3548; 0; 0; 0; 0; 0; 0{,}2258; 0; 0; 0; 0; 0; 0)$
  - $L_{ATE}(\widetilde{r}) = 0{,}6452$
  -

$m(p)$	$\widehat{m(p)}$	$m(p) - \widehat{m(p)}$	$M(p)$	$\widehat{M(p)}$	$M(p) - \widehat{M(p)}$
$-0{,}129$	$-0{,}129$	0	$0{,}6452$	$0{,}6452$	0

El trabajo de Pearl nos da las siguientes fórmulas explícitas para $m(p)$ y $M(p)$, representadas en el Cuadro 4.4 y el Cuadro 4.5, respectivamente:

Condiciones	$m(p)$
$p_{11,1} \geq p_{11,0}$ $p_{10,1} + p_{01,1} \geq p_{10,0}$ $p_{00,0} \geq p_{00,1}$ $p_{10,0} + p_{01,0} \geq p_{01,1}$	$p_{11,1} + p_{00,0} - 1$
$p_{11,0} \geq p_{11,1}$ $p_{10,0} + p_{01,0} \geq p_{10,1}$ $p_{00,1} \geq p_{00,0}$ $p_{10,1} + p_{01,1} \geq p_{01,0}$	$p_{11,0} + p_{00,1} - 1$
$p_{11,0} \geq p_{11,1} + p_{01,1}$ $p_{10,1} \geq p_{10,0} + p_{01,0}$	$p_{11,0} - p_{11,1} - p_{01,1} - p_{10,0} - p_{01,0}$
$p_{11,1} \geq p_{11,0} + p_{01,0}$ $p_{10,0} \geq p_{10,1} + p_{01,1}$	$p_{11,1} - p_{11,0} - p_{01,1} - p_{10,1} - p_{01,0}$
$p_{11,0} + p_{01,0} \geq p_{11,1} \geq p_{11,0}$ $p_{10,0} + p_{01,1} \geq p_{00,1} \geq p_{00,0}$	$-p_{10,1} - p_{01,1}$
$p_{11,1} + p_{01,1} \geq p_{11,0} \geq p_{11,1}$ $p_{10,1} + p_{00,1} \geq p_{00,0} \geq p_{00,1}$	$-p_{10,0} - p_{01,0}$
$p_{01,0} \geq p_{10,1} + p_{01,1}$ $p_{00,1} \geq p_{10,0} + p_{00,0}$	$p_{00,1} - p_{10,1} - p_{01,1} - p_{10,0} - p_{00,0}$
$p_{01,1} \geq p_{10,0} + p_{01,0}$ $p_{00,0} \geq p_{10,1} + p_{00,1}$	$p_{00,0} - p_{10,0} - p_{01,0} - p_{10,1} - p_{00,1}$

Cuadro 4.4: Cotas mínimas para $L_{ATE}(\widetilde{r})$, sujeto a $L_o(\widetilde{r}) = \widetilde{p}$

Condiciones	$M(p)$
$p_{10,1} \geq p_{10,0}$ $p_{11,1} + p_{00,1} \geq p_{11,0}$ $p_{01,0} \geq p_{01,1}$ $p_{11,0} + p_{00,0} \geq p_{00,1}$	$1 - p_{10,1} - p_{01,0}$
$p_{10,0} \geq p_{10,1}$ $p_{11,0} + p_{00,0} \geq p_{11,1}$ $p_{01,1} \geq p_{01,0}$ $p_{11,1} + p_{00,1} \geq p_{00,0}$	$1 - p_{10,0} - p_{01,1}$
$p_{10,0} \geq p_{10,1} + p_{00,1}$ $p_{11,1} \geq p_{11,0} + p_{00,0}$	$-p_{10,0} + p_{10,1} + p_{00,1} + p_{11,0} + p_{00,0}$
$p_{10,1} \geq p_{10,0} + p_{00,0}$ $p_{11,0} \geq p_{11,1} + p_{00,1}$	$-p_{10,1} + p_{11,1} + p_{00,1} + p_{11,0} + p_{00,0}$
$p_{10,0} + p_{00,0} \geq p_{10,1} \geq p_{10,0}$ $p_{11,0} + p_{01,0} \geq p_{01,1} \geq p_{01,0}$	$p_{11,1} + p_{00,1}$
$p_{10,1} + p_{00,1} \geq p_{10,0} \geq p_{10,1}$ $p_{11,1} + p_{01,1} \geq p_{01,0} \geq p_{01,0}$	$p_{11,0} + p_{00,0}$
$p_{00,0} \geq p_{11,1} + p_{00,1}$ $p_{01,1} \geq p_{11,0} + p_{01,0}$	$-p_{01,1} + p_{11,1} + p_{00,1} + p_{11,0} + p_{01,0}$
$p_{00,1} \geq p_{11,0} + p_{00,0}$ $p_{01,0} \geq p_{11,1} + p_{01,1}$	$-p_{01,0} + p_{11,0} + p_{00,0} + p_{11,1} + p_{01,1}$

Cuadro 4.5: Cotas máximas para $L_{ATE}(\widetilde{r})$, sujeto a $L_o(\widetilde{r}) = \widetilde{p}$

- $r^Z = (0{,}1; 0{,}9)$
- $\widetilde{r} = (0{,}3704; 0; 0; 0; 0{,}2222; 0; 0; 0; 0; 0; 0; 0; 0{,}4074; 0; 0; 0)$
- $L_{ATE}(\widetilde{r}) = 0$
-

$m(p)$	$\widehat{m(p)}$	$m(p) - \widehat{m(p)}$	$M(p)$	$\widehat{M(p)}$	$M(p) - \widehat{M(p)}$
$-0{,}4074$	$-0{,}4074$	0	0,3704	0,3704	0

4.4.5. Comparación con las cotas dadas por el caso simple

Recordemos el *toy example* presentado al principio de la sección, en la que tenemos sólo tratamiento y respuesta. Las cotas para el parámetro causal son las dadas por (1.18), donde llamaremos a $a(p)$ a la cota mínima y $b(p)$ a la cota máxima.

Una pregunta válida es: Si, en definitiva, se busca acotar el efecto del tratamiento que realmente **recibió** cada persona (dado por X) ¿Qué utilidad tiene incluir el tratamiento **asignado** en el modelo (dado por Z)? O, apuntando a nuestra resolución, ¿incluir a Z mejora las cotas que se obtienen en el caso simple?

En efecto, nuestras simulaciones demuestran que las cotas dadas por nuestro algoritmo quedan comprendidas entre las cotas del caso simple, resultando así cotas más finas para el ATE (también en el caso binario).

Compararemos entonces las cotas dadas por el estimador con las dadas en (1.18):

- $r^Z = (0{,}6; 0{,}4)$
- $\widetilde{r} = (0; 0{,}0667; 0{,}0444, 0{,}0778; 0{,}1222; 0{,}2222; 0; 0; 0; 0{,}0778; 0{,}1333; 0; 0{,}0778; 0; 0{,}1; 0{,}0778)$
- $L_{ATE}(\widetilde{r}) = 0{,}089$
-

$a(p)$	$m(p)$	ATE	$M(p)$	$b(p)$
$-0{,}48218$	$-0{,}2889$	0,089	0,4334	0,51782

- $r^Z = (0{,}3; 0{,}7)$
- $\widetilde{r} = (0; 0{,}4194; 0; 0{,}3548; 0; 0; 0; 0; 0; 0{,}2258; 0; 0; 0; 0; 0; 0)$
- $L_{ATE}(\widetilde{r}) = 0{,}6452$
-

$a(p)$	$m(p)$	ATE	$M(p)$	$b(p)$
$-0{,}3548$	$-0{,}129$	0,6452	0,6452	0,6452

- $r^Z = (0{,}1; 0{,}9)$
- $\widetilde{r} = (0{,}3704; 0; 0; 0; 0{,}2222; 0; 0; 0; 0; 0; 0; 0; 0{,}4074; 0; 0; 0)$
- $L_{ATE}(\widetilde{r}) = 0$
-

$a(p)$	$m(p)$	ATE	$M(p)$	$b(p)$
$-0{,}60738$	$-0{,}4074$	0	0,3704	0,39262

4.4.6. Error cuadrático medio

Luego de saber cómo generar muestras de tamaño n a partir de p_Z y p_R, hicimos una simulación de Monte Carlo.

- Tomando como parámetro $k = \#\{y_0, y_1, ..., y_{k-1}\}$ creamos la matriz de restricciones A
- Considerando los coeficientes $\{a_{j1}\}_{0\leq j\leq k-1}$ (para $P(Y_{x_1} = y_j)$) y $\{a_{j0}\}_{0\leq j\leq k-1}$ (para $P(Y_{x_0} = y_j)$), con ellos construimos el vector c que determina la función objetivo.
- Dado p_Z y p_R (también tomados como parámetros), construimos $\widetilde{p}$ el vector de probabilidades condicionales de $(X, Y)|Z$ y resolvemos los problemas de programación lineal dados en 4.13.
- Una vez resueltos estos problemas, tenemos $m(p)$ y $M(p)$ los respectivos valores óptimos para el funcional usando p_Z y p_R.
- Luego de obtener estos valores, generamos J muestras de tamaño n para (X, Y, Z) usando los mismos valores de p_Z y p_R (los valores de J y n son tomados también como parámetro del algoritmo). Es decir, para cada $1 \leq j \leq J$ tenemos una muestra $(X_i^{(j)}, Y_i^{(j)}, Z_i^{(j)})$, donde $1 \leq i \leq n$.
- Para cada muestra, calculamos la empírica $\widetilde{p}_n^{(j)}$ para las probabilidades condicionales de $(X, Y)|Z$ y resolvemos:

$$\left\{ \begin{array}{c} \min c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p}_n^{(j)} \end{array} \right\} \quad \left\{ \begin{array}{c} \max c \cdot \widetilde{r} \\ \\ A \cdot \widetilde{r} = \widetilde{p}_n^{(j)} \end{array} \right\}$$

- Este procedimiento nos da los óptimos $m(\widetilde{p}_n^{(j)})$ y $M(\widetilde{p}_n^{(j)})$ para cada problema de programación lineal.
- Como los datos fueron generados con p_Z y p_R, estos valores deberían estar cerca de $m(\widetilde{p})$ y $M(\widetilde{p})$. Para eso, estimamos el Error cuadrático medio de la siguiente forma:

$$\widehat{ECM}(m(\widetilde{p}), \widehat{m(\widetilde{p})}) = \sum_{j=1}^{J} \frac{(m(\widetilde{p}_n^{(j)}) - m(\widetilde{p}))^2}{J}$$

$$\widehat{ECM}(M(\widetilde{p}), \widehat{M(\widetilde{p})}) = \sum_{j=1}^{J} \frac{(M(\widetilde{p}_n^{(j)}) - M(\widetilde{p}))^2}{J}$$

Daremos los resultados para algunas de las pruebas hechas:

- $r^Z = (0{,}4; 0{,}6)$
- $\widetilde{r} = (0{,}0926; 0; 0{,}1111; 0; 0; 0; 0{,}1481; 0{,}1852; 0{,}1667; 0; 0; 0; 0; 0; 0{,}1481; 0{,}1482)$
- $Rango(Y) = \{0, 1\} = \{y_0, y_1\}$
- $\{a_{j1}\} = y_j,\ \forall 0 \leq j \leq 1$
- $\{a_{j0}\} = -y_j,\ \forall 0 \leq j \leq 1$
J	n	$\widehat{ECM}(m(\widetilde{p}), \widehat{m(\widetilde{p})})$	$\widehat{ECM}(M(\widetilde{p}), \widehat{M(\widetilde{p})})$
100	100	0,0049467092110743	0,0113445583536542

- $r^Z = (0{,}7; 0{,}3)$
- $\widetilde{r} = (0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0{,}09; 0; 0; 0; 0{,}06; 0; 0{,}07; 0{,}07; 0{,}07; 0{,}08;$ $0; 0; 0; 0{,}06; 0{,}07; 0{,}09; 0; 0{,}12; 0; 0; 0; 0{,}08; 0; 0; 0{,}14)$
- $Rango(Y) = \{0, 1, 2\} = \{y_0, y_1, y_2\}$
- $\{a_{j1}\} = y_j,\ \forall 0 \leq j \leq 2$
- $\{a_{j0}\} = -y_j,\ \forall 0 \leq j \leq 2$
J	n	$\widehat{ECM}(m(\widetilde{p}), \widehat{m(\widetilde{p})})$	$\widehat{ECM}(M(\widetilde{p}), \widehat{M(\widetilde{p})})$
100	100	0,0317473443343451	0,0300438179915405

- $r^Z = (0{,}2; 0{,}8)$
- $\widetilde{r} = (0; 0; 0; 0{,}2; 0{,}23; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0{,}3;$ $0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0{,}27; 0; 0)$
- $Rango(Y) = \{0, 1, 2\} = \{y_0, y_1, y_2\}$
- $\{a_{j1}\} = y_j,\ \forall 0 \leq j \leq 2$
- $\{a_{j0}\} = -y_j,\ \forall 0 \leq j \leq 2$
J	n	$\widehat{ECM}(m(\widetilde{p}), \widehat{m(\widetilde{p})})$	$\widehat{ECM}(M(\widetilde{p}), \widehat{M(\widetilde{p})})$
500	100	0,017098975959515	0,0394460172194809

- $r^Z = (0{,}1; 0{,}9)$
- $\widetilde{r} = (0; 0; 0; 0; 0{,}29; 0; 0; 0; 0; 0{,}31; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0{,}2;$ $0; 0; 0; 0{,}2; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0)$
- $Rango(Y) = \{0, 1, 2\} = \{y_0, y_1, y_2\}$
- $\{a_{j1}\} = y_j^2,\ \forall 0 \leq j \leq 2$
- $\{a_{j0}\} = -y_j^2,\ \forall 0 \leq j \leq 2$
J	n	$\widehat{ECM}(m(\widetilde{p}), \widehat{m(\widetilde{p})})$	$\widehat{ECM}(M(\widetilde{p}), \widehat{M(\widetilde{p})})$
500	500	0,108822931042428	0,093379333184444

Capítulo 5

Conclusiones y trabajos futuros

Hemos abarcado una ínfima parcela de la inmensa pradera que es la inferencia causal. Sin embargo, sale a las claras que es una área de numerosas aplicaciones. Éstas siempre requieren muchas consideraciones extras a la hora de coordinar el modelo con la realidad. Hemos tratado de esbozar algunos de los conceptos básicos de esta disciplina, para dar una idea de la vasta extensión de la misma.

Con las generalizaciones que hemos logrado para obtener las cotas, damos una mayor libertad al investigador a la hora de fijar tanto el parámetro causal como los valores de la variable respuesta. Dependiendo del tratamiento y la enfermedad que se quiera estudiar, puede ser muy útil esta variedad de opciones. Por ejemplo, el hecho de determinar sólo dos respuestas posibles a un tratamiento es un tanto restrictivo, ya que la respuesta de un paciente puede tener varios matices que necesiten ser considerados. Además, la generalización del parámetro causal permite asignar una función de peso para cada posible respuesta, priorizando algunas que puedan resultar de mayor importancia.

En los últimos años, el área ha adquirido un desarrollo descomunal. Esta tesis es un punto de partida para introducirse en la inferencia causal y a futuro, hay un mundo de extremo interés por descubrir.

Bibliografía

[1] A. Balke and J. Pearl, 1993. Nonparametric bounds on causal effects from partial compliance data. In *Journal of the American Statistical Association*.

[2] Laura Cacheiro, 2011. Introducción a la inferencia causal. Disponible en `http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2011`

[3] Z. Cai, M. Kuroki, J. Pearl, and J. Tian, 2008. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701.

[4] E. Castillo, A.J. Conejo, P. Pedregal, R. Garcıa, and N. Alguacil, 2002. Formulación y resolución de modelos de programación matemática en ingeniería y ciencia.

[5] A.P. Dawid, 1979. Conditional independence in statistical theory.

[6] C.A.P. Dawid et al, 2007. Fundamentals of statistical causality.

[7] M.A. Hernan, 2004. A definition of causal effect for epidemiological research. *Journal of epidemiology and community health*, 58:265–271.

[8] M.A. Hernán and J.M. Robins, 2006. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586.

[9] P.W. Holland, 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81:945–960.

[10] D. Geiger, T. Verma and J. Pearl, 1990. Identifying independence in Bayesian networks. *Networks*, 20:507–534.

[11] Steffen Lauritzen, 2011. Graphical models HT 2011. Disponible en `http://www.stats.ox.ac.uk/ steffen/teaching/grad/index.htm`

[12] S.L. Morgan and C. Winship, 2007. *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge University Press.

[13] J. Pearl, 2000. *Causality: models, reasoning and inference.* Cambridge Univ Press.

[14] J. Pearl, 2012. The causal foundations of structural equation modeling. Technical report, DTIC Document.

[15] J. Robins and T. Richardson, 2010. Alternative graphical causal models and the identification of direct effects.

[16] Andrea Rotnitzky, 2009. A tutorial on causal inference. Disponible en `http://www.matematica.uns.edu.ar/XCongresoMonteiro/spanish.htm`

[17] D.B. Rubin, 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688.

[18] D.B. Rubin, 1980. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

[19] J. Tian and J. Pearl, 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313.

[20] T. Verma, & J. Pearl, 1990. Causal networks: Semantics and expressiveness. In *Uncertainty in artificial intelligence*.