

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**MODELADO ESTADÍSTICO DEL VOTO FN EN FRANCIA:
Análisis de las configuraciones sociales de los movimientos
nativistas, autoritarios de corte populista.**

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN ACTUARÍA

PRESENTA

FERNANDO ANTONIO ZEPEDA HERRERA

ASESOR: DR. LUIS ENRIQUE NIETO BARAJAS

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**MODELADO ESTADÍSTICO DEL VOTO FN EN FRANCIA:
Análisis de las configuraciones sociales de los movimientos
nativistas, autoritarios de corte populista.**

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN ACTUARÍA

PRESENTA

FERNANDO ANTONIO ZEPEDA HERRERA

ASESOR: DR. LUIS ENRIQUE NIETO BARAJAS

“Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada “**MODELADO ESTADÍSTICO DEL VOTO FN EN FRANCIA: Análisis de las configuraciones sociales de los movimientos nativistas, autoritarios de corte populista.**”, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación”.

AUTOR

FECHA

FIRMA

A DIOS,
por todo

A MIS HERMANOS,
por tanto

A MIS PADRES,
por siempre

A TODAS LAS VÍCTIMAS DE
RACISMO, XENOFOBIA Y DISCRIMINACIÓN,
pero también a los que tienen miedo por su futuro,
para que puedan encontrar esperanza sin odio.

VSC

Agradecimientos

Índice general

Agradecimientos	III
Introducción	VII
I Los movimientos NAP y el caso del FN	1
1. Los movimientos NAP	2
1.1. Definiendo una familia política	5
1.2. El Front National como movimiento NAP	7
2. El FN dentro del sistema político francés	10
2.1. Recuento del sistema político	10
2.1.1. Semipresidencialismo francés	11
2.1.2. División territorial	14
2.1.3. Principales elecciones	17
2.2. Breve historia del Front National	19
2.2.1. El FN del desierto (1972-1983)	20
2.2.2. El FN Lepenista (1984-2010)	21
2.2.3. El FN Marinista (2011-2018)	24
3. Teorías sobre el voto NAP	28
3.1. Estructuralismo y Culturalismo	29
3.2. Clivajes y Escolaridad	30
3.3. La Clase Desplazada por la Globalización	31
3.4. Teorías del Conflicto	33

II	Paradigma estadístico bayesiano	37
4.	La probabilidad no existe	38
4.1.	¿Probabilidad o probabilidades?	38
4.2.	Aprendizaje bayesiano	41
4.3.	Distribuciones iniciales	44
4.3.1.	Distribuciones no informativas	45
4.3.2.	Distribuciones informativas	47
4.3.3.	Distribuciones Conjugadas	49
5.	Modelos de Regresión Bayesianos	52
5.1.	Regresión lineal	52
5.1.1.	Problema de multicolinealidad	55
5.2.	Modelos Lineales Generalizados	56
5.2.1.	Regresión logística	59
5.2.1.1.	Problema analítico	61
5.3.	Modelos Jerárquicos	63
5.3.1.	Intercambiabilidad	65
5.3.2.	Regresiones jerárquicas	67
6.	Cómputo bayesiano	70
6.1.	Monte Carlo	71
6.2.	MCMC	74
6.2.1.	Metropolis Hastings	75
6.2.2.	Gibbs Sampler	81
6.2.3.	Convergencia	86
6.3.	Hamiltonian Monte Carlo	92
6.3.1.	Sistemas físicos	94
6.3.2.	HMC ideal y en la práctica	96
III	Modelado de Datos Franceses	104
7.	Datos franceses	105
7.1.	Datos electorales	105
7.2.	Datos censales	110
7.3.	Otros datos a nivel comuna	117

7.4. Muestra de comunas	120
7.5. Asociaciones	122
8. Modelado	126
8.1. Distribuciones inciales	127
8.2. Modelos individuales	131
8.2.1. WAIC	135
8.3. Modelos Compuestos	136
Anexos	145
A. Análisis bayesiano del modelo lineal normal	147
B. Modelos Compuestos	154
B.1. Modelo A	154
B.2. Modelo B	155
B.3. Modelo C	156
B.4. Modelo D	157
B.5. Modelos E y F	158
B.6. Modelo G	159
B.7. Modelo H	161
Referencias	162

Introducción

Al cursar la Licenciatura en Actuaría elegí el área de concentración en estadística pues considero que esta es el puente perfecto entre mi inclinación cuantitativa y mi gusto por las ciencias sociales. Sin perder formalidad y rigor matemáticos, resulta ser ampliamente aplicable a problemas reales. Bien dijo John Tukey que la mejor parte de ser un estadístico es que puedes jugar en el patio trasero de todo mundo; el mío son los fenómenos políticos.

Además de Actuaría también cursé en el ITAM la Licenciatura en Relaciones Internacionales. Por ello, me resultó natural iniciar como tesis de ambas carreras un estudio estadístico relacionado con uno de los fenómenos políticos internacionales que más han despertado interés en los últimos años: aquellos movimientos populistas que usualmente son llamados de *derechas extremas o radicales*.

Desarrollada particularmente en Europa, ni la academia ni los medios de comunicación logran un consenso sobre cómo llamar o clasificar a esta familia política. A pesar de ello, de acuerdo con Cas Mudde, es posible definir a la familia política con base en tres características básicas (Mudde 2007; Beauchamp 2016a). En primer lugar, comparten el nativismo como forma xenófoba de nacionalismo. En segundo lugar, son movimientos autoritarios, que privilegian el discurso de la seguridad, la ley y el orden. Finalmente, los une el populismo como forma no solo de hacer política sino, más fundamentalmente, de ver a la sociedad: frente a las élites corruptas—*el ellos*—el pueblo puro—*el nosotros*—. Mudde los llama movimientos de derecha radical populistas aunque aquí los llamaré NAP—nativistas autoritarios de corte populista— con el objetivo de resaltar sus características fundamentales.

De manera particular, el partido que pudiera considerarse *pater familias* de esta

corriente es el *Front National* francés (Mudde 2007). Este fue fundado, entre otros personajes, por Jean-Marie Le Pen en 1972. Surgió en la escena política en 1984 al obtener algo más de 2 millones de sufragios en las elecciones europeas de ese año (Le Bras 2015). En 2002, Le Pen avanzó de manera sorpresiva a la segunda vuelta presidencial, aunque finalmente perdió frente a Jacques Chirac. En los últimos años, ya con Marine Le Pen—hija del fundador— como lideresa, el FN ha roto el bipartidismo en Francia (Le Bras 2016); en las elecciones presidenciales de 2017 ella también avanzó a la segunda vuelta presidencial. En junio de 2018 el partido cambió oficialmente de nombre a *Rassemblement National* con miras a seguir creciendo en medio de la ola nativista que pareciera darse en el mundo occidental.

Las preguntas respecto a los movimientos NAP son múltiples. ¿Por qué la gente vota por estas propuestas, siendo que algunas son abiertamente racistas? ¿Qué tipo de votante los ha apoyado y cuál los rechaza? ¿Dónde han tenido más votos? ¿Cuáles son los terrenos fértiles que han permitido este surgimiento y cuáles han fungido como barreras a su crecimiento? De hecho, es la familia política más estudiada en los últimos años (Mudde 2016).

Siendo el referente tradicional de los movimientos NAP, muchos libros y artículos se han publicado en la búsqueda por entender al votante y la sociología política del Front National. La mayoría de ellos se aproximan a las preguntas mediante estudios cualitativos o con base en datos individuales provenientes de encuestas de representatividad nacional.

Esta tesis pretende complementar el entendimiento del fenómeno mediante un estudio de caso, aunque con una estrategia distinta de modelado estadístico jerárquico de datos ecológicos. A partir de datos censales agregados, buscaré describir las *configuraciones sociales* que favorecieron o inhibieron el voto por la candidatura frontista de Marine Le Pen en la elección presidencial de 2012.

Este enfoque, como apunta Joël Gombin, tiene la ventaja de reconocer y modelar la variabilidad territorial de los fenómenos políticos, así como la importancia del contexto social en el que los individuos se desarrollan. Empero, se debe tener en cuenta que, a causa de la naturaleza agregada de los datos, las conclusiones derivadas del modelado estadístico de los mismos tiene que ser matizada para evitar caer en una falacia ecológica. No obstante lo anterior, los resultados pueden y deben considerarse a la luz de otras investigaciones que sí permiten obtener conclusiones a nivel individual (Gombin 2005).

Ahora bien, debido a que este trabajo es un proyecto de titulación para dos carreras pertenecientes a áreas del conocimiento distintas, decidí estructurarlo en 3 apartados relativamente independientes entre sí. Esto quiere decir que dependiendo del interés de cada lector, es posible concentrarse en cada uno de ellos por separado, estudiarlos en un orden diferente al presentado u omitir alguna de las 3 partes.

En primer lugar, considero que cualquier análisis estadístico aplicado tiene que estar acompañado de un conocimiento general del problema en cuestión. Es necesario entonces contar con un marco teórico sobre los movimientos NAP en general y el Front National en particular. Este es el objetivo de la primera parte. El capítulo 1 define en términos más formales a los movimientos NAP y presenta una primera referencia de lo que caracteriza al FN como un partido de dicha corriente política. El capítulo 2 tiene como objetivo familiarizar al lector con el sistema político francés así como con la historia del FN dentro de dicho sistema. El tercer capítulo, por su parte, recoge una serie de teorías presentes en la literatura sobre las razones por las cuales las personas votan por estos movimientos y partidos. Es con relación a estas teorías que busqué obtener las variables para el modelo estadístico.

La segunda parte de la tesis constituye el marco teórico estadístico en el que baso el análisis de los datos franceses. Así como es importante conocer el contexto al que aplicaremos la estadística— nuestro patio de juego— es fundamental el contexto estadístico por sí mismo, sobre todo en una tesis de Actuaría. Así pues, esta segunda parte me permitirá explorar un paradigma estadístico no siempre visto en la licenciatura, pero que tuve la fortuna de conocer a través de materias optativas: la estadística bayesiana.

En este sentido, el capítulo 4 es una introducción a dicho paradigma. En él presento la interpretación bayesiana de la probabilidad como medida de incertidumbre y no solo de variabilidad, comparándola con las más conocidas interpretaciones clásica y frecuentista. Asimismo, menciono algunas estrategias para asignar distribuciones iniciales que puedan ser actualizadas a la luz de los datos mediante el proceso de aprendizaje bayesiano y así analizar los fenómenos de interés con base en la distribución posterior.

El capítulo 5, por su parte, constituye una aproximación teórica a los modelos objeto de esta tesis. Comienzo exponiendo el modelo de regresión lineal y continúo presentando

la regresión logística como un caso particular de un modelo lineal generalizado. Finalmente discuto el modelado jerárquico o multinivel como importante alternativa a los más conocidos enfoques de regresión cuando se están estudiando fenómenos con subpoblaciones provenientes de una misma población general. En lugar de realizar una agrupación completa de los datos mediante una sola regresión poblacional o de presentar tantas regresiones independientes como subpoblaciones hayan, el modelado jerárquico busca realizar una agrupación parcial a través del concepto de intercambiabilidad.

Debido a las implicaciones que conlleva modelar desde la perspectiva bayesiana, el capítulo 6 recorre algunos de los métodos computacionales que permiten llevar a cabo inferencias estadísticas bayesianas en la práctica. De manera particular, recorro un poco la historia y características de los dos algoritmos más conocidos de MCMC: Metropolis-Hastings y *Gibbs Sampler*. Este recorrido busca llevar al lector a entender de mejor manera un atractivo método de MCMC, relativamente más reciente y desconocido, llamado *Hamiltonian Monte Carlo* y que será el utilizado en la última parte de la tesis.

Finalmente, la tercera parte de la tesis constituye propiamente el modelado de los datos franceses. Una vez que se cuenta con ambos marcos teóricos—el cualitativo y el cuantitativo—procedo al análisis. El capítulo 7 presenta los datos de manera descriptiva y mediante un análisis exploratorio. El capítulo 8 es el proceso de modelado en sí: contiene los diferentes modelos aplicados a los datos franceses y la discusión de por qué fueron considerados hasta llegar a un modelo final con base en el WAIC como medida de pérdida esperada. Por último, el noveno capítulo contiene la interpretación de los resultados de dicho modelo final y las conclusiones generales del proyecto.

Parte I

Los movimientos NAP y el caso del FN

Capítulo 1

Los movimientos NAP

En 2017, el diccionario Cambridge declaró al *populismo* como la palabra del año (Mudde 2017). Se percibiría entonces una *ola populista* a nivel mundial como uno de los temas más interesantes para estudiar sobre política internacional. El triunfo de Donald Trump en las elecciones de 2016 en Estados Unidos supuso, al menos en términos de percepción, su expansión y manifestación más importante no solo por la relevancia que reviste la nación norteamericana sino también por la atención mediática que generó en todo el mundo, las posibles consecuencias para el sistema internacional, el Estado y las vidas concretas de miles de seres humanos.

Ahora bien, Trump no es el único ni el primero de estos populistas. Los movimientos y partidos políticos de esta ola son altamente variables. La primera diferencia entre ellos es la cantidad de apelativos con la que son identificados. Es posible encontrar etiquetas tan variadas como *nacionalismo populista*, *tribalismo reaccionario*, *neopopulismo* o, incluso, *neofascismo*, *postfascismo* o *neonazismo* (Mudde 2007; Mammone, Godin y Jenkins 2012; Hainsworth 2016a); etiquetas que, en la mayoría de las ocasiones, los propios partidos rechazan (Le Parisien 2013; Hainsworth 2016a; Sputnik 2017).

Independientemente del debate sobre el término correcto, en general hay un acuerdo sobre cuáles movimientos pertenecen al núcleo de esta familia política. Esto permite ilustrar otra de sus características altamente variable: la geografía. Como mencionaba en la introducción, es un partido europeo—el Front National francés—el referente de esta familia política (Mudde 2007). Además, podemos mencionar dentro de la categoría a

varios partidos que han tenido éxitos electorales recientes y que motivaron la elección del populismo como palabra del año: el UKIP de Nigel Farage como impulsor del Brexit; la Lega de Matteo Salvini en Italia, socio principal del gobierno del M5S; el FPÖ austriaco, formando parte de la coalición gobernante surgida de las parlamentarias de 2017; el PVV holandés de Geert Wilders, quien llegó a ser líder en las encuestas hacia la elección de 2017 o la AfD alemana, misma que consiguió lugares en el Bundestag, gracias a sus buenos resultados particularmente en el este del país.

Otros ejemplos de partidos de esta corriente en regiones europeas tales como Escandinavia, Europa del Este o los Balcanes se pueden consultar en Mammone, Godin y Jenkins (2012) o Hainsworth (2016a). Pero, como sugieren el hablar de una *ola populista* o el triunfo de Trump, el fenómeno no es exclusivamente europeo, pues se da en partidos de Australia o Indonesia y líderes en el poder como Narendra Modi en India, Jair Bolsonaro en Brasil, Rodrigo Duterte en Filipinas o Benjamin Netanyahu en Israel.

A su vez, dicha extensión geográfica muestra que estos movimientos se pueden encontrar en países institucionalmente muy disímiles. Los hay bajo monarquías parlamentarias, sistemas federales o regímenes más centralizados; con elecciones de mayoría relativa, de representación proporcional o mixtas; con una o dos vueltas. Otra característica cambiante es que tampoco han estado exentos de evoluciones temporales, ni son un fenómeno completamente nuevo. Mientras que en los años 80 la mayoría de estos partidos tenían un programa económico marcadamente favorable al libre mercado, en los años recientes han migrado hacia posiciones mucho más proteccionistas (Mudde 2016).

A pesar de estas diferencias, es posible entender al fenómeno como una sola familia política trasnacional. Aunque hay que reconocer que, incluso si existen referencias que estudian su desarrollo fuera de Europa — por ejemplo, Cox y Durham (2016)— la realidad es que las caracterizaciones tradicionales que uno puede encontrar en la literatura frecuentemente se basan en elementos que aluden explícitamente al viejo continente.

Por ejemplo, Mammone, Godin y Jenkins (2012), siguiendo al sociólogo Alain Bihl, señalan dos aspectos que consideran fundamentales, aunque aparentemente contradictorios. Con la creciente globalización, los Estados-Nación se percibirían como inoperantes e impotentes ante las decisiones tomadas “en otro lado, por otros”; por ejemplo, en Bruselas por la Unión Europea. En consecuencia, estos partidos buscan resaltar a la Nación.

Sin embargo, al mismo tiempo añoran o reclaman a Europa como hogar. Esta sería una Europa blanca, opuesta a la globalización, la hegemonía de EUA, la pluralidad étnica y el Islam. Así, Mammone, Godin y Jenkins encuentran que los partidos han buscado dejar el simple nacionalismo para hablar de un “nacionalismo europeo”.

Por otro lado, estos movimientos podrían analizarse como una familia política porque comparten, según los autores, tres pilares básicos: la idea de la inequidad y la jerarquía, un nacionalismo étnico vinculado a la comunidad mono racial y la adopción de medios radicales para lograr sus objetivos y defender a la comunidad imaginada. De entre estos aspectos el más claro y relevante podría ser el etnonacionalismo. Los inmigrantes—no blancos y, en general, provenientes de ex colonias—serían vistos como poseedores de culturas diferentes que atentan contra la milenaria tradición nacional. Los autores consideran esta postura diferencialista como simple racismo pues cambiar el término biológico por el cultural es solo retórica.

Esta interpretación de que la retórica culturalista es más bien racismo cultural es compartida por muchos autores. Goodliffe (2017), en un foro sobre el crecimiento del populismo en el mundo, también señaló que las teorías e ideas diferencialistas son, en el fondo, manifestaciones de racismo. Hainsworth (2016a) refiere la propia visión al tiempo que presenta varias citas que apoyan esta argumentación. Por lo mismo, y sin que quede claro si el motivo es solamente aparentar ser una opción más aceptable, estos partidos habrían buscado bajar el tono de dicha retórica sustituyéndolo por un abierto populismo. El mismo Hainsworth (2016a), refiere otra serie de elementos que pudieran caracterizar a esta familia política: chauvinismo de bienestar, ley y orden, la búsqueda de un Estado fuerte, el carácter antisistema o incluso la antide democracia, entre otros.

A pesar de coincidir en algunas de estas características, no parece ser en movimientos como Syriza de Alexis Tsipras o Podemos de Pablo Iglesias en los que los medios y académicos internacionales están pensando cuando hablan de ola populista. Tampoco se piensa en populismos latinoamericanos como los de los países del ALBA ni en Morena de López Obrador. Estos cuatro ejemplos serían considerados populistas de izquierda (Mudde y Rovira Kaltwasser 2017). Por el contrario, el mayor interés está en aquellos movimientos que podríamos llamar *populismos de derecha, derechas extremas o derechas radicales*.

Por lo mismo, para poder tener un mayor entendimiento teórico del fenómeno se debe partir de una delimitación más puntual y formal de las características que hacen a un partido determinado pertenecer a esta familia política. Se debe pues, *definir* teóricamente a la familia política. Éste es el objetivo de Cas Mudde (2007).

1.1. Definiendo una familia política

Mudde busca encontrar la máxima cantidad de similitudes que puedan tener los partidos normalmente asociados al fenómeno para pertenecer a esta familia política. Así, logra construir de manera simple pero estructurada una caracterización teórica de lo que él llama los partidos de derecha radical de corte populista. Para ello, hace uso, principalmente, de tres definiciones básicas que presento a continuación.

■ **Nativismo**

El *nativismo* es una ideología que mantiene que los Estados deberán ser habitados exclusivamente por miembros de un grupo nativo— la Nación— y que los elementos no nativos— personas e ideas— amenazan fundamentalmente el Estado-Nación homogéneo.

■ **Autoritarismo**

El *autoritarismo* es una creencia en una sociedad estrictamente ordenada en la cual las violaciones a la autoridad deben ser castigadas severamente.

■ **Populismo**

El *populismo* es una ideología estrechamente centrada que considera que la sociedad está, al fin y al cabo, separada en dos grupos homogéneos al interior y antagónicos entre sí— “el pueblo puro” vs “la élite corrupta”— y que argumenta que la política debe ser una expresión de la voluntad general del pueblo.

Para Mudde, cuando se habla de nacionalismo no hay diferencia en el motivo del mismo. Este puede ser étnico o político y, de forma más usual, una mezcla de ambos. Por ello, el nacionalismo que interpreta Mudde es diferente al simple amor por la patria o patriotismo. Con esto no basta, pues, para hacer una diferencia entre los que podrían llamarse nacionalistas moderados y los nacionalistas radicales. Es necesario, entonces, definir un tipo específico de nacionalismo: el nativismo. Este es básicamente la unión de nacionalismo con xenofobia. El nativismo es una *definición mínima* que caracteriza a la familia política de interés. En este sentido, para este autor, la verdadera palabra del año

debió haber sido nativismo (Mudde 2017).

Sin embargo, se requieren dos elementos adicionales para llegar a una *definición máxima*. Los discursos de seguridad, ley y orden que imperan en esta familia política se incluyen dentro del término autoritarismo. Por su parte, el populismo es definido por Mudde como un componente ideológico y no solamente como un recurso retórico o un estilo político. Un populista venera el “sentido común” del pueblo y nada es más importante que esto, ni siquiera los derechos humanos o las garantías constitucionales.

Para Mudde, entonces, los partidos de derecha radical de corte populista son aquellos que cuentan con tres componentes ideológicos: nativismo, autoritarismo y populismo. El orden de los términos es importante pues el nativismo es la definición mínima, al agregársele el autoritarismo se obtendría la derecha radical y, entonces, el populismo debe fungir como componente adicional a dicho término principal.¹

No obstante, aquí los llamo NAP, aunque no para seguir contribuyendo a la falta de término común. Más bien me parece que es importante destacar siempre los conceptos básicos que fueron definidos sin obscurecerlos detrás de otros. No me parece necesario, para efectos de este trabajo, construir más términos sobre los ya definidos, al tiempo que es más transparente hacer mención constante al nativismo, autoritarismo y populismo que caracteriza a la familia política.² Es bajo esta delimitación de la familia política de los NAP que habré de analizar al Front National en Francia. Por ello, lo primero que debe hacerse es verificar que, efectivamente, el FN satisfaga las definiciones de un partido NAP.

¹ De hecho, para el autor, el populismo siempre es una especie de ideología huésped de otra ideología fundamental (Mudde y Rovira Kaltwasser 2017); en este caso la nativista y autoritaria.

² Más aún, la argumentación por la cual Mudde llega a la etiqueta de derecha radical de corte populista reconoce que los términos *derecha* y *radical* deben ser interpretados de manera muy particular. El término *radical*, por ejemplo, se refiere a una oposición fundamental a los valores de la democracia liberal. Desde mi punto de vista, ésta es una definición problemática pues el término es usado en política con un significado distinto. Ejemplifíco citando dos partidos políticos que enarbolan la etiqueta *radical* y que no la compartirían la visión de Mudde: la Unión Cívica Radical en Argentina y el Parti Radical de Gauche en Francia. Otra objeción más quisquillosa al uso del término escogido por Mudde es que me parece que las definiciones dadas para *derecha* y *radical* no se deducen lógicamente de las definiciones de *nativismo* y *autoritarismo*, lo que contradice el objetivo inicial de tener un mayor rigor teórico y contar con definiciones que no estiren conceptos, como diría Sartori. No obstante, reconozco la practicidad que el término colleva pues estos partidos son generalmente posicionados a la derecha de la derecha tradicional—sin importar la definición de derecha que se tenga—, como buscan Mammone, Godin y Jenkins (2012) o Hainsworth (2016a) al abogar por el término *derecha extrema*.

1.2. El Front National como movimiento NAP

El nativismo del Front National ha estado presente desde sus inicios. Por ejemplo, ya desde 1973 se observan tintes nativistas en sus diagnósticos que identificaban “la constitución de verdaderos barrios o ciudades extranjeras en Francia, elementos de fragmentación y que ponen en duda la unidad y la solidaridad de nuestro pueblo” y que exigían “poner fin a las políticas absurdas que toleran una inmigración salvaje, en condiciones materiales y morales desastrosas para los interesados y deshonrosas para nuestro país” (Mestre 2012, traducción propia). Otro ejemplo lo encontramos en un recorte de periódico sobre las elecciones legislativas de ese mismo año, en el que se lee “Contra la invasión a Francia por los indeseables” estableciendo que— a pesar de rechazar la idea de que los franceses sean xenófobos o racistas— la posición del FN es que “no es tolerable que nuestro país se haya convertido en un basurero abierto a los buenos para nada, a los defectuosos, a los delincuentes, a los criminales” (Delafoi 2017, traducciones propias).

Otro punto donde se refleja claramente el nativismo del FN es en sus frases o *slogans* de campaña. En 1978, presenta un cartel con la frase “Un millón de desempleados, son un millón de inmigrantes de más. Los franceses primero.”³. Después, en los años 80 surge el término *preferencia nacional* (Delafoi 2017). Este término incluye al chauvinismo de bienestar en el que la ayuda social está reservada a los miembros del grupo nativo. Esta posición continuó reflejada en el programa político del FN en 2007, pero reforzada, pues se propuso también incluirlo como un principio en el preámbulo de la constitución (L’Obs 2007). El término evolucionó después a *prioridad nacional* como puede leerse en la página 6 de la plataforma presidencial en 2012 (Front National 2012). Finalmente, tanto en 2007, 2012 como en 2017 las plataformas del FN propusieron eliminar la obtención de la nacionalidad francesa por derecho de suelo así como la binacionalidad, permitida únicamente para aquellos que posean otra nacionalidad europea (L’Obs 2007; Front National 2012; Nowak y Branford 2017).

Por su parte, el autoritarismo también ha estado presente en el programa político del FN. Íntimamente ligados a la inmigración dentro del discurso nativista de su líder histórico, Jean-Marie Le Pen, el crimen, la inseguridad, la ley y el orden han tenido también un lugar constante en la plataforma frontista. En 2001, a tan solo unos días de los atentados del 11 de septiembre en Nueva York, Le Pen aprovechó para señalar

³ *Un million de chômeurs, c'est un million d'immigrés en trop. Les Français d'abord.*” (Mestre 2012).

como amenazas, no realmente al terrorismo de Bin Laden sino todos los problemas de seguridad internos que él identificaba (Le Pen 2001). Baste el siguiente extracto de su discurso para exemplificarlo:

Los franceses enfrentan una dramática explosión de criminalidad, de violencia, de tráficos múltiples, el número de violaciones, de asesinatos y de actos de barbarie, así como el riesgo del terrorismo. La seguridad, sin embargo, es la primera de las libertades. Me comprometo, por una política de firmeza y de voluntad, basada sobre la tolerancia cero, a restaurar el orden y la ley y a organizar un referéndum sobre el restablecimiento de la pena de muerte para los crímenes más graves.

La propuesta de la pena de muerte— quizás la más viva prueba del autoritarismo— no desaparecería pronto del discurso del FN. El ofrecimiento de restablecerla continuó en la plataforma de Jean-Marie Le Pen en las elecciones de 2007 junto con otras propuestas como disminuir la edad de responsabilidad penal de menores a los 10 años (L'Obs 2007).

Asimismo, la sección sobre seguridad de la plataforma política de Marine Le Pen— nueva lideresa desde 2011— mantiene la línea dura de su padre: una política de tolerancia cero sería instaurada, los ataques organizados contra las fuerzas del orden serían fuertemente reprimidos, los efectivos policiales habrían de aumentar, las sanciones contra reincidentes serían acrecentadas, aquella persona condenada a un año o más de prisión por reincidencia perdería todas las prestaciones sociales y, de nueva cuenta, se propondría por referéndum reinstaurar la pena de muerte, así como la cadena perpetua como “alternativas para reforzar nuestro arsenal penal” (Front National 2012). En este mismo programa podemos encontrar más evidencia de que el autoritarismo y el nativismo están íntimamente ligados dentro del FN. Por ejemplo, el “racismo antifrancés” como motivación de un crimen o delito sería considerado como un fuerte agravante y debería acrecentar la pena.

Finalmente, quedan por presentar ejemplos del populismo frontista. En este sentido Jean-Marie Le Pen siempre buscó hacer referencia a una élite corrupta constituida por lo que llamó *le gang de l'établissement* (Leprince 2016), de manera particular por la *banda de los cuatro*, en referencia a los 4 partidos tradicionales en Francia— 2 de derecha y 2 de izquierda— (Boily 2005).

Debido a que el populista se considera intérprete de la voluntad del pueblo, misma que es un valor en sí misma, la predilección por métodos democráticos directos es un síntoma frecuente del populismo. En Jean-Marie Le Pen lo vemos con sus propuestas sobre el referendum, no solo para la pena de muerte sino para todas las reformas fundamentales (L'Obs 2007). Su carisma y autoproclamada vocación de darle voz al pueblo confirman ese populismo: en la elección presidencial del 2002 sus carteles rezaban *Le Pen, le peuple* (Gross 2016). Sin embargo, Jean-Marie Le Pen ha sido, desde mi punto de vista, ampliamente superado por su hija en términos populistas.

Marine Le Pen no solo ha mantenido el énfasis en la democracia directa y los referendums, sino que en su plataforma de 2012 propuso que el referendum fuera la *única* manera de modificar la constitución (Front National 2012, énfasis mío). La centralidad de Marine como la líder populista es clara. Los logos del partido y la flama tricolor han pasado a un segundo plano detrás del rostro omnipresente de Marine Le Pen. También decía yo, al explicar la terminología de Mudde, que el populista venera el sentido común del pueblo; nunca más clara esta posición en el FN como bajo el slogan *la force du bon sens* de las listas *Rassemblement Bleu Marine*: la fuerza del sentido común bajo el reagrupamiento azul marino, en clara referencia a la lideresa. Más aún, en los últimos años el slogan de las grandes campañas frontistas ha sido *Marine, au nom du peuple!* (Gross 2016). Uno solo puede conjeturar que el carácter personalista del movimiento se acentuará con el abandono del nombre Front National y su cambio por *Rassemblement National* que Marine Le Pen logró en 2018.

Capítulo 2

El FN dentro del sistema político francés

Una vez establecido, de manera general, el carácter NAP del Front National debemos estudiar de manera un poco más detallada a este partido. Para ello presento breves resúmenes del sistema político francés, así como de la historia del partido. Así podremos estar en mejores condiciones para entender dónde y cómo participa políticamente el FN y cuál ha sido su desarrollo general.

2.1. Recuento del sistema político

Los sistemas políticos tradicionales son el presidencialismo y el parlamentarismo (Carpizo 2004; Veser 1999). Mientras que en el primero el poder se concentra en un presidente, tradicionalmente electo por sufragio universal para un periodo fijo, en el segundo el depositario de la soberanía popular es un parlamento, por lo que el gobierno depende de la confianza de dicha asamblea (Linz 1990). Sin embargo, estos no son los únicos dos sistemas políticos de gobierno.

Existe también el llamado *régimen semi presidencial*. El término fue popularizado por el sociólogo francés Maurice Duverger y tradicionalmente se emplea para todos aquellos regímenes híbridos que no son sistemas totalmente presidenciales ni parlamentarios (Veser 1999; Carpizo 2004; Linz 1990). De entre estos, el más conocido es la actual Francia

(Carpizo 2004).

2.1.1. Semipresidencialismo francés

Para Duverger, un régimen semipresidencial como el francés tiene tres componentes principales (Veser 1999):

1. El presidente de la república es electo por sufragio universal;
2. posee considerables poderes;
3. tiene frente a él un primer ministro y gabinete que poseen poderes ejecutivos y gubernamentales y pueden permanecer en el poder solo si el parlamento no muestra su oposición a ellos.

Las primeras dos componentes tienen un carácter marcadamente presidencial. Sin embargo, el tercer punto es un contrapeso parlamentario que hace que el régimen no sea del todo presidencial. En general, los tres puntos se reflejan en la actual constitución, promulgada el 4 de octubre de 1958 y que instauró la llamada V^a República.

El primero, empero, no fue totalmente concretado sino hasta el referéndum plebiscitario de 1962 que estableció el sufragio universal directo para la elección a la presidencia de la república. Ya desde el texto original de 1958 la Asamblea Nacional era electa de manera directa por la ciudadanía, pero la primera elección presidencial de la V^a República fue mediante un colegio electoral.

Por otro lado, como sugeriría el segundo punto de Duverger, el presidente es la piedra angular del régimen francés (Assemblée National 2017a). Es el Jefe del Estado pues vela por el respeto a la Constitución, asegura el funcionamiento regular de los poderes públicos así como la continuidad del Estado, es el garante de la independencia nacional, de la integridad del territorio y del respeto de los tratados (*Constitution du 4 octobre 1958*). Algunos de los poderes más importantes del presidente francés son los siguientes (Assemblée National 2017a; Vie Publique):

- Nombra al Primer Ministro y acepta su dimisión.
- Puede disolver la Asamblea Nacional.

- En casos extremos puede tomar poderes extraordinarios.
- Puede someter a referéndum ciertos proyectos de ley.
- Tiene dos áreas de predominancia:
 - La Defensa, pues es comandante supremo de las fuerzas armadas y quien decide sobre el uso de la fuerza nuclear.
 - Las Relaciones Exteriores, pues es quien negocia y ratifica los tratados, acredita los embajadores franceses ante otros Estados y es frente a quien se acreditan los embajadores ante Francia.

Sin embargo, a diferencia de lo que sucede en un régimen estrictamente presidencial, el presidente no es el Jefe de Gobierno. De hecho, el poder ejecutivo en Francia está dividido entre la Presidencia y el Gobierno. El Primer Ministro es el Jefe de Gobierno y es este Gobierno formado por los ministros de las diferentes carteras el que determina y conduce la política, dispone de la administración pública y asegura la ejecución de las leyes (*Constitution du 4 octobre 1958*). Más aún, y en sincronía con el tercer componente de Duverger, el Gobierno es responsable frente a un Parlamento bicameral, compuesto por una Asamblea Nacional y un Senado. Sin embargo, esta división no es totalmente equitativa pues es solamente la Asamblea Nacional quien puede remover al Gobierno. Esta estructura general puede verse en el esquema de la **Figura 2.1**.

Existen tres mecanismos por los cuales la Asamblea Nacional puede obligar a un Gobierno a renunciar. El primero es usualmente llamado *voto de confianza* y se da cuando el Primer Ministro somete su programa o alguna política general a la confianza de la Asamblea que vota y si la niega, el Gobierno debe renunciar. La segunda es el *voto de censura* que se da por iniciativa de la Asamblea y en caso de aprobarse el Gobierno cae. Finalmente el Gobierno, bajo ciertas circunstancias, puede condicionar su responsabilidad a una iniciativa particular: si la Asamblea no está de acuerdo con el proyecto debe censurar al Gobierno y destituirlo o de lo contrario el texto propuesto es aprobado automáticamente (Assemblée National 2017c).

¿Qué implicaciones tiene esta configuración legal? Como bien apunta Carpizo (2004), el régimen semipresidencial francés en la práctica es un régimen de alternancias entre un sistema (casi) presidencial y uno (casi) parlamentario. Cuál de estos subsistemas imperará depende de lo que Marrani (2009) llama la sincronía dentro del ejecutivo y que

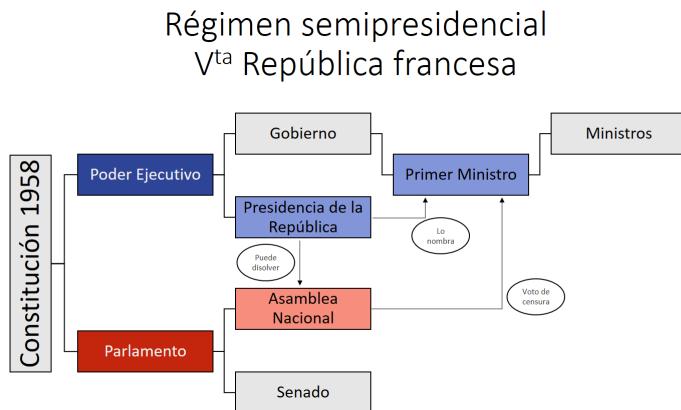


Figura 2.1: Esquema del régimen semipresidencial francés de la V^{ta} República. Los poderes ejecutivo—Presidencia de la República y Gobierno—y legislativo—Parlamento bicameral—están separados. El Presidente nombra al Primer Ministro, quien forma un Gobierno que depende de la confianza de la Asamblea Nacional, misma que puede ser disuelta por el Presidente. Fuente: elaboración propia.

depende de la configuración política de las tres instituciones resaltadas en la **Figura 2.1**: la Presidencia, la Asamblea Nacional y, por tanto, el(la) Primer(a) Ministro(a).¹

Recordemos que el Presidente es quien nombra al Primer Ministro. Cuando el Presidente logra tener de su lado a la mayoría de la Asamblea Nacional, este puede nombrar a quien quiera sin temor de que la Asamblea vaya a destituirlo. A pesar de que formalmente el Primer Ministro es quien gobierna, en la práctica este le debe deferencia al Presidente y se convierte solo en un facilitador de su política. Esta es la situación donde hay una sincronía al interior del ejecutivo y el sistema es marcadamente presidencial. Se da lo que Marrani llama *fait majoritaire* y que podríamos designar como el funcionamiento bajo mayoría presidencial.

Sin embargo, cabe la posibilidad de que la Asamblea Nacional tenga una mayoría que se oponga al Jefe de Estado. En este caso, conocido como *cohabitation*, el Presidente ya no puede nombrar a quien quiera. Debe nombrar a alguien que sea apoyado por la mayoría opositora de la Asamblea. El Primer Ministro, a diferencia de lo que sucede en un sistema puramente parlamentario, no precisa ser miembro del Parlamento; por el contrario, existe una incompatibilidad constitucional entre la función gubernamental y el mandato parlamentario (*Constitution du 4 octobre 1958*). Han habido Primeros

¹Hasta el momento solo ha habido una Primera Ministra, Édith Cresson (1991-1992).

Ministros que no eran diputados de la Asamblea y, cuando un diputado es nombrado ministro, este debe cesar sus funciones y su suplente lo reemplaza en la Asamblea. Se da una falta de sincronización al interior del ejecutivo y el Presidente debe cohabitar con un Jefe de Gobierno políticamente opuesto a él. Aquí es cuando estamos en un subsistema de corte parlamentario pues es la mayoría parlamentaria la que gobierna y, salvo en sus dominios de predominancia, los poderes del Presidente se ven fuertemente reducidos.²

En la historia de Francia han existido tres cohabitaciones: Miterrand-Chirac (1986-1988), Miterrand-Balladur (1993-1995) y Chirac-Jospin (1997-2002). Han habido también otros periodos en los que formalmente el Primer Ministro no pertenece al mismo partido político que el Presidente pero que sí forma parte de su mayoría presidencial, por lo que no podríamos llamarlos cohabitación. Estos han sido los cuatro gobiernos bajo la presidencia de Giscard d'Estaing—el primero de Chirac (1974-1976) y los tres de Barre (1976-1977, 1977-1978, 1978-1981)—así como los dos gobiernos de Philippe bajo la actual presidencia de Macron.

Una vez establecido el panorama general del sistema de gobierno en Francia, resulta pertinente hacer referencia a la división territorial del país galo.

2.1.2. División territorial

De acuerdo con el artículo 72 de la Constitución francesa,

Las colectividades territoriales de la República son las comunas, los departamentos, las regiones, las colectividades con estatus particular y las colectividades de ultramar...

Esta estructura general, puede verse en la **Figura 2.2**. Los primeros tres tipos de colectividades territoriales son los más usuales y se refieren a los 3 niveles de administración. El artículo después refiere la existencia de otros dos tipos de colectividades territoriales: las que tienen estatus particular y las de ultramar. No debe pensarse, sin embargo, que estas últimas son todas aquellas que no pertenecen a la metrópoli, identificada en la

²Su poder de veto, que no había mencionado, permanece. Sin embargo, el poder de veto en la práctica no es tan relevante como los otros puesto que en situación de mayoría presidencial no es usualmente necesario, mientras que en periodo de cohabitación es más un mecanismo de oposición legislativa que la mayoría de la Asamblea puede superar. El Presidente también conserva su poder de disolución, pero esta es una herramienta riesgosa y costosa políticamente.

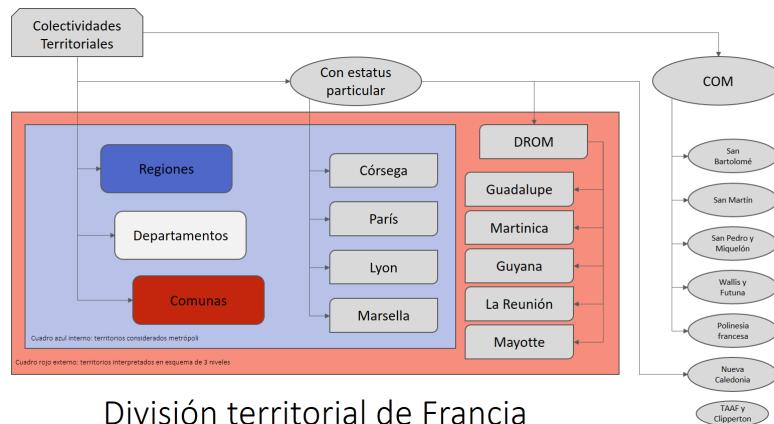


Figura 2.2: Esquema de la división territorial francesa. Fuente: elaboración propia a partir de Assemblée National (2017b).

Figura 2.2 mediante el recuadro azul interno.³

Más bien se refieren a 5 territorios específicos conocidos en Francia como *Collectivités d'outre-mer* (COM): San Bartolomé, San Martín, San Pedro y Miquelón, Wallis y Futuna, así como la Polinesia francesa. Existen otros territorios del ultramar francés diferentes a las COM. Hay territorios no habitados: la Isla de Clipperton y los Territorios australes y antárticos franceses (TAAF). Encontramos también a la Nueva Caledonia como una colectividad sui generis. Finalmente, existen otros 5 territorios de ultramar que son colectividades territoriales con estatus especial, los *Départements et Régions d'outre-mer* (DROM): Guadalupe, Martinica, Guyana, Reunión y Mayotte. Estos 5 DROM y otros 4 territorios en la metrópoli que también son colectividades con estatus especial—Córsega, París, Lyon y Marsella—usualmente son considerados dentro del sistema de 3 niveles, a pesar de las particularidades de cada uno. Así pues, el recuadro rojo externo de la **Figura 2.2** identifica a todas las colectividades territoriales que se interpretan bajo el esquema de 3 niveles.

Estos tienen las siguientes características generales (Asamblee National 2017b):

1. Comunas

Son el nivel más cercano a los ciudadanos. Existen desde 1789 cuando remplazaron

³La metrópoli incluye el territorio continental de Francia—coloquialmente conocido como el Hexágono, por su forma—y la isla de Córsega.

a las parroquias. Existen en Francia alrededor de 36,000, aunque año con año el número varía pues las comunas pueden fusionarse o separarse. Sus principales áreas de competencia son el urbanismo, la vivienda y el ambiente. Se gobiernan principalmente mediante un concejo municipal elegido popularmente y un *maire* designado por dicho concejo.

2. Departamentos

El segundo nivel también fue creado en 1789. Hasta el 31 de diciembre de 2017, existían 101 departamentos, de los cuales 96 formaban la metrópoli y los otros 5 son DROM. A partir del 1ro de enero de 2018 los dos departamentos de Córsega fueron sustituidos por una colectividad única, por lo que ahora hay 100 departamentos. Tienen dos dominios de responsabilidad: la acción social y el manejo de los espacios.⁴ Se gobiernan principalmente mediante un concejo departamental elegido popularmente y un presidente designado por dicho concejo.

3. Regiones

Es el nivel de más reciente creación, pues fue reconocido como colectividad territorial en 1982. Hasta 2015 existieron 27 regiones, 22 en la metrópoli— incluida Córsega— y las 5 DROM. En 2015 hubo una reforma que agrupó algunas, por lo que las nuevas regiones son solo 18— 13 en la metrópoli, contando Córsega, y las 5 DROM—. Las regiones están encargadas del desarrollo económico, la administración del territorio y los transportes no urbanos. Se gobiernan principalmente mediante un concejo regional elegido popularmente y un presidente designado por dicho concejo.

Las colectividades territoriales usualmente son de uno de los tres niveles. Por ejemplo, *Villemomble* es una de las 40 comunas que conforman el departamento de *Seine-Saint-Denis*. Este, a su vez, es uno de los 8 departamentos que conforman la región *Île de France*. Sin embargo, las colectividades con estatus especial pueden compartir, al mismo tiempo, dos niveles. Este es el caso de París, que también se encuentra en la región *Île de France* pero que es tanto un departamento como una comuna dividida en *arrondissements municipales*.⁵

⁴Estos términos incluyen, el primero, los temas relacionados con grupos vulnerables como la infancia, las personas con discapacidad o los adultos mayores, mientras que el segundo se refiere, por ejemplo, al control de puertos, aeropuertos o caminos dentro del territorio del departamento.

⁵Nótese que uso la palabra *pueden*; Marsella tiene un estatus especial pero es solamente una comuna.

Una vez que he presentado un resumen general del semipresidencialismo francés y de la organización del territorio, procedo a mencionar a grandes rasgos el sistema electoral del país.

2.1.3. Principales elecciones

Con base en lo que reporta el Ministerio del Interior francés (*Les différentes élections*), identifico cuatro categorías de elecciones francesas.

- Para conformar los poderes nacionales:
 - Presidenciales.
 - Legislativas, para la Asamblea Nacional.
 - Senatoriales.
- Para conformar las autoridades de los 3 niveles:
 - Municipales, a nivel comuna.
 - Departamentales, antes conocidas como cantonales.⁶
 - Regionales.
- Supranacionales:
 - Europeas, para el Parlamento europeo.
- Especiales, con diversos fines:
 - Referéndums.
 - Comunitarias, se dan al interior de las comunas.

De entre estas categorías, solo me concentraré en explicar un poco más las presidenciales y las legislativas por tratarse de las dos principales instituciones que determinan el funcionamiento del sistema semipresidencial francés. Ambas son elecciones de sufragio universal directo.

⁶Antes de una reforma de 2015 las elecciones para el nivel de departamento eran conocidas como elecciones cantonales pues los consejeros se eligen en circunscripciones llamadas cantones.

La Asamblea Nacional se conforma por 577 diputados que se eligen por períodos de 5 años con reelección.⁷ Se elige un diputado por circunscripción legislativa mediante un sistema de mayoría a doble vuelta. Esto significa que hay dos formas de ganar la elección:

1. **En la primera vuelta**, si se obtiene la mayoría absoluta de la votación efectiva, siempre y cuando los votos recibidos representen al menos un cuarto de los electores inscritos en las listas de votación.
2. **En la segunda vuelta**, donde basta la mayoría relativa. En caso de empate el(la) candidato(a) de mayor edad gana.

Pasan a la segunda vuelta todas las candidaturas que hayan obtenido un porcentaje de votación equivalente al menos al 12.5 % de los electores inscritos en las listas de votación. Esto significa que puede haber más de dos contendientes en la segunda vuelta.⁸

Por otro lado, tenemos la elección presidencial. En el siglo XXI han habido dos principales reformas constitucionales que han moldeado esta elección. La primera, en 2000, redujo el periodo presidencial de 7 a 5 años. Aunque esto pareciera a primera vista un control sobre el Presidente, en realidad es una reforma que refuerza al Jefe de Estado. Ahora los periodos presidenciales coinciden con los 5 años por los que son elegidos los diputados.⁹ Más aún, la elección presidencial se realiza unas semanas antes que la elección legislativa. Esto hace mucho menos probable que un presidente electo enfrente una cohabitación, pues su reciente triunfo tiende a impulsar una mayoría legislativa a su favor. Sin embargo, en 2008, se decidió limitar al Jefe del Estado prohibiendo que un presidente sirva más de dos mandatos.

Los presidentes también se eligen mediante mayoría a dos vueltas. Esto significa que, al igual que con los diputados, se puede ganar la elección de dos formas:

1. **En la primera vuelta**, si se obtiene la mayoría absoluta de la votación efectiva.
2. **En la segunda vuelta**, si se obtiene la mayoría absoluta de la votación efectiva.

⁷A menos que el Presidente disuelva la Asamblea antes del término de la legislatura, en cuyo caso no puede haber una nueva disolución en el año siguiente.

⁸Esto da lugar a las expresiones *duels*, entre dos candidaturas, y *triangulaires*, entre tres. Pueden también darse cuadrangulares o segundas vueltas entre más candidatos pero son mucho menos frecuentes y, cuando suceden, tienden a haber declinaciones y alianzas.

⁹A menos, claro, que el propio Presidente disuelva la asamblea, cosa que no ha ocurrido desde la reforma. De hecho, solo han existido 5 disoluciones en la V^a República, todas en el siglo XX, en 1962, 1968, 1981, 1988 y 1997.

Para garantizar que alguien gane la elección en la segunda vuelta, a diferencia de lo que pasa con las elecciones legislativas, solo pueden presentarse a la segunda vuelta las dos candidaturas más votadas en la primera. Sin embargo, existe un vacío legal en el caso de empate, pues el criterio de edad imperante en las elecciones legislativas no está contemplado para la elección presidencial (Lombart 2016).

Esta sección tuvo como objetivo proveer al lector del contexto político en Francia, presentar las estructuras básicas que se intentarán aprovechar en el análisis estadístico de los datos y facilitar la lectura de la historia del Front National, pues se hacen referencia a varios de los conceptos antes presentados y que en primera instancia pueden resultar extraños.

2.2. Breve historia del Front National

Como todo fenómeno social, el FN no ha sido un movimiento monolítico. Desde sus orígenes y hasta el día de hoy ha tenido momentos de menor y mayor éxito. Un esquema para estudiar la historia del partido, principalmente desde el punto de vista electoral, es señalar tres grandes etapas. En primer lugar, la primera década de su existencia estuvo marcada por la marginalidad política y es comúnmente llamada la *traversée du désert*. En segundo lugar, a partir de los años 80 el FN irrumpió en la escena política de la mano de su líder Jean-Marie Le Pen quien, en 2002, logró alcanzar su zenit al obtener el segundo lugar en la primera vuelta presidencial de ese año. No obstante la primera década del siglo XXI estuvo marcada por el declive de la figura de Le Pen. Por ello, en 2011 comienza la más reciente etapa del partido. Jean-Marie Le Pen pasó la batuta del liderazgo a su hija Marine Le Pen quien desde entonces ha implementado una estrategia de *dédiabolisation*, buscando acabar con la imagen negativa del partido, principalmente con respecto al carácter racista que caracterizó a su padre.

El recuento histórico que sigue está basado primordialmente en aquellos que hacen Hainsworth (2016b) y Stockemer (2017), así como en el apéndice cronológico al libro editado por Crépon, Dézé y Mayer (2015) pero algunas otras referencias se citan también.

2.2.1. El FN del desierto (1972-1983)

El Front National surgió como una iniciativa del movimiento *Ordre Nouveau* (ON) en 1972 para unir en una plataforma electoral a distintos grupúsculos¹⁰ políticos extremistas como los irredentistas de la colonización francesa en Algeria, los herederos del movimiento poujadista o del régimen de Vichy, intelectuales de la *Nouvelle Droite*, organizaciones juveniles como *Occident* o *L'Œuvre Française*, revisionistas del Holocausto, monarquistas, entre otros. Jean-Marie Le Pen fue elegido presidente del partido, en parte por ser una figura relativamente conocida— se había convertido en el diputado más joven en 1956 como parte del movimiento de Poujade y había sido el coordinador de campaña de Tixier-Vignancour en 1965— y en parte por, irónicamente, ser visto como un moderado.

Durante los años siguientes, sin embargo, el nuevo partido hiló una serie de fracasos electorales en las elecciones legislativas de 1973, las presidenciales de 1974, las legislativas de 1978 y, de manera más particular, en las presidenciales de 1981, en las cuales Le Pen no logró las firmas necesarias para participar.¹¹ Los problemas no se limitaron a las urnas. Ya desde 1973 la multiplicidad de facciones hizo que el partido se debilitara. Después de un mitin llamado *Alto a la inmigración salvaje* miembros de ON se enfrentaron en una batalla campal violenta con miembros de la Liga Comunista por lo que el movimiento fue proscrito. Esto le permitió a Le Pen tomar mayor control del partido, pero también hizo surgir en 1974 a un nuevo partido fundado por miembros de ON, el PFN (*Parti des Forces Nouvelles*). En 1976 Le Pen sobrevivió a un intento de asesinato pero su segundo al mando e ideólogo, François Duprat, fue asesinado por un coche bomba en 1978. El Front National parecía apenas sobrevivir.

Entonces, el partido comienza a poner en un segundo plano su carácter anticomunista para hacer mayor énfasis en la migración vinculando a los inmigrantes con los problemas de seguridad y desempleo. A inicios de los años 80, sus posturas nativistas, autoritarias y populistas empezaron a resonar en los votantes. El gobierno de Mitterrand implementó varias medidas para facilitar la inmigración, reducir las prerrogativas de la policía y otorgó amnistías a casi el 14 % de los prisioneros. La población francesa no fue del todo favorable a ello. Los partidos tradicionales de derecha comenzaron a endurecer sus posturas frente a la migración y la seguridad. Esto le dio mayor legitimidad al FN

¹⁰En el sentido de grupos más bien pequeños y dispersos.

¹¹En Francia, para que una persona se pueda presentar como candidata en las elecciones presidenciales requiere 500 apadrinamientos de parte de representantes electos como pueden ser alcaldes o legisladores.

pues le permitió normalizar, de cierta manera, su discurso.

A nivel local, el FN comienza a crecer en Dreux donde en las elecciones legislativas de 1981 logró obtener más de 10% de los votos. En las elecciones municipales de 1983 de París Le Pen es nombrado consejero municipal del distrito 20 al obtener 11.26% de los votos en la primera vuelta y 8.54% en la segunda. Jean-Pierre Stirbois en Dreux logra incluso un mejor resultado: 16.72% de los votos en la primera vuelta y una alianza con la derecha tradicional en la segunda vuelta lo llevan también a ser consejero municipal. El partido había atravesado un desierto político y estaba a punto de sorprender a propios y extraños.

2.2.2. El FN Lepenista (1984-2010)

La hora de la verdad de Le Pen, literalmente, llegó en 1984. El éxito en Dreux hizo que Le Pen fuera invitado en febrero de ese año al programa político de TV más visto en Francia entonces, *L'heure de verité*. La entrevista fue todo un éxito; las preferencias electorales del FN pasaron del 3.5% al 7% a la semana siguiente de la entrevista, por ejemplo (Stockemer 2017). Simbólicamente podría marcar el inicio de lo que aquí llamo la era lepenista del FN, cuya confirmación fue el 17 de junio de 1984. Las listas del FN lograron aproximadamente 11% de los votos en las elecciones al Parlamento Europeo que se transformaron, gracias al sistema proporcional, en 10 eurodiputados de extracción frontista. El FN ya no era un partido al margen del sistema político, lo que le permitió captar más liderazgos.

El más importante de esos nuevos adherentes fue, sin duda, Bruno Mégret. En 1985 se unió al partido y en 1988 se convirtió en director general. La adición de Mégret demostraría ser fundamental para la consolidación del FN como un actor político relevante en Francia. A él se le atribuye la profesionalización de los cuadros y estructuras del partido así como la popularización de la retórica frontista y, en gran medida, el que en las elecciones de la segunda mitad de los 80 el FN haya obtenido alrededor del 10% de manera consistente: 9% en las cantonales de 1985, 10% en las legislativas de 1986, 14% en las presidenciales de 1988, 9% en las legislativas de 1988, 11% en las europeas de 1989, así como el primer alcalde frontista en 1989.

En los años 90 el FN continuó— de la mano del populismo de Le Pen y la organización de Mégret— creciendo en las preferencias electorales. En 1995, los candidatos del FN se convirtieron en alcaldes en tres ciudades relevantes: Toulon, Marignane y Orange. En las elecciones presidenciales de 1995, las legislativas de 1997 y las regionales de 1998 el frontismo se afianzó alrededor del 15 % de los votos.

Empero, los éxitos no estuvieron alejados de los escándalos. A pesar de los intentos de Mégret por encuadrar el nativismo del FN en conceptos más políticamente correctos— que no más moderados— como la *preferencia nacional*, el antisemitismo de Le Pen y de otros miembros del FN era evidente. En varias ocasiones Le Pen fue centro de atención por sus comentarios antisemitas. Los más conocidos son quizás cuando dijo que el Holocausto era un “pequeño detalle” en la historia de la Segunda Guerra Mundial o cuando se burló de un ministro de nombre Michel Durafour llamándolo *Monsieur Durafour Crématoire*.¹²

Asimismo, con los éxitos también surgieron las rivalidades internas. El liderazgo de Mégret empezó a ser más evidente e incómodo para Le Pen. Ambos personajes tenían también visiones distintas sobre la estrategia del partido; por ejemplo, mientras Mégret abogaba por tejer alianzas con los partidos de la derecha tradicional, Le Pen siempre las rechazó, lo que a los seguidores de Mégret les hacía creer que en realidad no quería llegar al poder. Las tensiones fueron creciendo hasta llegar a un enfrentamiento directo entre 1998 y 1999.

En 1997, Le Pen había ido a Mantes-la-Jolie a apoyar a su hija en su campaña en las legislativas de ese año. Al llegar al lugar fue recibido por manifestantes que lo increparon, llamándolo fascista. Desde que bajaron del carro, los guardaespaldas de Le Pen comenzaron a golpear a los manifestantes y eso desató una pelea en la que el propio Le Pen se involucró. En un momento empujó contra un muro a la rival de su hija, la socialista Annette Peulvast-Bergeal (Ina Politique 2012). Por este ataque, Le Pen fue condenado en 1998 a un año de inelegibilidad, a tres meses de prisión suspendida condicionalmente y a 5,000 francos de multa (Les Echos 1999). Así pues, Le Pen estaba impedido para participar en las elecciones europeas de 1999. Mégret parecía la opción más natural para encabezar las listas del partido. Sin embargo, parafraseando a Le Pen, en el FN el único que importaba era el número uno, así que colocó a su esposa en el primer puesto en lugar de a Mégret.

¹²La palabra *four* en francés significa horno y agregó el término *crématoire*, es decir, crematorio.

En diciembre de 1998, Mégret criticó a Le Pen y dijo que se había convertido en un lastre para el partido. El comité ejecutivo del partido, controlado por Le Pen, votó para expulsar a Mégret del FN. En respuesta, Mégret fundó un nuevo partido, el MNR (*Mouvement National Républicain*). El cisma mégretista se llevó a casi la mitad de los miembros del FN y a su más importante operador político. Compitiendo contra el MNR, en las elecciones europeas de 1999 el frontismo sólo obtuvo el 5.7% de los votos.

No obstante, el más grande éxito de Jean-Marie Le Pen llegó apenas unos años después. Ante una izquierda sumamente dividida— hubo 8 candidatos de izquierda— en una elección presidencial atípica, Le Pen sacudió a propios y extraños el 21 de abril de 2002. En total 16 candidaturas lograron las firmas necesarias para competir, un récord en Francia. Nadie logró más de 20% de los votos en la primera ronda, algo también inédito. Los estrategas de los candidatos no creían lo que sus conteos rápidos les decían. A las 8 de la noche, las cadenas de televisión dieron la noticia: Jean-Marie Le Pen había quedado en segundo lugar, apenas por encima del candidato socialista, Lionel Jospin. La diferencia había sido de menos de 200 mil votos.

En Youtube se puede encontrar un documental en francés que refleja de muy buena manera cómo se vivió esa noche electoral en Francia (Capo 2017). Del lado socialista, incredulidad, sorpresa absoluta y llanto de tristeza. Del lado del FN, también hubo incredulidad, sorpresa absoluta y llanto, pero de emoción. Le Pen había avanzado a la segunda vuelta presidencial. Se enfrentaría al ganador de la primera, el presidente en funciones, Jacques Chirac. Hubo fuertes llamados a unirse en torno a Chirac para evitar que ganara el candidato FN. Ante la amenaza que— para la enorme mayoría de los franceses— representaba Le Pen, Chirac fue reelecto por un amplísimo margen: 82% vs 18%.

Así pues, 2002 fue el gran éxito de Jean-Marie Le Pen, pero también el inicio del fin: era claro que la mayor parte de la población no veía en él una opción aceptable. Los resultados electorales que siguieron fueron en declive: 10% en las europeas de 2004, 11% en las presidenciales de 2007, 4% en las legislativas de 2007, 6% en las europeas de 2009. Fiel a su costumbre, Le Pen siguió mostrando su revisionismo de la Segunda Guerra Mundial al declarar en 2005 que: “al menos en Francia, la Ocupación alemana no fue particularmente inhumana”. El partido enfrentó problemas económicos y deudas que lo obligaron a despedir a un tercio de los empleados, a cancelar su fiesta *Bleu Blanc*

Rouge, entre otras cosas. Era claro que Jean-Marie Le Pen ya no estaba en condiciones de liderar al Front National. Sin embargo, el líder histórico no habría de ceder el mando a cualquiera.

2.2.3. El FN Marinista (2011-2018)

Cuando fue momento de pasar la batuta de su movimiento, y a pesar de que existía mucho apoyo al interior del partido para que Bruno Gollnisch—un viejo lobo de mar del FN—fuera el nuevo líder, Jean-Marie Le Pen decidió apoyar a su hija, Marine Le Pen. En una elección interna anterior para el liderazgo del comité central del partido, Gollnisch había ganado mientras que Marine Le Pen había quedado en una lejana posición número 34 (Williams 2012). A pesar de ello, el 16 de enero de 2011, Marine Le Pen fue elegida presidenta del FN, mientras que Jean-Marie Le Pen fue nombrado presidente honorario del mismo.

La estrategia del *marinisme*, en contraposición a la del *lepenisme*, ha buscado *desatanizar* al partido. Marine Le Pen ha querido renovar la imagen que el FN presenta a los electores. Si el FN lepenista siempre fue muy vocal respecto a sus posiciones sobre los temas de prácticas y costumbres sociales o, como se les conoce en Francia, las *mœurs*, bajo el liderazgo marinista, la posición oficial del FN al respecto ha cambiado. El símbolo de esta inflexión en temas sociales quizás es Florian Philippot—vicepresidente del partido de 2012 a 2017—quien en 2014 declaró ser gay, algo impensable en el FN lepenista.¹³

Por otro lado, y de manera muy particular, Marine Le Pen ha cortado al interior del partido las expresiones antisemitas y abiertamente racistas. A diferencia de lo que dijo su padre, ella ha declarado que todo el mundo sabe lo que pasó en los campos nazis, en qué condiciones sucedió y cómo representó la máxima expresión de la barbarie. De hecho, a raíz de que Jean-Marie Le Pen reafirmara en 2015 el carácter de “pequeño detalle” del Holocausto, Marine Le Pen empujó y consiguió la expulsión de su padre del partido.

Esta estrategia parece haber rendido sus frutos con el electorado. A pesar de no avan-

¹³Digo inflexión porque no ha sido un cambio total: hay cuadros que mantienen posiciones firmes en contra del aborto o los matrimonios entre personas del mismo sexo, sin embargo, la postura oficial ha sido más ambigua y ha dejado de hacer énfasis en esos temas. El lector interesado en el tema de las *mœurs* dentro del FN puede consultar el texto de Crépon (2015).

zar a la segunda vuelta en 2012, Marine Le Pen superó a su padre al obtener 18 % de los votos en la primera vuelta. En las legislativas de ese año, el FN obtuvo dos diputaciones y el 14 % de los sufragios. En 2014, el partido conquistó 11 alcaldías, sus primeros dos senadores y fue la primera fuerza política votada en las elecciones europeas con 25 %. Este crecimiento ha continuado, pues en 2015 el FN obtuvo el 25 % en las elecciones departamentales y el 28 % en las regionales.

En las elecciones presidenciales de 2017, la lideresa del frontismo logró superar de nueva cuenta a su padre. Ella obtuvo 21 % de los votos de la primera vuelta, más que Jean-Marie Le Pen en 2002, lo que le permitió también avanzar a la segunda vuelta. Sin embargo en 2017, a diferencia del 2002, la presencia del FN en la segunda vuelta no fue sorpresiva sino más bien era lo esperado. Más aún, si Le Pen padre perdió dicha segunda vuelta por una diferencia de alrededor de 65 puntos porcentuales, Le Pen hija la perdió por poco más de 30. Los franceses volvieron a dejar claro que consideran, mayoritariamente, al FN como una amenaza; empero, esa percepción ha disminuido fuertemente en la era marinista del partido.

A pesar de estos éxitos de la estrategia de desdemonización, existen varios paralelismos con la historia de su padre. Así como Jean-Marie Le Pen tuvo a Bruno Mégret como su operador político, Marine Le Pen se apoyó en Florian Philippot para encuadrar su discurso y manejar al partido. Ambas manos derechas de los Le Pen, terminaron por salir del partido y fundar su alternativa política, aunque los motivos y circunstancias de los cismas fueron distintos. Como ya he dicho, Mégret fue expulsado después de desafiar directamente a Jean-Marie Le Pen por el liderazgo del partido. La renuncia de Philippot al FN, por el contrario, surge en medio de una introspección del partido y la búsqueda de culpables después de las elecciones de 2017.

Las elecciones presidenciales de 2017 marcaron la mayor cantidad de votos en la historia del FN. 33 % en la segunda vuelta no es un dato menor, sin embargo, para muchos significó un fracaso. Bruno Jeudy, Vanessa Schneider, Nonna Mayer y Jérôme Fourquet discutieron al respecto en el programa de televisión *C dans l'air (2017)*: muchas encuestas a lo largo del proceso apuntaban a que Le Pen lograría hasta 40 % de los sufragios hasta que la noche del debate entre Le Pen y Emmanuel Macron se dio un punto de inflexión. Marine Le Pen fue muy agresiva— minando, posiblemente, el esfuerzo de desdemonización— pero también se vio poco preparada. En un momento Macron le preguntó si ella

proponía abandonar el euro y regresar al franco como moneda francesa. Su respuesta fue ambigua y muy poco clara. El debate supuso un duro golpe que le habría hecho perder al menos 5 puntos porcentuales de intención de voto.

La postura de abandonar el euro no fue la más popular entre el electorado e incluso dentro de buena parte de los miembros del FN. Ésta propuesta, como la mayoría del programa económico del FN, se debe a Philippot y su corriente que podríamos llamar soberanista en contraposición a la línea identitaria más dura del partido (Marin 2017; Berteloot 2017; Europe 1 2017). Mientras que Philippot buscaba poner más énfasis en los temas económicos y sociales, que pudieran atraer a más electores— particularmente de izquierda, los llamados *gaucho-lepenistes*— otros cuadros del partido buscarían resaltar la línea de inmigración y seguridad de corte más derechista.

Philippot fue señalado como el culpable de la decepción. Formó una organización política interna al FN llamada *Les Patriotes* y amenazó con abandonar el partido si el FN renunciaba a su posición sobre el euro. Después de las elecciones legislativas, en las que el FN logró 8 escaños pero no los 15 necesarios para convertirse en grupo parlamentario, las presiones aumentaron. Marine Le Pen le pidió elegir entre la vicepresidencia del FN y la dirección de *Les Patriotes* y él respondió acusando al partido de estar cambiando de línea y regresando a un pasado absolutamente horripilante. Su destino estaba sellado: le fueron retiradas sus responsabilidades como vicepresidente y Philippot renunció al FN para convertir a *Les Patriotes* en un nuevo partido político (Galtier 2017; Zafimehy 2017).

A raíz del resultado electoral de 2017, Marine Le Pen impulsó la idea de una refundación del partido. El electorado francés había vuelto a dejar claro que la marca FN estaba vetada del Elíseo. Por ello, propuso un cambio de nombre. A partir de 2018, el Front National se convirtió oficialmente en Rassemblement National. Así pues, la era marinista del FN terminó en 2018.

En la primera elección del RN, las europeas de 2019, Marine Le Pen decidió no ser candidata. Esto no necesariamente quiere decir un abandono de la centralidad de Le Pen. De hecho, el nombre oficial de la lista fue *Prenez le pouvoir, liste soutenue par Marine Le Pen*— Tomar el poder, lista apoyada por Marine Le Pen—. Sin embargo, la estrategia fue continuar con la imagen del cambio de marca del partido, particularmente en términos generacionales. La lista fue encabezada por el joven portavoz del partido,

Jordan Bardella, de apenas 23 años. Con Bardella a la cabeza, RN obtuvo un cerrado primer lugar con 23.34 % de los votos efectivos contra 22.42 % de la lista del movimiento del presidente Macron, la República en Marcha.

Qué le depara al *pater familia* de los movimientos NAP es incierto. Los partidos políticos tradicionales de Francia han perdido su posición central. Pero con Trump en EUA o Salvini en Italia, Le Pen ya no es siquiera la cara mundial del nativismo autoritario de corte populista. No obstante, sigue siendo uno de los partidos importantes de esta corriente y estudiarlo es un buen punto de partida para aproximarse al fenómeno general de esta familia política.

Capítulo 3

Teorías sobre el voto NAP

En esta tesis, estoy interesado en explorar las *configuraciones sociales* donde se desarrolla o no el voto por un partido NAP específico. Si uno hace una revisión hemerográfica o en internet sobre los motivos de éxito de Trump o el Brexit encontramos que predominan fuertemente las explicaciones de carácter económico y/o cultural (Beauchamp 2016a, 2016b; Carney 2016; Tesler 2016; Sides y Tesler 2016; Arnade 2016).

¿Son los desplazados por la globalización los que votan por los NAP? Los trabajadores del Rust Belt le habrían costado la presidencia a Hillary Clinton, la escolaridad es un clivaje frecuentemente mencionado, existiría un voto de izquierda económica detrás de estos movimientos... ¿O es en realidad una reacción cultural frente a la presencia de aquellos que se consideran como el *Otro*: musulmanes, latinos, negros, inmigrantes en general? Estos fenómenos se están dando en sociedades con población predominantemente blanca y dentro de la cultura occidental, lo que podría interpretarse como un choque de culturas. Así pues, desde el estudio académico, el debate parece estar entre dos *resentimientos* distintos, el económico y el racial/cultural, sin que haya *a priori* uno incontrovertiblemente dominante sino que más bien están relacionados entre sí (Inglehart y Norris 2016; Ivarsflaten y Gudbransen 2014).

Para realizar un análisis de este tipo hay que recordar que dentro de la metodología de política comparada existen cuatro grandes paradigmas: institucional, instrumental, cultural y estructural (Uribe Coughlan 2016). Tomando en cuenta estas primeras ideas y la naturaleza del estudio de caso, me parece que un buen punto de partida es considerar

los enfoques estructural y cultural, más que las perspectivas institucional o instrumental.

3.1. Estructuralismo y Culturalismo

Desde el paradigma estructural, se dice que “las estructuras condicionan el resultado” (Balaam y Veseth 2008). El concepto tradicional de estructura se refiere a grupos con características materiales distintas. Por ejemplo, el estructuralista socioeconómico por antonomasia resulta Marx. Para él, las profundas fuerzas sociales seguirían un proceso histórico lineal que pasaría del feudalismo al capitalismo y después al socialismo, para culminar en el comunismo (Balaam y Veseth 2008). Los cambios tecnológicos impulsarían un sistema económico con nuevas clases sociales que desplazarían las viejas instituciones, refiere Heilbroner (1992) al hablar de la teoría marxista. Con la revolución industrial habrían llegado dos clases sociales encontradas: la burguesía y el proletariado. La Historia marxista sería inevitable, dice el Manifiesto Comunista (Heilbroner 1992):

[El desarrollo de la industria moderna] derriba los fundamentos mismos por los que la burguesía produce y apropiá la producción. Lo que la burguesía produce entonces son, sobre todo, sus propios sepultureros. Su caída y la victoria del proletariado son igualmente inevitables...

Hasta ahora no hemos verificado empíricamente esas afirmaciones de Marx y Engels, pero varias de sus herramientas teóricas siguen vigentes. Lo que la lucha de clases entre el proletariado y la burguesía ejemplifica es que, dentro de las sociedades, existen grupos cuyos intereses están enfrentados. La pertenencia a una de estas clases determinaría, primordialmente, las preferencias políticas de un individuo.

Por otro lado, el enfoque cultural tiene como unidad de análisis las ideas y valores existentes en una sociedad. De acuerdo con Sheri Berman (2001), una de las preguntas teóricas que este paradigma busca responder es cómo las variables culturales e ideológicas influyen el comportamiento político. Un ejemplo de esto es el libro *The Civic Culture* (Almond y Verba 1963)— generalmente considerado como el precursor de esta corriente— que propone cómo diferentes culturas políticas llevarían a distintos grados de estabilidad democrática. Otro ejemplo son los análisis que asocian las preferencias partidarias con ciertos valores e ideas, como el que realiza Ronald Inglehart en 1977 con su teoría de la revolución silenciosa. Para Inglehart, las condiciones económicas de la posguerra con la

que crecieron los jóvenes europeos— mejores con relación a las de sus padres— habrían hecho que se formaran en valores postmateriales como el sentido de pertenencia y la realización personal (Kesselman 1979). Esto, a su vez, se relacionaría con preferencias por partidos que él llama libertarios de izquierda (Inglehart y Norris 2016).

Considero, siguiendo en parte a Sewell Jr. (1992), que estos dos paradigmas pueden coexistir. Las estructuras pueden ser culturales, así como cambios en las condiciones materiales pueden llevar a cambios ideológicos. En el argumento de Inglehart esto es claro, pero la implicación va más allá. Los desarrollos económicos, así como los cambios en las relaciones de poder o la entrada de nuevos grupos sociales puede forzar un replanteamiento de las creencias existentes (Berman 2001). Cuando hablo de estructuras, entonces, me refiero en un sentido informal a las condiciones ideológicas o materiales que moldean diferentes grupos sociales con intereses distintos. Así pues, en lo subsecuente es bajo esta idea general de estructuras que planteo las posibles explicaciones del voto NAP.

3.2. Clivajes y Escolaridad

Un constructo teórico fundamental para mi propósito es el de clivaje político, pues el principal foco de la teoría de clivajes es la “evolución a largo plazo de la estructura social”, en donde yacen las “fuerzas más fundamentales de la política” (Bornschier 2009). Rokkan y Lipset son las referencias obligadas con respecto a los clivajes tradicionales, pero Bartolini presenta una conceptualización formal clara. Un clivaje es una división política que cuenta con tres elementos: división socioestructural, identidad colectiva y manifestación organizacional (Bornschier 2009).

Es decir, para que una división en la sociedad pueda considerarse como un clivaje, debe existir una diferencia clara en términos socioestructurales, y que los miembros de cada grupo que constituya dicha distinción formen una identidad colectiva y cierta capacidad de movilización. Estas tres características, la teoría indica, llevarían a que los miembros de cada grupo del clivaje voten de acuerdo a los intereses del mismo. La pertenencia a uno de los lados de la estructura determinaría las preferencias políticas del individuo, traducidas en su voto.

Con relación a los movimientos NAP, un clivaje determinante resultaría ser el de ciu-

dadanos escolarizados frente a los no escolarizados.¹ Por ejemplo, la primera realización de la segunda vuelta presidencial en Austria en 2016 arrojó en las encuestas una clara división entre aquellos con nivel escolar obligatorio y aquellos con niveles de escolaridad media superior y superior (Hoare 2016). En Estados Unidos, de acuerdo con Nate Silver (2016), la escolaridad fue el clivaje que mejor predijo quién votaba por Donald Trump. Rae (2016) encontró también una fuerte relación entre el porcentaje de voto antieuropeo en el referéndum británico de 2016 y el porcentaje a nivel local de personas poco escolarizadas.

Este clivaje, de acuerdo con Hervé Le Bras (2015), tiene consecuencias socioeconómicas directas al tiempo que contribuye a la formación de una clase— al menos en Francia— temerosa frente a su posición en la estructura social y susceptible de votar por el Front National. Desde los 90, se ha identificado a los individuos con escolaridad técnica o no universitaria, como un nicho electoral para el FN (Mayer y Perrineau 1990; Perrineau 1999; Gombin 2005; Mayer 2005, 2007; Perrineau 2007; Rivière y col. 2012; Perrineau 2012).

3.3. La Clase Desplazada por la Globalización

Por otro lado, en la discusión actual sobre los movimientos NAP, no es raro encontrar referencias al fascismo. Por ejemplo, el número de octubre de 2016 de la revista Letras Libres, dedicado a Donald Trump, presentó una sugerente portada con la foto del magnate y un bigote al estilo Hitler formado por las palabras *Fascista Americano* (*Letras Libres* 2016). Otro ejemplo es el carácter neofascista de Norbert Hofer, ver Hoare (2016). En este sentido resulta pertinente recordar el estudio que desarrolló Barrington Moore respecto del fascismo del siglo XX.

Desde una perspectiva estructural, Moore explica el surgimiento del fascismo en Alemania, Italia y Japón. La modernización conservadora y la industrialización, frutos de una *revolución desde arriba*, llevaron a contradicciones estructurales fuertes en estos países. El fascismo fue un intento de hacer *popular y plebeyo* el conservadurismo y el reaccionismo, que habían perdido su legitimidad, dice Moore (1966). La característica

¹Muchos hablan de ciudadanos educados y no educados, pero considero que el término correcto debe ser escolarizados o academizados pues la diferencia socioestructural es la asistencia a una institución académica de nivel medio superior o superior, cosa que no garantiza *per se* la *educación* del individuo, máxime que este es un concepto sujeto a debate, súmamente subjetivo, esquivo y cargado.

principal del fascismo del siglo XX es el anticapitalismo plebeyo. En términos de voto, por ejemplo, los Nazis fueron más populares entre aquellos que tenían menos y estaban más desfavorecidos *con relación al área particular en la que vivían* (Moore 1966).

Sides y Tesler (2016) presentan cierta evidencia de que este fue el caso en Estados Unidos con el apoyo hacia Donald Trump, Arnade (2016) habla de la concentración de la prosperidad en el Reino Unido como posible catalizador del voto para el Brexit y Le Bras (2015) examina el mapa del voto frontista en Francia frente a un índice de desigualdad socioeconómica en el mismo espíritu con el que Moore refiere el voto Nazi en el espacio rural alemán de inicios de los años treinta pero de una manera mucho más detallada.

Ciertamente Moore no es el único teórico que señala el vínculo entre una fuerte modernización económica y el voto hacia los movimientos extremos por parte de clases desfavorecidas:

Al evaluar su propia situación las personas hacen comparaciones con la situación de los demás. Los votantes con una situación socioeconómica débil pueden evaluar esta posición de forma más negativa en tanto vivan en una región acaudalada. Entonces, desigualdades en el ingreso pueden traducirse en apoyo a partidos de derecha extrema (Coffé, Heyndels y Vermeir 2007, traducción propia).

Esta línea de investigación argumentaría por qué estamos viviendo hoy estos movimientos NAP.

Para Branko Milanovic (2016), la globalización iniciada a finales del siglo XX—y vigente en nuestro siglo—constituye el más grande cambio en la distribución de los ingresos desde la revolución industrial. Esta redistribución ha traído crecimiento económico, pero de manera muy variada a lo largo de los percentiles poblacionales. Así como el 5% más rico ha seguido creciendo por encima del promedio, los primeros 70 percentiles lo han hecho a diferentes tasas. Sin embargo, el crecimiento que llama más la atención es el de la población entre el percentil 70 y 95: han crecido menos que la media o incluso han visto su ingreso real disminuir. ¿Quiénes son estas personas? Siete de cada 10 provienen de los *viejos países ricos* de la OCDE.² Precisamente muchos de los países en los cuales

²Esta evidencia ha llegado a ser conocida como la gráfica del elefante debido a su forma. Ver Milanovic (2016).

vemos estos movimientos NAP prosperar. Claro que deducir que esto es prueba de que son estas personas quienes votan por los movimientos NAP sería un error, pues se caería en un caso de falacia ecológica.

En el caso particular del FN, no obstante, hay varios estudios que sugieren que han sido las personas de la *clase media* las que han tendido a apoyar al partido. El modelo sociológico de Mayer y Perrineau (1990) refiere la existencia de una diferencia por nivel de sueldo y posesión de patrimonio. La pequeña burguesía ha sido históricamente asociada a movimientos de derecha (Mayer 1987; Le Bras 2015; Goodliffe 2019). Adicionalmente, en Francia la literatura consistentemente analiza el fenómeno electoral desde la perspectiva de las *categorías socioprofesionales* a las que pertenecen los individuos. Por ejemplo, la mayor o menor presencia de obreros o agricultores, han influido de distinta manera en el comportamiento de los pequeños empresarios en Francia (Mayer y Michelat 1981). Una lista de estudios que utilizan la variable de categorías socioprofesionales para estudiar al FN son Mayer (1987, 2005, 2007), Mayer y Perrineau (1990), Mayer y Boy (1987), Perrineau (1999, 2007, 2012), Gombin (2005, 2009, 2013b, 2013a), Rivière y col. (2012) y Gombin y Rivière (2013). En este sentido, esta tendría que ser una de las variables obligadas en mi análisis.

3.4. Teorías del Conflicto

Otros modelos que pudieran encajar con este comportamiento de crecimiento en las preferencias por movimientos NAP son la *teoría del conflicto* y la *teoría del interés económico*. La primera constituye una explicación consistente con ciertos argumentos culturales sobre la motivación racista. Esta teoría es expuesta por autores como Blalock (1967) y Olzac (1992)— citados por Coffé, Heyndels y Vermeir (2007)—. La violencia étnica, expresada en manifestaciones xenófobas propias de los movimientos NAP, proviene del resentimiento. La clase privilegiada de la sociedad— pensemos aquí en el tradicional grupo de hombres blancos que mencionan los culturalistas— comienza a acumular un “sentimiento de injusticia cuando ve su privilegio escurrirse hacia las manos de otro grupo que no lo tenía antes”, por lo que una “causa de la violencia étnica es el cambio en estatus legal y político de los grupos étnicos mayoritarios y minoritarios”— Petersen (2002) citado por Beauchamp (2016b), traducción propia—.

Blalock (1967) y Olzac (1992) teorizan sobre que el conflicto aumenta en zonas de

problemas económicos, como el desempleo, y ahí se suceden reacciones exclusionistas por parte de los grupos mayoritarios. Si al resentimiento sumáramos el miedo por competir por la escasez de los recursos, podría sugerirse el por qué los trabajadores manuales con menores niveles escolares tenderían a reaccionar frente a los grupos de inmigrantes. Esta es la lógica de la teoría del interés económico: la presencia de inmigrantes resulta en un aumento en la competencia por recursos escasos y, por lo tanto, en conflictos entre grupos sociales (Coffé, Heyndels y Vermeir 2007).

Estas teorías son consistentes con el concepto de chauvinismo del bienestar: la propuesta de que los beneficios que aporta el estado de bienestar deben estar reservados exclusivamente para el grupo nativo, no para los grupos de fuera.³ Esto encajaría con el nativismo que identifica Mudde en los grupos NAP (Mudde 2007; Beauchamp 2016a) y con las propuestas concretas de sus líderes, como el reclamo de Marine Le Pen para terminar con la educación gratuita a hijos de inmigrantes (Le Monde 2016).

Aunque probablemente menos verificable empíricamente, otra de las líneas teóricas estructurales sobre el voto por los partidos NAP lo encontramos en las referencias de Valentino, Brader y Jardina (2013). Aunado a las explicaciones que ya he mencionado—mismas que ellos agrupan dentro de la categoría de la *hipótesis de la competencia del mercado laboral*— encontramos las *hipótesis fiscales*. Esta hipótesis puede contribuir a explicar la oposición de la, así llamada, pequeña burguesía hacia los inmigrantes. Estos supondrían una presión adicional a la seguridad social y, consigo, la necesidad de incrementar los impuestos, un aumento en los costos de la educación y más cargas a la infraestructura (Valentino, Brader y Jardina 2013). La preferencia de clase de la pequeña burguesía, que tiene negocios propios y busca que sus hijos tengan grados universitarios, es altamente contraria al aumento de impuestos y costos. Si los inmigrantes son vistos como chivos expiatorios para estos riesgos, se entendería por qué habría un número elevado de votos por los NAP entre esta clase social.

Aquí hay que decir que estas teorías se refieren a percepciones, no necesariamente proponen la idea de que la competencia con los inmigrantes sea real, sino que se percibe como tal. En este sentido, el *miedo* a “ir cayendo” en la estructura social es lo que empujaría a muchos franceses a *apostar* por el Front National con su voto, de acuerdo a Le Bras

³Kitschelt (1995) citado por Coffé, Heyndels y Vermeir (2007). Observar también el paralelismo con el marco teórico estructural à la Tilly sobre grupos miembro y grupos retadores (Skocpol 1979).

(2015). Otro estudioso del FN que parece apoyar este tipo de teorías es Pascal Perrineau, quien también habla de conflictos de inseguridad o preocupaciones socioeconómicas y laborales (Perrineau 2007). Algo que normalmente se usa para apoyar estas tesis es que las principales motivaciones autoreportadas por los votantes FN constantemente han sido la inmigración, el desempleo y la delincuencia (Mayer 2007).

A pesar del tinte económico que sugeriría una primera lectura de estas teorías, existe literatura que caracteriza a los votantes nativistas y autoritarios con base en encuestas y que concluye muchas veces que el resentimiento no puede ser puramente económico y que las reacciones culturales se dan incluso en zonas económicamente beneficiadas. Estaríamos frente a una reacción cultural y de valores (Inglehart y Norris 2016). La teoría del conflicto, pues, no se reduciría a la pérdida de privilegios económicos sino importantemente, socioculturales. Por ejemplo, desde finales de los 80, los votantes FN han sido electores etnocéntricos y autoritarios que opinan que hay demasiados inmigrantes, que ya no se sienten como en casa y que apoyan la pena de muerte. También existe una diferencia de género, un *gender gap* histórico en los movimientos NAP que son más masculinos (Crépon 2015; Mayer 2015; Mudde 2018). La pérdida de referencias culturales con respecto a la religión o la familia es una característica de los electores del FN (Mayer y Perrineau 1990; Mayer y Boy 1987; Perrineau 1999, 2012). Incluso se llega a hablar de pérdida de referentes de civilidad (Perrineau 2007).

Lo que la variedad de teorías y enfoques utilizados para explicar el voto nativista y autoritario de corte populista refleja es que los individuos reaccionan ante los distintos contextos sociales. Ya sea que se perciban a sí mismos como una clase socioeconómica marginada o como un grupo cultural en peligro, las decisiones electorales de los individuos dependen fundamentalmente de la configuración social de su entorno. En este sentido, mi intención al aproximarme al fenómeno es reconocer que la relación que guardan las variables utilizadas por los diferentes autores citados con el voto frontista no es homogénea. En la búsqueda de *El votante FN*, parecería que perdemos de vista que no existe un solo tipo de elector.

Desde mi punto de vista, y siguiendo principalmente el ejemplo de Gombin (2009, 2013b, 2013a), el uso del modelado estadístico jerárquico de datos ecológicos ofrece una buena oportunidad de continuar el estudio de los fenómenos NAP que parecen estar tomando al mundo por sorpresa. Para ello entonces, procederé a presentar el marco

teórico estadístico que permitirá llevar a cabo el análisis de los datos franceses.

Parte II

Paradigma estadístico bayesiano

Capítulo 4

La probabilidad no existe

Cuando nos enfrentamos a situaciones inciertas, constantemente medimos nuestra incertidumbre de manera personal, con base en la información que tenemos disponible. Esta medición la expresamos coloquialmente mediante el término *probabilidad*. Por ejemplo, nos podemos preguntar la probabilidad de que en una elección popular, una candidatura haya recibido entre el 30 % y el 35 % de los votos. Más aún, como dice Berger (1985), es común “pensar en términos de probabilidades personales todo el tiempo: cuando se apuesta al resultado de un partido de fútbol americano, cuando se evalúa la posibilidad de lluvia al día siguiente...”.

Es decir, día con día utilizamos el término probabilidad para referirnos a apreciaciones subjetivas de la plausibilidad de cosas inciertas. A esta conceptualización de la probabilidad como una medida personal de la incertidumbre se le conoce como el paradigma subjetivo o *bayesiano*. Esta es una interpretación filosófica radicalmente distinta a las más tradicionales: la empírica o *frecuentista* y la *clásica* (Singpurwalla 2017; Gutiérrez Peña 2016).

4.1. ¿Probabilidad o probabilidades?

Desde la óptica bayesiana, la probabilidad en sentido matemático no sería, pues, un concepto ajeno ni incompatible al cotidiano; por el contrario, siendo en el fondo el mismo, permitiría extender su uso a todas aquellas situaciones en las que no aplican las otras perspectivas (Berger 1985; Gutiérrez Peña 2016). ¿Cuáles serían algunos ejemplos

de dichas situaciones?

El paradigma clásico está basado en “*simetrías* o en propiedades físicas” (Gutiérrez Peña 2016). En virtud de dichas simetrías las probabilidades asignadas a los eventos son iguales. Por ello, puede aplicarse a situaciones proclives al consenso, generalmente juegos de azar, y donde la regla general de probabilidades es: casos favorables entre casos totales. En lo que respecta a las probabilidades personales, estas pueden o no coincidir, incluso en las situaciones más proclives a un consenso.

Un lanzamiento de una moneda tiene dos posibles resultados igualmente probables... o no. Reducir la probabilidad a que *siempre* deba ser 0.5 es un error (Singpurwalla 2017). Si alguien tiene pruebas de que una moneda particular está cargada, pudiera asignar una probabilidad radicalmente distinta al águila que al sol, mientras que quien desconoce esta información medirá la incertidumbre que rodea al lanzamiento con una asignación equiprobable. Ambas mediciones son válidas, solamente difieren en que son realizadas por sujetos con distinto grado de incertidumbre o conocimiento sobre el fenómeno de interés. Lo que sí podemos decir es que esta interpretación clásica es la más restrictiva pues se reduce al hecho de que la probabilidad de algún evento siempre es equitativa y los casos en los que eso sea una buena aproximación a la realidad son relativamente pocos.

Por otro lado, para el paradigma frecuentista, la probabilidad es un límite de frecuencias relativas de eventos repetidos bajo condiciones similares. Es una extensión del concepto clásico: casos favorables entre casos totales, cuando los casos totales tienden a infinito. Así, “únicamente los sucesos que pueden ser repetidos tienen probabilidad” (Aquino Pérez 2010). Es decir, la probabilidad sería una propiedad de un colectivo y no de eventos individuales (Singpurwalla 2017).

¿Cuál sería pues la probabilidad de que cierto partido político gane las próximas elecciones presidenciales en México o de que la Selección Nacional logre jugar un quinto partido en el siguiente Mundial de futbol? Dichos eventos son únicos, no pueden repetirse infinitamente bajo las mismas circunstancias con tal de que seamos capaces de registrar empíricamente su límite de frecuencias relativas. Claro, podemos pensar en que hayan *suficientes* eventos bajo condiciones *similares* como para que la probabilidad sea aproximadamente, digamos, 30 %. Pero, ¿cuántos eventos son suficientes? y ¿qué tan similares deben ser? (Singpurwalla 2017). Por el contrario, para la perspectiva bayesiana sí que

podemos expresar, cada uno de nosotros, nuestras probabilidades subjetivas sobre ellos.

De manera general se podría decir que los paradigmas tradicionales interpretan la probabilidad exclusivamente como una medida de variabilidad. Dicha variabilidad es un concepto necesario, ciertamente, pero más limitado que aquel de incertidumbre: la incertidumbre puede derivarse de la variabilidad, pero existen muchos fenómenos invariantes que son inciertos. Pensemos en el ejemplo al que me refería al inicio de este capítulo. La proporción de votos de una candidatura en una elección es una proporción fija y que no varía una vez concluida la votación. Sin embargo, nos podemos hacer preguntas probabilísticas acerca de ella porque es desconocida, al menos mientras no se anuncian los resultados oficiales.

Es posible, incluso, llevar el argumento al extremo: ¿cuál es la probabilidad de que mis dos hermanos cumplan años el mismo día? Este es un evento cierto o falso, totalmente determinado en el pasado pues, o bien nacieron el mismo día del año o no. Por lo mismo, en estricto sentido sería difícil hacer una estimación probabilística de ello desde el punto de vista lógico o frecuentista. Alguien podría argumentar que, asumiendo equi-probabilidad e ignorando los años bisiestos, esta debería ser de una en 365. Alguien más, en un espíritu más frecuentista, podría decir que ésta debe ser la probabilidad de que sean gemelos y, por lo mismo, igual a la proporción de nacimientos gemelares. Si me preguntan a mí, les diría que esa probabilidad es 1: sé que ambos nacieron el 16 de diciembre.

El estadístico y actuaria italoaustríaco Bruno de Finetti, acuñó la frase que le da nombre a este capítulo y que sintetiza esta visión bayesiana de la probabilidad. Para los bayesianos, LA probabilidad única, absoluta y objetiva no existe; más bien es una medida subjetiva de la incertidumbre que un individuo particular tiene frente al fenómeno de interés.

Esta argumentación filosófica se ve reforzada por una justificación matemática formal: la teoría de decisión bayesiana. Con base en unos axiomas, llamados de coherencia , se puede demostrar que un decisor, ante un problema de decisión con incertidumbre, debe actuar de la siguiente manera:

1. Expresar su incertidumbre sobre los eventos inciertos relevantes en una medida llamada probabilidad subjetiva.

2. Reflejar sus preferencias sobre las posibles consecuencias mediante otra medida llamada utilidad.
3. Tomar la decisión que maximice la utilidad esperada o, equivalentemente, minimice la pérdida esperada.

La teoría de decisión bayesiana tiene el sentido normativo de que esta es la manera óptima de actuar *si se quiere evitar violar los axiomas de coherencia* y de ninguna manera en un sentido descriptivo que diga que los seres humanos actuamos de esa forma o incluso, que se tengan que aceptar los propios axiomas. Dicha justificación axiomática escapa los objetivos de esta tesis, pero una primera introducción a ella puede consultarse en referencias como Mendoza y Regueiro (2011) o Bernardo (1981).

Por otro lado, este procedimiento bayesiano está encaminado, de cierta forma, a reducir el riesgo de las consecuencias que puedan surgir en un problema de decisión. Cualquier decisor que *usa* la teoría bayesiana hará esto. Sin embargo, desde mi punto de vista, la diferencia entre ello y lo que hace un *bayesiano* estriba en que este último se preocupa también por minimizar la incertidumbre asociada al fenómeno de su interés. ¿Cómo podemos realizar un análisis que busque reducir la incertidumbre? Recabando mayor información. En el contexto de la estadística esto significa obtener datos.

Frecuentemente, en mi trabajo, bromeamos sobre el bayesiano que no necesita datos para estimar lo que sea, pues bastan sus probabilidades y pérdidas subjetivas. Sin embargo, al margen del buen humor, la realidad es que si solo tomamos decisiones *a priori* estamos siendo obtusos. En el mejor de los casos, diremos que ya contamos con la única información disponible; en el peor, estaríamos siendo víctimas de una ciega soberbia. Por lo mismo, el *estadístico bayesiano* debe primero preocuparse por los datos y, una vez recabados, proceder al análisis. Así pues, es necesario continuar esta introducción al paradigma bayesiano presentando el mecanismo por el cual un estadístico bayesiano aprende de los datos y actualiza sus creencias subjetivas.

4.2. Aprendizaje bayesiano

Un punto clave de la teoría de decisión bayesiana es que la definición matemática de la probabilidad subjetiva satisface los axiomas de Kolmogorov que dan sustento a la teoría usual de probabilidades. Es decir, las probabilidades subjetivas satisfacen los

mismos teoremas y resultados que se aprenden en los cursos de probabilidad. Uno de los más elementales es el teorema que hace posible este *aprendizaje bayesiano* de los datos: el Teorema de Bayes.

Teorema 4.1. Teorema de Bayes. (Versión probabilística)

Sean A y B eventos tales que $\mathbb{P}(B) \neq 0$. Se cumple que:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)},$$

donde $\mathbb{P}(A|B)$ y $\mathbb{P}(B|A)$ son probabilidades condicionales, mientras que $\mathbb{P}(A)$ y $\mathbb{P}(B)$ son las respectivas probabilidades marginales.

Vale la pena detenerse a pensar qué implica este teorema en nuestro contexto; no por nada el apelativo del paradigma hace honor a Thomas Bayes, mismo personaje al que se refiere el teorema. A partir de este momento utilizaré el término *distribución* de manera general para referirme, por ejemplo, a una ley de probabilidad o a sus funciones de densidad o masa de probabilidad, según sea el caso.

Cuando hacemos inferencia estadística paramétrica, usualmente estamos interesados en conocer sobre una cantidad desconocida θ , llamada parámetro. Bajo el paradigma bayesiano, este es tratado como si fuera una variable aleatoria y, por tanto, le asignamos una distribución de probabilidad *inicial* o *a priori* $f(\theta)$ que resume nuestro estado de conocimiento inicial sobre el parámetro. Posteriormente, para reducir nuestra incertidumbre procedemos a recabar información en forma de datos x . Finalmente determinamos una distribución sobre dichos datos dado el parámetro— $f(x|\theta)$ — y que, como función del parámetro, es llamada función de verosimilitud— $L(\theta)$ —.

Nuestro objetivo entonces es *invertir* las probabilidades y conocer la distribución del parámetro dados los datos $f(\theta|x)$. Precisamente una de las primeras motivaciones de la estadística bayesiana era justamente la de encontrar una forma de calcular estas *probabilidades inversas* (Robert 2007):

... el propósito de un análisis estadístico es fundamentalmente un propósito de *inversión*, puesto que busca recuperar las causas— reducidas a los parámetros del mecanismo probabilístico de generación— de los efectos— resumidos por las observaciones. En otras palabras, al observar un fenómeno aleatorio dirigido por un parámetro θ , los métodos estadísticos permiten deducir de estas

observaciones una *inferencia* (esto es, un resumen, una caracterización) sobre θ [...] La inferencia está basada entonces en la distribución de θ condicional en x [...] llamada la *distribución posterior*...

Esto lo hacemos mediante el teorema de Bayes:

Teorema 4.2. Teorema de Bayes. (Versión estadística)

Sean θ un vector de parámetros desconocidos, $f(\theta)$ su distribución inicial y x información adicional en forma de datos conocidos. Entonces, la distribución posterior de θ , una vez observados los datos, es:

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

donde $f(x|\theta) = L(\theta)$ es la función de verosimilitud para el vector de parámetros θ , considerando los datos x como fijos, y con $f(x)$ la distribución marginal de los datos.

Visto todo como función de θ , nuestra cantidad de interés, la expresión anterior se reduce a:

$$f(\theta|x) \propto L(\theta)f(\theta), \quad (4.1)$$

pues $f(x)$ depende solo de los datos ya observados y funge como la constante normalizadora que hace que la posterior integre a 1. Esta última representación resume el proceso de aprendizaje bayesiano: *la posterior es proporcional a la verosimilitud por la inicial*.

Otra forma de resumir el proceso de aprendizaje bayesiano es mediante lo que Gutiérrez Peña (2016) llama la *única receta de la inferencia bayesiana*:

... encontrar la distribución condicional de todas aquellas cantidades de interés cuyo valor desconocemos dado el valor conocido de las variables observadas.

Puesto de otra forma por el mismo autor: la distribución final es la inferencia. Esto resume el lugar central que ocupa la distribución posterior en el paradigma bayesiano. No obstante, llamarla única receta es más bien un recurso mnemotécnico, pues como bien advierte el autor, así como Berger (1985), es deseable resumir esta inferencia general en un proceso específico, como se puede intuir también de la cita anterior de Christian P. Robert (2007).

4.3. Distribuciones iniciales

En las secciones anteriores he tratado de exponer y enfatizar que un análisis bayesiano debería tener como objetivo práctico aprender de los datos, invirtiendo probabilidades, para calcular *distribuciones posteriores*. Sin embargo, para lograr esto es necesario empezar con una *distribución inicial* que refleje la incertidumbre que se tenga frente a un fenómeno de interés. Existen muchas formas diferentes de determinar la distribución inicial, en esta sección presentaré algunas de ellas con base en Berger (1985), Congdon (2006), Robert (2007) y Gelman y col. (2013).

Empecemos por considerar el caso cuando el espacio paramétrico sobre el que tenemos que determinar las probabilidades es discreto. Por ejemplo, pensemos en fijar una distribución inicial para un partido de fútbol, en el que hay 3 posibles resultados: victoria para el local, victoria para el visitante o empate. La primera distribución posible está basada en lo que se conoce como el *criterio de la razón insuficiente*. Fue propuesto por Laplace y está relacionado a la interpretación clásica de la probabilidad puesto que establece que, a falta de información adicional que permita invalidarlo, se debería tomar el supuesto de que todos los eventos son igualmente probables. Así pues, podríamos asignar una distribución discreta uniforme de $1/3$ a cada uno de los 3 resultados del partido.

Sin embargo, podría haber razones suficientes para que esta no sea la distribución inicial. Por ejemplo, podemos creer que el equipo local es mejor que el equipo visitante y, en consecuencia, asignar una probabilidad del 60 % a que el local gane, 30 % al empate y 10 % a una sorpresiva victoria de la visita. Es posible que otro aficionado, piense que un 30 % de probabilidad de empate es demasiado alto, en cuyo caso sería necesario ajustar las probabilidades así calculadas hasta encontrar una combinación aceptable. Pudiéramos también asignar una distribución inicial basados en algún tipo de frecuencia relativa histórica.

Otra forma de asignar probabilidades a eventos es pensar en apuestas que consideremos justas. Imaginemos un mercado como el del sitio web Predict It en el que podemos apostar una cantidad $p \in (0, 1)$ de centavos de dólar a que un evento suceda. Si el evento efectivamente sucede, nuestra apuesta vale 1 dólar y habremos ganado $1 - p$ centavos. Si el evento no sucede, perdemos p centavos. Por otro lado, pensemos en una apuesta justa, es decir, una cuya ganancia esperada fuera 0.

La ganancia esperada $\mathbb{E}[g]$ es igual a la ganancia cuando el evento sucede, por la probabilidad de que suceda, más la ganancia—en este caso, negativa—cuando el evento no sucede, por la probabilidad de que no suceda. Si denotamos al evento como A y a su probabilidad como $\mathbb{P}(A)$, tenemos que :

$$\begin{aligned}\mathbb{E}[g] = 0 &\Leftrightarrow (1 - p)\mathbb{P}(A) - p\mathbb{P}(A^c) = 0 \\ &\Leftrightarrow (1 - p)\mathbb{P}(A) - p(1 - \mathbb{P}(A)) = 0 \\ &\Leftrightarrow \mathbb{P}(A) = p.\end{aligned}$$

Si apostáramos una cantidad mayor estaríamos arriesgándonos de más, o dicho de otra forma, tendríamos una ganancia esperada negativa, por lo que no nos convendría la apuesta. Querríamos apostar menos, pues en ese caso nuestra ganancia esperada sería positiva. Esto quiere decir que podemos pensar en las probabilidades subjetivas de un evento como la máxima cantidad que estaríamos dispuestos a considerar en una apuesta de este tipo.

Pudiera haber cierta reticencia a este ejemplo pues sabemos que cuando se trata de apuestas, la gente tiende a ser aversa al riesgo, lo que podría requerir que la ganancia esperada fuera muy positiva para siquiera considerar una apuesta. De manera análoga, alguien amante al riesgo podría estar dispuesto a apostar aún con una ganancia esperada negativa. Sin embargo, quisiera decir que este tipo de apuestas son meramente imaginativas como mecanismo para dilucidar probabilidades subjetivas o que las cantidades apostadas serían en teoría lo suficientemente pequeñas como para pensar en una utilidad lineal del dinero.

Sin embargo, resulta más frecuente que la distribución inicial deba ser continua. ¿Cómo elegirla? A continuación presento algunos métodos.

4.3.1. Distribuciones no informativas

Al igual que con las distribuciones discretas podemos comenzar por el criterio de la razón insuficiente y proponer una distribución que busque reflejar una ignorancia general en la que no se prefieran *a priori* algunos resultados o valores de los parámetros. Así surgen las llamadas *distribuciones iniciales no informativas*. Estas se usan, generalmente,

por dos motivos. En primer lugar, habrá quien crea que no cuenta con información inicial y que, efectivamente, se encuentra en una situación de ignorancia total. Alternativamente, se podría ignorar la información existente porque se busca presentar una distribución relativamente más objetiva, ya sea porque el contexto así lo exige, porque se busca llevar a cabo un análisis de sensibilidad bajo diferentes supuestos o por algún otro motivo.

El criterio de la razón insuficiente buscaría proponer una asignación equiprobable a los eventos. Así pues, la primera distribución inicial no informativa que se puede considerar es la uniforme. Esto es que la distribución sea proporcional a una constante:

$$f(\theta) \propto c$$

Esta distribución, sin embargo, ha sufrido algunas críticas. La primera de ellas es que, si uno quiere reflejar incertidumbre total sobre un parámetro, esto querría decir que también deberíamos reflejar incertidumbre total sobre transformaciones a dicho parámetro. Sin embargo, la distribución uniforme no es *invariante ante transformaciones*. Por ello, se han buscado distribuciones no informativas diferentes a la uniforme y que sí cumplan con la invarianza ante transformaciones. La más utilizada es la *inicial de Jeffreys*.

Definición 4.1. Inicial de Jeffreys

La distribución *inicial de Jeffreys* para un parámetro θ se define en términos de la información de Fisher como:

$$f(\theta) \propto \sqrt{I_x(\theta)} \quad \text{donde} \quad I_x(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ln(f(x|\theta)) \right]$$

Un desarrollo de cómo obtener este resultado de invarianza se puede encontrar en Mendoza y Regueiro (2011).

Ciertamente la inicial de Jeffreys logra superar la primera crítica a las distribuciones iniciales uniformes no informativas, pero no otro tipo de críticas. Consideremos el caso en el que tenemos una observación x proveniente de una distribución normal con varianza conocida. En este caso tenemos que la inicial de Jeffreys es la siguiente:

$$f(\theta = \mu) \propto \sqrt{I_x(\theta = \mu)} = \sqrt{-\mathbb{E} \left[\frac{\partial^2}{\partial \mu^2} \ln(f(x|\mu)) \right]}$$

donde

$$\begin{aligned} \ln(f(x|\mu)) &= -\frac{\ln(2\pi\sigma^2)}{2} - \frac{(x-\mu)^2}{2\sigma^2} \\ \Rightarrow \quad \frac{\partial}{\partial\mu} \ln(f(x|\mu)) &= \frac{x-\mu}{\sigma^2} \\ \Rightarrow \quad \frac{\partial^2}{\partial\mu^2} \ln(f(x|\mu)) &= -\frac{1}{\sigma^2} \\ \Rightarrow \quad I_x(\theta = \mu) &= \frac{1}{\sigma^2}, \end{aligned}$$

por lo que,

$$f(\theta = \mu) \propto \frac{1}{\sigma^2} \propto 1.$$

Es decir, la inicial de Jeffreys en este caso coincide con la distribución “uniforme” para $\theta = \mu$. El problema radica en que, si integramos sobre todo el espacio paramétral \mathbb{R} , el resultado no es 1 como requiere una distribución de probabilidad; esta distribución es *impropia*.

Robert (2007) presenta algunas justificaciones para utilizar distribuciones iniciales impropias. La más razonable, desde mi punto de vista, resulta ser que en realidad cualquier inferencia o decisión se deberá tomar—como ya he mencionado—con la distribución posterior. Mientras esta distribución posterior sea propia, sería posible utilizar distribuciones iniciales impropias. Otra justificación es cuando la inicial impropia puede ser vista como un límite de distribuciones propias, pero de esto hablo un poco más adelante.

4.3.2. Distribuciones informativas

Cuando efectivamente se quiere reflejar la existencia de información inicial, es usual recurrir a una familia paramétrica conocida y, con base en algún criterio o proceso, determinar la distribución inicial. Uno de dichos métodos es cuando se tiene una idea general del valor esperado y la dispersión del parámetro y se utiliza esta información para fijar la distribución inicial como aquella que tiene como media y varianza estos valores.

Por ejemplo, si suponemos que la distribución inicial de un parámetro es una dis-

tribución normal, basta con especificar μ — el parámetro del valor esperado— y σ^2 — o alguna transformación de este como parámetro de dispersión—. Una alternativa más robusta— sobre todo en espacios no restringidos— es utilizar ciertos percentiles en lugar de momentos para fijar la distribución, pero la idea sigue siendo la misma: una vez elegida una familia paramétrica, basta determinar sus parámetros con base en información inicial resumida mediante alguna cantidad.

Hay ocasiones en las que en lugar de pensar directamente en los parámetros θ podemos pensar en términos del resultado x de nuestro experimento— o una observación futura \tilde{x} — como frecuentemente podría suceder con un modelo de regresión. De manera particular, podríamos utilizar la información inicial que tengamos para predecir alguna observación futura.

Si conociéramos el valor del parámetro θ , resultaría natural intentar predecir el resultado de una observación futura \tilde{x} mediante la distribución condicional $f(\tilde{x}|\theta)$. Sin embargo, en el contexto de incertidumbre sobre los parámetros en el que estamos trabajando, aunque tenemos cierta información, en realidad el valor específico de θ es desconocido. Por eso, debemos considerar más bien la distribución conjunta de las dos cantidades desconocidas, \tilde{x} y θ . A partir de esta distribución $f(\tilde{x}, \theta)$ podemos obtener la distribución marginal de la observación futura y con ella predecir:

$$f(\tilde{x}) = \int_{\Theta} f(\tilde{x}, \theta) d\theta = \int_{\Theta} f(\tilde{x}|\theta) f(\theta) d\theta.$$

Esta distribución marginal es conocida como *distribución predictiva inicial*. Al calcularla estamos promediando las distribuciones condicionales a través de los diferentes valores que el parámetro θ puede tomar.

Hoy en día, es relativamente sencillo visualizar mediante histogramas, densidades o resúmenes las implicaciones en la distribución de x o \tilde{x} que podrían tener diferentes distribuciones iniciales $f(\theta)$. Por eso podríamos utilizar un enfoque de momentos o cuantiles para elegir la distribución inicial $f(\theta)$ que produzca las mejores predicciones para \tilde{x} . Es decir, se puede, por ejemplo, tener una idea previa del valor esperado, dispersión o cuantiles de x y mediante prueba y error seleccionar parámetros de $f(\theta)$ consistentes con dicha información previa.

También es posible especificar *distribuciones débilmente informativas* basados en la idea que, aunque haya poca información, se tienen nociones generales de valores que resultarían muy poco razonables. Estas buscan ser un justo medio entre las distribuciones completamente informativas y aquellas no informativas. Este enfoque es particularmente útil en modelos de regresión cuando se pueden tener nociones de los tamaños de los efectos de ciertas variables (Gelman y col. 2013).

Por ejemplo, Regueiro Martínez (2012) busca definir distribuciones iniciales en un modelo predictivo del número de goles anotados en partidos de futbol. Él argumenta que la inicial debe concentrar la mayor parte de la probabilidad en valores que lleven a un número de goles entre 0 y 20, sin que esto signifique que no haya probabilidad de más de 20 goles. En efecto, la mayor cantidad de goles anotados por un equipo en un partido oficial de FIFA es de 31, pero ya 20 goles es una cantidad extrema. Ciertamente “una información inicial que permita que el número esperado de goles se encuentre en el orden de los millones es inapropiada”.

4.3.3. Distribuciones Conjugadas

Sobre todo antes de los avances computacionales, uno de los objetivos en un análisis bayesiano era simplificar al máximo el proceso de cálculo de la distribución posterior. Una forma de hacerlo es elegir una distribución inicial con una forma funcional que se conserve al ser multiplicada por la verosimilitud en el teorema de Bayes. Esto da lugar a lo que se conoce como *distribuciones iniciales conjugadas*:

Definición 4.2. Distribución conjugada

Sea x una variable aleatoria con distribución $f(x|\theta)$, entonces la familia de distribuciones iniciales $\mathcal{F} = \{f(\theta)\}$ es *conjugada* para $f(x|\theta)$ si $f(\theta|x) \in \mathcal{F}$.

Dejando de lado casos triviales como cuando se considera \mathcal{F} la familia de todas las distribuciones, la noción de distribuciones condicionales es útil cuando se hace uso de alguna forma paramétrica particular. Por ejemplo, supongamos que tenemos un experimento en el que consideramos que x , dado θ , se distribuye binomial. En esta situaciones tenemos que la función de verosimilitud, que es una función de θ , tiene la siguiente forma:

$$f(x|\theta) \propto \theta^x (1-\theta)^{n-x}.$$

Si queremos que al multiplicar esta expresión por una distribución inicial $f(\theta)$ el resultado

$f(\theta|x)$ siga perteneciendo a la misma familia de $f(\theta)$, lo que necesitamos es elegir la distribución inicial que tenga la misma forma— como función de θ — que $f(x|\theta)$. En el caso de la verosimilitud binomial, resulta que una distribución inicial beta es conjugada. En efecto, su forma es la misma de la binomial:

$$\begin{aligned}\theta &\sim \text{Beta}(\alpha, \beta) \Rightarrow f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{y} \\ x|\theta &\sim \text{Binom}(n, \theta) \Rightarrow f(x|\theta) \propto \theta^x(1-\theta)^{n-x},\end{aligned}$$

por lo que, al utilizar el teorema de Bayes tenemos que,

$$\begin{aligned}f(\theta|x) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^x(1-\theta)^{n-x} \\ &\propto \theta^{(\alpha+x)-1}(1-\theta)^{(\beta+n-x)-1} \\ &\Rightarrow \theta|x \sim \text{Beta}(\alpha+x, \beta+n-x)\end{aligned}$$

La gran ventaja de las distribuciones conjugadas es que, al mantener la familia paramétrica, la actualización bayesiana se reduce a actualizar los parámetros de la distribución inicial.

Otra justificación frecuente para utilizar distribuciones conjugadas es la interpretación de los parámetros iniciales como equivalentes a una *muestra previa*. Por ejemplo, en el caso binomial, x se interpreta como el número de éxitos, mientras que $n - x$ representa el número de fracasos en el experimento. Haciendo el paralelismo con los hiperparámetros de la distribución inicial, $(\alpha - 1)$ y $(\beta - 1)$ pueden interpretarse, respectivamente, como éxitos y fracasos previos. De esta manera, la distribución posterior *suma* los éxitos previos con los del experimento, $\alpha - 1 + x$, y los fracasos previos con los del experimento, $\beta - 1 + (n - x)$.

Al utilizar esta interpretación para especificar distribuciones iniciales informativas se debe cuidar que efectivamente la hipotética muestra previa refleje de manera aceptable la incertidumbre inicial. Puede no ser tan simple determinar realmente cuánta información conlleva una muestra previa de un determinado tamaño, por lo que se pudiera correr el riesgo de subestimar la influencia que tendrá una distribución inicial determinada de esta manera sobre la distribución posterior.

En ocasiones, por el contrario, esta interpretación de muestras previas puede utilizarse

también para buscar una distribución no informativa. Podemos, por ejemplo, especificar como inicial aquella distribución cuyos parámetros representen una muestra previa de tamaño cero, lo que se conoce como distribuciones iniciales mínimo informativas límites de conjugadas. En el caso de la distribución beta, una muestra previa de tamaño cero induce que ambos parámetros sean iguales a 1, por lo que la distribución límite de conjugadas resulta ser una $\text{Beta}(\alpha = 1, \beta = 1)$ que resulta ser equivalente a una distribución uniforme sobre el intervalo $[0, 1]$. Sin embargo, como discutía al presentar las distribuciones no informativas, es frecuente que estos límites de distribuciones sean distribuciones impropias, como sucede en el caso normal (Mendoza y Regueiro 2011). A pesar de ello, cuando llevan a una distribución posterior propia, estas son frecuentemente utilizadas.

Independientemente de cuál sea la distribución inicial que se proponga en un análisis—o incluso si se proponen varias con el fin de realizar un análisis de sensibilidad ante diferentes supuestos iniciales— es importante recordar que una de las ventajas de este paradigma bayesiano es que permite transparentar o enfatizar las decisiones subjetivas del estadístico y, por consiguiente, poner en contexto el alcance de sus conclusiones.

Ciertamente esta subjetividad es uno de los postulados más comentados por quienes no comparten el paradigma bayesiano. Sin embargo, como bien nos recuerda Berger (1985) al citar a Box y a Good, cualquier supuesto en un modelo puede considerarse como subjetivo y, siempre que existan justificaciones o consideraciones sobre los supuestos iniciales que tenga un análisis, es mejor ser explícito en ellos que correr el riesgo de ignorarlos “debajo del tapete” de la objetividad.

Capítulo 5

Modelos de Regresión Bayesianos

Parafraseando a Draper y Smith (1998), al realizar un análisis estadístico sobre algún fenómeno que presenta variabilidad, muchas veces lo que se busca es explorar los efectos que algunas *variables explicativas* ejercen— o parecen ejercer— sobre una *variable de interés*. En algunos casos puede darse que, efectivamente, exista una relación funcional simple entre ambos tipos de variables. No obstante es mucho más común que, o bien la relación sea mucho más compleja de lo que podemos entender o describir, o bien simplemente nos es desconocida. En ambos casos, lo que podemos hacer es *modelar*— esto es, aproximar— la relación mediante algunas funciones matemáticas. Por razones históricas relacionadas con el trabajo de Sir Francis Galton esta clase de modelos se conocen como *modelos de regresión* (Zepeda Herrera 2015).

5.1. Regresión lineal

En el caso más sencillo, tendríamos que $y = f(x)$. Sin embargo, en la mayoría de los casos este modelo tomado literalmente podría parecernos una mala aproximación. Pensemos en el caso en el que se busca describir el peso en kilogramos de una persona a partir de la estatura. Es claro que a una misma estatura le podrían corresponder distintos valores de peso. A pesar de ello, podemos observar empíricamente que a mayor estatura *esperamos* un mayor peso. Esto nos llevaría a pensar que podemos modelar el

valor esperado de nuestra variable de interés mediante alguna función de las variables explicativas.

Esto quiere decir que, bajo una perspectiva bayesiana paramétrica, modelaremos la incertidumbre sobre la variable de interés Y mediante una distribución de probabilidad condicional $f(y|\theta, x)$. Las variables explicativas X determinan esta distribución condicional a través del valor esperado mediante una función que depende también de un subconjunto de parámetros θ_d :

$$Y|\theta, X \sim f(y|\theta, x) \quad \mathbb{E}[Y|\theta, X] = h(\theta_d, X) \quad (5.1)$$

Las formas más simples de la función $h(\theta_d, X)$ son *lineales en los coeficientes*, esto es de la forma siguiente

$$h(\theta_d, X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{d-1} X_{d-1},$$

para $d - 1$ variables explicativas y donde $\theta_d = (\beta_0, \beta_1, \dots, \beta_{d-1})$. El vector general de parámetros θ entonces se descompone en dos— $\theta = (\theta_d, \theta_r)$ — con θ_r posibles parámetros adicionales de la distribución condicional de Y dado X pero que no determinan su esperanza. Las variables explicativas pueden llegar a ser transformaciones unas de otras como cuando se buscan ajustar relaciones de orden cuadrático o como cuando se incluyen interacciones.

Debido a que no conocemos la verdadera relación entre nuestras variables, tenemos incertidumbre sobre los valores de los coeficientes que determinan nuestra función $h(\theta_d, X)$ y, posiblemente, sobre el resto de parámetros θ_r . Así pues, bajo la perspectiva bayesiana debemos reflejar dicha incertidumbre también mediante alguna distribución de probabilidad $f(\theta)$. Buscaremos, entonces, reducirla mediante la recolección de datos (y, x) que nos permitan, a través del teorema de Bayes obtener la distribución posterior $f(\theta|y, x)$.

Como bien notan tanto Gelman y col. (2013) como Congdon (2006), el modelo más general debería incorporar también la incertidumbre que pudiera existir sobre las variables explicativas X derivada, por ejemplo, de posibles errores de medición. Sin embargo, si se puede asumir que los parámetros φ de la distribución de X , $f(x|\varphi)$, son independientes de θ —es decir $f(\varphi, \theta) = f(\varphi)f(\theta)$ —al aplicar el Teorema de Bayes veríamos que la distribución posterior $f(\theta|y, x)$ no dependería de φ por lo que podemos proceder

ignorando dicha incertidumbre para efectos de las inferencias sobre θ . Por eso— y para simplificar la notación— a partir de aquí omitiré la condicional en X , con lo que $f(\theta|y, x)$ se convierte en $f(\theta|y)$, por ejemplo.

El modelo de regresión más usual es cuando se asume que la variable de interés se distribuye normal con una media que depende linealmente de las variables explicativas. Esto es, supongamos que tenemos N conjuntos de observaciones, condicionalmente independientes, $\{(y_i, x_{i,1}, \dots, x_{i,d-1})\}_{i=1}^N$ donde nuestra variable de interés es Y y contamos con $d - 1$ variables explicativas $\{X_j\}_{j=1}^{d-1}$, la regresión lineal normal bajo los supuestos usuales es:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{d-1} x_{i,d-1} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad \forall i = 1, \dots, N.$$

Que en términos de (5.1) sería

$$y_i|\theta \sim N(\mu_i, \sigma^2) \quad \mu_i = \mathbb{E}[y_i|\theta] = \beta x_i \quad \forall i = 1, \dots, N,$$

donde $\beta = (\beta_0, \beta_1, \dots, \beta_{d-1})$, $x_i = (1, x_{i,1}, \dots, x_{i,d-1})$ y tal que $\theta = (\beta, \sigma^2)$ tenga alguna distribución inicial apropiada.

También es posible aprovechar la notación matricial para simplificar estas expresiones, así como trabajar con ellas. Definiendo lo siguiente,

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d-1} \end{pmatrix} \in \mathbb{R}_{N \times d},$$

$$y = (y_1, \dots, y_N)^T \in \mathbb{R}_{N \times 1},$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_{d-1})^T \in \mathbb{R}_{d \times 1},$$

tenemos que el modelo de regresión normal puede ser expresado de manera compacta de la siguiente forma:

$$y|\theta \sim N_N(X\beta, \sigma^2 \mathbb{I}_N) \quad \text{tal que } \theta = (\beta, \sigma^2) \sim f(\beta, \sigma^2),$$

donde $N_N(X\beta, \sigma^2 \mathbb{I}_N)$ representa una distribución normal N -variada con media $X\beta$ y varianzas individuales σ^2 y $f(\beta, \sigma^2)$ una distribución inicial para los parámetros desco-

nocidos.

5.1.1. Problema de multicolinealidad

Antes de continuar, debo señalar un posible problema cuando uno ajusta modelos de regresión. En el **Anexo A** se muestra el desarrollo del modelo lineal normal utilizando iniciales conjugadas—que pueden ser informativas o débilmente informativas—así como un caso límite que resulta ser la inicial no informativa de Jeffreys. El estimador máximo verosimil $b = (X^T X)^{-1} X^T y$ es clave para el análisis. En dicho desarrollo asumo que es posible calcular b , aún cuando no siempre es el caso.

El estimador involucra la matriz inversa $(X^T X)^{-1}$, pero no siempre es posible invertir una matriz. Para ello se requiere que todas sus columnas sean lo que en álgebra lineal se conoce como *linealmente independientes*. Esto no pasa si hay variables explicativas tales que exista una combinación lineal de ellas que sea igual a 0 para todos los datos (Gelman y Hill 2006). Cuando esto sucede se dice que dichas variables son linealmente dependientes o bien que son *colineales*.

El caso más sencillo de colinealidad es cuando se utiliza una variable categórica X a través de su representación como variables indicadoras. Esto quiere decir que si X tiene J categorías, podemos representarla como un vector de J variables diferentes $X \equiv (X_1, X_2, \dots, X_J)$. Cuando X toma el valor de la j -ésima categoría, todas las variables del vector son iguales a 0 y solamente hay un 1 en la correspondiente a la categoría de la observación: $(X_1 = 0, \dots, X_j = 1, \dots, X_J = 0)$. Así, la suma de las J variables dicotómicas siempre es igual a 1. El problema se materializa cuando consideramos que el intercepto en una regresión es equivalente a una variable explicativa ficticia igual a 1, por lo que al restarle la suma de las J variables dicotómicas se obtiene una combinación lineal de variables explicativas igual a 0 para todos los datos, i.e. dependencia lineal.

Una forma de evitar el problema es excluyendo el término del intercepto en la regresión. Sin embargo, si existe una segunda variable categórica, el problema se repite. Para resolverlo, entonces, se tienen que incorporar restricciones. Una de las más frecuentes es la llamada restricción de esquina (Regueiro Martínez 2012). Esta consiste en obligar a uno de los coeficientes de las categorías a que tome el valor de 0, lo que equivale a excluir una de las J categorías y solo incorporar $J - 1$ variables dicotómicas. Cuando la variable toma

el valor de la categoría excluida, el vector de las $J - 1$ restantes es un vector de ceros. Así, si existen otras variables categóricas, se puede proceder de la misma manera escogiendo, para cada una de ellas, una categoría de referencia para excluir (Gelman y Hill 2006).

Otra solución consiste en imponer una restricción de tipo suma cero. En este caso el valor de uno de los coeficientes se fija como el negativo de la suma del resto de los coeficientes de las variables linealmente dependientes:

$$\beta_j = - \sum_{k \neq j} \beta_k.$$

Esta restricción tiene como consecuencia que los efectos de las diferentes categorías no pueden ser todos positivos o todos negativos (Usi López 2014).

Ahora bien, en términos estadísticos normalmente se habla del problema de *multicolinealidad* porque incluso si no hay colinealidad exacta, hay ocasiones que los datos están altamente correlacionados y esto tiene como consecuencia que sea difícil invertir la matriz $X^T X$ (Usi López 2014). Una manera de atacar este problema es agregar información que permita facilitar la inversión, por ejemplo a través de la distribución inicial (Congdon 2006).

5.2. Modelos Lineales Generalizados

El modelo lineal normal es muy flexible—sobre todo aprovechando que puede construirse en términos de variables transformadas—pero, hay ocasiones en las que pudiera no ser el más adecuado. Por ejemplo, cuando la variable tiene restricciones pudiera no ser posible utilizar la regresión normal, incluso mediante una transformación, como cuando una variable no negativa puede tomar el valor de 0 y entonces aplicar el logaritmo no funciona (Gelman y col. 2013). Este caso se presenta con frecuencia en el estudio de fenómenos políticos relacionados con el voto, pues es posible que el número de votos sea realmente 0. De manera similar, si se estudian proporciones de votos, estas toman valores entre 0 y 1 lo que puede dificultar la aplicación del modelo tradicional que tiene soporte en todos los números reales. Cuando la variable de interés representa conteos, podría también dificultarse la aplicación del modelo lineal.

En estos casos existen otras alternativas de modelado. Una familia más general de modelos de regresión son los modelos lineales generalizados, mismos que constituyen un marco teórico general y unificado para pensar en la formulación de modelos estadísticos (Dobson 2001). Antes de introducirlos, sin embargo, necesitamos una definición dada por Nieto Barajas (2016) y que es un caso particular de la que utilizaron Nelder y Wedderburn (1972) al presentar originalmente esta clase de modelos.

Definición 5.1. Familia Exponencial

Sea Y una variable aleatoria con función de distribución $f(y|\theta, \phi)$ tal que

$$f(y|\theta, \phi) = b(y, \phi) \exp\{\phi[y\theta - a(\theta)]\}, \quad (5.2)$$

donde a y b son funciones conocidas. Se dice entonces que Y pertenece a la **familia exponencial**. Cuando el parámetro de dispersión ϕ es conocido, entonces Y pertenece a la **familia exponencial natural**.

Esta familia de distribuciones incluye a las más comunes, como la Normal, la Poisson o la Bernoulli (Nieto Barajas 2016). Con este tipo de distribuciones construimos los modelos lineales generalizados. Como se verá en la definición que sigue, la idea informal de la sección anterior también está presente en ellos. En efecto, los modelos lineales generalizados nos permitirán modelar una variable aleatoria de interés mediante una distribución condicional miembro de la familia exponencial, vinculando su valor esperado con las variables explicativas.

Definición 5.2. Modelo lineal generalizado (MLG)

Un modelo lineal generalizado, abreviado **MLG**, está compuesto por 3 elementos básicos:

1. **Variable aleatoria de interés:** se supone que la variable de interés Y se distribuye condicionalmente de acuerdo a alguna ley miembro de la familia exponencial. Esto es, $f(y|\theta, \phi)$ es alguna distribución de la forma de (5.2).

$$Y|\theta, \phi \sim f(y|\theta, \phi) = b(y, \phi) \exp\{\phi[y\theta - a(\theta)]\}.$$

A este elemento se le conoce también como *componente aleatoria*.

2. **Predictor lineal:** las variables explicativas X forman un predictor lineal en los coeficientes de la forma $\eta = X\beta$. Esto es, suponiendo que tenemos $d - 1$ variables

explicativas X e incluyendo quizás a un intercepto constante:

$$\eta = X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_{d-1} X_{d-1}.$$

Este elemento es llamado también *componente sistemática*.

- 3. Función liga:** el predictor lineal se vincula con nuestra variable de interés mediante una función liga invertible $g(\cdot)$. La forma específica del vínculo es que el valor del predictor lineal es el resultado de aplicar la función liga al valor esperado condicional de la variable de interés. Esto es, sea μ el valor esperado de $Y|\theta, \phi$, entonces

$$g(\mu) = \eta = X\beta.$$

Otra forma de ver la función liga es que el valor esperado de $Y|\theta, \phi$ es el resultado de aplicar al predictor lineal la función inversa de la liga:

$$\mu = g^{-1}(\eta) = g^{-1}(X\beta).$$

Bajo el paradigma bayesiano, además, un MLG debe incluir un cuarto elemento que refleje la incertidumbre existente sobre los parámetros del modelo:

- 4. Distribución Inicial:** la incertidumbre o el conocimiento inicial que se tenga sobre los parámetros θ y, en su defecto ϕ , se refleja en una distribución inicial de probabilidad $f(\theta, \phi)$. Nótese que en un MLG los parámetros θ de la variable de interés, incluyen a los coeficientes β del predictor lineal.

En lo que sigue, me referiré a N observaciones de una variable de interés Y , condicionalmente independientes dadas $d - 1$ variables explicativas, de manera tal que para cada individuo $i \in \mathbb{N}_N$, y_i representaría la observación de la variable de interés, $X_i = (1, x_{i,1}, \dots, x_{i,d-1})$ el correspondiente vector de variables explicativas y $\beta = (\beta_0, \beta_1, \dots, \beta_{d-1})^T$ el vector de coeficientes del predictor lineal.

Modelo Normal

El modelo usual de regresión lineal para variables de interés continuas en los reales puede expresarse como MLG de la siguiente manera:

$$y_i | \beta, \sigma^2 \sim N(\mu_i, \sigma^2) \quad \forall i \in \mathbb{N}_N$$

$$\text{con } \mu_i = X_i\beta \\ \beta, \sigma^2 \sim f(\beta, \sigma^2)$$

En este caso tenemos que la función liga resulta ser la identidad, lo que se conoce como *liga canónica*.

Modelo Poisson

Cuando nuestra variable de interés representa conteos, un modelo usual es la regresión Poisson o loglineal, en el que la liga resulta ser el logaritmo natural. En este caso el parámetro de dispersión ϕ no está presente; dicho de otra forma $\phi = 1$, por lo que la distribución Poisson es un caso de una distribución exponencial natural.

$$y_i|\beta \sim Poi(\lambda_i) \quad \forall i \in \mathbb{N}_N \\ \text{con } \ln(\lambda_i) = X_i\beta \\ \beta \sim p(\beta)$$

5.2.1. Regresión logística

A continuación presento uno de los MLG más conocidos y utilizados: la regresión logística. Esta busca modelar $f(y|\theta)$ mediante una distribución Bernoulli. La relación entre las variables explicativas y el valor esperado se construye con un predictor lineal para el logit de la probabilidad de éxito del ensayo de Bernoulli. Es decir, si $p = \mathbb{P}(Y = 1)$ es la probabilidad de éxito, entonces $\ln\left(\frac{p}{1-p}\right) = X\beta$.

En otras ocasiones, nuestra variable de interés podría ser binomial, es decir, los éxitos en una serie de ensayos Bernoulli independientes. Para cada observación y_i , además de las variables explicativas, conocemos también n_i , el número de ensayos Bernoulli para el i -ésimo individuo. Es decir, es posible generalizar una regresión logística para que el número de ensayos sea mayor a 1. Veamos ahora cómo construir el modelo como un MLG.

Modelo Binomial

En primer lugar debemos probar que una variable binomial, con parámetro n conocido, puede expresarse como miembro de la familia exponencial.

$$\begin{aligned} Y|p \sim \text{Binom}(n, p) &\Leftrightarrow f(y|p) = \binom{n}{y} p^y (1-p)^{n-y} \\ &\Leftrightarrow f(y|p) = \binom{n}{y} \exp \left\{ \ln [p^y (1-p)^{n-y}] \right\} \\ &\Leftrightarrow f(y|p) = \binom{n}{y} \exp \left\{ y \ln \left(\frac{p}{1-p} \right) + n \ln (1-p) \right\} \end{aligned}$$

Definiendo el logit de p como nuestro parámetro θ , tenemos que $\theta = \ln \left(\frac{p}{1-p} \right)$ y al despejar $p = \frac{e^\theta}{1+e^\theta}$, por lo que podemos sustituir:

$$\begin{aligned} \Rightarrow f(y|p) &= \binom{n}{y} \exp \left\{ y \theta + n \ln \left(1 - \frac{e^\theta}{1+e^\theta} \right) \right\} \\ &= \binom{n}{y} \exp \left\{ y \theta + n \ln \left(\frac{1}{1+e^\theta} \right) \right\} \\ &= \binom{n}{y} \exp \left\{ y \theta - n \ln (1+e^\theta) \right\} \end{aligned} \tag{5.3}$$

Así, tenemos que una variable binomial con parámetro n conocido se expresa de la forma de (5.2) tomando los siguientes valores:

$$\begin{aligned} \theta &= \ln \left(\frac{p}{1-p} \right) & \phi &= 1 \\ a(\theta) &= n \ln (1+e^\theta) & b(\theta, y) &= \binom{n}{y} \end{aligned}$$

Ahora bien, habiendo ilustrado la pertenencia a la familia exponencial, el MLG binomial normalmente se plantea en términos de p_i , el valor esperado de cada uno de los ensayos de Bernouilli (Gelman y col. 2013):

$$\begin{aligned} Y_i|\beta &\sim \text{Binom}(n_i, p_i) \quad \forall i \in \mathbb{N}_N \\ \text{con} \quad \ln \left(\frac{p_i}{1-p_i} \right) &= X_i \beta \\ \beta &\sim f(\beta) \end{aligned}$$

En este caso, debido a que cada valor esperado binomial μ_i es igual a $n_i p_i$, tenemos que $p_i = \mu_i/n_i$. Por lo que la tradicional función logística implícitamente refleja la siguiente función liga:

$$\begin{aligned} \ln\left(\frac{p_i}{1-p_i}\right) &= \ln(p_i) - \ln(1-p_i) \\ &= \ln\left(\frac{\mu_i}{n_i}\right) - \ln\left(1 - \frac{\mu_i}{n_i}\right) \\ &= \ln\left(\frac{\mu_i}{n_i}\right) - \ln\left(\frac{n_i - \mu_i}{n_i}\right) \\ &= \ln(\mu_i) - \ln(n_i - \mu_i) \\ \therefore \quad g(\mu_i) &= \ln\left(\frac{\mu_i}{n_i - \mu_i}\right) \end{aligned}$$

5.2.1.1. Problema analítico

Ahora supongamos que queremos realizar un análisis bayesiano y actualizar nuestras creencias mediante el teorema de Bayes. Podríamos llevarnos la desagradable sorpresa de que una regresión logística no tiene las mismas facilidades analíticas que sí tiene el modelo normal del **Anexo A**. De manera particular, no podemos encontrar una distribución conjugada.

En efecto, definiendo $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right) = X_i\beta$ y utilizando la forma de la función de verosimilitud en (5.3), donde y representa el vector de observaciones de nuestra variable de interés:

$$\begin{aligned} f(y|\beta) &= f(y_1, \dots, y_N|\beta) = \prod_{i=1}^N f(y_i|\beta) \\ &= \prod_{i=1}^N \binom{n_i}{y_i} \exp\{y_i X_i\beta - n_i \ln(1 + e^{X_i\beta})\} \\ &\propto \prod_{i=1}^N \exp\{y_i X_i\beta - n_i \ln(1 + e^{X_i\beta})\} \\ &\propto \exp\left\{\sum_{i=1}^N y_i X_i\beta - n_i \ln(1 + e^{X_i\beta})\right\} \end{aligned} \tag{5.4}$$

Esta no es una forma funcional que permita encontrar fácilmente una inicial conjugada

conocida.

En el mismo sentido, supongamos que tomamos como iniciales independientes para los coeficientes, distribuciones normales centradas en 0 con cierta desviación estándar común σ , conocida. Veamos cómo luce la actualización de la verosimilitud en (5.4) por esta inicial normal multivariada:

$$\begin{aligned}
 f(y|\beta) &\propto \exp \left\{ \sum_{i=1}^N y_i X_i \beta - n_i \ln(1 + e^{X_i \beta}) \right\} \\
 &\propto \exp \left\{ (\beta y X)^T - \sum_{i=1}^N n_i \ln(1 + e^{X_i \beta}) \right\}; \\
 f(\beta) &= \frac{1}{\sqrt{(2\pi)|\sigma^2 \mathbb{I}_N|}} \exp \left\{ -\frac{1}{2} \beta^T (\sigma^2 \mathbb{I}_N)^{-1} \beta \right\} \\
 &\propto \exp \left\{ -\frac{\beta^T \beta}{2\sigma^2} \right\}, \\
 \Rightarrow f(\beta|y) &\propto f(y|\beta) f(\beta) \\
 &\propto \exp \left\{ -\frac{\beta^T \beta}{2\sigma^2} + (\beta y X)^T - \sum_{i=1}^N n_i \ln(1 + e^{X_i \beta}) \right\} \\
 &\propto \exp \left\{ -\frac{1}{2\sigma^2} [\beta^T \beta - 2\sigma^2 (\beta y X)^T] \right\} \prod_{i=1}^N (1 + e^{X_i \beta})^{-n_i}
 \end{aligned}$$

Esta expresión tampoco permite un tratamiento analítico bajo una forma funcional conocida. No es posible integrarla para conocer la constante de normalización y así tener completa la distribución posterior.

¿Acaso es posible llevar a cabo una simple regresión logística bajo el paradigma bayesiano? Por mucho tiempo la respuesta a esta pregunta era prácticamente no. Como la integral no se puede calcular, la alternativa es obtenerla mediante aproximaciones analíticas como la llamada *aproximación de Laplace* o bien mediante integración numérica mediante métodos de cuadratura (Nieto Barajas 2016; Gutiérrez Peña 1997). Sin embargo, estos métodos son muy difíciles de aplicar cuando la dimensión del problema empieza a crecer.

Afortunadamente, el desarrollo de las computadoras permitió superar este obstáculo

y, desde la década de 1990, el paradigma bayesiano ha ido creciendo con fuerza. Hoy es posible aplicar modelos como la regresión logística y más complejos sin problemas. Por ello, pospongo el tema de cómo calcular las integrales al capítulo 6 y continúo con la discusión general de los modelos de regresión.

5.3. Modelos Jerárquicos

Los modelos de regresión—ya sean lineales o lineales generalizados—pueden interpretarse como un método que permite aproximar cómo cambia el valor esperado de una variable de interés a través de subpoblaciones definidas por funciones lineales de unas variables explicativas (Gelman y Hill 2006). En efecto, podemos pensar que diferentes valores de las variables explicativas definen diferentes subpoblaciones o grupos cuyos valores promedio en la variable de interés está determinado por la regresión. A pesar de esta variabilidad, la *forma específica* como cambian estos valores es la misma a través de las subpoblaciones pues está dada por los mismos coeficientes. De manera informal, podemos decir que las observaciones de todas las subpoblaciones tienen cierta simetría que las hace similares entre sí a nuestros ojos y por eso comparten los mismos parámetros.

No obstante, hay ocasiones en las que dicha simetría es más débil o, mejor dicho, podemos distinguir claramente subpoblaciones o grupos de observaciones como más homogéneas al interior que entre sí. Es decir, la simetría la encontramos para observaciones provenientes de la misma subpoblación y no entre aquellas que pertenezcan a grupos distintos. El ejemplo más claro es cuando, por diseño, nuestro estudio está compuesto por estratos o clusters de observaciones.

Uno esperaría mayor homogeneidad de los resultados en algún examen entre estudiantes de una misma escuela que entre aquellos de diferentes instituciones educativas (Ortiz Mancera 2012); de la misma manera, cuando se busca estimar el resultado de una elección con base en encuestas publicadas habrían más diferencias entre encuestas de diferentes casas que entre ejercicios de la misma organización (Zepeda Herrera 2018). En la práctica, este razonamiento implica que no queremos tratar a las distintas subpoblaciones o grupos con la misma cuchara y, por tanto, deben tener distintos parámetros o coeficientes; por ejemplo, cuando los efectos estacionales en la prevalencia de una enfermedad son distintos para diferentes regiones geográficas (Usi López 2014).

Una primera posibilidad es mantener una sola regresión incluyendo variables indicadoras de pertenencia al grupo. Sin embargo, este camino puede fallar incluso en situaciones más o menos sencillas. ¿Qué pasa si se tienen variables explicativas a nivel grupo? No es posible incluir al mismo tiempo tanto estas variables como las indicadoras pues tendríamos un problema de multicolinealidad (Gelman y Hill 2006). Otra alternativa es ajustar regresiones separadas para cada grupo. Sin embargo, el hecho de que se traten de *subpoblaciones* y no de *poblaciones* o fenómenos completamente distintos nos haría pensar que si bien los parámetros son diferentes, deben estar de todas formas relacionados. Más aún, ajustar regresiones separadas a cada grupo tiene el defecto de que cada una de ellas incorpora exclusivamente la información del grupo respectivo, desperdiando de alguna manera la información sobre el fenómeno o población general que los datos de las otras subpoblaciones pueden aportar.

Así pues, tenemos dos extremos que pudieran no parecernos ideales. Por un lado, podemos pensar que todos los datos son similares entre sí y, por tanto, ajustamos una sola regresión. Esta opción sería una *agrupación completa* o *complete-pooling*. El costo de tomar este camino podría ser subestimar la variabilidad originada por las diferentes subpoblaciones debido a que estamos sobresimplificando el modelo asumiendo la simetría total. En el otro extremo, podríamos suponer que cada subpoblación es completamente distinta y se requieren tantas regresiones independientes como grupos hayan. Un modelo así sería *sin agrupación* o de *no-pooling*. En este caso, un riesgo que corremos es que las estimaciones podrían ser demasiado ruidosas o inciertas debido a que estarían ignorando la información sobre la población general que comparten las observaciones de las distintas subpoblaciones o por el simple hecho de que hayan muy pocas observaciones por grupo, por ejemplo.

Existe otra alternativa conocida como modelos jerárquicos o multinivel y que representan un punto intermedio entre las dos anteriores mediante una *agrupación parcial* de los datos o *partial-pooling*. Su objetivo es reconocer las diferencias que existen a través de diferentes subpoblaciones mediante una estructura jerárquica que incorpore la información de todas ellas con relación a la población general.

5.3.1. Intercambiabilidad

Los modelos jerárquicos tienen su piedra angular en el concepto de intercambiabilidad que es la formalización de la idea de simetría u homogeneidad en los datos de la que hablaba. Su definición está basada en la de Bernardo y Smith (2000).

Definición 5.3. Intercambiabilidad

Sean X_1, \dots, X_n una sucesión finita de variables aleatorias. Se dice que son **finitamente intercambiables** si y solo si, para toda permutación π definida sobre el conjunto de índices \mathbb{N}_n , su distribución conjunta satisface que

$$p(X_1 = x_1, \dots, X_n = x_n) = f(X_1 = x_{\pi(1)}, \dots, X_n = x_{\pi(n)}).$$

La sucesión infinita X_1, X_2, \dots es **infinitamente intercambiable** si y solo si toda subsucesión finita es finitamente intercambiable.

Cuando asumimos que unas variables aleatorias son independientes, se cumple la propiedad de intercambiabilidad, pero el concepto de intercambiabilidad es un poco más general; queremos decir que los índices o etiquetas de las observaciones pueden cambiar y no los distinguiríamos. Para comprobar que independencia implica intercambiabilidad basta observar que, si hay independencia, la distribución conjunta se descompone en un producto y “el orden de los factores no altera el resultado”. Sin embargo, no es cierto que intercambiabilidad implique independencia, como puede verse en el contraejemplo de Gutiérrez Peña (1998). En este caso es natural asumir una representación de independencia condicional dado un parámetro común para las observaciones.

Dicha *naturalidad* puede justificarse a partir de un caso particular del Teorema de representación de De Finetti que establece que la densidad conjunta de unas variables aleatorias intercambiables puede representarse mediante el uso de la independencia condicional dado un parámetro:

$$f(x_1, \dots, x_n) = \int_{\Theta} \prod_{i=1}^n f(x_i | \theta) f(\theta) d\theta.$$

El lector interesado puede consultar Gutiérrez Peña (1998) o Bernardo y Smith (2000).

Cabe aclarar que la condición de intercambiabilidad que supone el Teorema de representación de De Finetti es de intercambiabilidad infinita, pero normalmente al modelar tratamos con observables que más bien satisfarían solamente una intercambiabilidad finita. A pesar de esto, si se da el caso en el que la secuencia finita de variables aleatorias observadas pueda representar una parte de una secuencia más larga de variables finitamente intercambiables, es posible asumir intercambiabilidad infinita como una aproximación suficientemente buena y proceder en consecuencia (Bernardo y Smith 2000). Este es el supuesto, a veces tácito, que normalmente se hace en un modelo estadístico (Gelman y col. 2013).

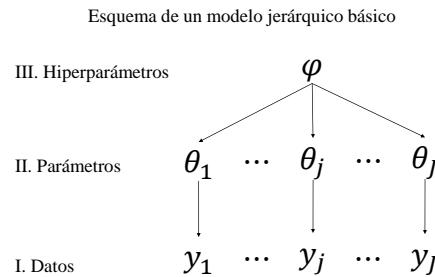


Figura 5.1: Esquema de un modelo jerárquico básico para J subpoblaciones. Los datos de cada subpoblación y_j son condicionalmente independientes dado un parámetro θ_j . Los parámetros θ son, a su vez, intercambiables y se modelan condicionalmente independientes dado un hiperparámetro poblacional φ . Fuente: elaboración propia.

Ahora bien, un modelo jerárquico establece la intercambiabilidad de las observaciones al interior de cada subpoblación mediante la independencia condicional dado un parámetro grupal. Esto haría también un modelo de no agregación, mientras que una agregación completa asume esto para toda la población. Lo que distingue a un modelo jerárquico es que adicionalmente supone intercambiabilidad para los parámetros de cada grupo de manera que estos sean condicionalmente independientes dado uno o más *hiperparámetros* poblacionales. Esto puede verse de manera esquemática en la **Figura 5.1**, mientras que en términos de las distribuciones del modelo tendríamos los siguientes niveles:

$$\text{I Datos} \quad f(y|\theta) = f(y_1, \dots, y_J | \theta_1, \dots, \theta_J) = \prod_{j=1}^J f(y_j | \theta_j)$$

$$\begin{array}{ll} \text{II Parámetros} & f(\theta|\varphi) = f(\theta_1, \dots, \theta_J|\varphi) = \prod_{j=1}^J f(\theta_j|\varphi) \\ \text{III Hiperparámetros} & f(\varphi) \end{array}$$

Un modelo jerárquico aumenta el número de parámetros a estimar al agregar los hiperparámetros. Entonces, tenemos que la distribución inicial de los parámetros depende de un hiperparámetro mismo que, para realizar el aprendizaje bayesiano, deberá tener una distribución *hiperinicial*. El problema consiste en inferir tanto las características de las subpoblaciones, θ_j , como aquellas poblacionales, φ (Gutiérrez Peña 1998):

$$\begin{aligned} f(\theta, \varphi|y) &\propto f(y|\theta, \varphi)f(\theta, \varphi) \\ &\propto f(y|\theta)f(\theta|\varphi)f(\varphi) \\ &\propto f(\varphi) \prod_{j=1}^J f(y_j|\theta_j) \prod_{j=1}^J f(\theta_j|\varphi) \\ &\propto f(\varphi) \prod_{j=1}^J f(y_j|\theta_j)f(\theta_j|\varphi) \end{aligned}$$

5.3.2. Regresiones jerárquicas

En términos de regresiones, los modelos jerárquicos son aquellos en los que los coeficientes también son modelados mediante hiperparámetros estimados con los mismos datos (Gelman y Hill 2006). Esto permite pensar en regresiones a distintos niveles y con variables grupales, estudios con base en muestreo estratificado e incluso estructuras no necesariamente anidadas. Las ventajas de los modelos jerárquicos hacen que, para algunos investigadores, las regresiones multinivel merezcan ser el enfoque predeterminado (McElreath 2015).

De manera general es posible clasificar las regresiones jerárquicas en tres grandes categorías: interceptos variables, pendientes variables e interceptos y pendientes variables. De manera gráfica estos pueden verse en la **Figura 5.2**. Para presentarlas, supongamos que tenemos N observaciones $\{(y_i, x_i, u_{j[i]})\}_{i=1}^N$ agrupadas en J grupos. Nuestra variable de interés es Y , contamos con una variable explicativa a nivel individuo, X , y una a nivel grupo, U . Debido a que ahora tenemos J grupos, indico con la notación $j[i]$ el grupo al que pertenece el i -ésimo individuo. En el **Cuadro 5.1** presento ejemplos esquemáticos de este tipo para modelos lineales en los que α representa el intercepto y

Nivel de jerarquía	Interceptos variables	Pendientes variables	Interceptos y pendientes variables
I. Datos	$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$	$y_i \sim N(\alpha + \beta_{j[i]} x_i, \sigma_y^2)$	$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$
II. Parámetros	$\alpha_{j[i]} \sim N(a + bu_{j[i]}, \sigma_\alpha^2)$	$\beta_{j[i]} \sim N(a + bu_{j[i]}, \sigma_\beta^2)$	$\alpha_{j[i]} \sim N(a_\alpha + b_\alpha u_{j[i]}, \sigma_\alpha^2)$ $\beta_{j[i]} \sim N(a_\beta + b_\beta u_{j[i]}, \sigma_\beta^2)$
III. Hiperparámetros	$f(\varphi) = f(\beta, \sigma_y^2, a, b, \sigma_\alpha^2)$	$f(\varphi) = f(\alpha, \sigma_y^2, a, b, \sigma_\beta^2)$	$f(\varphi) = f(\sigma_y^2, a_\alpha, b_\alpha, \sigma_\alpha^2, a_\beta, b_\beta, \sigma_\beta^2)$

Cuadro 5.1: Ejemplos esquemáticos de regresiones jerárquicas lineales. Fuente: elaboración propia.

β la pendiente de la recta; en el caso de los *hipercoeficientes* estos se representan por a y b .

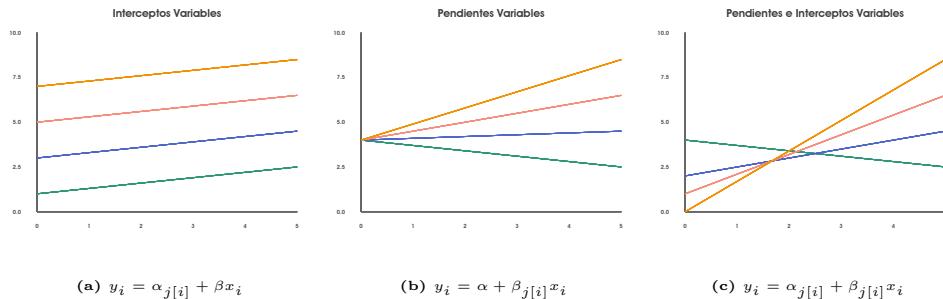


Figura 5.2: Ilustración de los tres tipos generales de modelos jerárquicos. Fuente: elaboración propia con base en la figura (11.1) de Gelman y Hill (2006).

Una de las ventajas que tienen los modelos jerárquicos es que cuando se tiene una variable categórica— a diferencia de lo que sucede en un modelo lineal normal (ver página 55)— no es necesario excluir una categoría de referencia. Por el contrario, es posible conservar el intercepto y todas las variables indicadoras dicotómicas con sus respectivos coeficientes. La clave está en que los coeficientes ahora tienen a su vez una distribución que los modela, misma que tiene el efecto de agregar un término a la matriz $X^T X$, lo que la convierte en una matriz invertible y elimina este caso particular de multicolinealidad (Gelman y Hill 2006).

Regresión logística jerárquica

Naturalmente, es posible proponer MLG jerárquicos. Por ejemplo, una regresión logística jerárquica con algunas *hiperiniciales* arbitrarias sería la siguiente:

$$y_i | \alpha_{j[i]}, \beta_{j[i]} \sim \text{Binom}(n_i, p_i)$$

$$\begin{aligned}
 \ln\left(\frac{p_i}{1-p_i}\right) &= \alpha_{j[i]} + \beta_{j[i]}x_i \\
 \alpha_{j[i]} &\sim N(a_\alpha + b_\alpha u_{j[i]}, \sigma_\alpha^2) \\
 \beta_{j[i]} &\sim N(a_\beta + b_\beta u_{j[i]}, \sigma_\beta^2) \\
 a_\alpha &\sim N(0, 5) \quad a_\beta \sim N(0, 5) \\
 b_\alpha &\sim N(0, 5) \quad b_\beta \sim N(0, 5)
 \end{aligned}$$

Como puede intuirse, la estructura de un modelo jerárquico crece rápidamente. Sin embargo, recordemos que la *única receta de la inferencia bayesiana* sigue siendo la misma: encontrar la distribución condicional de todas aquellas cantidades de interés cuyo valor desconocemos, dado el valor conocido de las variables observadas. Gracias a herramientas computacionales este cálculo es posible para modelos cada vez más complejos. En el siguiente capítulo, entonces, discutiré algunos métodos computacionales que permiten este aprendizaje bayesiano.

Capítulo 6

Cómputo bayesiano

De acuerdo con Ross (2010), los resultados más importantes y más conocidos de la teoría de probabilidad son los llamados teoremas límite, en particular aquellos conocidos como *leyes de los grandes números*. La idea general podemos tomarla de Jakob Bernoulli, el primero en presentar un teorema de este tipo, y estriba en que todos los hombres saben “por algún instinto de la naturaleza *per se* y sin ninguna instrucción previa, que entre más observaciones hay, menor es el peligro de alejarse del blanco” (Pulskamp 2009). Es decir, si tenemos suficientes realizaciones de un experimento, podemos estimar con mucha precisión aquello que buscamos.

Después de varios avances históricos que pueden consultarse en Seneta (2013), hoy contamos con las leyes débil y fuerte de los grandes números (Ross 2010). Ambas nos dicen que, conforme el tamaño de una muestra aleatoria aumenta, los promedios empíricos convergen a los promedios teóricos. Una manera común de ejemplificar este fenómeno es mediante el lanzamiento sucesivo de monedas. En este caso los volados *simulan* observaciones de una variable aleatoria de ensayos Bernoulli y, al tener una muestra suficientemente grande, se comienza a apreciar la convergencia hacia la probabilidad de éxito.

Gracias al avance tecnológico, hoy ya no tenemos necesariamente que lanzar volados físicamente sino que los simulamos desde una computadora, a partir de la generación de números pseudoaleatorios, diseñados de manera tal que satisfagan todas las propiedades básicas de números auténticamente aleatorios (Ross 2013). Podemos pedirle a la computadora que simule una gran cantidad de volados *justos* y registre la proporción empírica

acumulada de los que cayeron águila. Conforme más aumenta el número de volados, más nos acercamos a 0.5, la proporción teórica. Esto se puede repetir para otras series de volados y el comportamiento es el mismo, como puede apreciarse en la **Figura 6.1**.

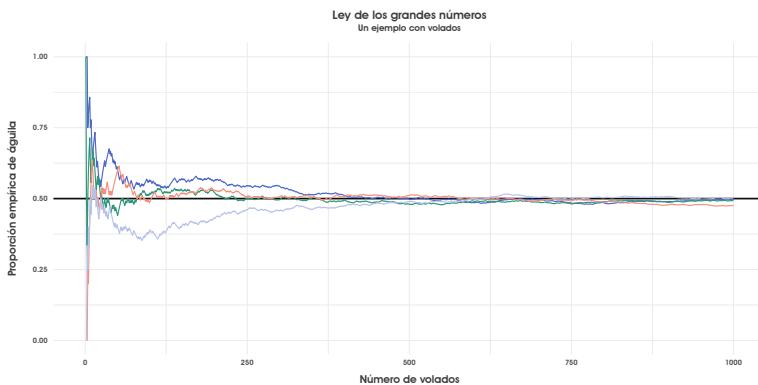


Figura 6.1: Ilustración de las leyes de los grandes números mediante volados simulados por computadora. Conforme el número de volados aumenta, la proporción empírica converge a la proporción teórica. Esto pasa para cada serie de volados. Fuente: elaboración propia.

Es cierto que esta práctica de aprender de un sistema, simulando con muestreo aleatorio no surge con las computadoras (Owen 2013). Ya desde 1812 Laplace había sugerido la posibilidad de estimar empíricamente el valor de π mediante el llamado problema de la aguja de Buffon (Ragheb 2013). Sin embargo, como el ejemplo de los volados muestra, sí han sido las computadoras las que han potenciado la utilidad de los métodos de simulación. ¿Cuánto hubiéramos tardado en lanzar los 4,000 volados que la computadora simuló al instante? Más aún, esta capacidad computacional consolidó la utilidad de la estadística bayesiana. ¡Y pensar que todo empezó con una guerra y un juego de solitario!, como explico a continuación.

6.1. Monte Carlo

La simulación por computadora ha permitido el desarrollo de la estadística bayesiana, particularmente desde la década de los 90 del siglo pasado (Robert y Casella 2011). Sin embargo, la semilla de este desarrollo ya había sido plantada medio siglo antes desde el terreno de la física, desafortunadamente a causa de la Segunda Guerra Mundial, por los científicos del laboratorio de Los Alamos, encargados del proyecto Manhattan y el

desarrollo de más armas de fisión nuclear.

Uno de esos científicos fue un matemático polaco-estadounidense llamado Stanislaw Ulam. En 1946, aburrido convaleciendo por una enfermedad, comenzó a preguntarse sobre la probabilidad de ganar en un juego de solitario. Después de mucho batallar con los cálculos de combinatoria se planteó si no sería más práctico estimarla simulando muchas partidas en una de las primeras computadoras electrónicas. Y ahí surgió el eureka: ¿por qué no hacer lo mismo para los problemas de física nuclear en los que estaban trabajando en Los Alamos? (Eckhardt 1987). Al igual que las partidas de solitario, podrían simular muchas realizaciones de los procesos físicos bajo estudio y estimar los resultados más probables.

Stan compartió su idea con John von Neumann quien, sorprendido y emocionado, le envió una carta a Richard Richtmayer—el líder del equipo en Los Alamos—con todos los cálculos necesarios para llevar a cabo el proyecto (von Neumann y Richtmayer 1947). El método fue rápidamente adoptado por todos en Los Alamos, tanto que otro físico, Nicholas Metropolis, sugirió llamarlo *Monte Carlo*, bromeando sobre un tío apostador de Stan, que vivía pidiendo prestado dinero porque “simplemente tenía que ir a Monte Carlo” (Metropolis 1987). Después de un arduo trabajo, el método pareció funcionar—gracias en buena medida al trabajo de programación de Klara von Neumann (Haigh, Priestley y Rope 2014)—y el propio Metropolis publicó, junto con Stan, un primer paper presentándolo a grandes rasgos (Metropolis y Ulam 1949).

De manera concreta, el método es la conjunción de la simulación con la ley de los grandes números. Las cantidades que requerían calcular eran valores esperados de la siguiente forma:

$$h^* = \mathbb{E}[h(Z)] = \int_Z h(z)f(z)dz, \quad (6.1)$$

donde h es una función de interés y $f(z)$ es la distribución de probabilidad sobre las *configuraciones* z en las que podía encontrarse el sistema físico. El gran problema era—y sigue siendo—que estas integrales típicamente no pueden calcularse ni analíticamente ni por métodos numéricos tradicionales. Sin embargo, si se tiene una muestra aleatoria de valores de Z provenientes de su distribución f , se puede aproximar h^* con el promedio

empírico, que en este contexto se conoce como *estimador de Monte Carlo*:

$$h^* \approx \hat{h} = \sum_{i=1}^N \frac{h(z_i)}{N}$$

Entonces, los científicos en Los Alamos se dedicaron a encontrar algoritmos eficientes para obtener una muestra aleatoria de variables provenientes de diferentes distribuciones f .

Lo importante para este trabajo es que este es exactamente el mismo problema que se tiene en la aplicación de la estadística bayesiana y que hizo que por décadas—por no decir siglos— fuera poco menos que imposible llevar a cabo análisis bayesianos no triviales, como ya anticipa en 5.2.1.1. Muchas de las integrales que surgen en la estadística bayesiana no pueden ser calculadas de manera analítica. El ejemplo más frecuente es la constante normalizadora del teorema de Bayes. ¿Cómo aplicar el teorema si el denominador no puede ser calculado?

No obstante, la forma de (6.1) permite llevar a cabo una gran variedad de *resúmenes inferenciales* (Gutiérrez Peña 1997). La posibilidad de realizar las correspondientes *integraciones por Monte Carlo* hace que, en la práctica, exista una dualidad entre una distribución o densidad y una muestra proveniente de ella (Smith y Gelfand 1992). La única receta de la inferencia bayesiana—ver la página 43—, en la práctica se convierte en obtener una muestra aleatoria de la distribución posterior lo suficientemente grande para, con ella, estimar por Monte Carlo los resúmenes inferenciales requeridos.

Hay varios métodos de simulación de variables aleatorias que permiten obtener muestras aleatorias independientes y calcular un estimador de Monte Carlo. Entre los más mencionados podemos encontrar el método de inversión, el de aceptación y rechazo o el muestreo por importancia (Ross 2013; Robert y Casella 2010). Sin embargo, los físicos de Los Alamos pronto se dieron cuenta que, en lugar de buscar realizar directamente simulaciones independientes, era más práctico hacer simulaciones secuenciales que dependieran entre sí.

6.2. MCMC

La forma de realizar simulaciones correlacionadas que logren simular de manera más eficiente que los métodos directos tradicionales de aceptación y rechazo o muestreo por importancia es utilizar *cadenas de Markov*.

Definición 6.1. Cadena de Markov

Una *cadena de Markov* $\{Z^{(n)}, n = 0, 1, 2, \dots\}$ es una secuencia de variables aleatorias tales que satisfacen la siguiente *propiedad de Markov* para toda n

$$Z^{(n+1)}|Z^{(n)}, \dots, Z^{(0)} \sim Z^{(n+1)}|Z^{(n)} \sim p(z^{(n+1)}|z^{(n)})$$

Si llamamos *estados* a los eventos que suceden en la cadena en cada punto en el tiempo, podemos decir que la distribución condicional del estado futuro de una cadena de Markov dada toda su historia depende exclusivamente del estado presente y no de los estados anteriores. Dicho de otra forma, *el futuro es independiente del pasado, dado el presente*. La distribución condicional se llama *kernel de transición*, mismo que normalmente es también independiente del índice n y depende exclusivamente del estado actual y el estado futuro. Esta propiedad se llama *homogeneidad* en el tiempo y permite simplificar la notación a $p(\tilde{z}|z)$.

La teoría de cadenas de Markov determina las condiciones bajo las cuales existen teoremas límites al estilo de las leyes de los grandes números y que en este contexto se conocen como *ergódicos*. Esta teoría escapa los objetivos particulares de la tesis pero, si es de interés para el lector, algunas referencias útiles son Rincón (2012), Neal (1993), Ross (1996) y Taylor y Karlin (1984). Baste decir por ahora que, bajo ciertas condiciones, sabemos que la distribución de $Z^{(n)}$ converge a una distribución límite conforme n tiende a infinito. Más aún, los *promedios ergódicos*— es decir los promedios acumulados de la cadena— también convergen al valor esperado de la distribución límite. Esto se puede expresar matemáticamente como sigue:

$$Z^{(N)} \xrightarrow[N \rightarrow \infty]{\mathcal{D}} Z, \quad Z \sim f(z) \tag{6.2a}$$

$$\frac{1}{N} \sum_{n=1}^N h(z^{(n)}) \xrightarrow[N \rightarrow \infty]{} \mathbb{E}_f[h(z)] \tag{6.2b}$$

Esto da lugar a los métodos de *Markov Chain Monte Carlo* o MCMC en los que el

objetivo es construir una cadena de Markov que satisfaga las condiciones necesarias y cuya distribución límite sea la distribución de la cual se quiere simular. Así, después de N transiciones de la cadena, la simulación convergería a la distribución objetivo y se podría estimar la cantidad de interés h^* mediante el estimador de MCMC:

$$h^* \approx \hat{h} = \sum_{n=1}^N \frac{h(z^{(n)})}{N}$$

6.2.1. Metropolis Hastings

El primero de estos métodos MCMC fue propuesto por las parejas de esposos Arianna y Marshall Rosenbluth y Augusta y Edward Teller junto con el propio Nicholas Metropolis en el *Journal of Chemical Physics* (Metropolis y col. 1953). Se conoce como el algoritmo de Metropolis, aunque hay quienes creen que en realidad el trabajo más fuerte lo hicieron el resto de los autores por lo que debería llamarse el algoritmo de Rosenbluth-Teller (Gubernatis 2005). Casi 20 años después, el estadístico canadiense Wilfred Keith Hastings lo generalizó (Hastings 1970), por lo que podemos hablar de algoritmos de Metropolis Hastings o MH.

Balance detallado, la clave de MH

Las cadenas de Markov para MH requieren ser homogéneas en el tiempo y que sea posible llegar a cualquier estado en un número finito de transiciones, algo que se conoce como *irreducibilidad*. Si además el kernel de transición satisface la siguiente *ecuación de balance detallado* para alguna distribución f , se dice que es *reversible* y podemos aplicar el teorema ergódico.

$$f(z)p(\tilde{z}|z) = f(\tilde{z})p(z|\tilde{z}) \quad (6.3)$$

El algoritmo de Metropolis Hastings busca construir cadenas de Markov homogéneas, irreducibles y reversibles que tengan como distribución límite a la distribución objetivo. ¿Cómo hacerlo? Para ello analicemos lo que la reversibilidad implica de manera intuitiva.

Siguiendo la argumentación de Chib y Greenberg (1995), (6.3) refleja que hay un balance entre las probabilidades de la cadena de estar en diferentes estados, de ahí el nombre. Supongamos que no se cumpliera la reversibilidad para algún kernel $q(\tilde{z}|z)$.

Entonces, sin pérdida de generalidad, para algunos estados pasaría que:

$$\frac{f(z)}{f(\tilde{z})} > \frac{q(z|\tilde{z})}{q(\tilde{z}|z)} \quad (6.4)$$

De manera un poco informal, tenemos que el miembro izquierdo de la desigualdad refleja las probabilidades relativas “necesarias” entre estar en el estado z y el estado \tilde{z} . El miembro derecho, por su parte, indica las probabilidades relativas de transitar a dichos estados bajo el kernel de transición de la cadena. La desigualdad indica que la cadena estaría transitando a \tilde{z} más de lo necesario y, equivalentemente, transitaría a z menos de lo necesario.

Para conseguir el balance requerido para aplicar el teorema ergódico, necesitamos hacer una *corrección de Metropolis* al kernel de transición, reduciendo el número relativo de veces que la cadena transite de z a \tilde{z} y aumentando el número relativo de transiciones de \tilde{z} a z . La forma de hacer la corrección es comenzar con un kernel, $q(\tilde{z}|z)$, que *proponga* un estado y agregar una *probabilidad de aceptación* $\alpha(\tilde{z}; z)$. Si la propuesta es rechazada, la cadena permanece en el mismo estado, reduciendo a la vez el número de transiciones hacia estados sobremuestreados y aumentando el número relativo de veces que estamos en el estado originalmente submuestreado. Estos dos pasos constituyen un kernel de transición de Metropolis Hastings de la siguiente forma:

$$\begin{aligned} p_{MH}(\tilde{z}|z) &= q(\tilde{z}|z)\alpha(\tilde{z}; z) \quad z \neq \tilde{z} \\ p_{MH}(z|\tilde{z}) &= 1 - \int_{\tilde{z}} p_{MH}(\tilde{z}|z)d\tilde{z} \end{aligned}$$

Queremos que el kernel p_{MH} satisfaga (6.3), esto es:

$$f(z)p_{MH}(\tilde{z}|z) = f(\tilde{z})p_{MH}(z|\tilde{z}) \quad \Leftrightarrow \quad f(z)q(\tilde{z}|z)\alpha(\tilde{z}; z) = f(\tilde{z})q(z|\tilde{z})\alpha(z; \tilde{z})$$

De acuerdo a nuestra desigualdad supuesta en (6.4), las transiciones de \tilde{z} a z se dan demasiado poco, por lo que siempre deberíamos aceptar este tipo de transiciones a fin de corregir el submuestreo. Tomemos entonces $\alpha(z; \tilde{z}) = 1$ y observemos que $\alpha(\tilde{z}; z)$ queda determinada de tal forma que logremos el balance necesario:

$$\alpha(\tilde{z}; z) = \frac{f(\tilde{z})q(z|\tilde{z})}{f(z)q(\tilde{z}|z)}$$

Si la desigualdad (6.4) fuera en el sentido contrario, i.e. $f(\tilde{z})q(z|\tilde{z}) > f(z)q(\tilde{z}|z)$, los roles de las probabilidades de aceptación se invertirían, por lo que de manera general tenemos que

$$\alpha(\tilde{z}; z) = \min \left\{ \frac{f(\tilde{z})q(z|\tilde{z})}{f(z)q(\tilde{z}|z)}, 1 \right\} \quad (6.5)$$

La utilidad de MH para la estadística bayesiana está en la forma de la probabilidad de aceptación en (6.5). Como la distribución objetivo se encuentra tanto en el denominador como en el numerador, no se requiere completa sino basta con conocerla salvo por una constante de proporcionalidad que desaparezca al realizar el cociente. Esta situación es exactamente la que impera en la aplicación de la estadística bayesiana. Recordemos el resumen del aprendizaje bayesiano en (4.1), *la posterior es proporcional a la inicial por la verosimilitud*:

$$f(\theta|y) \propto L(\theta)f(\theta).$$

Por esto, para simular valores de una distribución posterior $f(\theta|y)$ mediante MH solo necesitamos la distribución inicial $f(\theta)$, la verosimilitud $L(\theta) = f(y|\theta)$ y un kernel de propuestas $q(\tilde{\theta}|\theta)$, siguiendo el **Algoritmo 1**. Hay una última consideración que debe tenerse en cuenta y es que la posterior sea propia; si se inició con una inicial impropia es posible que la posterior no pueda integrar a 1, lo que haría que el algoritmo fallara (Robert y Casella 2010).

Algoritmo 1: Metropolis Hastings para el aprendizaje bayesiano

```

1 Valor inicial arbitrario o simulado  $\theta^{(0)}$ 
2 para  $n \leftarrow 1$  a  $N$  hacer
3    $\theta \leftarrow \theta^{(n-1)}$ 
4    $\tilde{\theta} \sim q(\tilde{\theta}|\theta)$ 
5    $\alpha(\tilde{\theta}; \theta) \leftarrow \min \left\{ \frac{f(y|\tilde{\theta})f(\tilde{\theta})q(\theta|\tilde{\theta})}{f(y|\theta)f(\theta)q(\tilde{\theta}|\theta)}, 1 \right\}$ 
6    $u \sim U[0, 1]$ 
7   si  $u \leq \alpha(\tilde{\theta}; \theta)$  entonces
8     |  $\theta^{(n)} \leftarrow \tilde{\theta}$ 
9   en otro caso
10    |  $\theta^{(n)} \leftarrow \theta$ 
11 fin
12 fin

```

Random Walk Metropolis

Si la posterior es propia, entonces el mayor problema, claro está, es el de encontrar un kernel de propuestas conveniente. Una alternativa son aquellos que exploran progresivamente el espacio de estados de manera local. Estos se conocen como *Random Walk Metropolis* o RWM (Robert y Casella 2010). De hecho, el algoritmo inicial de Metropolis y col. (1953) era de este tipo, usando una distribución uniforme en una vecindad del estado actual; pero también se pueden usar otras distribuciones, como una normal o una t de Student centrada en dicho estado. Robert (2015) hace la analogía con alguien que para ver una pintura en un cuarto oscuro tiene que ir alumbrando el cuadro con una antorcha, iluminando secuencialmente diferentes segmentos del lienzo.

Estos kérneles de caminata aleatoria resultan ser *simétricos*— esto es, $q(\tilde{\theta}|\theta) = q(\theta|\tilde{\theta})$ — lo que simplifica los cálculos de probabilidades de aceptación:

$$\alpha_{RWM}(\tilde{\theta}; \theta) = \min \left\{ \frac{f(\tilde{\theta}|y)q(\theta|\tilde{\theta})}{f(\theta|y)q(\tilde{\theta}|\theta)}, 1 \right\} = \min \left\{ \frac{f(\tilde{\theta}|y)}{f(\theta|y)}, 1 \right\}$$

Un ejemplo trivial puede ayudar a visualizar cómo funcionan. Supongamos que quisiéramos simular de una posterior que, en realidad, es una normal bivariada cuyas marginales son normales estándar. Un posible kernel de transición sería una distribución uniforme en un rectángulo de 3 de lado centrado en el estado actual de la cadena, por lo que tendríamos un kernel simétrico y un algoritmo de *Random Walk Metropolis*.

Podemos ver cómo se va explorando el espacio de estados en la **Figura 6.2**. Cuando una de las propuestas es rechazada, esta se marca como una tache. Empezando en un punto alejado del origen, la cadena comienza a acercarse a la región de mayor densidad. En el fondo— como espero explicar más claro después— encontrar esta *región crítica* y explorarla lo suficientemente bien para poder calcular los resúmenes inferenciales de interés es el verdadero objetivo de un algoritmo de integración por MCMC (Neal 1993; Betancourt 2017). Con tan solo 2,500 iteraciones la cadena ya la recorrió varias veces y los valores que alejarían a la cadena de la misma son normalmente rechazados. Podemos ver también cómo los promedios ergódicos van acercándose a 0 para ambas variables.

Pero la pregunta permanece, ¿cómo determinar el kernel de propuestas? ¿Por qué usar 3 como lado y no 0.5? ¿Qué hubiera pasado en ese caso? Podemos observarlo en la

Figura 6.3. El rectángulo de propuestas es más pequeño y eso tiene dos efectos: hay menos propuestas rechazadas pero la cadena avanza más lentamente. Vemos cómo aún después de las mismas 2,500 iteraciones la cadena no ha explorado por completo el área de mayor densidad y los promedios ergódicos apenas comienzan a converger.

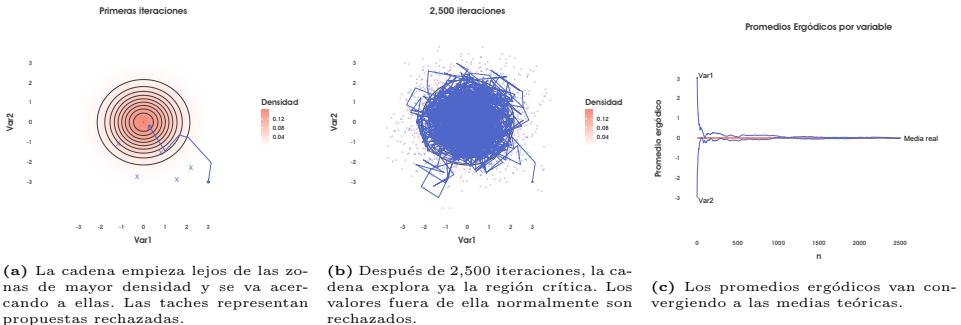


Figura 6.2: Ilustración de una cadena de Metropolis Hastings simulando de una normal bivariada sin correlación mediante una implementación de *Random Walk Metropolis*. Fuente: elaboración propia.

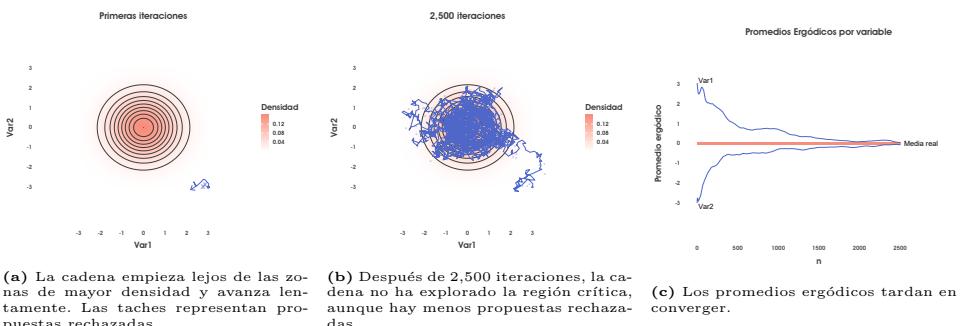


Figura 6.3: Ilustración de una cadena de Metropolis Hastings simulando de una normal bivariada sin correlación mediante una implementación de *Random Walk Metropolis* con un kernel de propuestas estrecho. Fuente: elaboración propia.

El rectángulo más amplio propone valores que se alejan de la región crítica, mismos que tienden a ser rechazados. Sin embargo, también permite proponer valores más distantes del punto actual, lo que en este caso hace que la cadena avance rápidamente. El rectángulo más pequeño avanza más bien “lento pero seguro”... quizás demasiado lento. Ambas cadenas convergerán, pero este ejemplo ilustra uno de los principales problemas de los algoritmos de MH, en general.

Elegir un kernel de propuestas eficiente no es sencillo. Si la escala del kernel es demasiado pequeña, los saltos son demasiado pequeños y la cadena avanza demasiado lento. Si la escala es demasiado grande, corremos el riesgo de que la cadena rechace casi todas las propuestas y quede “atorada” en algún lugar; si la cadena repite el mismo valor varias veces seguidas, también alenta su avance. Debe haber un justo medio entre ambos extremos. Esto podemos verlo comparando el comportamiento de tres kérneles cuando intentan simular normales ahora altamente correlacionadas, en la **Figura 6.4**.

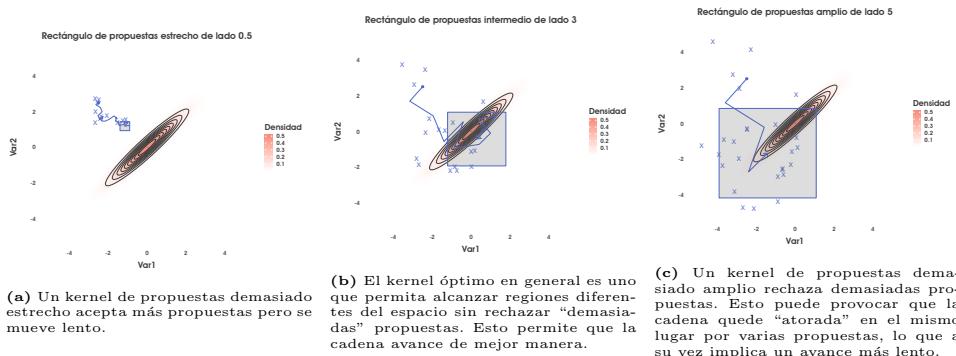


Figura 6.4: Comparación de kérneles uniformes de propuestas para simular una normal bivariada con correlación 0.95 mediante RWM. Las tres cadenas presentan las primeras 30 transiciones, aunque no se distingan las repeticiones de los puntos en los que la cadena permanece en el mismo estado. Fuente: elaboración propia.

Ciertamente hay investigación enfocada a proponer “reglas de dedo” que sean relativamente eficientes (Gelman, Roberts y W.R. 1996; Geyer 2005; Yang y Rodríguez 2013). Para el ejemplo original de normales bivariadas no correlacionadas elegí el rectángulo uniforme de lado 3 porque Yang y Rodríguez (2013) encuentran que este es el ancho óptimo para simular una normal estándar mediante un kernel uniforme. El uso del kernel uniforme se debe a que es fácil de entender y fue el utilizado por los científicos de Los Álamos, pero pueden haber kérneles más eficientes como reporta el mismo artículo.

El problema crece porque la dimensionalidad de los problemas hace que muy rápidamente nos alejemos de los ejemplos triviales y sea mucho más difícil proponer kérneles de transición que funcionen de manera adecuada. ¿Qué hacer? Por varios años esto desalentó el uso de los métodos MCMC, hasta la década de los 90 cuando hubo una “epifanía” en el mundo de la estadística bayesiana (Robert y Casella 2011).

6.2.2. Gibbs Sampler

Alan Gelfand y Adrian Smith presentaron en 1990 un artículo titulado *Sampling-Based Approaches to Calculating Marginal Densities* (Gelfand y Smith 1990). En él rescataron tres algoritmos de simulación para obtener densidades que no pueden ser conocidas de manera analítica. Sin embargo, el que provocó la explosión de los métodos bayesianos fue el llamado *Gibbs Sampler*. Este algoritmo de MCMC fue propuesto por los hermanos Stuart y Donald Geman mientras trabajaban en problemas de reconstrucción de imágenes (Geman y Geman 1984), pero ya estaba oculto en el artículo de Hastings (1970) cuando sugería que para simular de una distribución multidimensional se podría cambiar de manera secuencial una única coordenada o, incluso, subgrupos de coordenadas.

En otras palabras, “divide y vencerás”; si el problema es que la distribución objetivo es de alta dimensionalidad y compleja, ¿por qué no intentar simularla por partes más simples en lugar de querer simularla directamente? La gran contribución de los hermanos Geman fue encontrar una manera simple y poderosa de llevarlo a cabo. La chispa de Gelfand y Smith fue mostrarle al mundo estadístico que—sin ser el método óptimo para cualquier problema—podía ser bastante eficiente en un amplio rango de problemas comúnmente encontrados en la práctica y, mejor aún, era efectivamente simple y universal (Gelfand y col. 1990).

Supongamos que el vector que queremos simular es bivariado, digamos $Z = (X, Y)$. El algoritmo consiste en alternar simulaciones simples de cada variable condicional en la otra, conservando los vectores bivariados de cada ciclo de dos *pasos de Gibbs*. Primero simulamos de $X|Y$, luego de $Y|X$; volvemos a simular otra X dada la Y anterior, luego otra Y dada la nueva X y así sucesivamente. Una analogía imperfecta pero que me funciona como recurso mnemotécnico para recordar el *Gibbs Sampler* es imaginar que mi objetivo es cruzar de algún punto en una acera a otro en la acera contraria. En lugar de cruzar la calle de manera diagonal en un solo paso, cruzamos mediante dos. El primero es caminar por la misma banqueta hasta alguna esquina o lugar indicado y el segundo es cruzar la calle sobre el paso peatonal. Estos dos pasos constituirían una *transición de Gibbs* y repitiendo transiciones de este tipo podríamos caminar de una dirección alejada a algún punto de interés en la ciudad.

De manera más general, el *Gibbs Sampler* construye una cadena de Markov con base en las llamadas distribuciones *condicionales completas*. Estas no son otra cosa más que las distribuciones condicionales de una o más variables dentro de un vector, dado el resto. Supongamos que tenemos un vector de parámetros θ de dimensión d expresado mediante una partición $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ donde $\theta_j \in \mathbb{R}^{d_j}$ y $\sum_{j=1}^k d_j = d$. Para cada subvector θ_j , definimos la correspondiente condicional completa como $f(\theta_j | \theta_{-j}, y)$, donde θ_{-j} significa todos los componentes del vector menos el j -ésimo. Como estamos en un contexto de aprendizaje bayesiano, las distribuciones son posteriores; es decir, también condicionamos en los datos observados y .

Una transición de Gibbs se construye, al igual que en el caso bivariado y la analogía, mediante pasos de Gibbs intermedios, como puede verse en el **Algoritmo 2**. Empezando en un vector $\theta^{(0)}$ en el espacio de estados multidimensional, caminamos sobre la banqueta, simulando un valor $\theta_1^{(1)}$ dejando el resto de las variables fijas. Sustituimos este valor en el vector y , ahora, simulamos un valor $\theta_2^{(1)}$ de la correspondiente condicional completa. Así, cada subvector $\theta_j^{(1)}$ se simula con base en los nuevos valores de subvectores que van primero en la secuencia pero con los valores de la transición anterior para los subvectores que sigan en la secuencia. El último paso de la transición será entonces simular $\theta_k^{(1)}$ dados todos los nuevos valores simulados.

Algoritmo 2: Gibbs Sampler para el aprendizaje bayesiano

```

1 Valor inicial arbitrario o simulado  $\theta^{(0)} \leftarrow (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)})$ 
2 para  $n \leftarrow 1$  a  $N$  hacer
3    $\theta_1^{(n)} \sim f(\theta_1 | \theta_{-1}^{(n-1)}, y)$ 
4   para  $j \leftarrow 2$  a  $k - 1$  hacer
5      $\theta_j^{(n)} \sim f(\theta_j | \theta_1^{(n)}, \dots, \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, \dots, \theta_k^{(n-1)}, y)$ 
6   fin
7    $\theta_k^{(n)} \sim f(\theta_k | \theta_{-k}^{(n)}, y)$ 
8    $\theta^{(n)} \leftarrow (\theta_1^{(n)}, \theta_2^{(n)}, \dots, \theta_k^{(n)})$ 
9 fin

```

El algoritmo podría parecer a primera vista poco práctico porque, si no se conoce la distribución conjunta, ¿cómo conoceríamos las k condicionales completas? Sin embargo, como mostraron Gelfand y Smith (1990) o Gelfand y col. (1990), es factible aplicarlo a

modelos tan variados como multinomiales, normales multivariadas y modelos jerárquicos como regresiones de interceptos y coeficientes variables. La clave está en que las condicionales completas son proporcionales a la posterior pero simplificadas, pues al condicionar solo conservamos los términos que incluyen al respectivo subvector θ_j . Esto permite generalmente identificar una familia de distribuciones conocida o encontrar un algoritmo relativamente sencillo para simular de dicha distribución no normalizada.

Quizás la familia de modelos que mejor ejemplifican la utilidad del *Gibbs Sampler* son los modelos jerárquicos. La propia estructura jerárquica implica de manera natural que las condicionales completas se simplifiquen puesto que cada grupo de parámetros es condicionalmente independiente de los otros, dados los hiperparámetros correspondientes. Para verlo podemos seguir el ejemplo de Gutiérrez Peña (2016), suponiendo el siguiente modelo jerárquico simple de 3 niveles. Los datos y provienen de m subpoblaciones con el respectivo vector de parámetros ω que, a su vez, depende de un vector de hiperparámetros ϕ con su respectiva hiperinicial:

$$\begin{aligned} f(y|\omega) &= \prod_{i=1}^m f(y_i|\omega_i) \\ f(\omega|\phi) &= \prod_{i=1}^m f(\omega_i|\phi) \\ f(\phi) \end{aligned}$$

Tenemos en total $m+1$ parámetros $\theta = (\omega, \phi) = (\omega_1, \omega_2, \dots, \omega_m, \phi)$ cuya distribución posterior — $f(\theta = (\omega, \phi)|y) \propto f(\phi) \prod_{i=1}^m f(y_i|\omega_i) f(\omega_i|\phi)$ — presenta una factorización que simplifica las condicionales completas:

$$\begin{aligned} f(\theta_1|\theta_{-1}, y) &\propto f(\omega_1|\phi, y_1) \\ &\vdots \\ f(\theta_m|\theta_{-m}, y) &\propto f(\omega_m|\phi, y_m) \\ f(\theta_{m+1}|\theta_{-(m+1)}, y) &\propto f(\omega|\phi) f(\phi) \end{aligned}$$

La consagración del *Gibbs Sampler* como la principal herramienta de simulación y aprendizaje práctico bayesiano se dio cuando en la primera mitad de la década de los 90 fue presentado el software BUGS, que es un acrónimo para *Bayesian inference Using*

Gibbs Sampling (Betancourt 2018). Ya era posible automatizar la aplicación del algoritmo para una clase amplia de modelos que el usuario define mediante distribuciones iniciales, verosimilitudes y datos. Como bien dicen Casella y George (1992), al liberar a los estadísticos de tener que tratar con cálculos complicados, la atención principal se puede dedicar a los aspectos estadísticos de los problemas.

Debo decir que hasta ahora no he mencionado ninguna justificación teórica de convergencia para el algoritmo, pero el lector más inquieto puede consultar Geyer (2005), en donde se muestra que, aunque normalmente es considerado un algoritmo distinto por derecho propio y desarrollo histórico, el *Gibbs Sampler* es una implementación particular de Metropolis Hastings en la que la probabilidad de aceptación es siempre 1. Otra manera de justificar el algoritmo se puede consultar en el artículo de Casella y George (1992), donde además de discutir la convergencia también se da una explicación más detallada de por qué funciona.

Al igual que en el caso de *Random Walk Metropolis*, podemos observar el comportamiento del *Gibbs Sampler* para el ejemplo trivial de las normales independientes. El algoritmo funciona de manera mucho más eficiente que la implementación de *Random Walk Metropolis*—comparar **Figuras 6.5 y 6.2**. Esto no debería sorprendernos tanto, puesto que la independencia de normales implica que las condicionales completas son en realidad las verdaderas marginales.

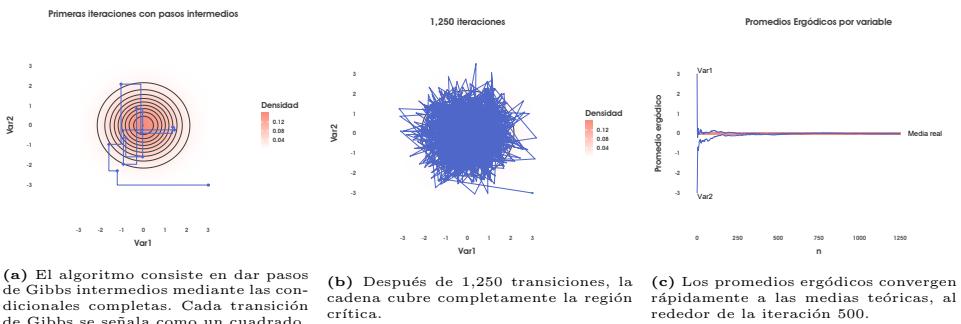


Figura 6.5: Ilustración de una cadena de *Gibbs Sampling* para una normal bivariada sin correlación. Fuente: elaboración propia.

En realidad, pocas veces contamos con tanta suerte como para tener completa independencia. Normalmente los parámetros que queremos simular están correlacionados

entre sí. En la **Figura 6.6** podemos ver cómo el **Gibbs Sampler** también se desempeña mucho mejor en este caso que RWM.

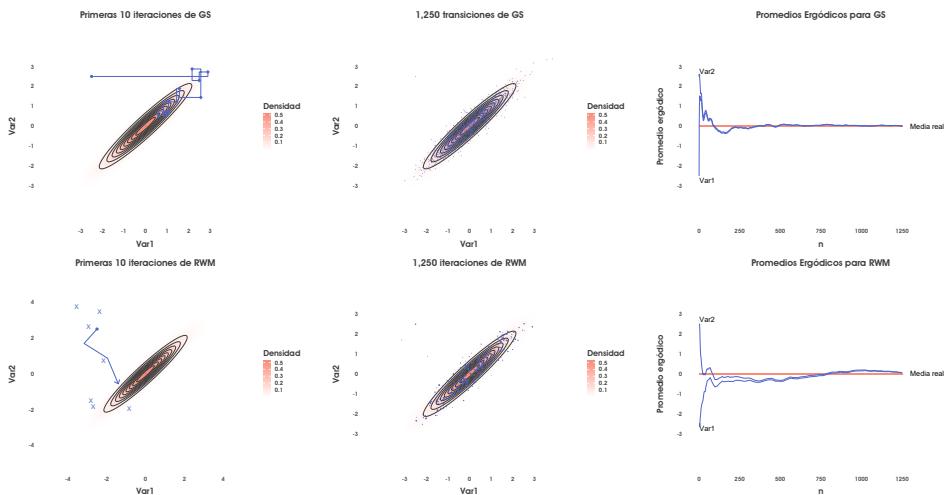


Figura 6.6: Comparación del *Gibbs Sampler* (GS) con *Random Walk Metropolis* (RWM) para simular de una normal bivariada de correlación 0.95. Las condicionales completas le permiten al GS llegar a la región crítica más rápidamente que al kernel uniforme de RWM que se retrasa en cada propuesta rechazada. Si comparamos todos los valores simulados después de 1,250 iteraciones vemos cómo aquellos de GS cubren de mejor manera la región crítica que los de RWM, mismos que se disturbaron ligeramente para presentar también los valores que se repiten por los rechazos. Asimismo, los promedios ergódicos convergen más rápidamente para GS que para RWM. Fuente: elaboración propia.

Por otro lado, si comparamos entre los dos casos del *Gibbs sampler*, vemos que el desempeño se vio afectado al agregar correlación. Este es un fenómeno que se debe tener siempre presente: cuando la parametrización del modelo genera alta correlación entre los componentes, el algoritmo se hace más lento. El motivo es un equivalente al caso de un kernel de transiciones más estrecho. La correlación de los parámetros implica que las condicionales completas están muy concentradas, determinadas fuertemente por el valor actual; es decir, la simulación no puede generar un valor muy alejado del actual y el avance de la cadena es más lento.

Por ello, el *Gibbs Sampler* puede sucumbir en la práctica ante problemas no triviales de alta correlación. Sabemos que convergerá, pero puede no hacerlo en un tiempo razonable, llegando a pasar cientos de miles de iteraciones antes de converger. Se pueden buscar parametrizaciones e implementaciones de Gibbs que mejoren el desempeño, pero como dice el Stan Development Team (2017),

...incluso una implementación eficiente y escalable no resuelve el problema subyacente de que *Gibbs sampling* no se desempeña bien con posteriores altamente correlacionadas. Finalmente nos dimos cuenta que necesitábamos un mejor muestrador, no una mejor implementación.

Ese mejor muestrador se materealizó en el software *Stan*— nombrado así en honor al propio Stan Ulam— cuya primera versión estable fue liberada en 2013 y que introduciré más adelante. Pero antes de hacerlo, quisiera detenerme en un tema que he omitido por ahora.

6.2.3. Convergencia

La convergencia de los promedios ergódicos que implica (6.2b) es la clave de un análisis de MCMC, pues nos permite aproximar los resúmenes inferenciales que necesitamos, tanto como nuestros recursos computacionales lo permitan. Sin embargo, como ya he mencionado, en un análisis bayesiano la única receta indica tener una muestra aleatoria de la distribución posterior. Si observamos con cuidado (6.2a), vemos que la última simulación de la cadena es la que podemos considerar como proveniente de la distribución objetivo límite. Si quisiéramos una muestra de tamaño N , podríamos inicialmente pensar que necesitaríamos correr N cadenas independientes.

Afortunadamente, en las cadenas de MCMC en general, la distribución límite es también lo que se conoce como *distribución estacionaria* de la cadena de Markov. Esto quiere decir que el kernel de transición de la cadena mantiene estable la distribución de las simulaciones, una vez que se alcanza la distribución estacionaria (Neal 1993). Esta característica estacionaria de la distribución límite implica que si reiniciamos la cadena una vez que se llega a la convergencia, podemos después de algún número de transiciones adicionales contar con otra observación prácticamente independiente proveniente de la misma distribución objetivo.

Por consiguiente, aunque los algoritmos de MCMC producen cadenas correlacionadas, es posible calcular un *tamaño efectivo de muestra* que aproxima el tamaño de una muestra aleatoria auténticamente independiente proveniente de la distribución objetivo cuyas estimaciones equivaldrían a las que hacemos con la muestra simulada. Los detalles de cómo se calcula dicha estadística puede consultarse en Gelman y col. (2013).

Por lo anterior, usualmente solo se conservan las simulaciones cada m transiciones de manera que la correlación entre ellas sea lo más cercana a cero posible. Buscaríamos descartar aquellas simulaciones que, debido a la correlación, no están aportando realmente mayor información sobre la distribución objetivo. Con este *adelgazamiento* de la cadena se busca ahorrar espacio de memoria en la computadora y conservar solo las simulaciones que aportan *nueva* información y aumentan el tamaño efectivo de muestra.

En este sentido, la forma usual de aplicar la única receta de la inferencia bayesiana mediante MCMC se resume de la manera siguiente:

1. Elegir un número c de cadenas para simular.
2. Iniciar cada cadena en un punto distinto y disperso del espacio paramétral.
3. Correr de manera independiente las c cadenas hasta que “alcancen la convergencia”.
4. Una vez que se considera que las cadenas convergieron, se desechan las transiciones iniciales que constituyen el *periodo de calentamiento*.
5. Se *adelgaza* cada cadena conservando las simulaciones solo cada m transiciones de manera que la correlación sea baja.
6. Se continúan realizando simulaciones después del calentamiento y el adelgazamiento hasta obtener una muestra “lo suficientemente grande”.

El número de cadenas se puede elegir en función del número de procesos paralelos que la computadora puede realizar para aprovechar la eficiencia que implica el cómputo paralelo. Usualmente se corren 3, 4 o 5 cadenas. Por su parte, seleccionar el espaciamiento m de las cadenas normalmente implica realizar gráficos de autocorrelación para algunas pruebas preliminares cuyo objetivo es también determinar la convergencia del algoritmo. En realidad, el punto más delicado es precisamente este último, declarar la convergencia.

Desafortunadamente, los teoremas ergódicos son asintóticos, por lo que solo se garantiza la convergencia en el infinito. Aunque hay cotas para los errores, normalmente no son muy útiles, salvo algunos casos (Smith y Roberts 1993). Por ello, en la práctica se han desarrollado diferentes técnicas de diagnóstico de convergencia. Sin embargo, ninguna puede garantizarnos totalmente que la cadena ya haya convergido. Entonces,

es recomendable considerar diferentes técnicas para disminuir la probabilidad de pasar por alto algún problema de convergencia. Una vez que estamos satisfechos con las pruebas de convergencia, podemos utilizar todas las simulaciones como una sola muestra que caracteriza la distribución posterior y realizar las estimaciones de Monte Carlo necesarias.

La primera técnica ya la he utilizado en los ejemplos de RWM y Gibbs: graficar los promedios ergódicos para diferentes variables o resúmenes inferenciales para observar en qué punto se estabilizan. La segunda es precisamente iniciar en puntos dispersos del espacio paramétral para evitar que las cadenas exploren solo una parte del espacio y, por ejemplo, queden atrapadas en una sola moda cuando la distribución posterior pueda ser multimodal. Al mismo tiempo, tener diferentes cadenas nos permite comparar las estimaciones dentro de cada cadena así como a través de las cadenas. Un ejemplo de promedios ergódicos para 4 cadenas de *Gibbs Sampler* (GS) iniciadas en puntos dispersos del espacio paramétral puede verse en la **Figura 6.7**.

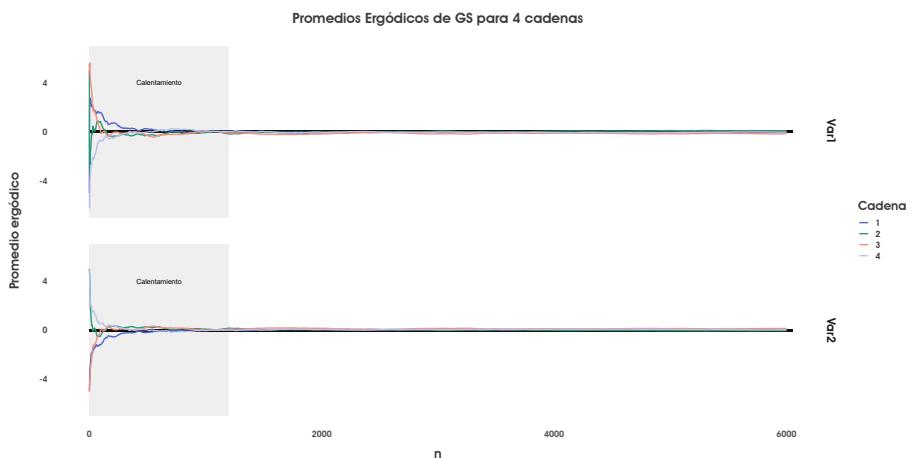


Figura 6.7: Ilustración de la convergencia de dos promedios ergódicos para 4 cadenas de *Gibbs Sampling* iniciadas en puntos dispersos del espacio paramétral. Vemos cómo las estimaciones de distintas cadenas se estabilizan y mezclan. Esto ayuda a detectar la fase de calentamiento. Fuente: elaboración propia.

En el fondo, hay dos características que se buscan para declarar la convergencia (Gelman y col. 2013). Queremos que las cadenas lleguen a la estacionariedad, pues la distribución límite es también estacionaria. Las estimaciones deben estabilizarse. También, y dicho de manera coloquial, buscamos que las cadenas *mezclen* bien. Es decir, que las estimaciones para cada cadena sean parecidas a las estimaciones de todas las

cadenas en su conjunto. Por ello un buen algoritmo de MCMC buscará recorrer y explorar todas las regiones críticas de la distribución objetivo (Neal 1993; Betancourt 2018).

Uno de los diagnósticos visuales más utilizados son los llamados gráficos de oruga o *trace plots*. En ellos se grafica la secuencia de iteraciones, es decir en el eje x se encuentra el número de iteración y en el eje y el valor de la cadena en dicha iteración. Si las cadenas son estacionarias entonces el gráfico saltará de un punto a otro dentro de un rango de valores que determinan la región crítica. Si las cadenas mezclan bien, entonces las diferentes secuencias estarán oscilando en la misma región crítica, intercalándose. El *trace plot* del ejemplo de esta sección puede verse en la **Figura 6.8**. También podemos ver en la **Figura 6.9**, el gráfico correspondiente a las iteraciones válidas para cada cadena, es decir, después del descarte del periodo de calentamiento y el espaciamiento para evitar la correlación.

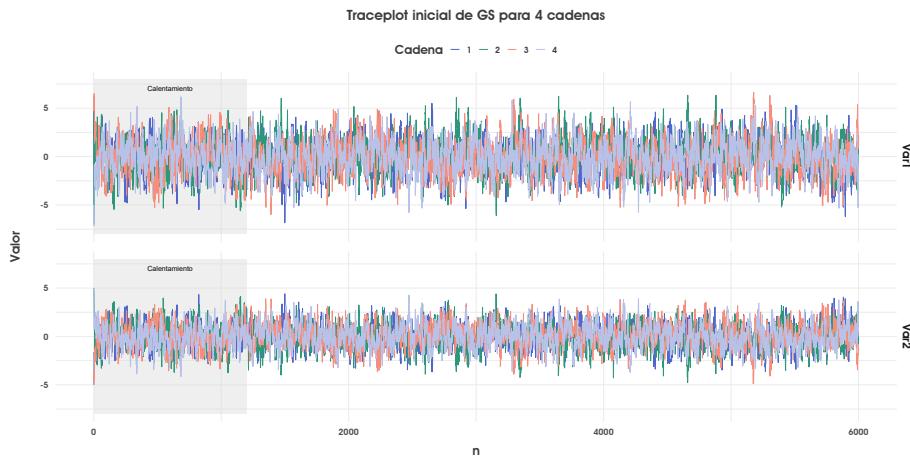


Figura 6.8: Ilustración de un *trace plot* para las 4 cadenas de *Gibbs Sampling*. Vemos el comportamiento de oruga, al rededor de una banda de valores. Fuente: elaboración propia.

En este caso, el ejemplo es lo suficientemente sencillo para permitirnos un adelgazamiento de 8. La reducción en la autocorrelación que esto tiene se puede apreciar en la **Figura 6.10**. En la práctica el adelgazamiento suele ser más chico.

Otro diagnóstico gráfico es el de comparar histogramas o densidades estimadas para varios parámetros y resúmenes inferenciales y verificar que estos son estables— com-

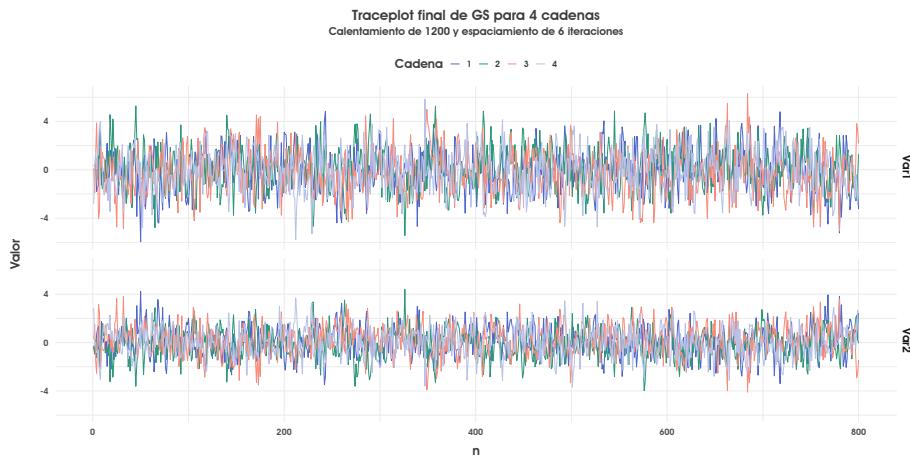


Figura 6.9: Ilustración de un *trace plot* para las 4 cadenas de *Gibbs Sampling* considerando solo las iteraciones válidas. Fuente: elaboración propia.

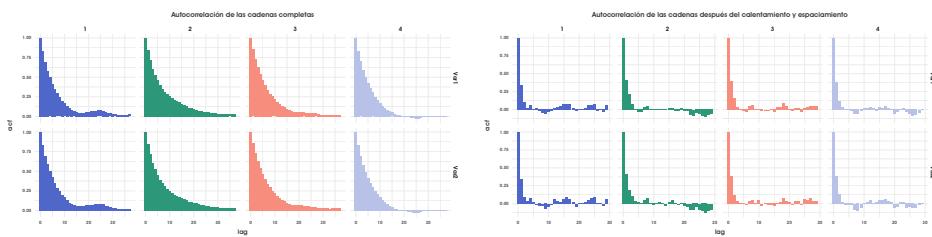


Figura 6.10: Gráficos de autocorrelación para las 4 cadenas de *Gibbs Sampling* antes y después de realizar un adelgazamiento de 8. Fuente: elaboración propia.

parándolos para la primera mitad de la muestra y para la segunda, por ejemplo— y mezclan bien— los gráficos para las diferentes cadenas se parecen al gráfico de todas juntas—. Para el ejemplo, vemos la comparación de los histogramas de cada cadena con el histograma de todas las cadenas en su conjunto en la **Figura 6.11**.

A pesar de su utilidad, siempre hay que tener presente que los diagnósticos gráficos pueden ser engañosos y no bastan por si solos. Si la distribución es multimodal, por ejemplo, los gráficos de oruga pueden no mezclarse porque cada cadena explora una moda distinta.

Por otro lado, existen también resúmenes estadísticos para evaluar convergencia. Una estadística frecuentemente utilizada y que es normalmente reportada en softwares de

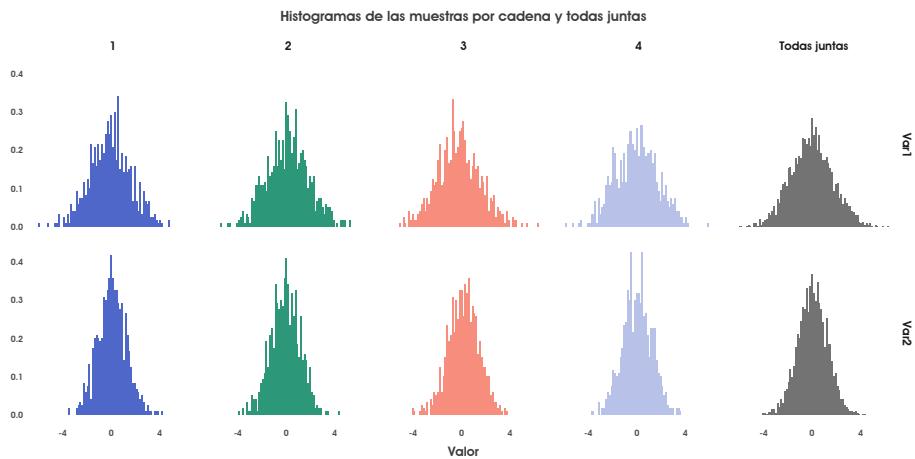


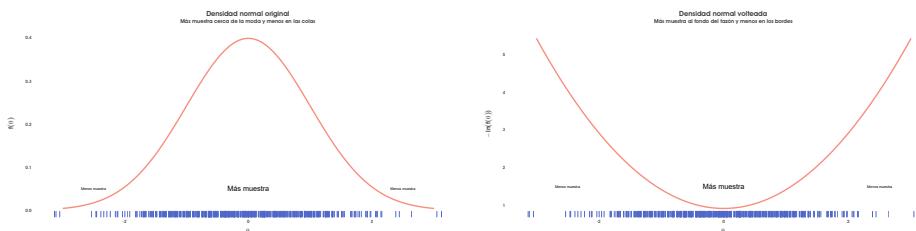
Figura 6.11: Comparación de histogramas para las 4 cadenas de *Gibbs Sampling* considerando solo las iteraciones válidas. Fuente: elaboración propia.

cómputo mediante MCMC es el *factor de reducción de escala* \hat{R} propuesto por Gelman y Rubin (1992). Este está basado en comparar las varianzas de los parámetros o resúmenes inferenciales a través de las cadenas con aquellos dentro de cada cadena. La \hat{R} , bajo supuestos de normalidad, tiende a 1 conforme la convergencia aumenta. Por ello, un valor empírico cercano a 1, es normalmente requerido antes de poder declarar la convergencia.

Finalmente, querría comentar que hay 3 factores principales que permiten tener un buen muestreador de MCMC (Neal 1993). El primero es la cantidad de cómputo requerida para simular cada transición; esta, por ejemplo, podría ser una ventaja de *Gibbs Sampling* sobre *Random Walk Metropolis* pues al aceptar todas las propuestas nos ahorraremos la necesidad de realizar la corrección de Metropolis. El segundo factor es el tiempo que le lleva a la cadena alcanzar la convergencia; esto indica de manera general la cantidad de cómputo invertido en el periodo de calentamiento que se descartará. El tercer y último factor está relacionado con el segundo pero es ligeramente diferente: las transiciones necesarias para movernos de un estado en la distribución objetivo a otro prácticamente independiente. Esto nos indicará el tamaño de la simulación requerido para realizar una estimación con alguna precisión deseada. Es normalmente este tercer factor el que motiva el uso de un muestreador distinto a RWM o a *Gibbs Sampler* y que introduzco a continuación.

6.3. Hamiltonian Monte Carlo

Una distribución posterior puede ser vista como una superficie de cimas y valles. Por ejemplo, la distribución normal— identificada por su típica forma de campana— sería una cima central. Imaginemos que queremos tener una muestra aleatoria de esta distribución. Esperaríamos tener más muestra en la región central— al rededor de la moda— que en las colas de la distribución, como puede verse en la **Figura 6.12a**. Si volteamos la función de densidad— tomando el menos logaritmo, por ejemplo— esta se convierte en una especie de tazón. Las regiones cercanas a la moda ahora son zonas bajas y las colas son zonas altas. La muestra se concentra en el fondo del tazón y no en los bordes, como puede verse en la **Figura 6.12b**.



(a) La densidad original puede verse como una superficie con una cumbre al rededor de la cual se concentra la mayoría de las observaciones provenientes de una muestra aleatoria.

(b) Si volteamos y suavizamos la densidad aplicando el menos logaritmo, vemos que ahora la gráfica parece un tazón. La muestra se concentra en el fondo del tazón, en lugar de en los bordes del mismo.

Figura 6.12: Una muestra aleatoria proveniente de una distribución tiende a concentrarse al rededor de las modas. En los gráficos, las observaciones de una muestra aleatoria particular proveniente de la distribución normal se muestran mediante un *rug plot* como marcas azules en el eje horizontal. Fuente: elaboración propia con base en la explicación de Lambert (2018).

Imaginemos una partícula en algún punto de la superficie del tazón. Si le aplicáramos una cantidad de movimiento o *momentum*, esta rodaría hacia el fondo del mismo, siguiendo las leyes de la física, como ilustro en la **Figura 6.13a**. Si asumimos que no existe fricción, la partícula continuaría subiendo y bajando, intercambiando energía cinética por energía potencial y viceversa (**Figura 6.13b**).

Si tomamos otra partícula con otro *momentum*, esta tendría una trayectoria distinta pero que también recorrería la superficie del tazón. Bajo esta dinámica, podemos conjutar que una partícula tendería a estar más cerca del fondo del tazón que de los bordes, justo como una observación muestreada de la distribución que diera origen al tazón.

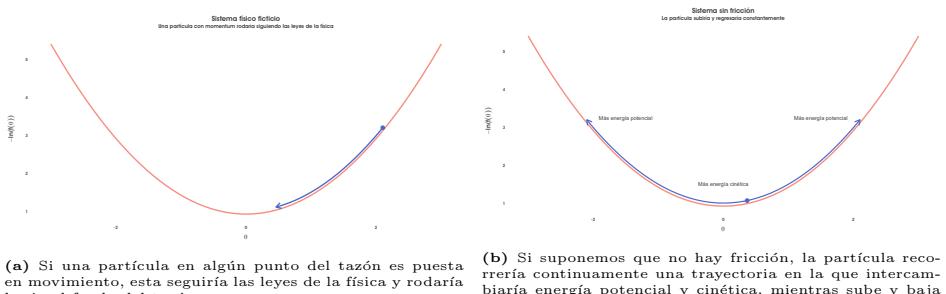


Figura 6.13: Para una densidad, podemos construir un sistema físico ficticio en el que partículas imaginarias recorren trayectorias sobre la superficie de la densidad volteada. Fuente: elaboración propia con base en la explicación de Lambert (2018).

Lo maravilloso de esta idea es que puede ser utilizada para obtener realizaciones provenientes de una distribución objetivo y realizar estimaciones de Monte Carlo. Es decir, podemos simular este sistema físico imaginario para construir una cadena de Markov y obtener una muestra de MCMC que no solo cumpla la intuición de tener más observaciones cerca de las modas que en las colas, sino que lo haga en la proporción correcta que marca la ley de probabilidad objetivo. Más aún, no requerimos conocerla *por completo*, pues “una pelota no conoce la superficie sobre la que rueda, sin embargo su camino está gobernado por ella” (McElreath 2017). Nos basta conocer la densidad objetivo salvo por una constante de normalización.

De manera informal, lo único que debemos hacer es imaginar una partícula y “empujarla” en alguna dirección. Luego seguimos por un tiempo la trayectoria que las leyes de la física determinarían sobre la superficie de nuestra densidad volteada y la “detenemos”. En este punto registramos su posición en el eje horizontal como una realización de nuestra variable aleatoria. Si volvemos a empujarla en otra dirección aleatoria y la seguimos por otro periodo de tiempo llegaríamos a una nueva posición que sería nuestra siguiente observación. Repitiendo este procedimiento obtendríamos una secuencia de observaciones que constituirían nuestra cadena de Markov.

Este método de MCMC se conoce como *Hamiltonian Monte Carlo* o HMC y procederé a exponerlo a continuación. Pero para entenderlo mejor, primero hay que introducir un poco de terminología del campo de la física estadística, para lo que seguiré la explicación de Neal (1993).

6.3.1. Sistemas físicos

Imaginemos que tuviéramos una descripción microscópica completa de un sistema físico. Por ejemplo, que conociéramos la posición y velocidad de todas las partículas de un sistema en un momento dado en el tiempo. Esta descripción se conoce como el *microestado* del sistema. Evidentemente, en la realidad no podemos observar dicho microestado; más bien observamos el *macroestado* del sistema, que incluye la temperatura o el volumen. Por ello, debemos asignar una medida de probabilidad sobre los microestados que podrían llevar al macroestado que observamos.

Por ejemplo, supongamos que tenemos un sistema que mantiene una temperatura constante T pero en el que cada microestado s tiene una cierta energía $H(s)$.¹ Normalmente se supone que se satisface la siguiente *distribución canónica*, también conocida como de Gibbs o de Boltzmann:

$$f(s) = \frac{\exp\{-H(s)/T\}}{Z}, \quad (6.6)$$

donde Z es la constante de normalización y $H(s)$ es la función de energía del sistema y que depende del microestado en el que este se encuentre.

Esta distribución canónica es particularmente útil porque *cualquier* distribución de probabilidad que no valga cero en ningún punto puede ser expresada de esta manera. Para ver por qué, consideremos que tenemos un vector de variables aleatorias s , cuya distribución de probabilidad es $f(s)$. Esta puede considerarse una distribución canónica como (6.6) si consideramos que s representa los microestados de un sistema físico ficticio con temperatura T y constante de normalización Z cuya función de energía esté dada por $H(s) = -T \ln [Z f(s)]$. Por construcción, podríamos decir que la temperatura constante es $T = 1$ y eliminarla de las ecuaciones.

Esto quiere decir que podemos interpretar a una variable aleatoria como microestado de un sistema físico imaginario con distribución canónica para alguna función de energía conveniente. En particular, regresemos al caso en que queremos muestrear de la distribución posterior $f(\theta|y)$ de nuestro vector de parámetros de interés $\theta = (\theta_1, \dots, \theta_d)$.

¹En física, y en la referencia de Neal (1993), se utiliza la notación $E(s)$, pero para evitar confusiones con el valor esperado de s y para facilitar la explicación de Hamiltonian Monte Carlo, utilizo la letra H . Asimismo, utilizaré otras notaciones no comunes en física, con la intención de ser más adecuadas al contexto estadístico.

Diremos que los parámetros representan las coordenadas de la *posición* de una partícula en un sistema físico y la expresamos como una distribución canónica con *energía potencial* $U(\theta|y) = -\ln [f(y|\theta)f(\theta)]$:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} = \frac{\exp \{-U(\theta|y)\}}{Z_y},$$

con $Z_y = f(y)$. Esta energía potencial es, precisamente, la forma de “voltrear” la densidad que mencionaba al inicio de la sección.

Por construcción, introduzcamos un vector adicional $m = (m_1, \dots, m_d)$, con una variable m_i para cada uno de los d parámetros θ_i , que represente la cantidad de movimiento o *momentum* de esta partícula ficticia. Este vector también tiene una distribución de probabilidad $f(m)$ que usualmente es independiente de θ y de los datos. En este caso, tendríamos una *energía cinética* $K(m)$ para la representación en distribución canónica:

$$f(m) = \frac{\exp \{-K(m)\}}{Z_m}$$

Usualmente, por conveniencia y eficacia esta distribución será una normal centrada en cero pero cabe aclarar que $f(m)$ no *necesita* ni ser independiente, ni ser gaussiana (Bertancourt 2017).

Así pues, tenemos un espacio ampliado $2d$ -dimensional y una distribución conjunta $f(\theta, m|y) = f(\theta|y)f(m)$. Dicho espacio para la posición y el *momentum* de una partícula se conoce en física como *espacio de fases*. De nueva cuenta, podemos construir la distribución canónica en este espacio; ahora, la función de energía total se llama el *hamiltoniano* del sistema y es igual a la suma de las energías potencial y cinética, $H(\theta, m|y) = U(\theta|y) + K(m)$. En efecto, ignorando las constantes de normalización,

$$\begin{aligned} f(\theta, m|y) &= f(\theta|y)f(m) \\ &\propto \exp \{-U(\theta|y)\} \exp \{-K(m)\} \\ &\propto \exp \{-[U(\theta|y) + K(m)]\} \\ &\propto \exp \{-H(\theta, m|y)\}. \end{aligned}$$

Notemos que si muestreamos de esta distribución canónica conjunta en el espacio de fases, automáticamente estaríamos muestreando de la distribución de nuestros paráme-

tros θ , pues solo debemos marginalizar ignorando los valores del vector auxiliar m . Para ello aprovecharemos que este sistema físico ficticio satisfaría las leyes de la mecánica hamiltoniana— que es solo una reformulación de la mecánica clásica— y utilizaremos una estrategia de dos pasos.

En ausencia de fricción o injerencia externa, tendríamos que la energía total de una partícula— i.e. el hamiltoniano que definimos como la suma de las dos energías— se mantiene constante. Para ello, una partícula en movimiento debe ir convirtiendo energía potencial en energía cinética y viceversa. Recordemos la concepción de una densidad como una superficie de cimas y valles con una partícula deslizándose sobre ella, como ilustraba en la **Figura 6.13b**. La energía potencial de la partícula en cada instante es proporcional a la altura de la superficie en la posición en que se encuentre. Conforme se desliza hacia abajo, la energía potencial se va convirtiendo en energía cinética hasta llegar al fondo del tazón y empezar a deslizarse hacia arriba. Ahora la energía cinética es la que se va convirtiendo en energía potencial hasta llegar a un punto donde la partícula se detiene momentáneamente y comienza a deslizarse de regreso, repitiendo el proceso.

La trayectoria que dicha partícula seguiría en el tiempo τ cumpliría las *ecuaciones de Hamilton* (Neal 1993, 2011; Betancourt 2017):

$$\frac{d\theta}{d\tau} = +\frac{\partial H}{\partial m} = \frac{\partial K}{\partial m} \quad \text{y} \quad \frac{dm}{d\tau} = -\frac{\partial H}{\partial \theta} = -\frac{\partial U}{\partial \theta} \quad (6.7)$$

Estas ecuaciones diferenciales nos dicen cómo cambiarían θ y m conforme avanza el tiempo en una simulación dinámica del sistema físico que construimos. Asimismo, podemos verificar que la energía se mantiene constante, viendo que el cambio del hamiltoniano en el tiempo es cero, sustituyendo (6.7):

$$\frac{dH}{d\tau} = \sum_{i=1}^d \frac{\partial H}{\partial \theta_i} \frac{d\theta_i}{d\tau} + \frac{\partial H}{\partial m_i} \frac{dm_i}{d\tau} = \sum_{i=1}^d \frac{\partial H}{\partial \theta_i} \frac{\partial H}{\partial m_i} - \frac{\partial H}{\partial m_i} \frac{\partial H}{\partial \theta_i} = 0$$

6.3.2. HMC ideal y en la práctica

La estrategia para muestrear de la distribución canónica en el espacio de fases es ir alternando pasos. Muestreamos primero de la distribución de m , dada una θ inicial; en el caso de independencia esta es simplemente la marginal de m . Luego, seguimos la trayectoria determinística que indican las ecuaciones de Hamilton.

Para ilustrarlo, recordemos que en el ejemplo con el que empezaba la sección, el parámetro θ se distribuía como una normal estándar. Introduzcamos una variable independiente de *momentum* m también distribuida $N(0, 1)$. En este caso, ambas energías son el menos logaritmo del kernel de una normal estándar, por lo que tendríamos que $U(\theta) = \frac{\theta^2}{2}$ y $K(m) = \frac{m^2}{2}$. Las ecuaciones de Hamilton (6.7) resultan entonces ser

$$\frac{d\theta}{d\tau} = \frac{dK}{dm} = m \quad \text{y} \quad \frac{dm}{d\tau} = -\frac{dU}{d\theta} = -\theta,$$

cuyas soluciones tienen una forma analítica conveniente, para algunas constantes r y a (Neal 2011):

$$\theta(\tau) = r \cos(a + \tau) \quad \text{y} \quad m(\tau) = -r \sin(a + \tau) \quad (6.8)$$

La partícula de la **Figura 6.13** en el espacio de fases se vería como en la **Figura 6.14a**. La distribución canónica es una normal bivariada cuyas curvas de nivel determinan órbitas con energía/hamiltoniano constante a lo largo de las cuales se desplazarían las partículas ficticias sobre las trayectorias definidas por las ecuaciones de Hamilton.

Si iniciamos con un valor de $\theta^{(0)}$, muestreamos un valor $m^{(0)} \sim f(m)$. Esto coloca a la partícula sobre una órbita particular con valor constante del hamiltoniano $H(\theta^{(0)}, m^{(0)})$. Ahora seguimos por un tiempo τ la trayectoria definida por (6.8). Esto nos lleva a una nueva coordenada $(\theta^{(1)}, \tilde{m}^{(0)})$, desde la que registramos la posición $\theta^{(1)}$ como nuestra propuesta de observación muestreada. En este punto repetimos el procedimiento: generamos un nuevo valor $m^{(1)} \sim f(m)$, saltamos a una nueva órbita con hamiltoniano $H(\theta^{(1)}, m^{(1)})$, seguimos la trayectoria por otro periodo τ para obtener $\theta^{(2)}$. En la **Figura 6.14b** se muestran 10 iteraciones de este procedimiento con los valores muestreados en los márgenes del gráfico como *rug plot*.

Esta estrategia de dos pasos funcionaría como un algoritmo de *Gibbs dentro de Metropolis-Hastings*. El primer paso—muestrear un nuevo valor de *momentum*—constituye un paso de Gibbs pues, en general, estaríamos muestreando de la condicional de m dada la θ anterior. Integrar las ecuaciones de Hamilton por un tiempo τ , constituiría un paso de Metropolis-Hastings, respecto a la distribución canónica $f(\theta, m)$. Es decir, podemos pensar que seguir la trayectoria de la partícula por un tiempo τ constituye un kernel de propuestas $q(\tilde{\theta}, \tilde{m}|\theta, m)$.

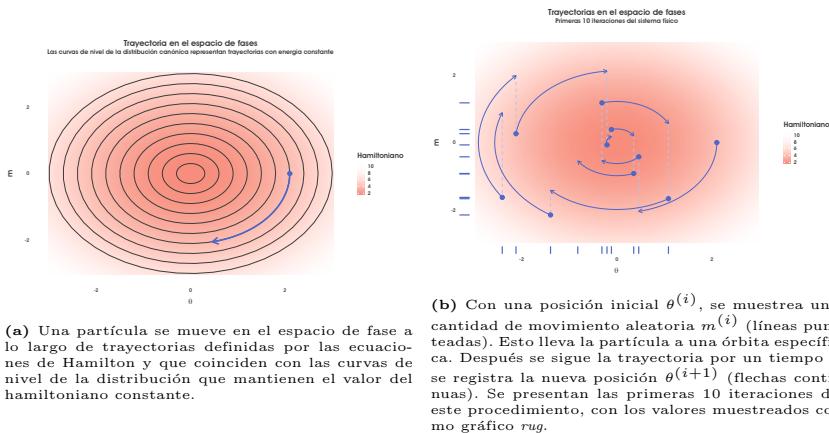


Figura 6.14: Ilustración de las trayectorias definidas por la distribución canónica de un espacio de fases en el que tanto la posición como el *momentum* se distribuyen normal estándar y en donde se muestrea mediante una estrategia de dos pasos. Fuente: elaboración propia con base en Betancourt (2017).

La única consideración teórica adicional es que, al finalizar de integrar la trayectoria por un tiempo τ , se deberían negar las variables de *momentum* para garantizar la reversibilidad de la cadena de Markov (Neal 1993, 2011; Betancourt 2017). Esta negación en la **Figura 6.13a** se vería simplemente como voltear la flecha. En las **Figuras 6.14** la partícula saltaría al lado contrario de la misma órbita. Ahora bien, debemos notar que esta negación del *momentum* vuelve al kernel de propuestas simétrico. Esta característica, junto con la propiedad de conservación del hamiltoniano— pues no hemos cambiado de órbita— lleva a que siempre aceptemos las propuestas.

En efecto, si comenzamos en un punto (θ, m) y seguimos la trayectoria un tiempo τ , llegamos a un nuevo punto (θ_τ, m_τ) ; negando el *momentum* llegaríamos al punto $(\theta_\tau, -m_\tau)$ de manera determinista, por lo que $q(\theta_\tau, -m_\tau | \theta, m) = 1$. Si ahora comenzamos en $(\theta_\tau, -m_\tau)$, la trayectoria “regresaría sobre sus propios pasos” excepto que con el signo contrario del *momentum* hasta $(\theta, -m)$. Al completar la propuesta, negando de nuevo cuenta el *momentum*, regresaríamos exactamente a (θ, m) . Así pues, resulta que también $q(\theta, m | \theta_\tau, -m_\tau) = 1$. Esto lleva a una probabilidad de aceptación (6.5) de 1:

$$\alpha(\theta_\tau, -m_\tau; \theta, m) = \min \left\{ \frac{f(\theta_\tau, -m_\tau) q(\theta, m | \theta_\tau, -m_\tau)}{f(\theta, m) q(\theta_\tau, -m_\tau | \theta, m)}, 1 \right\}$$

$$\begin{aligned}
&= \min \left\{ \frac{f(\theta_\tau, -m_\tau)}{f(\theta, m)}, 1 \right\} \\
&= \min \left\{ \frac{\exp[-H(\theta_\tau, -m_\tau)]}{\exp[-H(\theta, m)]}, 1 \right\} \\
&= \min \{ \exp[H(\theta, m) - H(\theta_\tau, -m_\tau)], 1 \},
\end{aligned}$$

y por conservación del hamiltoniano, $H(\theta, m) = H(\theta_\tau, -m_\tau)$, por lo que

$$\alpha(\theta_\tau, -m_\tau; \theta, m) = \min \{e^0, 1\} = 1. \quad (6.9)$$

Finalmente, como solo nos interesa conservar la nueva posición $\tilde{\theta} = \theta_\tau$ antes de muestrear un nuevo *momentum*, la negación al final de la trayectoria no necesita llevarse a cabo *en realidad*, por lo que el algoritmo puede proceder normalmente como los dos pasos antes mencionados.

Empero, todavía tenemos que considerar un detalle. Normalmente no tenemos tanta suerte como en el caso de la normal estándar y las ecuaciones de Hamilton no tienen una solución analítica simple. Esto quiere decir que tenemos que *aproximar* las trayectorias. Afortunadamente existe una familia poderosa de métodos, conocidos como *integradores simplécticos* que aproximan de buena manera. El lector interesado en la discusión sobre sus características y ventajas— incluyendo el cuidado que debe tenerse respecto a las divergencias— puede consultar Betancourt (2017) y Neal (1993). Para efectos de esta tesis, consideraré que se ha elegido ya una implementación de este tipo, siendo la más conocida el **Algoritmo 3**, conocido como *leapfrog*.

Algoritmo 3: Integrador simpléctico de *Leapfrog*

```

1 Valores iniciales  $\theta_0$  y  $m_0$  y parámetros  $\tau$  y  $\epsilon$ 
2 para  $t \leftarrow 0$  a  $\tau/\epsilon$  hacer
3    $m_{t+1/2} \leftarrow m_t - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta}(\theta_t)$ 
4    $\theta_{t+1} \leftarrow \theta_t + \epsilon m_{t+1/2}$ 
5    $m_t \leftarrow m_{t+1/2} - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta}(\theta_{t+1})$ 
6 fin

```

Cuando se utiliza un método simpléctico para aproximar las trayectorias de una

partícula, la propuesta no se mantiene exactamente sobre la misma órbita. Esto implica que el hamiltoniano al inicio de la trayectoria simulada y al final son distintos. Sin embargo, si la aproximación es buena, la propuesta estará en una órbita muy cercana a la verdadera y la diferencia en hamiltonianos será pequeña. Si no lo es, la propuesta estará más lejos y la diferencia en hamiltonianos será mayor. Afortunadamente, la corrección de Metropolis-Hastings, nos sirve precisamente para compensar estos sesgos. Las buenas aproximaciones se aceptan con alta probabilidad y las malas aproximaciones tienden a ser rechazadas.

Para ver por qué, notemos que el integrador simpléctico sigue siendo un algoritmo determinista y seguimos negando el *momentum* de la propuesta final. Esto quiere decir que podemos seguir considerando estos dos pasos— generación de trayectoria y negación de *momentum*— como un kernel de propuestas simétrico para Metropolis-Hastings y eliminarlo del cálculo de la probabilidad de aceptación, igual que en (6.9). La única diferencia es que ahora el hamiltoniano no se conserva exactamente, por lo que la probabilidad de aceptación depende del cambio en energía que se tiene al realizar la aproximación de la trayectoria mediante el integrador simpléctico y proponer $(\tilde{\theta}, \tilde{m})$:

$$\alpha(\tilde{\theta}, \tilde{m}; \theta, m) = \min \left\{ \exp \left[H(\theta, m) - H(\tilde{\theta}, \tilde{m}) \right], 1 \right\}. \quad (6.10)$$

Esta diferencia es positiva cuando la energía de la propuesta es menor a aquella inicial. Una energía menor corresponde a un valor mayor de la distribución canónica, por lo que en términos coloquiales estaríamos proponiendo un estado más probable. Así pues, la probabilidad de aceptación en estos casos es siempre 1. En el caso contrario, cuando la diferencia es negativa, diríamos que estamos proponiendo un estado menos probable por lo que no necesariamente queríamos movernos. Entre más negativa resulte la diferencia en energías, $\exp [H(\theta, m) - H(\theta_\tau, -m_\tau)]$ será más cercana a 0, por lo que tenderemos a rechazar las propuestas que nos llevarían a estados mucho menos probables.

De manera conceptual, entonces, el **Algoritmo 4** de *Hamiltonian Monte Carlo* o HMC para el aprendizaje bayesiano es un método de MCMC para simular de una distribución posterior $f(\theta|y)$ de acuerdo con los siguientes pasos:

1. La distribución $f(\theta|y)$ se representa de manera canónica (6.6) con función de energía potencial $U(\theta|y) = -\ln [f(y|\theta)f(\theta)]$.
2. Comenzando con un valor inicial $\theta^{(0)}$, se genera un valor auxiliar de *momentum*

$m^{(0)}$ de la misma dimensión que θ proveniente de una distribución especificada por el usuario y representada también de manera canónica para alguna energía cinética.

3. Mediante algún método válido simulamos una trayectoria de la dinámica hamiltoniana que seguiría una partícula en el espacio de fases iniciando en $(\theta^{(0)}, m^{(0)})$ y proponemos un nuevo estado $(\tilde{\theta}, \tilde{m})$.
4. Aceptamos $(\theta^{(1)}, m^{(1)}) = (\tilde{\theta}, \tilde{m})$ con probabilidad de aceptación (6.10); en caso de rechazar la propuesta, $(\theta^{(1)}, m^{(1)}) = (\theta^{(0)}, m^{(0)})$.
5. Repetimos el procedimiento para generar $(\theta^{(2)}, m^{(2)}), \dots, (\theta^{(N)}, m^{(N)})$.

Algoritmo 4: Hamiltonian Monte Carlo para el aprendizaje bayesiano

```

1 Valor inicial arbitrario o simulado  $\theta^{(0)} \in \mathbb{R}^d$ 
2 Para un momentum  $m \in \mathbb{R}^d$ , distribución conveniente  $f(m)$  con energía cinética
    $K(m)$ 
3 Algoritmo válido de simulación de trayectorias hamiltonianas  $\phi(\theta, m)$  para
    $n \leftarrow 1$  a  $N$  hacer
4    $\theta \leftarrow \theta^{(n-1)}$ 
5    $m \sim f(m)$ 
6    $(\tilde{\theta}, \tilde{m}) \sim \phi(\theta, m)$ 
7    $H(\theta, m) \leftarrow K(m) - \ln [f(y|\theta)f(\theta)]$ 
8    $H(\tilde{\theta}, \tilde{m}) \leftarrow K(\tilde{m}) - \ln [f(y|\tilde{\theta})f(\tilde{\theta})]$ 
9    $\alpha(\tilde{\theta}, \tilde{m}; \theta, m) \leftarrow \min \left\{ \exp [H(\theta, m) - H(\tilde{\theta}, \tilde{m})], 1 \right\}$ 
10   $u \sim U[0, 1]$ 
11  si  $u \leq \alpha(\tilde{\theta}, \tilde{m}; \theta, m)$  entonces
12     $\theta^{(n)} \leftarrow \tilde{\theta}$ 
13  en otro caso
14     $\theta^{(n)} \leftarrow \theta$ 
15  fin
16 fin
```

En estos pasos entonces hay dos grandes decisiones que el usuario debe tomar. En primer lugar, la elección de la distribución de *momentum* y que, general pero no necesariamente, resulta ser una normal centrada en cero independiente de θ y de los datos. En segundo lugar tenemos el paso de simulación de la dinámica hamiltoniana. Sobre este

punto vale la pena detenerse un poco.

El método más básico de HMC considera un tiempo fijo de integración τ , un tamaño de paso ϵ también constante y con estos parámetros genera trayectorias aproximadas de la misma longitud mediante el **Algoritmo 3**. No obstante, como mencionaba anteriormente, este no es el único método válido. Recientemente han surgido métodos que también simulan de una manera válida la dinámica hamiltoniana pero que, además, van adaptando el tiempo de integración τ y, por tanto, la longitud de las trayectorias generadas. El más famoso de estos algoritmos de HMC adaptativo se conoce como NUTS o *No-U-Turn Sampler* (Hoffman y Gelman 2011). La heurística detrás de NUTS es que se pueden generar trayectorias hamiltonianas eficientes si integramos no por un tiempo constante sino hasta el punto en que la partícula vaya a dar una “vuelta en U” (Betancourt 2017; McElreath 2017).

Por su parte, el *software* Stan también utiliza una implementación adaptativa de HMC. Además de ir modificando el tiempo de integración de las trayectorias, elige la propuesta mediante un muestreo multinomial dentro de la trayectoria generada y no necesariamente como el punto final de la misma (Betancourt 2016; Simpson 2018). Este nuevo software permite, al igual que hizo en su momento BUGS, definir modelos complejos mediante un lenguaje computacional sencillo sin que el usuario se preocupe por la implementación del algoritmo de MCMC. No obstante debe recalcarse que utilizar estas herramientas no significa abandonar las buenas prácticas de modelado y explorar distintas alternativas de parametrizaciones y modelos. Por el contrario, precisamente al liberar al usuario de la tarea más pragmática de cómputo, podemos dedicarnos a explorar los aspectos estadísticos más importantes y llegar a presentar mejores modelos que también se beneficien de mejor manera de las herramientas computacionales.

Para terminar este capítulo debo decir que, independientemente de la implementación específica que se tenga de HMC, la ventaja del algoritmo general sobre *Random Walk Metropolis* o *Gibbs Sampler* se relaciona con el último factor para el buen desempeño de MCMC que mencionaba en la página 91. Este es qué tantas iteraciones necesitamos para que, una vez que estamos en la región crítica de la distribución posterior donde queremos que la muestra se concentre, generemos otra observación casi independiente. Conforme aumenta la dimensión d de nuestros parámetros θ , esta región se vuelve cada vez más estrecha (Betancourt 2017; McElreath 2017). Podemos conceptualizarla como

una especie de circuito de carreras con diferentes curvas.

Random Walk Metropolis propone moverse en una dirección aleatoria y, si el kernel de propuestas es más “ancho” que la región crítica, la gran mayoría de las propuestas se “saldrían de la pista” y, por tanto, tenderían a ser rechazadas y la cadena se estancaría en la misma posición por un número importante de iteraciones. Como ya había mencionado anteriormente, una solución sería hacer más estrecho el kernel de propuestas, pero si es demasiado estrecho entonces las nuevas observaciones van a estar muy cerca de las anteriores y la cadena tardará en “dar la vuelta al circuito”. El problema es parecido para *Gibbs Sampler*, pues una región crítica estrecha significa que las condicionales completas tienden a estar fuertemente determinadas por las otras variables. Así pues, la cadena va avanzando dando pasos pequeños.

Por el contrario, las trayectorias de HMC constituyen un campo vectorial alineado con dicha región crítica (Betancourt 2017). Es decir, HMC funciona como un buen piloto de carreras que va siguiendo el circuito por lo que puede alcanzar más rápidamente zonas alejadas dentro de la región crítica de interés. El artículo de Betancourt (2017) es una gran referencia para observar con diagramas intuitivos este comportamiento. Termine recomendando al lector interesado que consulte y explore animaciones en internet como las que se encuentran en los *posts* de McElreath (2017) o Rogozhnikov (2016), donde se compara y discute el desempeño de HMC con los otros algoritmos de MCMC.

Parte III

Modelado de Datos Franceses

Capítulo 7

Datos franceses

Como adelantaba en la introducción, la tercera parte de este trabajo consiste concretamente en el modelado estadístico de los datos franceses que permitan la exploración de las configuraciones sociales que favorecieron o inhibieron el voto por el *Front National* en las elecciones Presidenciales de 2012. Por lo mismo, comienzo presentando los datos y realizando un análisis exploratorio; ese es el objetivo de este capítulo.

Debido a que el objetivo del modelado es explorar *configuraciones sociales*, la unidad básica de estudio son las *comunas*. Es decir, no estaré enfocándome en el nivel de individuo sino en los niveles de *colectividades territoriales* de Francia. Por ello, antes de comenzar, vale la pena recordar que los 3 niveles administrativos de Francia son, en orden de jerarquía, las regiones, los departamentos y las comunas; es decir, una comuna pertenece a un departamento, que a su vez pertenece a una región.

7.1. Datos electorales

En las elecciones de 2012, el presidente en funciones Nicolas Sarkozy perdió la reelección frente al socialista François Hollande, como puede verse en el **Cuadro 7.1**. En lo que respecta al FN, con Marine Le Pen como líderesa y candidata, el partido obtuvo 17.9 % de la votación efectiva en la primera vuelta presidencial. Dicho porcentaje representó el 3er lugar en la elección, mejorando el 4to puesto de 2007. No obstante, fue insuficiente para disputar la segunda vuelta electoral.

Elecciones Presidenciales 2012

Candidato(a)	Partido	Votos	% Ef.	Votos	% Ef.
		1ra vuelta		2da vuelta	
François Hollande	PS	10,272,705	28.63	18,000,668	51.64
Nicolas Sarkozy	UMP	9,753,629	27.18	16,860,685	48.36
Marine Le Pen	FN	6,421,426	17.90		
Jean-Luc Mélenchon	FG	3,984,822	11.10		
François Bayrou	MoDem	3,275,122	9.13		
Eva Joly	EELV	828,345	2.31		
Nicolas Dupont-Aignan	DLR	643,907	1.79		
Philippe Poutou	NPA	411,160	1.15		
Nathalie Arthaud	LO	202,548	0.56		
Jacques Cheminade	SP	89,545	0.25		
Votación efectiva		35,883,209		34,861,353	
Blancos o nulos		701,190		2,154,956	
Votación emitida		36,584,399		37,016,309	
Abstenciones		9,444,143		9,049,998	
Lista nominal		46,028,542		46,066,307	

Cuadro 7.1: Resultados de las elecciones presidenciales francesas de 2012, resultando presidente electo François Hollande (PS). Fuente: elaboración propia con los datos oficiales del Ministerio del Interior.

El análisis que realizaré se circunscribe a los resultados dentro de la metrópoli francesa. Esta decisión está basada en dos consideraciones. La principal es que no se cuentan con todos los datos de las variables explicativas para todos los territorios de ultramar. Adicionalmente, el fenómeno político es marcadamente distinto en los *DROM* que en el hexágono. Si el resultado nacional fue de 17.9 %, en la metrópoli llegó al 18.3 % mientras que fuera resultó ser de poco más del 7 %. El *DROM* en el que Marine Le Pen obtuvo el mayor porcentaje efectivo de votos fue la Guyana, con 10.5 %; este resultado fue casi dos puntos porcentuales menos que su peor región metropolitana, Île de France, con 12.3 %. En Mayotte, Le Pen se quedó incluso por debajo del 3 % de la votación efectiva. Más aún, si en lugar de observar la votación efectiva tomamos en cuenta los votos nulos y las abstenciones mediante la votación bruta—es decir los votos con respecto al número de inscritos en la lista nominal—la diferencia metrópoli-ultramericano es del 14.6 % contra 3.5 %. Las diferencias políticas no solo se reflejan en preferencias electorales y niveles de participación sino que los temas de discusión y alrededor de los cuales se desarrollan las

campañas también son diferentes. Así pues, deberá recordarse que cualquier conclusión que se tenga del análisis se refiere exclusivamente a la metrópoli francesa.

Esta última consideración nos hace notar que la distribución de los votos no es uniforme. Existen zonas de fortaleza y debilidad del FN. En 70 comunas Le Pen no recibió ningún sufragio mientras que la comuna en la que fue más votada obtuvo 62.5 % de la votación bruta¹. El rango intercuartílico va de 13.54 % a 21.44 % con una mediana de 17.46 %, estos valores están señalados en el eje del histograma de la **Figura 7.1**.

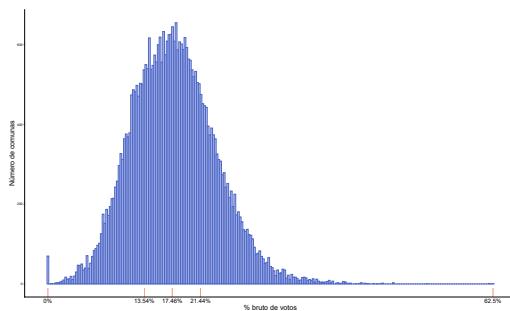
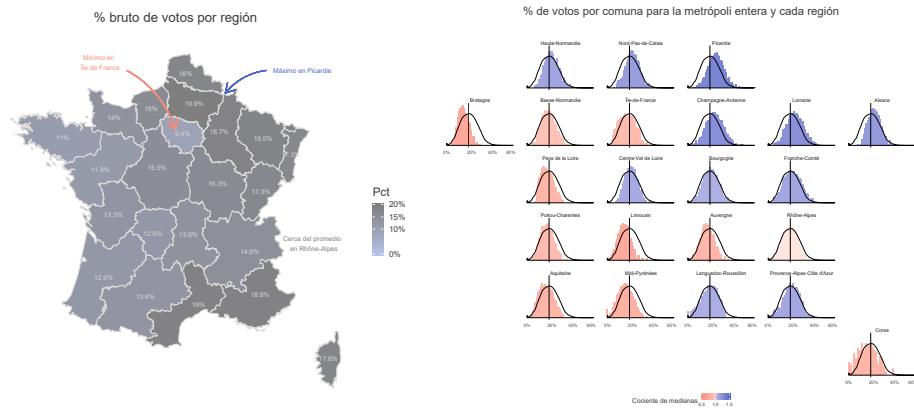


Figura 7.1: Histograma del % bruto de votos obtenido en cada comuna. Fuente: elaboración propia con base en los datos electorales oficiales del Ministerio del Interior francés.

Podemos agregar los datos a nivel región como en la **Figura 7.2a**, para describir de manera más general las zonas de fortaleza o debilidad. Mientras que el porcentaje de votos brutos en la metrópoli fue de 14.6 %, en la región île de France, Marine Le Pen obtuvo el 9.4 % y en Picardie llegó a obtener casi el 20 %.

Otra forma de valorar la presencia territorial del partido es considerar las comunas de cada región y comparar su distribución frente a la distribución agregada de toda la metrópoli. Podemos ver este ejercicio mediante el *geofacet* de la **Figura 7.2b**. Dentro de él comparo los histogramas de los porcentajes de votos que recibió la candidata frontista en las comunas de cada región con la distribución de los porcentajes para todas las comunas de la metrópoli. La intensidad del color de los histogramas depende del cociente del porcentaje mediano de votos en la región entre el porcentaje mediano de votos en toda la metrópoli, por lo que representa la fuerza relativa de cada etiqueta en la región. En azules encontramos las regiones con medianas superiores a la nacional y en rosas a aquellos con menores.

¹Hay que decir que recibió 5 votos de un listado nominal de 8 personas.



(a) Mapa del % bruto de votos obtenido por Marine Le Pen en las elecciones presidenciales del 2012 en cada región francesa.

(b) Los histogramas representan la distribución para cada una de las regiones y las densidades la de la metrópoli francesa.

Figura 7.2: Fuente: elaboración propia con base en los datos electorales oficiales del Ministerio del Interior francés y la cartografía de OpenStreetMap.

Vemos que, en las regiones del noreste francés como Picardie o Alsace, los histogramas del FN están desplazados a la derecha de la distribución de referencia y, por tanto, están coloreados con mayor intensidad de azul. Por el contrario, en regiones occidentales como Bretagne, Limousin o Aquitaine, los histogramas reflejan menores porcentajes de votos para el FN y se colorean más de rosa. Podemos empezar a conjeturar que una clave geográfica para el voto frontista es la diagonal que va de Normandía— Haute y Basse Normandie— hacia PACA²: si la comuna se encuentra al noreste de la diagonal, tendería a manifestar mayor apoyo al Front National que si se encuentra al sudoeste de la misma. Este sería a grandes rasgos el eje que Goodliffe (2016) identifica como la línea que une las ciudades de Cherbourg en Normandía y Valence en Rhône-Alpes y después con Perpignan en Languedoc-Rousillon.

Si descendemos un nivel y observamos el mapa a nivel departamento en la **Figura 7.3a**, la división este-oeste se ve de manera más clara. También podemos ver las distribuciones a nivel departamento mediante diagramas de violines. El *geofacet* respectivo se observa en la **Figura 7.3b**. En cada región se muestran 3 líneas como referencia del rango intercuartílico y la mediana considerando las comunas de toda la metrópoli. El diagrama

²Provence-Alpes-Côtes d'Azur.

de violín sin relleno muestra dicha distribución agrupada para toda la metrópoli. Los diagramas de violín con relleno representan, pues, las distribuciones del porcentaje de votos obtenido en cada comuna del departamento correspondiente, identificado mediante su código oficial geográfico. Dentro de cada región los departamentos están ordenados de menor a mayor apoyo al FN con base en las medianas. Al igual que en los histogramas a nivel región, la intensidad del relleno es el cociente de la mediana departamental respecto a la mediana global.

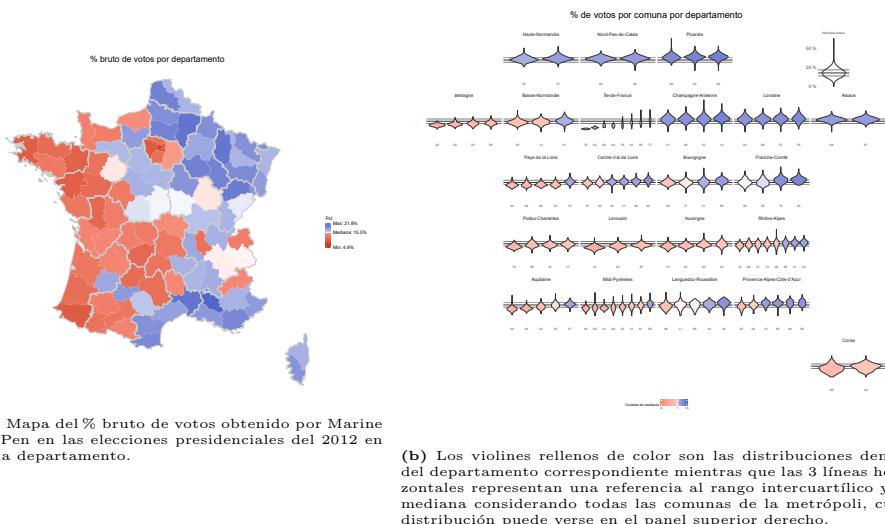


Figura 7.3: Fuente: elaboración propia con base en los datos electorales oficiales del Ministerio del Interior francés y la cartografía de OpenStreetMap.

Dentro de las regiones con mayor apoyo en general, este es relativamente homogéneo a través de los departamentos. En efecto, las distribuciones reflejadas en diagramas de violines para los departamentos dentro de Lorraine, Champagne-Ardenne o Picardie aparentan ser similares, siempre con cocientes de medianas mayores a 1. No obstante, existen regiones— como Île de France o Aquitaine— cuyos departamentos presentan distribuciones más variables. Este es un elemento que podría empezar a sugerir un modelado jerárquico de los datos por departamento.

Finalmente, observamos el mapa de los porcentajes de voto brutos que obtuvo Marine Le Pen en cada comuna en la **Figura 7.4**. De nueva cuenta distinguimos zonas de fuerza

o debilidad con respecto a la mediana. Es decir, en tonos de azul se encuentran la mitad de las comunas con mayor porcentaje de votos y en rojos la mitad de menor porcentaje. La variabilidad es clara. Observamos lo que autores como Le Bras (2015) llaman las capas del voto FN. En primer lugar la gran zona de fuerza del norte. También hay corredores de fortaleza: el litoral mediterráneo y dos diagonales que parecen desembocar en él; estas serían las márgenes del Ródano que viene del norte y del Garona que viene del Atlántico. La débilidad la encontramos al oeste y en zonas urbanas vistas como manchas rojas rodeadas de color azul, particularmente París.

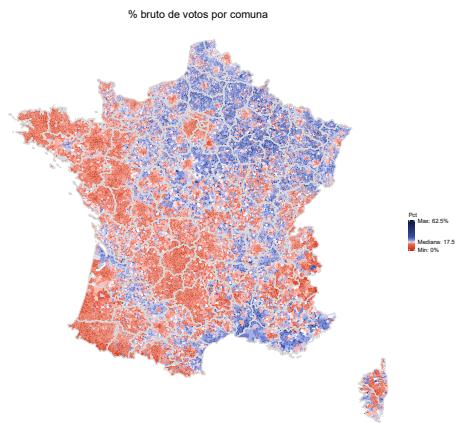


Figura 7.4: Mapa del % bruto de votos obtenido por Marine Le Pen en las elecciones presidenciales del 2012. Fuente: elaboración propia con base en los datos electorales oficiales del Ministerio del Interior francés y la cartografía de OpenStreetMap.

7.2. Datos censales

Por otro lado, para caracterizar a las comunas en términos de *configuraciones sociales* relacionadas con el voto nativista y autoritario de corte populista, requerimos su composición en términos de variables sociodemográficas. Afortunadamente, desde 2004, Francia cuenta con un sistema censal rotante que permite realizar estimaciones anuales para este tipo de variables.

De acuerdo con el INSEE (2017), el censo francés está compuesto por dos mecanismos distintos dependiendo del tamaño poblacional de la comuna. La comunas de menos de

10,000 habitantes realizan una encuesta censal a razón de una de cada cinco comunas todos los años. Las comunas de 10,000 habitantes o más realizan todos los años una encuesta por muestreo probabilístico del 8 % de sus hogares cada año. Acumulando cinco años, el total de los habitantes de las comunas “pequeñas” y al rededor del 40 % de la población de las comunas “grandes” son tomados en cuenta. Esta muestra acumulada se trata de manera estadística para estimar la población total en cada comuna al tercer año de una ventana de cinco. Así, para la estimación de la población al año n se consideran los datos de los años $n - 2$, $n - 1$, n , $n + 1$ y $n + 2$. Cada año se desecha la información más antigua y se incorpora la información del nuevo año. La primera estimación anual se tuvo en 2006 considerando la información de 2004 a 2008.

A partir de estos datos oficiales anuales que publica el INSEE podemos descomponer las distribuciones poblacionales en las comunas francesas a lo largo de algunas variables que la revisión de literatura sugiere: las categorías socioprofesionales, nacionalidad o condición migratoria, sexo y edad. Para ello he calculado el porcentaje de individuos como proporción de la población comunal de cada categoría dentro de estas 5 variables sociodemográficas básicas provenientes de 3 bases distintas que se pueden ver en el **Cuadro 7.2**.

Base de origen	Variable	Abreviatura	Categoría
POB1B	Sexo	Hom Muj	Hombres Mujeres
	Edad	Ed1	0 a 17 años
		Ed2	18 a 24 años
		Ed3	25 a 39 años
		Ed4	40 a 54 años
		Ed5	55 a 64 años
		Ed6	65+ años
NAT3A	Nacionalidad	Fra Ext	Franceses Extranjeros
	Categoría Socioprofesional	CSP1	Agricultores
		CSP2	Artesanos, comerciantes y empresarios
		CSP3	Cuadros y profesiones intelectuales superiores
		CSP4	Profesiones intermedias
		CSP5	Empleados
		CSP6	Obreros
		CSP7	Retirados
		CSP8	Otras personas sin actividad
IMG1A	Condición migratoria	Inm Loc	Inmigrantes Locales

Cuadro 7.2: Variables censales del INSEE a utilizar en el análisis.

Debido a que las variables de Sexo, Nacionalidad e Inmigración cuentan con dos ca-

tegorías, solo presento las distribuciones de una categoría de referencia, pues la otra solo es el complemento de esta.

Comenzando con la distribución del porcentaje de Mujeres en las comunas en 2012 vemos que, como es de esperarse, la mayoría de los departamentos tienen distribuciones muy concentradas al rededor del 50 % de la población. Las únicas distribuciones que llaman la atención en la **Figura 7.5** son las de los departamentos de Île de France, con porcentajes un poco mayores que el resto de la metrópoli.

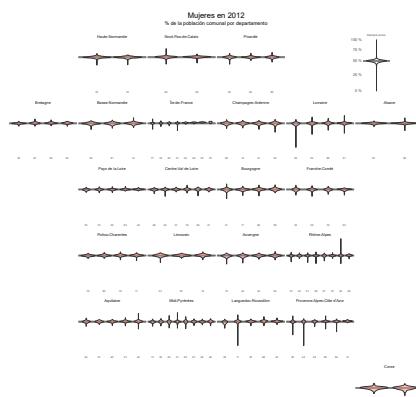


Figura 7.5: Distribuciones departamentales del porcentaje de mujeres como proporción de la población de las comunas en 2012. Fuente: elaboración propia con los datos censales.

Las variables de nacionalidad y condición migratoria son similares pero no idénticas. El INSEE (2019) define a un extranjero como un residente del territorio francés que no posee la nacionalidad francesa y a un inmigrante como aquella persona nacida extranjera en el extranjero; esto quiere decir, por ejemplo, que hay individuos franceses considerados inmigrantes pues pudieron haberse naturalizado, así como extranjeros considerados locales porque nacieron en territorio francés sin tener derecho a la nacionalidad.

En el proceso del modelado se deberá elegir la “mejor” de entre ambas. Por el momento, en la **Figura 7.6** vemos las distribuciones para los extranjeros y los inmigrantes en 2012. Debido a que la mayoría de inmigrantes son extranjeros las distribuciones siguen un patrón muy parecido, salvo que en general siempre hay más inmigrantes que

extranjeros— observar líneas de referencia—. Mientras que en regiones del norte las comunas tienden a tener pocos extranjeros e inmigrantes, vemos que en la región parisina de Île de France tienen una presencia considerable. Las regiones más sureñas como Corse o PACA también tienen comunas con mayor presencia de extranjeros e inmigrantes.

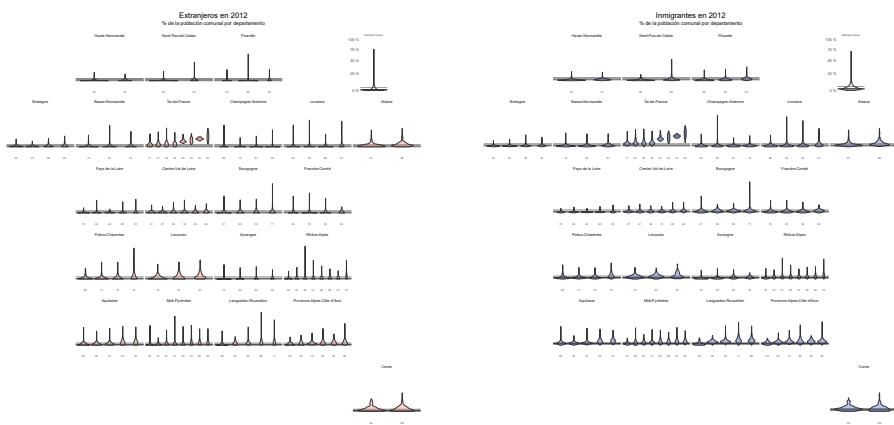


Figura 7.6: Distribuciones departamentales del porcentaje de extranjeros y de inmigrantes como proporción de la población de las comunas en 2012. Fuente: elaboración propia con los datos censales.

Ahora bien, la estructura generacional de las comunas francesas va cambiando. Si observamos las distribuciones para los diferentes grupos de edad en la **Figura 7.7** vemos que hay regiones cuyos departamentos tienden a tener comunas más *envejecidas* en el sentido de que hay comparativamente menor porcentaje de menores de edad o jóvenes que el resto de la metrópoli y mayores porcentajes de personas de 65 años o más que en el resto del hexágono francés. Esto se ve particularmente en Corse, pero también en otras regiones como Limousin, Midi-Pyrénées o Bourgogne. Por el contrario, el norte y particularmente Île de France presentan una estructura generacional más joven.

Finalmente, vemos en la **Figura 7.8** las distribuciones para las diferentes categorías socioprofesionales. El primer dato que salta a la vista es que la región parisina tiene una composición socioprofesional muy particular. En efecto, Île de France es una región que comparativamente hablando, casi no tiene agricultores, obreros o retirados. Por el contrario, tiene un fuerte componente de los llamados cuadros y profesiones intelectuales superiores, pero su distribución es distinta a través de los departamentos que conforman

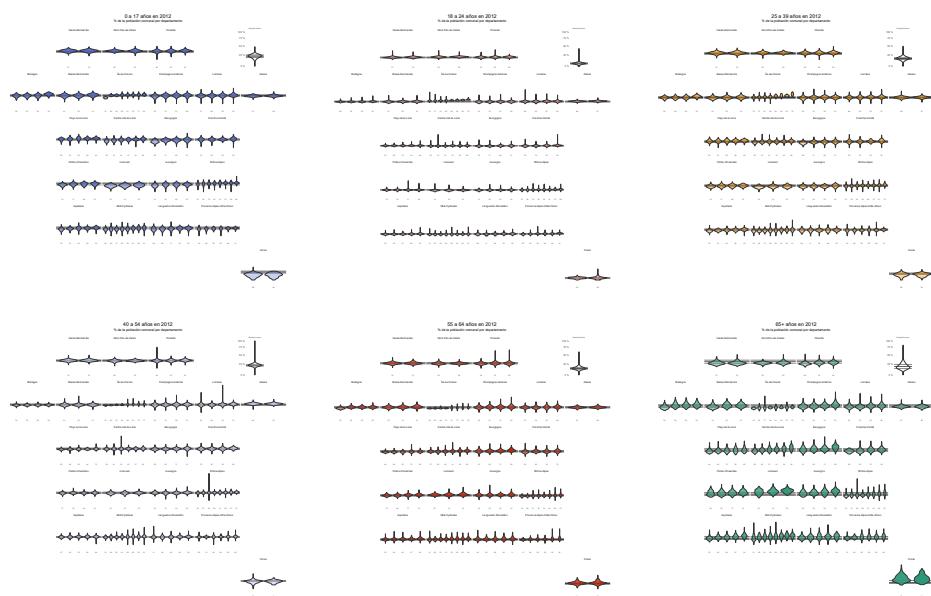


Figura 7.7: Distribuciones departamentales del porcentaje de los distintos grupos de edad como proporción de la población de las comunas en 2012. Fuente: elaboración propia con los datos censales.

la zona metropolitana. Dentro de la ciudad de París—departamento 75—representan el mayor porcentaje. Pero también forman un porcentaje importante de las poblaciones comunales de los departamentos al poniente y al sur de este departamento. Estos departamentos son el 92-Hauts-de-Seine y 94-Val-de-Marne, dentro de la llamada *pequeña corona* de París y los departamentos occidentales de la *gran corona* de París, que incluye 78-Yvelines, 91-Essonne y 95-Val-d'Oise. Por el contrario, en el oriente de la zona metropolitana, en los departamentos de 93-Seine-Saint-Denis y 77-Seine-et-Marne, viven los empleados de la metrópoli. Las profesiones intermedias viven más bien en la *gran corona*, es decir en los departamentos 77, 78, 91 y 95.

En el resto de la metrópoli francesa procederé a comentar categoría por categoría. Los agricultores están presentes sobre todo en las regiones centrales y sureñas. La pequeña burguesía—es decir los artesanos, comerciantes y empresarios—tiene presencia en general nacional, pero resaltan las regiones de la costa mediterránea como PACA y Languedoc-Rousillon. Los cuadros y las profesiones intelectuales superiores están más representadas en las grandes ciudades como son Lyon, Marsella, Toulouse, Niza o Lille—

departamentos 69 en Rhône-Alpes, 13 en PACA, 31 en Midi-Pyrénées, 06 en PACA y 59 en Nord-Pas-de-Calais respectivamente—.

Las profesiones intermediarias tienen una variabilidad intraregional que se puede observar por la progresión de intensidades en los colores de los violines; es decir, dentro de cada región tienden a haber departamentos con medianas superiores a la mediana nacional y otros con medianas inferiores. Los empleados, por el contrario, parecen distribuirse de manera relativamente homogénea a través de los departamentos de una misma región, salvo por algunas comunas atípicas que alargan las colas de algunos violines. El norte industrial es visible desde las distribuciones de los obreros, quienes están más presentes en las regiones del norte y oeste que en el sureste.

Las distribuciones de las proporciones de las poblaciones que representan las personas retiradas permite confirmar que el sur de Francia está más avejentado que el norte, aunque también observamos que dentro de algunas regiones como Aquitaine, Midi-Pyrénées o Bourgogne existe variabilidad a través de sus departamentos. Finalmente, la categoría de personas sin actividad es más bien residual y un poco más difícil de interpretar en este nivel, pues incorpora tanto menores de edad como personas sin trabajo, incluidos estudiantes. Podemos, sin embargo, observar similitudes con las distribuciones de los menores de edad y seguir pensando que las regiones del norte tienen una estructura más juvenil que regiones más avejentadas como Limousin y Auvergne.

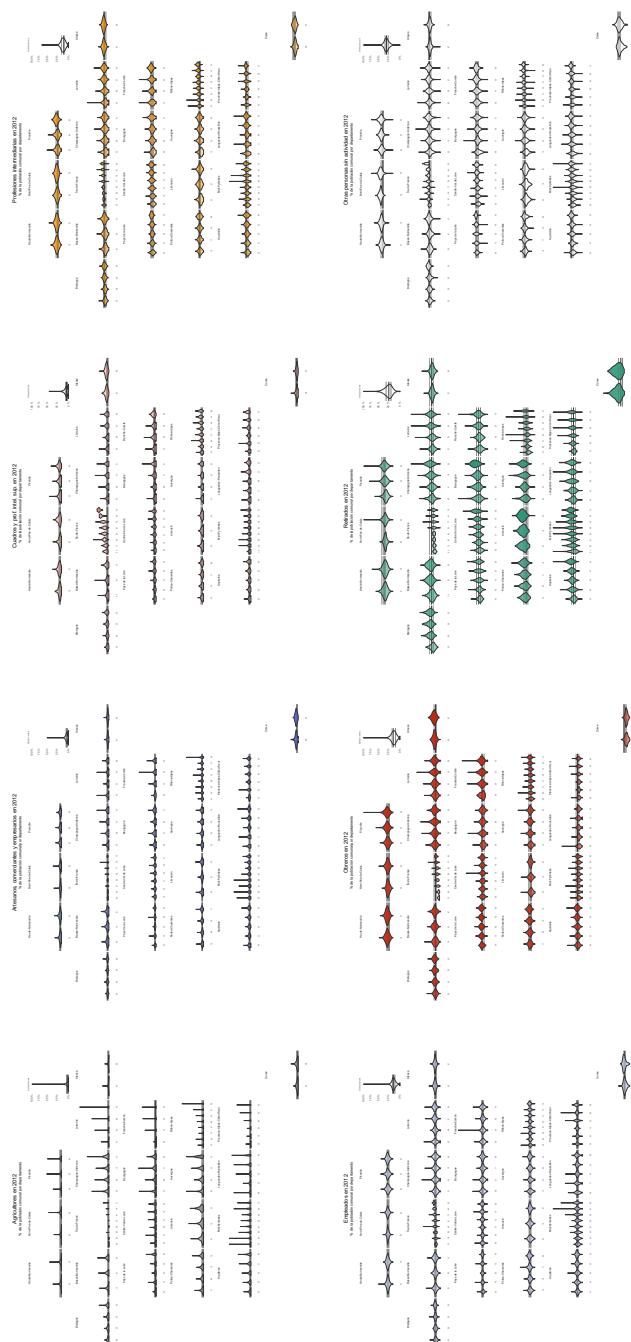


Figura 7.8: Distribuciones departamentales del porcentaje de los distintos grupos de categorías socioprofesionales como proporción de la población de las comunas en 2012. Fuente: elaboración propia con los datos censales.

7.3. Otros datos a nivel comuna

Derivado del marco teórico sobre el voto NAP existen otras variables explicativas de interés que ayuden a perfilar el clivaje de escolaridad o el estado económico del lugar. Afortunadamente el INSEE también publica datos sobre el máximo grado escolar que tienen los habitantes mayores de 15 años de cada comuna, así como el número de habitantes que siguen estudiando. El sistema educativo francés tiene una estructura distinta al mexicano, así que para las categorías en el **Cuadro 7.3** Dip2 incluye CEP, BEPC o Brevet; Dip3 se refiere a CAP, BEP, varios tipos de Baccalauréat, BEA, BEC, BEI, BEH, BTS, DUT, entre otros; para clasificarse en Dip4 se requieren al menos 2 años de ciclos universitarios. Por otro lado, también es posible obtener el número de personas empleadas y desempleadas dentro de las comunas para 3 grupos de edad. La distinción por grupos de edad es importante puesto que referencias como Le Bras (2016) o Perrineau (2007) tienden a mencionar la precariedad salarial juvenil como un catalizador del voto frontista.

Base de origen	Variable	Abreviatura	Categoría
Diplômes-formation	Escolaridad	Dip1	Personas sin escolaridad
		Dip2	Primaria o secundaria
		Dip3	Preparatoria o equivalente
		Dip4	Universidad o más
		Esc	Personas aún estudiando
Emploi-population active	Empleo	Ocu1	Empleados de 15 a 24 años
		Des1	Desempleados de 15 a 24 años
		Ocu2	Empleados de 25 a 54 años
		Des2	Desempleados de 25 a 54 años
		Ocu3	Empleados de 55 a 64 años
		Des3	Desempleados de 55 a 64 años

Cuadro 7.3: Otros datos a nivel comuna a utilizar en el análisis.

Si repetimos el análisis exploratorio hecho para los datos censales, vemos en la **Figura 7.9** las distribuciones departamentales por nivel escolar en 2012. En primer lugar, confirmamos la singularidad de Île de France respecto al resto del país. Es claramente la región con los departamentos más escolarizados de Francia, en el sentido de que tienden a tener una población con mayores niveles de población escolarizada y con diploma universitario—notablemente en 75-París intramuros y el acaudalado 92-Hauts-de-Seine—. Asimismo, observamos menores porcentajes de personas sin ningún diploma escolar, salvo el caso aparentemente atípico de 93-Seine-Saint-Denis, pero que veíamos

en la sección anterior que era el departamento de más inmigrantes, empleados y obreros. Resulta ilustrativo el patrón de personas con preparatoria pues hay un contraste entre París y su *petite courone* y la *grande courone*— 75, 92, 93 y 94 vs 78, 95, 91 y 77—.

Respecto al panorama general de la metrópoli, de nueva cuenta se refleja el patrón generacional de un norte más joven que el sur observando las distribuciones de personas todavía estudiando. Resaltan departamentos en Corse, Poitou-Charentes, Basse-Normandie o Picardie como lugares con altas poblaciones no escolarizadas. Por su parte, Limousin, Auvergne, Champagne-Ardenne y, en cierto sentido, Midi-Pyrénées son regiones con fuertes poblaciones cuyo máximo nivel de estudios es la educación básica.

En cuanto a las personas con preparatoria, ambos departamentos en Alsace se ubican por encima de la mediana nacional. En términos de población con diploma universitario encontramos lugares cercanos a grandes universidades francesas como la de Aix-Marsaille, 13 en PACA, o en Rhône-Alpes las de Grenoble-Alpes (38) y la ENS-Lyon (69). También en dicha región resalta que el departamento con mayor mediana es 74-Haute-Savoie, probablemente se debe a su carácter conurbado con la ciudad suiza de Ginebra.

En el caso del empleo, para cada grupo de edad, contamos con el porcentaje de desempleados y su complemento. Bastan presentarse entonces los gráficos para los desempleados en la **Figura 7.10**. Lo primero que llama la atención es el rango de variabilidad del porcentaje de desempleados juveniles. Mientras que en el resto de variables aquí consideradas las comunas que llegan a tener valores de 0 o 100 por ciento son más bien casos atípicos en las colas de las distribuciones, en la distribución nacional de desempleo juvenil observamos modas en 0 y en 100. Este carácter multimodal se conserva en la mayoría de los departamentos.

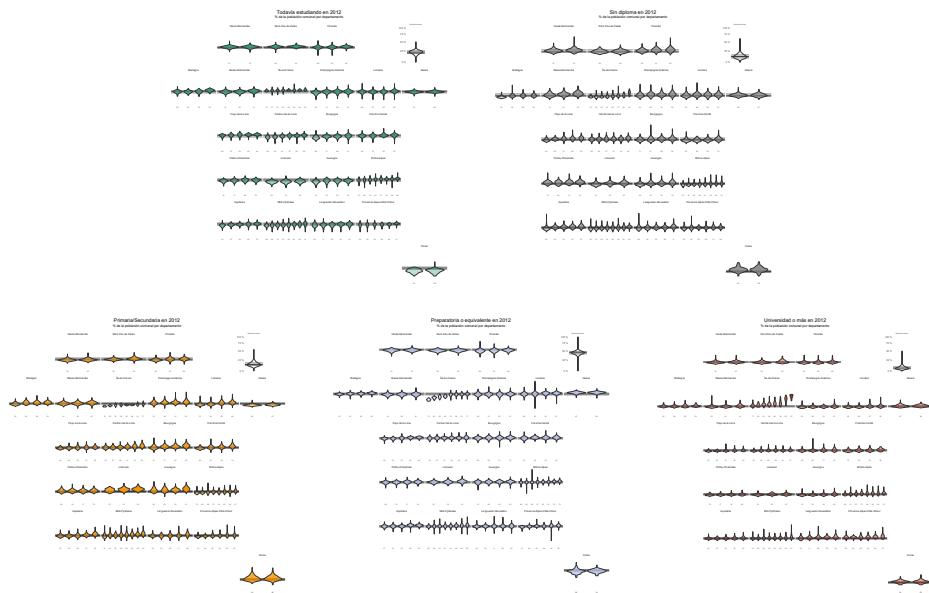


Figura 7.9: Distribuciones departamentales del porcentaje de los distintos grupos de escolaridad como proporción de la población de las comunas en 2012. Fuente: elaboración propia con los datos del INSEE.

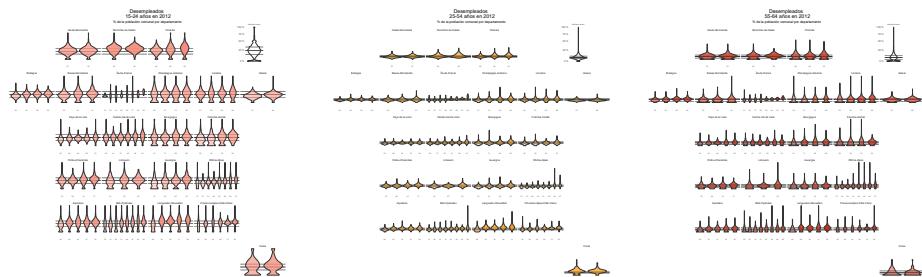


Figura 7.10: Distribuciones departamentales de los porcentajes de desempleados en 2012. Fuente: elaboración propia con los datos del INSEE.

Para el desempleo general, 25 a 54 años, las distribuciones son más homogéneas. Notamos algunos casos por encima de la referencia nacional como el popular 93-Seine-Saint-Denis o los departamentos sureños del Languedoc-Rousillon o Corse. En el desempleo de adultos cercanos al retiro, 55-64 años, las distribuciones son mucho más sesgadas a la derecha con algunas colas largas.

7.4. Muestra de comunas

En el proceso de modelado del siguiente capítulo se discutirán una serie de modelos. Debido al costo computacional que implica generarlos todos, decidí seleccionar una muestra aleatoria de comunas para poder llevarlo a cabo. Esta es una práctica recomendada cuando se intentarán explorar varios modelos con el objetivo de ir entendiendo de mejor manera el problema (Gelman y Hill 2006). Después de algunas pruebas— tomando en cuenta tanto el tiempo de ajuste como la precisión esperada dado el tamaño de muestra— decidí que muestrear alrededor de 4,000 comunas de las poco más de 36,000 existentes era adecuado. Debido a que el análisis exploratorio discutido hasta ahora comienza a sugerir un modelado jerárquico de los datos a nivel departamento, el diseño de muestreo considerado es estratificado por departamento.

París y su pequeña corona— departamentos 75, 92, 93 y 94— están conformados por pocas comunas. En efecto, de entre estos, el de mayor número de comunas era el 94-Val-de-Marne, con 46, mientras que del resto de departamentos el que menos comunas tenía era el 90-Territoire de Belfort con 101. Esta diferencia de más del doble motivó la decisión de que para estos 4 departamentos en lugar de muestrear se censaran.

Para el resto de departamentos se realizó un muestreo aleatorio simple dentro de cada departamento con un tamaño de muestra proporcional al número de comunas del mismo. Sin embargo, como el departamento 75-París se censaría en sus 20 comunas³, otra consideración fue que el tamaño mínimo de muestra en cada departamento fuera de 20 comunas. En total entonces se obtuvo una muestra de 4,157 comunas. **Tanto el listado de comunas como el script con el que se obtuvo la muestra están disponibles en el repositorio de Github de esta tesis.**

En la **Figura 7.11a** vemos un comparativo de las distribuciones de porcentajes brutos de votos para las comunas muestreadas y aquellas fuera de muestra. Vemos que la muestra cubre correctamente el rango de la distribución, sin embargo está más concentrada en el centro de la misma y no refleja de la mejor manera las colas, sobre todo la moda en 0 que sí presentan las comunas fuera de la muestra. Esto hay que tenerlo en

³Recordemos que técnicamente París es al mismo tiempo un departamento y una sola comuna. Sin embargo, tanto los resultados electorales como los datos del INSEE se tienen para cada uno de los 20 *arrondissements* en los que se divide. Por ello, cada uno de estos *arrondissements* se consideró como si fueran una comuna. Esto se repitió para los *arrondissements* de las ciudades de Lyon y de Marsella.

cuenta pues probablemente las predicciones que hagan los modelos serán más concentradas hacia la media o mediana de la distribución nacional y fallarán para comunas con porcentajes observados extremos. Sin embargo, para efectos de construcción del modelo es un sacrificio que aceptamos. La muestra logra delinear las grandes zonas de fuerza y debilidad del FN como vemos en el mapa de la **Figura 7.11b**.

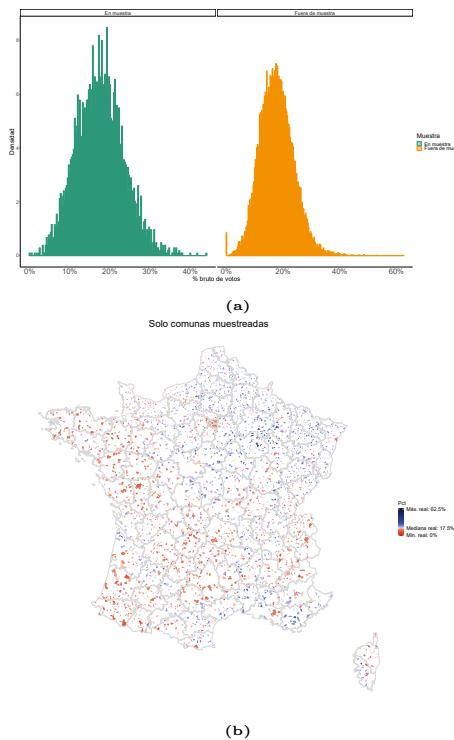


Figura 7.11: Distribuciones y mapa del % bruto de votos para las comunas muestreadas. Fuente: elaboración propia con base en los datos electorales oficiales del Ministerio del Interior francés y la cartografía de OpenStreetMap.

Si estimáramos el porcentaje de votos en cada departamento con base en el porcentaje observado en la muestra, tendríamos errores que van de 0 puntos porcentuales para los departamentos censados hasta un máximo de 4.2 pp de error en el Bas-Rhin. 86 de los 96 departamentos tienen errores menores a 2.5 pp, lo que se observa por la mayoría de tonos rosas y morados en la **Figura 7.12**. Este es un mapa *dorling* en el que cada bolita representa un departamento, identificado con su código geográfico INSEE. Los departamentos que pertenecen a una misma región están unidos mediante las líneas grises.

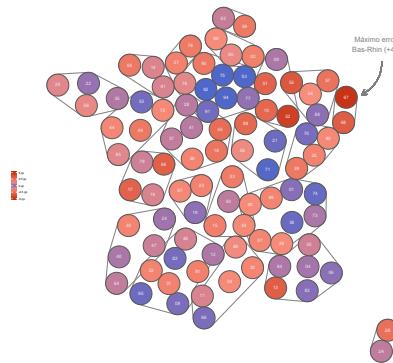


Figura 7.12: Errores departamentales de estimación con base en la muestra. Fuente: elaboración propia.

En general, los mayores errores se encuentran en departamentos del noreste francés, una zona de fortaleza del FN. También recordemos que los departamentos “sobremuestreados” por ser censados forman parte de Île de France, una región de debilidad frontista. Por lo mismo, podríamos esperar que las predicciones de los modelos ajustados a estos datos subestimen un poco los porcentajes realmente obtenidos. En este sentido podríamos conjeturar que un modelo ajustado con estos datos se desempeñe de mejor manera en departamentos señalados en azul que en aquellos en rojo.

7.5. Asociaciones

Con esta muestra podemos explorar las asociaciones entre el voto y las configuraciones sociales definidas por las variables seleccionadas. Debido a que los datos que tenemos son tanto los votos obtenidos como el número de inscritos, es decir el máximo de posibles votos, una buena alternativa de modelado será considerar los datos como binomiales. Por tanto, en lugar de intentar una asociación directamente en la escala de los votos o del porcentaje del voto, podemos buscarla en la escala logística.

En la **Figura 7.13** observamos diagramas de dispersión de los porcentajes brutos de votos observados en la escala logística⁴ con respecto a los distintos porcentajes de la po-

⁴Cuando los porcentajes observados son iguales a 0, la transformación logística no está definida por lo que los puntos son representados como si estuvieran por debajo de cualquier otro valor en el eje vertical. Recordemos que el modelo binomial sí permite una cantidad nula de votos, es solo el logit de

blación que representa cada categoría de las variables explicativas. Cuando las categorías son dicotómicas, como el caso del sexo, solo se presenta una de las dos.

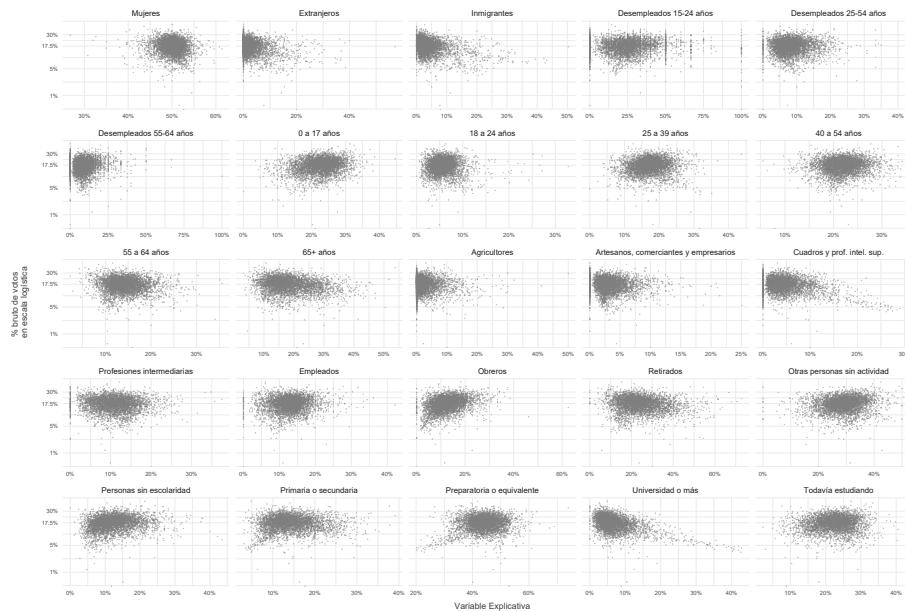


Figura 7.13: Diagramas de dispersión para los porcentajes brutos de voto y las distintas variables explicativas observados en la muestra de comunas. Fuente: elaboración propia.

Vemos que varias de las categorías presentan patrones de asociación tanto negativa—Mujeres, Extranjeros, Inmigrantes, Universidad o más— como positiva— 0 a 17 años, Obreros—. También observamos que las variables explicativas presentan observaciones en 0 que dificultan a primera vista la asignación de una tendencia.

Ahora bien, estas nubes de puntos son para el conjunto de la muestra. Debido a que el fenómeno electoral es variable por departamento, podríamos preguntarnos si existen distintas asociaciones a través de los diferentes departamentos. En caso de existir diferencias, esperaríamos que un modelado jerárquico de los datos fuera una mejor alternativa que un modelo de agregación completa. Para explorar esta hipótesis presento los gráficos de la **Figura 7.14** que llamo tendencias ingenuas con el objetivo de reafirmar que este es solamente un análisis exploratorio y no concluyente. Mediante el comando

un parámetro de probabilidad de éxito igual a 0 lo que no estaría definido.

`ggplot2::geom_smooth()` en R, podemos ajustar tendencias lineales a todos los datos de la muestra que no hayan presentado un porcentaje de votos igual a 0. Estas son las líneas azules gruesas. Asimismo podemos realizar el ajuste para cada departamento, obteniendo las 96 líneas delgadas en color rosa.

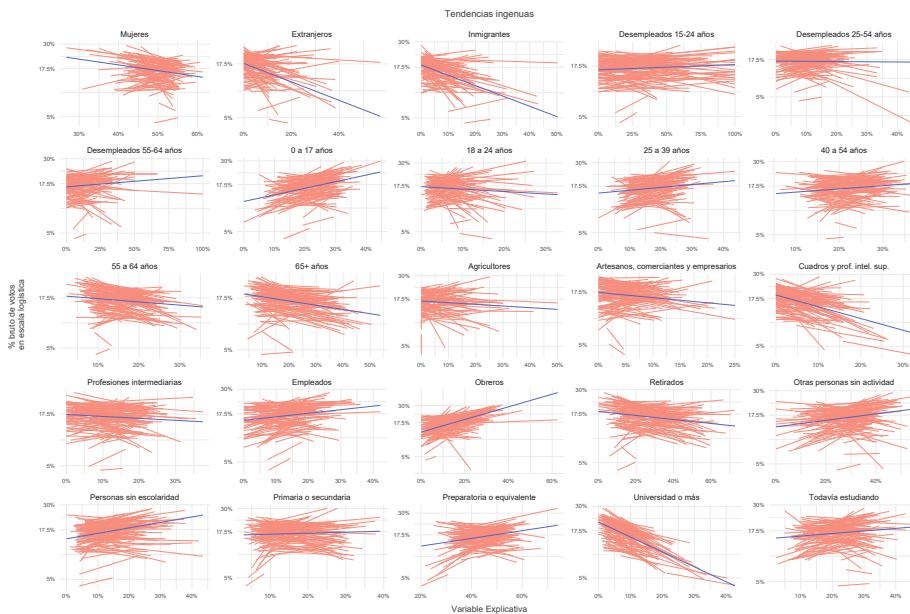


Figura 7.14: Tendencias ingenuas para la muestra entera y por departamento calculadas mediante ajustes lineales en la escala logística. Fuente: elaboración propia.

Efectivamente confirmamos que existirían indicios de asociaciones negativas para las categorías de Mujeres, Extranjeros, Inmigrantes y personas con diploma universitario. Pero también observamos pendientes ligeramente negativas para los grupos de edades 18 a 24, 55 a 64 y más de 65 años. Algunas categorías socioprofesionales como agricultores, artesanos, comerciantes y empresarios, cuadros y profesiones intelectuales superiores, profesiones intermedias y retirados parecen tener también asociaciones negativas. Salvo quizás los casos del desempleo juvenil (15-24 años) y general (25-54 años) o las personas con primaria y secundaria, el resto de las categorías muestran tendencias positivas.

Ahora bien, las tendencias generales no son homogéneas a través de los departamentos. Las líneas rosas muestran una gran variabilidad tanto en términos de interceptos

como de pendientes. Por ejemplo, aunque los desempleados de entre 25 y 54 años parecerían no tener ninguna relación tomando en cuenta todas las comunas, observamos algunas líneas rosas con claras pendientes. Quizás la categoría poblacional que muestra el comportamiento más homogéneo a través de los departamentos sea la población altamente escolarizada con diplomas de universidad y más.

Otro punto importante a notar es que estas tendencias ingenuas solamente se grafican para los rangos observados de los datos. Es decir, no se extrapolan las tendencias para todo el intervalo del $[0, 1]$. Esto es importante mencionarlo porque a la hora de analizar los modelos podría caerse en la tentación de querer predecir porcentajes de votos para valores de una categoría de alguna variable explicativa a lo largo de todo el intervalo. Extrapolar de esta manera es riesgoso y, además, en algunos casos podríamos estar cayendo en una falacia ecológica si tomáramos el valor de 1— es decir toda la población comunal perteneciente a la misma categoría— . Más aún, podríamos caer en situaciones ilógicas como hablar de votación emitida por individuos sin derecho al voto como los menores de edad o los extranjeros. No debemos perder de vista que el objetivo del análisis es la identificación de *configuraciones sociales* que favorecieron el voto, no la inferencia sobre qué individuos emitieron los sufragios.

Con este análisis exploratorio en mente, procedo a modelar los datos electorales franceses de 2012 mediante regresiones logísticas con la hipótesis de que se obtendrían mejores resultados con una estrategia jerárquica en lugar de agregación completa de los datos.

Capítulo 8

Modelado

En este capítulo desarollo el proceso de modelado de los datos franceses con tal de explorar las configuraciones sociales que favorecieron o inhibieron el voto por el FN en 2012. Para ello hay que notar que los datos electorales que considero son el número de votos por Marine Le Pen en cada comuna así como el número de inscritos en el listado nominal de las mismas. Tenemos entonces C pares de la forma $\{y_c, n_c\}$ donde y_c representa el número de votos y n_c el de inscritos en la comuna c . Este tipo de datos puede ser modelado como proveniente de una distribución binomial con número de ensayos conocido y parámetro de interés p_c :

$$y_c|p_c \sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C$$

Podemos interpretar cada parámetro p_c como la afinidad que se tuvo en la comuna c por Marine Le Pen en la primera vuelta presidencial de 2012. Esta afinidad es la que quisiéramos explicar en términos de configuraciones sociales.

Recordando la presentación de la regresión logística en la **Subsección 5.2.1**, si tenemos un vector x_c de variables explicativas en la comuna c , construimos un MLG de la siguiente manera:

$$y_c|\theta \sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C$$
$$\text{con } \ln\left(\frac{p_c}{1-p_c}\right) = \alpha + \beta x_c$$

$$\text{y } \theta = (\alpha, \beta) \sim f(\theta)$$

En nuestro caso, sin pérdida de generalidad, para la m -ésima variable explicativa tenemos un vector de proporciones $x_c = (x_{1,c}, \dots, x_{l_m,c})$ donde l_m es el número de categorías de la variable y $x_{j,c} = 1 - \sum_{k \neq j} x_{k,c}$ para toda $j \in \mathbb{N}_{l_m}$. Recordando el problema de multicolinealidad que esto ocasionaría al considerar la regresión con intercepto— Subsección 5.1.1— podemos definir una restricción de identificabilidad de suma cero para los coeficientes, de manera tal que $\beta_j = - \sum_{k \neq j} \beta_k$.

Para ilustrarlo, si tomáramos como variable explicativa la composición por edad de la población comunal, tendríamos $x_c = (Ed1_c, \dots, Ed6_c)$ donde Edj_c es la proporción de habitantes del grupo de edad j . En este caso, $\beta_6 = - \sum_{k=1}^5 \beta_k$, por lo que en realidad solo tenemos libres el intercepto y 5 coeficientes a la hora de asignar la distribución inicial $f(\theta)$.

$$y_c | \theta \sim \text{Binom}(n_c, p_c) \quad \forall \quad c \in \mathbb{N}_C$$

con $\ln\left(\frac{p_c}{1-p_c}\right) = \alpha + \beta_1 Ed1_c + \dots + \beta_6 Ed6_c$ tal que $\beta_6 = - \sum_{k=1}^5 \beta_k$

$$\text{y } \theta = (\alpha, \beta_1, \dots, \beta_5) \sim f(\theta)$$

8.1. Distribuciones inciales

¿Cómo asignamos distribuciones iniciales? Primero hay que tomar en cuenta que las distribuciones normalmente consideradas como mínimo informativas tienen implicaciones particulares en el contexto de la escala logística. Supongamos que quisiéramos asignar una distribución inicial al valor del predictor lineal $\eta = \alpha + \beta X$. Una distribución $N(\mu = 0, \sigma^2 = 100)$ sería normalmente considerada como una distribución mínimo informativa. Tendría una desviación estándar de 10, por lo que poco más del 95 % de sus observaciones oscilarían entre -20 y 20 . ¿Qué implicación tiene esto para una proporción p vinculada con η mediante la liga logística? Es decir, si asignamos esa distribución inicial a $\eta = \ln\left(\frac{p}{1-p}\right)$, ¿qué estamos diciendo sobre p ? Podemos observarlo en el primer histograma de la **Figura 8.1**. Simulando observaciones de $\eta \sim N(\mu = 0, \sigma_1 = 10)$ vemos que la mayoría de ellas llevan a valores extremos de p cercanos a 0 o a 1. Incluso

repetiendo el ejercicio para desviaciones estándar menores como $\sigma_2 = 5$ o $\sigma_3 = 2.5$, la distribución inicial lleva a un histograma para p en forma de U. Tendríamos que tener desviaciones estándar más pequeñas, como $\sigma_4 = 1.5$, o una normal estándar con $\sigma_5 = 1$ para estar asignando una distribución poco informativa para p .

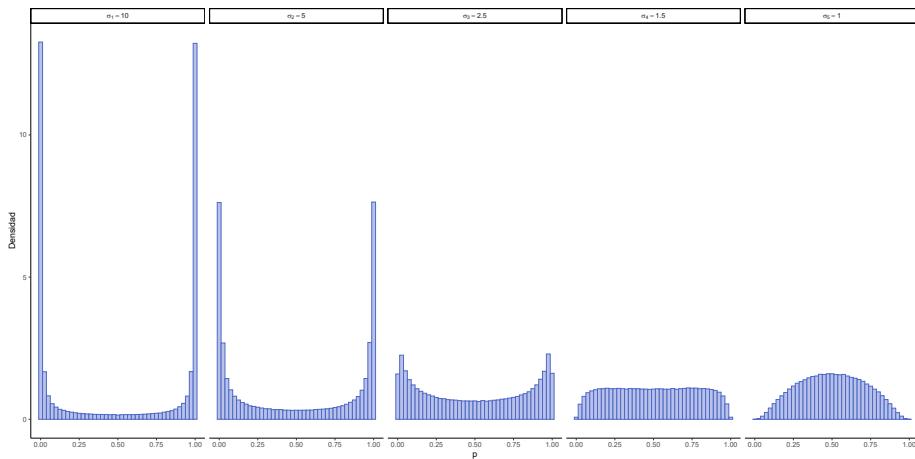
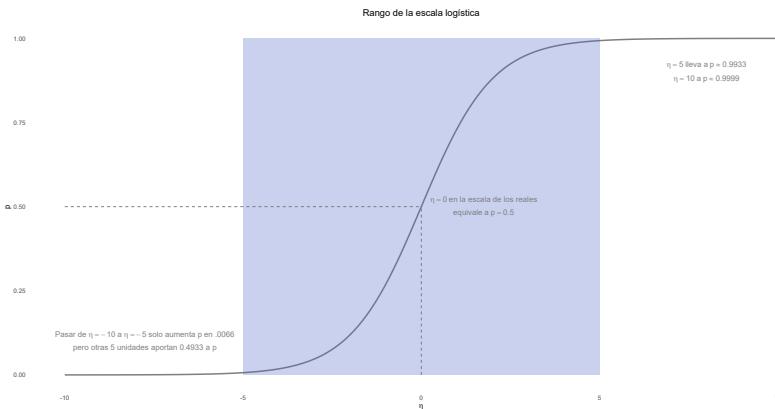


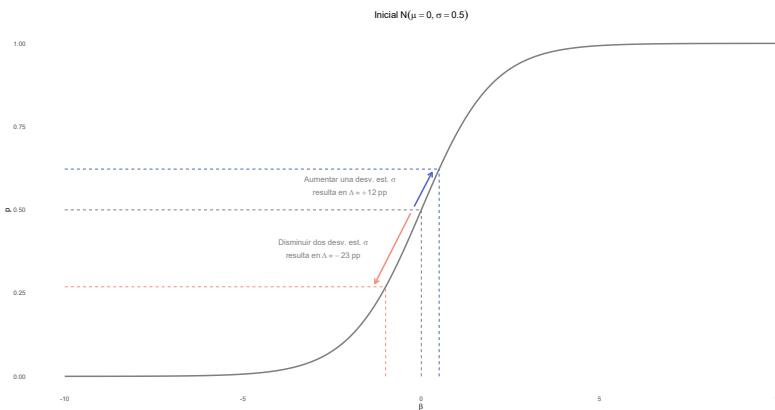
Figura 8.1: Ejemplo de implicaciones en p con distribuciones iniciales para un predictor lineal. Todas son normales centradas en 0 con distintos parámetros σ de desviación estándar. Fuente: elaboración propia.

El motivo es que la función logística no es lineal y tiene que “compactar” valores en los reales dentro del espacio confinado del $(0, 1)$. Por decirlo de una manera informal, el verdadero rango de variación de la escala logística son valores de η en $(-5, 5)$; valores fuera de este intervalo llevan prácticamente a los mismos valores extremos cercanos a 0 o a 1, como apreciamos en la **Figura 8.2**.

Ahora bien, nosotros no asignaremos distribuciones iniciales para el predictor lineal en su conjunto, sino para los coeficientes β de dicho predictor lineal. Optaré aquí por una posición un poco escéptica o conservadora. A pesar de lo que sugerían las tendencias ingenuas del capítulo anterior, *a priori*, no buscaría asignarle un sentido al efecto. Asimismo, los efectos de variables ecológicas o agregadas, no deberían ser tan grandes. No querría “inflar” la magnitud del valor explicativo mediante la distribución inicial pero tampoco negar la posibilidad de que existan algunos efectos importantes. Considero que una inicial para el coeficiente $\beta \sim N(\mu = 0, \sigma = 0.5)$ cumple relativamente bien con estas condiciones. Si tomamos el 0 como referencia y comparamos el valor correspondiente de

**Figura 8.2:** Fuente: elaboración propia.

p con aquel para una β una desviación estándar mayor— i.e. $\beta = 0.5$ — el efecto Δ sería de aproximadamente +12 puntos porcentuales. Si ahora disminuimos β en 2 desviaciones estándar, $\Delta \approx -23$ pp, como puede apreciarse en la **Figura 8.3**. Sin exagerar su interpretación, pues hay que multiplicar por el valor de la variable explicativa, estos podrían pensarse como los máximos efectos creíbles *a priori*.

**Figura 8.3:** Fuente: elaboración propia.

Por otro lado, debemos detenernos a pensar en la interpretación del intercepto α . Por un lado, α es el valor que tomaría el predictor lineal η_c , si todos los coeficientes fueran iguales a 0. Esto nos podría hacer pensar que querríamos que *a priori* su distribución nos lleve a valores cercanos al 17.9 %, porcentaje que obtuvo Le Pen en la elección. Sin

embargo, también hay que notar que con variables explicativas de configuraciones sociales, x_c , α es el valor que tomaría η_c si la población estuviera repartida equitativamente entre todas las categorías. En efecto, supongamos que para la m -ésima variable con l_m categorías cada $x_{j,c} = l_m^{-1}$. Tendríamos que

$$\eta_c = \alpha + \sum_{j=1}^{l_m} \beta_j x_{j,c} = \alpha + \sum_{j=1}^{l_m} \frac{\beta_j}{l_m} = \alpha + \frac{1}{l_m} \sum_{j=1}^{l_m} \beta_j$$

y por la restricción de identificabilidad de suma cero de los coeficientes,

$$= \alpha + \frac{1}{l_m} (0) = \alpha$$

Tomando en cuenta las teorías del conflicto discutidas en la revisión de literatura, podríamos pensar que si en lugar de tener grupos mayoritarios frente a grupos minoritarios hubiese una sociedad más “equilibrada”, el voto frontista disminuiría. Así pues, deberíamos buscar una distribución inicial con una media menor al 17.9 % obtenido en las elecciones. También, considerando que puede ser más robusto dilucidar una distribución inicial con base en cuantiles, más que igualar la media podríamos intentar aproximar el rango intercuartílico observado de entre 13.54 % y 21.44 %, pero sesgado un poco a la baja para tomar en cuenta la hipótesis anterior sobre una población equilibrada. Después de algunas pruebas elegí una distribución inicial para $\alpha \sim N(\mu = -1.7, \sigma = 0.25)$.

Consideremos ahora una “comuna promedio”, definida como aquella que tuviera valores promedio en las variables explicativas; podemos realizar simulaciones de la distribución predictiva con estas iniciales. Es decir, simulamos de $\alpha \sim N(\mu = -1.7, \sigma = 0.25)$ y para cada $\beta \sim N(\mu = 0, \sigma = 0.5)$. Calculamos el predictor lineal con base en los valores promedio de cada categoría—considerando también la restricción de suma cero de los coeficientes—y tomamos el logit inverso. Este proceso, por ejemplo para grupos de edad, nos lleva al histograma de la **Figura 8.4**, con el modelo

$$y_c | \theta = (\alpha, \beta) \sim \text{Binom}(n_c, p_c) \quad \forall \quad c \in \mathbb{N}_C$$

con $\ln \left(\frac{p_c}{1 - p_c} \right) = \alpha + \beta_1 Ed1_c + \cdots + \beta_6 Ed6_c$ tal que $\beta_6 = - \sum_{k=1}^5 \beta_k$

$$\begin{aligned} y - \alpha &\sim N(\mu = -1.7, \sigma = 0.25) \\ \beta_j &\sim N(\mu = 0, \sigma = 0.5) \quad j = 1, \dots, 5 \end{aligned}$$

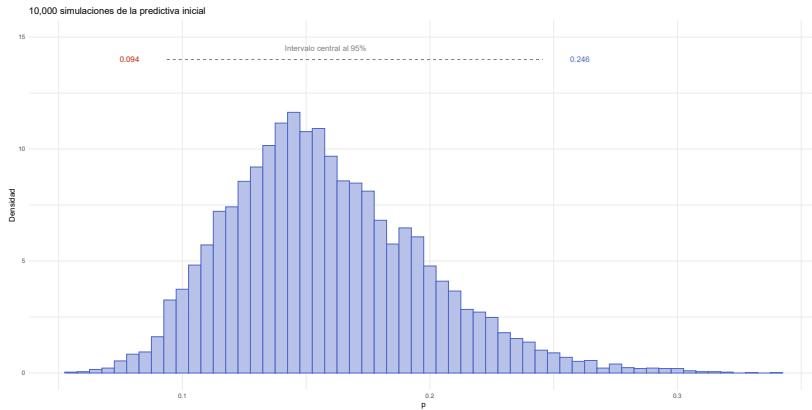


Figura 8.4: Fuente: elaboración propia.

Estas distribuciones iniciales, como ya he mencionado, son subjetivas. Alguien más podría no estar muy convencido de la elección que realizó, por lo que eran importantes los párrafos anteriores de manera que fuera más transparente el proceso por el cual llegó a elegirlas.

8.2. Modelos individuales

Consideremos entonces, para cada variable explicativa, modelos de la siguiente forma:

$$\begin{aligned} y_c | \theta = (\alpha, \beta) &\sim Binom(n_c, p_c) \quad \forall \quad c \in \mathbb{N}_C \\ \text{con} \quad \ln \left(\frac{p_c}{1 - p_c} \right) &= \alpha + \sum_{j=1}^{l_m} \beta_j x_{j,c} \quad \text{tal que} \quad \beta_{l_m} = - \sum_{k=1}^{l_m} \beta_k \\ y - \alpha &\sim N(\mu = -1.7, \sigma = 0.25) \\ \beta_j &\sim N(\mu = 0, \sigma = 0.5) \quad j \in \mathbb{N}_{l_m-1} \end{aligned} \tag{8.1}$$

A (8.1) le llamaré un **Modelo Nacional Individual** porque se tienen coeficientes

a nivel nacional¹ y solo incluyo una variable explicativa de manera individual. Tendría entonces 9 modelos nacionales individuales de esta forma. Sin embargo, el análisis exploratorio de datos sugería que se podría modelar de mejor manera dejando que los coeficientes e interceptos variasen para cada uno de los 96 departamentos. Es decir, podríamos construir igualmente 9 **Modelos Jerárquicos Individuales**. Si denotamos como $d[c]$ el *departamento* al que pertenece la comuna c , el modelo jerárquico individual sería:

$$\begin{aligned}
 y_c | \theta &\sim \text{Binom}(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\
 \text{con} \quad \ln\left(\frac{p_c}{1 - p_c}\right) &= \alpha_{d[c]} + \sum_{j=1}^{l_m} \beta_{d[c], j} x_{j,c} \\
 \text{tal que} \quad \beta_{d, l_m} &= - \sum_{k=1}^{l_m} \beta_{d, k} \\
 \alpha_d &\sim N(\mu_\alpha, \sigma = 1) \quad \forall d \in \mathbb{N}_{96} \\
 \beta_{d, j} &\sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_{l_m - 1} \quad \text{y} \quad d \in \mathbb{N}_{96} \\
 \mu_\alpha &\sim N(-1.7, \sigma = 0.25) \\
 \mu_\beta &\sim N(0, \sigma = 0.5)
 \end{aligned} \tag{8.2}$$

Se corrieron cada uno de los 18 modelos mediante el software Stan, simulando vía *Hamiltonian Monte Carlo* las distribuciones posteriores dada la muestra de datos discutida en la Sección 7.4. Una vez con dichas distribuciones posteriores, podemos hacer lo que Gelman y col. (2013) llaman *posterior predictive checks*. Podemos predecir el porcentaje esperado de votos en cada una de las comunas y calcular su error respecto al real. Para el modelo nacional de ocupación laboral juvenil, los resultados están en la **Figura 8.5**.

En el mapa de la izquierda las predicciones se colorean de acuerdo a la escala real que va de 0 % a 62.5 % y donde el cambio de tonos rojos a azules se da en la comuna mediana de 17.5 %. Observamos que la predicción en general subestima el verdadero porcentaje obtenido pues el mapa está prácticamente coloreado en su totalidad de tonos rojos. Sin embargo, las predicciones están cerca de la mediana. De hecho, el mapa de errores—verde significa poco error y naranjas y rojos más—visualmente es prácticamente un ne-

¹En realidad, como mencionaba en el análisis exploratorio de datos, son coeficientes para la metrópoli y no para todo el país.

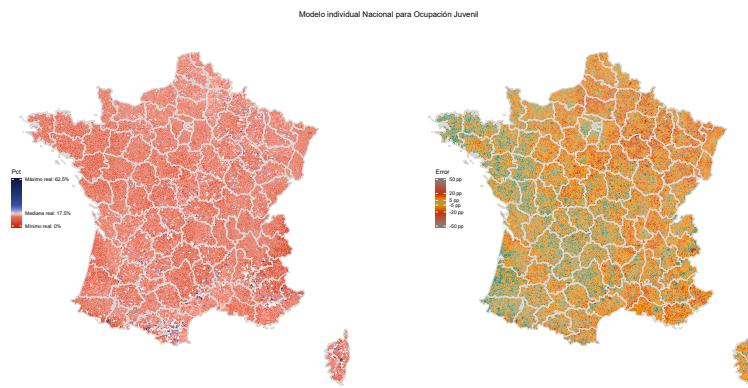


Figura 8.5: Mapa de predicciones medias del porcentaje bruto de votos obtenido por Marine Le Pen en las presidenciales 2012 mediante el Modelo Nacional por Ocupación Juvenil y mapa de los respectivos errores. Fuente: elaboración propia.

gativo del verdadero mapa de resultados que observábamos en la **Figura 7.4** del capítulo anterior.

En general, los modelos nacionales tienen el defecto de no reconocer la enorme variabilidad geográfica del fenómeno electoral. Si construimos distintas medidas de error de predicción vemos que pasar de un modelado nacional a un modelado jerárquico las reduce consistentemente. Para cada simulación posterior predecimos el número de votos en cada comuna, departamento y a nivel nacional. Luego lo convertimos en porcentaje bruto de votos predicho dividiendo entre el número de inscritos en la comuna. Así, podemos tomar los conocidos errores absoluto y cuadrático promedio para el porcentaje de votos a nivel comuna, departamento y nacional. Incluso podemos tomar una pérdida más arbitraria, pero ilustrativa, como el porcentaje promedio de estimaciones que se encuentran a más de 1.5 puntos porcentuales del verdadero valor en los 3 niveles de agregación; a esta medida la llamo tolerancia 1.5 pp. Finalmente se calcula el promedio de las medidas a través del total de simulaciones posteriores y se grafican en la **Figura 8.6**. El gráfico también incluye el cálculo de un cuarto error llamado WAIC para las comunas, pero a él me refiriré más adelante. En el gráfico, los puntos son las medidas para el modelo nacional y las flechas las de los modelos jerárquicos.

Como es de esperarse, los modelos estiman de mejor manera los porcentajes de niveles de mayor agregación que el del nivel comunal. Viendo la pérdida arbitraria llamada

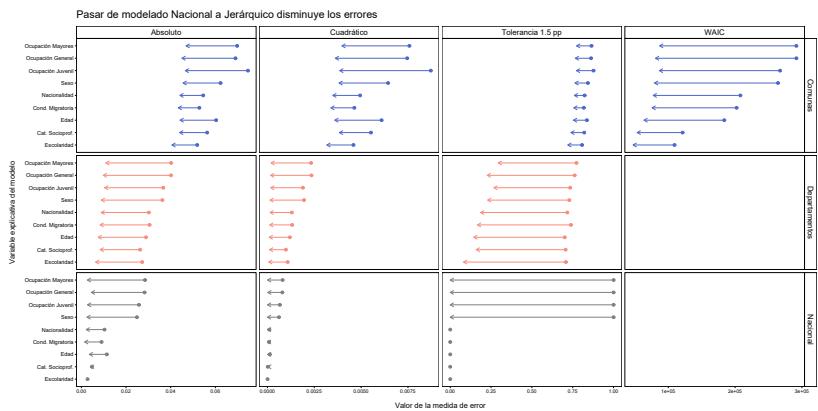


Figura 8.6: Comparación de los modelos nacionales y jerárquicos individuales bajo diferentes medidas de error. Fuente: elaboración propia.

de tolerancia 1.5, salvo las variables de sexo y ocupaciones, todas las predicciones estiman un porcentaje de votos nacional a menos de 1.5 puntos porcentuales del real. Sin embargo, vemos que para los 96 departamentos, los modelos nacionales solo logran esta precisión esperada de estimación en alrededor de 70. Esto refleja lo que había mencionado al ver el mapa de predicciones para el modelo nacional individual por ocupación juvenil. Al no permitir que los coeficientes varíen por departamento, el modelo ajusta para el país entero, sacrificando las estimaciones en las diferentes zonas geográficas que estos constituyen. Si observamos medidas de error más tradicionales como la absoluta o la cuadrática vemos que todas las flechas se dirigen a la izquierda, es decir que se reducen los errores al reconocer esa variabilidad geográfica. Ahora bien, estas 3 medidas tienen más un carácter ilustrativo y sirven para recordar que se pueden construir diferentes errores para diferentes estimadores.

Ahora notemos que las variables de solo 2 categorías como ocupación, sexo, nacionalidad y condición migratoria tienen peor desempeño que las de más categorías como edad, categoría socioprofesional y escolaridad. ¿Esto quiere decir que son peores variables explicativas? ¿No podría esto deberse a que las últimas tienen más parámetros y es más fácil ajustar mejor simplemente por introducir parámetros adicionales?

8.2.1. WAIC

Pensando en la posibilidad de mejorar un ajuste simplemente haciendo más complejo el modelo es que una mejor y más aceptada medida de error es el llamado WAIC por sus siglas en inglés *Widely Applicable* o *Watanabe-Akaike Information Criterion* (Vehtari, Gelman y Gabry 2016). El WAIC busca estimar la capacidad predictiva de un modelo vía las predictivas posteriores de los datos con los que se ajusta. La intuición es que a mayor valor de la predictiva posterior para el dato observado, mayor la posibilidad de predecir otros datos.

Para un conjunto de n datos $y = (y_1, \dots, y_n)$ y una muestra posterior de parámetros $\{\theta_{(s)}\}_{s=1}^S$, el WAIC se define de la siguiente manera:

$$\begin{aligned} WAIC(y|\theta) &= \widehat{lpd} + \widehat{par} \\ \text{con } \widehat{lpd} &= \sum_{i=1}^n \ln \left(\frac{1}{S} \sum_{s=1}^S f(y_i|\theta_{(s)}) \right) \\ \text{y } \widehat{par} &= \sum_{i=1}^n V(y_i|\theta), \end{aligned} \tag{8.3}$$

$$\text{donde } V(y_i|\theta) = \frac{1}{S-1} \sum_{s=1}^S \mathcal{L}(y_i; \theta_s)^2 \text{ y } \mathcal{L}(y_i; \theta_s) = \ln [f(y_i|\theta_{(s)})] - \frac{1}{S} \sum_{s=1}^S \ln [f(y_i|\theta_{(s)})].$$

El WAIC se conforma de dos sumas. La primera, \widehat{lpd} , es una aproximación de la log predictiva posterior, es decir una medida de ajuste. Por otro lado, \widehat{par} es normalmente llamado el *número efectivo estimado de parámetros* y es una medida que mediante sumas de varianzas busca medir la complejidad del modelo. En total, entre menor sea el valor del WAIC, esperaríamos un mejor desempeño predictivo tomando en cuenta la complejidad del modelo.

Notemos que con la definición dada, el WAIC es un estimador por lo que tiene una distribución muestral. Gracias al paquete *loo* de R, podemos estimar tanto los WAICs y sus errores estándar como la diferencia esperada entre dos de ellos y, argumentando un teorema central del límite, la probabilidad de que el WAIC de un modelo sea efectivamente menor o igual al WAIC de otro. Estas comparaciones las observamos en la **Figura COMPARA WAICS**.

Para la mayoría de las comparaciones es claro que un modelo tiene mejor ajuste que el otro. Podríamos pensar que efectivamente hay un ordenamiento en el poder predictivo de las distintas variables: escolaridad, categorías socioprofesionales, edad, condición migratoria, nacionalidad, sexo y, finalmente, las ocupaciones. Sin concluir todavía nada, parecería que a las hipótesis sobre inseguridad laboral no les corresponde el mejor de los poderes explicativos en términos de estos modelos.

8.3. Modelos Compuestos

Este ordenamiento preliminar tiene el objetivo de ayudarme a construir, secuencialmente, un modelo cada vez más complejo. Para comenzar, veamos los mapas de predicción que generó el modelo con menor WAIC, el jerárquico por escolaridad, en la **Figura 8.7**.

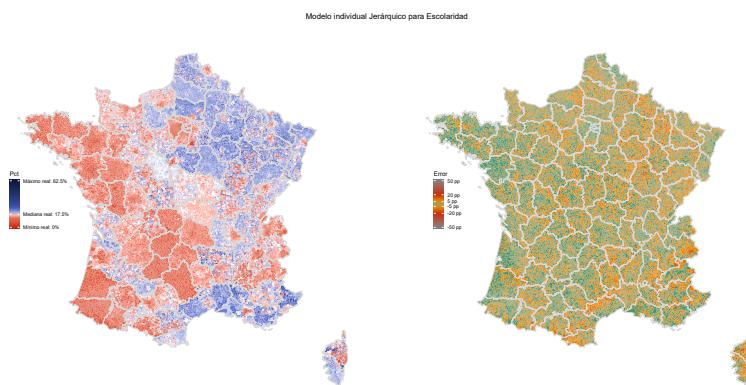


Figura 8.7: Mapa de predicciones medias del porcentaje bruto de votos obtenido por Marine Le Pen en las presidenciales 2012 mediante el Modelo Jerárquico por Escolaridad y mapa de los respectivos errores. Fuente: elaboración propia.

El mapa de errores ya se observa más verde y homogéneo. Las grandes zonas de fortaleza y debilidad del FN comienzan a ser identificadas por el modelo. Sin embargo, los tonos de algunas predicciones son todavía demasiado cercanos a la mediana. En el centro observamos una zona de tonos muy claros, casi blancos, que no corresponden totalmente a la realidad. En general, vemos que hay menor variabilidad dentro de los departamentos de la que se observó en la elección.

Estas reflexiones sugieren que, en lugar de tomar una a una las variables, podemos ir agregando variables a la regresión en modelos jerárquicos secuenciales que refinen el ajuste. Entonces, comenzando con la escolaridad, iré agrandando el modelo con la inclusión de una nueva variable cada vez.

El primer modelo compuesto incluye las variables de escolaridad y categorías socioprofesionales. A este modelo lo llamaré el **Modelo A**. Para distinguir las variables utilizaré diferentes letras griegas para sus coeficientes. Para la configuración social de escolaridad

$$x_{escol,c} = (Esc_c, Dip1_c, Dip2_c, Dip3_c, Dip4_c)^T$$

en la comuna c los coeficientes departamentales serán un vector $\beta_{d[c]} = (\beta_{d[c],1}, \dots, \beta_{d[c],5})$. Para las categorías socioprofesionales $x_{csp,c}$ los coeficientes serán $\gamma_{d[c]}$:

$$\begin{aligned} y_c | \theta &\sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\ \text{con} \quad ln\left(\frac{p_c}{1-p_c}\right) &= \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} \end{aligned} \tag{8.4}$$

tal que

$$\begin{aligned} \beta_{d,5} &= - \sum_{k=1}^4 \beta_{d,k}, \\ \gamma_{d,8} &= - \sum_{k=1}^7 \gamma_{d,k} \end{aligned}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\begin{aligned} \alpha_d &\sim N(\mu_\alpha, \sigma = 1) \\ \beta_{d,j} &\sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4 \\ \gamma_{d,j} &\sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7 \end{aligned}$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

Al estimar el modelo con ambas variables las predicciones efectivamente mejoran. La diferencia en WAIC entre el modelo A y el modelo jerárquico por escolaridad se estima en **AGREGAR ESTIMADO CON ERROR ESTÁNDAR**. Asimismo, observando los mapas de la **Figura 8.8**, vemos que se reduce la gran zona blanquiza del centro, se comienza a observar mayor variabilidad dentro de los departamentos y los tonos también se oscurecen más, sobre todo en el noreste.

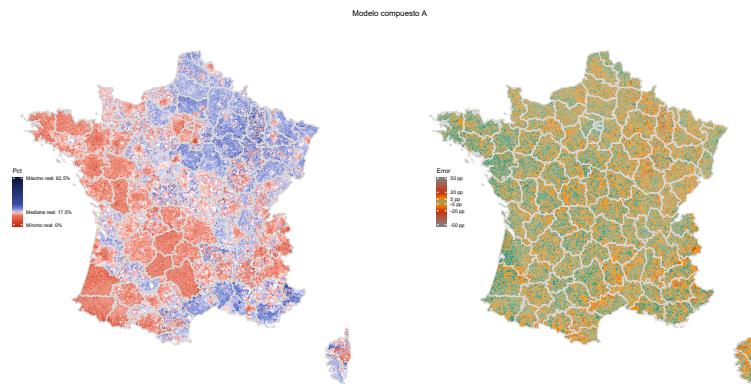


Figura 8.8: Mapas de predicciones medias y respectivos errores para el porcentaje bruto de votos obtenido por Marine Le Pen en las presidenciales 2012 mediante el Modelo Jerárquico por Escolaridad y Categorías socioprofesionales. Fuente: elaboración propia.

Los siguientes modelos se llamarán de manera progresiva por letras latinas y sus planteamientos en términos de ecuaciones pueden encontrarse en el **Apéndice B**. Por ejemplo, el **Modelo B** incorpora los grupos de edad $x_{edad,c}$. Al hacerlo, el WAIC vuelve a disminuir y los respectivos mapas corresponden a la **Figura 8.9**.

Debido a la similitud entre condición migratoria y nacionalidad, solamente buscaba incorporar una de las dos, tratando de identificar la que tuviera mejor desempeño. De acuerdo al ordenamiento previo en los modelos individuales, consideraré solo la variable migratoria $x_{migr,c}$ dentro del **Modelo C**. Por su parte, el **Modelo D** incluye la distribución comunal por sexo $x_{sexo,c}$.

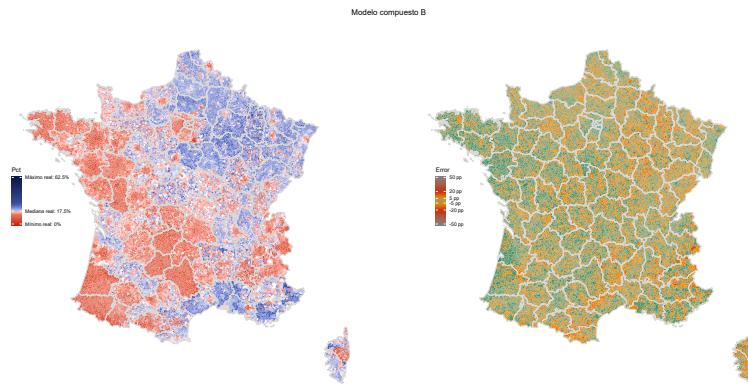


Figura 8.9: Mapas de predicciones medias y respectivos errores para el porcentaje bruto de votos obtenido por Marine Le Pen en las presidenciales 2012 mediante el Modelo Jerárquico por Escolaridad, Categorías socioprofesionales y Edad. Fuente: elaboración propia.

Después comenzaríamos a incluir las variables de ocupación. Como mencionaba al presentarlas, tengo un interés especial en la (des)ocupación juvenil, pues esta sería una variable que referencias como Le Bras (2016) y Perrineau (2007) favorecerían. Por ello, al margen del ordenamiento, vamos a construir dos modelos de 6 variables. De manera general ambos incorporarían una variable explicativa $x_{ocu,c}$. El **Modelo E** es el modelo D más la ocupación juvenil, mientras que el **Modelo F** es el modelo D más la ocupación general. Una vez generados por separado, podemos considerar un modelo que incorpore ambas variables, este sería el **Modelo G**. Finalmente agregamos la última variable considerada, la ocupación para personas de 55 a 64 años $x_{ocu.may,c}$, para obtener el **Modelo H**.

¿Cuál es la comparación de WAICs entre ellos? En la **Figura 8.10** podemos observarlo. En general, el WAIC mejora conforme se van agregando variables. Ciertamente las mayores ganancias se dan al agregar las primeras variables que el análisis individual sugería eran las más explicativas. Esto parece confirmar dicha hipótesis. Por el contrario, en términos de WAIC, la ocupación juvenil no parece ser la más poderosa de las variables. En efecto, la ganancia en WAIC al agregarla en el modelo E, es menor a la ganancia del modelo F. Más aún, al pasar del modelo F al modelo G agregándola de nuevo, la ganancia vuelve a ser poca. **Esto no quiere decir, sin embargo, que no aporte nada a la regresión. Aunque la mejora en WAIC sea pequeña y los intervalos con error estándar se traslapen, la probabilidad de que agregarla produce un mejor modelo es de INSERTAR PROBAS.**

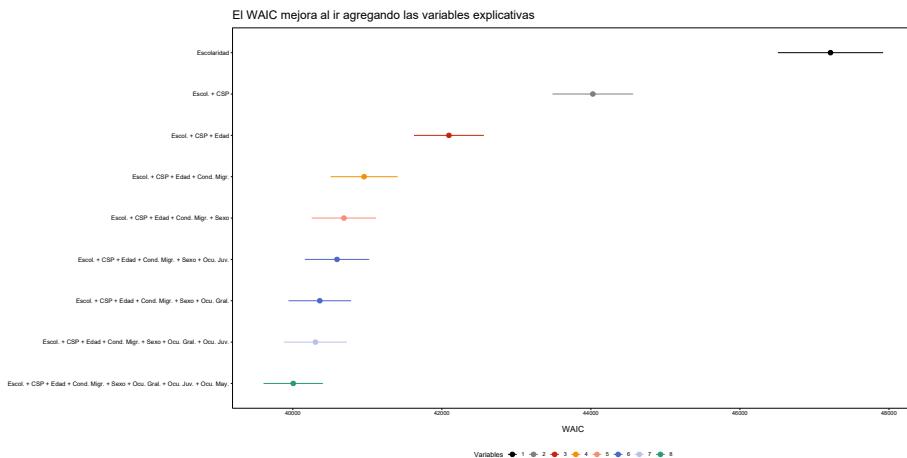


Figura 8.10: Comparativo de WAIC para distintos modelos. Comenzando por los modelos individuales en color rojo, ir agregando una variable adicional disminuye el WAIC. El Modelo de 5 variables ya no parece mejorar tanto. Fuente: elaboración propia.

8.3.1. Convergencia de HMC

Una vez ajustados los modelos, como adelantaba en **CONVERGENCIA**, habría que verificar que las cadenas construidas mediante la implementación de HMC dentro de Stan hubieran convergido. Debido a la cantidad de parámetros dentro de cada modelo, así como a la cantidad de modelos ajustados en sí, es difícil verificar a detalle todas y cada una de las muestras posteriores. Afortunadamente, Betancourt presenta un caso de estudio que permite realizar distintos diagnósticos útiles para todos los parámetros de interés dentro de todos los modelos. Entre ellos encontramos el factor de reducción de escala \hat{R} discutido en **CONVERGENCIA**, así como un cálculo del tamaño efectivo de muestra por iteración y 3 diagnósticos particulares de HMC. Utilizando su código abierto², verificamos que los modelos satisfacen los criterios planteados y podemos confiar en la convergencia de las muestras posteriores obtenidas. Esta comprobación puede reproducirse con el código del repositorio de Github de esta tesis y solicitándome acceso a los archivos .rds con los ajustes de todos los modelos.

Adicionalmente, podemos observar algunos diagnósticos gráficos para algunos parámetros del modelo H. Por ejemplo, en la **Figura 8.11a** observamos en los *traceplots* para algunos parámetros que hay una buena mezcla de las cadenas; esto también se confirma

²Los derechos de autor son de Michael Betancourt y la Universidad de Columbia y las licencias pueden verificarse en el link en las referencias.

viendo que las densidades por cadena de la **Figura 8.11b** son parecidas. Finalmente, en la **Figura 8.11c** vemos que las autocorrelaciones son pequeñas. De hecho, existe un poco de simulación antitética que permite una mejor eficiencia en el tamaño de muestra

CITAR DISCUSIÓN BLOG GELMAN.

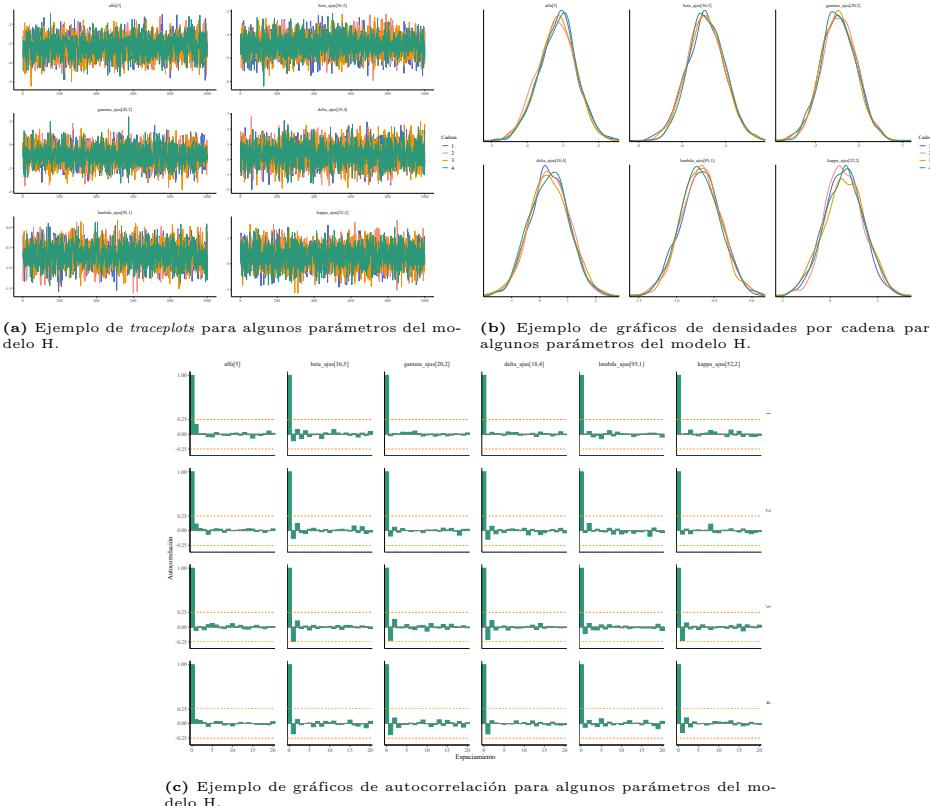


Figura 8.11: Fuente: elaboración propia.

También analicé la convergencia del modelo H mediante los diagnósticos del paquete shinystan, que permite de manera interactiva observar distintos gráficos y resúmenes respecto al ajuste del modelo vía HMC. De esta exploración, también concluí que el modelo satisfacía suficientemente bien la batería de diagnósticos para darnos confianza en que las muestras obtenidas provienen de la distribución posterior. Una vez verificada la convergencia, podemos proceder al análisis de los resultados.

8.3.2. Definiendo efectos

Interpretar los coeficientes de regresiones logísticas no es sencillo, pues estos se encuentran en la escala logística y muchas veces queremos interpretarlos en la escala de las proporciones o probabilidades que la liga logística define. Como explican Gelman y Hill (2006), además de las interpretaciones en términos de momios, un primer impulso a la hora de interpretar coeficientes es considerar el caso cuando la variable explicativa vale 1. Sin embargo, esto no siempre tiene sentido dependiendo del contexto en el que se realiza la regresión. Este es mi caso por lo que no podría interpretar los coeficientes mediante alguna estrategia de ese tipo, sobre todo para evitar alguna falacia ecológica—ver página 125—.

Por ello, resultan útiles lo que los mismos autores llaman *predictive comparisons*. Básicamente estos consisten en pensar en los efectos de los coeficientes en términos de una comparación entre dos valores distintos pero informativos de las variables explicativas de interés—en mi caso, las *configuraciones sociales* definidas por las categorías poblacionales de las variables censales—. Una alternativa de valores son la media y una desviación estándar por encima de ella. Esta elección tiene la ventaja de que, usualmente, se estaría realizando la interpretación en el rango de valores observado en la variable explicativa y no mediante una extrapolación. Hay que decir, no obstante, que una situación donde esto no sucede es cuando la variable explicativa observada es multimodal y la media cae en uno de los valles. Esta discusión general sobre alternativas de interpretación para regresiones logísticas puede consultarse en la referencia antes citada.

Por nuestra parte, para fijar ideas, supongamos por ahora que estamos analizando un modelo individual nacional para una variable con 3 categorías, $x = (x_1, x_2, x_3)$. Recordando (8.1) y omitiendo el subíndice de las comunas, tendríamos que para alguna comuna en particular,

$$\ln \left(\frac{p}{1-p} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Abusando de la notación, introduzco ahora unas expresiones. En primer lugar

$$\eta(\beta_i; x) = \alpha + \beta_i x_i + \sum_{k \neq i} \beta_k x_k,$$

para referirme al predictor lineal como función del coeficiente β_i , considerando la

configuración social x , el intercepto α y el resto de coeficientes como fijos. Asimismo tomemos

$$p(\beta_i; x) = \frac{e^{\eta(\beta_i; x)}}{1 + e^{\eta(\beta_i; x)}} = \frac{1}{1 + e^{-\eta(\beta_i; x)}}$$

como el valor de la afinidad por Marine Le Pen en la comuna como función del coeficiente β_i , de nuevo, con el resto del predictor lineal fijo. Buscaríamos entonces definir el efecto que tiene el coeficiente β_i como el cambio en puntos porcentuales para dicha afinidad cuando se pasa de una configuración a otra:

$$\Delta(\beta_i; x_{(0)}, x_{(1)}) = p(\beta_i; x_{(1)}) - p(\beta_i; x_{(0)})$$

Sin embargo, en nuestro caso no tenemos total libertad para elegir los valores $x_{(0)}$ y $x_{(1)}$ pues recordemos que son proporciones que deben sumar a 1. ¿Cómo elegir entonces los valores?

En primer lugar, definamos el punto de referencia $x_{(0)}$ como la media muestral, $x_{(0)} = \bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3)$. Posteriormente, como estamos interesados en el efecto del coeficiente β_i , empecemos a construir $x_{(1)}$ con base en $x_{i,(1)} = \bar{x}_i + s_i$ donde s_i es la desviación estándar muestral en la i -ésima categoría. Al agregarle una desviación estándar a x_i , para cumplir con la restricción debemos disminuir su valor del resto de x_k en su conjunto. Sin que sea la única forma de hacerlo, una propuesta es quitarle de manera proporcional a cada categoría restante una parte de la desviación estándar que sumamos a la categoría de interés. Esto sería de la siguiente manera:

$$x_{k,(1)} = \bar{x}_k - s_i \frac{\bar{x}_k}{1 - \bar{x}_i} \quad \forall k \neq i$$

Con este criterio podemos abusar un poco más de la notación. Si en lugar del caso específico de una configuración social de 3 categorías tenemos l categorías, el efecto sería

$$\Delta(\beta_i) = p(\beta_i; x^*) - p(\beta_i; \bar{x}) \quad \forall i \in \mathbb{N}_l \tag{8.5}$$

donde $\bar{x} = (\bar{x}_1, \dots, \bar{x}_l)$,

$$x^* = \begin{cases} \left(\bar{x}_1 + s_1, \bar{x}_2 - s_1 \frac{\bar{x}_2}{1 - \bar{x}_1}, \dots, \bar{x}_l - s_i \frac{\bar{x}_l}{1 - \bar{x}_1} \right) & i = 1, \\ \left(\bar{x}_1 - s_i \frac{\bar{x}_1}{1 - \bar{x}_i}, \dots, \bar{x}_i + s_i, \dots, \bar{x}_l - s_i \frac{\bar{x}_l}{1 - \bar{x}_i} \right) & 1 < i < l, \\ \left(\bar{x}_1 - s_l \frac{\bar{x}_1}{1 - \bar{x}_l}, \dots, \bar{x}_{l-1} - s_l \frac{\bar{x}_{l-1}}{1 - \bar{x}_l}, \bar{x}_l + s_l \right) & i = l. \end{cases}$$

Ahora pensemos que querríamos calcular efectos para modelos jerárquicos. En este caso, simplemente reemplazamos el parámetro de interés por su equivalente jerárquico $\beta_{d,i}$ y tendríamos que calcular la media y la desviación estándar correspondiente dentro de cada departamento d : \bar{x}_d y $s_{d,i}$. Esto lo indicamos mediante los subíndices:

$$\Delta(\beta_{d,i}) = p(\beta_{d,i}; x_d^*) - p(\beta_{d,i}; \bar{x}_d) \quad \forall i \in \mathbb{N}_l \quad (8.6)$$

¿Pero qué pasa cuando tenemos más de una variable? En esta situación, la decisión que tomé fue buscar una comparación *ceteris paribus* en la que los sumandos del resto de variables explicativas del predictor lineal toman sus valores promedio en el departamento. Podemos también pensar que estamos modificando el intercepto con los sumandos del resto de categorías con valores promedio. Por ello, si asociamos a x con la variable de la categoría de interés, β con sus coeficientes y z con el resto del predictor lineal, una notación más general podría ser

$$\Delta(\beta_{d,i}; \bar{z}_d) = p(\beta_{d,i}; x_d^*, \bar{z}_d) - p(\beta_{d,i}; \bar{x}_d, \bar{z}_d) \quad \forall i \in \mathbb{N}_l \quad (8.7)$$

$$\text{donde } p(\beta_{d,i}; x_d^*, \bar{z}_d) = \frac{1}{1 + e^{-\eta(\beta_{d,i}; x_d^*, \bar{z}_d)}}$$

$$\text{y } \eta(\beta_{d,i}; x_d, \bar{z}_d) = \bar{z}_d + \beta_{d,i} x_{d,i} + \sum_{k \neq i} \beta_{d,k} x_{d,k}.$$

Por ejemplo, en el caso del modelo compuesto A que incluye a la escolaridad y a las categorías socioprofesionales, el efecto para una categoría de escolaridad dependería de $\bar{z}_d = \alpha_d + \sum_{j=1}^8 \gamma_{d,k} \bar{x}_{csp,k}$, donde no hay que olvidar que $\bar{x}_{csp,k}$ se calcula para el departamento d .

Hasta aquí ya construimos una propuesta de efecto comparando dos predicciones específicas del modelo “enchufando” un valor específico o un estimador puntual de los coeficientes. Sin embargo, nuestro proceso de modelado bayesiano mediante HMC, nos indica que tenemos incertidumbre sobre su valor reflejada mediante una muestra posterior

de valores simulados. Por lo mismo, siguiendo la única receta de la estadística bayesiana tendríamos que obtener la distribución posterior de los efectos, calculando (??) para cada simulación posterior $\beta_{d,i}^{(s)}$ y, correspondientemente, $\bar{z}_d^{(s)}$:

$$\Delta(\beta_{d,i}^{(s)}; \bar{z}_d^{(s)}) = p(\beta_{d,i}^{(s)}; x_d^*, \bar{z}_d^{(s)}) - p(\beta_{d,i}^{(s)}; \bar{x}_d, \bar{z}_d^{(s)}) \quad \forall i \in \mathbb{N}_l, s \in \mathbb{N}_S.$$

Una vez con esta distribución posterior de los efectos podemos continuar el análisis de los modelos mediante distintos resúmenes inferenciales como el efecto medio o mediano. También podríamos reportar intervalos centrales de probabilidad para al 95 %. Siguiendo la práctica común, cuando dichos intervalos contienen al 0, diríamos que la categoría poblacional no tiene un efecto significativo. Si, por el contrario, se encuentran en su totalidad por encima o por debajo del 0, diremos que se tiene un efecto positivo o negativo, respectivamente.

Anexos

Anexo A

Análisis bayesiano del modelo lineal normal

Para realizar un análisis bayesiano del modelo lineal normal requerimos especificar una distribución inicial para θ y, mediante el teorema de Bayes, actualizarla para obtener una distribución posterior dados los datos observados. Entonces, primero presento una manipulación de la función de verosimilitud para después ver algunas distribuciones iniciales frecuentemente utilizadas y, finalmente, realizar la actualización de las mismas dados los datos.

Verosimilitud

Siguiendo a Gutiérrez Peña (1998) y Congdon (2006), manipulemos la función de verosimilitud de la normal multivariada para facilitar la actualización mediante el teorema de Bayes. Observemos que:

$$\begin{aligned} f(y|\theta) &= \frac{1}{\sqrt{(2\pi)|\sigma^2\mathbb{I}_N|}} \exp \left\{ -\frac{1}{2}(y - X\beta)^T(\sigma^2\mathbb{I}_N)^{-1}(y - X\beta) \right\} \\ &\propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \right\} \end{aligned} \quad (\text{A.1})$$

En el análisis clásico o frecuentista, el estimador máximo verosímil para los coeficientes β es $b = (X^T X)^{-1} X^T y$. Podemos manipular los términos dentro de la exponencial en la

distribución normal con este estimador b :

$$\begin{aligned} y - X\beta &= y - Xb + Xb - X\beta = (y - Xb) + X(b - \beta) \\ \Rightarrow (y - X\beta)^T(y - X\beta) &= \left\{ (y - X\beta)^T + [X(b - \beta)]^T \right\} \left\{ (y - Xb) + X(b - \beta) \right\} \\ &= (y - Xb)^T(y - Xb) + (y - Xb)^T X(b - \beta) + \\ &\quad [X(b - \beta)]^T(y - Xb) + [X(b - \beta)]^T X(b - \beta) \end{aligned}$$

y, agrupando los términos cruzados en $k(y, \beta)$,

$$\Rightarrow (y - X\beta)^T(y - X\beta) = (y - Xb)^T(y - Xb) + (b - \beta)^T X^T X(b - \beta) + k(y, \beta). \quad (\text{A.2})$$

En realidad, $k(y, \beta) = 0$:

$$k(y, \beta) = (y - Xb)^T X(b - \beta) + [X(b - \beta)]^T(y - Xb)$$

notando que el segundo término es igual al primero pero transpuesto,

$$(y - Xb)^T X(b - \beta) = (y^T - b^T X^T)(Xb - X\beta)$$

sustituyendo el valor de b y considerando que $Xb = y$

$$\begin{aligned} (y - Xb)^T X(b - \beta) &= \left\{ y^T - [(X^T X)^{-1} X^T y]^T X^T \right\} (y - X\beta) \\ &= \left\{ y^T - [y^T X(X^T X)^{-T}] X^T \right\} (y - X\beta) \\ &= [y^T - y^T X(X^{-1} X^{-T}) X^T] (y - X\beta) \\ &= (y^T - y^T)(y - X\beta) \end{aligned}$$

entonces,

$$(y - Xb)^T X(b - \beta) = 0 \implies k(y, \beta) = 0.$$

Podemos entonces sustituir (A.2) con $k(y, \beta) = 0$ en (A.1):

$$f(y|\theta) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} [(y - Xb)^T(y - Xb) + (b - \beta)^T X^T X(b - \beta)] \right\}$$

$$\propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} [(y - Xb)^T(y - Xb) + (\beta - b)^T X^T X (\beta - b)] \right\}$$

Igual que con el estimador b para los coeficientes, podemos utilizar el estimador máximo verosímil de la varianza, $\hat{\sigma}^2 = \frac{1}{N}(y - Xb)^T(y - Xb)$, para preparar la verosimilitud de $y|\theta$:

$$f(y|\theta) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} [N\hat{\sigma}^2 + (\beta - b)^T X^T X (\beta - b)] \right\}$$

Notemos ahora que si la varianza σ^2 fuera conocida podríamos descomponer esta distribución en dos partes, una de las cuales tiene la forma del kernel de una distribución normal para $\beta|\sigma^2$, lo que sugiere ya la familia conjugada de distribuciones iniciales:

$$f(y|\theta) \propto \exp \left\{ -\frac{1}{2\sigma^2} [(\beta - b)^T X^T X (\beta - b)] \right\} (\sigma^2)^{-N/2} \exp \left\{ -\frac{N\hat{\sigma}^2}{2\sigma^2} \right\}.$$

Finalmente, en este contexto resultará más fácil trabajar en términos de precisiones que de varianzas. Si definimos la precisión de una variable normal como $\tau = \frac{1}{\sigma^2}$, tenemos que la función de verosimilitud en el modelo normal se puede representar como sigue:

$$p(y|\theta) \propto \exp \left\{ -\frac{\tau}{2} [(\beta - b)^T X^T X (\beta - b)] \right\} \tau^{N/2} \exp \left\{ -\frac{N\hat{\sigma}^2\tau}{2} \right\}. \quad (\text{A.3})$$

Distribuciones iniciales

La primera distribución inicial que podríamos plantear sería la distribución conjugada. Recordemos que esta debe tener la misma forma funcional que la verosimilitud, por lo que (A.3) sugiere lo siguiente:

$$f(\theta) = f(\beta, \tau) \propto \exp \left\{ -\frac{\tau}{2} [(\beta - b_0)^T T_0 (\beta - b_0)] \right\} \tau^{a/2} \exp \left\{ -\frac{r\tau}{2} \right\},$$

donde b_0 , T_0 , a y r sean algunos parámetros convenientes. Con esta forma, podemos determinar la familia conjugada en un proceso de dos pasos. En primer lugar, asumimos que la varianza o precisión está dada, lo que permite definir una distribución inicial para $\beta|\tau$. Posteriormente, determinaremos la distribución inicial conjugada para τ . Es decir, separaremos la distribución inicial en dos: $f(\theta) = f(\beta, \tau) = f(\beta|\tau)f(\tau)$.

La distribución condicional resulta ser una normal centrada en b_0 y con precisión τT_0 , por lo que debemos completarla multiplicando por $1 = \tau^{(d-d)/2}$, donde d es el número

de coeficientes, incluyendo a β_0 . Así:

$$\begin{aligned} f(\theta) &= f(\beta|\tau)f(\tau) \propto \exp\left\{-\frac{\tau}{2} [(\beta - b_0)^T T_0 (\beta - b_0)]\right\} \tau^{a/2} \exp\left\{-\frac{r\tau}{2}\right\} \\ &\propto \tau^{(d-d)/2} \exp\left\{-\frac{\tau}{2} [(\beta - b_0)^T T_0 (\beta - b_0)]\right\} \tau^{a/2} \exp\left\{-\frac{r\tau}{2}\right\}. \end{aligned}$$

Con lo que

$$\begin{aligned} f(\beta|\tau) &\propto \tau^{d/2} \exp\left\{-\frac{\tau}{2} [(\beta - b_0)^T T_0 (\beta - b_0)]\right\} \text{ y} \\ f(\tau) &\propto \tau^{(a-d)/2} \exp\left\{-\frac{r\tau}{2}\right\}. \end{aligned} \quad (\text{A.4})$$

La distribución inicial de τ también ya tiene una forma conocida: es proporcional a una gamma. Para verlo solo basta con un poco de álgebra para verificar que el parámetro de forma debe ser $a_0 = (a-d+2)/2 = (a^*-d)/2$ con $a^* = a+2$ y el de tasa $r_0 = r/2$. Por lo tanto, en su conjunto, tenemos que θ tiene una distribución inicial *Normal-Gamma*:

$$\theta = (\beta, \tau) \sim NG_d \left(b_0, T_0, a_0 = \frac{a^* - d}{2}, r_0 = \frac{r}{2} \right)$$

de forma que

$$\beta|\tau \sim N_d(b_0, \tau T_0) \text{ y } \tau \sim \Gamma \left(a_0 = \frac{a^* - d}{2}, r_0 = \frac{r}{2} \right). \quad (\text{A.5})$$

Cabe hacer notar que esta distribución inicial conjugada es propia siempre que $a^* > d$, $r > 0$ y $B_0 = T_0^{-1}$ sea positiva definida.

Por otro lado, si se buscan distribuciones iniciales más vagas, resulta que también es posible obtener distribuciones mínimo informativas límites de esta conjugada. Por ejemplo, aunque es impropia, la inicial de Jeffreys es de esa forma con los siguientes límites: $a^* \rightarrow d$, $r \rightarrow 0$ y $B_0 = T_0^{-1} \rightarrow \mathbf{O}$. La (A.4) se reduce a la siguiente expresión (Gutiérrez Peña 1998):

$$f(\theta) = f(\beta, \tau) \propto \tau^{(d-2)/2} \quad (\text{A.6})$$

Distribuciones finales

Consideremos para la actualización el caso general de la distribución inicial normal gamma de (A.5).

$$\begin{aligned} y|\theta &\sim N_N(X\beta, \sigma^2 \mathbb{I}_N) \quad \text{tal que} \quad \theta = (\beta, \sigma^2) \sim f(\beta, \sigma^2) \\ \beta|\tau &\sim N_d(b_0, \tau T_0) \\ \tau &\sim \Gamma \left(a_0 = \frac{a^* - d}{2}, r_0 = \frac{r}{2} \right). \end{aligned} \quad (\text{A.7})$$

Aplicaremos el teorema de Bayes con base en (A.3) y (A.4) buscando, al tener una inicial conjugada, mantener la forma de normal gamma. Esto es, la verosimilitud la podemos ver también como el producto de dos distribuciones, una normal para $\beta|\tau$ centrada en el estimador máximo verosímil b y con precisión $\tau X^T X$ y una gamma para τ utilizando el estimador máximo verosímil de la varianza $\hat{\sigma}^2$.

$$\begin{aligned} f(\theta|y) &\propto f(y|\theta)f(\theta) \\ &\propto \exp \left\{ -\frac{\tau}{2} [(\beta - b)^T X^T X(\beta - b)] \right\} \tau^{N/2} \exp \left\{ -\frac{N\hat{\sigma}^2\tau}{2} \right\} \\ &\quad \tau^{d/2} \exp \left\{ -\frac{\tau}{2} [(\beta - b_0)^T T_0(\beta - b_0)] \right\} \tau^{(a-d)/2} \exp \left\{ -\frac{r\tau}{2} \right\} \\ &\propto \tau^{d/2} \exp \left\{ -\frac{\tau}{2} [(\beta - b)^T X^T X(\beta - b) + (\beta - b_0)^T T_0(\beta - b_0)] \right\} \\ &\quad \tau^{(N-d+a)/2} \exp \left\{ -\frac{N\hat{\sigma}^2 + r}{2}\tau \right\}. \end{aligned} \quad (\text{A.8})$$

Ahora simplifiquemos el término dentro de la primera exponencial para que coincida con el kernel de una distribución normal.

$$\begin{aligned} &(\beta - b)^T X^T X(\beta - b) + (\beta - b_0)^T T_0(\beta - b_0) \\ &= \beta^T X^T X \beta - \beta^T X^T X b - b^T X^T X \beta + b^T X^T X b + \\ &\quad \beta^T T_0 \beta - \beta^T T_0 b_0 - b_0^T T_0 \beta + b_0^T T_0 b_0 \end{aligned}$$

notando que todos estos términos son escalares de forma que sus transpuestos son ellos mismos, así como que $T_0^T = T_0$,

$$= \beta^T X^T X \beta - 2\beta^T X^T X b + b^T X^T X b + \beta^T T_0 \beta - 2\beta^T T_0 b_0 + b_0^T T_0 b_0$$

$$= \beta^T (X^T X + T_0) \beta - 2\beta^T X^T X b - 2\beta^T T_0 b_0 + b^T X^T X b + b_0^T T_0 b_0$$

definiendo $T_1 = X^T X + T_0$ y $g(X, y) = b^T X^T X b + b_0^T T_0 b_0$,

$$\begin{aligned} &= \beta^T T_1 \beta - 2\beta^T X^T X b - 2\beta^T T_0 b_0 + g(X, y) \\ &= \beta^T T_1 \beta - 2\beta^T [X^T X b + T_0 b_0] + g(X, y) \end{aligned}$$

definiendo $b_1 = T_1^{-1}(X^T X b + T_0 b_0)$ y completando el cuadrado:

$$\begin{aligned} &= \beta^T T_1 \beta - 2\beta^T T_1 b_1 + g(X, y) \\ &= (\beta - b_1)^T T_1 (\beta - b_1) + g(X, y) - b_1^T T_1 b_1. \end{aligned} \tag{A.9}$$

Con esta manipulación de términos, ya podemos tener la distribución posterior de $\beta|\tau$, sustituyendo (A.9) en (A.8), como una normal d -variada con media b_1 y precisión τT_1 :

$$\begin{aligned} f(\theta|y) &\propto \tau^{d/2} \exp \left\{ -\frac{\tau}{2} [(\beta - b_1)^T T_1 (\beta - b_1)] \right\} \\ &\quad \tau^{(N-d+a)/2} \exp \left\{ -\frac{N\hat{\sigma}^2 + g(X, y) - b_1^T T_1 b_1 + r}{2} \tau \right\}. \end{aligned}$$

La nueva media $b_1 = T_1^{-1}(X^T X b + T_0 b_0)$ puede verse como un promedio de las medias originales—la de la inicial y el estimador máximo verosímil— ponderadas por sus precisiones (Congdon 2006). La nueva precisión es simplemente la suma de las precisiones originales.

Ahora debemos encontrar los nuevos parámetros de forma y tasa para la distribución posterior de τ . Igualando el exponente de τ en la última expresión a $a_1 - 1$, donde a_1 es el nuevo parámetro de forma, para satisfacer la representación de una distribución gamma se llega a que $a_1 = (N - d + a^*)/2$. El nuevo parámetro de tasa r_1 requiere ser un poco más explícitos:

$$\begin{aligned} r_1 &= \frac{N\hat{\sigma}^2 + g(X, y) - b_1^T T_1 b_1 + r}{2} \\ &= \frac{(y - Xb)^T (y - Xb) + b^T X^T X b + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2}. \end{aligned}$$

Pero resulta que $(y - Xb)^T(y - Xb) + b^T X^T Xb = y^T y$:

$$\begin{aligned} (y - Xb)^T(y - Xb) + b^T X^T Xb &= y^T y - 2y^T Xb + b^T X^T Xb + b^T X^T Xb \\ &= y^T y - 2y^T Xb + 2b^T X^T Xb \\ &= y^T y - 2b^T X^T Xb + 2b^T X^T Xb \\ &= y^T y. \end{aligned} \quad (\text{A.10})$$

Por lo que, en realidad,

$$r_1 = \frac{y^T y + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2}.$$

Con esto tenemos que la actualización de las (A.7) nos llevan al siguiente modelo conjugado:

$$\begin{aligned} y|\theta &\sim N_N(X\beta, \sigma^2 \mathbb{I}_N) \quad \text{tal que} \quad \theta = (\beta, \sigma^2) \sim f(\beta, \sigma^2) \\ \beta|\tau &\sim N_d(b_0, \tau T_0) \quad \tau \sim \Gamma \left(a_0 = \frac{a^* - p}{2}, r_0 = \frac{r}{2} \right) \\ \beta|\tau, y &\sim N_p(b_1, \tau T_1) \quad \tau|y \sim \Gamma(a_1, r_1) \end{aligned}$$

con $a^* > d$, $r > 0$ y $B_0 = T_0^{-1}$ positiva definida y tal que

$$\begin{aligned} T_1 &= X^T X + T_0 & b_1 &= T_1^{-1}(X^T X b + T_0 b_0) = T_1^{-1}(X^T y + T_0 b_0), \\ a_1 &= \frac{N - d + a^*}{2} & r_1 &= \frac{y^T y + b_0^T T_0 b_0 - b_1^T T_1 b_1 + r}{2} \end{aligned} \quad (\text{A.11})$$

donde $b = (X^T X)^{-1} X^T y$ es el estimador máximo verosímil de β . Más aún, si en lugar de utilizar como distribución inicial una normal gamma de esta forma se utiliza la inicial de Jeffreys de (A.6), podemos utilizar estas expresiones para hacer la actualización—aprovechando el carácter que la inicial de Jeffreys tiene como límite de conjugadas—considerando $a^* \rightarrow d$, $r \rightarrow 0$ y $B_0 = T_0^{-1} \rightarrow \mathbf{O}$, por lo que se tendrían:

$$T_1 = X^T X \quad b_1 = b \quad a_1 = \frac{N}{2} \quad r_1 = \frac{y^T y - b^T X^T X b}{2} = \frac{N \hat{\sigma}^2}{2}$$

donde la equivalencia del estimador máximo verosímil $\hat{\sigma}^2$ puede verificarse con (A.10).

Anexo B

Modelos Compuestos

B.1. Modelo A

$$y_c | \theta \sim \text{Binom}(n_c, p_c) \quad \forall c \in \mathbb{N}_C$$
$$\text{con} \quad \ln \left(\frac{p_c}{1 - p_c} \right) = \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} \quad (\text{B.1})$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$
$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\alpha_d \sim N(\mu_\alpha, \sigma = 1)$$

$$\beta_{d,j} \sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4$$

$$\gamma_{d,j} \sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7$$

y

$$\begin{aligned}\mu_\alpha &\sim N(-1.7, \sigma = 0.25) \\ \mu_\beta &\sim N(0, \sigma = 0.5) \\ \mu_\gamma &\sim N(0, \sigma = 0.5)\end{aligned}$$

B.2. Modelo B

$$\begin{aligned}y_c | \theta &\sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\ \text{con} \quad ln\left(\frac{p_c}{1 - p_c}\right) &= \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} + \delta_{d[c]} x_{edad,c} \quad (\text{B.2})\end{aligned}$$

tal que

$$\begin{aligned}\beta_{d,5} &= - \sum_{k=1}^4 \beta_{d,k}, \\ \gamma_{d,8} &= - \sum_{k=1}^7 \gamma_{d,k} \\ \delta_{d,6} &= - \sum_{k=1}^5 \delta_{d,k}\end{aligned}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\begin{aligned}\alpha_d &\sim N(\mu_\alpha, \sigma = 1) \\ \beta_{d,j} &\sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4 \\ \gamma_{d,j} &\sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7 \\ \delta_{d,j} &\sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5\end{aligned}$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

$$\mu_\delta \sim N(0, \sigma = 0.5)$$

B.3. Modelo C

$$y_c | \theta \sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C$$

$$\text{con} \quad \ln \left(\frac{p_c}{1 - p_c} \right) = \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} + \delta_{d[c]} x_{edad,c} + \lambda_{d[c]} x_{migr,c} \quad (\text{B.3})$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$

$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

$$\delta_{d,6} = - \sum_{k=1}^5 \delta_{d,k}$$

$$\lambda_{d,2} = -\lambda_{d,1}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\alpha_d \sim N(\mu_\alpha, \sigma = 1)$$

$$\beta_{d,j} \sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4$$

$$\gamma_{d,j} \sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7$$

$$\delta_{d,j} \sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5$$

$$\lambda_{d,1} \sim N(\mu_\lambda, \sigma = 1)$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

$$\mu_\delta \sim N(0, \sigma = 0.5)$$

$$\mu_\lambda \sim N(0, \sigma = 0.5)$$

B.4. Modelo D

$$\begin{aligned}
y_c | \theta &\sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\
\text{con} \quad ln \left(\frac{p_c}{1 - p_c} \right) &= \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} + \delta_{d[c]} x_{edad,c} \\
&+ \lambda_{d[c]} x_{migr,c} + \kappa_{d[c]} x_{sexo,c}
\end{aligned} \tag{B.4}$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$

$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

$$\delta_{d,6} = - \sum_{k=1}^5 \delta_{d,k}$$

$$\lambda_{d,2} = -\lambda_{d,1}$$

$$\kappa_{d,2} = -\kappa_{d,1}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\alpha_d \sim N(\mu_\alpha, \sigma = 1)$$

$$\beta_{d,j} \sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4$$

$$\gamma_{d,j} \sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7$$

$$\delta_{d,j} \sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5$$

$$\lambda_{d,1} \sim N(\mu_\lambda, \sigma = 1)$$

$$\kappa_{d,1} \sim N(\mu_\kappa, \sigma = 1)$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

$$\mu_\delta \sim N(0, \sigma = 0.5)$$

$$\mu_\lambda \sim N(0, \sigma = 0.5)$$

$$\mu_\kappa \sim N(0, \sigma = 0.5)$$

B.5. Modelos E y F

$$y_c | \theta \sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C$$

$$\begin{aligned} \text{con} \quad \ln \left(\frac{p_c}{1 - p_c} \right) &= \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} + \delta_{d[c]} x_{edad,c} \\ &\quad + \lambda_{d[c]} x_{migr,c} + \kappa_{d[c]} x_{sexo,c} + \zeta_{d[c]} x_{ocu,c} \end{aligned} \tag{B.5}$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$

$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

$$\delta_{d,6} = - \sum_{k=1}^5 \delta_{d,k}$$

$$\lambda_{d,2} = -\lambda_{d,1}$$

$$\kappa_{d,2} = -\kappa_{d,1}$$

$$\zeta_{d,2} = -\zeta_{d,1}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\begin{aligned}\alpha_d &\sim N(\mu_\alpha, \sigma = 1) \\ \beta_{d,j} &\sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4 \\ \gamma_{d,j} &\sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7 \\ \delta_{d,j} &\sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5 \\ \lambda_{d,1} &\sim N(\mu_\lambda, \sigma = 1) \\ \kappa_{d,1} &\sim N(\mu_\kappa, \sigma = 1) \\ \zeta_{d,1} &\sim N(\mu_\zeta, \sigma = 1)\end{aligned}$$

y

$$\begin{aligned}\mu_\alpha &\sim N(-1.7, \sigma = 0.25) \\ \mu_\beta &\sim N(0, \sigma = 0.5) \\ \mu_\gamma &\sim N(0, \sigma = 0.5) \\ \mu_\delta &\sim N(0, \sigma = 0.5) \\ \mu_\lambda &\sim N(0, \sigma = 0.5) \\ \mu_\kappa &\sim N(0, \sigma = 0.5) \\ \mu_\zeta &\sim N(0, \sigma = 0.5)\end{aligned}$$

B.6. Modelo G

$$\begin{aligned}y_c | \theta &\sim Binom(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\ \text{con} \quad ln\left(\frac{p_c}{1-p_c}\right) &= \alpha_{d[c]} + \beta_{d[c]}x_{escol,c} + \gamma_{d[c]}x_{csp,c} + \delta_{d[c]}x_{edad,c} + \lambda_{d[c]}x_{migr,c} \\ &+ \kappa_{d[c]}x_{sexo,c} + \zeta_{d[c]}x_{ocu.gral,c} + \xi_{d[c]}x_{ocu.juv,c}\end{aligned}\tag{B.6}$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$

$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

$$\delta_{d,6} = - \sum_{k=1}^5 \delta_{d,k}$$

$$\lambda_{d,2} = -\lambda_{d,1}$$

$$\kappa_{d,2} = -\kappa_{d,1}$$

$$\zeta_{d,2} = -\zeta_{d,1}$$

$$\xi_{d,2} = -\xi_{d,1}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\alpha_d \sim N(\mu_\alpha, \sigma = 1)$$

$$\beta_{d,j} \sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4$$

$$\gamma_{d,j} \sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7$$

$$\delta_{d,j} \sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5$$

$$\lambda_{d,1} \sim N(\mu_\lambda, \sigma = 1)$$

$$\kappa_{d,1} \sim N(\mu_\kappa, \sigma = 1)$$

$$\zeta_{d,1} \sim N(\mu_\zeta, \sigma = 1)$$

$$\xi_{d,1} \sim N(\mu_\xi, \sigma = 1)$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

$$\mu_\delta \sim N(0, \sigma = 0.5)$$

$$\mu_\lambda \sim N(0, \sigma = 0.5)$$

$$\mu_\kappa \sim N(0, \sigma = 0.5)$$

$$\mu_\zeta \sim N(0, \sigma = 0.5)$$

$$\mu_\xi \sim N(0, \sigma = 0.5)$$

B.7. Modelo H

$$\begin{aligned}
 y_c | \theta &\sim \text{Binom}(n_c, p_c) \quad \forall c \in \mathbb{N}_C \\
 \text{con} \quad \ln \left(\frac{p_c}{1 - p_c} \right) = & \alpha_{d[c]} + \beta_{d[c]} x_{escol,c} + \gamma_{d[c]} x_{csp,c} + \delta_{d[c]} x_{edad,c} + \lambda_{d[c]} x_{migr,c} \\
 & + \kappa_{d[c]} x_{sexo,c} + \zeta_{d[c]} x_{ocu_gral,c} + \xi_{d[c]} x_{ocu_juv,c} + v_{d[c]} x_{ocu_may,c}
 \end{aligned} \tag{B.7}$$

tal que

$$\beta_{d,5} = - \sum_{k=1}^4 \beta_{d,k},$$

$$\gamma_{d,8} = - \sum_{k=1}^7 \gamma_{d,k}$$

$$\delta_{d,6} = - \sum_{k=1}^5 \delta_{d,k}$$

$$\lambda_{d,2} = -\lambda_{d,1}$$

$$\kappa_{d,2} = -\kappa_{d,1}$$

$$\zeta_{d,2} = -\zeta_{d,1}$$

$$\xi_{d,2} = -\xi_{d,1}$$

$$v_{d,2} = -v_{d,1}$$

donde $\forall d \in \mathbb{N}_{96}$

$$\alpha_d \sim N(\mu_\alpha, \sigma = 1)$$

$$\beta_{d,j} \sim N(\mu_\beta, \sigma = 1) \quad \forall j \in \mathbb{N}_4$$

$$\gamma_{d,j} \sim N(\mu_\gamma, \sigma = 1) \quad \forall j \in \mathbb{N}_7$$

$$\delta_{d,j} \sim N(\mu_\delta, \sigma = 1) \quad \forall j \in \mathbb{N}_5$$

$$\lambda_{d,1} \sim N(\mu_\lambda, \sigma = 1)$$

$$\kappa_{d,1} \sim N(\mu_\kappa, \sigma = 1)$$

$$\zeta_{d,1} \sim N(\mu_\zeta, \sigma = 1)$$

$$\xi_{d,1} \sim N(\mu_\xi, \sigma = 1)$$

$$v_{d,1} \sim N(\mu_v, \sigma = 1)$$

y

$$\mu_\alpha \sim N(-1.7, \sigma = 0.25)$$

$$\mu_\beta \sim N(0, \sigma = 0.5)$$

$$\mu_\gamma \sim N(0, \sigma = 0.5)$$

$$\mu_\delta \sim N(0, \sigma = 0.5)$$

$$\mu_\lambda \sim N(0, \sigma = 0.5)$$

$$\mu_\kappa \sim N(0, \sigma = 0.5)$$

$$\mu_\zeta \sim N(0, \sigma = 0.5)$$

$$\mu_\xi \sim N(0, \sigma = 0.5)$$

$$\mu_v \sim N(0, \sigma = 0.5)$$

Referencias

Referencias

- Almond, Gabriel A. y Sidney Verba. 1963. *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press.
- Aquino Pérez, Elizabeth. 2010. “Construcción de una tabla de mortalidad con un enfoque bayesiano”. Tesina para obtener el grado de Licenciada en Actuaría, Instituto Tecnológico Autónomo de México (ITAM).
- Arnade, Chris. 2016. *Attributions of causality*. Ver. 2. <http://www.interfluidity.com/v2/6602.html>.
- Asamblea National. 2017a. *Fiche de synthèse n1: Présentation synthétique des institutions françaises*. Consultado el 6 de enero de 2018. <http://www2.assemblee-nationale.fr/decouvrir-1-assemblee/role-et-pouvoirs-de-1-assemblee-nationale/les-institutions-francaises-generalites/presentation-synthetique-des-institutions-francaises>.
- . 2017b. *Fiche de synthèse n11: L'organisation territoriale de la France*. Consultado el 6 de enero de 2018. <http://www2.assemblee-nationale.fr/decouvrir-1-assemblee/role-et-pouvoirs-de-1-assemblee-nationale/les-institutions-francaises-generalites/l-organisation-territoriale-de-la-france>.
- . 2017c. *Fiche de synthèse n45: La mise en cause de la responsabilité du Gouvernement*. Consultado el 6 de enero de 2018. [http://www2.assemblee-nationale.fr/decouvrir-1-assemblee/role-et-pouvoirs-de-1-assemblee-nationale/les-fonctions-de-controle-et-l-information-des-deputes/la-mise-en-cause-de-la-responsabilite-du-gouvernement](http://www2.assemblee-nationale.fr/decouvrir-1-assemblee/role-et-pouvoirs-de-1-assemblee-nationale/les-fonctions-de-1-assemblee-nationale/les-fonctions-de-controle-et-l-information-des-deputes/la-mise-en-cause-de-la-responsabilite-du-gouvernement).

- Balaam, David y Michael Veseth. 2008. “Marx, Lenin, and the Structuralist Perspective”. En *Introduction to International Political Economy*, 4.^a ed. Upper Saddle River, Nueva Jersey: Pearson.
- Beauchamp, Zack. 2016a. *An expert on the European far right explains the growing influence of anti-immigrant politics*. Entrevista a Cass Mudde, de la Universidad de Georgia. Consultado el 30 de octubre de 2017. <http://www.vox.com/2016/5/31/11722994/european-far-right-cas-mudde>.
- . 2016b. *White Riot*. Vox.com. <http://www.vox.com/2016/9/19/12933072/far-right-white-riot-trump-brexit>.
- Berger, James O. 1985. *Statistical decision theory and Bayesian analysis: Springer series in statistics*. 2.^a ed. Estados Unidos: Springer-Verlag New York, Inc. ISBN: 0-387-96098-8.
- Berman, Sheri. 2001. “Review: Ideas, Norms, and Culture in Political Analysis”. *Comparative Politics* 33 (2): 231-250. <http://www.jstor.org/stable/422380>.
- Bernardo, José Miguel. 1981. *Bioestadística: Una Perspectiva Bayesiana*. 1.^a ed. Ed. por Albert Vicens. Barcelona: Ediciones Vicens-Vives. ISBN: 84-316-1889-2.
- Bernardo, José Miguel y Adrian F.M. Smith. 2000. “Bayesian Theory”. Cap. Modelling, 1.^a ed., 165-240. Reino Unido: John Wiley & Sons Ltd. ISBN: 978-0471494645.
- Berteloot, Tristan. 2017. *Pour Philippot, l'humiliation ou l'exil*. Consultado el 12 de noviembre de 2017. https://oeilsurlefront.liberation.fr/les-idees/2017/09/19/pour-philippot-l-humiliation-ou-l-exil_1597506.
- Betancourt, Michael. 2017. “A Conceptual Introduction to Hamiltonian Monte Carlo”. ArXiv e-print. <https://arxiv.org/abs/1701.02434>.
- . 2016. “Identifying the Optimal Integration Time in Hamiltonian Monte Carlo”. ArXiv e-print. <https://arxiv.org/abs/1601.00225>.
- . “Robust Statistical Workflow with RStan”. https://github.com/betanalpha/knitr_case_studies/tree/master/rstan_workflow.
- . 2018. “The Convergence of Markov chain Monte Carlo Methods: From the Metropolis method to Hamiltonian Monte Carlo”. ArXiv e-print. <https://arxiv.org/abs/1706.01520v2>.
- Blalock, Hubert M. 1967. *Toward a Theory of Minority-group Relations*. New York: John Wiley & Sons.

- Boily, Frédéric. 2005. “Aux sources idéologiques du Front national: Le mariage du traditionalisme et du populisme”. *Politiques et Sociétés* 24 (1): 23-47.
- Bornschier, Simon. 2009. “Cleavage Politics in Old and New Democracies”. *Living Reviews* (). <http://www.livingreviews.org/lrd-2009-6>.
- C dans l'air. 2017. *Règlement de comptes au FN*. Consultado el 12 de noviembre de 2017. https://www.youtube.com/watch?v=rdWfY7clIco&index=84&list=PL4_by8YNrJv-2o8t0vIHY0Q5TAhorDiin.
- Capo, Dominique. 2017. *21 Avril 2002, le choc, deuxième partie*. Consultado el 11 de noviembre de 2017. <https://www.youtube.com/watch?v=vOT-kmHUV1A>.
- Carney, Timothy. 2016. *Understanding Trump: Economic anxiety is racial anxiety*. The Washington Examiner. <http://www.washingtonexaminer.com/understanding-trump-economic-anxiety-is-racial-anxiety/article/2600612/>.
- Carpizo, Jorge. 2004. “México: ¿sistema presidencial o parlamentario?” *Revista Latinoamericana de Derecho* 1 (1): 1-37.
- Casella, George y Edward I. George. 1992. “Explaining the Gibbs Sampler”. *The American Statistician* 46 (3): 167-174.
- Chib, Siddhartha y Edward Greenberg. 1995. “Understanding the Metropolis-Hastings Algorithm”. *The American Statistician* 49 (4): 327-335.
- Coffé, Hilde, Bruno Heyndels y Jan Vermeir. 2007. “Fertile grounds for extreme right-wing parties: Explaining the Vlaams Blok’s electoral success”. *Electoral Studies* 26:142-155. doi:10.1016/j.electstud.2006.01.005.
- Congdon, Peter. 2006. *Bayesian Statistical Modelling*. 2.^a ed. Wiley series in probability and statistics. Reino Unido: John Wiley & Sons Ltd. ISBN: 978-0-470-01875-0.
- Constitution du 4 octobre 1958*. Texto vigente al 6 de enero de 2018. <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006071194&dateTexte=20180106>.
- Cox, Michael y Martin Durham. 2016. “Te politics of anger: the extreme right in the United States”. Cap. 13 en *The Politics of the Extreme Right: From the Margins to the Mainstream*, ed. por Paul Hainsworth, 287-311. Bloomsbury Academic.
- Crépon, Sylvain. 2015. “La politique des mœurs au Front National”. Cap. 8 en *Les faux-semblants du Front National: sociologie d'un parti politique*, ed. por Sylvain Crépon, Alexandre Dézé y Nonna Mayer, 185-205. Paris: Presses de SciencesPo.

- Crépon, Sylvain, Alexandre Dézé y Nonna Mayer, eds. 2015. “Repères chronologiques”. En *Les faux-semblants du Front National: sociologie d'un parti politique*, 545-557. Paris: Presses de SciencesPo.
- Delafoi, Florian. 2017. *Front national (1971-2017)*. <https://www.letemps.ch/monde/front-national-19712017>.
- Dobson, Annette J. 2001. *An introduction to generalized linear models*. Chapman & Hall. ISBN: 1-58488-165-8.
- Draper, Norman R. y Harry Smith. 1998. *Applied Regression Analysis*. 3.^a ed. Willey Series in Probability and Statistics. Estados Unidos: Wiley. ISBN: 978-0-471-17082-2.
- Eckhardt, Roger. 1987. “Stan Ulam, John von Neumann, and the Monte Carlo method”. Special issue dedicated to Stan Ulam, *Los Alamos Science*: 131-137.
- Europe 1. 2017. *Départ de Florian Philippot: «une mauvaise nouvelle pour le FN», selon Pascal Perrineau*. Entrevista a Pascal Perrineau. Consultado el 12 de noviembre de 2017. <http://www.europe1.fr/politique/depart-de-florian-philippot-une-mauvaise-nouvelle-pour-le-fn-selon-pascal-perrineau-3443705>.
- Front National. 2012. *Mon Projet: Pour la France et les français*. Consultado el 26 de marzo de 2018. http://www.frontnational.com/pdf/projet_mlp2012.pdf.
- Galtier, Ludovic. 2017. *Philippot: le FN «fait un retour en arrière terrifiant» sur sa ligne*. Consultado el 12 de noviembre de 2017. <http://www rtl.fr/actu/politique/philippot-le-front-national-fait-un-retour-en-arriere-terrifiant-sur-sa-ligne-7790173335>.
- Gelfand, Alan E. y Adrian F.M. Smith. 1990. “Sampling-Based Approaches to Calculating Marginal Densities”. *Journal of the American Statistical Association* 85 (410): 398-409.
- Gelfand, Alan E. y col. 1990. “Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling”. *Journal of the American Statistical Association* 85 (412): 972-985.
- Gelman, A., G.O. Roberts y Gilks W.R. 1996. “Efficient Metropolis Jumping Rules”. En *Bayesian Statistics 5*, 599-607.
- Gelman, Andrew y Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 5.^a ed. Analytical Methods for Social Science Research. Estados Unidos: Cambridge University Press. ISBN: 978-0-5216-8689-1.

- Gelman, Andrew y Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences". *Statistical Science* 7 (4): 457-472.
- Gelman, Andrew y col. 2013. *Bayesian Data Analysis*. 3.^a ed. Chapman & Hall/CRC texts in Statistical Science. Estados Unidos: CRC Press. ISBN: 978-1-4398-4095-5.
- Geman, Stuart y Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 370-384.
- Geyer, Charles J. 2005. *Markov Chain Monte Carlo Lecture Notes*. Course notes originally used Spring Quarter 1998. University of Minnesota, Twin Cities.
- Gombin, Joël. 2013a. «*Nouveau» FN, vieille carte électorale? Les territoires du vote pour le Front national de 1995 à 2002*. Congrès de l'Association française de science politique.
- . 2009. *Analyse écologique, modèles multi-niveaux et sociologie électorale: L'exemple des votes pour le Front national*. Congrès de l'Association française de science politique.
- . 2005. "Le vote pour le Front national dans le Vaucluse et les Bouches-du-Rhône". Mémoire pour l'obtention du grade de Master d'études politiques, Université de droit, d'économie et des sciences d'Aix-Marseille.
- . 2013b. *The Front National vote and its sectorial support*. ECPR General Conference.
- Gombin, Joël y Jean Rivière. 2013. *Éléments quantitatifs sur la dimension spatiale des effets électoraux des inégalités sociales dans les mondes périurbains français (2007-2012)*. Congrès de l'Association française de sociologie.
- Goodliffe, Gabriel. 2016. "From Political Fringe to Political Mainstream: The Front National and the 2014 Municipal Elections in France". *French Politics, Culture & Society* 34 (3): 126-147.
- . 2017. *Global Populism in the Rise*. Participación en un foro sobre populismo entre académicos del ITAM y la Universidad John Hopkins. Ciudad de México.
- . 2019. "Las clases medias tradicionales y la derecha radical en Francia: Una explicación cultural". *Estudios: Filosofía, Historia, Letras* 27 (128): 15-53.
- Gross, Estelle. 2016. *Marine Le Pen change de slogan : "Une ligne populiste assumée"*. <https://www.nouvelobs.com/politique/election-presidentielle-2017/20160919.OBS8333/marine-le-pen-change-de-slogan-une-ligne-populiste-assumee.html>.

- Gubernatis, J.E. 2005. "Marshall Rosenbluth and the Metropolis algorithm". *Physics of Plasmas* 12 (57303).
- Gutiérrez Peña, Eduardo. 1998. *Análisis Bayesiano de Modelos Jerárquicos Lineales*. Serie monografías, IIMAS. Universidad Nacional Autónoma de México (UNAM), Ciudad de México.
- . 1997. *Métodos Computacionales en la Inferencia Bayesiana*. Serie monografías, II-MAS. Universidad Nacional Autónoma de México (UNAM), Ciudad de México.
- . 2016. *Presentación Introductoria al curso Métodos Estadísticos Bayesianos*. Departamento Académico de Estadística. Instituto Tecnológico Autónomo de México (ITAM), Ciudad de México.
- Haigh, Thomas, Mark Priestley y Crispin Rope. 2014. "Los Alamos Bets on ENIAC: Nuclear Monte Carlo Simulations, 1947–1948". *IEEE Annals of the History of Computing* 36 (3): 42-63.
- Hainsworth, Paul. 2016a. "Introduction: the extreme right". Cap. 1 en *The Politics of the Extreme Right: From the Margins to the Mainstream*, ed. por Paul Hainsworth, 1-17. Bloomsbury Academic. ISBN: 978-1-4742-9095-1.
- . 2016b. "The Front National: from ascendancy to fragmentation on the French extreme right". Cap. 2 en *The Politics of the Extreme Right: From the Margins to the Mainstream*, ed. por Paul Hainsworth, 18-32. Bloomsbury Academic. ISBN: 978-1-4742-9095-1.
- Hastings, W.K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika* 57 (1): 97-109.
- Heilbroner, Robert. 1992. "The Worldy Philosophers: The Lives, Times, and Ideas of the Great Economic Thinkers". Cap. The Inexorable System of Karl Marx, 6.^a ed. Nueva York: Simon & Schuster.
- Hoare, Liam. 2016. *Austria's Incredibly Narrow Escape From Neo-Fascism*. Tablet Mag. <http://www.tabletmag.com/jewish-news-and-politics/202857/austrias-narrow-escape>.
- Hoffman, Matthew W. y Andrew Gelman. 2011. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". ArXiv e-print. <https://arxiv.org/pdf/1111.4246>.
- Ina Politique. 2012. *Marie Caroline Le Pen Mantes*. Consultado el 11 de noviembre de 2017. <https://www.youtube.com/watch?v=pja01SbmHE>.

- Inglehart, Ronald F. y Pipa Norris. 2016. "Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash". *Harvard Kennedy School Faculty Research Working Paper Series* ().
- INSEE. 2019. "Définitions des concepts - Recensement de la population". <https://www.insee.fr/fr/information/2383278>.
- . 2017. "Fiche thématique – La Précision des résultats du recensement". <https://www.insee.fr/fr/information/2383177>.
- Ivarsflaten, Elisabeth y Froy Gudbransen. 2014. *The populist radical right in Western Europe*.
- Kesselman, Mark. 1979. "Reviewed Work: The Silent Revolution: Changing Values and Political Styles Among Western Publics by Ronald Inglehart". *The American Political Science Review* 73, n.º 1 (): 284-286. <https://www.jstor.org/stable/pdf/1954818.pdf>.
- Kitschelt, Herbert. 1995. *The Radical Right in Western Europe: A Comparative Analysis*. Michigan: University of Michigan Press.
- L'Obs. 2007. *Le programme de Jean-Marie Le Pen*. Consultado el 26 de marzo de 2018. <https://www.nouvelobs.com/politique/elections-2007/20070224.OBS4060/le-programme-de-jean-marie-le-pen.html>.
- Lambert, Ben. 2018. "The intuition behind the Hamiltonian Monte Carlo algorithm". Youtube. <https://www.youtube.com/watch?v=a-wydhEuAm0>.
- Le Bras, Hervé. 2016. *Le nouvel ordre électoral: Tripartisme contre démocratie*. Chambray-lès-Tours: Éditions du Seuil et La République des idées. ISBN: 978-2-02-130028-4.
- . 2015. *Le Pari du FN*. París: Éditions Autrement. ISBN: 978-2-7467-4126-3.
- Le Monde. 2016. *Marine Le Pen prône la fin de l'éducation gratuite pour les enfants étrangers*. http://www.lemonde.fr/election-presidentielle-2017/article/2016/12/08/marine-le-pen-prone-la-fin-de-l-education-gratuite-pour-les-enfants-etrangers_5045648_4854003.html.
- Le Parisien. 2013. *Marine Le Pen : les origines du FN ne sont «absolument pas» à l'extrême droite*. <http://www.leparisien.fr/politique/marine-le-pen-les-origines-du-fn-ne-sont-absolument-pas-a-l-extreme-droite-03-10-2013-3192605.php>.

- Le Pen, Jean Marie. 2001. *Déclaration de M. Jean-Marie Le Pen, président du Front national et candidat à l'élection présidentielle de 2002 sur le terrorisme, l'identité nationale, l'immigration, le manque de sécurité, son programme électoral pour l'élection présidentielle de 2002.* <http://discours.vie-publique.fr/notices/013003238.html>.
- Leprince, Chloé. 2016. “Establishment”, une béquille lexicale populiste de Jean-Marie Le Pen à Donald Trump. <https://www.franceculture.fr/emissions/le-mot-de-la-semaine/establishment-une-bequille-lexicale-populiste-de-jean-marie-le-pen>.
- Les Echos. 1999. *La condamnation de Le Pen à un an d'inéligibilité confirmée*. Consultado el 11 de noviembre de 2017. https://www.lesechos.fr/24/11/1999/LesEchos/18032-181-ECH_la-condamnation-de-le-pen-a-un-an-d-ineligibilite-confirmee.htm.
- Letras Libres. 2016. <http://www.letraslibres.com/mexico/revista/fascista-americano>.
- Linz, Juan J. 1990. “The Perils of Presidentialism”. *Journal of Democracy* 1 (1): 51-69.
- Lombart, Gaël. 2016. *Présidentielle : et si deux candidats arrivaient à égalité...* Consultado el 7 de enero de 2018. <http://www.leparisien.fr/elections/presidentielle/pratique/presidentielle-et-si-deux-candidats-arrivaient-a-equalite-28-12-2016-6502290.php>.
- Mammone, Andrea, Emmanuel Godin y Brian Jenkins. 2012. “Introduction: mapping the ‘right of the mainstream right’ in contemporary Europe”. En *Mapping the Extreme Right in Contemporary Europe: From local to transnational*, ed. por Andrea Mammone, Emmanuel Godin y Brian Jenkins, 1-14. Routledge.
- Marin, Grégory. 2017. *Front national. Sylvain Crépon: «Les cadres du parti le disent, leur avenir est à droite»*. Entrevista a Sylvain Crépon. Consultado el 12 de noviembre de 2017. <https://www.humanite.fr/front-national-sylvain-crepon-les-cadres-du-parti-le-disent-leur-avenir-est-droite-637922>.
- Marrani, David. 2009. “Semi-Presidentialism à la française: the Recent Constitutional Evolution of the Two Headed Executive”. *Constitutional Forum* 18 (2): 55-67.
- Mayer, Nonna. 2007. “Comment Nicolas Sarkozy a rétréci l'électorat Le Pen”. *Revue française de science politique* 57 (3): 429-445.

- . 1987. “De Passy à Barbès: Deux visages du vote Le Pen à Paris”. *Revue française de science politique* 37 (6): 891-906.
- . 2015. “Le plafond de verre électoral entamé, mais pas brisé”. Cap. 13 en *Les faux-semblants du Front National: sociologie d'un parti politique*, ed. por Nonna Mayer, 297-322. Paris: Presses de SciencesPo.
- . 2005. “Votes populaires, votes populistes”. *Hermès, La Revue-Cognition, communication, politique* 42:161-166.
- Mayer, Nonna y Daniel Boy. 1987. “Les ‘variables lourdes’ en sociologie électorale”. *Enquête: anthropologie, histoire, sociologie* 5:109-122.
- Mayer, Nonna y Guy Michelat. 1981. “Les choix électoraux des petits commerçants et artisans en 1967: L’importance des variables contextuelles”. *Revue française de sociologie* 22 (4): 503-521.
- Mayer, Nonna y Pascal Perrineau. 1990. “Pourquoi votent-ils pour le Front national?” *Pouvoirs- Revue française d'études constitutionnelles et politiques* 55:163-184.
- McElreath, Richard. 2017. “Markov Chains: Why Walk When You Can Flow?” <http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/>.
- . 2015. “Statistical Rethinking: A Bayesian Course with Examples in R and Stan”. Cap. Multilevel Models, 1.^a ed., 355-386. CRC Press. ISBN: 978-1482253443.
- Mendoza, Manuel y Pedro Regueiro. 2011. *Estadística Bayesiana*. Departamento Académico de Estadística. Instituto Tecnológico Autónomo de México (ITAM), Ciudad de México.
- Mestre, Abel. 2012. *Le FN n'est plus le même, mais a-t-il vraiment changé?* Consultado el 26 de marzo de 2018. http://www.lemonde.fr/culture/article/2012/09/20/le-front-national-n-est-plus-le-meme-mais-a-t-il-vraiment-change_1763234_3246.html.
- Metropolis, Nicholas. 1987. “The beginning of the Monte Carlo Method”. Special issue dedicated to Stan Ulam, *Los Alamos Science*: 125-130.
- Metropolis, Nicholas y Stanislaw Ulam. 1949. “The Monte Carlo Method”. *Journal of the American Statistical Association* 44 (247): 335-341.
- Metropolis, Nicholas y col. 1953. “Equation of State Calculations by Fast Computing Machines”. *Journal of Chemical Physics* 21 (6): 1087-1092.

- Milanovic, Branko. 2016. *The greatest reshuffle of individual incomes since the Industrial Revolution*. VOX CEPRA. <http://voxeu.org/article/greatest-reshuffle-individual-incomes-industrial-revolution>.
- Ministère de l'Intérieur. 2011. *Les différentes élections*. Consultado el 6 de enero de 2018. <https://www.interieur.gouv.fr/fr/Elections/Les-elections-en-France/Les-modalites-d-elections/Les-differentes-elections>.
- Moore, Barrington. 1966. “Social origins of dictatorship and democracy: Lord and Peasant in the Making of the Modern World”. Cap. Revolution from Above and Fascism, 433-452. Boston: Beacon Press.
- Mudde, Cas. 2007. “Constructing a conceptual framework”. Cap. 1 en *Populist Radical Right Parties in Europe*, 11-31. Cambridge University Press. ISBN: 978-0-521-61632-4.
- . 2016. *The Study of Populist Radical Right Parties: Towards a Fourth Wave*. Center for Research on Extremism, University of Oslo.
- . 2018. *Why is the far right dominated by men?* <https://www.theguardian.com/commentisfree/2018/aug/17/why-is-the-far-right-dominated-by-men>.
- . 2017. *Why nativism, not populism, should be declared word of the year*. <https://www.theguardian.com/commentisfree/2017/dec/07/cambridge-dictionary-nativism-populism-word-year>.
- Mudde, Cas y Cristóbal Rovira Kaltwasser. 2017. *Populism: A Very Short Introduction*. Gran Bretaña: Oxford University Press. ISBN: 9780190234874.
- Neal, Radford M. 2011. “Handbook of Markov Chain Monte Carlo”. Cap. MCMC Using Hamiltonian Dynamics, 1.^a ed., 113-162. Florida: Chapman & Hall/CRC. ISBN: 978-1420079418.
- . 1993. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Technical Report CRG-TR-93-1. University of Toronto.
- Nelder, J.A. y R.W.M. Wedderburn. 1972. “Generalized Linear Models”. *Journal of the Royal Statistical Society* 135 (3): 370-384.
- Nieto Barajas, Luis Enrique. 2016. *Regresión Avanzada: con enfoque Bayesiano*. Maestría en Ciencia de Datos. Instituto Tecnológico Autónomo de México (ITAM), Ciudad de México.
- Nowak, Marysia y Becky Branford. 2017. *France elections: What makes Marine Le Pen far right?* <https://www.bbc.com/news/world-europe-38321401>.

- Olzac, Susan. 1992. *Dynamics of Ethnic Competition and Conflict*. Stanford: Stanford University Press.
- Ortiz Mancera, María Teresa. 2012. “Análisis bayesiano de desempeño escolar de acuerdo con los resultados de la prueba ENLACE”. Tesis para obtener el grado de Licenciada en Matemáticas Aplicadas, Instituto Tecnológico Autónomo de México (ITAM).
- Owen, Art. 2013. *Monte Carlo theory, methods and examples*. Sin publicar. <http://statweb.stanford.edu/~owen/mc/>.
- Perrineau, Pascal. 1999. “La peculiaridad de la extrema derecha francesa en Europa”. *Foro Internacional* 155 (1): 5-16.
- . 2012. *La renaissance électorale de l'électorat frontiste*. Centre de recherches politiques de Sciences Po.
- . 2007. *Qui sont les électeurs potentiels de Jean-Marie Le Pen?* Reporte sobre el Barómetro político francés 2007.
- Petersen, Roger D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Estados Unidos: Cambridge University Press. ISBN: 9780521007740.
- Pulskamp, Richard J. 2009. *Correspondence regarding the Art of Conjecturing*. Reporte de la correspondencia escrita entre Gottfried Leibniz y Jakob Bernoulli entre abril de 1703 y abril de 1705. <https://www.cs.xu.edu/math/Sources/JakobBernoulli/jakob%20and%20leibniz.pdf>.
- Rae, Alasdair. 2016. *What can explain Brexit?* <http://www.statsmapsnpix.com/2016/06/what-can-explain-brexit.html>.
- Ragheb, Magdi. 2013. *Monte Carlo Simulation: Science and Engineering Applications*. Lecture notes for Monte Carlo Simulation course taught at the University of Illinois at Urbana-Champaign. <http://mragheb.com/NPRE%20498MC%20Monte%20Carlo%20Simulations%20in%20Engineering/>.
- Regueiro Martínez, Pedro. 2012. “Análisis jerárquico bayesiano del futbol mexicano”. Tesis para obtener el grado de Licenciado en Matemáticas Aplicadas, Instituto Tecnológico Autónomo de México (ITAM).
- Rincón, Luis. 2012. *Introducción a los procesos estocásticos*. Notas para los cursos de procesos estocásticos en la Facultad de Ciencias de la UNAM.

- Rivière, Jean y col. 2012. “Des contrastes électoraux intra-régionaux aux clivages intra-urbains: Éléments sur le scrutin régional de 2010 dans le Nord-Pas-de-Calais”. *Territoire en mouvement* 16.
- Robert, Christian P. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. 2.^a ed. New York: Springer. ISBN: 978-0-387-71598-8.
- . 2015. “The Metropolis-Hastings algorithm”. *ArXiv e-prints* (). eprint: 1504.01896. <https://arxiv.org/abs/1504.01896>.
- Robert, Christian P. y George Casella. 2010. *Introducing Monte Carlo Methods with R*. Springer. ISBN: 978-1-4419-1575-7.
- Robert, Christian y George Casella. 2011. “A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data”. *Statistical Science* 26 (1): 102-115.
- Rogozhnikov, Alex. 2016. “Hamiltonian Monte Carlo explained”. http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html.
- Ross, Sheldon M. 2010. *A First Course in Probability*. 8.^a ed. Prentice Hall. ISBN: 978-0-13-603313-4.
- . 2013. *Simulation*. 5.^a ed. Estados Unidos: Academic Press. ISBN: 978-0-12-415825-2.
- . 1996. *Stochastic Processes*. 2.^a ed. John Wiley & Sons. ISBN: 0-471-12062-6.
- Seneta, Eugene. 2013. “A Tricentenary history of the Law of Large Numbers”. *Bernoulli* 19 (4): 1088-1121. <https://arxiv.org/pdf/1309.6488.pdf>.
- Sewell Jr., William H. 1992. “A Theory of Structure: Duality, Agency, and Transformation”. *American Journal of Sociology* 98, n.^o 1 (): 1-29. <http://www.jstor.org/stable/pdf/2781191.pdf>.
- Sides, John y Michael Tesler. 2016. *How political science helps explain the rise of Trump (part 3): It's the economy, stupid*. The Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/03/04/how-political-science-helps-explain-the-rise-of-trump-part-3-its-the-economy-stupid/?utm_term=.dd546463a844.
- Silver, Nate. 2016. *Education, Not Income, Predicted Who Would Vote For Trump*. FiveThirtyEight. <http://fivethirtyeight.com/features/education-not-income-predicted-who-would-vote-for-trump/>.

- Simpson, Dan. 2018. “Discusión respecto al muestreador específico de Stan”. Twitter. https://twitter.com/dan_p_simpson/status/1034099024486432770.
- Singpurwalla, Nozer D. 2017. *Subjective Probability: Its Axioms and Acrobatics*. Conferencia magistral dentro del International Workshop on Perspectives on High Dimensional Data VII (HDDA VII). Centro de Investigación en Matemáticas (CIMAT), Guanajuato. http://hddavii.eventos.cimat.mx/sites/hddavii/files/Nozer_Singpurwalla.pdf.
- Skocpol, Theda. 1979. “Old Regimes and Revolutionary Crises in France, Russia, and China”. Edición especial doble sobre el Estado y la Revolución, *Theory and Society* 7, n.º 1 (): 7-95. <http://www.jstor.org/stable/656999?origin=JSTOR-pdf>.
- Smith, A.F.M. y A.E. Gelfand. 1992. “Bayesian Statistics without Tears: A Sampling-Resampling Perspective”. *The American Statistician* 46 (2): 84-88.
- Smith, A.F.M. y G.O. Roberts. 1993. “Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods”. *Journal of the Royal Statistical Society* 55 (1): 3-23.
- Sputnik. 2017. *Austria's FPO Leader: 'We Are Not Extreme Right Party'*. Consultado el 30 de octubre de 2017. <https://sputniknews.com/europe/201710251058522359-austria-strache-fpo-ovp-coalition/>.
- Stan Development Team. 2017. *Stan Modeling Language: User's Guide and Reference Manual*. Version 2.17.0. <https://mc-stan.org/>.
- Stockemer, Daniel. 2017. *The Front National in France: Continuity and Change Under Jean-Marie Le Pen and Marine Le Pen*. Springer. ISBN: 978-3-319-49639-9.
- Taylor, Howard M. y Samuel Karlin. 1984. *An introduction to stochastic modeling*. 3.^a ed. Academic Press. ISBN: 978-0-12-684887-8.
- Tesler, Michael. 2016. *Economic anxiety isn't driving racial resentment. Racial resentment is driving economic anxiety*. The Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/08/22/economic-anxiety-isnt-driving-racial-resentment-racial-resentment-is-driving-economic-anxiety/?utm_term=.b93b1dde31b4.
- Uribe Coughlan, Alexandra. 2016. *Política Comparada*. Curso del Departamento de Ciencia Política del ITAM para el semestre de otoño.

- Usi López, María Andrea. 2014. “Análisis bayesiano de la influenza en México”. Tesis para obtener el grado de Licenciada en Matemáticas Aplicadas, Instituto Tecnológico Autónomo de México (ITAM).
- Valentino, Nicholas, Ted Brader y Ashley Jardina. 2013. “Immigration Opposition Among U.S. Whites: General Ethnocentrism or Media Priming of Attitudes About Latinos?” *Political Psychology* 34 (2): 149-166. doi:10.1111/j.1467-9221.2012.00928.x.
- Vehtari, Aki, Andrew Gelman y Jonah Gabry. 2016. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. ArXiv e-print. <https://arxiv.org/pdf/1507.04544.pdf>.
- Veser, Ernst. 1999. “Semi-presidentialism-Duverger’s Concept: A new political system model”. *Journal of Social Sciences and Philosophy* 11 (1): 39-60.
- Vie Publique. *Le président de la République*. Consultado el 6 de enero de 2018. <http://www.vie-publique.fr/découverte-institutions/institutions/fonctionnement/president-republique/>.
- von Neumann, John y Robert D. Richtmayer. 1947. *Statistical methods in neutron diffusion*. Correspondencia escrita entre von Neumann y Richtmyer en 1947 a propósito del método estadístico propuesto por Stan Ulam.
- Williams, Michelle Hale. 2012. “Downside after the summit: Factors in extreme-right party decline in France and Austria”. Cap. 16 en *Mapping the Extreme Right in Contemporary Europe: From local to transnational*, ed. por Andrea Mammone, Emmanuel Godin y Brian Jenkins, 254-269. Routledge.
- Yang, Ziheng y Carlos E. Rodríguez. 2013. “Searching for efficient Markov chain Monte Carlo proposal kernels”. *Proceedings of the National Academy of Sciences*. <https://www.pnas.org/content/early/2013/11/06/1311790110>.
- Zafimehy, Marie. 2017. *Florian Philippot quitte le Front National: retour sur un désamour en 6 dates*. Consultado el 12 de noviembre de 2017. <http://www rtl fr/actu/politique/florian-philippot-quitte-le-front-national-retour-sur-un-desamour-en-6-dates-7790176224>.
- Zepeda Herrera, Fernando Antonio. 2018. *Modelo de agregación de encuestas*. Resumen metodológico del modelo de agregación de encuestas para la elección presidencial de México para el periodo 2018-2024. Numérica, Ciudad de México. <https://www.numerika.mx/spanish/elecciones-2018/poll-of-polls/>.

- . 2015. “Sobre la conexión estadística de Gauss y Galton”. *Laberintos e Infinitos* 2 (37): 26-30.