

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**Diseño Bayesiano de experimentos
para modelos lineales
generalizados**

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA

GIAN CARLO DI-LUVI MARTÍNEZ

ASESOR: DR. ERNESTO JUVENAL BARRIOS ZAMUDIO

CIUDAD DE MÉXICO

2018

“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 “**Diseño Bayesiano de experimentos para modelos lineales generalizados**”, 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”.

GIAN CARLO DI-LUVI MARTÍNEZ

FECHA

FIRMA

“Essentially, all models are wrong, but some are useful.”
George E. P. Box

Dedicatoria

A Regina, Marcello, Regina, y Sarah.

Agradecimientos

A Sarah, a falta de palabras suficientes, por todo. “*Your mind is my treasure, and if it were broken, it would be my treasure still.*”

A mamá, porque sin tu apoyo incondicional nunca hubiera terminado esta aventura.

A mi familia, por siempre estar ahí para mí a lo largo de este trayecto.

A mi asesor, Ernesto Barrios, por guiarme en este camino e impulsarme a dar lo mejor de mí.

A mis sinodales, Manuel Mendoza, Luis Enrique Nieto y Gabriel Nuñez, por su tiempo, paciencia e interés. Sin ellos este trabajo no se hubiera completado.

A todos los profesores que compartieron sus conocimientos conmigo. Particularmente a Claudia Gómez, Manuel Mendoza, Ernesto Barrios, Vladimir Caetano, César Luis García, Marcela González, Guillermo Grabinsky, Ernesto Pérez Chavela y Julia Sierra.

A la Bandera y a $(\varepsilon - \delta)$.

Índice general

1. Introducción	1
1.1. Estructura de la tesis	3
2. Estadística Bayesiana	5
2.1. Problemas de decisión	7
2.2. Axiomas de coherencia	9
2.3. Probabilidad y utilidad	11
2.4. El Teorema de Bayes	13
2.5. Estadística Bayesiana en la práctica	15
3. Modelos lineales generalizados	19
3.1. Planteamiento	21
3.2. Funciones liga	24
3.3. Modelo	25
3.4. Estimación de los parámetros	25
3.5. Matriz de información de Fisher	28
3.6. Algunos modelos lineales generalizados	29
3.6.1. Modelo de regresión lineal	30
3.6.2. Modelo de regresión logística	33
3.6.3. Modelo de regresión Poisson	35
4. Diseño estadístico de experimentos	37
4.1. Diseño Bayesiano de experimentos	39
4.2. Funciones de pérdida	43

4.3. Método ACE	47
4.4. Comentarios finales	52
5. Implementación de diseños Bayesianos	55
5.1. Diseño para modelo de regresión logística	56
5.2. Diseño para modelo de regresión Poisson	66
6. Conclusiones	83
Apéndices	88
A. Métodos computacionales Bayesianos	89
A.1. Métodos de Monte Carlo	90
A.1.1. Muestreo por importancia	91
A.1.2. Muestreo-remuestreo	92
A.2. Monte Carlo vía cadenas de Markov	93
A.2.1. Muestreo de Gibbs	95
B. Procesos Gaussianos	97
C. Código	103
C.1. Diseño para modelo de regresión logística	103
C.2. Diseño para modelo de regresión Poisson	106
Referencias	119

Capítulo 1

Introducción

Statistics is the grammar of science.

— Karl Pearson

La idea de que un estadístico es aquel a quien se le proporcionan datos para que los analice no podría estar más lejos de la realidad. Para que un estudio se lleve a cabo satisfactoriamente un estadístico debe estar involucrado desde antes de la recolección de los datos. Más aún, el trabajo del estadístico no termina con la conclusión del estudio. La investigación e innovación científica no es estática ni aislada, sino un proceso de autocorrección en el cual la experiencia previa guía y sugiere el camino a seguir (Box y Liu, 1999).

El diseño de experimentos se refiere a la selección de todos los aspectos relevantes de un experimento y se realiza antes de la recolección de los datos. Debido a esto el paradigma Bayesiano resulta idóneo, ya que ofrece una manera de incorporar conocimientos previos al análisis (Bernardo y Smith, 2000; Gutiérrez-Peña, 2016b; Mendoza y Regueiro, 2011). Sin embargo, una de las desventajas de utilizar el enfoque Bayesiano es la latente complejidad computacional (Casella y George, 1992; Geman y Geman, 1984). En el caso del diseño de experimentos ésta se traduce en la dificultad de evaluar funciones de pérdida sofisticadas que engloben los objetivos del experimen-

to, así como integrales de dimensión sumamente alta (Woods et al., 2017).

Chaloner y Verdinelli (1995) publicaron un artículo de revisión que ha servido como punto de referencia para el diseño Bayesiano de experimentos. Sin embargo durante un tiempo hubo un progreso práctico poco sustancial en este respecto, logrando encontrar soluciones solamente a problemas particulares. No obstante los últimos años han visto el surgimiento de algoritmos complejos que pretenden atacar estos problemas y mantener una eficiencia computacional tolerable. Hoy en día el método de *Intercambio Aproximado de Coordenadas* (ACE por sus siglas en inglés) de Overstall y Woods (2016) se postula como la opción más efectiva para resolver problemas de diseño Bayesiano de experimentos, logrando encontrar diseños óptimos para una variedad de modelos y con la opción de personalizar la función de pérdida. Sin embargo su eficiencia depende altamente del número de ensayos y de la función de pérdida empleada. Más aún, su implementación no es trivial, lo que lo hace poco accesible para el público en general.

El diseño Bayesiano de experimentos generalmente aprovecha modelos populares y conocidos, como los modelos lineales generalizados (Nelder y Wedderburn, 1972). Estos son de particular interés porque permiten modelar una gran variedad de fenómenos, pero a la vez han sido estudiados con detalle (Gelfand y Gosh, 2000; McCullagh y Nelder, 1983; Myers et al., 2002). El método ACE es más sencillo de implementar si se utiliza en conjunto con este tipo de modelos, lo que se traduce en un menor costo computacional, particularmente si se utilizan las funciones de pérdida que éste ya incluye.

Hoy en día disciplinas como la ingeniería, la medicina, y las ciencias sociales realizan experimentos estadísticamente diseñados (Fisher, 1935; Woods et al., 2006). Por lo tanto avances en la literatura sobre diseño experimental se traducen en una mayor variedad de diseños al alcance de más investigadores, incrementando la validez y reproducibilidad de los estudios científicos. Es por ello que impulsar la investigación sobre diseño de experimentos es imperativo, más aún considerando que siguen existiendo multiplicidad de

complicaciones prácticas que restringen las opciones disponibles al diseñar un experimento.

El propósito de esta tesis es doble. Por un lado se desarrollarán los elementos teóricos detrás del diseño Bayesiano de experimentos. Esto es importante porque la gran mayoría de la literatura al respecto supone que el lector tiene conocimientos avanzados sobre diseño experimental y Estadística Bayesiana. Por lo mismo es difícil encontrar fuentes que contengan una revisión razonablemente completa del tema en cuestión. Más aún, se supondrá que el propósito último es modelar la relación entre algunas de las covariables del experimento y la variable respuesta mediante un modelo lineal generalizado.

Por otro lado se implementará el método ACE para resolver desde un enfoque Bayesiano dos problemas de diseño experimental, con el objetivo de constatar la eficacia de dicho algoritmo. En particular se replicarán dos de los resultados de Woods et al. (2017), los cuales suponen que la relación entre las covariables y la respuesta se describe mediante un modelo lineal generalizado. En ambos ejemplos se emplearán tanto las distribuciones iniciales propuestas por los autores como distribuciones diferentes, y además en el segundo experimento se utilizará también una función de pérdida distinta.

1.1. Estructura de la tesis

En el Capítulo 2 se plantean los fundamentos de la Teoría de la Decisión que dan lugar a la Estadística Bayesiana. Para ello se discuten las limitaciones del enfoque clásico o frecuentista, se habla de problemas de decisión, y se discute su solución. Además se habla de la importancia del Teorema de Bayes, así como de la Estadística Bayesiana en la práctica. También se discute la complejidad computacional inherente al cálculo de la distribución final, y en el Apéndice A se muestran algunos de los métodos computacionales más populares para sobrepasar dicha dificultad.

En el Capítulo 3 se derivan los modelos lineales generalizados desde la perspectiva de componentes sistemática y aleatoria. Se hace particular énfasis en los modelos de regresión lineal, logística, y Poisson, los cuales se estudian a detalle en el mismo capítulo. Se discute también la estimación de los parámetros desde la perspectiva Bayesiana, particularmente para el caso de regresión lineal.

En el Capítulo 4 se formula el problema de encontrar un diseño óptimo como uno de decisión utilizando las herramientas del Capítulo 2. Así pues, se deduce la solución genérica de cualquier problema de diseño experimental y se mencionan las dificultades prácticas de encontrar dicha solución. También se discute sobre la importancia de la función de pérdida o utilidad, además de que se revisan las elecciones más populares en la literatura. Finalmente se revisa con detalle el método ACE de Overstall y Woods (2016).

En el Capítulo 5 se reproducen dos ejemplos de diseño Bayesiano de experimentos para modelos lineales generalizados tomados de Woods et al. (2017). El primero supone un modelo de regresión logística y el segundo uno de regresión Poisson. Se utilizarán funciones iniciales para los parámetros diferentes a las propuestas por los autores para resaltar la importancia de éstas, y además para el segundo caso se utilizará una función de pérdida diferente también. El código utilizado para este capítulo se incluye en el Apéndice C.

En el Capítulo 6 se da un resumen global de la tesis y se discuten los principales hallazgos, particularmente los referentes al Capítulo 5. Además se comentan las principales dificultades de la implementación del método ACE, así como sugerencias para posibles investigaciones futuras.

Capítulo 2

Teoría de la Decisión y Estadística Bayesiana

Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterising how an individual should act in order to avoid certain kinds of undesirable behavioural inconsistencies.

— Bernardo y Smith, *Bayesian Theory* (2000)

Aunque no se ha llegado a un consenso sobre una definición general de la Estadística, para propósitos de esta tesis ésta se definirá como *un conjunto de técnicas cuyo propósito es describir fenómenos que se manifiestan a través de datos que presentan variabilidad* (Mendoza y Regueiro, 2011). Esta definición resalta el ámbito de estudio (datos que presentan variabilidad) y el propósito (describir) de la Estadística.

Si bien el razonamiento y estudio de la Estadística se remonta al Siglo XVII,¹ no fue sino hasta el Siglo XX que ésta se formalizó (o, quizás mejor dicho, se “matematizó”) gracias a las contribuciones de varios personajes,

¹Siglo en el que se comenzó a desarrollar la Teoría de la Probabilidad. La Estadística contemporánea no se desarrolló sino hasta el Siglo XIX.

entre los cuales destacan Karl Pearson (1857-1936), Ronald Fisher (1890-1962) y Jerzy Neyman (1894-1981), entre muchos otros. Fue en estos años cuando los académicos de la época decidieron tratar de encontrar soluciones que tuvieran fundamentos matemáticos sólidos a los problemas que comúnmente se enfrentaban en la Estadística. El resultado de esta hazaña es lo que hoy se conoce como *Estadística Matemática*; ésta surge como una serie de métodos o procedimientos que permiten resolver problemas como el de estimación puntual, estimación por regiones, pronósticos puntuales y por regiones, y contraste de hipótesis. A pesar de que la Estadística Matemática resultó ser (y sigue siendo) sumamente útil, algunos problemas comenzaron a surgir en los siguientes años; principalmente:

1. La Estadística Matemática no es una teoría, entendida en el sentido axiomático. No hay una serie de postulados de donde se puedan deducir simultáneamente todos los métodos que la conforman.²
2. Como consecuencia de 1. pueden existir inconsistencias entre métodos. Algunas soluciones a un problema pueden no tener sentido, o incluso pueden contradecir otras soluciones.

En años posteriores al desarrollo de la Estadística Matemática algunos estadísticos de la época trataron de solucionar estos problemas. Hay dos vertientes generales que logran este objetivo y desembocan en la Estadística Bayesiana: los teoremas de representación de Bruno de Finetti y la Teoría de la Decisión. En esta tesis se abordará la vertiente de la Teoría de la Decisión. El problema de diseño de experimentos (que se discute en el Capítulo 4) se trata también desde esta perspectiva.

El propósito de este capítulo es explicar los fundamentos de la Teoría de la Decisión y de la Teoría de la Inferencia Bayesiana necesarios para producir inferencia en el caso de los modelos lineales generalizados y para tratar el diseño de experimentos. Cabe mencionar que este capítulo no pretende

²Esto contrasta fuertemente con la Teoría de la Probabilidad, que tiene como postulados a los axiomas de Kolmogorov (1954).

desarrollar de manera profunda la Teoría de la Decisión; tampoco tiene como objetivo deducir la relación entre ésta y la Estadística Bayesiana más allá de lo necesario para su aplicación en capítulos subsecuentes. Si el lector está interesado en estos temas se le invita a consultar la bibliografía recomendada; en particular, ver (Bernardo y Smith, 2000; Mendoza y Gutiérrez-Peña, 2010; Mendoza y Regueiro, 2011).

2.1. Problemas de decisión

Para el ser humano siempre ha sido de interés estudiar el proceso de toma de decisiones. Como Mendoza y Gutiérrez-Peña (2010) mencionan, en la literatura generalmente hay dos formas de abordar este tema: una descriptiva y una normativa. La descriptiva pretende explicar cómo los agentes verdaderamente toman decisiones, y la normativa explica cómo *deberían* tomarlas para satisfacer algún o algunos criterios previamente establecidos. Esta tesis se centrará en una Teoría de la Decisión normativa.

Un problema de decisión se refiere a la situación en la que un *tomador de decisiones* debe seleccionar una, y solo una, opción de un conjunto de posibles decisiones, \mathbb{D} . Desde luego que esta decisión debe ser óptima en algún sentido. En particular las decisiones deben estar relacionadas con sus respectivas consecuencias, y es con éstas que la optimalidad de la decisión debe ser juzgada. De hecho debe existir una relación de orden \succ en el conjunto de consecuencias, \mathbb{C} , que permita determinar cuál de éstas es más preferible. La terna $(\mathbb{D}, \mathbb{C}, \succ)$ se conoce como *problema de decisión con certeza* cuando a cada decisión en \mathbb{D} se le asocia, sin incertidumbre, una consecuencia en \mathbb{C} . Sin embargo, en la práctica cuando una decisión ha sido seleccionada generalmente hay una serie de consecuencias que pueden ocurrir; más aún, no hay certeza sobre cuál consecuencia ocurrirá. Así pues a cada decisión se le asocia también un conjunto de eventos inciertos relevantes, cada uno de los cuales está a su vez asociado con una consecuencia. Aquí ya estamos hablando de un *problema de decisión en ambiente de incertidumbre*, cuya definición se enuncia a continuación para futura referencia.

Definición 2.1.1 (Problema de Decisión). Un problema de decisión en ambiente de incertidumbre es una cuarteta $(\mathbb{D}, \mathbb{E}, \mathbb{C}, \succ)$, donde:

1. $\mathbb{D} = \{d_1, d_2, \dots, d_k\}$ es el conjunto de posibles decisiones.
2. \mathbb{E} es el conjunto de eventos inciertos relevantes, de tal forma que a cada decisión d_i le corresponde un conjunto de eventos inciertos relevantes $\mathbb{E}_i = \{E_{i1}, E_{i2}, \dots, E_{in_i}\} \subset \mathbb{E}$, donde $E_{ij} \cap E_{ij'} = \emptyset$ para $j \neq j'$.
3. \mathbb{C} es el conjunto de consecuencias, de tal forma que a cada evento incierto relevante E_{ij} le corresponda unívocamente una consecuencia $c_{ij} \in \mathbb{C}$.
4. \succ es una relación de orden en \mathbb{C} que determina, dadas dos consecuencias, cuál es más preferible (o si son igualmente preferibles) para el tomador de decisiones.

Los eventos inciertos relevantes deben considerar todas las consecuencias, de manera que, para cualquier i ,

$$\bigcup_j E_{ij} = \Omega,$$

donde Ω es el evento seguro.³ En otras palabras, \mathbb{E}_i es una partición del evento seguro. En lo que resta de este capítulo siempre se pensará en un problema de decisión genérico como el recién definido.

Una manera gráfica de representar un problema de decisión es mediante un *árbol de decisión*. Dado un problema de decisión $(\mathbb{D}, \mathbb{E}, \mathbb{C}, \succ)$ cualquiera, su árbol de decisión asociado es el siguiente:

³El evento seguro es aquel que ocurre con probabilidad 1, y se refiere al hecho de que los eventos inciertos relevantes asociados a una decisión deben considerar todas las posibilidades a ocurrir.

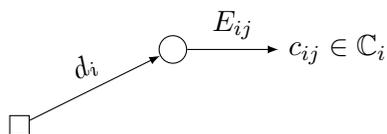


Figura 2.1: Árbol de decisión para un problema de decisión genérico como el de la Definición 2.1.1.

El nodo inicial es un cuadrado, lo que denota un *nodo de decisión*. El tomador de decisiones selecciona la opción d_i del conjunto de opciones. Sin embargo esta decisión tiene asociados los eventos inciertos relevantes en \mathbb{E}_i . En el diagrama esto se denota mediante un nodo circular llamado *nodo de incertidumbre*. Cada evento incierto relevante a su vez tiene asociada una consecuencia, la cual se encuentra al final del árbol de decisión. En resumen, el tomador de decisiones toma una decisión que se enfrenta a una incertidumbre y que deriva en una cierta consecuencia.

2.2. Axiomas de coherencia

El núcleo de la inferencia Bayesiana recae en los llamados *axiomas de coherencia*, pues son los postulados básicos de la Teoría de la Decisión. Estos se enunciarán de forma análoga a como lo hacen Mendoza y Regueiro (2011, Capítulo 3.1), aunque se recomienda revisar dicho reporte técnico, así como Bernardo y Smith (2000, Capítulo 2.3), para una discusión más detallada.

1. **Comparabilidad.** Dados cualesquiera $d_i, d_j \in \mathbb{D}$, debe ocurrir una, y solo una, de las siguientes:
 - $d_i \succ d_j$ (d_i es más preferido que d_j).
 - $d_j \succ d_i$ (d_j es más preferido que d_i).
 - $d_i \sim d_j$ (d_i y d_j son igualmente preferidos).

Además, existen $c_*, c^* \in \mathbb{C}$ tales que $c_* \preceq c \preceq c^*$ para todo $c \in \mathbb{C}$.

2. **Transitividad.** Dados $d_i, d_j, d_k \in \mathbb{D}$, si $d_i \succ d_j$ y $d_j \succ d_k$ entonces debe ocurrir que $d_i \succ d_k$. Análogamente, si $d_i \sim d_j$ y $d_j \sim d_k$ entonces necesariamente $d_i \sim d_k$.
3. **Sustituibilidad.** Si $d_i, d_j \in \mathbb{D}$ son opciones y A es un evento incierto tal que $d_i \succ d_j$ si ocurre A y $d_i \succ d_j$ si ocurre A^c , entonces $d_i \succ d_j$. Análogamente, si $d_i \sim d_j$ cuando ocurre A y también cuando ocurre A^c entonces $d_i \sim d_j$.
4. **Eventos de referencia.** Independientemente de los eventos inciertos relevantes, el tomador de decisiones puede imaginar un procedimiento para generar puntos en el cuadrado unitario I de manera que, para cualesquiera dos regiones $R_1, R_2 \in I$, el evento $A_1 = \{z \in R_1\}$ es más creíble que el evento $A_2 = \{z \in R_2\}$ si, y solo si, $\text{Área}(R_1) > \text{Área}(R_2)$.

Los primeros dos axiomas simplemente proveen de estructura a los conjuntos de consecuencias y opciones, mientras que el tercero define una forma precisa de coherencia en la toma de decisiones. El cuarto axioma, aunque puede parecer extraño a primera impresión, simplemente dice que el tomador de decisiones en cuestión puede concebir la idea de una distribución uniforme en el cuadrado unitario de \mathbb{R}^2 , la cual no está relacionada con los eventos inciertos relevantes.

Estos axiomas, o versiones similares, se han presentado en la literatura en distintas formas. La versión que aquí se presenta tiene como antecedente el libro de Leonard Jimmie Savage (1954). En cualquier caso, estos principios son los que garantizan que la Estadística Bayesiana tenga carácter de teoría, a diferencia de la Estadística Matemática. Los resultados que se discuten en lo que resta del capítulo pueden deducirse enteramente a partir de los axiomas de coherencia. Sin embargo, ya que esto rebasa el propósito de esta tesis, se recomienda al lector interesado revisar la bibliografía recomendada si desea ver los detalles de esas demostraciones.

2.3. Probabilidad y utilidad

Un tema que es de suma importancia y que aparece de manera natural en el enfoque Bayesiano es el de la probabilidad. En la vertiente Bayesiana la probabilidad tiene una interpretación *subjetiva*: cualquier evento sobre el que existe incertidumbre es sujeto a que se le asigne una probabilidad; además, ésta es personal y depende de la información disponible para quien la asigne. Los axiomas de coherencia ofrecen una manera objetiva de determinar la probabilidad que un agente atribuye a algún evento incierto. Algo que se puede demostrar es que los axiomas de Kolmogorov (1954) son consecuencia directa de la definición de probabilidad de Teoría de la Decisión. De esta manera, todos los resultados de la Teoría de la Probabilidad se pueden (y se deben) utilizar en el ámbito de la Teoría de la Decisión.

La idea natural es asignar una probabilidad de ocurrencia a cada uno de los eventos inciertos relevantes. Esto es, dado un problema de decisión $(\mathbb{D}, \mathbb{E}, \mathbb{C}, \succ)$, el tomador de decisiones debe asignar una distribución de probabilidades $p(\cdot)$ sobre cada \mathbb{E}_i . Por lo dicho anteriormente, p resulta ser una distribución en el sentido habitual de la Teoría de la Probabilidad.

Otro elemento importante en el enfoque Bayesiano es el de utilidad. Las consecuencias son un elemento fundamental en la solución de un problema de decisión; sin embargo la trascendencia específica de cada consecuencia solo se registra a través del beneficio que le produce al tomador de decisiones. De los axiomas se sigue que es posible y necesario definir una función $u : \mathbb{C} \rightarrow \mathbb{R}$ que asigna, a cada consecuencia, una utilidad numérica. De esta forma comparar consecuencias se reduce a comparar números reales. A veces es común escribir $u(d, E)$, con $d \in \mathbb{D}$ y $E \in \mathbb{E}$. También es posible (y a veces intuitivamente más sencillo) trabajar con funciones de *pérdida*, es decir, funciones que cuantifiquen el perjuicio que cada consecuencia produce al tomador de decisiones.

Es sencillo incorporar la probabilidad y la utilidad en los árboles de deci-

sión, y el árbol 2.1 revisado previamente tomaría la siguiente forma:

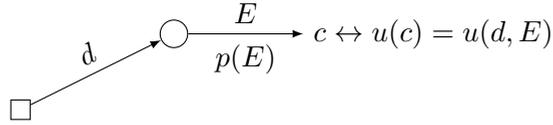


Figura 2.2: Árbol de decisión para un problema de decisión genérico incorporando la probabilidad y la utilidad.

No está de más insistir en que la utilidad, al igual que la probabilidad, se puede definir formalmente en el contexto de la Teoría de la Decisión. Además, si bien la utilidad que un tomador de decisiones asigna a cada consecuencia es subjetiva, la Teoría de la Decisión también ofrecen una manera de medir dicha utilidad. Estos dos conceptos se juntan para dar lugar a un tercer concepto que aparece de forma natural en el enfoque Bayesiano: el de utilidad esperada, el cual se define a continuación.

Definición 2.3.1 (Utilidad esperada). La utilidad esperada de una decisión $d_i \in \mathbb{D}$ se define como

$$\mathbb{E} [u(d_i)] = \sum_{E \in \mathbb{E}_i} u(d_i, E)P(E), \quad (2.1)$$

donde $P(E)$ denota la probabilidad que el tomador de decisiones asigna al evento incierto relevante E .

Esta definición se puede generalizar si el espacio de eventos inciertos relevantes es no numerable simplemente cambiando la suma por una integral. Además, la definición también es válida para funciones de pérdida en vez de utilidad (y se llama pérdida esperada).

Con estos elementos es posible caracterizar la solución de cualquier problema de decisión. Como (Bernardo y Smith, 2000, Capítulo 2.5) demuestran, los axiomas de coherencia implican que la solución de un problema de decisión es aquella decisión que maximice la utilidad esperada. En otras

palabras, si $(\mathbb{D}, \mathbb{E}, \mathbb{C}, \succ)$ es un problema de decisión con función de utilidad u , entonces la solución d^* es tal que

$$d^* = \arg \max_{d \in \mathbb{D}} \mathbb{E} [u(d)]. \quad (2.2)$$

Si en el problema se está trabajando con una función de pérdida, entonces se debe *minimizar* la pérdida esperada.

2.4. El Teorema de Bayes

Hasta ahora se ha discutido cómo resolver problemas de decisión genéricos. Sin embargo es claro que dicha solución depende de la información (o incertidumbre) que el tomador de decisiones tiene. Luego, la solución es en realidad estática y se refiere a un punto específico en el tiempo. Pero, citando a Dennis Lindley (2006), “uncertainty is a personal matter; it is not *the* uncertainty but *your* uncertainty.” ¿Qué pasa si el tomador de decisiones obtiene nueva información, en forma de datos por ejemplo? En esta sección se enunciará el Teorema de Bayes, que es el mecanismo que permite incorporar información adicional al análisis. La demostración del Teorema de Bayes se obviará, ya que es una consecuencia sencilla de los axiomas de Kolmogorov (y, por lo discutido en la sección anterior, lo es también de la definición de probabilidad en Teoría de la Decisión), y se puede encontrar en cualquier libro introductorio a Teoría de la Probabilidad (por ejemplo Ross, 2012).

Teorema 2.4.1 (de Bayes). *Sean E y F eventos inciertos tales que $P(F) > 0$. Entonces*

$$P(E | F) = \frac{P(F | E)P(E)}{P(F)}. \quad (2.3)$$

Pensando en un problema de decisión, F toma el lugar de la nueva información y $E \in \mathbb{E}_i$ será un evento incierto relevante para alguna decisión. En particular, como \mathbb{E}_i forma una partición del evento seguro, podemos utilizar el teorema de probabilidad total para escribir (2.3) como

$$P(E_{ij} | F) = \frac{P(F | E_{ij})P(E_{ij})}{\sum_k P(F | E_{ik})P(E_{ik})}.$$

Finalmente note que el denominador de la expresión anterior, cuando se considera como función de E_{ij} , no es más que una constante de normalización: no depende de E_{ij} . Es por ello que el Teorema de Bayes, en el contexto de la Estadística Bayesiana, generalmente se escribe

$$P(E_{ij} | F) \propto P(F | E_{ij})P(E_{ij}), \quad (2.4)$$

donde \propto se lee “es proporcional a”.

La interpretación de la expresión (2.4) es interesante. $P(E_{ij} | F)$ describe la probabilidad de los eventos inciertos relevantes *después* de haber observado la nueva información F , por lo que se conoce como *distribución final* o a posteriori de E_{ij} , y es proporcional al producto de dos factores. Uno de ellos, $P(E_{ij})$, es la distribución de probabilidad que describe la incertidumbre sobre los eventos inciertos *antes* de observar la nueva información, conocido como la *distribución inicial* o a priori de E_{ij} . El otro, $P(F | E_{ij})$, describe la probabilidad de haber observado la nueva información, dado el evento incierto relevante E_{ij} . Este factor se conoce como verosimilitud, ya que coincide con la función de verosimilitud encontrada en la literatura estadística. Note que es necesario conocer la forma de la verosimilitud, $P(F | E_{ij})$. De no ser así es imposible actualizar la incertidumbre que se tiene sobre los eventos inciertos relevantes vía el Teorema de Bayes. Esto tiene sentido si se piensa que el propósito de obtener nueva información es disminuir la incertidumbre. Luego, la información adicional *debe* estar relacionada con la fuente de la incertidumbre del problema de decisión, que son precisamente los eventos inciertos relevantes y, además, se debe conocer esta relación.

El Teorema de Bayes resulta ser entonces de importancia práctica en la Estadística Bayesiana. Se puede deducir de los axiomas de coherencia (ya que estos implican los axiomas de Kolmogorov), y es la forma de actualizar las probabilidades de los eventos inciertos relevantes que el tomador de decisiones asignó inicialmente a la luz de nueva información.

2.5. Estadística Bayesiana en la práctica

Hasta ahora se ha discutido el fundamento que proporciona a la Estadística Bayesiana el carácter de teoría axiomática, la Teoría de la Decisión, así como el mecanismo óptimo para incorporar información adicional al análisis. En esta sección se discutirá, brevemente, cómo se aplica esto en problemas de inferencia estadística simples. En particular, se discutirán problemas de inferencia paramétrica.

Como se discutió en la sección anterior, lo único necesario para resolver un problema de decisión es maximizar la utilidad esperada. Es entonces necesario definir las posibles opciones, los eventos inciertos relevantes con su distribución de probabilidades respectiva y las consecuencias con su respectiva función de utilidad o pérdida. En la práctica una decisión se refiere a la descripción del fenómeno que se desea utilizar. Por ejemplo, una decisión puede ser el valor del parámetro que se desea utilizar como estimador puntual; o los extremos de un intervalo si se desea realizar estimación por intervalos; también puede ser la acción de aceptar o rechazar una hipótesis nula en el contexto de contraste de hipótesis.

Los eventos inciertos relevantes generalmente son los posibles valores del parámetro. De esta forma la distribución de probabilidades se *debe* asignar sobre los valores del parámetro. Esto contrasta fuertemente con la vertiente frecuentista que, si bien reconoce la incertidumbre sobre los parámetros, no admite la idea de cuantificar dicha incertidumbre vía una distribución de probabilidades.

Finalmente la función de utilidad o pérdida generalmente mide la calidad de la decisión. Si, por ejemplo, para estimación puntual consideramos una función de pérdida, una posibilidad podría ser la desviación absoluta del valor que se elige como estimación con respecto al verdadero valor del parámetro.

Con estos elementos ya es posible resolver el problema. Sin embargo, hay

un tema importante que se ha obviado hasta el momento: la muestra. La definición de Estadística enunciada al inicio de este capítulo resalta que la Estadística tiene como objetivo describir fenómenos que se manifiestan a través de datos. Empero, no se ha hablado del papel que los datos juegan en el paradigma Bayesiano. Supongamos entonces que se tiene una muestra de estos datos. Ésta, que debe provenir de algún modelo de muestreo conocido, modifica la distribución de probabilidades que se asignó inicialmente sobre los valores del parámetro, vía el Teorema de Bayes. Es por ello que usualmente se habla de la distribución inicial o *a priori*, y la distribución final o *a posteriori*. Así pues, la solución al problema se encuentra maximizando la utilidad esperada posterior, es decir, aquella que se calcula con respecto a la distribución final.

Lo anterior se puede resumir en lo que Gutiérrez-Peña (2016b) denomina *el proceso de aprendizaje Bayesiano*, que consta de cuatro pasos:

1. Especificación de un modelo de muestreo o verosimilitud, $p(x | \theta)$.
2. Especificación de una distribución inicial, $p(\theta)$.
3. Cálculo de la distribución final, $p(\theta | x_1, \dots, x_n)$, vía el Teorema de Bayes.
4. Resolución del problema de decisión con la distribución final.

Este es el procedimiento Bayesiano por excelencia. Todos los problemas de inferencia se resuelven, en mayor o menor medida, de esta forma. No hay que olvidar que este proceso de aprendizaje tiene como fundamento a la Teoría de la Decisión. Los distintos valores del parámetro son los eventos inciertos y la función de utilidad o pérdida es la que determina el tipo de solución que se tiene. Por ejemplo, en el problema de estimación puntual una función de pérdida cuadrática deriva en el valor esperado de la distribución final como solución. En la práctica se calcula la distribución final independientemente del tipo específico de inferencia que se busque producir, ya que ésta contiene toda la información sobre los eventos inciertos. De

hecho, además de contribuir a la solución del problema de decisión específico, con ella se pueden calcular resúmenes para comprender la información que contiene, como su media, mediana, moda, quintiles, o intervalos de *probabilidad*.

Otro tema relevante es el de la especificación de la distribución inicial. Si bien en muchos casos la persona que realiza el estudio tiene relativamente claros los conocimientos iniciales,⁴ en muchos otros casos o bien no se tienen conocimientos iniciales, o bien estos se quieren omitir deliberadamente.⁵ En esta situación es de interés utilizar alguna distribución inicial cuyo impacto en el análisis sea mínimo. Aquí aparece el concepto de distribuciones iniciales *mínimo informativas*, a veces también llamadas de referencia. Mendoza y Regueiro (2011, Capítulo 6.4) discuten algunos de los métodos más populares para obtener este tipo de distribuciones, de entre los cuales destaca el método de Jeffreys. Éste asigna como distribución inicial para el parámetro la raíz cuadrada del determinante de la matriz de información de Fisher, la cual se estudiará en el siguiente capítulo con mayor detalle.

Finalmente es importante comentar que, si bien en principio el proceso de aprendizaje Bayesiano es sencillo, en la práctica calcular la distribución final puede no serlo. Existe toda un área de investigación que se dedica a diseñar métodos eficientes para realizar este cálculo, la mayoría de los cuales se desarrollaron a principios de los 90 y pertenecen a la familia conocida como *Markov chain Monte Carlo (MCMC) methods*. Para una discusión histórica sobre el primero de estos algoritmos (el de Metropolis-Hastings) se recomienda revisar (Hitchcock, 2003); para una deducción más formal del algoritmo, así como de otros métodos populares (e.g. el muestreo de Gibbs) se recomiendan (Chib y Greenberg, 1995; Gutiérrez-Peña, 2016c).

⁴Por ejemplo, si ya se había hecho un estudio similar es posible incorporar los resultados en la distribución inicial.

⁵Usualmente esto ocurre cuando el investigador o investigadora en cuestión desea realizar el análisis desde una postura neutral, como es el caso de los conteos rápidos en México.

Además, en el Apéndice A se discuten con mayor detalle algunos de estos métodos.

Capítulo 3

Modelos lineales generalizados

Thus the problem of looking intelligently at data demands the formulation of patterns that are thought capable of describing succinctly not only the systematic variation in the data under study, but also for describing patterns in similar data that might be collected by another investigator at another time and in another place.

— McCullagh y Nelder, *Generalized Linear Models* (1983)

Una de las familias de modelos de mayor importancia en la Estadística es la de los modelos lineales. El análisis de regresión lineal es útil porque permite modelar la variable de interés (o respuesta) como función lineal de una o más variables independientes (o covariables). Además, la teoría matemática y estadística detrás de los modelos lineales es elegante y ha sido muy estudiada.

El análisis de regresión se remonta al Siglo XIX, cuando Carl Friedrich Gauss y Adrien-Marie Legendre desarrollaron y aplicaron un primer método para datos astronómicos. Sus datos generalmente eran mediciones de cantidades continuas, como posiciones y magnitudes de diversos astros. La

variabilidad existente en dichos datos se derivaba de errores de medición, y fue por ello que Gauss introdujo la distribución Normal para describir el comportamiento de dichos errores. Cabe resaltar que Gauss notó que muchas de las propiedades de los estimadores del modelo de regresión lineal no dependen de la normalidad de los errores, sino de su independencia y homocedasticidad, resultado que hoy lleva su nombre.

Comúnmente, cuando se trabaja con datos, se busca algún patrón que los describa. Esta es la componente sistemática del análisis, y generalmente enfrenta a una variabilidad en los datos que en mayor o menor medida dificulta la percepción del patrón. Esta última es la componente aleatoria. Todos los modelos estudian ambas componentes, otorgando un resumen de los datos en sus componentes sistemáticas y un análisis de la componente aleatoria. En principio uno podría pensar que un buen modelo es simplemente aquel cuya predicción tiene el menor error posible. Se puede pensar en incluir un número muy alto de parámetros en nuestro modelo, lo que generalmente mejoraría cada vez más la calidad de la predicción en el conjunto de las observaciones; sin embargo, un modelo tal no garantiza que los pronósticos para datos no observados sean correctos. No obstante, ello aumentaría la complejidad del análisis, por lo que la simplicidad del modelo, entendida como *parsimonia*, es también algo importante a considerar. Es por ello que la idea clásica de un modelo lineal es atractiva: permite modelar una variable respuesta (que siga una distribución Normal) como función lineal de una serie de covariables.

En la práctica, sin embargo, muchas veces no ocurre que el comportamiento de la variable respuesta se describa adecuadamente mediante una distribución Normal, bien por las características de ésta (si tiene algún sesgo o es solo positiva, e.g.) o bien porque incluso puede ni siquiera ser continua. Los modelos lineales generalizados, introducidos por Nelder y Wedderburn (1972), permiten hacer análisis análogos a los de los modelos lineales clásicos, pero sin presuponer normalidad en la variable respuesta: basta que

ésta pertenezca a la familia exponencial de distribuciones.¹

El propósito de este capítulo es sentar las bases de los modelos lineales generalizados, de manera que en el siguiente capítulo se pueda ahondar en el diseño experimental con énfasis en dichos modelos. Por la naturaleza de la tesis, el análisis se hará desde un punto de vista Bayesiano, utilizando la notación e ideas introducidas en el capítulo anterior. Para este efecto se seguirán de cerca las notas de Gutiérrez-Peña (2016a) y Nieto (2017) así como los libros de modelos lineales generalizados de Gelfand y Gosh (2000) y de McCullagh y Nelder (1983).

3.1. Planteamiento

En el contexto de los modelos lineales, supondremos que se tiene una variable respuesta de interés Y con media μ y un conjunto de p covariables o regresores x . La componente sistemática del análisis corresponde a la descripción de μ en términos de una cantidad pequeña de parámetros desconocidos β_0, \dots, β_p . El primer paso es pensar en μ como función de x , es decir,

$$\mathbb{E}[Y | x] = \mu(x).$$

Aquí, $\mu(\cdot)$ es una función desconocida, por lo que el siguiente paso es preguntarse por su naturaleza. La manera más general de hacer esto es aproximarla mediante una función paramétrica conocida ψ tal que

$$\mu(x) = \psi(x; \beta),$$

donde $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ es el vector de parámetros desconocidos previamente mencionados. Aquí es donde entra la linealidad, pues por facilidad se supone que existe una función conocida h tal que

$$\psi(x; \beta) = h(\beta_0 + \beta_1 s_1(x) + \dots + \beta_p s_p(x)), \quad (3.1)$$

¹También existe otra sutil generalización que se discutirá más adelante.

donde las funciones s_1, \dots, s_p se suponen suaves y conocidas. Recapitulando, lo que se está planteando es el modelo

$$\mathbb{E}[Y | x] = h(\beta_0 + \beta_1 s_1(x) + \dots + \beta_p s_p(x)). \quad (3.2)$$

Ahora bien, podemos pensar que tenemos un vector de observaciones $y = (y_1, \dots, y_n)^T$, el cual es una realización de $Y = (Y_1, \dots, Y_n)^T$, donde las variables aleatorias que forman a Y son independientes y tienen media μ_i , $i = 1, \dots, n$. En ese caso podemos escribir la componente sistemática como

$$\mathbb{E}[Y_i | x_i] = \mu_i = h(\beta_0 + \beta_1 s_1(x_i) + \dots + \beta_p s_p(x_i))$$

para cada $i = 1, \dots, n$. Aquí, x_i se refiere a los valores observados de las covariables para la i -ésima observación. En la literatura se define el *predictor lineal* η_i como

$$\eta_i = \beta_0 + \beta_1 s_1(x_i) + \dots + \beta_p s_p(x_i).$$

La componente sistemática se resume en el predictor lineal, el cual es producido directamente por las covariables.

La componente aleatoria tiene como objetivo describir el comportamiento de la incertidumbre que naturalmente entra al análisis al trabajar con variables aleatorias. En el caso de modelos lineales generalizados, se supone que las variables respuesta son independientes, lo cual se debe corroborar tanto como sea posible con los datos. Otro supuesto es que además siguen una distribución perteneciente a la familia exponencial, la cual definimos a continuación.²

Definición 3.1.1 (Familia exponencial). La variable aleatoria Y pertenece a la familia exponencial de distribuciones si su función de probabilidad generalizada³ f_Y se puede escribir como

$$f_Y(y | \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

²Aunque la definición puede variar ligeramente según la fuente, para propósitos de esta tesis se seguirá la definición utilizada por (McCullagh y Nelder, 1983, Capítulo 2.2.2).

³Función de densidad si Y es continua, función masa de probabilidad si es discreta.

para funciones monótonas conocidas $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$. A θ se le conoce como parámetro canónico y a ϕ como parámetro de dispersión, ya que tiende a estar relacionado con la variabilidad del modelo.

Generalmente $a(\phi)$ es de la forma

$$a(\phi) = \frac{\phi}{w}, \quad (3.3)$$

donde w es un peso conocido que puede depender del tamaño de muestra. Como ejemplo, en el caso de una colección de n variables aleatorias independientes e idénticamente distribuidas $\mathcal{N}(\mu, \sigma^2)$, la media muestral \bar{Y} es una variable aleatoria Normal también y

$$a(\phi) = \frac{\phi}{n},$$

con $\phi = \sigma^2$. Por simplicidad se supondrá que $a(\phi)$ tiene la forma (3.3) y es siempre conocido, lo cual es válido para las distribuciones consideradas en esta tesis.

Además es posible mostrar que

$$\mathbb{E}[Y] = \mu(\theta) = b'(\theta), \quad (3.4)$$

y que

$$\text{Var}(Y) = b''(\theta)a(\phi) = \mu'(\theta)a(\phi),^4 \quad (3.5)$$

donde la diferenciación se toma con respecto a θ . Es interesante que la varianza de Y es el producto de dos funciones: $b''(\theta)$, que solo depende del parámetro canónico y comúnmente se llama *función de varianza*, y $a(\phi)$, que solamente depende de ϕ y no de θ .

⁴Generalmente existe una relación funcional entre μ y θ , la cual da lugar a la liga canónica mencionada más adelante.

3.2. Funciones liga

Las componentes sistemática y aleatoria se relacionan mediante una *función liga* $g(\cdot)$ que satisface

$$\eta_i = g(\mu_i). \quad (3.6)$$

En el caso de regresión lineal clásico, la variable respuesta sigue una distribución Normal (la cual es un miembro de la familia exponencial) y además la función liga es la función identidad, i.e., $\eta_i = \mu_i$. La generalización es doble: la respuesta puede venir de cualquier distribución de la familia exponencial, y la función liga puede ser cualquier función monótona diferenciable que respete los posibles valores que puedan tomar ambas componentes.

La importancia de la función liga recae en que ésta relaciona el predictor lineal η con la media μ de los datos y . Un caso especial de funciones liga ocurre cuando

$$\theta = \eta,$$

donde θ es el parámetro canónico definido en (3.1.1). Si dicha relación ocurre, la liga en cuestión se llama *liga canónica* y tiene la propiedad de que garantiza la existencia de una estadística suficiente para β , es decir, una función tal que la probabilidad condicional de los datos dada dicha función no depende de los parámetros β .

Sin embargo es claro que algunas restricciones deben existir sobre la función liga en general; particularmente hay que cerciorarse de que verdaderamente haga sentido igualar η y $g(\mu)$. Por ejemplo, si la distribución en cuestión tiene media estrictamente positiva pero el predictor lineal puede tomar valores negativos, claramente la función liga no puede ser la identidad (como es el caso del modelo Poisson). En particular puede ocurrir que la misma liga canónica no se pueda utilizar. A pesar de que utilizar una liga canónica lleve a un modelo con propiedades estadísticas favorables, no se debe sacrificar la validez del modelo en favor de dichas propiedades. La Tabla

3.1 muestra algunas de las ligas canónicas más comunes.

Distribución	Liga canónica
Normal	$\eta = \mu$
Poisson	$\eta = \log(\lambda)$
Binomial	$\eta = \log\left(\frac{p}{1-p}\right)$
Gamma	$\eta = -\frac{1}{\mu}$

Tabla 3.1: Algunas distribuciones importantes con sus respectivas ligas canónicas.

3.3. Modelo

En resumen, los modelos lineales generalizados están caracterizados por tres componentes.

- Componente aleatoria: se tiene una serie de variables aleatorias Y_1, Y_2, \dots, Y_n independientes provenientes de la familia exponencial, cada una con media μ_i .
- Componente sistemática: para cada una de dichas variables Y_i se tiene un vector de covariables $x_i = (x_{i1}, \dots, x_{ip})$ que da lugar al predictor lineal,

$$\eta_i = \beta_0 + \beta_1 s_1(x_i) + \dots + \beta_p s_p(x_i).$$

- Función liga: Las componentes aleatoria y sistemática se relacionan mediante una función liga g , de forma que

$$\eta_i = g(\mu_i).$$

3.4. Estimación de los parámetros

Si bien hasta ahora se ha seguido de cerca el planteamiento propuesto por McCullagh y Nelder (1983), la estimación de los parámetros discutida por

ellos se presenta desde una perspectiva clásica, no Bayesiana. Para este tema se seguirán el artículo de Gelfand y Gosh (2000) y las notas de Nieto (2017).

Consideremos una colección de variables aleatorias independientes Y_1, \dots, Y_n cada una proveniente de la familia exponencial, es decir, tales que

$$f_{Y_i}(y | \theta_i, \phi_i) = \exp \left\{ \frac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y, \phi_i) \right\}, \quad i = 1, \dots, n.$$

Supondremos que el parámetro de dispersión, ϕ_i , es conocido para cada i . En el contexto de los modelos lineales generalizados supondremos además que las componentes sistemática y aleatoria se relacionan mediante la liga canónica y que las funciones s_1, \dots, s_p del predictor lineal son tales que

$$\theta_i = \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}.$$

Se definen por simplicidad $y = (y_1, \dots, y_n)^T$ y $\beta = (\beta_0, \dots, \beta_p)^T$. En ese caso, la función de verosimilitud del modelo es de la forma

$$L(\beta | y) = \exp \left\{ \sum_{i=1}^n \left[\frac{y_i \eta_i - b(\eta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right] \right\} \quad (3.7)$$

o, alternativamente,

$$L(\beta | y) \propto \exp \left\{ \sum_{i=1}^n \frac{y_i \eta_i - b(\eta_i)}{a(\phi_i)} \right\}. \quad (3.8)$$

Desde la perspectiva clásica los parámetros del modelo se estiman vía máxima verosimilitud. En el enfoque Bayesiano se recurre al proceso de aprendizaje mencionado en el Capítulo 2. Las cantidades desconocidas del análisis son los parámetros del modelo, β y, en algunos casos, el parámetro de dispersión ϕ ; a dichas cantidades se les debe asignar una función de densidad inicial $p(\beta, \phi)$ y, utilizando el Teorema de Bayes, se debe encontrar la función de densidad final $p(\beta, \phi | y)$. Finalmente, dependiendo la función de

pérdida en cuestión, se debe resolver el problema de inferencia relevante minimizando la pérdida esperada, calculada con base en la distribución final.

De nuevo suponiendo ϕ_i conocido para cada i , una elección común para la distribución inicial de β (Gelfand y Gosh, 2000; Nieto, 2017) es una distribución $\mathcal{N}(b_0, T_0)$, es decir,

$$\begin{aligned} p(\beta) &= \mathcal{N}(\beta | b_0, T_0) \\ &\propto \exp \left\{ -\frac{1}{2}(\beta - b_0)^T T(\beta - b_0) \right\}. \end{aligned} \quad (3.9)$$

En ese caso la distribución final es

$$\begin{aligned} p(\beta | y) &\propto L(\beta | y)p(\beta) \\ &\propto \exp \left\{ \sum_{i=1}^n \frac{y_i \eta_i - b(\eta_i)}{a(\phi_i)} \right\} \exp \left\{ -\frac{1}{2}(\beta - b_0)^T T(\beta - b_0) \right\} \\ &\propto \exp \left\{ \sum_{i=1}^n \left[\frac{y_i \eta_i - b(\eta_i)}{a(\phi_i)} \right] - \frac{1}{2}(\beta - b_0)^T T(\beta - b_0) \right\}. \end{aligned} \quad (3.10)$$

Si la información inicial que se tiene sobre los parámetros β es mínima o, por alguna razón, no se desea incluirla en el análisis, es posible utilizar alguna distribución inicial mínimo informativa. Gelfand y Gosh (2000, Capítulo 2.2) mencionan que una posibilidad es utilizar la distribución inicial de Jeffreys,

$$p(\beta) \propto |I(\beta)|^{1/2}, \quad (3.11)$$

donde $I(\beta)$ es la matriz de información de Fisher, (3.13).

Es necesario recordar que, como se mencionó en el capítulo anterior, el cálculo de la distribución final tiende a ser sumamente complicado; en muchos casos puede no existir siquiera una forma analítica cerrada para ésta. Los métodos computacionales Bayesianos juegan un papel central en el

análisis Bayesiano de los modelos lineales generalizados, ya que permiten obtener una muestra de la distribución final. A partir de esa base se puede obtener cualquier resumen inferencial de interés (cuya precisión se puede fijar arbitrariamente).

3.5. Matriz de información de Fisher

Un concepto importante para el diseño de modelos lineales generalizados es el de la matriz de información de Fisher para β . En general, dada una muestra aleatoria Y_1, \dots, Y_n proveniente del modelo $f(y_i | \beta)$, la matriz de información de Fisher para β se define como

$$I(\beta) = -\mathbb{E} \left[\nabla^2 \ell(\beta | y) \right], \quad (3.12)$$

donde $\ell(\beta | y) = \log(L(\beta | y))$ es la log-verosimilitud del modelo y ∇^2 se refiere a la matriz Hessiana de esta última.

Proposición 3.5.1. *Consideremos un modelo lineal generalizado de n observaciones y p covariables en el cual la componente aleatoria y sistemática se relacionan mediante la liga canónica, $\eta_i = \theta_i$, y , además, el parámetro de dispersión ϕ_i es conocido para cada i . Entonces la matriz de información de Fisher para β está dada por*

$$I(\beta) = X^T W X, \quad (3.13)$$

donde X es una matriz de $n \times p$ cuyo i -ésimo renglón es x_i , y W es una matriz diagonal de $n \times n$, donde la i -ésima entrada de la diagonal es

$$W_{ii} = [\text{Var}(y_i)]^{-1} (b''(\theta_i))^2, \quad i = 1, \dots, n.$$

Demostración. Note que la log-verosimilitud del modelo está dada por

$$\ell(\beta | y) = \sum_{i=1}^n \left[\frac{y_i \eta_i - b(\eta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right].$$

El segundo término de la suma no depende de β por lo que, al derivar, éste se elimina. Así pues,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^n \frac{1}{a(\phi_i)} \frac{\partial}{\partial \beta_j} (y_i \eta_i - b(\eta_i)) \\ &= \sum_{i=1}^n \frac{x_{ij}}{a(\phi_i)} (y_i - b'(\eta_i)).\end{aligned}$$

Luego,

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \frac{\partial}{\partial \beta_k} \left(\frac{y_i x_{ij}}{a(\phi_i)} \right) - \sum_{i=1}^n \frac{\partial}{\partial \beta_k} \left(\frac{b'(\eta_i) x_{ij}}{a(\phi_i)} \right).$$

La primera suma es cero pues cada uno de los sumandos es a su vez cero: $y_i x_{ij} / a(\phi_i)$ no depende de β_k . Entonces

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n \frac{x_{ij} x_{ik}}{a(\phi_i)} b''(\eta_i).$$

Por lo dicho en la sección 3.1, $\text{Var}(y_i) = b''(\theta_i) a(\phi_i)$. Además $\theta_i = \eta_i$ ya que se está utilizando la liga canónica. Multiplicando la expresión anterior por $b''(\theta_i) / b''(\theta_i)$ obtenemos entonces que

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{Var}(y_i)} (b''(\theta_i))^2. \quad (3.14)$$

Ya que (3.14) no depende de y_i para ninguna i , no cambia al tomar valor esperado, y solo hay que multiplicar por -1 para obtener la expresión correspondiente a la información de Fisher. El resultado se deriva entonces expresando (3.14) matricialmente. \square

3.6. Algunos modelos lineales generalizados

A continuación se describirán tres de los modelos lineales generalizados más comunes: la regresión lineal clásica, la regresión logística y la regresión Poisson. En el caso de la primera, se mencionarán también algunas distribuciones inicial con sus correspondientes distribuciones finales.

3.6.1. Modelo de regresión lineal

Supongamos que $Y \sim \mathcal{N}(\mu, \sigma^2)$. Entonces

$$\begin{aligned} f_Y(y | \mu, \theta) &= \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\} \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}. \end{aligned}$$

Esto ya está en la forma de la Definición 3.1.1 de la familia exponencial, con $\theta = \mu$, $\phi = \sigma^2$ y

$$a(\phi) = \phi, \quad b(\theta) = \frac{1}{2}\theta^2, \quad c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right).$$

En este caso el parámetro canónico resulta ser precisamente μ y el parámetro de dispersión no es más que σ^2 . Por otro lado, como McCullagh y Nelder (1983, Capítulo 2.2.2) argumentan, es fácil comprobar que

$$\mathbb{E}[Y] = b'(\theta)$$

y

$$\text{Var}(Y) = b''(\theta)a(\phi),$$

donde la diferenciación se toma con respecto a θ .

De esta forma, se puede pensar en la familia de modelos lineales generalizados cuya respuesta Y sigue una distribución Normal. En particular, dado que $\theta = \mu$, podemos pensar en la liga canónica:

$$\eta = \mu.$$

Generalmente η es una combinación lineal de las covariables, x_1, \dots, x_p , y los parámetros desconocidos, $\beta_0, \beta_1, \dots, \beta_p$, es decir,

$$\mu = \beta_0 + \sum_{j=1}^p \beta_j x_j.$$

Sin embargo, no hemos considerado la componente aleatoria. Dado que $Y \sim \mathcal{N}(\mu, \sigma^2)$, podemos escribir

$$Y = \mu + \varepsilon,$$

donde $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.⁵ Esta estructura aditiva de errores no es general ni compartida con otros modelos lineales generalizados, pero es una de las características que hacen a este modelo particular tan atractivo. Así pues, el modelo de regresión lineal toma la forma

$$Y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \varepsilon.$$

En la práctica generalmente tendremos un vector de observaciones $y = (y_1, \dots, y_n)^T$, el cual se puede pensar como una realización de $Y = (Y_1, \dots, Y_n)^T$, donde las componentes de Y son variables aleatorias Normales independientes, cada una con media μ_i , $i = 1, \dots, n$. Así pues, si indexamos las observaciones con i y las covariables con j , el modelo de regresión lineal es

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.15)$$

donde x_{ij} se refiere a la i -ésima observación de la j -ésima covariable. Además, se supone que los errores son independientes y tienen la misma varianza.

Generalmente se define la matriz de datos como $X = (x_{ij})$, y se le agrega una columna de unos a la izquierda. De esta forma, X es $n \times (p+1)$, y debe ser de rango completo. Si $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ y $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, entonces podemos escribir (3.15) como

$$y = X\beta + \varepsilon. \quad (3.16)$$

⁵Se puede pensar que $\varepsilon = Y - \mu$, lo que hace evidente que es una variable aleatoria Normal centrada en cero y con la misma varianza que Y .

A (3.16) se le conoce como modelo de regresión lineal clásico, y es equivalente a decir que

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n), \quad (3.17)$$

donde I_n es la matriz identidad de dimensión n . Por simplicidad a partir de ahora se trabajará con la precisión en lugar de la varianza, $\tau = \frac{1}{\sigma^2}$. Además se define por facilidad $q = p + 1$ (el número de parámetros desconocidos).

Ahora bien, como Gutiérrez-Peña (2016a) argumenta, la verosimilitud del modelo (3.17) está dada por

$$L(\beta, \tau; y) \propto \tau^{n/2} \exp \left\{ -\frac{\tau}{2} \left[(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) + n\hat{\sigma}^2 \right] \right\},$$

donde

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad \hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

son los estimadores de máxima verosimilitud de β y σ^2 . Así pues, dada la forma de la verosimilitud, una familia de distribuciones iniciales conveniente es de la forma

$$p(\beta, \tau) = p(\beta | \tau) p(\tau) = \mathcal{N}_p \left(\beta | b_0, \tau^{-1} B_0^{-1} \right) \text{Gamma} \left(\tau \left| \frac{a}{2}, \frac{d}{2} \right. \right), \quad (3.18)$$

donde los *hiperparámetros*⁶ del modelo son tales que $b_0 \in \mathbb{R}^q$, $B_0 \in \mathbb{M}_{q \times q}(\mathbb{R})$ es una matriz simétrica positiva definida y $a, d > 0$. Esta distribución se conoce como *Normal-Gamma*, y es sumamente popular en el análisis Bayesiano del modelo Normal, por lo que no debe de sorprender que aparezca en el modelo de regresión lineal.

Más aún, esta familia es conjugada para β y τ (Gutiérrez-Peña, 2016a, Capítulo 3.3). En particular,

$$p(\beta, \tau | y) = p(\beta | \tau, y) p(\tau | y) = \mathcal{N}_p \left(\beta | b_1, \tau^{-1} B_1^{-1} \right) \text{Gamma} \left(\tau \left| \frac{a_1}{2}, \frac{d_1}{2} \right. \right),$$

⁶Así llamados porque son los parámetros de la distribución inicial del parámetro.

donde la actualización de los hiperparámetros está dada por

$$\begin{aligned} b_1 &= (X^T X + B_0)^{-1}(X^T y + B_0 b_0), \\ B_1 &= X^T X + B_0, \\ a_1 &= n + a, \\ d_1 &= (y - X b_1)^T (y - X b_1) + (b_1 - b_0)^T B_0 (b_1 - b_0) + d. \end{aligned}$$

Finalmente, es posible que quien está realizando el estudio en cuestión no tenga conocimientos iniciales, o desee representarlos como vagos. En este caso lo idóneo es utilizar alguna distribución mínimo informativa; por ejemplo, la distribución inicial de Jeffreys para el modelo (3.17) es

$$p(\beta, \tau) \propto \tau^{\frac{p-2}{2}}, \quad (3.19)$$

que no coincide con (3.13) porque aquí τ también es desconocido. Note que (3.19) es una distribución impropia, y se puede obtener a partir de (3.18) haciendo $a = 0$, $d = 0$ y $B_0 = 0$. Otra distribución inicial de referencia es la obtenida con el método de Bernardo, que da lugar a

$$p(\beta, \tau) \propto \tau^{-1}. \quad (3.20)$$

En (Gutiérrez-Peña, 2016a, pp. 16-18) se pueden encontrar las formas de las distribuciones finales para (3.19) y (3.20), así como algunos resultados para realizar inferencia sobre los parámetros de interés.

3.6.2. Modelo de regresión logística

Supongamos que $Y \sim \text{Binom}(n, p)$, donde $n \in \mathbb{N}$ es conocida pero $p \in [0, 1]$ es desconocido. Entonces, para $y \in \{0, 1, \dots, n\}$,

$$\begin{aligned} p(Y = y) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp \left\{ y \log \left(\frac{p}{1-p} \right) + n \log(1-p) + \log \binom{n}{y} \right\}. \end{aligned}$$

Así pues, la distribución binomial pertenece a la familia exponencial de distribuciones (3.1.1), con $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$ y

$$a(\phi) = 1, \quad b(\theta) = n \log(1 + e^\theta), \quad c(y, \phi) = \log\binom{n}{y}.$$

Notamos que, en efecto,

$$\begin{aligned} b'(\theta) &= n \frac{e^\theta}{1 + e^\theta} \\ &= np \\ &= \mathbb{E}[Y] \end{aligned}$$

y

$$\begin{aligned} b''(\theta)a(\phi) &= n \frac{e^{-\theta}}{(1 + e^{-\theta})^2} \\ &= np(1 - p) \\ &= \text{Var}(Y). \end{aligned}$$

De particular interés es la relación que existe entre θ y la media de la distribución, p . Ésta lleva el nombre de *función logit*, y es popular por su gran cantidad de aplicaciones. Además, note que es una función diferenciable, biyectiva y monótona creciente para valores de p en $(0,1)$. Más aún, esto nos dice que la liga canónica toma la forma

$$\eta = \theta = \log\left(\frac{p}{1-p}\right).$$

También es fácil ver que, si se despeja p como función de θ ,

$$p = \frac{1}{1 + e^{-\theta}}.$$

Esto quiere decir que el modelo lineal generalizado con distribución binomial y función liga logit es

$$p = \frac{1}{1 + e^{-\eta}}, \tag{3.21}$$

donde $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$. Si en particular la variable aleatoria de interés se distribuye Bernoulli (es decir, Binomial con $n = 1$), a (3.21) se le conoce como *modelo de regresión logística*. Note que, independientemente de los valores que tome η , la expresión del lado derecho siempre pertenecerá al intervalo $(0,1)$, donde debe estar p . Esto quiere decir que la función logit respeta el rango de valores que pueden tomar tanto η como p . En resumen, el modelo de regresión logística supone que la relación entre la media p de una variable aleatoria Bernoulli y una serie de q covariables x_1, \dots, x_q está dada por

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q. \quad (3.22)$$

3.6.3. Modelo de regresión Poisson

Consideremos una variable aleatoria $Y \sim \text{Poisson}(\lambda)$, donde $\lambda > 0$ es desconocido. Luego, la función masa de probabilidad de Y es, para $y \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned} P(Y = y) &= e^{-\lambda} \frac{\lambda^y}{y!} \\ &= \exp \{y \log \lambda - \lambda - \log y!\}. \end{aligned}$$

Así pues, la familia Poisson pertenece a la familia exponencial de distribuciones, con $\theta = \log \lambda$, $\phi = 1$ y

$$a(\phi) = 1, \quad b(\theta) = e^\theta, \quad c(y, \phi) = -\log y!.$$

Note también que

$$\mathbb{E}[Y] = \lambda = e^\theta = b'(\theta)$$

y

$$\text{Var}(Y) = \lambda = e^\theta = b''(\theta)a(\phi).$$

Con toda esta información podemos pensar en un modelo lineal generalizado cuya respuesta siga una distribución Poisson. Por lo anterior, la liga canónica es

$$\eta = \log \mu,$$

ya que $\lambda = \mu$. Despejando la media de la distribución, obtenemos que

$$\mu = e^\eta, \quad (3.23)$$

relación que se conoce como *modelo de regresión Poisson*. Particularmente para el caso en el que se tienen p covariables x_1, \dots, x_p el modelo de regresión Poisson supone que la relación entre la media μ de la variable respuesta y las covariables está dada por

$$\log \mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (3.24)$$

En resumen, los modelos lineales generalizados proporcionan un marco sumamente amplio para modelar problemas de prácticamente cualquier índole donde interese estudiar el efecto de una colección de covariables en una respuesta aleatoria. Desde que fueron introducidos han sido un objeto de estudio importante tanto desde el punto de vista teórico como práctico. Más aún, es un tema que ha atraído atención tanto por estadísticos clásicos como Bayesianos. Sin embargo, la vertiente Bayesiana sigue sufriendo de la complejidad computacional inherente al cálculo de la distribución final. Como se comentó en el capítulo anterior, hoy en día siguen desarrollándose algoritmos con el propósito de aumentar la eficiencia de este paso crucial del proceso de aprendizaje Bayesiano. En el siguiente capítulo se discutirá una de las muchas aplicaciones de este tipo de modelos: el diseño de experimentos para modelos lineales generalizados.

Capítulo 4

Diseño estadístico de experimentos

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

— Ronald Fisher

Para comenzar a discutir sobre diseño experimental es necesario primero definir qué es un experimento. Si el propósito de la Estadística es describir fenómenos que se manifiestan a través de datos, es de interés explorar cómo es que estos se obtienen. En muchos estudios estadísticos quien lleva a cabo la investigación simplemente observa los datos y realiza análisis con base en estas observaciones. En este tipo de estudios, comúnmente denominados *estudios observacionales*, no se tiene influencia sobre las variables que afectan la respuesta, bien porque es difícil o imposible cuantificarlas o incluso porque no es de interés.

Un experimento se diferencia de un estudio observacional porque intenta tener influencia o control sobre algunas de las variables que afectan al proceso que da lugar a los datos. Un experimento se lleva a cabo entonces bajo condiciones controladas, de manera que se pueda estudiar adecuadamente

la relación entre las variables independientes y la respuesta.

Bajo este contexto el diseño estadístico de experimentos se refiere al proceso de planeación de un experimento de manera que se recolecten los datos apropiados, susceptibles de ser analizados con métodos estadísticos. Por su naturaleza el diseño experimental se realiza antes de la recolección de los datos, y se ocupa de aspectos como qué tipo de modelo se planteará y, con base en él, cuántas observaciones se habrán de realizar; qué variables independientes se medirán; a qué niveles (si son factores, por ejemplo); así como cuántas observaciones hacer por cada variable y nivel. La importancia del diseño experimental recae en que permite extraer la mayor cantidad de información de calidad utilizando los recursos disponibles.

El diseño procura ser óptimo y para tal fin se establece un criterio de optimalidad. De particular importancia son los criterios de optimalidad alfabética, los cuales cuantifican alguna característica del diseño. Por ejemplo, el criterio de D -optimalidad está relacionado con el determinante de la matriz de covarianzas de la matriz de covariables (Box y Draper, 2007, Capítulo 14.4) y será estudiado más adelante.

Como Montgomery (2012) menciona, el diseño de experimentos ha pasado por varias etapas, comenzando con el trabajo de Ronald Fisher (1935) en las primeras décadas del Siglo XX. Hoy en día las técnicas del diseño experimental han sido estudiadas y desarrolladas por diversos investigadores, siendo ésta una disciplina frecuentemente abordada en diversos programas de licenciatura y de posgrado. Más aún, las aplicaciones del diseño de experimentos han evolucionado, superando por mucho sus orígenes ligados a la agricultura, las aplicaciones discutidas por Fisher. Prácticamente todas las áreas científicas y de ingeniería han realizado experimentos que se benefician del diseño estadístico.

En cuanto al enfoque Bayesiano para el diseño de experimentos, es importante destacar que éste en general ha tenido un desarrollo mucho más lento

que el de la contraparte clásica. Esto se debe principalmente a la complejidad tanto matemática como computacional que ocurre en el paradigma Bayesiano. Si bien con el desarrollo de los diversos métodos computacionales desde los años 90 se ha logrado sobrepasar esta complejidad, también es cierto que el avance ha sido menor que el de otras áreas de la Estadística Bayesiana.

El propósito de este capítulo es plantear el problema de diseño de experimentos como uno de decisión, para así resolverlo a nivel teórico con las técnicas del Capítulo 2, manteniendo por lo tanto un enfoque Bayesiano. En particular se discutirán algunas funciones de pérdida populares en la literatura Bayesiana. También se revisará el método ACE de Overstall y Woods (2016), uno de los métodos más recientes para encontrar diseños óptimos Bayesianos, tema principal de este trabajo. De esta forma se pretende desarrollar la teoría para que en el capítulo siguiente se resuelvan algunos ejemplos de problemas de diseño Bayesiano de experimentos.

Si se desea ahondar en alguno de los temas presentados, se recomienda ampliamente al lector revisar (Box et al., 2005; Montgomery, 2012) para un panorama general del tema y (Chaloner y Verdinelli, 1995) para una versión Bayesiana, así como la bibliografía ahí citada.

4.1. Diseño Bayesiano de experimentos

Ya que el diseño estadístico de experimentos se realiza antes de recopilar los datos, es natural utilizar los conocimientos que se tienen relativos al fenómeno a estudiar. Por lo mismo, el diseño experimental se puede considerar por lo menos implícitamente Bayesiano. Chaloner y Verdinelli (1995), siguiendo la idea de Lindley (1987, pp. 20-21), presentan un enfoque del diseño de experimentos basado en la Teoría de la Decisión.

La idea del diseño experimental es determinar qué valores de las covariables (si son continuas) se deben utilizar para obtener los valores de la variable

respuesta, o a qué niveles si éstas son variables categóricas, usualmente conocidas como factores. Para retomar la idea de Lindley (1987) y Chaloner y Verdinelli (1995), en este trabajo se supondrá que tanto el número de covariables como el de observaciones a registrar ya están determinados y, más aún, que las covariables ya están fijas. Esto quiere decir que no se abordará el tema de selección de modelos, aunque éste generalmente es una parte importante del diseño experimental (ver Montgomery, 2012). Más aún, también se supondrá que el modelo que se utilizará después de la recolección de datos es un modelo lineal generalizado. Esto tampoco suele ser así: muchas veces se realizan experimentos preliminares para determinar el tipo de modelo más conveniente (por ejemplo si hay interacciones o términos de orden alto) y se realizan pruebas de bondad de ajuste para posteriormente diseñar el experimento. Sin embargo, para efectos de esta tesis, se supondrá que se conocen qué variables se utilizarán y el tipo de modelo. Lo único que se deberá determinar son los valores que tomarán las covariables, \mathbf{x} .

Supongamos que nuestro experimento involucra p covariables y que el tamaño de muestra permitido es de n observaciones (usualmente llamadas *ensayos*), donde tanto n como p son fijos y conocidos. La i -ésima observación será y_i , la cual será registrada en las condiciones definidas por los valores $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathcal{X} \subset \mathbb{R}^p$ de las covariables *después* de realizar el experimento; por esto último y_1, \dots, y_n se supondrán desconocidas. Un diseño $\mathbf{x} = \{x_1, \dots, x_n\} \in \mathcal{H}$ se refiere a la colección de todos los valores que toman todas las covariables; son estos valores los que se desean determinar. La matriz de diseño de un experimento \mathbf{x} se define como $X = (x_{ij})$, donde i indexa las observaciones y j las variables. Dicho diseño ocasionará que se observen los datos $y = (y_1, \dots, y_n)^T \in \mathcal{Y}$. Además se supondrá que y tiene función de probabilidad generalizada $p(y | \theta, \mathbf{x})$, donde el parámetro¹ $\theta \in \Theta$ tiene distribución inicial $p(\theta)$. En este trabajo se supondrá que y es una observación de una variable aleatoria Y perteneciente a la familia

¹Puede ser un parámetro de varias dimensiones.

exponencial de distribuciones.

Una manera de identificar los distintos elementos del problema de decisión es refiriéndose a su correspondiente árbol de decisión. En el caso del diseño experimental éste toma la siguiente forma:

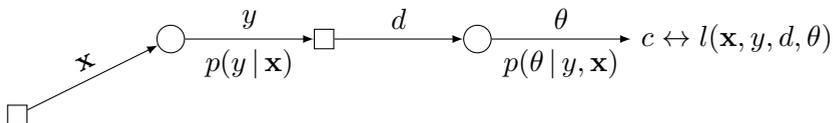


Figura 4.1: Árbol de decisión para un problema de diseño experimental.

El árbol de decisión 4.1 es similar a aquel introducido en el Capítulo 2, pero considera no solo la incertidumbre sobre los valores del parámetro θ sino también la referente a la obtención de los valores de la variable respuesta, y . El segundo nodo de decisión se refiere al tipo de descripción del fenómeno que se hará. Si se decide realizar una estimación por regiones entonces d se referirá a los extremos del intervalo, por ejemplo. Note que las consecuencias se estudian con la función de pérdida l , la cual depende del diseño elegido, los datos que se observaron, la descripción que se realizará del fenómeno, y los valores del parámetro. Esta función debe reflejar adecuadamente los objetivos del experimento (e.g. obtener la mejor predicción o minimizar la varianza de los estimadores).

Ahora bien, el hecho de que la función de pérdida dependa de d amerita mayor discusión. Como se mencionó en el Capítulo 2, en la práctica generalmente se hacen múltiples resúmenes inferenciales de la distribución final. En ese caso tendría que encontrarse un diseño óptimo por cada uno de dichos resúmenes, lo cual no es deseable. Una manera de darle la vuelta a este problema es pensar que se debe encontrar el diseño óptimo no para un problema de inferencia, sino para cualquiera. Es decir, la función de pérdida no debe depender del tipo de descripción que se haga del fenómeno, sino de los elementos que mejor cuantifican la incertidumbre posterior que

se tiene sobre el parámetro θ . Por lo tanto en este trabajo se supondrá que $l(\mathbf{x}, y, d, \theta) = l(\mathbf{x}, y, \theta)$, donde esta última se interpreta como la pérdida derivada de escoger el diseño \mathbf{x} que dio lugar a los datos y y considerando el valor θ del parámetro, independientemente de qué tipo de descripción d del fenómeno se desee hacer.

Como se discutió previamente, se desea determinar cuál es el diseño óptimo, \mathbf{x}^* . Dicha optimalidad debe determinarse con base en (i) la función de pérdida $l(\mathbf{x}, y, \theta)$ y (ii) la distribución que describe la incertidumbre del tomador de decisiones. Esta incertidumbre proviene de dos fuentes: de la incertidumbre sobre los valores de y que se observarán, la cual se cuantifica con $p(y | \mathbf{x})$; y de la incertidumbre sobre el verdadero valor del parámetro θ , la cual se cuantifica mediante $p(\theta | y, \mathbf{x})$ (que corresponde a la distribución final de θ). Entonces la distribución que describe la incertidumbre del tomador de decisiones está dada por

$$p(\theta, y | \mathbf{x}) = p(y | \mathbf{x}) p(\theta | y, \mathbf{x}).$$

Así pues juntando todo lo anterior tenemos que la pérdida esperada de un diseño cualquiera $\mathbf{x} \in \mathcal{H}$ es

$$\Phi_l(\mathbf{x}) = \mathbb{E} [l(\mathbf{x}, y, \theta)] = \int_{\mathcal{Y}} \int_{\Theta} l(\mathbf{x}, y, \theta) p(\theta, y | \mathbf{x}) d\theta dy.$$

Luego, el diseño óptimo \mathbf{x}^* es aquel que minimiza Φ_l ,

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{H}} \int_{\mathcal{Y}} \int_{\Theta} l(\mathbf{x}, y, \theta) p(\theta, y | \mathbf{x}) d\theta dy. \quad (4.1)$$

En principio esta es la solución del problema de diseño de experimentos con los supuestos previamente mencionados. Sin embargo claramente hay algunas observaciones que realizar. En primer lugar, la elección de la función de pérdida o utilidad es un tema sumamente delicado. No hay un consenso sobre una elección idónea, aunque en la literatura existen diversos artículos dedicados a discutir las opciones más populares. Se ahondará en este tema más adelante.

Por otro lado, hay que resaltar de nuevo la complejidad computacional en el cálculo de la pérdida esperada en (4.1). Como Woods et al. (2017) mencionan, existen algunos problemas que se presentan en la práctica:

- a) La evaluación de la función de pérdida l puede ser sumamente complicada, ya que puede depender de la distribución final de θ ; en muchas ocasiones solo es posible obtener valores de ésta numéricamente, e incluso en ese caso hacerlo no es trivial.
- b) Las integrales en (4.1) tienden a ser de dimensión alta (la dimensión del parámetro más el número de observaciones), por lo que su cálculo se debe realizar con métodos numéricos ingeniosos y que logren sobrepasar la *maldición de la dimensión*.²

Esto eleva la dificultad del cálculo de la solución, y es una de las razones por las que el diseño experimental desde un enfoque Bayesiano no ha ganado la popularidad del enfoque clásico. En particular, hasta recientemente no se había notado un avance importante en este tema por sobre los métodos explicados por Chaloner y Verdinelli (1995).

4.2. Funciones de pérdida

Hay una gran variedad de funciones de pérdida que se utilizan comúnmente para problemas de diseño experimental (ver Chaloner y Verdinelli, 1995, Capítulos 2.2, 2.4, 2.5). Para propósitos de esta tesis se estudiarán dos en particular. La primera es la *pérdida de auto-información* (SIL por sus siglas en inglés), definida como

$$l_{\text{SIL}}(\mathbf{x}, y, \theta) = \log \left(\frac{p(\theta)}{p(\theta | y, \mathbf{x})} \right). \quad (4.2)$$

Un diseño óptimo bajo dicha función de pérdida se llamará SIL-óptimo. Siguiendo a Woods et al., por simplicidad se define la pérdida esperada

²En el contexto de estimar integrales numéricamente, la maldición de la dimensión se refiere a que, conforme aumenta la dimensión de la integral a estimar, el trabajo computacional incrementa exponencialmente.

como

$$\Phi_{\text{SIL}}(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{p(\theta)}{p(\theta|y, \mathbf{x})} \right) p(\theta, y | \mathbf{x}) d\theta dy. \quad (4.3)$$

Un diseño SIL-óptimo es entonces aquel que minimice la pérdida esperada Φ_{SIL} .

La segunda pérdida que se estudiará es similar a l_{SIL} , pero sin considerar la distribución inicial de θ :

$$l_D(\mathbf{x}, y, \theta) = \log \left(\frac{1}{p(\theta|y, \mathbf{x})} \right) p(\theta, y | \mathbf{x}). \quad (4.4)$$

Así pues, la pérdida esperada a minimizar es

$$\Phi_D(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{1}{p(\theta|y, \mathbf{x})} \right) p(\theta, y | \mathbf{x}) d\theta dy. \quad (4.5)$$

Las pérdidas esperadas (4.3) y (4.5) merecen mayor discusión, para lo cual primero se enunciará una definición.

Definición 4.2.1 (Divergencia de Kullback-Leibler (1951)). Sean $p(x)$ y $q(x)$ densidades de probabilidad. Entonces la divergencia de Kullback-Leibler de q respecto a p es

$$KL(q || p) = \int_X q(x) \log \left(\frac{q(x)}{p(x)} \right) dx. \quad (4.6)$$

La divergencia de Kullback-Leibler cuantifica la pérdida derivada de aproximar $q(x)$ con $p(x)$, y está intrínsecamente relacionada con la teoría de la información (Kullback y Leibler, 1951). Bernardo y Smith (2000, Capítulos 2.7.3 y 2.7.4) argumentan que, en el paradigma Bayesiano, esta divergencia puede utilizarse para cuantificar la discrepancia entre la distribución inicial $p(\theta)$ y la distribución final $p(\theta|y)$. En ese caso la divergencia de Kullback-Leibler cuantifica la cantidad de información obtenida por actualizar la distribución de θ derivado de haber observado datos y .

La idea de utilizar esta divergencia como medida de ganancia en información es incluso más general. Como ejemplo, Bernardo (1979b) emplea esta idea para desarrollar un método que permita construir distribuciones iniciales mínimo informativas. Su idea consiste en pensar en la distribución final como fija y encontrar la distribución inicial que maximice la divergencia de Kullback-Leibler de ambas, pues esto implicaría que la cantidad de información obtenida al pasar de la distribución inicial a la distribución final fue máxima. Esto se puede interpretar como que los conocimientos iniciales reflejados por la distribución inicial eran tan vagos que los datos tuvieron un impacto máximo en dichos conocimientos.

No está de más resaltar la relevancia que la divergencia de Kullback-Liebler tiene en la Estadística, y particularmente en el paradigma Bayesiano. Hay un sinfín de literatura que explora con mayor profundidad las nociones previamente discutidas, y muchas otras. Particularmente se recomienda revisar Bernardo (1979a); Bernardo (1979b); Bernardo y Smith (2000); Kullback y Leibler (1951) y la bibliografía ahí citada para mayor información al respecto.

En el contexto de diseño Bayesiano de experimentos, la divergencia de Kullback-Leibler entre la distribución inicial y la distribución final de θ es entonces

$$KL(p(\theta | y, \mathbf{x}) || p(\theta)) = \int_{\Theta} \log \left(\frac{p(\theta | y, \mathbf{x})}{p(\theta)} \right) p(\theta | y, \mathbf{x}) d\theta. \quad (4.7)$$

La expresión (4.7) cuantifica la pérdida que se sufre al utilizar la distribución inicial, $p(\theta)$, en lugar de la distribución final, $p(\theta | y, \mathbf{x})$, después de haber observado datos y obtenidos bajo las condiciones definidas por los valores de \mathbf{x} . Puesto que en el caso de diseño experimental los datos y no han sido observados, puede pensarse en calcular el valor esperado de dicha expresión con respecto a $p(y | \mathbf{x})$, obteniendo así

$$\mathbb{E}_{p(y|\mathbf{x})} [KL(p(\theta | y, \mathbf{x}) || p(\theta))] = \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{p(\theta | y, \mathbf{x})}{p(\theta)} \right) p(\theta, y | \mathbf{x}) d\theta dy, \quad (4.8)$$

donde se utiliza el hecho de que $p(\theta, y | \mathbf{x}) = p(\theta | y, \mathbf{x}) p(y | \mathbf{x})$. Esta expresión es la que Bernardo (1979a) denomina la *información esperada acerca de θ obtenida al haber utilizado el diseño \mathbf{x}* .

Una estrategia es utilizar la esperanza (4.8) como función de utilidad con el propósito de maximizar la información obtenida al pasar de la distribución inicial a la distribución final de θ . Para continuar trabajando con funciones de pérdida es posible multiplicar esta expresión por -1 , lo cual se refleja cambiando el orden del numerador y el denominador dentro del logaritmo para obtener entonces

$$\Phi_{\text{SIL}}(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{p(\theta)}{p(\theta | y, \mathbf{x})} \right) p(\theta, y | \mathbf{x}) d\theta dy. \quad (4.9)$$

Así pues minimizar la pérdida esperada (4.9) es equivalente a maximizar la divergencia esperada de Kullback-Leibler entre las distribuciones inicial y final de θ . Alternativamente, dado que la distribución inicial $p(\theta)$ no depende del diseño \mathbf{x} , puede pensarse en omitir este término y obtener la pérdida esperada

$$\Phi_{\text{D}}(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\Theta} \log \left(\frac{1}{p(\theta | y, \mathbf{x})} \right) p(\theta, y | \mathbf{x}) d\theta dy. \quad (4.10)$$

Un diseño óptimo bajo dicha función de pérdida se conoce como D -óptimo; en general, dicha función da lugar a la D -optimalidad Bayesiana. El nombre se debe a la similitud con la contraparte clásica, la cual también cuenta con su propia D -optimalidad. En ese caso el criterio de optimalidad es minimizar el determinante de la matriz de información de Fisher del modelo, (3.13), que da lugar a la minimización de la varianza. Chaloner y Verdinelli (1995) argumentan que, en el caso de regresión normal clásico, la D -optimalidad Bayesiana es equivalente a minimizar el determinante de la matriz de información de Fisher del modelo, más la matriz de precisión inicial de θ .

Existen otras elecciones de funciones de pérdida que dan lugar a distintas familias de optimalidad alfabética similares a las de la contraparte clásica.

Éstas son revisadas con mayor detalle por (Chaloner y Verdinelli, 1995, Capítulo 2.3).

4.3. Método ACE

Si bien la teoría detrás del diseño Bayesiano de experimentos ha sido estudiada desde hace varias décadas, el desarrollo de métodos computacionales eficientes que permitan encontrar diseños que satisfagan (4.1) ha avanzado poco, por lo menos hasta años recientes. Por lo general se conocían soluciones a casos con características particulares. Sin embargo en los últimos años se ha logrado abordar, por lo menos en parte, este problema.

Overstall y Woods (2016) proponen su método de *intercambio aproximado de coordenadas* (ACE por sus siglas en inglés), que logra encontrar diseños óptimos para una gran variedad de problemas de diseño experimental. Además tiene la ventaja de no remontarse a estimaciones asintóticas de la distribución final o de la función de pérdida. La idea general es que el usuario ingrese un diseño inicial, cuyas entradas serán recorridas una por una, decidiendo en cada iteración si la entrada en cuestión debe o no ser cambiada y, en tal caso, por qué otra cantidad.

Consideremos el problema de diseñar un experimento de n ensayos y p covariables, donde tanto n como p se consideran fijas y conocidas. La matriz de diseño $X = (x_{ij})$ es entonces de dimensión $n \times p$; el algoritmo ACE recorrerá cada una de sus np entradas como sigue. Para cada $i = 1, \dots, n$ y $j = 1, \dots, p$, sea $\mathbf{x}_{ij}(x)$ el diseño igual a \mathbf{x} pero habiendo sustituido la ij -ésima entrada de \mathbf{x} por $x \in \mathcal{X}_j$, el espacio muestral de la j -ésima covariable. Considere ahora la pérdida esperada asociada a este diseño, la cual denotaremos por $\Phi_{ij}(x | \mathbf{x})$ para enfatizar el hecho de que se está partiendo de un diseño original \mathbf{x} . Se está pensando a la pérdida esperada como función solo de la coordenada ij -ésima del diseño.

El método ACE intenta minimizar $\Phi_{ij}(x | \mathbf{x})$ para cada i y j . Esto es, dada

una entrada de la matriz de diseño el método propone pensar en las demás entradas del diseño como fijas y luego encontrar el valor de dicha coordenada que minimice la pérdida esperada de todo el diseño. En principio ello se lograría determinando analíticamente la forma funcional de $\Phi_{ij}(x | \mathbf{x})$, y posteriormente minimizando dicha función (ya sea analíticamente o numéricamente) para cada entrada. Esto es prácticamente imposible de realizar por la complejidad analítica de la pérdida esperada: manipular $\Phi(\mathbf{x})$ para obtener $\Phi_{ij}(x | \mathbf{x})$ para cada i, j y x no es factible.

La forma en la que el método supera esta complicación es escoger, para cada i, j , m puntos distintos $x_1, x_2, \dots, x_m \in \mathcal{X}_j$.³ Con ellos se obtienen m aproximaciones de Monte Carlo a la pérdida esperada, $\tilde{\Phi}_{ij}(x_1 | \mathbf{x})$, $\tilde{\Phi}_{ij}(x_2 | \mathbf{x})$, ..., $\tilde{\Phi}_{ij}(x_m | \mathbf{x})$, mediante

$$\tilde{\Phi}_{ij}(x_l | \mathbf{x}) = \frac{1}{N} \sum_{k=1}^N l(\mathbf{x}_{ij}(x_l), y_k, \theta_k), \quad l = 1, \dots, m. \quad (4.11)$$

Aquí, $\mathbf{x}_{ij}(x_l)$ se refiere al diseño original pero habiendo cambiado la ij -ésima entrada por x_l , $l = 1, \dots, m$, $(y_k, \theta_k) \sim p(\theta, y | \mathbf{x})$ y N es grande (usualmente alrededor de 20,000). Para generar el vector (y, θ) se aprovecha el hecho de que $p(\theta, y | \mathbf{x}) = p(y | \theta, \mathbf{x}) p(\theta | \mathbf{x})$, y ambas distribuciones son totalmente conocidas. Más detalles sobre este tipo de estimaciones se pueden encontrar en el Apéndice A.

De esta manera para cada entrada i, j del diseño inicial \mathbf{x} se tienen m pares $\left\{x_l, \tilde{\Phi}_{ij}(x_l | \mathbf{x})\right\}_{l=1}^m$. Con ellos es posible estimar $\Phi_{ij}(x | \mathbf{x})$ mediante algún modelo estadístico. Particularmente Overstall y Woods (2016) proponen emplear un modelo de regresión de proceso Gaussiano, el cual deriva en una estimación $\tilde{\Phi}_{ij}(x | \mathbf{x})$ denominada *emulador Gaussiano* que pretende realizar una interpolación estadística de los puntos. Esto se discute con mayor detalle en el Apéndice B, aunque por el momento basta decir que el

³Overstall y Woods (2016) dividen \mathcal{X}_j en m intervalos y escogen aleatoriamente un punto dentro de cada uno de ellos.

método logra así estimar, para cada entrada i, j , la pérdida esperada como función de dicha entrada, $\tilde{\Phi}_{ij}(x | \mathbf{x})$.

Habiendo obtenido la estimación $\tilde{\Phi}_{ij}(x | \mathbf{x})$ para i, j el siguiente paso es encontrar aquella $x \in \mathcal{X}_j$ que la minimice. Para este propósito Overstall y Woods (2016) proponen generar 10,000 puntos uniformemente espaciados en \mathcal{X}_j y escoger aquel que minimice $\tilde{\Phi}_{ij}(x | \mathbf{x})$. De esta forma se obtiene x^* para cada entrada i, j del diseño inicial.

Como los autores discuten no es sensato simplemente reemplazar \mathbf{x}_{ij} por x^* , puesto que la obtención de esta última está sujeta a errores en la estimación de Monte Carlo (4.11) y en el ajuste del emulador. Sin embargo tampoco es factible emplear métodos que verifiquen la bondad de ajuste del modelo de proceso Gaussiano para cada i, j , pues esto incrementaría drásticamente el costo computacional del método. Por ello para minimizar estos errores se realiza un procedimiento de aceptación/rechazo, cuyo pseudocódigo se encuentra descrito más adelante en el Código 4.2.

Básicamente (Overstall y Woods, 2016, págs. 4, 6) llevan a cabo una prueba de t de Student para determinar si la diferencia entre $\tilde{\Phi}_{ij}(x^* | \mathbf{x})$ y $\tilde{\Phi}(\mathbf{x})$ es significativa. Particularmente los autores proponen lo siguiente:

1. Simular \tilde{B} observaciones independientes de $p(y, \theta | \mathbf{x}_{ij}(x^*))$ y obtener a partir de ellas \tilde{B} estimaciones de $\tilde{\Phi}_{ij}(x | \mathbf{x})$, denominadas $M_1, M_2, \dots, M_{\tilde{B}}$.
2. Simular \tilde{B} observaciones independientes de $p(y, \theta | \mathbf{x})$ y obtener a partir de ellas \tilde{B} estimaciones de $\tilde{\Phi}(\mathbf{x})$, denominadas $O_1, O_2, \dots, O_{\tilde{B}}$.
3. Suponer que $M_k \sim \mathcal{N}(\mu_M, \sigma^2)$ y $O_k \sim \mathcal{N}(\mu_O, \sigma^2)$ para cada $k = 1, 2, \dots, \tilde{B}$.
4. Obtener

$$T = \frac{\bar{O} - \bar{M}}{\sqrt{\frac{2}{\tilde{B}} \tilde{\sigma}}}, \quad (4.12)$$

donde

- a) $\bar{M} = \frac{1}{\tilde{B}} \sum_{k=1}^{\tilde{B}} M_k$ y $\bar{O} = \frac{1}{\tilde{B}} \sum_{k=1}^{\tilde{B}} O_k$ son las medias muestrales de las observaciones simuladas M y O , respectivamente.
- b) $\tilde{\sigma}^2 = \frac{1}{2\tilde{B}-2} \left[\sum_{k=1}^{\tilde{B}} (M_k - \bar{M})^2 + \sum_{k=1}^{\tilde{B}} (O_k - \bar{O})^2 \right]$ es la varianza muestral combinada (*pooled variance* en inglés) de las observaciones simuladas.

5. Calcular

$$\hat{p} = P(U \geq T), \quad (4.13)$$

donde $U \sim t_{(2\tilde{B}-2)}$.

6. Actualizar el diseño inicial \mathbf{x} al diseño modificado $\mathbf{x}_{ij}(x^*)$ con probabilidad \hat{p} .

La ecuación (4.12) es casi igual a la variable pivotal empleada para generar un intervalo de confianza de la diferencia de medias de dos muestras independientes provenientes de un modelo normal, con varianza igual y desconocida. La diferencia radica en que aquí no se incluye la diferencia de las medias poblacionales en el numerador, como si se supusiera que ésta es cero. Sin este supuesto, dicha variable pivotal seguiría una distribución t de Student con $2\tilde{B} - 2$ grados de libertad. Lo que los autores pretenden es dar por cierta la hipótesis de igualdad de medias, en cuyo caso la variable pivotal mencionada se convierte en el estadístico de la ecuación (4.12), y obtener la probabilidad de que la media de las observaciones simuladas del diseño modificado \bar{M} sea menor que la del diseño original \bar{O} . Posteriormente se actualiza el diseño precisamente con dicha probabilidad.

En esencia ese es el método ACE. El usuario ingresa un diseño inicial, el método recorre cada entrada de éste y encuentra cuál es el argumento dentro del espacio muestral de la covariable en cuestión que minimiza la pérdida esperada del diseño completo. Esto se realiza aproximando la pérdida esperada como función de la entrada en cuestión vía un emulador

y minimizando este último. Finalmente se emplea un algoritmo de aceptación/rechazo basado en una prueba t para diferencia de medias para decidir si cambiar la entrada original por la nueva.

Cabe mencionar que la convergencia del método se determina en el mismo espíritu que el utilizado para los métodos MCMC. En particular se utilizan análisis gráficos que comparan el número de iteración con alguna aproximación a la pérdida esperada del diseño en cuestión, $\tilde{\Phi}(\mathbf{x})$. Dicha aproximación generalmente se obtiene mediante métodos de Monte Carlo; por ejemplo, siguiendo la notación de la sección anterior, se puede estimar

$$\Phi_l(\mathbf{x}) = \int_{\mathcal{Y}} \int_{\Theta} l(\mathbf{x}, y, \theta) p(\theta, y | \mathbf{x}) d\theta dy$$

con la ecuación (4.11).

El Código 4.1 muestra los pasos básicos del método ACE, y fue tomado de (Woods et al., 2017).⁴ Este método será utilizado para encontrar diseños óptimos para algunos ejemplos en el Capítulo 5, donde se ejemplificará su implementación y se determinará la convergencia a la solución.

```

1  Input : Diseño inicial  $\mathbf{x} = (x_{ij})$  de tamaño  $n \times p$ .
2      Enteros positivos  $Q$  y  $B$ 
3  Output: Diseño  $\Phi$ -óptimo
4
5  repeat{ # Hasta convergencia
6      for(i in 1:n)
7          for(j in 1:p)
8              Genera  $Q$  puntos  $x_1, \dots, x_Q$  en  $\mathcal{X}_j \subset \mathbb{R}$ 
9                  for(k in 1:Q)
10                     Evalúa  $\tilde{\Phi}_{ij}(x_k | \mathbf{x})$  mediante una aproximación de Monte Carlo
11                         (4.11) con una muestra de tamaño  $B$ 
12                     end
13                     Construye un emulador unidimensional  $\tilde{\Phi}(x)$  utilizando la
14                         ecuación B.2
15                     Encuentra  $\hat{x} = \operatorname{argmin}_{x \in \mathcal{X}_j} \tilde{\Phi}(x)$  generando 10,000 puntos
16                         uniformemente distribuidos en  $\mathcal{X}_j$ 

```

⁴En el algoritmo se emplea Q en lugar de m . para tener consistencia con el algoritmo de aceptación/rechazo.

```

14     Encuentra  $\hat{p} = \hat{p}(\mathbf{x}, \hat{x})$  utilizando el Código 4.2
15     Define  $\mathbf{x}_{ij} = \hat{x}$  con probabilidad  $\hat{p}$ 
16     end
17 end
18 }

```

Código 4.1: Método de intercambio aproximado de coordenadas (ACE), propuesto por Overstall y Woods (2016).

```

1  Input : Diseño actual  $\mathbf{x} = (\mathbf{x}_{ij})$ 
2          Coordenada propuesta  $\hat{x}$ 
3          Entero positivo  $\tilde{B}$ 
4  Output: Probabilidad  $\hat{p}$  de aceptación de la nueva coordenada
5
6  Define  $\mathbf{x}_p$  como el diseño obtenido al reemplazar la  $ij$ -ésima
   entrada de  $\mathbf{x}$  con  $\hat{x}$ 
7  for (k in 1: $\tilde{B}$ )
8      Genera  $\tilde{\theta} \sim p(\theta)$ 
9      Genera  $y_1 \sim p(y|\theta, \mathbf{x}_p)$  y  $y_2 \sim p(y|\theta, \mathbf{x})$ 
10     Define  $M_k = l(\mathbf{x}_p, y_1, \tilde{\theta})$  y  $O_k = l(\mathbf{x}, y_2, \tilde{\theta})$ 
11 end
12 Supón que  $M_k \sim \mathcal{N}(\mu_M, \sigma^2)$  y  $O_k \sim \mathcal{N}(\mu_O, \sigma^2)$ 
13 Considerando  $M_1, \dots, M_k$  y  $O_1, \dots, O_k$  como si fueran datos, calcula  $\hat{p}$ 
   con la ecuación (4.13)

```

Código 4.2: Algoritmo de aceptación/rechazo utilizado para encontrar la probabilidad de aceptación de la línea 14 del Código 4.1, tomado de (Woods et al., 2017).

4.4. Comentarios finales

El enfoque Bayesiano es idóneo para el problema de diseño experimental, ya que este último se puede plantear de manera natural como un problema de decisión. Por lo mismo, es posible encontrar en forma genérica la solución de cualquier problema de esta índole. Sin embargo ésta involucra el cálculo de funciones de pérdida que suelen ser sumamente complejas y de integrales de dimensión alta. Es por ello que la investigación de diseño Bayesiano de experimentos en años recientes ha estado enfocada a encontrar

métodos numéricos que permitan sobrepasar estas complicaciones.

El método ACE es una de las más recientes de estas alternativas. La complejidad del algoritmo es evidente, pues combina y adapta una variedad de herramientas estadísticas, como el contraste de hipótesis para diferencia de medias y los modelos de regresión de procesos Gaussianos. Sin embargo, en el fondo éste simplemente trata de determinar, entrada por entrada, si las coordenadas del diseño inicial se pueden mejorar.

Algo que se debe de notar es la enorme cantidad de argumentos iniciales que recibe dicho método ($\mathbf{x}, B, Q, \tilde{B}$, la función de pérdida, las distribuciones iniciales de los parámetros desconocidos). Ello significa que no solo el método es sofisticado en sí, sino que implementarlo también lo es. Esto se hará evidente en el próximo capítulo, pero establece el difícil acceso a algoritmos de punta que permitan encontrar diseños óptimos con mayor flexibilidad.

Capítulo 5

Implementación de diseños Bayesianos para modelos lineales generalizados

To find out what happens when you change something, it is necessary to change it.

— Box, Hunter y Hunter, *Statistics for Experimenters* (2005)

En capítulos anteriores se han discutido los fundamentos de la Teoría de la Decisión, los modelos lineales generalizados, y la solución teórica del problema de diseño experimental. Estos tres temas se pueden conjuntar para dar lugar al diseño Bayesiano de experimentos para modelos lineales generalizados, el cual se ejemplificará en este capítulo. Como se discutió previamente, la implementación computacional puede ser compleja y no ha sido sino hasta en años recientes (Overstall y Woods, 2016) que se ha logrado atacar satisfactoriamente problemas en esta rama.

El propósito de este capítulo es abordar dos problemas de diseño experimental tomados de Woods et al. (2017). En ambos casos se replicarán los resultados, pero además se modificarán factores como las distribucio-

nes iniciales y la función de pérdida para estudiar el impacto que tienen en el diseño óptimo. La optimización de la pérdida esperada (4.1) se hará utilizando el paquete `acebayes` de R (Overstall et al., 2017).

5.1. Diseño para modelo de regresión logística

La distribución Bernoulli se utiliza para describir el comportamiento de respuestas binarias, es decir, que solo tomen dos valores: 0 (fracaso) y 1 (éxito). Esta distribución ha sido muy estudiada debido a la gran variedad de fenómenos que satisfacen esta característica. La familia de modelos lineales generalizados que utilizan una respuesta binaria son de especial interés. En particular el modelo de regresión logística, derivado en el Capítulo 3, es muy utilizado en la práctica por las propiedades estadísticas inherentes a utilizar la liga canónica en un modelo lineal generalizado.

Woods et al. (2006) encuentran el diseño óptimo para un ejemplo industrial en el que una empresa de tecnología alimentaria quería empaquetar papas en un ambiente de atmósfera protegida, con el objetivo de incrementar la vida útil de éstas. El experimento estudiaba el efecto de $p = 3$ covariables—concentración de vitaminas en el sumergimiento de pre-empacado, x_1 , y los niveles de dos gases en la atmósfera protectora, x_2 y x_3 —en diversas variables binarias, entre las que se encuentra la presencia o no de líquido en el paquete después de 7 días. El experimento consta de $n = 16$ ensayos.

En (Woods et al., 2017) los autores retoman el ejemplo del empaquetado de papas. Para ello suponen que la relación que existe entre la respuesta y las covariables se puede describir mediante un modelo de regresión logística. Sea $Y_i \sim \text{Bernoulli}(\pi(x_i))$ la respuesta del i -ésimo ensayo del experimento con valores $x_i = (x_{i1}, x_{i2}, x_{i3})^T$ de las covariables, $i = 1, 2, \dots, 16$. Los autores suponen que el predictor lineal η está dado por

$$\eta_i = \beta_0 + \sum_{j=1}^3 \beta_j x_{ij} + \sum_{j=1}^3 \sum_{k=j}^3 \beta_{jk} x_{ij} x_{ik}. \quad (5.1)$$

Aquí, $\beta_0, \dots, \beta_3, \beta_{11}, \beta_{12}, \dots, \beta_{33}$ son los coeficientes (desconocidos) a ser estimados después de la recolección de los datos. Note que, si bien el predictor lineal ya incluye términos cruzados, llamados también interacciones, éste sigue siendo *lineal en los parámetros*.

De esta forma el modelo propone

$$\pi(x_i) = \frac{1}{1 + e^{-\eta_i}} \quad (5.2)$$

o, alternativamente,

$$\log \left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \beta_0 + \sum_{j=1}^3 \beta_j x_{ij} + \sum_{j=1}^3 \sum_{k=j}^3 \beta_{jk} x_{ij} x_{ik}. \quad (5.3)$$

Recordemos que el objetivo es encontrar los valores de x_1, x_2 y x_3 (16 de cada uno de ellos) que deriven en un diseño óptimo, donde la optimalidad es resultado de la función de pérdida utilizada (la cual se comentará más adelante). Tanto $n = 16$ como $p = 3$ son fijos.

Con el propósito de ilustrar el funcionamiento del método, los autores asumen distribuciones iniciales uniformes independientes para los parámetros del modelo; en particular,

$$\begin{aligned} \beta_1, \beta_2 &\sim \mathcal{U}(2, 6), \\ \beta_0, \beta_3, \beta_{jk} &\sim \mathcal{U}(-2, 2) \quad \text{para } j, k = 1, 2, 3. \end{aligned} \quad (5.4)$$

Además el espacio de valores que pueden tomar las covariables considerado por los autores es

$$\mathcal{X} = [-1.2872, 1.2872] \times [-1.2872, 1.2872] \times [-1.2872, 1.2872],$$

abreviado $[-1.2872, 1.2872]^3$.

Se replicaron los resultados obtenidos por los autores utilizando el método previamente mencionado, así como los parámetros que proponen Woods

et al. Estos son $B = 1,000$ y $Q = 10$ en el método ACE (Código 4.1) y $\tilde{B} = 20,000$ en el algoritmo de aceptación y rechazo (Código 4.2). Se encontró el diseño D -óptimo Bayesiano para 16 ensayos, es decir, aquel diseño que minimiza la ecuación (4.5). La matriz de diseño óptima es

$$X^* = \begin{bmatrix} -0.53 & 0.46 & 1.29 \\ -1.25 & 0.20 & 1.29 \\ 1.28 & -0.14 & 1.12 \\ 0.17 & -1.29 & -1.29 \\ 1.29 & -0.32 & -1.26 \\ 1.29 & -1.29 & 0.18 \\ 1.10 & -1.20 & -1.29 \\ 0.37 & -1.29 & 1.29 \\ -1.29 & 0.33 & -1.29 \\ 0.67 & -0.62 & 1.29 \\ -1.29 & 1.29 & 0.34 \\ -0.14 & 0.10 & -0.01 \\ -0.35 & 1.27 & -1.29 \\ -1.24 & 1.26 & -1.15 \\ -0.34 & 1.19 & 1.29 \\ -0.20 & 0.28 & -1.26 \end{bmatrix} \quad (5.5)$$

Los números que aparecen como ± 1.29 en realidad son iguales a ± 1.2872 , el límite del espacio de diseño, pero se muestran redondeados. Esto nos dice que el diseño óptimo considera utilizar los extremos del cubo de diseño. La Figura 5.1 muestra las proyecciones en dos dimensiones de las variables, así como la proyección en una dimensión en forma de densidad, obtenida mediante un suavizamiento con un kernel Gaussiano. Esta gráfica tiene el propósito de ayudar a determinar, visualmente, en qué región del espacio muestral están concentrados los valores de cada covariable en el diseño óptimo. También se muestra la correlación entre las covariables derivada de emplear el diseño óptimo. Es interesante observar que la correlación entre x_3 y cualquiera de las otras covariables es cercana a cero, pero la correspondiente a x_1 y x_2 es -0.71 , un valor alto (en valor absoluto). Ello indica

que el diseño óptimo no considera emplear covariables no correlacionadas.

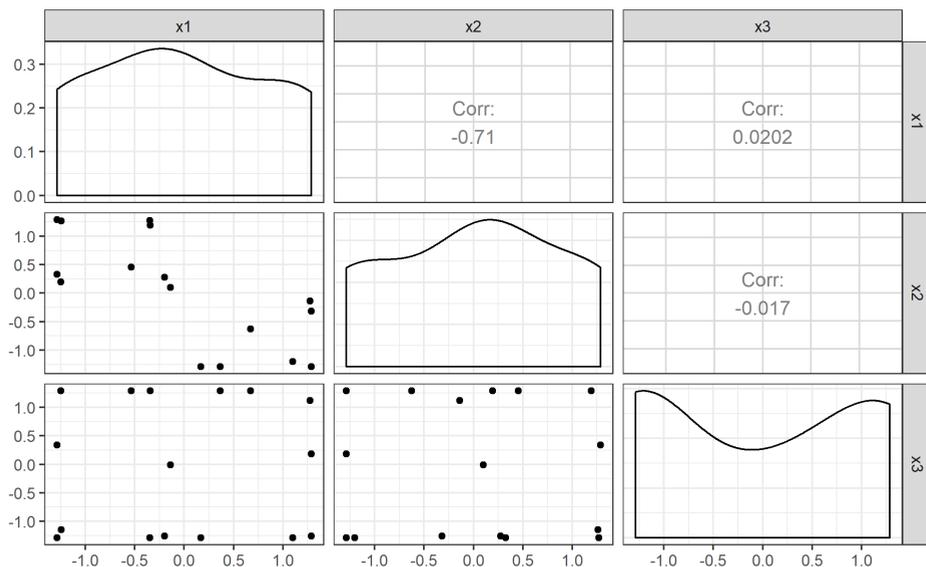


Figura 5.1: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión logística (5.3), así como su correlación.

Para corroborar la convergencia del algoritmo, la Figura 5.2 muestra una aproximación a la pérdida esperada del diseño en cuestión para cada una de las iteraciones de éste. Es inmediato observar que la pérdida esperada es prácticamente igual para las últimas iteraciones, lo que indica que, en efecto, el método convergió.

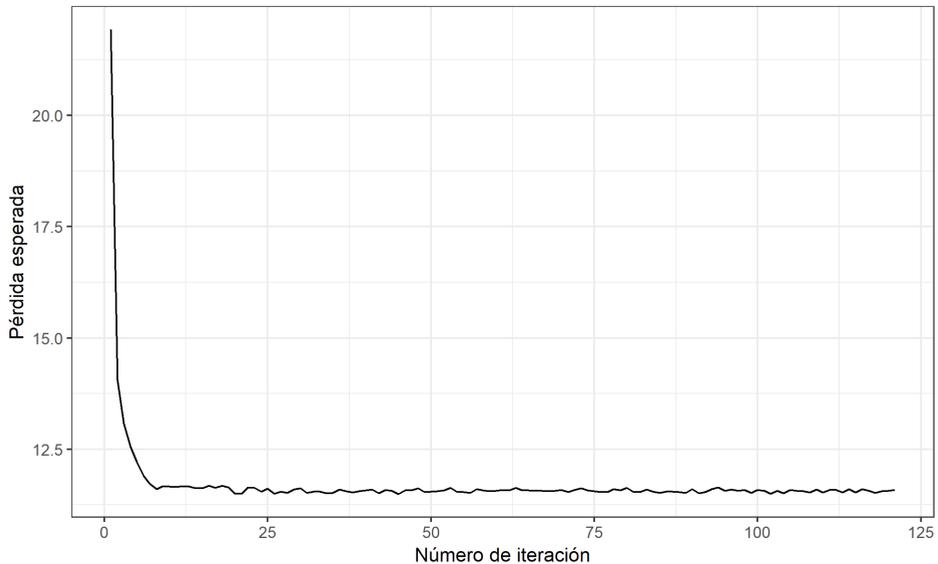


Figura 5.2: Para la regresión logística (5.3), se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

Para bosquejar la complejidad del problema es posible preguntarse qué forma toma la pérdida esperada. Dado que se encontró el diseño D -óptimo Bayesiano, la pérdida esperada que este diseño minimiza está dada por

$$\Phi_D(\mathbf{x}) = \int_{\mathbf{y}} \int_B \log \left(\frac{1}{p(\beta | \mathbf{y}, \mathbf{x})} \right) p(\beta, \mathbf{y} | \mathbf{x}) d\beta d\mathbf{y},$$

donde $B = S_1 \times S_2 \times \cdots \times S_{10}$ es el espacio parametral y S_i se refiere al soporte de β_i (uno de $(2, 6)$ o $(-2, 2)$). Hay dos factores que entran en juego en este caso: la distribución posterior de β y la distribución conjunta de β e \mathbf{y} .

La distribución posterior de β depende tanto del diseño en cuestión \mathbf{x} como de los datos \mathbf{y} que se observarán al realizar el experimento en las condiciones

especificadas por \mathbf{x} . Por lo mismo no es inmediato obtener una expresión para $p(\beta | y, \mathbf{x})$, sino que se debe emplear el teorema de Bayes:

$$p(\beta | y, \mathbf{x}) = \frac{p(y | \beta, \mathbf{x}) p(\beta)}{p(y | \mathbf{x})}.$$

Esta expresión a su vez se divide en tres factores:

1. $p(y | \beta, \mathbf{x})$ es la verosimilitud. Dado que los datos provienen independientemente de una distribución Bernoulli se tiene que

$$\begin{aligned} p(y | \beta, \mathbf{x}) &= \prod_{i=1}^{16} \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i} \\ &= \prod_{i=1}^{16} \left(\frac{1}{1 + e^{-\eta_i}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-\eta_i}} \right)^{1-y_i}. \end{aligned}$$

No es posible simplificar esta expresión sin conocer los valores de los predictores, η_i .

2. $p(\beta)$ es la distribución inicial conjunta de los parámetros. Por la independencia y distribuciones iniciales marginales de estos se tiene que

$$\begin{aligned} p(\beta) &= \prod_{i=1}^{10} p(\beta_i) \\ &= \left(\frac{1}{4} \right)^{10} \prod_{i=1}^{10} \mathbb{1}_{S_i}(\beta_i), \end{aligned}$$

donde $\mathbb{1}$ es la función indicadora.

3. $p(y | \mathbf{x})$ es la función de densidad de los datos y , la cual se obtiene multiplicando las distribuciones de los puntos 1. y 2. e integrando sobre el espacio parametral B .

La distribución conjunta de β e y también depende del diseño en cuestión y de los datos observados. Puntualmente

$$p(\beta, y | \mathbf{x}) = p(\beta | y, \mathbf{x}) p(y | \mathbf{x}).$$

Esta expresión se divide a su vez en dos expresiones más, las cuales ya fueron discutidas previamente. Luego, para obtener la distribución conjunta es necesario (i) obtener la distribución posterior de β conjuntando los puntos 1. a 3., (ii) obtener la densidad de y dado \mathbf{x} con el punto 3. y (iii) multiplicar ambas distribuciones.

Habiendo obtenido estas componentes es aún necesario calcular el logaritmo del recíproco de la distribución posterior de β , multiplicarlo por la distribución conjunta de β e y y, posteriormente, evaluar la integral sobre todo el espacio parametral y luego sobre todo el espacio muestral. Si se quisiera encontrar el diseño óptimo entonces tendría que encontrarse analíticamente el mínimo de dicha expresión.

Evaluar la integral y minimizarla es claramente imposible, pero incluso tratar de obtener una expresión para el integrando es claramente una tarea compleja. Es por ello que, si se desean encontrar diseños no triviales, se *debe* recurrir a herramientas computacionales como el método ACE.

Hay un punto que merece mayor discusión. La distribución inicial empleada por los autores para β_1 y β_2 es una distribución $\mathcal{U}(2, 6)$. Notoriamente el cero *no* está contenido en el soporte de esta distribución. Lo que ello implica es que, al momento de obtener datos y conjuntar esta distribución inicial con la verosimilitud para así obtener la distribución final, el cero tampoco estará contenido en el soporte de la distribución final. En otras palabras, los autores están suponiendo que tanto β_1 como β_2 son diferentes de cero a priori. Y, más aún, este supuesto no cambiará independientemente de los datos observados.

Por otro lado, ni Woods et al. (2006) ni Woods et al. (2017) dan alguna explicación clara sobre por qué supusieron que dichos coeficientes son distintos de cero. Por ello se volvió a encontrar el diseño D -óptimo Bayesiano pero cambiando las distribuciones iniciales de estos dos parámetros,

de forma que éstas fueran

$$\beta_1, \beta_2 \sim \mathcal{U}(-2, 6). \quad (5.6)$$

No hay razón para modificar las demás distribuciones iniciales, ya que éstas sí contienen al cero en su soporte. Además, se dejó igual el límite superior del intervalo de β_1 y β_2 , pero se modificó el límite inferior de forma que éste coincidiera con el de las demás distribuciones. Se utilizaron los mismos parámetros del método ACE que en el ejemplo anterior, es decir, $B = 1,000$ y $Q = 10$ en el método ACE (Código 4.1) y $\tilde{B} = 20,000$ en el algoritmo de aceptación y rechazo (Código 4.2). La matriz de diseño óptima es

$$X_{\text{Modificada}}^* = \begin{bmatrix} 1.29 & -1.29 & 1.23 \\ 1.29 & -1.27 & -1.27 \\ 0.12 & -0.19 & -1.23 \\ 0.44 & -1.29 & 0.06 \\ 0.18 & 1.29 & -1.29 \\ -1.29 & 0.89 & 1.29 \\ 1.27 & 0.55 & -1.27 \\ -1.15 & 1.29 & -1.11 \\ -0.16 & 0.66 & 1.25 \\ -0.33 & 0.13 & -0.05 \\ 1.28 & 1.29 & 1.29 \\ -0.44 & -1.29 & -1.29 \\ -1.29 & -0.03 & -1.29 \\ -1.29 & -0.86 & 1.21 \\ 0.09 & -0.46 & 1.29 \\ 1.13 & -0.09 & 0.42 \end{bmatrix} \quad (5.7)$$

Es de observar que el hecho de modificar las distribuciones iniciales derivó en un diseño óptimo diferente. Para obtener conclusiones de mayor profundidad la Figura 5.3 es análoga a la Figura 5.1, y muestra las proyecciones de las variables en una y dos dimensiones (con un suavizamiento para mostrar densidad en la primera de ellas) y las correlaciones.

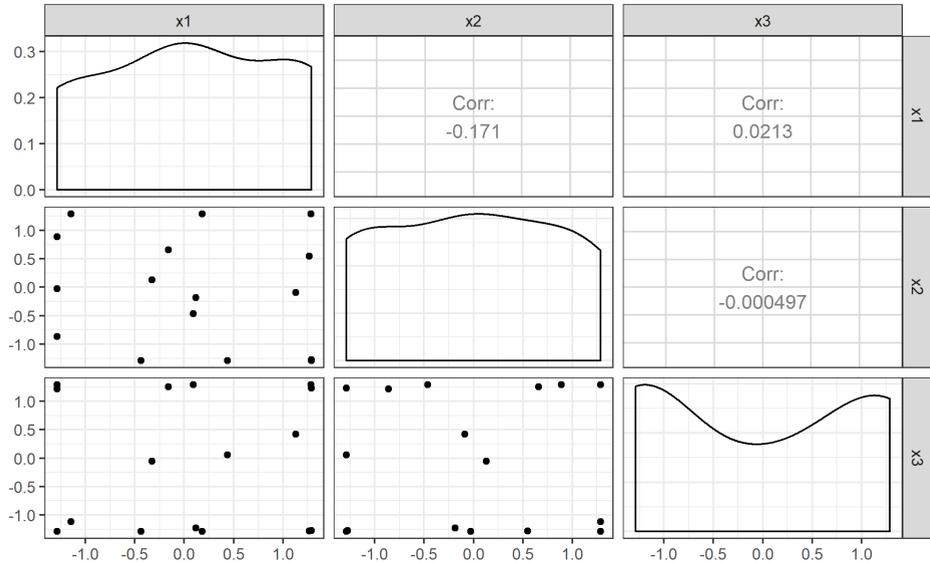


Figura 5.3: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión logística (5.3) con distribuciones iniciales modificadas (5.6), así como su correlación.

Hay algunas diferencias con respecto al diseño óptimo original que comentar. Primeramente la covariable x_3 sufrió pocos cambios, en tanto que su densidad suavizada se mantuvo prácticamente igual y sigue concentrada en los extremos del espacio muestral. En el caso de x_2 , sin embargo, la densidad pasó de estar concentrada alrededor del cero a estar distribuida uniformemente en el espacio muestral.

Por otro lado es interesante observar que las correlaciones entre x_3 y el resto de las covariables se mantuvieron relativamente similares, siendo ambas cercanas a cero. Sin embargo, la correlación entre x_1 y x_2 cambió drásticamente, puesto que dichas variables pasaron de estar altamente correlacionadas (-0.71) a estar solo un poco correlacionadas (-0.17). Resumiendo,

modificar las distribuciones iniciales de β_1 y β_2 tuvo un impacto directo en x_1 y x_2 , mas no en x_3 , lo cual tiene sentido.

De nuevo para corroborar la convergencia del algoritmo, la Figura 5.4 muestra una aproximación a la pérdida esperada del diseño en cuestión para cada una de las iteraciones de éste. Es inmediato observar que la pérdida esperada es prácticamente igual para las últimas iteraciones, señal de convergencia en el método.

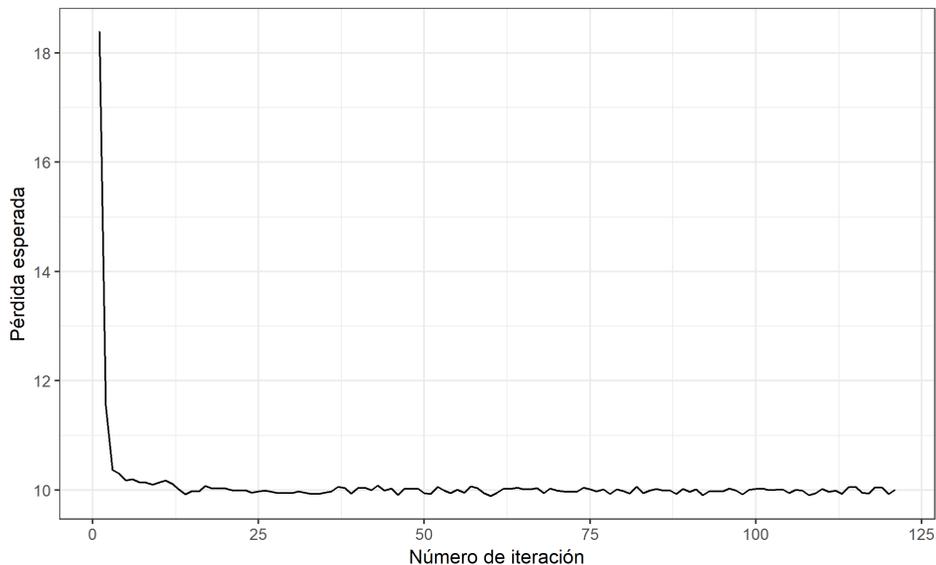


Figura 5.4: Para la regresión logística (5.3) con distribuciones iniciales modificadas (5.6), se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

5.2. Diseño para modelo de regresión Poisson

A lo largo de la historia siempre ha sido de interés modelar el número de ocurrencias de un fenómeno en cierto periodo de tiempo. La distribución Poisson es una de las alternativas más populares para este tipo de situaciones, y la familia de modelos lineales generalizados que suponen dicha distribución en la respuesta ha sido altamente estudiada. De particular interés es el modelo que utiliza la liga canónica (logarítmica), la cual se discutió con detalle en el Capítulo 3.

Atkinson y Woods (2013) encuentran resultados teóricos importantes para el diseño experimental para modelos de regresión Poisson. Woods et al. (2017) retoman y aprovechan dichos resultados, asumiendo que se tiene un experimento con $n = 6$ ensayos para cinco variables x_1, \dots, x_5 . Sea $Y_i \sim \text{Poisson}(\lambda(x_i))$. Entonces se pretende modelar

$$\log \lambda(x_i) = \beta_0 + \sum_{j=1}^5 \beta_j x_{ij}, \quad i = 1, \dots, 6, \quad (5.8)$$

aprovechando que $\lambda(x_i) = \mu(x_i)$ en el modelo Poisson.

Más aún, los autores asumen que $\beta_0 = 0$ es conocido. Las distribuciones iniciales para los demás parámetros son independientes y están dadas por

$$\begin{aligned} \beta_1, \beta_3, \beta_5 &\sim \mathcal{U}(1, 1 + \alpha), \\ \beta_2, \beta_4 &\sim \mathcal{U}(-1 - \alpha, -1), \end{aligned} \quad (5.9)$$

donde $\alpha > 0$ es un parámetro fijo. Las covariables toman valores en $[-1, 1]^5$, utilizando la notación del ejemplo anterior. Los demás argumentos del método ACE utilizados por los autores son $B = 1,000$ y $Q = 20$ en el método ACE (Código 4.1) y $\tilde{B} = 20,000$ en el algoritmo de aceptación y rechazo (Código 4.2).

Woods et al. encuentran el diseño SIL-óptimo para los valores $\alpha = 0.5$ y $\alpha = 0.75$. Para comparar con sus resultados se encontró el diseño D -óptimo.

La Tabla 5.1 muestra la comparación del diseño SIL-óptimo con el D -óptimo para $\alpha = 0.5$, y la Tabla 5.2 muestra lo respectivo para $\alpha = 0.75$.

Núm	Diseño D -óptimo					Diseño SIL-óptimo				
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	-0.42	-1	1	-1	1	-0.5	-1	1	-1	1
2	1	0.5	1	-1	1	1	0.56	1	-1	1
3	1	-1	-0.36	-1	1	1	-1	-0.31	-1	1
4	1	-1	1	0.51	1	1	-1	1	0.33	1
5	1	-1	1	-1	-0.39	1	-1	1	-1	-0.38
6	1	-1	1	-1	1	1	-1	1	-1	1

Tabla 5.1: Se muestran los diseños óptimos según los dos distintos criterios para $\alpha = 0.5$.

Núm	Diseño D -óptimo					Diseño SIL-óptimo				
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	-0.32	-1	1	-1	1	-0.22	-1	1	-1	1
2	1	0.34	1	-1	1	1	0.22	1	-1	1
3	1	-1	-0.34	-1	1	1	-1	-0.32	-1	1
4	1	-1	1	0.38	1	1	-1	1	0.11	1
5	1	-1	1	-1	-0.47	1	-1	1	-1	-0.31
6	1	-1	1	-1	1	1	-1	1	-1	1

Tabla 5.2: Se muestran los diseños óptimos según los dos distintos criterios para $\alpha = 0.75$.

En ambos casos las variables toman el mismo valor, ya sea ± 1 , en todos los ensayos menos uno: el que aparece en la diagonal.¹ Es ahí donde se aprecian las diferencias entre ambos diseños óptimos. Si bien es cierto que los signos son todos iguales, también se vislumbran pequeñas diferencias entre los valores de las variables del diseño óptimo. En particular, x_4 es la covariable que muestra mayor variación para ambos valores de α . Esto se debe a que distintos criterios de optimalidad llevan a distintos diseños ópti-

¹Aunque en estricto sentido no se puede hablar de una diagonal, se entiende que se refiere al valor x_{ii} de la covariable.

mos. Sin embargo cabe recordar que la D -optimalidad Bayesiana es similar a la SIL-optimalidad, lo que explica que ambos diseños sean relativamente similares.

Las Figuras 5.5 y 5.6 muestran las proyecciones en dos dimensiones de las variables junto con su proyección en una dimensión (en forma de densidad), para $\alpha = 0.5$ y $\alpha = 0.75$. Además se muestran las correlaciones que tendrán las covariables empleando el diseño óptimo. Es interesante que éstas son todas iguales a ± 0.2 para ambos valores de α . Ello indica que el diseño óptimo de nuevo no considera utilizar covariables no correlacionadas y, más aún, que la correlación entre las variables siempre será de ± 0.2 .

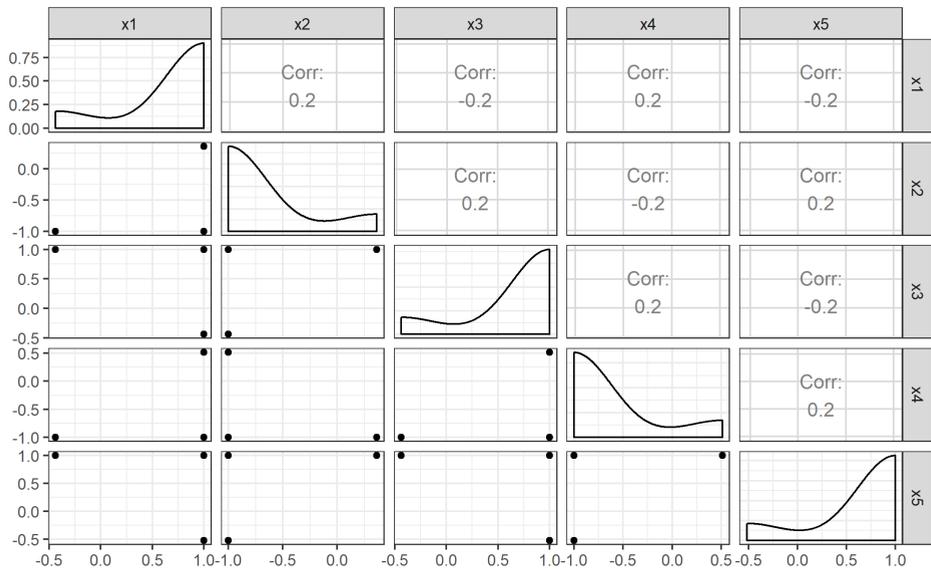


Figura 5.5: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión Poisson (5.8), así como su correlación, para $\alpha = 0.5$.

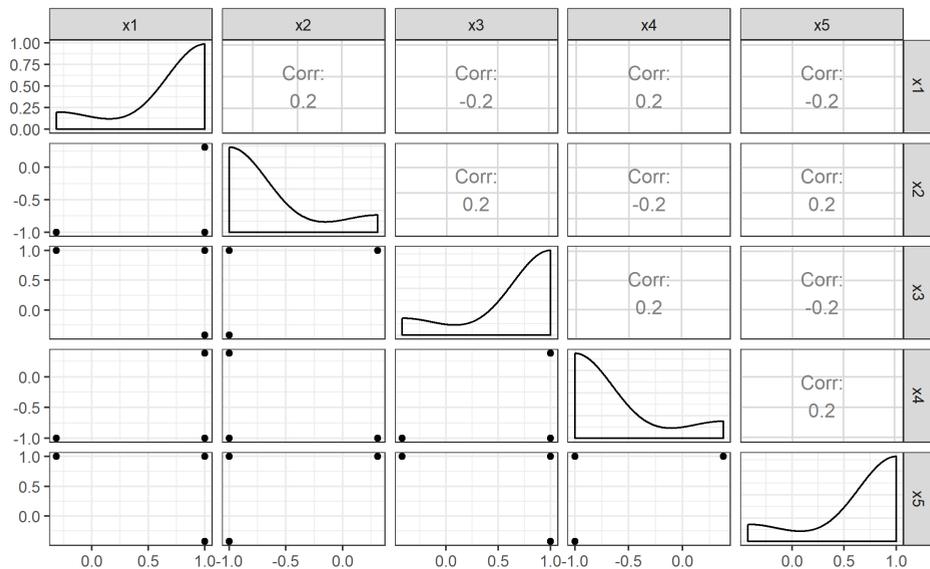


Figura 5.6: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión Poisson (5.8), así como su correlación, para $\alpha = 0.75$.

De nuevo para corroborar la convergencia del método, las Figuras 5.7 y 5.8 muestran las aproximaciones a las utilidades esperadas del diseño en cuestión para cada iteración del algoritmo, y para $\alpha = 0.5$ y $\alpha = 0.75$, respectivamente. En ambos casos se aprecia que dichos promedios se estabilizan, por lo que se puede concluir que el método en efecto convergió.

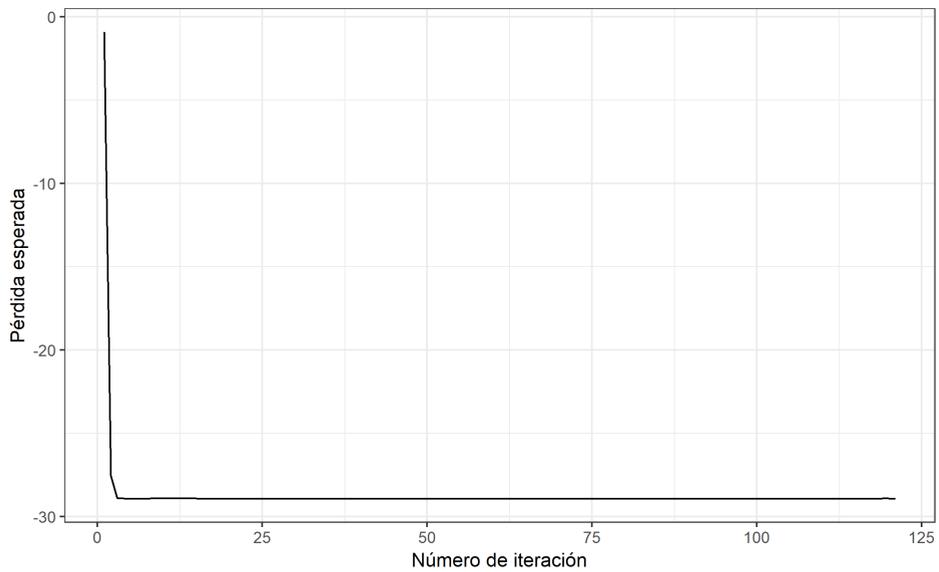


Figura 5.7: Para la regresión Poisson con $\alpha = 0.5$, se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

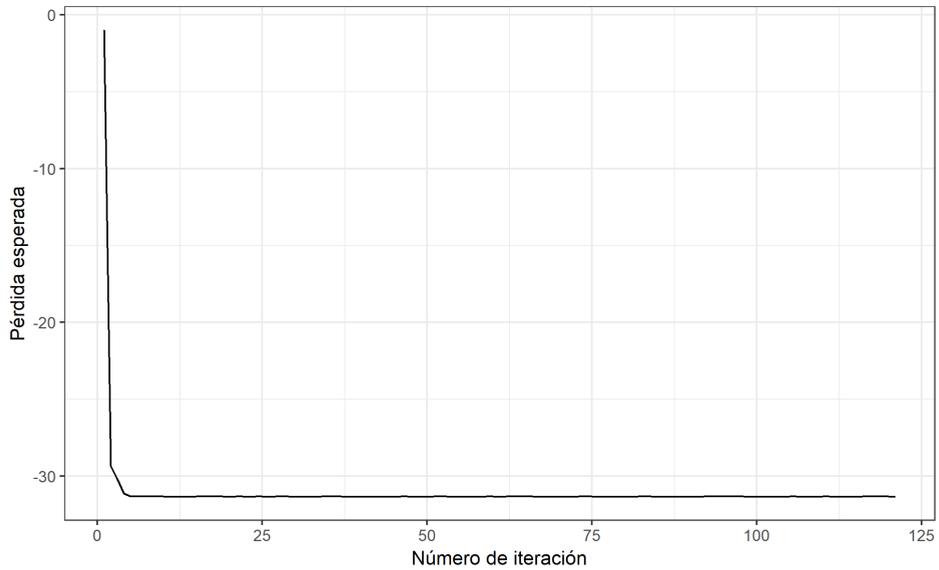


Figura 5.8: Para la regresión Poisson con $\alpha = 0.75$, se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

Otra manera de estudiar las diferencias entre los diseños D -óptimos y los diseños SIL-óptimos es comparando las distribuciones posteriores a las que dan lugar. Para ello es necesario conocer los verdaderos valores de los parámetros β , lo cual no ocurre en este caso. Sin embargo, para efectos prácticos, puede escogerse algún valor cualquiera dentro del soporte de cada parámetro y suponerlo como cierto. Aunque podría optarse por escoger la media de cada distribución, o bien la mediana o algún cuantil, para este trabajo se obtuvo una observación de cada una de las distribuciones iniciales marginales de estos parámetros; estas observaciones se fijaron y se supusieron como los valores reales de cada parámetro. Puntualmente se supuso que

$$\begin{aligned} \beta_1 &= 1.47, & \beta_2 &= -1.14, & \beta_3 &= 1.24, \\ \beta_4 &= -1.50, & \beta_5 &= 1.41. \end{aligned} \tag{5.10}$$

Es posible obtener datos simulados y conjuntando los valores (5.10) de β con algún diseño óptimo y utilizando la expresión del modelo (5.8). Para este ejercicio se fijó $\alpha = 0.5$ y se obtuvieron así dos series de $n = 6$ datos simulados: y_D , correspondiente a los datos obtenidos con el diseño D -óptimo, e y_{SIL} , correspondiente a los datos obtenidos con el diseño SIL-óptimo.

Posteriormente se obtuvo, utilizando el programa JAGS (Plummer, 2003), una muestra de tamaño 10,000 de la distribución posterior de β para cada serie de datos. La Figura 5.9 compara las densidades marginales posteriores de cada diseño óptimo, obtenidas suavizando los datos de las muestras de cada distribución posterior.

Si bien las distribuciones son similares para todos los parámetros, solo en el caso de β_5 éstas son prácticamente iguales. Las distribuciones marginales posteriores de $\beta_1, \beta_2, \beta_3$, y β_4 muestran diferencias entre los diseños óptimos.

En el caso de β_2 y β_3 las distribuciones tienen la misma forma pero las correspondientes al diseño SIL-óptimo están recorridas ligeramente hacia

la derecha, con mayor cercanía al verdadero valor del parámetro. Las distribuciones de β_1 y β_4 son interesantes porque corresponden a valores del parámetro cercanos al extremo de su respectivo intervalo. Es notorio que las distribuciones correspondientes al diseño SIL-óptimo tienen colas más pesadas y mayor varianza, mientras que las respectivas al diseño D -óptimo están más concentradas en el extremo del intervalo.

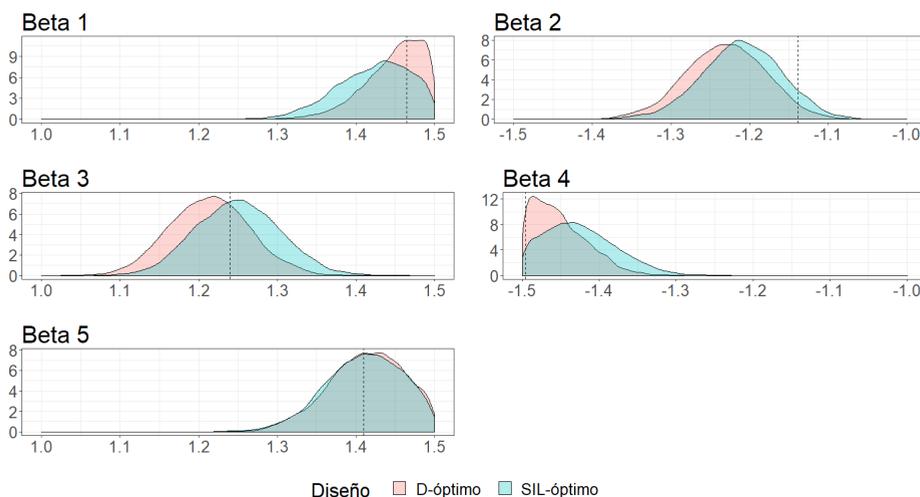


Figura 5.9: Densidades marginales posteriores de β_1, \dots, β_5 derivadas de las series de datos y_D y y_{SIL} . La línea punteada indica el valor simulado (5.10) del parámetro correspondiente.

Por otro lado, en el ejemplo de la regresión logística las distribuciones iniciales de dos de los parámetros no contenían al cero en su soporte, ocasionando que se diera por hecho que ambos parámetros eran diferentes de cero. En este caso ocurre algo similar, solo que *todos* los parámetros se suponen de dicha forma: para ambos valores de α (y para cualquier valor positivo de este parámetro, en realidad) las distribuciones iniciales no contienen al cero en su soporte.

De manera análoga al ejemplo anterior, se encontraron diseños óptimos pero habiendo modificado las distribuciones iniciales de forma que éstas sí contengan al cero en su soporte. En este caso, sin embargo, no solo se modificó el soporte de estas distribuciones, sino que además se escogió otra familia de distribuciones como sigue. Sea $\alpha > 0$ fijo. Las distribuciones iniciales propuestas por Woods et al. (2017) (5.9) se pueden clasificar en dos: aquellas que corresponden a β_i con i impar y aquellas que corresponden a β_i con i par. En el primer caso se considera una distribución uniforme en $(1, 1 + \alpha)$ y en el segundo una uniforme en $(-1 - \alpha, -1)$. Note que los autores suponen que todos los valores de β son mayores que $-1 - \alpha$ y menores que $1 + \alpha$.

Se puede entonces pensar en una distribución inicial que, independientemente de si i es par o impar, suponga que los parámetros se encuentran en el intervalo $(-1 - \alpha, 1 + \alpha)$. Además, para cambiar la familia de distribuciones y no solo el soporte, se puede utilizar alguna otra distribución que no sea uniforme. Para este propósito recordemos que la distribución $\text{Beta}(a, b)$ tiene soporte $(0, 1)$ y media igual a $a/(a+b)$. Luego, una variable aleatoria X que siga una distribución $\text{Beta}(2, 2)$ tendrá media igual a $1/2$. Sea $Y = 2(1 + \alpha)(X - \frac{1}{2})$ una transformación de la variable X . La función de densidad de Y tiene la misma forma que la de X , pero con soporte igual al intervalo $(-1 - \alpha, 1 + \alpha)$. Se encontraron los diseños D -óptimos Bayesianos para $\alpha = 0.5$ y $\alpha = 0.75$ utilizando esta distribución como inicial para todos los parámetros.

Las Tablas 5.3 y 5.4 muestran la comparación del diseño SIL-óptimo con el D -óptimo con las distribuciones iniciales modificadas y para $\alpha = 0.5$ y $\alpha = 0.75$, respectivamente.

Es evidente que los diseños óptimos bajo las nuevas distribuciones iniciales son diferentes a los diseños óptimos originalmente encontrados por Woods et al. (2017) (y también a los diseños óptimos previamente encontrados).

Núm	Diseño D -óptimo					Diseño SIL-óptimo				
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	1	1	0.96	-1	1	-0.5	-1	1	-1	1
2	1	1	1	1	-1	1	0.56	1	-1	1
3	-1	-0.99	-1	-1	-1	1	-1	-0.31	-1	1
4	-1	1	-1	1	1	1	-1	1	0.33	1
5	1	-1	-1	1	1	1	-1	1	-1	-0.38
6	-1	-1	1	1	1	1	-1	1	-1	1

Tabla 5.3: Se muestran los diseños óptimos según los dos distintos criterios para $\alpha = 0.5$ y con las distribuciones iniciales modificadas.

Núm	Diseño D -óptimo					Diseño SIL-óptimo				
	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
1	1	1	1	-1	1	-0.22	-1	1	-1	1
2	1	1	1	1	-1	1	0.22	1	-1	1
3	-1	-1	-1	-1	-1	1	-1	-0.32	-1	1
4	-1	1	-1	1	1	1	-1	1	0.11	1
5	1	-1	-1	1	1	1	-1	1	-1	-0.31
6	-1	-1	1	1	1	1	-1	1	-1	1

Tabla 5.4: Se muestran los diseños óptimos según los dos distintos criterios para $\alpha = 0.75$ y con las distribuciones iniciales modificadas.

Primeramente, independientemente del valor de α , ya no ocurre que todos los elementos de la matriz de diseño sean ± 1 exceptuando a los elementos de la diagonal. Más aún, en el diseño con las iniciales originales todos los valores de cada covariable (cada columna) siempre valían lo mismo, excepto en un ensayo. En el caso de las iniciales modificadas lo que ocurre es que (i) los elementos de la diagonal también valen ± 1 y (ii) las covariables ya no solo toman un valor (sin contar el valor de la diagonal que era distinto), sino que ahora valen 1 en algunos ensayos y -1 en otros.

Por otro lado al cambiar de $\alpha = 0.5$ a $\alpha = 0.75$ los diseños con las iniciales modificadas quedaron prácticamente iguales, salvo en dos entradas que son *prácticamente* ± 1 . En resumen, la modificación de distribuciones iniciales

introdujo cambios considerables en los diseño óptimos.

Por su parte, las Figuras 5.10 y 5.11 muestran las proyecciones en una y dos dimensiones de los diseños óptimos para $\alpha = 0.5$ y $\alpha = 0.75$, respectivamente (suavizando con un kernel Gaussiano para obtener la densidad de la proyección unidimensional).

Algo interesante es que las correlaciones entre las covariables también cambiaron: si bien antes eran ± 0.2 ahora no es así. En el caso de $\alpha = 0.5$ algunas correlaciones parecen estar alrededor de $1/3$, otras alrededor de 0 y una última igual a $1/4$. Esto se repite para el caso de $\alpha = 0.75$, donde además las correlaciones ya no están alrededor de estos valores sino que directamente eso valen. Puntualmente las correlaciones entre x_1, x_2 y x_3 son iguales a $1/3$ y las respectivas a x_4 y x_5 con las demás covariables son 0, excepto precisamente por la correlación entre estas dos covariables, la cual toma el valor de $1/4$. En resumen, pareciera que las covariables se separaron en dos grupos: x_1, x_2 y x_3 en un lado, correlacionadas entre sí por $1/3$ pero no correlacionadas con las últimas dos covariables; y por otro lado x_4 y x_5 , correlacionadas entre sí por $1/4$ pero no correlacionadas con las primeras tres covariables.

Finalmente las Figuras 5.12 y 5.13 muestran la evolución de la estimación de la pérdida esperada conforme el número de iteración avanza, para ambos valores de α respectivamente. En ambos casos la pérdida esperada parece haber convergido, por lo que se concluye lo mismo del método.

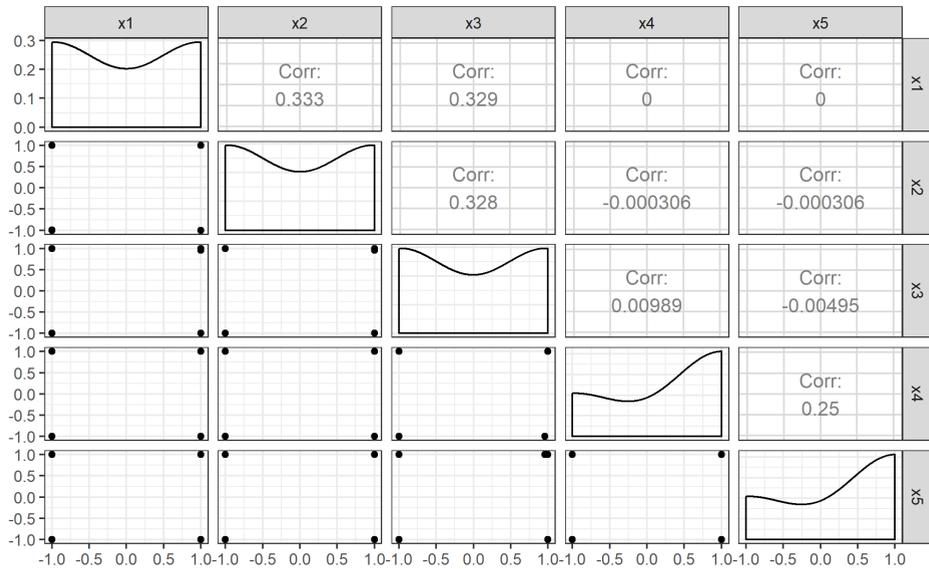


Figura 5.10: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión Poisson (5.8) con distribuciones iniciales modificadas, así como su correlación, para $\alpha = 0.5$.

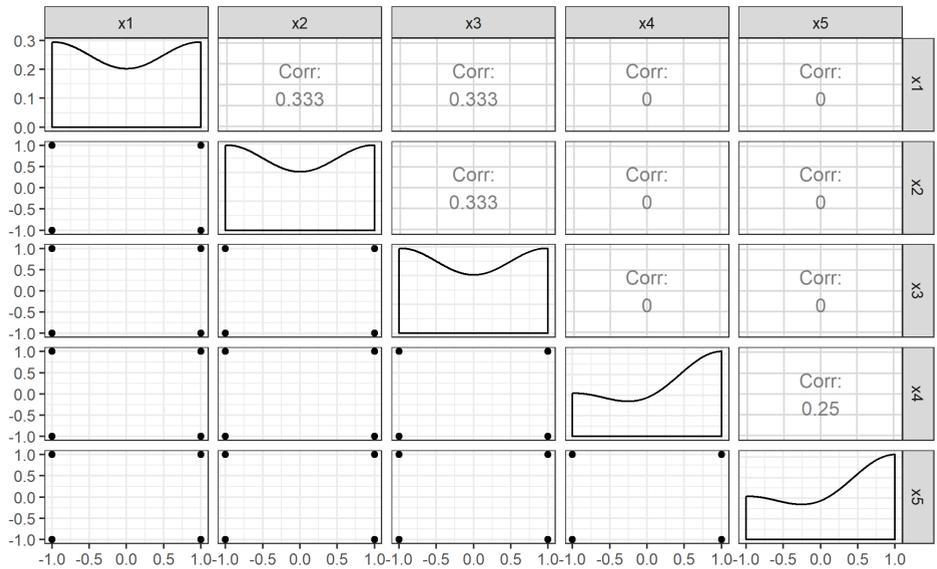


Figura 5.11: Se muestran proyecciones en una y dos dimensiones de las tres variables del modelo de regresión Poisson (5.8) con distribuciones iniciales modificadas, así como su correlación, para $\alpha = 0.75$.

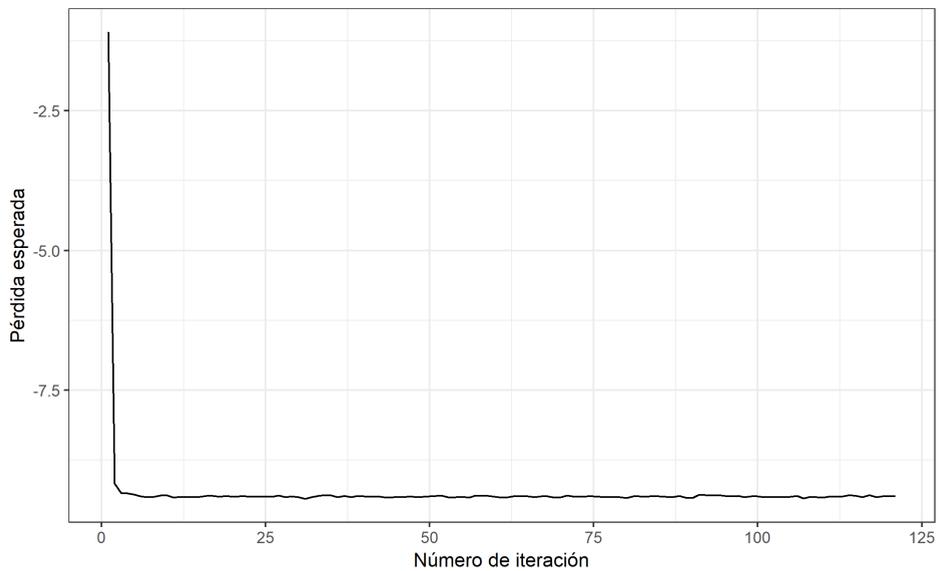


Figura 5.12: Para la regresión Poisson con distribuciones iniciales modificadas y $\alpha = 0.5$, se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

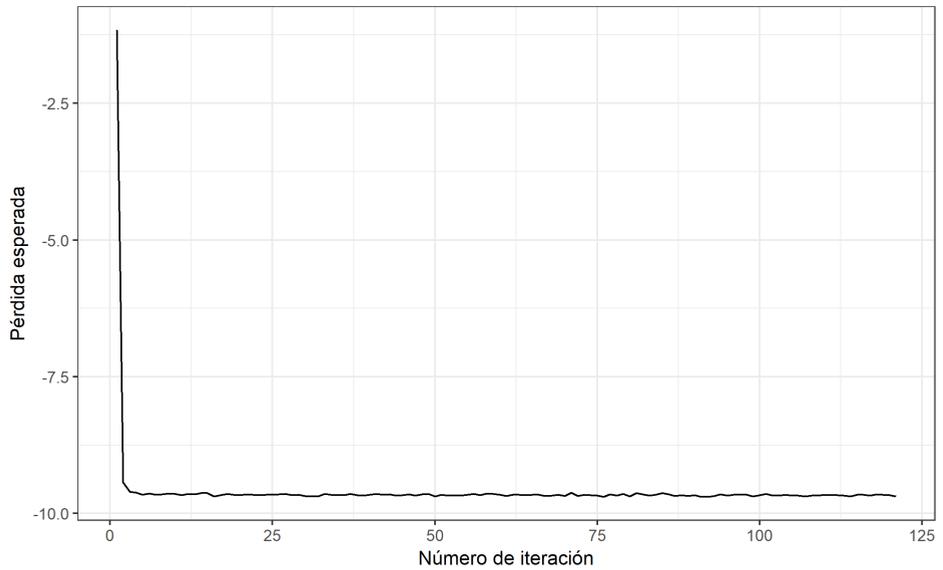


Figura 5.13: Para la regresión Poisson con distribuciones iniciales modificadas y $\alpha = 0.75$, se muestran aproximaciones a la utilidad esperada de los diseños en cada iteración del algoritmo para corroborar convergencia.

En el Apéndice C se incluye el código utilizado para encontrar los diseños óptimos para cada uno de los ejemplos, así como para generar las gráficas aquí mostradas.

Capítulo 6

Conclusiones

The concept of the statistician as one who analyzes someone else's data is flawed, but equally inappropriate is the idea of the statistician engaged only in the design and analysis of an individual experiment.

— Box y Liu, *Statistics as a Catalyst to Learning by Scientific Method Part I—An Example* (1999)

La importancia del diseño de experimentos radica en que éste permite que la recolección de datos se realice acorde a algún criterio de optimalidad, el cual debe reflejar el objetivo final del estudio en cuestión. Esto eleva la calidad de cualquier análisis inferencial hecho con base en los datos, pues permite que quien realiza el experimento modifique las covariables y observe el valor resultante de la variable respuesta en un ambiente controlado.

El problema de diseño experimental se puede plantear de manera natural como un problema de decisión, lo cual lo coloca en el marco de la Estadística Bayesiana. Sin embargo inmediatamente aparecen complicaciones en la formulación; en particular en la mayoría de los casos no existen soluciones analíticas. Más aún, dichas soluciones son difíciles de encontrar incluso mediante métodos computacionales. Una estrategia que ha resultado de utilidad es suponer conocido el modelo de muestreo de la variable respuesta,

pues esto puede simplificar algunos de los cálculos numéricos. Los modelos lineales generalizados son de gran ventaja en este caso, pues son lo suficientemente generales como para que muchos fenómenos se puedan modelar con ellos pero a la vez han sido estudiados con profundidad en la literatura.

El método ACE es uno de los algoritmos más recientes para encontrar diseños óptimos (Overstall y Woods, 2016), y es particularmente útil en el marco de modelos lineales generalizados y criterios de optimalidad Bayesiana populares. En esta tesis se utilizó este método para abordar dos problemas de diseño de experimentos para modelos lineales generalizados: una regresión logística y una regresión Poisson. Ambos provienen de un artículo publicado por Woods et al. (2017), en donde los autores muestran distintas aplicaciones del método ACE.

En el caso de la regresión logística el diseño encontrado fue consistente con el encontrado por Woods et al. Para la regresión Poisson se encontraron diseños óptimos utilizando un criterio de optimalidad distinto al propuesto por los autores, lo que implicó consecuencias interesantes en las que se ahondará más adelante. Por otra parte se discutió el hecho de que, en ambos ejemplos, Woods et al. utilizaron distribuciones iniciales que presuponían a los parámetros del modelo como diferentes de cero. Por ello se obtuvieron diseños óptimos con distribuciones iniciales modificadas que evitaran este supuesto.

No está de más recordar que, en el ejemplo de la regresión logística, se ilustró la dificultad que se encuentra al buscar una expresión analítica para la pérdida esperada. En el camino se resaltó la complejidad inherente a los cálculos relacionados con la optimización de la pérdida esperada. Este ejemplo ilustra por qué es necesario, en muchos casos, recurrir a métodos computacionales que asistan con la obtención del diseño óptimo. Por otra parte, al modificar el soporte de las distribuciones iniciales de forma que éste contuviera al cero, se encontró un diseño óptimo diferente al original. Puntualmente se observó que las únicas covariables afectadas eran las co-

rrespondientes a los parámetros cuyas distribuciones iniciales marginales fueron modificadas, y en ese caso ocurrió que la correlación entre estas covariables disminuyó considerablemente en el nuevo diseño óptimo.

El ejemplo de la regresión Poisson fue exhaustivo y dio lugar a algunos hallazgos que merecen mayor discusión. En primer lugar el hecho de utilizar criterios de optimalidad distintos derivó en diseños óptimos diferentes. Más aún, la distribución final del parámetro a la que cada diseño dio lugar difirió según el criterio de optimalidad empleado. Por otro lado se modificó la distribución inicial del parámetro. A diferencia del ejemplo de la regresión logística, en este caso se cambió tanto el espacio parametral como la forma de la distribución inicial. Ello derivó en diseños óptimos radicalmente distintos de los originalmente encontrados. Los diseños modificados no mostraban el patrón visto en los originales, es decir, columnas constantes (e iguales a ± 1) salvo en una entrada, sino que contenían combinaciones de ± 1 en todas sus entradas.

De estos dos ejemplos se concluye que tanto el criterio de optimalidad como la distribución inicial del parámetro son elementos de suma importancia y que pueden tener un impacto considerable en el diseño óptimo encontrado, además de que se resaltó la necesidad de emplear métodos computacionales debido a la complejidad analítica del diseño Bayesiano de experimentos.

Esta tesis logró entonces, por una parte, desarrollar la teoría detrás del diseño Bayesiano de experimentos y, por otra, ilustrar la implementación del método ACE para obtener diseños Bayesianos óptimos. Esto último se hizo reproduciendo satisfactoriamente dos resultados de la literatura, pero también modificando los factores que afectan los diseños óptimos encontrados. Finalmente hay algunos puntos importantes que merecen mayor discusión.

Primeramente, si bien el método ACE permite resolver problemas bastante generales de diseño experimental, también es cierto que la implementación de dicho método no es trivial. Muchos procedimientos frecuentistas, como

ajustar un modelo de regresión, son sumamente sencillos de realizar con ayuda de software especializado. Incluso algunos métodos computacionales Bayesianos, como el muestreo de Gibbs para generar muestras de una distribución final, son cada vez más accesibles. Sin embargo, debido a que muchas de las componentes del diseño de experimentos varían por estudio, éstas deben ser especificadas para cada análisis realizado. Es por ello que generalizar un método para este tipo de problemas es complicado, y la implementación termina siendo compleja en el mejor de los casos.

Por otro lado la eficiencia computacional del método depende altamente de las especificaciones del usuario. Si la función de pérdida es muy compleja el método puede tardar en terminar e, incluso cuando lo hace, la convergencia puede no haberse alcanzado. Esto se soluciona parcialmente utilizando el paquete en R de los autores (Overstall et al., 2017), pues incluye criterios de optimalidad predeterminados. Sin embargo no se debe olvidar que la función de pérdida debe reflejar los objetivos del experimento; si ninguna de las funciones incluidas en dicho paquete satisface esto, entonces se debe especificar una adecuada.

Un punto que es relevante en el diseño de experimentos pero que no fue abordado en esta tesis es el del costo del experimento. Se ha insistido en que una de las ventajas del diseño experimental es que permite obtener información de calidad satisfaciendo restricciones presupuestales. Sin embargo este punto no se consideró al momento de desarrollar la teoría detrás del diseño de experimentos, ni en las aplicaciones del Capítulo 5. No está de más mencionar que es posible incorporar este elemento en la función de pérdida, de manera similar a como lo hace Bernardo (1997) cuando considera el problema de la elección del tamaño de muestra.

El diseño Bayesiano de experimentos es un campo activo de investigación, particularmente por las dificultades prácticas inherentes a la solución de problemas reales. Sin embargo los últimos años han visto un despegue en el interés hacia esta área, acarreado el desarrollo de métodos ingeniosos

que permiten vencer dichas dificultades (Overstall y Woods, 2016; Woods et al., 2017). Si bien el enfoque clásico es aún mucho más popular, también es cierto que la contraparte Bayesiana ofrece ventajas sobre éste, como la capacidad de incorporar información preliminar en el análisis en forma de distribuciones iniciales.

Futuras investigaciones pueden estar enfocadas a desarrollar nuevos métodos que faciliten la resolución de este tipo de problemas, es decir, cuya implementación sea más accesible y, a la vez, se puedan utilizar para diseños con un mayor número de ensayos y funciones de pérdida más generales. Esto tendría un impacto relevante en todas las áreas que con regularidad diseñan estadísticamente sus experimentos; particularmente, áreas industriales y médicas en donde los recursos son limitados y se requiere obtener la mayor cantidad de información gastando la menor cantidad posible de presupuesto.

Apéndice A

Métodos computacionales Bayesianos

Este apéndice tiene como propósito desarrollar con mayor profundidad los métodos computacionales Bayesianos más actuales que se mencionaron en el Capítulo 2, particularmente aquellos que tienen relevancia en la implementación del método ACE discutido en el Capítulo 4. Puntualmente se discutirán varios métodos de Monte Carlo. Cabe mencionar que este apéndice no pretende demostrar que dichos algoritmos convergen, sino exponerlos para referencias en la tesis. Si el lector está interesado en dichas demostraciones se recomienda revisar las notas de Gutiérrez-Peña (2016c), las cuales se seguirán de cerca, así como la bibliografía ahí citada

Consideraciones iniciales

Supondremos que contamos con una muestra aleatoria $X = (X_1, \dots, X_n)^T$ de una variable aleatoria con función de probabilidad generalizada $f(x | \theta)$, donde $\theta \in \Theta$ es un parámetro (de cualquier dimensión) desconocido, pero la forma de f se supone totalmente conocida. La distribución inicial de θ será $p(\theta)$, la cual es totalmente conocida.

Al estimador de máxima verosimilitud de θ se le denotará por $\hat{\theta}$, y supondremos que es tal que

$$\nabla \ell(\hat{\theta}) = 0,$$

donde ℓ es la log-verosimilitud del modelo y la diferenciación se toma con respecto a θ .

A.1. Métodos de Monte Carlo

Los métodos de Monte Carlo son una vasta familia de algoritmos que utilizan la generación repetida de números aleatorios para producir valores numéricos de interés. Generalmente el propósito es estimar cantidades que, aunque pueden ser determinísticas en principio, estén sujetas a alguna interpretación probabilística. La idea que subyace en los orígenes de los métodos de Monte Carlo es la siguiente. Sea $\theta \in \mathbb{R}^m$ un vector que interesa calcular y que habitualmente resulta de evaluar una función H en un elemento w de \mathbb{R}^k . Esto es,

$$\begin{aligned} H : \mathbb{R}^k &\rightarrow \mathbb{R}^m \\ w &\mapsto \theta. \end{aligned}$$

Si para cada vector w es posible construir una distribución de probabilidad $p(u)$ en \mathbb{R}^k que satisfaga que $\mathbb{E}[H(U)] = \theta$, entonces es posible simular una muestra aleatoria u_1, u_2, \dots, u_N y aproximar θ con

$$\hat{\theta} = \frac{1}{N} \sum_{k=1}^N H(u_k).$$

Una de las aplicaciones principales de dichos métodos es la estimación de integrales. Por ejemplo, sea $g : [a, b] \rightarrow \mathbb{R}$ una función acotada y considere

$$I = \int_a^b g(x) dx.$$

Note entonces que

$$I = (b - a) \int_a^b \frac{1}{b - a} g(x) dx = (b - a) \mathbb{E}[g(X)],$$

donde $X \sim \mathcal{U}(a, b)$, pues así su función de densidad es $f_X(x) = \frac{1}{b-a} \mathbf{1}_{(a,b)}(x)$. Entonces es posible aproximar I con

$$\hat{I} = \frac{b - a}{N} \sum_{k=1}^N g(x_k),$$

donde $x_1, \dots, x_N \sim \mathcal{U}(a, b)$ de manera independiente. Por la Ley de los Grandes Números \hat{I} converge a I conforme N tiende a infinito, por lo que basta escoger N lo suficientemente grande para tener una aproximación de buena calidad.

Note que el problema de encontrar la integral definida de una función es puramente determinista y, sin embargo, se ofreció una solución (si bien aproximada) probabilística. La idea de pensar integrales como valores esperados es muy explotada por esta clase de métodos.

A.1.1. Muestreo por importancia

Otra aplicación muy popular de los métodos de Monte Carlo para estimar integrales es el muestreo por importancia. Si se desea estimar

$$I = \int_{\Theta} g(\theta) d\theta,$$

la idea es escribir

$$I = \int_{\Theta} \frac{g(\theta)}{s(\theta)} s(\theta) d\theta,$$

donde $s(\theta)$ es una densidad de probabilidades sobre Θ , llamada *distribución de muestreo*. Note que el ejemplo de la sección anterior es un caso particular de muestreo por importancia, en el que $\Theta = [a, b]$ y s es la densidad de

una variable aleatoria uniforme en Θ . Análogamente, se debe generar una muestra aleatoria $\theta_1, \dots, \theta_N$ de $s(\theta)$ y se aproxima I a través de

$$I \approx \frac{1}{N} \sum_{i=1}^N \frac{g(\theta_i)}{s(\theta_i)}.$$

Gutiérrez-Peña (2016b, Capítulo 4.2) discute las principales características de este estimador, como insesgamiento y varianza. Cabe resaltar que la elección de la distribución de muestreo es importante. En particular, ésta debe de tener una forma relativamente similar a la de g .

A.1.2. Muestreo-remuestreo

En muchas ocasiones (y ciertamente en problemas de inferencia Bayesiana) se requiere generar una muestra de una distribución que es conocida, salvo por cierta constante de normalización. Esto es, usualmente se conoce el *kernel* de la distribución (por ejemplo de la final de θ) sobre la cual se requiere hacer inferencia. En este caso, requerimos una muestra de

$$g(\theta) = \frac{g_0(\theta)}{\int g_0(\tilde{\theta}) d\tilde{\theta}},$$

donde la forma funcional de g_0 es totalmente conocida, pero no así la constante que aparece en el denominador.

Más aún, es común que conozcamos alguna densidad $s(\theta)$ de la cual sí podemos obtener muestras con facilidad.¹ El método de muestreo-remuestreo permite obtener una muestra de $g(\theta)$ a partir de $g_0(\theta)$ y una muestra de $s(\theta)$, y se divide en dos casos.

Caso 1

En este caso se supone que existe una constante conocida $M > 0$ tal que $g_0(\theta)/s(\theta) \leq M$ para todo $\theta \in \Theta$. El Código A.1 muestra el algoritmo de muestreo-remuestreo para este caso.

¹Desde luego que s y g tienden a ser similares.

```

1  Input : Funciones  $g_0$  y  $s$ , y número máximo de iteraciones  $N$ 
2  Output: Muestra de  $g$ 
3
4  for (i in 1:N)
5    Genera  $\tilde{\theta} \sim s(\theta)$ 
6    Genera  $u \sim \mathcal{U}(0,1)$ 
7    if (  $u \leq g_0(\tilde{\theta})/Ms(\tilde{\theta})$  )
8      Acepta  $\tilde{\theta}$  como una observación de  $g(\theta)$ 
9    end
10 end

```

Código A.1: Método de muestreo-remuestreo para el Caso 1.

Note que no se puede determinar de antemano el tamaño de la muestra generada. Si se desea fijar este valor, entonces el ciclo `for` del Código A.1 se debe reemplazar por un ciclo `while`, donde la condición de paro será que el tamaño de la muestra sea el requerido.

Caso 2

En este caso no existe dicha M (o no puede encontrarse). Una manera de obtener muestras de $g(\theta)$ es generar una muestra aleatoria $\theta_1, \dots, \theta_N$ de $s(\theta)$ y definir

$$w_i = \frac{v_i}{\sum_{k=1}^N v_k},$$

con $v_i = g_0(\theta_i)/s(\theta_i)$ para cada $i = 1, \dots, N$. Ahora podemos definir una distribución discreta de probabilidades en el conjunto $\{\theta_1, \dots, \theta_N\}$, de manera que $P(\theta = \theta_i) = w_i$ para cada i . Es posible probar que si se obtiene una observación $\tilde{\theta}$ de esta distribución discreta entonces ésta proviene (aproximadamente) de la distribución de interés, $g(\theta)$. Solo basta repetir este proceso hasta tener una muestra de tamaño suficiente.

A.2. Monte Carlo vía cadenas de Markov

Esta clase de métodos pretenden construir una cadena de Markov que converja a una distribución límite que sea igual a la distribución de la cual se

quiere muestrear (la distribución final, por ejemplo).

Uno de los algoritmos más populares de esta clase es el de Metropolis-Hastings (M-H). La idea es seleccionar una distribución de transición $Q(\theta^* | \theta)$ cuyo soporte sea al menos el de $p(\theta)$. Se define el *cociente de Hastings* como

$$\alpha(\theta^*, \theta) = \min \left\{ \frac{p(\theta^* | x) Q(\theta | \theta^*)}{p(\theta | x) Q(\theta^* | \theta)}, 1 \right\}. \quad (\text{A.1})$$

Dicho cociente siempre está entre 0 y 1, y servirá como la probabilidad de aceptar una observación generada en cada iteración. Note que solo es necesario conocer $p(\theta | x)$ hasta una constante de normalización, pues en el cociente de Hastings ésta se cancela. Es por ello que resulta sumamente útil para el cálculo de la distribución final. De forma análoga a las secciones anteriores se tiene que

$$p(\theta | x) = \frac{\pi_0(\theta, x)}{\int \pi_0(\tilde{\theta}, x) d\tilde{\theta}},$$

donde $\pi_0(\theta, x) = p(x | \theta)p(\theta)$. El Código A.2 muestra el algoritmo de Metropolis-Hastings.

```

1  Input : Funciones  $Q(\theta)$ ,  $p(\theta)$  y  $p(x|\theta)$ 
2          Muestra  $x_1, \dots, x_n$ 
3          Tamaño de muestra  $N$ 
4  Output: Muestra de  $p(\theta|x)$ 
5
6  Genera  $\theta_0 \in \Theta$  arbitrariamente
7
8  for(i in 1:N)
9    Genera  $\theta_s \sim Q(\theta_s|\theta_{i-1})$ 
10   Determina  $\alpha(\theta_s, \theta_{i-1})$ 
11   Genera  $u \sim \mathcal{U}(0,1)$ 
12   if(  $u < \alpha$  )
13     Define  $\theta_i = \theta_s$ 
14   else
15     Define  $\theta_i = \theta_{i-1}$ 
16   end
17 end
```

Código A.2: Algoritmo de Metropolis-Hastings.

A.2.1. Muestreo de Gibbs

Quizás el algoritmo más popular de los métodos MCMC es el muestreo de Gibbs, el cual resulta ser un caso particular de M-H que ganó popularidad a principios de los años 90 (ver Gelfand y Smith, 1990; Geman y Geman, 1984).

A diferencia del algoritmo de M-H, cada nuevo valor en la cadena se obtiene mediante un proceso iterativo que se basa en la generación de observaciones de distribuciones de menor dimensión que la de θ , digamos d . En particular, supongamos que $\theta = (\theta_1, \dots, \theta_k)^T$ es una partición de θ , con $\theta_i \in \mathbb{R}^{d_i}$ y $\sum_i d_i = d$. Se definen las *densidades condicionales completas* como

$$\begin{aligned} p(\theta_1 | \theta_2, \dots, \theta_k, x), \\ p(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k, x), \quad (i = 2, \dots, k-1), \\ p(\theta_k | \theta_1, \dots, \theta_{k-1}, x). \end{aligned}$$

Generalmente éstas son fáciles de identificar inspeccionando la forma de (proporcional) de $p(\theta | x)$.

En el muestreo de Gibbs la probabilidad de transición siempre es igual a 1 (es decir, el cociente de Hastings definido en (A.1) siempre valdrá 1). Así, dado un valor inicial $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$, el muestreo de Gibbs genera una cadena de Markov en la que $\theta^{(t+1)}$ se escoge a partir de $\theta^{(t)}$ como sigue:

$$\begin{aligned} \text{Generar } \theta_1^{(t+1)} \text{ de } p(\theta_1 | \theta_2^{(t)}, \dots, \theta_k^{(t)}, x), \\ \text{Generar } \theta_2^{(t+1)} \text{ de } p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, x), \\ \vdots \\ \text{Generar } \theta_k^{(t+1)} \text{ de } p(\theta_k | \theta_1^{(t)}, \dots, \theta_{k-1}^{(t+1)}, x). \end{aligned}$$

Si este proceso se itera un número suficientemente grande de veces, digamos M , entonces se puede considerar que $(\theta_1^M, \theta_2^M, \dots, \theta_k^M)$ es una observación que proviene aproximadamente de $p(\theta | x)$.

No se mostrará que dichos métodos en efecto estiman lo que pretenden, ya que esto rebasaría el propósito de esta tesis. Sin embargo, se reitera al lector interesado que consulte la bibliografía recomendada, en especial Chib y Greenberg (1995); Gutiérrez-Peña (2016c); Hitchcock (2003).

Apéndice B

Modelos de regresión de procesos Gaussianos

En este apéndice se discute con mayor detalle el tema de regresión de procesos Gaussianos. Es con un modelo de este estilo con el que se estima la pérdida esperada como función de una sola coordenada del diseño en el método ACE.

Modelo de regresión de proceso Gaussiano

Recordemos que en este punto del método se tienen m pares de observaciones de la forma $\left\{x_l, \tilde{\Phi}_{ij}(x_l | \mathbf{x})\right\}_{l=1}^m$. La idea es buscar una función que explique la relación entre las x_l y las estimaciones de la pérdida esperada evaluada en ellos. Por ejemplo, en primera instancia podría pensarse en ajustar un modelo de regresión lineal que explique dicha relación, aunque nada indica que exista una relación lineal entre las observaciones y las covariables. Sin embargo, en el mismo espíritu Overstall y Woods (2016) proponen ajustar un modelo de regresión basado en procesos Gaussianos, los cuales definimos a continuación.

Definición B.0.1 (Proceso Gaussiano). Un proceso Gaussiano es un proceso estocástico $(Y_t)_{t \in \mathcal{T}}$, con \mathcal{T} un conjunto de índices, tal que dada una colección finita de índices $t = (t_1, t_2, \dots, t_r)$ el vector aleatorio $Y = (Y_{t_1}, Y_{t_2}, \dots, Y_{t_r})$ sigue una distribución conjunta Normal multivariada.

Es decir, en el contexto de procesos Gaussianos todo vector aleatorio $Y = (Y_{t_1}, Y_{t_2}, \dots, Y_{t_r}) \sim \mathcal{N}_r(\mu, \Sigma)$. Note que el vector μ y la matriz Σ dependen del vector aleatorio Y en cuestión. En general es posible preguntarse por esta relación, de donde se derivan la función de media, $\mu(t)$, y la función de covarianzas, $k(t, t')$ para todo vector aleatorio Y . Es con ellas que se calcula, para cada vector Y de dimensión r , la media μ y la matriz de covarianzas $\Sigma = (\Sigma_{ij}) = (k(t_i, t_j))$ de la distribución normal que describe el comportamiento conjunto de Y .

La idea de la regresión de proceso Gaussiano es la siguiente. Dados m puntos $\{x_l, f(x_l)\}_{l=1}^m$, se quiere obtener la forma de f , o por lo menos sus valores para un conjunto predefinido de valores en el dominio. Recordemos que, en el caso de la regresión lineal, se plantea el modelo

$$f(x_l) = \beta_0 + \beta^T x_l + \epsilon_l, \quad l = 1, \dots, m.$$

Después se asigna una distribución inicial a los parámetros β y, con base en los datos observados y el teorema de Bayes, se obtiene la distribución final de los parámetros. Como se discutió previamente, en este caso no es sensato asumir que existe una relación lineal (ni de ninguna forma en particular) entre la covariable y la variable respuesta. La regresión de proceso Gaussiano lleva este razonamiento un nivel más arriba, reconociendo que la incertidumbre en este caso es sobre la misma forma de la función f y no sobre un conjunto de parámetros. Ergo, se asigna una distribución inicial a la función y, con base en los valores observados de ésta, se obtiene su distribución final.

La manera de asignar una distribución a una función es pensándola como una colección infinita de puntos, cada uno de los cuales es una observación de una cierta variable aleatoria. Luego, la función termina siendo una

realización de un proceso estocástico continuo. Si además suponemos, por ejemplo, que cada punto es una realización de una variable aleatoria Normal, la función será una realización de un proceso Gaussiano. Las funciones de media y covarianza son, en este contexto, los hiperpárametros del modelo, pues son los que determinan las características del proceso en cuestión.

Usualmente la función de media se supone, por lo menos a priori, igual a cero. Esto no tiene mayores implicaciones teóricas y se asume comúnmente. La razón es que la clave de la regresión de proceso Gaussiano es la función de covarianzas, k . Ésta determina, dados dos puntos, qué tan similares son entre sí. La idea es que dos puntos cercanos produzcan valores de la función cercanos. Overstall y Woods (2016) emplean una función de covarianzas cuadrática exponencial, definida por

$$k(x_1, x_2) = \exp \left\{ -\frac{(x_1 - x_2)^T(x_1 - x_2)}{2\ell^2} \right\}, \quad (\text{B.1})$$

donde ℓ es un parámetro que controla qué tanta influencia tienen los puntos entre sí.

Juntando lo anterior, lo que se propone es pensar a la función f como una realización de un proceso Gaussiano. Se le debe asignar una distribución inicial, es decir, un proceso Gaussiano a priori. Lo que Overstall y Woods (2016) proponen es suponer que la función de media inicial es igual a cero y la función de varianzas está dada por (B.1) con $\ell = 1$.

Ahora bien, dado que se tienen m observaciones de valores en el dominio y su correspondiente valor de la función, el siguiente paso es obtener el proceso Gaussiano posterior, el cual describe a la función f habiendo observado los datos. Sea $\hat{x} \in \mathcal{X}_j$ un valor cualquiera en el dominio de la función f . Note que, por todo lo anterior, el vector $(f(x_1), f(x_2), \dots, f(x_m), f(\hat{x}))$ sigue una distribución conjunta Normal multivariada, con vector de medias y matriz de covarianzas determinados por las funciones correspondientes.¹ Es posi-

¹El vector de medias tendrá las primeras m componentes iguales a cero.

ble entonces obtener la distribución condicional de $f(\hat{x}) | f(x_1), \dots, f(x_m)$, y esto se puede hacer empleando las propiedades de la distribución Normal multivariada. Particularmente dicha distribución será también una distribución Normal, con la media y la varianza determinadas por los valores observados.

Resumiendo, la manera en la que se obtiene $\tilde{\Phi}_{ij}(x)$ a partir de $\left\{x_l, \tilde{\Phi}_{ij}(x_l | \mathbf{x})\right\}_{l=1}^m$ es:

1. Pensar en $\tilde{\Phi}_{ij}(x)$ como un proceso Gaussiano, es decir, como una colección infinita de observaciones de variables aleatorias Normales.
2. Asignar una distribución inicial a $\tilde{\Phi}_{ij}(x)$, particularmente un proceso Gaussiano de función de media cero y función de varianzas como en (B.1) con $\ell = 1$.
3. Obtener la distribución posterior de $\tilde{\Phi}_{ij}(x)$ como la distribución marginal condicional de $\tilde{\Phi}_{ij}(x) | \tilde{\Phi}_{ij}(x_1), \tilde{\Phi}_{ij}(x_2), \dots, \tilde{\Phi}_{ij}(x_m)$.
4. La media de dicha distribución dependerá de x , y ésta se utilizará como la estimación puntual de $\tilde{\Phi}_{ij}(x)$, es decir, el emulador Gaussiano mencionado previamente es precisamente la función de media posterior de $\tilde{\Phi}_{ij}(x)$ (para una sola $x \in \mathcal{X}_j$).

Finalmente se muestran explícitamente las fórmulas para obtener el emulador, las cuales se derivan encontrando la media de la distribución condicional de una entrada de un vector aleatorio Normal multivariado. Como (Overstall y Woods, 2016; Woods et al., 2017) muestran, el emulador Gaussiano de la pérdida esperada como función de una coordenada i, j del diseño está dado por

$$\tilde{\Phi}_{ij}(x) = \hat{\mu}_{ij} + \hat{\sigma}_{ij} a(x, \zeta_{ij})^T A(\zeta_{ij})^{-1} z_{ij}. \quad (\text{B.2})$$

Aquí,

- $\hat{\mu}_{ij} = \frac{1}{m} \sum_{l=1}^m \tilde{\Phi}_{ij}(x_l | \mathbf{x})$.
- $\hat{\sigma}_{ij}^2 = \frac{1}{m-1} \sum_{l=1}^m \left(\tilde{\Phi}_{ij}(x_l | \mathbf{x}) - \hat{\mu}_{ij} \right)^2$.
- $z_{ij} \in \mathbb{R}^m$ es un vector cuya l -ésima entrada es igual a la pérdida esperada estandarizada,

$$\frac{\tilde{\Phi}_{ij}(x_l | \mathbf{x}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}.$$

- $\zeta_{ij} = \{x_1, \dots, x_m\}$ son los m puntos provenientes de \mathcal{X}_j .
- $a(x, \zeta_{ij}) \in \mathbb{R}^m$ es un vector cuya u -ésima entrada es

$$\exp \left\{ -\rho(x - x_u)^2 \right\}.$$

- A es una matriz de $m \times m$ cuya uv -ésima entrada es

$$\exp \left\{ -\rho(x_u - x_v)^2 \right\} + \xi \mathbf{1}(u = v),$$

donde $\xi > 0$ es un parámetro del emulador conocido como *nugget* en inglés y $\mathbf{1}(\cdot)$ es la función indicadora. La forma de este vector y matriz se deriva de la elección de la función de varianzas cuadrática exponencial.

El parámetro ξ se incluye para evitar que el algoritmo interpole los valores, y que más bien los suavice. La idea es que, en realidad, las observaciones de $\tilde{\Phi}_{ij}(x_l)$ son estimaciones de la verdadera pérdida $\Phi_{ij}(x_l)$; se desea reflejar esta incertidumbre adicional y ello se logra incorporando un factor aditivo de ruido en la diagonal de la matriz A . Tanto ρ como ξ se estiman en cada iteración vía máxima verosimilitud, particularmente utilizando el método de Newton (*Fisher's scoring* en inglés) (Overstall y Woods, 2016, pág. 5).

Finalmente cabe mencionar que el tema de regresión de procesos Gaussianos tiene mucho mayor profundidad que la aquí presentada. Además

ha ganado popularidad en años recientes, particularmente en áreas como aprendizaje de máquina y experimentos computacionales. Se recomienda ampliamente revisar el libro de Rasmussen y Williams (2006) y la bibliografía ahí citada para un análisis más detallado.

Apéndice C

Código

En este apéndice se muestra el código utilizado en la tesis; en particular, la implementación del algoritmo ACE para los ejemplos del Capítulo 5.

C.1. Diseño para modelo de regresión logística

El código para el modelo de regresión logística del Capítulo 5 fue programado en R 3.5.0 (R Core Team, 2018) y ejecutado en una computadora ASUS ROG GL552VX con sistema operativo Windows 10 64-bit, 8 GB de memoria RAM, y procesador Intel Core i7-6700HQ, con velocidad de reloj de 2.60GHz. El tiempo que tardó en correr fue de 703.70 segundos, es decir, 11 minutos, 43 segundos, y 70 centésimas de segundo.

```
1 # Semilla para reproducibilidad:
2 set.seed(10021995)
3
4
5 # Preámbulo ----
6 library(MASS)
7 library(acebayes)
8 library(tidyverse)
9 library(GGally)
10 ggplot2::theme_set(theme_bw())
```

```

11 |
12 | # Parámetros iniciales ---- ### ### ### ###
13 |
14 | # Tamaño de muestra ----
15 | ## Seleccionado acorde al artículo citado en la tesis
16 | n <- 16
17 |
18 |
19 | # Diseño inicial ----
20 | ## Éste se selecciona aleatoriamente. Es una matriz
21 | ## de 16 x 3, donde cada elemento es un número
22 | ## tomado de una distribución uniforme en (-1.2872, 1.2872)
23 | start.d <- matrix(runif(n=3*n, -1.2872, 1.2872),
24 |                 ncol = 3, nrow = n,
25 |                 byrow = TRUE,
26 |                 dimnames = list(as.character(1:n), c("x1", "x2"
27 |                                     , "x3")))
28 |
29 |
30 | # Distribución inicial ----
31 | ## Ésta es una función que regresa una matriz de B x 10,
32 | ## donde las entradas se distribuyen (independientemente)
33 | ## como se menciona en la tesis
34 | prior <- function(B){
35 |
36 |   # Inicializa como ceros
37 |   p <- matrix(rep(0, B*10), ncol = 10, nrow = B)
38 |
39 |   # Utiliza distribuciones definidas en la tesis
40 |   p[, c(1, 4:10)] <- runif(B*8, -2, 2)
41 |   p[, 2:3] <- runif(B*2, 2, 6)
42 |
43 |   return(p)
44 | }
45 |
46 |
47 | # Diseño óptimo ---- ### ### ### ###
48 | ## Se obtiene mediante el algoritmo ACE de Overstall y Woods
49 | ## con los parámetros especificados en la tesis
50 | eta <- acebayes::aceglm(formula=~x1+x2+x3+I(x1^2)+I(x2^2)+I(x3^2)+
51 |                         I(x1*x2)+I(x1*x3)+I(x2*x3),
52 |                         start.d = start.d,
53 |                         family = binomial,
54 |                         prior = prior,
55 |                         method = "MC",
56 |                         B = c(20000, 1000),

```

```
56         criterion = "D",
57         Q = 10,
58         lower = -1.2872,
59         upper = 1.2872)
60
61 # Muestra el resumen del diseño óptimo
62 eta
63
64 # Muestra el diseño óptimo
65 eta$phase2.d
66
67 # Compara gráficamente las variables
68 GGally::ggpairs(as.data.frame(eta$phase2.d))
69 ggplot2::ggsave("log_reg_fig.png")
70
71
72 # Guarda el diseño en otra variable
73 eta1 <- eta
74
75
76 # Determina convergencia ---- ### ### ### ####
77
78
79 eta1.phase <- -c(eta1$phase1.trace, eta1$phase2.trace)
80 m <- length(eta1.phase)
81 eta1.phase <- data.frame("Número" = 1:m, "Utilidad_esperada" =
      eta1.phase)
82
83 # Genera y guarda imagen
84 ggplot2::ggplot(eta1.phase, aes(x = Número, y = Utilidad_esperada))
85   +
86   ggplot2::geom_line()+
87   ggplot2::labs(x = "Número de iteración",
88                 y = "Pérdida esperada")
89 ggplot2::ggsave("fig_conv_log_reg.png")
```

Código C.1: Se muestra el código que implementa el algoritmo ACE de Overstall y Woods (2016) para replicar el ejemplo de regresión logística de Woods et al. (2017) del Capítulo 5

C.2. Diseño para modelo de regresión Poisson

El código para el modelo de regresión Poisson del Capítulo 5 fue programado en R 3.5.0 (R Core Team, 2018) y JAGS 4.2.0 (Plummer, 2003) y ejecutado en una computadora ASUS ROG GL552VX con sistema operativo Windows 10 64-bit, 8 GB de memoria RAM, y procesador Intel Core i7-6700HQ, con velocidad de reloj de 2.60GHz. El tiempo que tardó en correr fue de 459.45 segundos, es decir, 7 minutos, 39 segundos, y 45 centésimas de segundo.

```
1 ##
2 ptm <- proc.time()
3
4
5 # Semilla para reproducibilidad:
6 set.seed(19022013)
7
8
9 # Preámbulo ----
10 library(MASS)
11 library(acebayes)
12 library(tidyverse)
13 library(GGally)
14 library(rjags)
15 library(grid)
16 library(gridExtra)
17 ggplot2::theme_set(theme_bw())
18
19
20
21 # Parámetros iniciales ---- ### ### ### ####
22
23 # Tamaño de muestra ----
24 ## Seleccionado acorde al artículo citado en la tesis
25 n <- 6
26
27
28 # Diseño inicial ----
29 ## Éste se selecciona aleatoriamente. Es una matriz
30 ## de 6 x 5, donde cada elemento es un número
31 ## tomado de una distribución uniforme en (-1, 1)
32 start.d <- matrix(runif(5*n, -1, 1),
33                  ncol = 5, nrow = n,
34                  byrow = TRUE,
```

```
35         dimnames = list(as.character(1:n),
36                          c("x1", "x2", "x3", "x4", "x5"))
37     )
38 # Distribución inicial ----
39 ## Ésta es una función que regresa una matriz de B x 5,
40 ## donde las entradas se distribuyen (independientemente)
41 ## como se menciona en la tesis
42
43
44 for(alpha in c(0.5, 0.75)){
45
46
47     prior <- function(B){
48
49         # Inicializa como ceros
50         p <- matrix(rep(0, B*5), ncol = 5, nrow = B)
51
52         # Utiliza distribuciones definidas en la tesis
53         for(i in c(1, 3, 5)){
54             p[, i] <- runif(B, 1, 1+alpha)
55         }
56
57         for(i in c(2, 4)){
58             p[, i] <- runif(B, -1-alpha, -1)
59         }
60
61
62         return(p)
63     }
64
65
66 # Ahora modifica la distribución inicial
67 prior_2 <- function(B){
68
69     # Define matriz inicial
70     p <- matrix(2 * (1 + alpha) * (rbeta(5*B, 2, 2) - 0.5),
71                ncol = 5,
72                nrow = B)
73
74     return(p)
75 }
76
77
78 # Diseño óptimo ---- ### ### ### ####
79 ## Se obtiene mediante el algoritmo ACE de Overstall y Woods
80 ## con los parámetros especificados en la tesis
```

```

81  eta <- acebayes::aceglm(formula=~0+x1+x2+x3+x4+x5,
82                          start.d = start.d,
83                          family = poisson(link = "log"),
84                          prior = prior,
85                          method = "MC",
86                          B = c(20000, 1000),
87                          criterion = "D",
88                          Q = 20,
89                          lower = -1,
90                          upper = 1)
91
92  # Muestra el resumen del diseño óptimo
93  #eta
94
95  # Muestra el diseño óptimo
96  eta$phase2.d
97
98  # Compara gráficamente las variables
99  GGally::ggpairs(as.data.frame(eta$phase2.d))
100 ggplot2::ggsave(paste0("alpha_",
101                        as.character(round(100*alpha)),
102                        "_original.png"))
103
104
105  # Guarda el diseño en otra variable
106  ifelse(alpha == 0.5, eta_5_original <- eta, eta_75_original <-
      eta)
107
108
109  # Diseño óptimo con iniciales modificadas ---- ### ## ## ####
110  ## Se obtiene mediante el algoritmo ACE de Overstall y Woods
111  ## con los parámetros especificados en la tesis
112  eta <- acebayes::aceglm(formula=~0+x1+x2+x3+x4+x5,
113                          start.d = start.d,
114                          family = poisson(link = "log"),
115                          prior = prior_2,
116                          method = "MC",
117                          B = c(20000, 1000),
118                          criterion = "D",
119                          Q = 20,
120                          lower = -1,
121                          upper = 1)
122
123
124  # Muestra el resumen del diseño óptimo
125  #eta
126

```

```

127 # Muestra el diseño óptimo
128 eta$phase2.d
129
130 # Compara gráficamente las variables
131 GGally::ggpairs(as.data.frame(eta$phase2.d))
132 ggplot2::ggsave(paste0("alpha_",
133                     as.character(round(100*alpha)),
134                     "_modified.png"))
135
136
137
138 # Guarda el diseño en otra variable
139 ifelse(alpha == 0.5, eta_5_modified <- eta, eta_75_modified <-
      eta)
140
141 }
142
143
144 # Determina convergencia ---- ### ### ### ####
145
146 # Distribuciones iniciales originales
147 # alpha = 0.5
148 eta_5_original.phase <- -c(eta_5_original$phase1.trace,
149                          eta_5_original$phase2.trace)
150 m <- length(eta_5_original.phase)
151 eta_5_original.phase <- data.frame("Número" = 1:m,
152                                   "Utilidad_esperada" = eta_5_
153                                   original.phase)
154
155 # Genera y guarda gráfica
156 ggplot2::ggplot(eta_5_original.phase, aes(x = Número, y = Utilidad
157     _esperada))+
158   ggplot2::geom_line()+
159   ggplot2::labs(x = "Número de iteración",
160                 y = "Pérdida esperada")
161 ggplot2::ggsave("fig_conv_pois_reg_alfa5.png")
162
163 # alpha = 0.75
164 eta_75_original.phase <- -c(eta_75_original$phase1.trace,
165                          eta_75_original$phase2.trace)
166 m <- length(eta_75_original.phase)
167 eta_75_original.phase <- data.frame("Número" = 1:m,
168                                   "Utilidad_esperada" = eta_75_
169                                   original.phase)
170
171 # Genera y guarda gráfica

```

```

170 ggplot2::ggplot(eta_75_original.phase, aes(x = Número, y =
      Utilidad_esperada))+
171   ggplot2::geom_line()+
172   ggplot2::labs(x = "Número_de_iteración",
173                 y = "Pérdida_esperada")
174 ggplot2::ggsave("fig_conv_pois_reg_alfa75.png")
175
176
177
178 # Distribuciones iniciales modificadas
179 # alpha = 0.5
180 eta_5_modified.phase <- -c(eta_5_modified$phase1.trace,
181                            eta_5_modified$phase2.trace)
182 m <- length(eta_5_modified.phase)
183 eta_5_modified.phase <- data.frame("Número" = 1:m,
184                                   "Utilidad_esperada" = eta_5_
                                   modified.phase)
185
186 # Genera y guarda gráfica
187 ggplot2::ggplot(eta_5_modified.phase, aes(x = Número, y = Utilidad
      _esperada))+
188   ggplot2::geom_line()+
189   ggplot2::labs(x = "Número_de_iteración",
190                 y = "Pérdida_esperada")
191 ggplot2::ggsave("fig_conv_pois_reg_alfa5_modified.png")
192
193
194 # alpha = 0.75
195 eta_75_modified.phase <- -c(eta_75_modified$phase1.trace,
196                            eta_75_modified$phase2.trace)
197 m <- length(eta_75_modified.phase)
198 eta_75_modified.phase <- data.frame("Número" = 1:m,
199                                   "Utilidad_esperada" = eta_75_
                                   modified.phase)
200
201 # Genera y guarda gráfica
202 ggplot2::ggplot(eta_75_modified.phase, aes(x = Número, y =
      Utilidad_esperada))+
203   ggplot2::geom_line()+
204   ggplot2::labs(x = "Número_de_iteración",
205                 y = "Pérdida_esperada")
206 ggplot2::ggsave("fig_conv_pois_reg_alfa75_modified.png")
207
208
209
210 ### Análisis de la distribución posterior #### --- --- ---
211

```

```

212 # Ingresas de nuevo semilla inicial
213 set.seed(19022013)
214
215 # Simula betas
216 beta1 <- runif(1, 1, 1.5)
217 beta2 <- runif(1, -1.5, -1)
218 beta3 <- runif(1, 1, 1.5)
219 beta4 <- runif(1, -1.5, -1)
220 beta5 <- runif(1, 1, 1.5)
221 beta <- c(beta1, beta2, beta3, beta4, beta5)
222
223 # Modelo D-óptimo ###
224 # Simula datos
225 x_original <- eta_5_original$phase2.d
226 z_original <- x_original
227 for(i in 1:5){
228   z_original[ , i] <- beta[i] * x_original[ , i]
229 }
230 z_original <- rowSums(z_original)
231 l_original <- exp(z_original)
232 y <- NULL
233 for(i in 1:6){
234   y[i] <- rpois(1, lambda = l_original[i])
235 }
236
237
238 # Prepara datos para el modelo de JAGS
239 x1 <- x_original[ , 1]
240 x2 <- x_original[ , 2]
241 x3 <- x_original[ , 3]
242 x4 <- x_original[ , 4]
243 x5 <- x_original[ , 5]
244 N <- 6
245
246 # Ajusta modelo y obtén distribuciones posteriores con JAGS
247 cat("
248
249 #####model{
250 #####for(i in 1:N){
251 #####y[i]~ dpois(l[i])
252     log(l[i]) <- z[i]
253     z[i] <- b1*x1[i] + b2*x2[i] + b3*x3[i] + b4*x4[i] + b5*x5[i]
254   }
255     b1 ~ dunif(1, 1.5)
256     b2 ~ dunif(-1.5, -1)
257     b3 ~ dunif(1, 1.5)
258     b4 ~ dunif(-1.5, -1)

```

```

259     b5 ~ dunif(1, 1.5)
260   }
261   ", fill = TRUE, file = "Modelo_D.txt")
262
263 original <- jags.model("Modelo_D.txt",
264                       data = list('x1' = x1,
265                                   'x2' = x2,
266                                   'x3' = x3,
267                                   'x4' = x4,
268                                   'x5' = x5,
269                                   'y' = y,
270                                   'N' = N),
271                       n.chains = 1)
272
273 original <- coda.samples(original,
274                          c('b1', 'b2', 'b3', 'b4', 'b5'),
275                          10000)
276 # Guarda muestras
277 b1 <- original[[1]][, 1]
278 b2 <- original[[1]][, 2]
279 b3 <- original[[1]][, 3]
280 b4 <- original[[1]][, 4]
281 b5 <- original[[1]][, 5]
282
283
284 # Crea tibble con muestras de las distribuciones posteriores
285 posterior <- tibble(Diseño = rep("D-óptimo", 50000),
286                    Beta = rep(c("Beta1", "Beta2",
287                                "Beta3", "Beta4",
288                                "Beta5"), each = 10000),
289                    Valor = c(b1, b2, b3, b4, b5))
290
291
292 # Modelo SIL-óptimo ###
293 # Define diseño óptimo (tomado de Woods et al., 2017)
294 x_mod <- matrix(c(-0.5, rep(1, 5),
295                  -1, 0.56, rep(-1, 4),
296                  rep(1, 2), -0.31, rep(1, 3),
297                  rep(-1, 3), 0.33, rep(-1, 2),
298                  rep(1, 4), -0.38, 1),
299                 ncol = 5)
300
301 # Simula datos
302 z_mod <- x_mod
303 for(i in 1:5){
304   z_mod[, i] <- beta[i] * x_mod[, i]
305 }

```

```

306 z_mod <- rowSums(z_mod)
307 l_mod <- exp(z_mod)
308 y <- NULL
309 for(i in 1:6){
310   y[i] <- rpois(1, lambda = l_mod[i])
311 }
312
313
314 # Prepara datos para el modelo de JAGS
315 x1 <- x_mod[ , 1]
316 x2 <- x_mod[ , 2]
317 x3 <- x_mod[ , 3]
318 x4 <- x_mod[ , 4]
319 x5 <- x_mod[ , 5]
320 N <- 6
321
322 # Ajusta modelo y obtén distribuciones posteriores con JAGS
323 cat("
324
325     \\\model{
326     \\\for(i in 1:N){
327     \\\y[i] ~ dpois(l[i])
328         log(l[i]) <- z[i]
329         z[i] <- b1*x1[i] + b2*x2[i] + b3*x3[i] + b4*x4[i] + b5*x5[i]
330     }
331     b1 ~ dunif(1, 1.5)
332     b2 ~ dunif(-1.5, -1)
333     b3 ~ dunif(1, 1.5)
334     b4 ~ dunif(-1.5, -1)
335     b5 ~ dunif(1, 1.5)
336     }
337     ", fill = TRUE, file = "Modelo_SIL.txt")
338
339 mod <- jags.model("Modelo_SIL.txt",
340                 data = list('x1' = x1,
341                             'x2' = x2,
342                             'x3' = x3,
343                             'x4' = x4,
344                             'x5' = x5,
345                             'y' = y,
346                             'N' = N),
347                 n.chains = 1)
348
349 mod <- coda.samples(mod,
350                   c('b1', 'b2', 'b3', 'b4', 'b5'),
351                   10000)
352 # Guarda muestras

```

```

353 b1 <- mod[[1]][, 1]
354 b2 <- mod[[1]][, 2]
355 b3 <- mod[[1]][, 3]
356 b4 <- mod[[1]][, 4]
357 b5 <- mod[[1]][, 5]
358
359
360 # Crea tibble con muestras de las distribuciones posteriores
361 posterior_mod <- tibble(Diseño = rep("SIL-óptimo", 50000),
362                        Beta = rep(c("Beta1", "Beta2",
363                                    "Beta3", "Beta4",
364                                    "Beta5"), each = 10000),
365                        Valor = c(b1, b2, b3, b4, b5))
366
367
368 posterior <- bind_rows(posterior, posterior_mod)
369
370
371 # Grafica densidades marginales de
372 # las distribuciones posteriores
373 text_size <- 25
374
375 # Beta 1
376 beta1_graph <- posterior %>%
377   dplyr::filter(Beta == "Beta1") %>%
378   ggplot2::ggplot() +
379   ggplot2::geom_density(aes(x = Valor,
380                             fill = Diseño),
381                         alpha = 0.3) +
382   ggplot2::geom_vline(xintercept = beta1,
383                       linetype = "dashed") +
384   ggplot2::xlim(1, 1.5) +
385   ggplot2::labs(x = "",
386                y = "",
387                title = "Beta_1") +
388   ggplot2::theme(text = element_text(size = text_size))
389
390
391 # Beta 2
392 beta2_graph <- posterior %>%
393   dplyr::filter(Beta == "Beta2") %>%
394   ggplot2::ggplot() +
395   ggplot2::geom_density(aes(x = Valor,
396                             fill = Diseño),
397                         alpha = 0.3) +
398   ggplot2::geom_vline(xintercept = beta2,
399                       linetype = "dashed") +

```

```
400 ggplot2::xlim(-1.5, -1) +
401 ggplot2::labs(x = "",
402               y = "",
403               title = "Beta_2") +
404 ggplot2::theme(text = element_text(size = text_size))
405
406
407 # Beta 3
408 beta3_graph <- posterior %>%
409 dplyr::filter(Beta == "Beta3") %>%
410 ggplot2::ggplot() +
411 ggplot2::geom_density(aes(x = Valor,
412                           fill = Diseño),
413                       alpha = 0.3) +
414 ggplot2::geom_vline(xintercept = beta3,
415                    linetype = "dashed") +
416 ggplot2::xlim(1, 1.5) +
417 ggplot2::labs(x = "",
418               y = "",
419               title = "Beta_3") +
420 ggplot2::theme(text = element_text(size = text_size))
421
422
423 # Beta 4
424 beta4_graph <- posterior %>%
425 dplyr::filter(Beta == "Beta4") %>%
426 ggplot2::ggplot() +
427 ggplot2::geom_density(aes(x = Valor,
428                           fill = Diseño),
429                       alpha = 0.3) +
430 ggplot2::geom_vline(xintercept = beta4,
431                    linetype = "dashed") +
432 ggplot2::xlim(-1.5, -1) +
433 ggplot2::labs(x = "",
434               y = "",
435               title = "Beta_4") +
436 ggplot2::theme(text = element_text(size = text_size))
437
438
439 # Beta 5
440 beta5_graph <- posterior %>%
441 dplyr::filter(Beta == "Beta5") %>%
442 ggplot2::ggplot() +
443 ggplot2::geom_density(aes(x = Valor,
444                           fill = Diseño),
445                       alpha = 0.3) +
446 ggplot2::geom_vline(xintercept = beta5,
```

```

447         linetype = "dashed") +
448   ggplot2::xlim(1, 1.5) +
449   ggplot2::labs(x = "",
450                y = "",
451                title = "Beta_5") +
452   ggplot2::theme(text = element_text(size = text_size))
453
454
455
456
457 # Define función para crear grid.arrange
458 # con leyenda compartida (Wickham, 2016)
459 grid_arrange_shared_legend <- function(..., ncol = length(list
460   (...)),
461
462   nrow = 1,
463   position = c("bottom", "
464     right")) {
465
466   plots <- list(...)
467   position <- match.arg(position)
468   g <- ggplotGrob(plots[[1]] + theme(legend.position = position))$
469     grobs
470   legend <- g[[which(sapply(g, function(x) x$name) == "guide-box")
471     ]]
472   lheight <- sum(legend$height)
473   lwidth <- sum(legend$width)
474   gl <- lapply(plots, function(x) x + theme(legend.position="none"
475     ))
476   gl <- c(gl, ncol = ncol, nrow = nrow)
477
478   combined <- switch(position,
479     "bottom" = arrangeGrob(do.call(arrangeGrob,
480       gl),
481
482         legend,
483         ncol = 1,
484         heights = unit.c(unit
485           (1, "npc") -
486             lheight, lheight))
487       ,
488     "right" = arrangeGrob(do.call(arrangeGrob, gl
489       ),
490
491         legend,
492         ncol = 2,
493         widths = unit.c(unit(1,
494           "npc") - lwidth,
495             lwidth)))
481

```

```
482   grid.newpage()
483   grid.draw(combined)
484
485   # return gtable invisibly
486   invisible(combined)
487
488 }
489
490
491 # Genera grid de gráficas y guarda
492 graph_arrange <- grid.arrange_shared_legend(beta1_graph,
493                                             beta2_graph,
494                                             beta3_graph,
495                                             beta4_graph,
496                                             beta5_graph,
497                                             ncol=2, nrow=3)
498 ggsave(filename = "posterior_graphs.png",
499         plot = graph_arrange,
500         dpi = 300)
501
502
503
504 # Reporta el tiempo que tardó en correr el código
505 proc.time() - ptm
```

Código C.2: Se muestra el código que implementa el algoritmo ACE de Overstall y Woods (2016) para replicar el ejemplo de regresión Poisson de Woods et al. (2017) del Capítulo 5

Referencias

- Albert, J. et al. (2000). *Generalized Linear Models: A Bayesian Perspective* (1 ed.). Marcel Dekker Inc.
- Atkinson, A. C. y Woods, D. C. (2013). *Designs for Generalized Linear Models* (1 ed.). Handbook of Design and Analysis of Experiments. Chapman and Hall.
- Bernardo, J. M. (1979a). Expected Information as Expected Utility. *The Annals of Statistics*, 7(3), 686–690.
- Bernardo, J. M. (1979b). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2), 113–147.
- Bernardo, J. M. (1997). Statistical Inference as a Decision Problem: The Choice of Sample Size. *The Statistician*, 46(2), 151–153.
- Bernardo, J. M. y Smith, A. F. M. (2000). *Bayesian Theory* (3 ed.). Wiley Series in Probability and Statistics. John Wiley y Sons.
- Box, G. E. P. y Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses* (2 ed.). John Wiley y Sons.
- Box, G. E. P., Hunter, J. S., y Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2 ed.). Wiley Series in Probability and Statistics. John Wiley y Sons.

- Box, G. E. P. y Liu, P. Y. T. (1999). Statistics as a Catalyst to Learning by Scientific Method Part I—An Example. *Journal of Quality Technology*, 31(1), 1–15.
- Casella, G. y George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3), 167–174.
- Chaloner, K. y Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, 10(3), 273–304.
- Chib, S. y Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4), 327–335.
- Fisher, R. A. (1935). *The Design of Experiments* (8 ed.). Hafner Publishing Company.
- Gelfand, A. E. y Gosh, M. (2000). *Generalized Linear Models: A Bayesian View* (1 ed.). Marcel Dekker Inc.
- Gelfand, A. E. y Smith, A. F. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Geman, S. y Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721–741.
- Gutiérrez-Peña, E. (2016a). *Análisis Bayesiano de Modelos Jerárquicos Lineales*. Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México.
- Gutiérrez-Peña, E. (2016b). *Estadística Bayesiana*. Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México.
- Gutiérrez-Peña, E. (2016c). *Métodos Computacionales en la Inferencia Bayesiana*. Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México.

- Hitchcock, D. B. (2003). A History of the Metropolis-Hastings Algorithm. *The American Statistician*, 57(4), 254–257.
- Johnson, R. T. y Montgomery, D. C. (2009). Designing Experiments for Nonlinear Models — An Introduction. *Quality and Reliability Engineering International*, 26(8), 431–441.
- Khuri, A. I., Mukherjee, B., Sinha, B. K., y Ghosh, M. (2006). Design Issues for Generalized Linear Models: A Review. *Statistical Science*, 21(3), 376–399.
- Kolmogorov, A. N. (1954). *Foundations of the Theory of Probability* (2 ed.). Chelsea Pub. Co.
- Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS* (1 ed.). Academic Press.
- Kullback, S. y Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Statist.*, 22(1), 79–86.
- Lindley, D. V. (1987). *Bayesian Statistics. A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial Mathematics.
- Lindley, D. V. (1997). The choice of sample size. *The Statistician*, 46(2), 129–138.
- Lindley, D. V. (2006). *Understanding Uncertainty* (1 ed.). John Wiley y Sons.
- McCullagh, P. y Nelder, J. A. (1983). *Generalized Linear Models* (2 ed.). Chapman and Hall.
- Mendoza, M. y Gutiérrez-Peña, E. (2010). Decision Theory. *International Encyclopedia of Education*. Oxford: Elsevier, 7(1), 119–124.
- Mendoza, M. y Regueiro, P. (2011). *Estadística Bayesiana*. Instituto Tecnológico Autónomo de México.

- Montgomery, D. C. (2012). *Design and Analysis of Experiments* (8 ed.). Wiley.
- Montgomery, D. C., Peck, E. A., y Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5 ed.). Wiley Series in Probability and Statistics. Wiley.
- Myers, R. H., Montgomery, D. C., y Vining, G. G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences* (1 ed.). Wiley Series in probability and Statistics. John Wiley y Sons.
- Nelder, J. A. y Wedderburn, R. W. A. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 135(3), 370–384.
- Nieto, L. E. (2017). *Regresión avanzada (con enfoque Bayesiano)*. Instituto Tecnológico Autónomo de México.
- Overstall, A. M. y Woods, D. C. (2016). Bayesian Design of Experiments using Approximate Coordinate Exchange. *Technometrics In Press*, 1–13.
- Overstall, A. M., Woods, D. C., y Adamou, M. (2017). *acebayes: Optimal Bayesian Experimental Design using the ACE Algorithm*. R package version 1.4.1.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rasmussen, C. E. y Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning* (1 ed.). The MIT Press.
- Ross, S. M. (2012). *A First Course in Probability* (9 ed.). Pearson.
- Sacks, J., Welch, W. J., Mitchell, T. J., y Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, 4(4), 409–435.

- Savage, L. J. (1954). *The Foundations of Statistics* (1 ed.). Wiley Publications in Statistics. John Wiley y Sons.
- Schloerke, B. et al. (2017). *GGally: Extension to 'ggplot2'*. R package version 1.3.2.
- Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistics with S* (4 ed.). New York: Springer.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
- Woods, D. C., Lewis, S. M., Eccleston, J. A., y Russell, K. G. (2006). Designs for Generalized Linear Models With Several Variables and Model Uncertainty. *Technometrics*, *48*(2), 284–292.
- Woods, D. C., Overstall, A. M., Adamou, M., y Waite, T. W. (2017). Bayesian design of experiments for generalized linear models and dimensional analysis with industrial and scientific application. *Quality Engineering*, *29*(1), 91–103.