Índice general

Capítulo 1

Introducción

En luz de las nuevas y populares tendencias en el mundo de la estadística computacional, a veces llamada *aprendizaje de máquina o machine learning (ML)*, este trabajo, busca desarrollar y entender desde sus cimientos, un modelo aplicable a esta categoría. Este modelo, buscará hacer inferencia sobre una base de datos y *aprender* sobre los patrones subyacentes que estos puedan contener. Se busca revisar todos los aspectos de su construcción: tanto consideraciones teóricas e históricas hasta diferentes paradigmas de aprendizaje; así como su implementación y validación en una computadora. Todo esto, para dar un contexto sobre la, también llamda, *Inteligencia Artificial*, lo cual no es más que estadística computacional llevada al límite.

Este tipo de modelos, han resultado ser de enorme efectividad en ámbitos que van, desde la medicina hasta las finanzas. Sin embargo, en ocaciones, los métodos de ML son tratadas como *cajas negras*; se tienen datos que se alimentan a un modelo complejo y este arroja resultados. Aunque estos métodos son útiles, el tratamiento de los datos y el modelo en si, no se debe dejar de un lado, pues existen consideraciones técnicas y supuestos que se deben cumplir. Además, la interpretación, validación y análisis de los resultados, deben ser realizados por alguien que conozca, al menos de manera general, que está haciendo la computadora.

En particular, el modelo que se presenta a continuación, busca la predicción de variables binarias en un contexto de regresión bayesiana a través de un proyector aditivo con una transformación no lineal de los datos. Esta maraña de términos técnicos, se ira esclareciendo poco a poco conforme se construye el modelo. En el fondo, se busca clasificar cada observación i como: *éxito o fracaso*, *positivo o negativo*, *hombre o mujer* o cualquier otra respuesta binaria y_i, a través de información adicional $\mathbf{x}_i$ conocida como covariable. El problema radica en que es que está información adicional, puede contener un patrón complejo que es difícil de identificar a través métodos tradicionales. En la Figura 1.1, se tiene un ejemplo gráfico de este

tipo de clasificadores.

> [...], it is more common in machine learning to view the model as core, and how this is implemented is secondary. From this perspective, understanding how to translate a mathematical model into a piece of computer code is central.[1]

Usando esta cita como concepto rector del trabajo, además de una extensa discusión teórica, se hace énfasis en el desarrollo de un paquete en el software estadístico `R` para su implementación. Esto, en respuesta a que, pasar de la teoría a un código funcional, resultó ser más difícil de lo que se esperaba.

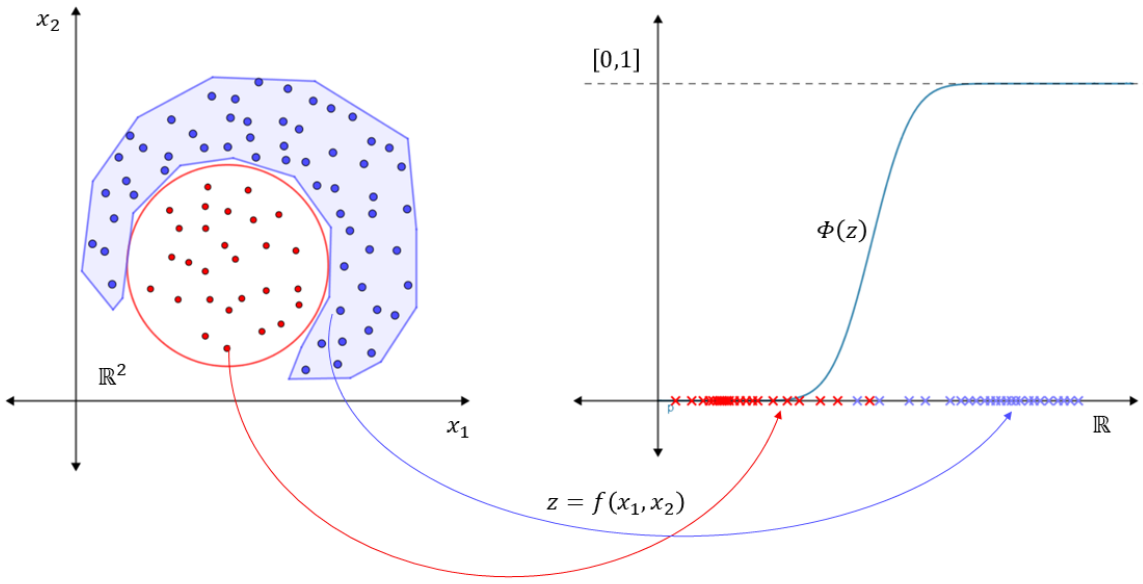

Figura 1.1: **Diagrama explicativo del modelo**. Se tienen observaciones del grupo azul y del grupo rojo con una clara separación no lineal en las covariables x_1 y x_2. El modelo busca *entrenar* una función f que logre separar lo mejor posible este espacio. Posteriormente, esta separación, induce una clasificación (0 y 1 correspondiendo a rojo y azul respectivamente) a través de la función de acumulación normal Φ, de ahí a que el modelo sea *probit*.

Es fundamental entender a fondo cada pedazo del modelo, por ello, en el Capítulo 2 se hace una exploración de su forma matemática más rigurosa. Dada su estructura, el modelo se puede estudiar de arriba hacia abajo, es decir, de la parte más general a la parte más profunda. Por lo tanto, primero se estudian los Modelos Lineales Generalizados (GLM), específicamente los modelos probit. Los GLM dan el salto de una regresión donde la respuesta y_i es real, a regresiones donde la respuesta, puede ser discreta o

1. (Barber 2012)

restringida a cierto dominio (MacCullagh y Nelder 1989). Los GLM, como su nombre lo indica, siguen siendo lineales, pero este proyector lineal, se puede flexibilizar un poco más usando las ideas de los Modelo Aditivo Generalizado (GAM) presentadas en (Hastie y Tibshirani 1986). Los GAM, buscan transformar a las covariables $\mathbf{x}_i$, previamente a la regresión, usando métodos no paramétricos. Este trabajo, toma esas ideas y las combina con las de (Denison, Mallick y Smith 1998), en el que se llevan un paso más allá la transformación para hacerla *tan flexible como sea necesaria*; todo bajo un paradigma de aprendizaje bayesiano. Esta transformación, corresponde a una serie de polinomios por partes de continuidad y grado arbitrarios, sujetos a ciertos nodos, lo cual representa la parte más profunda del modelo. La expansión que se presenta, resulta que conectan muchas disciplinas y ramas de las matemáticas que han sido de mucha utilidad no solo en el campo de la estadística. Al final del Capítulo 2, se verá que con estos principios se abre un mundo de posibilidades en cuanto a modelos y datos sobre los que se pueden aplicar.

Posteriormente en el Capitulo 3, se hace una breve introducción a la estadistica bayesiana, en particular al apredizaje bayesiano en el contexto de regresión lineal. Esto se debe a que la implementación algorítmica del modelo recae en una técnica fundamental de esta diciplina, el *Gibbs sampler*, usando las ideas de (Albert y Chib 1993). Con esta poderosa herramienta, se presenta los detalles y lógica del algoritmo. Además, se explica a grandes razgos como se hizo el desarrollo y de un paquete computacional de código abierto en el software `R` para el uso del modelo. El desarrollo de un paquete se detalla en el Apéndice B, y corresponde a que, no solo se simplifica la implementación, sino que se está fomentando el fácil uso del modelo y su validación a terceras personas que se puedan interesaran en el. Se hace notar, que al paquete se le añadió funcionalidad adicional para la visualización de ciertas partes del modelo bajo algunos supuestos facilitando su interpretación.[2]

Una vez que el modelo fue funcional y fácil de implementar, se probó y se validó contra una serie de bases de datos, tanto simulados como reales para probar su efectividad. En el Capítulo 4, se puede estudiar el modelo en una forma más pragmática, pues el uso del paquete lo facilita mucho. En particular, las bases de datos simuladas ejemplifican muy bien el modelo y muestran la flexibilidad de los polinomios por partes logrando encontrar fronteras de clasificación complejas y evidentemente no lineales. Uno de los ejemplos replica de forma fiel, la Figura 1.1.

Finalmente, en el Capítulo 5, se verán las consideraciones finales y limitaciones del modelo. Sin embargo, se abre una discusión a posibles extensiones para mejorarlo. Posteriormente, se da un vistazo a modelos más modernos los cuales han sido capaces de proezas computacionales que se creían imposibles hace algunas

2. El paquete se puede descargar libremente de: `https://github.com/PaoloLuciano/bpwpm`

décadas. Se verá, sin embargo, que muchos de los modelos más avanzados y usados hoy en día, son generalizaciones de los modelos tradicionales presentados en este trabajo. Si estos modelos, se comienzan a anidar unos dentro de los otros se logra extender el *aprendizaje* más allá de datos binarios y lograr clasificaciones de imágenes, sonidos y datos poco ortodoxos para la estadística.

Capítulo 2

Modelo en su forma matemática

Como base fundamental de este trabajo, a continuación, se expondrá a detalle el modelo. El objetivo, es construir un clasificador binario flexible con buena fuerza predictiva. La notación se irá explicando conforme aparece pero existe un compendio en el Apéndice C. En general, se trata de respetar la notación que usan en los libros (Hastie, Tibshirani y Friedman 2008) y (James y col. 2013)

Se supone la siguiente estructura en los datos:

- $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ con n el tamaño de la muestra.
- $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ variables de respuesta binarias o *output*.
- $\mathbf{x}_i \in \mathcal{X}^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ covariables, regresores o *input*.
- $d \in \mathbb{N}$ dimensionalidad de las covariables.

El modelo en si, se presenta a continuación de forma general para cualquier pareja de datos $(y, \mathbf{x})$:

$$y \,|\, z \;\sim \mathrm{Be}(y \,|\Phi(z)) \tag{2.1}$$

$$z \,|\, x \;\sim \mathrm{N}(z \,|f(\mathbf{x}), 1) \tag{2.2}$$

$$f(\mathbf{x}) \approx \sum_{j=0}^{d} \beta_j f_j(x_j) \tag{2.3}$$

$$f_j(x_j) \approx \sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_j, \mathcal{P}_j) \quad \forall j = 0, 1, \ldots, d \tag{2.4}$$

En las expresiones (2.1) y (2.3), dejando de un lado ecuación (2.2), se tiene una versión ligeramente modificada de un GLM (Sec. 2.1). Esto, pues la variable de respuesta y es binaria modelada con una distribución Bernoulli. Además, (2.3) es una función de proyección lineal como las que se usan el los

modelos tradicionales. En el contexto de un modelo probit, esta función f, busca separar el espacio d-dimensional de covariables $\mathcal{X}^d$ en regiones identificables en una sola dimensión $\mathbb{R}$. Esta función de proyección, asume que la dependencia entre covariables se puede modelar como la suma ponderada de los componentes f_j (Sec. 2.2). Para poder hacer la liga entre ambas ecuaciones, se requiere de la incorporación de una variable latente z, vista en la ecuación (2.2), esta variable es meramente estructural y será modelada a través de una distribución normal, lo cual lleva a tener un modelo probit. Finalmente (2.4) hace una transformación no lineal de cada dimensión j y trata de encontrar las tendencias individuales de cada una de las covariables. Esto se logra, haciendo un suavizamiento por medio de polinomios por partes que dependen de 3 objetos: una partición del intervalo $\mathcal{P}_j$, un vector de pesos w_j y parámetros que captura la N^* especificando la forma y grado de los polinomios. La forma funcional de Ψ es compleja y relativamente arbitraria dependiendo de la selección de la base, por lo tanto, no se especifican aún y se deja para la Sección 2.3. Se hace notar que el componente bayesiano se explora hasta el Capítulo 3 pues va estrechamente ligado con su implementación. En la Figura 2.1 se hace una representación visual del modelo para su mejor comprensión.

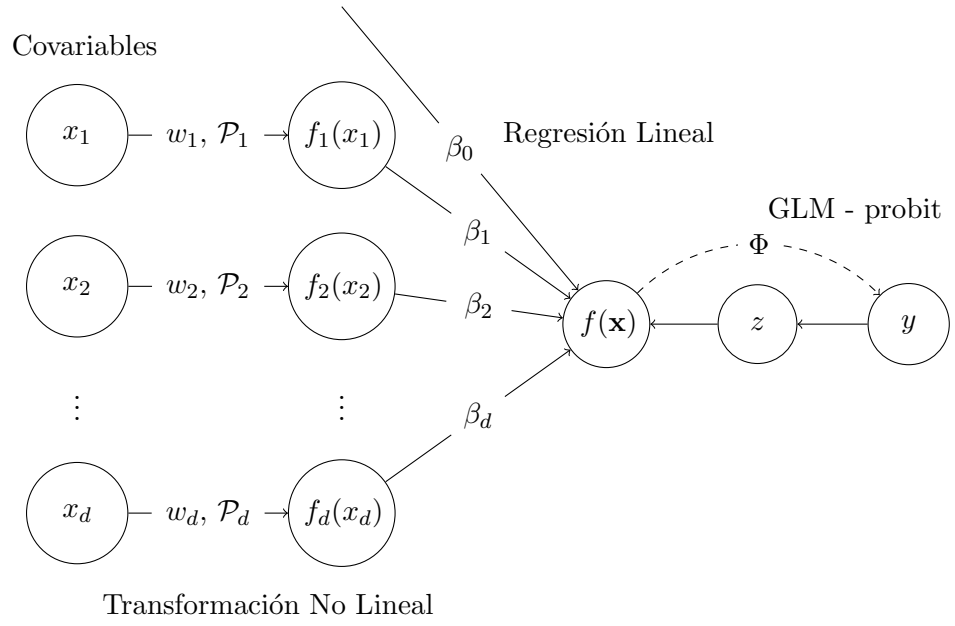

Figura 2.1: **Diagrama del modelo.** Se hace una transformación no lineal de las covariables x_j a través de los parámetros w_j y P_j. Con los datos transformados f_j, se lleva a cabo un modelo probit con función liga Ψ para lograr la clasificación binaria en y.

Antes de continuar, vale la pena recordar que:

> *All models are wrong but some are useful*[1]

Escoger un modelo que explique perfectamente los datos o que logre predecir todo sería una tarea inútil.

1. (Box 1979)

Sin embargo, no significa que no se pueda intentar discernir un patrón y es justamente lo que se busca con la construcción de este modelo. Además de entender a profundidad un modelo que sirve como base para modelos que se están usando en el mundo de la inteligencia artificial. En particular, este modelo tiene la ventaja que es flexible y, al menos en teoría, debería de servir para representar una gran cantidad de datos.

2.1. Modelos Lineales Generalizados (GLM)

Los modelos lineales generalizados, (Sundberg 2016) y (MacCullagh y Nelder 1989), surgen como una generalización del modelo lineal ordinario $y = \beta^t x + \epsilon$ donde $y \in \mathbb{R}$. En esta generalización, se busca darle otros rangos a y pues se tienen casos donde está restringida a un subconjunto de los reales como lo es el caso binario. Sin embargo, este cambio vuelve el modelo más complejo y lleva a técnicas diferentes en la estimación de los parámetros β. Además, se pierde algo de la interpretabiliad del modelo[2]. Sin embargo, han resultado ser realmente útiles.

Los GLM se especifican (de manera muy general) de la siguiente manera:

$$\begin{aligned} y &\sim F(\theta(x)) \\ z &= \beta^t x \\ \theta &= g^{-1}(z) \end{aligned} \tag{2.5}$$

con los siguientes tres elementos:

1. F: **Tipo de distribución** de la familia exponencial que describa el dominio de las respuestas y. Por ejemplo: Bernoulli si y es binaria, Poisson si $y \in \mathbb{Z}^+$ o una distribución Gamma si $y \in \mathbb{R}^+$

2. z: **Proyector lineal** que explique (linealmente) la variabilidad sistemática de tus datos. En el modelo tradicional $\dim(\beta) = d < n$.

3. g: **Función liga** que una la media (o los parámetros canónicos) θ de mi distribución con el proyector lineal. Es decir: $\theta(x) = \mathbb{E}[y|x] = g^{-1}(\beta^t x)$.

   Como ejemplos clásicos se tiene la función $\text{logit}(p) = \ln(p/(1-p))$ o la $\text{probit(p)} = \Phi^{-1}(p)$, donde $p = \mathbb{E}[y|x]$ y $\Phi(\cdot)$ la función de acumulación de una distribución normal estándar. En la Figura 2.2 se puede ver una representación gráfica para su mejor comprensión.

2. Dependiendo de la especificación, su interpretación puede ser complicada. Por ejemplo, cuando se tiene un modelo logit tradicional, se logra expresar el logaritmo de la proporción de probabilidades (*Log-Odds-Ratio*) como una combinación lineal de las covariables. $\ln(\pi_i/\pi_0) = \beta^t x$

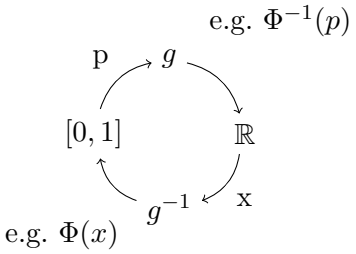

Figura 2.2: **Esquema de función liga** g

Para este trabajo, se busca construir un clasificador binario por lo que $y \in \{0, 1\}$, por lo cual, es natural modelar y como una distribución Bernoulli. Notese que si $Y \sim \text{Be}(y|p)$ se tienen las siguientes propiedades:

$$\mathbb{E}[Y] = p = P(y = 1)$$
$$\mathbb{V}[Y] = p(1 - p)$$

Por lo que solo se tiene un parámetro p y la varianza queda determinada automáticamente. Además, esta especificación deja como opción para las función liga, a las inversas de las funciones *sigmoidales* $s(x)$. Las funciones sigmoidales, son funciones $s : \mathbb{R} \to (0, 1)$, estrictamente monótonas y por ende, biyectivas. Algunos ejemplos son las ya mencionadas logit, probit y la curva de Gompertz[3]. Estas funciones cumplen un papel de activación, es decir, una vez que se rebase cierto umbral, crecen rápidamente y toman valores más cercanos a uno, *activando* asi la probabilidad de que y sea un éxito. [4]. Esto las hace perfectas herramientas para ligar el proyector lineal $z \in \mathbb{R}$ con una probabilidad $p \in [0, 1]$.

2.1.1. Uso de la Variable Latente

Ahora, para entender el papel que juega z, se necesita entender que es posible estructurar estos modelos como *modelos de variable latente* (Albert y Chib 1993). Bajo esta formulación, se asume que la relación entre y y x no es directa, sin embargo, existe una variable no observada z estructural que ayuda a discernir un vínculo entre ellas. En la Figura 2.3 se tiene una representación gráfica de esto.

Tradicionalmente, la normalidad en z es derivada de asumir normalidad en los errores; es decir, dada la regresión lineal $z = \beta^t x + e$ se asume (y se debe verificar) que $e \sim N(0, \sigma^2)$. Lo cual lleva a $z \sim N(\beta^t x, \sigma^2)$.

3. Para no caer en redundancia de notación para este trabajo se tiene a partir de ahora: $s(x) = g^{-1}(x) = \Phi(x)$

4. En un contexto de redes neuronales, lo que se activa es la neurona y recientemente, se usa la función *ReLu* $(x) :=$ máx $\{0, x\}$

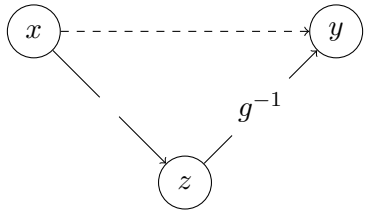

Figura 2.3: **Modelo de variable latente**

Además este supuesto facilita la estructura de los modelos y el algoritmo de ajuste. Bajo un paradigma frequentista, la estimación de los parámetros β se reduce a encontrar los estimadores de mínimos cuadrados. Sin embargo, bajo el paradigma bayesiano, dentro de un modelo probit como el de este trabajo, se adopta la normalidad en z pues en (Albert y Chib 1993), se sugiere un algoritmo *Gibbs sampler* con distribuciones truncadas de la normal para encontrar β.

La función liga probit se escoge como consecuencia de la normalidad en z, o viceversa, dependiendo de como se quiera ver. Esta función $\Phi^{-1}(p)$ es la inversa de la función de acumulación normal estándar:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

la cual no tiene forma cerrada. Sin embargo, tiene justamente las propiedades que se necesitan pues es claramente sigmoide. Esta función, cumple el propósito de modelar y cuantificar la incertidumbre pues está transformando la variable real z en una probabilidad p con su característica forma de "s". Se hace notar que se podría haber usado una función más flexible o que incluso se podría dejar la función como una parte del modelo a estimar. Sin embargo, al adoptar el algoritmo antes citado, se requiere esta especificación.[5]. La parte flexible de este modelo se encuentra en el proyector lineal.

Habiendo definido la función liga, distinguir entre si $y = 1$ ó 0, éxito o fracaso respectivamente, se reduce a distinguir en que área del espacio de covariables $\mathcal{X}^d$ se encuentra el dato. Esto se debe a que $y = 1$ cuando $\Phi(z) > 1/2$ que sucede *si y solo si* $z > 0$ lo cual, depende en gran media de su media; en este caso la función de proyección $f(\mathbf{x})$. Si esta función es muy positiva en alguna región, implicará que el modelo tiene mucha evidencia para confiar que, al menos en esa área, la respuesta y es un éxito. El razonamiento, funciona de forma análoga para los casos donde $y = 0$, claramente, para esas regiones, se busca que $f(\mathbf{x})$ sea negativa. Por lo tanto, es fundamental para el modelo que se realice una correcta estimación de los parámetros de la función de proyección. Nótese además, que z le agrega cierta *estocasticidad* al modelo.

5. En modelos multinomiales bayesianos, tomar esta decisión estructural lleva a que inclusive, se puede asumir una estructura de interdependencia en los errores aleatorios. **e** $N_k(0, \Sigma)$ con Σ una matriz de correlaciones

Bajo la suposición que existe una pareja $(y_i, \mathbf{x}_i)$ tal que $f(\mathbf{x}_i) = 0$; alrededor de una vecindad de este punto, no se tendría evidencia para clasificar a y_i como un éxito o como un fracaso; sería mejor un volado.

Otro factor importante a considerar, es que el modelo asume que la varianza de z es constante, específicamente $\sigma^2 = 1$. Dado que la escala de z es completamente arbitraria pues es una variable auxiliar, se puede *restringir* z al rango que se desee. El método de simulación para z usando una distribución normal truncada se simplifica ligeramente usando esta varianza unitaria. Se verá en los resultados, sin embargo, que dada la naturaleza global de los polinomios que se usan, la escala de z, o al menos la estimación de su media $\hat{f}(\mathbf{x})$, puede variar mucho dependiendo de los datos, mas esto no representa un problema. Pues, en la practica, al usar el algoritmo de Albert y Chibb z sirve mucho más, para hacer la ligadura de y hacia f y no viceversa. En z se codifica, mediante una normal truncada, los casos de éxito y de fracasos de y; posteriormente, se estima el vector β para la función f. Por ello, en la Figura (2.1), se representan las flechas de y a f por medio de z y solidas, en contraposición con la flecha punteada que va, directamente de f a y y pasa por la función Φ. Los detalles y su justificación probabilista, se tocan en detalle en el Capítulo 3.

Es importante mencionar, que el *corte* que se hace en $z = 0$ para la clasificación, es resultado de hacer una clasificación binaria. En modelos multinomiales también se debe tomar en cuenta los intervalos en $\mathbb{R}$ para los que la observación se clasificaría en alguna de las posibles clases y por ende, estimar los umbrales o usar una función diferente a las sigmoides. Este hecho, lleva a la realización de que z y su media f son *ajenas más no independientes*. Esto quiere decir que la parametrización de z como una normal $N(\mu, 1)$ es equivalente a la parametrización $N(0, 1)$. Este hecho se hará más claro cuando se hable del papel de β_0 en la función de proyección.

Finalmente, si se quisiera ver la relación de x en y directamente, se puede lograr usando el teorema de la probabilidad total. Se puede calcular (al menos de forma teórica), la distribución marginal de y dado $\mathbf{x}$ sumando sobre z:

$$\begin{aligned} P(y|x) &= \int_{-\infty}^{\infty} P(y|z)P(z|x)\,dz \\ &= \int_{-\infty}^{\infty} p(y;\Phi(z))p(z;f(\mathbf{x}),\sigma^2)\,dz \\ &= \int_{-\infty}^{\infty} \Phi(z)^y(1-\Phi(z))^{1-y}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-f(\mathbf{x}))^2}\,dz \end{aligned}$$

Sin embargo, está claro que esta derivación no lleva a ningún resultado analítico cerrado pues la relación es bastante más compleja como para resultar en una distribución tradicional; si lo hiciera, el propósito de

la z se perdería.

Recapitulando, mediante la función liga Φ se une la media p, la probabilidad de éxito o fracaso, de la respuesta y con los datos $\mathbf{x}$. Esto se logra, a través de una variable auxiliar z cuya media $f(\mathbf{x})$ es una función de proyección lineal.

$$P(y=1) = p(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g^{-1}(f(\mathbf{x})) = \Phi(f(\mathbf{x})) \tag{2.6}$$

2.2. Función de proyección f

En la sección anteriór, se vió que tradicionalmente, se asumía z como una combinación lineal de los parámetros β y las covariables x (segunda ecuación de (2.5)). Pero, como se explica en la página 6 de (James y col. 2013), conforme avanzarda on los métodos y el poder computacional disponible se fueron desarrollando técnicas cada vez más poderosas que permitieron romper la linealidad. En 1986, Hastie y Tibshirani introducen los modelos aditivos generalizados (GAM), una clase de modelos donde se rompe la linealidad en las covariables, flexibilizando aún más el modelo.

Se hace notar que esta generalización es sutil pues el modelo aún conserva una parte lineal *a lo largo*[6]. Se ve en la ecuación (2.3), que el modelo aún es lineal en las β_j y en las f_j. Donde se pierde la linealidad es *hacia abajo*,[7] pues cada f_j en realidad es la transformación no lineal de la covariables x_j. Además, las dimensiones j's se asumen independientes entre si, pues se busca que corresponden a diferentes *características* o variables con las que se planea modelar la variable de respuesta.

En este modelo, el proyector f de la ecuación (2.3), no hace otra cosa más que ir colapsando dimensiones, en particular: $\mathcal{X}^d \to \mathbb{R}$ que posteriormente se colapsa por medio de Φ en $[0,1]$. Es por esto que se le llama función de proyección, pues *proyecta* el espacio $\mathcal{X}^d$ en $\mathbb{R}$. Sin embargo, la forma en la que lo haga, debe de ser lo más precisa posible pues el modelo recae en que este colapso detecte los patrones correctos en las covariables que llevan a la correcta identificación de y. Por lo tanto, f como función de proyección es el corazón del modelo, por lo que su correcto entrenamiento es fundamental. La idea, recapitulando, es que f separe el espacio de covariables para que sea positiva en las regiones donde se tengan éxitos y negativa

6. Con esta frase se hace alusión a que, en una tabla de datos de tamaño $n \times d$, siendo cada fila una observación y cada columna una variable, el *largo* se piensa como la segunda dimensión de tamaño d. Por lo tanto, al usar esta expresión se busca considerar todas las variables. A lo largo de este trabajo y en su implementación, se usa el subíndice j para denotar esta idea.

7. De forma análoga, esta idea, hace alusión a las diferentes observaciones. Denotado por el subíndice i, el cual no se ha usado para simplificar la notación.

donde se tengan fracasos; para ello, es fundamental entender los GAM.

2.2.1. Modelos Aditivos Generalizados (GAM)

Un GAM como se introduce en (Hastie y Tibshirani 1986) remplaza la forma lineal $\sum_1^d \beta_j x_j = \beta^t x$ con una suma de funciones *suaves* $\sum_j^d f_j(x_j)$. Estas funciones no tienen una forma cerrada y son no especificadas, es decir, no hay tienen una forma funcional concreta y representable algebraicamente. Donde recae la fuerza de estos modelos, es que se estiman usando técnicas de suavizamiento no paramétricas[8] como lo sería un suavizamiento loess. En estos modelos, se asume que por más grande que sea $\mathcal{X}^d$, la relación que existe entre cada una de las dimensiones j, se puede explicar de manera aditiva, es por ello que cada función f_j tiene como argumento exclusivamente de x_j. Esta especificación fue revolucionaria pues no solo regresa interpretabilidad al modelo, sino que simplifica la estimación usando técnicas prácticamente automáticas con el algoritmo de *backfitting*. La idea fundamental de este algoritmo será de vital importancia para el ajuste de el modelo. Los principal ventaja de los GAM, es que logran descubrir efectos no lineales en las covariables, justamente lo que se busca.

La f de este trabajo, es una versión versión modificada de un GAM con tres cambios fundamentales. La primera modificación es que se ponderando cada f_j por un parámetro β_j, esto es, para suavizar aún más cada dimensión y captar el patrón general y no tanto los componentes individuales de cada x_j. Al entender que cada f_j es una transformación no-lineal de x_j (como lo sería una transformación logarítmica o una transformación Box-Cox) se le regresa cierta interpretabilidad al parámetro β_j como el efecto que tiene la dimensión i en particular para el modelo. Se deja espacio para un término independiente β_0 pues este ayuda a ajustar la escala en la estimación de los parámetros. La inclusión de este parámetro es fundamental para la correcta especificación del modelo pues ayuda a dar un *sesgo o nivel* base contra el cual comparar la suma y escalar la f para que sea compatible con el umbral de corte en 0 haciendo equivalente la parametrización de z con una normal estándar. Por convención $f_0(\cdot) = 1$ por lo que se puede re-expresar la ecuación (2.3) como:

$$f(\mathbf{x}) \approx \beta_0 + \sum_{i=1}^{d} \beta_i f_i(x_i) = \beta^t \mathbf{f}(\mathbf{x})$$

8. Las técnicas no paraméticas están fuera del alcance de este trabajo. Sin embargo, vale la pena una mención especial por su funcionalidad, practicidad y forma intuitiva, además del sinfín de aplicaciones que tienen. Una guía comprensiva de estas se encuentra en el libro (Wasserman 2007).

donde usando notación vectorial $\beta \in \mathbb{R}^{d+1}$ y $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{d+1}$:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_0 \\ f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_d(x_d) \end{bmatrix} = \begin{bmatrix} 1 \\ w_1^t \Psi_1(x_1, \mathcal{P}_1) \\ w_2^t \Psi_2(x_2, \mathcal{P}_2) \\ \vdots \\ w_d^t \Psi_d(x_d, \mathcal{P}_d) \end{bmatrix} \tag{2.7}$$

La segunda modificación, se ve en la expresión anterior (2.7). Aunque es muy práctico manejar las f_j's como indeterminadas y estimarlas con procedimientos de suavizamiento no paramétricos, también se puede optar por la vía en la que se especifica su forma funcional, en este caso $w_j^t \Psi_j(\cdot)$. No por ello, se le quita flexibilidad al procedimiento; esto se verá con todo detalle en la siguiente sección 2.3, donde se trata de adapart el procedimiento de (Denison, Mallick y Smith 1998). Esta modificación, obedece a que para ciertas aplicaciones, sirve hacer el modelo paramétrico en donde cada f_j se modela en su expansión de bases (Vease Capitulo 9.1 y Ejemplo 5.2.2 de (Hastie, Tibshirani y Friedman 2008)).

La tercera modificación, es que en la practica, se tiene una aproximación en vez de una igualdad para f. El simple hecho de asumir que existe una f que puede separar el espacio en regiones positivas y negativas es uno de los supuestos más fuertes del modelo, esto responde a que el *error aleatorio*, sistemático de los datos, está siendo capturado por una aproximación a la f real, dentro de cada una de las f_j's. En el apéndice A se explora el porqué de esta aproximación usando técnicas de análisis más avanzadas y se revisan cuestiones de convergencia del modelo.

En la peculiaridad de que $d = 2$, se podrá visualizar $f(\mathbf{x})$ en $\mathbb{R}^3$ como una serie de picos y valles donde será positiva en caso de ser éxito y negativa en caso contrario.

Modelos en bases de funciones lineales

Finalmente, se aclara que este modelo podría ser confundido con un modelo en bases de funciones lineales[9] como los presentados en Capitulo 3 de (Bishop 2006):

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{d} \beta_j f_j(\mathbf{x}) \tag{2.8}$$

9. *Linear basis function models* por falta de una mejor tradición.

claramente con una forma funcional simular. La diferencia radica en que cada f_j es función de todas las covariables en lugar de solo la que le corresponde. A estas funciones se les conocen como funciones base, sobre las que se hará una exposición más a detalle en la siguiente sección. La f una vez más es lineales en β, pero no-lineales para $\mathbf{x}$. Estos modelos en si, también son de gran utilidad pero la forma de f_j es de naturaleza global, algunos ejemplos son:

- **Bases gaussianas:**
$$f_j(\mathbf{x}) = \exp\left\{-\frac{(\mathbf{x}-\mu_j)^2}{2s^2}\right\}$$
- **Funciones sigmoidales:**
$$f_i(\mathbf{x}) = \sigma\left(\frac{\mathbf{x}-\mu_j}{s}\right)$$

Sin embargo, estos son un grupo de modelos completamente diferente cuyas aplicaciones usualmente son en estimación de curvas y no tanto inferencia como lo busca este trabajo.

2.3. Funciones f_j

Finalmente se trata la parte más profunda del modelo, las funciones f_j que, como se mencionó anteriormente, son transformaciones no lineales de cada componente x_j que buscan suavizar la nube de datos, para posteriormente sumarlas entre si y dar una medida f que resuma toda la información en un número real. Como se menciona en la introducción de (Härdle y col. 2004), el suavizamiento de los datos es central en la estadística inferencial. La idea es extraer la señal entre el ruido y para ello, se intenta estimar y modelar la estructura subyacente. Este suavizamiento, se llevará a cabo usando una expansión en bases funcionales, particularmente en polinomios por partes. Toda la siguiente sección se concentra en darle formas funcionales a Ψ y a explicar el papel de los pesos w. Se usa como referencia en la exposición, las primeras dos secciones de el captiulo 5 de (Hastie, Tibshirani y Friedman 2008).

Expansión en bases funcionales

Saliendo por un momento del domino de la estadística, se definen las expansiones en bases de funciones. Sin entrar mucho en los detalles técnicos, dado un espacio funcional, se puede representar cualquiera de sus elementos, en este caso una funciones arbitrarias h, como la combinación lineal de los elementos de la base Ψ (también funciones) y constantes w. En particular (y dados los objetivos de la exposición) se

considera el espacio funcional que mapea $\mathbb{R}^d$ a $\mathbb{R}$, quedando entonces la expansión:

$$h(\mathbf{x}) = \sum_{l=1}^{N} w_l \Psi_l(\mathbf{x}) = w^t \Psi(\mathbf{x}) \tag{2.9}$$

con el vector de funciones base $\Psi(\mathbf{x})^t = (\Psi_1(\mathbf{x}), \ldots, \Psi_N(\mathbf{x}))^t$, donde cada elemento Ψ_l es también una función con el mismo mapeado que h, $w^t = (w_1, \ldots, w_N)^t$ un vector de coeficientes constantes y N un entero mayor o igual a la dimensión del espacio funcional que se maneja[10].

Regresando a las regresiones en el mundo de la estadística. Se busca representar la media condicional de la respuesta y por una función que depende de los datos: $h(\mathbf{x}) = \mathbb{E}[y\,|\mathbf{x}]$. Se puede pensar, que esta h también puede ser expresada como su expansión en bases funcionales.[11] La idea, es que se remplace (o se aumente) la cantidad de covariables $\mathbf{x}$ con transformaciones de estas, capturadas en el vector $\Psi(\mathbf{x})$. Por ejemplo:

- $\Psi_l(\mathbf{x}) = x_j \quad \forall\, l = 1, \ldots, d$, donde se recupera un GLM tradicional.
- $\Psi_l(\mathbf{x}) = \ln x_j$ ó $x_j^{1/2}$ donde se tienen transformaciones no lineales en cada una (o algunas) de las covariables.
- $\Psi_l(\mathbf{x}) = ||\mathbf{x}||$ una transformación lineal de todas las covariables.[12]
- $\Psi_l(\mathbf{x}) = x_j^2$ donde se tiene una expansión en bases polinómicas.
- $\Psi_l(\mathbf{x}) = x_j x_k$ donde se incluyen términos de interacción.

Esta representación, engloba muchos de los modelos y transformaciones posibles en el mundo de las regresiones, uniendo temas de análisis funcional con estadística aplicada. Además de que en general, han resultado ser de gran utilidad en la practica. Se hace notar, que el último ejemplo, rompe con la aditividad inherente de las combinaciones lineales, demostrando que esta generalización, no está restringida a ser completamente aditiva.

Dependiendo del tipo de datos, puede ser conveniente usar forma sobre la otra pues existen muchas más posibles expansiones. Sin embargo, sobre todo cuando se tiene poca o ninguna experiencia con los datos, se busca una representación más flexible (por no decirla ingenua) de estos. El método más común, es tomar una familia grande de funciones que logre representar una gran variedad de patrones. Una de estas

10. Dependiendo de el espacio funcional, en ocasiones $N = \infty$, este tema se discute más a fondo en A
11. Un supuesto fuerte pero necesario en ocasiones.
12. Como se vio en la ecuación (2.8)

familias, es la de los polinomios por partes usadas en este modelo. Una desventaja de estos métodos, sin embargo, es que al contar con una cantidad muy grande de funciones base y por ende parámetros, se requiere controlar la complejidad del modelo para evitar el *sobreajuste*. Algunos de los métodos más comunes para lograrlo son los siguientes:

- **Métodos de restricción**: donde se selecciona un conjunto finito de funciones base y su tipo, limitando así las posibles expansiones. Los modelos aditivos como los usados en este trabajo, son un ejemplo perfecto de este tipo.
- **Métodos de selección de variables**: como lo son los modelos CART y MARS, donde se explora de forma iterativa las funciones base y se incluyen aquellas que contribuyan a la regresión de forma significativa.
- **Métodos de regularización**: donde se busca controlar la magnitud los coeficientes, buscando que la mayoría de ellos sean cero, como los son los modelos *Ridge* y *LASSO*.

Consideraciones para la expansión en bases de este trabajo

Para simplificar un poco la exposición y reducir la notación, se asume por lo pronto que $d = 1$, por lo tanto $\mathbf{x} = x$ y se puede pensar únicamente en funciones que mapean reales a reales. Esto permite librar el subíndice j para indicar componente del vector $\mathbf{x}$ y usarlo para otros fines.

Para el modelo de este trabajo, se aplicaron las ideas de (Denison, Mallick y Smith 1998) a los modelos GLM presentados con anterioridad. Los autores presentan un método revolucionario, automático y bayesiano, que permite estimar con un alto grado de precisión relaciones funcionales entre la variable de respuesta y y sus covariable $x \in \mathbb{R}$. En el trabajo original, se buscaba ajustar una curva tal que $y = h(x)$. El modelo en su forma estadística se plantea para un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$:

$$y_i = h(x_i) + e_i \quad i = 1, \ldots, n \tag{2.10}$$

con las e_i errores aleatorios de media cero. Este método, combina los procedimientos paramétricos y no paramétricos desarrollados antes para hacer más robusto el algoritmo de (Hastie y Tibshirani 1986). La idea, es ajustar un *polinomio por partes* muy flexible. Estos polinomios, se componen de partes de menor orden entre *nodos* adyacentes. Una de las muchas genialidad del su trabajo es que estos nodos, tradicionalmente fijos, se vuelven parámetros a estimar, usando un paradigma bayesiano. Y no solo eso, sino que permiten *aumentar o disminuir el número de nodos* desarrollando un algoritmo Gibbs sampler trans-dimensional. Esta generalización, logra estimaciones tan robustas, que logran aproximar funciones

continuas *casi en todas partes*, como lo son la función Doppler, funciones por bloques y funciones con picos pronunciados.

2.3.1. Polinomios por partes y splines

Antes de llegar a estos polinomios tan flexibles, se busca entender que son los polinomios por partes simplificando (bastante) el trabajo de (Wahba 1990). Sea $x \in [a, b] \subseteq \mathbb{R}$, se busca separar $[a, b]$ en J intervalos. Por lo tanto, se construye una partición correspondiente $\mathcal{P} = \{\tau_1, \tau_2, \ldots, \tau_{J-1}\}$ tal que $a \leq \tau_1 < \ldots < \tau_{J-1} \leq b$. Estas τ's son llamadas *nodos*. Se hace notar, que se puede incluir o no la frontera dependiendo de la especificación[13]. Con estos nodos seleccionados, se puede hacer una representación de h en su expansión de bases como en la ecuación (2.9), donde cada Ψ_j será una función que depende, tanto de la partición como de la variable x. Por ejemplo, se puede pensar en un caso sencillo donde se tiene que $J = 3$ y a cada uno subintervalos les corresponde una función Ψ_j donde $j = 1, \ldots, 3$. Simplificando aún más, se hacen que estas Ψ_j's sean funciones constantes en cada intervalo. Por lo tanto, las funciones base quedan:

$$\Psi_1(x, \mathcal{P}) = I(x < \tau_1)$$
$$\Psi_2(x, \mathcal{P}) = I(\tau_1 \leq x < \tau_2)$$
$$\Psi_3(x, \mathcal{P}) = I(\tau_2 \leq x)$$

Con $I(\cdot)$ la función indicadora que vale 1 si x se encuentra en la región y 0 en otro caso. Por lo tanto,

$$\begin{aligned} h(x) &= \sum_{j=1}^{J} w_j \Psi_j(x) \\ &= w_1 I(x < \tau_1) + w_2 I(\tau_1 \leq x < \tau_2) + w_3 I(\tau_2 \leq x) \end{aligned}$$

Lo cual es una función *escalonada*, en el sentido de que para cada región de x se un nivel w_j.[14] Esta aproximación, podría servir para datos que estén agrupados por niveles, sin embargo, rara vez será este el caso.

Con este ejemplo sencillo, se ilustra a grandes rasgos como funcionan los polinomios por partes. Sin embargo, en cada intervalo se puede ajustar un polinomio de grado arbitrario, aumentando así, el número de funciones base. Adicionalmente, se pueden añadir restricciones de continuidad en los nodos, y no

13. Esto se hace dependiendo de si se busca hacer inferencia fuera del intervalo.
14. Sin entrar en el detalle, usando una función de perdida cuadrática, es fácil demostrar que cada $\hat{w}_j = \bar{x}_j$ es decir, para cada región, el mejor estimador constante, es el promedio de los puntos de esa región.

solo continuidad entre los polinomios, sino continuidad en las derivadas, lo cual logra una estimación más robusta. Esta es la magia de los polinomios por partes, que se les puede pedir cuanta *suavidad* (o no) se requiera, entendido como la continuidad de la K-ésima derivada. Tradicionalmente, se construyen polinomios cúbicos con segunda derivada continua en los nodos. Esto, pues resulta en funciones suaves al ojo humano además de que logran aproximar una gran cantidad de funciones.

Orígenes y justificación de su uso

La palabra *spline* usualmente se usa para designar a un grupo particular de polinomios por parte. Sin embargo, no hay consenso en la literatura de su definición exacta. Dependiendo de las particularidades se pueden denotar funciones diferentes. Para este trabajo se usa la definición de (Wasserman 2007) y (Hastie, Tibshirani y Friedman 2008). Un *spline de grado M* es un polinomio por partes de grado $M-1$ y continuidad hasta la $(M-2)$-derivada. Se hace notar, que existen muchos tipos de splines, además de que pueden ser, puede ser más flexibles o más rápidos en su implementación computacional como los B-Splines. En (Boor 1978) y más recientemente (Wahba 1990) se hacen tratados extensivos sobre ellos.

Como breviario historico, los splines originales, surgen en (Schoenberg 1964) como la solución al problema de encontrar la función h en el espacio de Sobolev W_M de funciones con $M-1$ derivadas continuas y M-ésima derivada integrable al cuadrado que minimice:

$$\int_a^b (h^{(M)}(x))^2\,dx$$

sujeta a que interpole los puntos $h(x_i) = h_i \quad i = 1, 2, \ldots, n$. Posteriormente, la teoría sobre los splines se fue expandiendo y fueron adoptados por ramas de la matemática tan diversas como los gráficos por computadora y, como es el caso, la estadística comutacional. Bajo este contexto, los splines también surgen de forma orgánica pues, el problema (2.10) se puede plantear como encontrar la función h que minimiza:

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda \int_a^b (h^{(M)}(x))^2\,dx \tag{2.11}$$

para alguna $\lambda > 0$ donde la solución se demuestra que son *splines cúbicos naturales* ($M = 4$). Cabe mencionar, que esta formulación del problema engloba muchas de técnicas estadísticas interesantes además de conceptos de optimización. El lector reconocerá que el primer término claramente es la *suma de residuales cuadrados (RSS)* y el segundo término del sumando es un caso particular de los métodos de regularización vistos anteriormente. No es el enfoque entrar en el detalle pues cambios menores en la formulación y diferentes elecciones de λ llevan a modelos que cada uno merece una tesis por si mismo. Sin embargo, es importante mencionar que la regularización y modelos de este tipo, son algunos de los más usados y útiles

en ML, pues logran captar patrones muy complejos al incluir muchos términos de orden superior e interacciones sin sobreajustar los datos. Como ejemplo, se puede encontrar fronteras de clasificación circulares usando un modelo logistico normal en $\mathbb{R}^2$ al incluir todos los términos polinomiales y las interacciones hasta orden 6. Por lo pronto, lo esencial, en la expresión (2.11), es que al tratar de minimizar el RSS se puede caer en problemas de sobreajuste en donde los parámetros no estén capturando efectos y patrones subyacentes, sino solo se trata de seguir los datos. Para compensar la complejidad, se penaliza la función a minimizar con segundo termino que controla el número de parámetros y la suavidad deseada mediante λ. A este segundo término, se le conoce como *penalización* y crece a medida que h se vuelve más complicada.[15]

Posterior a estas formulaciones, los splines vuelven a ser relevantes con el modelo aditivo de Hastie y Tibshirani. Ellos extienden la formulación de un espacio de covariables en una sola dimensión, a muchas. La formulación del problema es prácticamente la misma que (2.11) pero ahora se busca estimar d funciones h, dando lugar a tener más parámetros λ:

$$y = \sum_{j=0}^{d} h_j(x_j) + \epsilon$$

$$\text{RSS}(h_0, h_1, \ldots, h_d) = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{d} h_j(x_{ij})]^2 \; + \; \sum_{j=1}^{d} \lambda_j \int h_j''(t_j)\, dt_j$$

con la convención de que h_0 es una constante. Ellos muestran que $h_j \quad j = 1, \ldots, d$ son splines cúbicos. Sin embargo, sin restricciones adicionales, el modelo no sería *identificable*, es decir, la h_0 podría ser cualquier cosa. Para asegurar la unicidad de la solución se añade la condición de que las funciones estimadas, promedien cero sobre los datos:

$$\sum_{i=1}^{n} h_j(x_{ij}) = 0 \quad \forall j \tag{2.12}$$

Esto lleva a la conclusión natural de que h_0 sea la media de las variables de respuesta, es decir: $h_0 = \bar{y}$. Por lo que si se ve cada dimensión j, se tiene que su función correspondiente h_j está centrada alrededor de la media $\bar{y}$. Este hecho es fundamental para el modelo de este trabajo, esto se debe a que en realidad, h_j es *arbitraria* para toda j y sólo se necesita que tenga la magnitud necesaria para ajustar los datos. Es decir, dada h_0, la estimación y entrenamiento de los parámetros que definen por completo a h_{j^*} (con j^*

15. Si el lector tiene una intuición de análisis, notará que integrar la función al cuadrado, corresponde con el producto interno de las funciones pertenecientes al espacio de Hilbert $\mathcal{L}_2([a, b])$. Más detalles de esto en el Apéndice A

alguna $j = 1, \ldots, d$) deben ser tales para que esta ajuste los *residuales parciales*:

$$\hat{h}_{j^*} = y - h_0 - \sum_{\substack{j=1 \\ j \neq k}}^{d} h_j \tag{2.13}$$

y se vaya captando en esta h_{j^*} la información aún captada por el modelo. Esta lógica, además de brillante, es la que le da fuerza a los GAM, pero solo se puede entender de forma completa hasta que se estudie el algoritmo de *backfitting* en el Capítulo 3.

Formalización matemática de splines

Retomando la discusión de la página 19, se está buscando definir un polinomio de grado $M-1$ por partes en J intervalos. Tomando una expansión de bases para cada intervalo, como en el primer ejemplo que se dio, el número de funciones base aumenta en J por cada grado que se agregue, dando un total de $J * M$ bases funcionales, y en consecuencia, el mismo número de parámetros por estimar. Esto ocurre porque se necesita definir una base de tamaño M para cada subintervalo $j = 1, \ldots, J$. Es decir: $\mathcal{B}_j = \left\{1, x, x^2, \ldots, x^{M-1}\right\} \quad j = 1, \ldots, J$. Sin embargo, esto lleva a polinomios que se comportan de forma independiente en cada intervalo y no se conectan. Naturalmente, la primera condición que se piensa en imponer es continuidad en los nodos, lo cual devuelve $J-1$ parámetros que corresponden a los $J-1$ nodos. De la misma forma, cada grado de continuidad nodal en las derivadas que se le pida al polinomio, restringe el modelo y por ende, devuelve el mismo número de funciones bases. Sea K este número, es decir, que se tiene continuidad hasta la K-ésima derivada, se tiene un total de:

$$N^*(M, J, K) = M * J - K * (J - 1) \tag{2.14}$$

bases funcionales y por ende, el mismo número de parámetros por estimar w.[16]. Es claro que N^* es la *dimensión mínima* necesaria para construir polinomios por partes con estas características. Pues, el número de funciones N^* está, a su vez, en función de M definiendo el grado, el número de intervalos J (por ende el número de nodos) y el número de restricciones K.

Por lo pronto, y para continuar con una exposición constructiva, se centra la discusión cuando $K = M-1$ devolviendo la definición de spline: polinomios de grado $M-1$ con continuidad hasta la $(M-2)$-derivada.

16. En ocaciones es más fácil pensar en K como el número de restricciones que se imponen en los nodos. Asi, $K = 0$ implica que los intervalos son independientes, $K = 1$, implica que los polinomios se conectan, $K = 2$ implica continuidad en la primera derivada y asi sucesivamente. Naturalmente $K < M$

Por ende: $N^* = M + J - 1$. Ahora, se recuerda que el objetivo es darle forma funcional a Ψ. Para lograr esto habiendo incorporado el número de bases, se define la función auxiliar *parte positiva*:

$$x_+ = \text{máx}\left\{0, x\right\}.$$

Esta función, ayuda a se puede representar la expansión en bases de una forma relativamente sencilla. A esta expansión, se le conoce como *expansión en bases truncada*:

$$\begin{aligned} h(x) &= \sum_{i=1}^{M+J-1} w_i \Psi_i(x, \mathcal{P}) \\ &= \sum_{i=1}^{M} w_i \; x^{i-1} + \sum_{j=1}^{J-1} w_{M+i} \; (x - \tau_i)_+^{M-1} \end{aligned} \tag{2.15}$$

El primer sumando de (2.15) representa el *polinomio base*[17] de grado $M - 1$ que afecta a todo el rango. El segundo sumando, está compuesto únicamente de funciones parte positivas que se van activando a medida que x recorre el rango $[a, b]$ a la derecha y va pasando por los nodos. Estas funciones parte positiva, capturan el efecto de todos los intervalos anteriores que, al combinarlos con el primer sumando definen un polinomio de grado $M - 1$ en todo el intervalo[18]. Esta derivación de las bases, surge cuando se integra un polinomio por partes constante $M - 1$ veces. En cada iteración, las constantes se juntan y se integran por si solas, independientemente de los intervalos, lo cual deriva en este polinomio base. De forma explicita, se tiene que $\Psi(x, \mathcal{P})$ es:

$$\begin{aligned} &\Psi_1(x, \mathcal{P}) = 1 \\ &\Psi_2(x, \mathcal{P}) = x \\ &\vdots \\ &\Psi_M(x, \mathcal{P}) = x^{M-1} \\ &\qquad \text{el } \textit{polinomio base} \\ &\Psi_{M+1}(x, \mathcal{P}) = (x - \tau_1)_+^{M-1} \\ &\vdots \\ &\Psi_{M+J-1}(x, \mathcal{P}) = (x - \tau_{J-1})_+^{M-1} \\ &\qquad \text{la base truncada} \end{aligned}$$

17. *Baseline*, una vez más a falta de una mejor traducción
18. En realidad lo hace en todo $\mathbb{R}$

las cuales forman un espacio lineal de funciones $(M+J-1)$-dimensional. En la particularidad que $M=4$, se les conoce como splines cúbicos y son los más usados cuando se buscan funciones suaves. En la práctica han resultado ser de gran utilidad pues el ojo humano no detecta la posición de los nodos

2.3.2. Polinomios por parte flexibles

Independientemente de la elección de parametros en la construcción del polinomio, se tiene el problema de seleccionar la posición de los nodos. Existen procedimientos adaptativos, como los propuestos en (Friedman 1991). No obstante, y como ya se mencionó anteriormente, (Denison, Mallick y Smith 1998), proponen un método bayesiano más atractivo, que aunqué no se implemente en este trabajo, se implementa su expansión en bases aún más general. Una ligera modificación en la ecuación (2.15) la convierte, de un spline, a un polinomio por partes más general, con grado arbitrario de continuidad en las derivadas. Dejando atrás el supuesto que $K=M-1$ y devolviendole esa flexibilidad al modelo. Su expansión en bases queda:

$$h(x)=\sum_{l=1}^{N^*} w_l\ \Psi_l(x,\mathcal{P})=w^t\Psi(x,\mathcal{P}) \qquad \text{con } N^*=J*M-K*(J-1) \tag{2.16}$$

$$=\sum_{i=1}^{M} w_{i,0}\ x^{i-1}+\sum_{i=K}^{M-1}\sum_{j=1}^{J-1} w_{i,j}\ (x-\tau_j)_+^i \tag{2.17}$$

la cual es la expansión de bases implementada en el modelo final.

Dado que se tiene una doble suma, es necesario incluir un segundo índice, al menos temporalmente, a los pesos. El primer índice, denotado por i está asociado al grado de su función base; si $i=2$ entonces, $w_{2,j}$ está asociado a una término de grado 1 cuando $j=0$, pero a uno de grado 2 si $j>0$.[19]. El segundo índice $j=1,\ldots,J-1$ denota el nodo al que está asociado el peso. Como convención, si $j=0$, se hace referencia al polinomio base que siempre tiene efecto. En el segundo sumando de (2.17) la primera suma comienza en K. Recordando, K es el número de restricciones de continuidad que se imponen al polinomio en los nodos. Por ejemplo, $K=0$ implicaría que cada polinomio es independiente; $K=2$, se tiene continuidad en la función y en la primera derivada, etc. En el caso que $K=M-1$ se regresa a la ecuación (2.15) y se recuperan los splines que, por construcción, son suaves. La suavidad, aunque útil, no siempre es necesaria. Existen muchas funciones con primera y segunda derivada que varían rápidamente e incluso funciones discontinuas que no se podrían estimar usando splines, todo depende de los datos. Esta construcción, con

19. Esta desgraciada disparidad surge para ser consistente con la notación anterior, y no se puede indexar directamente en el primer sumando.

su doble suma, permite tener $M - K$ términos por nodo, codificando así las continuidades arbitrarias en las derivadas[20]. La ecuación (2.16) es una vez más la expansión en bases arbitrarias, igual a (2.9) pero definiendo bien a N^*. Además, si finalmente en esta ecuación se deja que $h(x)$ sea igual a $f_j(x_j)$ para toda $j = 1, \ldots, d$, se regresa a la ecuación canónica del modelo (2.4) presentada al principio de este trabajo. Este era el último componente que quedaba por definir, completando así la exposición matemática del modelo.

Para ayudar con la interpretación (y lectura) de (2.17), la Tabla 2.1, de la página 26, hace un compendio de los polinomios por partes. Esto ayuda no solo a esclarecer la notación, sino a formar una biyección entre w_l, $w_{i,j}$ y Ψ_l que posteriormente ayudará a expresar todo de forma matricial en su implementación en código.

Antes de cerrar la sección, se centra la atención en los nodos τ. A estos, se les ha dado poca importancia hasta el momento. Como ya se mencionó antes, en (Denison, Mallick y Smith 1998) se desarrolla, además de la ecuación (2.17) un paradigma bayesiano en el que los nodos, son tratados como parámetros y por ende sus posiciones cambian. La ventaja de que estos estén indeterminados, es que se pueden concentrar en los lugares donde la función varia más. Y al contrario, si la función es relativamente suave para alguna sección, se usan pocos nodos. Aunque hubiera sido bueno implementar esto, el algoritmo que *mueve* los nodos va ligado directamente a un proceso de eliminación y nacimiento de estos, haciendo que la J sea variable. En el trabajo original, esto no era un problema pues solo se hacían estimaciones para una dimension, $d = 1$. En el contexto de este modelo probit, implementar el algoritmo trans-dimensional que los autores proponen, hubiera implicado que N^* estrella, no fuera constante para toda variable, $j = 1, \ldots, d$. Sino que se tendría N_j^*, incorporado otra capa de complejidad innecesaria. Además, la implementación habría sido radicalmente diferente. En el Capítulo 3 se detalla como la simplificación de no incorporar los nodos como parámetros ayuda bastante a la velocidad del algoritmo. Posteriormente en el Capítulo 4, se ve que para fines prácticos, el modelo funciona de maravilla y finalmente en el Capítulo 5 se discute que habría cambiado de haberse implementado.

20. Esta codificación es sutil pues, al hacer los cálculos de continuidad, hay que considerar los límites izquierdos y derechos, los cuales existen siempre. Sin embargo, los términos $(x - \tau)_+^K$ se desvanecen únicamente hasta la K-esima derivada. Para la $(K + 1)$-derivada, el coeficiente correspondiente se suma a la función y rompe la continuidad pues no corresponde con el límite izquierdo

w_l	$w_{i,js}$	$\Psi_l(x,\mathcal{P})$		
Subíndice l	Subíndices i,j	Función Base		
1	$1,0$	1	M elementos	
2	$2,0$	x		
$\vdots$	$\vdots$	$\vdots$		
M	$M,0$	x^{M-1}		
$M+1$	$K,1$	$(x-\tau_1)_+^K$	$M-K$	$J-1$ veces
$M+2$	$K+1,1$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(M-K)$	$M-1,1$	$(x-\tau_1)_+^{M-1}$		
$M+(M-K)+1$	$K,2$	$(x-\tau_1)_+^K$	$M-K$	
$M+(M-K)+2$	$K+1,2$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+2(M-K)$	$M-1,2$	$(x-\tau_1)_+^{M-1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-2)(M-K)+1$	$K,J-1$	$(x-\tau_{J-1})_+^K$	$M-K$	
$M+(J-2)(M-K)+2$	$K+1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-1)(M-K)$	$M-1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		

Tabla 2.1: Biyección entre w_l, $w_{i,j}$ y sus correspondientes funciones base Ψ_l. Se termina con $N^* = M+(J-1)(M-K) = J*M-K*(J-1)$ términos, ecuación (2.14). Por construcción, se es consistente con la definición (2.15) si $K = M-1$.

2.3.3. Consideraciones finales

A pesar de la utilidad de los splines (y los polinomios por parte), todos sufren de problemas más allá del rango de entrenamiento $[a, b]$. Pues, su naturaleza global hace que fuera de la región con nodos, los polinomios crezcan o decrezcan rápidamente. Por lo tanto, extrapolar con polinomios o splines es peligroso y podría llevar a estimaciones erróneas. Para corregir esto, en ocasiones, se puede imponer una restricción adicional para que el polinomio sea lineal en sus extremos. Se usa el adjetivo de *natural* para designarlos. Esta modificación, libera $2 * (M - 2)$ funciones bases, pues quita todas las bases de orden mayor a 1 en los dos nodos frontera. Su expansión en bases, también se deriva de la ecuación (2.15). Es razonable que esta modificación mejore la fuerza predictiva fuera de el dominio de entrenamiento. Sin embargo, en general, en un contexto de regresión, se recomienda no hacer inferencia fuera de el espacio de covariables $\mathcal{X}$, pues en realidad, no se tiene evidencia para tomar conclusiones en esta región. Todo depende de los datos y el objetivo del modelo.

Se han usando los parámetros M, J y K para hablar del número de funciones base N^*, ecuación (2.14), pero se recuerda que también, dictan el número de *grados de libertad* del modelo. Es decir, el número de pesos o coeficientes w, las cuales son igual o más importantes que las bases, no solo porqué son parámetros a estimar, sino que son los que dictan el *ajuste* a los datos a diferencia de Ψ que solo los operan.

Al estar trabajando en espacios funcionales, la elección de base es relativamente arbitraria y se podría cambiar como lo hace una transformación de coordenadas en un espacio euclidiano. Cada base tiene sus beneficios y desventajas. Para esta exposición, se escoge la expansión en bases truncadas pues es explicada fácilmente y tiene una forma funcional relativamente sencilla además, la interpretación de los coeficientes w es inmediata. Sin embargo, no es óptima computacionalmente cuando J es grande. En la practica, usualmente se implementan B-Splines[21] que se derivan de lo vistos anteriormente. No obstante, para no complicar más la exposición (y el algoritmo en si) se implementó una versión optimizada de (2.17) con base en la Tabla (2.1) que funciona bastante rápido inclusive cuando J grande.

En la practica, los parámetros M, J y K se calibran pues, como ya se mencionó anteriormente, hacer J variable y automático es muy complejo. Asimismo, la elección de M y K requeriría cierta exploración previa de los datos. No obstante existen algoritmos que realizan esto, para los fines de este trabajo no aportaría mucho, además de que los resultados que se obtuvieron por el método de calibración son bastante buenos.

21. Vease el Capítulo 5.5 de (Wasserman 2007) o el Apéndice del Capítulo 5 en (Hastie, Tibshirani y Friedman 2008)

Si se le da rigor al modelo, en realidad, hay dos expansiones en bases. La primera la primera *a lo largo* de la ecuación lineal (2.3) cuyos coeficientes son β y las funciones base $\mathbf{f}$. Posteriormente, se tiene la expansión de la ecuación de polinomios por partes de (2.4) cuyos coeficientes son w_j y las funciones base Ψ_j para toda j. Esto explica el salto conceptual (y notacional) que se da entre las representaciones de (2.8) y (2.9).

Al tener en mente que se tienen d covariables, y por ende d polinomios por partes, además de la estructura lineal de (2.16) se puede sustituir (2.4) dentro de (2.3) dando la siguiente estructura con doble suma:

$$\begin{aligned} f(\mathbf{x}) &\approx \sum_{j=0}^{d} \beta_j f_j(x_j) \\ &\approx \beta_0 + \sum_{j=1}^{d} \beta_j \left[\sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_j, \mathcal{P}_j) \right] \end{aligned}$$

Lo cual, es perfectamente lineal. Se tienen $1 + d * N^*$ términos que se pueden acomodar en un solo vector. Sin embargo, se tiene un cruce de parámetros interesante, la multiplicación de la $\beta_i \quad \forall i$ contra $w_{i,j} \quad \forall j$. Tradicionalmente, no se usan β's y se deja que se capture ese efecto dentro de las f_i como en los GAM. Sin embargo, dado que el objetivo de este trabajo es la predicción, más que la estimación de funciones, se opta por dar una nueva capa de suavizamiento con β. No existe forma de garantizar ortogonalidad de β contra las todas las w, por lo tanto, se le da prioridad a la correcta estimación de w pues captura un mayor efecto además de que, por la construcción de los polinomios por partes, si está garantizada la ortogonalidad contra las funciones bases Ψ.

Capítulo 3

Paradigma bayesiano e implementación

Pasar de un modelo tan estructurado a su implementación computacional no resulto fácil. Sin embargo, se logró desarrollar un algoritmo que estima todos los parámetros del modelo de una forma eficiente y que funciona en la práctica. En el fondo, el algoritmo recae en el método de Gibbs sampling propuesto en (Albert y Chib 1993), por lo que se hace una breve introducción a la escuela de inferencia bayesiana, y en el algortimo de backfitting descrito en (Hastie y Tibshirani 1986). Al algoritmo se le titula: *bayesian piece wise polinomial model (bpwpm)*. Para facilitar la utilización del modelo en diversas bases de datos, así como su validación y visualización, a la par del algoritmo se desarrolló un paquete de código abierto (con el mismo nombre) para el software estadístico `R`. Al darle un tratamiento bayesiano a los parámetros, más que estimarlos, se busca regresar una muestra de tamaño arbitrario de sus correspondientes distribuciones posteriores. La idea, es que estas distribuciones posteriores, se haya capturado toda la información de los datos de entrenamiento.

Se considera, que una buena forma de entender el algoritmo es *visualizando* tanto los datos como los objetos que componen el modelo, por lo tanto se hace un paréntesis notacional. De las ecuaciones del modelo: (2.1) a (2.4), se tienen dos grupos de parámetros por estimar, $\beta \in \mathbb{R}^{d+1}$ y $w_j \in \mathbb{R}^{N^*} \quad \forall j = 1, \ldots, d$. Donde:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} \quad \text{y} \quad w_1 = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{1,N^*} \end{bmatrix} \quad \ldots \quad w_d = \begin{bmatrix} w_{d,1} \\ \vdots \\ w_{d,N^*} \end{bmatrix}$$

Se hace énfasis en que existen d vectores w_j, cada uno de tamaño N^*. Por lo tanto, se tienen sun total de $1 + d + dN^*$ parámetros. Se usa el símbolo $\mathbf{w}$ para designar todos los vectores w_j, haciendo de este una

matriz, es decir: $\mathbf{w} \in \mathbb{R}^{d \times N^*}$. Cuando se habla de datos: $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, estos se pueden representar en una tabla (o matriz):

$$\left[\begin{array}{c|ccc} y_1 & x_{1,1} & \dots & x_{1,d} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,d} & \dots & x_{n,d} \end{array}\right]$$

Donde el vector de observaciones binarias $\mathbf{y} = (y_1, \dots, y_n)^t$ es la primer columna de la tabla, y la matriz de covariables $\mathbf{X}$ es el resto.

Bajo esta representación, se da contexto cuando se habla de que la estimación debe reflejar los patrones *hacia abajo* y *hacia lo largo*. Hacia abajo, se está captando la información existente entre las observaciones; cada f_j, mediante su parámetro w_j, representa una transformación no lineal de la variable (o dimensión) j. Hacia lo largo, la función de proyección f suma cada f_j a través de β, ponderando los efectos individuales de cada variable. Mantener el balance entre la estimación de β a lo largo y w_j hacia abajo, es fundamental para el algoritmo. Analizando este hecho, se concluye que la estimación de ambos grupos de parámetros, se puede ver como una regresión separada para cada uno, y por ende, estos pueden ser estimados por el mismo algoritmo. Esto responde a la dualidad que se exploró en el capítulo pasado de que ambas expresiones son expansiones en bases funcionales. El puente que conecta, y controla el balance entre ambas, son los residuales parciales. Los siguientes capítulos, se concentran en explicar e implementar este curioso patrón.

3.1. Fundamentos de la estadística bayesiana

Dado el problema de describir fenómenos bajo incertidumbre, existen dos escuelas dominantes de la estadística: la frecuentista y la bayesiana. La primera, aunque increíblemente útil, está hasta cierto punto limitada y en ocasiones termina derivando en colecciones de algoritmos. La teoría bayesiana, por el contrario, nombrada así en honor a Thomas Bayes (1702 - 1761), es una rama que enfatiza el componente *probabilista*, dando coherencia al proceso de inferencia (Mendoza y Regueiro 2011) y (Bernardo y Smith 2001). La estadistica bayesiana está axiomatizada bajo la *teoría de la decisión*. Esta teoría formaliza conceptos económicos como la *coherencia entre preferencias* y *utilidad*, sobre los que desarrolla un marco metodológico para la toma de decisiones.

Esta metodología, además de proveer técnicas concretas para resolver problemas, también formaliza en una forma de pensar sobre la probabilidad como una *medida racional para cuantificar la incertidumbre* condicionando sobre el conocimiento existente. Este paradigma es el que más corresponde con el sentido

que usualmente se le da a la palabra. La inferencia sobre creencias (o parámetros), se realiza mediante una *actualización* de estas en luz de nueva evidencia, modificando su medida de incertidumbre. El mecanismo que permite realizar esto, es la aclamada formula de Bayes. De manera informal se puede describir como: dado un evento E bajo condiciones C, la probabilidad *posterior* del evento, es proporcional a la probabilidad *previa* que se tiene sobre este, ponderado por la probabilidad de ocurrencia de las condiciones presentes, es decir:

$$P(E|C) \; \alpha \; P(C|E)P(E) \tag{3.1}$$

El término central $P(C|E)$ es una medida descriptiva de las condiciones (usualmente datos) llamada *verosimilitud*. Se hace notar que para poder hacer cualquier intento de descripción, se debe especificar el *modelo probablistico* que se asume describe el estado por el que se dan las condiciones C.
En un contexto matemático más formal, la cuantificación de la incertidumbre se da a través de medidas de probabilidad $\pi(\cdot)$, que describan el fenómeno observado. Estas medidas de probabilidad, usualmente son funciones que dependen de cantidades desconocidas llamados parámetros θ. Aunque desconocidas, se tienen ciertas creencias u conocimiento previo, *a priori*, sobre ellos, descritos por su correspondiente medida de probabilidad $\pi(\theta)$. Además, se tienen datos $\mathbf{X}$, interpretados como *evidencia*, a los cuales se les asigna un modelo de probabilidad dependiente de los parámetros, es decir, su verosimilitud: $\pi(\mathbf{X}|\theta)$. Usando la formula de Bayes, podemos actualizar el conocimiento que se tiene sobre los parámetros haciendo:

$$\pi(\theta|\mathbf{X}) \; \alpha \; \pi(\mathbf{X}|\theta)\pi(\theta) \tag{3.2}$$

La idea es que este proceso de actualización sea a la vez, un proceso de aprendizaje, en el cual los parámetros capturen la información contenida en los datos.

La teoría frecuentista, adopta un enfoque diferente para el aprendizaje. Se asume que no hay incertidumbre en los parámetros dado los datos y, por lo tanto, estos son tomados como fijos. El mecanismo que permite su estimación, usualmente consiste en plantear una función objetivo y optimizarla. Por ejemplo, si se escoge la verosimilitud $\pi(\mathbf{X}|\theta)$, se busca dar un estimador que la maximice, pues equivaldría a encontrar los parámetros que hagan más *posibles* los datos, bajo el modelo planteado. Si por el contrario, es escoge una función como la RSS de los modelos ANOVA (primer sumando de (2.11)), se busca la θ que minimice estos errores, así, el modelo logra capturar toda la variabilidad que puede sobre los datos. Independientemente del paradigma estadístico que se escoja, siempre es importante la validación del modelo y de sus supuestos. Además, tanto teoría bayesiana como frequentista han resultado de infinita utilidad en la practica y el avance de la estadística y ciencia en general.

Una de las dificultades que surgen en la estadística bayesiana, es que la obtención de resultados analíticos cerrados es difícil o muy tedioso una vez que los modelos se empiezan a complicar. Por ejemplo, en las ecuaciones anteriores, se ha usado el argumento de proporcionalidad α. Esto pues, para que se de la igualdad, el lado derecho de la ecuación (3.2) se debe de dividir entre $\pi(\mathbf{X}) = \int \pi(X|\tilde{\theta})\pi(\tilde{\theta})\, d\tilde{\theta}$, el cual usualmente es difícil, sino imposible, de calcular. A este término se le conoce como *constante de proporcionalidad* y su función es la de reescalar la expresión del lado derecho para que en realidad se tenga una distribución en el izquierdo. Usualmente, se escoger distribuciones *conjugadas*, para que tanto la distribución a priori como la posterior sean de la misma familia y por ende conocidas. Sin embargo, con los avances en el poder computacional disponible y técnicas numéricas para resolver integrales (Robert y Casella 2004), se han desarrollado muchos métodos para aplicar el proceso de aprendizaje, independientemente de que tan complejo sea el modelo o las distribuciones, iniciales y resultantes. Muchos de estos métodos recaen en la teoría de las *cadenas de Markov*, como lo es, el Gibbs sampler presentado en la sección. 3.2

Estimadores Bayesianos

Una vez realizado el proceso de actualización, el estadista se enfrenta con un problema. Se tiene una distribución posterior de probabilidad para los parámetros de interés, usualmente dada por una muestra y no por una distribución analítica. Sin embargo, por practicidad y utilidad, en ocaciones se busca dar un *estimador puntual*. Por ejemplo, si se necesita dar un estimador $\hat{\theta}$ para usarlo en otros cálculos, o si $\pi(\theta|\mathbf{X})$ es multidimensional. Para superar este problema teórico, se ha adoptado por usar *funciones de perdida* $L(\hat{\theta}, \theta)$[1]. Estas, miden las *consecuencias* que se dan, al tomar $\hat{\theta}$ como el verdadero valor del parámetro θ, es decir, las funciones de perdida evalúan que tan bien se está representando el valor de θ con un estimador puntual. Por ello, vale la pena usar funciones que penalicen la distancia entre θ y $\hat{\theta}$. Sin entrar mucho en los detalles técnicos, se tiene que calcular:

$$\hat{\theta} = \mathbb{E}[L(\hat{\theta}, \theta)] = \int_{\Theta} L(\hat{\theta}, \theta)\pi(\theta)\, d\theta \tag{3.3}$$

con Θ el espacio de todas las posibles valores de θ. Sin embargo, se demuestra que para funciones de perdida sencillas, pero intuitivas, se tiene que el estimador puntual posterior es alguna medida de centralidad de la distribución posterior. Por ejemplo:

- Función de pérdida cuadrática: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, deriva en la media posterior, es decir: $\hat{\theta} = \mathbb{E}[\theta|\mathbf{X}]$
- Función de perdida valor absoluto: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, deriva en la mediana de la distribución posterior.

1. Formalmente se tiene un problema de decisión.

- Función de pérdida 0-1: $L(\hat{\theta}, \theta) = I[\hat{\theta} \neq \theta]$, deriva en la moda de la distribución posterior.

En la práctica, estas cantidades son fáciles de calcular cuando se tiene una muestra simulada de θ proveniente de la distribución posterior. En el paquete, se implementa una forma sencilla de obtener estimadores puntuales con cualquiera de las 3 funciones de pérdida. Sin embargo, se verá que los resultados no varían mucho. Ver Apéndice B.

3.1.1. Funciones de probabilidad condicional completas

Retomando el modelo que concierne a este trabajo, se tienen dos grupos de parámetros, β y $\mathbf{w}$. Sin embargo, dados los supuestos del modelo, por el uso de la variable latente z, esta también se debe de incluir como parámetro pues es la liga entre la respuesta y y los datos $\mathbf{X}$, vista de forma bayesiana, también se debe de simular. Por lo tanto, los parámetros quedan: $\theta = (\mathbf{z}, \beta, \mathbf{w})$ con $\mathbf{z} = (z_1, \ldots, z_n)^t$. Esta sección concierne desglosar el proceso de aprendizaje sobre ellos; esta derivación es importante en si pues es la que induce el algoritmo. Usando la notación presentada al inicio de esta sección, los supuestos propuestos en las ecuaciones del modelo (2.1) a (2.4) y sustituyendo en (3.2) se tiene:

$$\begin{aligned}
\pi(\mathbf{z}, \beta, \mathbf{w}|\mathbf{y}, \mathbf{X}) \quad &\alpha \quad \pi(\mathbf{y}|\mathbf{X}, \mathbf{z}, \beta, \mathbf{w})\ \pi(\mathbf{z}, \beta, \mathbf{w}) \\
&\alpha \quad \pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\mathbf{X}, \beta, \mathbf{w})\ \pi(\beta, \mathbf{w}) \\
&\alpha \quad \pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\mathbf{X}, \beta, \mathbf{w})\ \pi(\beta)\ \pi(\mathbf{w}) \\
&\alpha \quad \prod_{i=1}^{n} \mathrm{Be}[y_i|\Phi(z_i))]\ \phi[z_i|f(\mathbf{x}_i), 1] \times \pi(\beta)\ \pi(\mathbf{w}) \\
&\alpha \quad \prod_{i=1}^{n} \mathrm{Be}[y_i|\Phi(z_i))]\ \phi[z_i|\beta^t\mathbf{f}(\mathbf{x}_i), 1] \times \pi(\beta)\ \pi(\mathbf{w}) \qquad (3.4)
\end{aligned}$$

donde $\phi(\cdot|\mu, \sigma^2)$ es la función de densidad de una variables aleatoria normal con media μ y varianza σ^2; asimismo $\mathrm{Be}(\cdot|p)$ es función de densidad de una variable Bernoulli con probabilidad de éxito p. Esta factorización es válida dados los supuestos, donde se hace notar, la forma que conecta z_i a las dos partes del modelo a través de la función de proyección $f(\mathbf{x}_i) = \beta^t\mathbf{f}(\mathbf{x}_i)$ que contiene tanto a β como $\mathbf{w}$. Esta derivación es una forma extendida (aunque simplificada) de la verosimilitud para todas las observaciones $i = 1, \ldots, n$. Aunque aún no se han especificado las formas funcionales para las distribuciones a priori $\pi(\mathbf{w})$ y $\pi(\beta)$, estas se pueden separar ya que se asumen independientes. Esta propiedad, combinada con la forma funcional en expansiones de bases, lleva a que se piense en hacer una estimación *por bloques*, es decir, se estima primero β y posteriormente $\mathbf{w}$ en un bucle iterativo, pues esta es la idea de un Gibbs sampler.

3.2. Simulación bayesiana: cadenas de Markov y el Gibbs sampler

Una vez establecida el proceso de actualización, el estadista se ve en la necesidad de tener que desarrollar técnicas para simular de la distribución posterior $\pi(\theta|\mathbf{X})$ sin importar que complejo sea el modelo. Desde principios de los años noventa, se desarrollaron muchos algoritmos y paquetería estadística al aumentar el poder computacional. La gran mayoría de los algoritmos recae en los *métodos Monte Carlo de cadenas de Markov* (MCMC). Estos métodos, como su nombre lo indica, hacen alusión a principios de aleatoriedad, como se daría en un casino. Usando ideas intuitivas de probabilidad y números pseudoaleatorios, se pueden generar muestras prácticamente de cualquier distribución, incluso si su forma funcional es desconocida. La simulación, como tal es un tema que merece un estudio más profundo (Robert y Casella 2004). Estas poderosas simulaciones, permitieron que los estadistas y experimentadores pudieran hacer el menor número de supuestos posibles sobre los modelos, puesto que ya no se buscan resultados analíticos sino más bien, se buscaba reflejar la realidad y dejar que los cálculos los hiciera una computadora.

Breve introducción a cadenas de Markov

Al final, la gran mayoría de estos métodos recaen sobre las la teoría de *cadenas de Markov*. Una cadena de Markov, es una secuencia de variables aleatorias: $X^{(1)}$, $X^{(2)}$, ... que cumplen la *propiedad Markoviana*:

$$P(X^{(t+1)}|X^{(t)} = x^{(t)}\,, X^{(t-1)} = x^{(t-1)}\,, \ldots, X^{(2)} = x^{(2)},\, X^{(1)} = x^{(1)}) = P(X^{(t+1)}|X^{(t)} = x^{(t)}) \quad \forall t$$

con t interpretado como *tiempo*. Por lo tanto, la siguiente variable de la cadena, $X^{(t+1)}$, únicamente depende de *el estado* actual $X^{(t)}$ y no de los anteriores. Usualmente esta propiead es expresada como: el futuro, condicionando al presente, es independiente del pasado. El ejemplo canónico que se presenta es la *caminata aleatoria*: $X^{(t+1)} = X^{(t)} + e^{(t)}$, con $e^{(t)}$ error aleatorio generado de forma independiente. De esta idea se desarrolla toda una rica teoría revisada en cursos de procesos estocásicos (Ross 2009), de donde surgen muchas propiedades aplicables a las cadenas. Una de las ideas más relevantes para lo que concierne este trabajo, es la de *matrices de transición*. Dada una cadena con n posibles estados, es decir, $X^{(t)}$ únicamente puede tomar valores de un subconjunto de cardinalidad n. Se puede construir una matriz cuadrada $P \in \mathbb{R}^{n\times n}$ donde cada entrada $0 \leq p_{i,j} \leq 1$ representa la probabilidad de transicionar del estado i al estado j. Se demuestra, que si una cadena es *ergodica*[2], entonces existe una *distribución límite* que es igual a la *distribución estacionaria*: $\exists\, \pi$ tal que $\pi P = \pi$. Sin entrar en los detalles técnicos, la ergodicidad es la propiedad que asegura que eventualmente se alcanza la convergencia de la cadena sin importar el

2. Aperiódica, irreducible y recurrente positiva. Para efectos de simplicidad en la exposición, la ergodicidad es tratada como una propiedad en si misma. Las definiciones formales, puede ser consultadas en cualquier texto de procesos estocásticos.

estado inicial tras repetidas aplicaciones de la matriz de transición P[3]. Esto es, dado un vector de estados inicial $X^{(0)} \in \mathbb{R}^n$ tal que $\mathbf{1}^t X^{(0)} = 1$, se puede encontrar la distribución estacionaria dejando:

$$\pi = \lim_{t \to \infty} X^{(t)} \quad \text{si} \quad X^{(t+1)} = P^t X^{(t)} \tag{3.5}$$

Esta idea se puede extender a casos más complejos donde se relajan o se cambian supuestos. Incluso, se extiende a casos donde el número de estados es no finito, pero la idea fundamental es la misma.

En el contexto de este trabajo la idea es poder simular *secuencialmente* cadenas de parámetros θ que estén ligados unos con los otros, que dependan únicamente de el presente y, sobre todo, que una vez simuladas un número arbitrario de estas, converjan a la distribución estacionaria. Precisamente lo que hace un Gibbs sampler.

Gibbs sampler

El Gibbs sampler como tal, es una técnica, para simular variables aleatorias de una *distribución conjunta* sin tener que calcularla directamente, análoga a lo que se vió en más arriba (Gelfand y Smith 1990) y (Casella y George 1992). Usualmente, el muestreo de Gibbs se usa dentro de un contexto bayesiano, aunque también funciona para otras aplicaciones. A primera vista, parece misterioso, pero en realidad, se basa únicamente en las propiedades revisadas (relativamente sencillas) de las cadenas de Markov. Sin perdida de generalidad, se busca simular una muestra de parámetros $\theta = (\theta_1, \ldots, \theta_p)$ que provienen de la distribución conjunta $\pi(\theta|\cdot)$. Esta distribución usualmente no es conocida analíticamente, sin embargo el Gibbs sampler, nos permite dar, no la distribución como tal, pero si una muestra arbitrariamente grande con la que se puede aproximar empíricamente $\hat{\pi}(\theta) \approx \pi(\theta)$. En la practica usualmente más que aproximar la distribución, se busca alguna función de los parámetros como la media o la varianza de la distribución.

Para llevar a cabo el muestreo, se intercambia el difícil cálculo de la distribución conjunta al cálculo de las distribuciones condicionales que usualmente son más fáciles de derivar. Las distribuciones condicionales

3. Esta convergencia es muy diferente a la presentada en el Apéndice **??**. Cuando se cambia de un paradigma frecuentista a uno bayesiano, los teoremas que aseguran la convergencia del modelo son radicalmente diferentes, tanto en forma como en fondo.

están dadas por:

$$\begin{aligned}
\theta_1 &\sim \pi(\theta_1|\theta_2,\ldots,\theta_p) \\
\theta_2 &\sim \pi(\theta_2|\theta_1,\theta_3,\ldots,\theta_p) \\
&\vdots \\
\theta_p &\sim \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})
\end{aligned} \tag{3.6}$$

Se comienza con una muestra inicial arbitraria $\theta^{(0)} = (\theta_1^{(0)},\ldots,\theta_p^{(0)})^t$, donde el superíndice $^{(k)}$ corresponde a la iteración k. Se comienza a simular de las correspondientes distribuciones condicionales, las cuales quedan especificadas para los valores iniciales. En este caso, para $k = 1, 2, 3, \ldots$, se tiene:

$$\begin{aligned}
\theta_1^{(k)} &\sim \pi(\theta_1|\theta_2^{(k-1)},\ldots,\theta_p^{(k-1)}) \\
\theta_2^{(k)} &\sim \pi(\theta_2|\theta_1^{(k)},\theta_3^{(k-1)},\ldots,\theta_p^{(k-1)}) \\
&\vdots \\
\theta_p^{(k)} &\sim \pi(\theta_p|\theta_1^{(k)},\ldots,\theta_{p-1}^{(k)})
\end{aligned} \tag{3.7}$$

Este proceso se itera hasta tener una muestra de tamaño arbitrario, que haya alcanzado la región de probabilidad donde se encuentra la distribución estacionaria, en este caso la distribución posterior $\pi(\theta|\cdot)$.

La convergencia no es intuitiva, es decir, no es trivial derivar que al muestrear de las distribuciones condicionales, se llegue (eventualmente) a un la distribución conjunta. Sin embargo, la prueba formal, aunque compleja, recae en que se puede formar una matriz de transición con las condicionales de θ_i, análoga a las matrices de las cadenas de Markov. Al dejar que $k \to \infty$, se llega a un resultado equivalente al de la ecuación (3.5), habiendo muestrado de la distribución posterior. Sin embargo, la ergodicidad, es un supuesto importante que se preserva y es necesario validad. Esto, pues se pueden dar casos donde la distribución posterior no existe.

El tener una muestra de la distribución final $\left\{\theta^{(k)}\right\}_{k=k^*}^{N_{\text{sim}}}$, donde N_{sim} es el número total de simulaciones arbitraria y k^* es el punto a partir del cual se obtiene la convergencia a $\pi(\theta|\cdot)$, tiene muchos beneficios en la practica. Se le pueden calcular momentos a la muestra, medidas de desviación y hacer representaciones gráficas para su análisis y evaluación. En la Figura 3.1, se tienen dos imágenes que de una muestra Gibbs para una simulación donde: $\theta = \beta \in \mathbb{R}^3$, $N_{\text{sim}} = 1000$ y $k^* = 500$. Para esta figura, se toman los últimos 500 valores de las cadenas. Se observan las *trazas* que se forman al ir simulando los parámetros y se hace

(a) Trazas (b) Histogramas

Figura 3.1: Muestro Gibbs para el ejemplo **??**

un histograma, dando una idea de las distribuciones subyacentes[4].

En la practica, muchos de los pormenores derivados del muestreo Gibbs, pueden ser mejorados. Dado que el valor inicial $\theta^{(0)}$ es dado por el estadista, en ocaciones el método tiene que explorar una región extensa de posibles valores para θ, por lo que podría tardar en converger. Esto deriva en que las primeras observaciones deban ser descartadas pues no son realizaciones de la distribución final buscada. A este periodo se le conoce como *burn-in*. En el ejemplo anterior, se utilizó k^* para designar la iteración a partir de cual se toman los valores simulados. En la practica, usualmente se guarda toda la cadena, se explora, tanto por resúmenes numéricos como con representaciones gráficas y se decide (de forma subjetiva) el corte k^*. Otro método ampliamente usado es el de adelgazamiento o *thinning*. Por las mismas características de la estimación por Gibbs, sobre todo en casos multivariados, los valores de las cadenas tienen correlaciones altas. Si se quieren muestras independientes, la bibliografía recomienda tomar cada k_{thin}-ésimo valor de la cadena generada para reducir la dependencia entre los parámetros. Usualmente se usan valores pequeños para k_{thin}. Estos sencillos pasos pasos para mejorar las cadenas, ya se encuentran implementados en el paquete *bpwpm* para `R` para el análisis rápido de las cadenas.

3.2.1. Algoritmo de Albert y Chibb

El algoritmo particular del Gibbs sampler que se usa en este trabajo, es una versión modificada del presentado en (Albert y Chib 1993). Este método ofrece varias ventajas para esta aplicación en particular,

4. Estas imágenes tienen como propósito ejemplificar el Gibbs sampler. El modelo que se usa es el presentado en este trabajo. Asimismo, en el Capítulo **??** se hará una exploración a fondo de este ejemplo en particular y se verá por que el parámetro β_2 es idénticamente cero, además de la razón por la que los histogramas se ven sospechosamente parecidos a una distribución normal. Asimismo, estas imágenes fueron generadas con la librería *ggplot2*, incorporada a las funcionalidades del paquete *bpwpm* desarrollado para este trabajo.

pues se desarrolló específicamente para regresiones probit usando la variables latente z. Además es un método muy eficiente pues usa distribuciones conjugadas, por lo que las distribuciones condicionales, se puede calcular directamente y la parte estocástica depende únicamente de simular valores de una distribución normal multivariada. Esto lleva a que los periodos de burn-in sean relativamente pequeños y que el adelgazamiento no sea fundamentalmente necesarios.

A su algoritmo, ellos lo llaman *Data Augmentation for Binary Data* y son los pioneros en el uso de variables latentes para unir la respuesta y con las covariables $\mathbf{x}$ como se vió en la sección 2.1. En su exposición, ellos no utilizan una función de proyección no lineal como la usada en este trabajo, sino, se restringen a la tradicional $f(\mathbf{x}) = \beta^t\mathbf{x}$. Por lo mismo (y por un breve momento) se utiliza la misma para explicar la idea fundamental. Usando la misma notación, se introducen n variables latentes $\mathbf{z} = (z_1, \ldots, z_n)^t$, con $z_i \sim N(\beta^t\mathbf{x}_i, 1)$. Se definen las respuestas:

$$y_i = \begin{cases} 1, & \text{si } z_i > 0 \\ 0, & \text{si } z_i \leq 0 \end{cases} \tag{3.8}$$

La simplificación del modelo, obliga a que, por el momento $\theta = (\mathbf{z}, \beta)$. Por lo tanto, la derivación bayesiana es:

$$\begin{aligned} \pi(\mathbf{z}, \beta|\mathbf{y}, \mathbf{X}) &\alpha \pi(\mathbf{y}|\mathbf{X}, \mathbf{z}, \beta)\ \pi(\mathbf{z}, \beta) \\ &\alpha \pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\beta, \mathbf{X})\ \pi(\beta) \\ &\alpha \prod_{i=1}^{n}[I(y_i = 1)I(z_i > 0) + I(y_i = 0)I(z_i \leq 0)] \times \phi(z_i|\beta^t\mathbf{x}_i, 1) \times \pi(\beta) \end{aligned} \tag{3.9}$$

Ahora, bajo los fundamentos del Gibbs Sampler, en lugar de querer encontrar la distribución posterior (3.9), se busca encontrar las distribuciones condicionales. Para β:

$$\pi(\beta|\mathbf{z}, \mathbf{y}, \mathbf{X}) = \frac{\pi(\mathbf{z}, \beta|\mathbf{y}, \mathbf{X})}{\pi(\mathbf{z})} \tag{3.10}$$

$$= \frac{\pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\beta, \mathbf{X})\ \pi(\beta)}{\pi(\mathbf{y}, \mathbf{X})\ \pi(\mathbf{z})}$$

$$= \frac{\pi(\mathbf{y}|\mathbf{z})}{\pi(\mathbf{y}, \mathbf{X})\ \pi(\mathbf{z})} \xrightarrow{C} \times \pi(\mathbf{z}|\beta, \mathbf{X})\ \pi(\beta) \tag{3.11}$$

$$= C\,\pi(\beta) \prod_{i=1}^{n} \phi(z_i|\beta^t\mathbf{x}_i, 1) \tag{3.12}$$

la densidad condicional completa es la misma que que la de una regresión lineal bayesiana donde $z_i = \beta^t \mathbf{x}_i + e_i$ con $e_i \sim N(0,1)$. Es decir, al estar usando z como regresor (dado) y condicionar sobre él, la simulación de β se reduce a la que se haría sobre un modelo lineal bayesiano tradicional, asimismo, para estos modelos, dependiendo de $\pi(\beta)$ existen resultados cerrados. Se hace notar que la ecuación (3.10) se toma de la definición de probabilidad condicional, y el paso de (3.11) a (3.12) se puede hacer ya que, al definir y como en la ecuación (3.8), sus representaciones son análogas y el cociente se desvanece, quedando únicamente la constante C que sale del término $\pi(\mathbf{y}, \mathbf{X})$. Ahora, únicamente falta definir $\pi(\beta)$. Es común en la práctica usar distribuciones *no informativas* sobre los parámetros, cuando no se tiene experiencia sobre ellos. Sin embargo, para el modelo lineal bayesiano, existe una familia de distribuciones conjugadas (Banerjee 2008). En particular, al dejar la varianza fija y eligiendo de distribución *a priori* $\pi(\beta)$:

$$\beta \sim N_{d+1}(\beta|\mu_\beta, \Sigma_\beta) \tag{3.13}$$

donde $\mu_\beta \in \mathbb{R}^{d+1}$ es el hiperparámetro de media y $\Sigma_\beta \in \mathbb{R}^{(d+1)^2}$ la matriz de covarianza, entonces, se tiene la distribución conjugada:

$$\beta|\mathbf{y}, \mathbf{z}, \mathbf{X} \sim N_{d+1}(\beta|\mu^*_\beta, \Sigma^*_\beta) \tag{3.14}$$

donde:

$$\mu^*_\beta = \Sigma^*_\beta \times (\Sigma^{-1}_\beta \mu_\beta + \mathbf{X}^t \mathbf{z})$$
$$\Sigma^*_\beta = (\Sigma^{-1}_\beta + \mathbf{X}^t \mathbf{X})^{-1}$$

Esta distribución es conjugada pues preserva la estructura normal de β. Asimismo, es fácil de simular usando cualquier software estadístico, calculando previamente todos los parámetros y dando un valor (o iteración) para $\mathbf{z}$[5].

Ahora, condicionar sobre $\mathbf{z}$, es más sencillo, pues la derivación es similar, comenzando con la expresión

5. Se hace notar, que este estimador, es relativamente parecido al estimador que se da en una regresión cordillera.

(3.9) y reordenando términos:

$$\begin{aligned}
\pi(\mathbf{z}|\beta,\mathbf{y},\mathbf{X}) &= \frac{\pi(\mathbf{z},\beta|\mathbf{y},\mathbf{X})}{\pi(\beta)} \\
&= \frac{\pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\beta,\mathbf{X})\ \pi(\beta)}{\pi(\mathbf{y},\mathbf{X})\ \pi(\beta)} \\
&= \cancel{\frac{1}{\pi(\mathbf{y},\mathbf{X})}}^{\,C}\ \pi(\mathbf{y}|\mathbf{z}) \times \pi(\mathbf{z}|\beta,\mathbf{X}) \\
&= C \prod_{i=1}^{n} [I(y_i=1)I(z_i>0) + I(y_i=0)I(z_i\leq 0)] \times \phi(z_i|\beta^t\mathbf{x}_i,1) \qquad (3.15)
\end{aligned}$$

De donde es claro ver que las z_i son independientes y tienen distribuciones normales truncadas en 0:

$$\begin{aligned}
&z_i|y_i=1,\beta \sim N(z_i|\beta^t\mathbf{x}_i,1) \text{ truncada a la izquierda} \qquad (3.16)\\
&z_i|y_i=0,\beta \sim N(z_i|\beta^t\mathbf{x}_i,1) \text{ truncada a la derecha}
\end{aligned}$$

las cuales también son fáciles de simular usando los algortimos de (Devroye 1986).

Finalmente, una vez que se tienen las distribuciones condicionales (3.16) y (3.14), la simulación de esto dos primeros grupos de parámetros se realiza con un muestreo de Gibbs dado por las ecuaciones (3.7) de la página 36. En forma de pseudo-código:

```
SampleoGibbs(y, X, N_sim, beta(0), mu_beta, sigma_beta):
    sigma_beta <- Inv(Inv(sigma_beta) + X'X)
    PARA k = 1 HASTA N_sim:
        z(k) <- NormTrunc(y, X, beta(k-1))
        mu_beta(k) <- sigma_beta*(Inv(sigma_beta)*mu_beta + X'z(k))
        beta(k) <- NormMulti(sigma_beta(k))
    REGRESAR beta
```

El valor de $\mathbf{z}^{(0)}$ en realidad no se tiene que dar; el valor inicial para $\beta^{(0)}$, puede ser dado por el estimador de máxima verosimilitud o el de mínimos cuadrados para las respuestas binarias $\beta^{(0)} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y$. Sin embargo, en la práctica el algoritmo, por default, los inicializa en ceros y tardan poco tiempo en converger a la distribución límite.

Consideraciones adicionales para la simulación de w

Para los parámetros, se usan las siguientes distribuciones *apriori*:

$$\beta = (\beta_0, \beta_1, \ldots, \beta_d)^t \sim N_d(\mu_0, \Sigma_0) \tag{3.17}$$

$$w^{(i)} = (w_1^{(i)}, \ldots, w_J^{(i)})^t \sim N_J(\mu_0^{(i)}, \Sigma_0^{(i)}) \quad i = 1, \ldots, d \tag{3.18}$$

3.3. Algoritmo *bpwpm*

En forma de pseudocódigo el algoritmo tiene la siguiente forma:
A diferencia de la exposición del modelo, el algoritmo debe de construir de abajo hacia arriba, pues se necesita tener una estimación puntual de los parámetros para poder calcular las funciones intermedias y que todo quede definido de forma numérica.

```
Parametros inciales:
WHILE (...)
    Transformación de X -> Phi -> F (Función: estimate_PWP)
    Simulación de betas (Función simulate_beta)
```

- Hacer énfasis en el apéndice y el paquete.

3.3.1. Algoritmo de *backfitting* para ajuste de modelos GAM

- Justificación final para las w's
- Usamos los nodos iniciales en cuantiles determinados.

- taus: HMC
- β Estimar por máxima verosimilitud pero dentro del Gibbs con el método ABC
- $w's$ BAyesianas + importantes que las betas.

- Explicar Alortimo y hacer pseudocódigo de cada sección - Explicar lógica del algoritmo - Explicar desarrolo de paquetes en R - Explicar bien la parte de los residuales y el algorimto backfitting, por que las f_j son arbitrarias y pueden interpolar a los residuales para hacer el ajuste. Esto también explica las β pues si se pueden capturar chigón los residuales con una sola dimensión, te vale verga la siguiente :). Yei bitches

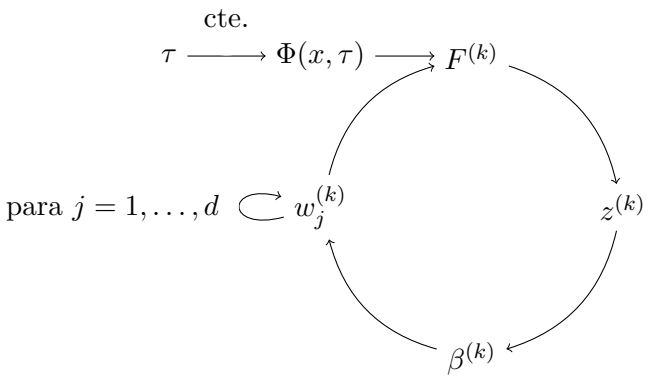

Figura 3.2: Esquema del algoritmo

Capítulo 4

Ejemplos y resultados

4.1. Análisis a fondo de un ejemplo sencillo

- Hacer todo el análisis de los grupos simulados con las normales bivariadas.

- Hacer varios tipos de polinomios con varias fronteras - Comparar contra un GLM probit normal

4.2. Otros resultados interesantes

- Dar gráficas y resultados de: la normal modificada, datos parabolicos, circulares y hopefully Ying-Yang.

4.3. Prueba con datos reales

Hacer prueba con datos SVSS y con datos de Hastie and Tibsh

Capítulo 5

Conclusiones

5.1. Consideraciones adicionales y posibles mejoras

- Y después? Mejoras al modelo - Seleccion de varaibles con SVSS - Automatización de selección de parámetros , $Jy\tau$ usando Mallik, M usando exploración previa de datos por dimensión.

- Problemas que tuve - Derivación bayesiana de las w0s

Listado de Assumptions - Assumptions: - 0. Y se distribuye bernoulli - 1. Existe y funciona f (aproximaxición) - 2. Las dimensioens son independientes entre si

5.2. Extensiones y alternativas al modelo

- Después: - Las redes neuronales se basan en coass parecidas. Ish hacer análogía - Cita de Articulo de ML para Cosas financieras

Apéndice A

Análisis Funcional

A.1. Convergencia del modelo

Habiendo entendido la estructura de las funciones que componen este proyecto, se busca demostrar (o dar una idea) por que se tiene una aproximación a f y a f_i's y no una igualdad escrita. Esto se logra, usando principios de álgebra lineal y análisis funcional. Esta discusión sigue los principios planteados por (Bergstrom 1985) en donde se demuestra el caso univariado y una pequeña discusión sobre los *Espacios de Hilbert de Kernel Reproductivo* (RKHS)[1]

Para entender el teorema de convergencia, necesitamos considerar los Espacios de Hilbert. Como introducción a estos, se usa el ejemplo clásico de plantearlo como una generalización del caso euclidiano. En este espacio euclidiano normal $\mathbb{R}^n$ podemos representar cualquier vector como una combinación lineal de un conjunto de bases ortogonales. En espacios abstractos de dimensión infinita, en particular en espacios de funciones, se busca representar *cualquier función*, en este caso f_i's, como una combinación lineal de bases. Los espacios de Hilbert dan idea de que los espacios vectoriales pueden ser suficientemente abstractos para que los vectores no sean simplemente listas ordenadas de números como lo son en $\mathbb{R}^n$. Los vectores pueden ser cualquier objeto en este caso, funciones. Una vez definido el espacio vectorial y sus objetos (que cumplan los correspondientes 8 axiomas presentados en el Apéndice D) se les puede denota de *producto interno* y por consecuente una *métrica* la cual induce una topología.

1. Reproducing Kernel Hilbert Space a falta de una mejor traducción.

Formalización matemática y teorema de convergencia

Se dice que $\mathcal{H}$ es un Espacio de Hilbert si $\mathcal{H}$ es un espacio vectorial con producto interno que también es un espacio metrico completo. (Rudin 1987).

Para hacer la prueba de convergencia, se considera únicamente a las funciones f_i y no a la f general. Se estudia en particular el espacio de Hilbert $\mathcal{H} = \mathbf{L}_2(\mathbf{R}, \mu)$ el *espacio de funciones integrables al cuadrado en* $\mathbf{R} = [a, b]$ con medida de Lesbegue ordinaria μ. Es decir:

$$f \in \mathcal{H} \iff \int_a^b f(x)^2 \, dx < \infty$$

Donde el producto punto es:

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx$$

y su norma inducida:

$$||f||_{\mathcal{H}} = \langle f, f \rangle = \int_a^b f(x)^2 \, dx$$

Antes de que se presente el teorema de Bergstrom, se tienen que presentar los tres supuestos fuertes que hace:

1. Las variables aleatorias $y_1, \ldots, y_n$ son generadas por la ecuación:

$$y_i = h(x_i) + e_i \quad \forall i = 1, \ldots, n$$

y los valores de las covariables están dados por las ecuaciones:

$$\begin{aligned} x_1 &= a + \frac{b-a}{2n} \\ x_{i+1} &= x_i + \frac{b-a}{n} \end{aligned} \tag{A.1}$$

Con a, b los extremos del intervalo $b > a$ y e_i son ruido aleatorio cumpliendo:

$$\begin{aligned} \mathbb{E}[e_i] &= 0, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i^2] &= \sigma^2, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i e_j] &= 0, \quad \forall i, j = 1, \ldots, n \; i \neq j \end{aligned} \tag{A.2}$$

2. La función $h(x)$, está definida en el intervalo cerrado $[a, b]$ es acotada y continua en casi todas partes.

3. El conjunto contables de funciones base $\Psi_1(x), \Psi_2(x), \ldots$ es un conjunto *ortonormal* en $\mathcal{H}$. Es decir, estas funciones cumplen:

$$\begin{aligned} \int_a^b \Psi_j^2(x)\, dx &= 1 \quad \forall j \\ \int_a^b \Psi_j(x)\Psi_k(x)\, dx &= 0 \quad \forall k, j \; k \neq j \end{aligned} \tag{A.3}$$

Estos supuestos son bastante fuertes y hay ciertas ecuaciones que no se cumplen *per se* en el modelo propuesto, sin embargo, vale la pena analizar el resultado pues lleva a cosas aún más interesantes. El primer supuesto es el más problemático. Aunque el modelo generador es idéntico a (2.10) y el ruido aleatorio es un supuesto aceptable (y común) el problema está en (A.2) pues para este trabajo no se asume que el estadista fija las x's sino que se asume una muestra aleatoria de datos. Sin embargo, en la prueba, este supuesto se usa para argumentar que, si $n \to \infty$, los datos cubren de manera homogénea todo el intervalo aproximando una integral. Aunque el propósito es completamente diferente que el de este trabajo en el que se busca suavizar sobre datos dispersos, se decide obviar por ahora el supuesto, en interés de presentar el teorema en su forma más rigurosa.

El segundo supuesto no es nada descabellado y se ha usado con anterioridad. Además, aún permite aproximar un número grande de funciones y es igual de flexible que el modelo anterior. Sin embargo, este supuesto si implica que $h \in \mathcal{H}$. Por lo que esta puede ser representada en su combinación lineal de bases funcionales, es decir:

$$h(x) = \sum_{i=1}^{\infty} w_i \Psi_i(x)$$

diferente a la expansión de bases en de la ecuación (2.9). Esto se deriva, de que ahora se busca encontrar

una representación *exacta* de h.[2] El último supuesto, implica la construcción de una base *ortonormal*. El trabajo original, sugiere que se puede lograr una base completa, aplicando un proceso de ortonormalización a las bases canónica polinomial $\{1, x, x^2, \ldots\}$. Por lo tanto, al menos de forma teórica, la base escogida para este trabajo definida en (2.17) también es ortonormalizable independientemente de la elección de J, N y K. Por lo que se cumple el supuesto. Finalmente:

Teorema 1 *Sea $\hat{h}_n^{N^*}(x)$ el estimador de h definido por:*

$$\hat{h}_n^{N^*}(x) = \hat{w}_1\Psi_1(x) + \ldots + \hat{w}_{N^*}\Psi_{N^*}(x) \tag{A.4}$$

que depende del número de datos n y el número de bases funcionales N^*. Y, con $\hat{w}_i \quad i = 1, \ldots, N^*$ los estimadores de mínimos cuadrados, es decir, los valores de $w_i \quad i = 1, \ldots, N^*$ que minimizan la expresión:

$$\sum_{i=1}^{n} [y_i - w_1\Psi_1(x_i) - \ldots - w_{N^*}\Psi_{N^*}(x_i)]^2 \tag{A.5}$$

Para todo $\epsilon > 0$, bajo los supuestos 1 a 3, existe un entero N^* y una función $n_\epsilon(N^*)$ tal que:

$$\mathbb{E}\left[\int_a^b \left(\hat{h}_n^{N^*}(x) - h(x)\right)^2 dx\right] < \epsilon \qquad \forall N \geq N^* \ y \ \forall n \geq n_\epsilon(N^*) \tag{A.6}$$

Detrás de toda esta verborrea y notación aparatosa, el corazón del teorema está en que, bajo ciertos supuestos, rigurosos más no descabellados, y en caso de existir una función h que genere los datos, está se puede aproximar a un grado de precisión arbitraria.

Aunque no es el objetivo del trabajo, vale la pena hacer una mención a lo sublime que es la demostración, pues utiliza conceptos de análisis, álgebra lineal, optimización y estadística. Los detalles y la utilización de los supuestos, son sutiles, sin embargo es una prueba rigurosa en todo el sentido de la palabra. Además, (Bergstrom 1985) va mucho más allá de únicamente demostrar la existencia. Se demuestran tres teoremas más, para dar estimaciones (bajo un supuesto adicional) de el tamaño de muestra necesario n y el número de bases N^* necesarias para la aproximación arbitraria de h. Sin embargo, este procedimiento, aunque elegante, no es nada practico pues depende de poder generar a merced las x's (con su correspondiente nivel y) aumentando y disminuyendo el tamaño de muestra. Además, se requiere ir generando todas las bases Ψ's de forma que sean ortonormales y sus correspondientes coeficientes de Fourier $w_j = \int_a^b h(x)\Psi_j(x)\,dx \quad \forall j$.

2. Antes se buscaba, más que aproximarla, suavizar los datos. Además, se usa el signo de igualdad para no introducir confusión en la exposición.

El resultado, es más bien teórico en el sentido de que, justifica que estos modelos tienen sentido. Para este trabajo, en especifico, da la *intuición* de que funcionará (obviando un poco el primer supuesto) pues, con una muestra suficientemente grande, las f_i's serán identificables y aunque sean aproximaciónes, estaremos captando los patrones subyacentes y con suerte, podremos hacer predicciones.

Se hace notar que el primer supuesto, da la intuición de *granularidad* de el intervalo $[a, b]$. Bajo la construcción de Bergstrom, tenemos las condiciones exactas para estimar $h(x)$. Sin embargo, en la practica es raro que el estadista tenga el control sobre las covariables y siempre tendremos datos aleatorios. A pesar de esto, al estar trabajando sobre intervalos cerrados, se puede suponer que $X \sim \mathrm{U}(a, b)$ de donde tenemos que si $n \to \infty$ cubrimos todo el intervalo y tenemos algo análogo al supuesto 1 y por ende, el teorema es válido.

Otro teorema de convergencia y RHKS

Este resultado de Bergstrom, es uno de los muchos teoremas que se probaron en la época para justificar la existencia de estos modelos. Otro resultado interesante del mismo año, viene dado por (Stone 1985). El, discute varios modelos posibles dada una estructura de datos practica. Plantea un GAM[3] tradicional $h^*(\mathbf{x}) = \sum h_i^*(x_i)$ con la restricción (2.12), y prueba que

$$\mathbb{E}[(h(x) - h^*(x))^2]$$

es mínimo, con la igualdad si h es en verdad aditiva. Además, los estimadores de h_j^* que da, son splines (los mismos que usan Hastie y Tibshirani).

Finalmente, se hace notar, que toda esta teoría y resultados, son casos específicos de una teoría más general y mucho más compleja, llamada *Espacios de Hilbert de Kernel Reproductivo* (RKHS) desarrollada en los años 90. Esta va mucho más allá del enfoque de este trabajo, sin embargo, vale la pena mencionarla pues engloba muchos de los modelos usados hoy en día en un marco matemático riguroso y basado en el análisis funcional. Muchas de las ideas presentadas en este trabajo, como lo son la regularización, las expansiones de bases y los espacios de Hilbert, se elevan varios niveles y llevan a resultados todavía más generales. Se recomienda el Capítulo 5.8 de (Hastie, Tibshirani y Friedman 2008) y el libro de (Wahba 1990) donde se discuten a detalle todas las consideraciones de los RKHS. Además, todos estos resultados

3. T. Hastie y R. Tibshirani publicaron preámbulos a los GAM antes del trabajo citado aquí de 1986. Además, de que la cercanía, Stone en Berkley y ellos en Stanford, ayudó a su colaboración.

son *deterministas* en sus parámetros, en contrapuesta de la filosofía bayesiana. Sin embargo, esto no le quita validez al modelo ni mucho menos a los resultados.

Apéndice B

Paquete en R. Desarrollo y Lista de Funciones

- Hacer un resumen de como se desarrolló todo el paquete. - Citar al buen H Wickham y su libro - Hacer un listado y un diagrama de las funciones que existen en el y como descargarlo - Mencionar los métodos S3 de `plot` y `summary`

Apéndice C

Notación

- **Datos:** $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$
- $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ variables de respuesta binarias.
- $\mathbf{x}_i \in X^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ covariables o regresores.
- $d \in \mathbb{N}$ dimensionalidad de mis regresores.
- $z_i \in \mathbb{R}$ variables latentes.
- $\Phi(\cdot) : \mathbb{R} \to (0, 1)$ la función de distrubución acomulada de una normal estandar.
- **Parámetros**: $\theta = (\sigma^2, \{\phi_i\}_{i=1}^d\}, \{\tau_j\}_{j=1}^J)$ donde,
- σ^2 es la varianza global de mis datos.
- $\{\phi_i\}_{i=0}^J$ los pesos de mi combinación lineal.
- $\{\tau_i\}$ los parámetros de mi función Ψ
- $f(\mathbf{x}_i)$ función de media para Z
- J orden de la expansión en bases.

Apéndice D

Definiciones

Espacio con producto interno Un espacio vectorial dotado de una estructura adicional llamada *producto interno*: $\langle \cdot, \cdot \rangle$, que asocia cada par de vectores con una cantidad escalar sobre F. Es decir, $\langle \cdot, \cdot \rangle : V \times V \to F$. Que cumple, para x, y, z vectores en V y a en F:

- $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- $\langle ax, y \rangle = a\langle x, y \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, x \rangle \geq 0$
- $\langle x, x \rangle = 0 \Leftrightarrow x = \mathbf{0}$

Espacio Funcional Un espacio funcional es un espacio vectorial cuyos elementos son funciones.

Espacio Metrico Un espacio métrico es un espacio donde la distancia (norma) inducida por el producto punto está definido sobre todos sus elementos. Norma: $||x|| = \sqrt{\langle x, x \rangle}$ la raiz no negativa del producto interno.

Espacio Métrico Completo Un espacio métrico es completo si todas las secuencias de Cauchy, convergen a puntos dentro del espacio.

Espacio Vectorial Un espacio vectorial sobre un campo F es un conjunto V, dotado de dos operaciones, *suma* $+$ y *multiplicación escalar* $\cdot$ que cumple los siguientes axiomas. Sean x, y, z vectores en V, y a, b escalares en F

1. $x + (y + z) = (x + y) + z$
2. $x + y = y + x$
3. $\exists 0 \in V$ tal que, $x + 0 = x$

4. $\forall x \in V \quad \exists -x \in V$ tal que, $x + (-x) = 0$
5. $a(bx) = (ab)x$
6. $\exists 1 \in F$ tal que, $1x = x$
7. $a(x + y) = ax + ay$
8. $(a + b)x = ax + bx$

Ortogonalidad Dos elementos son ortogonales (en cierto espacio) si $\langle x, y \rangle = 0$. Denotado $x \perp y$

Bibliografía

Albert, J.H., y S. Chib. 1993. "Bayesian analysis of binary and polychotomous response data". *Journal of the American Statistical Association:* 669-679.

Banerjee, Sudipto. 2008. *Bayesian Linear Model: Gory Details.* `http://www.biostat.umn.edu/~ph7440/pubh7440/BayesianLinearModelGoryDetails.pdf`. [En Linea; accedido el 10 de Mayo, 2018].

Barber, D. 2012. *Bayesian reasoning and machine learning.* Cambridge University Press.

Bergstrom, A. R. 1985. "The Estimation of Nonparametric Functions in a Hilbert Space". *Econometric Theory* 1 (01): 7-26.

Bernardo, José M, y Adrian FM Smith. 2001. *Bayesian theory.* IOP Publishing.

Bishop, C M. 2006. *Pattern Recognition and Machine Learning.* Springer.

Boor, C De. 1978. *A Practical Guide to Splines.* 346. New York, Springer-Verlag.

Box, George E. P. 1979. *Robustness in the Strategy of Scientific Model Building.* p. 74. May. RL Launer / GN Wilkinson.

Casella, George, y Edward I George. 1992. "Explaining the Gibbs sampler". *The American Statistician* 46 (3): 167-174.

Denison, DGT, BK Mallick y AFM Smith. 1998. "Automatic Bayesian curve fitting". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (2): 333-350.

Devroye, Luc. 1986. *Non-uniform random variate generation.* Volumen 4. Springer-Verlag New York.

Friedman, Jerome H. 1991. "Multivariate adaptive regression splines". *The Annals of Statistics:* 1-67.

Gelfand, A E, y A F M Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities". *Journal of the American Statistical Association* 85 (410): 398-409.

Härdle, Wolfgang, Marlene Müller, Stefan Sperlich y Axel Werwatz. 2004. *Nonparametric and semiparametric models.* Springer Verlag.

Hastie, T., R. Tibshirani y J. Friedman. 2008. *The elements of statistical learning,* volumen 1. Springer Series in Statistics.

Hastie, Trevor, y Robert Tibshirani. 1986. "Generalized additive models". *Statistical science:* 297-310.

James, Gareth, Daniela Witten, Trevor Hastie y Robert Tibshirani. 2013. *An introduction to statistical learning.* Springer.

MacCullagh, P., y J. A. Nelder. 1989. *Generalized Linear Models.* Chapman & Hall, London.

Mendoza, Manuel, y Pedro Regueiro. 2011. *Estadística Bayesiana.*

Robert, Christian P, y George Casella. 2004. *Monte Carlo statistical methods.* Volumen 319. Citeseer.

Ross, S.M. 2009. *Introduction to Probability Models.* Academic Press.

Rudin, Walter. 1987. *Real and complex analysis.* Tata McGraw-Hill Education.

Schoenberg, I J. 1964. "Spline Interpolation And The Higher Derivatives." *Proceedings of the National Academy of Sciences of the United States of America* 51, número 1 (): 24-8.

Stone, Charles J. 1985. "Additive regression and other nonparametric models". *The annals of Statistics:* 689-705.

Sundberg, Rolf. 2016. *Statistical Modelling by Exponential Families - Lecture Notes.* Stockholm University.

Wahba, G. 1990. *Spline Models for Observational Data.* Society for Industrial / Applied Mathematics.

Wasserman, Larry. 2007. *All of nonparametric statistics.* Springer.