COMPARACIÓN DE MODELOS DE REGRESIÓN SEMIPARAMÉTRICOS MIXTOS CON DISTRIBUCIÓN BETA

Por

Adriana Patricia Calvo Alfaro

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

2015

Aprobada por:

Edgardo Lorenzo González, Ph.D. Miembro, Comité Graduado	Fecha
Pedro A. Torres Saavedra, Ph.D. Miembro, Comité Graduado	Fecha
Raúl E. Macchiavelli, Ph.D. Presidente, Comité Graduado	Fecha
Orlando E. Ruiz Quiñones, Ph.D. Representante, Estudios Graduados	Fecha
Olgamary Rivera Marrero, Ph.D. Directora Interina del Departamento	Fecha

Abstract of Thesis Presented in Partial Fulfillment of
the Requirements for the Degree of Master of Science

COMPARISON OF SEMIPARAMETRIC MIXED REGRESSION MODELS WITH DISTRIBUTION BETA

By: Adriana P. Calvo Alfaro.

Chair: Raúl E. Macchiavelli
Major Department: Department of Mathematical Sciences

Some studies generate data that are rates, proportions or probabilities, continuously restricted in the range $(0, 1)$. The Beta regression allows to model this type of data and offers a number of advantages such as direct interpretation of results, since it is not necessary to use transformations, and the easy to model asymmetries because the distribution can take many forms if the scale and location parameters are changed.

The semiparametric regression models provide an effective tool for modeling complex data structures, since the main feature is that they do not assume a specific shape for the regression function, the real purpose is to build it by using the observations for a better use of the collected information. The implementation of smooth techniques like *splines* in the beta regression modeling and random effects allows having a better regression curve approximation which generates more precise results.

During the development of this work, three models of semiparametric regressions will be presented, and the matrices of design associated to the effects will be built by using *splines*. The conditional distribution from the variable of interest given the random effects is Beta and the random effects are assumed normally distributed; this Beta regression models are constructed by using the combination of *B-splines* in the fixed part; *B-splines*, *P-splines*

or Radial Smoothing in the random part.

By using simulations he chose the model that provides the better fit using Schwarz's Criteria and the better predictive model using the Mean Integrated Absolute Error (MIAE). The obtained results are used in a study of disease severity of banana crops in Puerto Rico.

Resumen de Tesis Presentada como Requisito Parcial de los
Requerimientos para el Grado de Maestría en Ciencias

COMPARACIÓN DE MODELOS DE REGRESIÓN SEMIPARAMÉTRICOS MIXTOS CON DISTRIBUCIÓN BETA

Por: Adriana P. Calvo Alfaro.

Consejero: Raúl E. Macchiavelli
Departamento: Departamento de Ciencias Matemáticas

Algunos estudios generan datos que son tasas, proporciones o probabilidades, restringidos de forma continua en el intervalo (0, 1). La regresión Beta permite modelar este tipo de datos y ofrece una serie de ventajas como la interpretación directa de los resultados, ya que no es necesario el uso de transformaciones, y la facilidad para modelar asimetrías puesto que la distribución puede tomar diversas formas si se varían sus parámetros de escala y localización.

Los modelos de regresión semiparamétricos proporcionan una herramienta eficaz en el modelado de datos con estructuras complejas, ya que su principal característica consiste en no asumir una forma específica para la función de regresión, si no construirla a través de las observaciones, permitiendo de esta forma un mayor aprovechamiento de la información. La incorporación de técnicas de suavizado como *splines* en el modelado con regresión beta y efectos aleatorios permiten realizar una mejor aproximación a la curva de regresión generando estimaciones más precisas.

En este trabajo se presentan tres modelos de regresión semiparámetricos cuyas matrices de diseño asociadas a los efectos se construyen mediante *splines*, la distribución condicional de la variable de interés dados los efectos aleatorios es Beta y se asume que los efectos aleatorios son normalmente distribuidos; estos modelos de regresión Beta se construyen mediante la combinación de *B-splines* en la parte fija; con *B-splines*, *P-splines* o Suavizado Radial en

la parte aleatoria.

Por medio de simulaciones se selecciona el modelo que mejor se ajuste a los datos según el Criterio de Schwarz y el modelo que mejor predice según el criterio de Error Absoluto Integrado Medio (MIAE). Finalmente se aplican los resultados obtenidos a un estudio de severidad de enfermedades en cultivos de banano en Puerto Rico.

Dedicado a:

A Dios, a mi esposo Jesús, a mis padres Xinia y Carlos. Por su amor, cooperación y apoyo en todo momento.

Agradecimientos

- Al Dr. Raúl E. Macchiavelli por su guía, paciencia y apoyo.
- Al Dr. Pedro Torres y al Dr. Edgardo Lorenzo por su colaboración.
- Al personal administrativo y docente del Departamento de Matemáticas.
- A todos mis compañeros estudiantes graduados, por su amistad.
- A mis amigos inolvidables Widad, Glorimar, Arlin, José y Paúl, por hacer este tiempo más agradable.

A todos Gracias!!!

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Justificación

Los modelos lineales se utilizan en situaciones en las cuales es posible explicar el comportamiento de la media de la variable dependiente o de interés a partir de una función lineal de variables independientes, ciertos parámetros a estimar y un error experimental.

Este tipo de modelos lineales son una herramienta ampliamente utilizada en diversos campos de investigación, como las ciencias médicas, agrícolas, sociales, educativas, entre otras. La necesidad de explicar un comportamiento o condición a través de una serie de variables que se relacionan mediante ciertos parámetros lineales con la variable de interés, nos da la justificación del uso de este tipo de modelos.

Aunque los modelos lineales son de gran utilidad en el análisis de datos, en algunas situaciones presentan deficiencias, ya que dependen de una serie de restricciones o supuestos que deben cumplirse para que el modelo genere resultados verídicos. Algunos de estos supuestos son: la normalidad, homocedasticidad e independencia entre las observaciones. En los casos donde se infringe alguna de estas pautas se puede recurrir a transformar las variables pero esto puede dificultar la interpretación de los resultados en términos de las variables originales. Otra alternativa es utilizar generalizaciones de estos modelos que permi-

tan estructuras de datos más complejas, es decir el incumplimiento de uno o varios de estos supuestos.

Para el análisis de datos con medidas repetidas se tiene una estructura de correlación, es decir con falta de independencia, esto se debe a que las observaciones provenientes de una misma unidad experimental tienden a ser más parecidos entre sí que entre las observaciones de unidades distintas, por lo que utilizar un modelo lineal ordinario no es apropiado. En estos casos suele ser de interés para el investigador analizar tanto la información de los individuos como entre los individuos y es por tal razón que se implementa un modelo que lo permita, como los modelos de efectos mixtos.

Según Pinheiro y Bates (2000), los modelos de efectos mixtos constituyen una herramienta flexible y potente para el análisis de datos agrupados, observaciones con medidas repetidas son un ejemplo de ello. La característica principal de estos modelos es que incluyen tanto efectos fijos como aleatorios.

Los efectos fijos son variables en las cuales el investigador ha incluido solo los niveles o tratamientos que son de su interés, mientras que los efectos se consideran aleatorios si solo es incluido en el modelo una representación o muestra aleatoria de los tratamientos (Durbán, 2010). En el caso de los datos de medidas repetidas las unidades experimentales suelen contener el efecto aleatorio del modelo, esto debido a que es de interés la población de los sujetos.

En el caso de los modelos mencionados hasta el momento es imprescindible la distribución normal de los datos, pero no en todas las ocasiones se cuenta con que los datos posean una distribución de este tipo. Las inferencias y estimaciones bajo los supuestos de normalidad, en un modelo cuyos datos no provienen realmente de esta distribución son completamente diferentes a las reales. Es por lo tanto que Nelder y Wedderburn (1989) presentan los Modelos Lineales Generalizados (MLG) como un nuevo método de modelación de datos, siempre y cuando el supuesto de independencia entre las observaciones se cumpla. Este tipo

de modelo permite ya no depender únicamente de que los datos provengan de una distribución normal, si no que ahora la distribución de los datos puede ser una distribución que pertenezca a la familia exponencial. Ejemplos de distribuciones provenientes de esta familia son: Binomial, Poisson, Gamma, Normal, Beta, entre otras.

Para modelar datos con distribución perteneciente a la familia exponencial pero que a su vez incumplen el supuesto de independencia, es posible generalizar los supuestos de este modelo permitiendo incluir la correlación. Esta generalización se conoce con el nombre de modelos lineales generalizados mixtos, los cuales vienen a ser un MLG al que se le incluyen efectos aleatorios.

Existen situaciones donde los datos que se desean modelar asumen valores en el intervalo (0,1). Un ejemplo de esto son el modelado de proporciones, porcentajes, etc. Para este tipo de datos la distribución beta es de gran utilidad, ya que es flexible con respecto a la diversidad de formas que puede tomar según los valores de sus parámetros. Ferrari y Cribari-Neto (2004) proponen un modelo de regresión beta y una re-parametrización de la función de densidad de la distribución, de forma que esta se pueda expresar en términos de la media y un parámetro de dispersión.

Por último, autores como Eilers y Marx (1996), Durbán (2009) y Ruppert et al. (2003) exponen técnicas para el suavizado de curvas de forma que es posible realizar una construcción que permita modelar de una mejor manera los datos, mejorando las estimaciones. En la actualidad un método que permite construir la función de regresión, mediante las observaciones de las covariables, es la técnica de regresión semiparamétrica utilizando *splines*. Estos se definen como una construcción formada de polinomios conectados entre sí (Eilers y Marx, 1996).

La motivación de esta investigación radica en proponer un modelo de regresión semiparamétrico cuyas matrices de diseño asociadas a los efectos fijos y aleatorios se construyan

mediante la combinación de tipos de *splines*, además la distribución condicional de la variable de interés dados los efectos aleatorios será beta, asumiendo que los efectos aleatorios son normalmente distribuidos.

1.2. Objetivos

Objetivo General

Examinar diferentes modelos mixtos de regresión beta con *Splines*.

Objetivos Especificos

- Comparar modelos mixtos de regresión Beta construidos mediante la combinación de *B-splines* en la parte fija del modelo con *B-splines*, *P-splines* o Suavizado Radial en la parte aleatoria para el modelado de curvas con intercepto y pendiente aleatorios.
- Seleccionar mediante simulaciones el modelo que mejor se ajuste a los datos y el que mejor prediga según algunos criterios de selección.
- Aplicar los resultados obtenidos a estudios de severidad de enfermedades en cultivos de banano en Puerto Rico.

1.3. Organización del Trabajo

En el Capítulo 2 se presenta una recopilación de los temas que constituyen la base de los modelos lineales generalizados mixtos, así como una breve introducción al estudio de datos medidos como proporciones en el intervalo $(0, 1)$ y el modelo de regresión Beta. Por último se presentan algunas nociones básicas sobre las técnicas de suavizado con *splines* como los *B-spline*, *P-splines* y Suavizado Radial.

En el Capítulo 3 se describe de forma detallada el modelo de regresión Beta lineal generalizado propuesto por Ferrari y Cribari-Neto (2004) y la reparametrización de la función de densidad. De igual manera se describen técnicas semiparámetricas para el suavizado de curvas. Por último se describen diferentes modelos mixtos de regresión Beta semiparamétricos construidos mediante la combinación de *B-splines* con *B-splines*, *P-splines* o Suavizado Radial para el modelado de datos con medidas repetidas cuya distribución condicional se asume Beta y la distribución de los efectos aleatorios se asume Normal.

En el Capítulo 4 se realizan simulaciones para estudiar, mediante algunos criterios de selección, cuál de los modelos propuestos es el que mejor se ajusta a los datos y cuál posee mejores carácteristicas para predicir. Además se analizan algunas propiedades importantes del modelo mixto de regresión Beta con *splines* seleccionado. En el Capítulo 5 se aplican los resultados obtenidos a un estudio de severidad de enfermedades en cultivos de banano en Puerto Rico. Por último en el Capítulo 6 se presentan las conclusiones generales y posibles trabajos futuros.

1.4. Implementación Computacional

Para la implementación computacional se utilizará el procedimiento GLIMMIX del software estadístico SAS 9.3, el cual permite ajustar el modelo lineal generalizado mixto condicional y el modelo marginal con estructura de correlación. A través de este procedimiento se pueden realizar las estimaciones de parámetros, efectos, medidas de ajuste; aplicar diferentes estrategias de optimización como por ejemplo Newton Raphson, Gradiente conjugado, Quasi-Newton; seleccionar el método de aproximación integral por cuadratura Gaussiana o Método de Laplace; establecer estructuras flexibles de covarianza para los efectos aleatorios e implementar métodos se suavizado como *Splines*, entre otras funciones (SAS Institute, 2011).

Capítulo 2

Revisión de Literatura

2.1. Modelos Lineales Mixtos

Los modelos estadísticos son una simple abstracción de la realidad ya que proporcionan una aproximación de un fenómeno relativamente más complejo (Gbur et al., 2010). Los modelos lineales (ML) representan una de las principales técnicas de modelado de datos. Su objetivo principal es analizar la relación entre la media de la variable de interés y un conjunto de covariables con el fin de proporcionar información afín a la realidad.

Su aplicación posee una gran trayectoria a lo largo de la historia ya que surge desde finales de los años 1800 cuando Francis Galton investigaba la relación entre los pesos de los padres y sus hijos (Casella y Berger, 2002). Diversas áreas de investigación como las ciencias biológicas, físicas y sociales hacen uso de técnicas estadísticas basadas en ML para desarrollar las etapas de planificación y análisis de resultados en sus investigaciones (Rencher y Schaalje, 2008).

Sin embargo, existen situaciones con estructuras complejas en las cuales utilizar ML ordinarios podrían generar resultados poco verídicos. Un ejemplo de esto son los estudios con datos correlacionados o desbalanceados ya que infringen los supuestos básicos de independencia. De manera similar si los datos no siguen una distribución normal y no se cuenta con

muestras lo suficientemente grandes no es apropiado hacer uso de las propiedades asintóticas de los estimadores y por lo tanto no es viable el uso de ML.

Los modelos lineales mixtos (MLM) son una generalización de los modelos lineales en la cual, los datos siguen una distribución normal pero puede que no sean independientes, es decir, este tipo de modelo permite estudiar datos que están correlacionados. Un ejemplo de ello son los datos con medidas repetidas o datos longitudinales.

Los MLM reciben ese nombre ya que contienen efectos fijos, como los modelos lineales ordinarios, pero también poseen efectos aleatorios. Los efectos fijos son parámetros asociados con una población entera o con ciertos niveles repetibles de factores experimentales, mientras que los efectos aleatorios se asocian con las unidades experimentales individuales extraídas al azar de una población (Durbán, 2010, Pinheiro y Bates, 2000, Ruppert et al., 2003)

En el contexto de un estudio longitudinal con N sujetos de estudio cada uno con un conjunto de observaciones n_i, $i = 1, \ldots, N$. Se define el MLM de la siguiente manera:

$$\begin{aligned} Y_i &= X_i\beta + Z_i u_i + e_i, \\ e_i &\backsim N(0, R_i) \\ u_i &\backsim N(0, G) \end{aligned} \tag{2.1}$$

donde $Y_i = (Y_{i1}, \ldots, Y_{in_i})$ y e_i son vectores de $(n_i \times 1)$ que contienen respectivamente la secuencia de medidas y errores del i-ésimo sujeto. La matriz X_i es la matriz de diseño de $(n_i \times p)$ con p covariables, β es el vector de dimensión p de parámetros de efectos fijos (parámetros de la población) asociados a las covariables, la matriz Z_i de dimensión $(n_i \times q)$ corresponde a la matriz de diseño de q efectos aleatorios (parámetros individuales) para cada i-ésimo sujeto.

La media condicional Y_i dado u_i describe la media específica del i-ésimo sujeto

$$E(Y_i \mid u_i) \quad = \quad X_i\beta + Z_i u_i \tag{2.2}$$

Mientras que la media marginal o media de la población de Y_i cuando se promedian sobre la distribución de los efectos aleatorios u_i es

$$\begin{aligned} E(Y_i) &= E\{E(Y_i \mid u_i)\} \\ &= E(X_i\beta + Z_i u_i) \\ &= X_i\beta + Z_i E(u_i) \\ &= X_i\beta. \end{aligned}$$

Una característica importante de este modelo es que permite considerar la correlación entre observaciones de un mismo sujeto, esto a través la matriz de covarianzas

$$\begin{aligned} cov(Y_i \mid u_i) &= cov(e_i) \\ &= R_i. \end{aligned}$$

Mientras que la covarianza marginal de los Y_i (promediada sobre la distribución de u_i) es dada por:

$$\begin{aligned} V_i &= cov(Y_i) \\ &= cov(X_i\beta + Z_i u_i + e_i) \\ &= cov(Z_i u_i) + cov(e_i) \\ &= Z_i G Z_i^t + R_i. \end{aligned}$$

Por lo tanto, en el modelo de efectos mixtos, el vector de parámetros β se supone que es el mismo para todos los individuos y tienen su interpretación en términos del promedio de la población. En cambio el vector de efectos aleatorios u_i cuando se combina con los efectos fijos correspondientes, se compone de los coeficientes de regresión de individuos específicos. Por lo que la respuesta media del i-ésimo individuo es dada por la expresión (2.2).

Resumiendo la distribución marginal de Y_i es:

$$Y_i \sim N(X_i\beta, V_i). \tag{2.3}$$

2.1.1. Estimación de Parámetros

I. Mínimos Cuadrados Generalizados

Este método de estimación es utilizado cuando el modelo incluye una estructura de correlación y consiste en hallar el valor $\hat{\beta}$ que minimiza la ecuación

$$S = (Y - X\beta)^t V^{-1}(Y - X\beta) \tag{2.4}$$

donde:
$Y = (Y_1, \ldots, Y_N)$: El conjuntos de observaciones de todos los sujetos de estudio
$X = (X_1, \ldots, X_N)$: La matriz de diseño asociada a los efectos fijos, que contiene la información para todos los sujetos.
$V = (V_1, \ldots, V_N)$: Es la matriz de covarianza marginal de Y.
Para obtener un estimador de β en el MLM se puede reescribir la ecuación del modelo de la siguiente manera:

$$Y = X\beta + \varepsilon^* \text{ donde } \varepsilon^* = Zu + \varepsilon.$$

Minimizando S se obtiene:

$$\widehat{\beta} = (X^t V^{-1} X) X^t V^{-1} Y, \tag{2.5}$$

Si los componentes de la matriz de varianza-covarianza V son conocidos el estimador $\hat{\beta}$ es el mejor estimador lineal insesgado y coincide con el estimador máximo verosímil. Sin embargo en la mayoría de las ocasiones la matriz V no es conocida por lo que las estimaciones deben realizarse usando $\hat{V}$.

II. Máxima Verosimilitud y Ecuaciones de Henderson

La estimación por Máxima Verosimilitud es un método de optimización cuyo supuesto base es que se conoce la distribución de probabilidad de las observaciones.

Para simplificar, considere la estimación de un solo parámetro θ. Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias independientes e idénticamente distribuidas con función de densidad $f(y_i; \theta)$. Se define la función verosimitud como:

$$L(\theta) = f(y_i, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(y_i, \theta). \tag{2.6}$$

En el caso de distribuciones discretas es posible interpretar esta función como la probabilidad de observar los datos que se tienen si los parámetros del modelo fuesen los propuestos, por tal motivo si para un valor estimado $\widetilde{\theta}$ la verosimilitud es "pequeña", es poco probable que ese valor sea el que generó los datos observados, caso contrario si la verosimilitud es un valor alto; así que el estimador óptimo es aquel que maximice $L(\theta)$.

Es decir, el estimador máximo verosímil $\widetilde{\theta}$ es aquel que satisface las siguientes condiciones:

- Condición de primer orden:

$$\left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\widetilde{\theta}} = 0$$

  Una manera más práctica y equivalente es hacer la estimación a partir del logaritmo de la función, $l(\theta) = \log[L(\theta)]$:

$$\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\widetilde{\theta}} = 0$$

- Condición de segundo orden:

$$\left.\frac{\partial^2 l(\theta)}{\partial^2 \theta^2}\right|_{\theta=\widetilde{\theta}} < 0$$

Los estimadores máximo verosímil son buenos estimadores si se utilizan muestras grandes. Un detalle importante es que estas estimaciones dependen de los supuestos sobre la distribución por lo tanto para buenas estimaciones se debe estar seguro de conocerla.

En el contexto de los modelos mixtos hay varias maneras de obtener predictores de u que tengan la propiedad de tener menor error cuadrático medio de predicción. Una de ellas es mediante lo que se llaman ecuaciones de modelos mixtos de Henderson (Durbán, 2010). Este método permite obtender el mejor estimador lineal insesgado de $X\beta$ y el mejor predictor lineal insesgado de u los cuales se obtienen maximizando la densidad conjunta $f(y,u)$ a través de Máxima Verosimilitud. Por lo tanto

$$f(y,u) \;=\; f(y \mid u)f(u). \tag{2.7}$$

En términos de verosimilitud

$$log(L) \;\propto\; -\frac{1}{2}\left[\log|R| + \log|G| + (Y - X\beta - Zu)'R^{-1}(Y - X\beta - Zu) + u'G^{-1}u\right] \tag{2.8}$$

Derivando con respecto a β y u obtenemos las siguientes ecuaciones:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix},$$

donde las soluciones de estas ecuaciones y por lo tanto los estimadores por maxíma verosimilitud son:

$$\widehat{\beta} \;=\; (X^tV^{-1}X)^{-1}X^tV^{-1}Y, \tag{2.9}$$

$$\widehat{u} \;=\; GZ^tV^{-1}(Y - X\widehat{\beta}). \tag{2.10}$$

III. Máxima Verosimilitud Restringida

La estimación de máxima verosimilitud restringida fue sugerida por Patterson y Thompson en el año de 1971 en el contexto de la estimación de componentes de la varianza. El método consiste en realizar las estimaciones basandose en los residuales obtenidos después de estimar los efectos fijos del modelo, es decir, los valores $Y - X\widehat{\beta}$.

El logaritmo de la función máximo verosímil restringida, para la estimación de los parámetros de varianza del modelo mixto es:

$$l_R(V) = l_P(V) - \frac{1}{2}\log|X^tV^{-1}X| \tag{2.11}$$

donde

$$l_P(V) = \frac{1}{2}\left[\log|V| + y^tV^{-1}(I - X(X^tV^{-1}X)^{-1}X^tV^{-1})y\right] \tag{2.12}$$

corresponde al logaritmo de la función máximo verosímil para la estimación de los parámetros de varianza.

Una de las principales ventajas de la estimación por máxima verosimilitud restringida es que tiene en cuenta los grados de libertad utilizados para estimar los efectos fijos en el modelo. Para tamaños de muestra pequeños proporciona mejores estimaciones que el método de máxima verosimilitud, mientras que si el tamaño de la muestra es grande no habrá prácticamente ninguna diferencia entre los métodos (Durbán, 2010).

Es importante tener presente que los modelos cuya parte fija es diferente no se deben comparar mediante métodos que dependan de la verosimilitud si las estimaciones se realizan a través de máxima verosimilitud restringida, ejemplos de estos métodos son las pruebas de razón de verosimilitud, BIC, AIC, entre otras (Pinheiro y Bates, 2000).

2.2. Modelos Lineales Generalizados

Los modelos lineales generalizados (MLG) fueron introducidos por Nelder y Wedderburn en el año 1972, como una alternativa para el tratamiento de datos independientes que provienen de otras distribuciones diferentes a la normal y cuyas varianzas no son constantes. La motivación para la implementación de este tipo de modelos es extender la teoría de los ML a distribuciones que sean una familia exponencial.

2.2.1. La Familia Exponencial

Considere la variable aleatoria y cuya distribución de probabilidad depende de los parámetros θ y ϕ. Una distribución de probabilidad es una familia exponencial si su función de densidad puede ser reescrita de la siguiente manera:

$$f(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right\}. \tag{2.13}$$

donde θ se conoce como el parámetro natural y ϕ es el parámetro de dispersión; $a(\cdot), b(\cdot)$ y $c(\cdot)$ son funciones conocidas que se definen según la distribución de la variable de interés.

Se establecen las siguientes características de las distribuciones que pertenecen a esta familia y pueden ser utilizadas para determinar la media y varianza de la variable y.

$$\mu = E(y) = b'(\theta), \tag{2.14}$$

$$Var(y) = b''(\theta)a(\phi), \tag{2.15}$$

donde $b^{'}$ y $b^{''}$ denotan la primera y segunda derivada de la función b con respecto a θ

2.2.2. Definición

Los MLG son una generalización de los ML en la cual la relación lineal entre la media de la variable repuesta y las covariables se obtiene a través de una función de $E(y)$, es decir

$$g[E(y)] = g(\mu) = \eta = X\beta, \tag{2.16}$$

donde η es el predictor lineal. La función g es monótona y diferenciable y se conoce como la función de enlace.

Por lo tanto los modelos lineales generalizados constan de tres componentes:

- *Componente aleatorio*: La variable de interés y la cual pertenece a una familia exponencial. Algunos ejemplos de distribuciones que son una familia exponencial son la Normal, Binomial, Poisson, Gamma, Binomial Negativa, entre otras.

- *Componente sistemático*: Las variables predictoras X, relacionadas linealmente con los β.

- *Función de enlace*: Relaciona la media $E(y)$ con la componente sistemática. Por ejemplo en el caso del modelo lineal estándar, $\mu = \eta$, por lo tanto la función de enlace es la identidad.

2.2.3. Estimación de Parámetros

El método utilizado para estimar el parámetro θ en un modelo lineal generalizado corresponde al método de máxima verosimilitud. Por motivos de simplicidad se suele considerar al parámetro ϕ como un parámetro de perturbación o ruido, esto con el objetivo de facilitar las estimaciones del parámetro de interés θ. Dado el vector de observaciones $y = (y_1, \ldots, y_n)'$, el logaritmo de la función de verosimilitud es:

$$l(\theta, \phi; y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi). \tag{2.17}$$

Sin embargo como $\theta_i = \eta_i = x_i'\beta$, es necesario estimar los parámetros β. El vector de primeras derivadas de la función de verosimilitud recibe el nombre de función *score* y se representa mediante la siguiente ecuación:

$$S(\beta) = \sum_{i=1}^{n} \frac{\partial l}{\partial \beta} \frac{(y_i - \mu_i)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_i = 0 \tag{2.18}$$

Usualmente, para encontrar las soluciones de estas ecuaciones se requiere la implementación de metodos iterativos como Newtow Raphson o Fisher-Scoring (McCullagh y Nelder, 1989).

2.3. Distribución Beta

La distribución Beta es una distribución de probabilidad continua cuyo dominio de valores se encuentra en el intervalo $(0, 1)$. Su función de densidad depende de los parámetros (p, q) y es dada por:

$$f(y; p, q) = \frac{y^{p-1}(1-y)^{q-1}}{B(p,q)}, \qquad 0 < y < 1, \qquad p > 0, \quad q > 0, \tag{2.19}$$

donde la $B(p, q)$ representa la función Beta y es definida por

$$B(p,q) = \int_0^1 y^{p-1}(1-y)^{q-1} dy = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},$$

y $\Gamma(x)$ representa la función Gamma definida como

$$\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy.$$

Una de las principales características de esta distribución es su capacidad de tomar diferentes formas según se varian sus parámetros, por lo que es versátil a la hora de modelar datos que se ajusten a su dominio. En la Figura 2.1 se muestra ejemplos de funciones de densidad Beta para diferentes combinaciones de valores en sus parámetros.

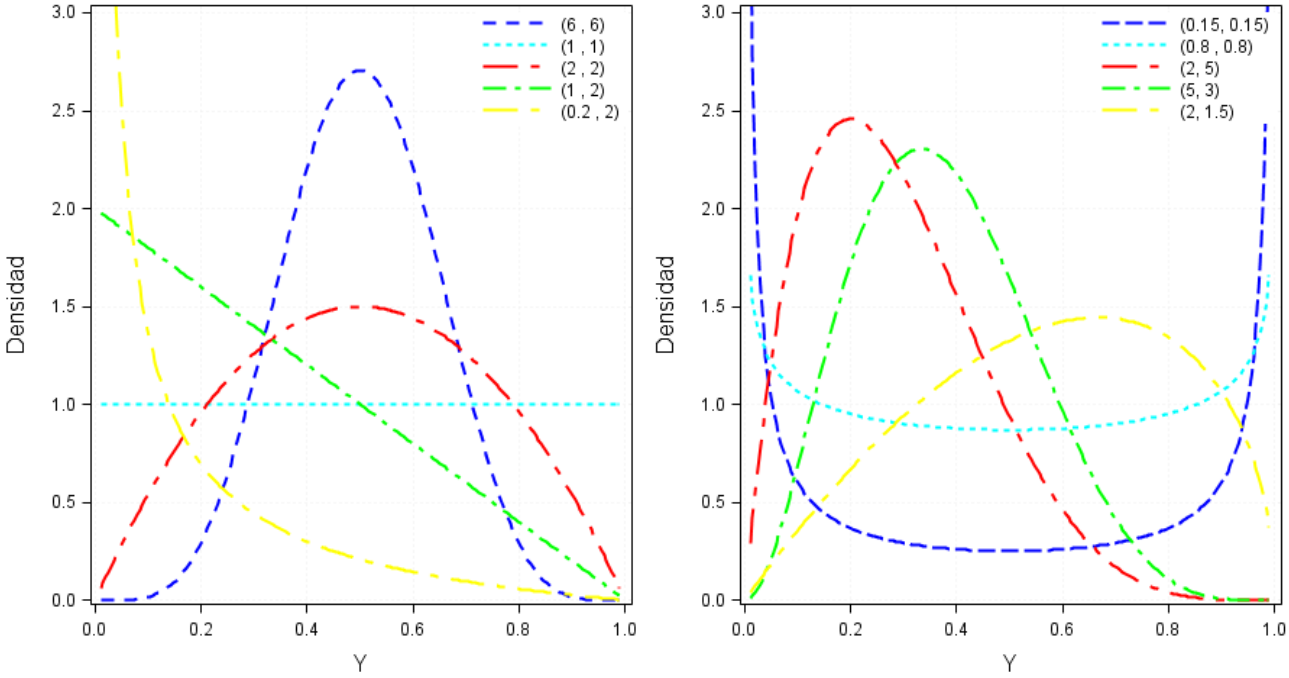

Figura 2.1: Función de densidad Beta para diferentes combinaciones de (p, q)

La media y varianza de la distribución se obtienen respectivamente a través las ecuaciones

$$E(y) = \frac{p}{p+q}, \tag{2.20}$$

$$Var(y) = \frac{p\,q}{(p+q)^2(p+q+1)}. \tag{2.21}$$

La distribución Beta es utilizada para modelar datos continuos que se restringen en el intervalo $(0, 1)$ como por ejemplo tasas, proporciones o probabilidades. Con el propósito de superar una serie de inconvenientes que surgían a la hora de modelar este tipo de observaciones mediante modelos lineales Ferrari y Cribari-Neto (2004) proponen un modelo de regresión que se basa en el supuesto de que la variable de interés sigue una distribución Beta. Este modelo se desarrolla en detalle en la Sección 3.1.

2.4. Modelos Lineales Generalizados Mixtos

Al igual que en el caso de los modelos lineales, a veces es útil la incorporación de efectos aleatorios en un modelo lineal generalizado a este tipo de modelos se les conoce como modelos lineales generalizados mixtos (MLGM) (Ruppert et al., 2003, Fitzmaurice et al., 2004, McCulloch y Searle, 2001).

Como en el caso de los modelos mixtos, en el MLGM surge el interés de analizar la información tanto al nivel de los individuos como de la población en general por lo que se estudian los modelos condicional y marginal. En el modelo condicional se define el predictor lineal y la media de la siguiente manera :

$$\begin{aligned} \eta &= g[E(Y \mid u)] &&= X\beta + Zu \\ \Rightarrow\ & E(Y \mid u) &&= g^{-1}[X\beta + Zu]. \end{aligned}$$

donde la distribución de la variable $(Y \mid u)$ describe a los sujetos de forma específica y pertenece a la familia exponencial, además se asume que los efectos aleatorios $u \sim N(0, G)$

En el caso del modelo marginal, el cual es utilizado para describir la población de individuos, debe estimarse a partir de la siguiente relación entre la distribución condicional y marginal.

$$f(Y) = \int_U f(Y \mid u)\, f(u)\, du$$

De acuerdo a lo anterior se puede calcular la media del modelo marginal a partir de:

$$E(Y) = E[E(Y \mid u)] = \int_u g^{-1}[X\beta + Zu]\, f(u)\, du. \tag{2.22}$$

Para el MLGM la media del modelo marginal no se relacionada con los regresores de la misma forma funcional como ocurre con los MLM, si la función enlace no es la identidad. Es por este motivo que las estimaciones a nivel del modelo marginal se dificultan requiriendo en la mayoría de los casos técnicas de integración numérica como Cuadratura Gaussiana o de Gauss-Hermite.

2.4.1. Estimación de Parámetros

Las estimación de los parámetros β y la matriz G del modelo se realizan mediante máxima verosimilitud. La función máximo verosímil marginal está dada por:

$$L(\beta, \phi, G) = \prod_{i=1}^{N} \int f(Y_i \mid u_i)\, f(u_i)\, du_i. \tag{2.23}$$

Métodos de integración numérica son requeridos y en el proceso de estimación. Si el número de efectos aleatorios es elevado es probable que se presentan dificultades computacionales. Una práctica alternativa es el método de cuasi-verosimilitud, el cual consiste en definir una aproximación de la función de verosimilitud especificando únicamente la relación entre media y varianza.

Las ecuaciones de estimación se obtienen a partir de la expresión:

$$S(\underline{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right) V_i^{-1} (Y_i - \mu_i) = 0, \tag{2.24}$$

donde $V_i = Cov(Y_i)$.

2.5. *Splines*

La construcción de modelos estadísticos que analicen la relación entre variables es una de las principales herramientas en el análisis estadístico. Para esto es necesario hacer uso de técnicas flexibles, potentes y eficientes. Una de estas técnicas es el modelado con *splines*, la cual es una práctica que ha tomado mucho auge en los últimos años, por sus facilidades de implementación computacional. (Wang, 2011, Durbán, 2009, Eilers y Marx, 1996, Ruppert et al., 2003).

En el caso bidimensional, dada una serie de observaciones (x_i, y_i) para $i = 1, \ldots, n$ se pretende ajustar un modelo de regresión de la forma:

$$E[y_i] = f(x_i)$$

donde $f(x_i)$ es una función suave y se conoce como la función de regresión. En el marco del análisis de regresión paramétrico su forma se asume como conocida por lo tanto lo único desconocido en el modelo son los parámetros asociados a la función. Por el contrario en el análisis semiparamétrico no se asume la forma de la curva, si no que se construye a partir de las observaciones. (Ruppert et al., 2003).

Por lo tanto en el enfoque semiparamétrico se aproxima a la función de regresión a través de la expresión

$$f(x) \approx \sum_{j=1}^{p} \beta_j B_j(x) \tag{2.25}$$

siendo β_j el parámetros de regresión asociado a la j-ésima base y $B_j(x)$ es una base para la curva de regresión, construida a partir de x mediante *splines*.

El objetivo es construir funciones base para las columnas de las matriz de diseño X, a partir de los valores observados de las covariables, para luego ajustar el modelo de regresión.

2.5.1. Definición

Un spline es una construcción formada por trozos de polinomios conectados en puntos, llamados nodos, y cumplen con las condiciones de continuidad y suavidad.
Considere $\{k_1, k_2, \ldots, k_K\}$ una secuencia de nodos y sea la función $S : [a, b] \to \mathbb{R}$ definida por:

$$S(x) = \begin{cases} P_0(x) & x < k_1 \\ P_i(x) & k_i \leq x < k_{i+1} \\ P_K(x) & x \geq k_K \end{cases}$$

$S(x)$ es una función de spline y cumple con las siguientes condiciones:

- $a < k_1, \ldots, k_K < b \qquad (k_0 = a, k_{K+1} = b)$
- En cada intervalo $[k_i, k_{i+1}]$ para $j = 0, \ldots, K$, $P_i(x)$ es un polinomio de grado d.
- La función $S(x)$ posee $d - 1$ derivadas continuas en $[a, b]$

Además el conjunto de los *splines* de grado d con nodos en $k_1, k_2, \ldots, k_K$ definido en $[a, b]$ es un espacio vectorial de dimensión $d + K + 1$.

2.5.2. *B-Splines*

Una de las formas para calcular la base para la curva de regresión mostrada en (2.25) es mediante el uso de las bases de *B-splines*. Estas son una construcción formada por trozos de polinomios conectados entre sí (Eilers y Marx, 1996). En la Figura 2.2 se muestra ejemplos de bases de B-spline.

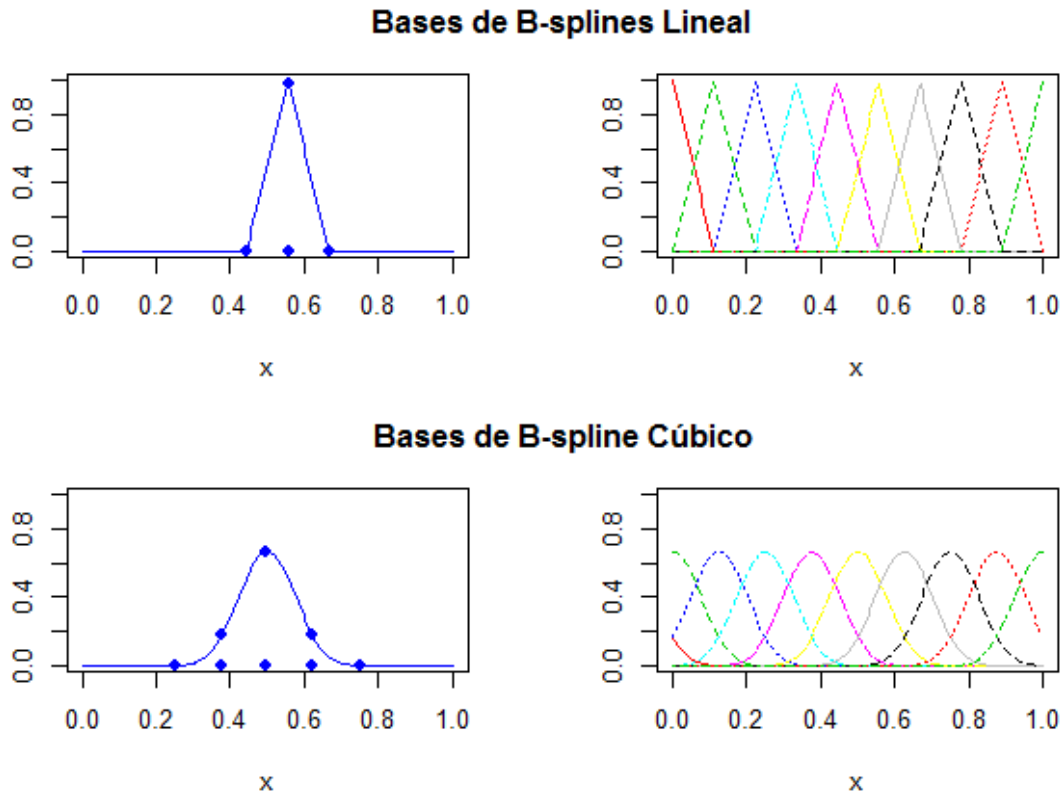

Figura 2.2: Bases de *B-Splines* de orden 1 y 3.

Las bases de B-spline de grado p construidas sobre $t_0, \ldots, t_K$ nodos, se definen de forma recursiva a través del siguiente algoritmo:

$$B_{i,0}(t) = \begin{cases} 1 & t_{i-1} \leq t < t_i \\ 0 & \text{en otro caso} \end{cases}$$

$$B_{i,p}(t) = \frac{t - t_{i-1}}{t_{i+p-1} - t_{i-1}} B_{i,p-1}(t) + \frac{t_{i+p} - t}{t_{i+p} - t_i} B_{i+1,p-1}(t) \tag{2.26}$$

Estas bases pueden ser facílmente calculadas a través del algoritmo de De Boor.

Las propiedades básicas de un B-spline de orden p son:

- Se construye a partir $p + 1$ trozos de polinomio de orden p.
- Se unen en p nodos internos.
- En los puntos de unión, las derivadas hasta el orden $p - 1$ son continuas.
- El B-spline es positivo en el dominio expandido por $p + 2$ nodos y 0 en el resto.
- Excepto en los extremos, se solapa con $2p$ trozos de polinomios de sus vecinos.
- Para cada valor de x, $p + 1$ *B-splines* no son nulos.

2.5.3. *Splines* Penalizados (*P-splines*)

Los *P-splines* fueron introducidos por Eilers y Marx en el año de 1996 con el objetivo de solucionar algunos problemas ocasionados por la dificultad de seleccionar el número óptimo y localización de los nodos, dado que al utilizar pocos en la construcción de los *splines* se obtiene un control limitado de la suavidad y el ajuste, mientras que al seleccionar un número excesivo se incurre en problemas de sobreajuste. La solución propuesta por los autores consiste en utilizar una base para la regresión e introducir una penalización basada en diferencias entre coeficientes adyacentes (Durbán, 2009).

Considere las observaciones (x_i, y_i) para $i = 1, \ldots, n$. La función objetivo a minimizar por mínimos cuadrados es dada por:

$$S = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \beta_j B_j(x_i) \right\}^2 + \lambda \sum_{j=k+1}^{p} (\Delta^k \beta_j)^2 \tag{2.27}$$

Minimizando S se obtiene:

$$B^t y = (B^t B + \lambda D_k^t D_k)\beta \tag{2.28}$$

donde D_k es la representación de la matriz del operador de diferencia Δ^k y los elementos de B son $b_{ij} = B_j(x_i)$. En el caso que λ sea cero se obtienen las ecuaciones normales estándar con bases de *splines*.

Según Eilers y Marx (1996) y Durbán (2009), algunas de las ventajas obtenidas a través de este enfoque son:

- La dimensión del problema se reduce al número de bases utilizadas en vez del número de datos.
- La penalización minimiza la importancia de la selección óptima del número y la localización de los nodos.
- Se cuenta con el parámetro de control sobre la suavidad del ajuste λ.
- Una penalización de este tipo preserva momentos (Media y Varianza).
- La penalidad de diferencia se incorpora con facilidad a las ecuaciones de regresión.
- Existe una correspondencia entre los *P-splines* y el mejor predictor lineal insesgado en un modelo mixto, por lo que es posible aplicar la metodología existente para estos modelos así como el software específico.

Selección del Parámetro de suavizado

El parámetro de suavizados de los *splines* penalizados, como en cualquier método de suavizado, se encarga de controlar qué tan suave será la curva. En este contexto se encarga

de penalizar los coeficientes que se encuentran muy separados, por lo que cuanto mayor sea el valor de λ más se aproximan los coeficientes a cero generando un ajuste casi polinómico, mientras que si el valor λ se acerca mucho a cero las estimaciones serán similares a las obtenidas por mínimos cuadrados ordinarios (Durbán, 2009).

Una forma eficaz y rápidad de seleccionar el valor de este parámetro, y que a su vez se aplica para cualquier método, es el uso de criterios de selección de modelos como el Criterio de Información Bayesiano (BIC), el cual se describe en la Sección 2.6; el Criterio de Información de Akaike (AIC), Validación Cruzada Generalizada (GCV), entre otros.

La expresión para el AIC es dada por:

$$AIC = 2\log\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) - 2\log(n) + 2\log(traza(H)) \tag{2.29}$$

donde

$$H = B(B^tB + \lambda D^tD)^{-1}\beta^t$$

Una ventaja de los *splines* penalizados sobre otros métodos de suavizado, es que el cálculo de la $traza(H)$ es más rápido (Durbán, 2009)

2.5.4. Suavizado Radial

Funciones de Base Radial

Las funciones de base radial son definidas para polinomios de grado impar p. Considere la secuencia de nodos $\{k_1, \ldots, k_K\}$ entonces se definen como:

$$1, x, \ldots, x^p, |x - k_1|^p, \ldots, |x - k_k|^p \tag{2.30}$$

con $|x - k_k|^p = \phi(|x - k_k|)$ donde $\phi(u) = u^p$.

Este tipo de bases poseen propiedades similares a las bases de *B-splines* con la ventaja

que solo dependen de la distancia $|x - k_i|$ y la función ϕ. Además son simétricas alrededor de los nodos k_i con $i = 1, \ldots, K$. Esta propiedad se extiende a variables predictoras de dimensiones superiores, por lo tanto si $x \in \mathbb{R}^n$ y $k_1, \ldots, k_K$ son nodos en el mismo espacio, las funciones de base radial estan dadas por $\phi(||x - k_K||)$ donde $\|v\| = \sqrt{v^t v}$ corresponde a la longitud del vector v. Las funciones ϕ conservan la propiedad de ser radialmente simétricas alrededor de k_K (Ruppert et al., 2003).

En la Figura 2.3 se muestra las funciones de base radial para $p = 1$ y 3 cuyos nodos fueron seleccionados en $x = \{0.1, 0.2, 0.4, 0.5, 0.8, 0.9\}$.

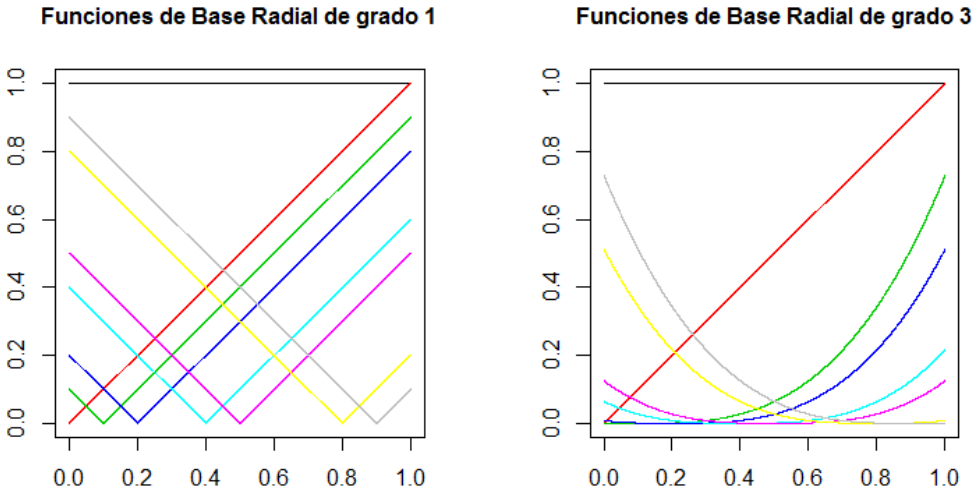

Figura 2.3: Ejemplos de Funciones de Base Radial

2.6. Selección de Modelos

Los modelos estadísticos son una herramientas para la extracción de información. Es importante entender que un modelo no es algo que se determina de manera única, sino que más bien es posible que asuma diversas formas dependiendo del punto de vista del investigador y la información disponible. Es decir, el propósito de modelado estadístico no es para estimar o identificar un modelo "único", sino más bien para construir un modelo que sea lo suficientemente bueno para la extracción de la información de acuerdo a las características del objeto y el propósito de la modelización (Konishi y Kitagawa, 2008).

Uno de los problemas más comunes en estadística es la selección del modelo que mejor describa el conjunto de datos y que a su vez no presente una cantidad excesiva de parámetros a estimar. A menudo se cuenta con modelos candidatos con número diferente de parámetros, por lo que existen varios criterios para la selección de modelos que proporcionan una alternativa a la hora de elegir un modelo parsimonioso, ya que se debe tener claro que al aumentar el número de parámetros en el modelo, también se incrementa la probabilidad de sobreajuste del mismo.

Cuando comparamos modelos basados en la verosimilitud, cuanto mayor sea esta mejor es el modelo. Los criterios de información indican, sin embargo, que dada una cantidad finita de datos disponibles para el modelado, un modelo que tiene una excesiva cantidad de grados de libertad conducirá a un aumento en la inestabilidad del modelo estimado, y esto dará lugar a una capacidad de predicción reducida (Konishi y Kitagawa, 2008). En otras palabras, no es beneficioso aumentar innecesariamente el número de parámetros libres sino se utiliza alguna restricción.

Uno de los criterios basados en la verosimilitud y que incluye una penalidad para el número de parámetros en el modelo es el criterio de información bayesiano (BIC) también conocido con el nombre de criterio de información de Schwarz. Este fue derivado por Schwarz en 1978 a partir de un enfoque bayesiano donde sus fundamentos se basan en la evaluación

de los modelos definidos en términos de su probabilidad posterior.

El BIC se define a través de la siguiente expresión:

$$BIC = -2l(\hat{\beta}) + n_p \log(N) \tag{2.31}$$

donde N representa el número de observaciones utilizadas para ajustar el modelo y n_p es el número de parámetros estimados. Entre un grupo de modelos ajustados al mismo conjunto de datos, aquel con menor BIC se considera el "mejor modelo".

Criterios basados en verosimilitud proporcionan una herramienta indispensable en la selección del "mejor" modelo para el ajuste. Sin embargo, esta no es la única finalidad de un modelo estadístico, también es de interés analizar su capacidad de predicción, es por esta razón que existen criterios que se enfocan en seleccionar el modelo con una mejor capacidad predictiva.

Los criterios de selección enfocados en la capacidad de predicción del modelo se basan en los errores. Uno de estos criterios es el Criterio de Error Absoluto Integrado Medio (MIAE por sus siglas en inglés Mean Integrated Absolute Error).

Para ejemplificar el concepto del Error Absoluto Integrado Medio considere que se conoce la curva verdadera $f(t)$ que describe la media de una población de sujetos en un estudio longitudinal. Sea $\hat{f}(t)$ una curva que estima a $f(t)$. Este estimador $\hat{f}$ se dice puntualmente insesgado si $E(\hat{f}(t_j)) = f(t_j)$.

La diferencia $\hat{f}(t_j) - f(t_j)$ es el error puntual de la estimación. Esta estimación será más precisa según sea más pequeño el error. El promedio de los errores obtenido debe estimarse, por ejemplo a tráves del valor absoluto, ya que estos errores son tanto positivos como negativos. Por lo tanto:

$$E\left[|\hat{f}(t_j) - f(t_j)|\right] \tag{2.32}$$

En el caso de estimar el cuadrado de los errores y mediante propiedades del valor esperado se puede obtener la varianza más el sesgo al cuadrado del estimador. Esta propiedad no se cumple de forma exacta con el uso del valor absoluto. Sin embargo, de la expresión (2.32) puede obtenerse una forma de acotar esta medida.

Como el interés no son las estimaciones puntuales de la curva, sino más bien la función completa se define el MIAE la cual es una medida de error que hace uso de la norma L_1, y el error global (IAE) obtenido integrando en el dominio de la función

$$IAE = \int |\hat{f}(t) - f(t)| \, dt. \tag{2.33}$$

El MIAE se obtiene promediando el error global

$$MIAE = E\left[IAE\left(\hat{f}(\cdot)\right)\right] = E\left[\int |\hat{f}(t) - f(t)| \, dt\right] = \int E\left[|\hat{f}(t) - f(t)|\right] dt$$

El criterio de selección elige como "mejor" modelo para la predicción al que tenga menor MIAE, ya que es el modelo que minimiza los errores.

Es importante tener presente que el MIAE es una medida de comparación que solo puede ser implementada en procesos de simulación, ya que en su definición se requiere conocer la curva verdadera.

Algunas ventajas del MIAE son:

- El una medida robusta a valores atípicos.
- El IAE tiene la interpretación atractiva de ser el área entre las curvas, lo que hace que sea fácil de visualizar.
- El estimador $\hat{f}$ es consistente si MIAE tiende a cero.

Capítulo 3

Regresión Beta y Métodos de Suavizado SemiParamétricos

3.1. Modelo de Regresión Beta

3.1.1. Introducción

La regresión Beta permite modelar datos restringidos de forma continua en el intervalo $(0, 1)$ como por ejemplo probabilidades, tasas y proporciones. Comúnmente el análisis de este tipo de datos se realizaba a través de transformaciones de la variable respuesta que permitieran el uso de modelos normales. Uno de los principales inconvenientes de esta práctica es la dificultad que surge al interpretar los coeficientes de regresión en términos de la variable original.

Ferrari y Cribari-Neto (2004) proponen un modelo de regresión que utiliza la distribución Beta a través de una re-parametrización de la función de densidad dada en (2.19) con el objetivo de que la distribución dependa de la media y un parámetro de dispersión.

Las ventajas que exponen los autores son por ejemplo, la interpretación directa de los resultados ya que no es necesario el uso de transformaciones para la variable de interés y la

facilidad para modelar asimetrías dado que la distribución puede adquirir gran variedad de formas como se observa en la Figura 3.1. Otra característica importante del modelo propuesto es que la varianza de la variable de interés depende de la media, por lo que no se asume homocedasticidad en el modelo.

3.1.2. Parametrización

Con el objetivo de definir un modelo de regresión Beta, para la media de la variable respuesta junto con un parámetro de dispersión constante, Ferrari y Cribari-Neto (2004) proponen utilizar la siguiente re-parametrización de la función de densidad de la distribución:

Sea la media μ y el parámetro de dispersión ϕ definidos de la siguiente manera:

$$\mu = \frac{p}{(p+q)} \tag{3.1}$$

$$\phi = p+q \tag{3.2}$$

Sustituyendo estas expresiones en (2.19) se obtiene:

$$f(y;\mu;\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\, y^{\mu\phi-1}\,(1-y)^{(1-\mu)\phi-1} \quad 0<y<1, \quad 0<\mu<1, \quad \phi>0 \tag{3.3}$$

Así, la media y varianza de la distribución Beta bajo esta parametrización se definen como:

$$E(y) = \mu, \tag{3.4}$$

$$Var(y) = \frac{\mu(1-\mu)}{(1+\phi)}. \tag{3.5}$$

Y el logaritmo de la función de densidad es:

$$\log f(y;\mu;\phi) = \log\Gamma(\phi) - \log\Gamma(\mu\phi) - \log\Gamma[(1-\mu)\phi] + (\mu\phi-1)\log(y) + [(1-\mu)\phi-1]\log(1-y) \tag{3.6}$$

Es importante destacar que la densidad de la distribución puede asumir diferentes formas según se varian los valores de los parámetros μ y ϕ. En particular para $\mu = \frac{1}{2}$ la forma será simetríca, mientras que para valores de $\mu \neq \frac{1}{2}$ se obtiene formas asimétricas. La Figura 3.1 muestra la densidad de la distribución para diferentes combinaciones de parámetros (μ, ϕ).

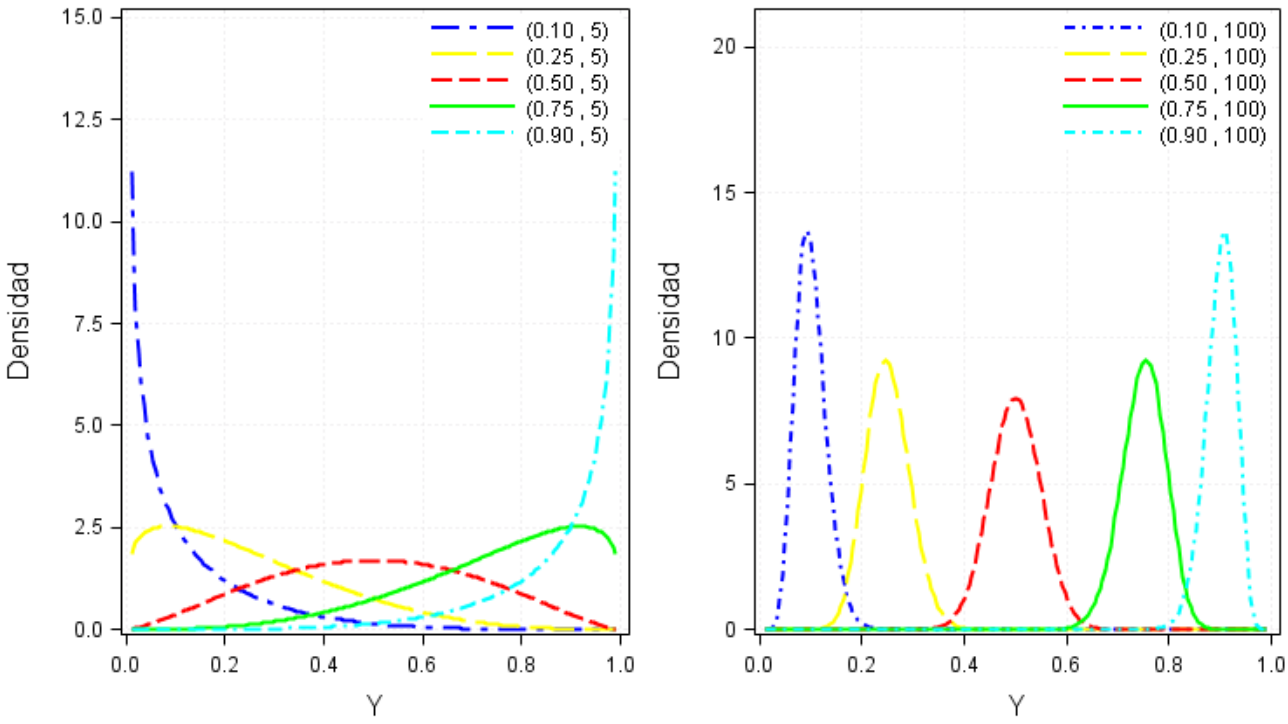

Figura 3.1: Función de densidad Beta para diferentes combinaciones de (μ, ϕ)

Pese a que se asume que la variable de interés para este modelo se restringe en el intervalo $(0, 1)$, es posible modelar datos restringidos en un intervalo conocido (a, b) si se utiliza la siguiente transformación:

$$\tilde{y} = \frac{(y - a)}{(b - a)}$$

Es importante tener claro que bajo esta transformación las interpretaciones serán en terminos de $\tilde{y}$, ya que es la variable que se está modelando.

3.1.3. Definición

Sea $y_1, y_2, \ldots, y_n$ variables aleatorios con función de densidad $f(y_i; \mu_i, \phi)$. Se define el modelo de regresión Beta como:

$$g(\mu_i) = x_i^t \beta = \eta_i \tag{3.7}$$

Por lo tanto:

$$\mu_i = g^{-1}(\eta_i) \tag{3.8}$$

donde $\beta = (\beta_1, \ldots \beta_k)^t$ representa el vector de parámetros de regresión $(k < n)$, $x_i = (x_{i1}, \ldots, x_{ik})^t$ es el vector de covariables, η_i es el predictor lineal y $g(\cdot)$ es la función de enlace tal que $g : (0, 1) \to \mathbb{R}$.

Típicamente las funciones de enlace más comúnmente utilizadas son:

- Función Logit: $g(\mu_i) = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$; $\qquad \mu_i = \dfrac{\exp(\eta_i)}{1+\exp(\eta_i)}$.
- Función Probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$, donde $\Phi^{-1}(\cdot)$ representa la función acumulada de la distribución normal; $\quad \mu_i = \Phi(\eta_i)$.
- Función Log-Log Complementario: $g(\mu_i) = \log[-\log(\mu_i)]$; $\qquad \mu_i = \exp[-\exp(-\eta_i)]$.

Estas funciones son monótonas y diferenciables en el intervalo $(0, 1)$. Ferrari y Cribari-Neto (2004) afirman que al modelar datos medidos como tasas o proporciones se obtienen resultados similares con cualquiera de estos enlaces. Pese a esto la función Logit es particularmente útil ya que es posible interpretar los parámetros de regresión en términos del cociente de chances (odds ratio).

La función de enlace Logit se define como:

$$\begin{aligned} g(\mu_i) &= \log\left(\frac{\mu_i}{1-\mu_i}\right) &&= x_i^t\beta \\ \Rightarrow \quad \frac{\mu_i}{1-\mu_i} &&&= \exp(x_i^t\beta) \\ \Rightarrow \quad \mu_i &&&= \frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)} \end{aligned}$$

con $x_i' = (x_{i1}, x_{i2}, \ldots, x_{ik})$ para $i = 1, \ldots, n$.

Ahora suponga que el valor de la i-ésima variable regresora se incrementa c unidades y las demás variables independientes permanecen constantes, y sea $\mu^\dagger$ la media de la variable respuesta bajo el nuevo valor de la covariable, y μ denota la media de la variable bajo los valores originales de las covariables, entonces el cociente de chances es:

$$e^{c\beta} = \frac{\dfrac{\mu^\dagger}{1-\mu^\dagger}}{\dfrac{\mu}{1-\mu}} \tag{3.9}$$

3.1.4. Estimación de Parámetros

Las estimaciones de los parámetros se obtienen por máxima verosimilitud.

Una función de log-verosimilitud basada en una muestra de n observaciones independientes está dada por:

$$l(\beta, \phi) = \sum_{i=1}^{n} l_i(\mu_i, \phi) \tag{3.10}$$

donde $l(\mu, \phi)$ corresponde al logaritmo de la función de densidad Beta expresado en (3.6).

Para obtener las estimaciones de los parámetros es necesario calcular la función de primeras derivadas, conocida comúnmente con el nombre de función *score*. Esta se obtiene derivando la log-verosimilitud con respecto a los parámetros desconocidos e igualando las ecuaciones a cero. Por lo tanto la función *score* calculada a partir de (3.10) corresponde a:

$$\frac{\partial l(\beta, \phi)}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i, \phi)}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_k} \tag{3.11}$$

donde $\dfrac{d\mu_i}{d\eta_i} = \dfrac{1}{g'(\mu_i)}$ y $\dfrac{\partial \eta_i}{\partial \beta_k} = x_{ik}$

Realizando los cálculos respectivos se obtiene:

$$\frac{\partial l_i(\mu_i, \phi)}{\partial \mu_i} = \phi \left[\log\left(\frac{y_i}{1-y_i}\right) - \{\psi(\mu_i\phi) - \psi[(1-\mu_i)\phi]\} \right] \tag{3.12}$$

donde la función $\psi(\cdot)$ corresponde a la función digamma definida por:

$$\psi(z) = \frac{\partial \log \Gamma(z)}{\partial z} = \frac{\Gamma'(z)}{\Gamma(z)}, \qquad z > 0$$

Al sustituir $y_i^* = \log\left(\frac{y_i}{1-y_i}\right)$ y $\mu_i^* = \psi(\mu_i\phi) - \psi[(1-\mu_i)\phi]$ en (3.12) se obtiene:

$$\frac{\partial l_i(\mu_i, \phi)}{\partial \mu_i} = \phi(y_i^* - \mu_i^*) \tag{3.13}$$

por lo tanto

$$\frac{\partial l(\beta, \phi)}{\partial \beta_k} = \phi \sum_{i=1}^{n} (y_i^* - \mu_i^*) \frac{1}{g'(\mu_i)} x_{ik} \tag{3.14}$$

La expresión anterior puede ser escrita en forma matricial de la siguiente manera:

$$\frac{\partial l(\beta, \phi)}{\partial \beta_k} = U_\beta(\beta, \phi) = \phi X^t T(y^* - \mu^*)$$

donde

X es una matriz de dimensión $n \times k$ cuya i-ésima fila es dada por x_i^t
$T = diag\left\{\frac{1}{g'(\mu_1)}, \ldots, \frac{1}{g'(\mu_n)}\right\}$
$y^* = (y_1^*, \ldots, y_n^*)^t$
$\mu^* = (\mu_1^*, \ldots, \mu_n^*)^t$

De igual manera la función *score* para el parámetro ϕ se obtiene derivando la función de log-verosimilitud con respecto a este parámetro, por lo tanto:

$$\frac{\partial l(\beta, \phi)}{\partial \phi} = \sum_{i=1}^{n} \frac{\partial l_i(\mu_i, \phi)}{\partial \phi} \tag{3.15}$$

donde

$$\begin{aligned} \frac{\partial l_i(\mu_i, \phi)}{\partial \phi} &= \mu_i \left[\log\left(\frac{y_i}{1-y_i}\right) - \psi(\mu_i\phi) + \psi[(1-\mu_i)\phi]\right] + \log(1-y_i) - \psi[(1-\mu_i)\phi] + \psi(\phi) \\ &= \mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi[(1-\mu_i)\phi] + \psi(\phi) \end{aligned}$$

por lo que se obtiene que:

$$\frac{\partial l(\beta,\phi)}{\partial \phi} \;=\; U_\phi(\beta,\phi) \;=\; \sum_{i=1}^{n} \mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi[(1-\mu_i)\phi] + \psi(\phi) \quad (3.16)$$

Finalmente la función *score* para los parámetros (β, ϕ) es dada por:

$$\left(U_\beta^t(\beta,\phi), U_\phi^t(\beta,\phi)\right) \quad (3.17)$$

Con el objetivo de determinar la variabilidad de las estimaciones de los parámetros del modelo de regresión Beta se obtiene la matriz de segundas derivadas, también conocida como matriz de información de *Fisher* (Ferrari y Cribari-Neto, 2004). La matriz de segundas derivadas de la función de log-verosimilitud con respecto al parámetro β es dada por:

$$\begin{aligned}
\frac{\partial^2 l(\beta,\phi)}{\partial \beta_k \partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial \mu_i}\left(\frac{\partial l_i(\mu_i,\phi)}{\partial \mu_i}\frac{d\mu_i}{d\eta_i}\right)\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}x_{ik} \\
&= \sum_{i=1}^{n}\left(\frac{\partial^2 l_i(\mu_i,\phi)}{\partial^2 \mu_i}\frac{d\mu_i}{d\eta_i} + \frac{\partial l_i(\mu_i,\phi_i)}{\partial \mu_i}\right)\frac{d\mu_i}{d\eta_i}x_{ik}x_{ij}
\end{aligned}$$

Similarmente se obtienen las matrices respecto a (β, ϕ) y respecto a ϕ. Para más detalle sobre el procedimiento para calcular la expresión de la matriz de información de Fisher, ver (Ferrari y Cribari-Neto, 2004, pág:15-16).
La matriz de información de Fisher es dada por la siguiente expresión:

$$K = K(\beta,\phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} \quad (3.18)$$

donde

$$\begin{aligned}
K_{\beta\beta} &= \phi X^t W X \\
K_{\beta\phi} &= K_{\phi\beta}^t \;=\; X^t T c \\
K_{\phi\phi} &= traza(D)
\end{aligned}$$

Siendo $W = diag\{w_1, w_2, \ldots, w_n\}$, $c = (c_1, c_2, \ldots, c_n)^t$ y $D = diag\{d_1, d_2, \ldots, d_n\}$ con:

$$\begin{aligned} w_i &= \phi\{\psi'(\mu_i\phi) + \psi'[(1-\mu_i)\phi]\}\frac{1}{\{g'(\mu_i)\}^2} \\ c_i &= \phi\{\psi'(\mu_i\phi)\mu_i - \psi'[(1-\mu_i)\phi](1-\mu_i)\} \\ d_i &= \psi'(\mu_i\phi)\mu_i^2 + \psi[(1-\mu_i)\phi](1-\mu_i)^2 - \psi'(\phi) \end{aligned}$$

donde $\psi'(\cdot)$ representa la función trigamma definida por:

$$\psi'(z) = \frac{\partial\psi(z)}{\partial z} = \frac{\partial^2 \log\Gamma(z)}{\partial^2 z} = \frac{\Gamma''(z)\Gamma(z) - [\Gamma(z)]^2}{[\Gamma(z)]^2} \qquad z > 0$$

A diferencia de los modelos lineales generalizados $K_{\beta\phi} = K_{\phi\beta}^t \neq 0$, por lo que los parámetros β y ϕ no son ortogonales.

3.2. Métodos SemiParamétricos para el suavizado de curvas

3.2.1. Introducción

Los modelos estadísticos basados en técnicas paramétricas son eficientes si se cuenta con información suficiente sobre el modelo subyacente a las variables, siendo así la principal tarea el determinar un número finito de parámetros (Durbán, 2009). Sin embargo, si no se cuenta con esta información, es posible que parte de la fuente de error del modelo sea la selección de una familia parámetrica no adecuada. Una alternativa eficiente para minimizar este tipo de error es la implementación de técnicas semiparamétricas. Su principal característica consiste en no asumir una forma específica para la función de regresión, sino construirla a través de las observaciones definiendo los parámetros y modelos de forma más general.

En el contexto semiparamétrico las técnicas de suavizado juegan un papel importante, ya que por la complejidad de los datos que se generan en muchos experimentos, es necesario

el uso de técnicas potentes que permitan construir una buena aproximación a la función de regresión, objetivo que en muchos casos es imposible bajo modelos de enfoque paramétrico.

Además de proporcionar una alternativa útil para realizar estimaciones y modelar datos con estructuras más complejas, las técnicas semiparámetricas pueden ser utilizadas como medio para explorar los datos y validar o realizar un diagnóstico de técnicas paramétricas. Desde el enfoque semiparamétrico la estimación de la función de regresión se realiza principalmente a través de métodos de regresión con *splines*. En la Sección 2.5 se presentan algunas nociones básicas. En la siguiente sección se desarrollan las estructuras de regresión de modelos con *splines*.

3.2.2. Regresión con *P-splines*

P-splines como Modelos Mixtos

La implementación de *splines* penalizados como modelos mixtos facilitan la incorporación de estructuras de correlación a la vez que se obtienen los beneficios del uso de curvas suaves para modelar. En el contexto de análisis de datos longitudinales, permiten ajustar modelos flexibles donde las diferencias individuales son funciones suaves del tiempo (Durbán, 2009).

Considere el modelo:

$$y = X\beta + Zu + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I)$$

Utilizando *P-splines* con bases de *B-splines* en la construcción de la matriz asociada a los efectos aleatorios Z se obtiene que:

$$Z = BU\Sigma^{-1/2}$$

Donde U y Σ son matrices que forman parte de la descomposición en valores singulares de

la matriz de penalidad $D'D$.

Ahora:

$$y = X\beta + Zu + \epsilon \qquad u \sim N(0, \sigma_u^2 I_{c-2}) \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I)$$

Donde c corresponde al número de columnas de la base.

Se obtiene que el parámetro de suavizado tiene la forma $\lambda = \sigma_\epsilon^2/\sigma_u^2$. La principal ventaja consiste en que el cálculo del parámetro de suavizado se realiza junto con la estimación de los otros parámetros del modelo. La estimación de estos componentes se realiza comúnmente a través del método de máxima verosimilitud restringida como se muestra a continuación,

$$l_R(\sigma_u^2, \sigma_\epsilon^2) = -\frac{1}{2}\log|V| - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}y'(V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})y$$

Con $V = \sigma_u^2 Z'Z + \sigma_\epsilon^2 I$. Los vectores de parámetros fijos y aleatorios se obtienen de:

$$\widehat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \tag{3.19}$$

$$\widehat{u} = \hat{\sigma}_u^2 Z'\hat{V}^{-1}(y - X\widehat{\beta}) \tag{3.20}$$

$$V^{-1} = \frac{1}{\sigma_\epsilon^2}(I - Z(Z'Z + \lambda I_{c-2})^{-1}Z') \tag{3.21}$$

3.2.3. Regresión con Suavizado Radial

Los *splines* de suavizado poseen una representación natural en términos de las funciones de base radial (Ruppert et al., 2003). Por ejemplo, para el parámetro de suavizado $\lambda > 0$ el spline de suavizado cúbico es definido a través de bases radiales de la siguiente forma:

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \sum_{j=1}^{n} \hat{\beta}_{1j} |x - x_j|^3 \tag{3.22}$$

donde $\hat{\beta}_0, \hat{\beta}_1$ y $[\hat{\beta}_{11}, \ldots, \hat{\beta_{1n}}]$ minimizan

$$||y - \boldsymbol{X_0}\boldsymbol{\beta_0} - \boldsymbol{X_1}\boldsymbol{\beta_1}||^2 + \lambda^3 \boldsymbol{\beta_1^t} \boldsymbol{K} \boldsymbol{\beta_1} \tag{3.23}$$

sujeto a la restricción $\boldsymbol{X_0^t}\boldsymbol{\beta_1} = 0$

donde

$$\begin{aligned}
\boldsymbol{\beta_0} &= [\beta_0, \beta_1]^t \\
\boldsymbol{\beta_1} &= [\beta_{11}, \ldots, \beta_{1n}]^t \\
\boldsymbol{X_0} &= [1, x_i]_{1 \leq i \leq n} \\
\boldsymbol{X_1} &= \boldsymbol{K} = \left[|x_i - x_j|^3\right]_{1 \leq i,j \leq n}
\end{aligned}$$

La ecuación (3.22) se generaliza mediante la sustitución de la potencia cúbica de la función radial a cualquier grado impar de la forma $2m - 1$ para $(m = 1, 2, \ldots)$ y la incorporación de los términos polinomicos hasta el grado $m - 1$.

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_{m-1} x^{m-1} + \sum_{k=1}^{K} \beta_{mk} |x - t_k|^{2m-1}$$

donde las estimaciones se obtiene minimizando

$$||y - X\beta||^2 + \lambda^{2m-1} \beta^t K \beta \tag{3.24}$$

con

$$X = \begin{pmatrix} 1 & x_1 & \dots & x^{m-1} & |x_1 - t_1|^{2m-1} & \dots & |x_1 - t_K|^{2m-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{m-1} & |x_n - t_1|^{2m-1} & \dots & |x_n - t_K|^{2m-1} \end{pmatrix}$$

$$K = \left[|t_k - t_{k'}|^{2m-1}\right]_{1 \leq k,k' \leq K}$$

Los valores ajustados son dados a tráves de la siguiente expresión:

$$\hat{y} = X(X^t X + \lambda^{2m-1} K)^{-1} X^t y \tag{3.25}$$

donde

$$S_\lambda = X(X^t X + \lambda^{2m-1} K)^{-1} X^t$$

es usualmente llamada matriz de suavizado.

3.3. Modelos de regresión semiparamétricos con distribución Beta y efectos aleatorios

3.3.1. Definición

Considere $y_i = (y_{i1}, y_{i2}, ..., y_{in_i})$ como el vector de n_i observaciones del i-ésimo sujeto.

El modelo de regresión que describe a los sujetos de forma específica es dado por

$$g(\mu_{ij}) = x_{ij}^t \beta + z_{ij}^t u_i$$
$$\mu_{ij} = E[y_{ij}|u_i]$$

donde $g(\cdot)$ la función de enlace y la distribución condicional es Beta definida como en

la Sección 3.1.

$$y_{ij}|u_i \sim Beta(\mu_{ij}, \phi) \qquad i = 1, \ldots, k. \qquad j = 1, \ldots, n_i.$$

Por lo que se describe a k sujetos de estudio, con n_i observaciones cada uno.
El vector u_i corresponde a los efectos aleatorios normalmente distribuidos:

$$u_i \sim N(0, G)$$

Por lo tanto, la verosimilitud marginal que describe los datos está dada por:

$$\prod_{i=1}^{k} \int \prod_{j=1}^{n_j} \left\{ \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma[(1-\mu_{ij})\phi]} y_{ij}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-\mu_{ij})\phi-1} \right\} \left\{ \frac{\exp\left(-\frac{1}{2}u_i^t G^{-1} u_i\right)}{|G|^{\frac{1}{2}}(2\pi)^{\frac{k}{2}}} \right\} du_i \tag{3.26}$$

Se define el modelo de regresión semiparamétrico con distribución Beta y efectos aleatorios como:

$$g(E[y_{ij}|u_i]) = f(t_{ij}, \gamma) + h_i(t_{ij}, \delta) \tag{3.27}$$

$$y_{ij}|u_i \sim B(\mu_{ij}, \phi)$$

donde f y h_i son funciones suaves del tiempo, γ es el parámetro asociado al efecto del tratamiento y $g(\cdot)$ es la función de enlace.

I. Modelos B-spline+B-spline

El modelo B-spline+B-spline utiliza bases de *B-splines* para la construcción de las funciones asociadas tanto a los efectos fijos como aleatorios del modelo, por lo que las funciones f y h_i del modelo (3.27) son de la forma:

$$\begin{aligned} f(t_{ij}, \gamma) &= \sum_{v=1}^{p} \gamma_v B(t_{ij}) \\ h_i(t_{ij}, \delta) &= \sum_{r=1}^{n_i} \delta_r B_{ir}(t_{ij}) \end{aligned}$$

II. Modelo B-spline+P-spline

El modelo B-spline+P-spline utiliza bases de *B-splines* para la construcción de la función asociada a los efectos fijos del modelo, mientras que para la función asociada a los aleatorios se utiliza *splines* penalizados descritos en la Sección 3.2.2. Por lo tanto la función h es de la forma:

$$h_i(t_{ij}, \vartheta) = \sum_{r=1}^{n_i} \vartheta_r B^*_{ir}(t_{ij})$$

III. Modelo B-spline+RS

En el modelo B-spline+RS la función asociada a los efectos fijos se obtiene a través de bases de *B-splines*, y la función asociada a los efectos aleatorios es construida a través del método de Suavizado Radial, el cual hace uso de funciones de base radial expresadas por (2.30). La función h definida en la secuencia de nodos $\{k_1, \ldots, k_t\}$ tiene la forma:

$$h_i(t_{ij}, \zeta) = \sum_{r=1}^{n_i} \zeta_r \phi_{ir}(t_{ij})$$

Capítulo 4

Simulaciones

4.1. Proceso de Simulación

En el estudio de modelos semiparamétricos con distribución Beta y efectos aleatorios, es de interés comparar los modelos propuestos en la Sección 3.3.1, a través de su capacidad de ajuste y predicción. Para este proceso se han generado datos simulados que correspondan a curvas de observaciones en el tiempo, de unidades experimentales similares clasificadas en dos tratamientos; estas observaciones siguen una distribución Beta.

Los escenarios consideramos en este proceso son los siguientes:

- Ajuste de los modelos a datos con tratamientos que presentan diferencias significativas.
- Ajuste de los modelos a datos con tratamientos sin diferencias.

Bajo cada escenerio se hace la selección del mejor modelo que se ajuste a los datos según el criterio de información bayesiano (BIC) y el mejor modelo de predicción según el criterio de error absoluto integrado medio (MIAE), ambos criterios presentandos en la Sección 2.6.

Antes de realizar el proceso de comparación de los modelos mediante los criterios de BIC y MIAE es necesario seleccionar el número óptimo de nodos para cada modelo. Por razones de costo computacional se decidió realizar esta selección a través de simulaciones

independientes al proceso de comparación de los modelos . La descripción y los resultados de ambos procesos se detallan a continuación.

4.1.1. Descripción de los datos.

Para el desarrollo del proceso de simulación se generan datos con una estructura similar a los datos reales con los que se llevará a cabo el proceso de aplicación. La descripción de cómo se obtuvieron los datos simulados es la siguiente:

En primer lugar se ajustarón a los datos de uno de los grupos del ejemplo a discutir en el próximo capítulo, modelos de regresión polinómica con el objetivo de establecer el grado de la curva que mejor los describe. En este caso se seleccionó el polinomio de grado $p = 6$. Se fijaron los paramétros asociados a los coeficientes del polinomio β_i para $i = 1, \ldots, 6$ y el parámetro de escala ϕ.

El polinomio seleccionado en el ajuste corresponderá a la curva típica del tratamiento, por lo tanto para el escenario con tratamientos que no presentan diferencias significativas, la curva será la misma para ambos tratamientos, mientras que en el escenario con diferencias entre tratamientos se modifican algunos de los coeficientes β_i de forma que las curvas presenten interacción. La Figura 4.1 muestra las curvas utilizadas en el escenario con diferencias entre tratamientos.

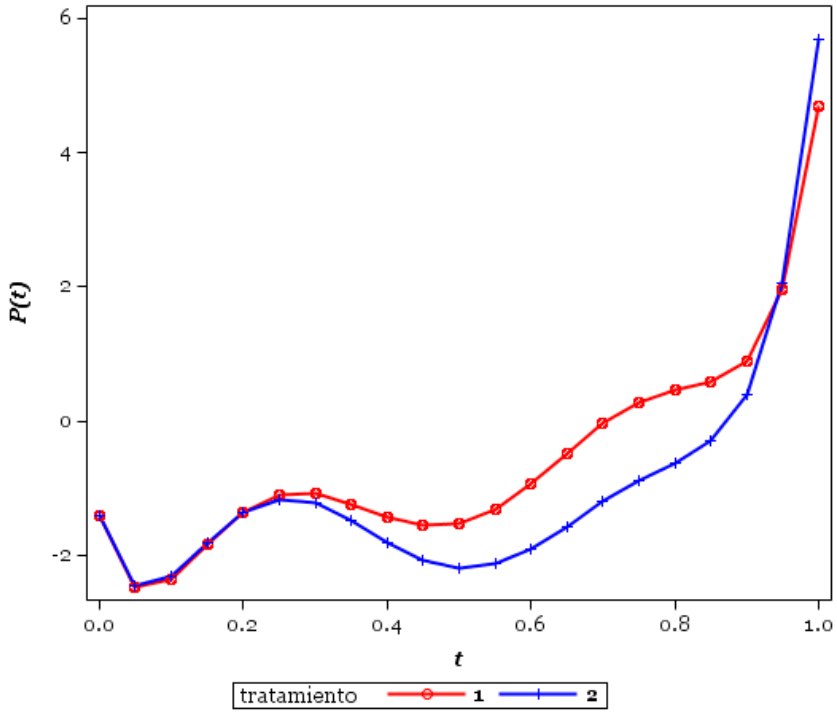

Figura 4.1: Curvas típicas para el proceso de simulación con tratamientos que presentan diferencia

Para generar las curvas para los sujetos específicos se agrega al polinomio una función sinusoidal que permita establecer las fluctuaciones como las observadas en los datos. Se simularon las curvas del proceso de la enfermedad de 20 plantas clasificadas en dos tratamientos con 10 plantas cada uno.

Para generar los datos distribuidos Beta, es decir valores restringidos de forma continua en $(0, 1)$, se utilizó la función **rannor** del programa SAS para generar los efectos aleatorios normalmente distribuidos, mientras que la variable respuesta y se generó con la función **rand('beta',p, q)**, donde **p** y **q** son los parámetros de la distribución.
Para el proceso se fijaron los paramétros:
β_0 = -1.42, β_1 = -40.59, β_2 = 476.05, β_3 = -1999.64, β_4 = 3847.88, β_5 = -3441.42, β_6 = 1163.85 y $\phi = 50$. Estos valores fueron obtenidos a través del ajuste polinómico a los datos de severidad en el ejemplo de aplicación.

Algoritmo

1. Inicia ciclo para cada tratamiento $k = 1, 2$.

   a) Para cada sujeto específico $i = 1, 2, 3, \ldots, 10$.

   1.1 Generar el efecto aleatorio $b_{ik} \sim N(0, \sigma_b)$

   1.2 Crear la variable fija $t = 0,\ 0.05,\ 0.1,\ 0.15,\ \ldots,\ 1$.

   1.3 Calcular la media $\mu_{ikt} = \dfrac{\exp\left(P_k + b_{ik}S_{ikt}\right)}{1 + \exp\left(P_k + b_{ik}S_{ikt}\right)}$

   donde P es el polinomio y S la función sinusoidal

   1.4 Calcular los parámetros $p_{ikt} = \mu_{ikt}\phi$ y $q_{ikt} = \phi(1 - \mu_{ikt})$

   1.5 Generar la variable $y_{ikt} \sim Beta(\mu_{ikt}, \phi)$

2. Fin del ciclo.

4.1.2. Proceso de selección del número óptimo de nodos en el modelo

El proceso implementado para la selección del número óptimo de nodos, en cada modelo propuesto, consistió en el ajuste de cada uno de los modelos para diferentes combinaciones de nodos en la parte fija y aleatoria del modelo, y la selección por medio del criterio de BIC de aquella combinación de valores que generó el mejor ajuste.

El proceso de simulación se llevo a cabo con ayuda de PROC GLIMMIX del programa estadístico SAS versión 9.3, el cual permite ajustar MLGM e incorporar las estructuras de regresión de *splines*. La función de enlace utilizada es la función Logit, las estimaciones se realizan a través del enfoque de máxima verosimilitud, ya que este permite la comparación de modelos con el criterio de BIC y finalmente métodos de apróximación como Laplace y de optimización como Newton Raphson son requeridos.

Para el proceso iterativo se definieron las variables $\boldsymbol{F}$ como el número de nodos utilizados en el ajuste da la parte fija del modelo y $\boldsymbol{A}$ como el número de nodos para la parte aleatoria. Ambas variables asumieron valores iniciales iguales a 2 y un valor final igual a 10 nodos, con un incremento de una unidad en cada proceso iterativo.

Un esquema general del algoritmo sería:
Para cada uno de los modelos propuestos realicé los siguientes pasos:
Paso 1: Generar los datos.
Paso 2. Ajustar el modelo con cada combinación de las variables $\boldsymbol{F}$ y $\boldsymbol{A}$.
Paso 3. Seleccionar a través del criterio de BIC, la combinación de nodos con la que se obtuvo el mejor ajuste.
Paso 4. Repetir 500 veces los pasos 1 al 3.

Los resultados obtenidos sobre la elección del número óptimo de nodos para cada uno de los modelos propuestos, bajo un escenario de tratamientos que presenta diferencias significativas, se muestran en la Tabla 4.1.

Tabla 4.1: Número óptimo de nodos para el ajuste de los datos simulados para cada modelo de regresión propuesto bajo el escenario de tratamientos con diferencias.

Modelo	**Nodos**		**Frecuencia**	**Porcentaje**	**Desviación Estándar**
	F	**A**			
B-spline + B-spline	4	2	304	60.80	0.69
B-spline + P-spline	4	3	94	18.80	0.55
B-spline + RS	4	–	320	64.00	0.68

El número óptimo de nodos para cada modelo ajustado a los datos simulados, bajo el escenario de tratamientos sin diferencias se muestran en la Tabla 4.2.

Tabla 4.2: Número óptimo de nodos para el ajuste de los datos simulados para cada modelo de regresión propuesto bajo el escenario de tratamientos sin diferencias

Modelo	**Nodos**		**Frecuencia**	**Porcentaje**	**Desviación Estándar**
	F	**A**			
B-spline + B-spline	4	2	353	70.60	0.64
B-spline + P-spline	4	4	108	21.60	0.58
B-spline + RS	4	–	324	64.80	0.68

Ahora que se conocen el número de nodos de cada modelo, es posible comparar los modelos de regresión semiparamétricos por medio de otras simulaciones y así seleccionar el modelo con mejores carácteristicas para el ajuste y el modelo con mejores condiciones para la predicción.

4.2. Descripción de los Escenarios

Escenario 1: Ajuste de los modelos a datos con tratamientos que presentan diferencias significativas.

Para el proceso de simulación se generan datos que son curvas de observaciones, clasificadas en dos tratamientos cuyas curvas típicas (los efectos aleatorios son cero) sean diferentes y presenten interacción significativa.

Resultados

La Tabla 4.3 presenta los porcentajes de elección de cada modelo, como el "mejor" modelo para el ajuste de los datos, a través del criterio de BIC. Se observa que el modelo que se elige con mayor frecuencia, en el proceso de simulación bajo este escenario, es el modelo que combina *B-splines* en la parte fija con *B-splines* en la parte aleatoria.

Por lo tanto el modelo que mejor se ajusta a los datos bajo un escenario de tratamientos que presentan diferencias es el modelo semiparamétrico B-spline+B-spline. También se puede

observar que el modelo B-spline+RS no es recomendable para el proceso de ajuste.

Tabla 4.3: Porcentajes de elección del modelo bajo un escenario de tratamientos con diferencia a través del criterio de BIC.

Modelo	**Porcentaje de aceptación BIC**	**Desviación Estandar**
B-spline + B-spline	54.62	0.70
B-spline + P-spline	43.72	0.70
B-spline + RS	1.66	0.18

Por otro lado la Tabla 4.4 muestra que el modelo B-spline + RS es el modelo con mejores características para la predicción, según el criterio de MIAE.

Tabla 4.4: Porcentajes de elección del modelo bajo un escenario de tratamientos con diferencia a través del criterio de MIAE.

Modelo	**Porcentaje de aceptación MIAE**	**Desviación Estandar**
B-spline + B-spline	30.02	0.65
B-spline + P-spline	25.02	0.61
B-spline + RS	44.96	0.70

Escenario 2: Ajuste de los modelos a datos con tratamientos sin diferencias significativas.

Para el proceso de simulación bajo este escenario se generan datos que son curvas de observaciones, clasificadas en dos tratamientos cuyas curvas típicas, a diferencia del escenario anterior, sean la misma, de esta forma no existe diferencias entre los tratamientos.

Resultados

Por medio de la información provista por la Tabla 4.5, se concluye que el modelo B-spline+B-spline es el modelo que mejor se ajusta a los datos generados bajo el escenario de tratamientos sin diferencia. De igual manera se establece que el modelo que combina

B-spline, en la parte fija del modelo y Suavizado Radial en la parte aleatoria, es el modelo que se selecciona la menor cantidad de ocasiones en el proceso de simulación, por lo que no es un modelo adecuado para el ajuste de este tipo de datos.

Tabla 4.5: Porcentajes de elección del modelo bajo un escenario de tratamientos sin diferencia a través del criterio de BIC.

Modelo	**Porcentaje de aceptación BIC**	**Desviación Estandar**
B-spline + B-spline	54.90	0.70
B-spline + P-spline	43.38	0.70
B-spline + RS	1.72	0.18

De la Tabla 4.6 se concluye que el modelo más adecuado para la predicción, bajo este escenerio es el modelo B-spline+RS, ya que se selecciona con más frecuencia, mediante el criterio de MIAE en el proceso de simulación.

Tabla 4.6: Porcentajes de elección del modelo bajo un escenario de tratamientos sin diferencia a través del criterio de MIAE.

Modelo	**Porcentaje de aceptación MIAE**	**Desviación Estandar**
B-spline + B-spline	32.64	0.66
B-spline + P-spline	28.06	0.64
B-spline + RS	39.30	0.69

Capítulo 5

Aplicaciones: Estudio de Severidad de Enfermedades en Cultivos de Banano en Puerto Rico

5.1. Enfermedad de Sigatoka Negra

La Sigatoka negra es una enfermedad causada por el hongo *Mycosphaerella fijiensis*. Este patógeno se encarga de destruir rápidamente el tejido foliar de la planta, como consecuencia el proceso fotosíntetico se ve afectando provocando un menor crecimiento tanto en la planta como en los racimos y frutos, en comparación con plantas sanas (Marengo, 2010, Álvarez et al., 2003). Infecciones severas de la enfermedad pueden causar la madurez prematura de los frutos entorpeciendo el proceso de recolección y generando perdidas económicas. Esta enfermedad constituye uno de los principales problemas fitopatológicos del cultivo de musáceas como el banano y el platano a nivel mundial. En la Figura 5.1 se observa plantas infectadas con la enfermedad.

El ciclo patológico de la Sigatoka negra comienza con el contagio de las hojas nuevas a través de esporas dispersadas normalmente por el viento o salpicadura de lluvia proveniente de hojas contaminadas. En un lapso aproximado de 2 a 4 días se produce la germinación

Figura 5.1: Plantas con síntomas de la enfermedad Sigatoka negra (Álvarez et al., 2003).

de las esporas infectando las hojas con la enfermedad, después de entre 10 a 30 días de la infección se comienzan a ver los sintomas del primer y segundo estado, a lo largo del tiempo estos estados evolucionan en la escala de severidad. Durante los estados 3 al 6 es donde se producen las esporas del hongo que afectarán las nuevas hojas, comenzando así el ciclo de la enfermedad (Álvarez et al., 2003).

La Sigatoka negra ha ocasionado graves pérdidas en la producción comercial de banano y ha modificado el manejo de las plantaciones, principalmente los programas de control químico. Actualmente el combate de la Sigatoka negra en las plantaciones de banano depende principalmente de la aplicación continua de fungicidas previo a la infección. Otra medida utilizada para el control y la mitigación de la enfermedad es el combate natural, este consiste en la implementación o modificación de ciertas prácticas de cultivo con la finalidad de generar un ambiente menos favorable para la enfermedad; o afectar la reproducción, diseminación e infección del patógeno. Dentro de las técnicas de combate cultural se encuentra el deshoje, deshije, embolsado de los racimos, construcción de drenajes, siembra de cultivos que sirvan de barrera biológica, aplicación de fertilizantes minerales, defoliación controlada a la floración, entre otras (Álvarez et al., 2003).

5.2. Descripción de los Datos

Los datos utilizados en este trabajo provienen del proyecto "Practices for the Control of Black Sigatoka in Puerto Rico"(Proyecto ZFIDA-01) a cargo del Dr. José A. Chavarría Carvajal, Estación Experimental Agrícola, Colegio de Ciencias Agrícolas, Recinto Universitario de Mayagüez, Universidad de Puerto Rico.

Con el objetivo de examinar el desarrollo de la enfermedad Sigatoka Negra se realizó una siembra de banano (*Musa acuminata*, AAA cv. "Grand Naine") por su importancia económica en Puerto Rico y la sensibilidad a la enfermedad y se recopiló información del índice de severidad de algunas plantas a lo largo del tiempo.

El diseño experimental utilizado fue de parcelas divididas con tres replicaciones por factor químico, en cada parcela experimental se realizó la siembra de 24 plantas de banano las cuales se distribuyeron en cuatro hileras de 6 plantas cada una. Con el objetivo de comparar diferentes prácticas para controlar la Sigatoka negra, se midió el índice de severidad de las tres plantas centrales de cada hilera. Las plantas de los extremos de cada hilera fuerón utilizadas como barrera con el objetivo de evitar la contaminación entre tratamientos.
El estudio conto con dos factores de interés:

1. Factor Químico
    - Ausencia
    - Presencia
2. Factor Cultural
    - Tratamiento 1: Desfoliación Mecánica
    - Tratamiento 2: No Desfoliación Mecánica

- Tratamiento 3: Deshije
- Tratamiento 4: No Deshije

Para cada combinación de factor químico con factor cultural se obtuvo la información de 9 plantas, las cuales fueron evaluadas a lo largo de 39 semanas, generando una base de datos con un total de 2807 observaciones. Es importante señalar que la base de datos muestra datos faltantes debido a la perdida de información de algunas plantas a lo largo del proceso. Por lo tanto se cuanta con 72 curvas de progreso cada una con aproximadamente 39 observaciones.

5.2.1. Índice de Severidad

El Método de Stover modificado por Gauhl estima de manera visual el área total, que presenta todos los síntomas de la enfermedad, en cada hoja de plantas próximas a la floración. Bajo este método existen siete grados de daño en el que se puede clasificar la hoja (Marengo, 2010), en la Figura 5.2 se puede visualizar un ejemplo, mientras que la Tabla 5.1 muestra la clasificación de la escala de Stover-Gauhl según los porcentajes del área afectada.

Tabla 5.1: Grados de severidad de la enfermedad Sigatoka negra según la escala de Stover-Gauhl

Grado	**Porcentaje del área foliar afectada**
0	0 %
1	Menos del 1 %
2	1 % - 5 %
3	6 % - 15 %
4	16 % - 33 %
5	34 % - 50 %
6	51 % - 100 %

Cada hoja de una planta es medida a través de la escala de Stover-Gauhl, el promedio de todas las hojas da como resultado el índice de severidad de la enfermedad. En el análisis con modelos semiparamétricos con distribución Beta se utiliza el índice de severidad en la

escala 0 - 1. El principal objetivo de este capítulo será la comparación, por separado, del factor químico de cada tratamiento, ajustando el modelo semiparamétricos con *B-splines* tanto en la parte fija como la aleatoria del modelo y el modelo con *B-splines* en la parte fija y Suavizado Radial en la aleatoria.

Figura 5.2: Ejemplo de hojas clasificadas según la escala de Stover-Gauhl (Marengo, 2010)

5.3. Métodos de Análisis

En las Figuras 5.3, 5.4, 5.5 y 5.6 se muestran las curvas de observaciones, a lo largo del tiempo, del índice de severidad de la enfermedad Sigatoka negra para cada uno de los cuatro tratamientos agrupados respecto al factor químico (NO = ausencia, SI= presencia). A continuación se presenta el análisis realizado mediantes los modelos de regresión semiparamétrica con distribución Beta y el efecto aleatorio de la planta, este efecto permite modelar la correlación entre las observaciones de la misma planta.

Los modelos para la media se ajustaron mediante el enlace Logit asumiendo distribución Beta para la variable respuesta. Además se permitió que las curvas tengan un intercepto y pendiente aleatoria. El objetivo del estudio de estos datos es comparar el efecto del factor químico en cada tratamiento por separado.

Cada uno de los tratamientos fue analizado de forma independiente, a través del modelo de regresión semiparamétrico B-spline+B-spline y el modelo semiparamétrico B-spline+RS, modelos que fueron seleccionados en ambos escenarios del proceso de simulación en la Sección 4.1.

En primer lugar se calcula el número óptimo de nodos para cada modelo dentro de cada tratamiento. Los resultados se muestran en las Tablas 5.2, para el modelo B-spline+B-spline y 5.3 para el modelo B-spline+RS. Este proceso se llevo a cabo ajustando el modelo para diferentes valores, y seleccionando aquel con menor BIC, en el caso del modelo con *B-splines* no se presentó ningún inconveniente, sin embargo para el modelo con B-spline y Suavizado Radial la estimación de la función de log-verisimilitud no convergió para la mayoría de los valores.

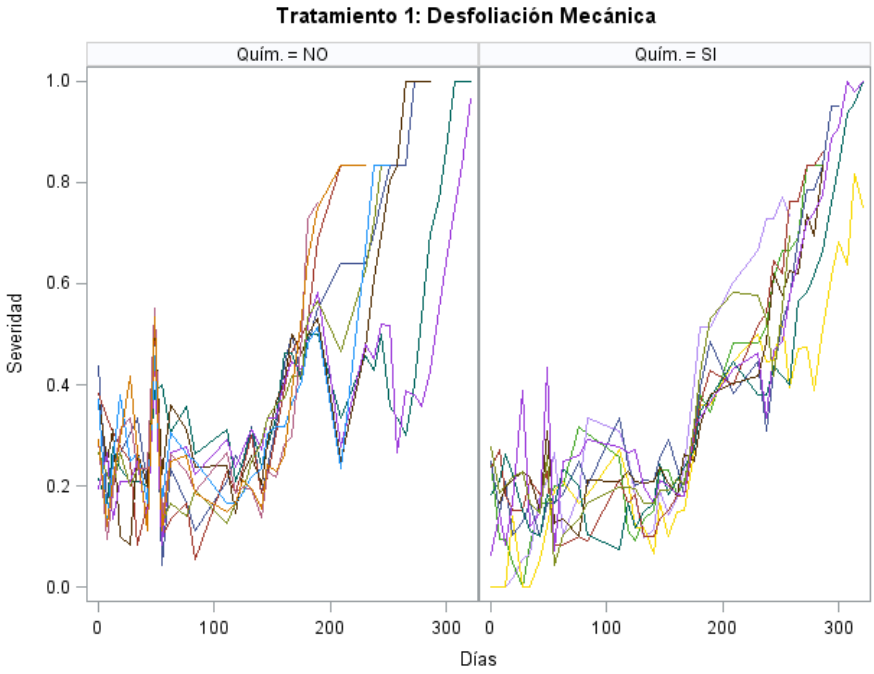

Figura 5.3: Curvas del proceso de la enfermedad Sigatoka negra en plantas del tratamiento Desfoliación Mecánica

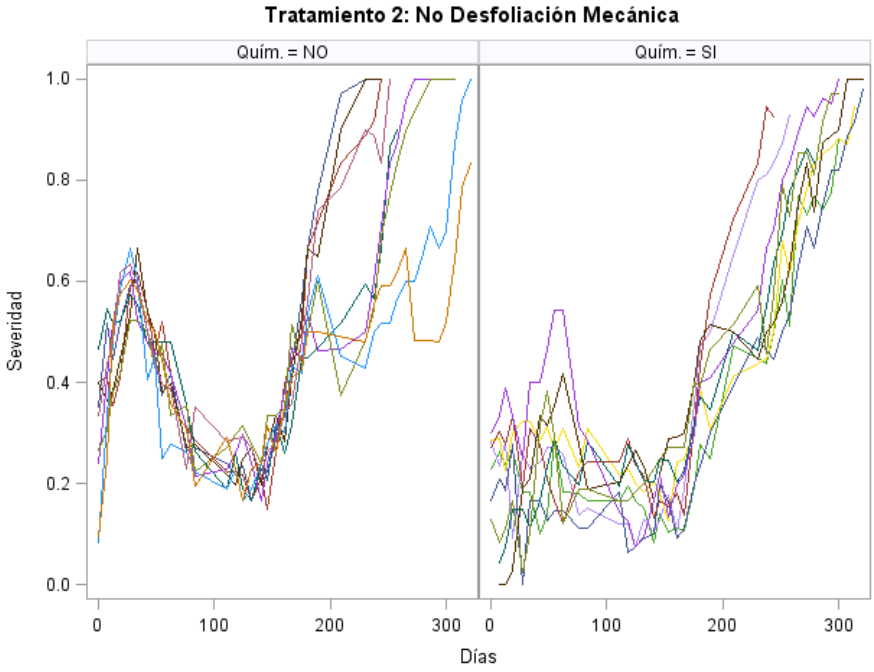

Figura 5.4: Curvas del proceso de la enfermedad Sigatoka negra en plantas del tratamiento No Desfoliación Mecánica

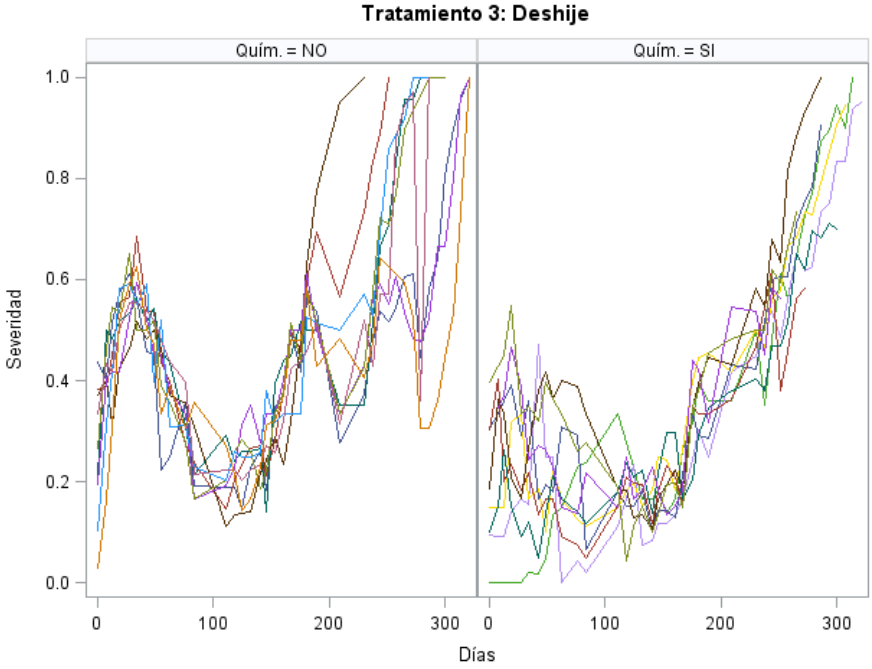

Figura 5.5: Curvas del proceso de la enfermedad Sigatoka negra en plantas del tratamiento Deshije

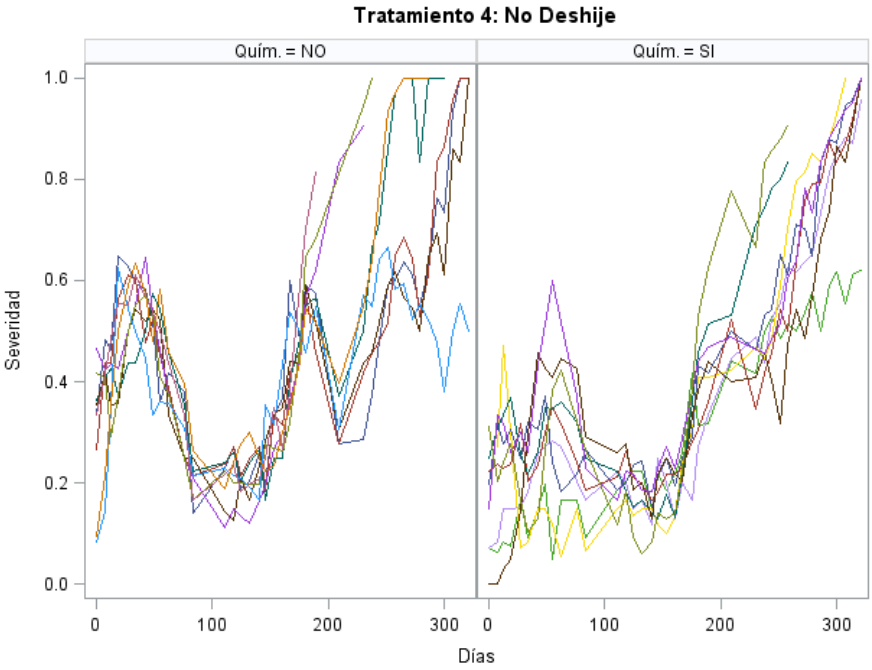

Figura 5.6: Curvas del proceso de la enfermedad Sigatoka negra en plantas del tratamiento No Deshije

5.3.1. Modelo de regresión semiparamétrico B-spline+B-spline

El modelo definido en la Sección 3.3.1 se puede reescribir como desviaciones con respecto a la media. El modelo para el análisis del índice de severidad en la enfermedad de Sigatoka negra, para cada tratamiento puede ser descrito de la siguiente manera:

$$logit\left[E(IS_{ijk} \mid planta_i)\right] = \underbrace{\mu + \kappa Q_k + \sum_{l=1}^{m} \delta_l B_l(t_j) + \sum_{r=1}^{n} \omega_r B_{kr}(t_j)}_{parte\quad fija} + \underbrace{\sum_{z=1}^{p} \xi_z B_{ikz}(t_j)}_{parte\quad aleatoria} \quad (5.1)$$

donde μ es la media general, Q_k representa el químico, t_j corresponde al j-ésimo tiempo y B son bases de *B-splines*.

Discusión de Resultados

Es importante aclarar que con la finalidad de reproducir el proceso llevado a cabo en las simulaciones, se decidió realizar el análisis de los tratamientos de forma completamente independiente, ya que los datos simulados se clasificarón en dos tratamientos y en los datos reales el factor químico cuenta con dos niveles en cada tratamiento.

Los resultados obtenidos mediante análisis separados en el cálculo del número óptimo de nodos en el modelo para cada tratamiento son los siguientes:

Tabla 5.2: Número de nodos óptimos por tratamiento para el modelo B-spline+B-spline, en el análisis del IS.

Tratamiento	**Nodos**	
	F	**A**
T1: Desfoliación Mecánica	8	4
T2: No Desfoliación Mecánica	8	4
T3: Deshije	7	4
T4: No Deshije	10	4

Una vez seleccionado el número de nodos óptimos para cada tratamiento se realizó el ajuste del modelo. En los resultados obtenidos en el ajuste del modelo para cada uno de los tratamientos (ver Anexo A) se observa que el modelo B-spline+B-spline se ajusta bien a los datos, ya que el estadístico Chi-cuadrado gener. / DF para todos los tratamientos es un valor cercano o igual a 1.

Las pruebas de efectos fijos muestran que para todos los tratamientos, el efecto de la interacción entre el químico y el *spline* es significativo, esto justifica la importancia de modelar curvas típicas diferentes para cada nivel del factor químico, permitiendo que tanto el intercepto como la pendiente sean aleatorias.

A través de los contrastes realizados se puedo observar lo siguiente:
Para el tratamiento 1: Desfoliación Mecánica no se detecta diferencias en el progreso de la enfermedad de Sikatoka negra, entre plantas tratadas y no tratadas químicamente a los 100, 200 y 320 días del estudio. Para el tratamiento 2 y 4: No Desfoliación Mecánica y No deshije, solo se detecto diferencias en el progreso de la enfermedad entre los niveles del factor químico a los 50 y 250 días. Finalmente para el tratamiento 3: Deshije solo se observan diferencias a los 50, 250 y 320 días.

En la Figura 5.7 se observan las curvas típicas de cada uno de los tratamientos, mediante estas imagenes es posible verificar los resultados obtenidos en los contrastes.

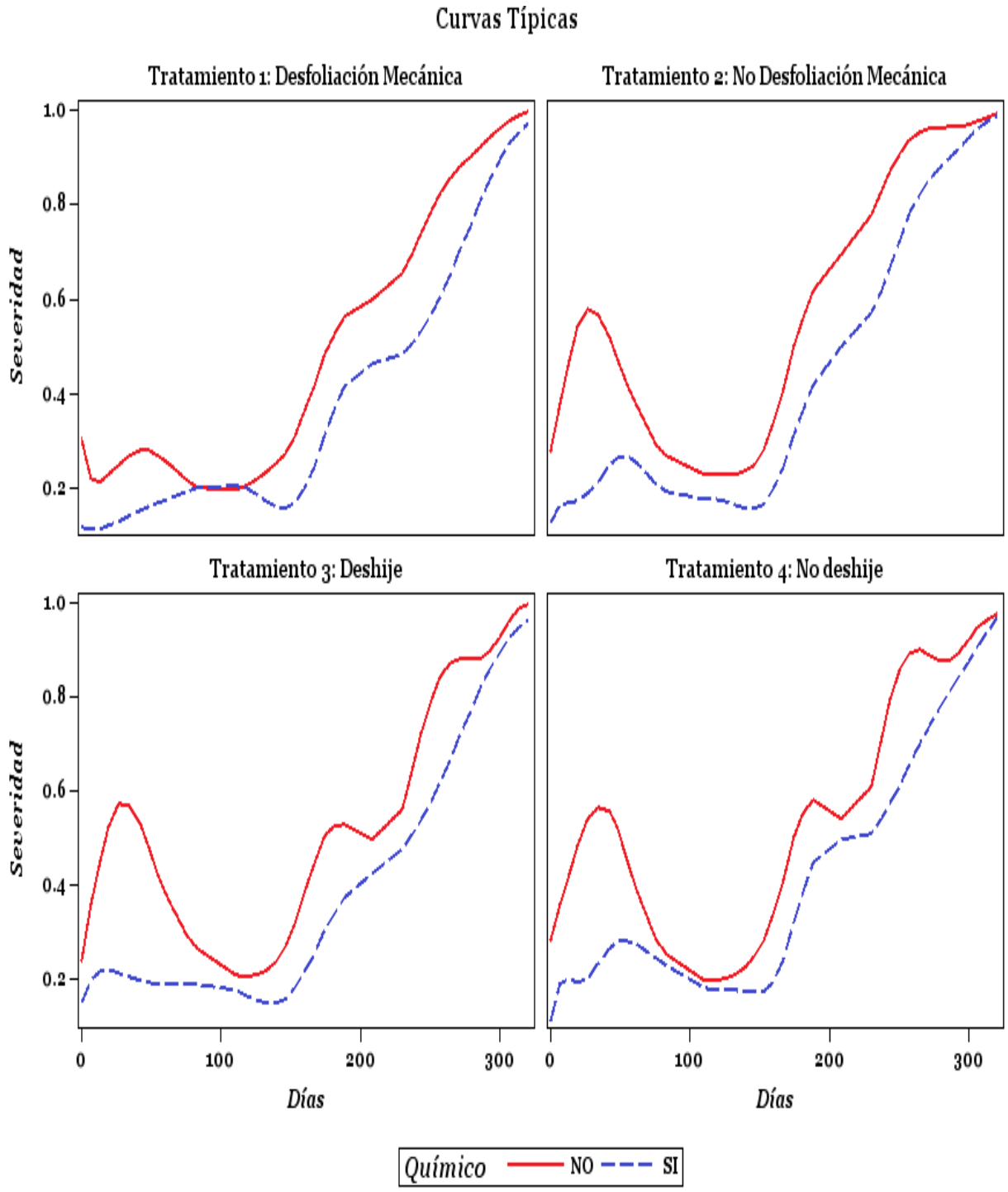

Figura 5.7: Curvas Típicas por tratamiento modelo B-spline+B-spline

5.3.2. Modelo de regresión semiparamétrico B-spline+RS

El modelo de regresión B-spline+RS implementado en el análisis de los datos de severidad es el siguiente:

$$logit\left[E(IS_{ijk} \mid planta_i)\right] = \mu + \kappa Q_k + \sum_{l=1}^{m} \delta_l B_l(t_j) + \sum_{r=1}^{n} \omega_r B_{kr}(t_j) + \sum_{z=1}^{p} \xi_z \phi_{ikz}(t_j) \quad (5.2)$$

Discusión de Resultados

En la Tabla 5.3 se muestra el número de nodos óptimos para la estimación de las bases de B-spline en el modelo B-spline+RS, según el proceso de selección, para cada uno de los tratamientos.

Tabla 5.3: Número de nodos por tratamiento para el modelo B-spline+RS, en el análisis del IS.

Tratamiento	**Nodos**
T1: Desfoliación Mecánica	8
T2: No Desfoliación Mecánica	5
T3: Deshije	4
T4: No Deshije	10

En el proceso de selección del número de nodos se presentarón problemas de convergencia en la estimación de la función de log-verosimilitud para este modelo, por ejemplo en el tratamiento 2: No Desfoliación Mecánica, solo se logró convergencia para el modelo con cinco nodos en la parte fija.

De los resultados obtenidos en el ajuste (ver Anexo B), se observa que el modelo B-spline+RS se ajusta bien a los datos, ya que el estadístico Chi-cuadrado gener. / DF es un valor cercano o igual a uno, sin embargo las pruebas de los efectos fijos reflejan que el modelo posee errores estándar mayores al modelo B-spline+B-spline, es por esta razón que las pruebas sugieren la no significancia de los efectos fijos.

En los contrates realizados no fue posible encontrar diferencias entre las curvas de progreso de plantas tratadas químicamente y las no tratadas cuando en la Figura 5.8 se reflejan diferencias en algunos intervalos de tiempo, esto se debe nuevamente a los errores del modelo. Es posible que el modelo B-spline+RS se sugiere como un buen modelo de predicción debido a que se generan errores altos, ya que por ejemplo al ser los errores mayores los intervalos de predicción son más anchos y es más probable que contengan las nuevas observaciones.

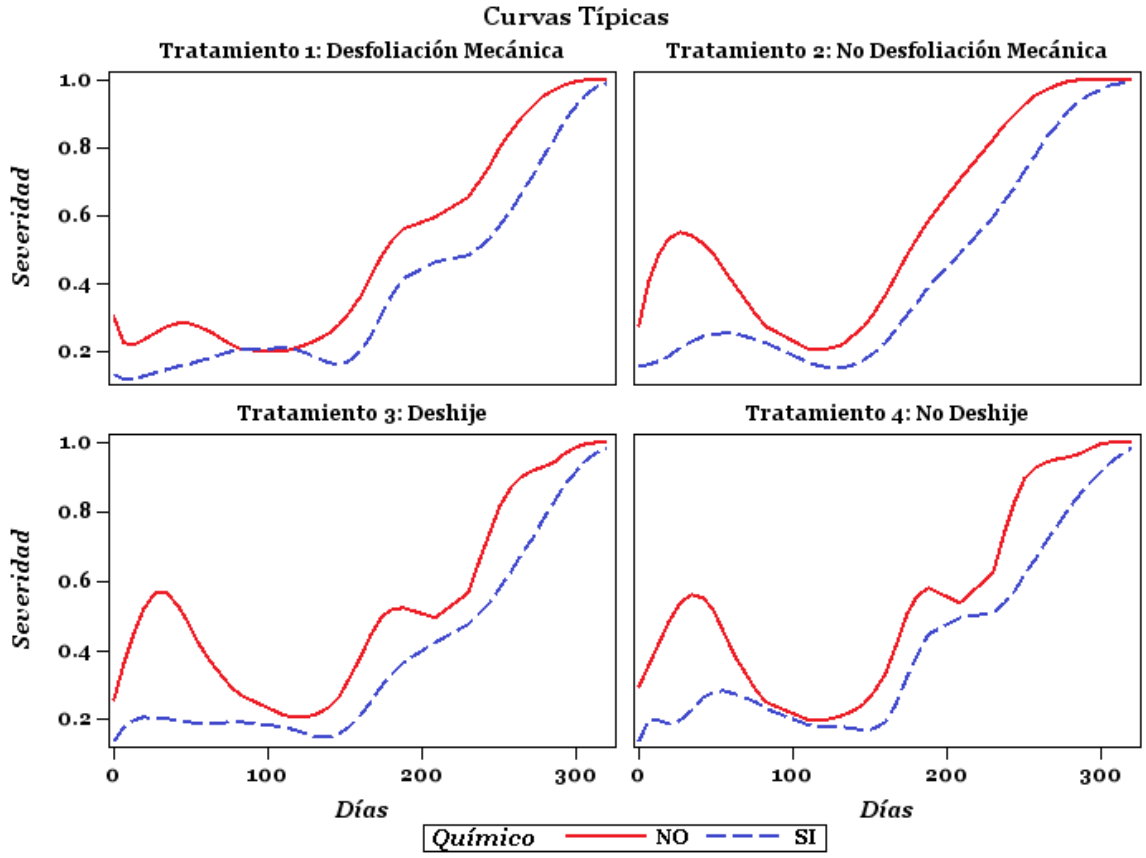

Figura 5.8: Curvas Típicas por tratamiento modelo B-spline+RS

Capítulo 6

Conclusiones Generales y Trabajos Futuros

6.1. Conclusiones Generales

La regresión Beta es de gran utilidad en el modelado de datos medidos como una proporción continua en el intervalo $(0, 1)$, ya que se cuenta con ventajas como; la facilidad de modelar diferentes formas que asumen los datos, principalmente las asimetrías; permite el modelado sin necesidad de aplicar transformaciones facilitando la interpretación de los resultados en términos de la variable de interés. Además, mediante la parametrización propuesta por Ferrari y Cribari-Neto (2004) es posible modelar los datos con varianzas heterogénas, ya que la varianza de la variable respuesta depende de la media.

En situaciones donde no se tiene una idea clara de la forma de la distribución de los datos a los que se les desea ajustar un modelo, es de vital importancia no asumir una función particular, ya que las estimaciones y resultados pueden ser muy diferentes a los reales. Una alternativa eficaz para modelar datos con estructuras complejas es a través de técnicas semiparamétricas, las cuales construyen la función de regresión a partir de las observaciones. La regresión con *Splines* es una técnica de este tipo.

De los resultados obtenidos en el proceso de simulación es posible concluir que tanto bajo un escenario de tratamientos que no presentan diferencias significativas, como un escenario donde si existan diferencias; el ajuste de un modelo de regresión semiparamétrica con combinaciones de *splines* es adecuado. Se observó que en ambos escenarios el modelo semiparámetrico B-spline+B-spline con distribución Beta y efectos aleatorios es adecuado para el ajuste, mientras que el modelo B-spline + RS se sugiere como un buen modelo para predecir.

De los resultados obtenidos en la aplicación podemos concluir que para la selección del número óptimo de nodos en el modelo B-spline+B-spline, a través del criterio de BIC, no se muestra ninguna dificultad, caso contrario con el modelo B-spline+RS el cuál no fue posible ajustar para la mayoria de los valores, esto es posiblemente debido a la estructura de los datos de severidad, ya que este problema no se presentó en el proceso simulado, pese a esto hay que tener claro que el trabajar con datos simulados es un poco más estable.

De los resultados de la aplicación, es posible que el modelo B-spline+RS se seleccionará como el "mejor" modelo para la predicción debido a los errores.

6.2. Trabajos Futuros

- En este trabajo se compararon modelos de regresión semiparamétricos asumiendo distribución Beta y efectos aleatorios normales, los modelos fueron construidos mediante la combinación de *B-splines* en la parte fija del modelo con *B-splines*, *P-splines* o Suavizado Radial en la parte aleatoria. Por lo que en un estudio posterior se podría estudiar otros modelos de regresión semi-paramétrica con distribución beta y efectos aleatorios.

- Extender el proceso de simulación a datos con más tratamientos y estudiar las características de los modelos para diferentes tamaños de muestra.

- La principal carácteristica de la regresión semi-paramétrica es construir la curva de regresión a través de las observaciones, por tal motivo se esperaría que los resultados no varien de forma significativa independientemente de la función de enlace utilizada. Un trabajo futuro es verificar esta hipótesis.

- Se debería desarrollar en forma detallada la teoría sobre la regresión semiparamétrica con Suavizado Radial, ya que la información disponible es escasa y se centra en estudios de alta dimensión.

Anexos

Anexo A: Resultados obtenidos en el ajuste del modelo B-spline+B-spline para cada uno los tratamientos en el estudio de aplicación

Resultados Tratamiento 1: Desfoliación Mecánica

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	903.89
Chi-cuadrado generalizado	569.00
Chi-cuadrado gener. / DF	1.00

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
sp_tr	planta(replica)	0.9905	0.1968
Scale		34.1481	2.4496

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	66.47	14.32	0.0003
sp_t	11	175.6	27.87	<.0001
sp_t*quimico	11	175.6	2.02	0.0286

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	1.1773	0.5293	80.06	2.22	0.0289	0.1660
Dif quim vs no quim a 50 días	0.6758	0.3201	66.29	2.11	0.0385	0.1661
Dif quim vs no quim a 100 días	-0.1089	0.3440	67.39	-0.32	**0.7526**	0.7526
Dif quim vs no quim a 150 días	0.7559	0.3443	60.65	2.20	0.0320	0.1660
Dif quim vs no quim a 200 días	0.5469	0.3315	61.07	1.65	**0.1040**	0.2743
Dif quim vs no quim a 250 días	1.0073	0.3283	65.71	3.07	0.0031	0.0229
Dif quim vs no quim a 300 días	1.0408	0.4967	93.59	2.10	0.0388	0.1661
Dif quim vs no quim a 320 días	1.6555	1.1859	149.9	1.40	**0.1648**	0.3001

Resultados Tratamiento 2: No Desfoliación Mecánica

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	596.39
Chi-cuadrado generalizado	576.00
Chi-cuadrado gener. / DF	1.00

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
sp_tr	planta(replica)	1.8647	0.2999
Scale		78.2254	5.3854

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	87.25	11.64	0.0010
sp_t	11	199	28.59	<.0001
sp_t*quimico	11	199	2.55	0.0049

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	0.9949	0.6713	90.33	1.48	**0.1418**	0.4910
Dif quim vs no quim a 50 días	0.8394	0.4204	81.31	2.00	0.0492	0.2595
Dif quim vs no quim a 100 días	0.3484	0.4448	82.61	0.78	**0.4358**	0.6778
Dif quim vs no quim a 150 días	0.6416	0.4530	80.87	1.42	**0.1605**	0.4910
Dif quim vs no quim a 200 días	0.8056	0.4459	81.23	1.81	**0.0745**	0.3248
Dif quim vs no quim a 250 días	1.2845	0.4264	85.73	3.01	0.0034	0.0240
Dif quim vs no quim a 300 días	0.7006	0.5815	109.8	1.20	**0.2309**	0.5064
Dif quim vs no quim a 320 días	0.2119	1.2301	148.9	0.17	**0.8634**	0.8634

Resultados Tratamiento 3: Deshije

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	851.22
Chi-cuadrado generalizado	598.00
Chi-cuadrado gener. / DF	1.00

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
sp_tr	planta(replica)	1.4639	0.2558
Scale		43.0386	2.9616

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	79.24	13.75	0.0004
sp_t	10	164.1	24.90	<.0001
sp_t*quimico	10	164.1	6.06	<.0001

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	0.5939	0.6092	85.62	0.97	**0.3324**	0.7273
Dif quim vs no quim a 50 días	1.2918	0.3766	71.52	3.43	0.0010	0.0077
Dif quim vs no quim a 100 días	0.2364	0.4073	77.05	0.58	**0.5634**	0.7273
Dif quim vs no quim a 150 días	0.7227	0.4045	72	1.79	**0.0782**	0.3000
Dif quim vs no quim a 200 días	0.4309	0.3978	71.38	1.08	**0.2823**	0.7273
Dif quim vs no quim a 250 días	0.9911	0.3770	73.12	2.63	0.0104	0.0630
Dif quim vs no quim a 300 días	0.3838	0.4832	99.8	0.79	**0.4290**	0.7273
Dif quim vs no quim a 320 días	2.7006	1.1942	181.6	2.26	0.0249	0.1258

Resultados Tratamiento 4: No Deshije

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	700.36
Chi-cuadrado generalizado	603.00
Chi-cuadrado gener. / DF	1.00

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
sp_tr	planta(replica)	1.5658	0.2506
Scale		62.0521	4.1697

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	84.89	8.90	0.0037
sp_t	13	259.5	24.80	<.0001
sp_t*quimico	13	259.5	8.70	<.0001

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	1.1583	0.6218	95.84	1.86	**0.0656**	0.2995
Dif quim vs no quim a 50 días	0.9395	0.3883	83.9	2.42	0.0177	0.1060
Dif quim vs no quim a 100 días	0.06529	0.4133	90.55	0.16	**0.8748**	0.8748
Dif quim vs no quim a 150 días	0.5526	0.4121	82.93	1.34	**0.1836**	0.5675
Dif quim vs no quim a 200 días	0.3117	0.4165	84.45	0.75	**0.4563**	0.8293
Dif quim vs no quim a 250 días	1.3138	0.4014	89.45	3.27	0.0015	0.0114
Dif quim vs no quim a 300 días	0.5255	0.4755	100.5	1.11	**0.2717**	0.6963
Dif quim vs no quim a 320 días	0.4092	0.8887	134.9	0.46	**0.6459**	0.8725

Anexo B: Resultados obtenidos en el ajuste del modelo B-spline+RS para cada uno los tratamientos en el estudio de aplicación

Resultados Tratamiento 1: Desfoliación Mecánica

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	884.94
Chi-cuadrado generalizado	569.00
Chi-cuadrado gener. / DF	1.00
Radial Smoother df(res)	502.82

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
Var[RSmooth(t)]	planta(replica)	0.000116	.
Scale		32.4275	2.3034

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	1	8.17	0.2142
sp_t	11	1	23.06	0.1612
sp_t*quimico	11	1	3.73	0.3853

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	1.0913	0.9978	1	1.09	0.4715	0.5970
Dif quim vs no quim a 50 días	0.6901	0.5530	1	1.25	0.4300	0.5767
Dif quim vs no quim a 100 días	-0.1254	0.3311	1	-0.38	0.7695	0.7695
Dif quim vs no quim a 150 días	0.7521	0.2181	1	3.45	0.1797	0.3392
Dif quim vs no quim a 200 días	0.5389	0.2629	1	2.05	0.2889	0.4565
Dif quim vs no quim a 250 días	1.0787	0.4478	1	2.41	0.2505	0.4237
Dif quim vs no quim a 300 días	2.6914	0.8937	1	3.01	0.2041	0.3674
Dif quim vs no quim a 320 días	5.1590	1.3993	1	3.69	0.1686	0.3253

Resultados Tratamiento 2: No Desfoliación Mecánica

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	702.85
Chi-cuadrado generalizado	582.00
Chi-cuadrado gener. / DF	1.00
Radial Smoother df(res)	476.64

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
Var[RSmooth(t)]	planta(replica)	0.000575	.
Scale		65.1222	5.7098

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	1	4.42	0.2825
sp_t	8	1	24.70	0.1545
sp_t*quimico	8	1	7.52	0.2752

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	0.7279	2.1671	1	0.34	0.7937	0.9134
Dif quim vs no quim a 50 días	0.9861	1.2063	1	0.82	0.5637	0.7058
Dif quim vs no quim a 100 días	0.1998	0.6879	1	0.29	0.8200	0.9134
Dif quim vs no quim a 150 días	0.6189	0.4349	1	1.42	0.3899	0.5771
Dif quim vs no quim a 200 días	0.8678	0.5459	1	1.59	0.3575	0.5771
Dif quim vs no quim a 250 días	1.5441	0.9525	1	1.62	0.3519	0.5771
Dif quim vs no quim a 300 días	4.4905	1.7994	1	2.50	0.2426	0.4244
Dif quim vs no quim a 320 días	7.6257	2.4614	1	3.10	0.1988	0.3600

Resultados Tratamiento 3: Deshije

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	946.36
Chi-cuadrado generalizado	604.00
Chi-cuadrado gener. / DF	1.00
Radial Smoother df(res)	519.74

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
Var[RSmooth(t)]	planta(replica)	0.000292	.
Scale		34.6064	2.7222

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	1	4.33	0.2853
sp_t	7	1	31.01	0.1374
sp_t*quimico	7	1	6.64	0.2906

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	0.6096	1.5551	1	0.39	0.7622	0.7622
Dif quim vs no quim a 50 días	1.2135	0.8629	1	1.41	0.3935	0.5949
Dif quim vs no quim a 100 días	0.3531	0.4950	1	0.71	0.6055	0.7432
Dif quim vs no quim a 150 días	0.5724	0.3129	1	1.83	0.3185	0.5186
Dif quim vs no quim a 200 días	0.7680	0.3915	1	1.96	0.3002	0.5087
Dif quim vs no quim a 250 días	0.7776	0.6786	1	1.15	0.4568	0.6711
Dif quim vs no quim a 300 días	2.5276	1.2585	1	2.01	0.2941	0.5087
Dif quim vs no quim a 320 días	4.4132	1.7760	1	2.48	0.2436	0.4351

Resultados Tratamiento 4: No Deshije

Estadísticos de ajuste	
Pseudo verosimilitud -2 Res Log	834.86
Chi-cuadrado generalizado	592.00
Chi-cuadrado gener. / DF	1.00
Radial Smoother df(res)	507.00

Estimaciones del parámetro de covarianza			
Cov Parm	Subject	Estimate	Standard Error
Var[RSmooth(t)]	planta(replica)	0.000259	.
Scale		42.8581	3.0898

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F-Valor	Pr >F
quimico	1	1	4.67	0.2760
sp_t	13	1	24.63	0.1566
sp_t*quimico	13	1	8.20	0.2675

Estimaciones Ajuste para multiplicidad: Holm-Simulated						
Etiqueta	Estimación	E.E	DF	Valor t	Pr>|t|	Adj P
Dif quim vs no quim a 0 días	0.8088	1.4659	1	0.55	0.6790	0.8195
Dif quim vs no quim a 50 días	1.2448	0.8137	1	1.53	0.3686	0.5639
Dif quim vs no quim a 100 días	0.2395	0.4821	1	0.50	0.7065	0.8195
Dif quim vs no quim a 150 días	0.7341	0.3034	1	2.42	0.2495	0.4476
Dif quim vs no quim a 200 días	0.3668	0.3752	1	0.98	0.5073	0.7359
Dif quim vs no quim a 250 días	1.2505	0.6425	1	1.95	0.3022	0.5149
Dif quim vs no quim a 300 días	2.0241	1.1896	1	1.70	0.3383	0.5617
Dif quim vs no quim a 320 días	5.5601	1.7759	1	3.13	0.1968	0.3688

Bibliografía

[Álvarez et al., 2003] Álvarez, E., Pantoja, A., Gañan, L. y Ceballos G. (2003). *La Sigatoka negra en plátano y banano: Guía para el reconocimiento y manejo de la enfermedad, aplicado a la agricultura familiar.*[disponible en: `http://www.fao.org/docrep/019/as089s/as089s.pdf`]. Accesado el 11 de febrero de 2015.

[Agresti, 2002] Agresti, A . (2002). *Categorical Data Analysis*. 2nd ed. John Wiley and Sons, Hoboken, New Jersey.

[Casella y Berger, 2002] Casella, G. y Berger, R. (2002). *Statistical Inference*. Duxbury Press, Belmont, CA.

[Durbán, 2010] Durbán M. (2010). *Introducción a los modelos mixtos*. Departamento de Estadística, Universidad Carlos III de Madrid, Madrid, España.

[Durbán, 2009] Durbán M. (2009). *Splines con Penalizaciones: Teoría y aplicaciones*. Departamento de Estadística, Universidad Carlos III de Madrid, Madrid, España.

[Cribari-Neto y Zeileis, 2010] Cribari-Neto, F. y Zeileis, A. (2010). *Beta Regression in R*. Journal of Statistical Software, $34(2) : 1 - 23$.

[Eilers y Marx, 1996] Eilers, P. y Marx, B. (1996). *Flexible Smoothing with B-splines and Penalties*. Statistical Science, $11(2) : 89 - 121$.

[Ferrari y Cribari-Neto, 2004] Ferrari, S. y Cribari-Neto, F. (2004). *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, 31(7) : 799–815.

[Fitzmaurice et al., 2004] Fitzmaurice, G., Laird, N. y Ware, J. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, New Jersey.

[García y Macchiavelli, 2012] García, Y. y Macchiavelli, R. (2012). *Modelos no lineales mixtos con variables de respuesta con distribución beta*. Tesis de Maestría, Universidad de Puerto Rico, Recinto Universitario de Mayagüez.

[Gbur et al., 2010] Gbur, E., Stroup, W., McCarter, K., Durham, S., Young, L., Christman, M., West, M. y Kramer, M. (2010). *Generalized linear models with Applications in Engineering and the Sciences*. 2nd ed. John Wiley & Sons, Hoboken, New Jersey.

[Konishi y Kitagawa, 2008] Konishi, S. y Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

[Marengo, 2010] Marengo, J. (2010).*Epidemiología de la Sigatoka Negra (Micosphaerella fijensis Morelet) en una Plantilla de Guineo en Puerto Rico*. Tesis de Maestría, Universidad de Puerto Rico, Recinto Universitario de Mayagüez.

[McCullagh y Nelder, 1989] McCullagh, P. y Nelder, J. (1989). *Generalized linear models*. 2nd ed. Chapman and Hall, New York.

[McCulloch y Searle, 2001] McCulloch, C. y Searle, S. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Hoboken, New Jersey.

[Pinheiro y Bates, 2000] Pinheiro, J. y Bates, D. (2000). *Mixed - Effect Models in S and S-plus*. Springer, New York.

[Rencher y Schaalje, 2008] Rencher, A., y Schaalje, G.(2008). *Linear models in statistics*. 2nd ed. John Wiley & Sons, Hoboken, New Jersey.

[Ruppert et al., 2003] Ruppert, D., Wand, M. y Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

[SAS Institute, 2011] SAS Institute.(2011). *SAS STAT 9.3 User's Guide: Mixed Modeling (Book Excerpt)*.Cary, NC, USA.

[Wang, 2011] Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Chapman and Hall, New York.