Trabajo Fin de Máster

Estudio de Modelos Estadísticos para la Degradación de Biomateriales

Jessica Méndez Rodríguez

Máster en Técnicas Estadísticas

Curso 2014-2015

Dirigido por:
Salvador Naya Fernández
Javier Tarrio Saavedra

Índice general

Capítulo 1

Introducción

El objetivo de este trabajo es modelizar la degradación de una familia de biopolímeros utilizados en odontología como andamios o "scaffolds"para promover la regeneración ósea después de la extracción de una pieza dental. Para ello se analizará la relación entre la pérdida de masa y variables críticas como la temperatura, el pH, el agua absorbida, etc. El diseño experimental se ha realizado en condiciones iniciales de temperatura y pH propias del cuerpo humano, con el objeto de evaluar esta familia de biopolímeros en condiciones de aplicación.

Este problema se enmarca dentro de la denominada ingeniería tisular y el interés estadístico del proyecto está en estudiar los modelos que se aplican partiendo de datos de experimentales. Uno de los objetivos es la propuesta de de modelos de regresión, tanto paramétricos como no paramétricos, para la estimación de la vida de estos materiales dentro del cuerpo humano.

El trabajo se enmcarca dentro del campo de la fiabilidad, que tiene una relación directa con el control de calidad, entendida esta fiabilidad como la calidad en el tiempo de estos biomateriales. Concretamente, el objetivo será estimar el tiempo que estos biomateriales tardan en degradarse dentro del cuerpo humano.

En este problema concreto la intención es estimar el tiempo que los materiales usados para la regeneración ósea (llamados comúnmente andamiajes o scaffolds) permanecen dentro del cuerpo del paciente, para ello se dispone de

datos reales de experimentos proporcionados por una colaboración con una empresa del sector que analiza y fabrica este tipo de materiales, *Developed-Biosystems*.

1.1. Biomateriales

En sentido amplio un biomaterial sería un material diseñado para actuar con sistemas biológicos con el fin de evaluar, tratar, aumentar o reemplazar algún tejido, órgano o función del cuerpo.

Los biomateriales están destinados a la fabricación de componentes, piezas o aparatos y sistemas médicos para su aplicación en seres vivos. Es bien sabido que la Ciencia y la Tecnología de los Biomateriales es muy reciente en cambio la investigación en este área se inició hace ya muchos siglos puesto que es posible encontrar trazas de prótesis implantadas en momias egipcias.

Sin embargo, es a partir de la segunda guerra mundial cuando, en el intento de resolver los problemas cotidianos asociados al tratamiento masivo de pacientes, aparece un vasto campo asociado a la Tecnología Médica y en concreto va tomando cuerpo una Ciencia de los Biomateriales.

Una de las características fundamentales es que deben ser biocompatibles. La biocompatibilidad se podría interpretar como la aceptabilidad biológica y el estudio de la interacción de los biomateriales con los tejidos susceptibles de estar en contacto con ellos, pero dicha propiedad no es intrínseca de un material, es decir un biomaterial no es biocompatible en cualquier condición.

La biocompatibilidad busca la aceptabilidad biológica, esta puede examinarse a varios niveles de interacción:

- La interacción entre el material y los tejidos.
- La reacción resultante de la degradación del material.
- Factores mecánicos (elasticidad, tenacidad etc.) o físicos.

Hay que tener en cuenta que existen dos tipos de biomateriales: bioinertes y bioactivos. Se llaman bioinertes a los que tienen una influencia nula o muy pequeña en los tejidos vivos que los rodean, mientras que son bioactivos

los que pueden enlazarse a los tejidos óseos vivos. Además, los biomateriales pueden ser de origen artificial (metales, cerámicas, polímeros), o biológico (colágeno, quitina, etc.).

Atendiendo a la naturaleza del material artificial con el que se fabrica un implante, se puede establecer una clasificación en materiales cerámicos (Figura 1.1), metálicos (Figura 1.2), poliméricos (Figura 1.3) o materiales compuestos. En lo que concierne a este trabajo, el interés se centra en los biomateriales poliméricos, que son ampliamente utilizados en clínicas, en implantes quirúrgicos, como en membranas protectoras, en sistemas de dosificación de fármacos o en cementos óseos acrílicos.

Figura 1.1: *Los biocerámicos son compuestos químicos complejos que contienen elementos metálicos y no metálicos. Sus principales aplicaciones están en el sistema óseo, con todo tipo de implantes y recubrimientos en prótesis articulares.*

En concreto se estudiarán varias formulaciones del ácido poliláctico (PLA). Este polímero se ha convertido en un material indispensable en la industria médica, donde se ha utilizado desde hace 25 años. El ácido poliláctico es un polímero biodegradable y bioabsorbible (que significa que puede ser asimilado por nuestro sistema biológico).

Sus características y absorbibilidad hacen del PLA un candidato ideal para implantes en el hueso o en el tejido (cirugía ortopédica, oftalmología,

Figura 1.2: *Los biomateriales metálicos se usan cuando es imprescindible soportar carga, como ocurre en las prótesis de cadera, para las que se utilizan aleaciones de cobalto (Co) con cromo (Cr) o de titanio (Ti) con aluminio (Al) y vanadio (V); el titanio también se usa en implantes dentales.*

Figura 1.3: *El desarrollo de los biopolímeros en las aplicaciones incluye prótesis faciales, partes de prótesis de oído, aplicaciones dentales; marcapasos, riñones, hígado y pulmones.*

ortodoncia, transporte de medicamentos para luchar contra las células cancerígenas), y para suturas (cirugía del ojo, cirugía del pecho y abdomen).

Las características mecánicas, farmacéuticas y de bioabsorción son dependientes de parámetros controlables tales como la composición química y

el peso molecular del polímero. El margen de tiempo para la bioabsorción del polímero puede ser de tan sólo unas semanas a algunos años y se puede regular por medio de diversas formulaciones y de la adición de radicales en sus cadenas.

Es aquí donde reside parte de la investigación de este trabajo: estudiar diferentes formulaciones del ácido poliláctico y dar con la que es mejor para el desarrollo de los "bio-andamios".

Los "Scaffolds" o "andamios" dentales son trozos de biopolímero formados para promover el crecimiento de nuevo tejido óseo. Se utiliza, por ejemplo, en odontología, después de la extracción y antes de la colocación de un implante dental de titanio.

El modelado de degradación es crucial para elegir el mejor polímero para equilibrar la degradación del andamio y el crecimiento de los huesos. El estudio de los materiales dentales incluye el análisis de su composición, propiedades e interacción con el medio ambiente.

En el momento de elegir un material dental se deben tener en cuenta además condiciones ambientales como son:

- Variación térmica: la temperatura de la cavidad oral varia entre 32 y 37^{o}C. Esta temperatura puede cambiar por la ingestión de alimentos, estados febriles, etc.
- El pH: la cavidad oral presenta un pH que va de 4 a 8.5. Con la ingestión de frutas o algunos fármacos este pH puede variar de 2 a 11.
- Esfuerzo masticatorio: al masticar los alimentos ejercen fuerzas que influyen sobre el material que se va a colocar en boca.

1.2. Fiabilidad: Modelos de degradación y pruebas aceleradas

Frecuentemente, las pruebas de fiabilidad deben realizarse con severas restricciones de tiempo y, por lo general, no se dan fallos durante esta pruebas. Por tanto es difícil evaluar la fiabilidad con este tipo de pruebas de vida

tradicionales, que registran únicamente tiempos de fallo.

Para algunos componentes pueden tomarse medidas de degradación respecto al tiempo. El uso de modelos de degradación, posible gracias a la relación existente entre el fallo del componente y la cantidad de degradación, permite hacer inferencias y predicciones sobre el tiempo de fallo.

1.2.1. Estimación de la distribución

Un modelo especificado para D(t) y Df define una distribución del tiempo de fallo. En general, esta distribución puede escribirse como una función de los parámetros del modelo de degradación.

Supongamos que una unidad falla en el instante t si el nivel de degradación alcanza primero Df en el instante t. Entonces:

$$F(t) = P(T \leq t) = P(D_t)$$

1.3. Pruebas aceleradas

El propósito principal del proceso de una prueba acelerada es alcanzar la mejora de la fiabilidad tan pronto como sea posible. Ya que no conocemos la naturaleza precisa de las debilidades futuras de un producto, debemos recurrir a la aplicación de un surtido variado de esfuerzos. La suposición básica es que sometiendo un producto a esfuerzo elevado provocará que las fallos ocurran más rápidamente.

Un ejemplo clásico de esto es una reacción química, donde se ha encontrado que la tasa de reacción se incrementa exponencialmente con la temperatura (de acuerdo con la relación de Arrhenius). Por tanto, la prueba acelerada puede considerarse como una herramienta de productividad, ya que ocurrirá un gran número de fallos en un tiempo más corto.

- Pruebas de Vida Acelerada (ALT): Se obtiene información sobre el tiempo de fallo (tiempo actual de fallo o un intervalo que contiene el tiempo de fallo) para unidades que fallaron y límites inferiores para el tiempo de fallo de las unidades que no fallaron.

- Pruebas de Degradación Acelerada (ADT) : Se observa en uno o más puntos en el tiempo, la cantidad de degradación para las unidades (tal vez con un error de medición).

1.3.1. Métodos de aceleración

- Incremento de tasa de uso del producto. Considerar la fiabilidad de una tostadora, que está diseñado para una mediana de 20 años como tiempo de vida, asumiendo una tasa de uso de 2 veces por día. Si, en su lugar, probamos el tostador 365 veces al día, podemos reducir la mediana del tiempo de vida a alrededor de 40 días. Ya que no es necesario que todas las unidades en una prueba fallen, se puede obtener información útil sobre la fiabilidad en cuestión de días en lugar de meses.

- Incremento de tasa de envejecimiento del producto. Por ejemplo, incrementando el nivel de variables experimentales como la temperatura o humedad puede acelerar el proceso químico de ciertos mecanismos de fallo, tales como degradación química (dando por resultado una debilidad eventual y fallo) de un adhesivo en una unión mecánica o el crecimiento de un filamento conductor a través de un aislante (causando eventualmente un corto circuito).

- Incremento del nivel de esfuerzo (ciclo de temperatura, voltaje o presión) bajo el cual operan las unidades de prueba. Una unidad fallará cuando su resistencia caiga por debajo del esfuerzo aplicado. Así una unidad a alto nivel de esfuerzo generalmente fallará más rápidamente que como habría fallado a bajo esfuerzo.

También pueden emplearse combinaciones de estos métodos de aceleración. Las variables como voltaje y ciclo de temperatura pueden incrementar la tasa de una reacción electromecánica (acelerando así la tasa de envejecimiento) e incrementar el esfuerzo relativo a la resistencia. En tales situaciones, cuando el efecto de una variable aceleradora es complicado, puede no existir suficiente conocimiento para proporcionar un modelo físico adecuado para la aceleración (y extrapolación). Los modelos empíricos podrdian ser útiles para la extrapolación en condiciones de uso determinadas.

La Prueba de Vida acelerada, a diferencia de la Prueba de Tortura, está diseñada para proveer información de la fiabilidad del producto, componente

o sistema.

Un dato básico es el tiempo para fallar. El tiempo de fallo puede estar en cualquier medida cuantitativa, tal como: horas, días, ciclos, actuaciones, etc. El fallo se puede deber a la fatiga mecánica, corrosión, reacción química, difusión, migración, etc. Estos son exactamente los mismos eventos conducentes a un fallo en esfuerzos mayores que en esfuerzos normales. Sólo cambia la escala del tiempo.

Un Factor de Aceleración es el multiplicador constante entre los dos niveles de esfuerzo. Cuando hay verdadera aceleración, cambiar los esfuerzos es equivalente a transformar la escala del tiempo usada para registrar cuando ocurren los fallos.

Las transformaciones usadas comúnmente son lineales, lo que significa que el tiempo para fallar en un esfuerzo alto sólo tiene que ser multiplicado por una constante (el factor de aceleración) para obtener el tiempo equivalente de fallo en el esfuerzo de uso.

Relaciones Lineales de Aceleración:

- Tiempo de fallo $t_u = AFxt_s$
- Probabilidad de fallo $Fu(t) = Fs(t/AF)$
- fiabilidad $Ru(t) = Rs(t/AF)$
- PDF o Función de Densidad $fu(t) = (1/AF)fs(t/AF)$
- Tasa de fallo $lu(t) = (1/AF)ls(t/AF)$

donde:

tu: tiempo de fallo en uso ts: tiempo de fallo en esfuerzo
Fu(t): CDF en uso Fs(t): CDF en esfuerzo
fu(t): PDF en uso fs(t): PDF en esfuerzo
lu(t): tasa de fallo en uso ls(t): tasa de fallo en esfuerzo
AF: Factor de Aceleración

Algunas consecuencias de las relaciones lineales es que:

- El parámetro de forma para los modelos clave de distribución de vida (Weibull y Lognormal) no cambia para las unidades operando bajo diferentes esfuerzos.
- Las gráficas en escala de probabilidad de los datos de diferentes condiciones de esfuerzo se alinearán aproximadamente paralelas.
- Los modelos de aceleración predicen el tiempo de fallo en función del esfuerzo.
- Los factores de aceleración muestran como el tiempo de fallo de un nivel particular de esfuerzo (para un modo o mecanismo de fallo) puede ser usado para predecir el tiempo equivalente de fallo en un nivel diferente de esfuerzo.

Un modelo que predice el tiempo de fallo como función del esfuerzo debiera ser mejor que una colección de factores de aceleración. Si escribimos $tf = G(S)$, donde $G(S)$ es la ecuación del modelo para un valor arbitrario de S, entonces el factor de aceleración entre los esfuerzos S1 y S2 puede evaluarse simplemente por:

$AF = G(S1)/G(S2)$

Ahora se puede probar, en el nivel de esfuerzo más alto S2, obtener un número suficiente de fallos para ajustar al modelo de distribución de vida y evaluar las tasas de fallo. Después se usa la Tabla de Relaciones Lineales de Aceleración para predecir lo que pasará en el nivel de esfuerzo menor S1.

Los modelos de aceleración se derivan a menudo de modelos físicos o cinéticos relacionados con el modo del fallo. Un modelo que predice el tiempo de fallo como función de los esfuerzos de operación se conoce como Modelo de Aceleración.

Algunos modelos útiles son los siguientes:

- Arrhenius
- Eyring

- Regla de Potencia Inversa para Voltaje
- Modelo exponencial de Voltaje
- Modelos de Dos: Temperatura / Voltaje o Temperatura/Humedad
- Modelo de Electromigración (Temperatura y Densidad)
- Modelos de tres esfuerzos (Temperatura, Voltaje y Humedad)
- Modelo Coffin-Manson de Crecimiento de Fracturas Mecánicas

1.3.2. El Modelo de Arrhenius

El Modelo de Arrhenius predice la aceleración de las fallos debido al aumento de temperatura. Es una de las primeras transformaciones y la de más éxito para predecir como varía el tiempo de fallo con la temperatura.

$$\begin{aligned} t_f &= E_a \exp\left(\frac{\triangle H}{kT}\right) \\ AF &= \exp\left(\frac{\triangle H}{k}\left[\frac{1}{T_1} - \frac{1}{T_2}\right]\right) \end{aligned}$$

Donde: AF = Factor de Aceleración

T= temperatura °K (273.16+°C)

k = Constante de Boltzmann (8.617E-05 eV/K)

$\triangle H$ = Energía de Activación del Modo de fallo (eV)

E_a = Constante de escala (se elimina en AF)

$\triangle H$ determina la pendiente del factor de aceleración con la temperatura. Una $\triangle H$ pequeña caracteriza una tasa de fallo que no es fuertemente dependiente de la temperatura. $\triangle H \geq 0{,}1eV$ significa fuerte dependencia de la temperatura.

Capítulo 2

Descripción del estudio

2.1. Objetivos del estudio

El objetivo principal del estudio es estimar la tendencia de degradación característica de algunos biopolímeros específicos para definir apropiadamente su futura aplicación como andamios.

Para ello se estudiará la degradación por hidrólisis del biopolímero en una solución con pH=7.4 con condiciones isotérmicas (37ºC) con el fin de modelar la pérdida de masa usando variables críticas.

El procedimiento a seguir será obtener y comparar la trayectoria de degradación para cada formulación de polímeros y así poder decidir, en base a las necesidades planteadas, cual es el polímero más fiable.

Además se realizará un estudio de análisis funcional de las curvas calorimétricas obtenidas, con el fin de proporcionar una alternativa al modelo de degradación propuesto. De este modo, si los resultados son los esperados y puede hallarse una buena relación entre las curvas calorimétricas y la pérdida de masa, se estaría ante un modo más eficiente de obtener predicciones.

2.2. Presentación de los biomateriales

Como ya se ha mencionado anteriormente, los biomateriales elegidos para este estudio son polímeros. En concreto polímeros del tipo ácido poli-(D,L-láctico-coglicólico).

En la Figura 4.1 se muestra la composición y estructura química de los mismos.

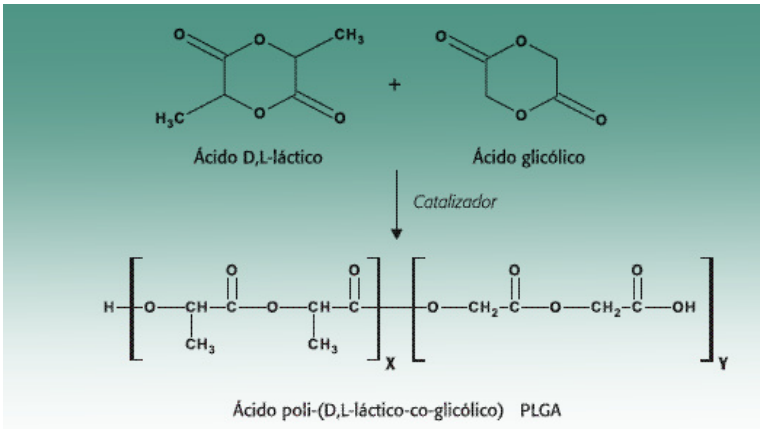

Figura 2.1: Composición y estructura química de los ácidos poli-(D,L-láctico-co-glicólico).

A continuación se presenta la Tabla 2.1 con los nombres y algunas características de los polímeros seleccionados para el estudio.

Se puede observar que todos son del tipo amorfo. Esto quiere decir que, a diferencia de los compuestos cristalinos (que presentan un estructura ordenada), las moléculas de los sólidos amorfos están distribuidas azarosamente y sus propiedades físicas son idénticas en todas las direcciones (isotropía). Constan de una temperatura característica conocida como *Temperatura de transición vítrea* (Tg) donde sus propiedades suelen experimentar cambios importantes.

Nombre	Polímero	Tipo	Peso Molecular
PDL 02	Poly(DL-lactide)	Amorfo	17
PDL 02A	Poly(DL-lactide)	Amorfo	17
PDLG 5002	50/50 DL-lactide/glycolide copolymer	Amorfo	17
PDLG 5004	50/50 DL-lactide/glycolide copolymer	Amorfo	17
PDLG 5010	50/50 DL-lactide/glycolide copolymer	Amorfo	17
PDLG 7502	50/50 DL-lactide/glycolide copolymer	Amorfo	17
PDLG 7502A	50/50 DL-lactide/glycolide copolymer	Amorfo	17

Cuadro 2.1: Nombre y propiedades de los polímeros incluídos en el estudio.

Una de las consecuencias que experimentan los sólido amorfos debido a la disposición de sus partículas, es la diferencia de intensidad que toman las fuerzas intermoleculares entre las mismas, alcanzándose la fusión a distintas temperaturas según la proporción de sus partículas, deduciéndose que estos no tienen un punto de fusión definido.

2.3. Procedimiento experimental

En primer lugar se procede a la preparación de los materiales. Se pesan las muestras de 0.3 gramos de los 7 polímeros PDLGA diferentes en frascos con 1ml de disolución salina taponada con fosfato. Dicha disolución tiene un pH 7.4 para simular las condiciones de la cavidad oral humana. Con el mismo fin se cierran y se mantienen a una temperatura de 37ºC.

Se dispone de 2 réplicas por cada polímero PDLGA, compuestas a lo sumo de 15 muestras cada una. Cada 3 días aproximadamente, se abre 1 muestra de cada polímero y se mide el pH de la disolución. Se pesa la muestra de polímero húmedo y se obtiene la masa de agua absorbida.

Posteriormente la muestra resultante se seca en un horno a 37-40ºC durante 24h. Se aprovecha este tiempo para obtener una curva calo-

rimétrica así como la temperatura de transición vítrea (Tg). Se pesa la masa y se obtiene la pérdida de masa del polímero.

Capítulo 3

Modelización de la degradación de los polímeros

En primer lugar se realiza un breve estudio descriptivo de las posibles variables de interés. En la Figura 3.1, que se muestra a continuación, se presetan las densidades de cada variable así como una nube de puntos que refleja la relación entre cada par de variables.

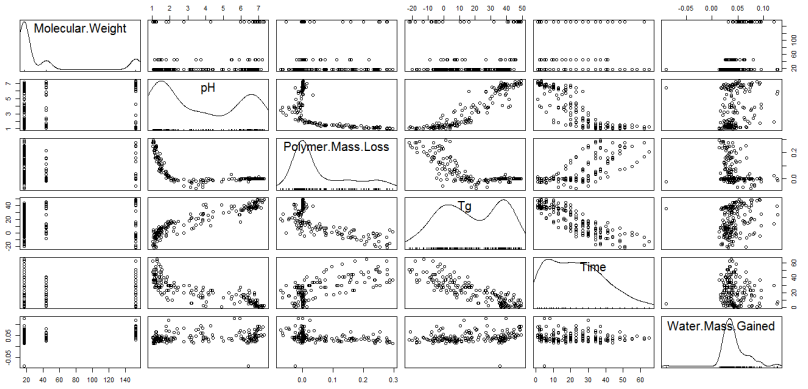

Figura 3.1: Relaciones entre las posibles variables de interés del estudio.

Se puede ver como una de las relaciones más acusadas es la existente entre la pérdida de masa y el pH. Este hecho se estudiará posterior-

mente con más detalle.

Llegados a este punto conviene recordar que el objetivo último del estudio es modelizar la degradación. Esta degradación viene definida en este caso por la pérdida de masa, por lo que en primer lugar se tratará de ver qué variables explican mejor dicha pérdida de masa.

3.1. Modelo Aditivo Generalizado (GAM)

3.1.1. Introducción

Los **Modelos Lineales Generalizados** (**GLM**, del inglés *Generalized Linear Models*), introducidos a principios de los años 70 ([1] Nelder y Wedderbur, 1972), sintetizan dentro de un marco homogéneo un conjunto muy amplio de métodos de regresión y se han convertido en una de las principales herramientas de análisis estadístico en toda clase de áreas.

La utilización de un mismo enfoque general para formular una serie de problema en principio heterogéneos no solo resulta interesante desde un punto de vista teórico, sino que permite trasladar con facilidad de unos a otros de estos problemas tanto algoritmos de estimación como software y métodos de cálculo numérico, herramientas de diagnóstico de resultados, etc.

Todo ello explica la amplia difusión que han tenido los GLM hasta el punto de convertirse en el enfoque estándar para abordar los métodos de regresión para los que reulsta aplicable, especialmente desde la aparición de la obra de referencia sobre la materia, McCullagh y Nelder, 1989 ([2]).

Los GLM constituyen una generalización de los clásicos **Modelos Lineales** (**LM** - del ingés *Linear Model*). En estos modelos se asume que el valor esperado de la variable dependiente condicionado a los valores

de las variables independientes se puede expresar como una combinación lineal de esos valores de las variables independientes:

$$E(y/x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \eta$$

Los parámetros desconociidos $(\beta_1, \beta_2, \ldots, \beta_p)$ de denominan **coeficientes de regresión** y determinan la *fuerza* y *dirección* de la influencia de la cada una de las covariables en la respuesta.
Puesto que la ecuación es lineal en los coeficientes de regresión, la suma de los efectos de las covariables, η, se denomina **predictor lineal**.

La diferencia entre los valores observados de y y los estimados por medio del predictor lineal serán unos términos de error ϵ que sigue una distribución gaussiana de media cero y varianza constsante σ^2. El papel privilegiado que juega en los LM el término de error se pierde en alguna medida en los GLM. Para formular éstos se parte del hecho de que para un LM la variable dependiente, condicionada a los valores de las x_i, sigue una distribución de probabilidad normal de media dada por el predictor lineal y varianza σ^2.

$$y|x_i \; N(\eta, \sigma^2)$$

Un GLM generaliza la expresión anterior en tres sentidos.

- Por una parte, la distribución de la variable dependiente condicionada a los valores de las independientes ya no tiene que ser normal, sino que puede ser cualquiera perteneciente a la familia exponencial, que incluye no solo a la normal, sino a muchas otras de las más usadas en las aplicaciones, como la binomial, Poisson, Gamma, etc.
- En el caso de respuestas **no gaussianas**, no es posible una conexión directa entre el valor esperado de y y el predictor lineal, η, puesto que el dominio de $E(y/x_1, x_2, \ldots, x_p)$ ya no es la recta real. Por tanto se necesita un transformación, h, para asegurar el dominio correcto de dicho valor esperado:

$$E(y/x_1, x_2, \ldots, x_p) = h(\eta)$$

Esto es, denotando por $\mu = E(y(/x_1, x_2, \ldots, x_p)$:

$$\mu = h(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) = h(\eta)$$

o equivalentemente

$$g(\mu) = \eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \beta_0 + \sum_{j=1}^{p} \eta_j$$

siendo

$g = h^{-1} =$ función link
$\eta =$ predictor lineal
$\eta_j = \beta_j x_j =$ efecto parcial (lineal) de $x_j \quad j = 1, \ldots, p$

- Por último, la varianza de $y|x_i$ deja de ser una constante σ^2 y se convierte en una función de μ:

  $Var(Y) = \phi \times V(\mu), \quad V$ una función de $\mu, \quad \phi =$ parámetros de dispersión.

Este plantemiento, es, como se puede ver, mucho más amplio y general que el de los modelos LM, pero también, como cabe esperar, complica la estimación de los modelos, ya que no puede hacerse por mínimos cuadrados (como en el caso de los LM), sino que requiere de una estimación por máxima verosimilitud a través de un procedimiento iterativo de tipo Newton-Rapson denominado *Iteratively Reweighted Least Squares* (McCullagh y Nelder, 1989 [2]).

Un exposición detallada pero muy accesible de todo lo anterior puede verse en Dobson, 2002 ([3]).

Pese a su notable generalidad, los GLM no están exentos de limitaciones. Una de las más relevantes deriva de su carácter de modelos "lineales". Las variables explicativas x_i entran en el momdelo a través del predictor lineal η que no es más que una combinación lineal de esas

variables explicativas: $\eta = \sum \beta_i x_i$.

Para las covariables continuas en el modelo la suposición de un efecto estrictamente lineal en el predictor puede no ser apropiada, por ejemplo:

- Algunos efectos pueden tener una forma desconocida.
- Las interacciones entre las covariables pueden adoptar una forma compleja.

Una alternativa a los GLM viene dadad por los **Modelos Aditivos Generalizados** (**GAM** - del inglés *Generalized Additive Models*) introducidos por Hastie y Tibshirani (1990).

Los GAM son una generalización de los GLM de los que se diferencian únicamente porque el predictor lineal ya no es simplemente una combinación lineal de las variables explicativas, sino que es una combinación lineal de funciones de dichas variables explicativas, lo que permite introducir en el modelo todo tipo de efectos y relaciones no lineales entre variables:

$$\eta = \mathbf{X}^*\mathbf{\Theta} + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

donde:

$\mathbf{X}^*\mathbf{\Theta}$ correponde a la parte estrictamente paramétrica del modelo (es decir, término independiente, factores, efectos lineales,...)

$f_j(x_j)$ = efecto parcial suave (desconocido) de x_j en el predictor.

La opción más frecuente para las funciones f_i consiste en utilizar *splines* ([4] Wood, 2006), pero no es la única. Las características concretas de las funciones f_i deben ser calculadas como parte del proceso de estimación del modelo.

Con esto el modelo pierde su carácter puramente paramétrico (ya no solo es necesario estimar los parámetros β_i), con lo que es frecuente referirse a los GAM como modelos "semiparamétricos"([5] Wang, 2011),

y suponen un buen compromiso entre los modelos lineales puramente paramétricos (fáciles de construir e interpretar, pero poco flexibles) y modelos totalmente no-paramétricos en los que las f_i serían completamente arbitrarias, lo que supondria modelos muy flexibles, pero difíciles de estimar y de interpretar ([6] Faraway, 2006).

3.1.2. Modelos Aditivos Generalizados

Los Modelos Aditivos Generalizados son modelos de regresión similares a los GLM pero en los que $Y|x_i$ sigue una distribución de la familia exponencial en la que la media viene dada por $\mu = h^{-1}\left(\sum_i \beta_i f_i(x_i)\right)$.

Las f_i son funciones suaves que permiten reflejar efectos no lineales de las variables x_i sobre la variable Y. Una solución sencilla para tener en cuenta estos efectos no lineales hubiese sido incorporar en el predictor lineal términos cuadráticos, cúbicos, etc., de las variables explicativas (es decir, términos de la forma x^2, x^3, etc.) de forma que el predictor deje de ser lineal y se convierta en un polinomio de las variables explicativas.

Sin embargo, esta solución plantea problemas bien conocidos derivados del hecho de que los polinomios son funciones de soporte acotado (su dominio se extiende sobre todo el eje real), de manera que, en general, cualquier intento de mejorar su ajuste en un punto determinado se consigue a expensas de empeorarlo en otros puntos muy alejados. Estos supone que los polinomios pueden proporcionar una solución adecuada cuando se busca un buen ajuste en el entorno de un punto concreto, pero no sobre todo un intervalo.

Una alternativa ampliamente utilizada consiste en emplear *splines* como se ha mencionado anteriormente. Los splines son polinomios definidos sobre intervalos y que toman un valor nulo fuera de esos intervalos. Con esto se consigue graduar el ajuste de la regresión de forma que los cambios que se produzcan para mejorar ese ajuste tengan un efecto local y no se extiendan más allá de los intervalos en los que están defi-

nidos los *splines* involucrados en cada caso.

Los intervalos sobre los que se definen los splines vienen determinados por un conjunto de puntos denominados *knots*. Un conjunto de $q-2$ knots determinan $q-1$ intervalos, que pueden estar contenidos dentro de un intervalo acotado [a,b] o extenderse hasta $-\infty$ y ∞. Habitualmente se utilizan splines cúbicos, que son polinomios de grado 3, con lo que cada uno de ellos queda determinado por 4 parámetros.

En consecuencia el espacio vectorial de los splines formados por combinaciones lineales de polinomios cúbicos definidos sobre cada uno de los $q-1$ intervalos determinados por $q-2$ knots y nulos en el resto de intervalos, es un espacio de dimensión $4(q-1)$. Sin embargo, se introducen una serie de restricciones que reducen la dimensión de este espacio. Estas restricciones consisten en que se exige que los splines y sus primeras y segundas derivadas sean continuas en los $q-2$ knots. Son entonces $3(q-2)$ condiciones que hacen que la dimensión del espacio vectorial de los splines cúbicos definidos sobre el intervalo $[a,b]$ sea $4(q-1)-3(q-2)=q+2$.

Para los splines denominados "naturales" también es habitual exigir que su derivada segunda se anule en los extremos del intervalo $[a,b]$, lo que da lugar a otras dos restricciones adicionales que hacen que finalmente la dimensión del espacio vectorial de los splines naturales cúbicos definidos por $q-2$ knots sea q.

3.1.3. Aplicación del GAM para la estimación de la pérdida de masa

El modelo GAM utilizado para la estimación de la pérdida de masa ha sido el siguiente:

$$\eta_x = \mu_x = \beta_0 + f_1(Time) + f_2(pH) + Polymer$$

En la Figura 3.2 se muestran los efectos del tiempo y el pH en la variable pérdida de masa, así como los de la variable polímero. Se puede

observar que el efecto más influyente en la pérdida de masa es el del pH.

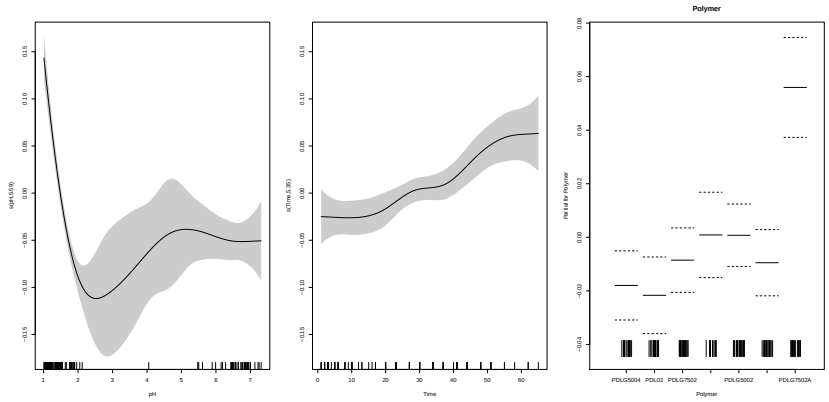

Figura 3.2: *Efectos del tiempo, el pH y la variable polímero en la pérdida de masa.*

Además el efecto del polímero también parece es también influyente puesto que el *PDLG7502A* presenta niveles muy diferenciados del resto.

En la Figura 3.3 se muestran las gráficas de vallidación de los residuos y parecen cumplir las hipótesis del modelo.

A continuación se estudiarán las relaciones entre las variables presentes en el modelo anterior. Para ello se utilizarán modelos no lineales de efectos mixtos (nlme). La aplicación de modelos NLME permite comparar el camino de degradación de la familia de polímeros estudiada, pudiendo estimar la variabilidad debida al tipo de biopolímero con un modelo más parsimonioso que el correspondiente a la aplicación de la regresión no lineal.

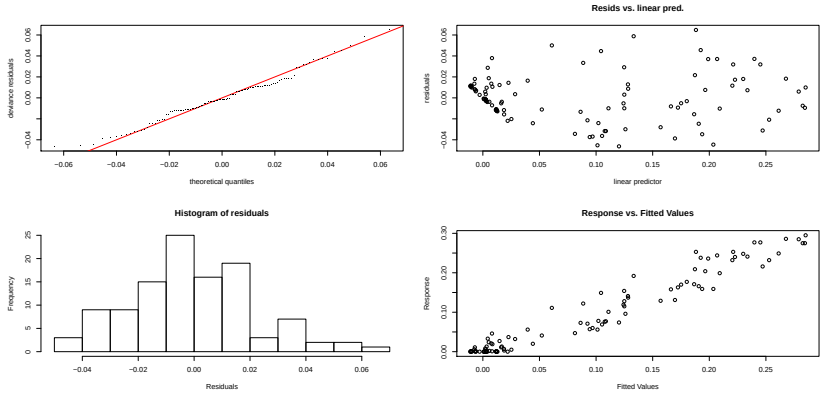

Figura 3.3: *Análisis de residuos del modelo GAM.*

3.2. Modelos No Lineales de Efectos Mixtos (NLME)

La primera pregunta que cabe plantearse sobre los modelos NLME (y posiblemente la más importante) es ¿por qué usarlos?

Esta pregunta, por supuesto, también se aplica a los modelos de regresión no lineal en general, al igual que la respuesta: interpretabilidad y validez más allá del rango de valores de los datos observados. Al elegir un modelo de regresión para describir como una variable respuesta varía en función de las covariables, siempre se tiene la opción de utilizar modelos como los polinomiales, que son lineales en los parámetros. Al aumentar el orden de un modelo polinomial, se puede obtener aproximaciones cada vez más precisas dentro del rango observado de los datos. Estos modelos empíricos se basan únicamente en la relación observada entre la respuesta y las covariables y no incluyen consideraciones teóricas sobre el mecanismo subyacente de los datos.

Los modelos no lineales, por otra parte, son a menudo mecánicos, es decir, basados en un modelo para la producción mecánica de una respuesta. Como consecuencia, los parámetros del modelo en un modelo no lineal tienen generalmente una interpretación física natural. Incluso

cuando los parámetros derivan empíricamente de los datos, los modelos no lineales suelen incorporar características teóricas conocidas de los datos, tales como asíntotas y monotonía, y en estos casos, pueden ser considerados como modelos semi-mecánicos.

Además los modelos no lineales también proporcionan predicciones más fiables de la variable de respuesta fuera del rango de valores observado de los datos que, por ejemplo, los modelos polinomiales.

Hay muchas más similitudes que diferencias entre los modelos LME y NLME. Ambos modelos se utilizan con datos agrupados y tienen el mismo propósito: describir una variable de respuesta en función de las covariables, teniendo en cuenta la correlación entre las observaciones en el mismo grupo. Los efectos aleatorios se utilizan para representar la dependencia dentro de los grupos en los dos modelos, y los supuestos sobre los efectos aleatorios y los errores dentro de los grupos son idénticos en ambos.

Los NLME son modelos de efectos mixtos en los que algunos (o todos) de los efectos fijos y aleatorios ocurren de forma no lineal en el modelo funcional. Pueden ser considerados bien como una extensión de los modelos lineales de efectos mixtos, o como una extensión de modelos de regresión no lineal para datos independientes ([7] Bates y Watts, 1988) en el que los efectos aleatorios se incorporan a los coeficientes para permitir la variación por grupo, induciendo así la correlación dentro de los grupos.

A continuación se presentará una formulación general de modelos NLME propuestas por Lindstrom y Bates (1990). Se describirá tanto el modelo NLME para datos agrupados de un solo nivel, que incluye medidas repetidas y los datos longitudinales, como el modelo NLME multinivel.

Un solo nivel de agrupación

La aplicación más común de los modelos NLME es para datos con medidas repetidas, más en particular para datos longitudinales. EL modelo de efectos mixtos no lineales propuesto por Lindstrom y Bates, 1990 ([8]) puede ser considerado como un modelo jerárquico.

En un primer nivel la observación j-ésima en el grupo i-ésimo se modela como:

$$y_{ij} = f(\phi_{ij}, \nu_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i, \tag{3.1}$$

donde M es el número de grupos, n_i es el número de observaciones en el grupo i-ésimo, f es una función general diferenciable, de valor real, de un grupo específico de vector de parámetros ϕ_{ij} y un vector de covariables ν_{ij}, y ϵ_{ij} es un término de error normalmente distribuído dentro del grupo.

La función f es no lineal al menos en una componente del vector específico del grupo, ϕ_{ij}, que se modela como:

$$\phi_{ij} = A_{ij}\beta + B_{ij}b_i, \quad b_i \backsim N(0, \Psi), \tag{3.2}$$

donde β es un vector p-dimensional de *efectos fijos* y b_i es un vector q-dimensional de *efectos aleatorios* asociado al grupo i-ésimo (no variable con la j) cuya matriz de varianzas y covarianzas es Ψ. Las matrices A_{ij} y B_{ij} son de dimensiones apropiadas y dependen del grupo y posiblemente de los valores de algunas covariables de la j-ésima observación.

Este modelo es una ligera generalización de la descrita en Lindstrom y Bates (1990) en la que A_{ij} y B_{ij} pueden depender de j. Esta generalización permite la incorporación de variables dependientes del tiempo tanto en los efectos fijos como en los aleatorios del modelo.

Se supone que las observaciones correspondientes a diferentes grupos son independientes y que los errores dentro de los grupos los errores ϵ_{ij} se distribuyen de forma independientes según una $N(0, \sigma^2)$ y son

independientes de los b_i.

Debido a que f puede ser una función no lineal de ϕ_{ij}, la representación de los coeficientes específicos de cada grupo, ϕ_{ij}, se eligen de modo que A_{ij} y B_{ij} siempre son matrices de incidencia simples.

Las ecuaciones 3.1 y 3.2 se puede escribir en forma matricial del siguiente modo:

$$y_i = f_i(\phi_i, \nu_i) + \epsilon_i, \tag{3.3}$$
$$\phi_i = A_i\beta + B_i b_i, \tag{3.4}$$

para $i = 1, \ldots, M$, donde

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \phi_i = \begin{bmatrix} \phi_{i1} \\ \vdots \\ \phi_{in_i} \end{bmatrix}, \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}, f_i(\phi_i, \nu_i) = \begin{bmatrix} f(\phi_{i1}, \nu_{i1}) \\ \vdots \\ f(\phi_{in_i}, \nu_{in_i}) \end{bmatrix},$$
$$\nu_i = \begin{bmatrix} \nu_{i1} \\ \vdots \\ \nu_{in_i} \end{bmatrix}, A_i = \begin{bmatrix} A_{i1} \\ \vdots \\ A_{in_i} \end{bmatrix}, B_i = \begin{bmatrix} B_{i1} \\ \vdots \\ B_{in_i} \end{bmatrix}$$

Varios niveles de agrupación

El modelo NLME de un solo nivel (3.1) se pueede extender a datos agrupados de acuerdo a varios factores anidados mediante la modificación del modelo para los efectos aleatorios (3.2). Por ejemplo, la versión de varios niveles de agrupación de Lindstrom y Bates (1990) para dos niveles anidados se escribe como un modelo de dos etapas en el que la primera etapa expresa la respuesta para la k-ésima observación en el grupo j-ésimo de segundo nivel del grupo i-ésimo de primer nivel:

$$y_{ijk} = f(\phi_{ijk}, \nu_{ijk}) + \epsilon_{ijk}, \tag{3.5}$$
$$i = 1, \ldots, M, \quad j = 1, \ldots, M_i, \quad k = 1, \ldots, n_{ij},$$

donde M es el número de grupos de primer nivel, M_i es el número de grupos de segundo nivel en el grupo i-ésimo de primer nivel, n_{ij} es el número de observaciones en el j-ésimo grupo de segundo nivel del i-ésimo grupo de primer nivel, y ϵ_{ijk} es el error que se distribuye normalmente dentro de cada grupo.

Como en el modelo de un solo nivel, f es, en general, una función diferenciable de valor real de un vector de parámetros ϕ_{ijk} específico de cada grupo y un vector de covariables ν_{ijk}. f es no lineal, al menos, en una componente de ϕ_{ij}.

La segunda etapa del modelo expresa ϕ_{ij} como:

$$\begin{aligned} \phi_{ijk} &= A_{ijk}\beta + B_{i,jk}b_i + B_{ijk}b_{ij}, \\ b_i &\backsim N(0, \Psi_1), \quad b_{ij} \backsim N(0, \Psi_2). \end{aligned} \tag{3.6}$$

Como en el modelo de un solo nivel (3.2), β es un vector p-dimensional de efectos fijos, con una matriz de diseño A_{ijk}, la cual debe incorporar covariables dependientes del tiempo. Los efectos aleatorios de primer nivel, b_i, son vectores q_1-dimensionales distribuídos de forma independiente, con matriz de varianzas y covarianzas Ψ_1. Los efectos aleatorios de segundo nivel, b_{ij}, son vectores q_2-dimensionales distribuídos independientemente, con matriz de variazans y covarianzas Ψ_2. Se asume que son independientes de los efectos aleatorios de primer nivel.

Las matrices de diseño de los efectos aleatorios, $B_{i,jk}$ y B_{ijk}, dependen de los grupos de primer y segundo nivel y posiblemente de los valores de algunas covariables de la k-ésima observación.

Los errores dentro de cada grupo, ϵ_{ijk}, se distribuyen de forma independiente según una $N(0, \sigma)$ y son independientes de los efectos aleatorios. Las hipóotesis de independencia y homocedasticidad de los errores dentro de cada grupo pueden relajarse como se muestra en 3.4.

(3.5) y (3.6) se pueden expresar en forma matricial del siguiente modo:

$$y_{ij} = f_{ij}(\phi_{ij}, \nu_{ij}) + \epsilon_{ij}, \tag{3.7}$$
$$\phi_{ij} = A_{ij}\beta + B_{i,j}b_i + B_{ij}b_{ij}, \tag{3.8}$$

para $i = 1, \ldots, M$, $j = 1, \ldots, M_i$, donde

$$y_{ij} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijn_{ij}} \end{bmatrix}, \phi_{ij} = \begin{bmatrix} \phi_{ij1} \\ \vdots \\ \phi_{ijn_{ij}} \end{bmatrix}, \epsilon_{ij} = \begin{bmatrix} \epsilon_{ij1} \\ \vdots \\ \epsilon_{ijn_{ij}} \end{bmatrix},$$
$$f_{ij}(\phi_{ij}, \nu_{ij}) = \begin{bmatrix} f(\phi_{ij1}, \nu_{ij1}) \\ \vdots \\ f(\phi_{ijn_{ij}}, \nu_{ijn_{ij}}) \end{bmatrix}, \nu_{ij} = \begin{bmatrix} \nu_{ij1} \\ \vdots \\ \nu_{ijn_{ij}} \end{bmatrix},$$
$$A_{ij} = \begin{bmatrix} A_{ij1} \\ \vdots \\ A_{ijn_{ij}} \end{bmatrix}, B_{i,j} = \begin{bmatrix} B_{i,j1} \\ \vdots \\ B_{i,jn_{ij}} \end{bmatrix}, B_{ij} = \begin{bmatrix} B_{ij1} \\ \vdots \\ B_{ijn_{ij}} \end{bmatrix}$$

La extensión de los modelos NLME a más de dos niveles de anidado es sencilla. Por ejemplo, con tres niveles de agrupación anidada, la segunda etapa del modelo para los coeficientes específicos de cada grupo sería:

$$\phi_{ijkl} = A_{ijkl}\beta + B_{i,jkl}b_i + B_{ij,kl}b_{ij} + B_{ijkl}b_{ijk},$$
$$b_i \backsim N(0, Psi_1), \quad b_{ij} \backsim N(0, \Psi_2), \quad b_{ijk} \backsim N(0, \Psi_3).$$

3.2.1. NLME para MassLoss vs. Time

En primer lugar se hará un representación gráfica de los datos. En la Figura 3.4 se muestran dos gráficos, en el de la izquierda está representada la pérdida de masa frente al tiempo por graficas según el tipo de polímero y en la de la derecha sin hacer distinción de gráfica. Parece que se observan diferencias en el comportamiento según el tipo de polímero.

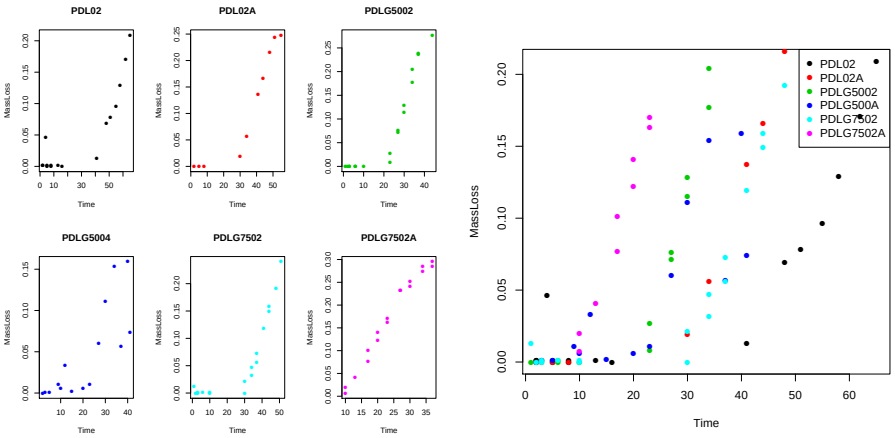

Figura 3.4: *Gráficas de la pérdida de masa frente al tiempo para los diferentes tipos de polímeros.*

En este caso se utilizará un modelo de regresión logistica "S-Shaped". Se trata de un modelo logístico de 4 parámetros en el que se tiene una respuesta y que depende de x:

$$y(x) = \phi_1 + \frac{\phi_2 - \phi_1}{1 + \exp[(\phi_3 - x)/\phi_4]} \tag{3.9}$$

Se tiene que ϕ_4 deberá ser mayor que 0 y entonces:

- ϕ_1 es la asíntota horizontal cuando $x \to \infty$
- ϕ_2 es la asíntota horizontal cuando $x \to -\infty$
- ϕ_3 es el valor de x en el punto de inflexión. En este valor de x la respuesta es el punto medio de las asintotas.
- ϕ_4 es el parámetro de escala en el eje x. Cuando $x = \phi_3 + \phi_4$ la respuesta es $\phi_1 + (\phi_2 - \phi_1)/(1 + e^{-1})$ o aproximadamente tres cuartas partes de la distancia entre ϕ_1 y ϕ_2.

Estos parámetros se muestran en la Figura 3.5.

En este cacso:

$$\text{MassLoss}_{ij} = \phi_{1i} + \frac{\phi_2 i - \phi_1 i}{1 + \exp\left(\frac{\text{Time}_i j - \phi_3 i}{\phi_{4i}}\right)} + \epsilon_{ij}$$

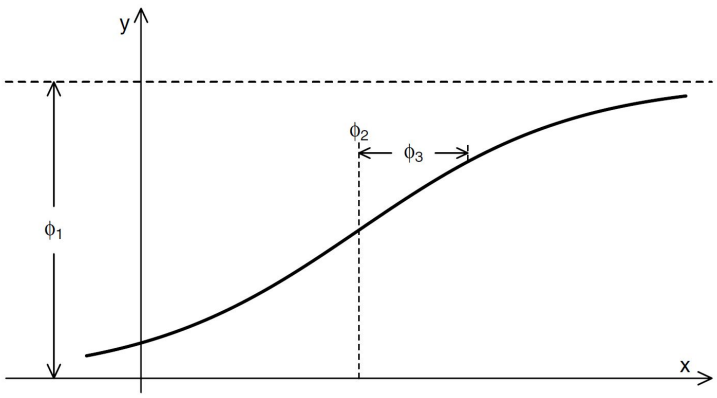

Figura 3.5: *El modelo logístico simple muestra los parámetros ϕ_1, la asíntota horizontal cuando $x \to \infty$; ϕ_2, el valor de x para el cual $y = \phi 1/2$; y ϕ_3, el parámetro de escala del eje x. Si $\phi_3 < 0$ la curvá será monótona decreciente en lugar de monótona creciente y ϕ_1 será la asíntota horizontal cuando $x \to \infty$.*

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} = \beta + b_i, \quad b_i \backsim N(0, \Psi), \quad \epsilon_{ij} \backsim N(0, \sigma^2)$$

donde i es el tipo de polímero y j cada observación. Además ϕ_{1i} es la pérdida de masa inicial, ϕ_{2i} es la pérdida de masa final, que se fija en 0.3, ϕ_{3i} representa el tiempo en la máxima velocidad de degradación y por último ϕ_{4i} es la relación con la tasa de descomposición. Por último, β representa los efectos fijos y b los efectos aleatorios.

En la Figura 3.6 se observa que, como era evidente, el tipo de polímero PDLG afecta significativamente al camino de degradación. El PDL02 comienza a degradarse más tarde y el proceso es más lento, mientras que el PDLG7502A ciomienza la pérdida de masa con anterioridad.

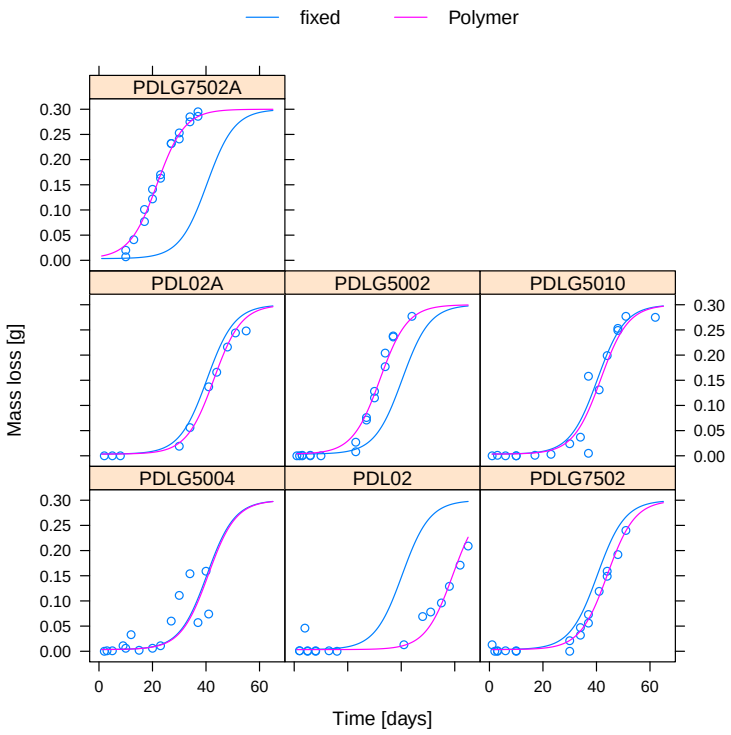

Figura 3.6: *Comparación del ajuste por cada tipo de polímero.*

3.2.2. NLME para MassLoss vs. pH

Como en el caso anterior, en la Figura 3.7 se realiza una representación gráfica de los datos. A la izquierda se muestran los gráficos de la pérdida de masa frente al pH, en diferentes ventanas según el tipo de polímero. A la derecha, las mismas representaciones pero en un mismo gráfico.

A simple vista se puede observar que conforme disminuye el pH, aumenta la pérdida de masa. Esto es debido a la descomposición del polímero, mediante la hidrólisis del mismo se libera ácido carboxílico que hace que el pH de la disolución disminuya. En resumen, un nivel bajo de pH podría ser un buen indicador de un alto grado de degradación del polímero, relacionado a su vez con un alto contenido en ácido carboxílico. Existe una degradación de tipo autocatalítica.

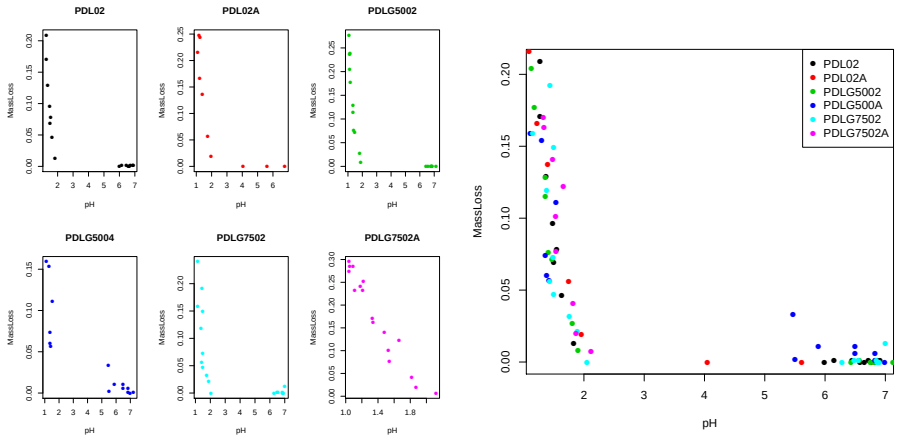

Figura 3.7: *Gráficas de la pérdida de masa frente al pH para los diferentes tipos de polímeros.*

Se utilizará un modelo de regresión asintótica. Dicho modelo se utiliza para modelar una respuesta y que se aproxima a una asíntota horizontal cuando $x \to \infty$.

$$y(x) = \phi_1 + (\phi_2 - \phi_1)\exp[-\exp(\phi_3)x], \tag{3.10}$$

donde ϕ_1 es la asíntota cuando $x \to \infty$ y ϕ_2 es $y(0)$. Estos parámetros se muestran en la Figura 3.8. El parámetro ϕ_3 es el logaritmo de la constante de velocidad. Se utiliza el logaritmo para hacer cumplir la positividad de la constante de velocidad de modo que el modelo se acerque a una asíntota. El correspondiente $t_{0,5} = \log 2/\exp(\phi_3)$, es decir, la mitad del tiempo de vida, se ilustra también en la Figura 3.8.

Aplicándolo a este caso concreto se tiene que:

$$\text{MassLoss}_{ij} = \phi_{1i} + (\phi_{2i} - \phi_{1i}) \cdot \exp(-\exp(\phi_{3i}) \cdot \text{pH}) + \epsilon_{ij}$$

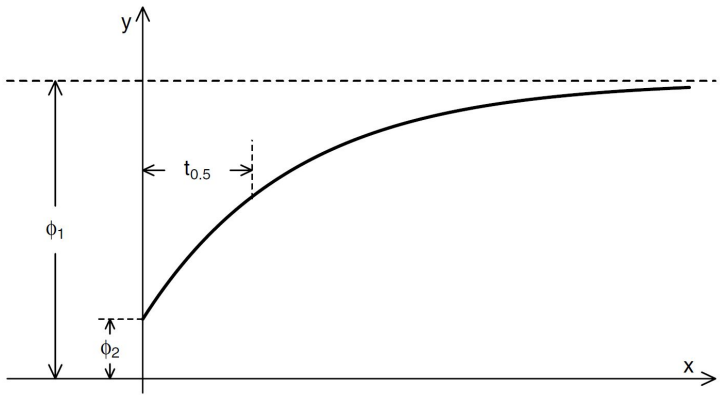

Figura 3.8: *El modelo de regresión asintótica muestra los parámetros ϕ_1, la respuesta asintótica cuando $x \to \infty$; ϕ_2, la respuesta cuando $x = 0$ y $t_{0,5}$, la mitad del tiempo de vida.*

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} = \beta + b_i, \quad b_i \backsim N(0, \Psi), \quad \epsilon_{ij} \backsim N(0, \sigma^2)$$

donde i es el tipo de polímero y j cada observación. Además ϕ_{1i} es la asíntota horizontal señalando la pérdida de masa inicial, ϕ_{2i} es la pérdida de masa estimada para pH=0 y ϕ_{3i} representa el logaritmo natural de la constante de velocidad de degradación. Además β representa los efectos fijos y b representa los efectos aleatorios.

En la Figura 3.9 se observa que, como se podía intuir desde un principio, no hay diferencias entre los polímeros. Las estimaciones son las mismas, es decir, la relación entre la pérdida de masa y el pH es independiente del tipo del polímero.

Como ya se ha mencionado anteriormente, el aumento de la concentración de ácido carboxílico acelera la hidrólisis. Cuanto mayor sea la

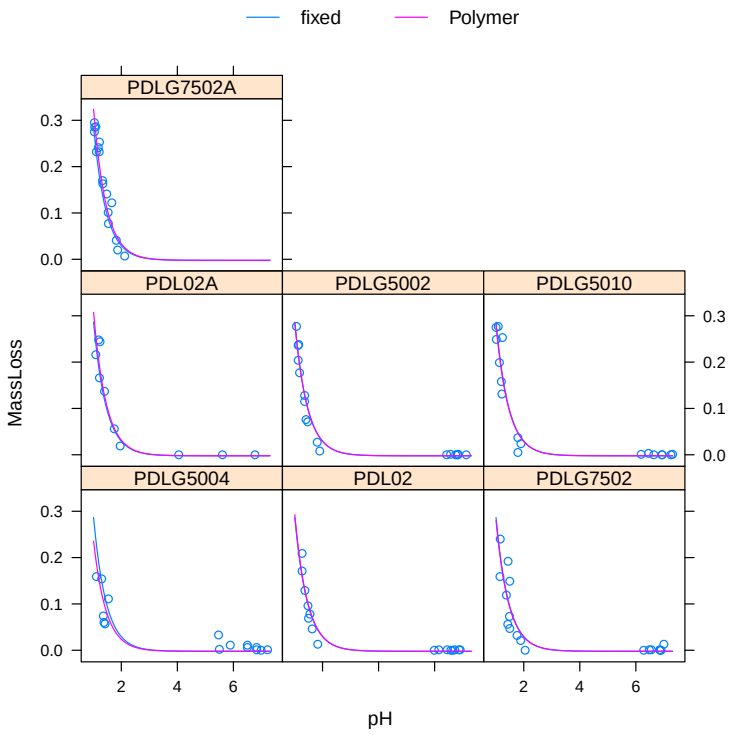

Figura 3.9: *Comparación del ajuste para cada tipo de polímero.*

concentración de ácido, menor es el pH y mayor la pérdida de masa.

3.2.3. NLME para pH vs. Time

En esta sección se estudiará la relación existente entre pH y el tiempo. En la Figura 3.10 se representan a la izquiera el pH frente al tiemop en diferentes gráficas para cada tipo de polímero. A la derecha, todo en la misma gráfica.

Observando estas gráficas puede intuirse que existirán diferencias en el comportamiento del pH a lo largo del tiempo según el tipo de polímero. Se va a ajustar a continuación un modelo de regresión logística para estudiar más en profundidad dicha dependencia.

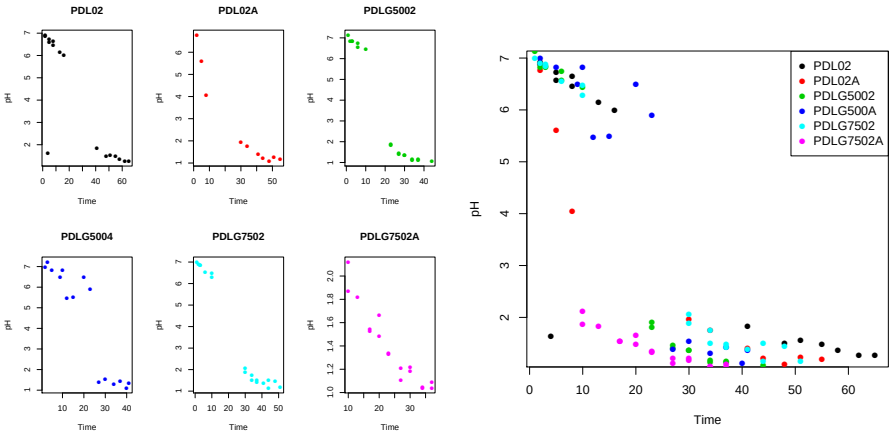

Figura 3.10: *Gráficas de la evolución del pH frente al tiempo para los diferentes tipos de polímeros.*

Dicho modelo se ha explicado previamente en el apartado 3.2.1, concretamente la expresión 3.9 presenta la formulación general del moodelo. Para este caso concreto se tiene que:

$$\mathrm{pH}_{ij} = \phi_{1i} + \frac{\phi_{2i} - \phi_{1i}}{1 + \exp\left(\frac{\mathrm{Time}_{ij} - \phi_{3i}}{\phi_{4i}}\right)} + \epsilon_{ij}$$

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} = \beta + b_i, \quad b_i \backsim N(0, \Psi), \quad \epsilon_{ij} \backsim N(0, \sigma^2)$$

donde ϕ_{1i} es el pH inicial: 7.4, ϕ_{2i} es la asíntota del pH final, ϕ_{3i} tiempo en la máxima velocidad de degradación del pH y ϕ_{4i} es la relación con la tasa de descomposición del pH. Además, como en los casos anteriores β representa los efectos fijos y b los efectos aleatorios.

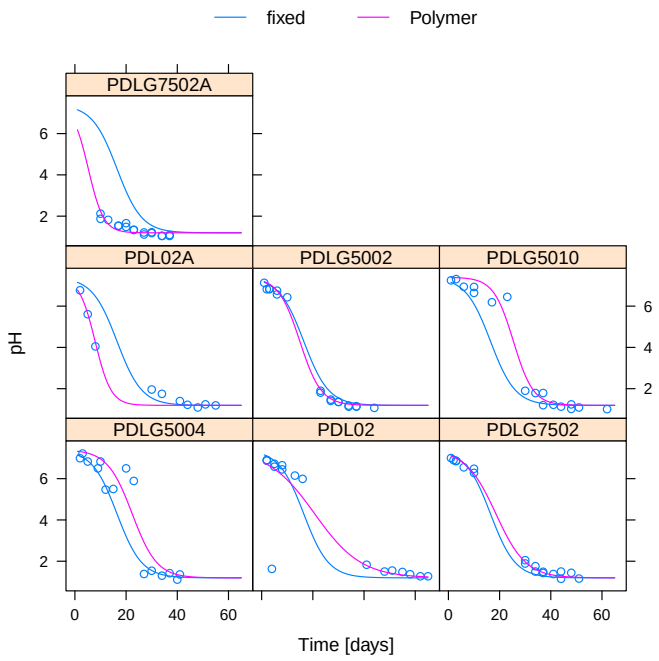

Figura 3.11: *Comparación del ajuste para cada tipo de polímero.*

En la Figura 3.11 se presentan los ajustes por tipo de polímero. Se observa que el pH depende del tiempo y también del tipo PDLG. Se había visto que la pérdida de masa dependía fuertemente del pH, lo que hace que la pérdida de masa también depende del tiempo.

Se tiene que los polímeros PDLG7502A y PDL02A tienen un comportamiento muy diferente al resto, con pH mucho más bajos desde tiempos bastante tempranos, es decir, comienza a descender antes y lo hace con una tendencia más fuerte. En cambio el polímero PDL02 comienza a descender más tarde y más lentamente.

En resumen, se observa que la pérdida de masa con respecto al tiempo, que caracteriza el grado de degradación de este tipo de materiales, se produce antes o después y a diferente velocidad dependiendo del tipo de polímero perteneciente a la familia estudiada. La diferente variación del pH con respecto al tiempo dependiendo del polímero (provocada

por la variación de la concentración de ácido carboxílico) condiciona muy probablemente la forma en la que este tipo de polímeros pierden masa, se degradan.

Capítulo 4

Análisis funcional

Uno de los objetivos de este trabajo es poder estimar el grado de degradación de una muestra a partir de una prueba acelerada. En este caso, se ha llevado a cabo, para cada muestra estudiada, a cada nivel de degradación, un análisis calorimétrico diferencial de barrido (DSC).

Ésta es una técnica experimental que permite caracterizar las propidades térmicas de un material, en particular permite estudiar procesos como la degradación, transición vítrea, cristalización, fusión etc. El output de esta técnica son las llamadas curvas DSC, que en este caso pueden tratarse como datos funcionales.

La técnica DSC permite tener una curva experimental en apenas unos minutos, calentando la muestra a una velocidad constante, en comparación. Podría por tanto considerse como una prueba de vida acelerada. Dicha técnica aporta información de la energía (llamada aquí flujo de calor) que se absorbe o se desprende al producirse ciertas reacciones químicas activadas al incrementar la temperatura del material estudiado.

En este capítulo se pretende realizar un estudio de los datos resultantes del estudio calorimétrico mediante la aplicación de técnicas de análisis de datos funcionales (FDA), dada la naturaleza de los datos resultantes. Principalmente se probarán diversos modelos de regresión FDA de

respuesta escalar y variable regresora funcional (curvas DSC) para, a partir de una prueba experimental rápida, poder estimar el grado de degradación de un biopolímero como los que aquí se estudian (en particular PDL02 y PDL5010).

Se tomaron los datos de la curva calorimétrica en un periodo de enfriamiento del polímero de 20 minutos, en el que la temperatura desciende desde $100°C$ hasta algo menos de $1°C$ aproximadamente.

4.1. Definición y representación funcional

Se define $\mathcal{X}$ como la variable funcional de interés, flujo de calor o ***Heat Flow*** que toma valores en un espacio normado (o semi-normado) $\mathcal{F}$; y se considera como los datos funcionales a analizar el conjunto $\{\S_1, \S_2, \ldots, \S_n\}$ que provienen de las n variable funcionales $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ idénticamente distribuidas como $\mathcal{X}$. Los datos funcionales se encuentran discretizados en un conjunto de punto $\{t_j\}_{j=1}^{d}$ no necesariamente equidistantes (como en este caso).

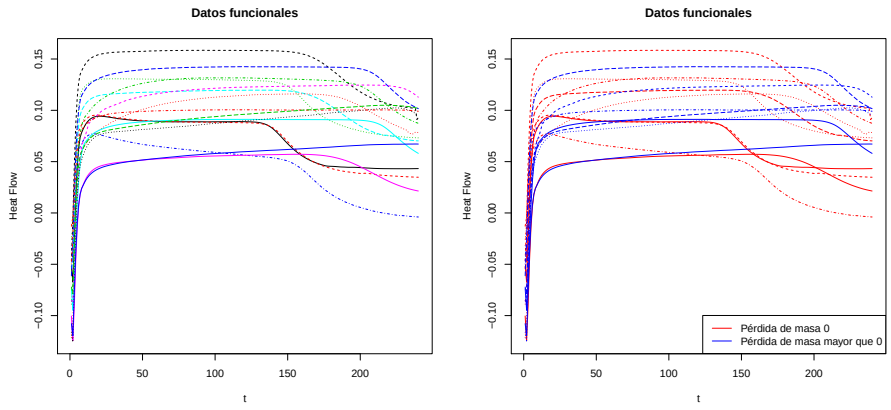

Figura 4.1: Gráficos para la representación de los datos funcionales.

Por tanto se cuenta con d evaluaciones (aproximadamente mediciones del flujo de calor cada 0,1 minutos ($6seg$)) para cada una de las $n = 16$

variables funcionales (observaciones). Cada una de ellas corresponde con uno de los días del estudio en que se realizaron las mediciones del polímero). Es decir se tendrá una matriz con 16 filas que representan las curvas discretizadas y 240 columnas que representan los puntos a evaluar. Las primeras 9 filas se correponden con una pérdida de masa 0, mientras que en las 7 restantes la pérdida de masa es mayor que 0.

En la Figura 4.1 se pueden apreciar los datos funcionales representados en la gráfica de la izquierda, y en la de la derecha se tiene en color rojo el flujo de calor para una pérdida de masa nula, y en color azul el fluj o de calor cuando existe pérdida de masa.

Para analizar si es necesario realizar alguna transformación de los datos que permitan una mejor visulización de los mismos (por ejemplo que arroje más claridad en cuanto a las diferencias entre flujo de calor con pérdida nula o mayor que cero), se ha realizado el cálculo de la primera y segunda derivada de los datos en base a varios métodos disponibles en la librería de `R`, `fda.usc`. Concretamente se han utilizado dos métodos: B-spline y Natural. Los resultados se presentan en la Figura 4.2.

La reprsentación realizada en la Figura 4.1 para los datos funcionales asume implícitamente un espacio $\mathcal{L}_2$, que a la vista está, es el más adecuado para la representación de estos datos (la transformación de la primera y segunda derivada no aporta una mejor representación de los mismos). Un posible motivo es que se trata de una variable (flujo de calor) representada a lo largo del tiempo.

En la Figura 4.3 se presentan las representaciones en bases realizadas para la primera curva de los datos funcionales de la muestra. Se consideran representaciones con bases B-spline (23,120) y Kernel Smoothing (KNN, LLR, NW). Como se puede observar en dichas gráficas, las respresentaciones en base están en el espacio $\mathcal{L}_2$; estas representaciones permitan trabajar el problema en menor dimensión.

Además en la Figura 4.4 se muestran las gráficas para los datos en bases tanto para B-splines-23, con $\lambda = 0{,}03125$ (derecha), como para

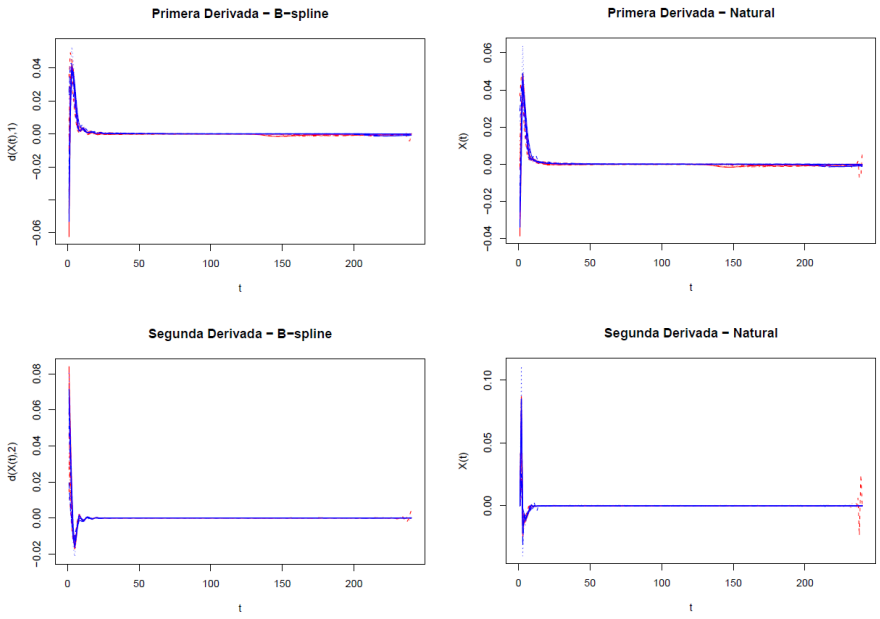

Figura 4.2: Primeras y segundas derivadas mediante B-spline y Natural. En color rojo las curvas correspondientes a una pérdida de masa nula y en azul a una pérdida de masa mayor que 0.

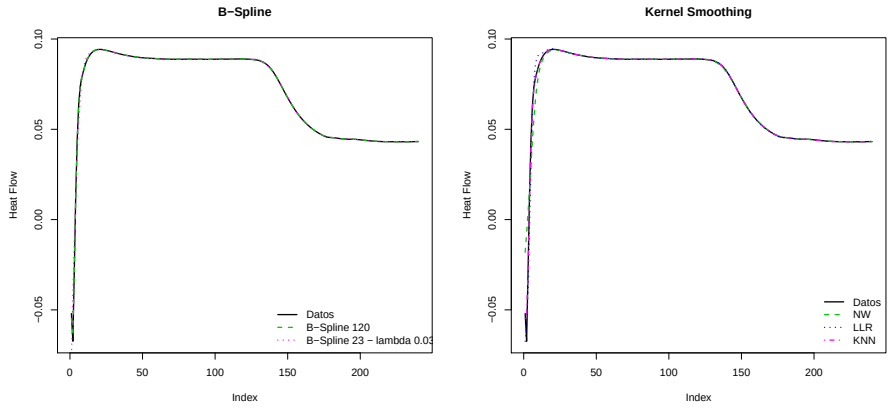

Figura 4.3: Representaciones en bases realizadas para la primera observación de los datos funcionales con bases: B-spline (23,120) y Kernel Smoothing (KNN, LLR, NW).

B-splines-120 (izquierda). Tanto el número de bases y valor de λ en el primer caso, como el número de bases en el segundo se obtuvieron con la función `min.basis`. Obviamente la representación con B-splines-120 es mejor, debido al mayor número de bases.

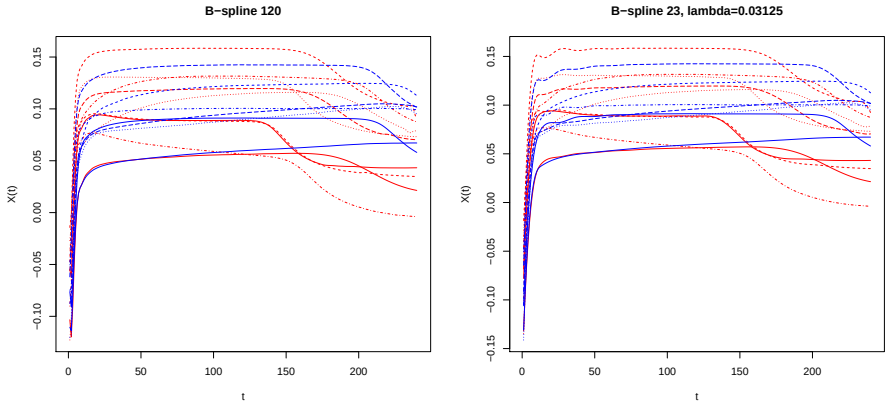

Figura 4.4: Representación de los datos en bases B-splines, 120 en el caso de la izquierda y 23 con $\lambda = 0{,}03125$ en el caso de la derecha.

También se realizó el método de componentes principales (FPCA) como el método de mínimos cuadrados parciales (FPLS). Para ambos casos se eligieron 2 componentes, respresentadas en la Figura 4.6.

En la Figura 4.5 se presenta una tabla con los siguientes indicadores: el porcentaje de varianza explicada por cada componente y la correlación con la pérdida de masa.

Alrededor del 83 % de la variabilidad total de los datos se explica con la primera componente con el método de componentes principales, mientras que este porcentaje bajas casi un 56 % cuando se utiliza el métodos de mínimos cuadrados parciales. La variabilidad explicada por la segunda componente es de casi un 15 % en el caso de FPCA, y sube hasta el 41'4 % en el caso de FPLS.

Con respecto a la correlación, es más fuerte con la segunda componente

Método	Índices	PC1	PC2
FPCA	Varianza Explicada	82.82%	14.85%
	Correlación con pérdida de masa	-0.113	0.717
		PLS1	PLS2
FPLS	Varianza Explicada	55.9%	41.4%
	Correlación con pérdida de masa	0.608	0.487

Figura 4.5: Tabla con los porcentajes de varianza explicada para cada componente y la correlación con la pérdida de masa.

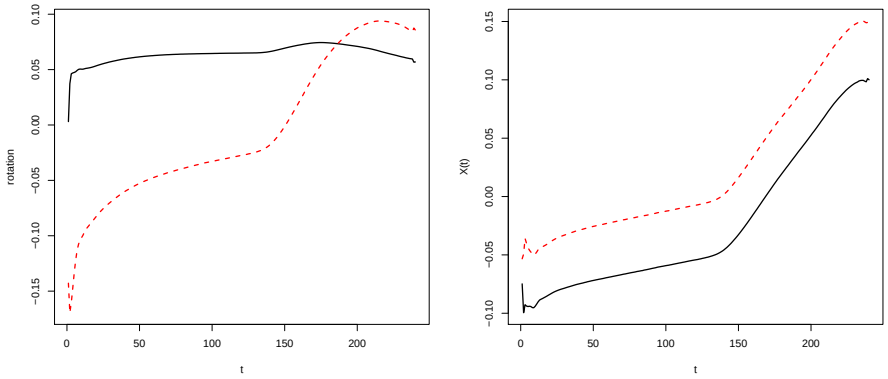

Figura 4.6: Representación de las tres primeras componentes tanta para FPCA (izquierda) como para FPLS (derecha)

del método de componentes principales y negativa más débil en caso de la primera componente. En el método de FPLS las correlaciones son positivas en torno a 0.5.

4.2. Análisis exploratorio de los datos funcionales

En esta sección, se comienza con un análisis descriptivo del flujo de calor mediante el cálculo de la media, y la varianza funcional de los datos. También se aplican otras variantes para el cálculo de la media y varianza (medidas acortadas).

En los datos univariantes, la mediana sería el punto más profundo del conjunto de puntos. Para este estudio, se han aplicado las siguientes medidas de profundidad que están incluidas en el paquete `fda.usc`: Moda (mode depth); Mediana definida por Friedman y Muniz (FM depth); y, Proyecciones Aleatorias (RP depth).

Posteriormente, se realiza un estudio sobre la presencia de datos atípicos ya que podrían influir en la estimación y desempeño del modelo. La profundidad es una medida cuyo concepto ha surgido en la literatura de robustez, mide que tan profundo (o central) es un punto de referencia respecto a una población (o muestra). Por tanto, para cuyos puntos que tengan valores grandes de profundidad, estarán más cerca del comportamiento central de los datos; y si son menos profundos, es decir, valores de profundidad pequeño son posibles candidatos a datos atípicos.

En la Figura 4.7, se puede apreciar las medias funcionales (normal y recortada) para los datos originales así como medianas y varianzas. Se puede observar que las medias y medianas son similares indpendientemente del método, en cambio las varianzas se diferencian más claramente.

Para tener una mejor reprsentación gráfica para observar los límites por donde oscilan la median funcional, media recortada y mediana se ha aplicado el método bootstrap suavizado mediante la función `fda.bootstrap` del paquete `fda.usc`. En la Figura 4.8 se pueden observar las bandas de confianza y se tiene que las más ajustadas se obtienen para la media funcional. La gran amplitud de las bandas es debida al

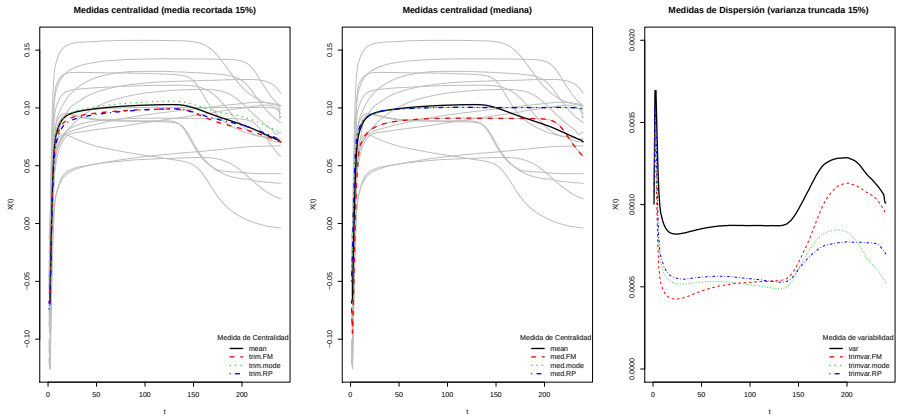

Figura 4.7: Curvas medias, medianas y de varianza funcional para los datos de flujo de calor.

reducido número de datos.

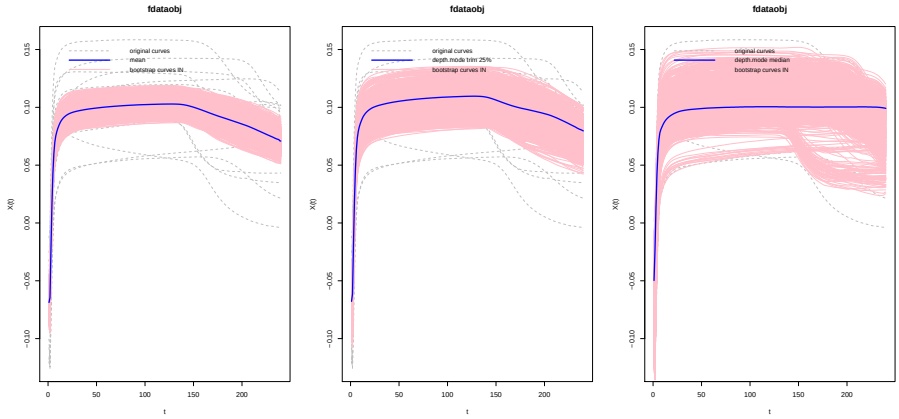

Figura 4.8: Bandas de confianza para la media funcional.

Una vez estudiado la tendencia central y la variabilidad de los datos se continua con la detección de datos atípicos en la muestra. Se inicia un análisis con todos los datos originales calculando tres medidas de profundidad (recortadas un 15 %) y se observa la diferencia de cada una con respecto a la mediana de los datos funcionales. Estas gráficas

se pueden observar en la Figura 4.9.

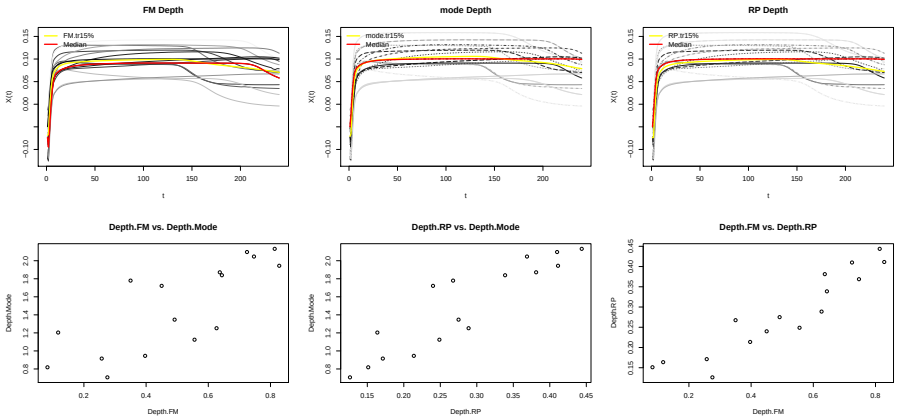

Figura 4.9: En la primera fila se tiene la representación de las medidas de profundidad consstrastadas con la mediana de la muestra total. En la segunda fila se tienen los gráficos de dispersión entre todas las medidas de profundidad.

Se han utilizado dos funciones de la librería `fda.usc` que permiten la detección de datos atípicos: `outliers.depth.trim` y `outliers.depth.pond`. Con la primera no se detectan outliers para ninguna de las medidas de profundidad, mientras que con la segunda se detecta como outlier o dato atípico la curva número 7, con la medida de profuncidad FM. Se representa en la Figura 4.10.

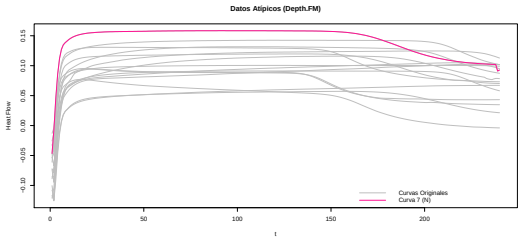

Figura 4.10: Outlier detectado con la función *outliers.depth.pond* y la medida de profundidad FM.

4.3. Distancia entre datos funcionales

En esta sección, se ha aplicado una métrica para el espacio $\mathcal{L}_2$y 4 semi-métricas para otros espacios semi-normados, con la finalidad de calcular la distancia entre los datos funcionales. Para el cálculo de estas medidas, se han aplicado las siguientes funciones desarrolladas en el paquete `fda.usc`:

- `metric.lp`: para datos funcionales representados en un espacio, con $p = 2$.
- `semimetric.deriv`: para datos funcionales en el espacio de funciones de la primera y segunda derivada.
- `semimetric.pca`: basado en el método de componentes principales (PCA), calcula una semi-métrica PCA entre los datos funcionales.
- `semimetric.mplsr`: basado en el método de mínimos cuadrados parciales (PLS), calcula una semi-métrica FPLS entre los datos funcionales.

El resultado de cada una de estas funciones (métricas y semi-métricas) es una matriz de dimensión 16 x 16 que contiene las distancias entre todas las curvas (datos funcionales).

Se puede utilizar esta información como regla de clasificación, ya que es de esperar que las curvas más próximas pertenezcan a una pérdida de masa similar.Se clasificará en dos grupos: pérdida de masa 0, y pérdida de masa mayor que 0.

En la Figura 4.11 se presentan los dendogramas para cada caso y en la Figura 4.12 se muestran los porcentajes de clasificaciones correctas para cada método.

De los resultados que se presentan en la Figura 4.12, el de menor porcentaje de una correcta clasificación es el de la segunda derivada. El de mayor porcentaje es el espacio de los mínimos cuadrados parciales, y el resto presentan el mismo porcentaje de aciertos.

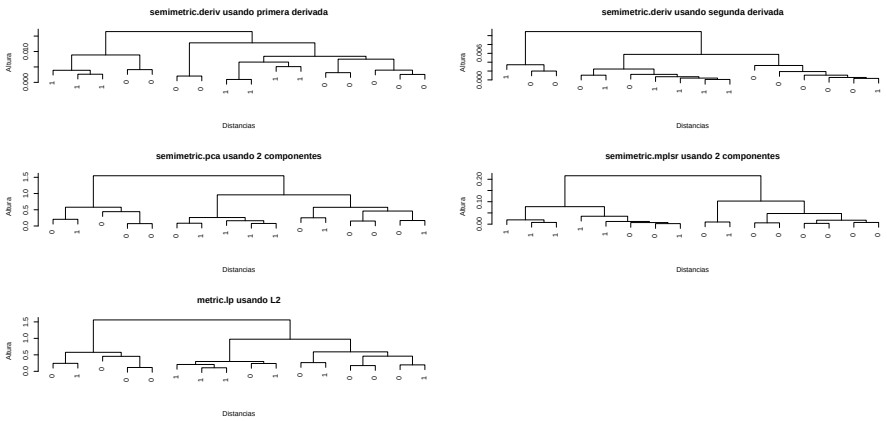

Figura 4.11: Dendogramas para las métricas $\mathcal{L}_2$ (esquina superior izquierda) y las cuatro semi-métricas en otros espacios.

Función	Espacio	% aciertos
metric.lp	L_2	62.5
semimetric.deriv	Primera derivada	62.5
semimetric.deriv	Segunda derivada	50
semimetric.pca	Componentes principales	62.5
semimetric.mplsr	Mínimos cuadrados parciales	81.5

Figura 4.12: Tabla con los porcentajes de acierto en la clasificación de la pérdida de masa para cada método.

Claramente, si el objetivo fuese identificar simplemente las diferencias del flujo del calor en función de si existe o no la pérdida de masa, a la vista de estos resultados sería más recomendable trabajar en espacios seminormados, específicamente en el espacio de mínimos cuadrados paricales. Pero en este trabajo se pretende ir más allá, realizando una regresión con la variable continua péridida de masa, sin discretizar.

4.4. Modelos de Regresión Funcionales

Para finalizar se han tratado de ajustar varios modelos ajusten la pérdida de masa (variable escalar) en función del flujo de calor (variable funcional). Lo ideal hubiese sido dividir los datos en un conjunto de entrenamiento y un conjunto de validación, haciendo predicciones y

pudiendo así comparar los resultados y ver la bondad del ajuste realizado, pero debido a los pocos datos funcionales (solo se dispone de 16 curvas) se ha preferido hacer el ajuste con todos los datos. Con vistas a la mejora del estudio para el trabajo de fin de máster, podría tratar de aumentarse el número de curvas para poder hacer un mejor ajuste.

Se han propuesto 4 modelos para el ajuste:

- Modelo lineal funcional (FLR) con base B-spline.
- Modelo lineal funcional (FLR) con base PCA.
- Modelo lineal funcional (FLR) con base PLS.
- Modelo de regresión funcional no paramétrico.

4.4.1. Modelo lineal funcional (FLR) con base B-spline.

En primer lugar se ajusta un modelo FLR con base B-spline. Este modelo asume que entre la respuesta escalar Y (*pérdida de masa*) y la covariable funcional $\mathcal{X}$ (*flujo de calor*) existe una relación lineal. Entonces el modelo lineal funcional bajo el enfoque paramétrico viene dado por la siguiente expresión:

$$y_i = \langle \mathcal{X}, \beta \rangle + \epsilon_i = \int_T \mathcal{X}_i(t)\beta(t)dt + \epsilon_i$$

donde $\langle \cdot, \cdot \rangle$ denota el producto interno con $\mathcal{L}_2$ y ϵ_i es una variable aleatoria con media cero y varianza finita σ^2.

[Ramsay y Silverman (2005)] modelizaron la relación entre una respuesta escalar y una covariable funcional mediante representaciones en base de los datos funcionales observados $\mathcal{X}(t)$ y el parámetro funcional desconocido $\beta(t)$. El anterior modelo funcional se estima mediante la expresión:

$$\hat{y}_i = \int_T \mathcal{X}_i(t)\beta(t)dt \approx \mathbf{C_i^T}\psi(\mathbf{t})\phi^{\mathbf{T}}(\mathbf{t})\hat{\mathbf{b}} = \tilde{\mathbf{X}}\hat{\mathbf{b}}$$

donde $\tilde{\mathbf{X}}_\mathbf{i}(\mathbf{t})$ es la puntuación tal que $\tilde{\mathbf{X}}_\mathbf{i}(\mathbf{t}) = \mathbf{C_i^T}\psi(\mathbf{t})\phi^\mathbf{T}(\mathbf{t})$ y $\hat{\mathbf{b}} = (\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}})^{-\mathbf{1}}\tilde{\mathbf{X}}^\mathbf{T}y$ y así, $\hat{y} = \tilde{\mathbf{X}}\hat{\mathbf{b}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}})^{-\mathbf{1}}\tilde{\mathbf{X}}^\mathbf{T}y = \mathbf{H}y$ donde $\mathbf{H}$ es la matriz con los grados de libertad: $df = trace(\mathbf{H})$.

Los resultados obtenidos en R mediante la función `fregre.basis` fueron los siguientes:

```
 *** Summary Functional Data Regression with representation in Basis ***

Call:
fregre.basis(fdataobj = x, y = y, basis.x = basis1, basis.b = basis2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.028284 -0.013721 -0.002884  0.013372  0.036682

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.067875   0.006493  10.453 6.09e-06 ***
x.bspl4.1    0.288814   0.136944   2.109  0.06798 .
x.bspl4.2   -0.494529   0.290713  -1.701  0.12734
x.bspl4.3    0.683947   0.438386   1.560  0.15735
x.bspl4.4   -0.660601   0.341826  -1.933  0.08937 .
x.bspl4.5    0.581291   0.263887   2.203  0.05874 .
x.bspl4.6   -0.621838   0.247267  -2.515  0.03610 *
x.bspl4.7    0.815994   0.205206   3.976  0.00408 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02597 on 8 degrees of freedom
Multiple R-squared:  0.9625,    Adjusted R-squared:  0.9296
F-statistic: 29.31 on 7 and 8 DF,  p-value: 4.311e-05

-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Y los gráficos resumen para le modelo ajustado se presentan en la Fi-

gura 4.13. Con estas gráficas no se puede suponer que se cumplan las condiciones del modelo lineal, aunque se dispone de pocas observaciones funcionales.

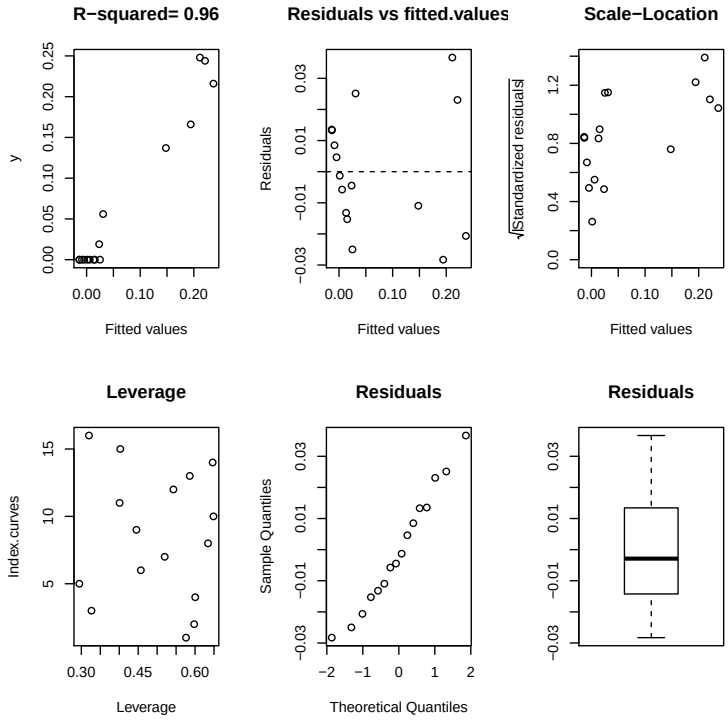

Figura 4.13: Gráficos para el ajuste lineal funcional con representacion en bases.

4.4.2. Modelo lineal funcional (FLR) con base PCA.

De un modo similar, [Cardot el al. (1999)] utilizan una base de componentes principales funcionales para reprsentar los datos funcionales $\mathcal{X}(t)$ y el parámetro funcional $\beta(t)$. Ahora la estimación de β se puede realizar mediante unas cuantas componentes principales de los datos

funcionales y la integral se podría aproximar así:

$$\hat{y}_i = \int_T \mathcal{X}_i(t)\beta(t)dt \approx \sum_{k=1}^{k_n} \gamma_{ik_n}\hat{\beta}_{k_n}$$

donde $\hat{\beta}_{1:k_n} = \left(\frac{\gamma_{.1}^T y}{n\lambda_1}, \ldots . \frac{\gamma_{.k_n}^T y}{n\lambda_{k_n}}\right)$ y $\gamma_{1:k_n}$ es la matriz $(n \times k_n)$ con las k_n estimaciones de componentes principales de las puntuaciones de β y λ_i los autovalores de las componentes principales.

Este modelo se puede expresar como: $\hat{y} = \mathbf{H}y$ donde $\mathbf{H} = \left(\frac{\gamma_{.1}\gamma_{.1}^T y}{n\lambda_1}, \ldots . \frac{\gamma_{.k_n}\gamma_{.k_n}^T y}{n\lambda_{k_n}}\right)$ con los siguientes grados de libertad: $df = trace(\mathbf{H}) = k_n$.

Los resultados obtenidos en R mediante la función `fregre.pc`, con las tres primera componentes principales, fueron los siguientes:

```
 *** Summary Functional Data Regression with Principal Components ***

Call:
fregre.pc(fdataobj = x, y = y, l = 1:3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.047356 -0.023702 -0.005114  0.024206  0.056104

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.067875   0.009084   7.472 7.51e-06 ***
PC1         -0.018382   0.020423  -0.900    0.386
PC2          0.354276   0.048824   7.256 1.01e-05 ***
PC3         -1.002100   0.159104  -6.298 3.96e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03634 on 12 degrees of freedom
Multiple R-squared:  0.8898,    Adjusted R-squared:  0.8623
F-statistic:  32.3 on 3 and 12 DF,  p-value: 4.992e-06
```

```
-With 3 Principal Components is  explained  99.04 %
 of the variability of explicative variables.

-Variability for each  principal components -PC- (%):
  PC1   PC2   PC3
83.12 14.54  1.37
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Y los gráficos resumen para le modelo ajustado se presentan en la Figura 4.14, donde se puede ver de nuevo que se trata de un buena ajuste.

Pero en el `summary` obtenido de `R`, se puede ver como la primera componente principal no es significativamente distinta de 0. El problema de decidir qué componentes a utilizar en la regresión se resuelve mediante la elección de un subconjunto óptimo componentes principales funcionales que mejor se estima la respuesta. La selección se realiza mediante validación cruzada (CV) o criterios de selección de modelos (MSC).

- Método de Validación Cruzada: $PCV(k_n) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle \mathcal{X}_i, \hat{\beta}_{(-i,k_n)}\rangle\right)^2$, `criteria="CV"`.
- Criterio de Selección de Modelos: $MSC(k_n) = log\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle \mathcal{X}_i, \hat{\beta}_{(i,k_n)}\rangle\right)^2\right] + p_n\frac{k}{n}$,
  $p_n = 2$, `criteria="AIC"`
  $p_n = \frac{2n}{n-k_n-2}$, `criteria="AICc"`
  $p_n = \frac{log(n)}{n}$, `criteria="SIC"`
  $p_n = \frac{log(n)}{n-k_n-2}$, `criteria="SICc"`

donde `criteria` es un argumento de la función de `R` que lo implementa `fregre.pc.cv`.

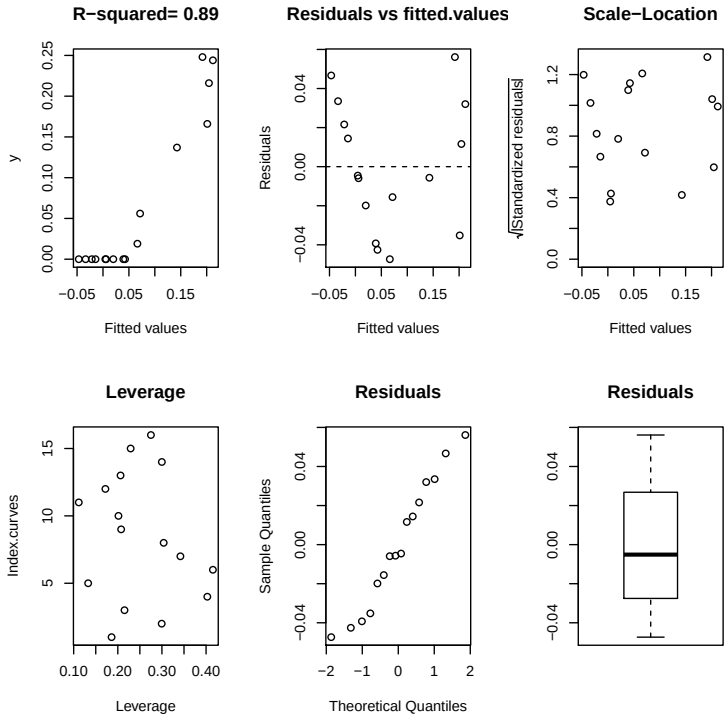

Figura 4.14: Gráficos para el ajuste lineal funcional con representacion mediante componentes principales.

Con dicha función se obtienen mediante el criterio SIC cuáles son las componentes óptimas para utilizar en el modelo,en este caso fueron la número 2, 3, 5 y 4. Una vez ajustado el modelo con dichas componentes los resultados fueron los siguientes:

```
 *** Summary Functional Data Regression with Principal Components ***

Call:
fregre.pc(fdataobj = x, y = y, l = c(2, 3, 5, 4, 7))

Residuals:
      Min        1Q    Median        3Q       Max
-0.031740 -0.015050 -0.004793  0.014209  0.050104
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.067875   0.006307  10.762 8.08e-07 ***
PC2          0.315299   0.046980   6.711 5.29e-05 ***
PC3         -1.081453   0.120330  -8.987 4.19e-06 ***
PC5         -0.839193   0.285554  -2.939   0.0148 *
PC4         -0.406961   0.204308  -1.992   0.0744 .
PC7         -1.090251   0.670620  -1.626   0.1351
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02523 on 10 degrees of freedom
Multiple R-squared:  0.9557,    Adjusted R-squared:  0.9336
F-statistic: 43.19 on 5 and 10 DF,  p-value: 1.883e-06


-With 5 Principal Components is  explained  16.61 %
 of the variability of explicative variables.

-Variability for each  principal components -PC- (%):
  PC2   PC3   PC5   PC4   PC7
14.54  1.37  0.22  0.41  0.08
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Con los gráficos del modelo presentados en la Figura 4.15.

4.4.3. Modelo lineal funcional (FLR) con base PLS.

Otra buena alternativa es utilizar el criterio que maximiza la covarianza entre $\mathcal{X}(t)$ y la respuesta escalar Y mediante las componentes PLS. La idea básica es construir un conjunto de componentes PLS en el espacio lineal engendrado por $\mathcal{X}(t)$, teniendo en cuenta la correlación entre Y y $\mathcal{X}(t)$. Estas componentes FPLS están relacionadas más directamente con la variabilidad en y que las componentes FPC.

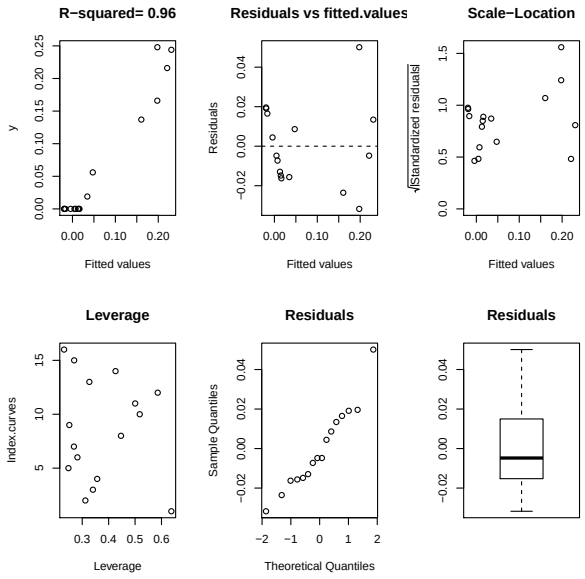

Figura 4.15: Gráficos para el ajuste lineal funcional con representacion mediante las componentes principales 2, 3, 5, 4 y 7.

Ahora la estimación de β se puede hacer mediante mínimos cuadrados parciales (factores PLS) de los datos funcionales y la integral se puede aproximar por:

$$\hat{y}_i = \int_T \mathcal{X}_i(t)\beta(t)dt \approx \sum_{j=1}^{q} c_{j,i}\nu_j$$

donde $c_{j,i} \in \mathcal{R}^p$, $i = 1, \ldots, p$, $j = 1, \ldots, q$ y ν es es autovector asociado al mayor autovalor de $W^X W^y$, tal que $\{\nu_i\}_1^q$ forma un sistema ortogonal en $\mathcal{L}_2$.

Los resultados obtenidos en R mediante la función "fregre.pls()çon las tres primeras componentes PLS, son los siguientes:

```
 *** Summary Functional Regression with Partial Least Squares***

-Call: fregre.pls(fdataobj = x, y = y, l = 1:3)

                Estimate  Std. Error  t value     Pr(>|t|)
```

```
(Intercept) 0.067875000 0.008146358 8.331944 2.718405e-06
PLS1        0.007921570 0.001074145 7.374767 9.273378e-06
PLS2        0.006890754 0.001245936 5.530586 1.364736e-04
PLS3        0.032004961 0.004986652 6.418127 3.534359e-05

-R squared:  0.9126224
-Residual variance:  0.00106181 on  11.83394  degrees of freedom
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Y los gráficos resumen para le modelo ajustado se presentan en la Figura 4.16, donde se puede ver de nuevo que se trata de un buena ajuste.

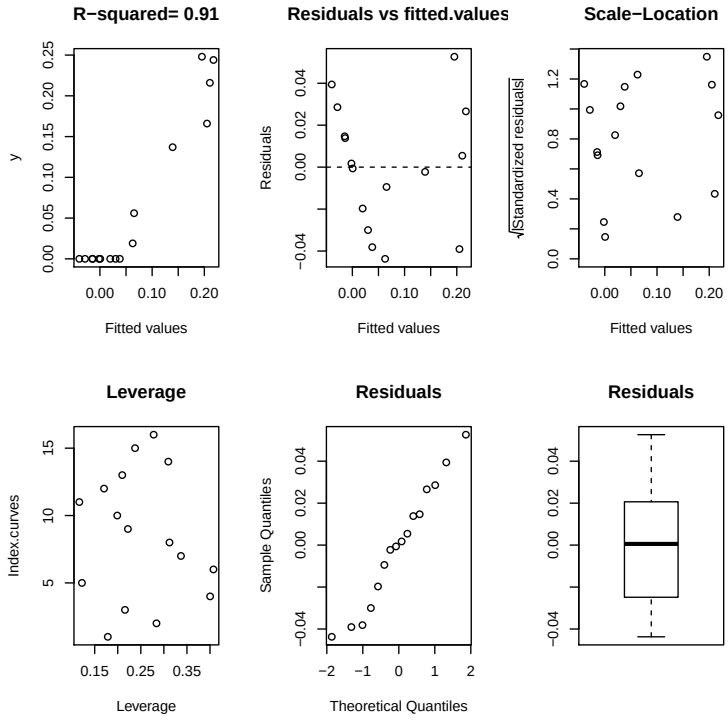

Figura 4.16: Gráficos para el ajuste lineal funcional con representacion mediante PLS.

En este caso todas las componentes son significativamente distintas de cero, pero aún así se utiliza la función `fregre.pls.cv` para obtener el número óptimo, como se hizo en el caso de componentes principales, mediante el criterio SIC. Se obtuvo que las componentes óptimas serían la 1, 2, 3 y 4. Así los resultados obtenidos para este modelo se muestran a continuación y las gráficas en la Figura 4.17.

```
 *** Summary Functional Regression with Partial Least Squares***

-Call: fregre.pls(fdataobj = x, y = y, l = 1:4)

              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 0.067875000 0.006532598 10.390199 8.503599e-07
PLS1        0.008769287 0.000911613  9.619528 1.760870e-06
PLS2        0.007261849 0.001007629  7.206868 2.431319e-05
PLS3        0.031708676 0.004000177  7.926818 1.044098e-05
PLS4        0.025576321 0.009005616  2.840041 1.701585e-02

-R squared:  0.9509194
-Residual variance:  0.0006827975 on  10.33699  degrees of freedom
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Al igual que ocurre con el model B-Spline, con estas gráficas no se puede suponer que se cumplan las condiciones del modelo lineal.

4.4.4. Modelo de regresión funcional no paramétrico.

Una alternativa a los modelos lineales funcionales es la regresión funcional no paramétrica estudiada por [Ferraty y Vieu (2006)]. En este caso, el modelo de regresión sería el siguiente:

$$y_i = r(\mathcal{X}_i(t)) + \epsilon_i$$

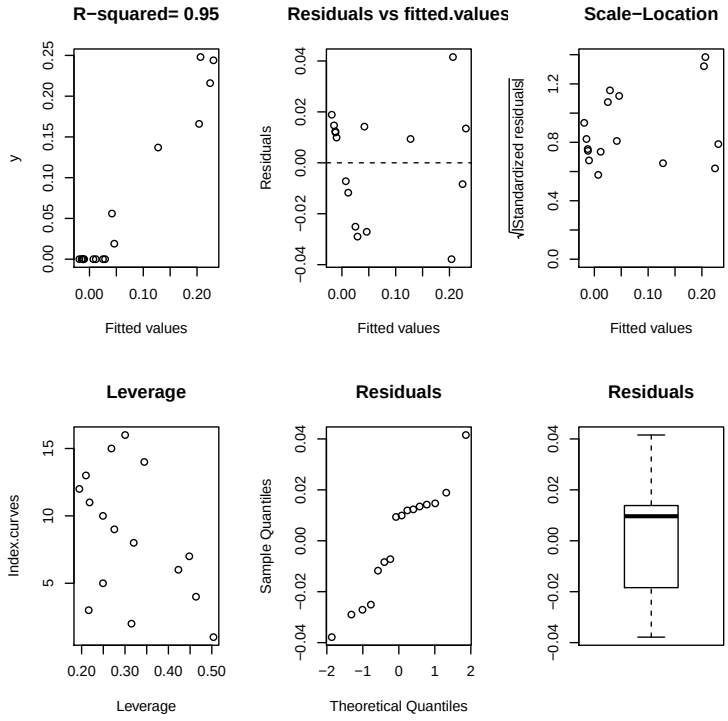

Figura 4.17: Gráficos para el ajuste lineal funcional con representacion mediante PLS (4 primeras componentes).

donde la función suave real r se estima mediante estimación kernel:

$$\hat{r}(\mathcal{X}) = \frac{\sum_{i=1}^{n} K(h^{-1}d(\mathcal{X}, \mathcal{X}_i))y_i}{\sum_{i=1}^{n} K(h^{-1}d(\mathcal{X}, \mathcal{X}_i))}$$

donde K es una función kernel asimétrica, h es el parámetro de suavizado y d es una métrica o semi-métrica.

Los resultados obtenidos en R mediante la función `fregre.np` se presentan a continuación:

```
 *** Summary Functional Non-linear Model ***

-Call: fregre.np(fdataobj = x, y = y)
```

```
-Bandwidth (h):  0.1660127
-R squared:  0.7511529
-Residual variance:  0.00457425 on  7.823282  degrees of freedom
-Names of possible atypical curves: No atypical curves
-Names of possible influence curves: No influence curves
```

Y los gráficos resumen para le modelo ajustado se presentan en la Figura 4.18. Hay que destacar que este es el peor ajuste de los 4 propuestos.

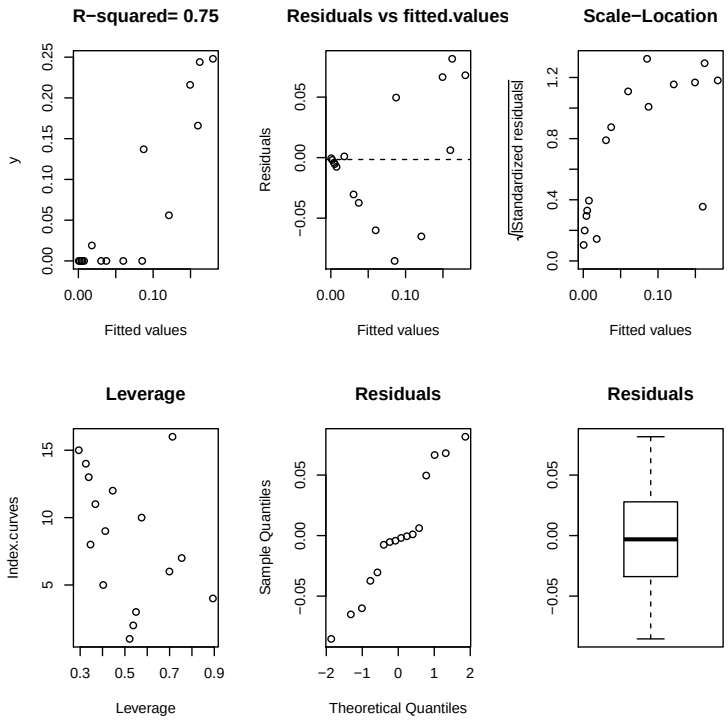

Figura 4.18: Gráficos para el ajuste del modelo de regresión funcional no paramétrico.

4.4.5. Resumen de resultados

En la Figura 4.19 se presenta una tabla resumen de los modelos ajustados:

Función	*df*	R^2	S^2_R
`fregre.basis(x,y)`	8	0.962	0.0259
`fregre.pc(x,y,l=1:3)`	12	0.889	0.0363
`fregre.pc(x,y,l=c(2,3,5,4,7))`	10	0.955	0.0252
`fregre.pls(x,y,l=1:3)`	11.83	0.912	0.0010
`fregre.pls(x,y,l=1:4)`	10.33	0.951	0.0006
`fregre.np(x,y)`	7.82	0.751	0.0045

Figura 4.19: Tabla con los resultados para los modelos de regresión funcional propuestos. *df* grados de libertad, S^2_R varianza residual, R^2 R-cuadrado.

Se han probado diversos modelos de regresión de variable respuesta escalar (masa perdida, grado de degradación) y variable regresora funcional (curvas DSC o calorimétricas). Como se ha comentado a lo largo del estudio funcional, estos modelos tienen problemas con algunas de las hipótesis de partida (ya sea en algunos casos la normalidad, la homocedasticidad o la independencia).

4.4.6. Modelo de regresión funcional lineal con una base PC funcional

Llegado este punto, se decidió hacer un nuevo estudio, tomando más datos, para poder obtener modelos que cumpliesen todas las hipótesis y evaluar así convenientemente su utilidad (el poder de predicción). Por ello se probó con otro de los polímeros, el *PDLG5010*.

En este caso se pretendíra realizar una estimación del tiempo de vida del material a partir de pruebas aceleradas DSC (de nuevo), siendo ahora la variación de energía dependietne de la temperatura/tiempo y con un calentamiento de 100ºC a 10ºC/min, caudal de 50ml/min de N_2 y 5 mg de muestra.

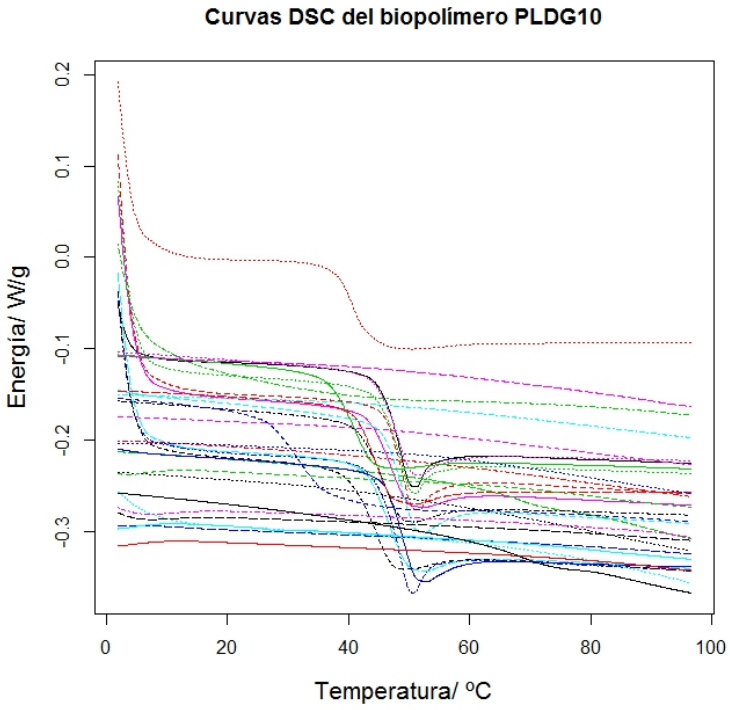

Figura 4.20: Pruebas aceleradas: curvas DSC.

En la Figura 4.20 se muestran las curvas DSC para las pruebas aceleradas representando la energía en W/g frente a la temperatura en ºC. Por otro lado, en la Figura 4.21 se representa la pérdida de masa (en mg) frente al tiempo (en días). Como se puede observar la pérdida de masa es mayor a medida que se avanza en el tiempo (como cabía esperar).

El modelo de regresión propuesto en este caso fue un modelo de regresión funcional lineal con una base PC funcional:

$$\hat{y} = \langle X, \hat{\beta} \rangle \approx \sum_{k=1}^{k_n} \gamma_{ik} \hat{\beta}_k$$

Dicho modelo se obtuvo con 4 componentes principales, seleccionadas por validación cruzada. El R^2 fue 0.94 y las 4 componentes fueron es-

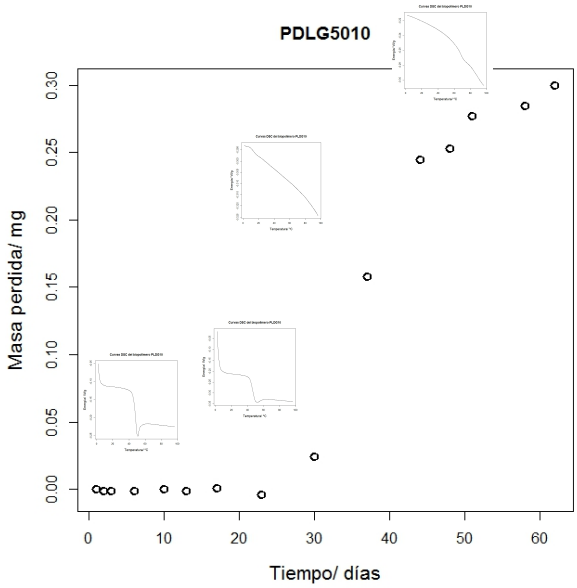

Figura 4.21: DSC por pérdida de masa.

tadísticamente significativas.

Este modelo además si cumple con las hipótesis previas (Figura 4.22), lo cual se refleja en gran medida en el alto nivel de acierto de las predicciones, mostradas en la Figura 4.23.

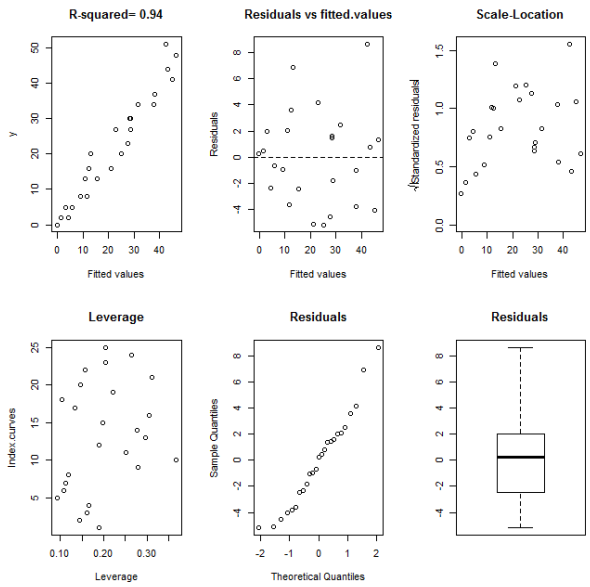

Figura 4.22: Análisis de residuos del modelo.

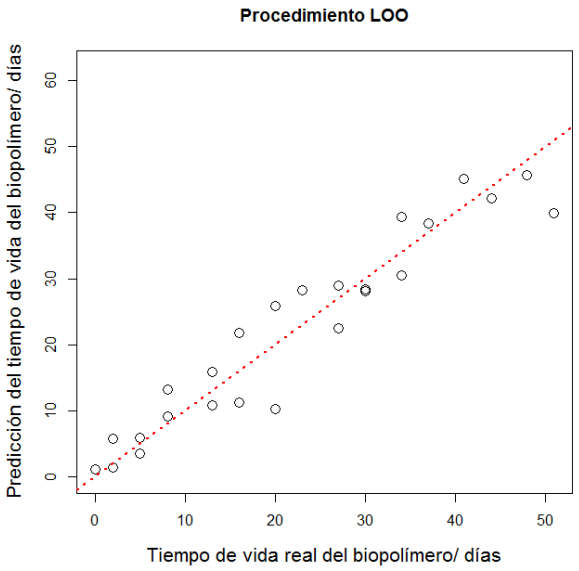

Figura 4.23: Predicciones.

Capítulo 5

Conclusiones finales

De la primera parte del estudio (modelización de la degradación de los polímeros) se pueden extraer las siguientes conclusiones:

- Los modelos GAM y modelos mixtos no lineales permiten estimar el pH y la correspondiente pérdida de masa de los polímeros PDLG a lo largo del tiempo.
- La dependencia entre la pérdida de masa y tiempo, y por otro lado entre el pH y el tiempo se puede modelizar utilizando una función logística. La relación entre la pérdida de masa y el pH se calcula mediante función asintótica. Es independiente del tipo de polímero, mientras que en los dos primeros casos no es así.
- Las diferencias en pérdida de masa están fuertemente relacionados con la forma en que el pH disminuye frente a tiempo.
- El polímero PDL02 comienza a degradarse más tarde y el proceso de degradación es más lento. El PDLG7502A comienza la pérdida de masa antes.

Mientras que del análisis funcional se concluye que, aunque los primeros estudios (para el polímero *PDL02*) no fueron todo lo exitosos que se espera (quizás por la escasez de los datos), al realizar el análisis funcional con el modelo de regresión funcional lineal y base PC funcional, sí se cumplen las hipótesis del modelo y se obtiene una buena predicción. Por tanto anima a seguir la línea de investigación del análisis funcional para la estimación de tiempos de vida de los biopolímeros.

Bibliografía

[1] Nelder J.A, Wedderburn R.M.W, (1972). Generalized Linear Models. *Journal of the Rroyal Statistical Society. Series A (General)*, **135** 3, 370-384.

[2] McCullagh P, Nelder J.A, (1989). *Generalized Linear Models*. Chapman and Hall, Florida.

[3] Dobson A.J, (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall, Florida.

[4] Wood S, (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall, Florida.

[5] Wang Y, (2011). *Smoothing Splines. Methods and Applications*. Chapman and Hall, Florida.

[6] Faraway J.J, (2006). *Extending the Lineal Model with R*. Chapman and Hall, Florida.

[7] Bates D.M, Watts D.G, (1988). Nonlinear Regreession Analysis: Its Applications. *Journal of Chemometrics*, **3** 3, 544-545.

[8] Lindstrom M.J, Bates D.M, (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *International Biometric Society*, **46** 3, 673-687.

[9] Pinheiro J.C, Bates D.M, (2000). *Mixed- Effects Models in S and S-PLUS*. Springer.

[10] Tarrío-Saavedra J, Naya S, López- Beceiro J, Zaragoza S, Álvarez A, Quintana-Pita S, García-Sebastián F.J, (2014). Degradation modelling of bio-polymers used as dental scaffolds. *Biodental Engineering*, **III** , 281-286.

[11] Oviedo de la Fuente M, (2011). *Utilities for Statistical Computing in Funcional Data Analysis: The R Package fda.usc*, Trabajo de Fin de Máster, Universidad de Santiago de Compostela.

[12] Febrero M, (2012/2015). Apuntes de la asignatura *Datos Funcionales*.

[13] Cadarso C, (2014). Apuntes de la asignatura *Estadística No Paramétrica*.