Trabajo Fin de Máster

Modelización estadística en pruebas de esfuerzo y rendimiento deportivo

Marcos Matabuena Rodríguez

Máster en Técnicas Estadísticas

Curso 2016-2017

Propuesta de Trabajo Fin de Máster

Título en galego: Modelización estatística en probas de esforzo e rendemento deportivo.

Título en español: Modelización estadística en pruebas de esfuerzo y rendimiento deportivo.

English title: Statistical modeling in stress test and sports performance.

Modalidad: Modalidad A.

Autor/a: Marcos Matabuena Rodríguez, Universidade de Vigo.

Director/a: Ricardo Cao Abad, Universidade da Coruña; Mario Francisco Fernández, Universidade da Coruña.

Tutor/a:

Breve resumen del trabajo:

La metodología más habitual para medir de manera objetiva los factores asociados al rendimiento deportivo es a través de los parámetros obtenidos con una prueba de esfuerzo. El interés práctico de lo anterior es doble: poder orientar el entrenamiento deportivo de una manera científica e individualizada y especificar las necesidades concretas para cada deporte o prueba desde un punto de vista fisiológico. En este trabajo se llevará a cabo, en primer lugar, una breve recopilación bibliográfica acerca de los estudios de pruebas de esfuerzo, así como su impacto sobre el entrenamiento deportivo. El objetivo fundamental del trabajo será el análisis de datos reales de pruebas de esfuerzo, aplicando distintas técnicas estadístico-matemáticas. Posteriormente se relacionarán, con las técnicas estadísticas adecuadas, los resultados obtenidos en las pruebas de esfuerzo con el rendimiento deportivo de un grupo de patinadores de alto nivel y deportistas del centro de alto rendimiento de Pontevedra.

Recomendaciones: Conocimientos de técnicas estadísticas presentes en el Master de Técnicas Estadísticas de las tres universidades gallegas. En concreto, se prevé utilizar métodos de regresión, de análisis de datos funcionales y de series temporales.

Otras observaciones: Propuesta de TFM presentada por el alumno Marcos Matabuena Rodríguez, con el visto bueno de los directores arriba indicados. El TFM requerirá de la colaboración con Daniel Ruiz Rivera de KIA Speed Skating Academy, que facilitará muchos de los datos necesarios para llevar a cabo los análisis estadísticos y colaborará en la interpretación de las conclusiones, desde el punto de vista de las Ciencias del Deporte.

Ricardo Cao Abad, Catedrático de la Universidade da Coruña y Mario Francisco Fernández, Profesor Titular de la Universidade da Coruña, informan de que el Trabajo Fin de Máster titulado

Modelización estadística en pruebas de esfuerzo y rendimiento deportivo

fue realizado bajo su dirección por Marcos Matabuena Rodríguez para el Máster en Técnicas Estadísticas, estimando que el trabajo está terminado y dando su conformidad para su presentación y defensa ante un tribunal.

En A Coruña / Santiago de Compostela, a 11 de Enero de 2017.

El director:

CAO ABAD RICARDO JOSE - DNI 32769018K

Firmado digitalmente por CAO ABAD RICARDO JOSE - DNI 32769018K
Nombre de reconocimiento (DN): c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=certificado electrónico de empleado público, ou=DEPARTAMENTO DE MATEMATICAS, ou=966, serialNumber=32769018K, sn=CAO ABAD, givenName=RICARDO JOSE, cn=CAO ABAD RICARDO JOSE - DNI 32769018K
Fecha: 2017.01.10 13:39:27 +01'00'

Ricardo Cao Abad

El director:

FRANCISCO FERNANDEZ MARIO - 32660323R

Firmado digitalmente por FRANCISCO FERNANDEZ MARIO - 32660323R
Nombre de reconocimiento (DN): c=ES, serialNumber=32660323R, sn=FRANCISCO FERNANDEZ, givenName=MARIO, cn=FRANCISCO FERNANDEZ MARIO - 32660323R
Fecha: 2017.01.10 12:40:25 +01'00'

Mario Francisco Fernández

El autor:

Marcos Matabuena Rodríguez

Agradecimientos

- Me gustaría mostrar mi agradecimiento de manera especial y sincera a los profesores Ricardo Cao y Mario Francisco Fernández, que han dirigido este trabajo de fin de master. Su apoyo y confianza en mi trabajo han sido fundamentales para el desarrollo del mismo.
- Quiero expresar también mi agradecimiento a Daniel Ruiz Rivera y al doctor Fernando Huelín Trillo. Sin su colaboración la realización de este trabajo hubiese sido imposible.
- A Mari Carmen García Noya, a José Rodríguez del Río y a Rosana Rodríguez López, grandes profesionales de la enseñanza, que me han inculcado el amor por el conocimiento, las Matemáticas y la investigación.
- A mis padres por apoyarme en todas las decisiones que he tomado a lo largo de mi vida. Y sobretodo escucharme, ayudarme y animarme constantemente.
- A mi abuela Tena ejemplo de superación y entrega.
- A todos gracias.

Índice general

Resumen

Resumen en español

La metodología más habitual para medir de manera objetiva los factores asociados al rendimiento deportivo es a través de los parámetros obtenidos con una prueba de esfuerzo. El interés práctico de lo anterior es doble: poder orientar el entrenamiento deportivo de una manera científica e individualizada y especificar las necesidades concretas para cada deporte o prueba desde un punto de vista fisiológico. En este trabajo se llevará a cabo, en primer lugar, una breve recopilación bibliográfica acerca de los estudios de pruebas de esfuerzo, así como su impacto sobre el entrenamiento deportivo. El objetivo fundamental del trabajo será el análisis de datos reales de pruebas de esfuerzo, aplicando distintas técnicas estadístico-matemáticas. Posteriormente se relacionarán, con las técnicas estadísticas adecuadas, los resultados obtenidos en las pruebas de esfuerzo con el rendimiento deportivo de un grupo de patinadores de alto nivel y deportistas del centro de alto rendimiento de Pontevedra.

English abstract

The most common methodology used to measure objectively factors related to sports performance is through the parameters obtained from athlete stress tests. The practical use of the aforementioned is twofold: being able to direct the sport training in a scientific and individual way, and specifying the particular needs for each sport or test from physiological point of view. In this project, we will primarily create a biographical compilation based on studies of treadmill tests, as well as its impact on sports training. The fundamental aim of the project will be to analyse real data on treadmill tests, applying differents mathematical and statistical techniques. Then, using suitable statistical techniques, the results obtained in treadmill test will be related to the sporting performance of a group of top level skaters and athletes of the high performance center of Pontevedra.

Prefacio

Uno de los principales objetivos de la ciencia del entrenamiento deportivo es la búsqueda de los factores influyentes en el rendimiento deportivo. Las primeras investigaciones en este campo aparecieron a partir del siglo XX con los estudios del matemático y premio nobel de medicina A.V Hill y del fisiólogo alemán Otto Meyerhoff. Ambos científicos trabajaron sobre los factores limitantes del ejercicio físico en base a diferentes estados cardiovasculares, respiratorios y bioquímicos. Desde entonces, multitud de publicaciones han plagado la literatura científica, pero, pese a todo, las ideas principales de estos autores continúan siendo la base de la teoría fisiológica actual y por el momento siguen sin producirse cambios significativos.

No obstante, muchas de las investigaciones en esta área han sido realizadas de manera imprecisa (con muy pocos datos) y simplificada, utilizando herramientas matemáticas y estadísticas demasiado simples, sin beneficiarse del uso de técnicas avanzadas como pueden ser medidas de dependencia no lineal, datos longitudinales, modelos de regresión no paramétricos, métodos estadísticos para alta dimensión, datos funcionales, o técnicas de series de tiempo no lineales basadas en la teoría del caos y de la información.

Por otra parte, cada vez es más frecuente que los profesionales del deporte y los deportistas, en general, demanden herramientas con las que poder cuantificar y regular el entrenamiento de manera precisa y segura. Para poder realizar esto es necesario hacer uso de herramientas matemáticas y estadísticas. Algunos modelos se han propuesto al respecto, por ejemplo, [2] propusieron un modelo basado en técnicas de ecuaciones diferenciales para predecir los efectos de las sesiones de entrenamiento deportivo, o más recientemente [18] han propuesto una versión mejorada del modelo, en la que se permite tener en cuenta el efecto residual de las sesiones de entrenamiento a varios lapsos de tiempo a través de un sistema de ecuaciones diferenciales funcionales. Sin embargo, dichos modelos son insuficientes en la práctica o difíciles de utilizar y, por tanto, se deben proponer nuevas metodologías para poder utilizar los modelos ya existentes o para la creación de nuevos modelos.

El principal obstáculo metodológico existente a día de hoy es el de predecir la forma física a través de mecanismos indirectos que no le produzcan fatiga al deportista. Resolviendo dicho problema, entrenadores y deportistas podrían conocer el efecto individual que proporcionan las sesiones de entrenamiento sobre la condición física, además de contar con una valiosa herramienta con la que controlar la fatiga acumulada, evitando así posibles estados de sobreentrenamiento.

Este proyecto tiene como objetivo principal resolver alguno de los problemas que se acaban de comentar a través del análisis de datos reales deportivos con las técnicas estadísticas más vanguardistas. Particularmente se pretenden abordar los siguientes problemas:

- Descubrir qué factores fisiológicos afectan al rendimiento en el patinaje sobre hielo en pruebas de corta distancia.
- Analizar las relaciones existentes entre los parámetros fisiológicos en pruebas de esfuerzo y predecirlos en situaciones de interés como es el caso del consumo de oxígeno a partir de la potencia y la frecuencia cardíaca.
- Proporcionar el primer test indirecto propuesto en la literatura científica que no cause fatiga al deportista y que permita, a su vez, predecir su forma física de forma fiable.

La estructura del trabajo es la siguiente:

- Capítulo 1: Se realiza una breve introducción bibliográfica acerca de la literatura existente sobre las pruebas de esfuerzo, así como una explicación de las variables fisiológicas que se

pueden medir en una prueba de esfuerzo y el impacto que tienen estas en el entrenamiento deportivo. Además se describen las bases de datos de trabajo y se realiza un análisis descriptivo en una de ellas.

- Capítulo 2: Se introducen las técnicas de estimación no paramétrica en modelos de regresión, para luego aplicarlas en un problema de interés: predecir el consumo de oxígeno a través de las mediciones de potencia y frecuencia cardíaca.
- Capítulo 3: Se describen las principales técnicas de análisis de datos funcionales para luego ilustrarlas en la determinación de un test indirecto para predecir la forma física. En particular, se utiliza regresión y supervivencia funcional.
- Capitulo 4: El objetivo de este capítulo es el ajuste de dos modelos predictivos en las distancias de 500 y 1000 metros (patinaje sobre hielo) en función de una serie de variables fisiológicas. Además se incluye un análisis multivariante con los datos de trabajo en donde se aplica análisis de componentes tipo sparse, medidas de dependencia no lineal y un análisis cluster.

Nomenclatura

DII	Derivación bipolar electrocardiograma
ECG	Electrocardiograma
HR	Frecuencia cardíaca
RPE	Método de valoración de la actividad física
TA	Tensión arterial
$V1$	Derivación electrocardiograma. Se coloca el electrodo en el cuarto espacio intercostal
$V5$	Derivación electrocardiograma. Se coloca el electrodo en el quinto espacio intercostal
$VT1$	Umbral Aeróbico
$VT2$	Umbral Anaeróbico
$VT3$	Velocidad aeróbica máxima

Capítulo 1

Pruebas de esfuerzo y parámetros fisiológicos asociados

1.1. Introducción a las pruebas de esfuerzo

Una prueba de esfuerzo es una prueba diagnóstica, realizada en un laboratorio, que estudia la respuesta del sistema cardiovascular y respiratoria cuando al cuerpo se le somete a un esfuerzo controlado y progresivo. Es la metodología más usual para medir de manera objetiva los factores asociados al rendimiento deportivo. El interés práctico de lo anterior es doble: poder orientar el entrenamiento deportivo de una manera científica e individualizada a través de la prescripción de unas intensidades de trabajo y especificar las necesidades concretas para cada deporte o prueba (para cada nivel de rendimiento deportivo específico) desde un punto de vista fisiológico.

A su vez, las pruebas de esfuerzo tienen especial interés en la predicción y control de diversos problemas cardíacos, específicamente aquellos relacionados con la cardiopatía isquémica u otros problemas cardíacos ocultos, susceptibles de manifestarse con la actividad física y no detectables en reconocimientos estáticos.

Otras aplicaciones, menos conocidas sobre la utilización de los datos asociados a una prueba de esfuerzo, pueden ser la detección de estados de sobreentrenamiento de un deportista o incluso el consumo de ayudas ergogénicas y similares.

Una prueba de esfuerzo se puede realizar en diferentes máquinas, como puede ser una cinta de correr, una bicicleta estática o un ergómetro. Dadas las especificaciones de cada deporte concreto, se opta por realizar la prueba a cada deportista en el aparato que más le convenga. Sin embargo, la metodología que se utiliza en el procedimiento de realización de la prueba es la misma: al sujeto se le obliga a realizar ejercicio físico de manera progresiva hasta una determinada intensidad. Teniendo en cuenta lo anterior, es natural clasificar las pruebas de esfuerzo en dos tipos:

- Pruebas de esfuerzo maximales: Son aquellas en las que se le exige al individuo llegar hasta el máximo agotamiento.
- Pruebas de esfuerzo submaximales: Se define una variable fisiológica[1] y se fija el test de tal forma que se finaliza el mismo cuando se alcanza un porcentaje sobre esa variable fisiológica.

Mientras el individuo realiza la prueba de esfuerzo, además de medir su frecuencia cardíaca y tener un control sobre la intensidad de trabajo (medida en unidades de potencia o de velocidad) se lleva a cabo una medición de distintos parámetros respiratorios y cardiovasculares. Toda esta información obtenida nos proporciona un punto de referencia para el estudio de las relaciones entre la frecuencia cardíaca, velocidad, potencia (medidas sobre las que podemos tener un control en el exterior e incluso en competición) y las diferentes fases fisiológicas y bioquímicas que se producen en el ejercicio físico y que determinan los cambios inducidos con el entrenamiento, así como el rendimiento final en competición.

[1] Normalmente el V_{O_2} máx (consumo máximo de oxígeno), que se define como la máxima cantidad de oxígeno que un individuo puede absorber, transportar y consumir en un tiempo determinado, es decir, el máximo volumen de oxígeno en la sangre que su organismo puede transportar y metabolizar.

Figura 1.1: Ejemplo de prueba de esfuerzo

En la figura 1.1 se muestra una imagen de una prueba de esfuerzo.

1.1.1. Significado fisiológico de los parámetros principales

En la tabla 1 se menciona una serie de medidas y parámetros fisiológicos. Vamos a explicar en algunas de ellas con más detalle su significado y relevancia.

$C(a-v)O_2$

La diferencia arteriovenosa de oxígeno es la diferencia en el oxígeno contenido de la sangre entre la gasometría arterial y la sangre venosa. Es una medida que indica cuánto oxígeno se elimina de la sangre por los tubos capilares, o también puede interpretarse como la capacidad de la sangre para circular por el cuerpo. La diferencia arteriovenosa de oxígeno y el gasto cardíaco son los principales factores que permiten la variación en el consumo total de oxígeno del cuerpo y son importantes en la medición del VO_2. Se mide generalmente en mililitros de oxígeno por 100 mililitros de sangre ($mL/100\ mL$).

FEV_1

El Volumen Espiratorio Forzado (FEV_1) es una medida obtenida por espirometría, que equivale al volumen de aire exhalado del pulmón de manera forzada durante un segundo, después de haber tomado aire al máximo. El resultado se expresa en porcentaje y el valor normal en sujetos sanos, tanto hombres como mujeres, equivale a un 75 % de su capacidad vital pulmonar. El resultado tiene aplicación en medicina para determinar ciertas enfermedades del pulmón y, además, es uno de los parámetros más importantes de la espirometría. En líneas generales, la FEV_1 refleja las condiciones de las vías aéreas más gruesas.

MET

El MET es la energía gastada en reposo. Se define como la energía consumida (medida en julios) a un consumo de oxígeno de 3.5 $mL/(min/kg)$.

Abreviatura	Parámetro respiratorio o medida
CO_2	Dióxido de Carbono
$C(a-v)O_2$	Diferencia arteriovenosa de oxígeno
EOB	Número de respiraciones del ejercicio
F_{ECO_2}	Fracción de dióxido de Carbono espirado
F_{IO_2}	Fracción de oxígeno inspirado
FEV_1	Volumen espiratorio forzado
HR	Frecuencia cardíaca
O_2/HR	Medida que relaciona el incremento del consumo de O_2 y HR
MET	Medida de equivalencia metabólica
MVV	Ventilación voluntaria máxima
O_2	Oxígeno
V_{O_2}	Oxígeno inspirado
$P_{ET}CO_2$	Presión parcial, final respiración, dióxido de carbono
$P_{ET}O_2$	Presión parcial, final respiración, oxígeno
RER	Ratio cantidad de CO_2 producido, O_2 usado
V_{CO_2}	Dióxido de Carbono expulsado
V_D	Aire que se inhala y no toma parte del intercambio de gases
V_E	Volumen de gas exhalado o inhalado por minuto
V_E/V_{CO_2}	Equivalencia ventilatoria del CO_2
V_E/V_{O_2}	Equivalencia ventilatoria del O_2
VT	Umbral respiratorio

Tabla 1.1: Medidas o parámetros más usuales en la literatura asociados a una prueba de esfuerzo.

MVV

El MVV se define como el volumen máximo que puede ser exhalado por minuto por la respiración del sujeto tan rápido y profundamente como sea posible.

$P_{ET}CO_2$

La $P_{ET}CO_2$ se define como el pico final de CO_2 que se alcanza al final de la espiración.

$P_{ET}O_2$

La $P_{ET}CO_2$ se define como el pico final de O_2 que se alcanza al final de la espiración.

RER

La relación de intercambio respiratorio (RER) se define como la relación entre la cantidad de dióxido de carbono (CO_2) producido en el metabolismo y el oxígeno (O_2) que se utiliza. La relación se determina mediante la comparación de los gases exhalados en el aire dentro de una habitación. La medición de esta relación se puede utilizar para estimar el cociente respiratorio, una medida que indica el combustible (hidratos de carbono o grasa) que se está metabolizando para suministrar energía al cuerpo. Esta estimación sólo es válida si el metabolismo se encuentra en un estado de equilibrio.

RER toma valores aproximadamente de 0.8 en reposo con una dieta moderna. Este valor, sin embargo, puede ser superior a 1 durante el ejercicio intenso, en el que el RER es de alrededor de 1.1 en el VO_2 máx.

V_E/V_{O_2}

El equivalente ventilatorio para el oxígeno (V_E/V_{O_2}) es el cociente entre la ventilación en litros por minuto y el consumo de oxígeno en litros por minuto. Es un parámetro que indica la cantidad de aire en cm^3 que debe ventilarse para que el organismo pueda utilizar un cm^3 de oxígeno. Expresa, por tanto, el grado de eficacia de la ventilación pulmonar.

V_E/V_{CO_2}

El equivalente ventilatorio para el CO_2 (V_E/V_{CO_2}) es el cociente entre la ventilación en litros por minuto y la cantidad de CO_2 expulsado en litros por minuto. Expresa la relación entre el aire ventilado y el CO_2 expulsado.

VT

El VT es la zona del ejercicio a partir de la cual tenemos dificultades para hablar.

1.2. Metodología y protocolos clásicos de una prueba de esfuerzo

Una prueba de esfuerzo es una prueba médica que puede llegar a ser peligrosa si no se realiza en las condiciones adecuadas. Para realizar el test de forma correcta, debemos contar, en un primer lugar, con unas instalaciones adecuadas, suficientemente habilitadas con el material de auxilio necesario y, a continuación, seguir un protocolo y una metodología clara y perfectamente establecida. En esta sección, pretendemos enunciar los principales requisitos operacionales para realizar una prueba de esfuerzo de manera correcta.

Laboratorio

- Local.
  - Suficientemente amplio (al menos 10 m^2).
  - Condiciones ambientales: bien ventilado e iluminado, con una temperatura entre los $20°C$ y los $23°C$ y una humedad relativa del 60 % al 65 %.
  - Disponer de servicios complementarios (vestuario, duchas y aseos).
- Instrumental.
  - Ergómetro: habitualmente cinta rodante y/o cicloergómetro.
  - Electrocardiógrafo: tricanal, osciloscopio o monitor, equipo computarizado.
  - Esfigmomanómetro.

  - Sistema de medida del volumen y composición del gas espirado: ergoespirómetro (saco de Douglas, respiración a respiración).
  - Para pruebas de tipo invasivo: analizador de gases sanguíneos, de ácido láctico, etc.
- Equipo de emergencias cardio-respiratorias.
  - Desfibrilador.
  - Material de intubación y ventilación.
  - Fuente de oxígeno, fármacos.
- Personal.
  - Médico.
  - Enfermera o asistente.

Protocolos pruebas de esfuerzo

- Cicloergómetro.
  - Ventajas: fácil control y seguridad, mejor registro ECG, manejabilidad en toma TA y muestras de sangre, menor error.
  - Inconvenientes: mayor fatiga muscular, incapacidad pedaleo por falta de coordinación, menor consumo de oxígeno.
  - Protocolos:
    - Incrementos de $10 - 20\ W$ cada min o $25\ W$ cada $2\ min$ (cardiología).
    - Incrementos de $15 - 25\ W$ cada min en rampa (ergoespirometría).
    - Incrementos de $30 - 50\ W$ cada $3 - 5\ min$ (valoración metabólica).

    En cicloergómetro (sólo para ciclistas), un protocolo muy habitual para la determinación del umbral anaeróbico por el método metabólico (análisis del ácido láctico), consiste en un inicio a $50\ W$ y aumento de la carga en $50\ W$ ($25\ W$ en mujeres) cada 3 minutos, tomando la muestra de sangre en los últimos segundos de cada escalón. En el caso de no precisar muestras sanguíneas, se realiza un protocolo en rampa con inicio a $50\ W$ e incrementos de $25\ W$ al minuto ($5\ W$ cada 12 segundos). Aunque existen diversas variaciones de este método.
- Cinta de correr.
  - Ventajas: ejercicio más natural, mayor consumo de oxígeno.
  - Inconvenientes: difícil transporte, peor registro ECG, difícil medición TA y toma de muestras de sangre, menor costo.
  - Protocolos.
    - Bruce y Bruce modificado (cardiología).
    - Incrementos de $1 - 2\ km/h$ cada $1 - 2\ min$ (valoración ergoespirométrica).
    - Incrementos de $1 - 2\ km/h$ cada $3 - 5\ min$ (valoración metabólica).

    En la cinta de correr, un protocolo muy habitual para la determinación del umbral anaeróbico es a través de los valores de lactato. Se inicia a una velocidad de $8\ km/h$ ($6\ km/h$ en mujeres), con duración de cada escalón de 3 minutos y aumento de la velocidad en $2\ km/h$. La duración de la pausa entre cada escalón de velocidad, para la toma de muestras sanguíneas, es de 30 segundos. El protocolo continuo, si no se realizan tomas de muestras de sangre, es en rampa con incrementos de $1\ km/h$ cada minuto ($0.25\ km/h$ cada 15 segundos) y no se realiza pausa.

  Metodologia de las pruebas de esfuerzo

  - Preparación del sujeto.

- Recomendaciones previas: ropa, comida, descanso, estimulantes, fármacos.
- Evaluación previa: historia clínica, exploración general y cardio-respiratoria.
- Explicación detenida de la prueba y familiarización con el ergómetro.
- Colocación electrodos y cables: rasurar, limpiar, fijar.
- Hiperventilación (si anomalías previas en la repolarización).

- Control de la prueba.
  - Determinaciones en reposo: HR, TA, ECG, determinaciones metabólicas.
  - Calentamiento.
  - Control paciente.
  - Control parámetros elementales: HR, TA, ECG (DII, $V1$ y $V5$).
  - Control otros parámetros: gases, lactato, etc.
- Periodo de recuperación.
  - Tiempo: mínimo 5 min (seguir ejercicio).
  - Posición: erecta, decúbito supino (ECO de esfuerzo).
  - Datos: determinaciones, síntomas y valoración esfuerzo (RPE).

1.3. Importancia de las pruebas de esfuerzo en el entrenamiento deportivo

El conocimiento actual sobre el cuerpo humano es cada vez más notable y no solo eso, evolucionamos a un contexto donde dicho conocimiento puede ser asimilable y transferible a otras áreas del conocimiento. Por otro lado, los métodos de preparación física siguen mejorando de forma significativa y esto es debido en gran medida a lo comentado anteriormente, junto a los diferentes avances tecnológicos producidos en los últimos tiempos.

Sin embargo, si la ciencia del entrenamiento deportivo quiere seguir mejorando, esta debe hacer uso de técnicas estadísticas y matemáticas. Precisamente existe una teoría, la llamada teoría de preparación funcional, que permite unificar la teoría clásica de entrenamiento con las matemáticas. De hecho, esta conexión es lo que determina las bases del entrenamiento deportivo moderno y futuro[2]. Para explicar esta relación, vamos a hacer uso del gráfico mostrado en la figura 1.2.

En la figura 1.2 podemos observar que, en función del grado de esfuerzo realizado con el ejercicio físico, se producen diferentes adaptaciones metabólicas y biomecánicas en el organismo. Además es importante recordar que estas adaptaciones están relacionadas temporalmente con los entrenamientos anteriores (pues dejan su huella en el organismo, en forma de fatiga, activando distintas vías metabólicas).

Como acabamos de ver, la preparación funcional es, por tanto, el punto de partida para relacionar, en términos cuantitativos, la carga de entrenamiento realizada con las adaptaciones producidas en el organismo, las cuales determinan, de manera bastante importante, el rendimiento final en competición. Una vez entendido esto, el siguiente paso es establecer una serie de variables fisiológicas de interés para poder organizar el entrenamiento deportivo (en función de sus adaptaciones). De esta cuestión nos encargaremos a continuación, donde proporcionaremos las variables de máximo interés. Una vez realizada esta tarea, ya tenemos una serie de variables con las que poder controlar el cansancio y las adaptaciones del entrenamiento deportivo y, por tanto, estamos en condiciones de realizar planes de entrenamiento individualizados y desde una perspectiva científica. Para estimar de manera precisa el valor de dichos parámetros necesitamos recurrir a una prueba de esfuerzo.

[2]Aunque no exista el concepto del entrenamiento deportivo basado en datos, existen profesionales que están trabajando en construir unas bases para poder realizar planes de entrenamiento inteligente únicamente con el conocimiento basado exclusivamente en datos, [10].

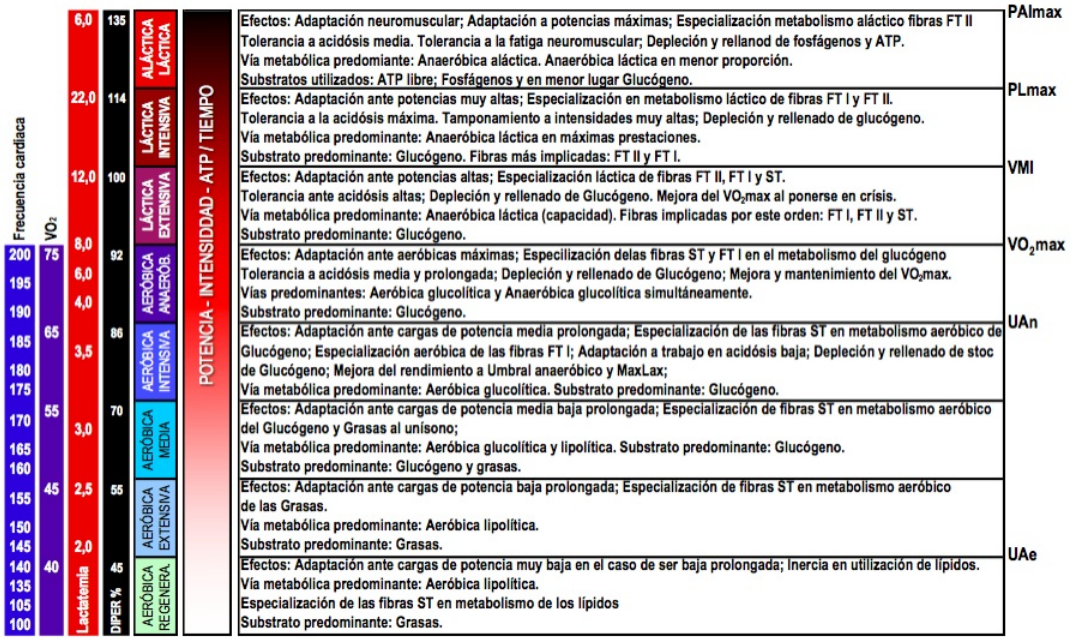


Figura 1.2: Adaptaciones del entrenamiento a diferentes intensidades.

1.4. Indicadores fisiológicos en el rendimiento deportivo

1.4.1. El consumo máximo de oxígeno: VO_2 máx

El consumo máximo de oxígeno está considerado como uno de los mejores parámetros posibles para indicar la capacidad de rendimiento en deportistas. La potencia aeróbica máxima, o consumo máximo de O_2, equivale a la máxima cantidad de oxígeno que un organismo puede extraer de la atmósfera y transportar hasta el tejido para allí ser utilizado. También se utilizan otros términos como consumo máximo de oxígeno, capacidad aeróbica de trabajo y capacidad de resistencia.

La capacidad aeróbica máxima es cuantitativamente equivalente a la cantidad máxima de oxígeno que un individuo puede consumir por unidad de tiempo durante una actividad que aumenta de intensidad progresivamente, realizada con un grupo muscular importante y hasta el agotamiento. Cuando es expresada en términos de oxígeno, suele escribirse como el máximo volumen de oxígeno por minuto y se abrevia como VO_2 máx. Es importante aclarar que en deportes como el remo, en los que es importante la respuesta del trabajo total, suele expresarse como volumen absoluto por minuto, mientras que en actividades como las carreras de fondo, en las que se soporta el peso del cuerpo durante la competición, se expresa como volumen por minuto en relación al peso corporal. Los buenos deportistas que participan en pruebas de resistencia son los que suelen tener asociados valores de consumo de oxígeno más altos. Los valores relativos más altos registrados aparecieron en deportes como el ski de fondo donde actúan de manera combinada pies y brazos y, a continuación, en deportes como el atletismo en pruebas de medio-fondo y fondo y ciclismo en ruta. Sin embargo, los valores absolutos más altos se han encontrado en deportistas de constituciones grandes y bien entrenados, como los remeros, que poseen una masa muscular importante y no tienen que soportar el peso de su cuerpo en la actividad específica de competición.

No se conoce con certeza el valor del VO_2 máx que se puede asociar al entrenamiento deportivo o a las cuestiones genéticas, sin embargo existen reglas empíricas que afirman que el incremento que podemos conseguir con el entrenamiento sobre el VO_2 máx, se encuentra en torno a 15 y 25 puntos (un sujeto adulto normal posee entre 35 y 40 puntos).

1.4.2. El umbral anaeróbico

En la década de los 60, Wasserman, Mclloroy (1964) [30] introdujeron el término umbral anaeróbico, el cual se definió como la carga de trabajo máxima que puede mantener un sujeto utilizando exclusivamente la fuente de energía obtenida por la vía aeróbica. Más adelante, se iría incrementando el interés de este concepto, sobre todo a raíz del trabajo de Mader, Heck (1986) [17], que indicaban que el umbral anaeróbico es la intensidad de trabajo más alta que el individuo puede soportar sin acumular lactato en concentraciones inferiores a 4 $mmol \cdot L^{-1}$. Este valor se encontraría entonces, entre concentraciones de alrededor de 2 $mmol \cdot L^{-1}$ y 4 $mmol \cdot L^{-1}$ por encima del umbral aeróbico que es la zona a partir de la cual el organismo empieza a tener dificultades para reciclar lactato.

El interés del umbral anaeróbico en el rendimiento deportivo es el identificar un parámetro que sea predictivo de la capacidad de rendimiento en disciplinas de larga duración. Según un gran número de autores, el umbral anaeróbico (AT) correlaciona mejor con la capacidad de rendimiento que con otros parámetros como el VO_2 máx, masa magra, porcentaje de fibras musculares, economía de carrera, etc., también relacionados con el metabolismo aeróbico. Farrel, Wilmore, Coyle, Billings, Costill (1975) [8] encontraron una correlación de $r = 0.98$ entre la velocidad de carrera en maratón y la velocidad en el umbral anaeróbica (tambien llamada velocidad en el AT), se encontró la misma correlación con la velocidad de una hora de carrera a máxima intensidad y también se encontró una fuerte relación ($r = 0.92$) con la velocidad de carrera en 10000 m [22].

El segundo aspecto que acentúa la importancia del AT es el gran potencial para la modificación de este parámetro, por ejemplo, Davis, Frank, Whipp, Wasserman (1984) [6] hallaron aumentos en el VO_2 máx de no más de un 25 %, mientras que en el umbral anaeróbico, se encontraron variaciones de un 44 % tras solo un periodo de entrenamiento de 9 semanas.

Es muy importante comentar que el porcentaje de AT respecto al VO_2 máximo varía considerablemente en función de si estamos hablando de un atleta de élite o de un deportista amateur, encontrando valores por encima del 80 % en el primer caso y en torno al 50 % en el segundo.

Uno de los grandes problemas del AT es encontrar un consenso acerca de sus bases bioquímicas, incluso hay autores que se niegan hablan del término umbral anaeróbico porque afirman que no existe, por lo menos bajo sus supuestos iniciales. La causa de fondo de este problema es identificar los factores responsables de la acidosis, un tema complejo dentro del campo de la bioquímica.

1.4.3. La velocidad aeróbica máxima: VAM

La VAM o velocidad aeróbica máxima, es un concepto que relaciona la velocidad de carrera con el consumo máximo de oxígeno. Los pioneros en su estudio fueron Astrand, Rodahl (1986) [1] que serían los primeros en investigar las relaciones existentes entre el máximo consumo de oxígeno y la capacidad para traducirlo en rendimiento mecánico.

La VAM es definida como la mínima velocidad de carrera a la que se obtiene el VO_2 máx y es uno de los mejores indicadores para predecir el rendimiento deportivo. Es una de las pocas medidas del rendimiento deportivo en las que entra en juego de manera conjunta la eficiencia metabólica y la mecánica del deportista.

Existen dos alternativas para estimar la VAM, o bien de forma casi exacta con una prueba de esfuerzo, o sino, con una determinación estadística, utilizando algún test físico como puede ser el test de Conconi [5].

1.5. Datos a analizar

Para realizar este trabajo, hemos contado con dos bases de datos:

- Una base de datos procedente de la localidad alemana de Inzell de la Academia KIA Speed skating Academy facilitada por Daniel Ruiz Rivera. La muestra contiene datos de patinadores de alto nivel. Para más detalles sobre la valoración funcional en este deporte puede verse [25].

  El objetivo es el de realizar un análisis descriptivo de los datos y, a su vez, diversos análisis multivariantes y de series de tiempo. Además se relacionarán estos análisis con los resultados en competición.

Abreviatura	Parámetro respiratorio o medida
HR	Frecuencia cardíaca medida en latidos minuto
$V'O2/kg$	Oxígeno por kg de peso cada min
$V'O2/HR$	Ratio entre oxígeno en ml y frecuencia cardíaca
WR	Potencia en W
$V'E/V'O2$	Ratio volumen espirado entre volumen oxígeno
$V'E/V'CO2$	Ratio volumen espirado entre volumen dióxido de carbono
RER	Ratio cantidad de CO_2 producido, O_2 usado
$V'E$	Volumen espirado en L/min
$V'T$	Volumen total en L
BF	Frecuencia de respiración

Tabla 1.2: Medidas o parámetros utilizados en las pruebas de esfuerzo del trabajo con los datos de Inzell.

- Una muestra de datos procedente del centro de tecnificación de Pontevedra recogida por el autor de este trabajo de fin de master con la autorización del jefe de servicios médicos del centro, Fernando Huelín Trillo.

  Estos datos permiten tratar las pruebas de esfuerzo mediante técnicas de análisis de datos funcionales.

1.5.1. Base de datos de Inzell

Se dispone de los resultados de 46 pruebas de esfuerzo de patinadores de alto nivel, de los cuales en 24 además se disponen de medidas en continuo. La mayor parte de los datos son de deportistas con edades comprendidas entre los 18 y 25 años, muchos de ellos deportistas en categoría Junior que están iniciando su etapa como profesionales. Las pruebas las realizan todos en bicicleta estática y el procedimiento es idéntico. Se comienza con 3 minutos de trabajo a una intensidad de 50 W y, a partir del minuto 3 se incrementa la potencia a razón de 2 W cada segundo, hasta que el deportista no puede más. Finalmente y una vez acabado este test, se le obliga al deportista a realizar 5 minutos más en la bicicleta a una potencia de 50 W.

En los resultados tenemos información para el umbral aeróbico, VT1, el umbral anaeróbico, VT2, y velocidad aeróbica máxima, VT3, de los parámetros incluidos en la tabla 1.2.

En 24 de las 46 pruebas de esfuerzo contamos además con mediciones (cada 3 segundos) de las medidas anteriores en continuo. En todas las pruebas con datos en continuo tenemos garantizado al menos 5 minutos de información sobre la capacidad de recuperación del deportista con los datos asociados a la frecuencia cardíaca y a los parámetros respiratorios comentados en la tabla 1.2. Nótese que esta es una importante medida sobre la forma física del deportista, pues cuanta mejor condición física, mayor capacidad de recuperación.

1.5.2. Base de datos de Pontevedra

Disponemos de alrededor de 400 pruebas de esfuerzo de deportistas de diferentes deportes: atletismo, bádminton, balonmano, baloncesto, ciclismo, fútbol, judo, lucha, natación, piraguismo, remo, squash, taekwondo, tenis, triatlón y vela.

La mayor parte de los datos son de jóvenes con edades comprendidas entre los 13 y 18 años. Además, en algunos casos podemos realizar estudios longitudinales (al disponer de varios registros por atleta, no más de 10), lo que nos permite tener información acerca de la evolución temporal del rendimiento de estos atletas.

Las pruebas se realizan tanto en cinta de correr, bicicleta estática como con ergómetro. Al tener deportistas de diferentes deportes y niveles, se intenta que las condiciones de las pruebas sean lo más homogéneas posibles: se intenta establecer el test de tal manera que los deportistas realicen un esfuerzo máximo de una duración aproximada de 13 minutos de tiempo. Se empieza a diferentes intensidades (en función del nivel) y se va incrementando la dureza de la prueba de manera constante (por ejemplo, en las pruebas realizadas en una cinta de correr, que será las que analicemos, 1 km/h cada minuto de media). En todas las pruebas de esfuerzo disponemos de mediciones (cada 5 segundos) de las siguientes variables que toman valores funcionales sobre el conjunto de instantes de tiempo en el que transcurre la prueba:

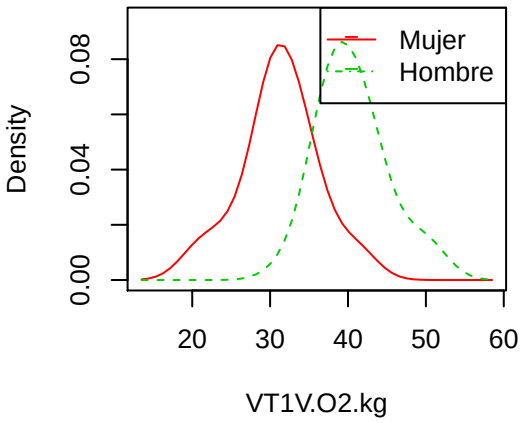
- Velocidad o Potencia.
- $EqCO_2$, V_{CO_2} en la tabla 1.1 anterior.
- EqO_2, V_{O_2} en la tabla 1.1 anterior.
- O_2/HR.
- HR.
- VE en L/min.
- VO_2 en ml/min.
- VCO_2 en ml/min.
- RER.
- VO_2 por kg de peso.

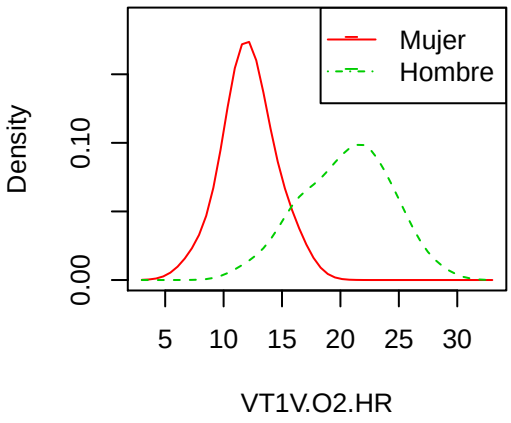
1.5.3. Estudio descriptivo de los datos de Inzell

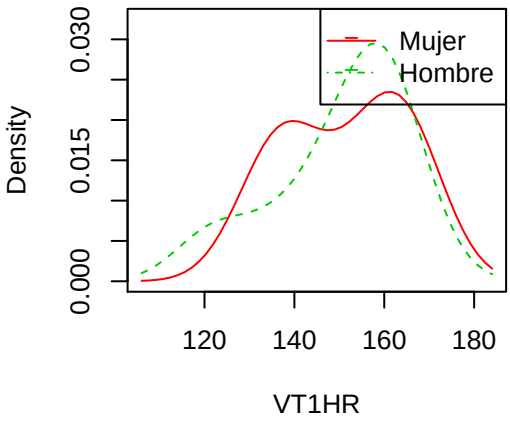
En esta sección realizaremos un análisis descriptivo de los datos asociados a los resultados de los patinadores de alto alto nivel de Inzell. Se dispone de una muestra de 33 variables, 3 categóricas y 30 continuas, las cuales se repiten en grupos de 10 (son las variables descritas anteriormente) tomando diferentes valores para los umbrales $VT1$, $VT2$ y $VT3$ (umbral aeróbico, anaeróbico y umbral láctico o velocidad aeróbica máxima) que serán descritos en el siguiente capítulo. En ellas tenemos un total de 46 individuos, 33 hombres y 13 mujeres. Para realizar el estudio descriptivo, en primer lugar representaremos los estimadores no paramétricos de las funciónes de densidad (figuras 1.3-1.5), para después pasar a calcular, para cada uno de los umbrales, los estadísticos más comunes por sexos (tablas 1.3-1.4). Por último, aplicaremos contrastes de hipótesis paramétricos y no paramétricos, para comprobar si existen o no diferencias significativas entre cada uno los grupos (tabla 1.5). Además de la inclusión de toda esta información, también se comentarán los resultados, explicando las diferencias en los valores entre hombres y mujeres y el posible significado fisiológico de los resultados obtenidos.

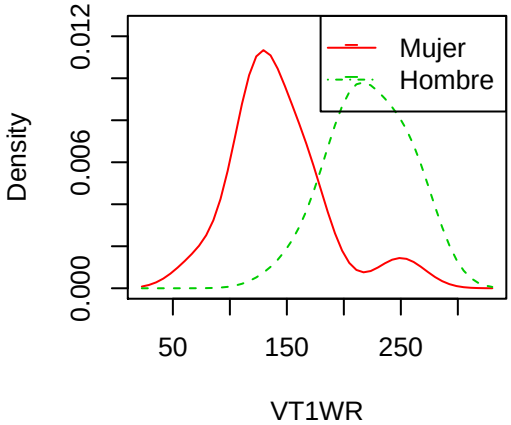
$VT1$

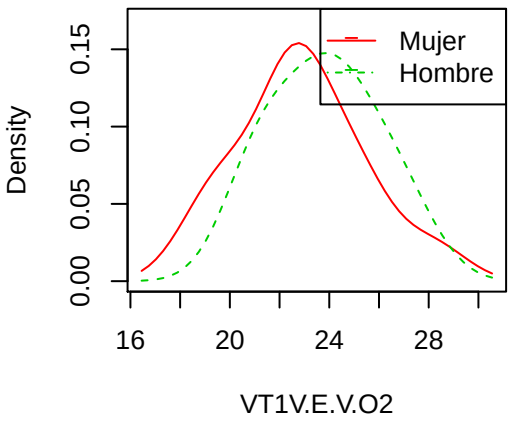
En las tablas asociadas al umbral $VT1$, vemos que, excepto en las variables asociadas a los índices respiratorios normalizados y a la HR y al BF, las diferencias estadísticas entre sexos son significativas. Las causas para estos resultados son razonables, pues las mujeres tienen un consumo

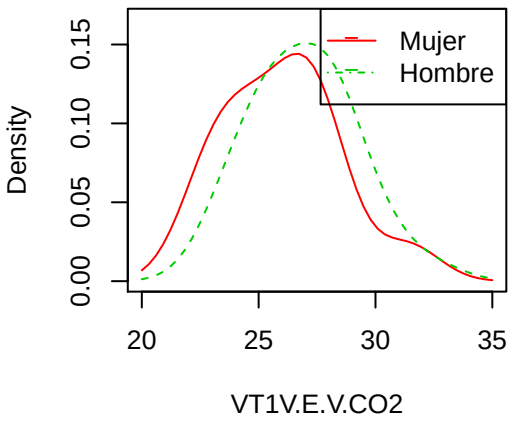

Mujer
Hombre
Density
0.00
0.04
0.08
20
30
40
50
60
VT1V.O2.kg


Mujer
Hombre
Density
0.00
0.10
5
10
15
20
25
30
VT1V.O2.HR


Mujer
Hombre
Density
0.000
0.015
0.030
120
140
160
180
VT1HR


Mujer
Hombre
Density
0.000
0.006
0.012
50
150
250
VT1WR


Mujer
Hombre
Density
0.00
0.05
0.10
0.15
16
20
24
28
VT1V.E.V.O2


Mujer
Hombre
Density
0.00
0.05
0.10
0.15
20
25
30
35
VT1V.E.V.CO2

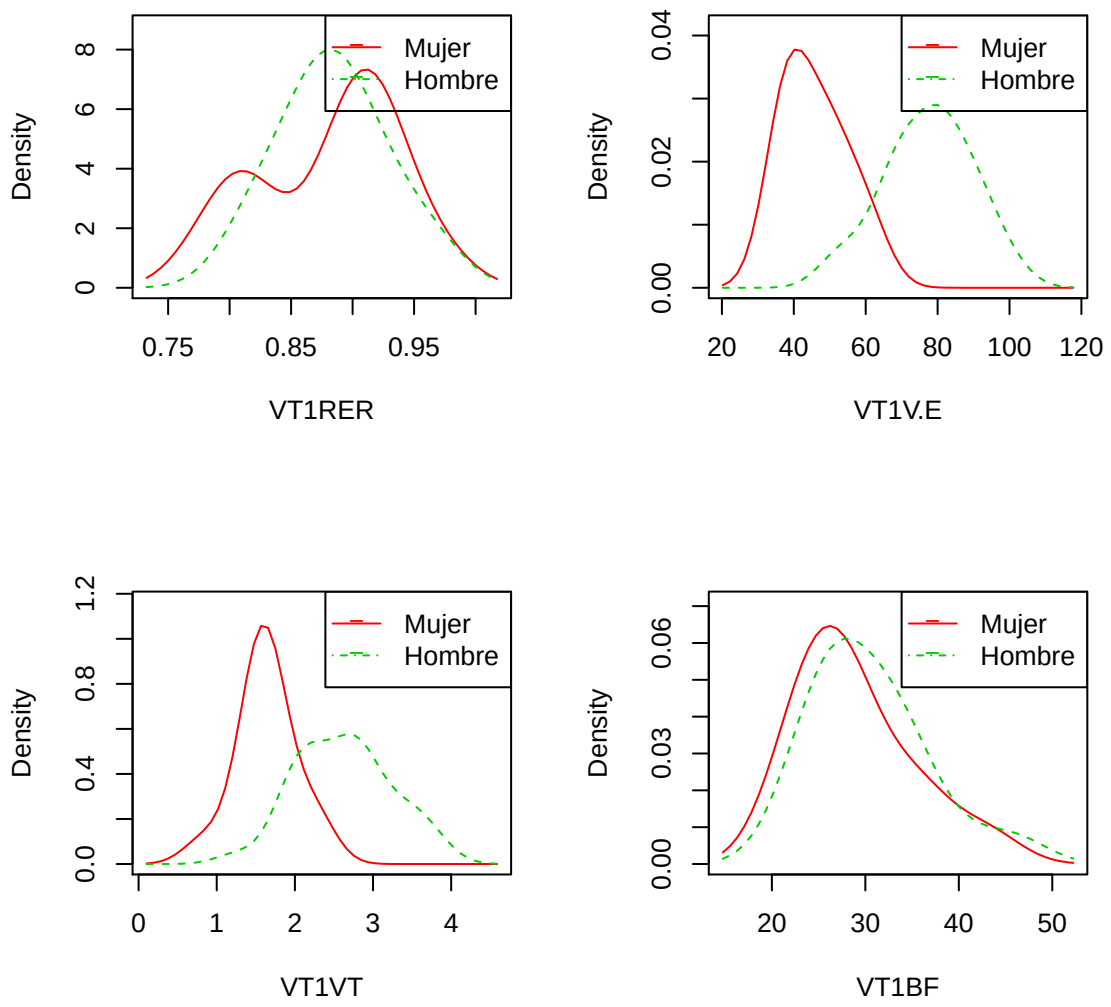
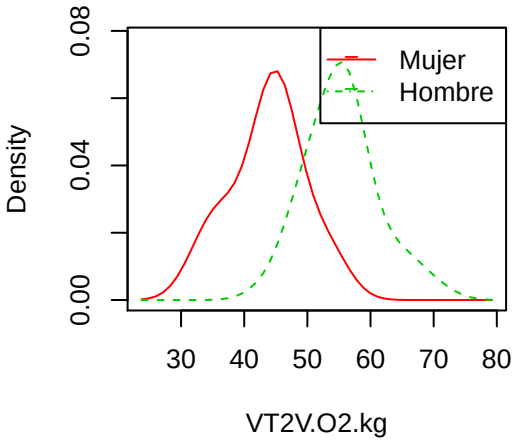
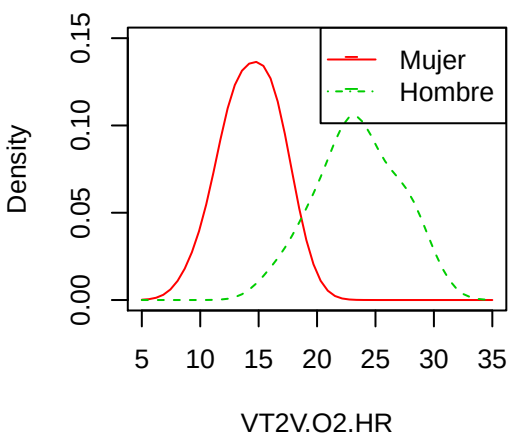
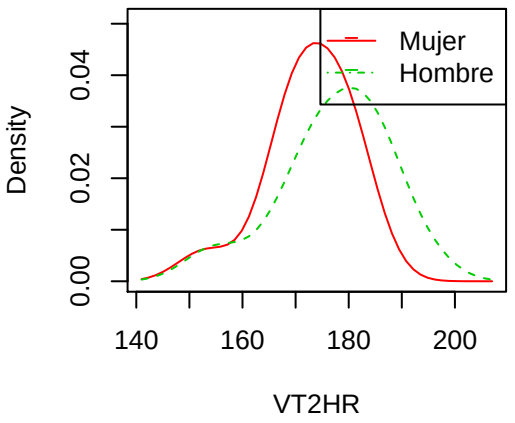
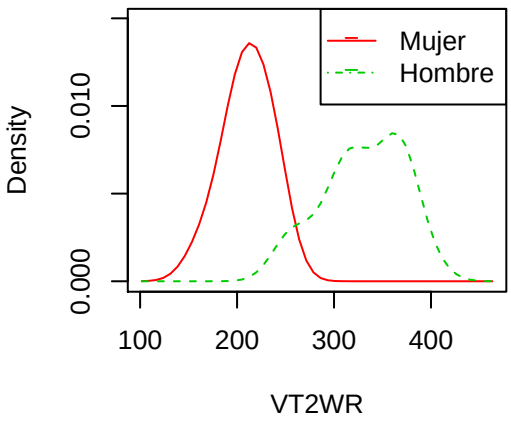


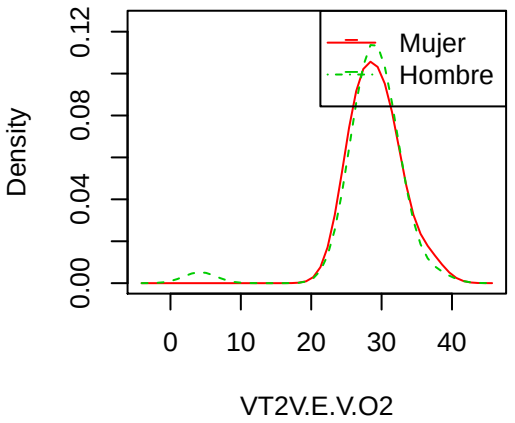
Figura 1.3: Gráficos por sexos de la densidad de las variables asociadas al umbral aeróbico $VT1$. Se muestran las densidades (por filas) de $V'O2/kg$, $V'O2/HR$, HR, WR, $V'E/V'O2$, $V'E/V'CO2$, RER, $V'E$, $V'T$ y BF.

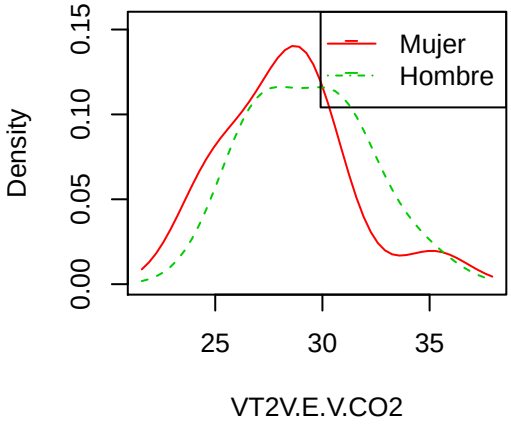

Density
0.00
0.04
0.08
30
40
50
60
70
80
VT2V.O2.kg
Mujer
Hombre


Density
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
VT2V.O2.HR
Mujer
Hombre


Density
0.00
0.02
0.04
140
160
180
200
VT2HR
Mujer
Hombre


Density
0.000
0.010
100
200
300
400
VT2WR
Mujer
Hombre


Density
0.00
0.04
0.08
0.12
0
10
20
30
40
VT2V.E.V.O2
Mujer
Hombre


Density
0.00
0.05
0.10
0.15
25
30
35
VT2V.E.V.CO2
Mujer
Hombre

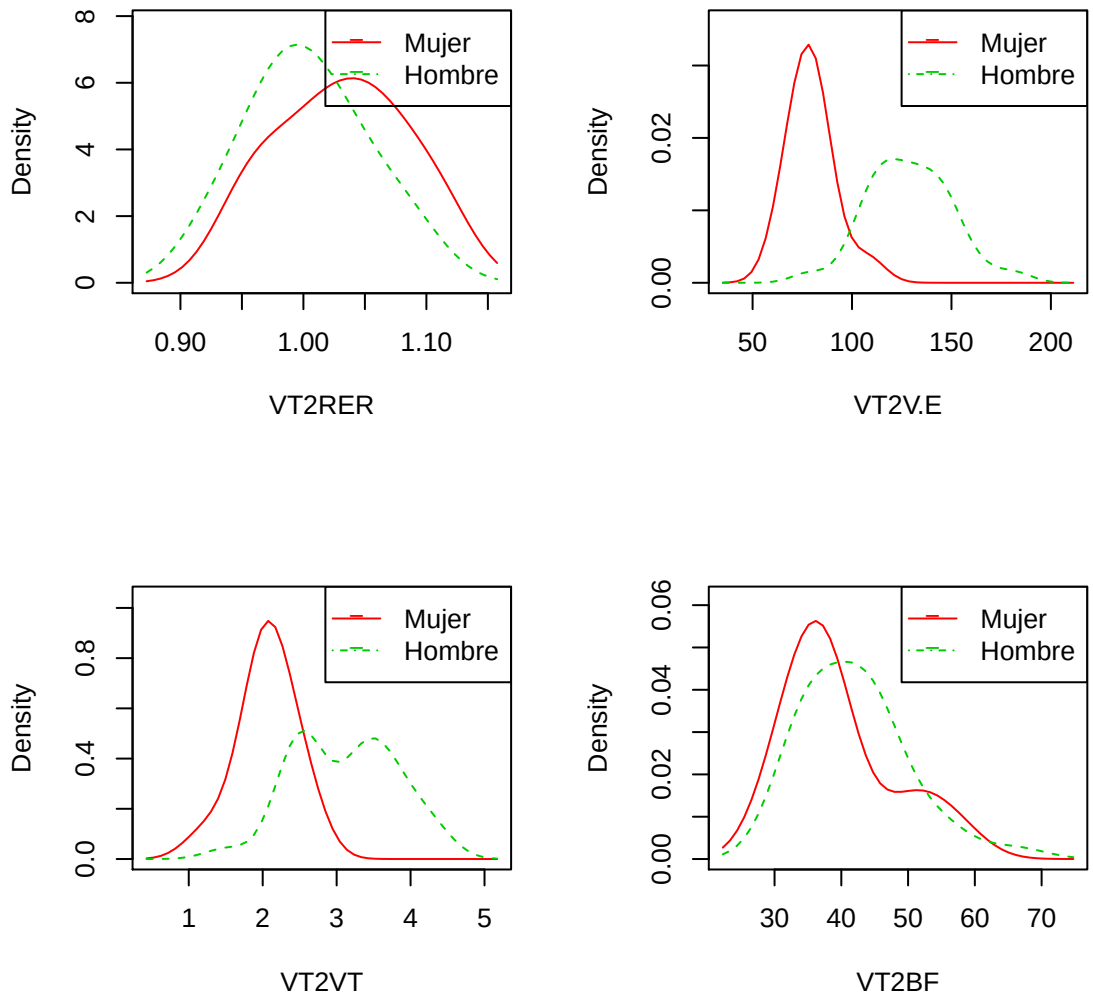
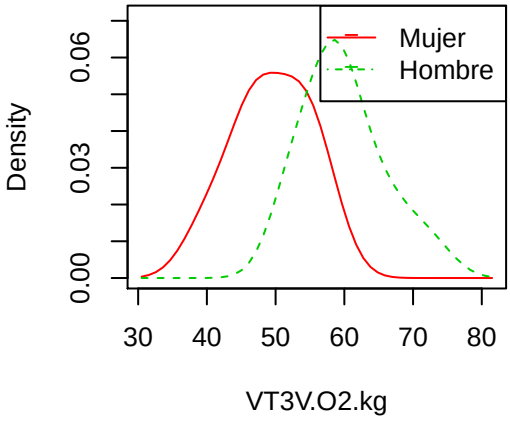
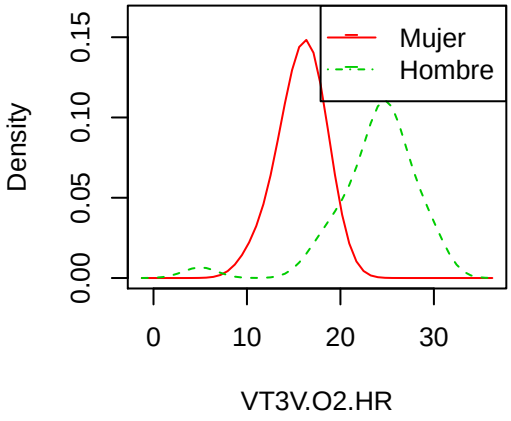
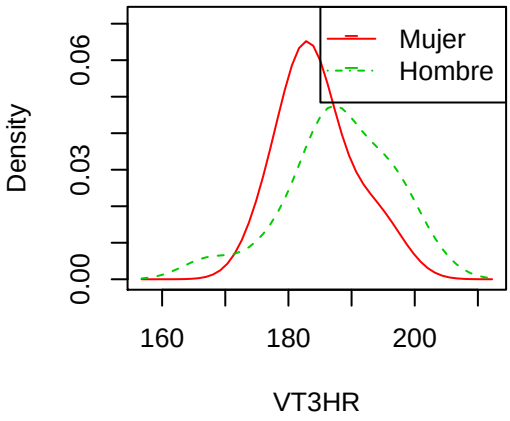
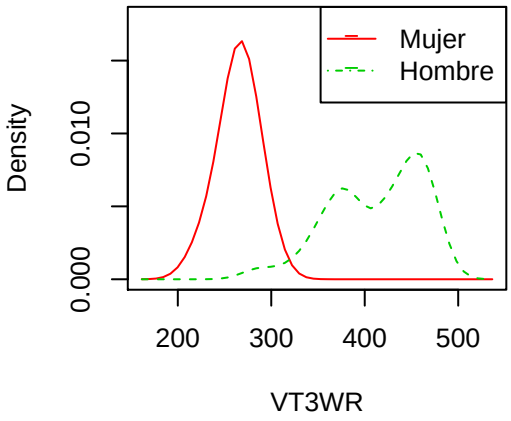


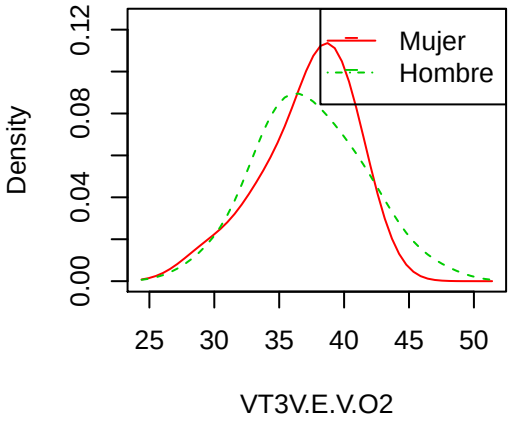
Figura 1.4: Gráficos por sexos de la densidad de las variables asociadas al umbral aeróbico $VT1$. Se muestran las densidades (por filas) de $V'O2/kg$, $V'O2/HR$, HR, WR, $V'E/V'O2$, $V'E/V'CO2$, RER, $V'E$, $V'T$ y BF.

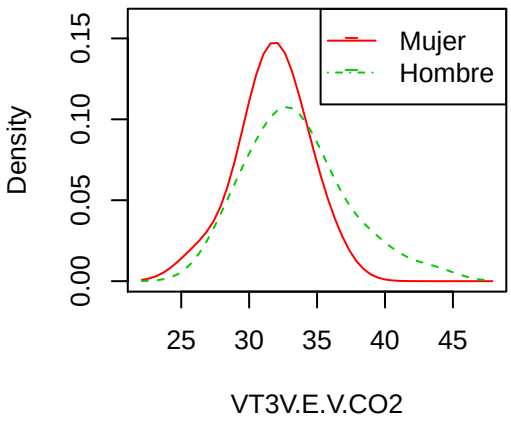

Density
0.00
0.03
0.06
Mujer
Hombre
30
40
50
60
70
80
VT3V.O2.kg


Density
0.00
0.05
0.10
0.15
Mujer
Hombre
0
10
20
30
VT3V.O2.HR


Density
0.00
0.03
0.06
Mujer
Hombre
160
180
200
VT3HR


Density
0.000
0.010
Mujer
Hombre
200
300
400
500
VT3WR


Density
0.00
0.04
0.08
0.12
Mujer
Hombre
25
30
35
40
45
50
VT3V.E.V.O2


Density
0.00
0.05
0.10
0.15
Mujer
Hombre
25
30
35
40
45
VT3V.E.V.CO2

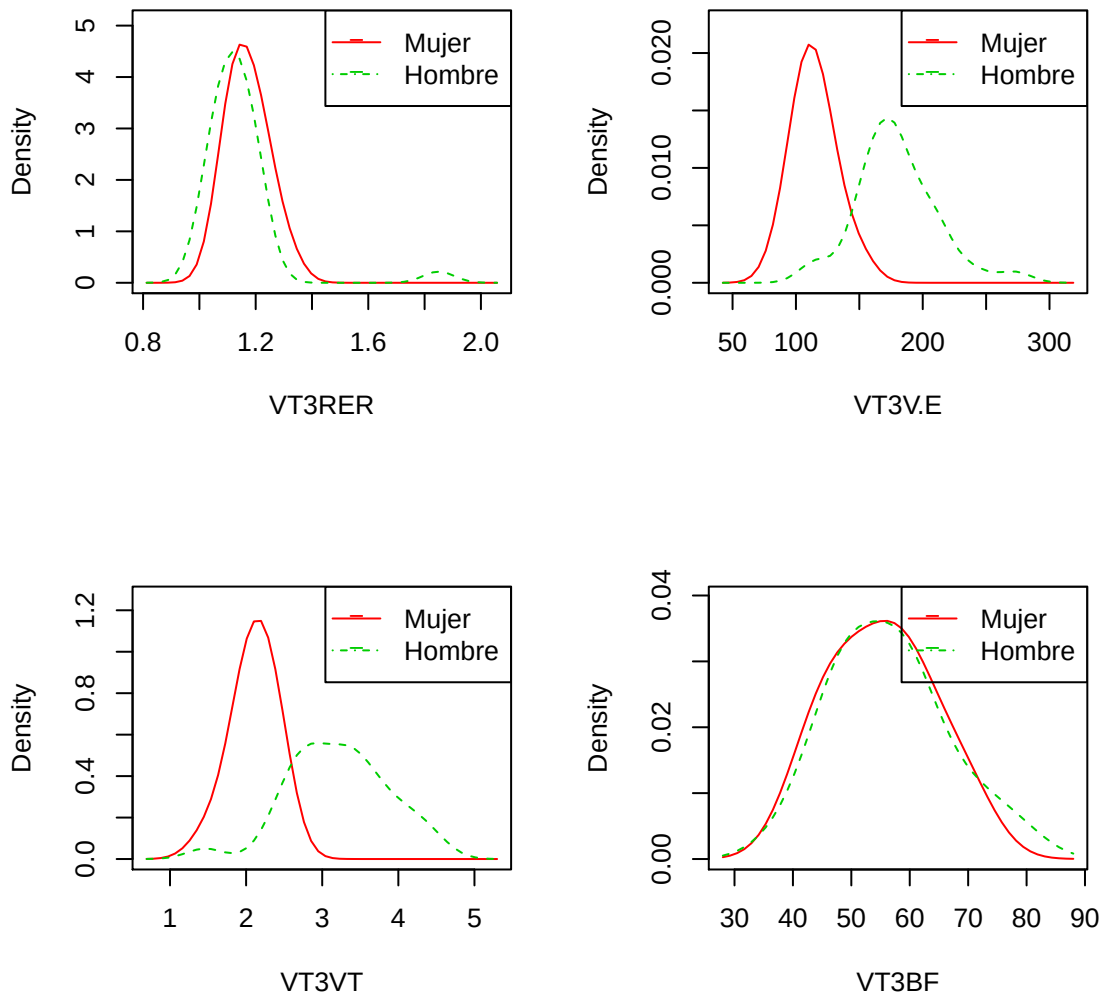


Figura 1.5: Gráficos por sexos de la densidad de las variables asociadas al umbral aeróbico $VT1$. Se muestran las densidades (por filas) de $V'O2/kg$, $V'O2/HR$, HR, WR, $V'E/V'O2$, $V'E/V'CO2$, RER, $V'E$, $V'T$ y BF.

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT1V.O2.kg	40.58	4.32	40.00	32.00	51.00
VT1V.O2.HR	20.45	3.72	20.00	12.00	28.00
VT1HR	150.58	14.17	156.00	119.00	171.00
VT1WR	221.67	32.87	216.00	143.00	279.00
VT1V.E.V.O2	23.79	2.11	24.00	20.50	27.80
VT1V.E.V.CO2	26.92	2.19	26.90	23.30	32.50
VT1RER	0.88	0.04	0.88	0.81	0.97
VT1V.E	77.36	12.30	78.80	50.80	101.50
VT1VT	2.62	0.60	2.64	1.26	3.84
VT1BF	30.48	6.11	29.00	22.00	46.00

(a) VT1 Hombres

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT1V.O2.kg	31.23	4.97	31.00	21.00	41.00
VT1V.O2.HR	12.15	1.99	12.00	8.00	16.00
VT1HR	151.23	13.57	155.00	130.00	170.00
VT1WR	142.38	41.51	127.00	74.00	250.00
VT1V.E.V.O2	22.89	2.56	22.60	18.80	28.20
VT1V.E.V.CO2	26.00	2.47	26.30	22.50	31.40
VT1RER	0.88	0.06	0.90	0.78	0.97
VT1V.E	45.87	8.66	44.50	36.50	61.70
VT1VT	1.64	0.35	1.63	0.85	2.27
VT1BF	28.92	6.24	27.00	21.00	43.00

(b) VT1 Mujeres

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT2V.O2.kg	55.42	5.52	56.00	44.00	70.00
VT2V.O2.HR	23.45	3.47	23.00	16.00	30.00
VT2HR	177.39	10.32	179.00	152.00	196.00
VT2WR	330.45	41.77	332.00	243.00	403.00
VT2V.E.V.O2	28.48	5.08	29.00	4.20	37.40
VT2V.E.V.CO2	29.32	2.63	29.50	24.80	35.20
VT2RER	1.00	0.05	1.00	0.92	1.10
VT2V.E	128.75	21.12	129.60	78.60	181.90
VT2VT	3.11	0.69	3.16	1.51	4.38
VT2BF	42.00	7.71	42.00	32.00	66.00

(c) VT2 Hombres

Tabla 1.3: Parámetros fisiológicos a distintos umbrales.

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT2V.O2.kg	43.69	5.84	45.00	33.00	54.00
VT2V.O2.HR	14.38	2.26	15.00	10.00	18.00
VT2HR	172.69	7.88	173.00	153.00	182.00
VT2WR	210.77	23.29	209.00	162.00	241.00
VT2V.E.V.O2	29.22	3.11	29.10	25.20	36.40
VT2V.E.V.CO2	28.28	2.88	28.50	24.30	35.20
VT2RER	1.03	0.05	1.03	0.95	1.11
VT2V.E	79.55	10.81	79.40	64.40	107.70
VT2VT	2.06	0.36	2.08	1.22	2.59
VT2BF	39.62	8.06	37.00	31.00	57.00

(a) $VT2$ mujeres

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT3V.O3.kg	59.88	5.82	59.00	52.00	73.00
VT3V.O3.HR	23.70	4.67	24.00	5.00	30.00
VT3HR	187.97	8.43	188.00	166.00	203.00
VT3WR	412.15	49.33	423.00	286.00	474.00
VT3V.E.V.O3	37.37	3.99	37.11	28.90	46.90
VT3V.E.V.CO3	33.52	3.70	32.70	27.30	43.60
VT3RER	1.14	0.14	1.13	1.02	1.85
VT3V.E	179.38	31.97	174.70	114.13	272.70
VT3VT	3.22	0.66	3.20	1.46	4.53
VT3BF	56.39	9.77	58.00	38.00	78.00

(b) $VT3$ hombres

	Media	Desviación típica	Mediana	Mínimo	Máximo
VT3V.O3.kg	49.46	5.67	49.00	39.00	58.00
VT3V.O3.HR	15.77	2.13	16.00	11.00	19.00
VT3HR	184.54	5.59	183.00	177.00	196.00
VT3WR	264.69	18.71	266.00	224.00	296.00
VT3V.E.V.O3	37.17	3.37	38.00	29.50	40.90
VT3V.E.V.CO3	31.72	2.39	31.80	26.40	35.80
VT3RER	1.17	0.07	1.15	1.08	1.31
VT3V.E	114.42	15.68	111.30	88.40	149.30
VT3VT	2.11	0.26	2.12	1.50	2.45
VT3BF	55.15	8.78	56.00	42.00	70.00

(c) $VT3$ mujeres

Tabla 1.4: Parámetros fisiológicos a distintos umbrales

	Shapiro	CVM	Liffolds
VT1V.O2.kg	0.11	0.12	0.19
VT1V.O2.HR	0.76	0.29	0.07
VT1HR	0.00	0.00	0.01
VT1WR	0.68	0.57	0.62
VT1V.E.V.O2	0.29	0.61	0.81
VT1V.E.V.CO2	0.56	0.59	0.73
VT1RER	0.47	0.71	0.74
VT1V.E	0.92	0.90	0.61
VT1VT	0.84	0.77	0.89
VT1BF	0.02	0.07	0.10
VT2V.O2.kg	0.15	0.04	0.04
VT2V.O2.HR	0.76	0.67	0.63
VT2HR	0.18	0.23	0.23
VT2WR	0.21	0.31	0.35
VT2V.E.V.O2	0.00	0.00	0.00
VT2V.E.V.CO2	0.32	0.16	0.12
VT2RER	0.46	0.54	0.49
VT2V.E	0.98	0.93	0.99
VT2VT	0.36	0.16	0.06
VT2BF	0.01	0.06	0.07
VT3V.O3.kg	0.05	0.14	0.18
VT3V.O3.HR	0.00	0.01	0.00
VT3HR	0.31	0.44	0.44
VT3WR	0.03	0.04	0.02
VT3V.E.V.O3	0.97	0.75	0.67
VT3V.E.V.CO3	0.34	0.40	0.49
VT3RER	0.00	0.00	0.00
VT3V.E	0.27	0.23	0.15
VT3VT	0.89	0.97	0.98
VT3BF	0.23	0.12	0.14

(a) hombres

	Shapiro	CVM	Liffolds
VT1V.O2.kg	0.89	0.51	0.36
VT1V.O2.HR	0.34	0.15	0.14
VT1HR	0.27	0.25	0.49
VT1WR	0.03	0.02	0.05
VT1V.E.V.O2	0.84	0.51	0.34
VT1V.E.V.CO2	0.64	0.81	0.87
VT1RER	0.22	0.08	0.03
VT1V.E	0.12	0.21	0.24
VT1VT	0.16	0.09	0.12
VT1BF	0.23	0.22	0.35
VT2V.O2.kg	0.78	0.37	0.53
VT2V.O2.HR	0.62	0.25	0.21
VT2HR	0.14	0.46	0.42
VT2WR	0.61	0.80	0.72
VT2V.E.V.O2	0.35	0.61	0.62
VT2V.E.V.CO2	0.29	0.54	0.44
VT2RER	0.66	0.93	0.95
VT2V.E	0.12	0.22	0.30
VT2VT	0.31	0.21	0.21
VT2BF	0.02	0.01	0.00
VT3V.O3.kg	0.94	0.91	0.75
VT3V.O3.HR	0.40	0.26	0.30
VT3HR	0.21	0.09	0.06
VT3WR	0.86	0.63	0.79
VT3V.E.V.O3	0.14	0.18	0.23
VT3V.E.V.CO3	0.88	0.74	0.72
VT3RER	0.65	0.59	0.46
VT3V.E	0.85	0.74	0.83
VT3VT	0.43	0.73	0.87
VT3BF	0.86	0.94	0.97

(b) mujeres

	t.test	KS	Wilcox-test
VT1V.O2.kg	0.00	0.00	0.00
VT1V.O2.HR	0.00	0.00	0.00
VT1HR	0.89	0.94	0.86
VT1WR	0.00	0.00	0.00
VT1V.E.V.O2	0.27	0.18	0.25
VT1V.E.V.CO2	0.26	0.66	0.17
VT1RER	0.83	0.60	0.86
VT1V.E	0.00	0.00	0.00
VT1VT	0.00	0.00	0.00
VT1BF	0.45	0.89	0.43
VT2V.O2.kg	0.00	0.00	0.00
VT2V.O2.HR	0.00	0.00	0.00
VT2HR	0.11	0.17	0.06
VT2WR	0.00	0.00	0.00
VT2V.E.V.O2	0.55	0.94	0.94
VT2V.E.V.CO2	0.27	0.31	0.20
VT2RER	0.11	0.46	0.12
VT2V.E	0.00	0.00	0.00
VT2VT	0.00	0.00	0.00
VT2BF	0.37	0.22	0.30
VT3V.O3.kg	0.00	0.00	0.00
VT3V.O3.HR	0.00	0.00	0.00
VT3HR	0.12	0.07	0.06
VT3WR	0.00	0.00	0.00
VT3V.E.V.O3	0.86	0.89	0.96
VT3V.E.V.CO3	0.06	0.26	0.17
VT3RER	0.34	0.37	0.04
VT3V.E	0.00	0.00	0.00
VT3VT	0.00	0.00	0.00
VT3BF	0.68	1.00	0.78

(c) Comparación hombres y mujeres.

Tabla 1.5: p-valores para los test de normalidad por sexos y contrastes de igualdad de distribuciones entre hombres y mujeres. CVM denota el test Crammer-Von Misses y KS al test Kolmogorov-Smirnov.

de oxígeno y una potencia inferior a los hombres y esto se ve reflejado claramente en los datos. Sin embargo en RER (explicada anteriormente), que es una variable normalizada, se eliminan estos efectos y, por tanto, no aparecen estas diferencias. Además, es conocida la relación entre el número de inspiraciones por minuto y la frecuencia cardíaca en el ejercicio físico. Esa relación se mantiene en este caso, pues en ninguna de las variables existen diferencias significativas entre los grupos, algo también conocido.

$VT2$

En este caso parece que existen ciertas diferencias entre hombres y mujeres en la frecuencia cardíaca (no estadísticamente significativas), la explicación para este fenómeno puede venir dada por el comportamiento no lineal y caótico de la frecuencia cardíaca cuando se aleja de la zona completamente aeróbica. Salvo en las medidas normalizadas y la frecuencia de respiración, en el resto de variables hay diferencias significativas entre ambos grupos.

$VT3$

Al igual que en los dos umbrales anteriores, vemos grandes diferencias en la potencia desarrollada, de hecho, estas diferencias se hacen cada vez mayores. Las diferencias en la frecuencia cardíaca se vuelven a estabilizar. En definitiva, al igual que ocurría para los dos umbrales anteriores, los resultados obtenidos concuerdan con la teoría fisiológica y, por tanto, no ha aparecido ningún suceso extraño. Es importante observar que la variabilidad en todas las tablas entre hombres y mujeres es mayor en el caso los varones, como vemos en los mayores valores de la desviación típica.

Los estimadores no paramétricos de la función de densidad muestran que la variable VE/VO_2 en los diferentes umbrales tiene un comportamiento similar por sexos, a veces casi idéntico. Esto ocurre porque se trata de un índice normalizado, esta variable suele usarse por ejemplo para identificar el umbral anaeróbico a través de las curvas obtenidas con una prueba de esfuerzo.

Capítulo 2

Modelos de regresión no paramétricos

2.1. Introducción al problema

En este capítulo se pretende construir un modelo de regresión de interés con la base de datos de Inzell descrita en el capítulo 1. Para ello, disponemos de 24 pruebas de esfuerzo en las que contamos con las medidas de multitud de variables respiratorias cada 3 segundos de tiempo.

Con el modelo de regresión se persigue determinar el consumo de oxígeno (una variable muy costosa de medir en el exterior y casi imposible en competición) a partir de la frecuencia cardíaca y la potencia (variables sobre las que podemos tener un control en el exterior o incluso en competición).

El interés práctico de lo anterior es el siguiente: poder construir una herramienta estadístico-matemática para predecir el consumo de oxígeno del deportista en tiempo real por mecanismos indirectos (algo casi inexistente entre los gadgets deportivos actuales, que cuando se realiza, se hace únicamente de manera imprecisa a través de modelos lineales, cuando es más que conocido el comportamiento no lineal entre estas variables), lo que permitiría cuantificar el esfuerzo del deportista de una manera fiable y además estimar su gasto energético, algo muy útil para las aplicaciones de bajada de peso por ejemplo.

2.2. Introducción a las técnicas de regresión no paramétricas

Las técnicas de estadística no paramétrica facilitan modelizar relaciones complejas entre los datos. Todos estas técnicas cumplen una propiedad común, no asumen ningún modelo particular para los datos y permiten que estos manifiesten sus características naturales por sí solos.

En este capítulo estudiaremos la predicción no paramétrica de una variable aleatoria respuesta, Y, condicionada a una variable aleatoria predictora, X. Se supondrá que se observan n pares de datos $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ que provienen del modelo de regresión no paramétrico:

$$y_i = m(x_i) + \epsilon_i,\ i = 1, 2, \ldots, n \tag{2.1}$$

donde $\epsilon_1, \ldots, \epsilon_n$ son variables aleatorias independientes con

$$E(\epsilon_i) = 0,\ \ Var(\epsilon_i) = \sigma^2,\ \ i = 1, 2, \ldots, n \tag{2.2}$$

y los valores de la variable explicativa x_1,..., x_n son conocidos, por lo que estamos bajo el contexto de un modelo de regresión no paramétrico con diseño fijo. Además, dado que hemos supuesto que la varianza del error es constante, el modelo será también homocedástico.

Una vez establecido el modelo, el paso siguiente será estimarlo a partir de las n observaciones disponibles. Es decir, se trata de construir un estimador, $\hat{m}(x)$, de la función de regresión y un estimador, $\hat{\sigma}^2$, de la varianza del error. A dicho procedimiento también se le conoce como aplicar métodos de suavización.

El conjunto de técnicas de suavización disponibles para estimar no paramétricamente la función de regresión es amplísimo e incluye, entre otras, las siguientes:

- Ajuste local de modelos paramétricos: Se basa en hacer varios ajustes paramétricos teniendo en cuenta unicamente los datos cercanos al punto donde se desea estimar la función.

- Métodos basados en series ortogonales de funciones: Se elige una base ortogonal en un espacio vectorial de funciones y se estiman los coeficientes del desarrollo a partir de los n pares de datos de las observaciones. Los ajustes por series de Fourier y mediante wavelets son los dos enfoques más utilizados.

- Suavización mediante splines: Se trata de buscar las funciones $\hat{m}(x)$ dentro de un espacio de funciones que minimicen la suma de los cuadrados de los residuos ($\hat{\epsilon}_i = y_i - \hat{m}(x_i)$) más un término que penalice la falta de suavidad de las funciones $\hat{m}(x)$.

- Técnicas de aprendizaje supervisado: Las redes neuronales, el random forest, los k-vecinos más próximos y los árboles de regresión también son técnicas de regresión no paramétricas utilizadas para estimar la función $m(x)$.

2.3. El estimador de regresión polinómico local

El estimador polinómico local de la regresión consiste en aproximar la función de regresión localmente por un polinomio, en cada punto de la muestra x_i ponderado por una función tipo núcleo K. El problema matemático de optimización a resolver es el siguiente:

$$\hat{\beta} = \arg \min_{\beta_0,\beta_1,\ldots,\beta_q} \sum_{i=1}^{n} w_i(y_i - (\beta_0 + \beta_1(x_i - t) + \cdots + \beta_q(x_i - t)^q))^2 \tag{2.3}$$

donde $w_i = w(t, x_i) = \frac{1}{h}K(\frac{x_i - t}{h})$.

Obsérvese que los coeficientes obtenidos dependen del punto t donde se realiza la estimación $\hat{\beta}_j = \beta_j(t)$. Finalmente, el estimador de $m(t)$ es el resultado de evaluar el polinomio $\sum_{j=0}^{p} \hat{\beta}_j(x-t)^j$ en $x = t$:

$$\hat{m}(t) = \hat{\beta}_0. \tag{2.4}$$

En el caso particular de que se ajuste localmente un polinomio de grado 0 se obtiene el conocido como estimador de Nadaraya-Watson (ver [20] y [31]) o estimador núcleo de regresión. Su expresión viene dada mediante:

$$\hat{m}(t) = \frac{\sum_{i=1}^{n} K(\frac{x_i - t}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x_i - t}{h})} = \sum_{i=1}^{n} w(t, x_i) y_i. \tag{2.5}$$

Obsérvese que $\hat{m}(t)$ es una media ponderada en este caso de los valores respuesta, donde el peso de cada punto depende de la distancia entre la variable explicativa y el punto t donde se está estimando la función de regresión.

Cálculo de los coeficientes del estimador lineal local

El problema de minimización (2.3) puede escribirse en notación matricial. Definimos la matriz:

$$X_t = \begin{pmatrix} 1 & (x_1 - t) & \ldots & (x_1 - t)^q \\ \vdots & \vdots & \ddots & \\ 1 & (x_n - t) & \ldots & (x_n - t)^q \end{pmatrix}$$

y los vectores $Y = (y_1, \ldots, y_n)^{'}$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{'}$, $\beta = (\beta_1, \ldots, \beta_q)^{'}$ y la matriz de pesos $W_t = diag(w(t, x_i), \ldots, w(t, x_n))$.

Se ajusta el modelo

$$Y = X_t\beta + \epsilon \tag{2.6}$$

por mínimos cuadrados generalizados:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{q+1}} (Y - X_t\beta)' W_t (Y - X_t\beta). \tag{2.7}$$

La solución es:

$$\hat{\beta} = (X^{'} W_t X)^{-1} (X^{'} W_t Y). \tag{2.8}$$

Elección del parámetro de suavizado

El parámetro de suavizado, h, controla el equilibrio que el estimador no parámetrico de la función de regresión debe mantener entre el buen ajuste a los datos observados y la capacidad para prededecir bien observaciones futuras.

Valores pequeños de h proporcionan mucha flexibilidad al estimador y le permite aproximarse a todos los datos observados (cuando h se aproxima a cero el estimador acaba por interpolar los datos), pero los errores de predicción asociados serán altos. Por tanto, se produce una situación donde el modelo esta sobreajustado (overfitting). En el caso de que tome un tamaño moderado no se ajustará tan bien a las observaciones (tampoco es necesario, dado que los datos pueden contener ruido aleatorio) pero predecirá mejor. El otro caso extremo, es cuando h toma valores muy altos y tendremos una falta de ajuste importante (underfitting). Buscar el valor adecuado del parámetro de suavizado persigue conseguir un equilibrio razonable entre el sesgo y la varianza del estimador. Para h pequeño el estimador es muy variable y tiene poco sesgo, sin embargo para h muy grande el estimador es poco variable, pero tiene mucho sesgo.

El parámetro de suavizado puede elegirse de forma manual: comparando los resultados obtenidos para distintos valores de h y eligiendo aquel que proporcione el resultado visualmente más adecuado. El problema de aplicar esta vía es que es un procedimiento subjetivo y no se puede automatizar, lo que lo convierte en un método desaconsejable cuando el número de estimaciones a realizar es grande. Existen diversos métodos para llevar a cabo la selección de ventana de manera automática:

- Minimizar la predicción en una muestra test: Si tenemos suficientes datos para construir una muestra de entrenamiento y otra muestra para evaluar el modelo, una opción es ajustar el modelo con la muestra de entrenamiento y elegir el parámetro de suavizado de tal forma que se minimice el error con la muestra test. Es un procedimiento muy habitual en el mundo del aprendizaje automático.

- Validación cruzada: Consiste en extraer de la muestra consecutivamente cada una de las observaciones x_i, estimar el modelo con los restantes datos (denotaremos por $\hat{m}^h_{-i}(x)$ el estimador obtenido con este procedimiento), y predecir el dato ausente con ese estimador. Finalmente se compara esa predicción con el dato real. Esto se hace con cada posible valor de h, lo que permite construir la siguiente función:

$$ECM_{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}^h_{-i}(x_i))^2 \tag{2.9}$$

  que mide el error de predicción del estimador para la muestra, dejando uno fuera, para cada h. El valor de h que minimice esta función será el valor del parámetro de suavización elegido.

- Plug-in: Es un método propuesto por [26]. Una medida global del ajuste del estimador $\hat{m}$ a la verdadera función m es el error cuadrático medio integrado:

$$MISE(\hat{m}) = E_{\mathbf{Z}}(ISE(\hat{m})) = E_{\mathbf{Z}} \int_a^b (\hat{m}(x) - m(x))^2 f(x) dx \tag{2.10}$$

donde **Z** representa la muestra aleatoria a partir de la cual se construye el estimador $\hat{m}$ y a y b son los extremos del soporte de la variable explicativa. Se puede probar que el valor de h que minimiza el término dominante del $MISE$ es:

$$h_0 = \left(\frac{R(K)\sigma^2}{\mu_2^2(K) \int_a^b m''(x)^2 f(x) dx} \right)^{1/5} \tag{2.11}$$

siendo:

$$R(K) = \int K(x)^2 dx \;\; y \;\; \mu_2(K) = \int x^2 K(x) dx. \tag{2.12}$$

El método plug-in de selección de h conduce a la selección de una ventana que consiste en substituir en esta expresión las cantidades desconocidas por estimaciones de ellas. En concreto, para dar un valor de h, necesitamos:

1. Estimar $\int_a^b m''(x)^2 f(x) dx = E[m''(X)^2]$. Por ejemplo, utilizando un modelo polinómico local de grado 3 y la muestra **Z** dada.
2. Estimar $\sigma^2 = Var(Y|X = x) = Var(\epsilon)$. Por ejemplo, a través del modelo ajustado, y después considerando un estimador insesgado de los residuos obtenidos.

2.4. Modelos de regresión polinómicos locales con más de una covariable

Anteriormente se han introducido los modelos de regresión polinómica con una covariable, la extensión al caso en el que hay p variables es directa:

$$y_i = m(x_{i1}, \ldots, x_{ip}) + \epsilon_i, \tag{2.13}$$

con $E(\epsilon_i) = 0$ y $Var(\epsilon_i) = \sigma^2$, para $i = 1, \ldots, n$. Para definir en este marco los estimadores de la función de regresión mediante polinomios locales necesitamos, por una parte, definir los pesos ω_i de cada observación y, por otra, especificar qué variables explicativas se incluyen en cada modelo de regresión local. Al igual que en el caso univariante los pesos ω_i se definen en función de la distancia entre las observaciones de las covariables y el punto $t = (t_i, t_2, \ldots, t_p)$ donde se quiere evaluar la función de regresión. Los pesos pueden fijarse de varias formas. Entre ellas, cabe destacar las dos siguientes:

- A través de un núcleo producto:

$$w_i = w(t, x_i) = \prod_{j=1}^{p} K\left(\frac{x_{ij} - t_j}{h_j} \right), \tag{2.14}$$

  donde K es un núcleo univariante y h_j es un parámetro de suavizado adecuado para la j-ésima variable explicativa. El problema de ajustar los pesos del estimador lineal local por esta vía, es que no tiene en cuenta la estructura de dependencia multivariante existente en el caso multidimensional.

- A través de un núcleo multivariante:

$$w_i = w(t, x_i) = \frac{1}{|H|} K\left(H^{-1}(x_i - t) \right), \tag{2.15}$$

donde K es un núcleo multivariante y H es un matriz $p \times p$ simétrica y definida positiva.

El estimador polinómico local múltiple es la generalización natural del estimador en el caso univariante definido en la sección 2.3. Si se desea ajustar los polinomios p-multivariantes de grado q, con $q \in \mathbb{N}$, se deben incluir todos los términos posibles de la forma

$$\beta_{s_1,\ldots,s_p} \prod_{j=1}^{p} (x_{ij} - t_j)^{s_j}, \tag{2.16}$$

cuyo grado,

$$\sum_{j=1}^{p} s_j \leq q. \tag{2.17}$$

La estimación de la función de regresión será el término independiente del polinomio ajustado alrededor del punto t:

$$\hat{m}(t) = \hat{m}(t_1, t_2, \ldots, t_p) = \beta_{0,\ldots,0}. \tag{2.18}$$

Por ejemplo, si hay dos variables explicativas el polinomio de grado 2 ajustado será:

$$\beta_{00} + \beta_{10}(x_{i1} - t_1) + \beta_{01}(x_{i2} - t_2) + \beta_{20}(x_{i1} - t_1)^2 + \beta_{02}(x_{i2} - t_2)^2 + \beta_{11}(x_{i1} - t_i)(x_{i2} - t_2) \tag{2.19}$$

y la estimación de $m(t)$ en $t = (t_1, t_2)$ será $\hat{\beta}_{00}$, el término independiente del polinomio.

La situación más habitual es utilizar el estimador local lineal. En el caso de bidimensional, este estimador es obtenido resolviendo el siguiente problema de optimización:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{Y_i - \alpha - \beta^t (X_i - x)\} K_H(X_i - x), \tag{2.20}$$

donde $K_H(u) = |H|^{-1} K(H^{-1}u)$, siendo K un núcleo multivariante, y H una matriz $p \times p$ no singular y simétrica. El estimador puede ser escrito explícitamente como:

$$\hat{m}_H(x) = e_1^t (X_x^t W_x X_x)^{-1} X_x^t W_x Y = s_x^t Y, \tag{2.21}$$

donde e_1 es un vector con 1 en la primera entrada y todos ceros en las restantes, X_x es una matriz con la i-ésima fila igual $(1, (x_i - x)^t)$, y

$$W_x = diag\{K_H(x_1 - x), \ldots, K_H(x_n - x)\}. \tag{2.22}$$

Uno de los problemas principales de la regresión polinómica múltiple es el efecto de la maldición de la dimensionalidad, que consiste en que, en alta dimensión, alrededor de un punto t no existan elementos de la muestra, al estar dichos entornos vacíos. Una manera de minimizar su efecto es a través de modelos que proyecten los datos en un espacio de dimensión inferior, por ejemplo, utilizando técnicas de *proyection pursuit* o mínimos cuadrados parciales. Para más detalles consultar [11].

2.5. Estimación de la regresión mediante splines

Consideremos el modelo del inicio de la sección:

$$y_i = m(x_i) + \epsilon_i,\ i = 1, 2, \ldots, n \tag{2.23}$$

donde $\epsilon_1, \ldots, \epsilon_n$ son variables aleatorias independientes con

$$E(\epsilon_i) = 0,\ \ Var(\epsilon_i) = \sigma^2,\ \ i = 1, 2, \ldots n \tag{2.24}$$

y los valores de la variable explicativa x_1,...,x_n son conocidos.

En la sección anterior se propusieron los modelos polinómicos locales para la función de regresión, $m(x)$, y a continuación se estudiaron los métodos de estimación y sus propiedades.

Ahora el enfoque es distinto. Planteamos un problema de optimización cuya solución dará lugar a una familia de estimadores no paramétricos. Por ejemplo:

$$\min_{\hat{m}:\mathbb{R}\to\mathbb{R}} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2. \tag{2.25}$$

La solución del problema anterior, es cualquier función que interpole al conjunto de datos $\{(x_i, y_i) : i = 1, \ldots, n.\}$. El problema de este enfoque es que una función $\hat{m}(x)$ que interpole a los

datos no es, en general, una función suave. Si queremos que esto ocurra, tenemos que incluir esa condición al problema de optimización, esto es:

$$\min_{\hat{m}\in M} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2 + \phi(\hat{m}), \tag{2.26}$$

donde M es un espacio de funciones suaves y ϕ es un funcional definido en M que penaliza la falta de suavidad.

Si los datos x_i se encuentran en un intervalo compacto $[a, b] \subset \mathbb{R}$ una elección usual es tomar como M el espacio de las funciones de cuadrado integrables en $[a, b]$ y que, además, su derivada segunda sea también de cuadrado integrable en $[a, b]$,

$$M = W_2^2[a, b] = \left\{ m : [a, b] \to \mathbb{R} : \int_a^b (m(x))^2 dx < \infty, \quad \int_a^b (m''(x))^2 dx < \infty \right\} \tag{2.27}$$

y como funcional de penalización

$$\phi(m) = \lambda \int_a^b m''(x)^2 dx, \quad \lambda > 0. \tag{2.28}$$

El espacio $W_2^2[a, b]$ recibe el nombre de espacio de Sobolev de segundo orden en $[a, b]$.

De esta forma, el problema anterior se escribe como:

$$\min_{m\in W_2^2[a,b]} \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2 + \lambda \int_a^b m''(x)^2 dx, \tag{2.29}$$

cuya solución corresponde a un spline, particularmente a una función spline cúbica con nodos en los valores observados $x_1, \ldots, x_n$.

2.6. Splines

Definición 2.1 *(Función spline) La función $s : [a, b] \to \mathbb{R}$ es una función spline de grado p con nodos $t_1, \ldots, t_k$ si se verifica lo siguiente:*

1. *$a < t_1 < \cdots < t_k < b$ (se denota $t_0 = a, t_{k+1} = b$)*
2. *En cada intervalo $[t_j, t_{j+1}]$, $j = 0, 1, \ldots, k$, $s(x)$ es un polinomio de grado p.*
3. *La función $s(x)$ tiene $p - 1$ derivadas continuas en $[a, b]$.*

Proposición 2.2 *Sea $S[a = t_0 < t_1 < \cdots < t_k < b, t_{k+1} = b]$ el conjunto de splines de grado p con nodos $t_1, t_2, \ldots, t_k$ definidos en $[a, b]$. $S[a = t_0 < t_1 < \ldots < t_k < b, t_{k+1} = b]$ es un espacio vectorial de dimensión $p + k + 1$.*

Este resultado nos permite identificar cada spline con una matriz finita y obtener una propiedad bastante más fuerte: se ha transformado el problema anterior de búsqueda del mínimo de un funcional en un espacio de funciones de dimensión infinita en el de buscar las coordenadas en una base finita de los splines de orden 3.

Sea $s(x)$ el spline buscado, este se puede expresar:

$$s(x) = \sum_{j=1}^{n} \alpha_j N_j(x) = \alpha' N(x) \tag{2.30}$$

donde $\alpha = (\alpha_1, \ldots, \alpha_n)'$ y $N(x) = (N_1(x), \ldots, N_n(x))'$ es una base del spline buscado. Por tanto:

$$s''(x) = \sum_{j=1}^{n} \alpha_j N''(x) \tag{2.31}$$

y

$$\int_a^b s^{''}(x)^2 dx = \int_a^b s^{''}(x) s^{''}(x)' dx = \tag{2.32}$$

$$\alpha^t \int_a^b N^{''}(x) N^{''}(x)^{'} dx \alpha = \alpha A \alpha^{'} \tag{2.33}$$

donde A es una matriz $n \times n$, cuyo elemento (i, j) es

$$\int_a^b N_i^{''}(x) N_j^{''}(x) dx. \tag{2.34}$$

Sea $Y = (y_1, y_2, \ldots, y_n)'$ y sea N_X la matriz $n \times n$ cuyo elemento (i, j) es $N_j(x_i)$. Entonces

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 = (Y - N_X \alpha)'(Y - N_X \alpha). \tag{2.35}$$

Si incluimos ahora la penalización, el problema de optimización resultaría:

$$\underset{\alpha \in \mathbb{R}^n}{\text{mín}} (Y - N_X \alpha)'(Y - N_X \alpha) + \lambda \alpha' A \alpha, \tag{2.36}$$

cuya solución explícita (resultante de igualar el gradiente a cero) es

$$\alpha = (N_X^{'} N_X + \lambda A)^{-1} N_X' Y. \tag{2.37}$$

Por tanto, el vector de valores y_i ajustados será

$$\hat{Y} = N_X \hat{\alpha} = N_X (N_X^{'} N_X + \lambda A)^{-1} N_X' Y. \tag{2.38}$$

En definitiva, el estimador spline es lineal en Y. Por consiguiente, toda la teoría construida para el caso del estimador polinómico local en cuanto a la elección del parámetro de suavizado es válida, donde, en este caso, el paramétro de suavización es el parámetro de penalización λ.

2.7. Modelos aditivos generalizados

Los modelos aditivos generalizados, GAM, han sido propuestos por [16] como una extensión de los modelos GLM (modelos lineales generalizados).

Estos modelos son menos flexibles que los vistos hasta el momento. Sin embargo, esta pérdida de flexibilidad se ve compensada por el hecho de que el modelo es más fácilmente interpretable y se puede estimar con buenos resultados, incluso con alta dimensión. El modelo aditivo tiene la siguiente estructrura:

$$Y_i = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \epsilon_i, \tag{2.39}$$

con $E(\epsilon_i) = 0$ y $Var(\epsilon_i) = \sigma^2$, $i = 1, 2, \ldots, n$, y, además, $E(g_j(x_j)) = 0$, $j = 1, \ldots, p$. Las funciones $g_j(x)$ tendrán que ser estimadas no paramétricamente puesto que no se especifica qué forma tienen. La única hipótesis adicional que se añade al modelo es que las funciones $g_j(x_j)$ se combinan aditivamente para dar lugar a la función conjunta que relaciona la variable respuesta con las p variables explicativas.

Para estimar el modelo aditivo se utiliza el algoritmo backfitting (una adaptación del teorema del punto fijo para el caso de métodos de estimación no paramétrica):

1. - Estimar α mediante $\hat{\alpha} = (1/n) \sum_{i=1}^{n} y_i$
   - Dar como estimaciones iniciales de las funciones g_k funciones cualesquiera, $\hat{g}_k = \hat{g}_k^0$ para $k = 1, \ldots, p$.
2. Repetir para cada $k = 1, 2 \ldots, p$

- Estimar g_k mediante el ajuste no paramétrico univariante de los datos (x_i, y_i^k) donde

$$y_i^k = y_i - \hat{\alpha} - \sum_{j=1, j\neq k}^{p} \hat{g_j}(x_{ij}). \tag{2.40}$$

3. Parar cuando se alcance la convergencia.

En [16] pueden encontrarse más detalles sobre este algoritmo y, en particular, sobre su convergencia y la unicidad de la solución a la que converge. El algoritmo backfitting fue propuesto por [4].

2.8. Resultados

En esta apartado se realizarán dos ajustes de modelos de regresión, uno con modelos aditivos generalizados con splines y otro con estimación polinómica local. Se mostrará la representación de las covariables ajustadas, además de varias medidas de bondad de ajuste.

Las variables explicativas serán la potencia y la frecuencia cardíaca, mientras la variable respuesta es el consumo de oxígeno por kg de peso.

Modelos aditivos generalizados

Se muestran los resultados de aplicar un modelo aditivo generalizado con splines en tabla 2.1 y se representan las covariables ajustadas del modelo en la figura 2.1.

Error absoluto medio	2.953
R^2 ajustado	0.929
Log verosimilitud	−22,711.670
UBRE (AIC reescalado)	13.169

Tabla 2.1: Resultados de ajustar un modelo aditivo generalizado para predecir el consumo de oxígeno a partir de la potencia y la frecuencia cardíaca. Se muestran varias medidas de bondad de ajuste del modelo.

Los resultados muestran una relación no lineal entre las variables, el ajuste es bueno, un R^2 de 0.929. Al ser el intercepto alto, 42.21, a esfuerzos de baja intensidad, donde la frecuencia cardíaca y la potencia es baja, los coeficientes de las variables explicativas toman valores negativos y, de hecho, tienen una tendencia decreciente hasta cierto umbral de intensidad, aproximadamente a una frecuencia cardíaca de 90 latidos por minuto y una potencia de 75 W. A partir de ese punto, como vemos en la figura 2.1, la tendencia de los coeficientes es creciente pero de forma no lineal. La potencia vuelve a decrecer a partir de los 450 W.

Los cálculos se han realizado con el software estadístico R, utilizando el paquete $mgcv$.

Modelos de regresión polinómicos locales

El estimador polinómico local ha sido calculado utilizando el paquete $locfit$ de R. Se ha empleado un método de validación cruzada para la selección de la ventana de suavizado, h, y el núcleo selecionado es el de Epanechnikov.

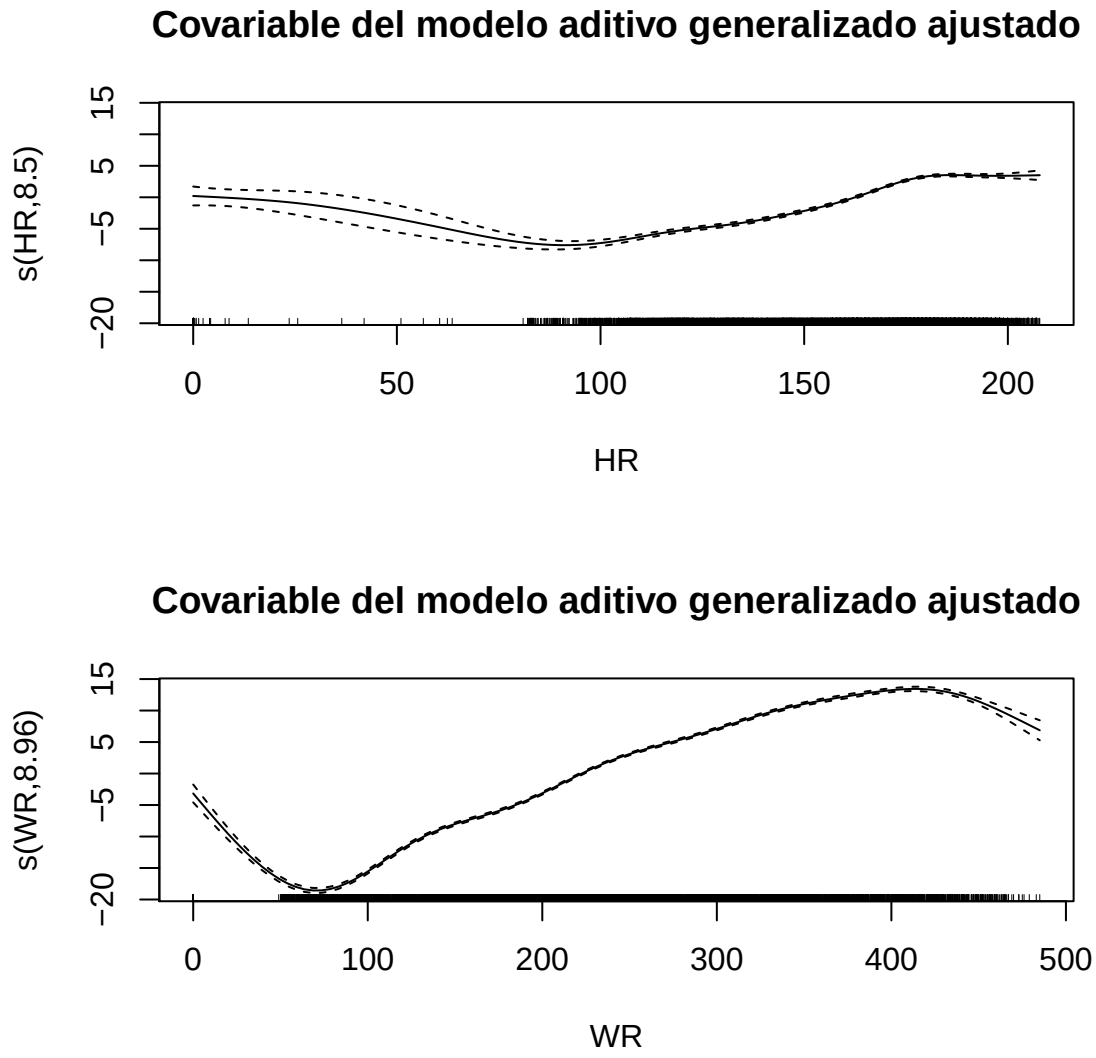


Figura 2.1: Resultados de ajustar un modelo aditivo generalizado para predecir el consumo de oxígeno a partir de la potencia (WR) y la frecuencia cardíaca (HR). Se muestra el ajuste de cada una de las covariables. En el eje vertical de cada gráfica se indican los grados de libertad del spline ajustado.

El error absoluto medio es de 2.929, superior al obtenido con el modelo aditivo generalizado ajustado con splines. En las figuras 2.2 y 2.3 podemos ver que las relaciones son de nuevo no lineales y como es el comportamiento de cada una de las variables explicativas en relación a la variable respuesta. Se puede ver además, un claro efecto frontera en el comportamiento de las variables en los valores extremos.

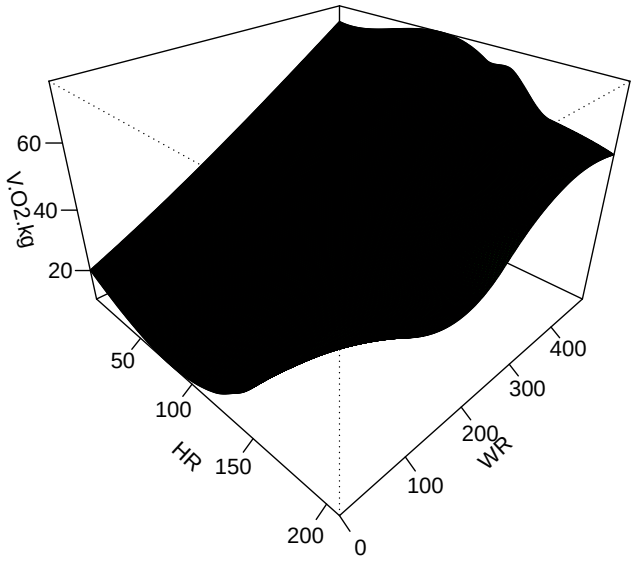


Figura 2.2: Representación gráfica del ajuste del modelo polinómico local. Se muestra la relación del consumo de oxígeno con la frecuencia cardíaca y potencia.

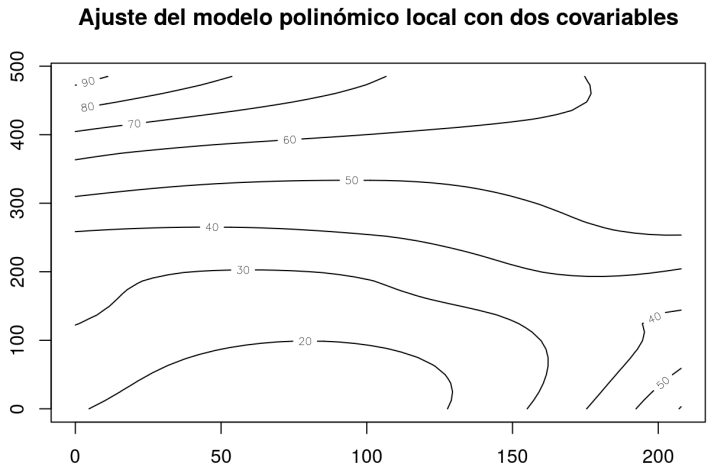


Figura 2.3: Representación gráfica del ajuste del modelo polinómico local. Se muestra la relación del consumo de oxígeno con la frecuencia cardíaca y la potencia en un gráfico de contorno. En el eje x está representada la frecuencia cardíaca y en el eje y la potencia.

Capítulo 3

Análisis de datos funcionales

3.1. Introducción

El análisis de datos funcionales (FDA) es una área relativamente nueva dentro del mundo de la estadística que consiste en analizar realizaciones muestrales de curvas, recogidas estas sobre cierto intervalo de la recta real. Debido a los incesantes cambios tecnológicos en la medición de los aparatos y al incremento de la potencia computacional, los datos de alta frecuencia están siendo un importante tema en la investigación estadística actual. Por alta frecuencia entendemos medidas u observaciones observadas sobre una rejilla de puntos en un intervalo. Una manera de modelizar este tipo de datos es asumir que son realizaciones de un proceso estocástico en tiempo continuo $\{X(t) : t \in T\}$ que toma valores en un espacio de dimensión infinita, por tanto, el conjunto de observaciones en la rejilla está asociado a un espacio de funciones concreto.

Aunque las técnicas de datos funcionales y del análisis multivariante tienen muchas similitudes, la naturaleza infinito dimensional de los datos funcionales presenta nuevos retos ausentes en la estadística multivariante clásica. De hecho, si aplicamos técnicas multivariantes a datos funcionales, el análisis no será el adecuado. Una de las ventajas del análisis de datos funcionales es que el tiempo de las observaciones no tiene porque ser igualmente espaciado y puede variar el rango del dominio de un sujeto a otro. Además, los datos funcionales no asumen que las observaciones de cada individuo sean independientes entre dos instantes de tiempo, por ello a veces también son conocidos como datos longitudinales. Lo que sí se asume normalmente es la independencia entre los datos funcionales.

En todo el capítulo, consideraremos que disponemos de curvas en el espacio funcional L^2 sobre cierto intervalo de la recta real $[a, b]$ esto es, $\{X(t) : t \in [a, b]\}$ como variable aleatoria sobre $L^2[a, b]$ de cierta σ-álgebra de Borel. Además supondremos que el proceso estocástico $\{X(t) : t \in [a, b]\}$ está definido en un espacio de probabilidad común (Ω, A, P). Para cada $t \in [a, b]$, $X(t)$ es una función medible de Ω en $\mathbb{R}$. Para un $\omega \in \Omega$, $X(\omega, .)$ es una trayectoria de la variable funcional, es decir, una función de $L^2[a, b]$. Las funciones en el espacio $L^2[a, b]$ se pueden expresar en términos de una base de funciones (al ser $L^2[a, b]$ un espacio de Hilbert con producto interior) y, además, dado que es un espacio separable, cada función se puede expresar a través de una colección numerable de elementos de una base. Sea $\{\phi_k\}_{k\in\mathbb{N}}$ una base de funciones de L^2, entonces para un ω fijo de Ω, existe una sucesión de números $c_1(\omega)$, $c_2(\omega)$,..., tales que:

$$X(t, \omega) = \sum_{i=1}^{\infty} c_k(\omega)\phi_k(t).$$

En la expresión anterior, el proceso estocástico $X(t, \omega)$ es descompuesto en dos partes, los $c_k(\omega)$ y los ϕ_k, la primera es la componente estocástica y la segunda la determinista.

El análisis de datos funcionales incluye las siguientes fases. Para más información consultar, por ejemplo, [24] o [29].

- En primer lugar, los datos se limpian y organizan. Los datos funcionales son observados únicamente es un conjunto discreto de puntos $\{t_j\}_{j\in\{1,2,\ldots,k\}}$, pudiendo ser este equiespaciado

o no. Quizás en cada cada dato funcional X_i (asumimos que disponemos de N observaciones funcionales) las observaciones se encuentren en diferentes localizaciones de la rejilla, por tanto, el dominio total de observaciones sería $\{t_{ij}\}_{i\in\{1,2,\ldots,N\},j\in\{1,2,\ldots,k_i\}}$.

- La siguiente tarea es la de convertir las observaciones funcionales en una función (normalmente cada dato i-ésimo, se denota mediante $X_i(t)$). Para realizar esto, se elige una base de funciones o bien se aplican técnicas de suavización no paramétrica. Una base es una colección de funciones de tal forma que aproximan las observaciones funcionales tan bien como se quiera:

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t), \quad i = 1, \ldots, N \tag{3.1}$$

  Cuando la base de funciones $\phi_k(t)$ es especificada, la transformación de las observaciones funcionales a datos funcionales involucra el cálculo de los coeficientes c_{ik} de la expresión (3.1). El criterio que se establece para seleccionar los c_{ik} es normalmente el de minimizar la suma residual de cuadrados sobre las observaciones funcionales y a veces, además, se le añade un término de penalización para tener en cuenta la suavidad de la aproximación. Las bases de funciones más usadas en la práctica son las siguientes:

  - Bases de Fourier. Usadas normalmente con datos periódicos que no presentan ninguna característica local específica.
  - B-Spline o bases de wavelets. Usados típicamente con datos no periódicos.
  - Bases de componentes principales. Son las bases resultantes de aplicar un análisis de componentes principales funcional ($FPCA$) que se describirá más adelante.
  - Bases exponenciales. Un conjunto de funciones exponenciales, $e^{\lambda_k t}$ con diferente parámetro λ_k.
  - Bases de polinomios. Consiste en considerar como elementos de la base el conjunto de potencias de t: 1, t, t^2,..., t^n.
  - Bases potenciales. Consiste en una secuencia de potencias no enteras, incluyendo potencias negativas de un argumento t.

- Si las curvas no están alineadas deberíamos aplicar una transformada DTW[1] para corregir este problema. Además, si el dominio de cada dato funcional no es el mismo existen técnicas para transformar los datos funcionales en un mismo intervalo común.

- El siguiente paso es realizar un análisis descriptivo de los datos funcionales. En él se contempla calcular los estadísticos principales tales como media, varianza, medidas de profundidad, así como realizar una representación gráfica de los datos funcionales.

- A continuación se procede a aplicar técnicas de reducción de la dimensión como pueden ser los componentes principales funcionales ($FPCA$), u otras técnicas del análisis exploratorio como el análisis canónico de correspondencias o el análisis de diferencias principal (PDA).

- En la última parte del análisis se aplican distintos modelos de predicción, clasificadores, series de tiempo, modelos de regresión con respuesta funcional, o no, o sistemas dados por ecuaciones diferenciales con coeficientes cambiantes a lo largo del tiempo. Para concluir se validan los modelos a través de contrastes de hipótesis y análisis gráfico.

El término análisis de datos funcionales fue acuñado por [23], no obstante existen trabajos previos de [13] y [28]. Desde entonces, su popularidad ha ido en creciente aumento, siendo hoy en día una herramienta imprescindible en el análisis masivo de datos (el paradigma conocido como big data), apareciendo de forma casi obligatoria en problemas reales de ámbitos tan diversos coma la economía, en la predicción de cotizaciones bursátiles, la meteorología, predicción de sequías, o la

[1] El dynamic time warping (DTW) es un algoritmo basado en técnicas de optimización dinámica con la que se puede alinear una secuencia de curvas.

medicina en el contexto del procesamiento de señales como los electrocardiogramas o el tratamiento médico de las imágenes médicas con la resonancia magnética nuclear.

En el ámbito del deporte las aplicaciones de FDA son escasas, reduciéndose a dos, un análisis descriptivo de las curvas de ácido láctico y un análisis de la coordinación de la rodilla en saltos verticales. Para más información consultar [15], [21].

En este capítulo se pretende realizar una nueva aplicación práctica, inédita del FDA, pero esta vez con los datos asociados a una prueba de esfuerzo. Para ello se cuenta con la base de datos de Pontevedra descrita en el capítulo 1. El objetivo final que se quiere alcanzar con estos análisis es el de conseguir evitar al deportista que realice toda la prueba de esfuerzo completa y, por tanto, que se fatigue, pero que igualmente seamos capaces de caracterizar su rendimiento deportivo, medido en este caso como consumo máximo de oxígeno (VO_2 máx) y con la variable aleatoria T : "tiempo de duración de la prueba de esfuerzo". Para alcanzar dichos propósitos se realizará en primer lugar un suavizado a las observaciones con métodos no paramétricos, para luego aplicar técnicas de reducción de la dimensión y finalmente aplicar técnicas de regresión funcional y de supervivencia funcional.

La estructura del capítulo es la siguiente, se comenzará con una breve introducción a las técnicas de suavizado, después se introducirá las técnicas de componentes principales, y, a continuación, se explicaran las técnicas de supervivencia y regresión utilizadas específicamente en el trabajo. Por último, se aplicará dicho conjunto de técnicas sobre la base de datos comentada anteriormente.

3.2. Representación de datos funcionales

3.2.1. Bases para datos funcionales

Una base de un espacio funcional es un conjunto de funciones linealmente independientes, $\{\phi_k\}_{k\in\mathbb{N}}$, tales que cualquier función del espacio funcional dado puede ser aproximada por combinaciones lineales de los elementos de la base:

$$X(t) = \sum_{k=1}^{\infty} c_k \phi_k(t) \tag{3.2}$$

Si los elementos de la base son diferenciables y también los datos funcionales, pongamos hasta un cierto orden q, se puede aproximar a partir de los elementos de la base la derivada de la observación funcional:

$$X^{(k)}(t) = \sum_{k=1}^{\infty} c_k \phi_k^{(k)}(t) \ \ , k = 0, 1, \ldots, q. \tag{3.3}$$

En la práctica, en las aproximaciones de las expresiones (3.2) y (3.3) se suele realizar una aproximación con un número finito K de funciones:

$$X(t) \approx \sum_{k=1}^{K} c_k \phi_k(t), \tag{3.4}$$

$$X^{(k)}(t) \approx \sum_{k=1}^{K} c_k \phi_k^{(k)}(t) \ \ , k = 0, 1, \ldots, q. \tag{3.5}$$

Precisamente, uno de los grandes problemas es el de seleccionar el número de elementos K, intentando evitar los conocidos problemas de overfitting y underfitting, con los clásicos inconvenientes computacionales asociados. Los otros problemas inherentes al proceso de selección de la base son los siguientes:

- ¿Qué tipo de base de funciones es la más adecuada?
- ¿Cómo se determinan los coeficientes asociados a la base?

Para la primera pregunta la respuesta se obtiene teniendo en cuenta las características de cada una de las bases y de la forma funcional visible en las observaciones funcionales. Mientras que para la segunda, existen múltiples criterios, uno puede ser a través de mínimos cuadrados ordinarios, por ejemplo. A continuación se explicarán con detalle las características de las bases de funciones más habituales.

Bases Spline

Una opción es expresar cada observación funcional como un spline. Un spline (de orden p) es un conjunto de polinomios definidos en un conjunto de subintervalos construidos de tal modo que en el extremo superior de cada subintervalo el valor del polinomio coincida con el valor del polinomio en el extremo inferior del siguiente subintervalo (hasta la derivada de orden $p-1$).

La definición de un spline ha sido introducida en el capítulo anterior, además se ha enunciado una propiedad que permite identificar el conjunto de los splines con un espacio vectorial de dimensión finita.

Por tanto, cada spline $S(t)$ se escribe:

$$S(t) = \sum_{k=1}^{p+k+1} c_k B_k(t), \tag{3.6}$$

siendo $B_k(t)$ un elemento de la base de splines, que existe al ser el conjunto de splines un espacio vectorial de dimensión finita.

Las bases de spline se utilizan normalmente cuando las observaciones funcionales no son periodicas y presenten cierta estructura de regularidad.

Bases de Fourier

Una base de Fourier está compuesta por un conjunto de funciones periódicas de periodo $\frac{2\pi}{\omega}$. Cuando las observaciones funcionales se encuentran igualmente espaciadas en un intervalo $[0,T]$, la base está compuesta por el siguiente conjunto de funciones ortogonales:

$$\phi_0(t) = \frac{1}{\sqrt{T}},$$

$$\phi_{2r-1}(t) = \frac{\sin(r\omega t)}{\sqrt{\frac{T}{2}}},$$

$$\phi_{2r}(t) = \frac{\cos(r\omega t)}{\sqrt{\frac{T}{2}}}.$$

Por tanto, un dato funcional $X(t)$ en el intervalo $[0,T]$ se expresa:

$$X(t) = c_0 + \sum_{r=1}^{\infty} c_{2r-1}\phi_{2r-1}(t) + \sum_{r=1}^{\infty} c_{2r}\phi_{2r}(t) \tag{3.7}$$

donde $c_r =< X, \phi_r > \forall r \in \mathbb{N}$.

Dicha base, es utilizada frecuentemente cuando las observaciones funcionales presentan una estructura periódica regular.

Bases de wavelets

La base de wavelet se construye a partir de dos funciones con soporte compacto en $[a,b]$, el wavelet padre ϕ y el wavelet madre ψ que verifican respectivamente $\int \phi(t)dt = 1$ y $\int \psi(t)dt = 0$. Los elementos de la base se obtienen a partir de la translación y del cambio de escala de las funciones padre y madre:

$$\phi_{j,r}(t) = 2^{-\frac{j}{2}}\phi(2^{-j}t - k),$$

$$\psi_{j,r}(t) = 2^{-\frac{j}{2}}\psi(2^{-j}t - k).$$

Además se verifican las siguientes propiedades de ortogonalidad:

$$< \phi_{j,k}(t), \phi_{j,k'}(t) >= \int \phi_{j,k}(t), \phi_{j,k'}(t)dt = \delta_{k,k'},$$

$$< \phi_{j,k}(t), \psi_{j,k'}(t) >= \int \phi_{j,k}(t), \psi_{j,k'}(t)dt = 0,$$

$$< \psi_{j,k}(t), \psi_{j',k'}(t) >= \int \psi_{j,k}(t), \psi_{j',k'}(t)dt = \delta_{j,j'}\delta_{k,k'},$$

con el producto interior de $L^2[a, b]$.

Dado un dato funcional $X(t)$ este se expresa en una base de wavelets mediante:

$$X(t) \approx \sum_k S_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \cdots + \sum_k d_{1,k}\psi_{1,k}(t), \tag{3.8}$$

donde $S_{J,k} =< X, \phi_{J,k} >$ y $d_{s,k} =< X, \psi_{s,k} >$, $\forall k$ y $s \in \{1, \ldots, S\}$.

A los términos de la expresión anterior, que depende de los wavelets padre, se los conoce como señal suave, mientras que a los términos del wavelet madre se los denota por funciones detalle. A esta descomposición en algunos libros se le conoce como descomposición multiresolución, por ejemplo, en el ámbito de procesamiento de señales.

3.2.2. Técnicas de suavizado no paramétricas de datos funcionales

En la práctica, es muy frecuente que las observaciones funcionales obtenidas se encuentren convolucionadas por el error de medición, es decir, nos encontramos en el siguiente escenario:

$$y(t_j) = x(t_j) + \epsilon(t_j), \quad j = 1, \ldots, k$$

donde $y(t_j)$ es lo observado, $x(t_j)$ es la señal real y $\epsilon(t_j)$ denota el error, como variable aleatoria de media cero. Para recuperar la señal original, se suele aplicar un suavizador lineal, esto es:

$$\hat{x}(t_j) = \sum_{i=1}^{k} s_j(t_i)y(t_i) \implies \hat{X} = SY.$$

En el caso de la suavización tipo núcleo, se toma como matriz $S \in M_{N\times N}(\mathbb{R})$, la dada por la siguiente expresión:

$$(S_{ji}) = s_i(t_j) = \frac{K(\frac{t_i - t_j}{h})}{\sum_{s=1}^{k} K(\frac{t_s - t_j}{h})},$$

donde K es la función núcleo y h el parámetro de suavizado.

El parámetro h de suavizado se suele elegir de tal forma que se minimize el error cuadrático medio:

$$MSE[\hat{x}(t)] = E[(\hat{x}(t) - x(t))^2].$$

Uno de los grandes problemas, todavía sin resolver en datos funcionales, es el de proporcionar mecanismos automáticos para seleccionar la ventana de suavizado h, ver [9].

3.3. Análisis de componentes principales

El análisis de componentes principales funcionales es la principal herramienta para analizar la variabilidad de los datos funcionales. Al igual que en el contexto multivariante, el análisis de componentes principales es usado para obtener un conjunto ortogonal de elementos del espacio (vectores en multivariante y funciones en análisis de datos funcionales) que describa la variabilidad de los datos de la forma más eficiente posible. La gran complejidad de los datos funcionales, al trabajar en espacios de dimensión infinita, obliga a los componentes principales a ocupar un papel muy importante en muchas tareas estadísticas, cosa que no ocurría en dimensión finita. En

el caso funcional, no tenemos la posibilidad de caracterizar el mecanismo generador del proceso estocástico (función de distribución, función de densidad o función característica) lo que representa un obstáculo importante, y por tanto los componentes principales tienen que jugar ese papel, [7]. Los componentes principales funcionales aparecen explicitamente en la descomposición de Karhunen-Loeve.

Teorema 3.1 *Sea* $\{X(t) : t \in [a,b]\}$ *un proceso estocástico en* $L^2[a,b]$ *tal que* $E(X(t)) = 0 \ \forall t \in [a,b]$*, definido sobre un espacio de probabilidad* (Ω, A, P)*, con función de covarianzas* $K_X(s,t)$*. Consideremos la siguiente aplicación:*

$$T_{K_X} : L^2([a,b]) \to L^2([a,b]) : f \to T_{K_X} f = \int_a^b K_X(s,.) f(s) ds. \tag{3.9}$$

Entonces, existe una base ortonormal $\{e_k\}_{k\in\mathbb{N}}$ *en* $L^2[a,b]$ *formada por los autofunciones de la aplicación* T_{K_X} *con sus respectivos autovalores* λ_k*, tal que* $X(t)$ *admite la siguiente representación:*

$$X(t) = \sum_{k=1}^{\infty} Z_k e_t, \tag{3.10}$$

donde la convergencia de la serie anterior es uniforme en L^2 *para cada* t *y*

$$Z_k = \int_a^b X(t) e_k(t) dt =< X, e_k > . \tag{3.11}$$

Además, las variables aleatorias Z_k *tienen media cero, son incorreladas y tienen varianza* λ_k*, esto es:*

$$E(Z_k) = 0, \forall k \in \mathbb{N} \ \ y \ \ E(Z_i Z_j) = \delta_{ij}\lambda_j, \ \ \forall i,j \in \mathbb{N} \tag{3.12}$$

con

$$\delta_{ij} = \begin{cases} 1 & si\ i = j \\ 0 & si\ i \neq j \end{cases}.$$

El teorema de Karhunen-Loeve proporciona el algoritmo para calcular los componentes principales: basta con realizar la descomposición espectral del operador de covarianzas $K_X(s,t)$ del proceso estocástico $\{X(t) : t \in [a,b]\}$. Las autofunciones son los componentes principales y los autovalores la varianza de cada una de los componentes.

En la práctica, se tiene un conjunto de observaciones funcionales $X_1(t)$, $X_2(t)$,..., $X_n(t)$ y, a partir de ellas, se estima el operador de covarianzas, $\hat{K}_X(s,t)$, esto es:

$$\hat{K}_X(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i(s) - \overline{X(s)})(X_i(t) - \overline{X(t)}),$$

donde $\overline{X(t)}$ y $\overline{X(s)}$ denotan la media en los puntos s y t respectivamente de las observaciones funcionales. Supongamos que el dominio de discretización es igual para todos los datos funcionales: $\{t_j\}_{j\in\{1,2,...,k\}}$. El operador covarianza para las observaciones funcionales se puede expresar por una matriz $\Sigma \in M_{k\times k}(\mathbb{R})$ que además es simétrica y definida positiva.

Los componentes principales aproximados para las observaciones funcionales resultan de realizar la descomposición espectral sobre la matriz Σ, obteniéndose una base de vectores sobre $\mathbb{R}^k$, $\{v_i\}_{i\in\{1,2,...,k\}}$ y un conjunto de autovalores $\{\lambda_i\}_{i\in\{1,2,...,k\}}$. El último paso es expresar la descomposición obtenida como un funcional, como hemos visto anteriormente. Existen diversas alternativas al método aquí propuesto, ellas se basan en estimar directamente la matriz de covarianzas desde su representación en funciones y no desde las observaciones funcionales.

El análisis de componentes principales es mucho más rico que en el caso de dimensión finita, algunas de sus características principales son las siguientes:

- Es el mejor método para resumir la información sobre la variabilidad de los datos funcionales.
- Permite obtener una base ortogonal empírica que es la que más rápido converge y, además, con unos pocos términos es capaz de presentar una representación adecuada del dato funcional.

- Puede servir para detectar datos atípicos (aunque también esconderlos).

- La rotación de las componentes principales puede ayudar a encontrar mejores explicaciones de los componentes principales y es útil si queremos aplicar técnicas de multidimensional scaling sobre datos funcionales o técnicas de clustering funcionales.

3.4. Modelos de regresión funcionales

3.4.1. El modelo de regresión lineal

Supongamos que $\mathbf{X}$ es una variable funcional en $L^2[a,b]$ e Y es una variable aleatoria real. Supondremos además que ambas variables están centradas, i.e., $E(X(t)) = 0\ \forall t \in [a,b]$ y $E(Y) = 0$. El modelo de regresión funcional lineal (FLM) establece una relación entre ambas variables de la siguiente forma:

$$Y =< \mathbf{X}, \beta > +\epsilon = \int_a^b \mathbf{X}(t)\beta(t)dt + \epsilon,$$

donde la función $\beta(t) \in L^2[a,b]$ y ϵ es una variable aleatoria con media cero y varianza σ^2 que además verifica que $E(X(t)\epsilon) = 0$, $\forall t \in [a,b]$. La predicción de Y es realizada a partir de la esperanza condicional de Y sobre $\mathbf{X}$:

$$m(\mathbf{X}) = E(Y|\mathbf{X}) =< \mathbf{X}, \beta > .$$

El que la relación entre los elementos del par $(\mathbf{X}, Y)$ sea un modelo funcional lineal es equivalente a establecer que en el modelo de regresión funcional de Y sobre $\mathbf{X}$, la esperanza condicional, pertenece a la siguiente familia paramétrica $M = \{< \cdot, \beta >: \beta \in L^2[a,b]\}$.

Dada una muestra $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, la estimación de los parámetros puede ser realizada minimizando la suma residual de cuadrados

$$\hat{\beta} = \arg \min_{\beta \in L^2[a,b]} \sum_{i=1}^{n} (Y_i - < \mathbf{X}_i, \beta >)^2.$$

Un posible método para buscar el parámetro β es representar los datos funcionales y el parámetro funcional (β) en una base de funciones truncada $\{\psi_i\}_{i=1}^m$, y, $\{\theta_i\}_{i=1}^k$, respectivamente, y exigirle que minimize la suma residual de cuadrados:

$$\mathbf{X}_i = \sum_{j=1}^{m} c_{ij}\psi_j, \quad \beta = \sum_{j=1}^{k} b_j\theta_j, i = 1, \ldots, n.$$

Usando la notación vectorial $x = (\mathbf{X}_i)_i$, $C = (c_{i,j})_{i,j}$, $\psi = (\psi_j)_j$, $b = (b_j)_j$ y $\theta = (\theta_j)_j$. De forma matricial $x = C\psi$ y $\beta = \theta^T b$. El modelo funcional resulta:

$$Y =< \mathbf{X}, \beta > +\epsilon \approx CJb + \epsilon = Zb + \epsilon \tag{3.13}$$

donde $J = (< \psi_i, \theta_j >)_{i,j}$.

La representación en bases del modelo de regresión funcional permite expresar el problema como uno matricial, por tanto, podemos aplicar las ecuaciones normales del modelo de regresión lineal múltiple y obtener las cordenadas de β en la base escogida $\{\theta_i\}_{i=1}^k$. Estas vienen dadas por $b = (Z'Z)^{-1}Z'Y$. Normalmente se escogen las bases $\{\psi_i\}_{i=1}^m$, y, $\{\theta_i\}_{i=1}^k$ ortogonales entre sí, con lo que la matriz J resulta ser diagonal.

Para elegir el número de elementos de las bases se han propuesto diversas alternativas, desde PCV (validación cruzada), GCV (validación cruzada generalizada) a las técnicas basadas en teoría de la información como el AIC (el criterio de información de Akaike), el $AICc$ (el criterio de información de Akaike corregido) y por último BIC (el criterio de información bayesiana).

3.4.2. Modelo GSAM: modelos de regresión funcionales espectrales

El anterior modelo representa de manera adecuada las relación entre las variables $\mathbf{X}$ e Y, si estas son lineales. Sin embargo, en la práctica es de esperar que muchas veces estas sean no lineales. Una manera de solventar dicho problema es a través de la utilización de modelos aditivos generalizados. Supongamos que $\mathbf{X} = (X_1, \ldots, X_p)$ es una variable aleatoria funcional p-dimensional en $L^2[a, b]$. La relación entre las variables $\mathbf{X}$ e Y en un modelo aditivo generalizado espectral ($GSAM$) es de la forma:

$$E(Y|X) = \beta_0 + \sum_{k=1}^{p} f_k(X_k) \tag{3.14}$$

donde cada f_k es una función suave.

Por otra parte, cada variable funcional X_k (con $k = 1, \ldots, p$), utilizando Karhunen-Loeve, se puede expresar de la siguiente forma:

$$X_k(t) = \mu(t) + \sum_{j=1}^{\infty} x_{kj} v_{kj}(t), \tag{3.15}$$

donde cada $v_{kj}(t)$ es la j-ésima autofunción para la k-ésima variable funcional y el término x_{kj} representa la puntuación.

Por tanto, una aproximación de la expresión (3.14) puede ser la siguiente (para más detalles consultar [19]):

$$E(Y|X) \approx \beta_0 + \sum_{k=1}^{p} \sum_{m=1}^{K_k} f_k(x_{km}), \tag{3.16}$$

en la cual f_k denota a la k-ésima funcion suavizadora y x_{mk} la m-puntuación de los componentes principales de la k-ésima covariable funcional.

La estimación de la funciones suaves es realizada por la técnica conocida como componentes principales con esperanza condicional ($PACE$). Además, dicho método también establece de manera automática el número de autovectores elegidos para cada covariable funcional a través del criterio AIC [19].

3.5. Supervivencia con covariable funcional

3.5.1. Conceptos básicos

El análisis de supervivencia abarca un conjunto de técnicas estadísticas cuya finalidad es analizar el tiempo de ocurrencia de un suceso de interés, una vez definido un tiempo de inicio. Ejemplos de esta situación pueden ser caracterizar el tiempo de nacimiento de una persona hasta su muerte, o, en el caso de las pruebas de esfuerzo, establecer con cierta probabilidad el tiempo que transcurre desde que el deportista comienza la prueba de esfuerzo hasta que la finaliza. A la ocurrencia del evento de interés se le suele denominar fallo o muerte, de ahí el nombre de supervivencia (se cuantifica el tiempo hasta la muerte). Antes de comenzar con el contenido propiamente dicho de esta sección vamos a introducir una serie de conceptos previos, que ayudarán a clarificar el resto de la sección.

- Tiempo de vida: Es el tiempo que transcurre desde un suceso inicial hasta la ocurrencia del suceso final. Lo representamos mediante la variable aleatoria T y será una variable continua y no negativa, $T \geq 0$.

- Tiempo de observación o seguimiento: Es el tiempo que transcurre desde la fecha de entrada en el estudio hasta la fecha registrada en la última observación del individuo.

- Función de supervivencia: Se define como la probabilidad de que el tiempo de vida estudiado sea mayor que un tiempo t dado. Es decir, si T es la variable aleatoria que representa el

tiempo de vida, su función de distribución es $F(t)$ y su función de densidad $f(t)$. La función de supervivencia $S(t)$ se define mediante:

$$S(t) = P(T > t) = 1 - F(t) \tag{3.17}$$

Sus propiedades son:

- Es una función monótona decreciente y continua.
- $S(0) = 1$, $\lim_{t\to\infty} S(t) = 0$.
- $S(t) = P(X > t) = \int_t^\infty f(x)dx$, $0 \leq t \leq \infty$ donde f es la función de densidad asociada a la variable aleatoria T.

- Función de riesgo: la función de razón de fallo, de riesgo o tasa instantánea de fallos, $\lambda(t)$, se define como el cociente entre la función de densidad y la función de supervivencia:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{3.18}$$

Se interpreta como el límite, cuando δ tiende a cero, de la probabilidad de que el evento de interés ocurra en el intervalo $[t, t + \delta]$ asumiendo que no ha ocurrido en el intervalo $[0, t]$, siendo δ un incremento infinitesimal.

3.5.2. Plantamiento del problema

Dada una variable aleatoria unidimensional T: "tiempo de duración de la prueba de esfuerzo" y una prueba de esfuerzo, el objetivo es calcular la función de supervivencia condicionada a la información funcional observada en la prueba de esfuerzo en los 6 primeros minutos de la misma. En términos más formales, se trata de calcular:

$$P(T > t|X = X_u) = E(1_{\{T>t\}}|X = X_u) \tag{3.19}$$

para un t fijo y la observación funcional X_u dada.

Para ello se utiliza una muestra aleatoria simple $(X_1, T_1), (X_2, T_2), \ldots, (X_n, T_n)$, donde X_k denota la k-ésima observación funcional en el intervalo $[0, s]$ (en este caso s=6 minutos) y T_k es el tiempo de vida de la k-ésima prueba de esfuerzo, con la que estimar la función de supervivencia. Un estimador de la expresión (3.19) puede ser el siguiente:

$$\hat{P}(T > t|X = X_u) = \hat{E}(1_{\{T>t\}}|X = X_u) = \frac{\sum_{i=1}^n K(\frac{d(X_u,X_i)}{h})1_{\{T_i>t\}}}{\sum_{i=1}^n K(\frac{d(X_u,X_i)}{h})} \tag{3.20}$$

siendo $d(X_u, X_i)$ es la distancia en L^2 entre las observaciones funcionales X_u y X_i (con $i = 1, 2, \ldots, n$), 1 indica la función indicadora, K es una función tipo núcleo y h es el parámetro de suavizado. El problema se encuentra en seleccionar la ventana de suavizado, h, de tal forma que se obtenga una estimación adecuada (ver en el capítulo anterior la subsección dedicada al parámetro de suavizado) de:

$$E(1_{\{T>t\}}|X = X_u)$$

por ejemplo, se puede seleccionar $h(t)$ (para cada t fijo) con el método de validación cruzada:

$$\hat{h}(t) = \arg\min_{h\in\mathbb{R}} \sum_{i=1}^n \left(1_{\{T_i>t\}} - \frac{\sum_{k\neq i}^n K(\frac{d(X_i,X_k)}{h})1_{\{T_k>t\}}}{\sum_{k\neq i}^n K(\frac{d(X_i,X_k)}{h})}\right)^2. \tag{3.21}$$

Utilizando las propiedades del estimador de Nadaraya-Watson, se obtiene que la función de supervivencia así estimada, $\hat{P}(T > t|X = X_u)$, cumple las propiedades de una función de

supervivencia, para más detalles consultar [14]. Además en [9] se prueba la consistencia de la estimación condicional de la función de distribución introduciendo ciertas condiciones de regularidad, por tanto también va a ser consistente la función de supervivencia (bajo esas hipótesis), esto es:

$$\hat{P}(T > t|X = X_u) \xrightarrow{P} P(T > t|X = X_u), \quad n \to \infty \tag{3.22}$$

donde P denota la convergencia en probabilidad.

3.6. Resultados

3.6.1. Resultados de los modelos de regresión

En esta sección se mostraran los resultados de realizar un análisis de datos funcionales para predecir el consumo máximo de oxígeno del deportista (una variable escalar) en función de el consumo de oxígeno y la frecuencia cardíaca (dos covariables funcionales) y la frecuencia cardíaca máxima (una variable escalar). Para ello utilizaremos un modelo de regresión funcional espectral con las mediciones obtenidas en los 6 primeros minutos de la prueba de esfuerzo. Una prueba de esfuerzo suele tener una duración media de alrededor de 13 minutos. La intensidad máxima obtenida en los 6 primeros minutos respecto a la máxima velocidad que alcanzan los deportistas en la prueba (su velocidad aeróbica máxima) es de media el 61 %, lo que representa un esfuerzo insignificante. Aproximadamente se corresponde con la intensidad que los deportistas pueden soportar si estuviesen preparados para disputar una carrera a pie de 100 km.

Los gráficos asociados a las observaciones funcionales, suavizados por una metodología no paramétrica, se representan en las figuras 3.1 y 3.2. A continuación, en las figuras 3.3 y 3.4 se muestran los resultados de aplicar un análisis de componentes principales funcional a dichas covariables. Se observa que con unos pocos componentes principales somos capaces de representar gran parte de la variabilidad (más de un 96 % en ambos casos solo con 3 componentes), además se concluye que, en ambas covariables funcionales, la variabilidad se concentra por debajo de la media al inicio de la prueba y, a medida que esta avanza, los datos de mayor variabilidad se encuentran cada vez en valores más bajos de consumo de oxígeno y frecuencia cardíaca respecto a la media.

La interpretación de los componentes principales en este caso es la siguiente: la variabilidad del consumo de oxígeno y la frecuencia cardíaca al inicio de la prueba se encuentra concentrada por debajo de la media. En el caso del consumo de oxígeno es en los deportistas, de menos nivel que son los que más les cuesta mantener el esfuerzo, y se mantiene esta tendencia en el resto de la prueba observada. Sin embargo, en la frecuencia cardíaca no se puede atribuir a ningún grupo de deportistas, debido a que no hay relación entre la frecuencia cardíaca y el nivel deportivo, por lo menos a las intensidades que estamos observando en la prueba de esfuerzo. La frecuencia cardíaca en el esfuerzo depende de la frecuencia cardíaca máxima y de la frecuencia cardíaca en reposo y, en menor medida, de la forma física del deportista.

Los resultados del modelo de regresión arrojan un R^2 de 0.803 y el error medio en valor absoluto es de 2.754. Dichos resultados son satisfactorios: en comparación con otros tests, el error cometido es muy bajo, el resto de tests tienen errores superiores al aquí mostrado y, además, le obliga al deportista a realizar un esfuerzo máximo, mientras que aquí únicamente se está utilizando 6 minutos de la prueba de esfuerzo, una actividad que no le produce sufrimiento y desgaste al deportista.

El test ha sido validado con una cantidad importante de individuos (341) de diferente nivel, peso, talla y edad. Esto es muy importante, pues otros tests se validan con un conjunto de estudiantes con características similares y, por tanto, únicamente pueden ser aplicados a un conjunto restrictivo de individuos. Debido a esto y a que muchas veces se obtengan buenos ajustes, muchos autores seleccionan una muestra sin apenas variabilidad y, por tanto, cualquier aproximación va a ser buena, esto es visible en el estudio comparativo mostrado en [12].

En conclusión, los resultados demuestran que con técnicas estadísticas basadas en datos funcionales y la motorización del deportista con datos en continuo podemos caracterizar el rendimiento del deportista con métodos indirectos de manera bastante fidedigna. Esto es una auténtica novedad en el campo del entrenamiento deportivo. En futuros trabajos se estudiarán las intensidades óptimas de toma de las mediciones para conseguir los mejores resultados en la predicción del consumo

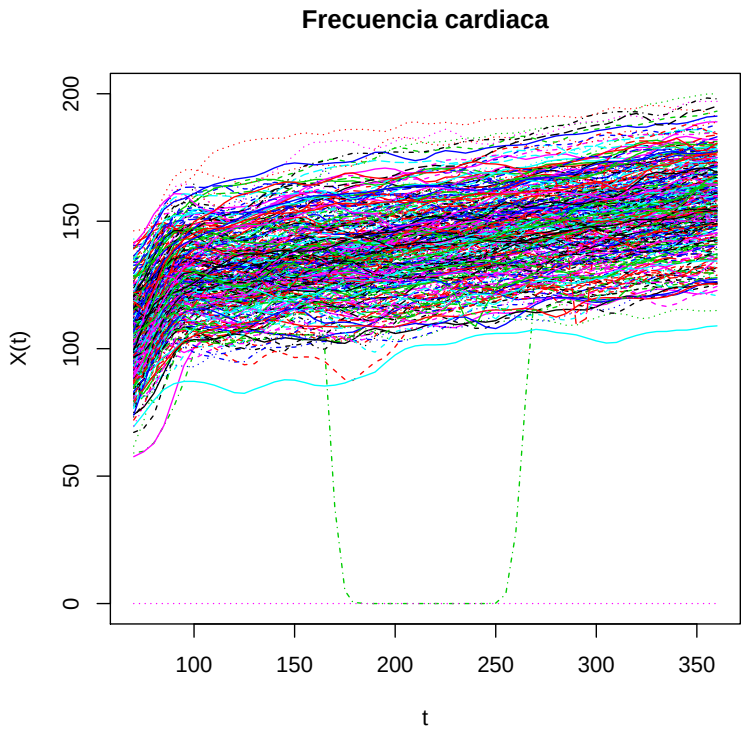


Figura 3.1: Frecuencia cardíaca suavizada en los 6 primeros minutos de la prueba de esfuerzo.

de oxígeno, buscando el beneficio entre la mejora de la predicción (a mayor intensidad y duración de la prueba mejor ajuste) y la fatiga generada en el deportista. En este trabajo hemos sido demasiado conservadores en este aspecto, pero pese a todo, somos capaces de superar con creces el resto de tests para predecir el consumo máximo de oxígeno, uno de los mejores predictores del rendimiento deportivo, lo cual nos hace ser optimistas acerca del potencial de esta línea de investigación.

3.6.2. Resultados de la supervivencia con covariable funcional

Se representan los resultados de estimar la función de supervivencia $P(T > t|X = X_u)$ en la figura 3.5 para cada uno de los 341 datos de la muestra descrita en el capítulo 1.

Únicamente se utiliza una covariable funcional, el consumo de oxígeno. Observamos una enorme variabilidad de los datos. Ello dificulta la utilización de este modelo desde el punto de vista de las aplicaciones prácticas, pues el tiempo que transcurre desde que la probabilidad es cero hasta que esta es uno es en la mayoría de casos superior a cinco minutos, lo que indica que la información proporcionada por esta única covariable no es suficiente para construir un modelo adecuado.

En un futuro, se pretende realizar la extensión multivariante de dicho procedimiento o extenderla con la regresión polinómico local funcional de tal manera que se preserven las propiedades de la función de supervivencia, como se procede en [14] en el caso de dimensión finita.

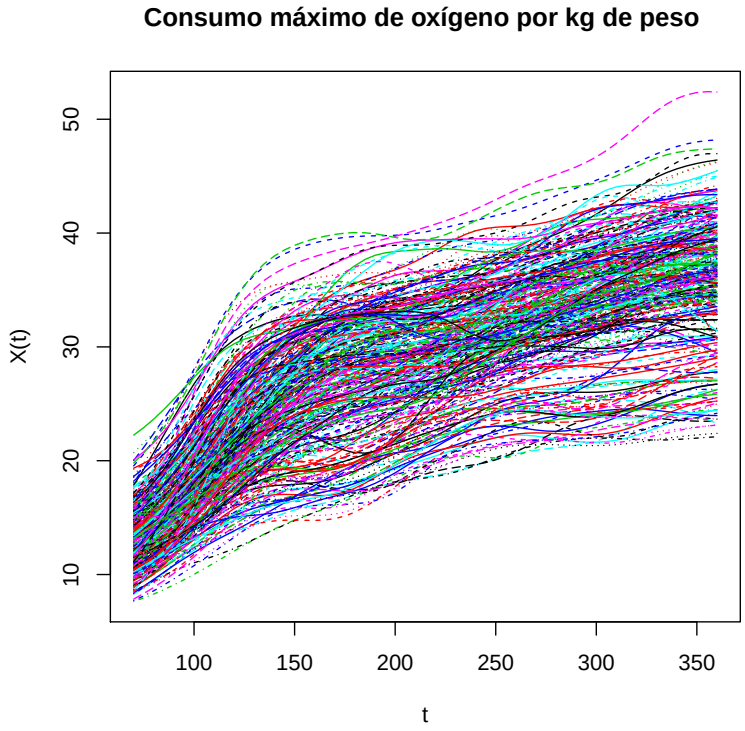


Figura 3.2: Consumo máximo de oxígeno suavizado en los 6 primeros minutos de la prueba de esfuerzo.

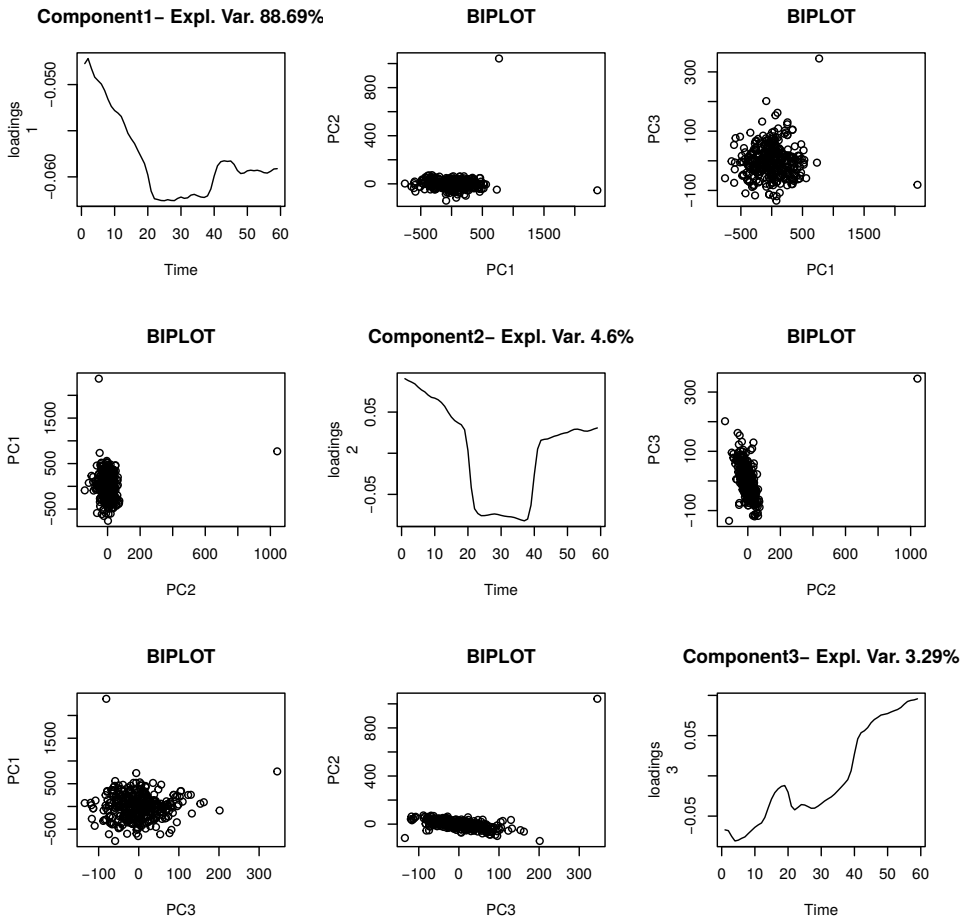


Figura 3.3: Componentes principales funcionales de la frecuencia cardíaca en los en 6 minutos de esfuerzo.

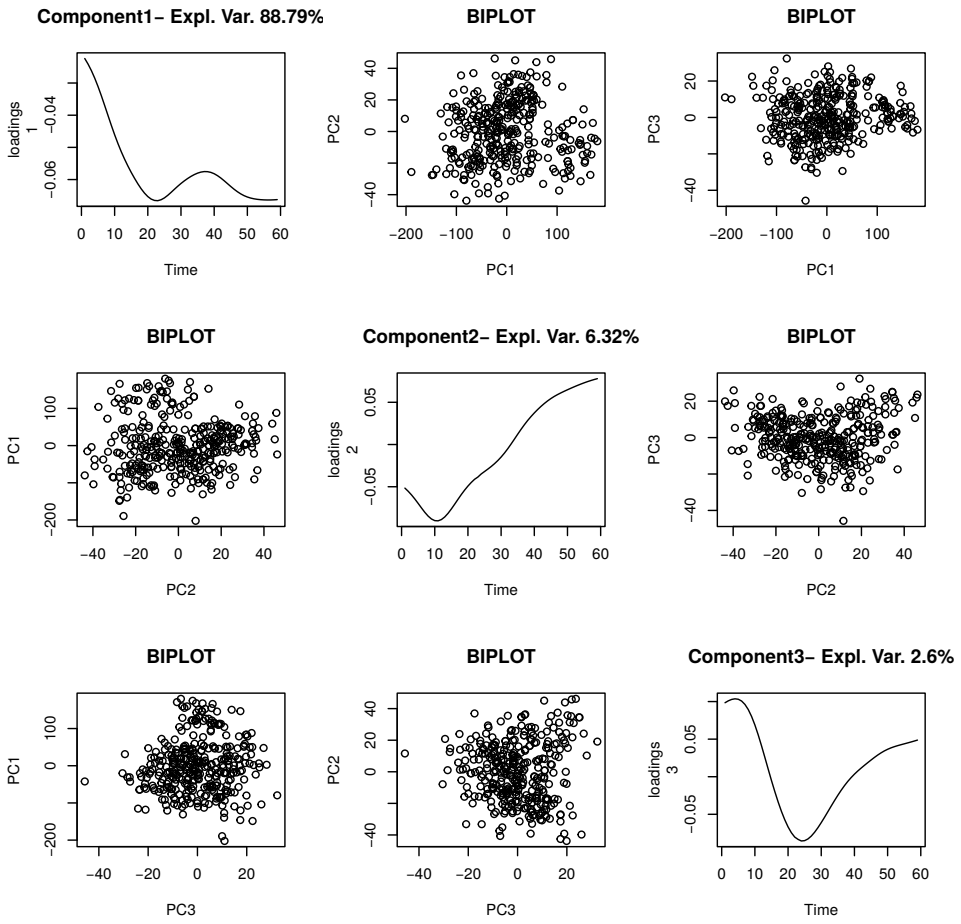


Figura 3.4: Componentes principales funcionales del consumo de oxígeno por kg de peso en los 6 minutos de esfuerzo.

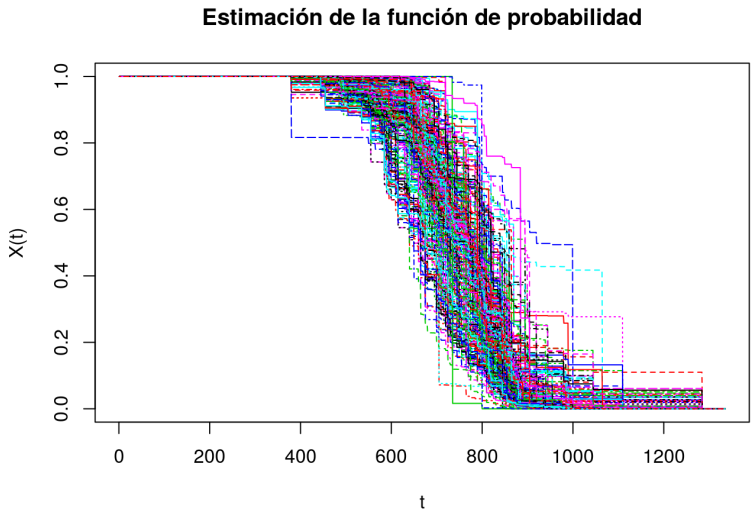


Figura 3.5: Estimación de la función de supervivencia de la variable duración del tiempo de la prueba de esfuerzo condicionada a la variable consumo de oxígeno para cada dato de la muestra.

Capítulo 4

Análisis de los resultados en competición

4.1. Análisis multivariante

En esta sección se aplicarán varias técnicas de análisis multivariante a los datos, de dimensión 30, obtenidos de los registros de las pruebas de esfuerzo realizadas en Inzell. Para ello contamos con la muestra compuesta por 46 individuos, 33 hombres y 13 mujeres, descrita en el capítulo 1.

4.1.1. Análisis de componentes principales

Las técnicas de reducción de la dimensión incluyen diversos procedimientos que pretenden reducir el número de variables a considerar. La necesidad de la utilización de estas técnicas aparece de manera natural cuando se necesita analizar una gran cantidad de datos multidimensionales. Lo que se consigue con estas técnicas es encontrar una representación de nuestros datos sin apenas perder información, evitando así muchos problemas de la alta dimensión como puede ser el efecto de la maldición de la dimensionalidad. La causa común de estos problemas es que cuando aumenta la dimensionalidad, el volumen del espacio aumenta exponencialmente haciendo que los datos disponibles se vuelvan dispersos. Esta dispersión es problemática para cualquier método que requiera significación estadística. Con el fin de obtener un resultado estadísticamente sólido y fiable, la cantidad de datos necesarios para mantener el resultado a menudo debe crecer también exponencialmente con la dimensionalidad. El problema de la maldición de la dimensionalidad implica que no podamos usar técnicas estadísticas a nivel local o flexibles para inferir resultados generales en espacios de gran dimensión.

El análisis de componentes principales es la técnica más conocida y más utilizada en la reducción de la dimensión. Ha tenido éxito en muchas ramas del conocimiento, como por ejemplo en el ámbito de la genética y en el de la inteligencia artificial, usada, entre otras cosas para la digitalización de documentos manuscritos, el reconocimiento facial o la compresión de información.

En líneas generales, el análisis de componentes principales consiste en proyectar las variables originales sobre otras nuevas, en menor número, de tal forma que estas últimas capturen la máxima varianza posible del conjunto de datos originales. El éxito de esta técnica reside en que recoge la máxima variabilidad posible, aunque supone pérdida de información.

Dada la alta dimensión de los datos en relación con el tamaño muestral (46 datos y 30 variables), en esta sección se va a aplicar una técnica no recogida en la literatura clásica como es la de componentes principales tipo sparse. Para ello, en primer lugar se introducirán los componentes principales tipo SCoTLAS, por ser la primera técnica de componentes principales en alta dimensión que permitía calcular componentes principales imponiendo que bastantes coordenadas sean iguales a cero. A continuación se explicará el desarrollo metodológico de los componentes principales tipo sparse [32]. Finalmente, se buscarán qué variables explican la posible mayor variabilidad de los datos, sobre la matriz de correlaciones, evitando así la posible influencia de los cambios de escalas.

Los cálculos se realizarán en R, con el paquete elasticnet.

Compontes principales tipo SCoTLAS

Definición 4.1 *Dada una variable aleatoria X, con valores en $\mathbb{R}^n$, con media $\mu = 0$, se definen los componentes principales tipo SCoTLAS como el conjunto de n vectores $\nu_1, \nu_2, \ldots, \nu_n \in \mathbb{R}^n$, resultantes de resolver los siguientes n problemas de optimización:*

$$\nu_k = \arg\max_{\nu} \nu'(X'X)\nu, \quad con \;\; k \in \{1, \ldots n\},$$

sujeto a:

$$\nu_k'\nu_k = 1 \quad \nu_k'\nu_h = 0, \;\; \forall\, h < k$$

y la restricción extra

$$||\nu_k||_1 < \delta,$$

para un $\delta > 0$ fijado de tal forma que algunas coordenadas de los componentes principales resulten ser cero.

Compontes principales tipo sparse

Definición 4.2 *Dada una matriz $A \in M_{n\times m}(\mathbb{R})$, se define la norma de Frobenius de la matriz A, mediante:*
$||A||_F = \sqrt{tr(A^tA)}$

Definición 4.3 *Dado un vector $x \in \mathbb{R}^n$, se define la norma uno por:*
$||x||_1 = \sum_{i=1}^{n} |x_i|$.

Definición 4.4 *Para calcular los componentes principales sparse, se fija en primer lugar un número entero $k > 0$ de componentes principales y, a continuación, se resuelve el siguiente problema de optimización:*

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} ||X - AB^tX||_F^2 + \epsilon_2 \sum_{j=1}^{k} ||B_j||^2 + \sum_{j=1}^{k} \epsilon_{1,j}||B_j||_1$$

sujeto a que A es una matriz $k \times k$ ortogonal, esto es:

$$AA^t = I,$$

donde I es la matriz identidad $k \times k$, B_j la columna j de la matriz $B \in M_{n\times m}(\mathbb{R})$, $||.||_F$ denota la norma de Frobenius, $||.||$ la norma euclidea, $||.||_1$ la norma uno, y ϵ_2, ϵ_{1k} son números positivos fijados de tal forma que algunas coordenadas de los componentes principales sean cero.

El primer sumando del problema de optimización representa la captura de variabilidad, mientras que los posteriores representan las respectivas penalizaciones del método para los k componentes principales. Los componentes principales vienen dadas por las columnas de la matriz B.

Resultados

En los resultados recogidos en la tabla 4.1, vemos que la primera componente principal es capaz de capturar gran parte la variabilidad: un 87 % y, además, se concentra en las variables asociadas a la potencia en los diferentes umbrales (umbral aeróbico, anaeróbico y láctico) y en el volumen expirado. En la segunda componente son las mismas variables las que explican la variabilidad de nuestros datos, salvo la frecuencia cardíaca en el umbral aeróbico. Es importante señalar que en la primera componente principal, donde se concentra la mayor parte de la variabilidad, todas las coordenadas son no negativas, lo que quiere decir que el valor de la componente principal se incrementa cuanto más aumenta el valor de los distintos parámetros.

A la vista de los resultados, es adecuado quedarnos con los dos primeros componentes principales, pues explican un 93 % de la variabilidad a partir de unas pocas de las variables originales.

Los resultados se han obtenido sobre la matriz de correlación, evitando así, cualquier variación del cambio de escala entre variables en los resultados finales.

4.1.2. Análisis cluster

El análisis cluster es una técnica multivariante que permite agrupar los datos, de tal forma que en un mismo grupo se encuentren los datos más similares entre sí y que, además, los distintos grupos formados sean lo más diferentes entre sí. Con nuestros datos, lo que nos interesa es saber qué deportistas tienen características fisiológicas similares. Para ello realizamos un análisis cluster k-medias para 2 y 3 grupos.

Algoritmo k medias.

Partimos de un conjunto de datos $X_1, \ldots, X_n$ y de k grupos, cuyos centroides iniciales son $m_1^{(0)}, \ldots, m_k^{(0)}$. El algoritmo consta de dos pasos (para cada índice t de iteración):

- Paso de asignación: Asigna cada observación al grupo con la media más cercana, es decir, la partición de las observaciones de acuerdo con el diagrama de Voronoi generado por los centroides:

$$S_i^{(t)} = \{X_p : ||X_p - m_i^{(t)}|| \leq ||X_p - m_j^{(t)}|| \ \forall j \in \{1, 2, \ldots, k\},\ p \in \{1, \ldots n\}\} \tag{4.1}$$

- Calcular los nuevos centroides como el centroide de las observaciones en el grupo,

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{X_j \in S_i^{(t)}} X_j. \tag{4.2}$$

El algoritmo se considera que ha convergido cuando las asignaciones ya no cambian.

Resultados

Los resultados referentes al análisis de cluster con los datos de Inzell pueden observarse en la tabla 4.2. Para el caso de dos grupos, $k = 2$, son interesantes, todas las mujeres aparecen en el mismo cluster. En el caso $k = 3$, se produce la separación en tres grupos, hombres de nivel intermedio, mujeres y hombres de nivel bajo y hombres rápidos.

4.1.3. Análisis de dependencia

En esta sección vamos aplicar un análisis de correlación lineal a las variables iniciales, además de un análisis de dependencia no lineal utilizando la distancia de correlación.

El coeficiente de correlación de Fisher

El coeficiente de correlación de Fisher es una medida lineal de correlación entre dos variables aleatorias X e Y, dando un valor en $[-1, 1]$, donde 1 significa correlación positiva total, 0 ninguna correlación, -1 es correlación negativa total.

Definición 4.5 *Dadas dos variables aleatorias reales X e Y con momentos de orden dos, al menos, se define el coeficiente de correlación lineal de Pearson de las variables X e Y, como:*

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{4.3}$$

Definición 4.6 *Dadas dos variables aleatorias reales X e Y, con momentos de orden dos, al menos, y dos muestras $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ respectivamente de X e Y, se define el coeficiente de correlación de Pearson muestral de las variables X e Y y de las muestras $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$, como:*

$$\hat{r}_{X,Y} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}} \tag{4.4}$$

Distancia de correlación

Uno de los problemas más importantes dentro de la estadística es estudiar si dadas dos variables X e Y, estas son independientes, o bien, por el contrario existe relación entre ellas, y cómo de grande es esta relación.

Tradicionalmente, esta relación se ha medido a través del coeficiente de correlación de Fisher definido en la expresión 4.3. Sin embargo, esta medida tiene varias limitaciones: solo es capaz de detectar relaciones lineales, y muchas veces refleja situaciones de posible incorrelación, cuando en realidad las variables son totalmente dependientes, un ejemplo de esta situación puede ser el caso de las variables $X \in U(-1,1)$ e $Y = X^2$.

El problema de conseguir una medida que sea capaz de detectar relaciones más generales ha tenido un seguimiento muy importante en los últimos años, apareciendo diversas medidas, motivadas especialmente por problemas de genética y microbiología.

En este capítulo vamos a presentar una nueva medida, que ha tenido mucho interés dentro de la comunidad científica, como es la distancia de correlación, V^2, propuesta por Szeley y Rizzo en 2007 [27]. Esta medida toma valores en $[0,1]$ y además es capaz de solventar muchos problemas de la correlación de Fisher, como por ejemplo, el de captar relaciones no lineales entre los datos, pudiendose aplicar a datos multivariantes de distinta dimensión o incluso a datos funcionales.

Definición 4.7 *Dados dos vectores X e Y que toman valores en $\mathbb{R}^p$ y $\mathbb{R}^q$, respectivamente y dadas $\phi_{X,Y}$, ϕ_X y ϕ_Y las funciones características asociadas a (X,Y), X, Y respectivamente, de las cuales asumimos que sus distribuciones marginales tienen momentos en sus primeros ordenes, la distancia correlación entre X e Y se define por:*

$$V^2(X,Y) = \int_{R^{p+q}} |\phi_{X,Y}(u,v) - \phi_X(u)\phi_Y(v)|^2 w(u,v)dudv, \tag{4.5}$$

con $w(u,v) = (c_p c_q |u|_p^{1+p} |v|_q^{1+q})^{-1}$, donde $c_d = \frac{\pi^{1+d/2}}{\Gamma((1+d)/2)}$ es el volumen de media esfera en $\mathbb{R}^{d+1}$ y $|.|_d$ denota la norma euclídea estándar en $\mathbb{R}^d$, finalmente definimos la distancia correlación normalizada mediante:

$$R^2(X,Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{V^2(X,X)V^2(Y,Y)}} & si\ V^2(X,X)V^2(Y,Y) \neq 0 \\ 0 & si\ \ V^2(X,X)V^2(Y,Y) = 0 \end{cases}$$

Observación 4.8 *Con la misma notación de la definición 4.7 se verifica:*

$$V^2(X,Y) = E(|X-X'|_p \cdot |Y-Y'|_q) + E(|X-X'|_p) \cdot E(|Y-Y'|_q) \\ - 2 \cdot E(|X-X'|_p \cdot |Y-Y''|_q), \tag{4.6}$$

siendo (X',Y') y (X'',Y'') vectores aleatorios identicamente distribuidos al (X,Y) de partida y de tal forma que los vectores, (X,Y), (X',Y') y (X'',Y''), son independientes.

Dicho resultado establece la relación existente entre la definición de la distancia de correlación poblacional dada a partir de la función característica y la muestral que veremos a continuación.

Definición 4.9 *Sean X e Y vectores aleatorios y sea $\{(X_i,Y_i)\}_{i=1}^n$ una muestra de sus distribuciones conjuntas, definimos la distancia de correlación empírica mediante:*

$$V_n^2(X,Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{i,j}B_{i,j}, \tag{4.7}$$

donde $A_{ij} = a_{ij} - \overline{a_{i.}} - \overline{a_{.j}} + \overline{a}$ con $a_{ij} = |X_i - X_j|_p$, donde $\overline{a_{i.}}$ denota la media por filas y $\overline{a_{.j}}$ denota la media por columnas respectivamente de la matriz (a_{ij}) mientras que $\overline{a}$ es la media global de dicha matriz. Los B_{ij} se definen de manera totalmente análoga a los A_{ij} a partir de los a_{ij} con los b_{ij} pero en este caso $b_{ij} = |Y_i - Y_j|_q$.

Análogamente, se define la distancia de correlación empírica normalidada por:

$$R_n^2(X,Y) = \begin{cases} \frac{V_n^2(X,Y)}{\sqrt{V_n^2(X,X)V_n^2(Y,Y)}} & si\ V_n^2(X,X)V_n^2(Y,Y) \neq 0 \\ 0 & si\ \ V_n^2(X,X)V_n^2(Y,Y) = 0 \end{cases}$$

La distancia de correlación está implementada en R en el paquete energy a través de la función dcor.

Resultados

Los resultados se encuentran en las tablas 4.3 y 4.4 en donde se muestra una matriz con las relaciones de dependencia para el coeficiente de correlación de Fisher y la distancia de correlación, respectivamente. Vemos que hay diferencias significativas entre aplicar una medida u otra. Por ejemplo, en la sexta variable más relacionada con la variable $VT3RER$, observamos que, si aplicamos una técnica lineal, el resultado es que apenas existe relación con la sexta variable ($VT1V.E.V.CO2$), sin embargo, si utilizamos una métrica que sea capaz de detectar las no linealidades de los datos, el resultado es totalmente opuesto. La situación es todavía mas desfavorable si buscamos un conjunto de variables mayor que seis, pues muchas variables resultan tener coeficientes de correlación cercanos a cero, existiendo una relación de dependencia real como muestra la distancia de correlación.

Algunos resultados bastante significativos desde el punto de vista fisiológico pueden ser los siguientes:

- La dificultad de encontrar variables que estén relacionadas con la frecuencia cardíaca salvo la propia frecuencia cardíaca ($VT1HR$, $VT2HR$, $VT3HR$), siendo la siguiente relación de intensidad leve, el número de inspiraciones por minuto (BR), algo en parte conocido debido a la sincronización entre el número de inspiraciones por minuto y el pulso (HR).
- Las variables más relacionadas con los componentes no nulos de los componentes principales son los propios componentes no nulos.
- Las relaciones más intensas con el consumo de oxígeno máximo vienen dadas por las variables de potencia.

	PC1	PC2	PC3	PC4
VT1V.O2.kg	0.00	0.00	0.00	0.00
VT1V.O2.HR	0.00	0.00	0.00	0.00
VT1HR	0.00	-0.19	-0.21	0.38
VT1WR	0.37	-0.79	-0.32	-0.31
VT1V.E.V.O2	0.00	0.00	0.00	0.00
VT1V.E.V.CO2	0.00	0.00	0.00	0.00
VT1RER	0.00	0.00	0.00	0.00
VT1V.E	0.13	-0.05	-0.01	-0.03
VT1VT	0.00	0.00	0.00	0.00
VT1BF	0.00	0.00	0.00	0.10
VT2V.O2.kg	0.05	0.00	0.07	0.05
VT2V.O2.HR	0.00	0.00	0.00	0.00
VT2HR	0.00	0.00	0.00	0.35
VT2WR	0.53	-0.09	0.72	0.03
VT2V.E.V.O2	0.00	0.00	0.00	0.00
VT2V.E.V.CO2	0.00	0.00	0.00	0.00
VT2RER	0.00	0.00	0.00	0.00
VT2V.E	0.23	0.08	0.30	-0.09
VT2VT	0.00	0.00	0.00	0.00
VT2BF	0.00	0.00	0.00	0.17
VT3V.O3.kg	0.03	0.00	0.00	0.09
VT3V.O3.HR	0.00	0.00	0.00	0.00
VT3HR	0.00	0.00	-0.07	0.24
VT3WR	0.65	0.27	-0.39	0.44
VT3V.E.V.O3	0.00	0.00	0.00	0.00
VT3V.E.V.CO3	0.00	0.00	0.00	0.00
VT3RER	0.00	0.00	0.00	0.00
VT3V.E	0.30	0.51	-0.28	-0.58
	1	2	3	4
Autovalores	0.87	0.06	0.02	0.02

(a) Tabla de autovalores y autovectores

Tabla 4.1: Resultado del análisis de componentes principales tipo sparse.

	Deportista	Sexo	Grupo(k=2)	Grupo(k=3)
1	A1	0	1	3
2	A2	1	1	1
3	A3	1	2	1
4	A4	1	2	2
5	A5	1	1	3
6	A6	0	1	3
7	A7	1	2	2
8	A8	1	2	2
9	A9	1	2	2
10	A10	1	2	1
11	A11	1	2	1
12	A12	1	2	2
13	A13	0	1	3
14	A14	1	2	1
15	A15	0	1	3
16	A16	1	2	1
17	A17	0	1	3
18	A18	0	1	3
19	A19	1	1	3
20	A20	0	1	3
21	A21	1	2	1
22	A22	1	1	1
23	A23	0	1	3
24	A24	1	2	2
25	A25	0	1	3
26	A26	1	2	2
27	A27	1	2	2
28	A28	1	2	2
29	A29	1	2	1
30	A30	1	2	1
31	A31	0	1	3
32	A32	1	2	2
33	A33	0	1	3
34	A34	1	2	2
35	A35	1	2	2
36	A36	1	2	2
37	A37	1	2	2
38	A38	1	2	1
39	A39	1	2	2
40	A40	0	1	3
41	A41	1	2	2
42	A42	1	2	2
43	A43	1	2	2
44	A44	1	2	2
45	A45	0	1	3
46	A46	1	2	2

Tabla 4.2: Resultado del análisis cluster k-medias (con k=2 y k=3).

Variable a estudiar	1 relación	2 relación	3 relación	4 relación	5 relación	6 relación
VT1V.O2.kg	VT2V.O2.kg	VT3V.O3.kg	VT1WR	VT1V.E	VT2WR	VT2V.O2.HR
1.00	0.85	0.79	0.78	0.77	0.71	0.71
VT1V.O2.HR	VT2V.O2.HR	VT3VT	VT1VT	VT2VT	VT2WR	VT3WR
1.00	0.97	0.87	0.87	0.86	0.86	0.84
VT1HR	VT2HR	VT3HR	VT1BF	VT2BF	VT1V.O2.kg	VT3BF
1.00	0.62	0.56	0.33	0.31	0.30	0.28
VT1WR	VT1V.E	VT2WR	VT3WR	VT1V.O2.HR	VT2V.O2.HR	VT1VT
1.00	0.89	0.87	0.83	0.80	0.79	0.79
VT1V.E.V.O2	VT1V.E.V.CO2	VT2V.E.V.CO2	VT1BF	VT3V.E.V.CO3	VT1RER	VT3V.E.V.O3
1.00	0.84	0.66	0.53	0.49	0.44	0.42
VT1V.E.V.COA	VT1V.E.V.O2	VT2V.E.V.CO2	VT3V.E.V.CO3	VT1BF	VT3V.E.V.O3	VT1V.E
1.00	0.84	0.83	0.59	0.42	0.37	0.35
VT1RER	VT2RER	VT1V.E.V.O2	VT1BF	VT2V.E.V.O2	VT1V.E	VT3V.E.V.O3
1.00	0.62	0.44	0.26	0.20	0.18	0.16
VT1V.E	VT2V.E	VT1WR	VT2WR	VT3WR	VT2V.O2.HR	VT1V.O2.HR
1.00	0.90	0.89	0.89	0.88	0.83	0.82
VT1VT	VT2VT	VT3VT	VT1V.O2.HR	VT2V.O2.HR	VT2WR	VT2V.E
1.00	0.93	0.89	0.87	0.83	0.80	0.79
VT1BF	VT2BF	VT3BF	VT1V.E.V.O2	VT1V.E.V.CO2	VT1HR	VT1RER
1.00	0.72	0.63	0.53	0.42	0.33	0.26
VT2V.O2.kg	VT3V.O3.kg	VT1V.O2.kg	VT2V.O2.HR	VT2WR	VT3WR	VT2V.E
1.00	0.93	0.85	0.75	0.73	0.72	0.69
VT2V.O2.HR	VT1V.O2.HR	VT2WR	VT3WR	VT2V.E	VT3V.E	VT3VT
1.00	0.97	0.90	0.89	0.87	0.86	0.85
VT2HR	VT3HR	VT1HR	VT2V.E.V.CO2	VT2V.O2.kg	VT1V.O2.kg	VT2BF
1.00	0.78	0.62	0.31	0.31	0.30	0.26
VT2WR	VT3WR	VT2V.E	VT2V.O2.HR	VT1V.E	VT1WR	VT1V.O2.HR
1.00	0.94	0.93	0.90	0.89	0.87	0.86
VT2V.E.V.O2	VT2RER	VT3RER	VT1RER	VT2V.E.V.CO2	VT2BF	VT1V.E.V.O2
1.00	0.30	0.22	0.20	0.20	0.17	0.13
VT2V.E.V.CO2	VT1V.E.V.CO2	VT3V.E.V.CO3	VT1V.E.V.O2	VT3V.E.V.O3	VT2V.E	VT1V.E
1.00	0.83	0.73	0.66	0.50	0.45	0.35
VT2RER	VT1RER	VT2V.E.V.O2	VT2BF	VT3RER	VT3BF	VT3V.E.V.O3
1.00	0.62	0.30	0.22	0.17	0.12	0.11
VT2V.E	VT2WR	VT1V.E	VT3WR	VT3V.E	VT2V.O2.HR	VT1V.O2.HR
1.00	0.93	0.90	0.89	0.88	0.87	0.83
VT2VT	VT3VT	VT1VT	VT1V.O2.HR	VT2V.O2.HR	VT2WR	VT2V.E
1.00	0.97	0.93	0.86	0.84	0.80	0.79
VT2BF	VT1BF	VT3BF	VT2V.E.V.CO2	VT1V.E.V.O2	VT1HR	VT1V.E.V.CO2
1.00	0.72	0.66	0.34	0.33	0.31	0.27
VT3V.O3.kg	VT2V.O2.kg	VT1V.O2.kg	VT2V.O2.HR	VT3WR	VT2WR	VT1V.O2.HR
1.00	0.93	0.79	0.70	0.68	0.63	0.61
VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR	VT1VT	VT2VT	VT2V.E	VT2WR
1.00	0.83	0.82	0.78	0.77	0.77	0.77
VT3HR	VT2HR	VT1HR	VT3V.E.V.CO3	VT3WR	VT2V.E.V.CO2	VT1V.E.V.CO2
1.00	0.78	0.56	0.26	0.25	0.25	0.20
VT3WR	VT2WR	VT2V.E	VT2V.O2.HR	VT3V.E	VT1V.E	VT1V.O2.HR
1.00	0.94	0.89	0.89	0.88	0.88	0.84
VT3V.E.V.O3	VT3V.E.V.CO3	VT3V.E	VT2V.E.V.CO2	VT3BF	VT1V.E.V.O2	VT1V.E.V.CO2
1.00	0.83	0.51	0.50	0.47	0.42	0.37
VT3V.E.V.CO3	VT3V.E.V.O3	VT2V.E.V.CO2	VT1V.E.V.CO2	VT3V.E	VT1V.E.V.O2	VT2V.E
1.00	0.83	0.73	0.59	0.59	0.49	0.46
VT3RER	VT2V.E.V.O2	VT2RER	VT3V.E.V.O3	VT1RER	VT1V.E.V.O2	VT1V.E.V.CO2
1.00	0.22	0.17	0.15	0.12	0.10	0.05
VT3V.E	VT2V.E	VT3WR	VT2V.O2.HR	VT1V.O2.HR	VT2WR	VT1V.E
1.00	0.88	0.88	0.86	0.83	0.82	0.80
VT3VT	VT2VT	VT1VT	VT1V.O2.HR	VT2V.O2.HR	VT2WR	VT3V.E
1.00	0.97	0.89	0.87	0.85	0.80	0.79
VT3BF	VT2BF	VT1BF	VT3V.E.V.O3	VT3V.E.V.CO3	VT1HR	VT1V.E.V.O2
1.00	0.66	0.63	0.47	0.40	0.28	0.28

Variables con mayor relación

(a) Tabla lineal variables

Tabla 4.3: Resultado del análisis de dependencia con el coeficiente de correlación. En el primer elemento de cada fila se muestra cada variable. Las seis variables más relacionadas con ella aparecen por orden decreciente en los sucesivos elementos de la fila. En la parte inferior de cada fila, para cada variable, se proporciona el valor numérico de dicha relación.

Variables con mayor relación

Variable a estudiar	1 relación	2 relación	3 relación	4 relación	5 relación	6 relación
VT1V.O2.kg	VT1V.E	VT2V.O2.kg	VT1WR	VT2V.E	VT3WR	VT3V.O3.kg
1.00	0.78	0.78	0.74	0.71	0.71	0.70
VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT3VT	VT1VT	VT2WR	VT3V.E
1.00	0.97	0.90	0.87	0.85	0.85	0.84
VT1HR	VT2HR	VT3HR	VT3RER	VT2BF	VT1BF	VT1V.O2.kg
1.00	0.67	0.52	0.40	0.38	0.38	0.37
VT1WR	VT1V.E	VT2WR	VT3WR	VT2V.O2.HR	VT1V.O2.HR	VT2V.E
1.00	0.89	0.84	0.83	0.80	0.80	0.78
VT1V.E.V.O2	VT1V.E.V.CO2	VT2V.E.V.CO2	VT3V.E.V.CO3	VT2V.E.V.O2	VT1BF	VT1RER
1.00	0.81	0.62	0.54	0.54	0.50	0.45
VT1V.E.V.CO2	VT1V.E.V.O2	VT2V.E.V.CO2	VT3V.E.V.CO3	VT2V.E.V.O2	VT1V.E	VT1BF
1.00	0.81	0.77	0.62	0.54	0.41	0.40
VT1RER	VT2RER	VT1V.E.V.O2	VT3RER	VT1BF	VT2HR	VT2BF
1.00	0.62	0.45	0.36	0.35	0.31	0.31
VT1V.E	VT2V.E	VT3WR	VT2WR	VT1WR	VT2V.O2.HR	VT3V.E
1.00	0.91	0.90	0.89	0.89	0.86	0.86
VT1VT	VT2VT	VT3VT	VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT2WR
1.00	0.94	0.92	0.85	0.82	0.82	0.78
VT1BF	VT2BF	VT3BF	VT1V.E.V.O2	VT2VT	VT1VT	VT3VT
1.00	0.71	0.62	0.50	0.47	0.44	0.44
VT2V.O2.kg	VT3V.O3.kg	VT1V.O2.kg	VT2WR	VT2V.E	VT3WR	VT2V.O2.HR
1.00	0.88	0.78	0.74	0.73	0.72	0.71
VT2V.O2.HR	VT1V.O2.HR	VT3V.O3.HR	VT2WR	VT3WR	VT2V.E	VT3V.E
1.00	0.97	0.92	0.90	0.89	0.88	0.88
VT2HR	VT3HR	VT1HR	VT1V.O2.kg	VT2V.E.V.CO2	VT2V.O2.kg	VT1V.E.V.O2
1.00	0.76	0.67	0.40	0.37	0.36	0.34
VT2WR	VT3WR	VT2V.E	VT2V.O2.HR	VT1V.E	VT1V.O2.HR	VT1WR
1.00	0.94	0.94	0.90	0.89	0.85	0.84
VT2V.E.V.O2	VT2V.E.V.CO2	VT3V.E.V.CO3	VT1V.E.V.CO2	VT1V.E.V.O2	VT3V.E.V.O3	VT2BF
1.00	0.74	0.57	0.54	0.54	0.49	0.39
VT2V.E.V.CO2	VT1V.E.V.CO2	VT3V.E.V.CO3	VT2V.E.V.O2	VT1V.E.V.O2	VT3V.E.V.O3	VT2V.E
1.00	0.77	0.76	0.74	0.62	0.52	0.49
VT2RER	VT1RER	VT3RER	VT2V.E.V.O2	VT3HR	VT3WR	VT1WR
1.00	0.62	0.50	0.38	0.36	0.35	0.33
VT2V.E	VT2WR	VT1V.E	VT3WR	VT3V.E	VT2V.O2.HR	VT3V.O3.HR
1.00	0.94	0.91	0.91	0.90	0.88	0.85
VT2VT	VT3VT	VT1VT	VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT2WR
1.00	0.96	0.94	0.82	0.82	0.80	0.78
VT2BF	VT1BF	VT3BF	VT2V.E.V.O2	VT1HR	VT3VT	VT2V.E.V.CO2
1.00	0.71	0.63	0.39	0.38	0.37	0.37
VT3V.O3.kg	VT2V.O2.kg	VT1V.O2.kg	VT3WR	VT2V.O2.HR	VT2WR	VT2V.E
1.00	0.88	0.70	0.65	0.64	0.62	0.61
VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR	VT2V.E	VT2WR	VT3VT	VT3V.E
1.00	0.92	0.90	0.85	0.84	0.84	0.83
VT3HR	VT2HR	VT1HR	VT3V.E.V.CO3	VT2V.E.V.CO2	VT2RER	VT3WR
1.00	0.76	0.52	0.38	0.37	0.36	0.34
VT3WR	VT2WR	VT2V.E	VT3V.E	VT1V.E	VT2V.O2.HR	VT1V.O2.HR
1.00	0.94	0.91	0.90	0.90	0.89	0.84
VT3V.E.V.O3	VT3V.E.V.CO3	VT2V.E.V.CO2	VT2V.E.V.O2	VT3V.E	VT3BF	VT1V.E.V.O2
1.00	0.79	0.52	0.49	0.48	0.45	0.44
VT3V.E.V.CO3	VT3V.E.V.O3	VT2V.E.V.CO2	VT1V.E.V.CO2	VT3V.E	VT2V.E.V.O2	VT1V.E.V.O2
1.00	0.79	0.76	0.62	0.59	0.57	0.54
VT3RER	VT2RER	VT1WR	VT1V.O2.kg	VT1HR	VT3V.E.V.O3	VT1V.E
1.00	0.50	0.42	0.40	0.40	0.39	0.36
VT3V.E	VT3WR	VT2V.E	VT2V.O2.HR	VT1V.E	VT1V.O2.HR	VT3V.O3.HR
1.00	0.90	0.90	0.88	0.86	0.84	0.83
VT3VT	VT2VT	VT1VT	VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT2WR
1.00	0.96	0.92	0.87	0.86	0.84	0.80
VT3BF	VT2BF	VT1BF	VT3V.E.V.O3	VT3VT	VT2VT	VT3V.E.V.CO3
1.00	0.63	0.62	0.45	0.42	0.40	0.38

(a) Tabla no lineal variables

Tabla 4.4: Resultado del análisis de dependencia con el coeficiente de correlación. En el primer elemento de cada fila se muestra cada variable. Las seis variables más relacionadas con ella aparecen por orden decreciente en los sucesivos elementos de la fila. En la parte inferior de cada fila, para cada variable, se proporciona el valor numérico de dicha relación.

4.2. Relación entre los parámetros fisiológicos y las marcas en competición

En esta sección se estudiará la relación entre los parámetros obtenidos en las pruebas de esfuerzo y los resultados en competición. Para ello contamos con información de los resultados deportivos en distancias de 500 y 1000 m para 18 pruebas de esfuerzo.

El análisis estadístico que vamos a realizar será, de nuevo, un análisis de dependencia, pero esta vez teniendo en cuenta el efecto de los tiempos en las distancias de 500 m y 1000 m y, a continuación, ajustar un modelo de regresión simple con un método de selección de variables maxima-hunting con tres variables explicativas, utilizando la distancia de correlación. La razón para usar un modelo de regresión lineal es la escasez en el número de datos, además únicamente utilizamos tres variables explicativas, por la misma razón: en caso de usar más covariables estaríamos sobreajustando el modelo y no podríamos llevar a cabo técnicas de la inferencia estadística, como la de predicción.

4.2.1. El modelo de regresión lineal general

Dada una variable aleatoria $X \in \mathbb{R}^p$ y una variable aleatoria $Y \in \mathbb{R}$. Uno los principales objetivos de la estadística es ser capaces de construir modelos, a partir de una muestra aleatoria simple $X_1, \ldots, X_n \in \mathbb{R}^p$ de X, e $Y_1, \ldots, Y_n \in \mathbb{R}$ de Y, que permitan predecir, a partir de los valores de X, el valor de la variable Y, esto es, ajustar modelos de la siguiente forma:

$$Y = m(X_1, \ldots, X_p) + \epsilon, \tag{4.8}$$

donde ϵ denota el error.

Una familia importante de estos modelos, es aquella en la que la función $m : \mathbb{R}^p \to \mathbb{R}$ resulta ser una aplicación lineal. En este caso, nos encontramos ante modelos de la forma:

$$Y = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \epsilon, \tag{4.9}$$

y en esta situación se suele considerar $\epsilon \in N(0, \sigma^2)$. Si tenemos en cuenta la existencia de la muestra aleatoria simple anterior, podemos reescribir las ecuaciones del modelo de manera matricial mediante:

$$Y = X\beta + \epsilon, \tag{4.10}$$

donde $Y \in \mathbb{R}^n$, $X \in M_{n\times p}(\mathbb{R})$, $\beta \in \mathbb{R}^p$, $\epsilon \in N_n(0, \sigma^2 I)$, X_i denota la i-ésima observación de la muestra, la fila i de la matriz X, mientras Y_i la i-ésima observación de la muestra y componente del vector Y.

El estimador mínimo cuadrático será aquel que resuelva:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}(\sum_{i=1}^{n}(Y_i - X_i'\beta)^2) = \arg\min_{\beta\in\mathbb{R}^p}(||Y - X\beta||^2). \tag{4.11}$$

Es claro que el mínimo anterior, se verifica, cuando:

$$X^{'}(Y - X\beta) = 0,$$

por tanto, $\hat{\beta} = (X'X)^{-1}X'Y$. y $\hat{Y} = X(X'X)^{-1}X'Y$. Además, es fácil comprobar que: $E(\hat{\beta}) = \beta$, $cov(\hat{\beta}, \hat{\beta}) = (X'X)^{-1}\sigma^2$.

4.2.2. Modelo maxima-hunting

El método maxima-hunting es un método de selección de variables en problemas de clasificación y regresión, tanto de variables multivariantes como funcionales [3].

La idea del método es sencilla. Dado un conjunto de covariables $X = (X_1, \ldots, X_p)$, una variable respuesta, Y, y una medida de asociación $M(X, Y)$, el problema consiste en encontrar un subconjunto, S, de partes de X, que maximice la cantidad $M(S, Y)$. Para aplicar dicho método es necesario exigir ciertas condiciones al subconjunto S de X, como el número de elementos.

4.2.3. Resultados

Análisis dependencia

Las tablas 4.5 y 4.6 muestran el resultado del análisis de dependencia con el coeficiente de correlación y la distancia de correlación respectivamente. Vemos que las variables más relacionadas con los resultados deportivos son los propios resultados deportivos ($X500$, $X1000$), algo más que razonable: las necesidades energéticas de las pruebas de 500 m y 1000 m son bastante similares, además de que es normal una mayor relación entre un movimiento específico, como es realizar un esfuerzo en una pista de patinaje, que otro realizado en una bicicleta estática.

Las siguientes relaciones más significativas con el rendimiento deportivo son las asociadas a las variables de potencia (WR) y, además, estas no son lineales como podemos ver comparando los resultados obtenidos entre las tablas 4.5 y 4.6. Es importante señalar que las variables más relacionadas con los parámetros fisiológicos no aparecen los resultados deportivos, lo cual es bastante significativo.

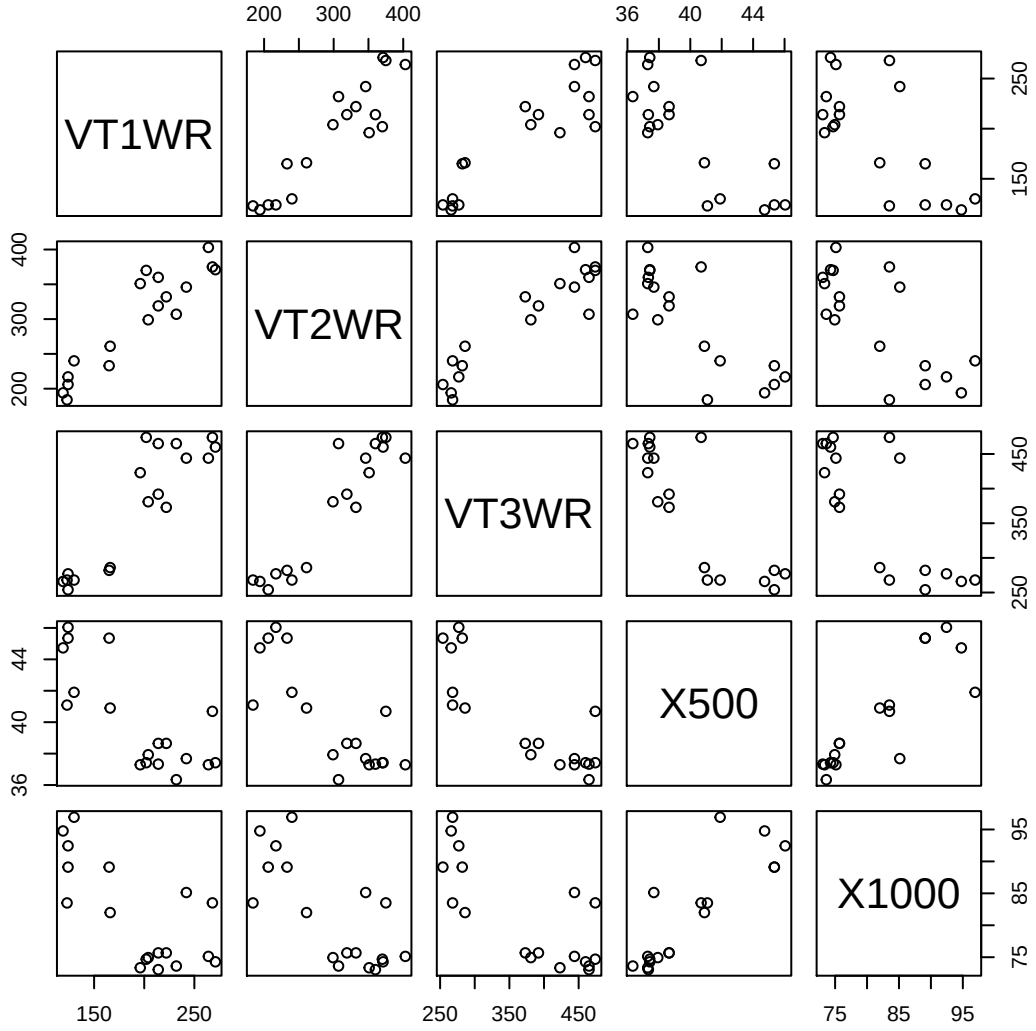


Figura 4.1: Gráfico de pares de variables entre variables de potencia y tiempos en competición.

En un gráfico conjunto (figura 4.1) entre pares de variables vemos que, en efecto, las relaciones entre la potencia a los distintos umbrales ($VT1WR, VT2WR, VT3WR$) y las marcas de compe-

Variables con mayor relación

Variable a estudiar	1 relación	2 relación	3 relación	4 relación	5 relación	6 relación
VT1V.O2.kg	VT3V.O3.kg	VT2V.O2.kg	VT1WR	VT1V.E	VT2V.O2.HR	VT3V.O3.HR
1.00	0.81	0.80	0.79	0.78	0.62	0.60
VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT1VT	VT3VT	VT2VT	VT3V.E
1.00	0.96	0.96	0.94	0.90	0.90	0.88
VT1HR	VT2HR	VT2BF	VT1BF	VT1V.E.V.O2	VT2V.E.V.O2	VT3BF
1.00	0.72	0.67	0.64	0.54	0.42	0.37
VT1WR	VT1V.E	VT2WR	VT3WR	VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR
1.00	0.93	0.92	0.90	0.90	0.90	0.86
VT1V.E.V.O2	VT1V.E.V.CO2	VT1BF	VT2V.E.V.O2	VT2BF	VT1HR	VT2V.E.V.CO2
1.00	0.80	0.67	0.63	0.57	0.54	0.54
VT1V.E.V.CO2	VT1V.E.V.O2	VT2V.E.V.CO2	VT2V.E.V.O2	VT2BF	VT1BF	VT3V.E.V.CO3
1.00	0.80	0.78	0.62	0.46	0.45	0.44
VT1RER	VT2RER	VT3RER	VT1V.E.V.O2	VT1BF	VT3HR	VT3BF
1.00	0.67	0.65	0.48	0.44	0.37	0.36
VT1V.E	VT1WR	VT2V.E	VT3WR	VT2WR	VT2V.O2.HR	VT3V.O3.HR
1.00	0.93	0.90	0.89	0.89	0.82	0.80
VT1VT	VT2VT	VT1V.O2.HR	VT3VT	VT3V.O3.HR	VT2V.O2.HR	VT1WR
1.00	0.94	0.94	0.92	0.90	0.88	0.85
VT1BF	VT2BF	VT3BF	VT1V.E.V.O2	VT1HR	VT2HR	VT1V.E.V.CO2
1.00	0.92	0.73	0.67	0.64	0.46	0.45
VT2V.O2.kg	VT3V.O3.kg	VT1V.O2.kg	VT1WR	VT2WR	VT2V.O2.HR	VT1V.E
1.00	0.92	0.80	0.73	0.67	0.66	0.64
VT2V.O2.HR	VT3V.O3.HR	VT1V.O2.HR	VT2WR	VT3WR	VT1WR	VT3V.E
1.00	0.97	0.96	0.91	0.91	0.90	0.89
VT2HR	VT1HR	VT3HR	VT2V.E.V.O2	VT1V.E.V.O2	VT2BF	VT1BF
1.00	0.72	0.69	0.54	0.53	0.48	0.46
VT2WR	VT2V.E	VT3WR	VT1WR	VT2V.O2.HR	VT3V.O3.HR	VT1V.E
1.00	0.94	0.93	0.92	0.91	0.90	0.89
VT2V.E.V.O2	VT2V.E.V.CO2	VT1V.E.V.O2	VT1V.E.V.CO2	VT2V.E	VT2HR	VT3V.E.V.CO3
1.00	0.82	0.63	0.62	0.54	0.54	0.51
VT2V.E.V.CO2	VT2V.E.V.O2	VT1V.E.V.CO2	VT3V.E.V.CO3	VT2V.E	VT1V.E.V.O2	VT1V.E
1.00	0.82	0.78	0.69	0.60	0.54	0.49
VT2RER	VT1RER	VT3RER	VT1BF	VT2V.E.V.O2	VT2BF	X500
1.00	0.67	0.64	0.43	0.38	0.36	0.35
VT2V.E	VT2WR	VT1V.E	VT3WR	VT1WR	VT3V.E	VT2V.O2.HR
1.00	0.94	0.90	0.86	0.84	0.84	0.83
VT2VT	VT3VT	VT1VT	VT1V.O2.HR	VT3V.O3.HR	VT2V.O2.HR	VT2WR
1.00	0.98	0.94	0.90	0.87	0.86	0.80
VT2BF	VT1BF	VT3BF	VT1HR	VT1V.E.V.O2	VT2HR	VT2V.E.V.O2
1.00	0.92	0.76	0.67	0.57	0.48	0.47
VT3V.O3.kg	VT2V.O2.kg	VT1V.O2.kg	VT1WR	VT3V.O3.HR	VT3WR	VT2V.O2.HR
1.00	0.92	0.81	0.73	0.71	0.69	0.68
VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR	VT3WR	VT2WR	VT1VT	VT1WR
1.00	0.97	0.96	0.92	0.90	0.90	0.90
VT3HR	VT2HR	VT1V.E.V.O2	VT1V.E	VT2V.E	VT3WR	VT3V.E
1.00	0.69	0.52	0.49	0.45	0.43	0.42
VT3WR	VT2WR	VT3V.O3.HR	VT2V.O2.HR	VT1WR	VT1V.E	VT3V.E
1.00	0.93	0.92	0.91	0.90	0.89	0.86
VT3V.E.V.O3	VT3V.E.V.CO3	VT3V.E	VT3RER	VT3BF	VT2V.E.V.O2	VT2V.E
1.00	0.83	0.65	0.52	0.42	0.40	0.40
VT3V.E.V.CO3	VT3V.E.V.O3	VT3V.E	VT2V.E.V.CO2	VT2V.E	VT2V.E.V.O2	VT3WR
1.00	0.83	0.73	0.69	0.58	0.51	0.50
VT3RER	VT1RER	VT2RER	VT3V.E.V.O3	VT3BF	X500	X1000
1.00	0.65	0.64	0.52	0.37	0.32	0.14
VT3V.E	VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR	VT3WR	VT2V.E	VT2WR
1.00	0.90	0.89	0.88	0.86	0.84	0.81
VT3VT	VT2VT	VT1VT	VT1V.O2.HR	VT3V.O3.HR	VT2V.O2.HR	VT2WR
1.00	0.98	0.92	0.90	0.88	0.86	0.80
VT3BF	VT2BF	VT1BF	VT3V.E.V.O3	VT1V.E.V.O2	VT1HR	VT3RER
1.00	0.76	0.73	0.42	0.39	0.37	0.37
X500	X1000	VT2RER	VT3RER	VT3BF	VT2BF	VT1HR
1.00	0.86	0.35	0.32	0.17	0.07	0.07
X1000	X500	VT2RER	VT3RER	VT3BF	VT1BF	VT1HR
1.00	0.86	0.17	0.14	-0.02	-0.10	-0.11

(a) Tabla lineal variables

Tabla 4.5: Resultado del análisis de dependencia con el coeficiente de correlación. En el primer elemento de cada fila se muestra cada variable. Las seis variables más relacionadas con ella aparecen por orden decreciente en los sucesivos elementos de la fila. En la parte inferior de cada fila, para cada variable, se proporciona el valor numérico de dicha relación.

Variables con mayor relación

	Variable a estudiar	1 relación	2 relación	3 relación	4 relación	5 relación	6 relación
	VT1V.O2.kg	VT1V.E	VT1WR	VT3V.O3.kg	VT2V.O2.kg	VT2V.O2.HR	VT3V.E
	1.00	0.78	0.77	0.77	0.77	0.66	0.66
	VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT1VT	VT3VT	VT2VT	VT3V.E
	1.00	0.96	0.95	0.94	0.91	0.88	0.87
	VT1HR	VT2HR	VT2BF	VT1BF	VT2V.E.V.O2	VT1V.E.V.O2	VT2V.E.V.CO2
	1.00	0.77	0.75	0.66	0.61	0.60	0.52
4	VT1WR	VT1V.E	VT3V.O3.HR	VT2V.O2.HR	VT2WR	VT3WR	VT2V.E
	1.00	0.94	0.93	0.92	0.92	0.92	0.88
	VT1V.E.V.O2	VT1V.E.V.CO2	VT1BF	VT2V.E.V.O2	VT2V.E.V.CO2	VT2BF	VT3HR
	1.00	0.82	0.69	0.69	0.62	0.62	0.61
	VT1V.E.V.CO2	VT1V.E.V.O2	VT2V.E.V.CO2	VT2V.E.V.O2	VT3V.E.V.CO3	VT2BF	VT3RER
	1.00	0.82	0.81	0.70	0.59	0.48	0.47
	VT1RER	VT2RER	VT3RER	VT1BF	VT3V.E.V.O3	VT1V.E.V.O2	VT2HR
	1.00	0.68	0.65	0.54	0.53	0.50	0.48
	VT1V.E	VT1WR	VT3WR	VT2V.E	VT2WR	VT3V.E	VT3V.O3.HR
	1.00	0.94	0.92	0.91	0.91	0.89	0.87
	VT1VT	VT2VT	VT3VT	VT1V.O2.HR	VT3V.O3.HR	VT2V.O2.HR	VT1WR
	1.00	0.96	0.95	0.94	0.91	0.91	0.86
	VT1BF	VT2BF	VT3BF	VT1V.E.V.O2	VT1HR	VT2VT	VT1VT
	1.00	0.90	0.72	0.69	0.66	0.64	0.59
	VT2V.O2.kg	VT3V.O3.kg	VT1V.O2.kg	VT1WR	VT1V.E	VT2WR	VT2V.E
	1.00	0.91	0.77	0.73	0.70	0.70	0.68
	VT2V.O2.HR	VT3V.O3.HR	VT1V.O2.HR	VT1WR	VT2WR	VT3WR	VT3V.E
	1.00	0.98	0.96	0.92	0.92	0.92	0.92
	VT2HR	VT1HR	VT3HR	VT2BF	VT2V.E.V.O2	VT1V.E.V.O2	VT1BF
	1.00	0.77	0.69	0.61	0.61	0.59	0.55
	VT2WR	VT2V.E	VT3WR	VT1WR	VT2V.O2.HR	VT3V.O3.HR	VT1V.E
	1.00	0.95	0.95	0.92	0.92	0.92	0.91
	VT2V.E.V.O2	VT2V.E.V.CO2	VT1V.E.V.CO2	VT1V.E.V.O2	VT1HR	VT2HR	VT3V.E.V.CO3
	1.00	0.81	0.70	0.69	0.61	0.61	0.60
	VT2V.E.V.CO2	VT2V.E.V.O2	VT1V.E.V.CO2	VT3V.E.V.CO3	VT1V.E.V.O2	VT2V.E	VT1HR
	1.00	0.81	0.81	0.74	0.62	0.58	0.52
	VT2RER	VT1RER	VT3RER	VT3V.O3.kg	VT1V.O2.HR	VT2V.E.V.O2	VT3V.O3.HR
	1.00	0.68	0.60	0.52	0.51	0.51	0.50
	VT2V.E	VT2WR	VT1V.E	VT3WR	VT3V.E	VT1WR	VT2V.O2.HR
	1.00	0.95	0.91	0.90	0.89	0.88	0.87
	VT2VT	VT3VT	VT1VT	VT1V.O2.HR	VT2V.O2.HR	VT3V.O3.HR	VT2WR
	1.00	0.97	0.96	0.88	0.87	0.86	0.82
	VT2BF	VT1BF	VT1HR	VT3BF	VT1V.E.V.O2	VT2HR	VT2V.E.V.O2
	1.00	0.90	0.75	0.72	0.62	0.61	0.58
	VT3V.O3.kg	VT2V.O2.kg	VT1V.O2.kg	VT1WR	VT1V.E	VT3V.O3.HR	VT2V.O2.HR
	1.00	0.91	0.77	0.75	0.74	0.72	0.71
	VT3V.O3.HR	VT2V.O2.HR	VT1V.O2.HR	VT3WR	VT1WR	VT3V.E	VT2WR
	1.00	0.98	0.95	0.93	0.93	0.92	0.92
	VT3HR	VT2HR	VT1V.E.V.O2	VT2V.E	VT3V.E.V.CO3	VT3V.E	VT1V.E
	1.00	0.69	0.61	0.55	0.54	0.54	0.54
	VT3WR	VT2WR	VT3V.O3.HR	VT2V.O2.HR	VT1V.E	VT3V.E	VT1WR
	1.00	0.95	0.93	0.92	0.92	0.92	0.92
	VT3V.E.V.O3	VT3V.E.V.CO3	VT3RER	VT2V.E.V.O2	VT3V.E	VT3BF	VT1RER
	1.00	0.79	0.68	0.59	0.58	0.55	0.53
	VT3V.E.V.CO3	VT3V.E.V.O3	VT2V.E.V.CO2	VT3V.E	VT2V.E.V.O2	VT1V.E.V.CO2	VT2V.E
	1.00	0.79	0.74	0.69	0.60	0.59	0.59
	VT3RER	VT3V.E.V.O3	VT1RER	VT2RER	VT3BF	VT2V.E.V.CO2	VT2HR
	1.00	0.68	0.65	0.60	0.55	0.50	0.49
	VT3V.E	VT3WR	VT3V.O3.HR	VT2V.O2.HR	VT2V.E	VT1V.E	VT2WR
	1.00	0.92	0.92	0.92	0.89	0.89	0.88
	VT3VT	VT2VT	VT1VT	VT3V.O3.HR	VT1V.O2.HR	VT2V.O2.HR	VT2WR
	1.00	0.97	0.95	0.91	0.91	0.89	0.84
	VT3BF	VT2BF	VT1BF	VT2VT	VT3VT	VT3V.E.V.O3	VT3RER
	1.00	0.72	0.72	0.56	0.56	0.55	0.55
	X500	X1000	VT3WR	VT2WR	VT3V.O3.HR	VT1V.E	VT1WR
	1.00	0.89	0.86	0.86	0.81	0.81	0.81
	X1000	X500	VT3WR	VT2WR	VT1V.E	VT3V.O3.HR	VT1WR
	1.00	0.89	0.78	0.78	0.78	0.77	0.76

(a) Tabla no lineal variables

Tabla 4.6: Resultado del análisis de dependencia con el coeficiente de correlación. En el primer elemento de cada fila se muestra cada variable. Las seis variables más relacionadas con ella aparecen por orden decreciente en los sucesivos elementos de la fila. En la parte inferior de cada fila, para cada variable, se proporciona el valor numérico de dicha relación.

tición ($X500, X1000$) son no lineales (no se ajustan por una recta), sin embargo, se observa una relación importante a través del ajuste de polinomios de grado 2 y 3.

Modelo de regresión lineal con maxima-hunting

Aplicando el método máxima-hunting con tres variables y un modelo de regresión lineal múltiple para ambas distancias obtenemos los siguientes resultados recogidos en la tabla 4.7. La elección del número de tres covariables ha sido determinada de manera arbitraria, eligiendo un compromiso razonable entre la cantidad escasa de datos disponibles y el número de covariables.

	Variables Dependientes:	
	X500	X1000
Variables explicativas:	(1)	(2)
VT1V.E.V.CO2	-0.398^{**}	
VT3V.O3.HR	-0.239^{**}	
X1000	0.223^{***}	
VT2RER		-12.710
VT3RER		-12.532
X500		2.286^{***}
Término independiente	37.706^{***}	17.026
Observaciones	18	18
R^2	0.849	0.769
R^2 ajustado	0.817	0.720
Desviación típica residuos (df = 14)	1.406	4.299
F test (df = 3; 14)	26.329^{***}	15.540^{***}
Nota:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$	

Tabla 4.7: Resultados de los modelos de regresión lineales. En la tabla se muestran los estadísticos asociados a cada uno de los dos modelos de regresión, así como los coeficientes de los modelos ajustados.

Los resultados son bastante aceptables, dado el escaso número de variables que utilizamos para predecir el rendimiento: cometemos un error medio de 0.95 segundos en la prueba de 500 m, mientras que un error de 2.6 segundos en la prueba de 1000 m.

En cuanto a los coeficientes, observamos que aparecen en ambas distancias el tiempo en la otra prueba sobre la que tenemos información y, además, es interesante señalar que en la prueba de mayor duración ($X1000m$) las otras dos covariables presentes están vinculadas a la tasa de

intercambio respiratorio ($VT2RER, VT3RER$), una medida asociada a la capacidad del cuerpo de como consume energía, obteniendo coeficientes negativos, lo que indica que el tiempo es menor en cuanto más eficiente sea el deportista. Es importante comentar que en la prueba de 1000 m entra en juego el metabolismo aeróbico de manera bastante destacable.

Capítulo 5

Conclusiones

En este trabajo se han aplicado variadas técnicas estadísticas para la resolución de diversos problemas del ámbito deportivo. Algunos de dichos procedimientos es la primera vez que se aplican en el mundo del deporte. Por ejemplo, es el caso de problemas encuadrados dentro de la estadística de alta dimensión como los componentes principales tipo sparse, medidas de dependencia multivariante como la distancia de correlación en lugar del coeficiente de correlación de Pearson, el método de selección de variables maxima-hunting o los modelos de regresión funcionales.

Los resultados obtenidos nos indican que es necesaria la utilización de técnicas estadísticas avanzadas, algo no realizado hasta el momento y que, además, como hemos visto, muchas de estas técnicas aparecen de forma casi natural como es el caso de los datos funcionales.

En líneas generales, en este trabajo se han encontrado resultados científicos valiosos:

- Se ha demostrado la validez de la utilización de técnicas de datos funcionales como un instrumento adecuado para establecer tests que predigan la forma física de manera indirecta, además del enorme potencial que puede tener su utilización en múltiples problemas deportivos. De hecho, quizás el futuro del entrenamiento deportivo dependa precisamente de la utilización de este abanico de técnicas.

- Se han establecido las relaciones de dependencia entre variables mecánicas de rendimiento y los parámetros fisiológicos, algo apenas sin estudiar (lo que sí esta más estudiado es en relación a los resultados de competición, pero no entre los parámetros predictivos entre sí). Los pocos casos estudiados con estas técnicas se tratan exclusivamente con medidas de dependencia lineal, lo cual también hemos visto que es un procedimiento inadecuado con este tipo de datos.

- Se ha creado una herramienta para predecir el consumo de oxígeno a través de la potencia y la frecuencia cardíaca, algo muy interesante para los gadgets deportivos actuales y, sin duda, mucho más preciso que las herramientas implementadas hasta el momento y necesarias, por ejemplo, en las aplicaciones de bajada de peso, pues a partir del consumo de oxígeno se estima la energía gastada con el ejercicio.

A la vista de los resultados alcanzados, creemos que podemos ser optimistas acerca del futuro que puede tener la inclusión de metodologías estadísticas dentro de la investigación y entrenamiento deportivo.

Bibliografía

[1] Åstrand, P.O., Rodahl, K., (1980). Fisiología del trabajo físico: bases fisiológicas del ejercicio. Buenos Aires. Editorial Médica Panamericana.

[2] Banister, E. W., Calvert, T. W., Savage, M. V., Bach, T. (1975). A systems model of training for athletic performance. Australian Journal of Sports Medicine, 7(3), 57-61.

[3] Berrendero, J. R., Cuevas, A., Torrecilla, J. L. (2013). Variable selection in functional data classification: a maxima-hunting proposal. arXiv preprint arXiv:1309.6697.

[4] Breiman, L., Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80(391), 580-598.

[5] Conconi, F., Grazzi, G., Casoni, I., Guglielmini, C., Borsetto, C., Ballarin, E., Mazzoni, G., Patracchini, M., anfredini, F. (1996). The Conconi test: Methology after 12 years of application. International Journal of Sports Medicine, 17, 509-519.

[6] Davis, J. A., Frank, M. H., Whipp, B. J., Wasserman, K. (1979). Anaerobic threshold alterations caused by endurance training in middle-aged men. Journal of Applied Physiology, 46(6), 1039-1046.

[7] Delaigle, A., Hall, P. (2010). Defining probability density for a distribution of random functions. The Annals of Statistics, 1171-1193.

[8] Farrel, P.A., Wilmore, J.H., Coyle, E.F., Billings, J.E., Costill, D.L. (1975). Plasma lactate accumulation and distance running performance. Medicine & Science in Sports, 11: 338-344.

[9] Ferraty, F., Vieu, P. (2006). Nonparametric functional data analysis: theory and practice. Springer Science & Business Media.

[10] Fister, I., Ljubi, K., Suganthan, P. N., Perc, M. (2015). Computational intelligence in sports: challenges and opportunities within a new research domain. Applied Mathematics and Computation, 262, 178-186.

[11] Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

[12] García, G. C., Secchi, J. D. (2014). Test course navette de 20 metros con etapas de un minuto. Una idea original que perdura hace 30 años. Apunts. Medicina de lÉsport, 49(183), 93-103.

[13] Grenander, U. (1950). Stochastic processes and statistical inference. Arkiv för matematik, 1(3), 195-277.

[14] Hall, P., Wolff, R. C., Yao, Q. (1999). Methods for estimating a conditional distribution function. Journal of the American Statistical Association, 94(445), 154-163

[15] Harrison, A. J., Ryan, W., Hayes, K. (2007). Functional data analysis of joint coordination in the development of vertical jump performance. Sports Biomechanics, 6(2), 199-214

[16] Hastie, T. J., Tibshirani, R. J. (1990). Generalized additive models (Vol. 43). CRC Press.

[17] Mader, A., Heck, H. (1986). Atheory of the origin of anaerobic threshold. International Journal of Sports Medicine. Suppl. 7: 45-65.

[18] Matabuena, M., Rodríguez, R. (2016). A new approach to predict changes in physical condition: A new extension of the classical Banister model. ArXiv e-prints arXiv:1612.08591.

[19] Müller, H. G., Yao, F. (2012). Functional additive models. Journal of the American Statistical Association, 103 (484), 1534-1544.

[20] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability & Its Applications, 9(1), 141-142.

[21] Newell, J., McMillan, K., Grant, S., McCabe, G. (2006). Using functional data analysis to summarise and interpret lactate curves. Computers in Biology and Medicine, 36(3), 262-275.

[22] Powers, S.K., Dodd, S., Garner, R. (1984): Precision of ventilator and gas exchange alterations as a predictor of the anaerobic threshold. European Journal of Applied Physiology. 52: 173-177

[23] Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47(4), 379-396.

[24] Ramsay, J. O. Silverman, B.W. (2006). Functional data analysis. John Wiley & Sons, Inc..

[25] Ruiz Rivera, D. (2015). Valoración funcional en patinadores de velocidad de lato nivel: determinación de forma indirecta, mediante una prueba de campo, de la Velocidad Aeróbica Máxima patinando. Tesis Doctoral. Universidade da Coruña.

[26] Ruppert, D., Sheather, S. J., Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 90(432), 1257-1270.

[27] Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6), 2769-2794.

[28] Rao, C. R. (1958). Some statistical methods for comparison of growth curves. Biometrics, 14(1), 1-17.

[29] Wang, J. L., Chiou, J. M., Müller, H. G. (2015). Review of functional data analysis. arXiv preprint arXiv:1507.05135.

[30] Wasserman, K., McIlroy, M.B. (1964). Detecting the threshold of anaerobic metabolism in cardiac patients during exercise. American Journal of Cardiology. 14: 844-852

[31] Watson, G. S. (1964). Smooth regression analysis. Sankhya: The Indian Journal of Statistics, Series A, 359-372.

[32] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265-286.