



MODELOS ESTADÍSTICOS PARA LA PREDICCIÓN DE LESIONES MEDIANTE DATOS PROCEDENTES DE GPS

María del Carmen Montañés Palma

Grado en Matemáticas

Universidad de Sevilla
Facultad de Matemáticas
Departamento de Estadística e Investigación Operativa

MODELOS ESTADÍSTICOS PARA LA PREDICCIÓN DE LESIONES MEDIANTE DATOS PROCEDENTES DE GPS

Trabajo realizado por:

María del Carmen Montañés Palma

Tutor:

José María Fernández Ponce

Trabajo de fin grado presentado como requisito para optar al
título del Grado en Matemáticas

Resumen

En este trabajo se exponen modelos estadísticos aplicados a la predicción de lesiones en futbolistas, mediante el uso de la tecnología GPS para obtener los datos necesarios. En el primer capítulo se introducen los conocimientos sobre las lesiones y sobre los dispositivos GPS y su utilidad para el registro de datos y la posterior interpretación de los mismos. Los tres capítulos posteriores se centran en tres modelos matemáticos y sus correspondientes aplicaciones prácticas. En el segundo capítulo se lleva a cabo el Análisis de Componentes Principales, con el objetivo de reducir el número de variables a algunos factores significativos que se correlacionan con las lesiones. En el tercer capítulo se presenta el ANOVA de Medidas Repetidas para estudiar el efecto de la mejora del equilibrio en la reducción del número de lesiones. Por último, en el capítulo cuatro se trata el modelo de Regresión de Cox para estudiar los factores de riesgo que afectan a lesiones en una parte concreta del cuerpo: el muslo.

Abstract

In this final project for graduation are developed some statistical models applied to injury prevention in football players. These models use GPS technology in order to obtain the necessary data. The first chapter introduces some knowledges about injuries and GPS devices, which are useful for data recording and subsequent interpretation. The next three chapters focuses on three mathematical models and their corresponding practical applications. In the second chapter the Principal Components Analysis is carried out with the aim of reducing the number of variables to some significant factors that correlate with injuries. The third chapter presents the ANOVA of Repeated Measures with the goal of studying if the effect of equilibrium's improvement helps to reduce the number of injuries. Finally, chapter four deals with the Cox Regression model to study the risk factors that affect injuries in a specific part of the body: the thigh.

Contenido

1. Introducción	1
1.1. Conceptos previos	1
1.1.1. Ejemplos de registro de lesiones según su definición	3
1.2. Clasificación de lesiones	4
1.3. Implementación de cuestiones	6
1.4. Tecnología GPS	8
1.4.1. Funcionamiento y validez de la tecnología GPS	9
1.4.2. Aplicaciones prácticas de la tecnología GPS	11
2. Análisis de Componentes Principales	15
2.1. Desarrollo teórico	15
2.1.1. Cálculo de las componentes	16
2.1.2. Propiedades de los componentes	18
2.1.3. Interpretación de los componentes	20
2.2. Aplicación práctica	21
2.2.1. Variables	21
2.2.2. Datos	22
2.2.3. Definición del problema	24
2.2.4. Modelo	25
2.2.5. Resultados	25
2.2.6. Otros comentarios sobre CP	27
2.2.7. Conclusiones	32
3. ANOVA de medidas repetidas	34
3.1. Desarrollo teórico	34
3.1.1. Supuestos	34
3.1.2. Diseño de medidas repetidas con un factor intra-sujetos	36
3.1.3. Diseño de medidas repetidas con dos factores intra-sujetos	39
3.1.4. Diseño de medidas repetidas con un factor intra-sujetos y un factor inter-sujetos	45

3.2. Aplicación práctica	52
3.2.1. ¿Cómo afecta la mejora del equilibrio en la reducción del número de lesiones?	53
3.2.2. Material y métodos	53
3.2.3. Análisis	56
4. Modelo de Regresión de Cox	59
4.1. Desarrollo teórico	59
4.1.1. Estimador de Kaplan-Meier	62
4.1.2. Modelo de Cox	64
4.2. Aplicación práctica	68
4.2.1. Métodos	68
4.2.2. Análisis	71
A. Anexo	76
Bibliografía	79

1. Introducción

*“Un hombre con ideas nuevas es un loco,
hasta que estas triunfan”
Marcelo Bielsa (1955-)*

En este capítulo se tratan conceptos básicos sobre las lesiones en futbolistas y sobre los dispositivos GPS. Estos conocimientos son necesario para una mejor recogida de datos y para una mejor interpretación de los mismos, una vez aplicados los modelos estadísticos, que se desarrollaran en posteriores capítulos.

1.1. Conceptos previos

En esta sección se expondrán algunos conceptos básicos sobre las lesiones en el fútbol (ver Fuller y cols. (2006))

Entrenamiento. Un entrenamiento se define como un conjunto de actividades físicas, tanto individuales como en equipo, bajo el control del entrenador y del cuerpo técnico, que están enfocadas a mantener o mejorar las habilidades o la condición física de los futbolistas.

El calentamiento pre-partido y la sesión de enfriamiento post-partido deben registrarse como entrenamiento. Las charlas de motivación del equipo, los debates sobre las tácticas, y las sesiones con psicólogos deportivos, nutricionistas, y cualquier otro tipo que no requieran de actividad física no deben registrarse como entrenamiento. Las actividades de entrenamiento personal llevadas a cabo por los jugadores fuera del equipo y que no son planificadas por el entrenador o el equipo técnico no deben ser registradas como entrenamiento. Cualquier actividad que forma parte de la rehabilitación de una lesión no debe ser registrado como entrenamiento.

Partido. Un partido se define como un juego entre dos equipos de diferentes clubes.

Un partido entre dos equipos del mismo club debe ser considerada como un entrenamiento.

Cualquier partido que forme parte de la rehabilitación de una lesión no debe ser registrado como partido.

Lesión. Definimos una lesión como cualquier daño que sufre un jugador como consecuencia de un partido o un entrenamiento, independientemente de la necesidad de atención médica requerida o el tiempo perdido en las futuras actividades futbolísticas.

La atención médica se refiere a una evaluación de la condición médica de un jugador por un médico cualificado. El hecho de que un jugador no pueda participar plenamente en el entrenamiento o en futuros partidos es independiente de que la sesión de entrenamiento tenga lugar el día después de la lesión o de si el jugador ha sido seleccionado para jugar en el siguiente partido. El término futuro se refiere a cualquier hora después de la aparición de lesiones, incluyendo el día de la lesión. Es importante reconocer que la tolerancia de un jugador al dolor puede crear diferencias en la incidencia de lesiones registradas en los estudios. Si un jugador sufre múltiples lesiones en un mismo evento, se debe registrar como un lesionado con múltiples diagnósticos. Las lesiones que no están relacionados con el fútbol no deben registrarse en estos estudios. Los jugadores también pueden experimentar otros problemas médicos, que pueden ser necesarios en estudios específicos para registrar estos problemas. Sin embargo, su incidencia se debe informar por separado de la incidencia de problemas físicos.

Lesión recurrente. Definimos una lesión recurrente como una lesión del mismo tipo y en el mismo sitio que una lesión previa y que ocurre después de que un jugador regrese a la práctica de fútbol completa después de dicha lesión previa. Una lesión recurrente se clasifica según el instante de aparición en:

- **Recurrencia temprana:** Si la lesión recurrente se da durante los dos meses posteriores al regreso completo a la actividad futbolística.
- **Recurrencia tardía:** Si la lesión se produce entre los 2 y los 12 meses posteriores.
- **Recurrencia demorada:** Si es después de los 12 meses.

Es esencial adoptar el mismo tipo de registro para lesiones recurrentes que para otro tipo de lesiones. Cuando las lesiones recurrentes se registran, se debe indicar el músculo específico o ligamento donde se ha producido, en lugar de la ubicación general de la lesión. Las lesiones sufridas antes del inicio de un estudio también se deben considerar cuando se identifica una lesión recurrente. Lesiones tales como contusiones, conmociones cerebrales

y secuelas que resultan de una lesión anterior no deben registrarse como recurrentes.

Gravedad de la lesión. Definimos la gravedad de la lesión como el número de días transcurridos desde la fecha en la que se produce la lesión hasta la fecha en la que el jugador participa en los entrenamientos y está en condiciones para disputar un partido. El día en que se produce una lesión es el día cero y no se cuenta cuando se determina la gravedad de la lesión. Por lo tanto, si un jugador no puede participar plenamente en el día que se produce la lesión, pero está disponible para la participación completa al día siguiente, el incidente debe ser registrado como una lesión de pérdida de tiempo con una gravedad de cero días. Si un jugador lesionado deja un estudio o un estudio termina antes de que la lesión de un jugador se haya resuelto, el médico responsable debe estimar el número total de días que transcurrirán antes de que el jugador regrese a la participación plena si el jugador hubiera permanecido en el estudio. Si un jugador se retira de la actividad profesional, como consecuencia directa de una lesión sufrida durante un estudio antes de regresar a la plena participación, la lesión se debe informar por separado como una “lesión que termina su carrera”.

1.1.1. Ejemplos de registro de lesiones según su definición

En los estudios se llevará a cabo un registro de las lesiones en base a las definiciones dadas anteriormente. A continuación se exponen algunos ejemplos:

1. Un defensa sufrió una lesión en el muslo durante un partido que requiere 30 días de rehabilitación antes de poder volver a los entrenamientos. El jugador sufrió otra lesión en los isquiotibiales del mismo músculo (misma pierna) 3 semanas más tarde y que requiere otros 50 días de rehabilitación antes poder volver a los entrenamientos. Con respecto a la atención médica el primer incidente debe ser registrado como una lesión (severidad: 30 días); el segundo incidente como una lesión recurrente (severidad: 50 días). Respecto al tiempo perdido, el primer incidente debe ser registrado como una lesión (severidad: 30 días); el segundo incidente como una lesión recurrente (severidad: 50 días).

2. Un portero empezó a sentir una molestia en el hombro y buscó atención médica; la cual no le impidió entrenar normalmente o jugar partidos a pesar de que le causaba un poco de dolor. El fisioterapeuta del equipo recomendó un programa de entrenamiento individual para el portero para evitar agravar la situación. Con respecto a la atención médica, el episodio debe registrarse como una lesión (gravedad: 0 días). Respecto al tiempo perdido,

como el portero fue capaz de realizar los entrenamientos y los partidos con normalidad, no se debe registrar como una lesión con pérdida de tiempo.

3. Un delantero sufrió un esguince de tobillo durante un partido, pero siguió jugando; recibió la atención médica después del partido. El jugador terminó el entrenamiento completo con un vendaje en el tobillo (con un poco de dolor) durante 6 días, pero la lesión se agravó durante el partido. Se requiere 15 días de rehabilitación. Con respecto a la atención médica, el primer incidente debe ser registrado como una lesión (gravedad: 0 días) y cuando ocurrió el segundo incidente, la gravedad de la lesión debe ser reclasificado a 15 días. Respecto al tiempo perdido, el primer incidente no debe ser registrado y el segundo incidente debe ser registrado como una lesión (15 días).

1.2. Clasificación de lesiones

Agrupación principal	Categoría	Equivalente OSICS Ochar (1996) al área del cuerpo
Cabeza y cuello	Cabeza/Cara	N
	Cuello/Columna vertical	H
Miembros superiores	Hombro / Clavícula	S
	Parte superior del brazo	U
	Codo	E
	Antebrazo	R
	Muñeca	W
	Mano / Dedo de la mano / Pulgar	P
Tronco	Esternón / Costillas / Parte superior de la espalda	C,D
	Abdomen	O
	Zona lumbar / Pelvis / Sacro	B,L
Miembros inferiores	Cadera/Ingle	G
	Muslo	T
	Rodilla	K
	Parte inferior de la pierna / Tendón de Aquiles	Q,A
	Tobillo	A
	Pie / Dedo del pie	F

Tabla 1-1.: Principales grupos y categorías para clasificar la localización de una lesión.

Cuando se requiere un diagnóstico de una lesión, un médico cualificado debe proporcionar un diagnóstico específico por escrito o utilizar un sistema de codificación de lesiones específica del deporte, tales como el sistema Orchard (véase Ochar (1996)). Las lesiones

se pueden clasificar de diferentes modos.

Según la ubicación. La ubicación de las lesiones debe ser registrada utilizando las categorías de ubicación individuales indicadas en la Tabla 1.1. En un estudio donde el número de lesiones es pequeño, puede ser necesario agrupar las categorías de ubicación individuales en la categoría de principales grupos con fines de análisis.

El tipo de lesión. El tipo de lesión debe ser registrado usando las categorías enumeradas en la Tabla 1.2. En un estudio donde el número de lesiones es pequeño, puede ser necesario agrupar las categorías de ubicación individuales en la categoría de principales grupos con fines de análisis. Si el propósito de un estudio es investigar otros problemas médicos, estos deben ser registrados en encabezados separados de los que se enumeran en la Tabla 1.2.

Agrupación principal	Categoría	Equivalente OSICS Ochar (1996) del carácter de la patología
Fracturas y tensión ósea	Fractura	F
	Otras lesiones óseas	G,Q,S
Articulación (no óseo) y ligamento	Dislocación / Subluxación	D,U
	Esguince / Lesión del ligamento	J,L
	Lesión de menisco o cartílago	C
Músculo y tendón	Rotura del músculo / Desgarro / Deformación / Calambres	M,Y
	Lesión en el tendón / Ruptura / Tendinosis / Bursitis	T,R
	Hematoma / Contusión / Moretón	H
	Abrasión	K
Desgarre y lesión de piel	Laceración	K
	Conmoción cerebral (con o sin pérdida de la conciencia)	N
Sistema nervioso central/periférico	Lesión nerviosa	N
	Lesiones dentales	G
Otros	Otras lesiones	

Tabla 1-2.: Principales grupos y categorías para la clasificación del tipo de lesiones

Otras cuestiones sobre la clasificación de lesiones. Las lesiones también deben clasificarse en cuanto a si se produjo durante una sesión de entrenamiento o partido, y si

fueron el resultado del contacto con otro jugador u otro objeto. También puede ser apropiado registrar si la acción causante de la lesión es una violación del reglamento.

Estudio de población La población de estudio normalmente debe consistir en más de un equipo de jugadores y el estudio debe durar un período mínimo de una temporada (incluyendo pretemporada), 12 meses, o la duración de un torneo. Los jugadores que se unen/abandonan el estudio deben ser incluidos/excluidos desde la fecha de incorporación/salida. Los jugadores que tienen una lesión existente en el inicio de un estudio no deben ser excluidos del estudio, pero dichas lesiones no se incluirán como parte del estudio.

1.3. Implementación de cuestiones

Los formularios de recogida de datos estandarizados, que pueden ser en formato electrónico o papel, se deben mantener en el mismo formato a lo largo del estudio.

Formulario de datos “básicos” del jugador. Dependerá del propósito principal del estudio. Se requerirán parámetros como la edad del jugador, el género, su posición en el campo, la estatura y la masa corporal. Para otros estudios tales como las intervenciones e investigaciones de las relaciones entre los factores de riesgo y la incidencia de las lesiones, se necesitará más información de base.

Formulario de lesiones. Debe incluir los números de referencia del equipo y los jugadores, la fecha de la lesión, si la lesión ha sido durante un partido o un entrenamiento, información describiendo las circunstancias que rodean a la lesión, y la fecha que el futbolista vuelve por completo a entrenamientos y partidos. La naturaleza de la lesión (localización, tipo, parte lesionada, recurrencia) también debe ser incluida. La información sobre lesiones debe actualizarse si se obtiene más información en una fecha posterior.

Formulario de entrenamientos y partidos. El formulario debe registrar la fecha, tipo y duración (en minutos) de cada entrenamiento/partido. Estos datos pueden ser recogidos para cada jugador o para un grupo de jugadores. La recogida de los datos individuales de los jugadores, especialmente para las sesiones de entrenamiento, implica mucho tiempo, y este nivel de detalle sólo debe ser registrado si la información es requerida específicamente para satisfacer los objetivos del estudio. Sin embargo, para los estudios donde los jugadores están expuestos en varias de situaciones, como club y la selección, o cuando la intención es investigar las relaciones entre la incidencia de lesiones y los factores de riesgo individuales,

los datos reales de cada jugador si deben ser registrados. Si se recogen datos sobre un grupo, el número de partidos y sesiones de formación, la duración de cada sesión de entrenamiento, y el número de jugadores que asisten a cada sesión de entrenamiento deben ser registrados. Los datos de los entrenamientos deben recogerse preferentemente para cada sesión de entrenamiento, pero, como mínimo, la información debe ser recogida semanalmente.

Para los estudios de registro de datos de partido por equipos, el tiempo de exposición total de un jugador en horas para un equipo está dado por:

$$T = \frac{N_M \cdot P_M \cdot D_M}{60},$$

donde N_M es el número de partidos jugados por el equipo, P_M es el número de jugadores en el equipo (normalmente 11), y D_M es la duración (en minutos) de los partidos (normalmente 90). La recopilación de datos por equipo no tiene en cuenta el tiempo perdido porque los jugadores están temporal o permanentemente ausentes durante un partido, por ejemplo, cuando están siendo atendidos o cuando les sacan una tarjeta roja. Para los estudios de recogida de datos de entrenamiento, el tiempo total del entrenamiento en horas está dado por:

$$T_{TOT} = \sum_{\forall T} \frac{P_T \cdot D_T}{60},$$

por cada sesión de entrenamiento durante todo el estudio, donde P_T es el número de jugadores que asisten a una sesión de entrenamiento y D_T es la duración de la sesión de entrenamiento en minutos. La recopilación de datos basados en un equipo en lugar de los datos individuales no tiene en cuenta el tiempo perdido en el entrenamiento (cuando un jugador no completa una sesión de entrenamiento completo).

Otras cuestiones sobre implementación. Sea cual sea el método de recoger datos de la lesión que se adopte, es esencial que se utilice consistentemente para todos los aspectos del estudio. El formato y contenido de cada estudio deben ser aprobados y deben seguir los protocolos definidos por un comité de ética institucional apropiado. Se deben asignar códigos de referencia individuales a cada jugador, que son conocidos sólo por el personal de recogida de información con el fin de mantener la confidencialidad de los jugadores cuando se analizan los datos. Los jugadores que participan en un estudio deben ser informados del propósito y el formato del estudio. Los jugadores deben conservar siempre el derecho de no participar y retirarse de un estudio en cualquier momento. Es beneficioso tener a una persona de contacto en cada equipo, que no tiene que ser el propio jugador. Se debe mantener un contacto regular con dicha persona para compartir información relevante a lo largo del estudio.

1.4. Tecnología GPS

A continuación se introducirán algunas nociones básicas sobre la tecnología “Global Positional System” (GPS) y su aplicación a los deportes de equipo (veáse Twist, Waldron, Worsfold, y Gabbett (2013)). La tecnología GPS (Figura 1-1) es un sistema de posicionamiento que mediante la tecnología disponible de satélites construye la ubicación geográfica y los movimientos de un individuo. Una integración adicional de varios Sistemas Mecánicos Micro Electromecánicos (MEMS) y la telemetría de la frecuencia cardíaca permiten a los expertos cuantificar la carga interna y externa a la que están expuestas los jugadores de deportes de equipo (en concreto nos centraremos en los jugadores de fútbol) durante el entrenamiento y la competición.



Figura 1-1.: Dispositivos GPS.

La validez y fiabilidad de los sistemas de GPS al evaluar el movimiento de los futbolistas están influenciados por varios factores, entre ellos la velocidad de movimiento, la distancia recorrida, la trayectoria del movimiento individual y la recepción de GPS debido al ambiente.

Los sistemas GPS proporcionan una serie de datos que se analizarán con el objetivo de beneficiar al equipo en ámbitos como la prevención de lesiones o la estrategia de cambios durante los partidos, como se explica más adelante.

A la hora de analizar los datos hay que tener en cuenta que el contacto físico aumenta las demandas fisiológicas y la fatiga de los jugadores, además de la duración e intensidad del trabajo y los períodos de descanso. Además hay que observar que con el fin de preparar al futbolista para la competición, el tipo, la intensidad, el volumen y la frecuencia de la formación dependerán de la fase del año. Con esta información se puede gestionar

la fatiga puntual y a largo plazo, minimizar lesiones y optimizar la recuperación de un jugador. En el caso de un equipo de fútbol profesional, este puede jugar unos 38-64 partidos por temporada, por lo general con uno o dos partidos por semana compaginados con entrenamientos regulares que suelen consistir en el acondicionamiento metabólico, el entrenamiento de resistencia, habilidades y elementos tácticos.

En la actualidad, todos los equipos disponen de la tecnología GPS. Cada uno de ellos trabaja con diferentes marcas, entre ellas destacan GPSPORTS, que trabaja con equipos como el Real Madrid, Atlético de Madrid o Sevilla, y “RealTrack Systems Wimbu, la cual tiene un acuerdo con LaLiga que ofrece a los equipos informes con datos como las distancias grabadas a alta demanda metabólica (HMLD), valores de las acciones cinemáticas de alta densidad (KHIA), potencia metabólica, distancia recorrida a alta velocidad (HSR), distancia recorrida de alta intensidad (HIBD), aceleraciones/desaceleraciones de alta y media intensidad y la distancia explosiva entre otros.

1.4.1. Funcionamiento y validez de la tecnología GPS

Los dispositivos GPS se fabrican actualmente con frecuencias de muestreo de 1, 5 y 10 Hz (la velocidad a la que la unidad recopila datos). La literatura sugiere que el GPS con una tasa de frecuencia más alta proporciona una mayor validez para la medición de la distancia. Al comparar la precisión de la adquisición de distancias entre un GPS de 1 y 5 Hz, el error estándar de un sprint de 10 m fue de 32.4 y 30.9 %, respectivamente (véase Jennings, Cormack, Coutts, y cols. (2010)). Por el contrario, un GPS de 10 Hz mostró un error estándar de 10.9 % en un sprint de 15 m (véase Aughey (2011)). Recientemente, se ha conocido que los dispositivos GPS a 1 Hz pueden no ser capaces de registrar movimientos que tardan menos de un segundo en completarse (véase Johnston, Watsford, Pine, y cols (2012)). Las unidades más nuevas de 10 Hz son capaces de medir el cambio más pequeño que vale la pena en la aceleración y desaceleración, mientras que la unidad de 5 Hz no puede hacerlo (véase Varley y Aughey (2012)). Los mayores errores que se producen en la medición de la distancia con los dispositivos GPS de 1 y 5 Hz frente a los de 10 Hz indican que la frecuencia de muestreo puede limitar la precisión de las mediciones de distancia y la velocidad.

La velocidad de un movimiento afecta la validez de la distancia medida por GPS. La primera validación de un dispositivo GPS mostró varias velocidades de marcha y carrera (2-20 km \cdot h₋₁) que estaban altamente correlacionadas ($r= 0,99$) con un cronómetro

(véase Schutz y Chambaz (1997)). Un estudio más reciente (véase Portas, Rush, Barnes, y cols (2009)) muestra que el error de medición de la distancia del GPS es más bajo durante la marcha (aprox. $1.8 ms_{-1}$; error estándar de estimación [SEE] 0.7 %) y más alto durante la carrera (aprox. $6 ms_{-1}$; VER 5.6 %). Del mismo modo otros estudios (véase Johnston y cols. (2012)) informaron que el GPS es capaz de medir patrones de velocidad de trabajo realizados a velocidades menores de $20 km \cdot h_{-1}$; sin embargo, se recomienda precaución al analizar los patrones de velocidad de trabajo a velocidades mayores de $20 km \cdot h_{-1}$. Estos resultados indican que la velocidad de movimiento afecta a la precisión.

La validez de las medidas de distancia mejora con actividades de mayor duración (véase Jennings y cols. (2010)); por ejemplo, el coeficiente de variación (CV) disminuyó de 32.4 a 9.0 % para distancias de sprint de 10 y 40 m, respectivamente. El CV se redujo aun más al 3.8 % para un rango de velocidades completado en un circuito de carrera de deportes de equipo modificado de 140 m (véase Jennings y cols. (2010)).

Los factores de frecuencia de muestreo, distancia y velocidad, que afectan la validez del GPS, también afectan la exactitud del GPS. El impacto de la frecuencia de muestreo aun no está claro; por ejemplo, el CV de una tarea de fútbol lineal se ha mostrado como del 4.4 a 4.5 % para un GPS de 1 Hz y del 4.6 a 5.3 % para un 5 Hz (véase Portas, Rush, Barnes, y cols (2010)). Sin embargo, otro estudio (véase Jennings y cols. (2010)) informó el CV de un sprint de 10 m como del 77 y 39 % para 1 y 5 Hz, respectivamente. Más recientemente, una tasa de muestreo más alta de 10 Hz ha demostrado una mejora en la exactitud durante la velocidad constante y la fase de aceleración o desaceleración (CV menor al 5.3 y 6 %, respectivamente) (véase Varley y Aughey (2012)). Si bien los datos son actualmente ambiguos y pueden explicarse a través del uso de diferentes fabricantes y modelos de GPS (véase Aughey (2011)), parece que un aumento de la frecuencia de muestreo parece mejorar la exactitud del GPS.

La exactitud del GPS disminuye con la mayor velocidad de movimiento. El CV de paseo a pie para un GPS de 5 Hz fue del 1.4–2.6 %, mientras que el CV de carrera a una distancia de 20 m fue del 19.7–30 %. De manera similar, se anotaron los CV del 30.8 y 77.2 %, respectivamente, para caminar y correr a una distancia de 10 m con un GPS de 1 Hz (véase Jennings y cols. (2010)). La exactitud de los dispositivos GPS también se ve afectada negativamente por los movimientos que requieren cambios en la dirección. El CV para cambios graduales y ajustados de movimientos de dirección al caminar se ha mostrado como el 11.5 y 15.2 %, respectivamente (véase Jennings y cols. (2010)).

En general, los estudios concluyen que los dispositivos GPS tienen un nivel aceptable de validez y exactitud para evaluar los patrones de movimiento a velocidades más bajas y a mayores esfuerzos de distancia. La menor exactitud de las unidades de GPS para medir con precisión los patrones de movimiento durante carreras de línea recta de alta velocidad y corta duración, y los esfuerzos que requieren cambios en la dirección pueden limitar la precisión para evaluar estos aspectos en los deportes de equipo. Sin embargo, las unidades de GPS con mayor frecuencia de muestreo demuestran mejoras en la exactitud y en la validez y pueden utilizarse en el monitoreo de la actividad física en situaciones como los deportes de equipo, siempre que se tome precaución al interpretar los sprints individuales y los cambios rápidos en dirección y velocidad.

1.4.2. Aplicaciones prácticas de la tecnología GPS

Uso del GPS para evaluar los patrones de movimiento de los equipos

La tecnología GPS se puede utilizar para cuantificar el movimiento de los futbolistas tanto en el entrenamiento como en la competición. Si bien el uso de dispositivos GPS portátiles en “deportes de colisión” como el rugby es ahora una práctica común, su aplicación en el ámbito del fútbol profesional sigue estando limitada debido a las preocupaciones sobre la seguridad de los jugadores. Dichas preocupaciones se relacionan con el tamaño físico de los dispositivos GPS portátiles y la posición en la que se colocan al jugador (columna torácica), como se puede ver en la siguiente imagen (Figura 1-2).



Figura 1-2.: Dispositivos GPS.

Con la aparición de la tecnología GPS se ha mejorado la comprensión de las exigencias asociadas a los deportes de equipo, en particular del fútbol, y a aumentar la información de las actividades diarias de los deportistas.

La aplicación de la tecnología GPS en los deportes de equipo también ha manifestado los diferentes patrones de carrera observados entre los distintos deportes. Por ejemplo, un informe de un partido de fútbol amistoso demostró que los jugadores recorren aproximadamente 10-11 km (aprox. $111-122 \text{ m} \cdot \text{min}^{-1}$), de los cuales aproximadamente 2,5 km ($27,7 \text{ m} \cdot \text{min}^{-1}$) se cubren a alta intensidad. Si bien las diferencias de posición no se han evaluado, en el fútbol para adultos, utilizando el GPS, hay informes de partidos juveniles (16-18 años) que han demostrado que los jugadores de medio campo cubren hasta 8,6 km ($107,5 \text{ m} \cdot \text{min}^{-1}$), con aproximadamente 1,1 km cubiertos a alta intensidad. Lo que es más que lo que recorren los defensas centrales (7,6 km) y los delanteros (7,8 km), que también realizaron carreras de menor intensidad (aprox. 0.7 km). Los estudios también señalan que los jugadores adultos de fútbol australiano son los que mayor distancia recorren de todos los deportes de equipo estudiados (12,3 km), lo que equivale a $124 \text{ m} \cdot \text{min}^{-1}$, que es similar a los jugadores adultos de fútbol. Se ha especulado que el mayor grado de actividad de los movimientos en el fútbol y el fútbol australiano es una consecuencia de la naturaleza abierta de estos deportes, en contraposición el rugby, donde las leyes de fuera de juego sirven para restringir el espacio de paso disponible. El mismo razonamiento se puede explicar a las diferencias de posición, con lo que los jugadores nómadas (fútbol australiano) y los mediocentros (fútbol) poseen mayor libertad de movimiento por el campo de juego.

Es importante señalar la necesidad de llegar a un consenso con respecto a la clasificación de movimiento de alta intensidad para permitir una verdadera comparación entre los estudios.

Sprints repetidos y esfuerzos repetidos en deportes de equipo

El análisis de la actividad de “sprint repetido” y “esfuerzo repetido”, mediante GPS durante las competencias, ha mejorado la comprensión del rendimiento deportivo en equipo. La actividad repetida del sprint se define a menudo como la ocurrencia de tres o más sprints, separados por menos de 21 segundos de ejercicio de intensidad moderada a baja. Esfuerzos repetidos difieren de los análisis tradicionales de sprints repetidos, al utilizar la integración de GPS portátil con sistemas microelectromecánicos (MEMS). En

el entrenamiento, los acelerómetros triaxiales integrados, encargados de medir la aceleración y vibración en múltiples sectores, se usan para generar una medida conocida como “carga de cuerpo”, que se deriva de la raíz cuadrada de la suma de la tasa instantánea al cuadrado del cambio en la aceleración en cada uno de los tres ejes ortogonales (anterior-posterior [x], medio-lateral [y] y vertical [z]). Los métodos que capturan los casos de aceleración/deceleración utilizando el acelerómetro integrado en los GPS deben ser considerados como indicadores útiles de esfuerzo durante la actividad del deporte en equipo.

En deportes donde las colisiones son menos frecuentes, como es el caso del fútbol donde de hecho están penalizadas, la actividad de sprint repetida ha surgido como un indicador útil de rendimiento. El análisis de GPS en el fútbol juvenil ha informado que los jugadores ofensivos son los que realizan más sprints repetidos (aprox. 14) por partido y los defensas centrales (jugadores defensivos) los que menos (aprox. 5). Durante un episodio de seis sprints repetidos, la duración promedio de un sprint se fue mayor durante los partidos internacionales (2,32 s) en comparación con los partidos nacionales (1,83 s), con un período de recuperación más corto de 11.9 s en partidos internacionales en comparación con 14.6 s en partidos nacionales.

Información sobre prácticas de entrenamiento usando datos de GPS de la competencia

Los resultados de los estudios que utilizan dispositivos GPS para monitorear a los jugadores de deportes de equipo durante la competición han ayudado a proporcionar directrices más claras para trabajar los entrenamientos. Recientemente, los datos de GPS se han utilizado para diseñar simulaciones específicas de partidos que puede ser utilizado tanto para fines de acondicionamiento y como de monitorización. Por otra parte, el rendimiento de los jugadores durante escenarios de entrenamiento se puede controlar utilizando los mismos dispositivos GPS utilizados durante la competición; algo que no se puede hacer cuando se utilizan otros modos de análisis de movimiento, tales como sistemas de cámaras múltiples.

Monitoreo de la fatiga transitoria usando dispositivos GPS en el deporte de equipo competencia

La fatiga, en el caso de un partido, se considera tradicionalmente como la disminución de intensidad del movimiento en función del tiempo del partido.

La fatiga transitoria se refiere a los períodos de intensidad significativamente más baja que siguen inmediatamente a un período de máxima intensidad. A menudo se utiliza la tecnología GPS para medir la capacidad en tiempo real, es posible que esta información pueda ayudar en la estrategia de cambios durante los partidos.

Tecnología GPS aplicada a las lesiones.

Varios estudios han investigado la influencia del volumen de entrenamiento, la intensidad y la frecuencia en el rendimiento de los deportistas, mejorando el rendimiento, generalmente, mejorando con aumentando la carga de entrenamiento (véase Foster, Daines, Hector, Snyder, y Welsh (1996), Stewart y Hopkins (2000)). Sin embargo, también se ha demostrado que una mala actividad en el entrenamiento produce una mayor incidencia de lesiones que ocurren cuando las cargas de entrenamiento son más altas (véase Foster (1998), Gabbett (2004a)).

Estudios previos de atletas de deportes de equipo han mostrado relaciones positivas significativas entre las cargas de entrenamiento y la cantidad de lesiones de entrenamiento (véase JGabbett y Jenkins (2011), Gabbett (2004a)), lo que sugiere que cuanto más se entrene, más lesiones se sufrirán. Además, se ha demostrado que las reducciones en las cargas de entrenamiento reducen la cantidad de lesiones de entrenamiento y dan como resultado mayores mejoras en la potencia aeróbica máxima (véase Gabbett (2004b)). Claramente, el entrenamiento para el fútbol refleja un equilibrio entre la carga mínima de entrenamiento requerida para obtener una mejora en la condición física y la carga máxima de entrenamiento tolerable antes de incrementar la cantidad de lesiones. Una proporción considerable de las lesiones sufridas por los deportistas de deportes de equipo sin contacto, en particular de los futbolistas, son problemas de los tejidos blandos que se producen como resultado de cargas de entrenamiento excesivas, recuperación inadecuada y sobreentrenamiento (véase Gabbett (2003), Gabbett (2005), Gabbett y Domrow (2007)). Estas lesiones, que en gran medida se pueden prevenir, pueden afectar negativamente a los equipos. La tecnología GPS tiene aplicaciones para los entrenadores encargados del acondicionamiento físico y que son responsables de diseñar y entregar periódicamente programas de entrenamiento. Los entrenadores pueden conocer rápidamente la carga de entrenamiento de los jugadores, y así saber cuáles pueden ser susceptibles de sufrir lesiones o enfermedades debido al sobreentrenamiento. De hecho, los entrenadores a menudo usan datos de GPS para restringir la cantidad de alta intensidad en carrera de los deportistas que presentes en una sesión de entrenamiento determinada o en varias sesiones de entrenamiento.

2. Análisis de Componentes Principales

“Cada derrota es una victoria en sí misma”

Francisco Maturana (1949-)

En este capítulo se desarrolla el Análisis de Componentes Principales. Posteriormente se aplica a unos ciertos datos con el objetivo de reducir el número de variables a algunos factores significativos que se correlacionan con las lesiones.

2.1. Desarrollo teórico

El análisis de componentes principales tiene como objetivo reducir la dimensionalidad, es decir, conseguir explicar un conjunto p de variables usando un número menor de ellas. Dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables, construidas como combinaciones lineales de las originales. Las nuevas variables que sustituyen a las originales tienen que tener máxima correlación con las variables originales.

Este método es de gran utilidad, por una parte permite representar óptimamente en un espacio de dimensión pequeña observaciones de un espacio general p -dimensional. Por otra parte, permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

Es importante señalar que en el análisis de componentes principales se trabaja con variables tipificadas. Por lo tanto, en lo que sigue, se supone que a cada variable previamente se le ha restado su media y dividido por su desviación típica.

2.1.1. Cálculo de las componentes

Cálculo de la primera componente

Sea

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

la matriz de datos multivariantes donde las filas representan los individuos y las columnas las p variables continuas analizadas. La primera componente principal es una combinación lineal de las variables originales y viene dada por:

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1.$$

\mathbf{z}_1 tiene media 0 porque las variables originales también tienen media 0. Su varianza vendrá dada por:

$$\frac{1}{n}\mathbf{z}'_1\mathbf{z}_1 = \frac{1}{n}\mathbf{a}'_1\mathbf{X}'\mathbf{X}\mathbf{a}_1 = \mathbf{a}'_1\mathbf{S}\mathbf{a}_1, \quad (2-1)$$

donde \mathbf{S} es la matriz de varianzas y covarianzas de las observaciones.

Se supone $\mathbf{a}'_1\mathbf{a}_1 = 1$ y usando el multiplicador de Lagrange quedaría:

$$M = \mathbf{a}'_1\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a}'_1\mathbf{a}_1 - 1).$$

A continuación se maximiza esta expresión. Para ello se deriva respecto a la componente \mathbf{a}_1 y posteriormente se iguala a 0 de la siguiente forma:

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0,$$

siendo su solución:

$$\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}_1. \quad (2-2)$$

Por tanto \mathbf{a}_1 es un vector propio de la matriz \mathbf{S} , y λ su correspondiente valor propio. Seguidamente se determina qué valor propio de \mathbf{S} es la solución de (2-2), para ello se usa la siguiente ecuación,

$$\mathbf{a}'_1\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}'_1\mathbf{a}_1 = \lambda,$$

y se concluye, por (2-1), que λ es la varianza de \mathbf{z}_1 . Como esta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz \mathbf{S} . Su vector asociado, \mathbf{a}_1 , define los coeficientes de cada variable en la primera componente principal.

Cálculo de la segunda componente

En este caso la cuestión es conseguir el mejor plano de proyección \mathbf{X} . Se calcula estableciendo como función objetivo que la suma de las varianzas de $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$ y $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$ sea máxima, donde \mathbf{a}_1 y \mathbf{a}_2 son los vectores que definen el plano. La función objetivo será:

$$\phi = \mathbf{a}'_1 \mathbf{S} \mathbf{a}_1 + \mathbf{a}'_2 \mathbf{S} \mathbf{a}_2 - \lambda_1 (\mathbf{a}'_1 \mathbf{a}_1 - 1) - \lambda_2 (\mathbf{a}'_2 \mathbf{a}_2 - 1) \quad (2-3)$$

que tiene las restricciones de que las direcciones deben de tener módulo unitario $(\mathbf{a}'_i \mathbf{a}_i) = 1$, $i = 1, 2$. Derivando e igualando a cero:

$$\begin{aligned} \frac{\partial \phi}{\partial \mathbf{a}_1} &= 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0 \\ \frac{\partial \phi}{\partial \mathbf{a}_2} &= 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 = 0 \end{aligned}$$

La solución de este sistema es:

$$\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1,$$

$$\mathbf{S}\mathbf{a}_2 = \lambda_2 \mathbf{a}_2$$

que implica que \mathbf{a}_1 y \mathbf{a}_2 deben ser vectores propios de \mathbf{S} . Sustituyendo en (2-3) los vectores propios de norma uno, se tiene que, en el máximo, la función objetivo es

$$\phi = \lambda_1 + \lambda_2.$$

Además λ_1 y λ_2 son los dos autovalores mayores de la matriz \mathbf{S} y \mathbf{a}_1 y \mathbf{a}_2 sus respectivos autovectores. Se observa que la covarianza entre \mathbf{z}_1 y \mathbf{z}_2 , dada por $\mathbf{a}'_1 \mathbf{S} \mathbf{a}_2$ es cero debido a que $\mathbf{a}'_1 \mathbf{a}_2 = 0$, y las variables \mathbf{z}_1 y \mathbf{z}_2 están incorreladas.

Generalización

Los r mayores valores propios de \mathbf{S} definen al espacio de dimensión r que mejor representa a los puntos. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables por ellas definidas componentes principales.

Se calculan los valores propios o raíces características $\lambda_1, \dots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables, \mathbf{S} , mediante:

$$|\mathbf{S} - \lambda\mathbf{I}| = 0,$$

se tiene que existen tantas componentes principales como variables, ya que generalmente, la matriz \mathbf{X} (y por tanto la \mathbf{S}) tiene rango p .

Los vectores asociados a $\lambda_1, \dots, \lambda_p$ verifican:

$$(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{a}_i = 0$$

Como \mathbf{S} es simétrica, los términos λ_i son reales, y positivos por ser \mathbf{S} definida positiva. Como \mathbf{S} es simétrica los vectores asociados a λ_j y λ_h , raíces distintas, son ortogonales. En el caso en el que \mathbf{S} sea semidefinida positiva de rango $r < p$, lo que ocurre si $p - r$ variables son combinación lineal de las demás, habrá solamente r raíces características positivas y el resto serán ceros.

Las nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z} = \mathbf{XA},$$

siendo \mathbf{Z} la matriz cuyas columnas son los valores de los p componentes en los n individuos, y donde $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Se calculan los componentes principales aplicando una transformación ortogonal \mathbf{A} a las variables \mathbf{X} (ejes originales) para obtener unas nuevas variables \mathbf{Z} incorreladas entre sí.

2.1.2. Propiedades de los componentes

Las componentes principales son nuevas variables con las siguientes propiedades:

1. La suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
2. El cociente entre la varianza de un componente, el valor propio asociado a su vector propio, y la suma de los valores propios de la matriz da la proporción de la variabilidad explicada.
3. Sea \mathbf{a}_i es el vector de coeficientes de la componente \mathbf{z}_i , las covarianzas entre cada componente principal y las variables X vienen dadas por:

$$Cov[z_i; x_1, \dots, x_p] = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

4. Las correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
5. Las r componentes principales proporcionan una predicción exacta de las r variables que las conforman.
6. La estandarización multivariante de los datos originales se obtiene estandarizando las componentes principales.
7. Las componentes principales no están correlacionadas, es decir, las direcciones de las componentes principales son ortogonales (véase Figura 2-1).

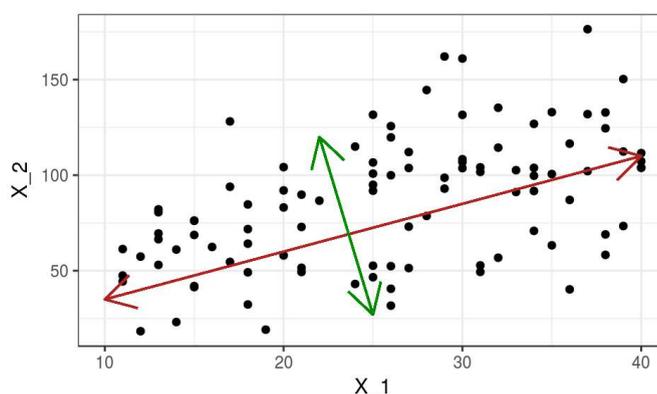


Figura 2-1.: Representación de dos componentes principales.

2.1.3. Interpretación de los componentes

Cuando la primera componente principal tiene todas sus coordenadas positivas significa que existe una alta correlación positiva entre todas las variables originales. En este caso, la primera componente principal puede interpretarse como un promedio ponderado de todas las variables. El resto de componentes, por lo general, tienen coordenadas positivas y negativas y esto implica que el conjunto de variables originales se dividan en grupos. En este caso los factores pueden escribirse como medias ponderadas de dos grupos de distinto signo.

Selección del número de componentes

Entre las reglas para seleccionar el número de componentes principales destacan principalmente las siguientes:

1. Se realiza un gráfico de λ_i frente a i . Se seleccionan componentes hasta que tengan un valor aproximado a λ_i .
2. Seleccionar componentes hasta cubrir un porcentaje determinado de varianza.
3. Fijada la varianza media, $\sum \frac{\lambda_i}{p}$, se eliminan los componentes que tengan valores propios menores que ésta.

Valores atípicos

Antes de empezar con el análisis de componentes principales se debe comprobar que no existen valores atípicos, ya que estos pueden modificar la matriz de covarianzas.

El efecto que tienen estos valores atípicos grandes es aumentar considerablemente las varianzas de esas variables y disminuir las covarianzas con las restantes. Por tanto, un valor atípico suficientemente grande modifica todos los componentes que podemos obtener.

Distribución de los componentes

Los componentes principales pueden verse como un conjunto nuevo de variables y estudiar su distribución individual y conjunta. Por construcción estarán incorrelados, pero pueden

existir relaciones no lineales entre ellos.

2.2. Aplicación práctica

El propósito de Kampakis (2016) es construir un modelo para predecir lesiones intrínsecas en base a los datos del GPS. Se usa para ello el método de componentes principales.

2.2.1. Variables

Se describen a continuación las principales variables del conjunto de datos.

Distancia recorrida: Valor para la distancia recorrida en un partido de 90 minutos, varía entre 9 y 13 km.

Alta velocidad en carrera: Mide la distancia recorrida cuando un jugador se mueve en las zonas 5 o 6.

Intensidad de la velocidad: Mide la cantidad de esfuerzo físico total para la sesión basada en la velocidad y es una variable producida por el GPS.

Aceleraciones y desaceleraciones: Un movimiento se registra como una aceleración o desaceleración solamente cuando la medición excede $0,5 \text{ m/s}^2$ y cuando su duración es mayor a medio segundo.

Metabolismo: La energía metabólica mide la velocidad a la que se gasta energía.

Carga de estrés dinámico: Es una variable que sirve para medir la fatiga.

Impactos: El propósito de la variable impactos^{es} detectar cada movimiento, ya sea saltando o dando un solo paso.

Carga total: Esta variable utiliza los datos del acelerómetro para medir las fuerzas aplicadas al jugador durante toda la sesión.

Baja velocidad de carga: Esta variable es equivalente a “carga total”, pero captura

única actividad que está por debajo de un cierto umbral y la actividad estática. Para estos datos, el umbral se ha establecido como a cualquier actividad en las zonas 1 o 2.

Sprints: Se basan en acciones por encima de cierto umbral de velocidad que duran al menos 1 segundo. El sprint se detiene cuando la velocidad cae por debajo de 80 % del umbral de sprint. Debido a eso, no todas las aceleraciones o desaceleraciones cuentan como carreras cortas.

Gasto de energía: Esta medida proporciona la energía total asociada con la ejecución, incluida la actividad de aceleración y desaceleración.

Equilibrio: Mide el impacto que tiene lugar en cada lado del cuerpo. Un valor de 0 indica una sesión perfectamente equilibrado.

2.2.2. Datos

Los datos se obtuvieron de los registros de GPS de 29 jugadores de fútbol profesional desde julio a diciembre de 2014 y de un registro de los días que estuvieron lesionados. Se utilizó el sistema GPS sólo durante el entrenamiento y no durante los partidos, debido a las regulaciones de primera división.

Había 68 variables de GPS registradas junto con la duración de la sesión de entrenamiento, dando 69 variables en total. El gran número de variables se debe al uso de distintas zonas para la medición de una sola variable, donde cada zona representa un rango de velocidad respecto a la velocidad máxima para cada futbolista. En la Tabla 2-2 se muestra una división en seis zonas, donde la primera zona corresponde al rango de velocidad de entre el 0 % y el 35 % respecto a la velocidad máxima de cada futbolista y la última corresponde a un rango de entre el 75 % y el 100 % respecto a la velocidad máxima de cada jugador.

Zona	Z1	Z2	Z3	Z4	Z5	Z6
Límite	<35 %	\ <35-45 %	<45-55 %	<55-65 %	<65-75 %	>75 %

Tabla 2-1.: Ejemplo de diferentes zonas definidas como regiones de la velocidad máxima.

La Figura 2-2 (Kampakis (2016)) muestra la correlación para las variables, en los que las variables con alta correlación han sido agrupadas juntas. Hay varios grupos que pueden ser identificados a lo largo del gráfico. En primer lugar, hay grupos de variables que

están midiendo la misma variable en diferentes zonas. Por ejemplo, los impactos en las zonas 1 a 6 están correlacionados entre sí, así como desaceleraciones de las zonas uno a seis.

Hay también algunas otras correlaciones que se pueden observar. Aceleraciones y desaceleraciones están correlacionados. Del mismo modo, el número total de pasos y la carga total se correlacionan con la distancia. Además, los impactos se correlacionan con la carga de la tensión dinámica.

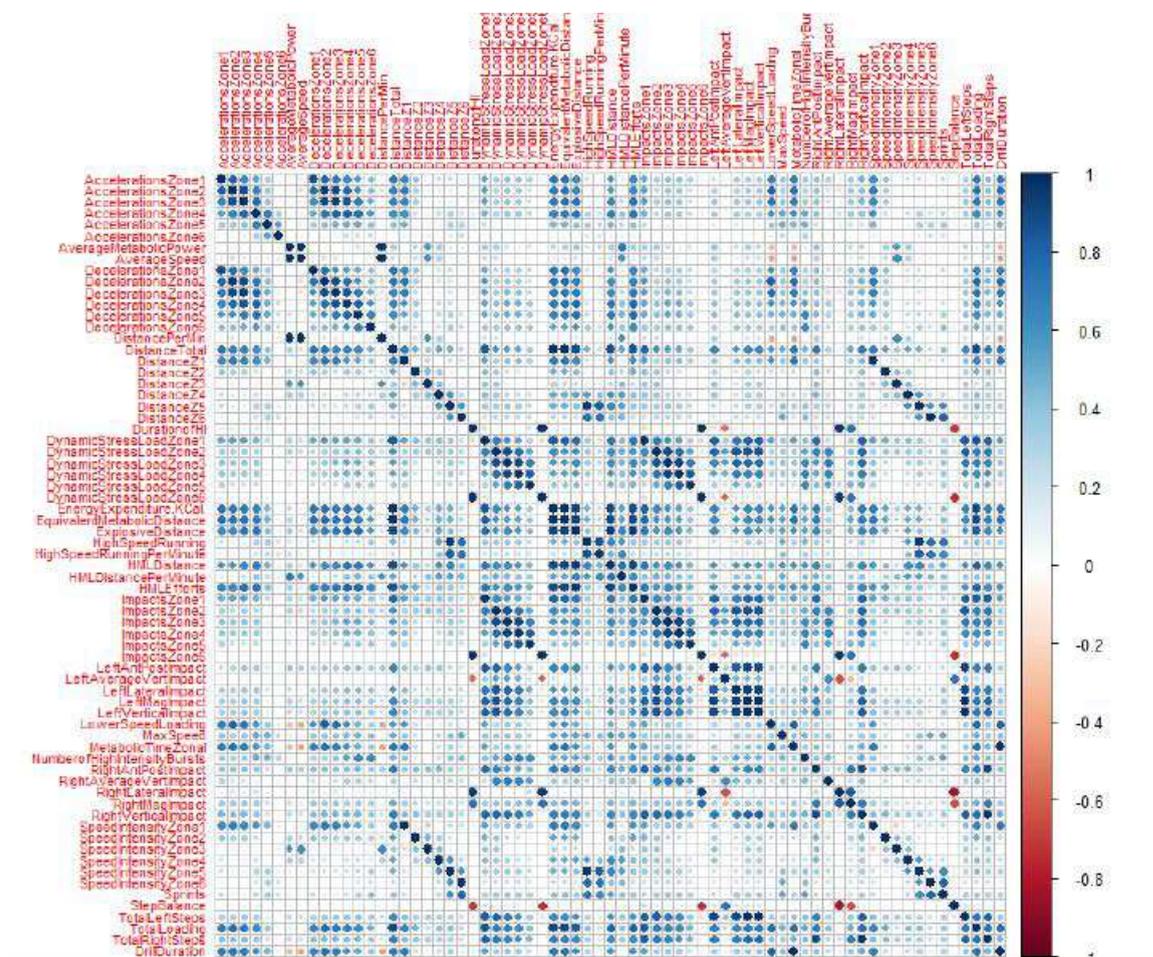


Figura 2-2.: Correlaciones de las variables del GPS

Fuente: Kampakis (2016)

2.2.3. Definición del problema

Enfoque A

En primer lugar, se eliminaron todos los jugadores que no habían sido lesionados. En segundo lugar, de los jugadores que habían sido lesionados, sólo se estudiaron aquellos que tenían lesiones no producidas por contacto. Estas lesiones son susceptibles de ser predichas por variables aportadas por los datos GPS. Esto dejó un total de 11 jugadores en el conjunto de datos.

El conjunto de datos se agrupó por jugador y semana, de manera que cada variable correspondía a un jugador i y una semana j . Se tiene una variable binaria que indica si el jugador se lesionó una semana o no.

La agregación por semanas fue elegida porque la semana es una división natural para la programación en el fútbol ya que cada semana se caracteriza por al menos un partido, y el programa de entrenamiento está diseñado sobre una base semanal.

La media fue utilizada como un resumen de agregación para agregar los casos durante la semana. Para cada jugador lesionado, había por lo menos 2 sesiones de entrenamiento, antes de que se produjera la lesión.

El último conjunto de datos consistió en 206 filas y 68 variables de GPS, además de la variable de respuesta y la duración de la sesión de entrenamiento. De las 206 filas, hubo 11 casos de lesiones (correspondientes a las semanas que cada jugador había estado lesionado).

El problema planteado consistía en saber si es posible predecir si se va a producir una lesión en una semana en base a los datos GPS recogidos: se centró sólo en las lesiones que se producen por un factor identificable durante el entrenamiento.

Enfoque B

En este caso el proceso es idéntico al anterior, pero están incluidos todos los jugadores, incluso los no lesionados. Esto condujo a un conjunto de datos de 426 filas y 70 las variables en total (el mismo que en el enfoque A), incluyendo la variable de clase.

2.2.4. Modelo

Se utiliza el método de componentes principales (CP), el cual sirve para reducir la dimensionalidad del problema con respecto a una variable de respuesta dada. En la Tabla 2-2 se exponen los parámetros del modelo.

Modelo	Parametros
Componentes principales	$\alpha \in \{0, 0.01, \dots, 0.5\}$, num_components $\in \{2:50\}$

Tabla 2-2.: Parámetros del modelo.

2.2.5. Resultados

Enfoque A

La Tabla 2-3 muestra los mejores parámetros del modelo para el enfoque A. La precisión y kappa se muestran como media \pm desviación típica.

Modelo	Parámetros	Exactitud	Kappa	Precisión	Recall
CP	$\alpha=0.01$ num_components=41	88.8% \pm 1.8%	0.21 \pm 0.04	0.55 \pm 0.09	0.33 \pm 0.07

Tabla 2-3.: Mejores parámetros del modelo para el enfoque A.

La Figura 2-3 muestra el *levelplot* del parámetro α y el número de componentes principales, con el color que representa el estadístico kappa. Parece que los mejores valores se consiguen cuando $\alpha < 0,05$ y cuando se usan más de 20 componentes.

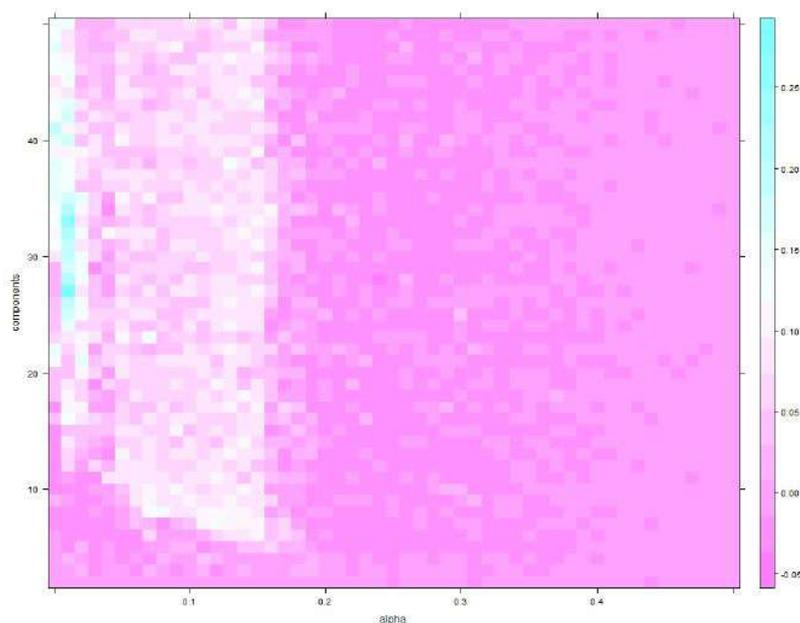


Figura 2-3.: Levelplot del enfoque A

Fuente: Kampakis (2016)

Enfoque B

La Tabla 2-4 muestra los mejores parámetros del modelo para el enfoque B. La precisión y kappa se muestran como media \pm desviación típica.

Modelo	Parámetros	Exactitud	Kappa	Precisión	Recall
CP	$\alpha=0.09$ num_components=40	97.07% \pm 3.31%	0.14 \pm 0.09	0.19 \pm 0.06	0.20 \pm 0.8

Tabla 2-4.: Mejores parámetros del modelo para el enfoque B

Se observa que el rendimiento de estos clasificadores es peor que el rendimiento de los del enfoque A. Esto se debe probablemente al hecho tener lesionados y no lesionados, los cuales presentan datos significativamente diferentes entre sí.

La Figura 2-4 muestra el *levelplot* para la búsqueda de parámetros para ACP para la aproximación del enfoque B. Podemos ver que las mejores soluciones se encuentran en la esquina superior izquierda para valores $\alpha < 0,2$ y cuando se usan más de 30 componentes.

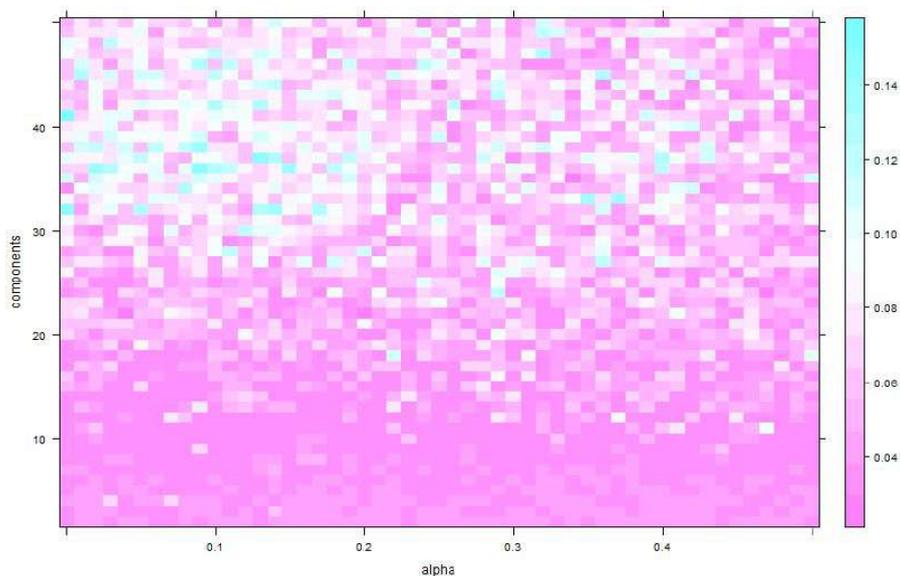


Figura 2-4.: Levelplot del enfoque B

Fuente: Kampakis (2016)

2.2.6. Otros comentarios sobre CP

Los componentes principales se utilizan con el fin de obtener una mejor comprensión del conjunto de datos e información de cómo las variables se relacionan con lesiones. En esta sección se analizarán las componentes principales para ambos enfoques.

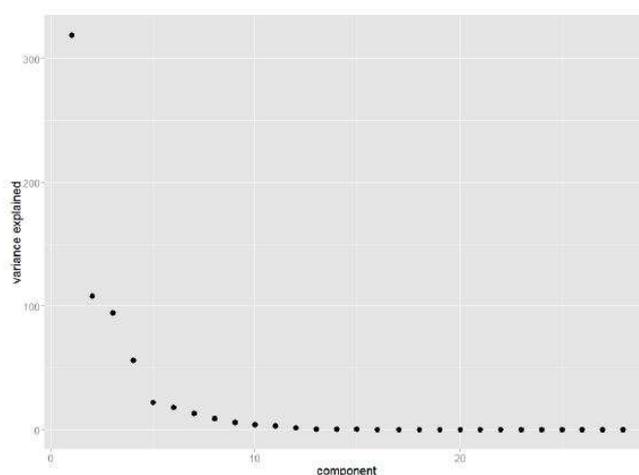


Figura 2-5.: Varianza explicada para cada componente en el enfoque A.

Fuente: Kampakis (2016)

Enfoque A

La Figura 2-5 muestra el gráfico de sedimentación para los componentes. Parece que los cuatro primeros componentes explican el mayor porcentaje de la varianza, por lo tanto, se decidió que los cuatro primeros componentes eran los adecuados para el análisis.

Observamos en la Tabla 2-5 que el primer componente está dominado por “Zona Distancia 5” con la que se correlaciona negativamente. Un valor mayor en este componente indica que el jugador no cubre una gran distancia en funcionamiento en altas velocidades. Por lo tanto, este componente parece ser un buen indicador de “Ausencia de correr en velocidades altas”. El segundo componente tiene una gran correlación negativa con “Distancia en la Zona 5” y una gran correlación positiva con aceleraciones y deceleraciones en la zona 3. Un valor grande en este componente parece indicar la ausencia de funcionamiento en las zonas de mayor velocidad y muchos sprints en la zona 3. Por lo tanto, este componente parece representar “Velocidad media de sprint”.

	PC1	PC2	PC3	PC4
Aceleración Zona 3	-0.22842	0.47640	0.02004	-0.18828
Desaceleración Zona 3	-0.22305	0.43085	0.07311	-0.07318
Desaceleración Zona 4	-0.09469	0.14696	-0.00709	-0.00245
Distancia por minuto	0.00317	-0.1908	0.43666	-0.19408
Distancia Zona 5	-0.76398	-0.48592	0.22479	-0.11721
Carga de estrés dinámico Zona 2	-0.18703	0.04578	-0.39766	0.37565
Carga de estrés dinámico Zona 3	-0.16017	0.06806	-0.17699	0.40768
Impactos Zona 4	-0.17954	0.08479	-0.16280	0.47654
Intensidad de velocidad Zona 3	-0.02075	-0.13704	-0.59876	-0.33671

Tabla 2-5.: Los dos primeros componentes principales del modelo con $\alpha=0.01$ para el conjunto de datos en el enfoque A.

El tercer componente está altamente correlacionado con “Distancia por minuto” y una correlación negativa con la “carga dinámica de la tensión en la zona 3” y “Intensidad velocidad en la zona 3”. Un gran valor en este componente parece indicar la ausencia de correr en zonas de mayor velocidad y muchos sprints en zona 3, por lo tanto, este componente puede ser un buen indicador de “Alta Velocidad promedio”.

El cuarto componente tiene una alta correlación con las variables que indican el estrés físico en las zonas de velocidad media (zonas 2-4). Por lo tanto, un buen nombre para este componente es “la intensidad del trabajo medio”.

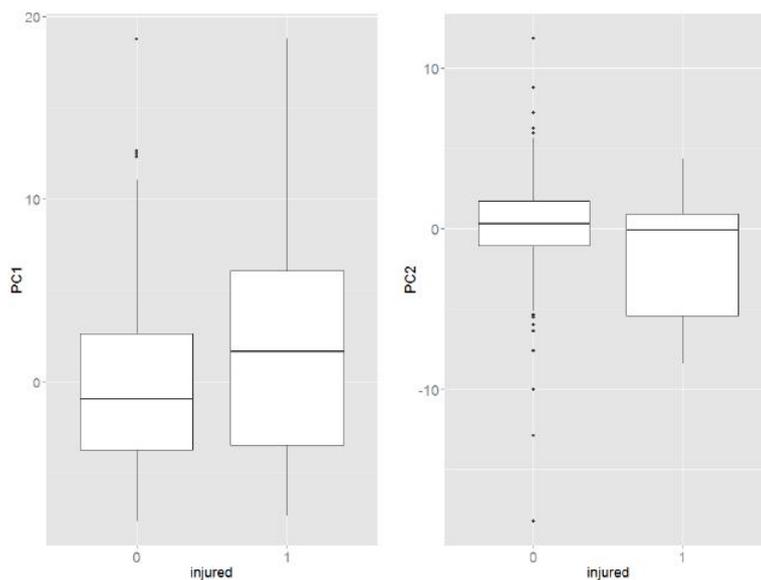


Figura 2-6.: Boxplot de los dos primeros componentes principales frente a la variable de respuesta.

Fuente: Kampakis (2016)

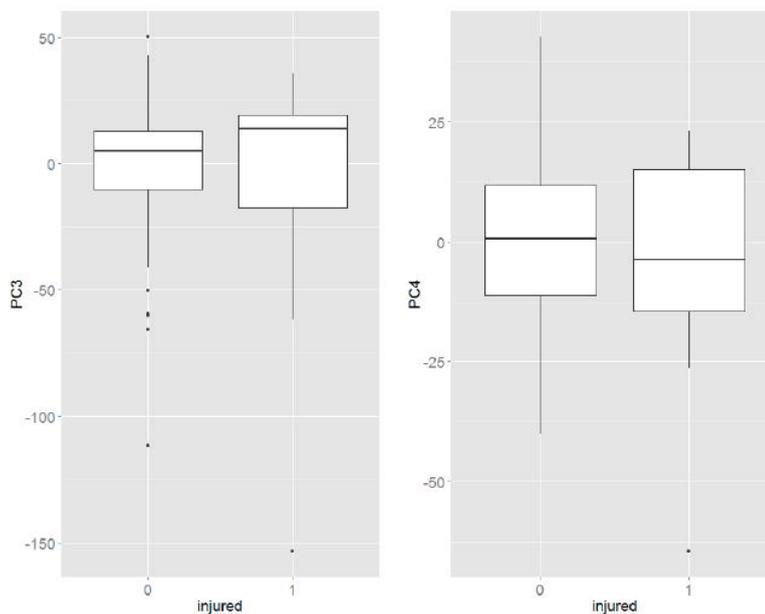


Figura 2-7.: Boxplot los componentes principales 3 y 4 frente a la variable de respuesta.

Fuente: Kampakis (2016)

Las Figuras 2-6 y la Figura 2-7 muestran los diagramas de caja de los componentes principales frente a la lesión. Existen diferencias en la mediana y el rango de los grupos lesionados y no lesionados para todos los componentes. Los componentes uno y tres tienen medianas más altas para los futbolistas lesionados, mientras que los componentes dos y cuatro tienen menor.

La prueba de rangos con signo de Wilcoxon indica que la mediana de las diferencias entre los heridos y los grupos no lesionados es cero y tiene asociado un para la segunda y la cuarto componentes. Esto refuerza la evidencia de que los componentes están asociados con la lesión.

Basado en el diagrama de caja para los componentes uno y tres, parece que una distancia máxima por minuto se correlaciona con la lesión, pero esto no se cumple cuando un jugador cubre una larga distancia en una velocidad alta. Tal vez la capacidad de un jugador para cubrir una gran distancia con las altas velocidades es una indicación de que su cuerpo no está en riesgo de lesión.

Los componentes dos y cuatro de la actividad medida de las zonas intermedias, en forma de aceleraciones/desaceleraciones, los impactos y la carga de estrés dinámico tienen una correlación negativa con la lesión. Esta relación podría deberse a una actividad de estrés medio que tiene un efecto preventivo sobre las lesiones o debido al hecho de que las sesiones de baja intensidad pueden no ser lo suficientemente estresantes como para causar lesiones.

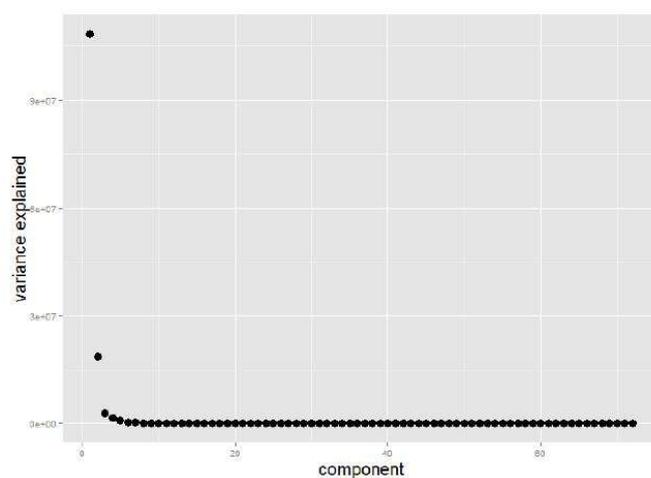


Figura 2-8.: Varianza explicada para cada componente para el enfoque B.

Fuente: Kampakis (2016)

Enfoque B

La Figura 2-8 muestra la varianza por componente para la aproximación B. Se puede observar que los dos primeros componentes explican el mayor porcentaje de la varianza.

La Tabla 2-6 muestra las correlaciones para los dos primeros componentes de este modelo. El primer componente se correlaciona negativamente con la “carga de la tensión dinámica en la zona 6” y “Impactos en la zona 6”. Parece representar la ausencia de altas velocidades en una sesión. Por lo tanto, un buen nombre para este componente es “ausencia de estrés de alta velocidad”.

El segundo componente tiene una alta correlación con la magnitud y el número de impactos verticales derecha e izquierda. Por lo tanto, este componente es un buen indicador de “impactos”.

	PC1	PC2
Estrés dinámico zona 6	-0.7336374014	-0.141046
Impactos zona 6	-0.2726364018	-0.083711
Impacto lateral derecho	-0.1904632684	-0.00039
Impacto de magnitud derecho	-0.1737929754	0.3375606
Metabolismo zona 1	0.0075269287	0.2539652
Impacto vertical derecho	0.0037938173	0.3199798
Impacto vertical izquierdo	0.0442883442	0.3203094
Impacto de magnitud izquierdo	0.0477825359	0.3527447

Tabla 2-6.: Los dos primeros componentes principales del modelo con $\alpha=0.01$ para el conjunto de datos en el enfoque B.

La Figura 2-9 muestra los diagramas de caja para los dos componentes frente a la variable de respuesta. Las diferencias entre los lesionados y los grupos no lesionados son destacadas. El primer componente tiene una mediana inferior para futbolistas lesionados, mientras que el segundo componente tiene claramente una mayor mediana para los jugadores lesionados. La interpretación de los componentes en este caso es que el estrés en las zonas de mayor velocidad y un gran número de impactos se correlaciona con la lesión.

La prueba de los rangos con signos de Wilcoxon que la mediana de las diferencias entre

los heridos y los grupos no lesionados tiene un p -value para la hipótesis nula de menos de 0,01 para ambos componentes.

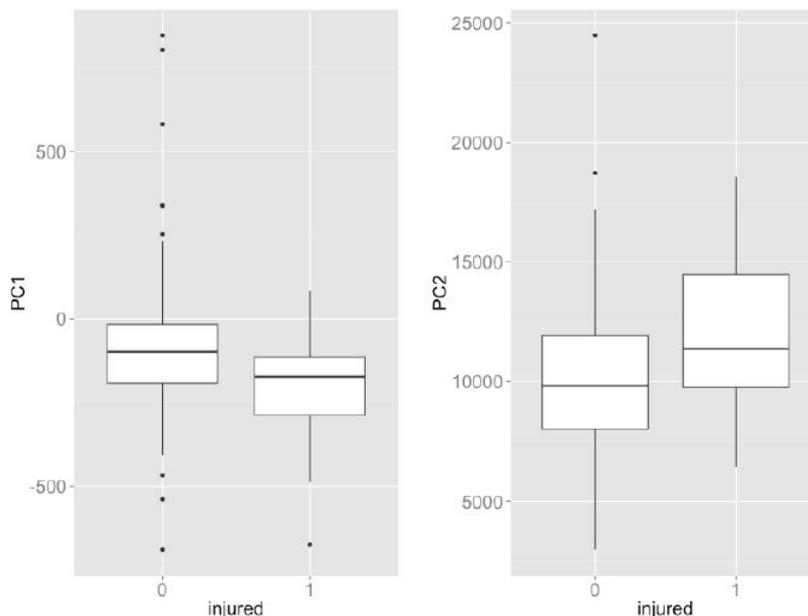


Figura 2-9.: Boxplot de los dos componentes principales frente a la variable de respuesta.

Fuente: Kampakis (2016)

2.2.7. Conclusiones

En este estudio se planteó el problema de la predicción de lesiones sin contacto de los datos GPS recogidos de entrenamiento. Los resultados demostraron que los datos GPS recogidos en los entrenamientos contienen información que puede ser utilizada para predecir lesiones.

En primer lugar, un problema con estos datos fue que las variables estaban altamente correlacionados entre sí, por ello se usó el método de componentes principales.

El objetivo del estudio fue, dado un conjunto de variables, aplicar el método de componentes principales para reducir el número de variables a aquellas que mas relación tengan con las lesiones. Esto es muy útil porque facilita la interpretación de los medicos.

Sin embargo, en este estudio se encuentran varias limitaciones. Otra es que muchas de las lesiones tienen lugar durante los partidos, o no están directamente relacionados con la fatiga. Este es el caso de las lesiones traumáticas, por ejemplo, que suelen venir como

resultado de un solo evento (por ejemplo, un sprint muy rápido).

Otro problema es el hecho de que los datos del GPS no se recogen de partidos, y estos tienen un gran impacto en la condición física de los futbolistas. Esto podría ser planteando un inconveniente en el rendimiento del modelo predictivo que tiene este estudio como objetivo conseguir.

3. ANOVA de medidas repetidas

“Ningún jugador es tan bueno como todos juntos”

Alfredo Di Stéfano (1926-2014)

En este capítulo se presenta el ANOVA de Medidas Repetidas para estudiar el efecto de la mejora del equilibrio en la reducción del número de lesiones.

3.1. Desarrollo teórico

Se entiende por medidas repetidas las múltiples medidas tomadas sobre un mismo sujeto. Estas medidas no son independientes ya que se caracterizan porque las mediciones son observadas en los mismos sujetos.

Los modelos de análisis de varianza (ANOVA) con medidas repetidas sirven para estudiar el efecto de uno o más factores cuando al menos uno de ellos es un factor intra-sujeto. Se entiende por factor intra-sujeto aquel cuyos niveles del factor se aplican a todos los sujetos de un mismo grupo. Puede ocurrir que además del factor intra-sujeto se disponga de un factor inter-sujeto, caracterizado porque cada nivel del factor se asigna a un grupo diferente de sujetos.

Antes de exponer los distintos modelos es necesario comprobar que se cumplen tres hipótesis: independencia, normalidad y esfericidad o circularidad.

3.1.1. Supuestos

- Independencia.

Cada sujeto de la muestra se selecciona aleatoriamente de su población y es asignando a cada uno de los T tratamientos del factor, por lo que sus datos son independientes de los obtenidos por el resto de sujetos. En consecuencia, lo que unos datos se desvíen del promedio de su grupo es independiente de lo que otros datos se desvíen de ese mismo grupo. Este supuesto suele comprobarse mediante la prueba de rachas.

- Normalidad.

La muestra aleatoria de t sujetos se extrae de una población normal. Para comprobar este supuesto se suele emplear el test de Shapiro-Wilk.

- Esfericidad o circularidad.

Sea la matriz de varianzas-covarianzas:

$$\Sigma_{t \times t} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1t} \\ & \sigma_2^2 & \cdots & \sigma_{2t} \\ & & \ddots & \vdots \\ & & & \sigma_t^2 \end{pmatrix}, \text{ donde } \sigma_{ij} \text{ es la covarianza entre } i \text{ y } j$$

entonces, una condición suficiente para que se verifique que el estadístico F se distribuya mediante el modelo probabilístico F -Snedecor con $t - 1$ y $(n - 1)(t - 1)$ grados de libertad si la matriz de varianzas-covarianzas verifica que las varianzas de las diferencias de todos los pares de niveles de tratamiento son homogéneos (iguales), es decir:

$$\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij} = 2\lambda, \text{ para todo } i \neq j.$$

La condición mínima exigible para que el estadístico F sea válido es que se cumpla el supuesto de esfericidad o circularidad en la matriz de varianzas-covarianzas de la población. Para ello vamos a considerar la matriz C como la matriz de transformación, que se encarga de transformar el conjunto de variables originales en un conjunto de $t - 1$ variables ortonormales (Z), es decir, ortogonales entre sí y que además la suma de cuadrados de los coeficientes para cada variable es 1. Por la hipótesis de esfericidad la matriz de varianzas-covarianzas para las nuevas variables transformadas es una matriz diagonal, con un escalar $\lambda > 0$ igual en la diagonal. Luego,

$$\Sigma_Z = \lambda I_{t-1} = C' \Sigma C = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ & \lambda & 0 & 0 \\ & & \ddots & 0 \\ & & & \lambda \end{pmatrix} \in M_{t-1 \times t}.$$

Por lo tanto, antes de realizar el análisis de la varianza conviene contrastar la hipótesis nula de que la matriz de varianzas-covarianzas es circular, para ello utilizamos el Test de Mauchly. Hay que tener en cuenta que este Test es una prueba sensible al incumplimiento

de la hipótesis de normalidad.

En los siguientes apartados se expondrán los distintos modelos de ANOVA con medidas repetidas.

3.1.2. Diseño de medidas repetidas con un factor intra-sujetos

En este diseño se dispone de un único factor T con t niveles, los cuales se aplican a los n individuos de una muestra seleccionada aleatoriamente. El objetivo será estudiar el efecto del factor intra-sujetos (con t niveles) sobre la variable respuesta Y , donde y_{ij} es el valor de la variable respuesta del j -ésimo individuo en el nivel i -ésimo.

Sea μ la media poblacional de la variable respuesta, α_i el efecto medio adicional debido al tratamiento i , S_j el efecto aleatorio del sujeto j y e_{ij} el error aleatorio. Suponiendo que no existe interacción entre los sujetos y el factor se tiene que el modelo lineal asociado a este diseño viene dado por:

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + S_j + e_{ij} \quad i = 1, 2, \dots, t \quad j = 1, 2, \dots, n; \\ \sum_{i=1}^t \alpha_i &= 0; \quad S_j \sim N(0, \sigma_1); \quad e_{ij} \sim N(0, \sigma_2); \\ S_j \text{ y } e_{ij} &\text{ son mutuamente independientes.} \end{aligned}$$

Se dispone de t niveles y n réplicas por cada nivel.

Análisis de la varianza y esperanzas de los cuadrados medios

Descomponiendo la varianza (SC) según la variación entre sujetos (SC_E), encargada de medir la diferencia entre los sujetos, y la variación dentro de los sujetos (SC_D), que se ocupa de medir las diferencias dentro de los sujetos, se tiene:

$$SC = SC_E + SC_D,$$

donde:

$$SC = \sum_{i=1}^t \sum_{j=1}^n (y_{ij} - \bar{y})^2$$

tiene $nt - 1$ grados de libertad.

$$SC_E = \sum_{i=1}^t \sum_{j=1}^n (y_j - \bar{y})^2$$

tiene $n - 1$ grados de libertad.

$$SC_D = \sum_{i=1}^t \sum_{j=1}^n (y_{ij} - \bar{y}_j)^2$$

tiene $n(t - 1)$ grados de libertad.

Además, la variación dentro de los sujetos se puede descomponer en dos, la variación debida a los niveles (SC_T), encargada de medir las diferencias entre los tratamientos, y la debida a la aleatoriedad (error), encargada de recoger la variación no explicada por los tratamientos (SC_{Error}).

$$SC_D = SC_T + SC_{Error},$$

donde:

$$SC_T = \sum_{i=1}^t \sum_{j=1}^n (y_i - \bar{y})^2$$

tiene $t - 1$ grados de libertad.

$$SC_{Error} = \sum_{i=1}^t \sum_{j=1}^n (y_{ij} - \bar{y}_j - \bar{y}_i + \bar{y})^2$$

tiene $(n - 1)(t - 1)$ grados de libertad.

Por tanto la descomposición de la varianza queda:

$$SC = SC_E + SC_T + SC_{Error}.$$

Los cuadrados medios representan una estimación de la varianza de la población, en este caso se utilizan para determinar si los factores (tratamientos) son significativos.

En la variación entre sujetos los cuadrados medios vienen dados por $\frac{SC_E}{n - 1}$ y la esperanza de estos por $\sigma_2^2 + t\sigma_1^2$.

En la variación dentro de los sujetos los cuadrados medios vienen dados por $\frac{SC_D}{n(t - 1)}$ y

la esperanza de estos por $\sigma_2^2 + \frac{1}{t-1} \sum \alpha_i^2$.

En la variación debida a los tratamientos los cuadrados medios vienen dados por $CM_T = \frac{SC_T}{t-1}$ y la esperanza de estos por $\sigma_2^2 + \frac{n}{t-1} \sum \alpha_i^2$.

En la variación debida a la aleatoriedad (error) los cuadrados medios vienen dados por $CM_{Error} = \frac{SC_{Error}}{(n-1)(t-1)}$ y la esperanza de estos por σ_2^2 .

Usando la descomposición de la varianza y las esperanzas de los cuadrados medios nos planteamos el contraste de igualdad de tratamientos:

$$H_0 : \alpha_i = 0, \quad i = 1, \dots, t \iff H_0 : \mu_1 = \dots = \mu_t,$$

el cual tiene asociado un estadístico:

$$F_0 = \frac{CM_T}{CM_{Error}},$$

cuya distribución, bajo la hipótesis nula y algunas suposiciones de la matriz de varianzas-covarianzas que se expondrán un poco más adelante, es una F -Snedecor con $t-1$ y $(n-1)(t-1)$ grados de libertad.

Se necesitan los estimadores de σ_1^2 y σ_2^2 . Para ello, teniendo en cuenta las esperanzas de los cuadrados medios de SC_E y SC_{Error} , se tiene que los estimadores vienen dados por:

$$\hat{\sigma}_1^2 = \frac{CM_E - CM_{Error}}{t},$$

$$\hat{\sigma}_2^2 = CM_{Error} = \frac{SC_E}{(n-1)(t-1)}.$$

Si se verifica la condición de circularidad, entonces el estadístico $F_0 = \frac{CM_T}{CM_{Error}}$, bajo la hipótesis nula, tiene asociada una distribución F -Snedecor con $t-1$ y $(n-1)(t-1)$ grados de libertad y realizamos un análisis univariante.

Si no se cumple la condición de esfericidad se planten dos opciones:

1. Ajustar los grados de libertad del estadístico F .

La idea es ajustar los grados de libertad mediante la prueba épsilon (ϵ) que es una corrección que se aplica en los grados de libertad del numerador y del denominador del estadístico F de forma que este pasa a distribuirse según F con $\epsilon(t-1)$ y $\epsilon(t-1)(n-1)$ grados de libertad tal y como se explica en Ximénez y Martín (2000)

2. Aplicar un análisis de varianza multivariado (o MANOVA) basado en $t - 1$ contrastes independientes.

En este caso el objetivo es transformar el ANOVA de un factor en un análisis multivariado (MANOVA), como se desarrolla en Ximénez y Martín (2000).

Comparaciones múltiples de medias

Si al realizar el Análisis de la Varianza en un modelo de medidas repetidas con un factor intra-sujeto aparecen diferencias significativas, deberemos localizar dónde se encuentran esas diferencias.

En particular nos interesa comparar los niveles del factor T (factor intra-sujetos). Para ello debemos comparar los valores μ_i estimador por \bar{y}_i . Un estimador insesgado de la diferencia de medias es

$$\bar{y}_i - \bar{y}_{i'} = \hat{\mu}_i - \hat{\mu}_{i'},$$

cuya varianza es $\frac{2\sigma_2^2}{n}$, y el estimador del error estándar de esta diferencia es:

$$\hat{\sigma}(\bar{y}_i - \bar{y}_{i'}) = \sqrt{\frac{2CM_{Error}}{n}}.$$

En consecuencia a la hora de hacer comparaciones múltiples entre los niveles i y i' del factor T , según el método LSD (Least Significance Method) de Fisher, la mínima diferencia significativa viene dada por

$$LSD_{\alpha} = t_{(n-1)(t-1); \frac{\alpha}{2}} \sqrt{\frac{2CM_{Error}}{n}},$$

donde $t_{(n-1)(t-1); \frac{\alpha}{2}}$ representa una t-Student con $(n - 1)(t - 1)$ grados de libertad y nivel de confianza $\frac{\alpha}{2}$.

3.1.3. Diseño de medidas repetidas con dos factores intra-sujetos

Se dispone de dos factores intra-sujetos T y V , con t niveles respectivamente, de forma que tenemos tv tratamientos los cuales se aplican a una muestra de n sujetos seleccionados aleatoriamente. El objetivo será estudiar el efecto de los factores intra-sujetos T y V sobre la variable respuesta Y , donde y_{ijk} es el valor de la variable respuesta del k -ésimo individuo bajo el nivel i -ésimo del factor T y el nivel j -ésimo del factor V .

Sea μ la media poblacional de la variable respuesta, α_i el efecto medio adicional debido al nivel i del factor T , β_j el efecto adicional debido al nivel j del factor V , $(\alpha\beta)_{ij}$ el efecto

medio adicional debido a la interacción entre el nivel i del factor T y el nivel j del factor V , S_k el efecto aleatorio del sujeto k , S_{ik} el efecto aleatorio del nivel i del factor T y el sujeto k , S_{jk} el efecto aleatorio del nivel j del factor V y el sujeto k y e_{ijk} representa el error aleatorio. Suponiendo que no existe interacción de orden tres entre los sujetos y los factores intra-sujetos se tiene que el modelo lineal asociado a este diseño viene dado por:

$$\begin{aligned}
 y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + S_k + S_{ik} + S_{jk} + e_{ijk} \\
 &\text{con } i = 1, 2, \dots, t \quad j = 1, 2, \dots, v \quad k = 1, 2, \dots, n; \\
 \sum_{i=1}^t \alpha_i &= 0; \quad \sum_{j=1}^v \beta_j = 0; \quad \sum_{i=1}^t (\alpha\beta)_{ij} = 0 \quad \forall j; \quad \sum_{j=1}^v (\alpha\beta)_{ij} = 0 \quad \forall i; \\
 S_k &\sim N(0, \sigma_1); \quad S_{ik} \sim N(0, \sigma_2); \quad S_{jk} \sim N(0, \sigma_3); \quad e_{ij} \sim N(0, \sigma_4); \\
 S_j, S_{ik}, S_{jk}, &\text{ y } e_{ij} \text{ son mutuamente independientes.}
 \end{aligned}$$

Se dispone de tv tratamientos y n réplicas por tratamiento.

Análisis de la varianza y esperanzas de los cuadrados medios

Descomponiendo la varianza (SC) según la variación entre sujetos (SC_E), encargada de medir la diferencia entre los sujetos, y la variación dentro de los sujetos (SC_D), que se ocupa de medir las diferencias dentro de los sujetos, se tiene:

$$SC = SC_E + SC_D,$$

donde:

$$SC = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (y_{ijk} - \bar{y})^2$$

tiene $nvt - 1$ grados de libertad.

$$SC_E = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_k - \bar{y})^2$$

tiene $n - 1$ grados de libertad.

$$SC_D = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_{ijk} - \bar{y}_k)^2$$

tiene $vt(n - 1)$ grados de libertad.

Además, la variación dentro de los sujetos se puede descomponer en tres, la variación debida a los factores (SC_T), la debida a las interacciones entre los sujetos y cada factor (SC_V) y la debida a la aleatoriedad (error) (SC_{Error}).

$$SC_D = SC_T + SC_V + SC_T * V + SC_T * S + SC_V * S + SC_{Error},$$

donde:

$$SC_T = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_i - \bar{y})^2$$

es la suma de cuadrados asociada al factor T y tiene $t - 1$ grados de libertad.

$$SC_V = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_j - \bar{y})^2$$

es la suma de cuadrados asociada al factor V tiene $v - 1$ grados de libertad.

$$SC_T * V = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

es la suma de cuadrados asociada a la interacción entre los dos factores intra-sujetos T y tiene $(t - 1)(v - 1)$ grados de libertad.

$$SC_T * S = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_{ik} - \bar{y}_i - \bar{y}_k + \bar{y})^2$$

es la variabilidad correspondiente a la interacción entre los factores intra-sujetos y tiene $(t - 1)(n - 1)$ grados de libertad.

$$SC_V * S = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})^2$$

es la variabilidad correspondiente a la interacción entre el factor T y los sujetos y tiene $(v - 1)(n - 1)$ grados de libertad.

$$SC_{Error} = \sum_{i=1}^t \sum_{j=1}^v \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij} - \bar{y}_{ik} - \bar{y}_{jk} + \bar{y}_i + \bar{y}_j + \bar{y}_k - \bar{y})^2$$

que recoge la variación no explicada por los tratamientos y las interacciones de orden tres, es decir, el error aleatorio. Tiene $(n-1)(t-1)(v-1)$ grados de libertad.

Por tanto la descomposición de la varianza queda:

$$SC = SC_E + SC_T + SC_V + SC_T * V + SC_T * S + SC_V * S + SC_{Error}.$$

En la variación entre sujetos los cuadrados medios vienen dados por $\frac{SC_E}{n-1}$.

En la variación dentro de los sujetos los cuadrados medios vienen dados por $\frac{SC_D}{vt(n-1)}$.

En la variación asociada al factor T los cuadrados medios vienen dados por $CM_T = \frac{SC_T}{t-1}$

y la esperanza de estos por $\sigma_4^2 + v\sigma_2^2 + \frac{nv}{t-1} \sum \alpha_i^2$.

En la variación asociada al factor V los cuadrados medios vienen dados por $CM_V = \frac{SC_V}{v-1}$

y la esperanza de estos por $\sigma_4^2 + t\sigma_3^2 + \frac{nt}{v-1} \sum \beta_j^2$.

En la variación asociada a la interacción entre los factores intra-sujetos los cuadrados medios vienen dados por $CM_T * V = \frac{SC_T * V}{(t-1)(v-1)}$ y la esperanza de estos por

$$\sigma_4^2 + \frac{n}{(t-1)(v-1)} \sum \sum (\alpha\beta)_{ij}^2.$$

En la variación asociada a la interacción entre el factor T y los sujetos los cuadrados medios vienen dados por $CM_T * S = \frac{SC_T * S}{(t-1)(n-1)}$ y la esperanza de estos por $\sigma_4^2 + v\sigma_2^2$.

En la variación asociada a la interacción entre el factor V y los sujetos los cuadrados medios vienen dados por $\frac{SC_V * S}{(v-1)(n-1)}$ y la esperanza de estos por $\sigma_4^2 + t\sigma_3^2$.

En la variación debida a la aleatoriedad (error) los cuadrados medios vienen dados por $\frac{SC_{Error}}{(n-1)(t-1)(v-1)}$ y la esperanza de estos por σ_4^2 .

Se necesitan los estimadores de σ_2^2 , σ_3^2 y σ_4^2 . Para ello, teniendo en cuenta las esperanzas de los cuadrados medios de $SC_T * S$, $SC_V * S$ y SC_{Error} y se tiene que los estimadores vienen dados por:

$$\begin{aligned} \hat{\sigma}_2^2 &= \frac{CM_T * S - CM_{Error}}{v}, \\ \hat{\sigma}_3^2 &= \frac{CM_V * S - CM_{Error}}{t}, \\ \hat{\sigma}_4^2 &= CM_{Error} = \frac{SC_E}{(n-1)(t-1)(v-1)}. \end{aligned}$$

Usando la descomposición de la varianza y las esperanzas de los cuadrados medios y suponiendo que las matrices de varianzas-covarianzas $\Sigma_{tv \times tv}$, correspondiente al vector

aleatorio (Y_1, \dots, Y_t) , $\Sigma_{t \times t}$, asociada al vector que recoge las respuestas bajo los niveles de factor T , y $\Sigma_{v \times v}$, asociada al vector que recoge las respuestas bajo los niveles de factor V , verifican la propiedad de circularidad, se plantean los contrastes de hipótesis sobre la interacción entre ambos factores y cada factor por separado:

- Efectos de la interacción.

$$\begin{cases} H_0 : (\alpha\beta)_{ij} = 0; \\ H_1 : \text{Existe al menos un par } (i, j) \text{ tal que } (\alpha\beta)_{ij} \neq 0. \end{cases}$$

Este contraste tiene asociado el estadístico:

$$F_0 = \frac{CM_T * V}{CM_{Error}},$$

con $(t-1)(v-1)$ y $(n-1)(t-1)(v-1)$ grados de libertad.

- Efectos del factor T .

$$\begin{cases} H_0 : \alpha_i = 0 \quad i = 1, \dots, t; \\ H_1 : \text{Existe al menos un par } (i, i') \text{ tal que } \alpha_i \neq \alpha_{i'}. \end{cases}$$

Este contraste tiene asociado el estadístico:

$$F_0 = \frac{CM_T}{CM_T * S},$$

con $(t-1)$ y $(n-1)(t-1)$ grados de libertad.

- Efectos del factor V .

$$\begin{cases} H_0 : \beta_j = 0 \quad j = 1, \dots, v; \\ H_1 : \text{Existe al menos un par } (j, j') \text{ tal que } \beta_j \neq \beta_{j'}. \end{cases}$$

Este contraste tiene asociado el estadístico:

$$F_0 = \frac{CM_V}{CM_V * S},$$

con $(v-1)$ y $(n-1)(v-1)$ grados de libertad.

De forma análoga al modelo de medidas repetidas con un factor intra-sujetos si no se verifica la condición de circularidad de la matriz de varianzas-covarianzas se presentan dos soluciones alternativas:

1. Realizar un análisis univariante “aproximado” aplicando la prueba ϵ a los grados de libertad del estadístico F_0 .
2. Aplicar un análisis de varianza multivariado (o MANOVA) con dos factores.

Comparaciones múltiples de medias.

Supongamos que se cumple la condición de circularidad en las matrices de varianzas-covarianzas. Si una vez realizado el Análisis de la Varianza en este modelo aparecen diferencias significativas, el siguiente objetivo será localizar donde se encuentran esas diferencias. Se considerarán dos situaciones en función de si existe o no interacción significativa entre los factores T y V .

Si no existe interacción significativa nos interesa comparar los niveles del factor T , cuando existan diferencias significativas entre ellos, y los niveles del factor V cuando se dé esta situación.

Se ven las diferencias entre los estimadores de μ_i , las medias de \bar{y}_i para comparar los niveles del factor T . El estimador de insesgado de las diferencias de medias es:

$$\bar{y}_i - \bar{y}_{i'} = \hat{\mu}_i - \hat{\mu}_{i'},$$

que tiene por varianza:

$$\frac{2}{vn}(\sigma_4^2 + v\sigma_2^2),$$

y por estimador del error:

$$\hat{\sigma}(\bar{y}_i - \bar{y}_{i'}) = \sqrt{\frac{2}{vn}(\hat{\sigma}_4^2 + v\hat{\sigma}_2^2)}.$$

Por tanto, según el método LSD de Fisher, al hacer comparaciones múltiples entre los niveles i y i' del factor T , la mínima diferencia significativa viene dada por:

$$LSD_\alpha = t_{(t-1)(n-1); \frac{\alpha}{2}} \sqrt{\frac{2}{vn}(\hat{\sigma}_4^2 + v\hat{\sigma}_2^2)}.$$

En el caso de en el que se comparan los niveles del factor V se deben estudiar las diferencias entre los estimadores de μ_j y las medias \bar{y}_j . El estimador insesgado de la diferencias de medias es:

$$\bar{y}_j - \bar{y}_{j'} = \hat{\mu}_j - \hat{\mu}_{j'},$$

cuya varianza es:

$$\hat{\sigma}(\bar{y}_j - \bar{y}_{j'}) = \sqrt{\frac{2}{tn}(\hat{\sigma}_4^2 + t\hat{\sigma}_3^2)}.$$

Por lo tanto al hacer comparaciones múltiples entre los niveles j y j' del factor V , atendiendo al método LSD de Fisher, la mínima diferencia significativa viene dada por:

$$LSD_{\alpha} = t_{(v-1)(n-1); \frac{\alpha}{2}} \sqrt{\frac{2}{tn} (\hat{\sigma}_4^2 + t\hat{\sigma}_3^2)}.$$

Si no existe interacción significativa es posible realizar comparaciones de los tv tratamiento de forma análoga que en un diseño de dos factores de los factores inter-sujetos.

En general, aparece interacción significativa de los dos factores cuando no se cumple el supuesto de circularidad. En ese caso se utiliza para cada nivel del factor T (o cada nivel del factor V) el test t para dos muestras relacionadas de forma que se comparan dos a dos entre los niveles v niveles del factor V (o los t niveles del factor T) controlando el error experimental por el procedimiento de Bonferroni. Si no hay interacción entre los dos factores se pueden estudiar las diferencias para cada uno de los factores utilizando el test t para dos muestras relacionadas de forma que se comparan dos a dos entre los t o v niveles de cada factor controlando el error experimental por el método de Bonferri.

3.1.4. Diseño de medidas repetidas con un factor intra-sujetos y un factor inter-sujetos

Se dispone de un factor A (factor inter-sujetos) con a niveles que se asignan a grupos independientes de n sujetos, de forma que se tiene a muestras aleatorias independientes de n sujetos a los que se somete a los t niveles del factor T (factor intra-sujetos). El objetivo será caracterizar el comportamiento de la variable respuesta Y sobre los a grupos a lo largo de los t niveles del factor T , donde y_{ijk} es el valor de la variable respuesta del k -ésimo individuo bajo el i -ésimo nivel del factor A y el j -ésimo nivel del factor T . Hay que tener en cuenta que al considerar los sujetos fuente de variación sus niveles están anidados en el factor inter-sujetos A .

Sea μ la media poblacional de la variable respuesta, α_i el efecto medio adicional debido al nivel i del factor A , $S_{k(i)}$ el efecto aleatorio del sujeto k en el nivel i del factor A , β_j el efecto adicional debido al nivel j del factor T , $(\alpha\beta)_{ij}$ el efecto medio adicional debido a la interacción entre el nivel i del factor A y el nivel j del factor T , μ_{ij} la respuesta media bajo el nivel i -ésimo del factor A y e_{ijk} representa el error aleatorio. Este diseño viene dado por:

$$\begin{aligned} y_{ijk} &= \mu + S_{k(i)} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} = \mu_{ij} + S_{k(i)} + e_{ijk} \\ i &= 1, 2, \dots, a \quad j = 1, 2, \dots, t \quad k = 1, 2, \dots, n; \\ \sum_{i=1}^a \alpha_i &= 0; \quad \sum_{j=1}^t \beta_j = 0; \quad \sum_{j=1}^t (\alpha\beta)_{ij} = 0 \quad \forall i; \quad \sum_{i=1}^a (\alpha\beta)_{ij} = 0 \quad \forall j; \\ S_{k(i)} &\sim N(0, \sigma_1); \quad e_{ij} \sim N(0, \sigma_2) \\ S_{k(i)} &\text{ y } e_{ij} \text{ son mutuamente independientes.} \end{aligned}$$

Análisis de la varianza y esperanzas de los cuadrados medios

Descomponiendo la varianza (SC) según la variación entre sujetos (SC_E), encargada de medir la diferencia entre los sujetos, y la variación dentro de los sujetos (SC_D), que se ocupa de medir las diferencias dentro de los sujetos, se tiene:

$$SC = SC_E + SC_D,$$

donde:

$$SC = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (y_{ijk} - \bar{y})^2$$

tiene $nat - 1$ grados de libertad.

$$SC_E = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_{ik} - \bar{y})^2$$

tiene $na - 1$ grados de libertad.

$$SC_D = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_{ijk} - \bar{y}_{ik})^2$$

tiene $na(t - 1)$ grados de libertad.

Además, la variación entre sujetos se puede descomponer en dos, la variación debida al factor A (SC_A), y la debida a los sujetos en cada nivel A (SC_{SA}),

$$SC_E = SC_A + SC_{SA},$$

donde:

$$SC_A = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_i - \bar{y})^2$$

que mide las diferencias entre los niveles del factor A , y tiene $a - 1$ grados de libertad.

$$SC_{SA} = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_j - \bar{y})^2$$

que mide las diferencias entre los sujetos para cada nivel del factor A y tiene $a(n - 1)$ grados de libertad.

Por otro lado, la variación dentro de los sujetos se puede descomponer en distintas fuentes de variación, la debida al factor T (SC_T), la debida a las interacciones entre el factor A y el factor T ($SC_A * T$) y la debida a la aleatoriedad (error) (SC_{Error}),

$$SC_D = SC_T + SC_A * T + SC_{Error}.$$

$$SC_T = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_j - \bar{y})^2$$

que mide las diferencias entre los niveles del factor T , y tiene $(t - 1)$ grados de libertad.

$$SC_A * T = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$$

es la suma de cuadrados asociada a la interacción entre el factor inter-sujetos y el factor intra-sujetos, y tiene $(a - 1)(t - 1)$ grados de libertad.

$$SC_{Error} = \sum_{i=1}^a \sum_{j=1}^t \sum_{k=1}^n (y_{ijk} - \bar{y}_{ik} - \bar{y}_{ij} + \bar{y}_i)^2$$

que recoge la variación no explicada por los tratamientos y las interacciones de orden tres, es decir, el error aleatorio. Tiene $a(n - 1)(t - 1)$ grados de libertad.

Por tanto la descomposición de la varianza queda:

$$SC = SC_A + SC_{SA} + SC_T + SC_A * T + SC_{Error}.$$

Para SC_E los cuadrados medios vienen dados por $\frac{SC_E}{an - 1}$.

Para SC_A los cuadrados medios vienen dados por $CM_A = \frac{SC_A}{a - 1}$, y la esperanza de estos por $\sigma_2^2 + t\sigma_1^2 + \frac{nt}{a - 1} \sum \alpha_i^2$

Para SC_{SA} los cuadrados medios vienen dados por $\frac{SC_{SA}}{a(n - 1)}$ y la esperanza de estos por $\sigma_2^2 + t\sigma_1^2$.

Para SC_D los cuadrados medios vienen dados por $\frac{SC_D}{na(t - 1)}$.

Para SC_E los cuadrados medios vienen dados por $CM_T = \frac{SC_T}{(t-1)}$ y la esperanza de estos por $\sigma_2^2 + \frac{na}{(t-1)} \sum \sum (\alpha\beta)_{ij}^2$.

En la variación debida a la aleatoriedad (error) los cuadrados medios vienen dados por $\frac{SC_{Error}}{a(n-1)(t-1)}$ y la esperanza de estos por σ_2^2 .

Se necesitan los estimadores de σ_2^2 y σ_1^2 . Para ello, teniendo en cuenta las esperanzas de los cuadrados medios de $SC_{SA} * S$ y SC_{Error} y se tiene que los estimadores vienen dados por:

$$\hat{\sigma}_1^2 = \frac{CM_{SA} * S - CM_{Error}}{t},$$

$$\hat{\sigma}_2^2 = CM_{Error} = \frac{SC_E}{a(n-1)(t-1)}.$$

Usando la descomposición de la varianza y las esperanzas de los cuadrados medios y suponiendo que las matrices de varianzas-covarianzas Σ_i $i = 1, \dots, a$ deben ser iguales en cada uno de los niveles del factor inter-sujetos y que la matriz de varianzas-covarianzas combinada, $\Sigma_{t \times t}$, definida como:

$$\Sigma_{t \times t} = \frac{1}{\sum_{i=1}^a (n_i - 1)} \sum_{i=1}^a (n_i - 1) \Sigma_i,$$

es circular, se obtienen los contrastes de hipótesis sobre la interacción entre ambos factores y cada factor por separado.

- Efectos de la interacción.

$$\begin{cases} H_0 : (\alpha\beta)_{ij} = 0; \\ H_1 : \text{Existe al menos un par } (i, j) \text{ tal que } (\alpha\beta)_{ij} \neq 0. \end{cases}$$

Este constraate tiene asociado el estadístico:

$$F_0 = \frac{CM_T * A}{\hat{\sigma}_2^2},$$

con $(a-1)(t-1)$ y $a(n-1)(t-1)$ grados de libertad.

- Efectos del factor inter-sujetos A .

$$\begin{cases} H_0 : \alpha_i = 0 \quad i = 1, \dots, t; \\ H_1 : \text{Existe al menos un par } (i, i') \text{ tal que } \alpha_i \neq \alpha_{i'}. \end{cases}$$

Este constraate tiene asociado el estadístico:

$$F_0 = \frac{CM_A}{\widehat{\sigma}_1^2},$$

con $(a - 1)$ y $a(n - 1)$ grados de libertad.

- Efectos del factor intra-sujetos T .

$$\begin{cases} H_0 : \beta_j = 0 \quad j = 1, \dots, v; \\ H_1 : \text{Existe al menos un par } (j, j') \text{ tal que } \beta_j \neq \beta_{j'}. \end{cases}$$

Este contraste tiene asociado el estadístico:

$$F_0 = \frac{CM_T}{\widehat{\sigma}_2^2},$$

con $(t - 1)$ y $a(n - 1)(t - 1)$ grados de libertad.

Si no se verifican los supuestos para la matriz de varianzas-covarianzas, de forma análoga a los modelos anteriores se puede optar por dos soluciones alternativas:

1. Realizar un análisis univariante “aproximado” aplicando la prueba épsilon a los grados de libertad del estadístico F_0 .
2. Aplicar un análisis de varianza multivariado (o MANOVA) con dos factores A y T .

Comparaciones múltiples de medias.

Si se cumplen los supuestos en las matrices de varianzas-covarianzas, y una vez realizado el Análisis de la Varianza en este modelo aparecen diferencias significativas, el siguiente objetivo será localizar donde se encuentran esas diferencias. Se considerarán dos situaciones en función de si existe o no interacción significativa entre los factores A y T .

Si no existe interacción significativa nos interesa comparar los niveles del factor A , cuando existan diferencias significativas entre ellos, y los niveles del factor T cuando se dé para él esta situación.

Para comparar los niveles del factor A se estudia las diferencias estimadas \bar{y}_i . Teniendo en cuenta la expresión del modelo asociado al diseño se obtiene un estimador insesgado de la diferencia de medias, que es:

$$\bar{y}_i - \bar{y}_{i'} = \widehat{\mu}_i - \widehat{\mu}_{i'},$$

que tiene por varianza:

$$\frac{2}{vn}(\sigma_2^2 + t\sigma_1^2),$$

y por estimador del error:

$$\hat{\sigma}(\bar{y}_i - \bar{y}_{i'}) = \sqrt{\frac{2\hat{\sigma}_1^2}{tn}}.$$

Por tanto, según el método LSD de Fisher, al hacer comparaciones múltiples entre los niveles i y i' del factor A , la mínima diferencia significativa viene dada por:

$$LSD_\alpha = t_{a(n-1); \frac{\alpha}{2}} \sqrt{\frac{2\hat{\sigma}_1^2}{tn}}.$$

En el caso de en el que se comparan los niveles del factor T se deben estudiar las diferencias estimadas \bar{y}_j . Un estimador insesgado de las medias viene dado por:

$$\bar{y}_j - \bar{y}_{j'} = \hat{\mu}_j - \hat{\mu}_{j'},$$

cuya varianza es:

$$\hat{\sigma}(\bar{y}_j - \bar{y}_{j'}) = \sqrt{\frac{2\hat{\sigma}_2^2}{an}}.$$

Por lo tanto al hacer comparaciones múltiples entre los niveles j y j' del factor T , atendiendo al método LSD de Fisher, la mínima diferencia significativa viene dada por:

$$LSD_\alpha = t_{a(n-1)(t-1); \frac{\alpha}{2}} \sqrt{\frac{2\hat{\sigma}_2^2}{an}}.$$

Si no existe interacción significativa se pueden realizar dos tipos de análisis basados en medias de tratamientos μ_{ij} .

Uno de ellos consiste en comparar los niveles del factor T en un nivel i del factor A . Para ello se debe comparar los valores mu_{ij} , para i fijo, estimador por \bar{y}_{ij} . Un estimador insesgado de la diferencias de medias para el nivel i del factor A es:

$$\bar{y}_{ij} - \bar{y}_{ij'} = \hat{\mu}_{ij} - \hat{\mu}_{ij'},$$

cuya varianza es:

$$\frac{2\hat{\sigma}_2^2}{n},$$

y el estimador del error estándar de esta diferencia es:

$$\hat{\sigma}(\bar{y}_{ij} - \bar{y}_{ij'}) = \sqrt{\frac{2\hat{\sigma}_2^2}{n}}.$$

Conforme al método LSD de Fisher, para hacer comparaciones múltiples entre los niveles j y j' del factor T para el nivel i del factor A , la mínima diferencia significativa viene dada por:

$$LSD_{\alpha} = t_{a(n-1)(t-1); \frac{\alpha}{2}} \sqrt{\frac{2\widehat{\sigma}_2^2}{n}}.$$

El otro tipo de análisis consiste en comparar los niveles del factor A en un nivel j del factor T . Para ello, se debe comparar los valores de μ_{ij} , para j fijo, estimados por \bar{e}_{ij} . Un estimador insesgado de las diferencias de medias para el nivel j del factor T es

$$\bar{y}_{ij} - \bar{y}_{i'j} = \widehat{\mu}_{ij} - \widehat{\mu}_{i'j},$$

cuya varianza es

$$\frac{2}{n}(\sigma_1^2 + \sigma_2^2).$$

Un estimador insesgado de $\sigma_1^2 + \sigma_2^2$ es:

$$\widehat{\sigma_1^2 + \sigma_2^2} = \frac{SC_{SA} + SC_{Error}}{a(n-1) + a(n-1)(t-1)} = \frac{\widehat{\sigma}_1^2 + (t-1)\widehat{\sigma}_2^2}{t}.$$

La distribución de este estimador no es una χ^2 sino una combinación lineal de χ^2 con lo que no hay un valor LSD exacto para estas comparaciones. Por el método de los momentos se puede calcular aproximadamente el valor LSD:

$$LSD_{\alpha} = t_{\frac{\alpha}{2}}^* \sqrt{\frac{2(\widehat{\sigma_1^2 + \sigma_2^2})}{n}},$$

siendo $t_{\frac{\alpha}{2}}^*$ una media ponderada de $t_{a(n-1); \frac{\alpha}{2}}$ y $t_{a(n-1)(t-1); \frac{\alpha}{2}}$ de la forma:

$$t_{\frac{\alpha}{2}}^* = \frac{t_{a(n-1); \frac{\alpha}{2}} \widehat{\sigma}_1^2 + t_{a(n-1)(t-1); \frac{\alpha}{2}} (t-1) \widehat{\sigma}_2^2}{\widehat{\sigma}_1^2 + (t-1) \widehat{\sigma}_2^2}.$$

Si no se cumplen los supuestos en las matrices de varianzas-covarianzas y en el análisis aparecen diferencias significativas entre los tratamientos se realizan los siguientes análisis:

1. Comparaciones dos a dos de los niveles del factor intra-sujetos T mediante el test t para dos muestras relacionadas controlando el error experimental por el procedimiento de Bonferroni.

2. Comparaciones dos a dos de los niveles del factor inter-sujetos A en cada nivel del factor intra-sujetos T mediante Análisis de la Varianza del factor inter-sujetos A en cada

nivel del factor intra-sujetos T .

3. Contrastes del factor inter-sujetos A repetidos en el factor intra-sujetos T mediante análisis univariante del factor inter-sujetos A entre dos niveles consecutivos del factor intra-sujetos T .

3.2. Aplicación práctica

La práctica del fútbol conlleva una gran exigencia física que tiene como consecuencia lesiones en los jugadores. Para evitar este tipo de lesiones, los especialistas en rehabilitación proponen llevar a cabo programas de ejercicios específicos para mejorar el equilibrio y con ello la propiocepción.

El equilibrio se define generalmente como la capacidad de mantener el centro de gravedad del cuerpo dentro de su base de apoyo y se diferencia entre equilibrio estático y equilibrio dinámico. El equilibrio estático es la capacidad de mantener el cuerpo en equilibrio mientras no se está realizando ningún movimiento. El equilibrio dinámico requiere la capacidad de mantener el equilibrio durante la transición del movimiento a un estado estático.

Es importante diferenciar la propiocepción y el equilibrio. La propiocepción es la recepción de estímulos producidos dentro organismo, mientras que el equilibrio es equilibrio físico. Esto significa que la propiocepción es un proceso neurológico, mientras que el equilibrio es la capacidad de permanecer en una posición vertical. Los ejercicios de equilibrio van destinados a desarrollar la propiocepción, consisten en entrenar el cerebro para reconocer la posición de cada parte del cuerpo en cada momento. El objetivo de los ejercicios de equilibrio debe ser reducir el tiempo entre estímulos neuronales y respuesta muscular.

Por otra parte, es importante que los programas de ejercicios equilibrio ayuden a la mejor de la propiocepción, no sólo durante la fase de rehabilitación, sino también durante el período de competición. Es decir, el desarrollo del equilibrio puede proteger a los futbolistas de posibles lesiones futuras.

En Cummins, Orr, O'Connor, y cols (2013) se compararon dos programas de equilibrio con diferente frecuencia de aplicación para determinar su eficacia en la mejora de la propiocepción de las extremidades inferiores: un programa de balance diario que duró tres semanas o un programa que se aplicó tres por semana durante un período de seis semanas.

3.2.1. ¿Cómo afecta la mejora del equilibrio en la reducción del número de lesiones?

Wedderkopp, Kaltof, Lundgaard, Rosendahl, y Froberg (s.f.) tiene como objetivo aplicar un programa de entrenamiento en el que se incluye el uso de la tabla de equilibrio durante 10-15 minutos cada vez que se llevara a cabo el entrenamiento. Este programa se diseñó para reducir el número de lesiones en la zona inferior de las extremidades y se tuvo en cuenta la incidencia de lesiones y los factores de riesgo entre el grupo que llevo a cabo el entrenamiento y el que no. Como conclusión se obtuvo que el grupo que llevaba a cabo el entrenamiento redujo el número de lesiones y que por lo tanto el entrenamiento de equilibrio ayudó reducir el número de lesiones.

Por otro parte, una de las lesiones más comunes en el fútbol es la lesión en el ligamento cruzado anterior (LCA). Según Alentorn-Geli y cols. (2009) los ejercicios de equilibrio y propiocepción constituyen una parte importante en el entrenamiento a la hora de reducir los factores de riesgo de lesión del LCA sin contacto.

Se puede concluir que una mejora en el equilibrio conlleva una reducción en el número de lesiones, por lo tanto se puede considerar que el equilibrio es un factor de riesgo para las lesiones. Por lo tanto, los resultados obtenidos en este estudio, relativos a como mejora el equilibrio y la propiocepción en cada grupo de estudio, en función a que programa de ejercicios lleva a cabo cada uno, ayudará a saber que programa es más eficiente para mejorar el equilibrio y con ello reducir el número de lesiones.

3.2.2. Material y métodos

Teniendo en cuenta que siempre hay una necesidad de mejorar la propiocepción a través de programas de ejercicios específicos, en este estudio se examina qué frecuencia de realización de ejercicios del balance sería más eficaz en la mejora de la capacidad de equilibrio:

- a) seis veces a la semana durante un período de tres semanas (grupo A).
- b) tres veces a la semana durante un período de seis semanas (grupo B).

En otras palabras, los dos grupos experimentales siguieron el mismo protocolo (20 minutos), pero en diferentes frecuencias de ejercicio. El grupo C no siguió un entrenamiento del equilibrio específico, sólo un entrenamiento de fútbol estándar y se estableció como grupo de control. El grupo de control se utilizó con el fin de comprobar si el entrenamiento de fútbol por sí solo puede mejorar la capacidad de equilibrio significativamente. La cuestión principal del estudio fue, cuál es la frecuencia de ejercicio que va a ser más eficaz para mejorar el control postural y entrenar el cerebro para reconocer la posición de cada parte del cuerpo, el que se aplicó diariamente durante un período de 3 semanas, o el otro que se aplicó 3 veces a la semana durante un período de 6 semanas.

■ Participantes

Los sujetos tenían una edad media de $22,7 \pm 3,5$ años, un peso medio de $76,2 \pm 4,9$, y una altura media de $1,79 \pm 6.36$. Ninguno de los sujetos estaban participando en cualquier otra actividad física, salvo el entrenamiento de fútbol y el entrenamiento del equilibrio. Además, estaban libres de lesiones en sus extremidades inferiores y no tenía la inestabilidad mecánica o funcional en su rodilla o el tobillo en ese período, por lo que estaban participando regularmente en los entrenamientos.

Se determinó la pierna dominante a través de la prueba de tiro pelota. Se pidió a los jugadores patear un balón de fútbol lo más fuerte posible, la pierna que se utiliza para lanzar el balón se registró como la pierna dominante.

■ Medidas

Los equipos de prueba que se seleccionaron para evaluar a los tres grupos consistían en un sistema electrónico de estabilidad (sistema de estabilidad Biodex), una tabla de equilibrio de madera con superficie inferior hemisférica, y un cronómetro.

■ Las desviaciones del plano horizontal

El sistema de estabilidad Biodex es un sistema de evaluación de la estabilidad postural que evalúa la capacidad del cuerpo para mantener el equilibrio sobre una plataforma inestable. Todos los sujetos completaron una evaluación del equilibrio dinámico con ambas piernas (dominantes y no dominantes). Para cada ensayo de prueba, se pidió a los sujetos que se

colocaran en la plataforma sobre una sola pierna, con los dos brazos en el pecho y con la extremidad que no apoyaba en una posición cómoda, para no ponerla en contacto con la plataforma de pruebas. La posición de prueba elegida por el sujeto se utilizó para todos los ensayos y recolección de datos. Se dieron instrucciones a los sujetos para centrarse en una pantalla de información visual directamente en frente de ellos.

Los sujetos trataron de mantener la plataforma de equilibrio inestable en la posición horizontal. Cualquier desviación de la plataforma de equilibrio se registró numéricamente, en grados. Más específicamente, el sistema proporciona tres índices diferentes de acuerdo con la dirección de las desviaciones desde el plano horizontal; el índice total estabilidad (SI), el índice anterior-posterior (API) y el índice medio-lateral (MLI). Los participantes realizaron el nivel de estabilidad 1 (menos estable) y efectuaron tres ensayos de prácticas de 20 segundos y tres ensayos de prueba 20 segundos de los cuales sólo se registró la mejor puntuación. La fiabilidad se estableció con coeficientes de correlación intra-clase (CCI) para los jugadores de fútbol. Los CCI para el primer objetivo fue de 0,80, para el tercero de 0,77 y para el sexto de 0,67.

■ Estudio

Los 38 jugadores de fútbol fueron asignados al azar en 3 grupos, 2 grupos experimentales (13 participantes cada uno) se sometieron al entrenamiento de fútbol y el programa de equilibrio adicional y 1 grupo de control (12 participantes) no siguieron un entrenamiento del equilibrio altamente específico, solo un entrenamiento estándar de fútbol. El procedimiento de ensayo fue seguido de un período de entrenamiento para los 2 grupos (por un periodo de entrenamiento de 3 semanas para el grupo A y de 6 semanas para el grupo B). La prueba post-formación se realizó al final de la tercera semana para el grupo A, y al final de la sexta semana para el grupo B y C.

Descripción del programa de ejercicio de equilibrio: El programa de ejercicios fue diseñado usando los principios de los programas de entrenamiento de un futbolista de élite y los destinados a la rehabilitación de deportistas lesionados con inestabilidad funcional de sus tobillos o ruptura del ligamento cruzado anterior. El objetivo principal fue incluir ejercicios que mejoraran el conocimiento y control de la rodilla al estar de pie, teniendo en cuenta el salto y el aterrizaje, que son elementos técnicos importantes de fútbol. Los ejercicios de equilibrio se realizaron durante 20 minutos en dos tablas de equilibrio de madera: una placa con superficie inferior semiesférica y la otra con la superficie inferior semicilíndrica.

3.2.3. Análisis

Se recogen los mejores valores de tiempo para cada prueba y se calculan las medias y las desviaciones típicas, como se muestran en Tabla 3-1 y Tabla 3-2.

	Grupo A		Grupo B		Grupo C	
	Antes	Tres semanas después	Antes	Seis semanas después	Antes	Seis semanas después
Desviaciones	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$
SI dom	9,6 \pm 2,4	6,1 \pm 0,8*	9,0 \pm 2,2	7,6 \pm 1,6*	9,2 \pm 2,4	8,7 \pm 1,5
SI no-dom	8,1 \pm 2,5	6,2 \pm 2,4*	8,4 \pm 3,0	6,6 \pm 2,0*	8,5 \pm 3,1	8,6 \pm 2,1
API dom	7,7 \pm 2,3	5,6 \pm 1,1*	8,2 \pm 3,1	6,3 \pm 1,6*	8,1 \pm 3,3	8,1 \pm 1,5
API no-dom	6,9 \pm 1,9	5,3 \pm 2,1*	8,0 \pm 2,8	6,4 \pm 1,8*	8,4 \pm 2,9	8,2 \pm 1,9
MLI dom	4,9 \pm 1,2	3,7 \pm 1,0*	5,4 \pm 1,3	3,5 \pm 0,9*	5,3 \pm 1,5	5,2 \pm 0,8
MLI no-dom	4,4 \pm 1,2	3,3 \pm 1,2*	4,9 \pm 1,6	3,1 \pm 0,8*	4,7 \pm 1,4	4,8 \pm 0,9

Tabla 3-1.: Medias (μ), desviaciones típicas ($\pm SD$), y total (Si), anterior-posterior (API) y medio-lateral (MLI), dominante (dom) , no-dominante (no-dom) extremidad.

	Grupo A		Grupo B		Grupo C	
	Antes	Tres semanas después	Antes	Seis semanas después	Antes	Seis semanas después
Desviaciones	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$	$\mu \pm SD$
APM dom	3,6 \pm 2,4	10,6 \pm 7,3*	2,5 \pm 0,7	9,6 \pm 3,1*	2,6 \pm 0,7	2,6 \pm 1,1
APM no-dom	3,0 \pm 1,6	7,1 \pm 2,2*	2,6 \pm 1,8	11,4 \pm 6,7*	2,8 \pm 1,8	2,4 \pm 2,1
MLI dom	2,7 \pm 1,6	15,2 \pm 8,1*	3,3 \pm 1,7	17,1 \pm 12,1*	3,3 \pm 1,7	3,1 \pm 1,1
MLI no-dom	3,0 \pm 1,6	17,8 \pm 7,8*	3,5 \pm 1,8	19,3 \pm 14,2*	3,5 \pm 1,8	3,3 \pm 1,2

Tabla 3-2.: Medias (μ), desviaciones típicas ($\pm SD$) tiempo de mantenimiento en la tabla de equilibrio para los movimientos anterior-posterior (APM), y los movimientos medio-lateral (MLM), para la pierna dominante (dom) y la no-dominante (no-dom).

Se realiza un ANOVA de medidas repetidas con un factor intra-sujetos y un factor inter-

sujeto. El factor intra-sujetos (variable dependiente) que consta de la pierna dominante y la no dominante. El factor inter-sujeto (variable independiente) consta de cada uno de los tres grupos en los que se dividió a los futbolistas: grupo A (13 jugadores), grupo B (13 jugadores) y grupo C (12 jugadores).

Variable a estudiar o variable respuesta:

Y = mejora del control postural y capacidad del cerebro para reconocer cada parte del cuerpo.

Cada componente de la variable Y viene dada por y_{ijk} , donde i representa el i -ésimo nivel del factor inter-sujeto, j representa el j -ésimo nivel del factor intra-sujeto y k el k -ésimo individuo.

El modelo matemático asociado es:

$$\begin{aligned}
 y_{ijk} &= \mu + S_{k(i)} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}; \\
 \text{con } i &= 1, 2, \dots, a \quad j = 1, 2, \dots, t \quad k = 1, 2, \dots, n; \\
 \sum_{i=1}^a \alpha_i &= 0; \quad \sum_{j=1}^t \beta_j = 0; \quad \sum_{j=1}^t (\alpha\beta)_{ij} = 0 \quad \forall i; \quad \sum_{i=1}^a (\alpha\beta)_{ij} = 0 \quad \forall j; \\
 S_{k(i)} &\sim N(0, \sigma_1); \quad e_{ij} \sim N(0, \sigma_2) \\
 S_{k(i)} \text{ y } e_{ij} &\text{ son mutuamente independientes.}
 \end{aligned}$$

Donde μ la media poblacional de la variable respuesta, α_i el efecto medio adicional debido al grupo en el que se encuentra el jugador $S_{k(i)}$ el efecto aleatorio del jugador k en el grupo en el que se encuentra, β_j el efecto adicional debido a si estudiamos la pierna dominante o la no dominante, $(\alpha\beta)_{ij}$ el efecto medio adicional debido al grupo en el que se encuentra el jugador junto con la pierna que se estudia (la dominante o la no dominante) y e_{ijk} representa el error aleatorio.

Una vez comprobados los supuestos de independencia, normalidad y esfericidad se plantean los siguientes contrastes de hipótesis:

- Efectos de la interacción.

$$\begin{cases} H_0 : (\alpha\beta)_{ij} = 0; \\ H_1 : \text{Existe al menos un par } (i, j) \text{ tal que } (\alpha\beta)_{ij} \neq 0. \end{cases}$$

Para el cual se obtiene un p -valor < 0.05 con lo que se concluye que se detectaron diferencias en cada grupo para cada extremidad en la mejora del equilibrio.

- Efectos del factor inter-sujetos.

$$\begin{cases} H_0 : \alpha_i = 0 \quad i = 1, \dots, t; \\ H_1 : \text{Existe al menos un par } (i, i') \text{ tal que } \alpha_i \neq \alpha_{i'}. \end{cases}$$

Para el cual se obtiene un p -valor > 0.05 con lo que se concluye que no existen diferencias significativas entre los grupos A, B y C en la mejora del equilibrio.

- Efectos del factor intra-sujetos.

$$\begin{cases} H_0 : \beta_j = 0 \quad j = 1, \dots, v; \\ H_1 : \text{Existe al menos un par } (j, j') \text{ tal que } \beta_j \neq \beta_{j'}. \end{cases}$$

Tanto en las pruebas de estabilidad Biodex como en las tablas de equilibrio se obtuvo un p -valor < 0.05 para la pierna dominante y la no dominante. Por tanto se concluye que existen mejoras significativa para ambas piernas.

4. Modelo de Regresión de Cox

“No hay nada más peligroso que no arriesgarse”

Pep Guardiola (1971-)

En este último capítulo se trata el modelo de Regresión de Cox para estudiar los factores de riesgo que afectan a lesiones en una parte concreta del cuerpo: el muslo.

4.1. Desarrollo teórico

En esta sección nos centraremos en el análisis de supervivencia, en concreto en la regresión de Cox. Primero se expondrán algunos conceptos básicos del análisis de supervivencia para luego profundizar en la regresión de Cox.

El análisis de supervivencia es una técnica inferencial que consiste en seguir a una serie de individuos y modelar el tiempo que transcurre desde que comienza el estudio, punto inicial, hasta que ocurre el suceso de interés, punto final. La variable de estudio es el tiempo de supervivencia.

Puede ocurrir que haya individuos que desaparezcan del estudio, que entren en él. También puede que el estudio finalice antes de que suceda el evento de interés. En estos casos se habla de observaciones o individuos censurados.

A continuación se presentan varios tipos de censura:

- **Censura de tipo I:** Sea un conjunto de individuos y un intervalo de tiempo fijado $[0, t_f]$. El conjunto de observaciones censuradas estará formado por los individuos que no presentan fallo en ese intervalo.
- **Censura de tipo II:** En este caso se fija el final del estudio en el momento en el que ocurren r fallos de n posibles ($r < n$). Por lo tanto habrá $n - r$ observaciones

censuradas.

- **Censura de tipo aleatorio:** Este tipo puede ocurrir por diversas causas que no están influenciadas por la persona que realiza el estudio. Por ejemplo:
 - Abandono: el individuo se ve obligado a dejar el estudio (por ejemplo un futbolista que abandona el club).
 - Fin del ensayo: termina el estudio y no se produce el suceso de interés.
- **Censura por la izquierda:** El tiempo de supervivencia real es menor o igual al tiempo de supervivencia observado.
- **Censura por la derecha:** El tiempo de supervivencia real es igual o mayor que el tiempo de supervivencia observado, es decir, si el individuo presenta fallo sabemos que este ocurre después del tiempo de censura observado.
- **Censura por intervalo:** El tiempo real de supervivencia está dentro de un intervalo de tiempo conocido, pero se desconoce el tiempo exacto.

Sea T el tiempo de supervivencia, consideramos T como una variable aleatoria continua. Denotaremos $f(x)$ como su función de densidad y $F(x)$ como su función de distribución.

A continuación se exponen funciones que intervienen en el análisis de supervivencia.

Función de supervivencia

La función de supervivencia es la probabilidad de que el sujeto estudiado sobreviva pasado un tiempo t .

$$S_T(t) = P[T > t] = 1 - F_T(t) \text{ para todo } t > 0$$

donde

$$F_T(t) = P[T \leq t] = \int_0^t f(x)dx.$$

Función de riesgo

La función de riesgo es la probabilidad de que un sujeto experimente el suceso de interés durante un intervalo infinitesimal de tiempo dado que el individuo ha sobrevivido hasta el comienzo de ese intervalo. La expresión matemática de la función de riesgo es

$$\begin{aligned} h_T(t) &= \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < (t + \Delta t) | T \geq t]}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{F_T(t + \Delta t) - F_T(t)}{\Delta t} = \frac{f_T(t)}{S_T(t)}. \end{aligned}$$

Tasa de fallo acumulado

La tasa de fallo acumulado se define integrando la función de riesgo.

$$H_T(t) = \int_0^t h_T(u) du.$$

Por lo tanto, la función de riesgo es la derivada, o pendiente, de la función de riesgo acumulado. La función de riesgo acumulado se relaciona con la función de supervivencia acumulada por la expresión

$$S_T(t) = e^{-H_T(t)} \quad \text{o} \quad H_T(t) = -\ln(S_T(t)).$$

Tiempo esperado de vida

El tiempo esperado de vida se refiere a la esperanza matemática de T , siempre que exista,

$$E[T] = \int_0^{\infty} t f(t) dt.$$

Vida media residual

La vida media residual es una función que mide la esperanza de vida restante en los individuos de edad t o el tiempo esperado de vida después de t , hasta que ocurre el fallo.

$$vmr(t) = E[T - t | T > t] = \frac{1}{S(t)} \int_t^{\infty} S(x) dx,$$

para todo $t > 0$.

4.1.1. Estimador de Kaplan-Meier

En el análisis de supervivencia se puede distinguir entre métodos paramétricos y métodos no paramétricos.

Entre los métodos no paramétricos, uno de los métodos más utilizados es el estimador de Kaplan-Meier.

La finalidad del estimador de Kaplan-Meier es estimar la probabilidad de que un conjunto de individuos sobreviva en un intervalo de tiempo dado. El interés de este método reside en que tiene en cuenta las observaciones o individuos censurados para estimar las funciones antes mencionadas.

En el caso en el que no haya observaciones censuradas, se estima la función de supervivencia mediante:

$$S_n = \frac{1}{n} \text{Card}\{i : t_i > t\}$$

donde n es el tamaño de la muestra.

Supongamos ahora que existen observaciones censuradas y se considera la censura a la derecha. Se observan los pares formador por (Y_i, δ_i) $i = 1, \dots, n$, donde,

$$Y_i = \min\{T_i, C_i\}, \quad \delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \quad (\text{Observado}); \\ 0 & \text{si } T_i > C_i \quad (\text{Censurado}). \end{cases}$$

Se construye un conjunto de $n' \leq n$ intervalos usando los tiempos esperados observados.

$$I_i = (y_{(i-1)}, y_{(i)}) \text{ con } i=1, \dots, n'.$$

Sean los parámetros:

n_i = número de individuos vivos al inicio del intervalo $(y_{(i-1)}, y_{(i)})$ con $i=1, \dots, n'$,

d_i = número de individuos que fallecen en el intervalo $(y_{(i-1)}, y_{(i)})$ con $i=1, \dots, n'$,

$$p_i = P[T > y_{(i)} | T > y_{(i-1)}],$$

$$q_i = P[T \leq y_{(i)} | T > y_{(i-1)}].$$

A continuación se muestran los estimadores de Kaplan-Meier para:

■ **Función de supervivencia**

Aplicando de forma reiterada la probabilidad condicionada se obtiene la siguiente expresión para $S_T(t)$.

$$S_T(t) = P[T > t] = \prod_{i:y_{(i)} \leq t} p_i$$

Por tanto considerando los estimadores:

$$\hat{q}_i = \frac{d_i}{n_i}, \quad \hat{p}_i = 1 - \hat{q}_i = \frac{n_i - d_i}{n_i},$$

se obtiene el estimador no paramétrico de la función de supervivencia propuesto por Kaplan-Meier (1958)

$$\hat{S}_T(t) = \prod_{i:y_{(i)} \leq t} \hat{p}_i = \prod_{i=1}^k \frac{n_i - d_i}{n_i},$$

donde $k = y_{(k)} \leq t < y_{(k+1)}$.

■ **Función de riesgo**

La estimación de la tasa de fallo instantánea mediante el estimador de Kaplan-Meier en un intervalo, es el siguiente:

$$\hat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)},$$

donde $i : t_i \leq t < t_{i+1}$.

■ **Estimación del fallo acumulado**

Estimación de fallo acumulada mediante los estimadores de Kaplan-Meier:

$$\hat{H}_T(t) = -\ln[\hat{S}_T(t)] = -\ln \prod_{i:y_{(i)} \leq t} \frac{n_i - d_i}{n_i},$$

$$\hat{V}[\hat{H}_T(t)] = \sum_{i:y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

- **Estimación del tiempo medio de vida truncada**

Habitualmente, el tiempo medio de vida se estima a través de:

$$\hat{E}[T] = \int_0^{y_{(n)}} \hat{S}(t) dt$$

Pero en el caso de $y_{(n)} = \max(y_i)$ no esté censurada, la integral anterior coincide con la que obtenemos en el intervalo $[0, \infty)$; pero en el caso de que esté censurada, el $\lim_{t \rightarrow \infty} \hat{S}_T(t) \neq 0$, por lo que la integral en el intervalo definido anteriormente no está definida. Para intentar evitar este problema, consideramos $y_{(n)}$ como si no estuviera censurada obteniendo un estimador de la media, el cual coincide con la curva de Kaplan-Meier:

$$\hat{E}[T] = \sum_{i=1}^{n'} [y_{(i)} - y_{(i-1)}] \hat{S}(y_i)$$

4.1.2. Modelo de Cox

En el análisis de supervivencia el modelo de regresión más utilizado es el modelo paramétrico de Cox, también denominado como modelo de riesgos proporcionales.

Este procedimiento puede ser utilizado para estudiar el impacto de varios factores sobre la supervivencia, es decir, la regresión de Cox modela la relación entre un conjunto de una o más covariables y la función de riesgo y por tanto sobre la función de supervivencia. Las covariables pueden ser discretas o continuas. El modelo de regresión de riesgos proporcionales de Cox se resuelve utilizando el método de probabilidad marginal.

Se consideran las covariables x_1, \dots, x_p y β_1, \dots, β_p los coeficientes de regresión a estimar. Además, t representa el tiempo transcurrido, y h_0 la tasa de riesgo de referencia cuando todas las covariables son iguales a cero.

- **Función de riesgo**

La función de riesgo para este modelo es:

$$h(t, \mathbf{x}) = \int_0^t h_T(u, x) du = \int_0^t h_0(u) e^{\sum_{i=1}^p x_i \beta_i} du$$

$$= e^{\sum_{i=1}^p x_i \beta_i} \int_0^t h_0(u) du = H_0 e^{\sum_{i=1}^p x_i \beta_i},$$

Hay que notar que para este modelo la función de fallo base h_0 no es conocida.

■ Función de supervivencia

Bajo el modelo de regresión de riesgos proporcionales, la supervivencia acumulada es

$$\begin{aligned} S_T(t, x) &= e^{-h_T(t)} = \exp \left[-h_0(t) e^{\sum_{i=1}^p x_i \beta_i} \right] \\ &= \left[e^{-h_0(t)} \right] e^{\sum_{i=1}^p x_i \beta_i} = S_0(t) e^{\sum_{i=1}^p x_i \beta_i}, \end{aligned}$$

Estimación

Sea $t = 1, \dots, M$ el subíndice de los M tiempos de fallos completos sin censuras. Hay que tener en cuenta que M no incluye los tiempos duplicados o las observaciones censuradas. El conjunto de todas los fallos que se producen en el momento T_t se denotan como D_t . Sean c y $d = 1, \dots, m_t$ índice de individuos de D_t . El conjunto de todos los individuos que están en riesgo inmediato antes del tiempo T_t se denota como R_t . Este conjunto, a menudo se llama conjunto de riesgos e incluye todos los individuos que fracasan en el tiempo T_t así como aquellos que están censurados o no en un momento posterior a T_t . Sea $r = 1, \dots, n_t$ índice de individuos de R_t . Sea X referido a un conjunto de p covariables. Estas covariables tienen como subíndices i, j , o k . Los valores de las covariables en un tiempo de fallo en particular T_d se escriben como $x_{1d}, x_{2d}, \dots, x_{pd}$ o x_{id} en general. Los coeficientes de regresión a estimar son $\beta_1, \beta_2, \dots, \beta_p$.

Teniendo en cuenta lo anterior, cuando no hay relación entre los tiempos de fallo, el logaritmo de verosimilitud viene dado por:

$$LL(\beta) = \sum_{t=1}^M \left[\sum_{i=1}^p x_{it} \beta_i - \ln \left(\sum_{r \in R_t} e^{\sum_{i=1}^p x_{ir} \beta_i} \right) \right] = \sum_{t=1}^M \left[\sum_{i=1}^p x_{it} \beta_i - \ln(G_{R_t}) \right], \quad (4-1)$$

donde

$$G_R = \sum_{r \in R} e^{\sum_{i=1}^p x_{ir} \beta_i}.$$

Maximizando (4-1) se obtiene los estimadores de β . Para ello se aplica el método de Newton-Raphson, el cual usa las derivadas parciales y la resolución del sistema:

$$\hat{\beta} : \frac{\partial LL(\beta)}{\partial \beta} = 0.$$

En los casos en los que coinciden los tiempos de fallo se suele aplicar el método de Breslow o el método de Efron.

El logaritmo de verosimilitud de aproximación de Breslow viene dado por:

$$\begin{aligned} LL(\beta) &= \sum_{t=1}^M \left[\sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - m_t \ln \left(\sum_{r \in R_t} e^{\sum_{i=1}^p x_{ir} \beta_i} \right) \right] \\ &= \sum_{t=1}^M \left[\sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - m_t \ln(G_{R_t}) \right], \end{aligned}$$

donde

$$G_R = \sum_{r \in R} e^{\sum_{i=1}^p x_{ir} \beta_i}.$$

El logaritmo de verosimilitud de aproximación de Efron viene dado por:

$$\begin{aligned} LL(\beta) &= \sum_{t=1}^M \left[\sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - \sum_{d \in D_t} \ln \left(\sum_{r \in R_t} e^{\sum_{i=1}^p x_{ir} \beta_i} - \frac{d-1}{m_t} \sum_{c \in D_t} e^{\sum_{i=1}^p x_{ic} \beta_i} \right) \right] \\ &= \sum_{t=1}^M \left[\sum_{d \in D_t} \sum_{i=1}^p x_{id} \beta_i - \sum_{d \in D_t} \ln \left(G_{R_t} - \frac{d-1}{m_t} G_{D_t} \right) \right]. \end{aligned}$$

En ambos casos se aplicaría el método de de Newton-Raphson para hallar el estimador de β .

Contraste de hipótesis e intervalos de confianza

Esta sección se centra en la realización de contraste de hipótesis y de intervalos de confianza.

Para contrastar la hipótesis nula $H_0 : \beta = \beta_0$ se pueden aplicar tres contrastes: el test de razón de verosimilitud, el test de Wald y contraste “score”. En caso de aceptar la hipótesis nula, se tiene que el estadístico sigue una distribución chi-cuadrado con p grados de libertad.

1. Test de razón de verosimilitud. Este test se define como:

$$2[\log(LL(\beta_0)) - \log(LL(\hat{\beta}))]$$

donde β_0 son los valores iniciales de los coeficientes y $\hat{\beta}$ es la solución después de ajustar el modelo.

2. Test de Wald. Proporciona un contraste por variables en vez de una medida de significación global. El estadístico de contraste se define mediante:

$$(\hat{\beta} - \beta_0)' \hat{\Sigma}^{-1} (\hat{\beta} - \beta_0)$$

donde $\hat{\Sigma}$ es la matriz de varianzas y covarianzas estimada.

3. Contraste “score”. Se define utilizando las derivadas del logaritmo de verosimilitud evaluada en la hipótesis nula ($H_0 : \beta = \beta_0$) y suponiendo que bajo la hipótesis nula el vector scores es:

$$X_s = \left(\frac{\partial LL(\beta_0)}{\partial \beta} \right)' \left(\frac{\partial^2 LL(\beta_0)}{\partial \beta \partial \beta'} \right) \left(\frac{\partial LL(\beta_0)}{\partial \beta} \right).$$

Por otro lado, para contrastar la hipótesis nula $H_0 : \beta = 0$ se usa el estadístico de Wald:

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{V}[\hat{\beta}_j]}}. \quad (4-2)$$

El intervalo de confianza para β_j viene dado por:

$$\hat{\beta}_j \pm z_{\frac{\alpha}{2}} \sqrt{\hat{V}[\hat{\beta}_j]}.$$

Interpretación

La interpretación del modelo de Cox no se hace directamente a través del coeficiente estimado sino del exponencial de la estimación del coeficiente estimado.

Para variables dicotómicas, el estimador $e^{\hat{\beta}}$ es la razón de riesgos (hazard ratio) y se interpreta como la cantidad de riesgo que se tiene con la presencia de cada covariable en relación a la ausencia del resto de las la covariables.

Para el caso de covariables continuas $e^{\hat{\beta}}$ representa la razón de riesgos (hazard ratio) al incrementar en una unidad la covariable.

4.2. Aplicación práctica

El estudio Venturelli, Schena, Zanolla, y Bishop (2011) investiga los factores de riesgo para las lesiones de futbolistas jóvenes mediante la regresión de Cox.

Las distensiones musculares son muy frecuentes entre los jugadores de fútbol, por ello este estudio se centrará particularmente en examinar qué factores (edad, categoría, posición en el campo, índice de masa corporal, etc.) podrían ser identificados como factores de riesgo para distensiones musculares en los jugadores jóvenes.

La identificación temprana de factores de riesgo, seguida de un control regular, puede ayudar a los médicos de los equipos y entrenadores a identificar a los jugadores de fútbol jóvenes que presentan un riesgo alto de lesión para que estos sigan programas de formación de prevención.

4.2.1. Métodos

El estudio contó con 96 jugadores en total, quienes participaban en el torneo de élite nacional de Italia. Los jugadores se dividieron en tres grupos de edad: CA (17-18 años), CB (15-16 años) y CC (13-14 años). En el grupo CA estaba compuesto por 33 jugadores, el CB por 31 y el CC por 32.

En primer lugar se realizó una encuesta sobre lesiones anteriores y como consecuencia se excluyeron a 7 jugadores antes de la primera sesión de prueba ya que se estaban sometiendo a tratamientos fisioterapéuticos para una lesión previa (censura a la izquierda). También se excluyó al portero de cada grupo debido a que ellos realizan una formación especializada para su posición en el campo.

El estudio se llevó a cabo desde agosto hasta mayo, es decir, un temporada completa. Este período se dividió en tres: los 40 primeros días, que corresponden a la pretemporada, los 170 días siguientes conformarían la segunda división y finalmente los últimos 130 días.

Se realizaron a los futbolistas ciertas pruebas y se tomaron los datos necesarios ayudándose de la tecnología GPS cuando fue necesario. El primer día de pretemporada se tomaron los datos de las pruebas de salto, el tercero los de la prueba de flexibilidad y el cuarto los de la prueba de resistencia. Estos datos se registraron en un formulario junto con la información de las lesiones anteriores (categoría, posición, gravedad y fecha).

El entrenador y el preparador físico de cada grupo registró el tiempo de exposición individual para cada jugador (horas de participación) durante todas las sesiones de entrenamiento y partidos (jugados y no jugados).

Todas las lesiones musculares se registraron como una lesión estándar durante toda la temporada por un fisioterapeuta experto. Para la evaluación específica de las distensiones musculares se utilizó una evaluación ecográfica del daño muscular para determinar la severidad de la lesión. Todos los jugadores que sufrieron una lesión que no fue una distensión muscular (lesión de grado dos), o una distensión menor, (lesión de grado uno), durante la temporada observada, se censuraron en el modelo estadístico de supervivencia al momento de la lesión: dos jugadores por distensiones menores, cuatro jugadores por contusiones, dos jugadores por inflamación del pubis y uno por un esguince de ligamento de rodilla. En consecuencia, el grupo se redujo a 84 jugadores que fueron seguido durante la temporada juvenil de fútbol, 28 de la categoría CA, 29 de la categoría CB, 27 de la categoría CC.

Se determinaron las medidas antropométricas (masa y altura), el índice de masa corporal (IMC, $\text{kg}\cdot\text{m}^{-2}$) y la densidad corporal. Para calcular la composición corporal, se utilizaron diferentes fórmulas dependiendo de la edad y la raza (caucásica, negra e hispánica) del jugador.

Antes de realizar las dos pruebas de salto todos los jugadores realizaron de calentamiento estándar: diez minutos de carrera lenta, evitando el estiramiento de las extremidades inferiores. Se realizaron tres saltos y se registró el mejor de ellos. Se probaron dos tipos de saltos. El primero fue un salto de potencia (SJ) que requería que los jugadores flexionaran sus rodillas hasta 90° , para detenerse dos segundos y luego extienden las rodillas y caderas para saltar lo mas alto posible (ver Figura 4-1).

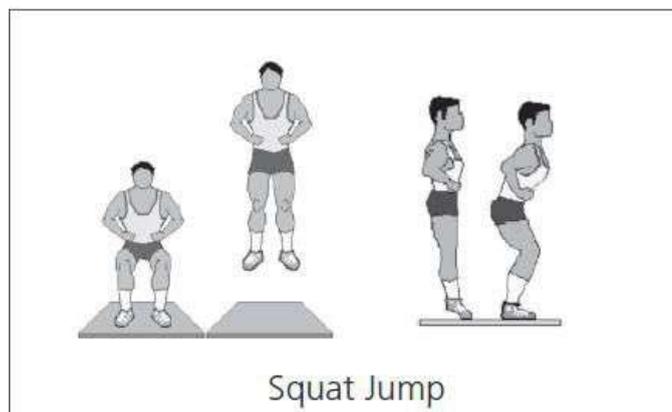


Figura 4-1.: SJ

El salto de contramovimiento (CMJ) fue realizado de la misma manera pero sin detenerse en la posición más baja (ver Figura 4-2).



Figura 4-2.: CMJ

La diferencia porcentual entre los dos tipos de saltos se calculó también para cada jugador:

$$\Delta JH = (CMJ - SJ) \cdot SJ^{-1} \cdot 100.$$

Para la prueba de flexibilidad (SR) se colocó una caja de SR en el suelo y los jugadores se sentaron sin zapatos con la piernas extendidas y los pies en flex pegados a la caja SR. Luego extendieron sus brazos hacia adelante, con las palmas hacia abajo, moviendo la escala de medición lo más lejos posible sin flexionar las rodilla (ver Figura 4-3). Se realizaron tres ensayos que fueron utilizados para análisis.

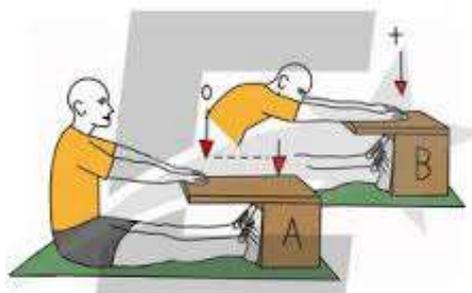


Figura 4-3.: CMJ

Para la prueba de resistencia, la sesión comenzó con un calentamiento consistente en diez minutos de carrera lenta. La prueba se consideró terminada cuando el jugador no pudo alcanzar dos veces la línea final.

4.2.2. Análisis

Se usó el análisis de supervivencia (regresión de Cox) para estimar la proporción de los jugadores sin distensiones en el muslo durante la temporada. La variable de estudio fue el tiempo que transcurre desde la evaluación hasta la ocurrencia de la lesión o hasta el final del estudio (censurado). Se censuraron todos los jugadores lesionados, incluyendo distensiones musculares o torceduras leves.

En los 40 primeros días 5 jugadores resultaron lesionados en el muslo en los 170 días siguientes 14 jugadores y en los 130 últimos 8. En total se registraron 27 jugadores con distensión muscular en el muslo, 6 en los cuádriceps, 7 los abductores y 14 en los isquiotiviales.

La curva de Kaplan-Meier (4.1.1) representa la probabilidad supervivencia durante la pretemporada y la temporada. En este caso la variable de estudio fue tiempo de evaluación (días) para la ocurrencia de una lesión (evento sin censura) o al final del seguimiento (censurado). Para los datos de este estudio la curva de Kaplan-Meier dada por la Figura 4-4

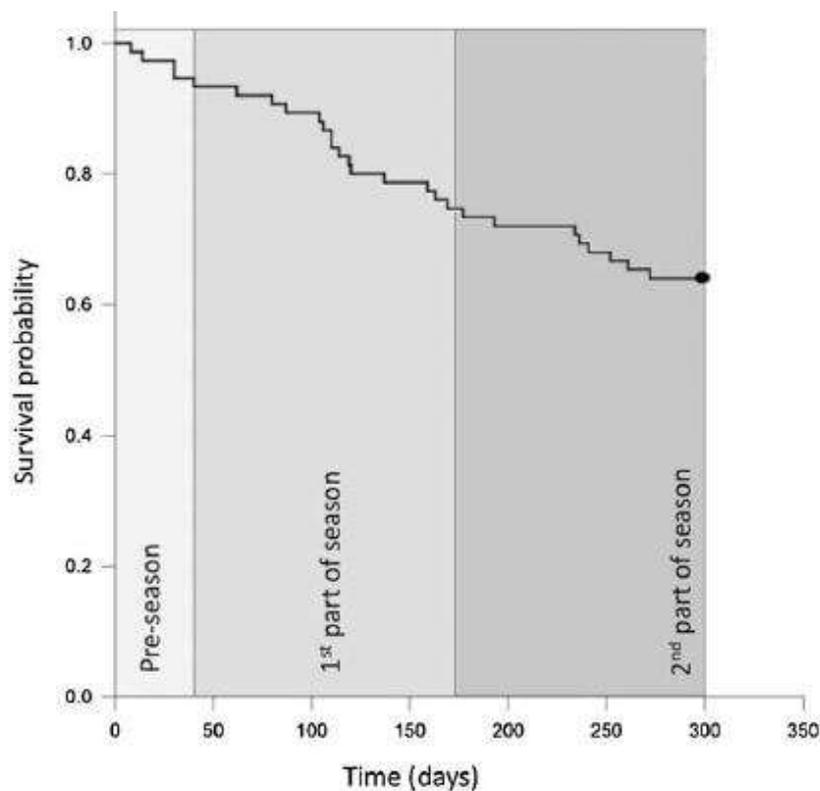


Figura 4-4.: Curva de Kaplan-Meier

Fuente: Venturelli y cols. (2011)

El objetivo del estudio fue evaluar dentro de un conjunto de covariables (los factores tales como la altura, la posición en el campo o el índice de masa corporal entre otros) cuáles tienen influencia sobre la función de riesgo (4.1.2). Como ya se vió en (4.1) la función de riesgo y la función de supervivencia están relacionadas, por tanto el estudio lleva a conocer como afectan los factores a la función de supervivencia (4.1.2)

Aplicando los contrastes de hipótesis explicados en (4.1.2) se observa que factores tales como una lesión anterior, la posición en el campo (defensores vs mediocampistas), la altura, el índice de masa corporal, el porcentaje de grasa, paso yo-yo, SJ y JH presentan una alta significación.

	Rango	Haz. ratio \pm SE	I.C.(95 %)	z
Lesiones previas (N)	0 a 13	3.55 \pm 1.46*	1.58-7.95	3.09
Partidos jugados (h)	24 a 68	0.977 \pm 0.02	0.95-1.01	-1.33
Entrenamientos (h)	150 a 314	0.996 \pm 0.00	0.99-1.00	-0.91
Entenamientos y partidos (h)	185 a 382	1.154 \pm 0.22	0.79-1.66	0.76
Categoría CA, CB o CC (N)	28, 29, 27	0.87 \pm 0.21	0.55-1.39	0.56
Edad (años)	14 a 18	1.07 \pm 0.13	0.85-1.34	0.55
Defensas vs delanteros (N)	35 a 14	0.34 \pm 0.16*	0.14-0.85	-2.31
Defensas vs mediocentros (N)	35 a 35	0.61 \pm 0.17	0.36-1.10	-1.66
Mediocentros vs delanteros	35 a 14	1.18 \pm 0.75	0.35-4.10	0.28
Alura (cm)	163 a 191	1.19 \pm 0.05*	1.11-1.29	4.58
Peso (kg)	47 a 88	1.16 \pm 0.03*	1.09-1.23	5.28
Índice de masa corporal (kgm^{-1})	18 a 26	1.59 \pm 0.18*	1.27-1.99	4.05
Porcentaje de grasa	6 a 16	1.37 \pm 0.11*	1.16-1.61	3.80
Paso yo-yo	9 a 15	0.67 \pm 0.09*	0.50-0.89	-2.79
SJ (cm)	26 a 45	1.13 \pm 0.06*	1.02-1.25	2.39
CMJ (cm)	28 a 51	0.99 \pm 0.04	0.92-1.08	-0.07
ΔJH (%)	-7 a 11	0.76 \pm 0.04*	0.68-0.84	-5.35
Flexibilidad	-11 a 16	0.98 \pm 0.03	0.92-1.05	-0.57

* p -value $<$ 0.05

SE se refiere al error estándar

Tabla 4-1.: Factores de riesgo intrínsecos y extrínsecos univariados que utilizan las regresiones de Cox para la probabilidad de supervivencia de la distensión del muslo

Tomando, por ejemplo, el factor peso, se tiene (de acuerdo a la Tabla 4-1) que los jugadores tienen un peso entre los 47 y los 88 kg. Como el peso es una variable continua, “Haz. ratio” representa la razón de riesgos al incrementar una unidad la covariable, en este caso al aumentar un kg la razón de riesgo aumenta $1,16 \pm 0,03$, donde 0.03 es el error estándar. (1.09,1.23) es el intervalo de confianza al 95 % y $z=5.28$ el valor del estadístico de Wald (4-2) que sirva para contrastar la hipótesis nula $H_0 : \beta_j = 0$.

El principal hallazgo de este estudio fue que, durante una temporada de fútbol (pretemporada + temporada de competición), no hubo efecto significativo de la exposición al entrenamiento y partidos en la probabilidad de distensión del muslo.

Con el fin de interpretar con precisión los resultados, se deben considerar algunos factores metodológicos. La principal limitación es el número relativamente bajo de casos y eventos. Sólo 84 de los 96 jugadores de fútbol jóvenes fueron seguidos durante la temporada, y sólo 27 incurrieron en una lesión muscular.

Los datos del presente estudio sugieren que la estatura del jugador es un fuerte factor de riesgo para las distensiones musculares.

A. Anexo

En este anexo se explican algunos tests usados en el trabajo.

Prueba de rachas

La prueba de rachas permite contrasta la hipótesis nula de que la muestra sea aleatoria, es decir, si las observaciones son independientes. Una racha es una secuencia de observaciones similares. Una muestra con un número demasiado grande o demasiado pequeño de rachas sugiere que la muestra no es aleatoria.

Test de Shapiro-Wilk

El test de Shapiro-Wilk contrasta la hipótesis nula de que la muestra de estudio esté normalmente distribuida. El estadístico del test viene dado por:

$$W = \frac{D^2}{nS^2},$$

donde D es la suma de las diferencias corregidas, n el tamaño muestral y S^2 es la varianza muestral.

Test de Mauchly

Este test comprueba la esfericidad de las matrices de varianzas-covarianzas. Este test usa estimación muestral de la matriz de varianzas-covarianzas poblacional, Σ , con el propósito de probar si $C^*'\Sigma C^*$ cumple el supuesto de esfericidad.

Test LSD de Fisher

El Test LSD (Least significant difference) de Fisher es un test de comparaciones múltiples. Se aplica cuando se rechaza la hipótesis nula de igualdad de medias en el ANOVA y en este caso permite comparar las medias de los t niveles de un factor.

El test se basa en crear un valor común usando el test t -Student. Este valor viene dado por:

$$T = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\hat{S}_E^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \equiv t_{N-t}$$

$$LSD = t_{N-t, \frac{\alpha}{2}} \sqrt{\hat{S}_E^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

donde N el número total de observaciones, t el número de niveles del factor, n_i, n_j tamaños muestrales de los niveles i y j , \hat{S}_E^2 la estimación de la varianza del error o residual, \bar{Y}_i, \bar{Y}_j las medias muestrales de los niveles i y j y $t_{N-t, \frac{\alpha}{2}}$ es la distribución t -student con $N - t$ grados de libertad y a un nivel de significación $\frac{\alpha}{2}$.

Si $|\bar{Y}_i - \bar{Y}_j| > LSD$, entonces hay diferencia significativa.

Test de Bonferroni

Al igual que el test LSD de Fisher es un test de comparaciones múltiples y se aplica cuando se rechaza la hipótesis nula de igualdad de medias en el ANOVA y en este caso permite comparar las medias de los t niveles de un factor.

Del mismo modo que en el test LSD de Fisher se calculará un valor, BSD , por encima del cual la diferencia entre las dos medias será significativa y por debajo del cual esa diferencia no será significativa. El valor BSD viene dado por:

$$BSD = t_{N-t, \frac{\alpha}{2M}} \sqrt{\hat{S}_E^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

donde N el número total de observaciones, t el número de niveles del factor, n_i, n_j tamaños muestrales de los niveles i y j , \hat{S}_E^2 la estimación de la varianza del error o residual, $t_{N-t, \frac{\alpha}{2M}}$ es la distribución t -student con $N - t$ grados de libertad y a un nivel de significación $\frac{\alpha}{2}$ y M son las combinaciones posibles de niveles tomados de dos en dos.

Test de rangos con signo de Wilcoxon

El test de rangos con signo de Wilcoxon es una prueba no paramétrica, la cual permite comparar poblaciones cuando sus distribuciones no satisfacen las condiciones necesarias

para otros test paramétricos.

Se utiliza para contrastar la hipótesis nula H_0 : No hay diferencias significativas entre las observaciones.

Referencias

- Alentorn-Geli, E., Myer, G. D., Silvers, H. J., Samitier, G., Romero, D., Lazaro-Haro, C., y Cugat, R. (2009). Prevention of non-contact anterior cruciate ligament injuries in soccer players. part 2: A review of prevention programs aimed to modify risk factors and to reduce injury rates. *Knee Surg Sports Traumatol Arthrosc*, *17*, 859-897.
- Aughey, R. J. (2011). Applications of gps technologies to field sports. *Int J Sports Physiol Perform.*, *6*(3), 295-310.
- Cummins, C., Orr, R., O'Connor, H., y cols. (2013). Global positioning systems (gps) and microtechnology sensors in team sports: A systematic review. *Sports Med.*, *43*, 1025-1042.
- Foster, C. (1998). Monitoring training in athletes with reference to overtraining syndrome. *Med. Sci. Sports Exer.*, *30*, 1164-1168.
- Foster, C., Daines, E., Hector, L., Snyder, A. C., y Welsh, R. (1996). Athletic performance in relation to training load. *Wisc. Med. J.*, *95*, 370-374.
- Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., . . . Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Clin J Sport Med*, *Volume 16*, *Number 2*, 97-106.
- Gabbett, T. J. (2003). Incidence of injury in semi-professional rugby league players. *Br. J. Sports Med.*, *37*, 36-43.
- Gabbett, T. J. (2004a). Influence of training and match intensity on injuries in rugby league. *J. Sports Sci.*, *22*, 409-417.
- Gabbett, T. J. (2004b). Reductions in pre-season training loads reduce training injury rates in rugby league players. *Br. J. Sports Med.*, *38*, 743-749.
- Gabbett, T. J. (2005). Changes in physiological and anthropometric characteristics of rugby league players during a competitive season. *J. Strength Cond. Res.*, *19*, 400-408.
- Gabbett, T. J., y Domrow, N. (2007). Relationships between training load, injury, and fitness in sub-elite collision sport athletes. *J. Sports Sci.*, *25*, 1507-1519.
- Hernanz, M. L. V., Daviña, P. G., Nieto, C., y Pérez, T. P. (2005). *Elaboración de material*

- didáctico de modelos avanzados de diseños de experimentos y sus aplicaciones*. (1.^a ed.). La Muralla, S.A.
- Jennings, D., Cormack, S., Coutts, A. J., y cols. (2010). The validity and reliability of gps units for measuring distance in team sport specific running patterns. *Int J Sports Physiol Perform.*, 5(3), 328-341.
- JGabbett, T., y Jenkins, D. G. (2011). Relationship between training load and injury in professional rugby league players. *J. Sci. Med. Sport.*, 26, 953-960.
- Johnston, R., Watsford, M., Pine, M., y cols. (2012). The validity and reliability of 5-hz global positioning system units to measure team sport movement demands. *J Strength Cond Res.*, 26(3), 758-765.
- Kampakis, S. (2016). *Predictive modeling of football injuries* (Tesis Doctoral no publicada). University College London.
- Kleinbaum, D., y Klein, M. (1996). *Survival analysis. a self-learning text*. Springer.
- Ochar, J. (1996). Orchard sports injury classification system (osics). *Sport Health*, 313:570-571.
- Peterson, L., Junge, A., Chomiak, J., y cols. (2000). Incidence of football injuries and complaints in different age groups and skill-level groups. *Am J Sports Med*, 28(Suppl), S51-S57.
- Peña, D. (2002). *Análisis de datos multivariantes* (1.^a ed.). McGraw-Hill.
- Portas, M., Rush, C., Barnes, C., y cols. (2009). Method comparison of linear distance and velocity measurements with global positioning satellite (gps) and the timing gates techniques. *J Sci Med Sport.*, 4, 381-393.
- Portas, M., Rush, C., Barnes, C., y cols. (2010). The validity and reliability of 1-hz and 5-hz global positioning systems for linear, multidirectional, and soccer-specific activities. *Int J Sports Physiol Perform.*, 5, 448-458.
- Schutz, Y., y Chambaz, A. (1997). Could a satellite-based navigation system (gps) be used to assess the physical activity of individuals on earth? *Eur J Clin Nutr.*, 51(5), 338-339.
- Stewart, A. M., y Hopkins, W. G. (2000). Seasonal training and performance of competitive swimmers. *J. Sports Sci.*, 18, 873-884.
- Twist, C., Waldron, M., Worsfold, P., y Gabbett, T. J. (2013). Monitoring team sport athletes using global positioning system technology. *Nova Science Publishers, Inc.*, 25, 95-113.
- Varley, M., y Aughey, R. (2012). Validity and reliability of gps for measuring instantaneous velocity during acceleration, deceleration, and constant motion. *J Sports Sci.*, 30(2), 121-127.

- Venturelli, M., Schena, F., Zanolla, L., y Bishop, D. (2011). Injury risk factors in young soccer players detected by a multivariate survival model. *Journal of Science and Medicine in Sport*, 1.
- Wedderkopp, N., Kaltof, M., Lundgaard, B., Rosendahl, M., y Froberg, K. (s.f). , 41-47.
- Ximénez, C., y Martín, R. S. (2000). *Cuadernos de estadística 8. análisis de varianza con medidas repetidas* (1.^a ed.). La Muralla, S.A.
- Álvarez, P. V. (2015/16). *Modelo de regresión de cox y sus aplicaciones biosanitarias* (Tesis Doctoral no publicada). Universidad de Sevilla.