Análisis Estadístico - Actividad 4

Semestre 2025.1

Índice

Introducción

El conjunto de datos Findata.csv incluye información de calorías quemadas durante entrenamientos basado en el conjunto de datos Final_data.csv disponible en la plataforma Kaggle:

https://www.kaggle.com/datasets/jockeroika/life-style-data/data.

Este conjunto de datos se basa en 20.000 registros reales de fitness con 54 características. La calidad y la naturaleza de la información recopilada tanto su significado son responsabilidad de Kaggle.

Las principales variables que utilizaremos en esta actividad son:

- Age: Edad del participante (en años).
- Gender: Género biológico (Hombre/Mujer).
- Weight (kg): Peso de la persona en kilogramos.
- Height (m): Altura de la persona en metros.
- Max_BPM: Frecuencia cardíaca máxima registrada durante la sesión de entrenamiento.
- Avg_BPM: Frecuencia cardíaca media mantenida durante la sesión.
- Resting_BPM: Frecuencia cardíaca en reposo antes de empezar el entrenamiento.
- Session_Duration (hours): Duración de la sesión de entrenamiento en horas.
- Calories_Burned: Total de calorías quemadas durante la sesión.
- Workout_Type: Tipo de entrenamiento realizado (p. ej., Fuerza, HIIT, Cardio).
- Workout_Frequency: Número de días de entrenamiento por semana.
- Experience_Level: Nivel de experiencia en fitness (1 = Principiante, 2 = Intermedio, 3 = Avanzado).
- BMI Índice de masa corporal.

El resto de la información se puede consultar en https://www.kaggle.com/datasets/jockeroika/life-style-data/data

En esta actividad se analizará si las calorías quemadas están influenciadas por el tipo de entrenamiento y el BMI. Para ello, se aplican diferentes tipos de análisis, revisando el contraste de hipótesis de dos muestras y después realizando análisis ANOVA unifactorial y multifactorial.

Nota aclaratoria sobre la variable Gender En el dataset original, la variable Gender aparece definida como Biological Gender, una denominación ambigua que no especifica si se refiere al sexo biológico o al género. Además, la clasificación binaria utilizada constituye una simplificación que no refleja la realidad intersexual ni la diversidad cromosómica existente

Notas importantes a tener en cuenta para la entrega de la actividad:

- Es necesario entregar el archivo Rmd y el archivo de salida (PDF o html). El archivo de salida debe incluir el código y el resultado de su ejecución (paso a paso). Se debe incluir un índice o tabla de contenidos y se debe respetar la numeración de los apartados del enunciado.
- En la evaluación se valorará especialmente la interpretación de resultados.
- No realicéis listados de los conjuntos de datos, ya que estos pueden ocupar varias páginas. Si queréis comprobar el efecto de una instrucción sobre un conjunto de datos podéis usar la función **head** y **tail** que muestran las primeras o últimas filas del conjunto de datos.

1. Preprocesamiento

1.1. Nombre de las variables

Carga el archivo de datos "Findata.csv". Comprueba los tipos de datos de las variables. Elimina los puntos al final de los nombres de las variables. Sustituye los puntos dentro del nombre de la variable por un guion bajo. Si hay dos puntos seguidos, reemplazadlos por un único guion bajo. Comprobad que no haya dos guiones bajos seguidos.

1.2. Variables Weight_kg y Height_m

Consulta los tipos de datos de las variables `Weight_kg` y `Height_m`. Si es necesario, aplica las transformaciones apropiadas. Averiguad posibles inconsistencias en los valores de las variables. En caso de que existan inconsistencias, corrigelas. Define las variables como numéricas.

Nota: Si los valores de `Weight_kg` están en libras se deben pasar a kg. Si los valores de `Height_m` están en pies y pulgadas se deben pasar a metros. Considera los siguientes factores de conversión: 1 kg = 2.20 lb y 1 m = 39.37 pulgadas, 12 pulgadas = 1 pie.

1.3. Valores ausentes y atípicos

Determina el número de valores ausentes y el número de valores atípicos de las variables `Weight_kg` y `Height_m`.

---

2. Estadística inferencial

2.1. Contraste de hipótesis

¿Podemos aceptar que **los hombres queman en promedio más calorías en los entrenamientos que las mujeres**? Responder a la pregunta utilizando un nivel de confianza del 97,5 %.

Nota:

Se deben realizar los cálculos de comparación de la media manualmente. Se pueden usar funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`. Usar funciones de `R` que calculen directamente el contraste como `t.test` o similar para comparar los resultados.

Sigue los pasos que se detallan a continuación.

2.1.1. Escribid la hipótesis nula y la alternativa

2.1.2. Justificación del test a aplicar

2.1.3. Cálculos

Realiza los cálculos del estadístico de contraste, valor crítico y p-valor a un nivel de confianza del 97,5 %.

2.1.4. Interpretación del test

---

3. Modelo de regresión

3.1. Regresión lineal múltiple

Queremos investigar qué variables explican el valor de las calorías quemadas (`Calories_Burned`). Estima un modelo de regresión lineal múltiple que tenga como variables explicativas: `Session_Duration`, `Experience_Level`, `Workout_Frequency`, `Height_m`, `Weight_m`, `Gender` y `Workout_Type`.

Interpreta el modelo lineal ajustado. ¿Cómo es la calidad del ajuste? Interpretad brevemente la contribución de la variable `Experience_Level` sobre la variable dependiente.

3.2. Multicolinealidad

Analiza posibles problemas de multicolinealidad (alta correlación entre variables explicativas) mediante la interpretación de la matriz de correlaciones de las variables explicativas cuantitativas y del factor de inflación de la varianza (VIF). Se puede utilizar la función **vif** de la librería **car**.

3.3. Análisis visual de los residuos

Analiza gráficamente los residuos del modelo para comprobar la linealidad, la homocedasticidad, la normalidad y la presencia de outliers. ¿A qué conclusión llegáis?

3.4. Regresión logística

Ajusta un modelo predictivo basado en la regresión logística para predecir la probabilidad de ser una persona sin experiencia en fitness.

A partir de la variable `Experience_Level`, clasifica las personas sin experiencia (valor 1) en comparación con el resto (valor 2 o 3). Como regresores considera las mismas variables que en el apartado anterior. Nota: `Calories_Burned` ahora es una variable explicativa del modelo de regresión.

Muestra el resultado del modelo e interpretalo en términos de: cuáles son las variables significativas y cómo es la calidad del modelo.

3.5. Matriz de confusión

A continuación analizad la precisión del modelo, comparando la predicción del modelo sobre los mismos datos del conjunto de datos.

Asumiremos que la predicción del modelo es 1 (Sin experiencia) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario.

Calcula la matriz de confusión. Interpreta los resultados. Indicad los valores de sensibilidad y especificidad e interpretadlos. Se puede utilizar la función confusionMatrix de la librería **Caret**.

4. Análisis de la varianza (ANOVA) de un factor

En este apartado analizaremos si existen diferencias significativas en la variable `Calories_Burned` en función del tipo de entrenamiento (`Workout_Type`).

4.1. Visualización

Visualizar los datos por tipo de entrenamiento con un boxplot

4.2. Hipótesis nula y alternativa

Escribe la hipótesis nula y la alternativa.

4.3. Modelo

Calcula el análisis de varianza, usando la función `aov` o `lm`. Interpreta el resultado del análisis, teniendo en cuenta los valores: Sum Sq, Mean SQ, F y Pr (> F).

4.4. Efectos de los niveles del factor

Proporciona la estimación del efecto de los niveles del factor `Workout_Type`.

4.5. Interpretación de los resultados

Interpreta los resultados obtenidos en los anteriores apartados.

4.6. Adecuación del modelo

Muestra visualmente la adecuación del modelo ANOVA. Verificar los supuestos de normalidad y homoscedasticidad

Puedes usar **plot** sobre el modelo ANOVA calculado.

En caso de encontrar diferencias en las medias procede hacer análisis post hoc donde se hacen las comparaciones múltiples de todos los pares de medias posibles. Si tuviéramos homogeneidad de varianzas, prueba de Scheffé, si las varianzas no son iguales optar por la prueba T2 de Tamhane.

4.6.1. Normalidad de los residuos

Interpreta la normalidad de los residuos a partir del gráfico Normal Q-Q que habéis mostrado en el apartado anterior y verificadla con una prueba de contraste (Anderson-Darling).

Aun así, recordamos que el ANOVA es un test muy robusto a la violación de supuestos de normalidad si tenemos $n \geq 30$ debido al Teorema del Límite Central.

4.6.2. Homocedasticidad de los residuos

El gráfico "Residuals vs Fitted" proporciona información sobre la homocedasticidad de los residuos.

Muestra e interpreta este gráfico y verificadla con una prueba de contraste. Si no se cumple la homocedasticidad, recuerda que se recomienda una ANOVA de Welch. La ANOVA de Welch no asume la igualdad de varianzas y utiliza una corrección en los grados de libertad para calcular el valor F y el valor p. Confirma con la prueba de contraste de igualdad de varianzas de levene.

---

5. ANOVA multifactorial

Ahora analizaremos si las calorías quemadas dependen del tipo de entrenamiento (Workout_Type), de tener un índice de masa corporal saludable y si existe interacción entre estos dos factores.

En primer lugar se creará una variable binaria llamada BMI_bin que indique un índice de masa corporal alto si es mayor a 25 y normal/bajo si es igual o menor a 25.

También analizaremos si existe interacción entre estos dos factores. Si no existiera interacción entre los factores, se debe crear un modelo sin interacción.

Sigue los pasos que se indican a continuación.

5.1. Crea la variable BIM para mayores de 25 igual a "Alto"

5.2. Análisis visual de los efectos principales y posibles interacciones

Dibuja un gráfico que permita evaluar si hay interacción entre ambos factores. Haga un gráfico que muestre las calorías medias quemadas según tipo de entrenamiento diferenciando entre aquellas personas con un índice de masa corporal alto y las que tienen un índice de masa corporal normal o bajo.

5.3. Hipótesis nula y alternativa

Escribe las hipótesis nulas correspondiente a cada variable

5.4. Cálculo del modelo

Calcula el modelo, verifica si existe iteracción entre las variables

5.5. Adecuación del modelo

Comprueba supuestos de manera gráfica y a través de contrastes usados anteriormente.

5.6. Interpretación de los resultados

---

6. Conclusiones

Resume las conclusiones principales del análisis. Para ello, tambien debes de resumir las conclusiones de cada uno de los apartados.

Puntuación de la actividad

Se valorará especialmente la interpretación de los resultados

- Apartados 1 y 2 (20 %)
- Apartado 3 (20 %)
- Apartado 4 (20 %)
- Apartado 5 (20 %)
- Apartado 6 (10 %)
- Calidad del informe dinámico (10 %)