A4 - Análisis de varianza y repaso del curso

Enunciado

Semestre 2020.1

Índex

Introducción

El conjunto de datos Fifa.csv se encuentra disponible en la plataforma Kaggle: https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global.

Este conjunto de datos contiene el estilo de juego del videojuego de consola Fifa 2017, así como estadísticas reales de los jugadores de futbol. El conjunto de datos contiene más de 17,500 registros y 53 variables.

Las principales variables que se usarán en esta actividad son:

- Name (Nombre del jugador)
- Nationality (Nacionalidad del jugador)
- National_Position (Posición de juego en equipo nacional).
- National_Kit (Número de equipación en equipo nacional)
- Club (Nombre del club)
- Club_Position (Posición de juego en club)
- Club_Kit (Número de equipación en club)
- Club_Joining (Fecha en la que empezó en el club)
- Contract_Expire (Año finalización del contrato)
- Rating (Valoración global del jugador, entre 0 y 100)
- Height (Altura)
- Weight (Peso)
- Preffered_Foot (Pie preferido)
- Birth_Date (Fecha de nacimiento)
- Age (Edad)
- Preffered_Position (Posición preferida)
- Work_Rate (valoración cualitativa en términos de ataque-defensa)
- Weak_foot (valoración de 1 a 5 de control y potencia de la pierna no preferida)
- Skill_Moves (valoración de 1 a 5 de la habilidad en movimientos del jugador)
- El resto de variables hacen referencia a atributos del jugador.

La descripción de los atributos se puede consultar en https://www.fifplay.com/encyclopedia. La descripción de las abreviaturas de la posición del jugador en el campo se puede consultar en https://www.dtgre.com/2016/10/fifa-17-position-abbreviations-acronyms.html.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, preprocesado, análisis descriptivo e inferencia estadística.

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.

- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

1 Lectura del fichero y preparación de los datos

Leed el fichero `fifa.csv` y guardad los datos en un objeto con identificador denominado fifa. A continuación, verificad el tipo de cada variable. ¿Qué variables son de tipo numérico? ¿Qué variables son de tipo categórico?

1.1 Preparación de los datos

Las variables `Weight` y `Height` están clasificadas como factor. Para poder trabajar con ellas hay que convertirlas en numéricas.

- Convertir el peso de los jugadores en un valor numérico, eliminando el texto "kg" de los datos.
- Convertir la altura de los jugadores en un valor numérico, eliminando el texto "cm" de los datos.

1.2 Clasificación de jugadores

La variable `Rating` indica la calidad del jugador de la siguiente forma: Excelente de 90 a 99, Muy bueno de 80 a 89, Bueno de 70 a 79, Regular de 50 a 69, Malo de 40 a 49, Muy malo de 0 a 39. Cread una variable categórica denominada `clasificacion`, que clasifique al jugador en una de estas categorías.

---

2 Estadística descriptiva y visualización

2.1 Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas). Mostrad el número de observaciones y el número de variables.

Contad cuántos clubs distintos y cuántas nacionalidades distintas hay representados en los datos.

2.2 Valores ausentes

- Eliminad los valores ausentes del conjunto de datos. Denominad al nuevo conjunto de datos fifaNet (Nota: En las variables 'National_Kit' y 'National_Position' se observan muchos casos sin valor. No eliminéis estas observaciones ya que no son verdaderos missings, sino que simplemente indican que el jugador no ha jugado nunca con el equipo nacional).
- Comprobad cuántas observaciones no tienen valores ausentes y sacad conclusiones sobre cómo de serio es el problema de valores ausentes en estos datos.

2.3 Visualización

1. Cread una variable denominada 'portero' que indique si el jugador juega de portero en su club o juega en otra posición (categoría "GK" en 'Club_Position').
2. Mostrad con diversos diagramas de caja la distribución de la variable 'Weight' según la variable 'portero', según 'Preffered_Foot', según 'clasificacion' y según 'Age'.
3. Dibujad un diagrama de barras que muestre el porcentaje de jugadores que finalizan el contrato en cada uno de los años.
4. Interpretad los gráficos brevemente.

2.4 Comprobación de normalidad

¿Podemos asumir que la variable `Weight` tiene una distribución normal? Debéis justificar la respuesta a partir de métodos visuales.

---

3 Estadística inferencial

Suponemos que los jugadores del año 2017 son una muestra representativa de los jugadores de la última década (población). Utilizamos el conjunto de datos fifaNet.

3.1 Intervalo de confianza de la media poblacional de la variable `Weight`

a) Calculad manualmente el intervalo de confianza al 95% de la media poblacional de la variable `Weight` de los jugadores (No se pueden utilizar funciones como t.test o z.test para el cálculo). A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

b) Calculad los intervalos de confianza al 95% de la media poblacional de la variable `Weight`, en función de si los jugadores son de campo o porteros. ¿Qué conclusión se puede extraer de la comparación de los dos intervalos, en relación a si existe solapamiento o no en los intervalos de confianza? Justificad la respuesta.

Nota: es aconsejable definir una función que dada una muestra y un nivel de confianza dado, calcule el intervalo de confianza asociado. Esta función se puede usar tantas veces como sea necesario. Su uso simplificará las instrucciones necesarias.

3.2 Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que la altura de los porteros supera en más de 5 centímetros la altura de los jugadores de campo? Responded a la pregunta utilizando un nivel de confianza del 95%.

Nota: se deben realizar los cálculos manualmente. No se pueden usar funciones de `R` que calculen directamente el contraste como `t.test` o similar. Sí se pueden usar funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`.

Seguid los pasos que se detallan a continuación.

3.2.1 Escribid la hipótesis nula y la alternativa

3.2.2 Justificación del test a aplicar

3.2.3 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%.

3.2.4 Interpretación del test

---

4 Modelo de regresión lineal

Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: `Age`, `portero`, `Weight`, `Preffered_Foot`, `Vision` y `Ball_Control`, y como variable dependiente el `Rating` de los jugadores.

Especificad el nivel base de referencia de las variables cualitativas, usando la función relevel:

– Para la variable `portero`, la categoría "Portero".

– Para la variable `Preffered_Foot`, la categoría "Left".

4.1 Interpretación del modelo

Interpretad el modelo lineal ajustado:

– ¿Cuál es la calidad del ajuste?

– Explicad la contribución de las variables explicativas en el modelo.

4.2 Predicción

Aplicad el modelo de regresión para predecir el rating de un jugador de campo con pie izquierdo preferido, con un peso de 70, edad de 24, control del balón de 80 y visión de 60.

---

5 Regresión logística

5.1 Modelo predictivo

Ajustad un modelo predictivo basado en la regresión logística para predecir la probabilidad de jugar en la selección nacional en función de las variables: `portero`, `Rating`, `Age` y `Work_Rate`.

Para ello, cread una variable `internacional` que indique si el jugador es internacional, es decir, si está en la selección nacional. La variable `internacional` debe codificarse como una variable dicotómica, que toma el valor 0 cuando el jugador no tiene dorsal en la selección (valor ausente en `National_Kit`) y 1 cuando tiene dorsal (valor en `National_Kit`).

La variable `internacional` será la variable dependiente del modelo. Concretamente, se quiere evaluar la probabilidad de ser un jugador internacional en función de las variables: `portero`, `Rating`, `Age` y `Work_Rate`. Analizad la calidad del modelo y las variables que son relevantes.

5.2 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 (internacional) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Analizad la matriz de confusión y las medidas de 'sensitivity' y 'specificity'.

Nota: Tomad como variable de interés ser jugador internacional. Por tanto, internacional igual a 1 será el caso positivo en la matriz de confusión y 0 el caso negativo.

5.3 Interpretación

Interpretad el modelo ajustado. Concretamente, explicad la contribución de las variables explicativas para predecir si el jugador juega en la selección o no.

5.4 Interpretación de la variable Work_Rate

La variable `Work_Rate` es una variable categórica con 9 categorías diferentes. Volved a ajustar el modelo logístico con las variables `portero`, `Rating`, `Age` y `Work_Rate`, pero ahora considerad como categoría de referencia de la variable `Work_Rate` la categoría 'Medium / Medium'. Interpretad las diferencias en los resultados.

5.5 Importancia de ser portero

En el modelo anterior, interpretad los niveles de la variable `portero` a partir del **odds ratio**. ¿En qué porcentaje se ve aumentada la probabilidad de ir a la selección si eres portero? Proporcionad intervalos de confianza del 95% de los odds ratio.

Realiza el mismo análisis para la variable 'Work_Rate'.

5.6 Predicción

¿Con que probabilidad un portero de 25 años, con un rating de 95 puntos y una clasificación de Work_Rate como High/High irá a la selección?

---

6 Análisis de la varianza (ANOVA) de un factor

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable `Rating` en función del grupo de edad al que pertenecen los jugadores. Seguid los pasos que se indican.

6.1 Visualización gráfica

En primer lugar, a partir de la variable `Age` cread una variable categórica denominada `Age_Int`, que clasifique al jugador en una de estas tres categorías: `Junior` (edad menor o igual a 20), `Middle` (edad entre 21 y 27), `Senior` (edad mayor o igual a 28).

Mostrad gráficamente la distribución de `Rating` según los valores de `Age_Int` ordenados: `Junior`, `Middle`, `Senior`.

6.2 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

6.3 Modelo

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

6.4 Efectos de los niveles del factor

Proporcionad la estimación del efecto de los niveles del factor `Age_Int`.

6.5 Interpretación de los resultados

Interpretad los resultados obtenidos en los apartados anteriores.

6.6 Adecuación del modelo

Mostrad visualmente la adecuación del modelo ANOVA. Podéis usar **plot** sobre el modelo ANOVA calculado.

6.6.1 Normalidad de los residuos

Interpretad la normalidad de los residuos a partir del gráfico Normal Q-Q que habéis mostrado en el apartado anterior.

6.6.2 Homocedasticidad de los residuos

El gráfico "Residuals vs Fitted" proporciona información sobre la homcedasticidad de los residuos. Mostrad e interpretad este gráfico.

---

7 ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre `Rating` del grupo de edad combinado con el factor tipo de jugador (`portero`). Seguid los pasos que se indican a continuación.

7.1 Análisis visual de los efectos principales y posibles interacciones

Dibujad en un gráfico la variable `Rating` en función de `Age_Int` y en función de `portero`. El gráfico debe permitir evaluar si hay interacción entre los dos factores. Por ello, se recomienda seguir estos pasos:

1. Agrupad el conjunto de datos por `Age_Int` y por `portero`. Calculad la media de rating para cada grupo. Para realizar este proceso, se puede hacer con las funciones **group_by** y **summarise** de la librería `dplyr`.
2. Mostrad el conjunto de datos en forma de tabla (data frame), donde se muestre la media de cada grupo según `Age_Int` y `portero`.
3. Mostrad en un gráfico el valor medio de la variable `Rating` para cada factor. Podéis inspiraros en los gráficos de López-Roldán y Fachelli (2015), p.38. Podéis realizar este tipo de gráfico usando la función `ggplot` de la librería `ggplot2`.
4. Interpretad el resultado sobre si sólo hay efectos principales o hay interacción entre los factores. Si hay interacción, explicad cómo se observa esta interacción en el gráfico.

7.2 Cálculo del modelo

Podéis usar la función **aov**.

7.3 Interpretación de los resultados

7.4 Adecuación del modelo

Interpretad la adecuación del modelo ANOVA obtenido usando los gráficos de los residuos.

---

8 Conclusiones

Resumid las conclusiones principales del análisis. Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Puntuación de la actividad

- Apartados 1 y 2 (10%)
- Apartado 3 (10%)
- Apartado 4 (10%)
- Apartado 5 (15%)
- Apartado 6 (20%)
- Apartado 7 (15%)
- Apartado 8 (10%)
- Calidad del informe dinámico (10%)