Proyecto de Estadística
2da Fase

Eric Martin Garcia
Grupo C411
Alberto Helguera Fleitas
Grupo C412
Jose Gabriel Navarro Comabella
Grupo C412

Tutor(es):

Introducción

En este trabajo se realizará un estudio de los datos de los jugadores del FIFA19 usando las técnicas de regresión, reducción de dimensión y de ANOVA.

1. Se elegirán las variables a las se cuales les aplicara cada técnica y se explicará el por qué.
2. En las técnicas que lo requieran, se realizará el análisis de los supuestos y se explicará si es válida la aplicación de la técnica en esa variable.

Desarrollo

Regresión Múltiple

En este apartado se trabajará sobre el subconjunto de jugadores de campo del FIFA19 que pertenecen al club Inter de Milán. Con el objetivo de analizar como se evoluciona el promedio de un jugador con respecto a su edad y potencial.
En primer lugar se analiza la relación entre las variables utilizando la matriz de correlación "`cor(dataset)`", resultando:

```
> # Calcular la matriz de correlacion
> cor(dataset)
              overall         age  potential
overall    1.00000000 -0.07330373  0.9208640
age       -0.07330373  1.00000000 -0.4157898
potential  0.92086401 -0.41578978  1.0000000
```

El modelo elegido buscará analizar la relación de la variable (*overall*) con las variables (*age*) y (*potential*), quedando representado de la siguiente forma:

$$overall = \beta_0 + potential\beta_1 + age\beta_2 + e$$

Para investigar los resultados de la regresión y determinar el valor de los β_j se utilizó la función "`summary(regression_model)`" de `R` y se obtuvo la salida mostrada en la Figura 1

```
> # Averiguar si overall tiene alguna relacion con la edad y el potencial
> regression_model <- lm(overall ~ potential + age, data=dataset)
> #Mostrar los resutados de la regresion
> summary(regression_model)

Call:
lm(formula = overall ~ potential + age, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6051 -0.8864  0.1065  0.8380  1.9154 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.76790    4.55955  -3.897 0.000831 ***
potential     1.01151    0.04286  23.601  < 2e-16 ***
age           0.53539    0.06524   8.206 5.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.141 on 21 degrees of freedom
Multiple R-squared:  0.9639,	Adjusted R-squared:  0.9604 
F-statistic: 280.1 on 2 and 21 DF,  p-value: 7.217e-16
```

Sustituyendo los valores obtenidos resulta el modelo:

$$\widehat{overall} = 1.01151potential + 0.53539age - 17.76790$$

Coeficientes: Los coeficientes son significativos al 0% inclusive el intercepto lo que es bueno para el modelo, los valores de $Pr(>|t|)$ son menores que 0.05 por lo tanto todas las variables aportan información al modelo.

Adjusted R-Square: El valor del R-Cuadrado es 0.9604 por lo tanto el modelo se considera muy bueno en cuanto a la realización de predicciones

F-Statistic: Su valor menor que 0.05 nos indica la existencia de al menos una variable que esta siendo significativa para el modelo

Analizando los Residuos:

1. **La media de los errores es 0 y la suma de los errores es 0**:

```
> #1: La media y la suma de los errores es 0
> mean(regression_model$residuals)
[1] -4.857226e-17
> sum(regression_model$residuals)
[1] -1.165734e-15
```

2. **Errores normalmente distribuidos**: Se muestra el histograma y el Normal Q-Q Plot ,en el histograma se puede apreciar un parecido a una distribución normal y al observar el QQ Plot se aprecia como la mayoría de los puntos de residuo se encuentran sobre la recta, por lo tanto se asume una normalidad en los errores del modelo (**Figura 2**)

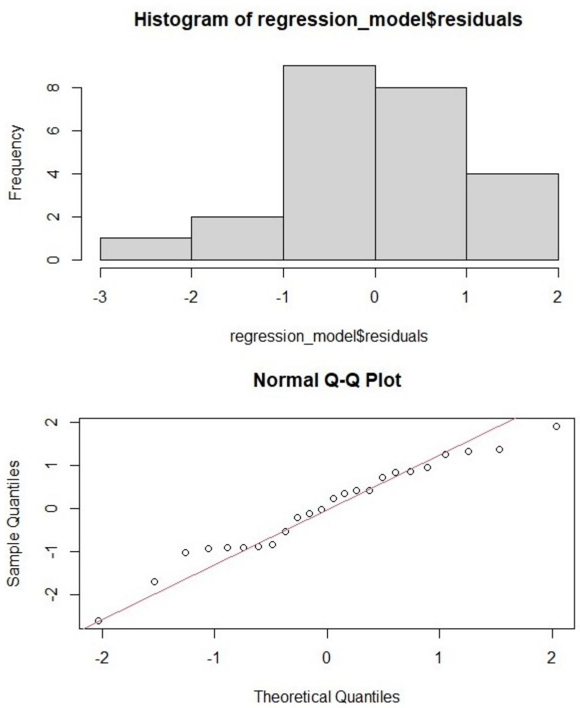


Figure 1:

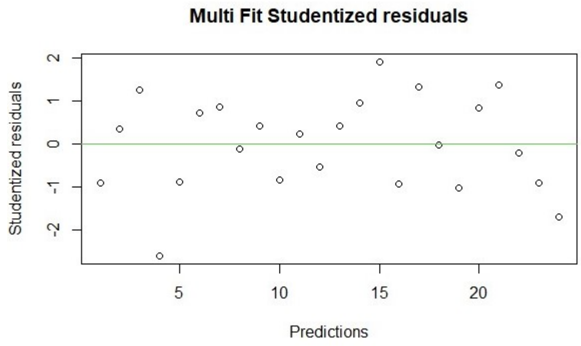
3. **Independencia de los residuos**: Para la prueba de independencia se emplea la prueba Durbin-Watson:

```
> #3: Los errores son independientes
> dwtest(regression_model)

        Durbin-Watson test

data:  regression_model
DW = 1.9316, p-value = 0.3463
alternative hypothesis: true autocorrelation is greater than 0
```

Como el p-value es mayor que 0.05 no se puede rechazar la hipótesis nula, por lo tanto se puede afirmar que los errores son independientes.

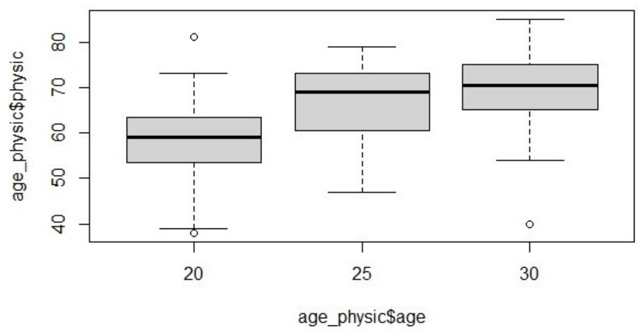
4. **La varianza de los errores es constante (Homocedasticidad)**:

Se puede observar en el gráfico el cumplimiento de la homocedasticidad:



ANOVA

En este apartado se trabajará sobre el subconjunto de jugadores de campo del FIFA19 nacidos en Argentina con edades de 20, 25 y 30 años (factor edad), con el objetivo de reconocer como el físico de los jugadores varía con respecto a su edad.
Luego de reorganizar los datos para un correcto análisis podemos comparar las medidas de los 3 niveles del factor "edad" realizando un gráfico de cajas con las medias de cada nivel:

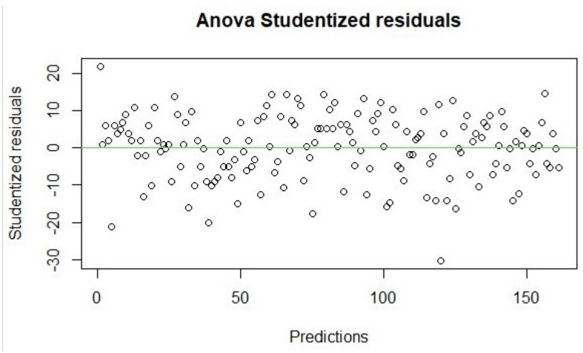


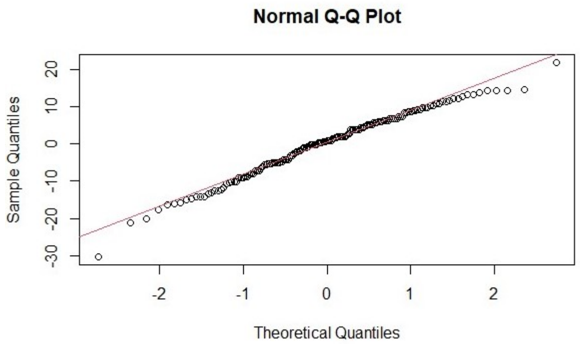
El siguiente paso sería realizar el análisis de varianza, para ver si la prueba de hipótesis es válida. En este caso el análisis de varianza es el siguiente:

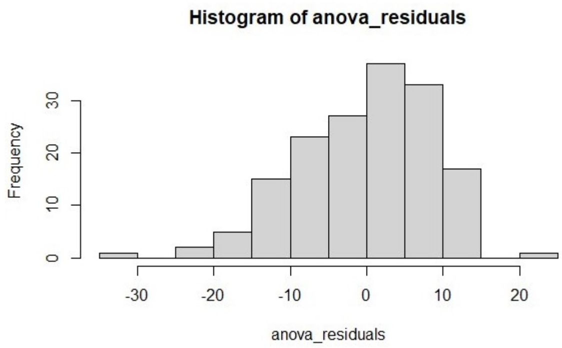
```
> #Analisis de Varianza
> physic_anova <- aov(age_physic$physic ~ age_physic$age, data = df)
> summary(physic_anova)
               Df Sum Sq Mean Sq F value   Pr(>F)    
age_physic$age   1   3272    3272   43.74 5.42e-10 ***
Residuals      159  11896      75                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como p-value = 0.0000 es menor que la significación prefijada $\alpha = 0.05$, se rechaza H_0 y se acepta que al menos un grupo de futbolistas de cierta edad, tiene un físico promedio diferente.
Por último, es necesario verificar que se cumplen los supuestos del modelo:

1. Los e_{ij} siguen una distribución normal con media cero.
2. Los e_{ij} son independientes entre sí.
3. Los residuos de cada tratamiento tienen la misma varianza σ^2







Como se puede observar en el gráfico estandarizado de residuos, tienen varianza constante, el qq-plot y el histograma de residuos muestran un comportamiento de forma normal, sin embargo, se realizarán las pruebas Shapiro-Wilcox, Bartlet y DurbinWatson para constatar el cumplimiento de los supuestos.

1. Test de Shapiro-Wilcox

```
> shapiro.test(anova_residuals)

        Shapiro-Wilk normality test

data:  anova_residuals
W = 0.98383, p-value = 0.05751
```

   El test de Shapiro-Wilcox no es significativo (p-value $= 0.057 > 0.05$), no podemos rechazar H_0 por lo que se puede confirmar la hipótesis de normalidad en los residuos.

2. Test de Bartlett

```
> bartlett.test(anova_residuals, df$age_physic.age)

        Bartlett test of homogeneity of variances

data:  anova_residuals and df$age_physic.age
Bartlett's K-squared = 0.055737, df = 2, p-value = 0.9725
```

   El test de Bartlett no es significativo (p-value $= 0.97 > 0.05$), no podemos rechazar H_0 por lo que se puede confirmar la hipótesis de homogeneidad de las varianzas.

3. Test de Durbin-Watson

```
> dwtest(physic_anova)

        Durbin-Watson test

data:  physic_anova
DW = 2.0854, p-value = 0.6792
alternative hypothesis: true autocorrelation is greater than 0
```

   El test de Durbin-Watson no es significativo (p-value $= 0.67 > 0.05$), no podemos rechazar H_0 por lo que se puede confirmar la hipótesis de independencia de los errores.

ACP

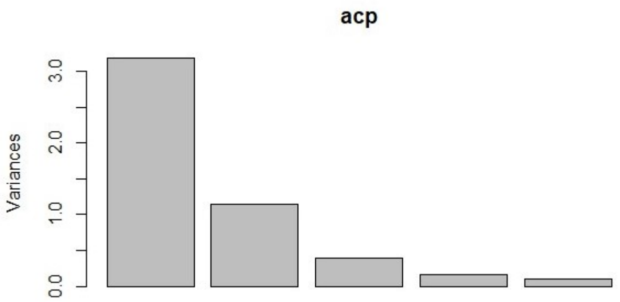
En este apartado se trabajará sobre el subconjunto de jugadores de campo del FIFA19 que pertenecen al club FC Barcelona.
En primera instancia se hace un análisis de la correlación en la muestra utilizando la matriz de correlación "`cor(acp_dataset)`" y la función "`symnum(tp)`" presente en `R`, que retorna de forma gráfica si dicha matriz esta o no, altamente correlacionada, resultando:

```
> #Calcular la matriz de correlacion
> tp = cor(acp_dataset)
> #Chequear de forma grafica si existe correlacion
> symnum(tp)
          pc s ps dr df
pace      1
shooting     1
passing      , 1
dribbling .  , ,  1
defending .  , .  ,  1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

Se puede observar que no existe una relación entre las variables por lo tanto se procede a reducir dimensión. Para esto se seleccionan las componentes principales:

```
> #Calculo de ACP
> acp <- prcomp(acp_dataset, scale = TRUE)
> summary(acp)
Importance of components:
                          PC1    PC2     PC3    PC4     PC5
Standard deviation     1.7838 1.0727 0.63018 0.4037 0.32752
Proportion of Variance 0.6364 0.2301 0.07942 0.0326 0.02145
Cumulative Proportion  0.6364 0.8665 0.94594 0.9786 1.00000
```

Para la selección de las componentes principales se emplea la proporción acumulativa, la primera componente PC1 es 0.61364 , que solo explicaría un 61% por lo que se necesita al menos otra componente para tener un valor mayor al 70%. Si se incluye PC2 se alcanza un 86%. De acuerdo con el criterio de Kaiser se tiene que las dos componentes con valores propios superiores a 1 son la primera y la segunda. Por lo que PC1 y PC2 son las componentes principales a elegir. Otra forma de corroborar dicha elección sería graficar todas las componentes principales como se muestra en la siguiente gráfica:



Para interpretar los datos se debe calcular la matriz de valores propios y así se sabrá que variable es importante para cada componente y en qué medida:

```
> #Matriz de valores propios
> acp$rotation[,1:5]
                 PC1         PC2        PC3         PC4        PC5
pace      -0.2954208 -0.73010330  0.4552518 -0.38168633  0.1635147
shooting  -0.4908186  0.31706862 -0.2671513 -0.68896123 -0.3354508
passing   -0.4402481  0.49896167  0.3553318  0.08878313  0.6504427
dribbling -0.5258033 -0.01108467  0.3021175  0.53755759 -0.5858027
defending  0.4489278  0.34253621  0.7098171 -0.28774533 -0.3074006
```

Para esto se analiza por cada componente el mayor valor propio λ_i, este se divide entre 2 y la variable asociada a cada valor propio cuyo valor absoluto este por encima de $\lambda_i/2$, pertenecerá a la componente.

PC1 ($\lambda_{\max} = 0.52$): Esta componente esta caraterizada por una muestra de jugadores con estadísticas ofensivas bajas y buenas capacidades defensivas. (Defensas)

PC2 ($\lambda_{\max} = 0.73$): Esta componente esta caracterizada por una muestra de jugadores con poco ritmo pero buenos pasadores y disparando al arco. (Delantero)

PC3 ($\lambda_{\max} = 0.70$): Esta componente esta caracteriza por una muestra de jugadores con buen ritmo, buenas capacidades defensivas y pasando el balón. (Medio Centro Defensivo)

PC4 ($\lambda_{\max} = 0.68$): Esta componente esta caracteriza por una muestra de jugadores con buenas habilidades en el dominio del balón, con poco ritmo y no muy finos en los disparos. (Medio Centro)

PC5 ($\lambda_{\max} = 0.65$): Esta componente esta caracteriza por una muestra de jugadores con poca habilidad en el dominio del balón y en el disparo pero buenos pasadores. (Medio Ofensivo)

La figura siguiente muestra un biplot para las 2 primeras componentes (PC1, PC2) con la representación de los jugadores según sus habilidades: