Proyecto de Estadística
2da Fase

Eric Martin Garcia
Grupo C411
Carlos Rafael Ortega Lezcano
Grupo C411
Harold Rosales Hernandez
Grupo C411

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. 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 poten
> 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**)

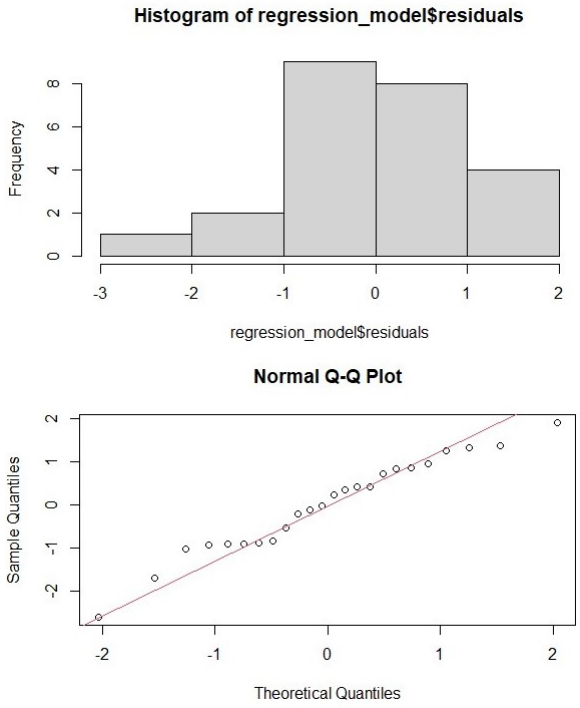


Figure 1:

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

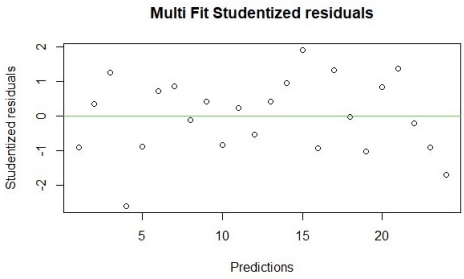
```
> #3: Los errores son independientes
> dwtest(regression_model)

        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:



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 procede a graficar los valores propios asociados a cada una ellas:

Para la selección de las componentes principales empleamos la proporción acumulativa y nos quedamos con aquellas que su porciento acumulatico es menor que 0.70 y la primera que supera este valor, según lo visto en el criterio del porcentaje, además si observamos detenidamente los valores propios las tres primeras componentes tienen valor propio mayor a 1 por tanto si hubieramos seleccionado el criterio de Kaiser obtendriamos las 3 primeras componentes como principales.
Los valores propios de las 3 primeras componentes son los siguientes:

Ahora pasemos a analizar cuales variables son importantes en cada componente y en que medida, para esto tomamos por cada componente el mayor valor propio λ_i, dividimos entre 2 y todo valor propio cuyo valor

absoluto este por encima de $\lambda_i/2$, la variable asociada a este conformará la componente.

PC1 ($\lambda_{\max} = 0.61$): Esta componente esta caraterizada por una muestra de jugadores de bajo *overall*, con poco desarrollo en el juego (*potential*) y de baja reputación internacional, o sea no serán convocados a la selección con mucha frecuencia

PC2 ($\lambda_{\max} = 0.65$): Esta componente esta caracterizada por una muestra de jugadores jóvenes con buenas habilidades en el dominio del balón, además el valor negativo en la variable altura puede asociarse a jugadores del mediocampo, en su mayoria en lugar de defensores y delanteros centros

PC3 ($\lambda_{\max} = 0.72$): Esta componente esta caracteriza por una muestra de jugadores jóvenes con gran potencial (a medida que avancen las temporadas su promedio y valor aumentará en gran medida). Esta componente resulta interesante porque describe a aquellos jugadores que podríamos elegir en el juego para el pasar las temporadas sea de gran valor para nuestro equipo

La figura siguiente muesta los biplot para las 2 primeras componentes (PC1, PC2) y para las dos últimas (PC2, PC3):

Conclussion

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

./imgs/main.png

```
> biplot(acp, choices = 1:2)
```

./imgs/biplotp1p2.jpeg

```
> biplot(acp, choices = 2:3)
```

./imgs/biplotp2p3.jpeg

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

References