Proyecto de Estadística
2da Fase

Oscar Luis Hernandez Solano
Grupo C411
Carlos Rafael Ortega Lezcano
Grupo C411
Harold Rosales Hernandez
Grupo C411

Tutor(es):

Intro

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Regresión Múltiple

En primer lugar se analiza la relación entre las variables mediante `cor(dt)`, resultando:

```
> cor(dt)
                              age    overall    potential skill_moves    height_cm international_reputation
age                    1.00000000 0.46184226 -0.21890445 -0.03064357  0.07722374                0.24068105
overall                0.46184226 1.00000000  0.68087194  0.22459966  0.04556033                0.53206418
potential             -0.21890445 0.68087194  1.00000000  0.21711797  0.00719362                0.42839324
skill_moves           -0.03064357 0.22459966  0.21711797  1.00000000 -0.44392635                0.15435409
height_cm              0.07722374 0.04556033  0.00719362 -0.44392635  1.00000000                0.04404545
international_reputation 0.24068105 0.53206418 0.42839324  0.15435409  0.04404545                1.00000000
```

Plantearemos un modelo que busca estimar el valor de la variable (*overall*) mediante el resto de las variables, al observar la matriz de correlación descartamos las variables (*height_cm* y *skill_moves*), por tanto el modelo se escribe de la siguiente forma:

$$overall = \beta_0 + potential\beta_1 + age\beta_2 + irep\beta_3 + e$$

Usando `R` para determinar el valor de los β_j se obtiene la salida mostrada en la Figura 1

Sustituyendo los valores obtenidos resulta el modelo:

$$\widehat{overall} = 0.92potential + 0.95age + 0.62irep - 23.91$$

```
Call:
lm(formula = overall ~ age + potential + international_reputation,
    data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7323  -1.4928   0.5371   1.9026  10.2298

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -23.919360   0.310488  -77.04   <2e-16 ***
age                      0.950447   0.004783  198.71   <2e-16 ***
potential                0.920688   0.003898  236.19   <2e-16 ***
international_reputation 0.628999   0.058258   10.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.644 on 17588 degrees of freedom
Multiple R-squared:  0.8565,    Adjusted R-squared:  0.8565
F-statistic: 3.499e+04 on 3 and 17588 DF,  p-value: < 2.2e-16
```

Coeficientes: Los coeficientes son significativos al 0% inclusive el intercepto lo que es bueno para el modelo, los valores de $Pr(> |t|)$ son menores por lo tanto no existe variable que no aporte información al modelo

Adjusted R-Square: El valor del R-Cuadrado es 0.8565 por lo tanto el modelo se considera bastante bueno en cuanto a la realización de predicciones

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

Ahora pasemos a analizar los residuos para el modelo:

Analizando los Residuos:

1. **La media de los errores es 0 y la suma de los errores es 0**:
   Empleando `R` se obtiene:

   ```
   > mean(model$residuals)
   [1] -3.006259e-15
   > sum(model$residuals)
   [1] -5.330704e-11
   ```

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, pero al observar el QQ Plot se aprecia como la mayoria de los puntos de

residuo se encuentran sobre la recta, por lo tanto se asume una normalidad en los errores del modelo (**Figura 2**)

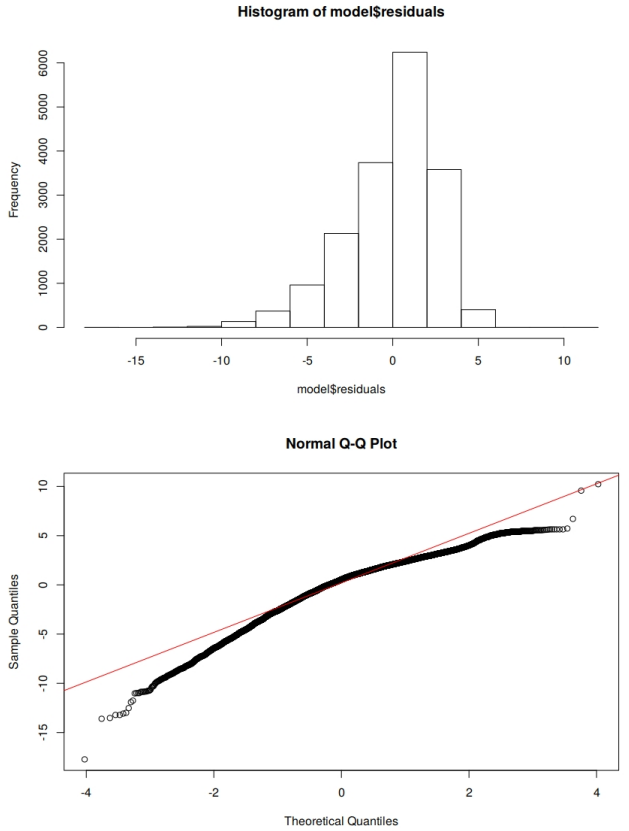


Figure 1:

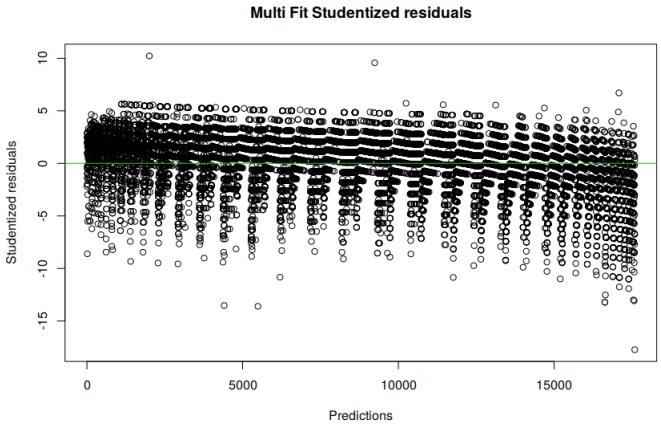
3. **Independencia de los residuos**:
   Para la prueba de independencia se emplea la prueba Durbin-Watson:

```
> dwtest(model)

Durbin-Watson test

data:  model
DW = 0.64356, p-value < 2.2e-16
alternative hypothesis: true
autocorrelation is greater than 0
```

   Como el p-value no es mayor que 0.05 se rechaza la hipotesis nula por tanto no podemos afirmar que los errores sean independientes, el incumplimiento de este residuo hace que nuestro modelo no sea el idóneo para estimar valores de la variable *overall*

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

   Se puede observar en el gráfico que se cumple la homocedasticidad:



ACP

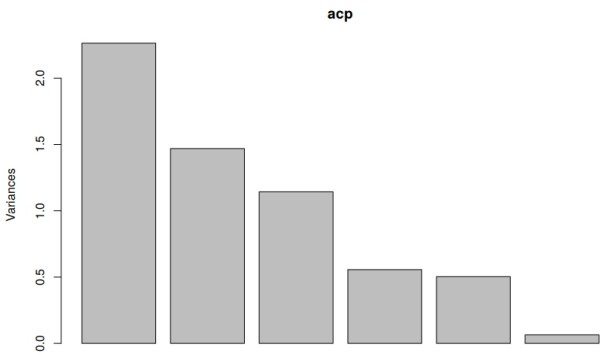
Primeramente podemos observar que la matriz de correlación a simple vista no nos muestra la relación que se establece entre las variables por tanto empleamos la función `synnum` a la matriz obteniendo:

```
> symnum(tp)
                          a o p s h i
age                       1
overall                   . 1
potential                   , 1
skill_moves                     1
height_cm                       . 1
international_reputation    . .     1
```

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

```
> acp <- prcomp(dt, scale = TRUE)
> summary(acp)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.5046 1.2122 1.0692 0.74568 0.70940 0.25352
Proportion of Variance 0.3773 0.2449 0.1905 0.09267 0.08387 0.01071
Cumulative Proportion  0.3773 0.6222 0.8127 0.90541 0.98929 1.00000
```

Para la selección de las componentes principales pasemos a graficar los valores propios asociados a cada una:



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:

```
> acp$rotation[,1:3]
                              PC1        PC2         PC3
age                   -0.22233237 -0.4148444 -0.727129418
overall               -0.61028910 -0.1460054 -0.055460503
potential             -0.50543502  0.1288750  0.529422985
skill_moves           -0.26408031  0.5910827 -0.239311927
height_cm              0.03681673 -0.6514440  0.361345070
international_reputation -0.50155468 -0.1273569  0.008819649
```

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.

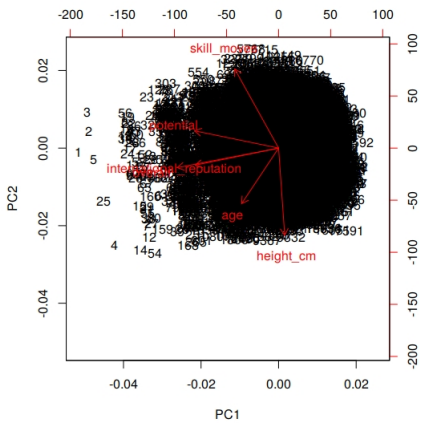
PC1 ($\lambda_{\max} = 0.61$): Esta componente esta caraterizada por una muestra de jugadores de bajo *overall*, con poco desarrollo en el juego (*potential*) y de baja reputación internacional, o sea no serán convocados a la selección con mucha frecuencia

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

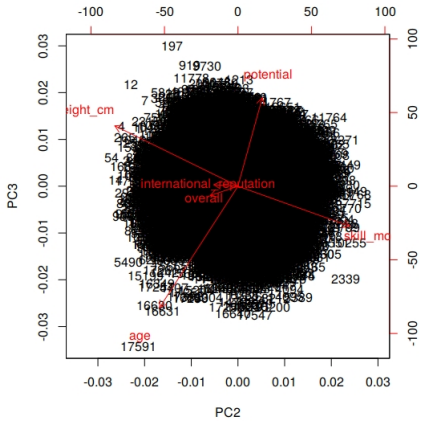
PC3 ($\lambda_{\max} = 0.72$): Esta componente esta caracteriza por una muestra de jugadores jóvenes con gran potencial (a medida que avancen las temporadas su promedio y valor aumentará en gran medida). Esta componente resulta interesante porque describe a aquellos jugadores que podríamos elegir en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

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




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



Development

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

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.

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