UOC - M2.951 Tipologia i cicle de vida de les dades

Pràctica 2: Neteja i anàlisi de les dades

Nicolás González Soler

22 de desembre de 2021

Índex

1 Dataset seleccionat

1.1 Font del dataset

El dataset objecte d'estudi s'ha obtingut de la plataforma Kaggle i es denomina Complete FIFA 2017 Player dataset (Global). Conté informació de més de 17500 jugadors d'aquest clàssic videojoc, per a cadascun dels quals es faciliten 53 característiques. La descripció dels atributs es pot consultar a la web https://www.fifplay.com/encyclopedia.

Aquest volum i diversitat d'informació el fan especialment atractiu a efectes d'estudis estadístics. Aquest fet, juntament amb l'alt interès de l'autor pels esports, ha propiciat la seva selecció.

1.2 Objectius d'investigació

Dels múltiples centres d'interès possibles, en aquest exercici ens centrarem en l'anàlisi comparativa *porter vs jugador de camp* des de diverses vessants: l'índex de massa corporal, la valoració com a jugador, o el fet de jugar o no amb la *selecció nacional*. En particular, intentarem donar resposta a les preguntes següents:

1. La característica que està més relacionada amb la valoració, és la mateixa per als porters i que per als jugadors de camp?

2. Podem afirmar que la mitjana de l'IMC de porters i jugadors de camp és igual?

3. Quina combinació de característiques esportives explica millor la valoració d'un jugador? Com afecta a la valoració ser o no porter?

4. Quina és la probabilitat de ser internacional en funció de la valoració, l'edat i la posició? En quin percentatge augmenta o disminueix aquesta probabilitat el fet de ser o no porter?

1.3 Determinació de les dades d'interès

Ateses les preguntes d'investigació formulades, i després d'estudiar conceptualment les característiques dels jugadors presents al dataset, s'ha optat per seleccionar-ne un subconjunt. A la taula següent s'enumeren i descriuen les variables escollides.

Observació: s'especifica també el tipus de la variable atenent a la seva naturalesa, tot i que estigui expressada i emmagatzemada en un format diferent (en tal cas eventualment caldria realitzar una transformació).

Variable	Descripció	Tipus
Name	Nom del jugador	categòric
National_Position	Posició del jugador a la selecció nacional	categòric
Club_Position	Posició del jugador al seu equip	categòric
Rating	Valoració global del jugador, entre 0 i 100	numèric
Height	Alçada del jugador, en centímetres	numèric
Weight	Pes del jugador, en quilograms	numèric
Age	Edat del jugador, en anys	numèric
Work_Rate	Valoració qualitativa del jugador en termes d'atac-defensa	categòric
Ball_Control	Habilitat del jugador per controlar la pilota, entre 0 i 100	numèric
Vision	Habilitat del jugador en termes de visió de joc, entre 0 i 100	numèric

2 Càrrega i selecció

El fitxer *csv* proporcionat per Kaggle s'anomena *FullData.csv*. Atès que l'enunciat de la pràctica ens insta a gravar el fitxer resultant del procés de neteja, per major claredat s'ha canviat el nom del fitxer d'entrada a *Fifa2017_original.csv*.

```
# Càrrega del fitxer csv
df <- read.csv("Fifa2017_original.csv",    # per defecte header=TRUE i sep=","
               strip.white = TRUE,         # espais eliminats a principi i final d'string
               stringsAsFactors = TRUE,    # strings convertits a factor
               na.strings = c(""))         # strings buits convertits a valor NA
```

Després de la càrrega seleccionem només les característiques d'interès.

```
# Selecció de característiques d'interès
df <- df %>% select(Name, National_Position, Club_Position, Rating,
                    Height, Weight, Age, Work_Rate, Ball_Control, Vision)
```

Mostrem l'estructura inicial del data frame creat, conformada per 17588 observacions i 10 variables.

```
# Estructura del data frame
str(df)
```

```
## 'data.frame':    17588 obs. of  10 variables:
##  $ Name             : Factor w/ 17341 levels "A.J. DeLaGarza",..: 3270 9925 12459 10269 10555 3900
##  $ National_Position: Factor w/ 27 levels "CAM","CB","CDM",..: 13 24 14 13 5 5 13 23 NA 5 ...
##  $ Club_Position    : Factor w/ 29 levels "CAM","CB","CDM",..: 15 26 15 28 6 6 28 26 28 6 ...
##  $ Rating           : int  94 93 92 92 92 90 90 90 90 89 ...
##  $ Height           : Factor w/ 50 levels "155 cm","157 cm",..: 30 15 19 27 38 38 30 28 40 44 ...
##  $ Weight           : Factor w/ 56 levels "100 kg","101 kg",..: 37 29 25 42 49 39 36 31 52 48 ...
##  $ Age              : int  32 29 25 30 31 26 28 27 35 24 ...
##  $ Work_Rate        : Factor w/ 9 levels "High / High",..: 2 9 3 3 9 9 3 3 8 9 ...
##  $ Ball_Control     : int  93 95 95 91 48 31 87 88 90 23 ...
##  $ Vision           : int  85 90 80 84 70 68 78 79 83 44 ...
```

3 Conversió

3.1 Tractament de les variables *Height* i *Weight*

Aquestes dues variables són de naturalesa numèrica. Per tant, caldrà eliminar el text associat a les unitats i fer un canvi de tipus.

```
# Tractament de Height
df$Height <- gsub(" ", "", df$Height)    # Treure blancs
df$Height <- gsub("cm$", "", df$Height)  # Treure text "cm"
df$Height <- as.numeric(df$Height)       # Convertir a numèric

# Tractament de Weight
df$Weight <- gsub(" ", "", df$Weight)    # Treure blancs
df$Weight <- gsub("kg$", "", df$Weight)  # Treure text "kg"
df$Weight <- as.numeric(df$Weight)       # Convertir a numèric
```

3.2 Tractament de la variable *Work_Rate*

Aquesta variable categòrica ja és de tipus factor, però ens assegurarem que el valor de referència sigui *High / High*.

```
# Tractament de Work_Rate
df$Work_Rate <- relevel(df$Work_Rate, ref="High / High")  # Establir valor referència
```

4 Creació de variables

Per tal de donar resposta a les preguntes d'investigació serà convenient crear dues variables. D'una banda classificarem els jugadors en *porters* o *no porters*. D'una altra identificarem si un jugador és *internacional* o no. Això es farà examinant el valor de la variable *National_Position*: un valor *NA* no és indicatiu de valor perdut, sinó que identifica el fet que el jugador no ha jugat amb la selecció nacional.

4.1 Creació de la variable *Goalkeeper*

Un porter és aquell que juga en la posició *GK* (goal keeper). Crearem una variable categòrica, de tipus factor, que identificarà si un jugador és porter o no. El valor de referència serà *YES*.

```
# Tractament de Goalkeeper
df$Goalkeeper <- as.factor(ifelse(df$Club_Position == "GK", "YES", "NO"))  # Creació
df$Goalkeeper <- relevel(df$Goalkeeper, ref="YES")                         # Valor referència
```

4.2 Creació de la variable *International*

Crearem una variable categòrica, de tipus factor, que identificarà si un jugador és internacional o no.

```
# Tractament de International
df$International <- as.factor(ifelse(is.na(df$National_Position), "NO", "YES"))  # Creació
```

4.3 Creació de la variable *IMC*

Crearem una variable numèrica que identificarà l'*índex de massa corporal* (IMC) del jugador.

```
# Tractament de IMC
df$IMC <- df$Weight/((df$Height/100)^2)
```

5 Exploració de les dades

5.1 Registres duplicats

En primer lloc comprovem si existeixen registres duplicats. Tal com veiem a continuació, aquest no és el cas.

```
# Comprovació de registres duplicats
sum(duplicated(df))
```

```
## [1] 0
```

5.2 Resum estadístic general

Obtenim un resum estadístic del data frame.

```
# Resum estadístic
summary(df)
```

```
##                Name       National_Position Club_Position      Rating
##  Felipe          :    6   Sub    :  556     Sub    :7492   Min.   :45.00
##  Danilo          :    5   LCB    :   48     Res    :3146   1st Qu.:62.00
##  Gabriel         :    5   GK     :   47     RCB    : 633   Median :66.00
##  Carlos Rodríguez:    4   RCB    :   46     GK     : 632   Mean   :66.17
##  Roberto         :    4   LB     :   39     LCB    : 631   3rd Qu.:71.00
##  Álvaro          :    3   (Other):  339     (Other):5053   Max.   :94.00
##  (Other)         :17561   NA's   :16513     NA's   :   1
##      Height          Weight            Age                  Work_Rate
##  Min.   :155.0   Min.   : 48.00   Min.   :17.00   Medium / Medium:9897
##  1st Qu.:176.0   1st Qu.: 70.00   1st Qu.:22.00   High / Medium  :2918
##  Median :181.0   Median : 75.00   Median :25.00   Medium / High  :1534
##  Mean   :181.1   Mean   : 75.25   Mean   :25.46   Medium / Low   : 845
##  3rd Qu.:186.0   3rd Qu.: 80.00   3rd Qu.:29.00   High / High    : 747
##  Max.   :207.0   Max.   :110.00   Max.   :47.00   High / Low     : 730
##                                                   (Other)        : 917
##   Ball_Control       Vision      Goalkeeper   International      IMC
##  Min.   : 5.00   Min.   :10.00   YES :  632   NO :16513     Min.   :17.09
##  1st Qu.:53.00   1st Qu.:43.00   NO  :16955   YES: 1075     1st Qu.:22.06
##  Median :63.00   Median :54.00   NA's:    1                 Median :22.92
##  Mean   :57.97   Mean   :52.71                              Mean   :22.92
##  3rd Qu.:69.00   3rd Qu.:64.00                              3rd Qu.:23.77
##  Max.   :95.00   Max.   :94.00                              Max.   :34.72
##
```

5.3 Tractament dels valors perduts

A partir de l'anàlisi del resum estadístic observem que:

- Hi ha 16513 valors *NA* a la variable *National_Position*. Tal com ja s'ha mencionat, un *NA* indica que el jugador no ha jugat amb la selecció nacional i, per tant, aquests no són veritables valors perduts. Conseqüentment, no procedeix fer cap acció.

- Hi ha 1 valor *NA* a la variable *Club_Position*. Aquest cas s'ha d'investigar.

- Hi ha 1 valor *NA* a la variable *Goalkeeper*. Podria ser que fos conseqüència del *NA* a *Club_Position*. Caldrà examinar-ho.

- Hi ha 917 valors *Other* a la variable *Work_Rate*. Aparentment no són valors *NA*, però ens assegurarem que són valors vàlids.

Començarem examinant la variable *Work_Rate*.

```
# Examen de Work_Rate
kable(t(table(df$Work_Rate)), booktabs = TRUE,) %>%
  kable_styling(latex_options = c("HOLD_position","scale_down"), position = "center")
```

High / High	High / Low	High / Medium	Low / High	Low / Low	Low / Medium	Medium / High	Medium / Low	Medium / Medium
747	730	2918	438	30	449	1534	845	9897

Constatem que són valors vàlids, així que prosseguim l'anàlisi examinant les variables *Club_Position* i *Goalkeeper*.

```
# Examen de Club_Position i Goalkeeper
kable(filter(df, is.na(Club_Position) | is.na(Goalkeeper)), booktabs = TRUE) %>%
  kable_styling(latex_options = c("HOLD_position","scale_down"), position = "center")
```

Name	National_Position	Club_Position	Rating	Height	Weight	Age	Work_Rate	Ball_Control	Vision	Goalkeeper	International	IMC
Didier Drogba	NA	NA	81	189	80	39	Medium / Low	80	76	NA	NO	22.39579

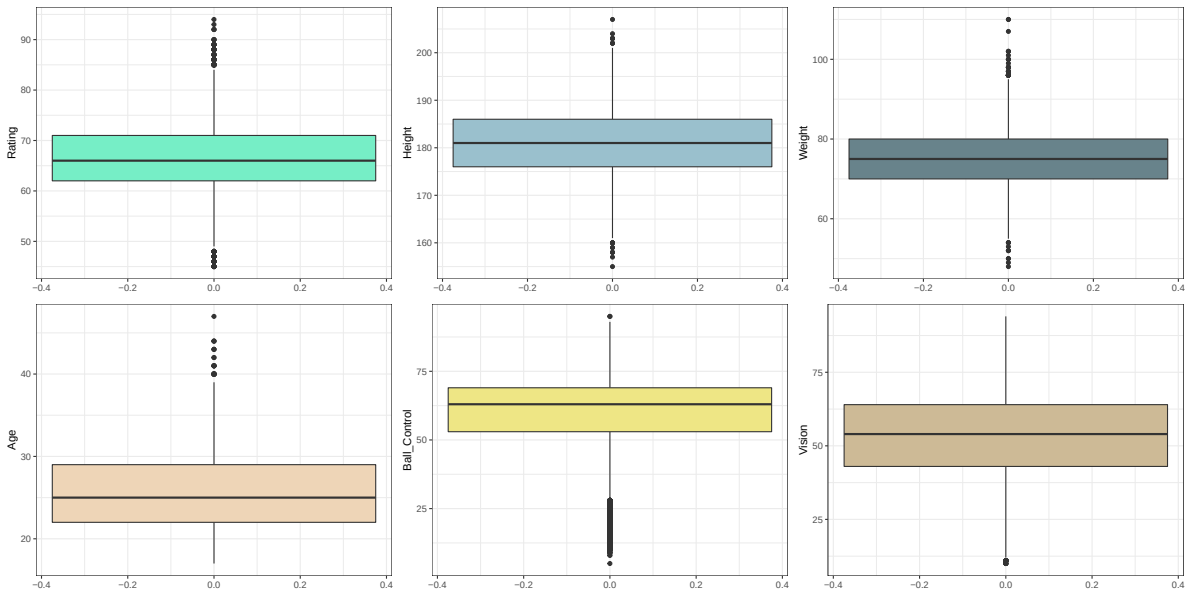
Comprovem que només hi ha un jugador afectat. Per tant, decidim eliminar aquest registre.

```
# Eliminació dels registres missing identificats
df <- filter(df, !is.na(Club_Position))
```

5.4 Tractament dels valors extrems

Per avaluar l'existència de valors extrems dibuixarem els diagrames de caixa de les variables numèriques.

```
# Boxplot (comprovació d'outliers)
bp1 <- ggplot(df, aes(y=Rating)) + geom_boxplot(fill="aquamarine2") + theme_bw()
bp2 <- ggplot(df, aes(y=Height)) + geom_boxplot(fill="lightblue3") + theme_bw()
bp3 <- ggplot(df, aes(y=Weight)) + geom_boxplot(fill="lightblue4") + theme_bw()
bp4 <- ggplot(df, aes(y=Age)) + geom_boxplot(fill="bisque2") + theme_bw()
bp5 <- ggplot(df, aes(y=Ball_Control)) + geom_boxplot(fill="khaki2") + theme_bw()
bp6 <- ggplot(df, aes(y=Vision)) + geom_boxplot(fill="wheat3") + theme_bw()
grid.arrange(bp1,bp2,bp3,bp4,bp5,bp6, nrow=2, ncol=3)
```



S'observen *outliers* en totes les variables. Això no obstant, prendrem en consideració que:

- Basant-nos en el resum estadístic general, els valors extrems estan dins del rang possible en cada cas, i no són fruit d'una mala recol · lecció de dades.

- En una primera aproximació als problemes d'investigació plantejats ens pot interessar no prescindir-ne. Fer-ho podria esbiaixar alguna conclusió: així, per exemple, és obvi que hi ha uns pocs jugadors que són *cracs mundials* i tenen *ratings* extraordinaris.

- En el cas de la variable *Ball_Control*, una anàlisi detallada de les dades ens indica que els valors més baixos corresponen als *porters*, la qual cosa és perfectament normal.

- Només en cas de que la normalitat de les dades estigués *molt seriosament compromesa*, i en una eventual segona iteració, entenc que seria apropiat prescindir-ne.

En conseqüència, mantindrem aquests valors.

6 Estudi de la normalitat

6.1 Tests de normalitat

Començarem l'estudi aplicant tests de normalitat. Val a dir que la literatura ens prevé sobre aquests tipus de test, en el sentit que s'han de complementar amb estudis gràfics, atès que alguns són molt sensibles a aspectes com els *outliers* i el volum de dades. Atès l'alt volum d'observacions emprarem *Lilliefors*, *Anderson-Darling* i *Cramer-von Mises*.

```
# Test de normalitat

# Funció que aplica els tests sobre una variable passada per paràmetre
Test_Normalitat <- function(Camp, Nom, Matriu) {
  ll <- lillie.test(Camp)$p.value
  ad <- ad.test(Camp)$p.value
  cm <- cvm.test(Camp, "pnorm")$p.value
  Matriu <- rbind(Matriu, as.character(c(ll, ad, cm)))
  rownames(Matriu)[nrow(Matriu)] <- Nom
  return(Matriu)
}

# Matriu amb el resultat dels tests
M <- matrix(nrow=0, ncol=3)
colnames(M) <- c("Lilliefors", "Anderson-Darling", "Cramer-von Mises")

# Bucle que aplica els tests a les variables numèriques
for (i in 1:ncol(df)) {
  if (class(df[,i]) %in% c("integer", "numeric")) {
    M <- Test_Normalitat(df[,i], colnames(df)[i], M)
  }
}

# Presentació del resultat
kable(M, booktabs = TRUE, caption = "Test de normalitat") %>%
  kable_styling(latex_options = "HOLD_position", position = "center")
```

Test de normalitat

	Lilliefors	Anderson-Darling	Cramer-von Mises
Rating	2.07269187727758e-89	3.4116108493798e-08	0
Height	1.21708098637971e-109	3.4116108493798e-08	0
Weight	5.01278218468443e-148	3.4116108493798e-08	0
Age	0	3.4116108493798e-08	0
Ball_Control	0	3.4116108493798e-08	0
Vision	2.96295849275147e-152	3.4116108493798e-08	0
IMC	5.59647899733609e-25	3.4116108493798e-08	0

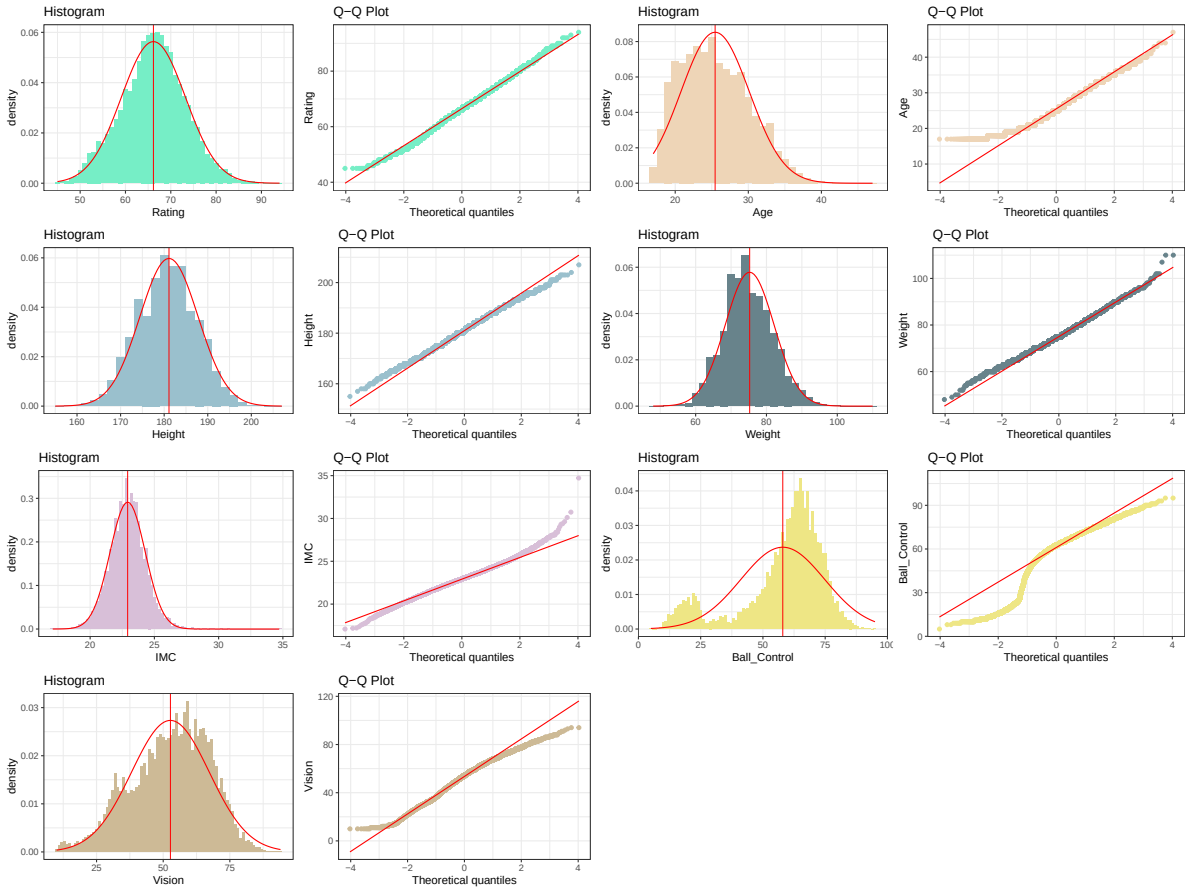
Constatem que, per a totes les variables i en tots els mètodes, el p-valor és significatiu. Per tant, rebutgem la hipòtesi nul · la i podem concloure amb un nivell de confiança del 95% que les distribucions no són normals.

En tot cas, no ens conformarem amb aquesta informació i tot seguit farem una anàlisi gràfica.

6.2 Histogrames i gràfics Q-Q

```
# Histogrames i gràfics Q-Q (estudi de la normalitat)

h1 <- ggplot(data=df, aes(x = Rating)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="aquamarine2",  binwidth=1) +
      geom_vline(aes(xintercept=mean(Rating)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Rating), sd=sd(df$Rating)))

q1 <- ggplot(df, aes(sample=Rating)) + theme_bw() +
      stat_qq(col='aquamarine2') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Rating")

h2 <- ggplot(data=df, aes(x = Age)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="bisque2",  binwidth=1) +
      geom_vline(aes(xintercept=mean(Age)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Age), sd=sd(df$Age)))

q2 <- ggplot(df, aes(sample=Age)) + theme_bw() +
      stat_qq(col='bisque2') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Age")


h3 <- ggplot(data=df, aes(x = Height)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="lightblue3",  binwidth=2) +
      geom_vline(aes(xintercept=mean(Height)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Height), sd=sd(df$Height)))

q3 <- ggplot(df, aes(sample=Height)) + theme_bw() +
      stat_qq(col='lightblue3') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Height")
```

```
h4 <- ggplot(data=df, aes(x = Weight)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="lightblue4",  binwidth=2) +
      geom_vline(aes(xintercept=mean(Weight)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Weight), sd=sd(df$Weight)))

q4 <- ggplot(df, aes(sample=Weight)) + theme_bw() +
      stat_qq(col='lightblue4') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Weight")

h5 <- ggplot(data=df, aes(x = IMC)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="thistle",  binwidth=0.2) +
      geom_vline(aes(xintercept=mean(IMC)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$IMC), sd=sd(df$IMC)))

q5 <- ggplot(df, aes(sample=IMC)) + theme_bw() +
      stat_qq(col='thistle') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("IMC")

h6 <- ggplot(data=df, aes(x = Ball_Control)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="khaki2",  binwidth=1) +
      geom_vline(aes(xintercept=mean(Ball_Control)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Ball_Control), sd=sd(df$Ball_Control)))

q6 <- ggplot(df, aes(sample=Ball_Control)) + theme_bw() +
      stat_qq(col='khaki2') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Ball_Control")

h7 <- ggplot(data=df, aes(x = Vision)) + theme_bw() + ggtitle("Histogram") +
      geom_histogram(aes(y=..density..), fill="wheat3",  binwidth=1) +
      geom_vline(aes(xintercept=mean(Vision)), color="red", size=0.2) +
      stat_function(fun=dnorm, lwd=0.5,
                    col='red', args=list(mean=mean(df$Vision), sd=sd(df$Vision)))

q7 <- ggplot(df, aes(sample=Vision)) + theme_bw() +
      stat_qq(col='wheat3') + stat_qq_line(lwd=0.5, col='red') +
      ggtitle("Q-Q Plot") + xlab("Theoretical quantiles") +ylab("Vision")

grid.arrange(h1,q1,h2,q2,h3,q3,h4,q4,h5,q5,h6,q6,h7,q7, nrow=4, ncol=4)
```



Els histogrames i els gràfics Q-Q mostren que, en la majoria de casos, i sobretot en els valors centrals de cada distribució, la normalitat pot arribar a acceptar-se. Això no és ben bé així per a les variables *Age*, per l'alta concentració al voltant dels 20 anys, *Ball_Control*, per la clara diferenciació entre porters i jugadors de camp (aquesta anàlisi no es reprodueix aquí per la limitació del màxim nombre de pàgines del treball), i *Vision*.

6.3 Teorema del Límit Central (TLC)

Tot i el resultat de l'anàlisi de normalitat de la secció anterior, és molt rellevant esmentar que en les condicions d'aquest dataset és aplicable el *Teorema del Límit Central*. En virtut del TLC, amb independència de la distribució de la població d'estudi, la distribució de la mitjana mostral és aproximadament normal per a mostres grans (N>30), i ho és cada vegada més a mesura que N augmenta.

En aquest cas totes les *mostres aleatòries simples*, que considerarem en les proves estadístiques que es realitzaran a continuació, tenen un tamany bastant més alt del límit fixat (N>30).

7 Resolució dels reptes d'investigació

7.1 Pregunta 1: correlació

La característica que està més relacionada amb la valoració, és la mateixa per als porters i que per als jugadors de camp?

En primer lloc haurem de fer dos grups: porters i jugadors de camp, seleccionant en ambdós casos les variables numèriques.

```
# Segmentació de variables numèriques per a porters i jugadors
Porters  <- df %>% filter(Goalkeeper=="YES") %>%
            select(Rating, Height, Weight, Age, Ball_Control, Vision, IMC)

Jugadors <- df %>% filter(Goalkeeper=="NO") %>%
            select(Rating, Height, Weight, Age, Ball_Control, Vision, IMC)
```

Ara executarem un test de correlació en cada grup per determinar la significació estadística; així que ens fixarem en els *p-values*, tenint en compte que $H_0 : \rho = 0$ i $H_1 : \rho \neq 0$, és a dir, H_1 implica correlació. En concret analitzarem la relació de *Rating* amb cadascuna de la resta de variables. Val a dir que, tot i les consideracions expressades en l'anàlisi de normalitat, aplicarem el mètode de *Spearman* per a major seguretat.

Començarem pels porters.

```
# Anàlisi de correlació per als porters
kable(rcorr(as.matrix(Porters), type = "spearman")$P, booktabs = TRUE) %>%
  kable_styling(latex_options = c("HOLD_position","scale_down"), position = "center")
```

	Rating	Height	Weight	Age	Ball_Control	Vision	IMC
Rating	NA	0.0101108	0.0000023	0.0000000	0.0000101	0.0000000	0.0005132
Height	0.0101108	NA	0.0000000	0.1756720	0.9589791	0.5714616	0.0094613
Weight	0.0000023	0.0000000	NA	0.0000024	0.7456266	0.1521294	0.0000000
Age	0.0000000	0.1756720	0.0000024	NA	0.0482957	0.0069533	0.0000000
Ball_Control	0.0000101	0.9589791	0.7456266	0.0482957	NA	0.0004065	0.6913692
Vision	0.0000000	0.5714616	0.1521294	0.0069533	0.0004065	NA	0.0119231
IMC	0.0005132	0.0094613	0.0000000	0.0000000	0.6913692	0.0119231	NA

Observem que, respecte de *Rating*, en tots els casos $p - value < \alpha = 0.05$. Per tant, rebutgem la hipòtesi nul · la i acceptem la correlació. Ara continuarem amb els jugadors de camp.

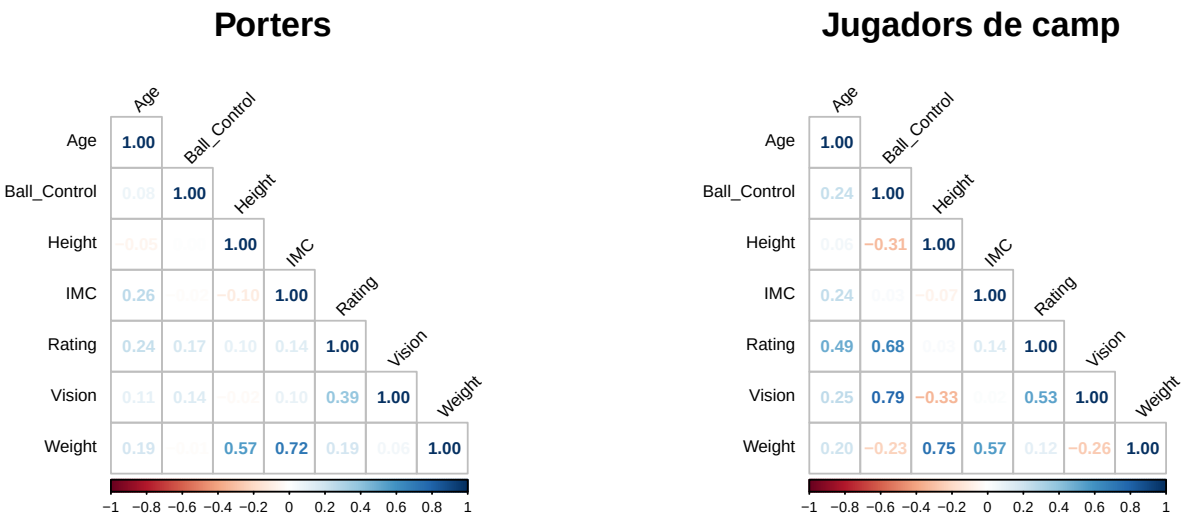
```
# Anàlisi de correlació per als jugadors de camp
kable(rcorr(as.matrix(Jugadors), type = "spearman")$P, booktabs = TRUE) %>%
  kable_styling(latex_options = c("HOLD_position","scale_down"), position = "center")
```

	Rating	Height	Weight	Age	Ball_Control	Vision	IMC
Rating	NA	0.0003704	0	0	0.0000000	0.0000000	0.0000000
Height	0.0003704	NA	0	0	0.0000000	0.0000000	0.0000000
Weight	0.0000000	0.0000000	NA	0	0.0000000	0.0000000	0.0000000
Age	0.0000000	0.0000000	0	NA	0.0000000	0.0000000	0.0000000
Ball_Control	0.0000000	0.0000000	0	0	NA	0.0000000	0.0008995
Vision	0.0000000	0.0000000	0	0	0.0000000	NA	0.0196772
IMC	0.0000000	0.0000000	0	0	0.0008995	0.0196772	NA

En aquest cas també els *p-value* són significatius i, per tant, acceptem la correlació.

Atès que la resposta dels test de correlació ha estat favorable, tot seguit mostrarem de forma gràfica els coeficients de correlació.

```r
# Coeficients de correlació
par(mfrow=c(1,2))

corrplot(corr = cor(x = Porters, method = "spearman"),
         method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, order = "alphabet",
         mar=c(0,0,1,0), title="Porters")

corrplot(corr = cor(x = Jugadors, method = "spearman"),
         method = "number", type = "lower", cl.cex = 0.5, number.cex = 0.6,
         tl.col = "black", tl.srt = 45, tl.cex = 0.6, order = "alphabet",
         mar=c(0,0,1,0), title="Jugadors de camp")
```



En el cas dels *porters* la relació més important de *Rating* és amb *Vision* que, amb coeficient 0.39, és manifestament baixa. Al seu torn, en el cas dels *jugadors de camp* la relació més important és amb *Ball_Control* que, amb coeficient 0.68, és moderada-alta. En conclusió, es constata que hi ha una diferència entre *porters* i *jugadors de camp*.

7.2 Pregunta 2: contrast d'hipòtesis

Podem afirmar que la mitjana de l'IMC de porters i jugadors de camp és igual?

Considerarem la v.a. X_P definida sobre la població de l'IMC dels porters, amb mitjana poblacional μ_P. Al seu torn, també considerarem la v.a. X_J definida sobre la població de l'IMC dels jugadors de camp, amb mitjana poblacional μ_J. Això ens porta a la formulació de les següents hipòtesis nul · la H_0 i alternativa H_1:

$$\begin{cases} H_0: & \mu_P = \mu_J \\ H_1: & \mu_P \neq \mu_J \end{cases}$$

Sabem que les variàncies poblacionals són desconegudes, però haurem d'estimar si són iguals o diferents de cara a saber quin test aplicar. Per tal de determinar-ho farem un test d'igualtat de variàncies de dues mostres (o *test d'homoscedasticitat*). Usarem la funció *var.test*, i tindrem en compte que $\alpha = 0.05$ i estem treballant amb un test sobre variàncies bilateral (ja que $H_0 : \sigma_P^2 = \sigma_J^2$ i $H_1 : \sigma_P^2 \neq \sigma_J^2$).

Primer segmentarem les mostres aleatòries simples de les dues poblacions.

```
# Segmentació d'IMC de porters i jugadors
IMC_porters <-  df$IMC[df$Goalkeeper=="YES"]
IMC_jugadors <- df$IMC[df$Goalkeeper=="NO"]
```

I ara farem el test d'homoscedasticitat.

```
# Test d'homoscedasticitat per a IMC, porters vs. jugadors
var.test(x=IMC_porters, y=IMC_jugadors, alternative="two.sided", conf.level=0.95)
```

```
##
##  F test to compare two variances
##
## data:  IMC_porters and IMC_jugadors
## F = 1.0648, num df = 631, denom df = 16954, p-value = 0.2611
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9543848 1.1951119
## sample estimates:
## ratio of variances
##           1.064816
```

Atès que el p-valor obtingut és $p = 0.2611 > \alpha = 0.05$ no podem rebutjar la hipòtesi nul · la i podem concloure que les variàncies de les dues poblacions són iguals amb un nivell de confiança del 95%.

En conclusió, per tal de respondre a la pregunta d'investigació, haurem d'aplicar un test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals. A més, tal com es desprèn de la formulació de les hipòtesis, es tracta d'un test bilateral.

```
# Càlcul del test d'hipòtesis amb t.test (IMC, porters vs. jugadors)
t.test(x=IMC_porters, y=IMC_jugadors,
       alternative="two.sided", var.equal=TRUE, conf.level=0.95)
```

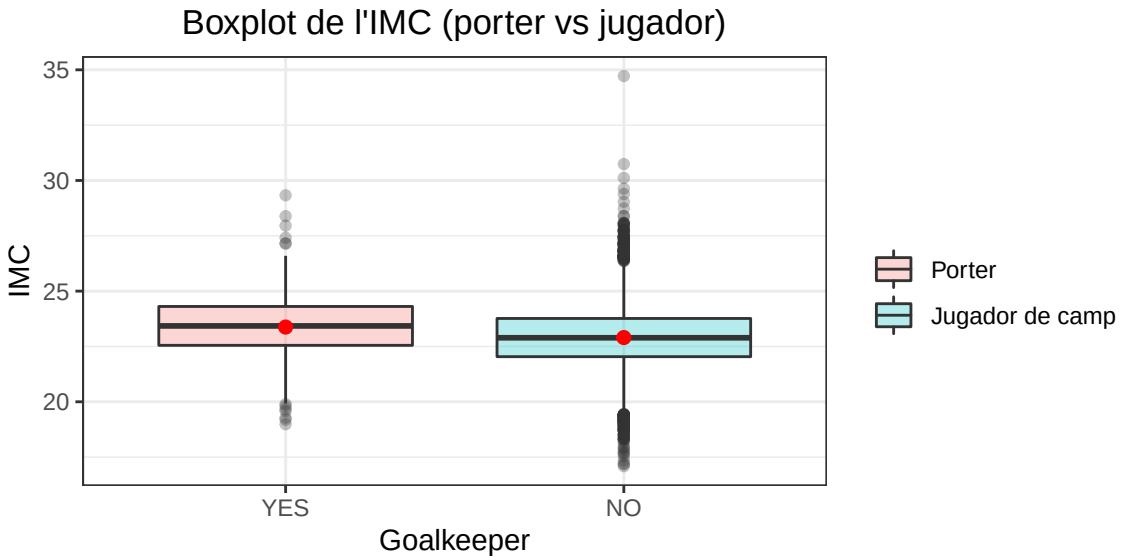
```
##
##  Two Sample t-test
##
## data:  IMC_porters and IMC_jugadors
## t = 8.5682, df = 17585, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3659286 0.5830138
## sample estimates:
## mean of x mean of y
##  23.37706  22.90259
```

Atès que el p-valor obtingut és $p < 2.2 \times 10^{-16} < \alpha = 0.05$ rebutgem la hipòtesi nul · la. Consegüentment, amb un nivell de confiança del 95%, podem concloure que les mitjanes poblacionals de l'IMC de porters i jugadors de camp són diferents.

En termes gràfics això significa que, tot i que ens sembli que les mitjanes són molt i molt properes, la diferència de valors és estadísticament significativa.

```
# Diagrames de caixa de l'IMC per a Goalkeeper

ggplot(df, aes(x=Goalkeeper, y=IMC, fill = Goalkeeper)) +
geom_boxplot(alpha=0.3, show.legend = TRUE) + theme_bw() +
```

```
stat_summary(fun=mean, geom="point", size=2, color="red", fill="red") +
scale_fill_discrete(name = "", labels = c("Porter", "Jugador de camp")) +
ggtitle("Boxplot de l'IMC (porter vs jugador)") + xlab("Goalkeeper") +
theme(plot.title = element_text(hjust = 0.5))
```



7.3 Pregunta 3: regressió lineal múltiple

Quina combinació de característiques esportives explica millor la valoració d'un jugador?

Per respondre aquesta pregunta plantejarem models de regressió lineal sobre la variable *Rating*, considerant les variables associades a les característiques esportives: *Work_Rate*, *Ball_Control*, *Vision* i *Goalkeeper*.

```
# Models de regressió lineal múltiple per a 'Rating'
m1  <- lm(data=df, Rating ~ Work_Rate)
m2  <- lm(data=df, Rating ~ Work_Rate + Ball_Control)
m3  <- lm(data=df, Rating ~ Work_Rate + Vision)
m4  <- lm(data=df, Rating ~ Work_Rate + Goalkeeper)
m5  <- lm(data=df, Rating ~ Work_Rate + Ball_Control + Goalkeeper)
m6  <- lm(data=df, Rating ~ Work_Rate + Vision + Goalkeeper)
m7  <- lm(data=df, Rating ~ Work_Rate + Ball_Control + Vision)
m8  <- lm(data=df, Rating ~ Work_Rate + Ball_Control + Vision + Goalkeeper)
m9  <- lm(data=df, Rating ~ Ball_Control)
m10 <- lm(data=df, Rating ~ Ball_Control + Vision)
m11 <- lm(data=df, Rating ~ Ball_Control + Goalkeeper)
m12 <- lm(data=df, Rating ~ Ball_Control + Vision + Goalkeeper)
m13 <- lm(data=df, Rating ~ Vision)
m14 <- lm(data=df, Rating ~ Vision + Goalkeeper)
m15 <- lm(data=df, Rating ~ Goalkeeper)
```

Ara analitzarem la bondat d'ajust per determinar la millor combinació, comparant el valor del coeficient de determinació $R^2_{ajustat}$ i també l'índex *AIC* (Akaike Information Criterion).

```
# Anàlisi dels coeficients de determinació
R <- matrix(c(
```

```
              1, round(summary(m1)$adj.r.squared,4), round(AIC(m1),3),
              2, round(summary(m2)$adj.r.squared,4), round(AIC(m2),3),
              3, round(summary(m3)$adj.r.squared,4), round(AIC(m3),3),
              4, round(summary(m4)$adj.r.squared,4), round(AIC(m4),3),
              5, round(summary(m5)$adj.r.squared,4), round(AIC(m5),3),
              6, round(summary(m6)$adj.r.squared,4), round(AIC(m6),3),
              7, round(summary(m7)$adj.r.squared,4), round(AIC(m7),3),
              8, round(summary(m8)$adj.r.squared,4), round(AIC(m8),3),
              9, round(summary(m9)$adj.r.squared,4), round(AIC(m9),3),
              10, round(summary(m10)$adj.r.squared,4), round(AIC(m10),3),
              11, round(summary(m11)$adj.r.squared,4), round(AIC(m11),3),
              12, round(summary(m12)$adj.r.squared,4), round(AIC(m12),3),
              13, round(summary(m13)$adj.r.squared,4), round(AIC(m13),3),
              14, round(summary(m14)$adj.r.squared,4), round(AIC(m14),3),
              15, round(summary(m15)$adj.r.squared,4), round(AIC(m15),3)
             ), ncol = 3, byrow = TRUE)
colnames(R) <- c("Núm. model", "R2 ajustat", "AIC")

kable(t(R), booktabs = TRUE,
      caption="Comparació de models mitjançant R2 ajustat i AIC") %>%
  kable_styling(latex_options = c("HOLD_position","scale_down"), position = "center")
```

Comparació de models mitjançant R2 ajustat i AIC

Núm. model	1.0000	2.000	3.000	4.0000	5.0000	6.0000	7.0000	8.0000	9.0000	10.000	11.0000	12.0000	13.0000	14.0000	15.0000
R2 ajustat	0.0644	0.227	0.262	0.0845	0.3371	0.3044	0.2785	0.3623	0.2145	0.263	0.3216	0.3439	0.2393	0.2739	0.0099
AIC	117605.8790	114249.962	113434.420	117225.3230	111547.8930	112395.4330	113038.3280	110869.2450	114523.9140	113404.223	111948.3600	111359.4210	113959.0170	113143.3440	118595.4510

Pel que fa a la qualitat de l'ajust tenim que $0 \leq R^2 \leq 1$, i s'estableix que la bondat d'ajust és millor quan més proper a 1 és el valor de R^2. En aquest cas el millor model és el 8è, amb un valor $R^2_{ajustat} = 0.3623$: això significa que el model només és capaç d'explicar el 36.23% de la variabilitat observada a *Rating*. Aquest valor es podria considerar com a *molt pobre*. Al seu torn, l'índex AIC també indica que el 8è model és el millor, ja que obté el menor valor d'AIC: 110869.2450.

En tot cas, analitzarem aquest model.

```
# Anàlisi del millor model de regressió lineal múltiple
summary(m8)
```

```
##
## Call:
## lm(formula = Rating ~ Work_Rate + Ball_Control + Vision + Goalkeeper,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7060  -3.7218  -0.2137   3.4540  27.1399
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               63.944370   0.353002 181.144  < 2e-16 ***
## Work_RateHigh / Low       -2.596560   0.294621  -8.813  < 2e-16 ***
## Work_RateHigh / Medium    -1.551691   0.232320  -6.679 2.48e-11 ***
## Work_RateLow / High        0.977799   0.345262   2.832  0.00463 **
## Work_RateLow / Low        -2.813280   1.053165  -2.671  0.00756 **
```

```
## Work_RateLow / Medium      0.162126   0.343450   0.472  0.63690
## Work_RateMedium / High    -0.086245   0.254125  -0.339  0.73433
## Work_RateMedium / Low     -2.757295   0.284239  -9.701  < 2e-16 ***
## Work_RateMedium / Medium  -2.389535   0.219640 -10.879  < 2e-16 ***
## Ball_Control               0.170624   0.004272  39.943  < 2e-16 ***
## Vision                     0.117787   0.004473  26.335  < 2e-16 ***
## GoalkeeperNO             -12.501641   0.260157 -48.054  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.656 on 17575 degrees of freedom
## Multiple R-squared:  0.3627, Adjusted R-squared:  0.3623
## F-statistic: 909.2 on 11 and 17575 DF,  p-value: < 2.2e-16
```

El p-valor del *test F* de contrast conjunt del model ens dona un valor estadísticament significatiu ($p-value < \alpha = 0.05$, essent $H_0 : \beta_i = 0, \forall i$) i, per tant, el model sembla vàlid en conjunt. Els p-valors corresponents al *test t-Student* de contrast dels coeficients individuals són significatius ($p-value < \alpha = 0.05$, essent $H_0 : \beta_i = 0$) en tots els casos, excepte per a certes combinacions de la variable *Work_Rate*.

Així, l'equació de la recta de regressió obtinguda és (valors truncats a 2 decimals):

$$\begin{aligned}\widehat{Rating} = {} & 63.94 - 2.59 \cdot Work\_Rate\_HL - 1.55 \cdot Work\_Rate\_HM \\ & + 0.97 \cdot Work\_Rate\_LH - 2.81 \cdot Work\_Rate\_LL + 0.16 \cdot Work\_Rate\_LM \\ & - 0.08 \cdot Work\_Rate\_MH - 2.75 \cdot Work\_Rate\_ML - 2.38 \cdot Work\_Rate\_MM \\ & + 0.17 \cdot Ball\_Control + 0.11 \cdot Vision - 12.50 \cdot Goalkeeper\end{aligned}$$

tenint en compte que, a la fórmula:

- *Goalkeeper = 1* quan pren el valor *NO* (posició de jugador de camp), i 0 en cas contrari.
- *Work_Rate_XY = 1* quan la variable categòrica pren el valor *"X / Y"*, i 0 en cas contrari.

En aquest punt podem respondre ja l'altra pregunta d'aquesta secció: *Com afecta a la valoració ser o no porter?* D'acord al coeficient de la recta de regressió vinculat a la variable *Goalkeeper*, que té valor -12.501641, el fet de ser porter proporciona una millor valoració. Això és així perquè un jugador de camp té associat el valor *Goalkeeper = NO*, fet que es tradueix en el valor *«1»* en la variable *Goalkeeper* de la fórmula. Llavors, en un tal cas, la valoració es veu disminuïda en 12.501641 unitats.

Un cop disposem de la fórmula de l'equació també podem fer prediccions. Per exemple, podem predir el *Rating* per a la combinació següent: *Work_Rate* = High / High, *Ball_Control* = 60, *Vision* = 80 i *Goalkeeper* = NO. Per dur-ho a terme usarem la funció *predict*.

```
# Predicció amb el model de regressió lineal múltiple
P <- predict(m8,
     data.frame(Work_Rate="High / High", Ball_Control=60, Vision=80, Goalkeeper="NO"),
     interval = "prediction")

kable(P, booktabs = TRUE, caption = "Interval de predicció al 95 per cent") %>%
  kable_styling(latex_options = "HOLD_position")
```

Interval de predicció al 95 per cent

fit	lwr	upr
71.10309	60.00774	82.19844

La conclusió respecte del valor predit de la valoració és: $\widehat{Rating} = 71.10$

7.4 Pregunta 4: regressió logística

Quina és la probabilitat de ser internacional en funció de la valoració, l'edat i la posició?

Per donar resposta a aquesta pregunta ajustarem un model de regressió logística: la variable dependent serà *International*, i les variables independents seran *Rating, Age* i *Goalkeeper*. La regressió logística ens proporcionarà la fórmula mitjançant la qual s'obtindrà la probabilitat.

```
# Regressió logística per a 'International'
rl <- glm(formula = International ~ Rating + Age + Goalkeeper,
          data = df, family = binomial(link = logit))
summary(rl)
```

```
##
## Call:
## glm(formula = International ~ Rating + Age + Goalkeeper, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6412  -0.3487  -0.2249  -0.1361   3.5426
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -15.796227   0.466298 -33.876  < 2e-16 ***
## Rating         0.202598   0.005822  34.797  < 2e-16 ***
## Age           -0.029574   0.008503  -3.478 0.000505 ***
## GoalkeeperNO  -0.356958   0.140013  -2.549 0.010789 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8091.8  on 17586  degrees of freedom
## Residual deviance: 6477.4  on 17583  degrees of freedom
## AIC: 6485.4
##
## Number of Fisher Scoring iterations: 6
```

Observem com els p-valors corresponents al *test de Wald* de contrast dels coeficients individuals són significatius ($p - value < \alpha = 0.05$, essent $H_0 : \beta_i = 0$) per a totes les variables independents. Per tant, totes les variables contribueixen a l'explicació del model.

En termes de *logit*, la fórmula del model de regressió estimat és (valors truncats a 2 decimals):

$$\ln\left(\frac{P(International = 1 \mid X)}{1 - P(International = 1 \mid X)}\right) = -15.79 + 0.20 \cdot Rating - 0.02 \cdot Age - 0.35 \cdot Goalkeeper$$

tenint en compte que, a la fórmula, *Goalkeeper = 1* quan pren el valor *NO* (posició de jugador de camp) i és 0 en cas contrari.

Per tant, la fórmula que ens proporciona la probabilitat és la següent:

$$P(International = 1 \mid X) = \frac{\exp\left(-15.79 + 0.20 \cdot Rating - 0.02 \cdot Age - 0.35 \cdot Goalkeeper\right)}{1 + \exp\left(-15.79 + 0.20 \cdot Rating - 0.02 \cdot Age - 0.35 \cdot Goalkeeper\right)}$$

Respecte a l'efecte de les variables comprovem que: *Age* i *Goalkeeper* tenen coeficients negatius i, per tant, augments en aquestes variables provoquen una disminució en la probabilitat; *Rating* te coeficient positiu i, per tant, augments en aquesta variable provoca un augment en la probabilitat. En resum, observem que l'única variable que contribueix positivament a la probabilitat de ser internacional és la variable *Rating*.

En aquest punt podem respondre ja l'altra pregunta d'aquesta secció: *En quin percentatge augmenta o disminueix aquesta probabilitat el fet de ser o no porter?*

Per respondre la pregunta haurem de calcular els *Odds Ratio* de les variables explicatives. De forma genèrica, podem expressar els *Odds Ratio* de de la manera següent.

$$OR(International \mid X) = \frac{ODD(International \mid X = 1)}{ODD(International \mid X = 0)} = \frac{\frac{P(International=1|X=1)}{1-P(International=1|X=1)}}{\frac{P(International=1|X=0)}{1-P(International=1|X=0)}} = e^{\beta_X}$$

on β_X és el coeficient estimat de la variable explicativa X que s'està analitzant. Per tant, cada *Odd Ratio* s'obté fent l'exponencial del coeficient respectiu.

```
# Càlculs dels ODD RATIO
kable(exp(coefficients(rl)),
      booktabs = TRUE, caption = "Odd Ratio estimat", col.names=c("OR")) %>%
  kable_styling(latex_options = "HOLD_position")
```

Odd Ratio estimat

	OR
(Intercept)	0.0000001
Rating	1.2245802
Age	0.9708590
GoalkeeperNO	0.6998021

I ara analitzarem l'*Odd Ratio* de la variable *Goalkeeper*: atès que $OR = 0.6998021 < 1$, ens trobem davant del que s'anomena *factor de protecció*: el fet de ser internacional és menys probable en presència d'aquesta variable. Ara bé, el nivell de referència de la variable categòrica és *ser porter* (valor "YES" en el dataset). Per tant, quan es passa de *ser porter* a *ser jugador de camp* tenim que *ODD(International)* es multiplica per 0.6998021, és a dir, és 0.6998021 vegades menor: la probabilitat de ser internacional (respecte a no ser-ho) disminueix un 30.02% (aprox).

En sentit contrari, quin és l'augment de la probabilitat de ser internacional si s'és porter respecte a ser jugador de camp (i.e. quan *es passa de ser jugador de camp a ser porter*)? Per respondre la pregunta cal fer la inversa de l'*Odd Ratio*: $1/OR = 1/0.6998021 = 1.4289 \implies 42.89\%$.

En tot cas la conclusió és que, fixades les altres variables del model, la probabilitat de ser internacional és major si s'és porter.

També en aquest cas podem fer prediccions. En particular, en ajustar el model de regressió logística ja s'obté el valor de probabilitat per a cada observació. Aquest valor es troba a la variable *fitted.values* del model generat. Per tant, gairebé ja estem en condicions de generar una *taula de confusió*.

Per tal d'elaborar-la, primer categoritzarem els valors de les probabilitats predites pel model prenent com a llindar el valor 0.5 (probabilitat del 50%): si la probabilitat és superior a 0.5 assignarem la categoria "YES" (assumirem que s'ha predit ser internacional), i en cas contrari la categoria "NO" (assumirem que s'ha predit una no internacionalitat). De tal forma que, per a la variable *International*, confrontarem els valors *observats* amb els valors *predits*.

```
# Selecció de prediccions i observacions
prediccions  <- ifelse(test = rl$fitted.values > 0.5, yes = "YES", no = "NO")
observacions <- as.character(df$International)

# Creació de la matriu (reordenem files i columnes)
matriu_confusio <- table(prediccions, observacions)
matriu_confusio <- matriu_confusio[2:1, 2:1]
```

I ara, amb l'ajuda de la funció *confusionMatrix* de la biblioteca *caret*, obtenim els principals indicadors.

```
# Indicadors de la matriu de confusió
confusionMatrix(matriu_confusio, positive = "YES")
```

```
## Confusion Matrix and Statistics
##
##            observacions
## prediccions   YES    NO
##         YES    91    38
##         NO    984 16474
##
##                Accuracy : 0.9419
##                  95% CI : (0.9383, 0.9453)
##     No Information Rate : 0.9389
##     P-Value [Acc > NIR] : 0.04838
##
##                   Kappa : 0.1399
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.084651
##             Specificity : 0.997699
##          Pos Pred Value : 0.705426
##          Neg Pred Value : 0.943636
##              Prevalence : 0.061125
##          Detection Rate : 0.005174
##    Detection Prevalence : 0.007335
##       Balanced Accuracy : 0.541175
##
##        'Positive' Class : YES
##
```

Observem que el model té una exactitud (*accuracy*) del 94%; una sensibilitat (*sensitivity*) del 8%; i una especificitat (*specificity*) del 99%. Així, tot i que l'exactitud és molt alta, el model fa prediccions molt dolentes pel que respecta a predir com a internacionals (i.e. *veritables positius*) els que realment ho són (i.e. *total de positius observats*).

8 Conclusió de l'estudi

Essent conscients de les assumpcions fetes al llarg del desenvolupament de l'exercici, podem concloure que hem pogut donar una resposta satisfactòria a cadascuna de les preguntes d'investigació plantejades a l'inici. Així, amb un nivell de confiança del 95% segons el cas, es constata que:

1. En el cas dels *porters* la relació més important de *Rating* és amb *Vision*, per bé que en el cas dels *jugadors de camp* la relació més important és amb *Ball_Control*.
2. Les mitjanes poblacionals de l'IMC de porters i jugadors de camp són diferents.
3. La millor explicació de la valoració d'un jugador s'obté considerant la combinació de totes les característiques esportives. A més, en aquest context, el fet de ser porter proporciona una millor valoració.
4. Ha estat possible l'obtenció d'una fórmula que calculi la probabilitat de ser internacional, en funció de les característiques fixades, malgrat que la capacitat predictiva en termes de sensibilitat és extremadament baixa. Addicionalment es conclou que, fixades les altres variables del model, la probabilitat de ser internacional és major si s'és porter.

9 Gravació del dataset processat

D'acord als requeriments de la pràctica, procedirem a gravar en un fitxer *csv* el dataset resultant del procès de neteja. S'anomenarà *Fifa2017_final.csv*.

```
# Gravació del dataset processat (net) en fitxer CSV
write.csv(df, "Fifa2017_final.csv", row.names = FALSE)
```

10 Bibliografia

1. Calvo, M., Pérez, D., Subirats, L. (2019). *Introducció a la neteja i anàlisi de dades*. FUOC.
2. Bernadó, E. (2020). *Tests d'hipòtesis*. FUOC.
3. Gibergans, J. (2009). *Regressió lineal múltiple*. FUOC.
4. Guillén, M., Alonso, M. (2020). *Models de regressió logística*. FUOC.
5. Squire, Megan (2015). *Clean Data*. Packt Publishing Ltd.
6. Dalgaard, Peter (2008). *Introductory statistics with R*. Springer Science & Business Media.
7. Wikipedia. *Normality test* [en línia]. Actualitzada: 2021. [Data de consulta: 21 de desembre de 2021]. Disponible a: https://en.wikipedia.org/wiki/Normality_test

11 Taula de contribucions al treball

La realització de totes les tasques directes i indirectes necessàries per a elaborar tots els apartats que constitueixen aquest document, així com la resta d'elements pertanyents a la pràctica 2, s'ha fet exclusivament de forma individual pel seu autor: **Nicolás González Soler**.

La totalitat del treball corresponent a aquesta pràctica es troba disponible al repositori Github https://github.com/ngonzalezs-UOC/Fifa2017-GK-vs-Others.