GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

CURS 2017-2018 Q1 – EXAMEN PARCIAL : MODEL LINEAL GENERALITZAT

(Data: 10 d'Octubre del 2017 a les 15:00 Aula S02-FME)

Professors:	Erik Cobo – Jordi Cortés – Josep Anton Sanchez
Localització:	ETSEIB 6a Planta 6-67
Normativa de l'examen:	ES POT DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 24 d'Octubre al Web Docent de MLGz
Revisió de l'examen:	24 d'octubre a 12 h a Sala Professors FME– Campus Sud

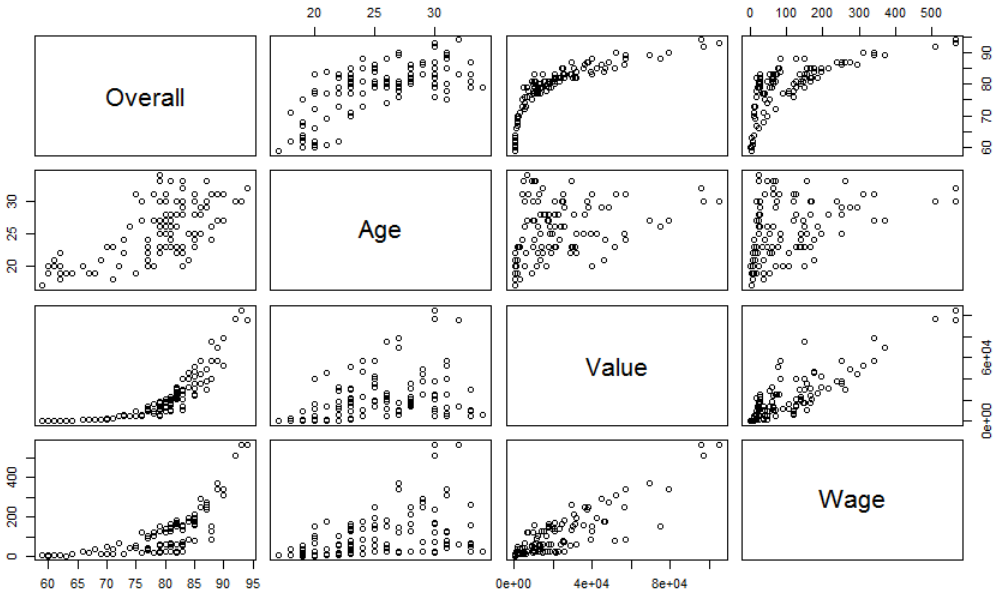
El viernes 29 de septiembre de 2017 se puso a la venta el juego FIFA 18©. A los futbolistas del juego se les asignan unas valoraciones en diferentes características que se resumen en una valoración final (Overall, escala entre 0 y 100). No se dispone del modelo que asigna esta valoración final, la cual determina la posición del jugador en el ranking. En esta edición, el jugador situado en primera posición es Cristiano Ronaldo con un score global de 94, seguido de Messi con 93.

En la página de Kaggle https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset es posible acceder al detalle de las valoraciones de los atributos de los más de 17.000 jugadores incluidos en el juego. El objetivo consiste en explicar qué atributos están relacionados con una mejor valoración global del jugador.

Para ello, se ha seleccionado la información de los jugadores de los equipos españoles que acabaron en las cuatro primeras posiciones en la Liga 2016-17 (Real Madrid, FC Barcelona, At. Madrid y Sevilla). La base de datos estadística tiene la siguiente información de 109 jugadores:

```
Overall:    "Score" global (puntuación entre 0 y 100)
Age:        Edad (años)
Club:       Club de futbol al que pertenece (RMD, FCB, ATM y SEV)
Value:      Valor de la ficha del jugador(millones de euros)
Wage:       Sueldo anual del jugador (miles de euros anuales)
Attacking:  Score de Ataque (Remates, pases cortos, voleas, tiros de Cabeza) (0-500)
Skill:      Score de Habilidad (Regates, pases largos, control de balón) (0-500)
Movement:   Score de Movimiento (Aceleración, agilidad, sprint, reacción) (0-500)
Power:      Score de Potencia (Fuerza, tiros largos, salto, potencia de tiro) (0-500)
Mentality:  Score de Mentalidad (Agresión, intercepción, posición, visión) (0-600)
Defending:  Score de Defensa (Marcaje, faltas) (0-300)
```

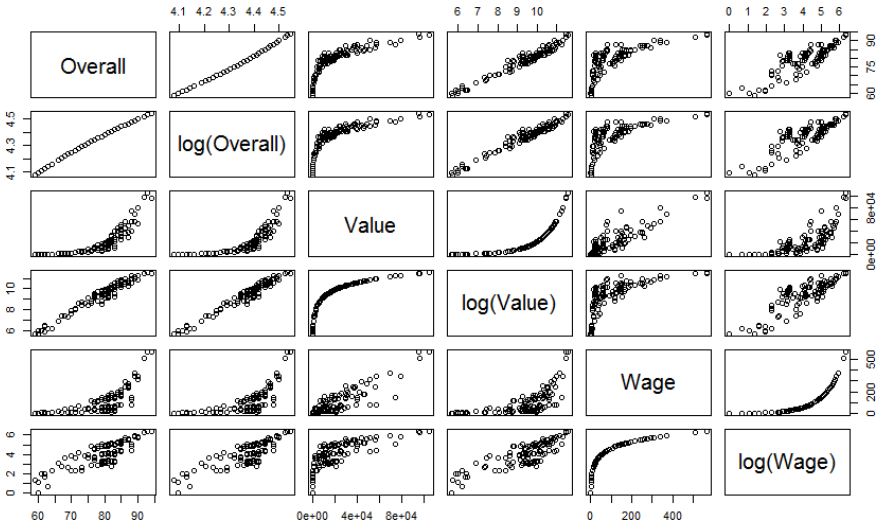
La estadística descriptiva que incluye el Overall Score, así como la edad, el valor de la ficha y el sueldo del jugador es la siguiente:



1- **De forma clara, las relaciones entre las predictoras propuestas y la respuesta no presentan comportamiento lineal. Interpreta dicha relación e indica a qué puede ser debida, teniendo en cuenta la definición de la variable respuesta. ¿Es correcto concluir que, debido a que las tres variables presentan un patrón similar de relación con la respuesta, los modelos que las incluyan presentarán multicolinealidad? Razona la respuesta**

En las tres variables predictoras propuestas (Age, Value y Wage) la relación es directa (un incremento de estas variables supone un incremento de la puntuación (Overall). Sin embargo estas relaciones no son lineales ya que el crecimiento de la respuesta se estabiliza a partir de un cierto punto en cada variable. Este tipo de relación es logístico, donde el crecimiento de la variable se atenúa para acercarse a un valor asintótico. En este caso, puesto que la variable respuesta es un índice que va de 0 a 100, es razonable encontrar este comportamiento, ya que la respuesta no puede superar el valor de 100 y se observan casos próximos a este valor, con lo que la relación puramente lineal no es razonable.
Si bien una relación similar, en la mayoría de situaciones implica una alta correlación en los predictores, es posible encontrar casos donde esto no se cumpla. La multicolinealidad es una propiedad ligada a la matriz de diseño (X) y es independiente de la variable respuesta. Por ello, no es adecuado inferir de esta configuración que hay necesariamente multicolinealidad.

Se plantea la transformación de variables para mejorar la linealidad de las relaciones. En el siguiente gráfico aparecen las variables Overall, valor de la ficha y sueldo, en las escalas originales y transformadas mediante logaritmos.



2- Justifica que transformaciones de las variables consideras adecuadas en el modelo lineal a ajustar

Respecto a las variables Value y Wage, resulta evidente que su relación con la respuesta es de tipo logístico (no lineal) y en cambio, mediante la transformación logaritmo obtenemos una relación claramente lineal, tanto si la respuesta está en la escala original como en la logarítmica. El rango de valores de la respuesta es pequeño, sin valores extremos o próximos a cero, por lo que la transformación logarítmica apenas cambia la métrica. Es por ello que la respuesta se puede incluir en el modelo con cualquiera de las dos escalas de forma indistinta.

Independientemente de la respuesta anterior, se ajusta un modelo que incluye como predictores la edad y los logaritmos del valor y sueldo del jugador, con el logaritmo del Overall Score como variable respuesta:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.017964 -0.007481 -0.001096  0.006321  0.028949

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.6426901  0.0085188 427.608   <2e-16 ***
Age         0.0049737  0.0002648  --1--    <2e-16 ***
log(Value)  0.0620323  0.0013137  47.221   <2e-16 ***
log(Wage)     --2--    0.0014242   2.207   0.0295 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01023 on 105 degrees of freedom
Multiple R-squared:    --3-- ,   Adjusted R-squared:  --4--
F-statistic:  --5-- on  --6-- and --7-- DF,  p-value: < 2.2e-16
```

3- Sabiendo que la desviación estándar de la variable “logaritmo de Overall” vale 0.1060, calcula los valores no incluidos en el listado anterior, indicando el proceso de cálculo.

1- T-value de $\hat{\beta}_{age}$: $\frac{0.0049737}{0.002648} = 18.78285$

2- Coeficiente $\hat{\beta}_{logWage} = t - value * s_{\hat{\beta}} = 2.207 * 0.0014242 = 0.003143209$

3- Coeficiente de Determinación R^2 : $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{105*0.01023^2}{108*0.106^2} = 0.9909447$

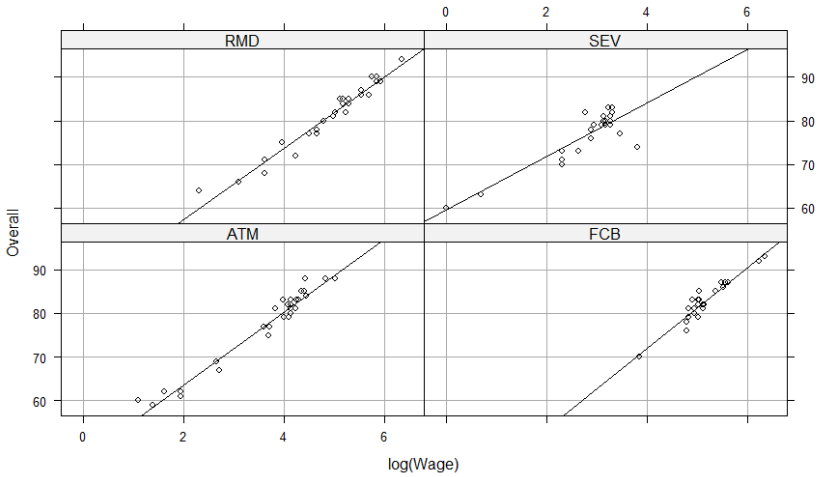
4- Coeficiente de Determinación Ajustado: $R^2_{adj} = 1 - \frac{S_R^2}{S_{LBM}^2} = 1 - \frac{0.01023^2}{0.106^2} = 0.9906859$

5- Estadístico F=$\frac{(TSS-RSS)/(p-1)}{RSS/(N-p)} = \frac{(108*0.106^2-105*0.01023^2)/3}{0.01023^2} = \frac{0.4008331}{0.01023^2} = 3830.12$

6- Grados de libertad del numerador: p-1 = 4 = 3

7- Grados de libertad del denominador: N-p = 109-4 = 105

Se analiza la relación entre el Overall Score y el sueldo del jugador. En el siguiente gráfico se representan los gráficos que relacionan el logaritmo del sueldo con la puntuación final, segmentado por club e incluyendo la recta de ajuste de mínimos cuadrados.



El modelo ajustado, con contraste de tipo “suma cero”, tiene el siguiente resultado:

```
    [,1] [,2] [,3]
ATM    1    0    0
FCB    0    1    0
RMD    0    0    1
SEV   -1   -1   -1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.7123     1.3030  35.082  < 2e-16 ***
log(Wage)         7.9804     0.2802  28.483  < 2e-16 ***
Club1             1.1935     1.5413   0.774  0.44053
Club2           -10.6854     3.3518  -3.188  0.00191 **
Club3            -4.3745     1.9964  -2.191  0.03074 *
log(Wage):Club1   0.4046     0.3603   1.123  0.26415
log(Wage):Club2   1.2869     0.6617   1.945  0.05458 .
log(Wage):Club3   0.1508     0.4137   0.365  0.71623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.041 on 101 degrees of freedom
Multiple R-squared:  0.9375,     Adjusted R-squared:  0.9332
F-statistic: 216.5 on 7 and 101 DF,  p-value: < 2.2e-16
```

Se añaden los siguientes test de tipo lineal sobre los coeficientes:

```
Linear hypothesis test

Hypothesis:
- Club1 - Club2 - Club3 = 0

Model 1: restricted model
Model 2: Overall ~ log(Wage) * Club

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    102 708.74
2    101 420.57  1    288.17 69.203 4.391e-13 ***
```

```
Linear hypothesis test

Hypothesis:
- log(Wage):Club1 - log(Wage):Club2 - log(Wage):Club3 = 0

Model 1: restricted model
Model 2: Overall ~ log(Wage) * Club

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    102 490.11
2    101 420.57  1    69.531 16.698 8.802e-05 ***
```

4- Para este modelo escribe los modelos estimados que relacionan la puntuación global con el logaritmo del sueldo según el club al que pertenece. Indica si hay diferencias significativas de esta relación comparando el promedio global con cada uno de los equipos. ¿De qué tipo son estas diferencias?

Como el contraste active es de tipo suma cero, el intercept y la pendiente se interpretan como el modelo promedio de todas las observaciones. Las dummies y las interacacciones con éstas corresponden al cambio de pasar del modelo promedio a cada categoría.

Modelo para Club=ATM

$$Overall = (45.7123 + 1.1935) + (7.9804 + 0.4046) * \text{logWage} + \text{e}$$

Comparándolo con el modelo promedio ($Overall = 45.7123 + 7.9804 * \text{logWage} + \text{e}$), tanto el p-valor del parámetro `Club1` como el de la interacción `log(Wage):Club1` tienen p-valores superiores a 0.05, indicando que no se detectan diferencias significativas entre el modelo para el Atlético y el promedio, ni en nivel ni en pendiente.

Modelo para Club=BCN

$$Overall = (45.7123 - 10.6854) + (7.9804 + 1.2869) * \text{logWage} + \text{e}$$

Comparándolo con el modelo promedio, el p-valor del parámetro `Club2` es significativo, y siendo estricto en el límite del 0.05 para la significación, el de la interacción `log(Wage):Club2` tienen p-valor ligeramente superior a 0.05. En este caso la recta difiere en el nivel (decremento significativo de más de 10 unidades) pero la pendiente se puede considerar la misma. Por ello se puede interpretar que la recta para el Barcelona es paralela pero con una ordenada en el origen inferior a la recta promedio.

Modelo para Club=RMD

$$Overall = (45.7123 - 4.3745) + (7.9804 + 0.1508) * \text{logWage} + \text{e}$$

Comparándolo con el modelo promedio, el p-valor del parámetro `Club3` es significativo el de la interacción `log(Wage):Club2` no lo es. La interpretación es la misma que la del club anterior, aunque la diferencia de nivel es menor (-4 unidades)

Los coeficientes para la última categoría (Club=SEV) se calculan sumando el resto de coeficientes y cambiando el signo al resultado.

Modelo para Club=SEV

$$Overall = (45.7123 - 1.1935 + 10.6854 + 4.3745) + (7.9804 - 0.4046 - 1.2869 - 0.1508) * \text{logWage} + \text{e}$$

Para compararlo con el modelo promedio, la tabla summary del modelo no nos indica el p-valor de significación de las variaciones en el intercept y la pendiente. Por ello, se añaden los dos tests lineales sobre los coeficientes. El primero hace referencia a la significación del coeficiente para el intercept en el caso del Sevilla:

$$H_0: \beta_{0,SEV} = 0 \iff H_0: -\beta_{0,ATM} - \beta_{0,BCN} - \beta_{0,RMD} = 0 \iff H_0: -Club_1 - Club_2 - Club_3 = 0$$

El p-valor de 4.391e-13 indica que hay diferencia significativa en el intercept del modelo del Sevilla respecto al del modelo promedio. De la misma manera se procede con la pendiente:

$$H_0: \beta_{1,SEV} = 0 \iff H_0: -\beta_{1,ATM} - \beta_{1,BCN} - \beta_{1,RMD} = 0 \iff H_0: -\log(Wage):Club_1 - \log(Wage):Club_2 - \log(Wage):Club_3 = 0$$

El p-valor muy inferior al 5% indica que las rectas no son paralelas.

Se realiza el siguiente contraste simultáneo sobre los parámetros del modelo:

```
Linear hypothesis test
```

```
Hypothesis:
Club2 - Club3 = 0
log(Wage):Club2 - log(Wage):Club3 = 0

Model 1: restricted model
Model 2: Overall ~ log(Wage) * Club

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    103 430.15
2    101 420.57  2    9.5785 1.1501 0.3207
```

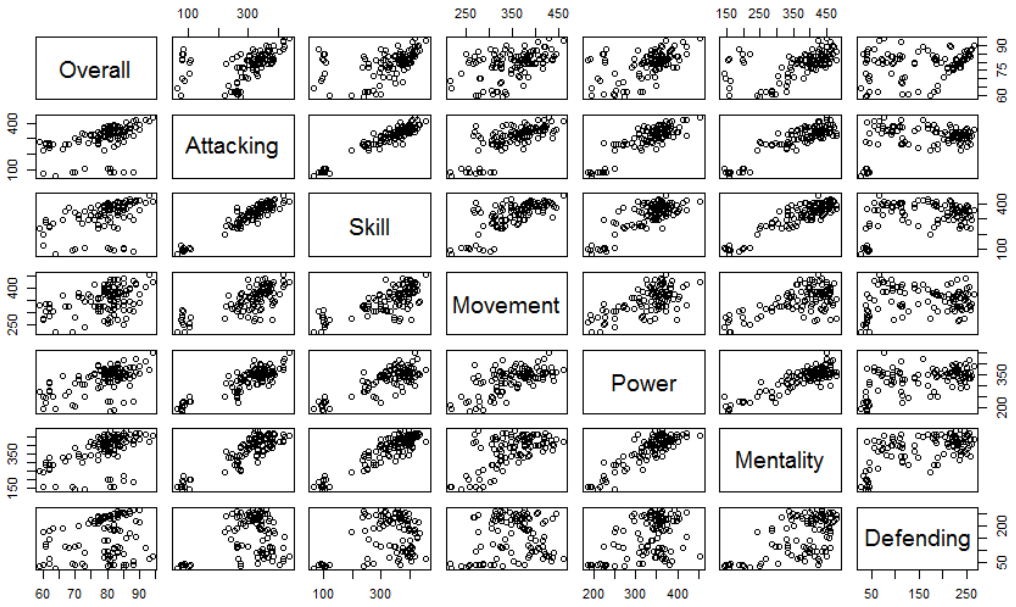
5- ¿Qué conclusión práctica se obtiene de dicho test?

El test que se está aplicando compara de forma simultánea los siguientes parámetros:

$$H_0: \begin{cases} Club_2 = Club_3 \\ -\log(wage) : Club_2 = -\log(wage) : Club_3 \end{cases}$$

Es decir, las variaciones en el intercept y en la pendiente respecto al modelo promedio en el Barcelona y en el Real Madrid son del mismo orden y no hay diferencias significativas entre ambos clubes. Desde el punto de vista práctico, corresponde a que no hay diferencias estadísticamente significativas entre ambas rectas, por lo que no se detectan diferencias en la relación entre el logaritmo del sueldo y la puntuación overall en ambos clubes.

A continuación, se explora la relación entre los diferentes scores calculados y la puntuación global (sin tomar logaritmos). La descriptiva gráfica es la siguiente:



El modelo estimado con estas variables, ajustando por edad y logaritmo del sueldo es el siguiente:

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.5855 -2.3667 -0.2388  2.3959 11.0273

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.485741   3.590690  11.554  < 2e-16 ***
Age          0.510580   0.102689   4.972 2.75e-06 ***
logWage      3.843699   0.351978  10.920  < 2e-16 ***
Attacking   -0.023240   0.014787  -1.572   0.1192
Skill        0.001262   0.015346   0.082   0.9346
Movement     0.003132   0.010403   0.301   0.7640
Power        0.028980   0.014705   1.971   0.0515 .
Mentality    0.017552   0.017537   1.001   0.3193
Defending   -0.010711   0.007031  -1.523   0.1308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.481 on 100 degrees of freedom
Multiple R-squared:  0.8199,     Adjusted R-squared:  0.8055
```

```
F-statistic: 56.92 on 8 and 100 DF,  p-value: < 2.2e-16
```

6- En base al listado anterior, ¿se puede concluir que ninguno de los scores parciales está relacionado de forma significativa con el cálculo del Score "Overall" si ajustamos por edad y logaritmo del sueldo?

Aunque los p-valores de los scores parciales son todos superiores al 5%, indicando que no son significativos, hay que tener en ecuenta que si se procede a la simplificación del modelo, eliminando parámetros no significativos, se debe realizar de forma secuencial uno a uno. La eliminación de un coeficiente debe dar lugar a una re-estimación del modelo, donde los nuevos p-valores pueden poner de manifiesto alguna relación significativa. Por ello, no es conveniente eliminar más de un parámetro de forma simultánea, ni concluir que ninguno de ellos está relacionado de forma significativa con la respuesta. Es especialmente crítico en presencia de multicolinealidad. La correlación entre los predictores hace que la eliminación de uno de ellos del modelo suponga importantes cambios en la significación del otro. Por otro lado, si estuvieramos ante un diseño experimental configurado de forma ortogonal (todos los predictores incorrelados y por lo tanto, todos los VIFs igual a 1) se garantizaría que la eliminación de un coeficiente no afecta al resto. En este caso sí que se podría concluir que ningún parámetro es significativo.

Los VIFs obtenidos del anterior modelo son los siguientes:

```
      Age    logWage Attacking     Skill  Movement     Power Mentality Defending
 1.751353  1.866002 15.363888 19.414518  2.702504  6.081867 21.768463  2.677711
```

7- ¿Qué representan los estadísticos VIF que aparecen en el listado? (no hace falta dar la expresión matemática) ¿Por qué estos valores indican que no sería adecuado interpretar directamente cada uno de los coeficientes del modelo, en este caso?

Los VIFs (Variance Inflation Factor) o factores de inflamiento de la varianza son estadísticos que permiten diagnosticar la presencia de multicolinealidad en el modelo. Este fenómeno implica que las variables predictoras puedan presentar correlación alta que impida una interpretación simple de cada coeficiente del modelo, manteniendo el resto de predictores fijo (ceteris paribus). Así mismo, la estimación en presencia de multicolinealidad deja de ser eficiente. Para cada coeficiente, el VIF calculado es una medida de cuan correlacionada está ese predictor con el resto de predictores (Si R2 es el coeficiente de determinación en tanto por 1 del modelo que se ajusta con cada variable como si fuera la respuesta en base al resto de predictores, entonces $VIF=1/(1-R^2)$). En el caso en que los predictores sean independientes (y no haya multicolinealidad) el valor de los VIFs será 1. Cuanto más se aleje de 1 indica una mayor presencia de multicolinealidad. Un VIF superior aproximadamente a 8 indica que el problema de multicolinealidad puede ser importante.
En este caso, las variables Attacking, Skill y Mentality poseen unos VIFs altos que indican que el modelo presenta multicolinealidad. Si interpretamos el primer coeficiente como el cambio producido en el valor esperado de la respuesta al aumentar en una unidad el score de Ataque, no tendríamos en cuenta que a la vez los otros scores pueden aumentar y se debería también tener en cuenta sus coeficientes y el grado de correlación con esa variable. No sería realista hacer una interpretación donde el resto de coeficientes se mantuviera constante (ceteris paribus)

A continuación, aplicamos el mecanismo de selección de variables "stepwise" con la siguiente sintaxis:

```
> mod2<-step(mod, direction="both", k=log(nrow(dades))
```

donde mod corresponde al modelo completo anterior.

8- Explica brevemente en qué consiste este procedimiento, indicando los pasos de forma esquemática y explicitando el criterio que se utiliza en este caso

El método "stepwise" permite seleccionar las variables a introducir en el modelo. Se especifica un modelo de partida (puede ser el modelo nulo o el completo con todas las variables propuestas). En cada etapa se puede plantear tres opciones:

-Eliminar alguna de las variables del modelo: -var
-Incluir alguna variable descartada, que no está en el modelo actual: +var
-No realizar ningún cambio: <none>

En cada etapa lista las diferentes posibilidades en orden creciente del criterio de información especificado:

K=2 (por defecto) ➔ AIC=RSS+2*#param
K=log(N) (utilizado aquí) ➔ BIC=RSS+log(N)*#param

La opción listada en primera instancia es la alternativa adecuada para mejorar el modelo. En la siguiente etapa se sigue con el modelo modificado según esa opción, y se vuelven a analizar las opciones hasta el momento en que la mejor sea no modificar el modelo (<none>).

Si se parte del modelo nulo y solo se permite añadir variables, el método se denomina "forward". Si se parte del modelo completo y sólo se permite eliminar variables, el método se llama "backward". Indicando el parámetro direction="both" se contempla tanto la eliminación como la inclusión de variables, y el método se denomina "stepwise".
Si bien, no se garantiza el mejor modelo, si que da buenos modelos y permite explorar las relaciones entre los predictores en presencia de multicolinealidad.

El modelo obtenido, los VIFs asociados y los plots de validación son los siguientes:

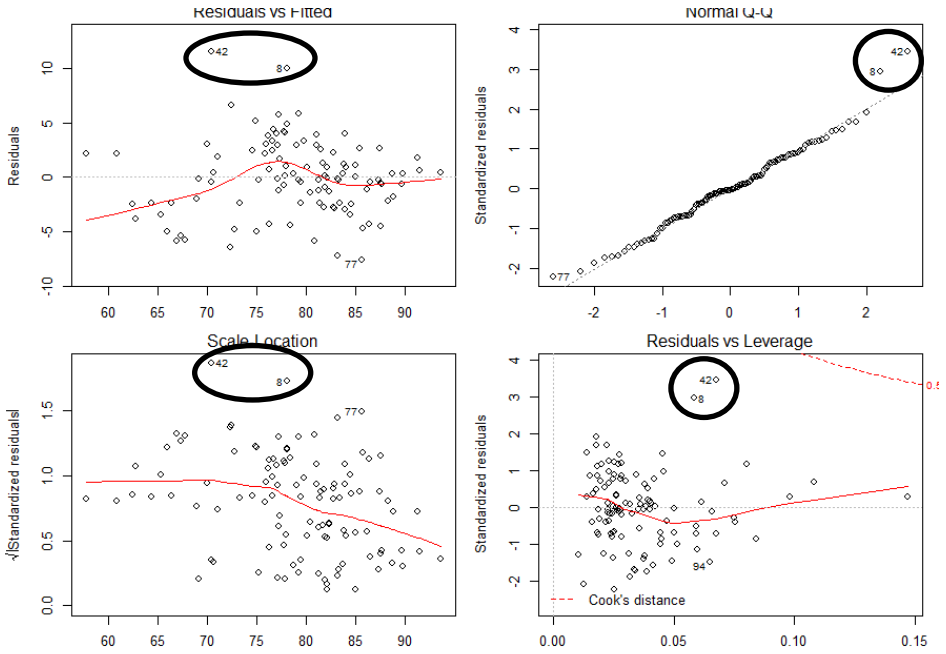
```
Residuals:
    Min      1Q  Median      3Q     Max
-7.5968 -2.3588 -0.1384  2.2825 11.5709

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.735661   2.464235  17.748  < 2e-16 ***
Age          0.554248   0.088358   6.273 8.05e-09 ***
logWage      3.978937   0.318586  12.489  < 2e-16 ***
Power        0.015176   0.006965   2.179   0.0316 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.469 on 105 degrees of freedom
Multiple R-squared:  0.8122,     Adjusted R-squared:  0.8069
F-statistic: 151.4 on 3 and 105 DF,  p-value: < 2.2e-16

VIFS:

     Age  logWage    Power
1.305651 1.539393 1.374052
```



9- Realiza la validación del modelo, indicando en cada gráfico las premisas que se analizan. Caracteriza en base a si son datos atípicos y/o influyentes las observaciones 8 (Oblak, portero del ATM) y 42 (Sergio Rico, portero del Sevilla) que aparecen señaladas en los gráfico.

El primer plot es el de los residuos frente las predicciones, permite ver si la disposición de los residuos es aleatoria alrededor del cero, sin que se observe claramente ningún patrón que indique desviaciones de la relación lineal. El ajuste local (línea roja) es prácticamente horizontal, confirmando en este caso no parece haber patrones de no linealidad (se podría comentar una cierta curvatura en el centro). En este plot también se puede verificar descriptivamente si la varianza puede considerarse constante, frente a las predicciones. En este caso, no se observa incremento de la variabilidad de los residuos a medida que

aumenta la predicción, indicando que se puede asumir homocedasticidad. También en este plot, aparecen etiquetadas las observaciones con residuos estandarizados superior a 2 (aprox) en valor absoluto (valores atípicos).

El segundo plot es el plot de normalidad, que permite determinar si podemos considerar que la distribución Normal es adecuada para los residuos. Si los puntos están alineados podemos asumir Normalidad de los residuos. Este plot permitiría ver patrones de asimetría o colas pesadas en los residuos que irían en contra de la hipótesis de normalidad. También se etiquetan los atípicos. En este caso, la disposición de los puntos está claramente alineada lo que permite asumir normalidad en los residuos.

El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos frente a las predicciones. Es un plot que permite determinar de forma más clara la presencia de heteroscedasticidad. El ajuste local mediante la recta no indica un claro descenso de los valores que constituyen una estimación de la varianza de los residuos. No es concluyente para confirmar la presencia de varianza no constante y además se puede ver influido por la poca presencia de observaciones que estén relacionados con valores altos de las predicciones, lo cual puede suponer una peor estimación de la variabilidad. El descenso de la varianza para predicciones altas (próximas a 100) es habitual cuando la variable respuesta está truncada en ese valor.

El cuarto gráfico permite identificar y caracterizar los datos influyentes. Representa los residuos estandarizados frente al factor de anclaje/apalancamiento (leverage). Además incluye curvas de nivel para indicar la distancia de Cook de las observaciones. Valores con una distancia de Cook alta pueden ser valores influyentes y se debe analizar su efecto en el ajuste del modelo. La distancia de Cook es una función creciente de los residuos al cuadrado y del leverage. Las observaciones que tienen un valor alto de la distancia de Cook aparecen etiquetadas (pueden ser por tener muy leverage, o tener un residuos alto en valor absoluto o una combinación de ambas situaciones no tan extremas). Las observaciones etiquetadas como influyentes parece que tienen un leverage alto ya la vez tienen un residuo de magnitud elevada. Habría que analizar qué efecto tienen en la estimación del modelo.

Las observaciones 8 y 42 corresponden a valores atípicos (error estandarizado superior a 3) con un factor de apalancamiento moderado. Se interpreta como que el modelo no explica adecuadamente estas observaciones. El hecho de que los residuos sean positivos indican que el modelo predice una valoración muy inferior a la que poseen. Es decir, hay información no recogida en el modelo que justifica la alta puntuación de estos jugadores. Hay que tener en cuenta que la mayoría de scores parciales recogidos están relacionados con habilidades de jugadores de campo (Ataque, Movimiento, Potencia). Es lógico pensar que para los porteros, las puntuaciones de estos scores sea baja (en realidad en la base de datos original se incluye un score parcial relacionado con la habilidad de los porteros, pero este score no se ha incluido en el estudio). No parecen las líneas de nivel de la distancia de Cook en el último gráfico, por lo que posiblemente no se trate de datos influyentes.

Las entradas en la base de datos de Cristiano Ronaldo y Messi son las siguientes:

	Overall	Age	logWage	Attacking	Skill	Movement	Power	Mentality	Defending
C.Ronaldo	94	32	6.336826	438	418	428	453	452	76
L.Messi	93	30	6.336826	416	458	459	373	423	67

10- De acuerdo al último modelo, calcula sus predicciones puntuales e interpreta el resultado indicando qué factores determinan su posición final (teniendo en cuenta que sólo se utiliza esta información).

Usando el último modelo:

Para C. Ronaldo:

$$Overall = 43.7356 + 0.5542 * 32 + 3.9789 * 6.3368 + 0.015 * 452 = 93.48$$

Para Messi:

$$Overall = 43.7356 + 0.5542 * 30 + 3.9789 * 6.3368 + 0.015 * 423 = 91.92$$

Las diferencias se deben a la edad (2 años supone un incremento de una unidad) y a la potencia, donde Ronaldo aparece mejor puntuado.