Instituto Tecnológico Autónomo de México

Implementación de un Modelo de Regresión Lineal para predecir la temporada 19-20 de la Premier League

Trabajo Final

estadística aplicada 2

PABLO LÓPEZ LANDEROS 178863

FERNANDO STEIN VALLARTA 165455

FRANCISCO GABRIEL HUERTA FERNANDEZ 166040

Profesor: LUIS ENRIQUE NIETO BARAJAS

México, D.F. 2020

1. Introducción

La motivación de este trabajo nace del interés compartido por los tres estudiantes en el campo de los *sports analytics*, particularmente en el futbol. Comencemos entonces por contestar la pregunta ¿Qué son los sports analytics? Podríamos definir el concepto como una nueva forma de estudiar el deporte. Consiste en hacer un análisis estadístico avanzado de los datos de juego para identificar áreas de oportunidad y obtener una ventaja competitiva. El objetivo de los sports analytics es usar datos para reducir el desempeño de el/la atleta a un número. Si bien este campo es joven, presenta todo un nuevo panorama dentro del mundo deportivo. Por ejemplo, en el año 2015, un artículo publicado en el portal SportTechie revela que utilizando datos de juego y combinándolo con variables como el patrón de sueño, niveles de estrés, cansancio, dieta e intensidad de entrenamiento, podemos predecir cuándo un jugador es más propenso a lesionarse.

A partir del año 2003, este campo ha crecido a un ritmo muy acelerado y este crecimiento no parece que vaya a disminuir. En 2015, este mercado tenía un valor estimado de 125 millones de dólares. Para 2021 se espera que valga 4.7 billones de dólares. Existen ya casos de éxito como los Oakland Athletics(MLB), los Houston Rockets (NBA) o el Liverpool F.C. (Premier League). Estos equipos han logrado mejorar su despempeño gracias al uso de la Ciencia de Datos. Particularmente, interesaba el estudio de la Premier League por ser una liga con equipos competitivos pero sin un par de equipos dominantes como sería La Liga española.

Cada verano, los equipos de fútbol invierten cientos de millones de dólares compitiendo por los mejores jugadores para las próximas temporadas. Sin embargo, el valor de los futbolistas y de las transferencias se ha incrementado en los últimos años. Desde 2013, diez transferencias realizadas superaron los 100 millones de euros. La más cara fue el traspaso por 222 millones de euros del astro brasileño Neymar desde Barcelona al Paris Saint-Germain. El precio de los jugadores entre 2017 y 2018 aumentó en un 31 por ciento, y desde 2014 la tasa de crecimiento de la inflación en el mercado de transferencias de las 5 grandes ligas (LaLiga, English Premier League, Bundesliga, Serie A y Ligue 1) ha sido del 26 por ciento. (CIES, 2019)

En este trabajo pretendemos implementar varios modelos de regresión lineal tomando diversas variables explicativas para intentar contestar la pregunta: **¿cuál es la mejor variable a tomar en cuenta si queremos predecir que equipo ganará la liga?** Tomamos en cuenta aquellas que a nuestro criterio más influyen. Una cosa que vale la pena notar es que las valuaciones de los jugadores no tienen que ver únicamente con su desempeño en el juego, muchas veces factores como la popularidad del jugador, la edad del jugador, los equipos en los que ha jugado y la nacionalidad del jugador influyen sobre la misma. Por ejemplo, algunos jugadores que tienen menos de 21 años y ya han jugado algunos partidos de fútbol suelen tener valoraciones más altas que jugadores de 30 años que son muy efectivos en el campo. De ahí que existan historias de jugadores que tienen un valor de merado muy elevado, pero que no suelen ser los mejores jugadores del equipo.

2. Descripción de la información

Explicación de variables

Primero, debemos de establecer que las siguientes métricas fueron tomadas como el promedio de los 15 jugadores con más minutos durante la temporada. La razón de hacerlo de esta forma en lugar de tomar a todos los jugadores es porque algunos de los jugadores registrados en la plantilla del equipo pueden estar lesionados o no participar en la temporada. Esto sería menos representativo ya que pudiera haber jugadores con pocos minutos que afectarían en el cálculo de valores para el equipo pero juegan poco. Las métricas de eficiencia son las siguientes:

- AvgValue: Valor de mercado medido en Euros para cada jugador. Los valores que utiliza la página Transfer Markt se calculan a grosso modo de la siguiente manera: $contrato = duracion + salarios$. Al comprar un jugador, escencialmente lo que un equipo hace es pagar el resto de sus salarios.

- AvgFifa: Promedio de los ratings asignados en el videojuego FIFA. Cada año EA studios saca un videojuego en el que los jugadores tienen un rating dependiendo de que tan bien o mal estén jugando. Este rating va desde el 0 y hasta el 100.

- AvgRating: La calificación asignada por la página Whoscored. Esta medida es de las más aceptadas en el negocio y reconocida por expertos. Consta de una califiación del 0 al 10 que empieza cada partido en 6. Cada jugada que pasa se va actualizando en vivo y al final tomamos aquel valor con el que el jugador termina el partido. Para cada año tomamos el promedio de los 38 partidos por cada jugador.

- Avg90 : La diferencia entre goles anotados menos goles recibidos mientras el jugador está en cancha cada 90 minutos

- AvgGoal: El cociente de goles anotados dividido entre los goles recibidos mientras el jugador está en cancha.

También tomamos otras variables que no dependían de los jugadores sino que se calculan por equipo:

- Points: Número de puntos sumados por temporada.

- Wins: Número de victorias por temporada.

- Draws: Número de empates por temporada.

- Losses: Número de derrotas por temporada.

- PPG: Una métrica calculada para cada equipo. Son los puntos que se espera que sume por juego.

- LogPoints: Logaritmo natural de la variable *Points*. Esta es la variable sobre la cual se hará la regresión. Utilizamos logaritmo para mejorar los supuestos de normalidad y tener un soporte real en la variable dependiente.

EDA

Crear la base de datos fue la parte más desafiante del trabajo. Lo que realmente complicó la extracción de datos fue el hecho de que decidimos utilizar los 15 jugadores con más minutos jugados por temporada. El proceso se puede resumir así:

1. Tuvimos que extraer para cada año y cada equipo, cuáles eran los 15 jugadores con más minutos en cada temporada.

2. Una vez que logramos obtener estos datos, hubo que extraer individualmente para cada jugador su Valor de mercado(TransferMarkt),rating del Fifa (Kaggle) y el Whoscored rating. Los demás datos se obruvieron de la página Football Reference. Obtuvimos esta información para todos los jugadores de los equipos.

3. Por último en Python y Stata extrajimos los datos necesarios para los 15 jugadores con más minutos.

Para darnos una perspectiva global de de la base de datos construida, creamos un script en Python que itera sobre el dataset utilizado. El output nos muestra la siguiente tabla:

	metric	Year	AvgValue	AvgFifa	AvgRating	Avg90	AvgGoal	Points	Wins	Draws	Losses	PPG	LogPoints
0	Valor máximo	18.000000	64.000000	84.333336	7.217333	2.022667	4.430206	100.000000	32.000000	17.000000	32.000000	2.631579	4.605170
1	Valor Mínimo	14.000000	2.600000	69.533333	6.440667	-1.396000	0.302640	16.000000	3.000000	2.000000	1.000000	0.421053	2.772589
2	Media	16.000000	14.772738	77.277333	6.854167	0.015587	1.192279	52.460000	14.440000	9.080000	14.480000	1.380526	3.900249
3	Desv.Estandar	1.421338	12.559789	3.435037	0.157568	0.718022	0.784844	18.118811	6.328260	2.897857	6.024328	0.476811	0.353833
4	Q1	15.000000	5.733333	75.000000	6.736000	-0.502333	0.697949	40.000000	10.000000	7.000000	10.000000	1.052632	3.688879
5	Mediana	16.000000	10.100000	76.699997	6.819000	-0.176333	0.885215	47.000000	12.000000	9.000000	16.000000	1.236842	3.850148
6	Q3	17.000000	21.983332	79.383335	6.974500	0.568667	1.522700	66.000000	19.000000	11.000000	19.000000	1.736842	4.189655
7	Kurtosis	-1.304934	3.178537	-0.613972	-0.582471	0.031695	4.783145	-0.072237	0.217014	0.131378	0.019513	-0.072237	0.580902
8	Oblicuidad	0.000000	1.721732	0.405516	0.279824	0.652104	2.013946	0.639549	0.784959	-0.015069	-0.065383	0.639549	-0.335506
9	# de valores unicos	5.000000	98.000000	78.000000	97.000000	100.000000	100.000000	52.000000	25.000000	15.000000	26.000000	52.000000	52.000000
10	% de datos faltantes	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
11	top1_repetidos	18.000000	2.600000	74.400002	6.810667	2.022667	4.430206	44.000000	12.000000	9.000000	16.000000	1.157895	3.784190
12	top2_repetidos	17.000000	9.133333	75.266670	6.986000	-0.317333	0.794585	45.000000	11.000000	7.000000	19.000000	1.184211	3.806662
13	top3_repetidos	16.000000	7.316667	76.266670	6.822000	-0.490000	0.708140	47.000000	9.000000	11.000000	15.000000	1.236842	3.850148

La tabla anterior es un resumen del dataset que utilizamos para las regresiones. Este tenía datos para cada equipo de la Premier League y para todas las temporadas desde el año 2014 y hasta los últimos partidos del 2020 (que fue suspendida por la contingencia sanitaria.).

Para darnos una idea de los valores de cada equipo durante las 6 temporadas, hicimos dos cosas:

- Promediamos los valores de cada variable por equipo e hicimos un diagrama de dispersión.

- Hicimos un diagrama de caja para cada variable por equipo durante las 6 temporadas (2014-2020). Estos diagramas los incluimos en el **Anexo**

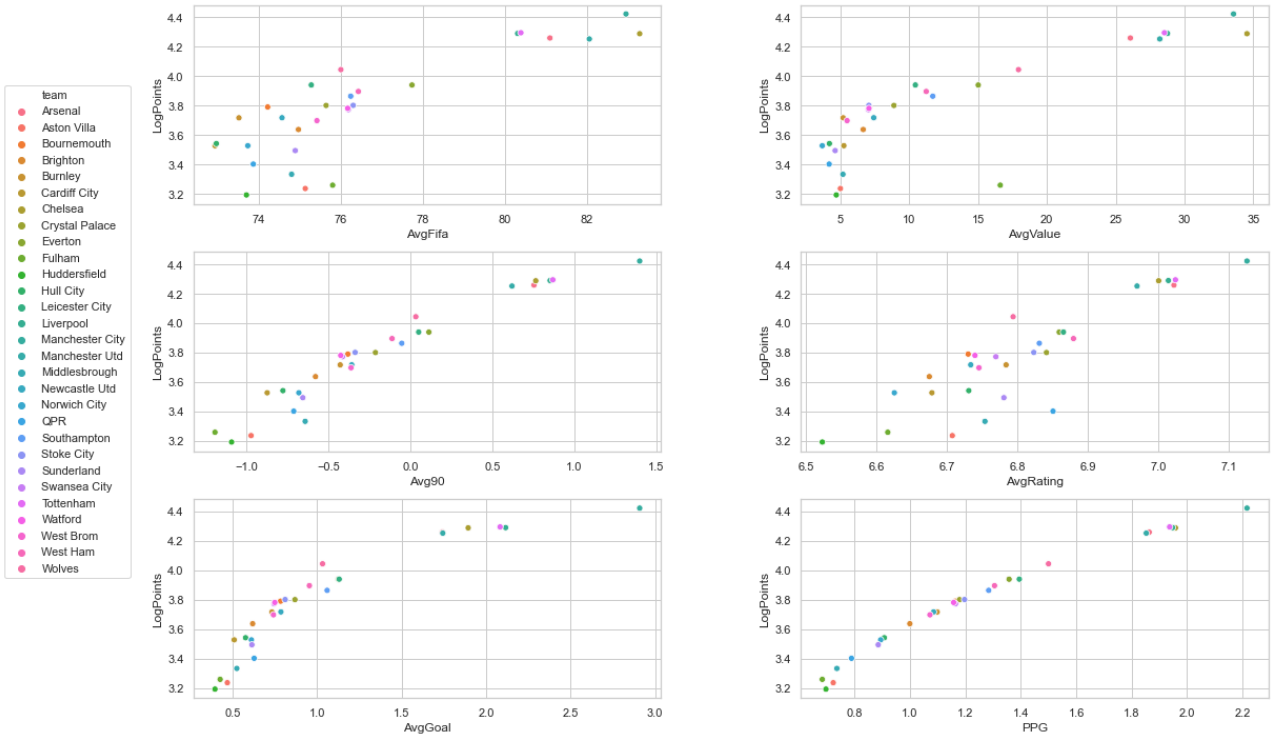


Figura 1: Diagrama de dispersión para cada una de las variables explicativas utilizadas en el modelo y la variable dependiente *LogPoints*

3. Modelos

Para el análisis de la Premier League optamos por realizar el siguiente análisis: buscamos analizar factores de los equipos que los llevan a obtener una mayor cantidad de puntos, es decir buscamos responder a la siguiente pregunta: ¿qué variables nos podrían ayudar a predecir el orden de la tabla de la Premier League?

¿Qué equipo ganará la Premier League?

Para responder a la anterior pregunta, planteamos múltiples modelos para observar que variables resultan mejores para poder predecir al ganador y la clasificación de los equipos de la Premier League. En el fútbol, a pesar de que han habido muchos intentos, no existe una medida de eficiencia clara de los jugadores como la existe en el basquetbol con la métrica de "Player Efficiency Rating"(PER) o en el beisbol con la métrica de .ºn-Base Percentage"(OBP). Esto se debe a que las posiciones de fútbol tienen objetivos fundamentalmente diferentes, un defensa tiene un rol diferente al rol que un delantero tiene dentro de la cancha. Además, no es clara la forma en la que el juego defensivo, el número de pases, los tiros o la presión ejercida se puedan traducir en mejores resultados.

3.1 Valor de mercado vs. Puntos

El primer modelo de estudio es de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

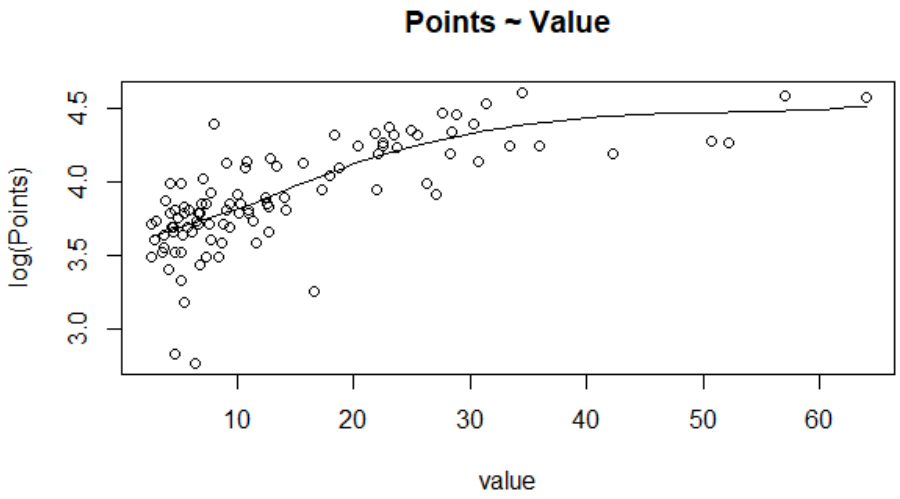
Donde Y_i es la variable dependiente que hace referencia al número total de puntos obtenidos durante una temporada, mientras que X_i es el valor de mercado promedio del equipo.

Notemos que las variable dependiente son los puntos, de donde nuestra varaiable dependiente tiene un soporte discreto. De ahí que proponemos una transformación del modelo anterior al siguiente:

$$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

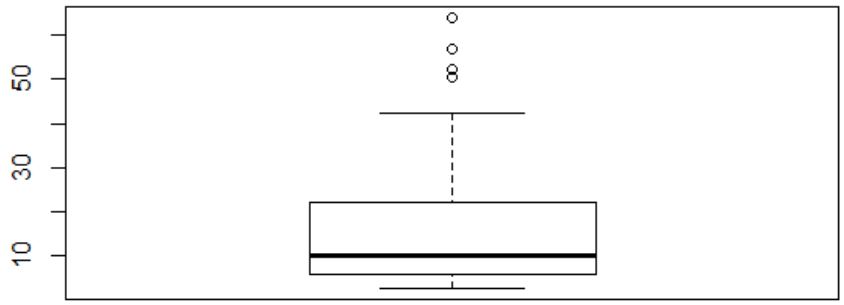
Antes de continuar con el modelo anterior, utilizamos un scatter plot para poder observar el supuesto de linealidad. El scatter plot nos queda de la siguiente forma:

3.1.1 Scatter Plot Valor de Mercados, log(Puntos)



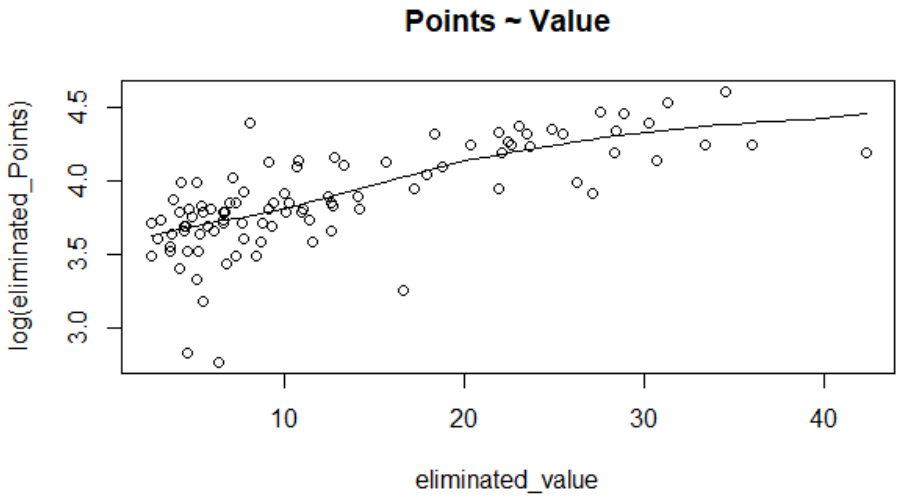
De acuerdo al anterior scatterplot no parece que un modelo de regresión lineal pudiese ajustar los datos. Interpretamos que tal vez existen datos que pueden ser considerados outliers que inhiben nuestro estudio. Luego entonces, realizamos un diagrama de caja utilizando el Rango Intercuartilico. El Rango Intercuartilico es el 50 % central o el area entre los percentiles 75 % y 25 % de una distribución. Dicho rango nos permite determinar si un punto es un *outlier*. Si el valor está por arriba del percentil 75 o debajo del percintil 25 por un factor de 1.5 veces el RIC, este valor se considera entonces un *outlier*.

3.1.2 Box Plot Valor de Mercados



Notemos que tenemos 4 datos que pueden ser considerados como outliers. De ahí que decidimos eliminarlos para poder realizar un mejor estudio. Al eliminar los datos anteriores tenemos el siguiente scatter plot:

3.1.3 Scatter Plot Valor de Mercados, log(Puntos) sin Outliers



La gráfica anterior exhibe una relación lineal más clara que la que exhibía el primer scatter plot. De ahí que procedemos a plantear un modelo de regresión lineal planteado en la ecuación 2.

Notemos que en este caso buscamos refutar la hipótesis nula de que nuestros parametros $\beta_0 = \beta_1 = 0$. Al

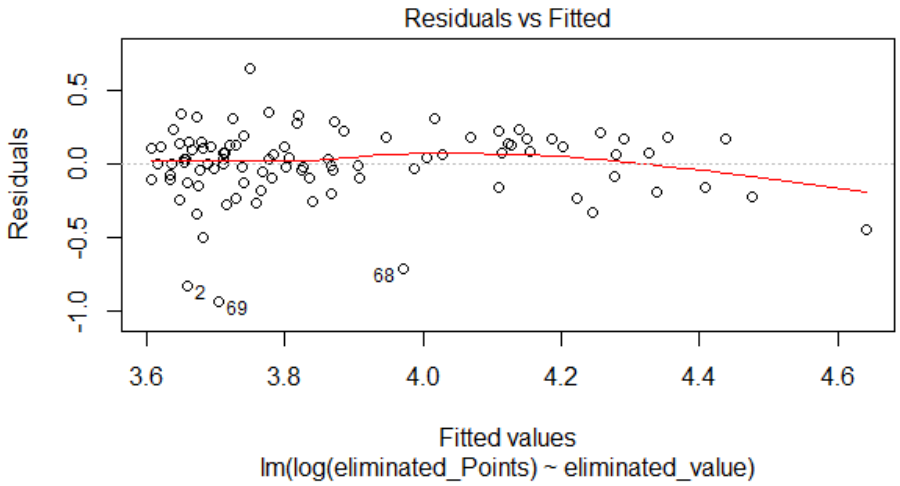
simular el modelo anterior, obtenemos los siguientes resultados:

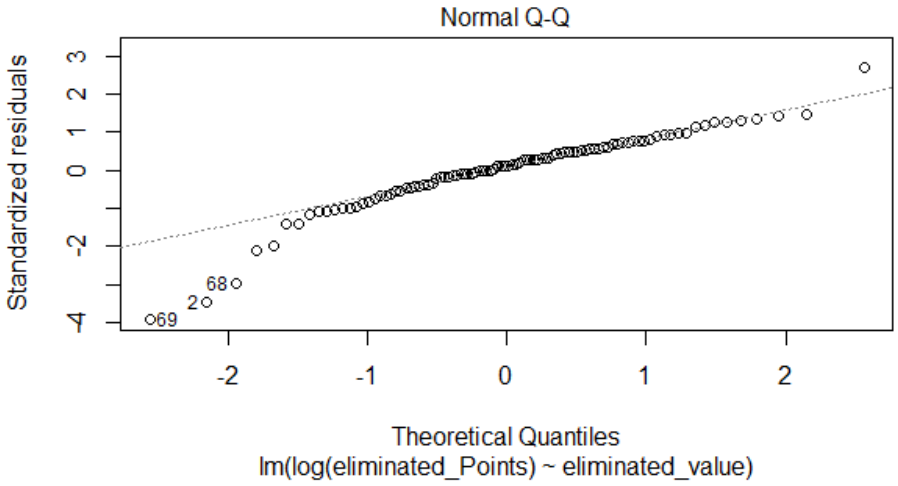
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.538332   0.042045  84.156  < 2e-16 ***
eliminated_value 0.026046   0.002616   9.957 2.26e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.2403 on 94 degrees of freedom
Multiple R-squared:  0.5133,  Adjusted R-squared:  0.5082
F-statistic: 99.15 on 1 and 94 DF,  p-value: 2.264e-16
```

El resumen estadísitico anterior nos da una R^2 cercana a .6 lo cual podemos decir es un resultado estadísticamente moderado. Mientras que el estadístico F de la prueba de Wald confirma que lo anterior.

Ahora bien, para validar que se cumplen los supuestos necesarios para el modelo de regresión lineal, tenemos los siguientes resultados:

3.1.4 Residuals vs Fitted



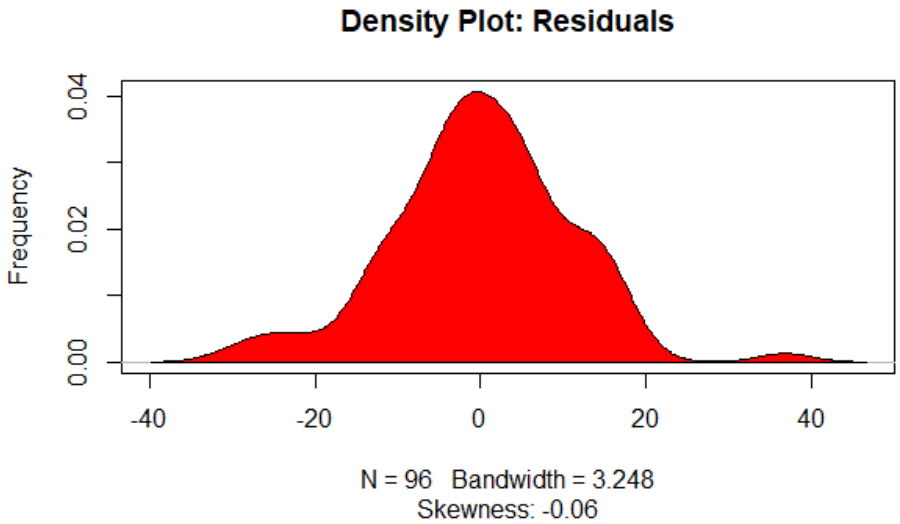


2.1.5 Normal Q-Q

Aqui podemos ver que la linearidad parece ser respetada, pues la linea roja parece estar cerca de la linea punteada. También podemos observar la heterocedasticidad, ya que al movernos a la derecha sobre el eje horizontal x la propagación de los residuos parece ser creciente.

Observando la gráfica QQ, podemos inferir que los puntos caen aproximadamente sobre la línea, pero igual podemos ver que existen puntos del lado izquierdo que insinuan que las colas de la distribución de los residuos no son completamente planas. Lo anterior lo podemos comprobar con una gráfica de la densidad de los residuos.

3.1.6 Densidad Residuos



Sin embargo, podemos intuir que el problema no parece ser significativo, además de que el número de datos

nos permite establecer que por el Teorema del Límite Central nuestra $\hat{\beta}$ se distribuirá de forma normal lo que nos permite construir los intervalos de confianza y realizar la prueba de hipótesis.

3.2 Rating WhoScored vs. Puntos

Nuestro segundo modelo de estudio es de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{3}$$

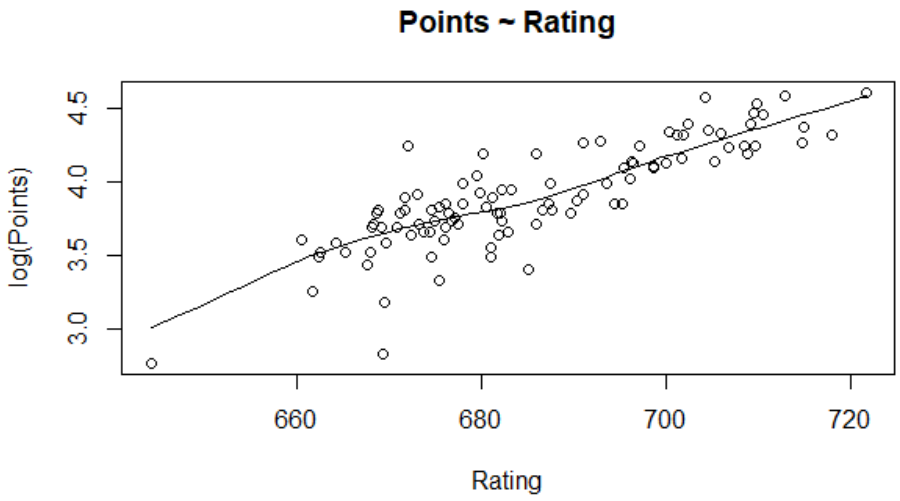
Donde Y_i es la variable dependiente que hace referencia al número de puntos, mientras que X_i es el rating promedio del equipo.

Notemos que las variable dependiente son los puntos, de donde nuestra varaiable dependiente tiene un soporte discreto. De ahí que proponemos una transformación del modelo anterior al siguiente:

$$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4}$$

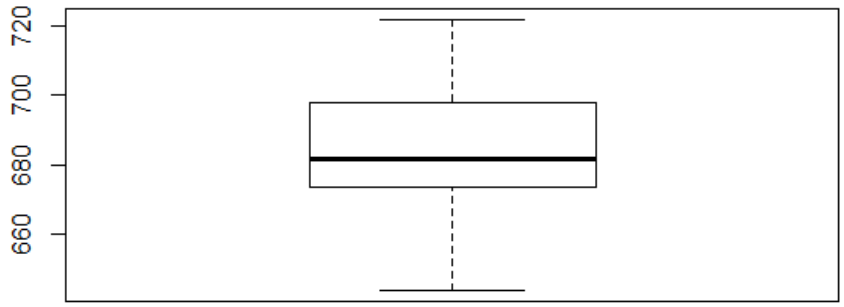
En primera instancia presentamos un scatterplot donde se visualicen las estadísticas de los puntos frente a las de Rating.

3.2.1 Scatter Plot Rating, log(Puntos)



Notemos que se puede observar una relación bastante lineal. Sin embargo, buscaremos outliers dentro de los datos utilizando un diagrama de caja para hacer un estudio más robusto.

2.2.2 Box Plot Rating



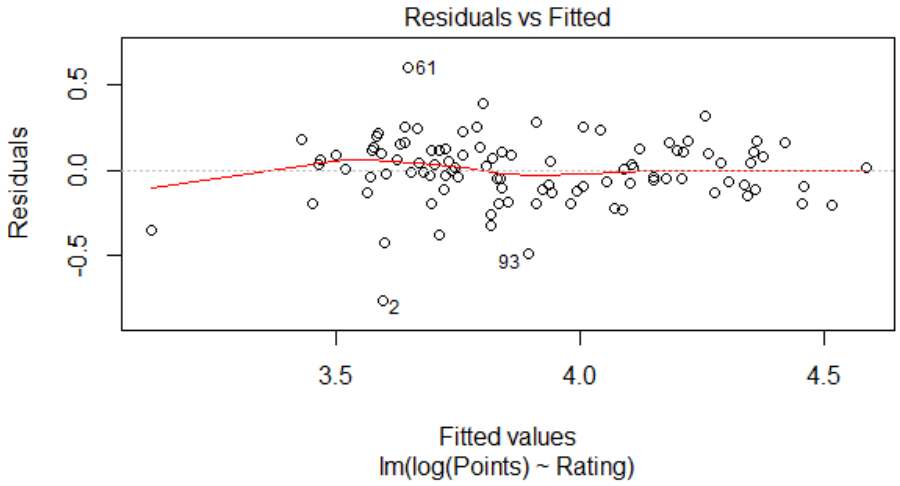
Notemos que no existen outliers dentro de los datos, lo que nos lleva entonces a continuar con nuestro análisis planteando el modelo como se describió en la ecuación 4.

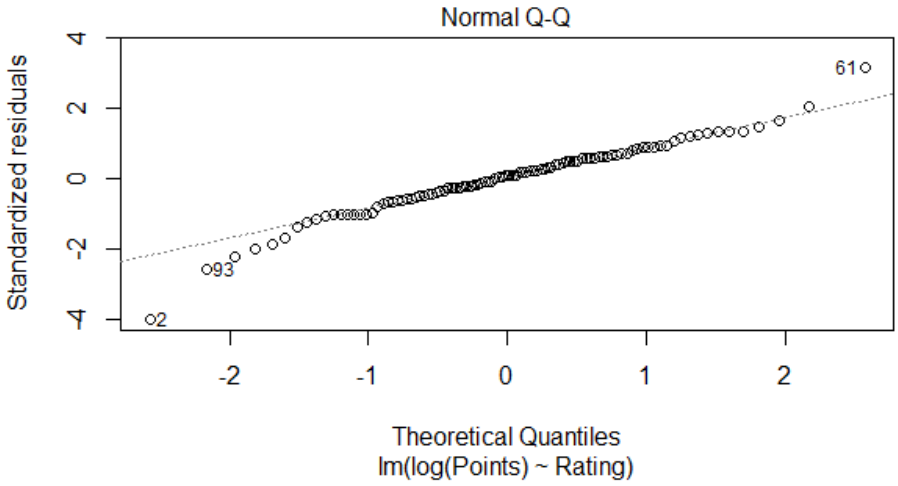
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.047533   0.840883  -10.76   <2e-16 ***
Rating       0.018890   0.001226   15.40   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.1923 on 98 degrees of freedom
Multiple R-squared:  0.7077,  Adjusted R-squared:  0.7047
F-statistic: 237.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Podemos ver del resumen anterior que el modelo lineal nos da un resultado significativo con una R^2 ajustada de 0.7047. La significancia del estadístico F que proviene de la prueba de Wald indica que el modelo lineal ajusta los datos y podría ser potencialmente empleado para realizar predicciones.

Las gráficas siguientes buscan robustecer el argumento de que los supuestos de linealidad y normalidad de los residuos se cumplen.

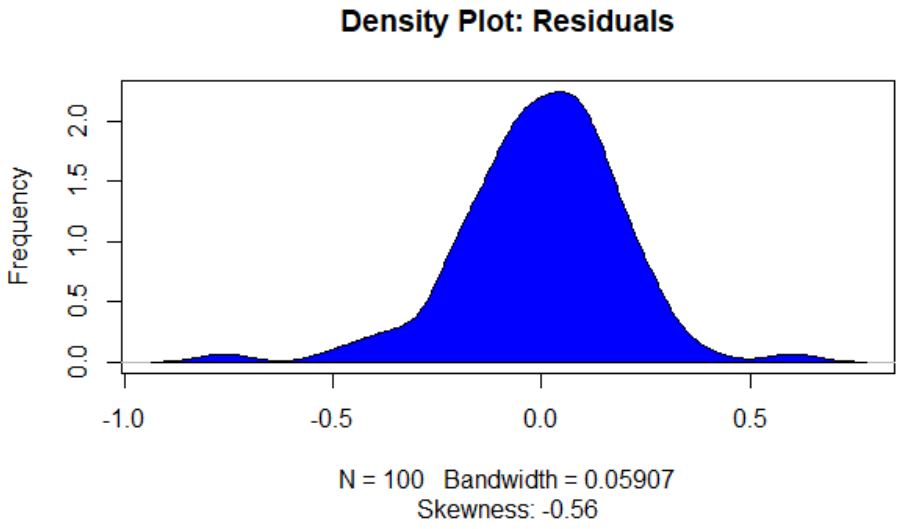
3.2.3 Residuals vs Fitted





3.2.4 Normal QQ

3.2.5 Densidad Residuos



Notemos que en la primera gráfica donde se presentan Residuals vs Fitted observamos que la línea roja está prácticamente montada encima de la línea punteada. Mientras que la gran mayoría de los puntos se encuentran sobre la línea en la gráfica de Q-Q. La normalidad de los residuos puede ser ratificada observamos la gráfica, lo cual sabemos que se cumple debido al Teorema del Límite Central.

3.1.3 Puntos vs Avg90

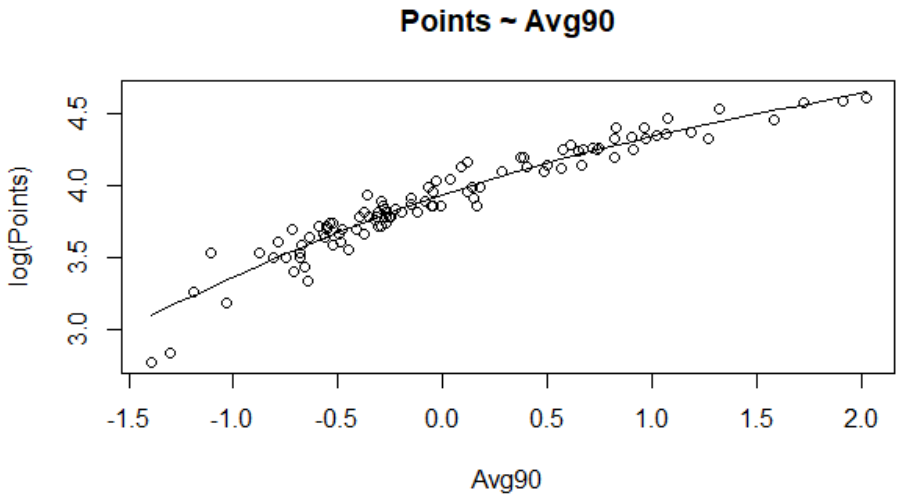
La métrica que ahora empleamos es otra métrica de eficiencia comúnmente empleada en el fútbol. La anteriór busca explicar la relación entre los goles recibidos y los goles anotados mientras un jugador se encuentra en el terreno de juego extendida a 90 minutos.

Nuevamente, como se ha descrito en los anteriores dos incisos tenemos que utilizar un modelo de la siguiente forma.

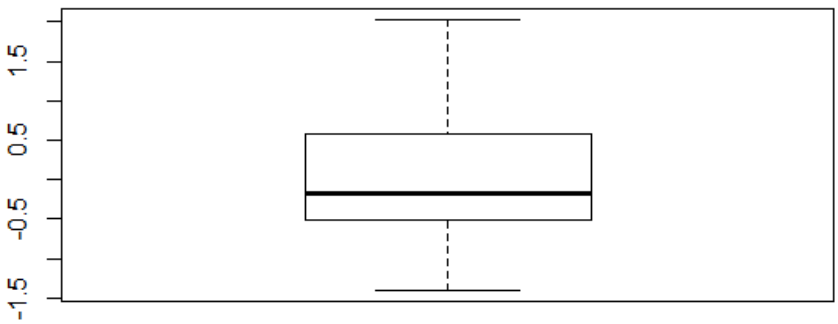
$$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \tag{5}$$

Donde Y_i son los puntos obtenidos y X_i es la métrica anteriormente descrita.
A continuación presentamos los resultados del anterior modelo. En primera instancia, tenemos el scatterplot.

3.3.1 ScatterPlotAvg90, log(Puntos)



Aqui podemos ver una relación lineal que no parece ser tan evidente, por lo que se justificará a continuación. En segundo lugar, buscamos .ºutliers"de la misma forma en que hemos realizado con los modelos anteriores y obtenemos el siguiente boxplot, que no indica la existencia de .ºutliers".



3.3.2 BoxPlot Avg90

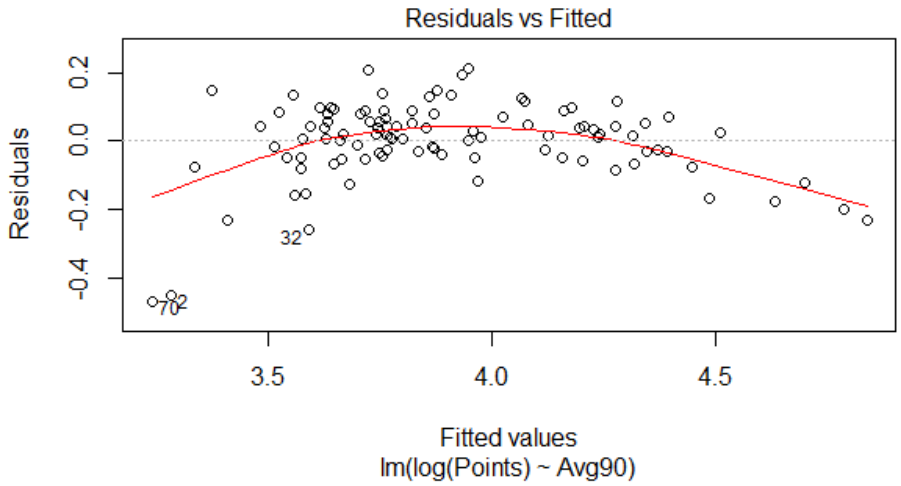
Los resultados del modelo son los siguientes:

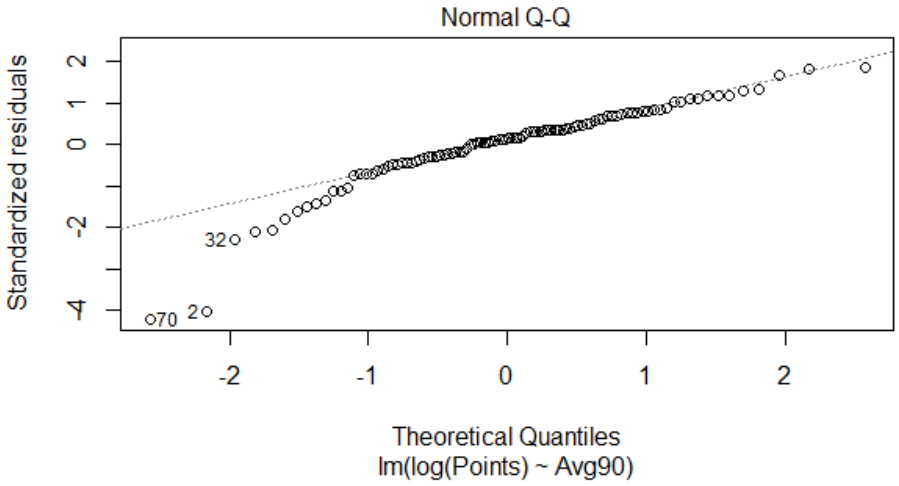
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.89298    0.01146  339.73   <2e-16 ***
Avg90        0.46652    0.01604   29.09   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.1146 on 98 degrees of freedom
Multiple R-squared:  0.8962,  Adjusted R-squared:  0.8952
F-statistic: 846.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

En este caso obtenemos una R^2 ajustada muy significativa ya que es de practicamente .9, lo que nos indica que podemos explicar casi el 90 % de la variabilidad de los puntos usando la metrica Avg90.

Lo único que nos falta es la validación de supuestos lo cual se realiza a través de los siguientes gráficos.

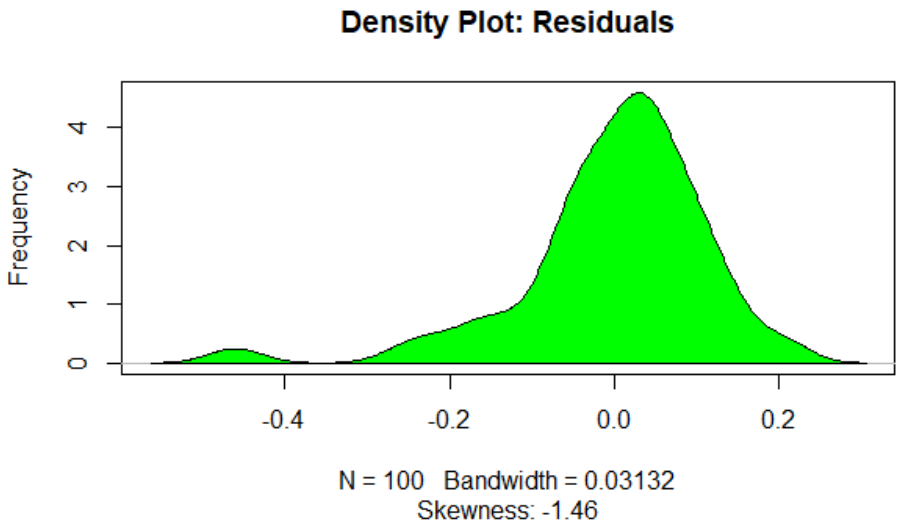
3.3.3 Residuals vs Fitted





3.3.4 Normal QQ

2.3.5 Densidad Residuos



Con las gráficas anteriores podríamos validar los supuestos que son necesarios para poder realizar el análisis. Notamos que el supuesto de varianza constante no se cumple, pues tenemos heterocedasticidad al observar el diagrama de residuals vs fitted. Sin embargo, el supuesto de normalidad se cumple, por lo que continuaremos con este análisis.

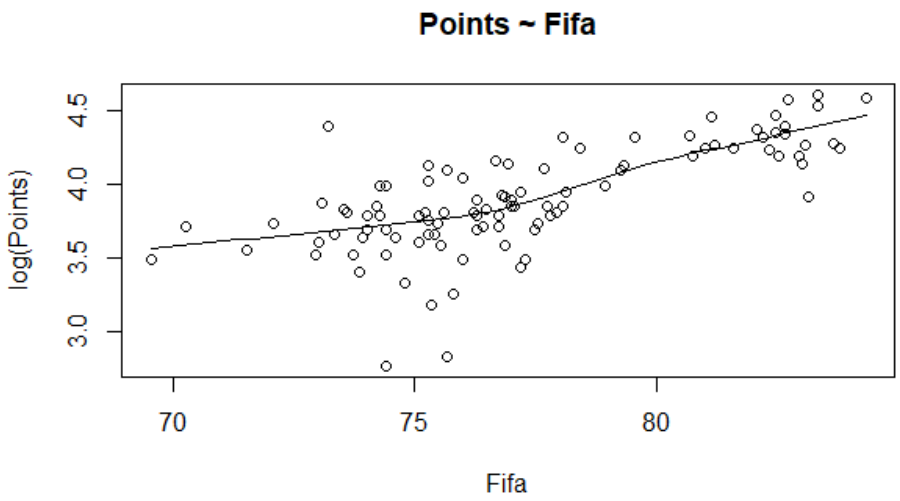
3.4 Puntos vs Fifa

Ahora se buscará presentar un modelo de regresión lineal simple empleado las calificaciones dadas por el FIFA. Ahora bien, siguiendo la lógica presentada en los modelos anteriores tenemos un modelo de la siguiente forma.

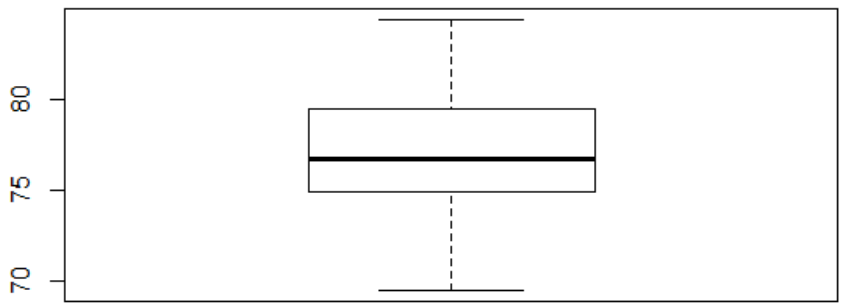
$$log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \tag{6}$$

Donde Y_i son los puntos obtenidos y X_i es la métrica anteriormente descrita. Observemos el scatter plot y el boxplot del modelo.

3.4.1 ScatterPlot Fifa, log(Puntos)



3.4.2 BoxPlot Fifa



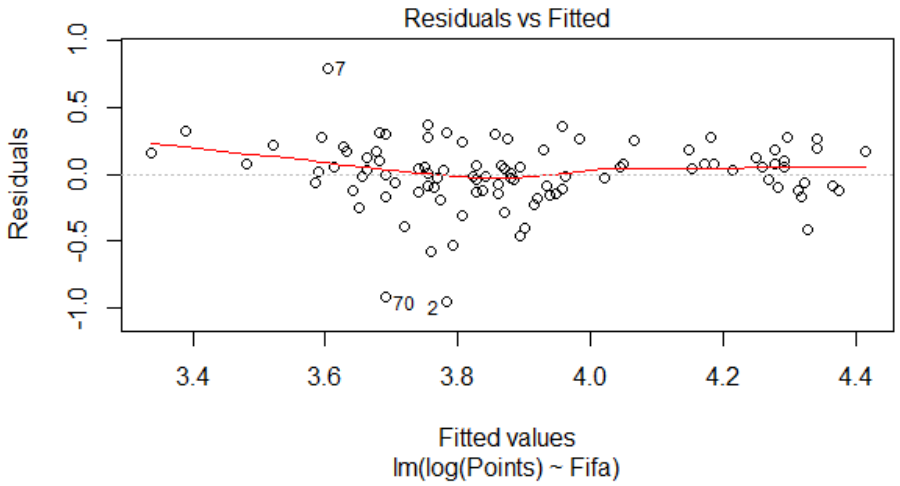
Con las anteriores gráficas podemos justificar la relación lineal que presentan los datos, y la ausencia de datos .ºutliers". Ahora bien, los resultados estadísticos obtenidos en el modelo son los siguientes:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.727211   0.569254  -3.034  0.00309 **
Fifa         0.072822   0.007359   9.895  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.2515 on 98 degrees of freedom
Multiple R-squared:  0.4998,  Adjusted R-squared:  0.4947
F-statistic: 97.92 on 1 and 98 DF,  p-value: < 2.2e-16
```

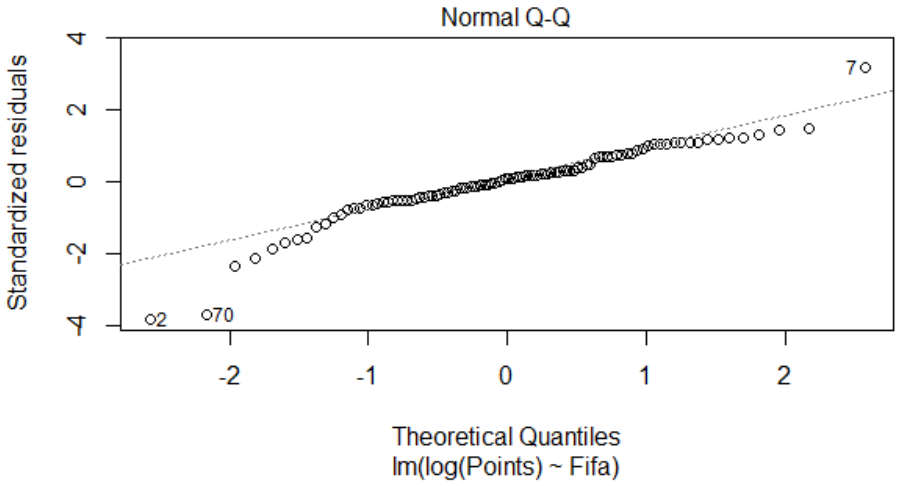
Notemos que la R^2 nos otorga un resultado moderado, pues es de aproximadamente 0.5, es decir solamente se explica el 50 % de la variabilidad del modelo.

Ahora bien, buscamos validar los supuestos del modelo de regresión lineal, como hemos hecho anteriormente.

3.4.3 Residuals vs Fitted

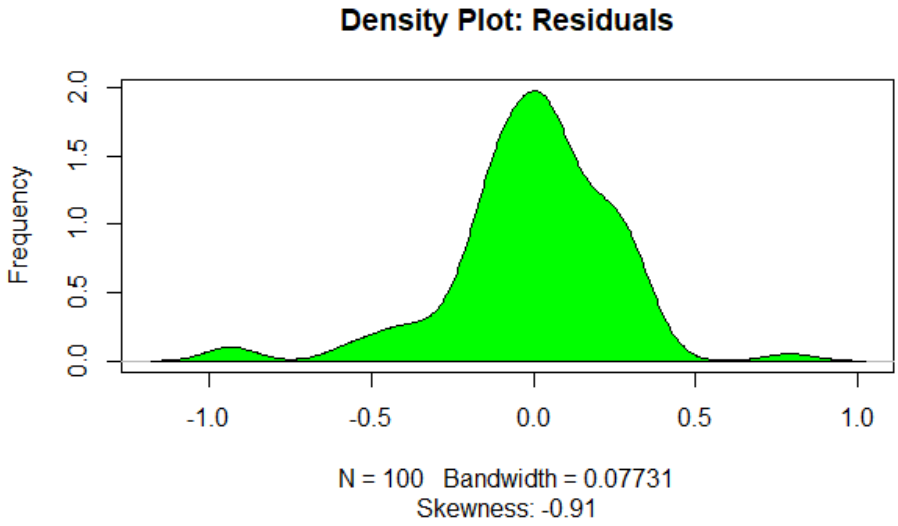


3.4.4 Normal QQ



Residuos

3.4.5 Densidad



Podemos observar entonces, que los supuestos de normalidad y linealidad se cumplen.

3.5 Conclusiones de los Modelos de Regresión Simple

Realizamos 4 modelos de regresión simple diferentes. ¿Qué podemos concluir a partir de los resultados obtenidos?

1. La métrica de Avg90 parece ser la mejor de todas, ya que ambos coeficientes β_0 y β_1 son significativos con un $\alpha = 0{,}001$ y la R^2 es mayor para este modelo. Además de que satisface todos los supuestos necesarios para el modelo de regresión lineal.

2. Notando que las variables de Avg90 y Rating son medidas expost, mientras que las variables de valor de mercado y Fifa son ex ante, los resultados dados por el valor de mercado parecen ser sumamente interesantes, ya que antes de comenzar la temporada se podría utilizar el valor de mercado para realizar predicciones.

3. El modelo del Fifa parece ser el peor, de donde deberíamos de desconfiar de las simulaciones del anterior juego para realizar diagnósticos.

3.6 Modelos de Regresión Lineal Múltiple

Observando los resultados obtenidos a través de los modelos de regresión lineal simple, surge la duda de si aumentando el número de variables independientes o variables explicativas podríamos obtener mejores resultados.

Luego entonces, empezamos por realizar un breve diagnóstico de las correlaciones entre las variables explicativas para saber que variables podemos emplear en el modelo lineal múltiple para no tener un problema de multicolinealidad. Recordemos, que cuando exista una correlación mayor a ,8 debemos de preocuparnos de que exista multicolinealidad. Ahora bien las correlaciones obtenidas son como sigue:

1. Cor(Fifa, Avg90) = 0.7515644
2. Cor(Fifa, Rating) = 0.6032466
3. Cor(Fifa, value) = 0.8680784
4. Cor(value,Rating) = 0.5570597
5. Cor(value, Avg90) = 0.7697626
6. Cor(Rating, Avg90) = 0.8854012

3.6.1 Puntos = WhoScored + Value

Interpretativamente consideraremos un modelo donde podamos estudiar la relación de una variable que se observe antes de comenzar la temporada y una que se observe después de comenzar la temporada. De ahí que consideremos Rating y value como nuestras variables explicativas, ya que la correlación entre dichas variables es menor a ,8

Al igual que en los modelos de regresión lineal simple, Y_i es la variable dependiente que representa el número total de puntos obtenidos durante una temporada, por otro lado, X_{i1} representa el rating promedio del equipo y X_{i2} representa el valor promedio en el mercado. Así, llegamos al siguiente modelo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{7}$$

Al igual que en regresión lineal simple, se aplica la misma transformación con *log* ya que la variable Y_i tiene un soporte discreto. De ahí se llega al siguiente modelo:

$$log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{8}$$

Para dicho modelo aplicamos el mismo procedimiento que en regresión lineal simple para identificar los *outliers* y eliminarlos.

Utilizando intervalos de confianza al 97.5 % llegamos a los siguientes resultados:

$$\beta_0 \in (-7{,}541, -4{,}008) \tag{9}$$

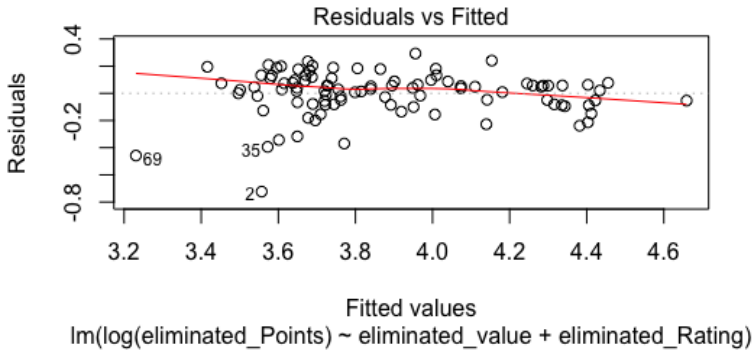
$$\beta_1 \in (0{,}008, 0{,}016) \tag{10}$$

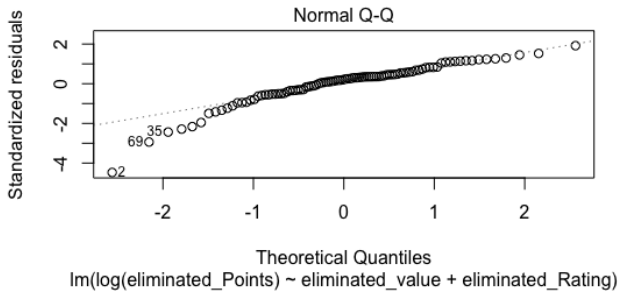
$$\beta_2 \in (1{,}123, 1{,}648) \tag{11}$$

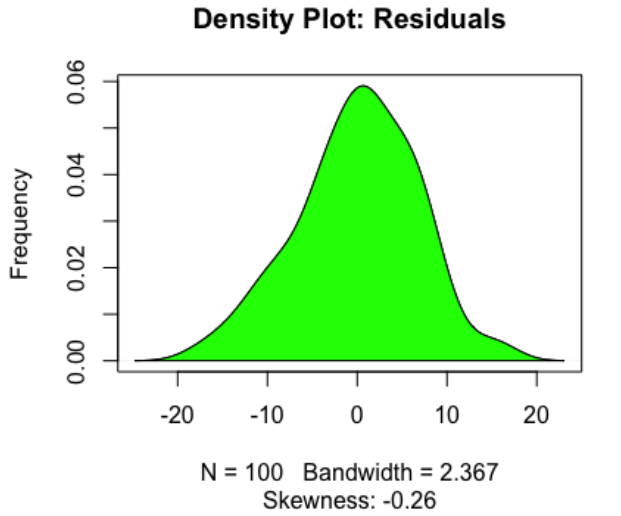
Con estos resultados β_0 , β_1 y β_2 son significativas para el modelo ya que 0 no pertenece a sus respectivos intervalos.

Aplicando el resumen llegamos a una R^2 de .78 lo cual es mejor que los modelos de regresión donde unicamente se utilizaba una de las variables (valor promedio en el mercado o rating promedio del equipo)

Ahora hay que verificar los supuestos. Tenemos los siguientes resultados:







En la gráfica Residuals vs Fitted se puede notar que la línea roja no esta montada sobre la línea punteada, de hecho la línea roja parece una recta con pendiente negativa, con este detalle ya no se cumple con el supuesto de varianza constante. Sin embargo, no parece ser tan grave por lo que podemos notar que tenemos heterocedasticidad pero no resulta tan significativa. Por otro lado, con las gráficas Normal Q-Q y Density Plot podemos notar que los residuos se distribuyen normal. A continuación analizaremos otro modelo donde con dos variables explicativas.

3.6.2 Puntos = Fifa + Rating

Para el siguiente modelo utilizamos las variables explicativas Fifa y Rating ya que su correlación es menor a ,8. Al igual que en el modelo anterior, Y_i representa los puntos pero ahora X_{i1} representa la variable Fifa y X_{i2} representa a Rating. Siguiendo el mismo procedimiento que los otros modelos se aplicó la transformación *log* y llegamos a la siguiente modelo:

$$log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{12}$$

Proseguimos a eliminar los *outliers* y a realizar los intervalos de confianza al 97.5 % para cada β_i. Llegamos a los siguientes intervalos:

$$\beta_0 \in (-9{,}990, -6{,}910) \tag{13}$$

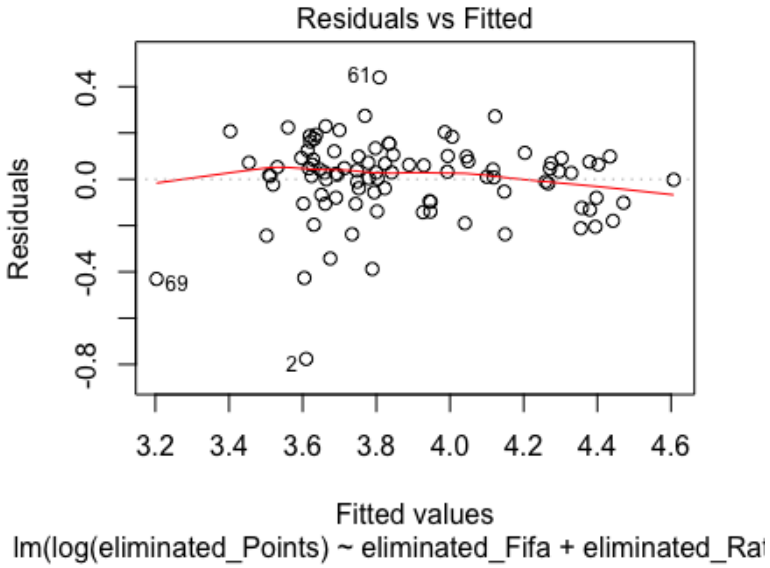
$$\beta_1 \in (0{,}016, 0{,}042) \tag{14}$$

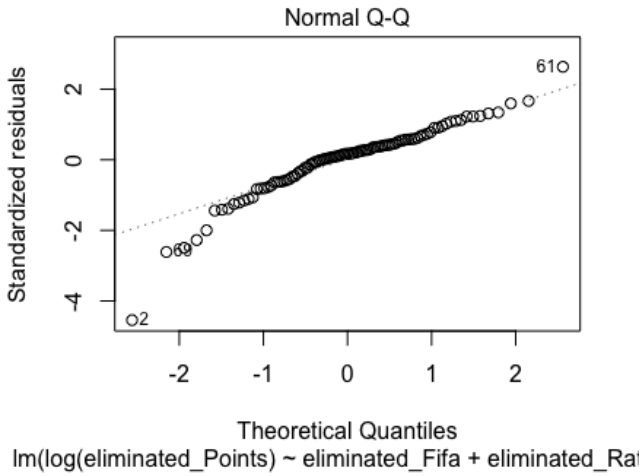
$$\beta_2 \in (1{,}190, 1{,}744) \tag{15}$$

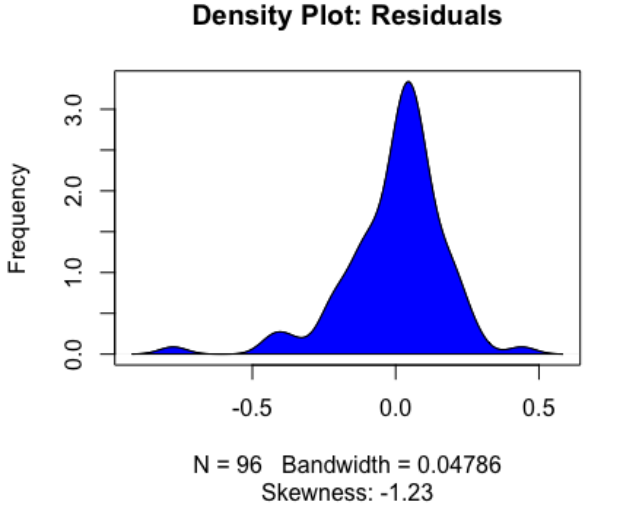
Con estos resultados β_0 , β_1 y β_2 son significativas para el modelo ya que 0 no pertenece a sus respectivos intervalos.

Aplicando el resumen llegamos a una R^2 de .75 lo cual es mejor que los modelos de regresión donde unicamente se utilizaba una de las variables.

Ahora hay que verificar los supuestos. Tenemos los siguientes resultados:





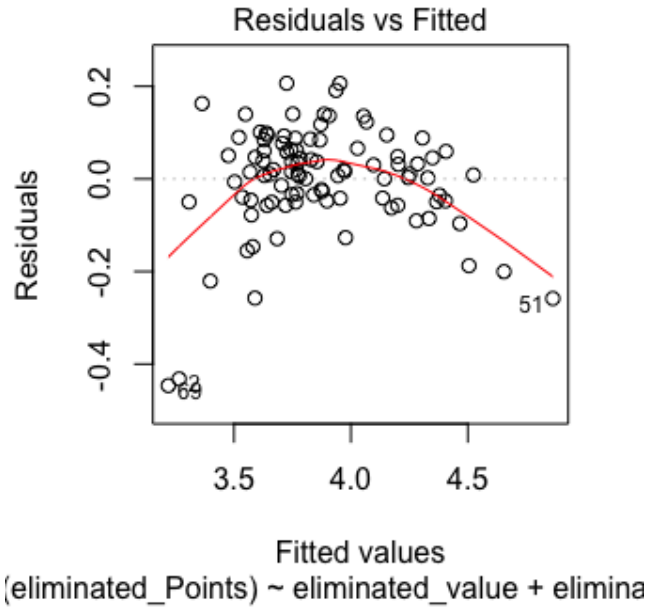


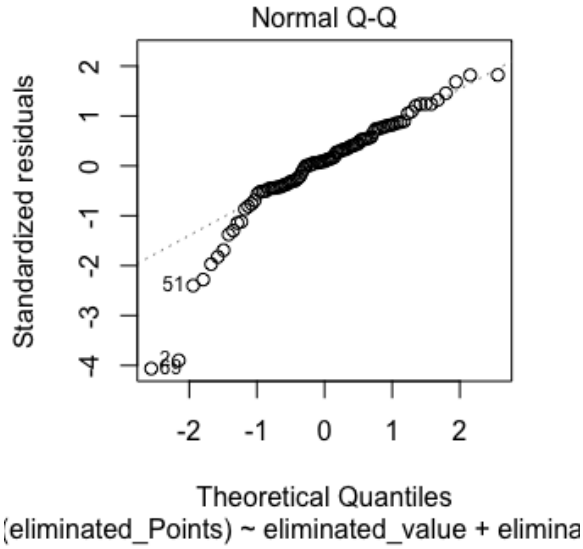
A diferencia del primer modelo con dos variables explicativas, en la gráfica Residuals vs Fitted se puede notar que la línea roja esta un poco mas montada hacia la línea punteada, esto nos asegura una varianza constante. Por otro lado, las gráficas Normal Q-Q y Density Plot nos muestran que los residuos tienen una distribución normal.

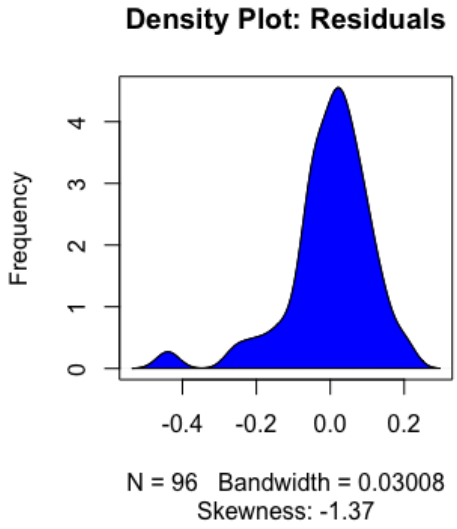
3.6.3 Puntos = Value + Avg90

Para el siguiente modelo hicimos algo diferente a todos los modelos, en lugar de hacer la estimación de los puntos hicimos una estimación de cuantos puntos por partido obtiene cada equipo, ya teniendo el cuantos puntos se ganan por partido multiplicamos por el número de juegos (38) y así obtuvimos la suma de los puntos por

temporada. Para este modelo también se aplicó la transformación *log* para que la variable puntos por partido tenga soporte real. A continuación realizaremos la validación de supuestos.







Aunque los residuos parecen tener una distribución normal, la varianza no es constante por su forma cuadrática; sin embargo, no podíamos descartar este modelo ya que tiene resultados muy buenos. A continuación mostraremos los resultados

4. Resultados

Después de analizar los modelos propuestos utilizamos las estadísticas de la temporada 19-20 para predecir los puntos de cada equipo y comparamos contra los puntos obtenidos hasta la fecha. A continuación mostraremos la tabla general hasta la fecha faltando 9 jornadas por disputarse o 27 puntos por jugarse. Para comparar los resultados contra los puntos obtenidos hasta la fecha clasificamos a los equipos en cuatro categorías. La primera categoría son los primeros cuatro lugares, son los equipos que clasifican a la Champions League. La segunda categoría esta formada por los siguientes tres equipos, estos son los equipos que clasifican a la Europa League. La tercera categoría esta conformada por los equipos que están en las posiciones ocho a diecisiete, estos son los equipos de media tabla. La cuarta categoría esta formada por los últimos tres lugares de la tabla, son los equipos que descienden.

Equipo	Puntos
Liverpool	82
Manchester City	57
Leicester City	53
Chelsea	48
Manchester United	45
Wolves	43
Sheffield Utd	43
Tottenham	41
Arsenal	40
Burnley	39
Crystal Palace	39
Everton	37
Newcastle United	35
Southampton	34
Brighton	29
West Ham	27
Watford	27
Bournemouth	27
Aston Villa	25
Norwich City	21

Cuadro 1: Tabla Premier League 2019-2020

4.1 Points vs Avg90

Equipo	Puntos
Liverpool	101
Manchester City	94
Leicester City	76
Manchester United	62
Chelsea	57
Tottenham	57
Wolves	56
Sheffield Utd	54
Arsenal	53
Crystal Palace	44
Brighton	44
Everton	43
Burnley	42
Newcastle United	40
Southampton	40
West Ham	39
Watford	38
Bournemouth	38
Aston Villa	36
Norwich City	32

Cuadro 2: Tabla de Resultados Points vs Avg90

4.2 Points vs Value

Equipo	Puntos
Liverpool	111
Manchester City	105
Tottenham	90
Chelsea	83
Arsenal	73
Manchester United	72
Leicester City	66
Everton	63
Wolves	58
West Ham	55
Bournemouth	52
Aston Villa	51
Newcastle United	48
Crystal Palace	47
Southampton	47
Brighton	46
Watford	45
Burnley	45
Norwich City	43
Sheffield Utd	42

Cuadro 3: Tabla de Resultados Points vs Value

4.3 Points vs Rating

Equipo	Puntos
Manchester City	70
Liverpool	67
Leicester City	59
Chelsea	56
Manchester United	50
Tottenham	50
Wolves	49
Arsenal	44
West Ham	43
Crystal Palace	42
Newcastle United	42
Sheffield Utd	40
Aston Villa	39
Brighton	39
Everton	39
Watford	38
Burnley	37
Southampton	36
Bournemouth	35
Norwich City	34

Cuadro 4: Tabla de Resultados Points vs Rating

4.4 Points vs Fifa

Equipo	Puntos
Manchester City	88
Liverpool	81
Tottenham	78
Arsenal	62
Chelsea	60
Manchester United	59
Everton	55
West Ham	54
Leicester City	52
Watford	51
Wolves	50
Crystal Palace	47
Newcastle United	47
Southampton	46
Burnley	45
Bournemouth	44
Aston Villa	41
Brighton	39
Sheffield Utd	35
Norwich City	34

Cuadro 5: Tabla de Resultados Points vs Fifa

4.5 Points = Value + Rating

Equipo	Puntos
Liverpool	105
Manchester City	102
Chelsea	71
Tottenham	69
Leicester City	63
Manchester United	59
Arsenal	55
Wolves	52
West Ham	45
Everton	45
Newcastle United	42
Crystal Palace	42
Aston Villa	41
Brighton	39
Sheffield Utd	39
Burnley	39
Bournemouth	39
Watford	37
Southampton	37
Norwich City	35

Cuadro 6: Tabla de Resultados Points = Value + Rating

4.6 Points = Fifa + Rating

Equipo	Puntos
Manchester City	82
Liverpool	76
Tottenham	59
Chelsea	58
Leicester City	58
Manchester United	53
Arsenal	50
Wolves	49
West Ham	45
Crystal Palace	43
Newcastle United	42
Everton	42
Watford	40
Burnley	39
Aston Villa	38
Southampton	37
Brighton	37
Sheffield Utd	37
Bournemouth	36
Norwich City	32

Cuadro 7: Tabla de Resultados Points = Fifa + Rating

4.7 Points = Value + Avg90

Equipo	Puntos
Liverpool	100
Manchester City	94
Leicester City	78
Manchester United	62
Chelsea	56
Wolves	56
Tottenham	55
Sheffield Utd	55
Arsenal	52
Brighton	44
Crystal Palace	44
Everton	42
Burnley	42
Newcastle United	40
Southampton	39
West Ham	39
Watford	38
Bournemouth	37
Aston Villa	35
Norwich City	32

Cuadro 8: Tabla de Resultados Points = Value + Avg90

- Podemos observar los resultados anteriores y comparar más que los puntos las posiciones que predice el modelo, y sobre todo realizaremos una mayor clasificación de los dato. Como sabemos los primeros 4 equipos de la tabla participarán en la UEFA Champions League, mientras que el 5to, 6to y 7mo lugar jugarán Europa League, finalmente los últimos 3 equipos de la tabla descenderán.

- Con lo anterior en mente buscaremos interpretar que modelo ajusta mejor los datos.

	Actual	Avg90	Valor	Rating	Fifa	Valor + Rating	Fifa + Rating	Valor + Avg90
CAMPEON	Liverpool	Liverpool	Liverpool	Manchester City	Manchester City	Liverpool	Manchester City	Liverpool
CHAMPIONS LEAGUE	Manchester City	Manchester City	Manchester City	Liverpool	Liverpool	Manchester City	Liverpool	Manchester City
	Leicester	Leicester	Tottenham	Leicester	Tottenham	Chelsea	Tottenham	Leicester City
	Chelsea	Manchester United	Chelsea	Chelsea	Arsenal	Tottenham	Chelsea	Manchester United
EUROPA LEAGUE	Manchester United	Chelsea	Arsenal	Manchester United	Chelsea	Leicester	Leicester	Chelsea
	Wolves	Tottenham	Manchester United	Tottenham	Manchester United	Manchester United	Manchester United	Wolves
	Sheffield United	Wolves	Leicester	Wolves	Everton	Arsenal	Arsenal	Tottenham
DESCENSO	Bournemouth	Bournemouth	Burnley	Southampton	Brighton	Watford	Sheffield United	Bournemouth
	Aston Villa	Aston Villa	Norwich City	Bournemouth	Sheffield United	Southampton	Bournemouth	Aston Villa
	Norwich City	Norwich City	Sheffield United	Norwich City	Norwich City	Norwich City	Norwich City	Norwich City

- CAMPEÓN: El Liverpool hubiera quedado campeón de la temporada 19-20.Notemos que 4 de 7 modelos predicen al campeón de forma correcta: Avg90, Valor, Valor+Rating y Valor + Avg90.

- CHAMPIONS LEAGUE: Los equipos que jugarían Champions League son: Liverpool, Manchester City, Leicester y Chelsea. Notemos que el modelo empleando Rating es el único que predice de forma correcta a los 4 equipos que participan. Los modelos de Avg90, Valor, Valor + Rating y Valor+Avg90 predicen 3 de 4 equipos que jugarían la Champions League. Finalmente los modelos de Fifa y Fifa + Rating tienen una efectividad del 50 %

- EUROPA LEAGUE: Los equipos que jugarían la Europa League Manchester United, Wolves y Sheffield United. En este caso, ninguno de los modelos predice de forma correcta todos los equipos. Sin embargo, el modelo de Rating obtiene 2 de 3 correctos, y todos los demás solamente tienen 1.

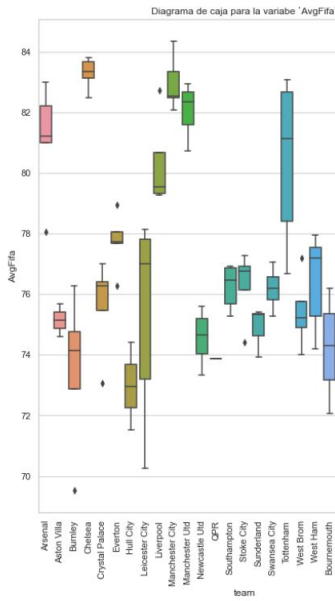
- DESCENSO: Los modelos de Avg90 y Valor+Avg90 predicen de forma correcta los 3 equipos que descenderán: Bournemouth, Aston Villa y Norwich City. Mientras que los modelos de Rating y Fifa + Rating predicen 2 de 3 de forma correcta.

- Construimos la siguiente métrica para poder determinar cuál es el mejor modelo: 3 puntos por determinar de forma correcta al campeón; 3 puntos por equipo que se determine de forma correcta que jugará Champions League; 2 puntos por equipo en Europa League, y 1 un punto por equipo de descenso.

- Si utilizamos la anterior métrica podemos clasificar los modelos de la siguiente forma.

1. Rating
2. Avg 90
3. Valor + Avg90
4. Valor
5. Valor + Rating
6. Fifa
7. Fifa + Rating

5. Anexo



A continuación anexamos los diagramas de caja por equipo para todas las variables explicativas.

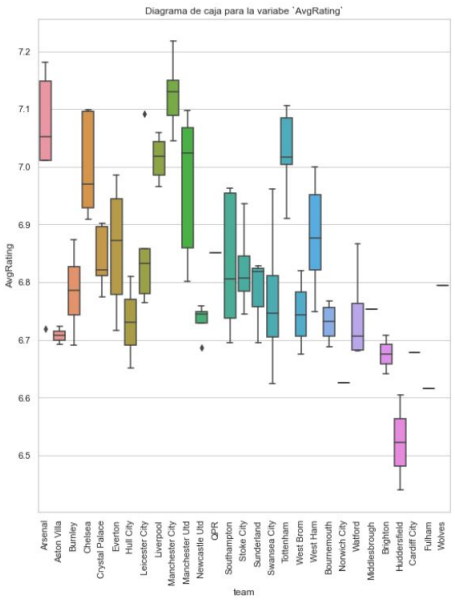


Diagrama de caja para la variabe `AvgRating`
AvgRating
team
Arsenal
Aston Villa
Burnley
Chelsea
Crystal Palace
Everton
Hull City
Leicester City
Liverpool
Manchester City
Manchester Utd
Newcastle Utd
QPR
Southampton
Stoke City
Sunderland
Swansea City
Tottenham
West Brom
West Ham
Bournemouth
Norwich City
Watford
Middlesbrough
Brighton
Huddersfield
Cardiff City
Fulham
Wolves

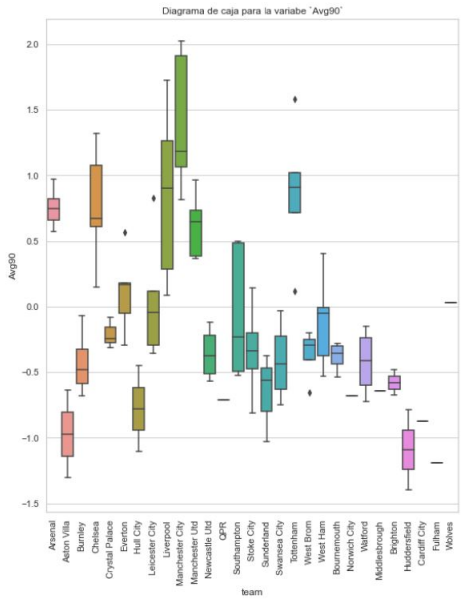


Diagrama de caja para la variabe `Avg90`
Avg90
team
Arsenal
Aston Villa
Burnley
Chelsea
Crystal Palace
Everton
Hull City
Leicester City
Liverpool
Manchester City
Manchester Utd
Newcastle Utd
QPR
Southampton
Stoke City
Sunderland
Swansea City
Tottenham
West Brom
West Ham
Bournemouth
Norwich City
Watford
Middlesbrough
Brighton
Huddersfield
Cardiff City
Fulham
Wolves

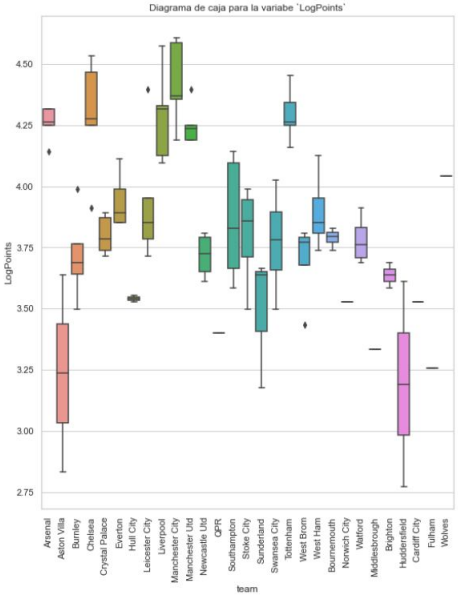


Diagrama de caja para la variabe `LogPoints`
LogPoints
team
Arsenal
Aston Villa
Burnley
Chelsea
Crystal Palace
Everton
Hull City
Leicester City
Liverpool
Manchester City
Manchester Utd
Newcastle Utd
QPR
Southampton
Stoke City
Sunderland
Swansea City
Tottenham
West Brom
West Ham
Bournemouth
Norwich City
Watford
Middlesbrough
Brighton
Huddersfield
Cardiff City
Fulham
Wolves

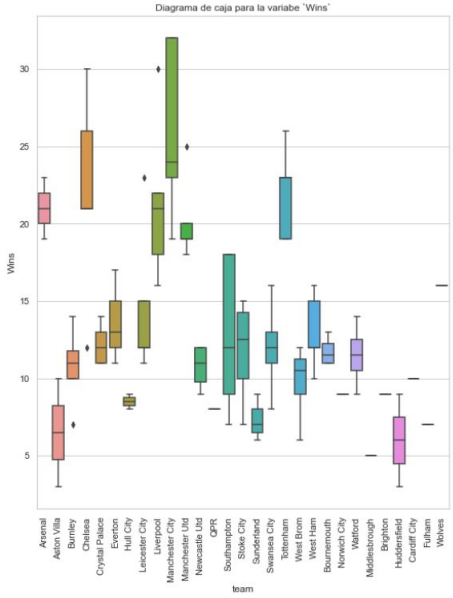
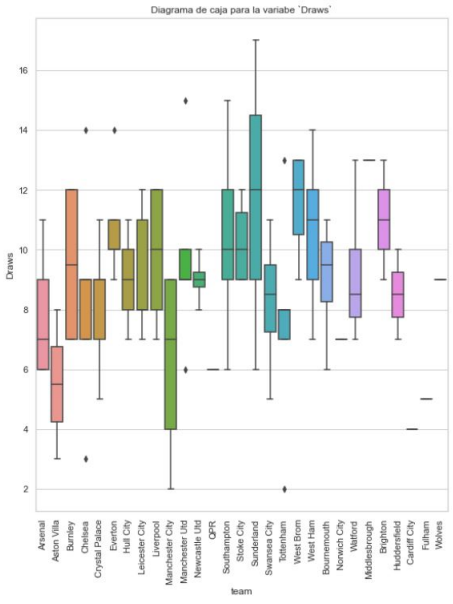
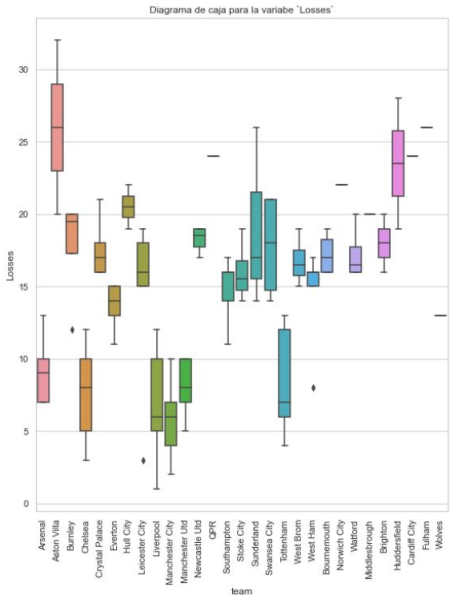
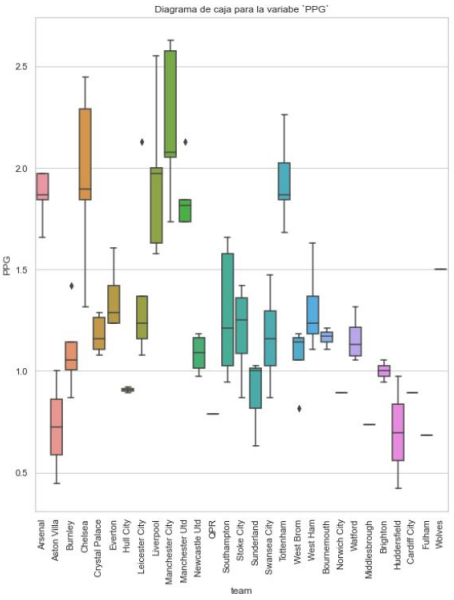


Diagrama de caja para la variabe `Wins`
Wins
team
Arsenal
Aston Villa
Burnley
Chelsea
Crystal Palace
Everton
Hull City
Leicester City
Liverpool
Manchester City
Manchester Utd
Newcastle Utd
QPR
Southampton
Stoke City
Sunderland
Swansea City
Tottenham
West Brom
West Ham
Bournemouth
Norwich City
Watford
Middlesbrough
Brighton
Huddersfield
Cardiff City
Fulham
Wolves







Referencias

[1] Football Reference Premier League Stats. `https://fbref.com/en/comps/9/Premier-League-Stats`. Consultado el: 15-05-2020.

[2] Kaggle Datasets for Fifa Ratings. `https://www.kaggle.com/karangadiya`. Consultado el: 16-05-2020.

[3] Transfer Markt values for Premier League. https://www.transfermarkt.us/premier-league/marktwerte/wettbewerb/GB1/pos//detailpos/0/altersklasse/alle/plus/1. Consultado el: 16-05-2020.

[4] Whoscored Ratings for Premier League Players. https://www.whoscored.com/Regions/252/Tournaments/2/England-Premier-League. Consultado el: 15-05-2020.

[5] *Whoscored Ratings Explained*. https://www.whoscored.com/Explanations. Consultado el: 20-05-2020.

[6] MarinStatsLectures. *Multiple Linear Regression in R.* https://www.youtube.com/watch?v=eTZ4VUZHzxw&feature=youtu.be, 2013. Consultado el: 12-05-2020.

[7] Basketball Reference. Calculating PER. https://www.basketball-reference.com/about/per.html. Consultado el: 10-05-2020.