

Regresion Lineal Múltiple

Ramírez Montes Jonathan Natael (50) & Sánchez Romero Paulina Michelle (51)

31/1/2021

Introducción.

Dragon Ball es uno de los animes más populares de la historia. Creado por Akira Toriyama en 1984 como manga, y llevado a la televisión en 1986. Este fenómeno ha traspasado fronteras y plataformas (papel, televisión, merchandising, videojuegos y cine). El impacto de Dragon Ball y parte de su éxito en Japón se origina con la influencia que tomó de la novela clásica de la literatura china "Viaje al Oeste", de gran popularidad en Asia. Este anime trata sobre una raza denominada "Saiyajin" con aspecto humano que cuentan con gran cantidad de energía denominada "ki". A través de las Artes Marciales los personajes logran desarrollar técnicas que les permiten emitir energía a través de su cuerpo y lanzar esta misma a sus rivales. A lo largo de varias series de Dragon Ball los personajes evolucionan, mejorando sus habilidades mientras pelean con villanos extraterrestres que quieren destruir la tierra y apoderarse del universo. Dragon Ball es un símbolo importante en la cultura pop pues ha marcado tendencia y moda desde los años 90. En la actualidad lo sigue siendo con la última creación de Akira Toriyama, Dragon Ball Super. En esta ocasión abordaremos un data set sobre uno de los videojuegos de Dragon Ball producido por la empresa BAN DAI, este videojuego se encuentra disponible en la play store del sistema operativo Android con el nombre de "Dragon Ball Legends".

Propósito.

Trataremos de predecir el nivel de poder en los personajes dadas las cualidades de defensa y ataque proporcionadas por el data set. Procederemos a través de una regresión lineal múltiple. Una vez conseguido el modelo clásico procederemos a utilizar la herramienta "rjags" proporcionada por el software de R para aplicar metodologías bayesianas para la estimación de los parámetros de nuestra regresión.

Objetivo del documento.

Mostrar la diferencia entre un análisis estadístico clásico y uno bayesiano utilizando una regresión lineal múltiple.

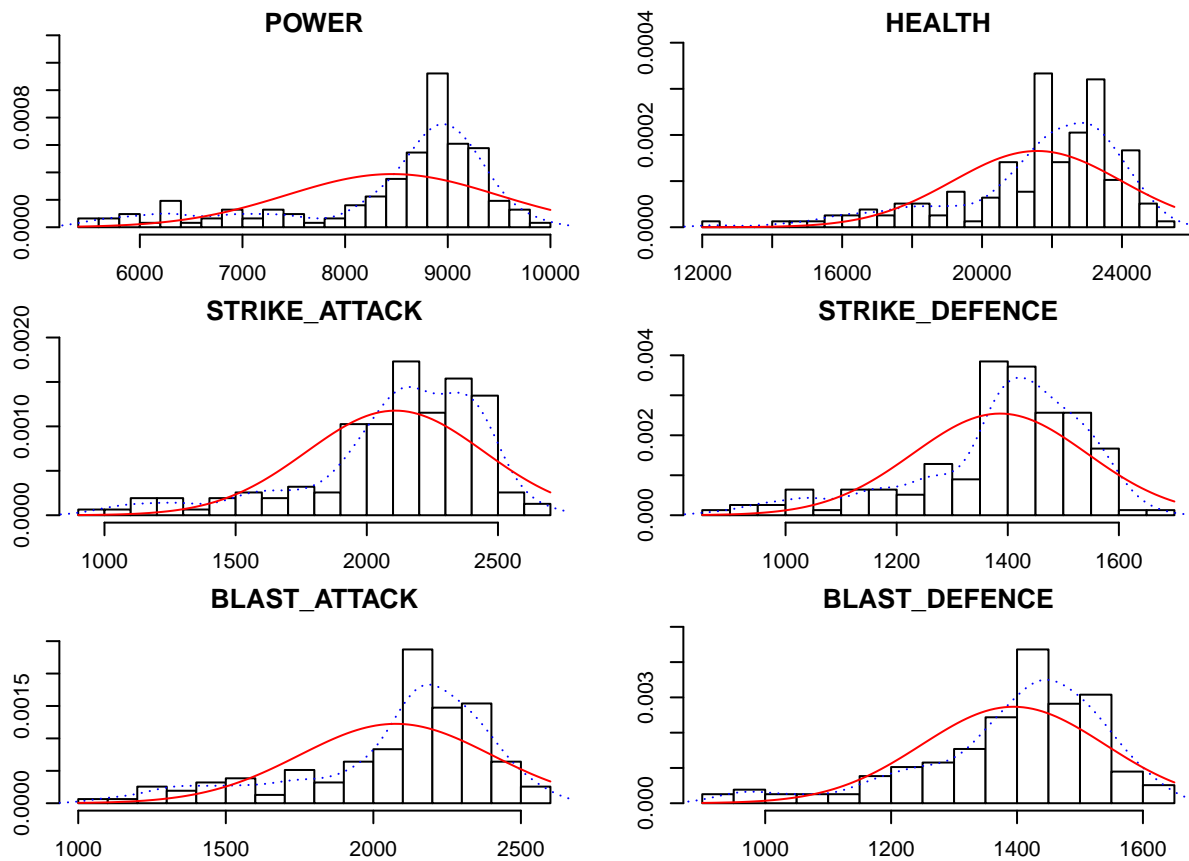
Entendimiento de los datos.

En este módulo se describirá mediante una tabla la información que fue obtenida vía la página web kaggle, el conjunto de datos estaba conformado por: Una tabla con la siguiente taxonomía: 290 observaciones con 8 columnas, cada registro contaba con un identificador CARD.NUMBER. los 290 registros inicialmente estaban representados por un tipo de dato carácter, no obstante esto no era correcto pues el data set presenta las siguientes variables: "CHARACTER", "CARD.NUMBER", "POWER", "HEALTH", "STRIKE.ATTACK", "STRIKE.DEFENCE", "BLAST.ATTACK", "BLAST.DEFENCE", de las cuales las únicas que eran tipo carácter son "CHARACTER" y "CARD.NUMBER", las demás son de tipo numérico pero estaban representadas de una manera incorrecta. A continuación, se muestra una tabla con la descripción de las variables.

Análisis descriptivo

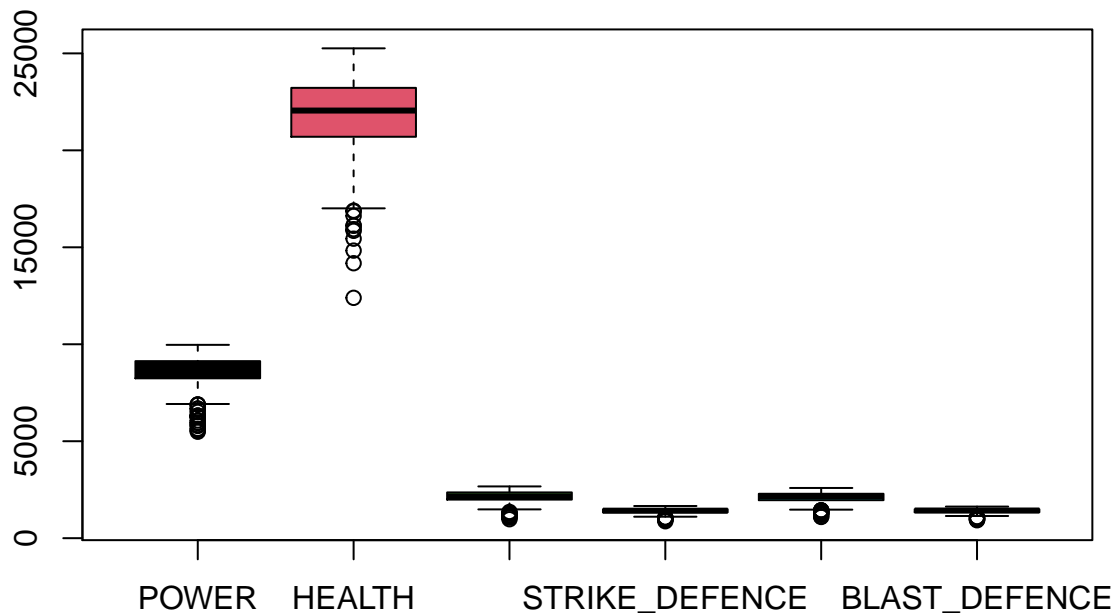
```
dball <- db[c(3:8)]
```

```
multi.hist(x = dball, dcol = c("blue", "red"), dlty = c("dotted", "solid"),  
  main = names(dball) , mar = c(2.5, 2.5, 1, 2))
```



Con estas gráficas se puede observar el comportamiento de las diferentes variables. Se puede notar que la variable HEALTH es una de las que pareciera que no le podríamos ajustar algún tipo de distribución, pues al menos gráficamente, se puede destacar que hay valores atípicos. En relación con las otras variables podemos ver que muestran colas pesadas. Nota: la línea roja es una distribución normal ajustada a los datos. Pero, por otra parte, podemos observar en las variables STRIKE se acumula más información en los valores más grandes, resaltando STRIKE_ATTACK, en cuanto a las variables BLAST, podemos ver que tienen un comportamiento similar a pesar de que, al menos intuitivamente, las variables parecieran no tener relación pues una es de ataque y la otra es de defensa.

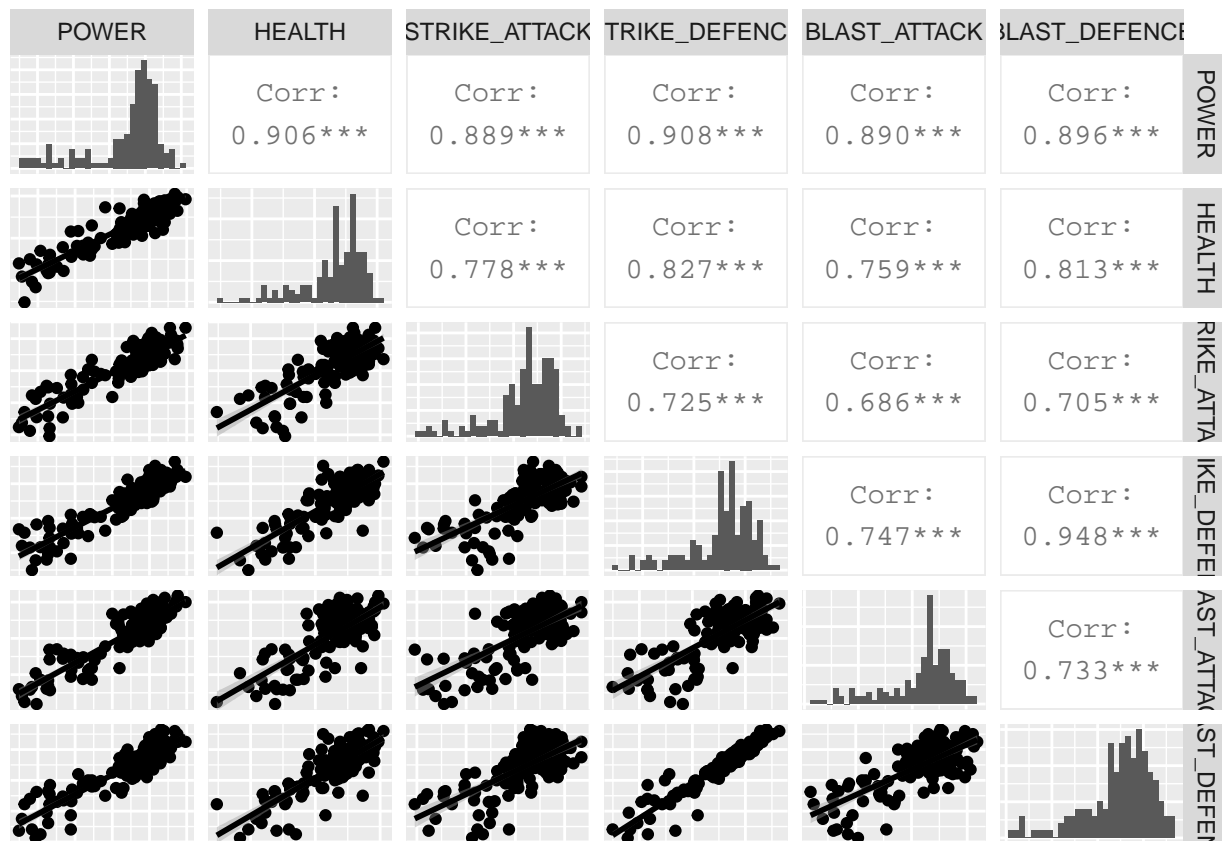
```
boxplot(dball, col = c(1:5))
```



Con esta gráfica podemos notar en mayor proporción los datos atípicos de nuestras variables, vemos que en todas las variables tenemos datos atípicos, pero que en donde se concentran los de mayor “lejanía” o mayor diferencia es en la variable HEALTH, pues como lo habíamos mencionado anteriormente, es la que tiene un intervalo de mayor magnitud, en comparación de BLAST_DEFENCE y STRIKE_DEFENCE. También podemos ver que POWER es la segunda variable que tiene más valores atípicos, y también un intervalo más grande que en las demás variables a excepción de HEALTH. Lo que nos indica que en POWER, además de tener personajes con menor poder que otros se ve una mayor diferencia, con los valores atípicos, entonces observamos que hay personajes que tienen mucho menos poder que otros debido a que todos los valores atípicos que tenemos en todas las variables son en la parte inferior (o en el lado menor del intervalo de valores), y lo mismo sucede con HEALTH.

```
ggpairs(dball, lower = list(continuous = "smooth"),
        diag = list(continuous = "barDiag"), axisLabels = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Estos valores nos indican la correlación que existe entre variables, por ejemplo, vemos que las correlaciones de mayor valor las tienen las demás variables con POWER (aunque no tienen la máxima), y que la menor correlación se puede observar en BLAST_ATTACK con STRIKE_ATTACK de 0.686, lo cual es un poco lógico tomando en cuenta que son diferentes tipos de ataques, uno directo y el otro a larga distancia. La correlación de mayor valor la tiene BLAST_DEFENCE con STRIKE_DEFENCE, con 0.948, que realmente tomaríamos la misma lógica que en las variables comentadas anteriormente, pero tal vez tengan mayor correlación por ser defensa en lugar de ataque.

Resúmenes estadísticos

```
summary(dball)
```

```
##      POWER      HEALTH  STRIKE_ATTACK  STRIKE_DEFENCE  BLAST_ATTACK
##  Min.   :5490   Min.   :12395   Min.   : 978.5   Min.   : 874   Min.   :1095
##  1st Qu.:8248   1st Qu.:20700   1st Qu.:1990.0   1st Qu.:1336   1st Qu.:1960
##  Median :8835   Median :22053   Median :2157.5   Median :1420   Median :2158
##  Mean   :8460   Mean   :21575   Mean   :2108.6   Mean   :1386   Mean   :2076
##  3rd Qu.:9130   3rd Qu.:23216   3rd Qu.:2360.0   3rd Qu.:1492   3rd Qu.:2290
##  Max.   :9970   Max.   :25260   Max.   :2670.0   Max.   :1660   Max.   :2590
##  BLAST_DEFENCE
##  Min.   : 926
##  1st Qu.:1340
##  Median :1428
##  Mean   :1394
##  3rd Qu.:1495
##  Max.   :1630
```

En cuanto a estos diferentes resúmenes, vemos que la variable que tiene valores más grandes es HEALTH, y que también es la que tiene un intervalo mucho más grande, pues varían en aproximadamente 13,000 unidades/puntos, lo que no pasa con STRIKE_DEFENCE, por ejemplo, que sólo va de 874 a 1660, o BLAST_DEFENCE de 926 a 1630. Por otra parte, analizando las medias y las medianas, nos interesa que sean el mismo valor, o que al menos se le acerque mucho una a otra, esto para poder inferir algún tipo de distribución que sea simétrica, como lo es en el caso de casi todas nuestras variables, exceptuando POWER, ya que es la de mayor diferencia entre la mediana y la media. También vemos que otra variable que se dispersa más con los valores, es nuestra variable dependiente POWER, y la que tiene el intervalo más pequeño de valores es la BLAST_DEFENCE, lo que nos puede decir que los diferentes personajes pueden tener una defensa de ataque directo más parecida entre ellos, y lo contrario sucede con la mencionada antes POWER, por lo que entre un personaje y otro puede haber una mayor diferencia de poder, lo cual puede poner en ventaja o desventaja en algún enfrentamiento.

Regresión Lineal Múltiple (análisis clásico)

```
##          HEALTH  STRIKE_ATTACK STRIKE_DEFENCE  BLAST_ATTACK  BLAST_DEFENCE
##          4.410738      2.756429      11.075798      2.729663      10.109834

##
## Call:
## lm(formula = POWER ~ HEALTH + STRIKE_ATTACK + STRIKE_DEFENCE +
##      BLAST_ATTACK + BLAST_DEFENCE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -380.04  -16.76   -3.52    9.63   374.27
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -115.580530   55.145857  -2.096      0.0378 *
## HEALTH         0.064932    0.004712  13.779 < 0.0000000000000002 ***
## STRIKE_ATTACK  0.991071    0.026531  37.355 < 0.0000000000000002 ***
## STRIKE_DEFENCE 0.981622    0.114861   8.546  0.00000000000000134 ***
## BLAST_ATTACK   0.994235    0.027460  36.206 < 0.0000000000000002 ***
## BLAST_DEFENCE  1.190618    0.118013  10.089 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.41 on 150 degrees of freedom
## Multiple R-squared:  0.9958, Adjusted R-squared:  0.9957
## F-statistic: 7186 on 5 and 150 DF, p-value: < 0.00000000000000022

##              2.5 %      97.5 %
## (Intercept)  -224.54352275 -6.61753780
## HEALTH         0.05562124  0.07424331
## STRIKE_ATTACK  0.93864832  1.04349423
## STRIKE_DEFENCE 0.75466789  1.20857648
## BLAST_ATTACK   0.93997620  1.04849368
## BLAST_DEFENCE  0.95743481  1.42380106

##          HEALTH  STRIKE_ATTACK STRIKE_DEFENCE
##          4.016370      2.673862      3.339840

##
## Call:
## lm(formula = POWER ~ HEALTH + STRIKE_ATTACK + STRIKE_DEFENCE)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -714.62 -112.97    3.63  150.32  541.84
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -183.31736   168.02532  -1.091      0.277
## HEALTH        0.11819    0.01465    8.069 0.000000000000000199 ***
## STRIKE_ATTACK  1.16015    0.08512   13.630 < 0.00000000000000002 ***
## STRIKE_DEFENCE 2.63095    0.20546   12.805 < 0.00000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 219.6 on 152 degrees of freedom
## Multiple R-squared:  0.9553, Adjusted R-squared:  0.9544
## F-statistic: 1083 on 3 and 152 DF, p-value: < 0.000000000000000022

##              2.5 %      97.5 %
## (Intercept)  -515.28395489 148.6492406
## HEALTH        0.08924651   0.1471258
## STRIKE_ATTACK  0.99198312   1.3283254
## STRIKE_DEFENCE 2.22502822   3.0368816

##      HEALTH  BLAST_ATTACK BLAST_DEFENCE
##      3.555027    2.600505    3.261961

##
## Call:
## lm(formula = POWER ~ HEALTH + BLAST_ATTACK + BLAST_DEFENCE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -710.14 -155.34    4.35  154.10  518.63
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -503.33348   180.87484  -2.783    0.00607 **
## HEALTH        0.14512    0.01413   10.272 < 0.00000000000000002 ***
## BLAST_ATTACK  1.21585    0.08950   13.585 < 0.00000000000000002 ***
## BLAST_DEFENCE 2.37272    0.22384   10.600 < 0.00000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 225.1 on 152 degrees of freedom
## Multiple R-squared:  0.953, Adjusted R-squared:  0.9521
## F-statistic: 1028 on 3 and 152 DF, p-value: < 0.000000000000000022

##              2.5 %      97.5 %
## (Intercept)  -860.6868025 -145.9801548
## HEALTH        0.1172055    0.1730257
## BLAST_ATTACK  1.0390295    1.3926776
## BLAST_DEFENCE 1.9304729    2.8149608

##      HEALTH STRIKE_ATTACK BLAST_ATTACK
##      3.341665    2.679233    2.493785
```

```
##
## Call:
## lm(formula = POWER ~ HEALTH + STRIKE_ATTACK + BLAST_ATTACK)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -675.48  -93.44  -15.79   113.08   471.87
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   568.93258   132.90984    4.281    0.0000329 ***
## HEALTH         0.13663     0.01109   12.320 < 0.0000000000000002 ***
## STRIKE_ATTACK  1.12091     0.07073   15.849 < 0.0000000000000002 ***
## BLAST_ATTACK   1.24266     0.07097   17.510 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 182.3 on 152 degrees of freedom
## Multiple R-squared:  0.9692, Adjusted R-squared:  0.9686
## F-statistic: 1594 on 3 and 152 DF, p-value: < 0.00000000000000022

##              2.5 %      97.5 %
## (Intercept)  306.3434097 831.5217420
## HEALTH       0.1147168   0.1585392
## STRIKE_ATTACK 0.9811828   1.2606468
## BLAST_ATTACK  1.1024465   1.3828718

##      HEALTH STRIKE_DEFENCE  BLAST_DEFENCE
##      3.252783    10.774356    10.071037

##
## Call:
## lm(formula = POWER ~ HEALTH + STRIKE_DEFENCE + BLAST_DEFENCE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -970.84 -191.81   42.11   217.20   660.75
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -842.6051   254.5905  -3.310    0.001167 **
## HEALTH        0.2016     0.0192   10.502 < 0.0000000000000002 ***
## STRIKE_DEFENCE 2.0617     0.5375   3.836    0.000183 ***
## BLAST_DEFENCE  1.5019     0.5588   2.688    0.007998 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.8 on 152 degrees of freedom
## Multiple R-squared:  0.9052, Adjusted R-squared:  0.9033
## F-statistic: 483.6 on 3 and 152 DF, p-value: < 0.00000000000000022

##              2.5 %      97.5 %
## (Intercept) -1345.5979240 -339.6123270
## HEALTH       0.1637083    0.2395728
## STRIKE_DEFENCE 0.9997598    3.1235679
## BLAST_DEFENCE 0.3978648    2.6060268
```

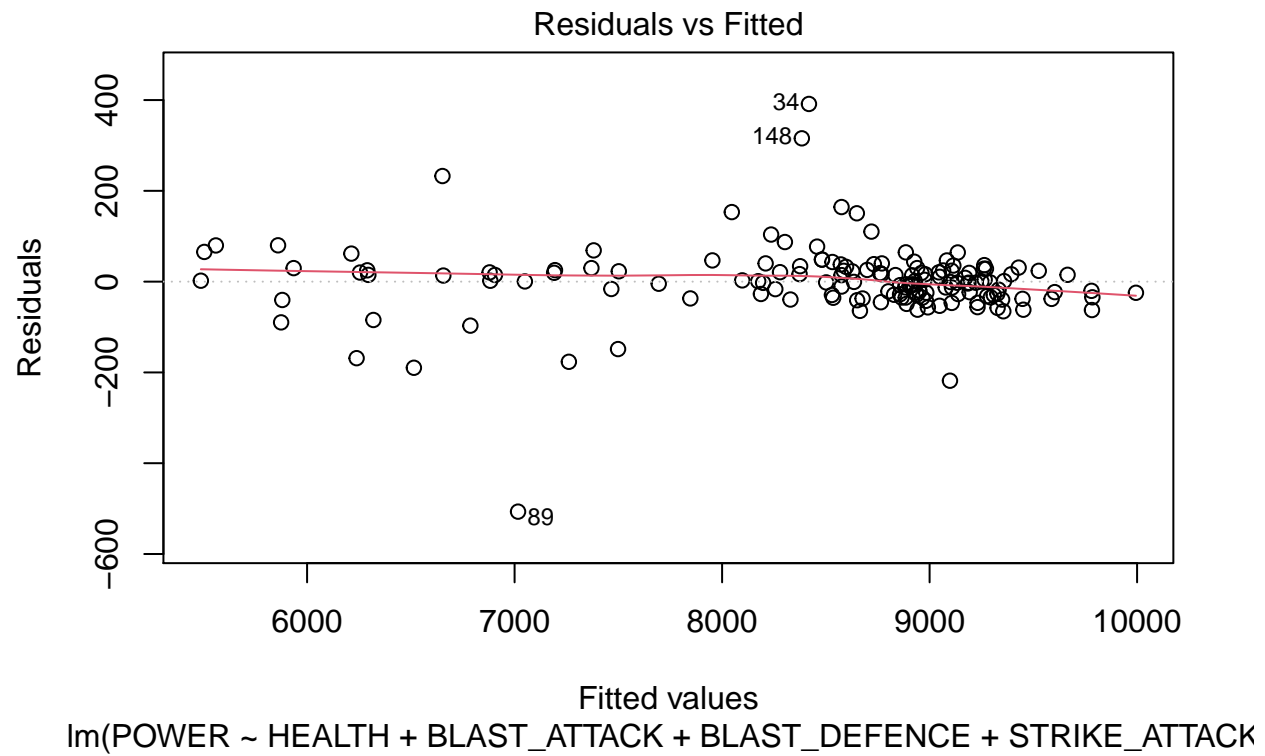
```

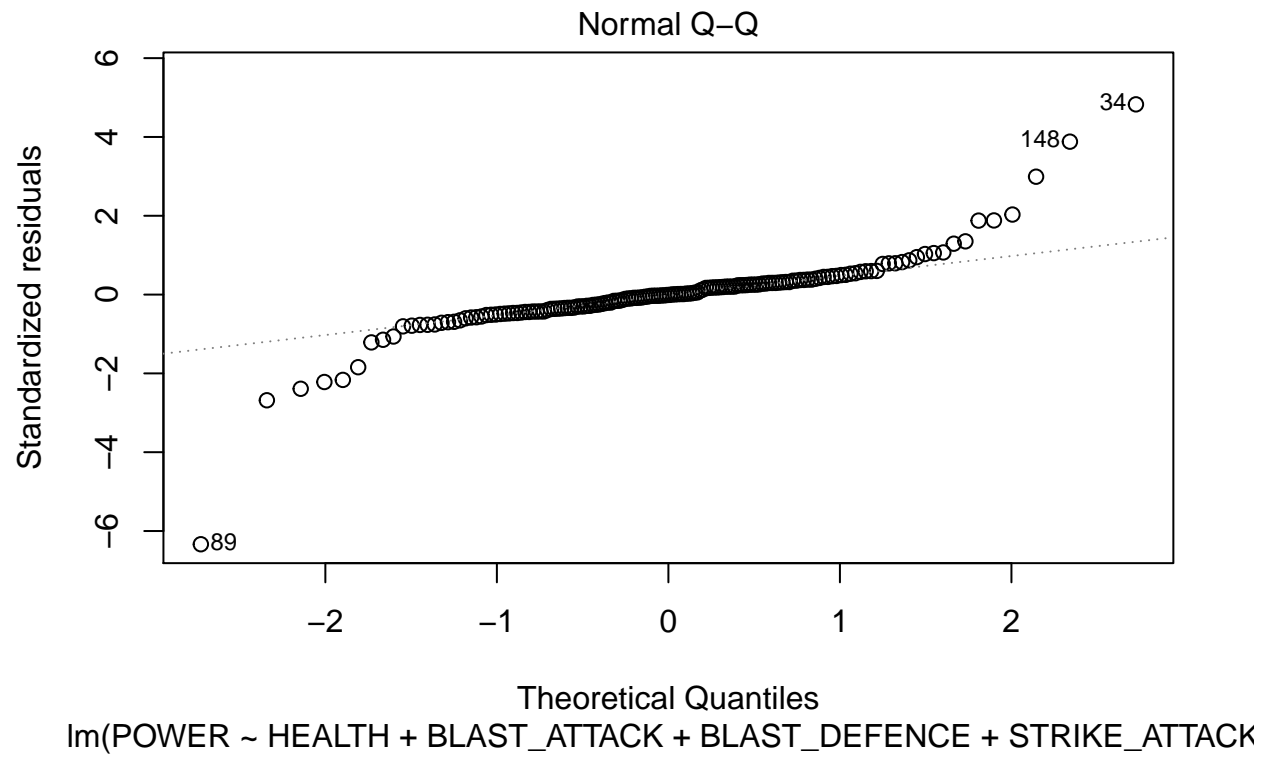
##          HEALTH  BLAST_ATTACK BLAST_DEFENCE STRIKE_ATTACK
##      4.278208      2.693735      3.324409      2.730526

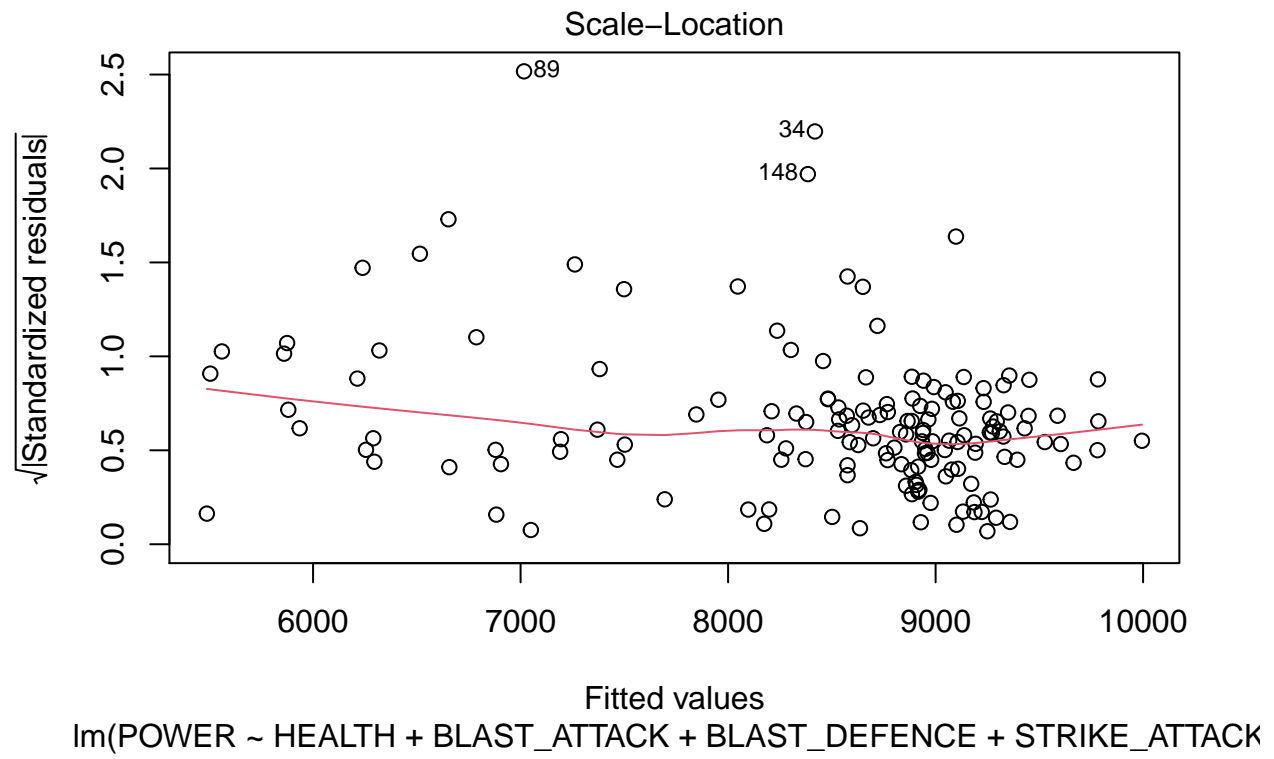
##
## Call:
## lm(formula = POWER ~ HEALTH + BLAST_ATTACK + BLAST_DEFENCE +
##     STRIKE_ATTACK)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -506.63  -29.31   -0.73   25.03  391.39
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -159.87119    66.72472  -2.396      0.0178 *
## HEALTH         0.07191     0.00564   12.750 <0.0000000000000002 ***
## BLAST_ATTACK   1.02116     0.03315   30.801 <0.0000000000000002 ***
## BLAST_DEFENCE  2.01688     0.08225   24.522 <0.0000000000000002 ***
## STRIKE_ATTACK  1.01305     0.03209   31.566 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.93 on 151 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9937
## F-statistic: 6069 on 4 and 151 DF, p-value: < 0.00000000000000022

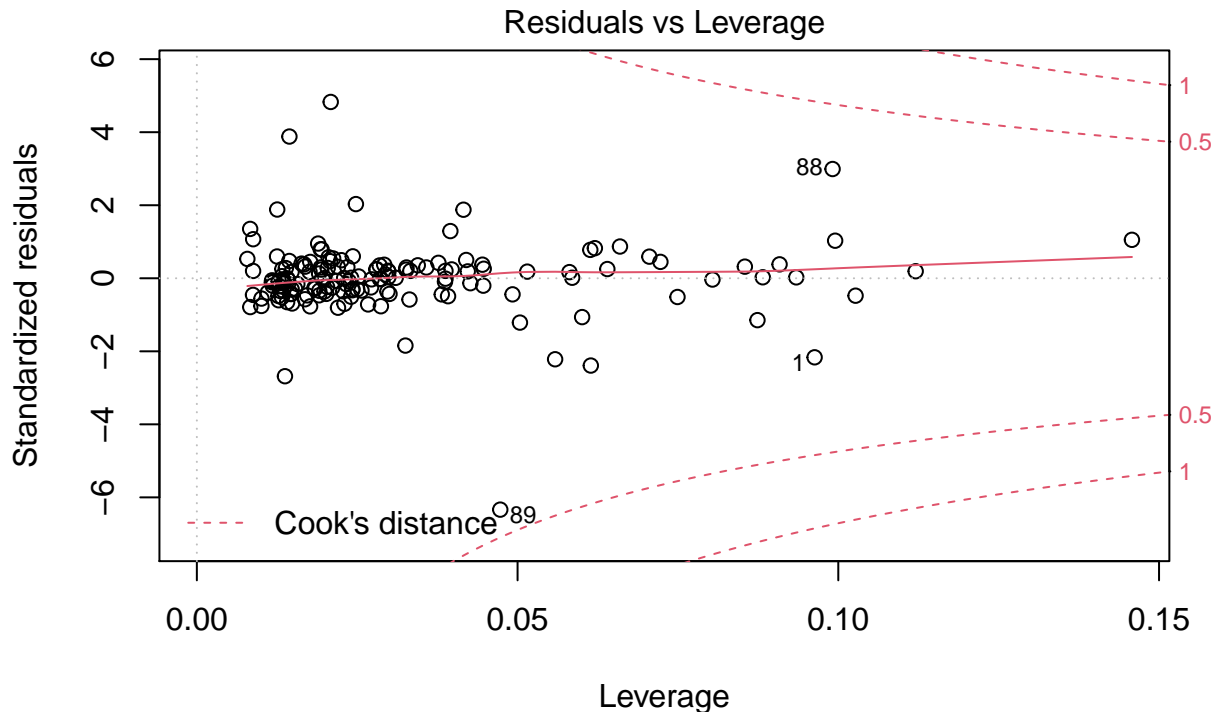
##              2.5 %      97.5 %
## (Intercept) -291.70581076 -28.03656583
## HEALTH       0.06076886   0.08305731
## BLAST_ATTACK  0.95565458   1.08666306
## BLAST_DEFENCE 1.85438118   2.17938551
## STRIKE_ATTACK 0.94964283   1.07646000

```







$\text{lm}(\text{POWER} \sim \text{HEALTH} + \text{BLAST_ATTACK} + \text{BLAST_DEFENCE} + \text{STRIKE_ATTACK})$

Para nuestro Data Set y nuestro modelo, recordemos que la variable dependiente es POWER. Después de probar con varias combinaciones de variables y aplicar las pruebas sugeridas para el modelo de regresión lineal múltiple, el modelo elegido fue el siguiente; A pesar de que sus residuales no distribuyan de forma normal, se intentó ajustar otra distribución a los residuales, no obstante, al realizar pruebas estadísticas para corroborar que los residuales seguían la distribución que se propuso, la estadística rechazaba que los residuales siguieran alguna distribución, por lo que podemos decir que los residuales no siguen una distribución.

```
m5 <- lm(formula = POWER ~ HEALTH + BLAST_ATTACK
          + BLAST_DEFENCE + STRIKE_ATTACK)
vif(m5)
```

```
##          HEALTH  BLAST_ATTACK BLAST_DEFENCE STRIKE_ATTACK
##          4.278208      2.693735      3.324409      2.730526
```

```
summary(m5)
```

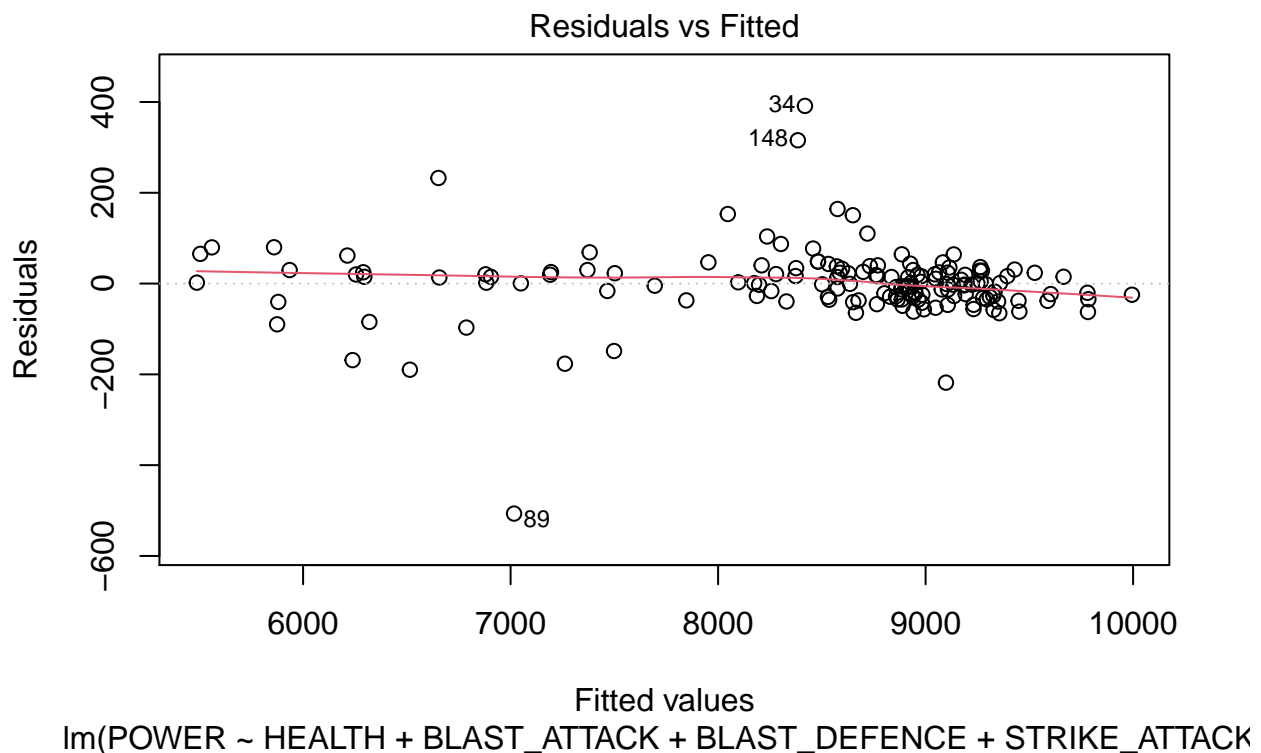
```
##
## Call:
## lm(formula = POWER ~ HEALTH + BLAST_ATTACK + BLAST_DEFENCE +
##     STRIKE_ATTACK)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -506.63  -29.31   -0.73   25.03  391.39
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  -159.87119   66.72472  -2.396    0.0178 *
```

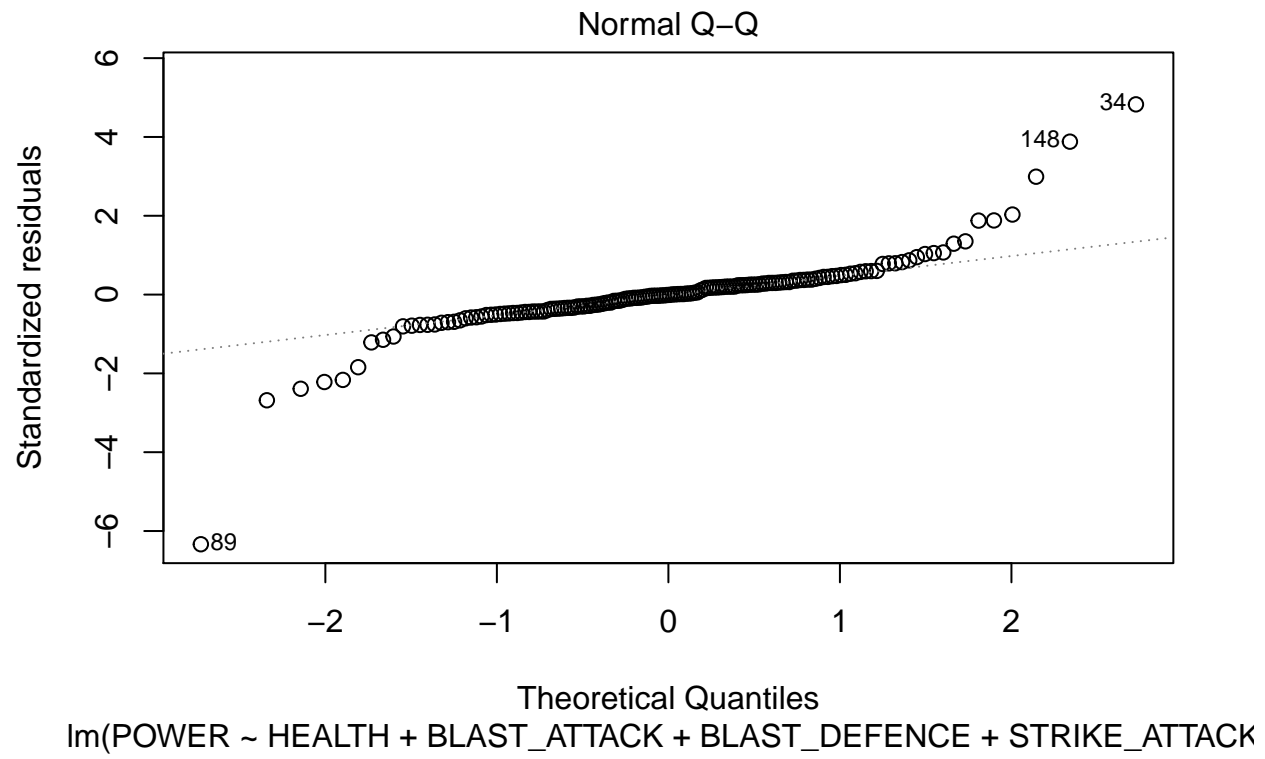
```
## HEALTH      0.07191    0.00564   12.750 <0.0000000000000002 ***
## BLAST_ATTACK 1.02116    0.03315   30.801 <0.0000000000000002 ***
## BLAST_DEFENCE 2.01688    0.08225   24.522 <0.0000000000000002 ***
## STRIKE_ATTACK 1.01305    0.03209   31.566 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.93 on 151 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9937
## F-statistic: 6069 on 4 and 151 DF,  p-value: < 0.00000000000000022
```

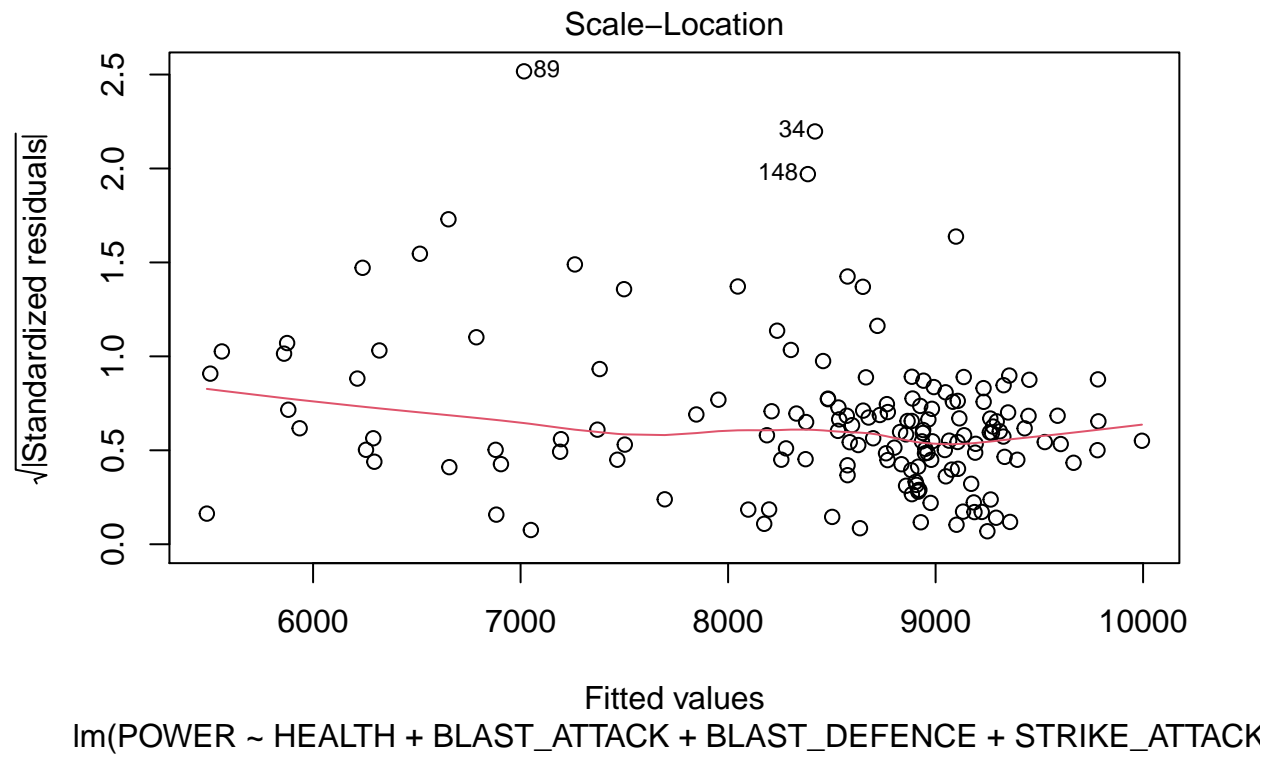
```
confint(m5)
```

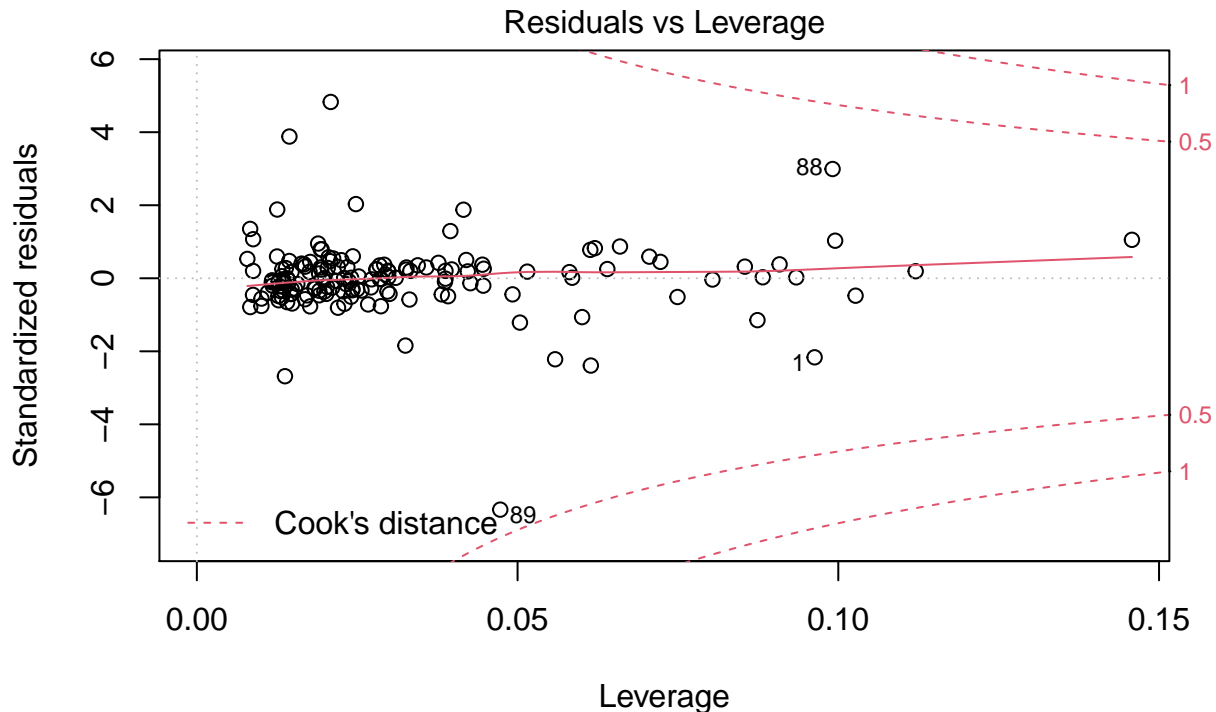
```
##              2.5 %      97.5 %
## (Intercept) -291.70581076 -28.03656583
## HEALTH      0.06076886   0.08305731
## BLAST_ATTACK 0.95565458   1.08666306
## BLAST_DEFENCE 1.85438118   2.17938551
## STRIKE_ATTACK 0.94964283   1.07646000
```

```
plot(m5)
```









$\text{lm}(\text{POWER} \sim \text{HEALTH} + \text{BLAST_ATTACK} + \text{BLAST_DEFENCE} + \text{STRIKE_ATTACK})$

Este modelo pasó las pruebas de varianza constante y no autocorrelación arrojando los siguientes resultados:

#varianza constante

`bptest(m5)`

```
##
## studentized Breusch-Pagan test
##
## data: m5
## BP = 4.8925, df = 4, p-value = 0.2985
```

#no autocorrelacion

`dwtest(m5)`

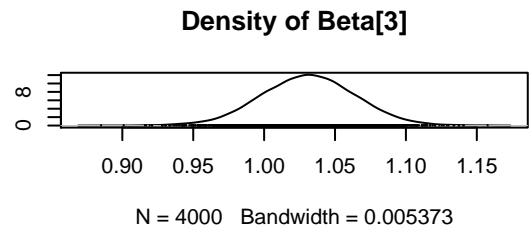
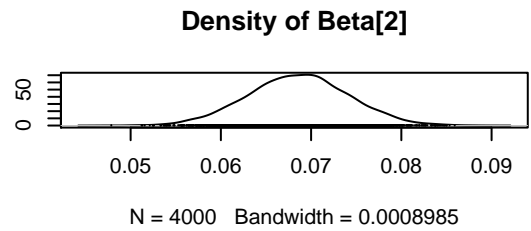
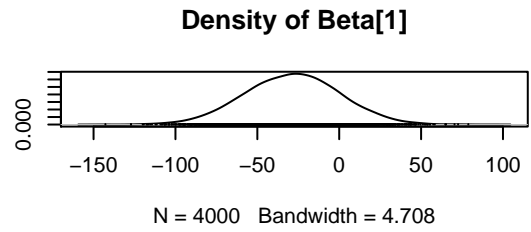
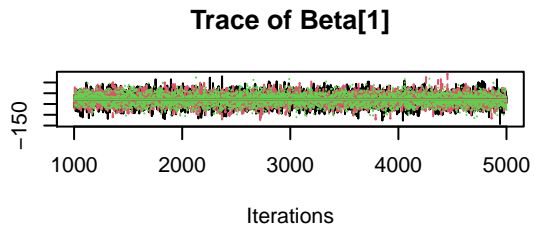
```
##
## Durbin-Watson test
##
## data: m5
## DW = 2.0283, p-value = 0.5679
## alternative hypothesis: true autocorrelation is greater than 0
```

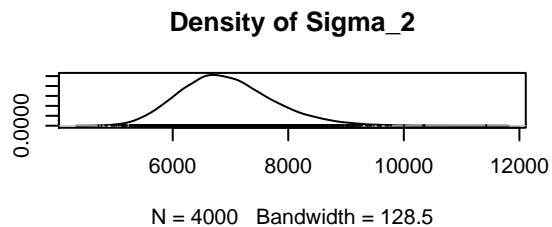
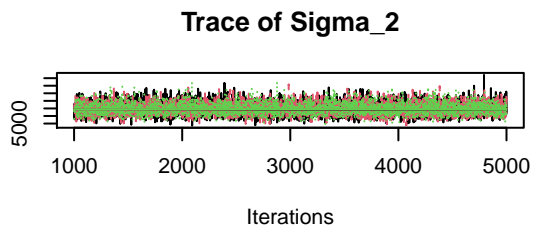
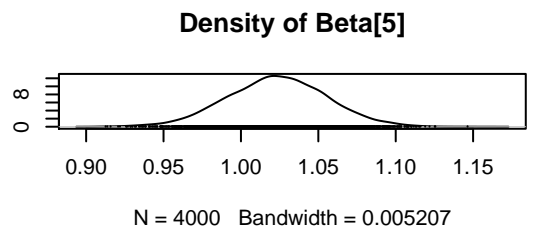
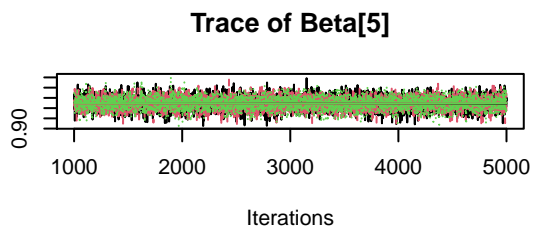
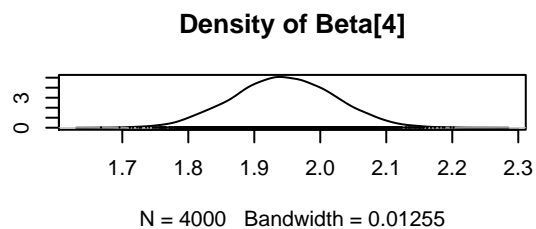
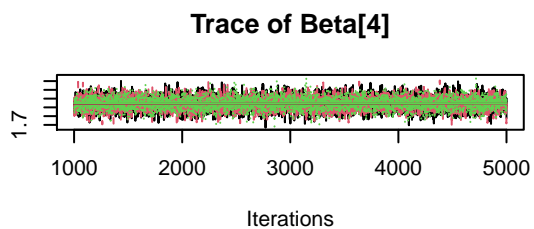
Modelo Bayesiano

```
## Compiling model graph
## Resolving undeclared variables
## Allocating nodes
## Graph information:
## Observed stochastic nodes: 156
## Unobserved stochastic nodes: 2
```

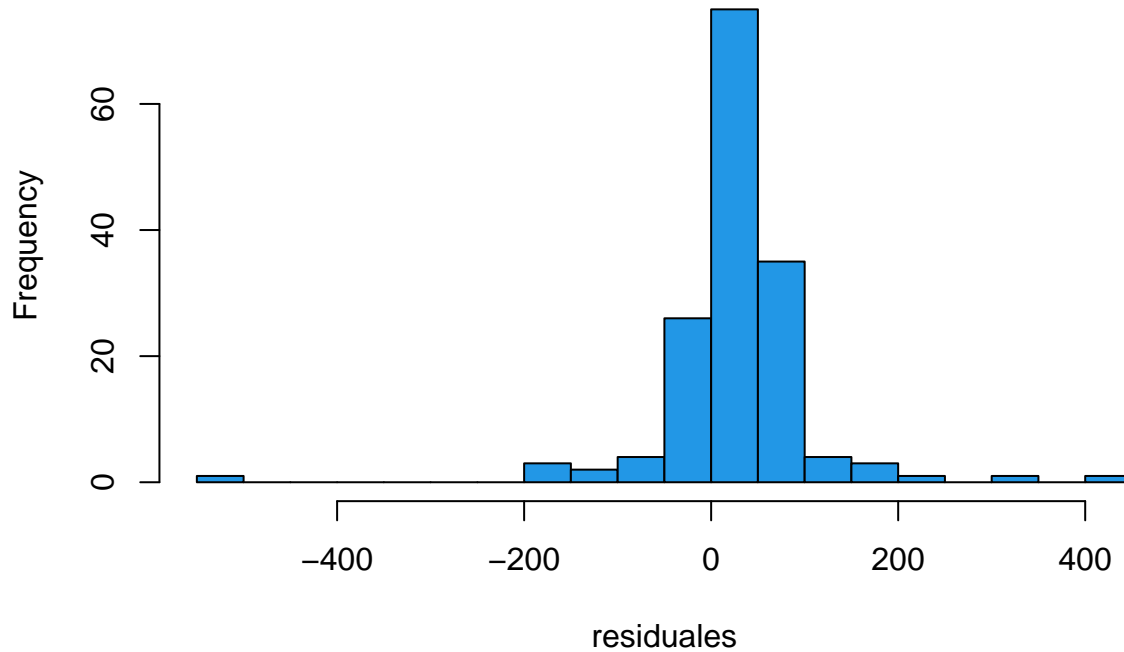


```
## Total graph size: 1371
##
## Initializing model
```





Residuales Bayesianos



```
##
## Jarque Bera Test
##
## data:  residuales
## X-squared = 1573.4, df = 2, p-value < 0.00000000000000022
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  residuales
## D = 0.19645, p-value = 0.0000000000000003469
```

Para las trazas se puede ver que todas convergen, a pesar de que nuestra varianza tiene valores muy grandes, podemos ver que también converge. Lo que hace que nuestra varianza tome intervalos tan grandes es la dimensión de los datos, pues las variables están representadas en miles y diez miles.

En relación con los residuales del modelo bayesiano se aplicaron 2 pruebas que también se rechaza la hipótesis de que los residuales distribuyan de manera normal.

Comparacion de los estimadores

Estimadores clásicos

Intercepto	HEALTH	BLAST ATTACK	BLAST DEFENCE	STRIKE ATTACK	VARIANZA
-159.87119	0.07191	1.02116	2.01688	1.01305	6712

Estimadores bayesianos

Intercepto	HEALTH	BLAST ATTACK	BLAST DEFENCE	STRIKE ATTACK	VARIANZA
-29.64673	0.06862	1.03118	1.94535	1.02303	6930.97264

Se puede observar que los estimadores, a excepción del intercepto, se parecen demasiado, incluso la varianzas también se encuentran relativamente cercanas, esto se debe a que se asignaron distribuciones poco informativas a las variables predictoras del modelo.