Facultad de Ciencias

Estadística Bayesiana

Proyecto

Profesora: Lizbeth Naranjo Albarrán

Ayudantes: Yadira Rivas Godoy

Edgar Gerardo Alarcón González

Equipo:

* Ramírez Montes Jonathan Natael 49 (?)
* Sánchez Romero Paulina Michelle 51

Contenido

[**Introducción.** 3](#_Toc62510951)

[**Propósito.** 3](#_Toc62510952)

[**Objetivo del documento.** 3](#_Toc62510953)

[**Entendimiento de los datos.** 3](#_Toc62510954)

[**Análisis descriptivo.** 5](#_Toc62510955)

[**Histogramas de datos.** 5](#_Toc62510956)

[**Diagrama de dispersión y correlación entre variables.** 6](#_Toc62510957)

[**Resúmenes estadísticos de las variables.** 6](#_Toc62510958)

[**Datos atípicos.** 7](#_Toc62510959)

[**Regresión lineal múltiple (análisis clásico).** 7](#_Toc62510960)

[**Coeficientes:** 8](#_Toc62510961)

[**Summary del modelo:** 8](#_Toc62510962)

[**Intervalo de los estimadores:** 8](#_Toc62510963)

[**Prueba de los supuestos del modelo.** 9](#_Toc62510964)

[**Normalidad.** 9](#_Toc62510965)

[**Varianza constante:** 10](#_Toc62510966)

[**Autocorrelación:** 10](#_Toc62510967)

[**Regresión lineal múltiple (Modelo Bayesiano).** 11](#_Toc62510968)

[**Trazas:** 11](#_Toc62510969)

[**Residuales** 12](#_Toc62510970)

[**Comparación de los estimadores Clásicos y Bayesianos.** 12](#_Toc62510971)

# **Introducción.**

Dragon Ball es uno de los animes[[1]](#footnote-1) más populares de la historia. Creado por **Akira Toriyama** en 1984 como manga[[2]](#footnote-2), 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 “Saiayjin” 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.**

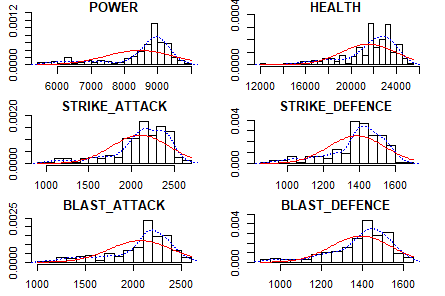
En este módulo se describirá mediante una tabla la información que fue obtenida vía la página web kaggle, el conjunto de datos estaba conformado por:

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.

|  |  |  |
| --- | --- | --- |
| **Nombre de Variable** | **% NA** | **Descripción** |
| "CHARACTER" | 0% | Denota el nombre de los personajes en el video juego.   * 156 valores distintos. |
| "CARD.NUMBER" | 0% | Identificador único para los personajes.   * 290 registros únicos. |
| "POWER" | 0% | Se reconoce como el nivel de poder (ki) que tienen los personajes. |
| "HEALTH" | 0% | Se reconoce como el nivel de salud que tienen los personajes. |
| "STRIKE.ATTACK" | 0% | Se reconoce como el nivel de un ataqué directo. |
| "STRIKE.DEFENCE" | 0% | Se reconoce como el nivel de defensa ante un ataque directo. |
| "BLAST.ATTACK" | 0% | Se reconoce como el nivel de un ataqué a larga distancia. |
| "BLAST.DEFENCE" | 0% | Se reconoce como el nivel de defensa ante un ataque a larga distancia. |

# **Análisis descriptivo.**

## **Histogramas de datos.**

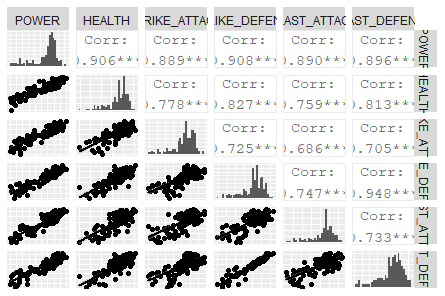


Con estas gráficas se puede observar el comportamiento de las diferentes variables. Se puede notar que la variable HEALTH es una de las que pareciera que no le podríamos ajustar algún tipo de distribución, pues al menos gráficamente, se puede destacar que hay valores atípicos. En relación con las otras 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.

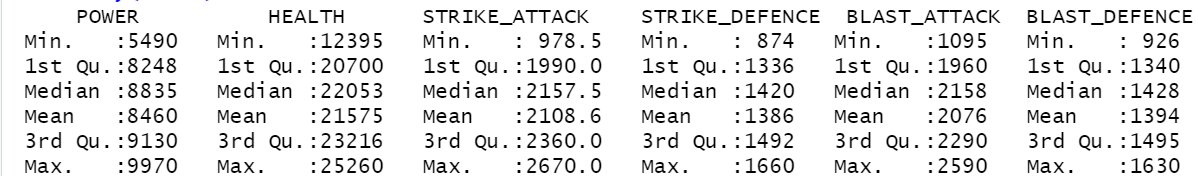
## **Diagrama de dispersión y correlación entre variables.**



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 de las variables.**

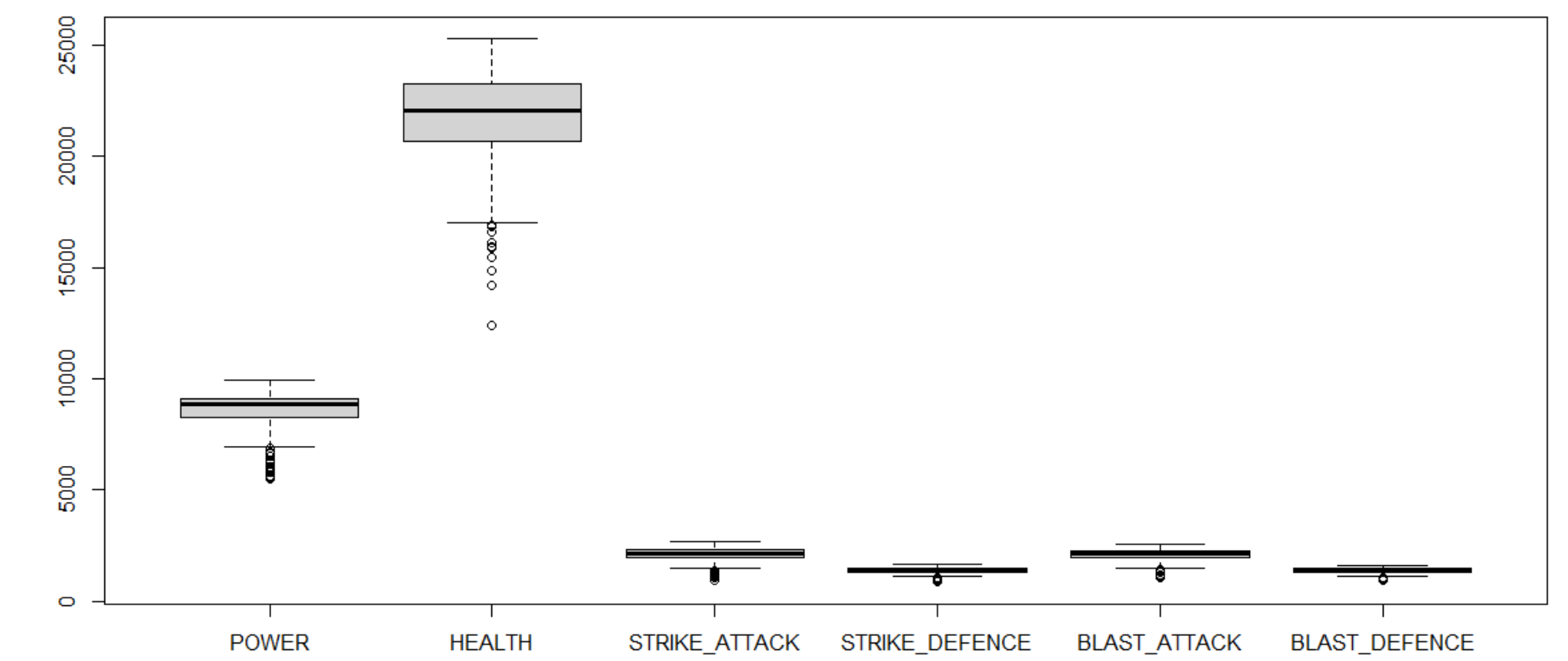


En cuanto a estos diferentes resúmenes, vemos que la variable que tiene valores más grandes es HEALTH, y que también es la que tiene un intervalo mucho más grande, pues varían en aproximadamente 13,000 unidades/puntos, lo que no pasa con STRIKE\_DEFENCE, por ejemplo, que sólo va de 874 a 1660, o BLAST\_DEFENCE 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.

## **Datos atípicos.**



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.

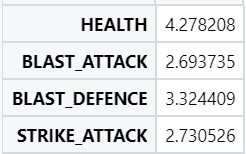
# **Regresión lineal múltiple (análisis clásico).**

Para nuestro Data Set y nuestro modelo, recordemos que la variable dependiente es POWER. Después de probar con varias combinaciones de variables, el modelo elegido fue el siguiente:

## **Coeficientes:**

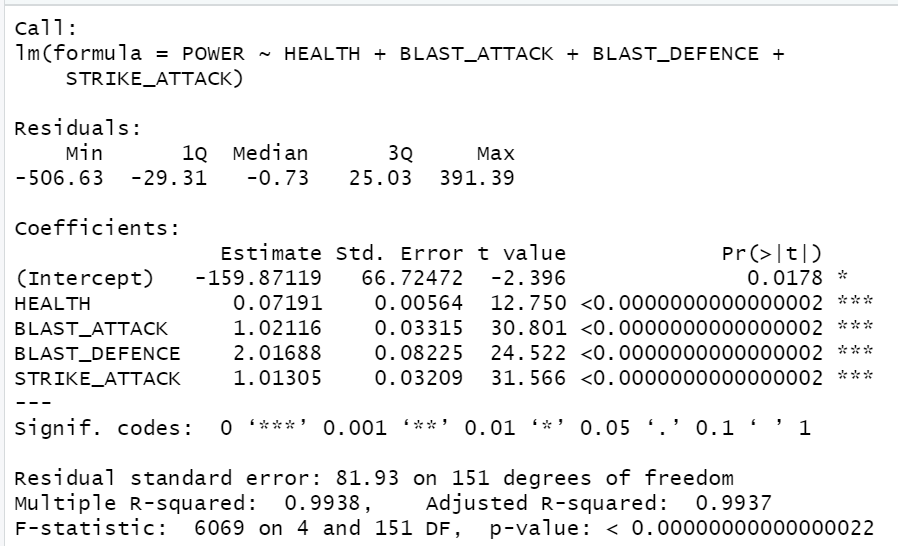
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Intercepto** | **HEALTH** | **BLAST\_ATTACK** | **BLAST\_DEFENCE** | **STRIKE\_ATTACK** |
| **-159.87119** | **0.07191** | **1.02116** | **2.01688** | **1.01305** |

También se realizó el factor de inflación de la varianza para cuantificar la multicolinealidad en entre las variables predictoras, el resultado para modelo fue el siguiente:

Como estos índices no son relativamente grandes podemos decir que no tendremos problemas de multicolinealidad, es decir, no tenemos redundancia entre las variables predictoras.

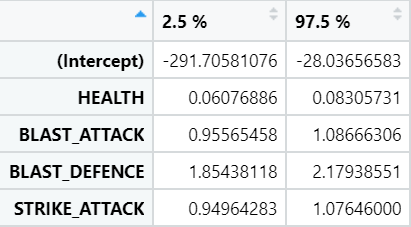
Después obtuvimos un resumen del modelo propuesto, teniendo el siguiente resultado:

## **Summary del modelo:**



Para este modelo podemos observar que explica a los datos con un 99.37% lo que nos indicaría un buen modelo, pues en principio no se muestra multicolinealidad en las variables predictoras.

## **Intervalo de los estimadores:**

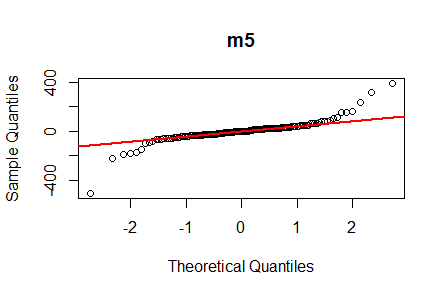
Podemos observar que los intervalos de confianza para los estimadores no atraviesan el cero, por lo que, ahora con esta información pareciera que las variables escogidas representan bien a los datos.

A continuación, realizaremos las pruebas al modelo para conocer si es un buen ajuste o no.

## **Prueba de los supuestos del modelo.**

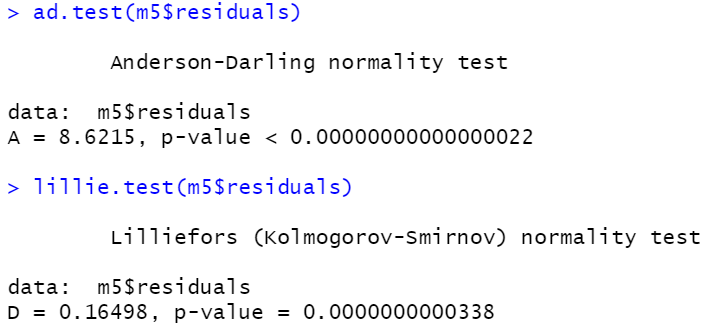
### **Normalidad.**

Se analizará el comportamiento de la distribución de los residuales.



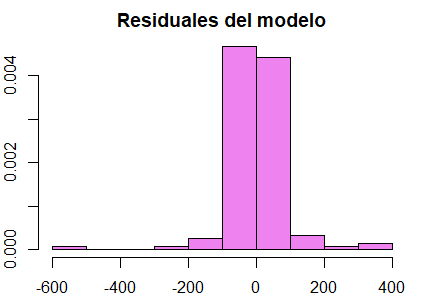
Podemos ver que gráficamente pareciera que los residuales no tendrán una distribución normal como se pide en uno de los supuestos, pues podemos ver que hay valores que se alejan mucho de la línea roja y estos podrían causar problemas en las pruebas de hipótesis.

Para probar estadísticamente que los residuales distribuyen de forma normal se aplicaron dos pruebas de normalidad Anderson-Darling y Kolmorov-Smirnov.

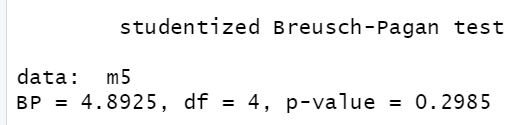


Se puede observar que en ambas pruebas se rechaza la hipótesis nula, por lo tanto, los residuales no distribuyen normal.

Ahora se quiso ajustar una distribución a los residuales, se probó con una distribución Cauchy y una T-student, ambas fueran rechazados con un p\_value muy cercano a cero.

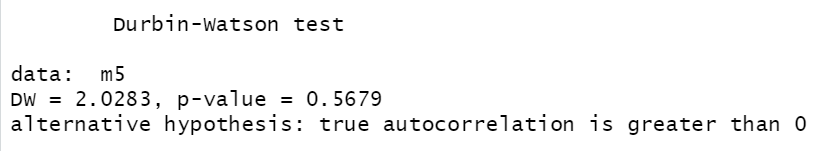


### **Varianza constante:**



Para probar que el modelo presenta una varianza constante se realizó la prueba Breusch-Pagan el resultado es que aceptamos la hipótesis nula, es decir, el modelo tiene varianza constante

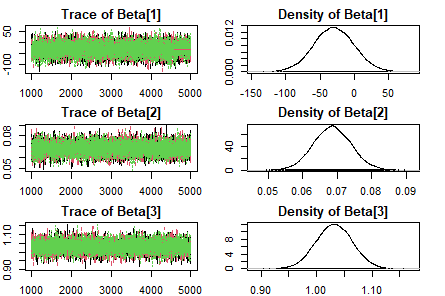
### **Autocorrelación:**

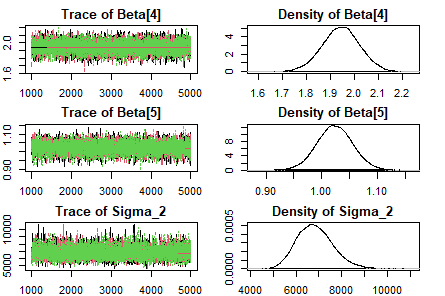


Por último, para el análisis clásico, se realizó una prueba de Durbin Watson, la cual nos indica que si el valor es cercano a 2 quiere decir que que no se presenta autocorrelación en los datos y se presenta un p\_value del 0.56, por lo tanto aceptamos la hipótesis nula, es decir el modelo no presenta problemas de autocorrelación.

## **Regresión lineal múltiple (Modelo Bayesiano).**

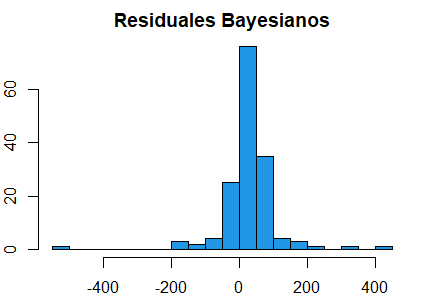
## **Trazas:**



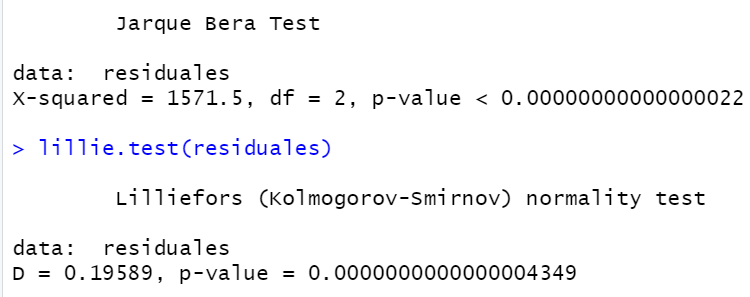


A pesar de que nuestra varianza tiene valores muy grandes, podemos ver que sí converge. Lo que hace que nuestra varianza tome intervalos tan grandes es la dimensión de los datos, pues las variables están representadas en miles y diez miles.

## **Residuales**



Se aplicaron 2 pruebas a los residuales del modelo Bayesiano y también se rechaza la hipótesis de que los residuales distribuyan de manera normal.



# **Comparación de los estimadores Clásicos y Bayesianos.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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

1. Género de animación de origen japonés que se caracteriza por un grafismo crudo y argumentos que frecuentemente tratan temas fantásticos o futuristas. [↑](#footnote-ref-1)
2. Es una palabra japonesa utilizada para denotar historietas. [↑](#footnote-ref-2)