Análisis estadístico sobre jugadores la MLB

Miguel Cordero
Universidad Simón Bolívar
Caracas, Venezuela
15-10326@usb.ve

Eduardo Gavazut
Universidad Simón Bolívar
Caracas, Venezuela
13-10524@usb.ve

Luis Riera
Universidad Simón Bolívar
Caracas, Venezuela
16-10976@usb.ve

8 de abril de 2022

RESUMEN: El deporte como actividad social del ser humano no es ajena a la ciencia, en particular, a la matemática y la estadística. En béisbol, por ejemplo, se recopila cada mínimo de información de todo lo que sucede durante el juego, aspectos como: tasa de bateo, carreras anotadas o ponches. Y es que estos datos permiten medir cuán bueno o acertado es el desempeño de cada jugador. Debido a la gran cantidad de información, que además se ha incrementado con los años, es necesario recurrir a la ciencia y a modelos computacionales de predicción que ofrezcan un punto de objetividad que permita a los equipos mejorar su competitividad. En este trabajo, se mostrará como realizar un análisis estadítico sobre una base de datos de jugadores de la Major League Baseball (MLB), relativo a la tasa de bateo, carreras anotadas, triples, dobles y ponches por veces al bate. De este análisis se destaca el comprobar cómo la tasa de ponches es mayor a la tasa de jonrones de un jugador y que es posible hallar una relación lineal entre las tasas de bateo y las tasas de carreras anotadas, de dobles y ponches por veces al bate. Esto permitirá predecir con un nivel 0.8589 de error cuadrático ajustado, cual será la tasa de carreras anotadas por jugador según su desempeño en el campo. Más aún, un análisis de varianza (ANOVA), permite demostrar que no hay mayor distinción entre entre jugadores con diferentes tasas de bateo.

Palabras clave: Proyecto, Estadística, R, RStudio, Baseball, MLB, Predicción, ANOVA

0.1. PLANTEAMIENTO DEL PROBLEMA

En el presente proyecto, el objetivo es tomar una base de datos (en este caso de diversas métricas que corresponden a jugadores de la MLB) y realizar distintos estudios sobre ella, los cuales serán:

1. Análisis descriptivo.
2. Intervalo de confianza ($97\,\%$) para la media de cada variable.
3. Probar (a nivel de 0.05) que el promedio de bateo es inferior a 0.300.
4. Estudiar si la tasa de ponches y de jonrones son iguales.
5. Prueba de bondad de ajuste para la tasa de bateo para determinar si tiene distribución normal.
6. Gráfico de dispersión y matriz de correlación para las variables.
7. Modelo de regresión lineal que prediga la tasa de bateo en función al resto de las variables.
8. Separar a la tasa de bateo en tres grupos: los que tienen menos de 0.200, los que tienen entre 0.200 y 0.300, y los que tienen más de 0.300, y realización de un análisis de varianza para estudiar si los promedios de tasas de las otras variables son iguales.

0.2. Descripción de la base de datos

La base de datos a estudiar cuenta con 45 observaciones de 6 variables, las cuales son:

1. X1 = tasa de bateo, calculada como hits/veces al bate. Entiéndase la conexión efectuada por el bateador que coloca la pelota dentro del terreno de juego, permitiéndole alcanzar al menos una base, sin que se produzca un error de defensa del equipo contrario o algún otro jugador sea declarado como fuera de juego.
2. X2 = carreras anotadas/veces al bate. Entiéndase carrera por anotación, y se logra al recorrer un corredor la totalidad de las bases volviendo al home, bien de manera continua (por medio de un jonrón) o de forma alternada consecutiva antes de que se realicen 3 outs.
3. X3 = dobles/veces al bate. Entiéndase por doble como un hit en el que el bateador logra llegar a segunda base sin ser puesto out y sin que haya error alguno de la defensiva.
4. X4 = triples/veces al bate. Entiéndase por triple como un hit en el que el bateador logra llegar satisfactoriamente a tercera base, sin que ocurra ningún error por parte de la defensiva.
5. X5 = jonrones/veces al bate. Un jonrón se da cuando el bateador hace contacto con la pelota de una manera que le permita recorrer las bases y anotar una carrera (junto con todos los corredores en base) en la misma jugada, sin que se registre ningún out ni error de la defensa.
6. X6 = ponches/veces al bate. Por último, un ponche es la acción de retirar a un bateador con una cuenta de 3 strikes, al que la recibe se le suele llamar ponchao o ponchado.

De esta forma, vemos que cada una de las variables miden números bastante relevantes para cada jugador. Como cada una de estas estadísticas pueden ocurrir una sola vez mientras se está al bate, cada una será un numéro entre el 0 y el 1.

0.3. Metodología

Para la realización de esta investigación se hará uso del sofware estadístico `R` en el entorno de desarrollo intergrado (IDE) `RStudio`. En este se iniciará por una descripción de los datos y variables almacenadas en el archivo fuente *Baseball.xlsx*, tales como: mínimo, media, cuantiles y desviación estándar. Para la media de las variables se obtendrá un intervalo de confianza del $95\,\%$. Como se desea estudiar la relación de la tasa de bateo respecto al resto de las variables, se buscará determinar la mejor distribución de probabilidad que se ajuste a esta variable. Finalmente, se estudiará la eficiencia del mejor modelo lineal de predicción que se ajuste a los datos y permita establecer si en efecto existe tal relación entre las variables y las implicaciones que tendría en las estrategias para futuros juegos de béisbol.

0.4. Análisis de los datos

Para la realización de este proyecto se contó con una archivo de excel con la información de algunos jugadores de la Major League béisbol o MLB, el cual se almacenó en una variable llamada `Baseball`. De esta archivo podemos realizar el siguiente análisis de datos.

0.4.1 ¿Qué clase es la base de datos?

Con el comando `class`, se pudo determinar el tipo de base de datos utilizada o lo que es equivalente, la clase de la variable `Baseball`. El resultado que se obtuvo indica que es del tipo `tbl_df`, que es una subclase de la clase `data.frame`. `tbl_df` cumple con tener propiedades diferentes por defecto y se suele referir a ellas como `tibble`. Es una clase eficiente para trabajar con bases de datos grandes y su visualización.

0.4.2 Variables en la base de datos

Si se desea saber que tipo de variables están almacenadas en la base de datos, se puede utilizar el comando `str`. Esta función nos indica que se cuentan con 6 variables denominadas `X1,X2,X3,X4,X5,X6`, y distribuidas de tal manera que representan la columnas de la base de datos. Cada una de estas variables tienen 45 valores de tipo `double` o número decimal, que representan las 45 observaciones aleatorias (una por fila) realizadas a jugadores de la MLB.

0.4.3 Estadísticos

Para obtener los estadísticos de las seis (6) variables de esta base de datos, se inicia por guardar las 45 observaciones en un vector que represente a cada variable.

Con los datos vectorizados se pueden aplicar las siguientes funciones: `mean` que permite obtener la media de los datos, `median` para obtener la mediana, `quantile` para retornar los cuantiles al $0{,}25\,\%$, $0{,}50\,\%$ y $0{,}75\,\%$ de cada variable, `min` para el valor mínimo, `max` para el valor máximo, `var` para la varianza, `sd` que es para la desviación estándar, `IQR` es para el rango intercuartil y finalmente, el coeficiente de variación obtenido como `stad/media`.

	Mínimo	25 %	Media	Mediana (50 %)	75 %	Máximo	RIC	Varianza	Desv. Estándar	Coef. Variación
X1	0.188	0.248	0.2805	0.290	0.308	0.367	0.060	0.0019	0.0440	0.1569
X2	0.064	0.119	0.1509	0.150	0.189	0.259	0.070	0.0018	0.0420	0.2784
X3	0.025	0.039	0.0464	0.045	0.053	0.068	0.014	0.0001	0.0105	0.2255
X4	0.001	0.007	0.0113	0.009	0.016	0.030	0.009	0.0000	0.0070	0.6165
X5	0.000	0.009	0.0243	0.013	0.039	0.085	0.030	0.0005	0.0223	0.9173
X6	0.000	0.062	0.1043	0.095	0.138	0.264	0.076	0.0040	0.0631	0.6044

Tabla 1: Resumen Estadístico de las variables

Los resultados pueden ser apreciados en la tabla 1. De estos resultados hay varios puntos que podemos destacar. La varianza de los datos es muy baja, indicativo que entre los datos hay pocos valores atípicos o muy dispersos, lo que se refleja en valores más cercanos a la media. La misma interpretación se puede extender a la desviación estándar pues esta es la raíz cuadrada de la varianza.

Una consecuencia de la baja varianza es que la media y la mediana son valores muy cercanos. Esto es particularmente útil al analizar el valor del `RIC`, que toma como medida central la mediana de los datos. Es decir, nos indica donde se encuentra el $50\,\%$ de los datos, cuánto mas bajo es el valor del RIC menos dispersos están los datos.

0.4.4 Diagramas e histograma de los datos por cada variable

De la figura 4, podemos establecer: para la variable `X1`, que los valores máximos de los datos se obtienen luego de la media, pero el mayor volumen de ellos se encuentra antes como bien se observa en el diagrama de caja que permite confirmar, además, la ausencia de datos atípicos. Para la variable `X2`, se puede comprobar que ver simetría de los datos que se infería de la tabla 1, particularmente respecto al valor $0{,}15$ que coincide a su vez con la media de los datos. El diagrama de caja permite confirmar la ausencia de los valores atípicos.

Por su parte, para la variable `X3` y `X4`, Vemos que en general, ambos diagramas de caja son bastante parecidos, con la única diferencia siendo que el de triples está $0{,}03$ puntos corrido hacia arriba y los datos desde el primer cuartil hasta la mediana están muchos más dispersos. Otra diferencia es que el diagrama de cajas para los triples no cuenta con datos atípicos, en cambio los dobles si, que corresponde a $0{,}3$. Todo esto hace que el diagrama de los triples sea casi simétrico, y el de los dobles sea más chato entre el valor mínimo y la mediana, en comparación con lo que tenemos entre la mediana y el máximo valor.

De la gráfica para la variable `X5` podemos ver como a medida que nos vamos acercando a 1, la frecuencia de jonrones decae rapidamente, mientras que al incio es muy alta. De la gráfica para la variable `X6` podemos ver que la mayoría de los jugadores se ponchan menos de un $15\,\%$ de las veces que estan al bate.

0.5. Intervalo de confianza para la media de las variables

Con el uso de la función `t.test()` se puede encontrar el intervalo de confianza con una significancia de $0{,}03$ o $97\,\%$ de confianza para las variables estudiadas. Los resultados de aplicar esta función, se pueden visualizar en la tabla 2.

	Limite inferior	Promedio	Limite Superior
Tasa de bateo	0.2658	0.2805	0.2952
Carreras anotadas	0.1368	0.1509	0.1649
Dobles	0.0429	0.0464	0.0498
Triples	0.0090	0.0113	0.0136
Jonrones	0.0168	0.0243	0.0317
Ponches	0.0833	0.1043	0.1254

Tabla 2: Intervalos de confianza para las medias de las variables

Se pueden visualizar mejor estos intervalos de confianza, en la figura 5 de los anexos.

Note que en general, los intervalos de confianza más estrechos son los de dobles y triples, lo que nos indica que en general, con una probabilidad del $97\,\%$, podemos asegurar que los jugadores de la MLB tendrán un promedio de triples y dobles que puede ser estimado con bastante certeza, pero vemos que las carreras anotadas, los ponches y la tasa de bateo tienen un intervalo de confianza mucho más grande, por lo que no podemos asegurar que el promedio será estimado de forma tan certera.

0.6. Promedio de bateo

Con lo obtenido en los intervalos de confianza del apartado anterior se tiene que la tasa de bateo toma valores por debajo de $0{,}300$. Para corroborar este resultado, se realizará un prueba de hipótesis con un nivel de significancia de $\alpha = 0{,}05$.

Entonces, como hipótesis nula H_0 y como hipótesis altenativa H_a tenemos: vamos a suponer que la media de bateo, $H_0 : \mu_{\text{bateo}} \leq 0{,}3 \quad \text{vs} \quad H_\alpha : \mu_{\text{bateo}} > 0{,}3$.

Si suponemos que los datos presentan una distribución normal, podemos aplicar el comando `t.test` de `R`, que permite realizar pruebas de hipótesis sobre las medias de los datos cuando se trabaja con una sola variable.

Con esta función, se obtuvo que el valor para el estadístico t es $-23{,}811$, con 44 grados libertad. Como el $p - valor$ es bastante alto, de hecho es igual $0,9976$ (que representa un $99{,}76\,\%$), se cumple que $\alpha = 0{,}05 < 99{,}76$ y por lo tanto la hipótesis alternativa se rechaza, mas aún, se rechaza para todo nivel de significancia porque se necesita un valor para α más alto que el $p - valor$ para rechazar la hipótesis nula.

Se afirma entonces, con total seguridad, que la tasa de bateo es inferior a $0{,}300$, tal como se podía con el intervalo de confianza.

0.7. Comparación entre las tasas de ponches y las de jonrones

Ahora, deseamos comparar las tasas de ponches y de jonrones para determinar si o no parecidas. Como no tenemos conocimiento acerca de las varianzas poblacionales, usaremos el test de Welch tal y como es explicado en Heumann, Schomaker (2017) para comparar las medias. En este caso, haremos una prueba de hipótesis, donde tomaremos como hipótesis nula, H_0 e hipótesis alternativa H_a las dadas por: $H_0 : \mu_{\text{jonrones}} - \mu_{\text{ponches}} = 0 \quad$ vs. $\quad H_a : \mu_{\text{jonrones}} - \mu_{\text{ponches}} \neq 0$

Es decir, queremos determinar si las tasas de jonrones y ponches son distintas. Ahora, con apoyo del comando anterior `t.test()`, pero esta vez para comparar dos variables, podremos determinar cuál de estas hipótesis es aceptada.

Como resultado se obtuvo que el $\text{p-valor} = 1{,}112\times10^{-8}$, que es extremadamente pequeño, mucho más que el nivel de significancia $\alpha = 0{,}01$ que es razonable utilizar para nuestra prueba de hipótesis. Adicionalmente, el intervalo de confianza que se obtuvo fue de $(-0{,}1068, -0{,}0593)$ que no incluye el cero. Otra cosa que podemos hacer es evaluar el estadístico de prueba con el comando `qt()` (vemos por lo anterior que $dt = 55$ y $\alpha = 0{,}05$).

Como $t = -8$, vemos que el estadístico cae en la región de rechazo (porque es de cola doble).

Para cualquiera de estos casos, podemos concluir que la hipótesis nula se rechaza, es decir que hay suficiente evidencia para creer que $\mu_{\text{jonrones}} - \mu_{\text{ponches}} \neq 0$. Y además, como el intervalo de confianza es negativo, concluimos que $\mu_{\text{ponches}} > \mu_{\text{jonrones}}$ con un nivel de confianza del $95\,\%$, como se podía apreciar de la figura 5

0.8. Prueba de bondad de ajuste para la distribución de X1

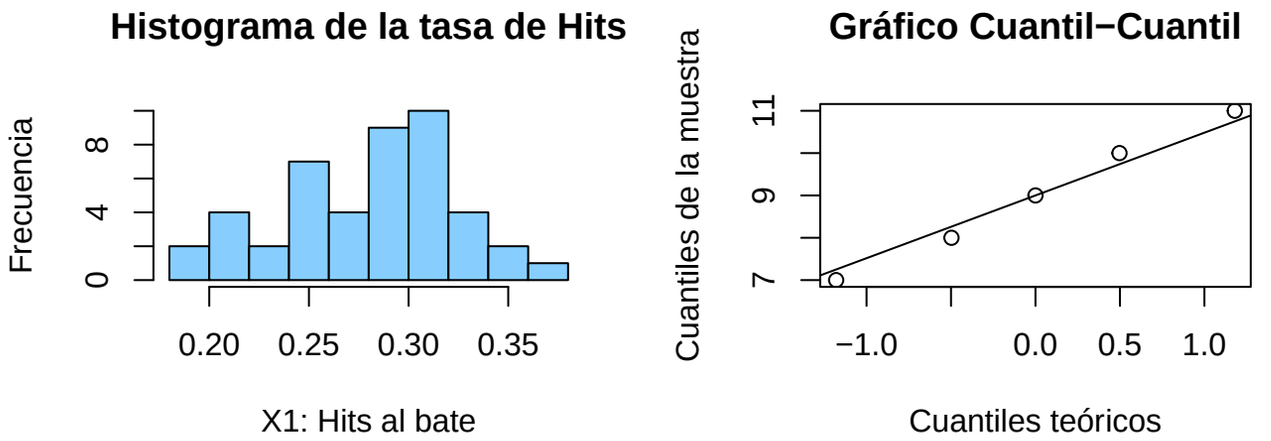

Figura 1: Histograma y Gráfico cuantil-cuantil de la variable X1

Para continuar con el análisis a un nivel más profundo, resulta conveniente determinar si los datos en la variable `X1`, sobre la tasa de bateos, sigue una distribución normal.

Como se señaló en la figura 4, del histograma para la variable, obtenemos entonces que si subdividimos en intervalos de longitud $0{,}02$, las frecuencias son como las descritas en la tabla 3.

Ahora agruparemos los datos en categorías de frecuencia mayor o igual a 5 (para poder aplicar el método de bondad de ajuste) tal como puede apreciarse en la tabla 4.

Intervalo	0,18- 0,20		0,22- 0,24		0,26- 0,28		0,30-0,32		0,34-0,36		0,38
Frecuencia	2	4	2	4	7	9	10	4	2	1	

Tabla 3: Tabla de clases y frecuencias

Clases	[0,18	0,24)	[0,24	0,28)	[0,28	0,30)	[0,30	0,32)	[0,32	0,38)
Frecuencia	8		11		9		10		7	

Tabla 4: Nueva agrupación en clases con frecuencia mayor o igual a 5

Con la gráfica cuantil-cuantil de la figura 1, podemos ver que esta agrupación se ajusta bien a un distribución normal (representada por la recta).

Vamos a proceder a realizar una prueba χ^2, que es una prueba de hipótesis que compara la distribución observada de los datos con la distribución esperada de los datos. Para este tipo de pruebas, el estadístico de χ^2 cuantífica que tanto varía la distribución respecto a la distribución hipotética. La hipótesis nula H_0 y la hipótesis altenativa H_a vienen dadas por: H_0 : Los datos siguen una distribución normal H_a : Los datos no siguen una distribución normal

Como estadístico χ^2 tenemos: $X^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)}$ con $k = 5$ el número de clases o categorías, n_i las frecuencias de cada categoría, $E(n_i) = n * p_i$ el valor esparado con n el número total de datos y p_i la probabilidad de cada clase n_i.

Para calcular las probabilidades p_i se obtuvo la media y la desviación estándar de los datos agrupados como $\bar{x} = 0{,}2822$ y $\sigma = 0{,}045$, respectivamente. Con $\bar{x}$ y σ se obtuvieron las siguientes probabilidades para cada clase: $p_1 = 0{,}172,\ p_2 = 0{,}3082,\ p_3 = 0{,}1747,\ p_4 = 0{,}1466,\ p_5 = 0{,}1986$.

Sustituyendo los datos en el estadístico tenemos que: $\chi^2 = 2{,}9421$, y el $p - valor$ viende dado por $1 - P(\chi^2 < 2{,}9421) = 0{,}2297$. El $p - valor$ es bastante alto por lo que la hipótesis nula no se rechaza para ningún nivel de significancia. Por tanto los datos siguen una distribución normal con media $0{,}2822$ y deviación estándar $0{,}045$.

0.9. Gráfico de dispersión y matrix de correlación

Es ahora, de nuestro interés estudiar la relación entre las variables de la base de datos. Esto lo podemos observar en la figura 2. Note que las gráficas de dispersión de la mitad inferior de la figura 2, se puede apreciar que para carreras anotadas, dobles y triples tenemos algo que se asemeja a una relación lineal positiva, mientras que para los ponches, estos disminuyen a medida que la tasa de bateo aumenta. La única variable que no parece tener ninguna relación clara con la tasa de bateo es la tasa de jonrones, por lo que es una variable que probablemente no nos ofrezca mayor información si queremos establecer un modelo lineal que relacione a las variables.

Por otro lado, con la parte superior de la figura 2 se tienen los coeficientes de correlación por pares de variables. Estos coeficientes nos indican que, efectivamente, para las carreras, dobles y triples, tenemos una correlación positiva (siendo las carreras la que tiene mayor correlación, y los triples la menor). Además, para los ponches tenemos una correlación negativa bastante significativa, y entre todas las variables, los jonrones tienen la menor correlación.

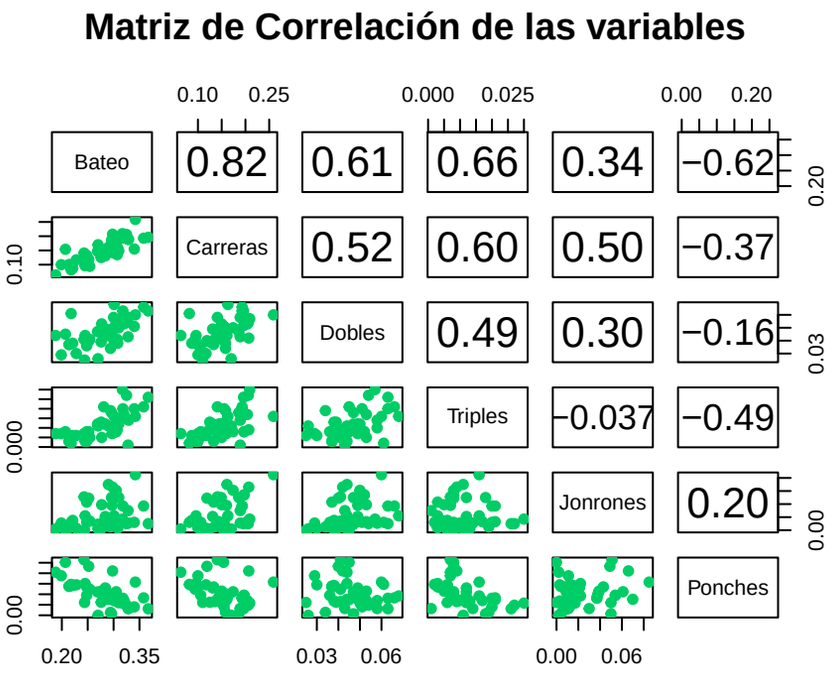

Figura 2: Matrix de correlación y dispersión de las variables

0.10. Muestreo 80 %-20 %

Como por la figura 2, parece existir una relación lineal entre las variables, particularmente vamos a estar interesados en ver como se relaciona cada campo información (carreras, dobles, etc.) con la variable `X1` que es la tasa de bateo. Con `R` tenemos la posibilidad de obtener un modelo de regresión lineal con la función `lm`.

Pero para asegurarnos que el modelo sea el más adecuado, priemero necesitamos extraer una muestra que permita que entrenar al modelo de predicción, y con los datos restantes probar que tan eficiente es el modelo. Con este objetivo, se dividen los datos en un $80\,\%$ para el entrenamiento y en un $20\,\%$ para las pruebas.

Como la base base de datos consta de 45 observaciones por variable, el $80\,\%$ representa tomar una muestra aleatoria de 36 observaciones, por los que el $20\,\%$ restante serán las 9 no tomadas en la muestra. Vale la pena resaltar que se habla de observaciones, o las filas de la base de datos y no de las entradas particulares de cada variable porque se busca estudiar la relación por jugador, de su tasa de bateo, respecto a su tasa de carreras, dobles, triples, jonrones y ponches. En otras palabras, las filas son independientes entre sí y por eso se pueden tomar muestras al azar, pero las columnas no lo son por ser datos relativos a un jugador en particular.

0.11. Modelo de regresión lineal para la variable X1

Ahora, teniendo seleccionado nuestros datos, podemos pasar a realizar el modelo.

La mejor manera de realizar un modelo de regresión lineal es seguir el método de **regresión paso a paso**, de esta manera determinar cuáles variables son significativas o no al tomar en cuenta la tasa de bateo.

Ahora, pasemos a realizar el modelo lineal utilizando el comando `lm()` de R. Se desarrolla primero el modelo dado por $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$. Suponiendo que $E(\epsilon) = 0$, buscamos estimar los parámetro β_i para los cuales $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$.

Con el comando `lm()` se obtuvo que el único valor no significativo (y de hecho el p-valor más alto) fue la tasa de triples, seguido de la tasa de jonrones que era significativa a nivel $0{,}05$.

De esta forma, realicemos de nuevo el modelo pero sin la variable $X4$ correspondiente a los triples. Es decir, el modelo a estimar es: $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5$.

Con `R` se obtuvo en esta prueba, que la tasa de jonrones es la variables con p-valor mas alto, con $0{,}0611$. A pesar, de ser significativa a nivel de $0{,}1$ procedemos a realizar una nueva prueba, esta vez sin la tasa de jonrones.

EL nuevo modelo, consiste en estimar $E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5$. Ahora, todas nuestras variables son bastante significativas, por lo que sus p-valores son bastante pequeños, significativos a nivel $0{,}001$. Los valores estimados fueron: $\beta_0 = 0{,}1630$, $\beta_1 = 0{,}5192$, $\beta_2 = 1{,}3650$ y $\beta_5 = -0{,}2451$.

Como medida del error, tenemos el R^2_{ajus}, con valor $0{,}8489$, indicando que hay un buen ajuste de los datos al modelo.

Además, tenemos que:

1. Para los estimadores, los dobles es el mayor de todos, y este nos indica que por cada aumento del $1\,\%$ en la tasa de dobles, hay un aumento correlacionado del $136\,\%$ en la tasa de bateo. Es interesante ver que este estimador es muchísimo mayor que el de las carreras.
2. La varianza es estimada como $\hat{\sigma}^2 = 0{,}01705^2$.
3. Para el error estándar (Std. Error), podemos construir los intervalos de confianza para las variables. Primero, tenemos que $t_{32,0{,}975} = 2{,}0369$: $I_{carreras} = 0{,}5192 \pm 2{,}0369 * 0{,}0868 = (0{,}3424, 0{,}6960)$. $I_{dobles} = 1{,}3650 \pm 2{,}0369 * 0{,}3471 = (0{,}6580, 2{,}0720)$; $I_{ponches} = -0{,}2451 \pm 2{,}0369 * 0{,}0460 = (-0{,}3388, -0{,}1514)$

   Como ninguno de estos intervalos incluye el 0, se puede concluir que efectivamente hay una relación existente entre estas variables seleccionadas y la tasa de bateo.

Ahora, veamos que efectivamente se cumple con las característica de un buen modelo apoyándonos en las gráficas de la figura 3.

- Cuando vemos la gráfica de "Residuos vs Ajustados", nos damos cuenta de que la línea azúl es bastante horizontal, y esta además está centrada alrededor del cero, es decir que podemos asumir que no hay independencia entre las variables y la tasa de bateo.
- Al ver el gráfico "Normal Cuantil-Cuantil", vemos que todos los valores están bastante cercanos a la recta, lo que nos confirma la normalidad.
- En "Escala-Localización" no vemos ningún patrón, lo que nos indica que los valores presentan homocedasticidad.
- Y por último, en "Residuos vs Apalancamiento", no hay ningún valor que esté fuera de las líneas rayadas, por lo que no parece haber valores que generen apalancamiento.

En conclusión, podemos ver que este es un buen modelo, cuyas variables son todas significativas, no tiene datos que generen apalancamiento y cumple con homocedasticidad.

Sintetizando, nuestro modelos es: $\hat{Y} = 0{,}1630 + (0{,}5192)x_2 + (1{,}3605)x_3 - (0{,}2451)x_5$

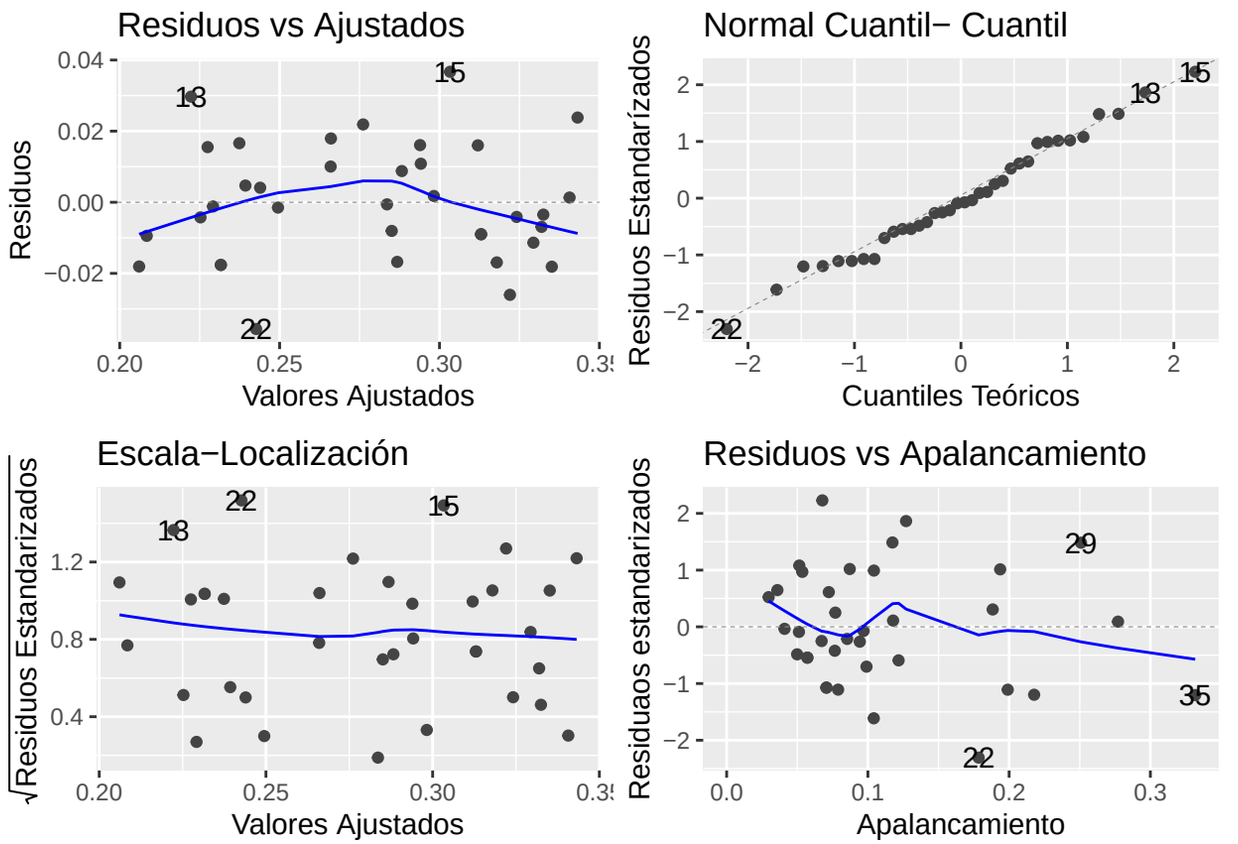

Figura 3: Graficos descriptivos del modelo

0.12. Prueba y predicción del modelo lineal

Ahora, haremos uso del comando `predict` para hacer la predicción de la variable X1 (tasa de hits), utilizando las 9 observaciones que se seleccionaron previamente.

Luego calculamos la diferencia entre los valores reales y los valores estimados por el modelo. Los resultados se muestran en la tabla 5. Es claro que los residuos son bastante pequeños, así que se considera que el modelo es suficientemente bueno para predecir la tasa de hits.

Tabla 5: Hits reales vs Hits predichos

Tasa de hits real	Tasa de hits predicha	Diferencia
0.281	0.2817371	-0.0007371
0.290	0.2638287	0.0261713
0.269	0.3130776	-0.0440776
0.307	0.3310601	-0.0240601
0.308	0.2455406	0.0624594
0.358	0.2242784	0.1337216
0.245	0.3030698	-0.0580698
0.294	0.2396552	0.0543448
0.218	0.2968785	-0.0788785

0.13. ANOVA para la Tasa de Bateo

Para finalizar, estamos interesados en realizar un análisis de varianza sobre la variable `X1` O la tasa de bateo, para compararla con el resto de los variables. Particularmente queremos realizar, el estudio sobre 3 categorías o grupos:

- Grupo 1: los bateadores con una tasa de bateo igual a $(X1 < 0{,}200)$.
- Grupo 2: los bateadores con una tasa de bateo igual a $(0{,}200 \leq X1 < 0{,}300)$.
- Grupo 3: los bateadores con una tasa de bateo igual a $(0{,}300 \leq X1)$

Con esta agrupación se opta por realizar un análisis de varianza con bloques aleatorizados, donde los bloques serán los grupos y los tratamientos o métodos serán las distintas variables de la base de datos.

Con la tabla 6 podemos apreciar las medias de los valores agrupados. Con estos valores, se puede aplicar el comando `anova` de `R` para obtener la tabla ANDEVA, tal y como se detalla en la tabla 7.

En esta tabla, se aprecia que el p-valor para lo grupos es de $0{,}6198$ que es alto, indicando que la hipótesis nula para los grupos no se puede rechazar, es decir, que las medias por grupos son iguales. Sin embargo, para las medias clasificadas por variable o método se obtuvo un p-valor de $0{,}0004$ que es bastante bajo, incluso significativo indicando que las medias son distintas tal como se esperaba por los datos analizados.

Con esto podemos afirmar con seguridad, que los promedios de las tasas son iguales por cada grupo.

Tabla 6: Tabla a dos factores para las medias de las variables

	X1	X2	X3	X4	X5	X6
Grupo1	0.1935000	0.0820000	0.0365000	0.0070000	0.00650	0.1965000
Grupo2	0.2580400	0.1327600	0.0414400	0.0087200	0.02228	0.1204800
Grupo3	0.3212778	0.1837222	0.0542778	0.0153333	0.02900	0.0716667

Tabla 7: Tabla ANDEVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupos	2	0.0019826	0.0009913	0.5020692	0.6197546
variables	5	0.1334024	0.0266805	13.5132135	0.0003524
Residuals	10	0.0197440	0.0019744	NA	NA

0.14. Conclusiones

De todo el análisis anterior se deduce que:

1. No hay demasiada variabilidad entre las diferentes tasas de bateo, salvo jonrones, asi que en general los jugadores de la MLB proectan rendimientos parecidos (aunque esto depende de la exactitud con la que se quiera medir).
2. La tasa de bateo es en media al menos mas del doble que la tasa de ponches para cualquier jugador (esto se sigue de la tabla 2).
3. La tasa de hits sigue aproximadamente un distribución normal centrada en $0{,}2822$ y con desviación estándar de $0{,}045$.
4. Las variables más significativas (entre las estudiadas), para predecir la tasa de hits o bateos son la tasa de carreras, la tasa de dobles, y la tasa de ponches, con estas se puede lograr un buen modelo lineal.

0.15. Referencias

Referencias

Albert, J. (2022). Comparing Home Run Rates for Two Seasons. https://baseballwithr.wordpress.com/

Barrendero, J. (2016). Regresión lineal simple con R.

Bentley, J. (s.f.). http://facweb1.redlands.edu/fac/jim_bentley/Downloads/R/TSUBUsingR.html

Community, C. (2010). Baseball analytics with R – the science of baseball. https://web.colby.edu/baseball/baseball-analytics-with-r/

Esteva, A. B. (2019). *Análisis y predicción del rendimiento ofensivo de debutantes en las grandes ligas de beisbol.* (Tesis doctoral). http://openaccess.uoc.edu/webapps/o2/bitstream/10609/98686/6/abenavideseTFM0619memoria.pdf

Heumann, C., y Schomaker, M. (2017). *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*. Springer International Publishing. https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.filepicker.io/api/file/TgilBNiCSk26axWSRDVA&ved=2ahUKEwiHr4z-lun2AhXFTDABHafBB24QFnoECBMQAQ&usg=AOvVaw3BrPvJxckh0b3geLuJEoCJ

Mora, W., y Borbón, A. (2010). *Edicion de Textos Cientificos Latex. Composición, Diseño Editorial, Gráficos, Inkscape, Tikz y Presentaciones Beamer* (2da). Escuela de Matemática,Instituto Tecnológico de Costa Rica. https://cristiancastrop.files.wordpress.com/2013/04/latex_febrero_2012_composicion_disenoeditorial_graficos_inkscape_tikz_beamer.pdf

R coder la mejor forma de empezar a Aprender Programación en r. (2020). https://r-coder.com/inicio/

Ramírez–Hassan, A., y Graciano-Londoño, M. (2021). A GUIded tour of Bayesian regression. *The R Journal*, *13*(2), 135-152. https://doi.org/10.32614/RJ-2021-081

Ramos-López, D., y Maldonado, A. D. (2021). Analysis of Corneal Data in R with the rPACI Package. *The R Journal*, *13*(2), 321-335. https://doi.org/10.32614/RJ-2021-099

Rodrigo, J. A. (2016). Correlación lineal y Regresión lineal simple. https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal#Informaci%C3%B3n_sesi%C3%B3n

Santo, S. (2019). Baseball and R Markdown Introduction. https://rstudio-pubs-static.s3.amazonaws.com/458464_993172e6aa3742e088eb86cc1b818eec.html#4_references

Tibau, M. (2017). Exploratory data analysis and baseball. https://rpubs.com/marcelo_tibau/exploratory-and-baseball

Xie, Y., Allaire, J. J., y Grolemund, G. (2019). *R markdown: The definitive guide*. CRC press. https://bookdown.org/yihui/rmarkdown/

Zu, T., y Yu, Y. (2021). SIQR: An R Package for Single-index Quantile Regression. *The R Journal*, *13*(2), 460-470. https://doi.org/10.32614/RJ-2021-092

0.16. ANEXO

Codigos utilizados en este informe:

Obtener los estadísticos

```
# Función para obtener un resumen estadístico completo de cada variable
estadisticos<- function(variables){
  # Inicializamos las variables
  k<- length(variables)
  # Minimo
  minimo <- rep(0,k)
  # Media
  media <- rep(0,k)
  # Mediana
  mediana<- rep(0,k)
  # Cuartile 1: 25 %
  q1 <-rep(0,k)
  # Cuartile 3: 75 %
  q3 <- rep(0,k)
  # Maximo
  maximo <- rep(0,k)
  # Rango Intercuartile
  ric <- rep(0,k)
  # Varianza
  varianza <- rep(0,k)
  # Desviación estándar
  stad <-rep(0,k)
  # Coeficiente de variación
  coef_var <-  rep(0,k)

  for(i in 1:k){
    # Minimo
    minimo[i] <- min(variables[,i])
    # Media
    media[i] <- mean(variables[,i])
    # Mediana
    mediana[i]<- median(variables[,i])
    # Cuartile 1: 25 %
    q1[i] <- quantile(variables[,i],0.25)
    # Cuartile 3: 75 %
    q3[i] <- quantile(variables[,i],0.75)
    # Maximo
    maximo[i] <- max(variables[,i])
    # Rango Intercuartile
    ric[i] <- IQR(variables[,i])
    # Varianza
```

```
    varianza[i] <- var(variables[,i])
    # Desviación estándar
    stad[i] <- sd(variables[,i])
    # Coeficiente de variación
    coef_var <-  stad/media
  }

  # Unimos los valores obtenidos
  estadisticos <- cbind(round(minimo, digits=4),round(q1, digits = 4),
                        round(media, digits=4), round(mediana,digits=4),
                        round(q3, digits=4), round(maximo, digits=4),
                        round(ric, digits=4),round(varianza, digits=4),
                        round(stad, digits=4), round(coef_var, digits=4))
  # Definimos los nombres de las columnas y filas
  rownames(estadisticos) <- c("X1", "X2", "X3", "X4", "X5", "X6")
  colnames(estadisticos) <- c("Minimo", "25%", "Media", "Mediana / 50" ,
                              "75%", "Máximo", "RIC","Varianza",
                              "Desv. Estándar","Coef. Variación")
  # Mostramos el arreglo
  return(estadisticos)

}
variables <-as.data.frame(Baseball)
```

Obtener las gráficas

```
# Se crea una matriz que defina el layout de las graficas
par(mfrow=c(6,2), mai=c(0.4,0.4,0.4,0.4), mgp=c(1.8,0.6,0.3))

hist(X1, main = "Histograma: Tasa de hits", ylab = "Frecuencia",
     xlab = "X1: Hits al bate", col="skyblue4")
boxplot(X1, main = "Gráfico de Cajas: Tasa de hits", xlab="Frecuencia",
        col="springgreen4", horizontal = T)

hist(X2, main = "Histograma: Tasa carreras anotadas", ylab = "Frecuencia",
     xlab = "X2: Tasa de carreras anotadas", col="skyblue3")
boxplot(X2, main = "Gráfico de Cajas: Tasa carreras anotadas", xlab="Frecuencia",
        col="springgreen3", horizontal = T)

hist(X3, main = "Histograma: Tasa de dobles", ylab= "Frecuencia",
     xlab= "X3: Dobles por veces al bate", col="skyblue2")
boxplot(X3, main = "Gráfico de Cajas: Tasa de dobles", xlab = "Frecuencia",
        col = "springgreen2", horizontal = T)

hist(X4, main="Histograma: Tasa de triples", ylab ="Frecuencia",
     xlab="X4: Triples por veces al bate", col ="skyblue1")
```

```
boxplot(X4, main = "Gráfico de Cajas: Tasa de triples", xlab = "Frecuencia",
        col = "springgreen1", horizontal = T)

hist(X5, main = "Histograma: Tasa de jonrones", ylab = "Frecuencia",
     xlab = "X5: Jonrones por veces al bate", col="skyblue")
boxplot(X5, main = "Gráfico de Cajas: Tasa de jonrones",xlab="Frecuencia",
        col="springgreen", horizontal = T)

hist(X6, main = "Histograma: Tasa de ponches", ylab = "Frecuencia",
     xlab = "X6: Ponches por veces al bate", col="skyblue")
boxplot(X6, main = "Gráfico de Cajas: Tasa de ponches", xlab="Frecuencia",
        col="springgreen", horizontal = T)
```

Intervalo de confianza para la media de las variables

Prueba de hipotesis para la tasa de bateo

```
t.test(X1, alternative = "greater", mu=0.3, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  X1
## t = -2.9779, df = 44, p-value = 0.9976
## alternative hypothesis: true mean is greater than 0.3
## 95 percent confidence interval:
##  0.2694453       Inf
## sample estimates:
## mean of x
## 0.2804667
```

Prueba de hipotesis para la tasa de jonrones y ponches

```
t.test(jonrones, ponches, alternative='two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  jonrones and ponches
## t = -7.0502, df = 42.827, p-value = 1.112e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10681617 -0.05929494
## sample estimates:
##  mean of x  mean of y
## 0.02544444 0.10850000
```

Prueba de bondad de ajuste

```
# Tamaño de los fi
(k<- length(fi))
```

```
## [1] 5
```

```
# Numero total de datos
n<- sum(fi)
# Puntos medios de los intervalos
(mi<-c(0.18+(0.24-0.18)/2,0.24+(0.28-0.24)/2,0.28+(0.30-0.28)/2,
       0.30+(0.32-0.30)/2, 0.32+(0.38-0.32)/2))
```

```
## [1] 0.21 0.26 0.29 0.31 0.35
```

```
# Media de los datos
(xbarra<-sum(fi*mi)/n)
```

```
## [1] 0.2822222
```

```
# Vector con las medias
x_barra<-rep(xbarra,k)

# Varianza
(S_cuadrado<-sum(fi*(mi-x_barra)^{2})/(n-1))
```

```
## [1] 0.001990404
```

```
# Desviación estandar
(S<-sqrt(S_cuadrado))
```

```
## [1] 0.04461394
```

```
#calculemos los pi
# P(Z<0.24)
(p1<-pnorm(0.24,mean= xbarra,sd=S))
```

```
## [1] 0.1719747
```

```
# P(0.24 < Z < 0.28)
(p2<-pnorm(0.28,mean= xbarra,sd=S)-pnorm(0.24,mean= xbarra,sd=S))
```

```
## [1] 0.3081622
```

```
(p3<-pnorm(0.30,mean= xbarra,sd=S)-pnorm(0.28,mean= xbarra,sd=S))
```

```
## [1] 0.174725
```

```
(p4<-pnorm(0.32,mean= xbarra,sd=S)-pnorm(0.30,mean= xbarra,sd=S))
```

```
## [1] 0.1465766
```

```
(p5<-pnorm(0.32,mean= xbarra,sd=S, lower.tail = F))
```

```
## [1] 0.1985615
```

```
# Vector con las probabilidades
(pi<-c(p1,p2,p3,p4,p5))
```

```
## [1] 0.1719747 0.3081622 0.1747250 0.1465766 0.1985615
```

```
# Suma de las probabilidades
(sum(pi))
```

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

```
# Estadistico
(t<-sum(((fi-n*pi)^{2})/(n*pi)))
```

```
## [1] 2.942129
```

```
# p-valor
(p_Valor<- 1-pchisq(t,k-1-2))
```

```
## [1] 0.2296808
```

Matriz de correlación

```
## Codigo de R-coder https://r-coder.com/grafico-correlacion-r/
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    Cor <- cor(x, y) # Elimina la función abs si lo prefieres
    txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])
    if(missing(cex.cor)) {
        cex.cor <- 0.4 / strwidth(txt)
```

```
    }
    text(0.5, 0.5, txt,
         cex = 1 + cex.cor) # Escala el texto al nivel de correlación
}
# Dibujamos la matriz de correlación
pairs(Baseball,
      label=c("Bateo","Carreras","Dobles", "Triples", "Jonrones", "Ponches"),
      upper.panel = panel.cor,    # Panel de correlación
      col = c("springgreen3"),  # Colores de los puntos
      bg = c("springgreen3"),   # Colores de los puntos
      pch = 21,                   # Símbolo pch
      main = "Matriz de Correlación de las variables",  # Titulo
      cex.labels = NULL,         # Tamaño del texto de la diagonal
      font.labels = 1           # Estilo de fuente del texto de la diagonal
)
```

Modelo lineal

```
n <- 36
set.seed(777)

elegidos <- sort(sample(seq_len(nrow(Baseball)),size = n))
Baseball_80 <- Baseball[elegidos, ]
Baseball_20 <- Baseball[-elegidos, ]
```

1 era prueba

```
tasa_de_bateo <- Baseball_80$X1
carreras <- Baseball_80$X2
dobles <- Baseball_80$X3
triples <- Baseball_80$X4
jonrones <- Baseball_80$X5
ponches <- Baseball_80$X6

m1 <- lm(tasa_de_bateo ~ carreras + dobles + triples + jonrones + ponches)
summary(m1)
```

```
##
## Call:
## lm(formula = tasa_de_bateo ~ carreras + dobles + triples + jonrones +
##     ponches)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038609 -0.011625  0.001590  0.008057  0.034985
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18250    0.01816  10.049 4.08e-11 ***
## carreras     0.35996    0.11455   3.142  0.00376 **
## dobles       1.22453    0.37152   3.296  0.00253 **
## triples      0.55428    0.62532   0.886  0.38245
## jonrones     0.39297    0.18458   2.129  0.04158 *
## ponches     -0.28815    0.05097  -5.653 3.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01641 on 30 degrees of freedom
## Multiple R-squared:  0.8879, Adjusted R-squared:  0.8692
## F-statistic: 47.53 on 5 and 30 DF,  p-value: 2.308e-13
```

2da prueba

```
m2 <- lm(tasa_de_bateo ~ carreras + dobles + jonrones+ ponches)
summary(m2)
```

```
##
## Call:
## lm(formula = tasa_de_bateo ~ carreras + dobles + jonrones + ponches)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.035060 -0.012486  0.001372  0.007376  0.037906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.17715    0.01707  10.379 1.31e-11 ***
## carreras     0.40656    0.10142   4.009 0.000357 ***
## dobles       1.36860    0.33292   4.111 0.000268 ***
## jonrones     0.32041    0.16486   1.944 0.061079 .
## ponches     -0.29278    0.05053  -5.794 2.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01635 on 31 degrees of freedom
## Multiple R-squared:  0.885,  Adjusted R-squared:  0.8701
## F-statistic: 59.63 on 4 and 31 DF,  p-value: 4.07e-14
```

3ra prueba

```
m3 <- lm(tasa_de_bateo ~ carreras + dobles + ponches)
summary(m3)
```

```
##
## Call:
## lm(formula = tasa_de_bateo ~ carreras + dobles + ponches)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.035643 -0.009929 -0.001336  0.012005  0.036693
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.16301    0.01610  10.126 1.67e-11 ***
## carreras     0.51923    0.08676   5.985 1.13e-06 ***
## dobles       1.36501    0.34706   3.933 0.000423 ***
## ponches     -0.24507    0.04604  -5.323 7.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01705 on 32 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8589
## F-statistic:    72 on 3 and 32 DF,  p-value: 2.555e-14
```

Predicciones

```
predicciones <- predict(m2,Baseball_20, interval="confidence")
pre <- predicciones[c(1:9)]
diferencia <- Baseball_20$X1 - pre
tabla_pre <- cbind(Baseball_20$X1 , pre ,  diferencia)
colnames(tabla_pre) <- c("Tasa de hits real "," Tasa de hits predicha " , " Diferencia ")
knitr::kable(tabla_pre, align = "ccc" , caption = "Hits reales vs Hits predichos")
```

Modelo Anova

```
grupo1<- subset(Baseball, (X1<0.200) )
grupo2<- subset(Baseball, (X1 >= 0.200 & X1 <=0.300))
grupo3<- subset(Baseball, (X1>=0.300))

media1<- colMeans(grupo1[sapply(grupo1, is.numeric)])
media2<- colMeans(grupo2[sapply(grupo1, is.numeric)])
media3<- colMeans(grupo3[sapply(grupo1, is.numeric)])

medias <- rbind(media1,media2,media3)
```

```
datos<- c(medias[,1] , medias[,2], medias[,3],medias[,4],medias[,5],medias[,6])

variables <- gl(6,3, labels=c("X1","X2","X3","X4","X5","X6"))
grupos = factor(rep(1:3,6), labels=c("Grupo1","Grupo2","Grupo3"))
xtabs(datos~grupos+variables )
```

```
##         variables
## grupos           X1         X2         X3         X4         X5         X6
##   Grupo1 0.19350000 0.08200000 0.03650000 0.00700000 0.00650000 0.19650000
##   Grupo2 0.25804000 0.13276000 0.04144000 0.00872000 0.02228000 0.12048000
##   Grupo3 0.32127778 0.18372222 0.05427778 0.01533333 0.02900000 0.07166667
```

```
library(knitr)
library(kableExtra)

kable(xtabs(datos~grupos+variables ),
      caption = "Tabla a dos factores para las medias de las variables") %>%
  kable_styling(full_width = F)
```

```
# Una variables
modelo.lineal = lm(datos~variables)
anova(modelo.lineal)
```

```
## Analysis of Variance Table
##
## Response: datos
##           Df   Sum Sq   Mean Sq F value    Pr(>F)
## variables  5 0.133402 0.0266805  14.736 9.136e-05 ***
## Residuals 12 0.021727 0.0018105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Dos variables
modelo.lineal2 = lm(datos~grupos+variables)
prueba<-anova(modelo.lineal2)
```

```
kable(prueba,caption = "Tabla a dos factores para las medias de las variables") %>%
  kable_styling(full_width = F)
```

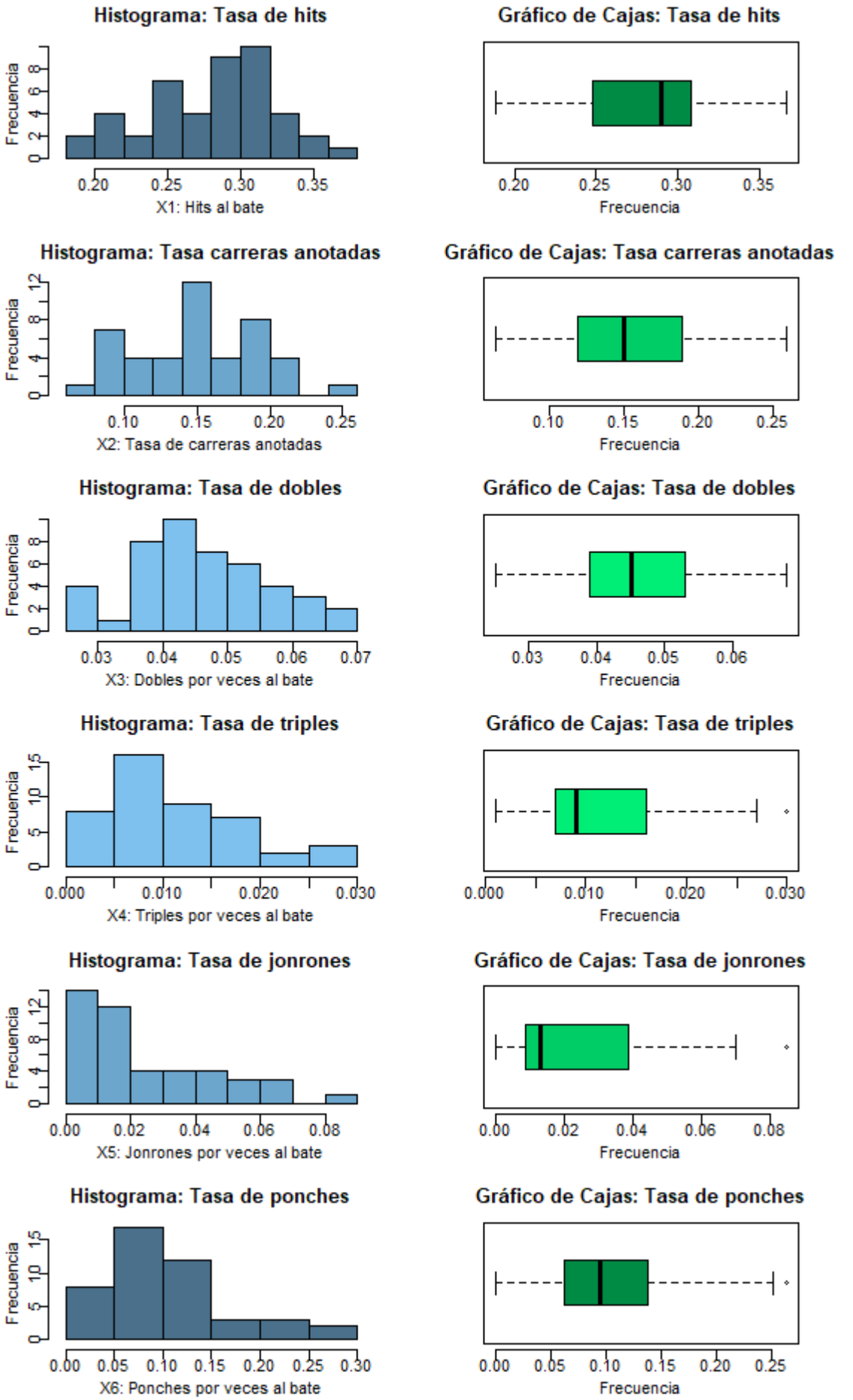

Figura 4: Histograma y gráfico de cajas para las variables

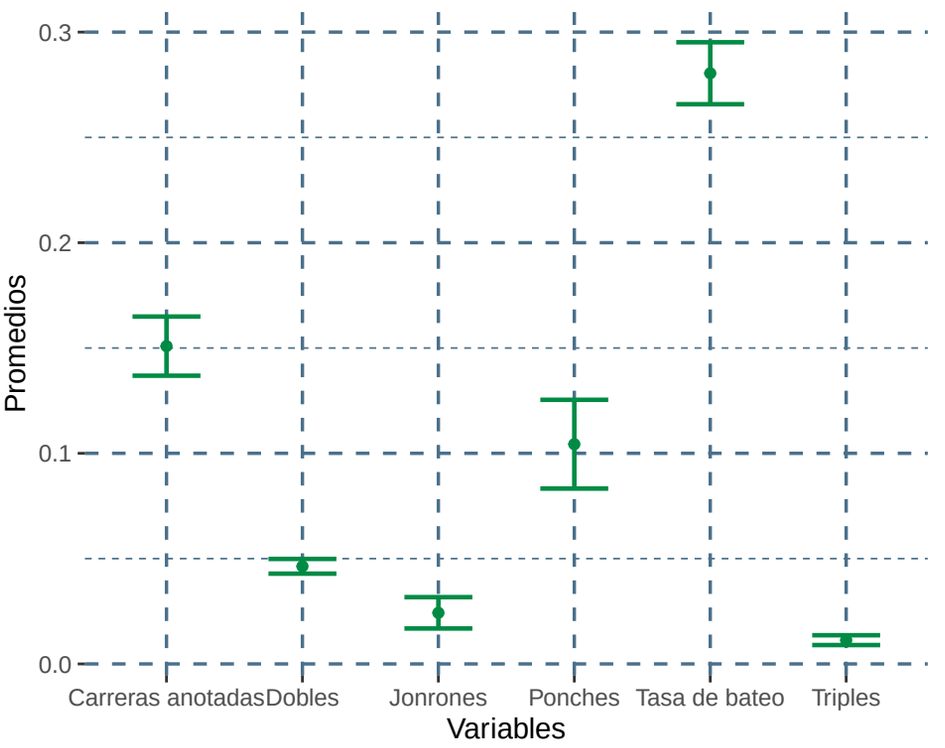

Figura 5: Rpresentación gráfica de los intervalos de confianza para las medias