Proyecto Final
Modelo de Regresión Lineal Para Salarios Anuales y Producción Ofensiva en MLB

Andrés Moguel López Jensen
Regresión
29 de noviembre de 2022
Licenciatura en Actuaría
Otoño 2022
Universidad Iberoamericana

Contenido

Índice de Figuras, Tablas y Ecuaciones

Figuras

Tablas

Ecuaciones

Introducción

El béisbol es un deporte donde las estadísticas son el rey y mandan en el día a día de las distintas ligas y sus jugadores. La liga profesional de los Estados Unidos *Major League Baseball* (MLB) es la liga en la cuál más se ha profundizado el análisis estadístico de este deporte. Las estadísticas dan argumentos para que los fanáticos y analistas debatan quien es el mejor jugador de una temporada, quién pertenece en el salón de la fama o quién está teniendo un año para el olvido. Estas estadísticas no solo permiten y abren este tipo de debates y conversaciones, sino que atraen o repelen a los equipos de la liga. Si un jugador tiene una temporada impresionante o buena a la ofensiva en términos de producción, medida a través de distintas estadísticas, su equipo buscará mantenerlo en su roster mientras que los demás equipos de la liga buscarán firmarlo al suyo. Los equipos mostrarán este interés ofreciéndole al jugador un contrato o extensión de contrato.

Los mejores jugadores recibirán contratos más atractivos, es decir tendrán una mayor duración y valuación en términos monetarios. Jugadores con un mal desempeño o desempeño promedio se esperaría que reciban contratos relativos a lo que ofrecen dentro del terreno de juego. Es decir que la valuación de los contratos que ofrecen los equipos se basa en su nivel de interés de traerlos a su equipo y cuanto piensan que vale la producción del jugador.

Los equipos y su gerencia no son adivinos entonces no pueden predecir si el jugador que firmaron realmente valdrá, en términos de producción, lo que le están pagando. Este es un problema para los equipos ya que los contratos en MLB son completamente garantizados. Es decir, que los equipos están obligados a pagar todo el dinero acordado al jugador sin importar que pase a futuro.

Al combinar las ideas presentadas en los últimos dos párrafos nace la idea de este proyecto de investigación. Intentar explicar el salario anual de un jugador a través de sus estadísticas ofensivas de ese año. Se busca contestar: ¿Un jugador realmente vale lo que se le está pagando? ***¿La directiva del equipo le ofreció el contrato correcto?*** A continuación, se explora, utilizando modelos de regresión lineal, la complicada relación que genera tantos dolores de cabeza a las directivas entre producción y valuación de sus jugadores.

Alcance

Este trabajo se enfoca en usar la regresión lineal para explicar la relación que existe entre el salario anual de un bateador y su desempeño y producción de esa temporada. No se pretende crear un modelo para predecir salarios a base del desempeño. De hecho, hacer dicha predicción sería incongruente con la realidad dado que el salario para la temporada se establece previo a iniciar la temporada cuando las estadísticas de todos los jugadores son desconocidas. Bajo el supuesto de que solamente se toma el salario anual por temporada, el estudio no contempla ni el tipo de contrato ni la duración del mismo. Ya que se ignoran las dos variables anteriores y solamente es de interés las estadísticas productivas, también se excluye la variable de edad del jugador. Se reconoce que existe, en todos los deportes, una relación importante entre edad y

productividad. En béisbol también existe una relación entre edad y salario percibido que va de la mano con el tipo de contrato; existe un contrato para novatos y para veteranos. Al intentar explicar el salario con base en estadísticas productivas se podría contestar la siguiente pregunta ***¿Un jugador realmente vale lo que se le está pagando?***, esto en términos de productividad en el terreno de juego.

Solamente se trabajará con bateadores, es decir que no se consideran lanzadores para este estudio. Esto se debe a que son estadísticas distintas las que se miden para los lanzadores que para los bateadores; i.e. se tendría que desarrollar otro modelo para poder explicar el salario de los lanzadores. El modelo desarrollado para los bateadores solamente se enfoca en producción ofensiva, es decir que se ignora todo lo que el bateador hace fuera del plato, i.e. fuera de un turno al bate. No se considera ningún tipo de estadística defensiva en este estudio. Al ignorar las estadísticas defensivas, no se tomará en cuenta la posición de campo de los jugadores (receptor, primera base, jardinero, campo corto, etc.) y solamente se contemplará que son bateadores. Es importante notar que es sabido que los equipos esperan producciones ofensivas distintas para cada posición en el campo debido a la dificultad defensiva que cada posición tiene. Se sugiere para un estudio futuro contemplar este hecho.

2093 temporadas individuales (datos de bateadores) fueron consideradas para este estudio. Se tomaron los datos de productividad de todos los bateadores calificados de la temporada 2003 en adelante. Tanto en el marco teórico como en el desarrollo y la justificación se explica a detalle la decisión de tomar solamente a los bateadores calificados, así como elegir la temporada 2003 como punto de partida.

En MLB no existe un tope salarial y por ende cada equipo puede gastar lo que desee para armar el roster para cada temporada. No todos los equipos tienen la misma capacidad financiera, esto se cubrirá más a detalle en el marco teórico. Esto genera que ciertos equipos le puedan pagar salarios mucho más altos a sus jugadores que otros. El factor de equipo al que perteneces claramente afecta el monto de salario que un jugador percibe. Este trabajo solamente contempla explicar salarios a base de productividad en el diamante y por ende no se contempla el equipo al que el jugador pertenece; se sugiere contemplar dicho factor para un análisis futuro.

Justificación

En *Moneyball: The Art of Winning an Unfair Game,* Michael Lewis narra la historia de la temporada 2002 de los Atléticos de Oakland. Los Atléticos, con la tercera nómina más baja de la liga, buscaban una manera de mantenerse competitivos contra equipos de mercados y nóminas mucho mayores. De la mano de su gerente general Billy Beane y su asistente Paul de Podesta, se construyó un roster basado en sabermetría, i.e. un enfoque analítico y estadístico para encontrar jugadores productivos a precios accesibles. Los Atléticos lograron ganar 20 juegos consecutivos y avanzar a la postemporada en 2002[1]. Tras el éxito de este experimento, el béisbol cambió para siempre. Poco a poco las 30 organizaciones fueron cambiando su estrategia para afrontar las nuevas temporadas y construir sus rosters. Hoy en día la sabermetría y sus terminología son muy comunes y están integradas completamente en la vida cotidiana

del deporte: desde los aficionados, jugadores y la gerencia de los equipos. La productividad de los jugadores está a la orden del día en MLB.

Los contratos de larga duración, y, por ende, mayor valuación han ido a la alza en MLB[2,3]. Al querer asegurar a sus estrellas jóvenes a mediano y largo plazo, las organizaciones han otorgado, con mayor frecuencia este tipo de contratos. Se considera que estos contratos tienen un potencial de "auge o pérdida". Esto se debe a que los contratos en MLB son completamente garantizados y en caso de que los jugadores no produzcan al nivel que sus equipos esperaban, recibirán el monto acordado en sus contratos. Algunos ejemplos recientes de los contratos previamente mencionados son: Mike Trout: $426.5 millones por 12 años, Fernando Tatís Jr: 340 millones por 12 años y Corey Seager: $325 millones por 10 años[4]. La tendencia a la alza de este tipo de contratos junto con el nuevo enfoque analítico de construcción de rosters y composición de nómina justifican la idea de explicar el salario percibido de los bateadores a través de su producción ofensiva.

Marco Teórico

Nóminas y Tamaños de Mercado

Para la temporada 2022 se estimó que MLB alcanzó niveles de ingresos alrededor de $11 mil millones, a través de mercados de distintos tamaños[5]. Los mercados van desde las megalópolis de Los Ángeles y Nueva York, que cuentan con dos franquicias cada uno, a mercados muy pequeños como Tampa Bay y Oakland (véase Anexo 1 para la lista completa de valuaciones)[6]. El tamaño del mercado, que afecta la valuación del equipo, influye en que tanto gastan los equipos al construir su roster.

A diferencia de otras ligas y deportes profesionales en los Estados Unidos, como la NFL o la NBA, en MLB no existe un tope salarial. Es decir, que los equipos pueden gastar y manejar su nómina sin restricción alguna. Sin embargo, para mantener un cierto nivel de competitividad considerando los distintos tamaños de mercado, MLB intenta desalentar el gasto excesivo en nómina de jugadores con la implementación del Impuesto de Balance Competitivo (*Competitive Balance Tax)* conocido coloquialmente como el Impuesto de Lujo. Esta regla establece que los equipos que sobrepasen el límite establecido por el impuesto deberán pagar una multa. Para la temporada 2022 este límite se estableció en $230,000,000; se proyectó que 7 equipos rebasaron dicho umbral para ver las nóminas de cada equipo en 2022 (véase Anexo 2)[4].

Tipos de Contratos y Salarios

Para cada temporada, MLB establece un salario mínimo. Para la temporada 2022 el salario mínimo fue de $700,000; para ver el desglose de salarios mínimos desde el año 2003 véase Anexo 3[7]. El salario máximo, no ajustado, en los años contemplados para el estudio es de $43,300,00 ; ya que el estudio considera 20 años de juego, se toma en consideración el valor del dinero en el tiempo para poder comparar salarios y que el ajuste a las categorías ofensivas, que no han sufrido cambios, sea el correcto. Todos los salarios se ajustan al año base de 2012; es decir que las cifras con las que se trabajan son en dólares americanos de 2012. Se utiliza el año base de 2012 ya que este es el que considera la *Federal Reserve Economic Data* (FRED) para el índice de precios al consumidor[8] (véase Anexo 18 para ver los factores de transformación). Hay un rango

considerablemente grande para los salarios considerados; el objetivo de este trabajo es intentar explicar los salarios anuales a través del desempeño y producción en el terreno de juego.

Un elemento importante para la determinación de salarios anuales son las audiencias de arbitraje. Un jugador es elegible para dicha audiencia para determinar su salario si ha acumulado entre 3 y 6 años de tiempo de juego en MLB[9]. Esta audiencia determina el salario del jugador para la próxima temporada en caso de que no se pueda poner de acuerdo con su equipo para llegar a una extensión de contrato o no logren determinar una cifra anual que le parezca justa a ambos.

Así como no todos los jugadores perciben el mismo salario, es importante notar que no todas las posiciones de campo a la defensiva valen o perciben el mismo salario. En béisbol hay 9 posiciones de campo a la defensiva: lanzador, receptor, primera base, segunda base, tercera base, campo corto y 3 jardineros (izquierdo, derecho y central). Hasta el 2021 existió una diferencia principal entre las dos ligas, Liga Americana y Liga Nacional, que conforman MLB. En la Liga Americana había un bateador designado (DH) y en la Liga Nacional no. El DH tomaba turnos al bate en lugar del lanzador. A partir de 2022, MLB adoptó al DH universal; tanto en la Liga Americana como en la Liga Nacional. El tipo de producción que un equipo busca de un jugador depende de la posición en la que jueguen en el campo y su nivel defensivo que se requiera. Se considera que jugar primera base no es tan complicado defensivamente hablando y es por esto por lo que los equipos buscan jugadores que sean bateadores de poder con muchas carreras producidas y cuadrangulares. La posición de campo corto es mucho más demandante defensivamente y por lo tanto los equipos buscan buenos defensas con pocos errores a cambio de una menor producción ofensiva. Es por esto por lo que es de esperarse que el salario de un buen primera base sea distinto al de un buen campo corto, para ver salarios promedios por posición véase Anexo 4[4].

Se considera relevante que el lector tenga esta información a la mano para entender como funcionan los salarios en MLB a grandes rasgos. Sin embargo, este estudio no contempla los factores de tipo de salario, salario mínimo, arbitraje o posición de campo para explicar los salarios. Se recomienda que estos sean explorados en un estudio futuro.

Estudios y Casos Anteriores

El béisbol de MLB se ha estudiado a detalle desde que firmó el *Basic Agreement* entre el sindicato de jugadores y la liga en 1976. Al entrar en vigor, esto permitió a los jugadores convertirse en agentes libres y les otorgó el poder de negociar nuevos contratos con todos los equipos de la liga. Desde entonces los contratos otorgados, así como el desempeño de los jugadores y la relación que existe entre ambos se ha estudiado por varios investigadores.

Moneyball dirigida por Bennet Miller (2011) se basa en el libro de Michael Lewis y narra la historia de la temporada 2002 de los Atléticos de Oakland. En la película Billy Beane, gerente general del equipo decide tomar un enfoque analítico, sugerido por Peter Brand,

para la construcción del roster de la temporada. Los Atléticos lograron imponer la marca de partidos ganados de manera consecutiva para la Liga Americana con 20 y lograron ganar 103 encuentros al finalizar la temporada[10]. Oakland tenía una nómina de $39,679,476, la tercera más baja de MLB para el 2002[4]. El enfoque de Brand consintió en encontrar jugadores subvalorados por el sistema para conseguir máxima productividad al menor costo. Aunque esto es posible, y ciertos equipos intentan negociar con jugadores salarios anuales para intentar traerlos a un menor costo, la mayoría de los equipos de MLB tienen por lo menos un contrato de valuación alta. Esto lleva a la idea de polarización de nóminas descrita por Staudohar (1997) fenómeno en el cual 20% de los jugadores de un equipo componen el 80% de la nómina[11]. Esto indica que ciertos buenos jugadores que tenga el equipo, los intentará mantener a mayores costos y años de contrato[12].

Estudios anteriores, utilizando distintas metodologías, indicaron una relación significativa entre la duración del contrato y el rendimiento del jugador. Meltzer (2005), realizó una regresión de mínimos cuadrados en dos etapas para analizar cómo las variaciones en el rendimiento de los jugadores y otros factores, como lesiones, premios o reconocimientos e incidentes fuera del campo, tienen un impacto en el valor del contrato y en la duración del contrato de un jugador[13]. Meltzer concluye que los jugadores jóvenes obtienen acuerdos de bajo dinero a largo plazo para alentarlos a mejorar su rendimiento. Los jugadores veteranos y aquellos propensos a lesionarse reciben contratos cortos con un salario que los equipos consideran adecuado contemplando los riesgos de un bajo desempeño[13].

Stankiewicz (2009) explora con más detalle la relación entre la duración del contrato y el desempeño del jugador[14]. Stankiewicz estudia los contratos en función de la opinión de los jugadores y explicaque los jugadores de MLB prefieren un contrato a largo plazo ya que hay un ingreso garantizado durante un largo período de tiempo. Stankiewicz luego respalda esto explicando que un contrato de varios años es mejor que un contrato de un año, ya que aumenta el rendimiento del jugador[14]. Aparentemente, los jugadores respaldan ese argumento para acuerdos a largo plazo al generar una mayor producción ofensiva a largo plazo para el equipo. Stankiewicz sugiere que este tema se estudie más a fondo ya que sus hallazgos se realizaron solo en algunos jugadores con un enfoque muy general.

Otros estudios que exploran la relación entre la duración del contrato y el rendimiento del jugador contradicen la conclusión de Stankiewicz de que los acuerdos de varios años resultan en una mayor producción ofensiva para el equipo. Cahill (2014) y Sturman y Thibodeau (2001) concluyen que el rendimiento de los jugadores tiende a disminuir una vez que firman un contrato a largo plazo[15,16]. Es importante señalar que Stankiewicz utiliza teorías económicas en su estudio mientras que Cahill, Sturman y Thibodeau utilizan modelos de regresión para obtener sus hallazgos[14,15,16].

Judge y colaboradores (2010) realizaron un metanálisis con datos de los últimos 120 años explorando la relación entre remuneración y satisfacción laboral en todos los ámbitos de trabajo[17]. Concluyen que menos del 2% de las personas están realmente

motivadas por los salarios que perciben. Los investigadores también explican que la gente no relacionaba su desempeño y/o satisfacción laboral con el aumento en sus salarios.

Desarrollo

Este estudio busca explicar el salario anual de un bateador a través de su producción y desempeño en el terreno de juego tomando en consideración diferentes estadísticas y categorías ofensivas. Es decir que se busca contestar la siguiente pregunta ***¿Qué variables de producción ofensiva son significativas para explicar el salario anual de un bateador?*** Esto se hará a través de un modelo de regresión simple utilizando el lenguaje de programación y la plataforma de SAS. A continuación, se desarrolla toda la metodología de investigación y desarrollo del Modelo.

Recopilación y Selección de Datos

Se recolectaron los datos de bateadores para la temporadas 2003 a 2022 de la popular página de internet, utilizada en diferentes estudios, Baseball-Reference[7]. Se toma la temporada 2003 como el punto de inicio del estudio ya que se considera que la temporada de 2002 de los Atlético de Oakland y el libro de Michael Lewis cambiaron, desde entonces, la forma de pensar en MLB acerca de la productividad y salarios de jugadores por parte de los equipos[1,10].

Como primer paso para evitar puntos anómalos, de apalancamiento o de influencia el estudio solamente considera bateadores calificados, es decir, que tienen por lo menos 3.1 apariciones al plato (PA) por juego; el promedio de la liga[7]. Esto también va de acuerdo con el método que utiliza MLB para determinar a los líderes en las distintas categorías ofensivas[9].

Los salarios anuales fueron obtenidos de Spotrac[4], página que se especializa en contratos y salarios de las 4 ligas mayores de los Estados Unidos (béisbol, basquetbol, hockey y futbol americano. Al considerar 20 temporadas de estudio, se considera la idea del valor del dinero en el tiempo y todos los salarios se convierten a dólares del 2012. Para esta conversión se usa el índice de Precios al Consumidor para Consumidores Urbanos en Estados Unidos de acuerdo con FRED[8]. El rango de salarios es muy amplio [555,879.94, 35,740,592.80] y por lo tanto para manejar con mayor facilidad estas cifras se transforman tomando el logaritmo natural de los salarios; nuevo rango [13.2283, 17.3918] esto convierte al modelo en uno log-lineal. Es importante considerar que debido a la pandemia de COVID-19, la temporada 2020 consistió en 60 juegos en lugar de 162. Es por esto por lo que se eliminan los datos de la temporada 2020; para que exista consistencia de las observaciones.

Análisis Exploratorio de Datos

La base de datos con la que se realizó el estudio se compone de las siguientes variables (por orden de aparición): nombre del jugador, edad, equipo, liga, juegos, apariciones al plato, turnos al bate, carreras anotadas, imparables, dobles, triples, cuadrangulares (jonrones), carreras producidas, bases robadas, intentos fallidos de robo, bases por bolas, ponches, promedio de bateo, porcentaje en base, slugging, porcentaje en base

más slugging, porcentaje en base más slugging avanzado, bases totales, doble matanzas, golpes, sacrificios, elevados de sacrificio, bases por bolas intencionales, posición, año, salario, salario ajustado a 2012 y los logaritmos naturales de ambas variables de salario. Para ver una descripción más detallada de cada variable y sus abreviaciones, véase Anexo 5. A partir de este punto en adelante se mencionarán estas abreviaciones para las estadísticas ofensivas.

Para el desarrollo del modelo se comienza descartando variables que no son relevantes para el estudio o que aportan información redundante o que se sabe que, por su definición o composición, tienen una alta correlación con otras variables represoras. El nombre del jugador es irrelevante para el estudio. Como se menciona en alcance, el modelo no contemplará la edad, el equipo (y por ende la liga) ni la posición de campo a la defensiva. Las estadísticas de AB y H son redundantes ya que la variable de BA se calcula con el cociente de las mismas y aporta más información; es de las tres estadísticas convencionales junto con HR y RBI. Las variables de AB, HBP y BB se consideran para el calculo de OBP y presentarían una alta correlación. IBB también se elimina ya que es un subconjunto de BB. Tanto SB como CS no se contemplan ya que esta no es producción ofensiva en el turno al bate (no es de interés para el estudio) y depende de la estrategia del equipo; ya se había eliminado la variable de equipo. TB no se considera ya que es el numerador de SLG. OPS es la suma de OBP y SLG y se considera una estadística más robusta para la producción ofensiva; no se contemplan OBP y SLG individualmente. OPS+ no se contempla al ser una transformación muy compleja de OPS. SH no se consideran ya que dependen de la estrategia del equipo. En cuanto a la respuesta (salario) se contemplará el logaritmo natural del salario ajustado a 2012; se descartan las variables de año, salario y ln(salario). Tras esta simplificación inicial nos quedan las variables predictores: R, 2B, 3B, HR, RBI, SO, BA, OPS, GDP y SF con las que se buscará explicar la variable respuesta de: logaritmo natural del salario ajustado.

Antes de comenzar con la construcción y definición del modelo, es importante considerar las relaciones que existen entre las variables regresoras y la variable respuesta, así como la relación entre las variables regresoras. Esto se hará a través de un análisis exploratorio de datos.

Coeficientes de Correlación de Pearson, N = 2093											
	G	R	2B	3B	HR	RBI	SO	BA	OPS	GDP	SF
Salario	0.01293	0.15708	0.04625	-0.18853	0.24103	0.25088	0.03871	0.09600	0.23600	0.14930	0.10173

Tabla 1 ***Coeficientes de correlación de Pearson de Salario contra variables regresoras***

La Tabla 1, véase Anexo 6 para su generación en SAS, muestra un vector de correlaciones de Pearson entre la variable respuesta (salario) y las variables regresoras.

La correlación más alta (y más fuerte) es entre Salario y RBI (0.25088). Todas las correlaciones son positivas excepto con 3B (-0.18853) esto puede sugerir multicolinealidad ya que esta es una estadística de producción positiva. Tanto SO como GDP son estadísticas de producción negativas; su signo de correlación positiva puede indicar multicolinealidad. La correlación más débil es con G (0.01293). La tabla deja observar que no parece haber una sola variable que pudiese explicar el salario anual por si sola; es importante recordar que correlación no implica causalidad.

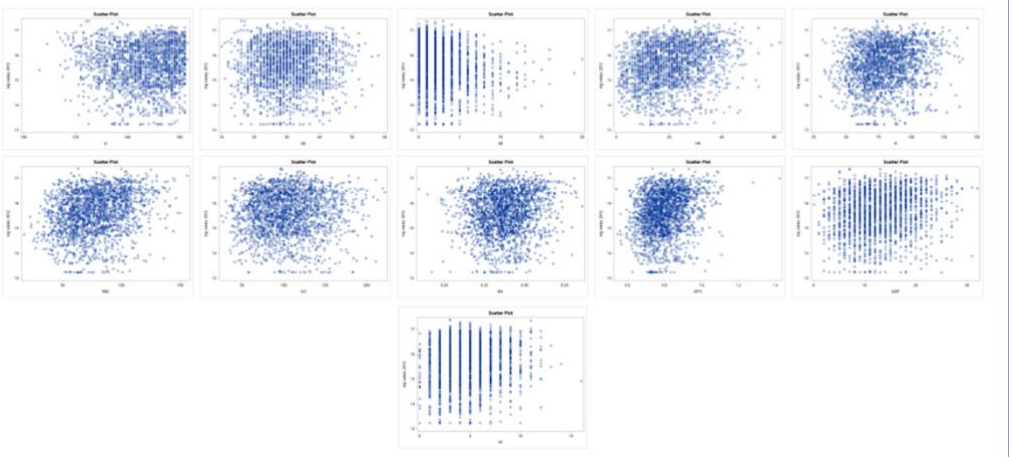


Figura 1 ***Matriz de Gráficos de dispersión, Salario vs. Variables Regresoras***

La Figura 1 muestra una matriz de gráficos de dispersión de salarios contra todas las variables regresoras a considerar. Para ver los gráficos de manera individual y a mayor detalle, así como el código en SAS véase Anexo 7. De manera general, en los gráficos de R, HR, RBI, y OPS parece existir una relación lineal (o un indicio de un patrón lineal) entre estas variables y el salario percibido. Para las variables de G, 2B, 3B, SO, BA, GDP y SF los puntos parecen estar repartidos aleatoriamente y no parece que exista una relación lineal. Al correr el modelo inicial con todas las variables regresoras se harán pruebas de significancia de la regresión tanto para el modelo (con la estadística F) y las variables regresoras de manera individual (con la estadística t).

En todos los gráficos se podría considerar que se pueden observar valores atípicos o *outliers* los más notorios son los puntos para G (una menor cantidad), R (a la extrema derecha), 3B (a la extrema derecha), OPS (arriba a la extrema derecha) y SF (extrema derecha en medio). A continuación, se identifican apropiadamente y se detalla una breve explicación de los mismos.

Identificación de Datos Influyentes y *Outliers*

Para identificar los *outliers* se utilizarán los residuales r-estudentizados y para los puntos de influencia se utiliza la D de Cook; el desarrollo se hace en SAS (véase Anexo 8).

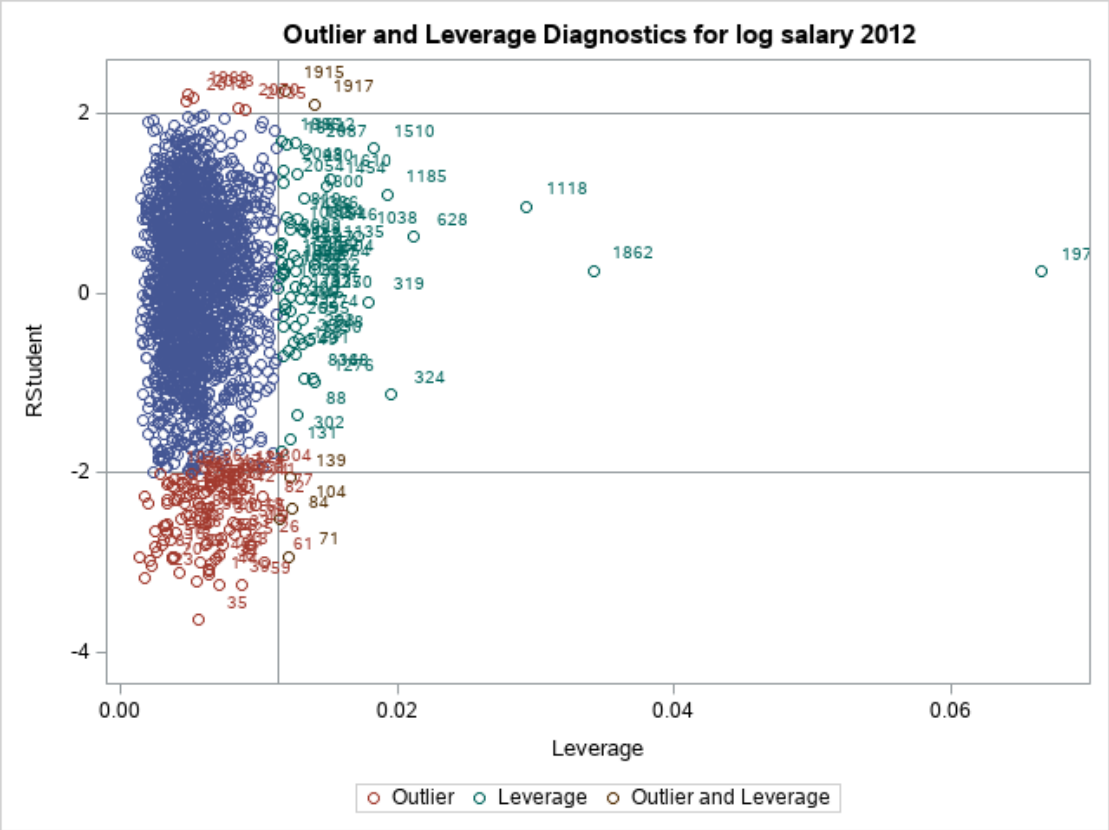


Figura 2 ***Detección de Outliers utilizando residuales R-Estudentizados***

Observación	Residuales R-Estudentizados
1	-3.191
4	-3.095
6	-3.087
20	-3.03
23	-3.157
34	-3.061
35	-3.623
37	-3.015
39	-3.232
44	-3.128
59	-3.238

Tabla 2 ***Outliers Detectados por valor de R-Student***

La Figura 2 muestra los puntos que potencialmente son de influencia y outliers según los valores de los residuales r-estudentizados. Los puntos que nos interesan son los de color naranja/rojo ya que solamente se utilizarán este tipo de residuales para detectar posibles *outliers* y no los puntos de influencia. La Tabla 2 muestra las observaciones que superaron el valor de corte de $|\pm 3|$ para residuales r-estandarizados que deben ser

mencionadas o investigadas más a detalle. Las observaciones corresponden a las siguientes temporadas individuales (en orden de aparición de arriba abajo según la Tabla 2): Brendan Rodgers 2022, Alejandro Kirk 2022, Alec Bohm 2022, Will Smith 2022, Ty France 2022, Kyle Tucker 2022, Yordan Álvarez 2022, Buster Posey 2012, Carlos Peña 2007, Adrián González 2008 y Ryan Howard 2007.

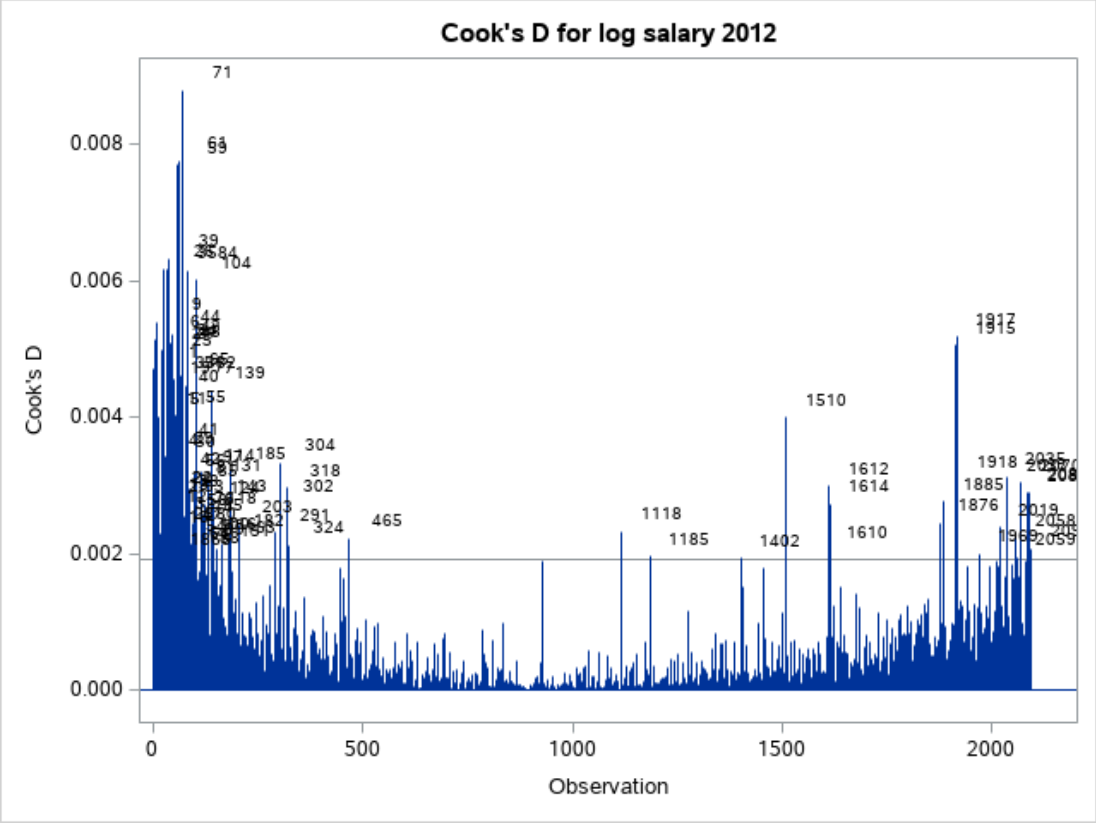


Figura 3 ***Puntos de Influencia con D de Cook***

Observación	Cook's D
71	0.009
61	0.008
59	0.008
104	0.006
84	0.006
39	0.006
35	0.006
26	0.006

Tabla 3 ***Observaciones con D de Cook > 0.005***

La Figura 3 muestra la D de Cook de todas las observaciones. La línea horizontal es el valor de corte para detectar puntos de influencia; se establece en 0.19 ($\frac{4}{n} = \frac{4}{2093}$). Las observaciones que se encuentran señaladas con su respectivo número son aquellas que están por encima del valor de corte y por ende pueden ser puntos de influencia. Para simplificar un poco la visualización de datos seleccionaron las observaciones que tienen una D de Cook mayor que 0.005 las cuales se muestran en la Tabla 3. Estas observaciones corresponden a las siguientes temporadas individuales (en orden de aparición de arriba abajo según la Tabla 3): Ronald Acuña Jr. 2019, Albert Pujols 2003, Ryan Howard 2007, Casey McGehee 2014, Troy Tulowitzki 2009, Carlos Peña 2007, Yordan Álvarez 2022 e Ian Kinsler 2008.

Hay tres temporadas que se repiten en las Tablas 2 y 3, es decir que son influyentes y outliers. En 2007 Ryan Howard quedó en quinto lugar en la votación para el jugador más valioso de la Liga Nacional y fue el líder en SO de MLB. En 2007 Carlos Peña quedó en noveno lugar en la votación para el jugador más valioso de la Liga Americana y ganó el premio *Silver Slugger* para su posición de campo: primera base. En 2022 Yordan Álvarez fue seleccionado al juego de estrellas, quedó en tercer lugar en la votación para el jugador más valioso de la Liga Americana y ganó el premio *Silver Slugger* para los bateadores designados.

La Figura 2 muestra ciertas observaciones (1118, 1862,1977) que podrían también ser de interés por su influencia extrema basada en residuales r-estudentizados. La observación 1118 corresponde a la temporada de 2007 de Jimmy Rollins, quién ganó el premio al jugador más valioso de la Liga Nacional y el premio *Silver Slugger* y fue líder de MLB en: G, PA, AB, R y 3B. La observación 1862 corresponde a la temporada de 2003 de Barry Bonds. En 2003 Bonds ganó el premio al jugador más valioso de la Liga Nacional y el premio *Silver Slugger* y lideró MLB en: BB, OBP, SLG, OPS, OPS+ y IBB. La observación 1997 corresponde a la temporada 2004 de Barry Bonds, en la cuál ganó el premio al jugador más valioso de la Liga Nacional y el premio *Silver Slugger* y lideró MLB en: BB, BA, OBP, SLG, OPS, OPS+, TB y IBB. Este es el dato anómalo que aparece en el gráfico de dispersión de salario y OPS; es el OPS más alto de la historia en MLB para una sola temporada.

Todas estas observaciones, tanto *outliers* como puntos de influencia, son mencionadas y destacadas para que el lector sepa que existen y que han sido identificadas apropiadamente y si se desea hacer un análisis propio pueda tomarlas en consideración. Para este trabajo, dichas observaciones solamente se comentan; no se eliminarán o se ajustarán de ninguna manera para los ajustes del modelo de regresión. Se toma esta decisión considerando que al solamente considerar jugadores calificados y temporadas regulares completas ya se evitaron los puntos más extremos de la información disponible.

Modelo Completo

Tras el análisis exploratorio y la primer determinación de variables con las que se trabajarán se hará una primer versión del modelo de regresión para explicar salarios con base en estadísticas de producción ofensiva. Recordemos que trabajamos con las

siguientes variables: G, R, 2B, 3B, HR, RBI, OPS, SO, BA, GDP y SF. Para esta primer versión del modelo se considerarán todas las variables, así como el intercepto; no se requiere ni se utiliza ningún método de selección del modelo para subconjuntos de regresores. Tras correr el modelo en SAS (véase Anexo 9) se obtiene lo siguiente:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	11	183.61135	16.69194	26.52	<.0001
Error	2081	1309.58109	0.62930		
Corrected Total	2092	1493.19244			

Tabla 4 ***Análisis de Varianza del Modelo Completo***

La Tabla 4 muestra el análisis de varianza del modelo en donde se prueba la significancia de la regresión. Al tener una estadística F grande, 26.52 y un p-valor menor a alfa (se hace con un alfa de 0.05) se concluye que si hay significancia en la regresión. Es decir que el modelo, inicialmente se ajusta bien a los datos y puede explicar significativamente el salario anual percibido por los jugadores.

Root MSE	0.79329	**R-Square**	0.1230
Dependent Mean	15.63467	**Adj R-Sq**	0.1183
Coeff Var	5.07389		

Tabla 5 ***Ajuste del Modelo Completo***

La Tabla 5 muestra distintas métricas de ajuste del modelo. Nos interesa tanto la R-cuadrada como la R-cuadrada ajustada. El valor de la R-cuadrada es de 0.1230, esto quiere decir que las variables regresoras en el modelo completo, o simplemente el modelo completo, explica la variabilidad en el salario anual percibido en un 12.30%. Este es un valor muy bajo; no es muy bueno el ajuste del modelo. El valor de R-cuadrada ajustada de 11.83% indica que no es muy bueno el ajuste del modelo y además hay regresores que no ayudan a explicar de mejor manera el modelo. Esto lo sabemos ya que el valor de R-cuadrada ajustada es menor al de la R-cuadrada.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	15.16108	0.39769	38.12	<.0001
G	G	1	-0.00518	0.00221	-2.34	0.0192
R	R	1	0.00909	0.00164	5.55	<.0001
2B	2B	1	-0.01217	0.00297	-4.10	<.0001
3B	3B	1	-0.05455	0.00799	-6.83	<.0001
HR	HR	1	-0.01040	0.00541	-1.92	0.0549
RBI	RBI	1	0.00475	0.00189	2.51	0.0121
SO	SO	1	-0.00067088	0.00071810	-0.93	0.3503
BA	BA	1	-4.00980	1.45886	-2.75	0.0060
OPS	OPS	1	2.15177	0.59922	3.59	0.0003
GDP	GDP	1	0.02004	0.00380	5.28	<.0001
SF	SF	1	0.01350	0.00827	1.63	0.1027

Tabla 6 ***Coeficientes de Regresión Modelo Completo***

La Tabla 6 indica que esta es la ecuación para el modelo de regresión completo:

$$Salario = 15.16108 - 0.00518 * G + 0.00909 * R - 0.01217 * 2B - 0.05455 * 3B - 0.01040 * HR + 0.00475 * RBI - 0.00067088 * SO - 4.00980 * BA + 2.15177 * OPS + 0.02004 * GDP + 0.01350 * SF$$

Ecuación 1 ***Modelo de Regresión Completo***

El valor del intercepto, G, R, 2B, 3B, RBI, BA, OPS y GDP son variables significativas al 100(1-0.05)% en este modelo completo. Las variables de HR, SO y SF no son significativas para explicar la variable salario. Esto se puede concluir observando los valores de la estadística t y los p-valores correspondientes de cada variable. Recordemos que los coeficientes miden el aumento en términos de logaritmo natural del salario. Por ejemplo, si todas las estadísticas productivas fuesen 0, es decir si un jugador se pierde toda la temporada, se esperaría que el logaritmo natural del salario anual sea de 15.16108. Para llegar a unidades del salario tenemos que evaluar este valor con la

función exponencial (lo mismo para los coeficientes de las variables regresoras) y esto nos da un salario de $3,840,372.788; dólares del 2012. Al estar tratando solamente con jugadores calificados, es imposible que se obtenga 0 en los parámetros, pero es importante considerar que, en una temporada de béisbol, un jugador puede estar fuera los 162 partidos y recibir su salario para esa temporada.

Análisis de Residuales Modelo Completo

Tras ajustar el modelo, es importante verificar que cumpla con los siguientes supuestos:

1. La relación entre la variable respuesta y los regresores es aproximadamente lineal.
2. La media del término del error es cero.
3. La varianza del término del error es constante.
4. Los errores están normalmente distribuidos.
5. Los errores son independientes.

El modelo cumple con el primer supuesto dado que al tener una estadística F grande, 26.52 y un p-valor menor a alfa (se hace con un alfa de 0.05), se concluye que si hay significancia en la regresión y se puede decir que el modelo si tiene una relación aproximadamente lineal. Para los supuestos 2, 3, 4 y 5 (y un poco más de detalle en el 1) se realiza un análisis de residuales.

Para verificar el supuesto 4 (y consecuentemente el 2) se realiza una prueba de normalidad sobre los residuales ajustándolos a una distribución normal con media igual a cero y varianza constante estimada por el sistema (véase Anexo 10). En este caso la prueba que se tiene es la siguiente:

$$H_0: Los\ residuales\ se\ distribuyen\ de\ manera\ normal\ con\ media\ 0$$

vs.

$$H_1: Los\ residuales\ no\ se\ distribuyen\ de\ manera\ normal\ con\ media\ 0$$

Ecuación 2 ***Prueba de Normalidad***

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0556229	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.9895592	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	13.2611890	**Pr > A-Sq**	<0.005

Tabla 7 ***Pruebas de Normalidad Residuales Modelo Completo***

La Tabla 7 muestra que todos los p-valores de las 3 pruebas de normalidad realizadas son menores al alfa de 0.05 con el que se realiza la prueba. Por lo tanto, se rechaza la hipótesis nula y se concluye que los residuales no se distribuyen normalmente con media 0. Este hecho viola los supuestos de los residuales para que el modelo de regresión y las conclusiones del mismo sean válidas. Para poder seguir trabajando se tendrá que normalizar la variable respuesta como se muestra en la ecuación 3.

$$Z = \frac{Y - \bar{Y}}{s_Y}$$

Ecuación 3 ***Normalización de la variable respuesta***

Modelo Completo Transformado

Al normalizar la variable respuesta, se tiene que volver a correr el modelo completo ya que dicha transformación afectará los valores de los coeficientes de la regresión. Al correr en SAS se obtiene lo siguiente (véase Anexo 11):

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	11	257.24409	23.38583	26.52	<.0001
Error	2081	1834.75591	0.88167		
Corrected Total	2092	2092.00000			

Tabla 8 ***Análisis de Varianza del Modelo Completo Transformado***

La Tabla 8 muestra el análisis de varianza del modelo transformado en donde se prueba la significancia de la regresión. Al tener una estadística F grande, 26.52 y un p-valor menor a alfa (se hace con un alfa de 0.05) se concluye que si hay significancia en la regresión. Es decir que el modelo, inicialmente se ajusta bien a los datos y puede explicar significativamente el salario anual percibido por los jugadores.

Root MSE	0.93897	**R-Square**	0.1230
Dependent Mean	-2.4672E-14	**Adj R-Sq**	0.1183
Coeff Var	-3.80581E15		

Tabla 9 ***Ajuste del Modelo Completo Transformado***

La Tabla 9 muestra distintas métricas de ajuste del modelo. Nos interesa tanto la R-cuadrada como la R-cuadrada ajustada. El valor de la R-cuadrada es de 0.1230, esto quiere decir que las variables regresoras en el modelo completo, o simplemente el modelo completo, explica la variabilidad en el salario anual percibido en un 12.30%. Este es un valor muy bajo; no es muy bueno el ajuste del modelo. El valor de R-cuadrada

ajustada de 11.83% indica que no es muy bueno el ajuste del modelo y además hay regresores que no ayudan a explicar de mejor manera el modelo. Esto lo sabemos ya que el valor de R-cuadrada ajustada es menor al de la R-cuadrada.

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-0.56057	0.47073	-1.19	0.2338
G	G	1	-0.00613	0.00262	-2.34	0.0192
R	R	1	0.01076	0.00194	5.55	<.0001
2B	2B	1	-0.01441	0.00352	-4.10	<.0001
3B	3B	1	-0.06457	0.00945	-6.83	<.0001
HR	HR	1	-0.01231	0.00641	-1.92	0.0549
RBI	RBI	1	0.00562	0.00224	2.51	0.0121
SO	SO	1	-0.00079409	0.00084998	-0.93	0.3503
BA	BA	1	-4.74620	1.72678	-2.75	0.0060
OPS	OPS	1	2.54695	0.70927	3.59	0.0003
GDP	GDP	1	0.02373	0.00450	5.28	<.0001
SF	SF	1	0.01598	0.00979	1.63	0.1027

Tabla 10 ***Coeficientes de Regresión Modelo Completo Transformado***

La Tabla 10 indica que esta es la ecuación para el modelo de regresión completo:

$$Salario = -0.56057 - 0.00613 * G + 0.01076 * R - 0.01441 * 2B - 0.06457 * 3B - 0.01231 * HR + 0.00562 * RBI - 0.00079409 * SO - 4.74620 * BA + 2.54695 * OPS + 0.02373 * GDP + 0.01598 * SF$$

Ecuación 4 ***Modelo de Regresión Transformado***

G, R, 2B, 3B, RBI, BA, OPS y GDP son variables significativas al 100(1-0.05)% en este modelo completo. Las variables del valor del intercepto, HR, SO y SF no son significativas para explicar la variable salario. Esto se puede concluir observando los valores de la estadística t y los p-valores correspondientes de cada variable. Recordemos que los

coeficientes miden el aumento en términos de logaritmo natural del salario y su consecuente normalización.

Análisis de Residuales Modelo Completo Transformado

Tras correr el modelo completo transformado, se deben verificar los supuestos del mismo y de los residuales. El modelo cumple con el primer supuesto dado que al tener una estadística F grande, 26.52 y un p-valor menor a alfa (se hace con un alfa de 0.05), se concluye que si hay significancia en la regresión y se puede decir que el modelo si tiene una relación aproximadamente lineal.

Se realizan en SAS (véase Anexo 12) las pruebas de normalidad, según la Ecuación 2, para el modelo completo transformado:

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0556229	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.9895592	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	13.2611890	**Pr > A-Sq**	<0.005

Tabla 11 ***Pruebas de Normalidad Residuales Modelo Completo Transformado***

La Tabla 11 muestra que todos los p-valores de las 3 pruebas de normalidad realizadas son menores al alfa de 0.05 con el que se realiza la prueba. Por lo tanto, se rechaza la hipótesis nula y se concluye que los residuales no se distribuyen normalmente con media 0. Este hecho viola los supuestos de los residuales para que el modelo de regresión y las conclusiones del mismo sean válidas. Es decir, que, si se desea usar este modelo de regresión para explicar o predecir, los resultados que se obtengan deben utilizarse y analizarse con cautela. Es importante reportar que los valores de los estadísticos son idénticos a los reportados en la Tabla 7. Se continuará con el análisis de residuales para verificar los otros supuestos de los residuales mencionados en la sección de Análisis de Residuales Modelo Completo. Se deben revisar los supuestos 2, 3 y 5; se hará un análisis gráfico en SAS (véase Anexo 12).

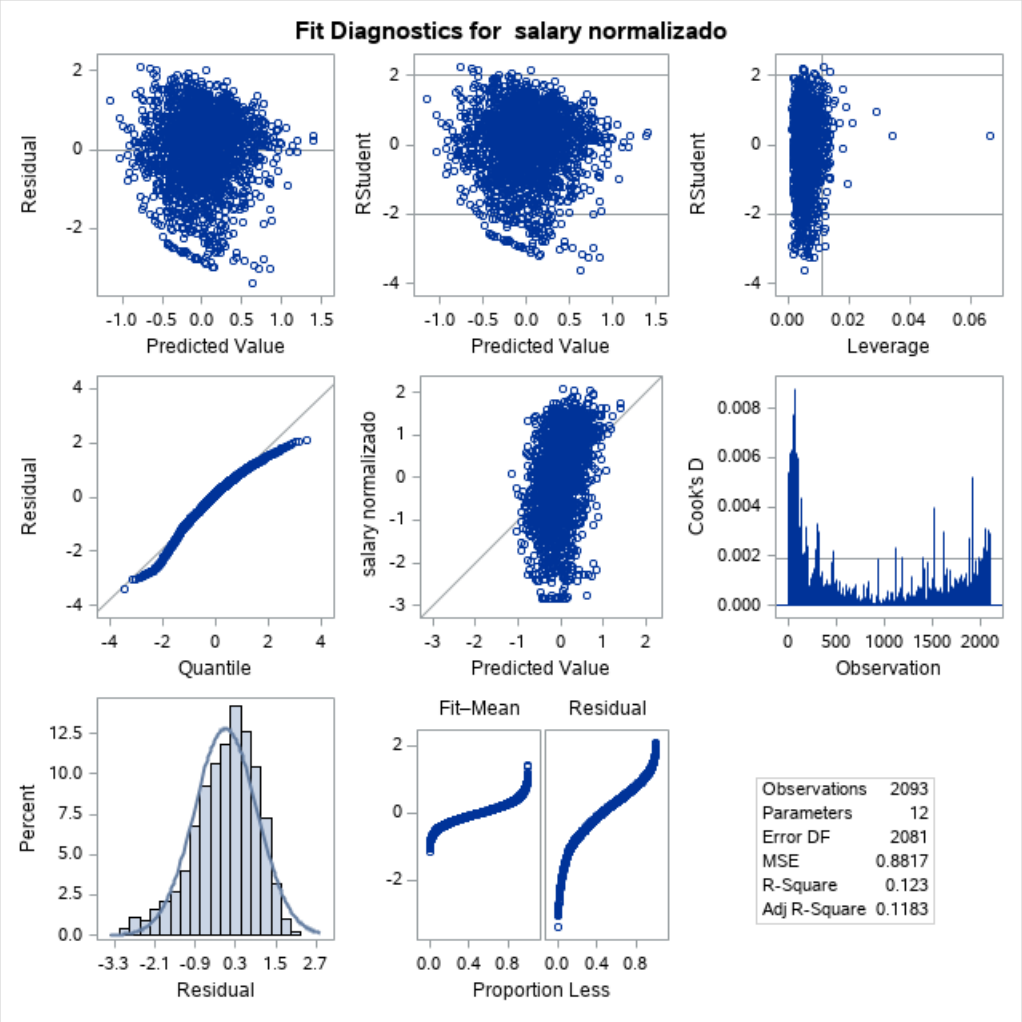


Figura 4 ***Diagnostico de Residuales del Modelo Completo Transformado***

La Figura 4 muestra un diagnóstico de residuales y ajuste para el modelo completo transformado. En la primer gráfica de la figura (*Residual vs. Predicted Value*) los residuales parecen estar dispersos aleatoriamente alrededor del 0. La varianza de los residuales es constante de izquierda a derecha en esta misma gráfica y no parece que los errores sigan algún patrón o tendencia. El estar dispersos aleatoriamente alrededor del 0 indica que los errores si tienen media cero y se cumple el supuesto 2. La varianza constante cumple el supuesto 3 y al no seguir tendencias o patrones se cumple con el supuesto 5 de independencia.

Este análisis se debe aplicar también a los residuales para cada variable regresora. Para ver estos gráficos de dispersión véase el Anexo 12. Para todos los regresores, así como en el modelo completo transformado, los residuales están muy concentrados en una parte del gráfico de dispersión. Todos los errores de las variables están distribuidos aleatoriamente alrededor del 0; es decir que se cumple con el supuesto 2 y se concluye que estos errores tienen media igual a 0. Igualmente, todos los residuales parecen tener una varianza constante, cumpliendo con el supuesto 3. Sería interesante investigar un poco más a fondo acerca de la varianza de los residuales de las variables regresoras de

G y 3B ya que sus gráficos de dispersión tienen leves indicios visuales de una posible varianza no constante. Ninguno de los residuales de las variables regresoras parece seguir algún patrón o tendencia y por lo tanto se cumple que estos son independientes (supuesto número 5).

Multicolinealidad en el Modelo Completo Transformado

Tras correr el modelo completo es importante revisar la posible existencia de multicolinealidad en el mismo. En este paso del desarrollo del modelo se buscan eliminar variables regresoras que tengan un alto grado de correlación entre ellas; así eliminando posibles redundancias de información en el modelo. Para detectar multicolinealidad se utilizarán factores de inflación de varianza o VIFs por sus siglas en inglés. El criterio que se usará para determinar si una variable contribuye a la multicolinealidad será usando lo siguiente:

$$Eliminar\ si: VIF > 10$$

Ecuación 5 ***Criterio de determinación de multicolinealdiad***

Se utiliza SAS (véase Anexo 13) para calcular los VIFs de las variables regresoras:

Variable	Label	Variance Inflation
Intercept	Intercept	0
G	G	1.93009
R	R	2.91161
2B	2B	1.78809
3B	3B	1.42695
HR	HR	10.59717
RBI	RBI	5.84071
SO	SO	1.93716
BA	BA	5.51756
OPS	OPS	10.85189
GDP	GDP	1.40000
SF	SF	1.31923

Tabla 12 ***VIFS Modelo Completo Transformado***

De la Tabla 12 y siguiendo la regla de eliminación de variables basado en la Ecuación 5 se eliminan las variables HR (10.597) y OPS (10.852) del modelo completo transformado para correr un nuevo modelo actualizado. Sin embargo, OPS es la estadística más completa para la explicación de la producción ofensiva de un jugador en una temporada y por lo tanto esta variable se mantendrá en el modelo actualizado. Además de revisar VIFs es importante considerar las correlaciones que existen entre las variables regresoras; para esto se genera la siguiente matriz de correlación (véase Anexo 14 para el código en SAS):

Pearson Correlation Coefficients, N = 2093											
	G	R	2B	3B	HR	RBI	SO	BA	OPS	GDP	SF
G	1.00000	0.45579	0.39497	0.13305	0.25643	0.38826	0.23392	0.09859	0.13212	0.17584	0.18601
R	0.45579	1.00000	0.46642	0.25705	0.52616	0.52897	0.19123	0.46804	0.66014	-0.05682	0.11756
2B	0.39497	0.46642	1.00000	0.04314	0.20062	0.40845	-0.03293	0.48038	0.44112	0.18826	0.21236
3B	0.13305	0.25705	0.04314	1.00000	-0.21588	-0.18455	-0.06036	0.14140	-0.03013	-0.26775	-0.08188
HR	0.25643	0.52616	0.20062	-0.21588	1.00000	0.82891	0.49277	0.08886	0.73610	0.06009	0.15318
RBI	0.38826	0.52897	0.40845	-0.18455	0.82891	1.00000	0.29380	0.30600	0.71958	0.23665	0.36927
SO	0.23392	0.19123	-0.03293	-0.06036	0.49277	0.29380	1.00000	-0.37674	0.13251	-0.18183	-0.05445
BA	0.09859	0.46804	0.48038	0.14140	0.08886	0.30600	-0.37674	1.00000	0.64530	0.19423	0.09229
OPS	0.13212	0.66014	0.44112	-0.03013	0.73610	0.71958	0.13251	0.64530	1.00000	0.05812	0.12354

Pearson Correlation Coefficients, N = 2093											
	G	R	2B	3B	HR	RBI	SO	BA	OPS	GDP	SF
GDP	0.17584	-0.05682	0.18826	-0.26775	0.06009	0.23665	-0.18183	0.19423	0.05812	1.00000	0.13599
SF	0.18601	0.11756	0.21236	-0.08188	0.15318	0.36927	-0.05445	0.09229	0.12354	0.13599	1.00000

Tabla 13 ***Matriz de Correlaciones entre variables regresoras***

La Tabla 13 muestra la matriz de correlación entre las variables regresoras, en rojo se marcan las correlaciones cuyo valor es mayor a 0.5. De esta matriz destacan las correlaciones altas que tienen las variables de R, HR, RBI y OPS con las demás variables. Previamente ya se trataron tanto HR y OPS. Al tener correlaciones elevadas con 3 otras variables se deciden eliminar R y RBI.

Modelo Actualizado

Tras la eliminación de variables del modelo completo transformado nos quedamos con las siguientes variables regresoras: G, 2B, 3B, SO, BA, OPS, GDP Y SF. Estas 8 variables regresoras serán utilizadas para ajustar un modelo actualizado. Para dicho ajuste se utilizará un método de selección de subconjuntos de variables regresoras, así como un criterio de selección determinado.

Para el método de selección se utilizará el procedimiento *stepwise selection* que combina tanto *forward selection* como *backward selection.* Dado que tanto la R-cuadrada y la R-cuadrada ajustada son muy bajas para el modelo original, y se desea buscar un mejor ajuste y mejor explicación de la variabilidad de los salarios por parte del modelo, se utilizará el criterio de la R-cuadrada ajustada para añadir y eliminar variables y determinar el mejor modelo actualizado. Este proceso se hace de manera automatizada en SAS (véase Anexo 15):

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square
* Optimal Value of Criterion				
0	Intercept		1	0.0000
1	OPS		2	0.0552
2	3B		3	0.0878

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square
3	GDP		4	0.0956
4	2B		5	0.1005
5	SF		6	0.1040
6	BA		7	0.1056*

Tabla 14 ***Stepwise Selection con criterio de R-cuadrada Ajustada***

La Tabla 14 muestra los pasos y los regresores añadidos y eliminados durante el proceso de stepwise para la selección del modelo (para ver esta selección gráficamente véase el Anexo 15) que maximice la R-cuadrada ajustada. Se selecciona un modelo de 6 variables regresoras más el intercepto con una R-cuadrada ajustada de 0.1056. Las variables que no se consideran son G y SO. Es importante notar que, aunque se hizo lo estadísticamente apropiado al eliminar la multicolinealidad en el modelo, el modelo completo transformado tiene una R-cuadrada mayor a la de este modelo actualizado.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	226.32716	37.72119	42.18	<.0001
Error	2086	1865.67284	0.89438		
Corrected Total	2092	2092.00000			

Tabla 15 ***Análisis de Varianza Modelo Actualizado***

La Tabla 15 muestra el análisis de varianza del modelo en donde se prueba la significancia de la regresión. Al tener una estadística F grande, 42.18 y un p-valor menor a alfa (se hace con un alfa de 0.05) se concluye que si hay significancia en la regresión. Es decir que el modelo, inicialmente se ajusta bien a los datos y puede explicar significativamente el salario anual percibido por los jugadores. Dado el valor de estadística F, este modelo es más significativo que el modelo completo (F de 26.52).

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-1.782724	0.216332	-8.24	<.0001
2B	1	-0.010572	0.003154	-3.35	0.0008
3B	1	-0.050237	0.008643	-5.81	<.0001
BA	1	-2.349337	1.062786	-2.21	0.0272
OPS	1	3.108150	0.297943	10.43	<.0001
GDP	1	0.021656	0.004196	5.16	<.0001
SF	1	0.025791	0.008860	2.91	0.0036

Tabla 16 ***Coeficientes de Regresión Modelo Actualizado***

La Tabla 16 indica que esta es la ecuación para el modelo de regresión completo:

$$Salario = -1.782724 - 0.010572 * 2B - 0.050237 * 3B - 2.349337 * BA + 3.108150 * OPS + 0.021656 * GDP + 0.025791 * SF$$

Ecuación 6 ***Modelo de Regresión Actualizado***

Todas las variables y el intercepto son significativas al 100(1-0.05)% en este modelo completo. Esto se puede concluir observando los valores de la estadística t y los p-valores correspondientes de cada variable. Recordemos que los coeficientes miden el aumento en términos de logaritmo natural del salario y su consecuente normalización. La ecuación 6 se explicará detalladamente en las conclusiones.

Análisis de Residuales Modelo Actualizado

Tras correr y seleccionar el modelo actualizado, se deben verificar los supuestos del mismo y de los residuales. El modelo cumple con el primer supuesto dado que al tener una estadística F grande, 42.18 y un p-valor menor a alfa (se hace con un alfa de 0.05), se concluye que si hay significancia en la regresión y se puede decir que el modelo si tiene una relación aproximadamente lineal.

Se realizan en SAS (véase Anexo 16) las pruebas de normalidad, según la Ecuación 2, para el modelo actualizado:

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.0535762	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.9449186	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	13.0405696	**Pr > A-Sq**	<0.005

Tabla 17 ***Pruebas de Normalidad Residuales Modelo Actualizado***

La Tabla 17 muestra que todos los p-valores de las 3 pruebas de normalidad realizadas son menores al alfa de 0.05 con el que se realiza la prueba. Por lo tanto, se rechaza la hipótesis nula y se concluye que los residuales no se distribuyen normalmente con media 0. Este hecho viola los supuestos de los residuales para que el modelo de regresión y las conclusiones del mismo sean válidas. Al igual que con el modelo de regresión completo, si se desea utilizar este modelo actualizado para explicar o hacer predicciones o algún otro efecto, sus resultados deben tomarse e interpretarse con cautela. Se continuará con el análisis de residuales para verificar los otros supuestos de los residuales mencionados en la sección de Análisis de Residuales Modelo Completo. Se deben revisar los supuestos 2, 3 y 5; se hará un análisis gráfico en SAS (véase Anexo 16).

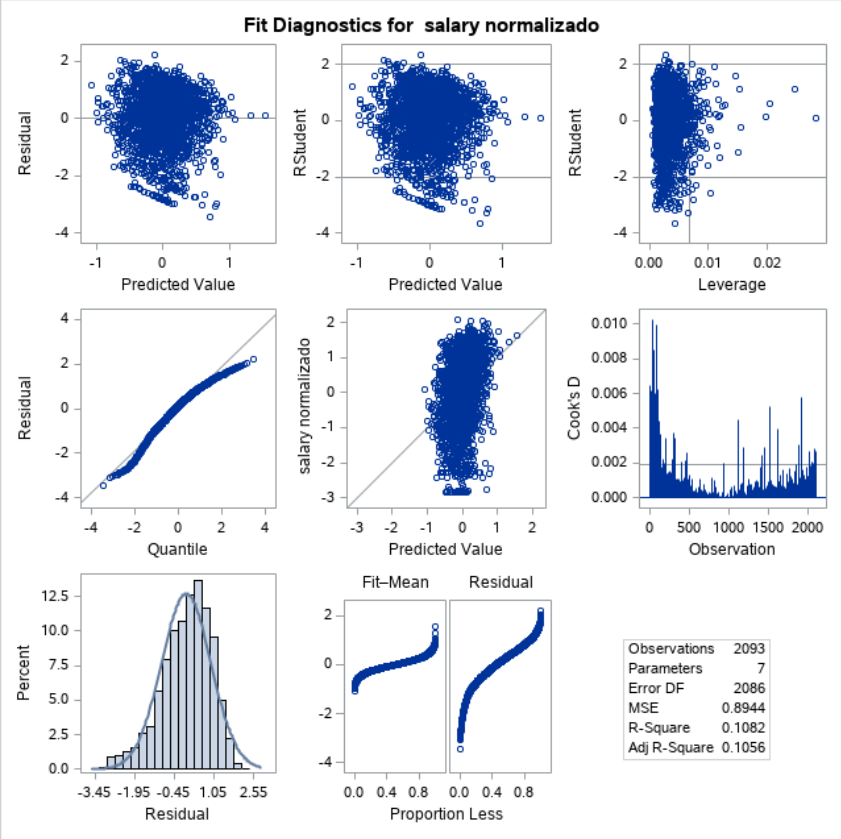


Figura 5 ***Diagnóstico de Residuales Modelo Actualizado***

La Figura 5 muestra un diagnóstico de residuales y ajuste para el modelo actualizado. La gráfica de arriba a la derecha (*Residual vs. Predicted Value*) muestra que los residuales parecen estar dispersos aleatoriamente alrededor del 0. La varianza de los residuales es constante de izquierda a derecha en esta misma gráfica y no parece que los errores sigan algún patrón o tendencia. El estar dispersos aleatoriamente alrededor del 0 indica que los errores si tienen media cero y se cumple el supuesto 2. La varianza constante cumple el supuesto 3 y al no seguir tendencias o patrones se cumple con el supuesto 5 de independencia.

Este análisis se debe aplicar también a los residuales para las seis variables regresoras. Para ver estos gráficos de dispersión véase el Anexo 16. Para todos los regresores, así como en el modelo completo transformado, los residuales están muy concentrados en una parte del gráfico de dispersión. Todos los errores de las variables están distribuidos aleatoriamente alrededor del 0; es decir que se cumple con el supuesto 2 y se concluye que estos errores tienen media igual a 0. Igualmente, todos los residuales parecen tener una varianza constante, cumpliendo con el supuesto 3. Sería interesante investigar un poco más a fondo acerca de la varianza de los residuales de la variable de 3B ya que su gráfico de dispersión tiene leves indicios visuales de una posible varianza no constante. Ninguno de los residuales de las variables regresoras parece seguir algún patrón o tendencia y por lo tanto se cumple que estos son independientes (supuesto número 5).

Multicolinealidad en el Modelo Actualizado

Se procede a revisar la posible existencia de multicolinealidad y consecuente eliminación de variables en el modelo actualizado. Esto se hace según el criterio establecido en la ecuación 5 (véase Anexo 17 para el código en SAS de los VIFs).

Parameter Estimates	
Variable	**Variance Inflation**
Intercept	0
2B	1.41886
3B	1.17629
BA	2.06039
OPS	1.88770
GDP	1.20237
SF	1.06541

Tabla 18 ***VIFs Modelo Actualizado***

La Tabla 18 muestra que los valores de los VIFs de todas las variables que componen el modelo son pequeñas y por lo tanto no existe multicolinealidad en el mismo. Todos los VIFS están por debajo del valor de corte de 10. Al no existir multicolinealidad, este modelo actualizado se considera el modelo final del trabajo.

Conclusiones

El objetivo del trabajo es intentar explicar el salario anual de un bateador de MLB a través de distintas estadísticas cuantitativas de su desempeño y producción ofensiva. Tras la definición del problema y el análisis exploratorio de las estadísticas ofensivas, y la selección y eliminación de variables regresoras y la subsecuente selección de distintos modelos se llegó al modelo descrito por la ecuación 6. Recordemos:

$$Salario = -1.782724 - 0.010572 * 2B - 0.050237 * 3B - 2.349337 * BA + 3.108150 * OPS + 0.021656 * GDP + 0.025791 * SF$$

Ecuación 7 ***Modelo de Regresión Actualizado***

En la ecuación 6 se describe lo siguiente. Es importante recordar que todas las estadísticas en una temporada empiezan en 0 y van aumentando una unidad o entero a la vez (2B, 3B, GDP y SF) o cambiando dentro de un rango a partir del 0 (BA y OPS); véase Anexo 5 para más sobre esto. Por cada 2B que pegue un jugador se espera que su salario disminuya en 0.010572 unidades y por cada 3B que pegue, que el salario disminuya 0.050237 unidades. Por cada aumento en una unidad de cambio en el BA se espera que el salario disminuya 2.349337 unidades. Por cada unidad positiva de cambio en su OPS se espera que el salario anual del jugador aumente 3.108150 unidades. Por cada GDP que genere, su salario anual aumenta en 0.021656 unidades. Por último, por cada SF que batee se espera que su salario anual aumente 0.025791 unidades. Para la correcta interpretación del modelo es importante recordar que la variable respuesta del salario en el modelo actualizado en realidad representa la normalización del logaritmo natural del salario. Se deben hacer las inversas de estas transformaciones para llegar a una figura correcta del salario anual percibido.

Para interpretar al intercepto se debe pensar en el caso en el que todas las estadísticas productivas sean iguales a 0, es decir si un jugador se pierde toda la temporada, se esperaría que la variable transformada del salario sea -1.782724. Para llegar a unidades del salario se tiene que des-normalizar la variable y luego evaluarla en la función exponencial. Esto nos da un salario de $1,367,572.788 dólares de 2012. Utilizando el factor de conversión del Anexo 18 esto equivale a $1,746,738.07 en 2022. Al estar tratando solamente con jugadores calificados, es imposible que se obtenga 0 en los parámetros, pero es importante considerar que, en una temporada de béisbol, un jugador puede estar fuera los 162 partidos y recibir su salario para esa temporada.

De la Tabla 15 recordemos que este modelo es significativo en la regresión (dada la estadística F y el p-valor) y por lo tanto explica adecuadamente el salario transformado a través de las variables regresoras. La Tabla 16 indica, dado los valores de las estadísticas t y los p-valores de cada variable regresora y el intercepto, se concluye que todas son significativas.

Root MSE	0.94572
Dependent Mean	-2.4672E-14
R-Square	0.1082
Adj R-Sq	0.1056

Tabla 19 ***Ajuste del Modelo Actualizado***

La Tabla 19 muestra estadísticas de ajuste del modelo. Interesa tanto la R-cuadrada como la R-cuadrada ajustada. El valor de la R-cuadrada es de 0.1082, esto quiere decir que el modelo actualizado explica en un 10.82% la variabilidad de la variable respuesta. Este es un valor muy bajo; no es muy bueno el ajuste del modelo a los datos. El valor de R-cuadrada ajustada de 10.56% indica que no es muy bueno el ajuste del modelo. Estos dos valores nos indican que utilizar un modelo de regresión lineal no es el más adecuado para explicar la relación que existe entre la producción ofensiva de un jugador y el salario anual que recibe (salario transformado). Se sugiere para investigación futura intentar utilizar otros modelos para explicar esta relación. Por ejemplo, regresión local (LOESS) o regresión polinómica.

Es importante reiterar que la Tabla 17 indica que los residuales del modelo ajustado no cumplen con el supuesto de normalidad. Si se desea utilizar el modelo actualizado que se desarrolló en este trabajo sus resultados deben ser tomados e interpretados con mucha cautela.

Así como ya se propuso utilizar otros modelos para intentar explicar la relación de salario y producción ofensiva se sugiere que se intenten utilizar otras variables de producción ofensiva más avanzadas como WAR o WRC+. También se sugiere realizar un estudio con un giro distinto o de mayor alcance considerando variables como edad del jugador, equipo con el que juega y su tamaño de mercado, posición de campo y tipo de contrato.

Referencias

1. Lewis, M. M. (2003). *Moneyball: The art of winning an unfair game*. W.W. Norton.

2. Turvey, J. (2013 Fall). The Future of Baseball Contracts: A Look at the Growing Trend in Long-Term Contracts. *Baseball Research Journal*. Retrieved October 24, 2022, from https://sabr.org/journal/article/the-future-of-baseball-contracts-a-look-at-the-growing-trend-in-long-term-contracts/.

3.Rattner, N. (2021, December 6). *Baseball's record-setting free agency spending spree, in charts*. CNBC. Retrieved October 24, 2022, from https://www.cnbc.com/2021/12/06/baseballs-record-setting-free-agency-spending-spree-in-charts.html

4. Spotrac. (2022). *MLB*. Spotrac.com. Retrieved October 24, 2022, from https://www.spotrac.com/Mlb/

5. Brown, M. (2022, November 8). *How Major League Baseball could crack $11 billion in revenues in 2022*. Forbes. Retrieved November 11, 2022, from https://www.forbes.com/sites/maurybrown/2022/04/07/how-major-league-baseball-could-crack-11-billion-in-revenues-in-2022/?sh=78a59f027f63

6. Ozanian, M. (2022, September 22). *Baseball's most valuable teams 2022: Yankees hit $6 billion as new CBA creates new revenue streams*. Forbes. Retrieved October 20, 2022, from https://www.forbes.com/sites/mikeozanian/2022/03/24/baseballs-most-valuable-teams-2022-yankees-hit-6-billion-as-new-cba-creates-new-revenue-streams/?sh=46d5169c600a

7. Sports Reference. (2022). *MLB stats, scores, History, & Records*. Baseball Reference. Retrieved October 12, 2022, from https://www.baseball-reference.com/

8. St. Lous FED Economic Research. (2022, November 10). *Consumer price index for all urban consumers: All items in U.S. city average*. FRED. Retrieved November 14, 2022, from https://fred.stlouisfed.org/series/CPIAUCSL#0

9. MLB. (2022). *Glossary*. MLB.com. Retrieved October 30, 2022, from https://www.mlb.com/glossary

10. Universal. (2011). *Moneyball*. Estados Unidos.

11. Magel, R., & Hoffman, M. (2015). Predicting Salaries of Major League Baseball Players. *International Journal of Sports Science*, *5*(2), 51–58. https://doi.org/10.5923/j.sports.20150502.02

12. Staudohar, P. D. (1997 Fall). *Baseball's changing salary structure*. Compensation and Working Conditions. Retrieved November 20, 2022, from https://www.bls.gov/opub/mlr/cwc/baseballs-changing-salary-structure.pdf

13. Meltzer, J. (2005, May). Average Salary and Contract Length in Major League Baseball: When Do They Diverge. *Stanford University Department of Economics,* 1-40. Retrieved November 15, 2022, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.650.468&rep=rep1&type=pdf

14. Stankiewicz, K. (2009). Length of Contracts and the Effect on the Performance of MLB Players. *The Park Place Economist, XVII*, 76-82. Retrieved November 15, 2022, from https://www-test.iwu.edu/economics/PPE17/stankiewicz.pdf.

15. Cahill, M. J. (2014). Change in Major League Baseball player performance after signing a Long-Term Deal. *Sport Management Undergraduate ,* 2-24. Retrieved November 6, 2022 from http://fisherpub.sjfc.edu/cgi/viewcontent.cgi?article=1082&context=sport_undergrad

16. Sturman, T. S., & Thibodeau, R. (2001). Performance-Undermining Effects of Baseball Free Agent Contracts. *Journal of Sport and Exercise Psychology,23*(1), 23-36. doi:10.1123/jsep.23.1.23

17. Judge, T. L. (2010, October). The Relationship between Pay and Job Satisfaction. *Journal of Vocational Behavior, 77*(2), 157-167. Retrieved November 15, 2022, from http://www.timothy-judge.com/Judge, Piccolo, Podsakoff, et al. (JVB 2010).pdf

Anexos

Anexo 1: Valuación de Franquicias de MLB según Forbes para la temporada 2022

Posición	Equipo	Ciudad (Mercado)	Valuación en miles de millones de dólares americanos
1	Yankees	Nueva York	6
2	Dodgers	Los Ángeles	4.075
3	Red Sox	Boston	3.9
4	Cubs	Chicago	3.8
5	Giants	San Francisco	3.5
6	Mets	Nueva York	2.65
7	Cardinals	San Luis	2.45
8	Phillies	Filadelfia	2.3
9	Angels	Los Ángeles/Anaheim	2.2
10	Braves	Atlanta	2.1
11	Rangers	Dallas – Fort Worth	2.05
12	Nationals	Washington D.C.	2
13	Astros	Houston	1.98
14	Blue Jays	Toronto	1.78
15	White Sox	Chicago	1.76
16	Mariners	Seattle	1.7
17	Padres	San Diego	1.575
18	Tigers	Detroit	1.4
19	Twins	Minneapolis – St. Paul	1.39
20	Rockies	Denver	1.385
21	Diamondbacks	Phoenix - Glendale	1.38
22	Orioles	Baltimore	1.375
23	Pirates	Pittsburgh	1.32
24	Guardians	Cleveland	1.3
25	Brewers	Milwaukee	1.28
26	Reds	Cincinnati	1.19
27	Athletics	Oakland	1.18
28	Royals	Kansas City	1.11
29	Rays	Tampa Bay – St. Petersburg	1.1
30	Marlins	Miami	0.990

Fuente: Forbes, 2022

Anexo 2: Nómina de equipos de MLB para el Impuesto de Lujo temporada 2022

Posición	Equipo	Nómina para el Impuesto (USD)
1	New York Mets	$313,728,454
2	Los Angeles Dodgers	$283,244,118
3	New York Yankees	$276,505,154
4	Philadelphia Phillies	$267,052,173
5	Boston Red Sox	$257,018,718
6	Atlanta Braves	$233,893,638
7	San Diego Padres	$233,672,865
8	Chicago White Sox	$228,417,890
9	Los Angeles Angels	$212,338,707
10	Houston Astros	$212,018,922
11	Toronto Blue Jays	$204,379,060
12	Chicago Cubs	$194,613,370
13	Colorado Rockies	$190,037,118
14	St. Louis Cardinals	$188,770,649
15	Minnesota Twins	$188,741,945
16	San Francisco Giants	$183,459,445
17	Texas Rangers	$181,691,784
18	Detroit Tigers	$175,341,045
19	Washington Nationals	$174,580,783
20	Milwaukee Brewers	$168,917,134
21	Seattle Mariners	$161,685,752
22	Cincinnati Reds	$141,451,245
23	Tampa Bay Rays	$133,466,946
24	Kansas City Royals	$126,207,790
25	Miami Marlins	$125,234,488
26	Arizona Diamondbacks	$121,158,582
27	Cleveland Guardians	$107,963,917
28	Baltimore Orioles	$93,782,150
29	Pittsburgh Pirates	$92,751,663
30	Oakland Athletics	$86,639,109

Fuente: Spotrac 2022

Anexo 3: Salario mínimo MLB por temporada 2003-2022

Temporada	Salario (USD)
2003	300,000
2004	300,000
2005	316,000
2006	327,000
2007	380,000
2008	390,000
2009	400,000
2010	400,000
2011	414,000
2012	480,000
2013	490,000
2014	500,000
2015	507,500
2016	507,500
2017	535,000
2018	545,000
2019	555,000
2020	563,500 (208,704)
2021	570,500
2022	700,000

Fuente: Baseball Reference 2022

Nota: en 2020 MLB utilizó salarios prorrateados al solamente llevarse a cabo 60 juegos por la pandemia de coronavirus. Una temporada normal tiene 162 juegos.

Anexo 4: Salario Promedio por posición MLB temporada 2022

Posición	Salario Promedio (USD)
Receptor	1,735,000
Primera Base	5,628,161
Segunda Base	3,791,500
Campo Corto	2,822,747
Tercera Base	6,065,122
Jardinero Izquierdo	4,504,611
Jardinero Central	3,213,757
Jardinero Derecho	8,245,555
Bateador Designado	9,416,667

Fuente: Spotrac 2022

Anexo 5: Diccionario de Datos

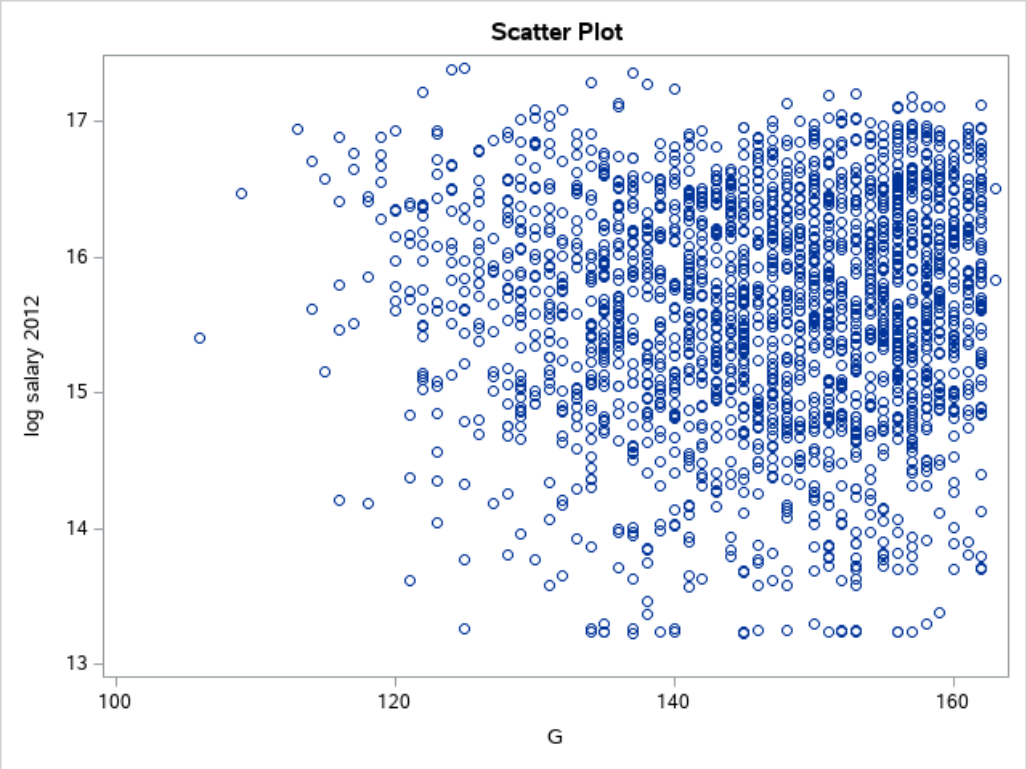
Name: Nombre del jugador
Age: Edad cumplida del jugador
Tm: Equipo al que pertenece el jugador
Lg: Liga a la que pertenece el equipo
G: Juegos jugados
PA: Apariciones al plato
AB: turnos al bate
R: Carreras anotadas
H: Imparables conectados
2B: Dobles conectados
3B: Triples conectados
HR: cuadrangulares conectados
RBI: carreras remolcadas
SB: bases robadas
CS: Intentos de robo de base fallidos
BB: bases por bolas
SO: ponches
BA: promedio de bateo (H/PA)
OBP: porcentaje en base
SLG: Bases por turno al bate, (TB/AB)
OPS: OBP + SLG
OPS+: OPS ajustado al estadio del jugador
TB: bases totales
GDP: rodados de doble-play
HBP: golpeado por lanzamiento
SH: toque de sacrificio
SF: elevado de sacrificio
IBB: Bases por bolas intencionales
Pos Summary: posición de campo jugada durante la temporada
Year: año o temporada
Salario: salario en dólares para la temporada
Salary in 2012 USD: salario en dólares ajustados al año base 2012
Log salary: ln(Salary) nota en el trabajo
Log salary 2022 ln(Salary in 2012 USD)

Anexo 6: Tabla de Correlación de Pearson en SAS

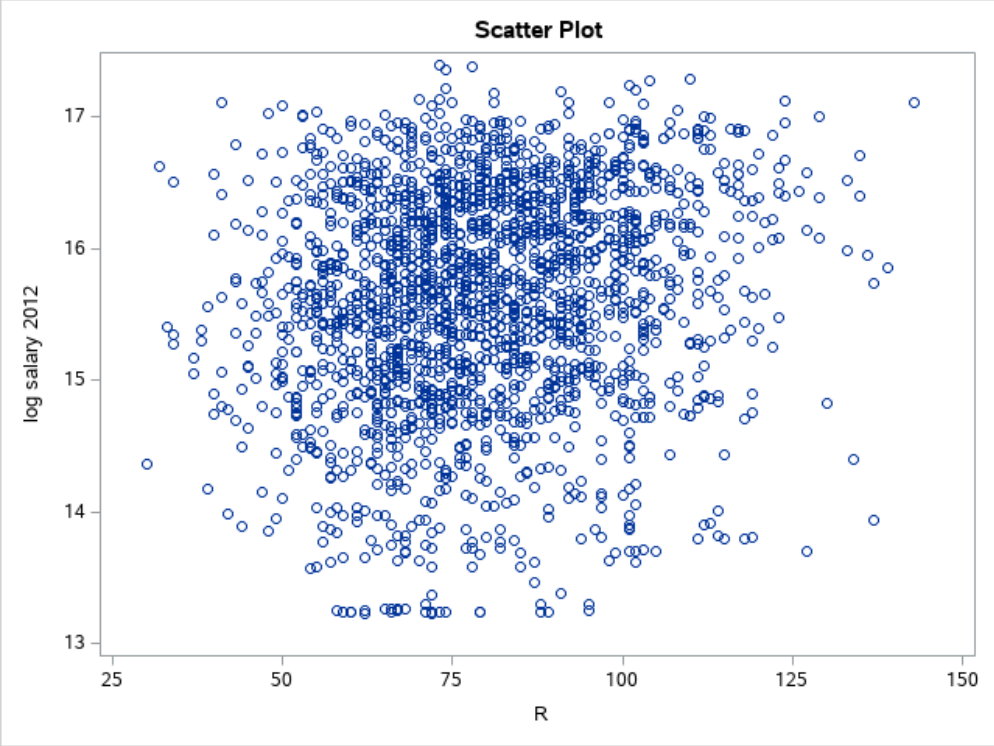
```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=PROYFIN.DATOS2 pearson nosimple noprob plots=none;
    var G R '2B'n '3B'n HR RBI SO BA OPS GDP SF;
    with 'log salary 2012 usd'n;
run;
```

Anexo 7: Gráficos de Dispersión Salario vs. Variables Regresoras

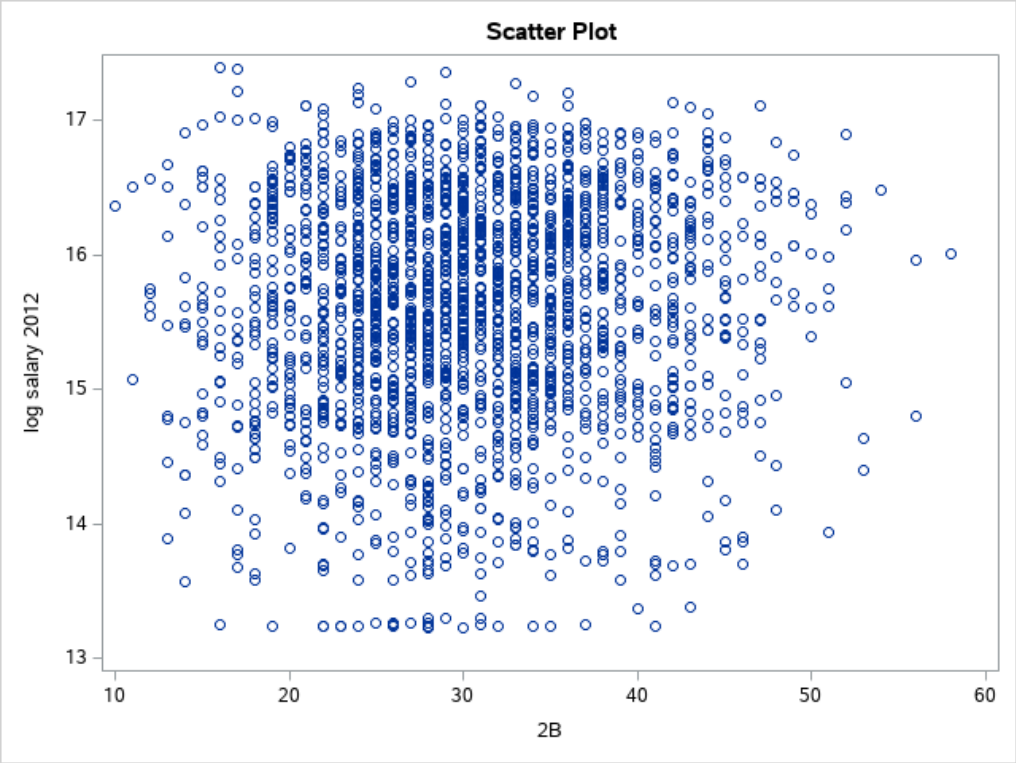
G vs. Salario



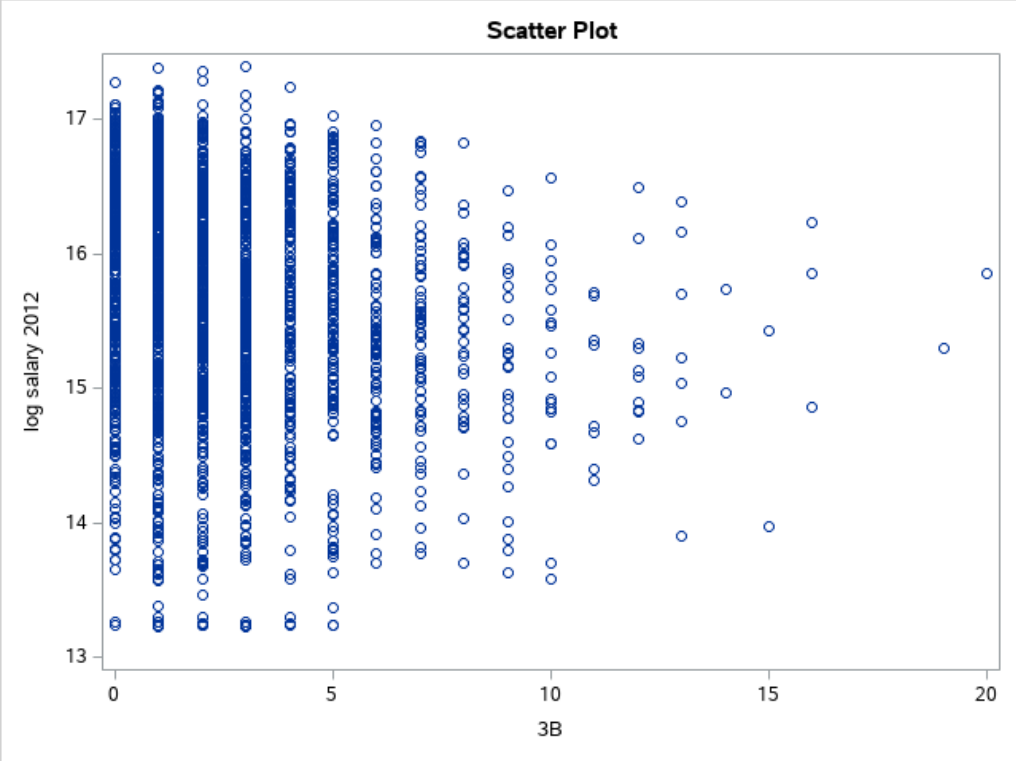
R vs. Salario



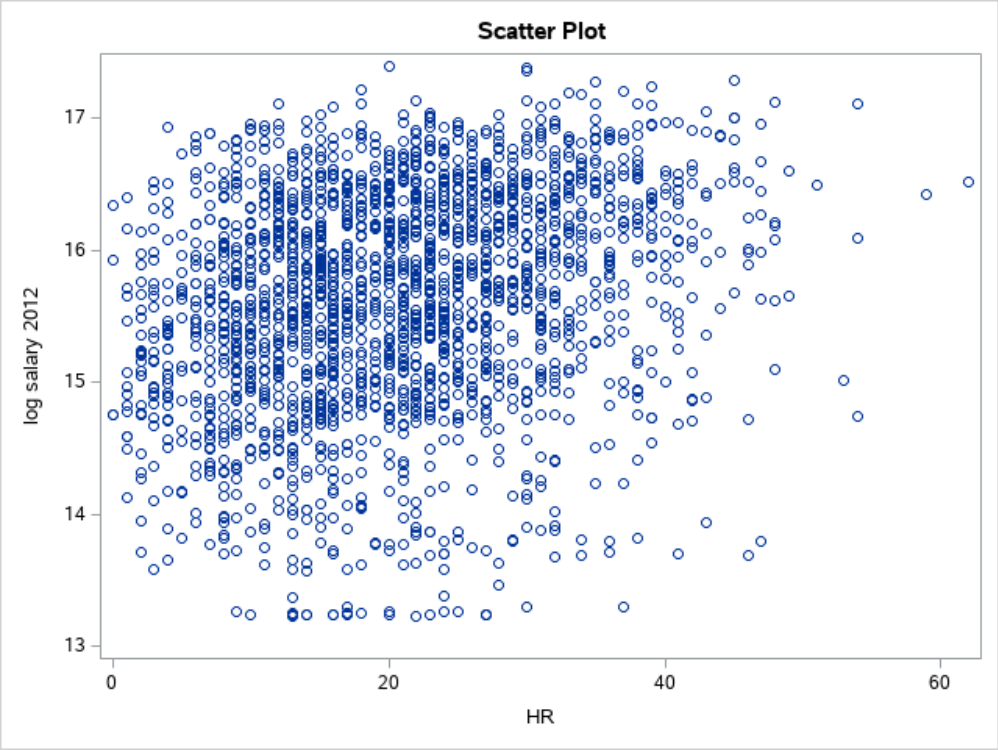
2B vs. Salario



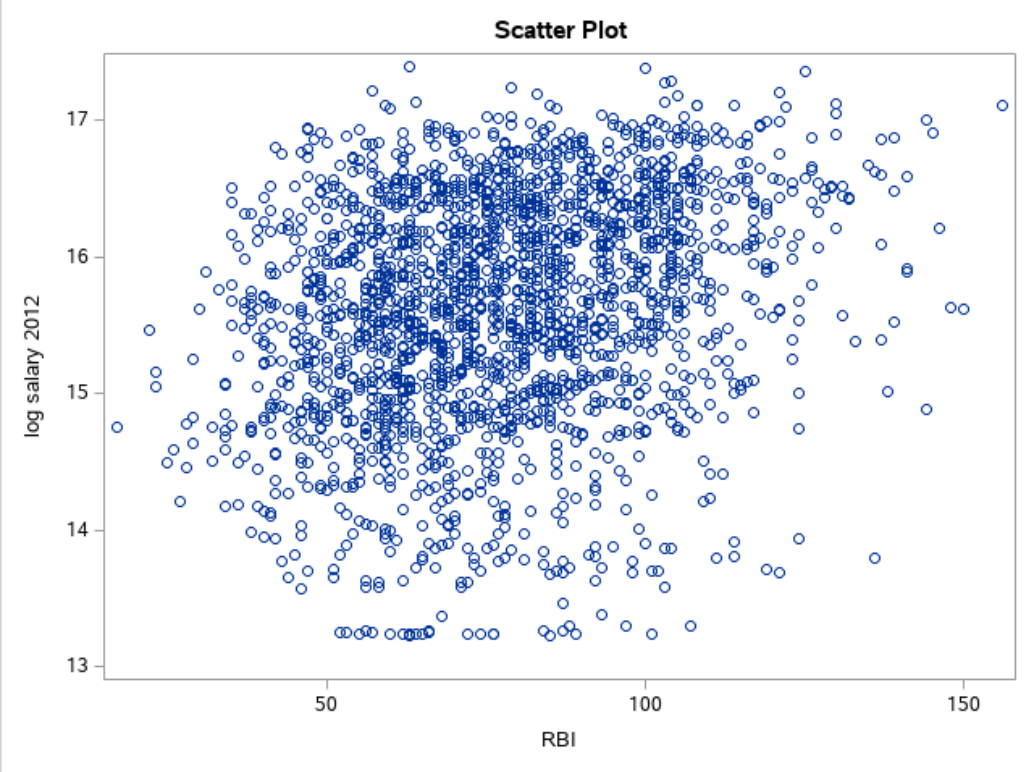
3B vs. Salario



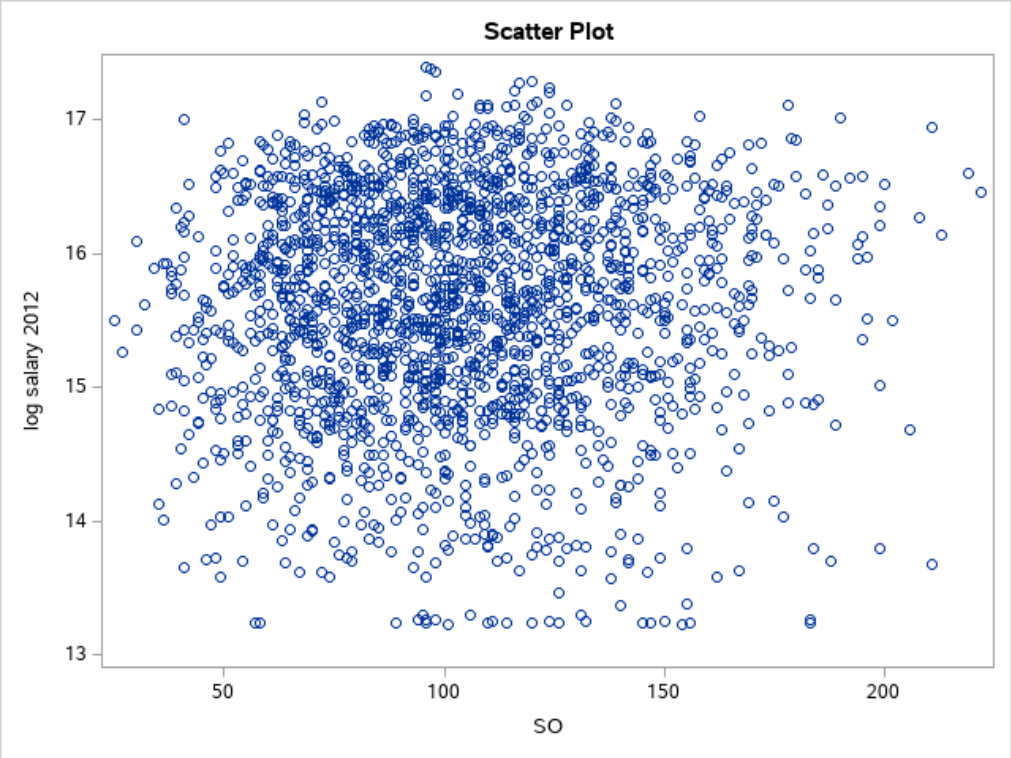
HR vs. Salario



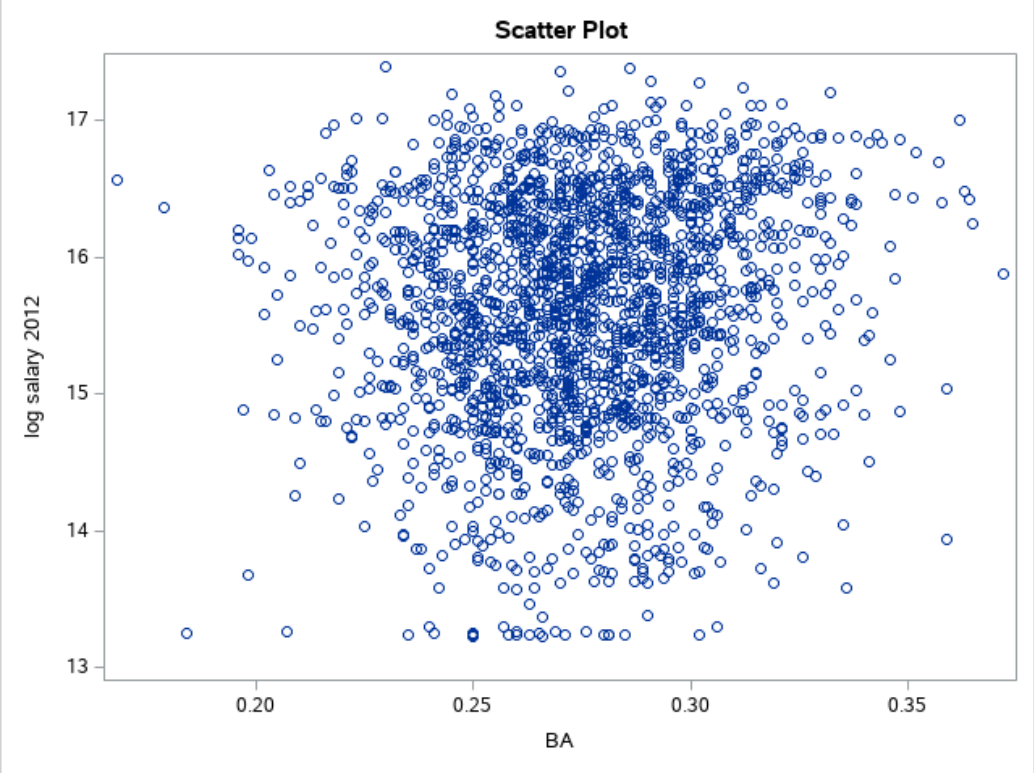
RBI vs. Salario



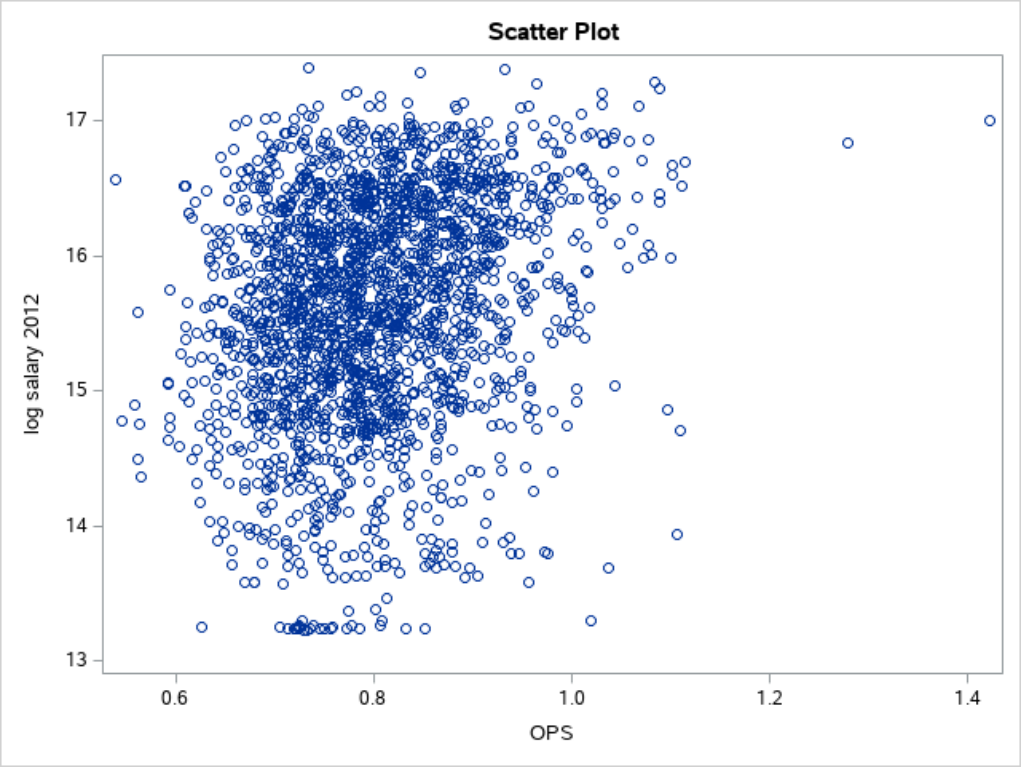
SO vs. Salario



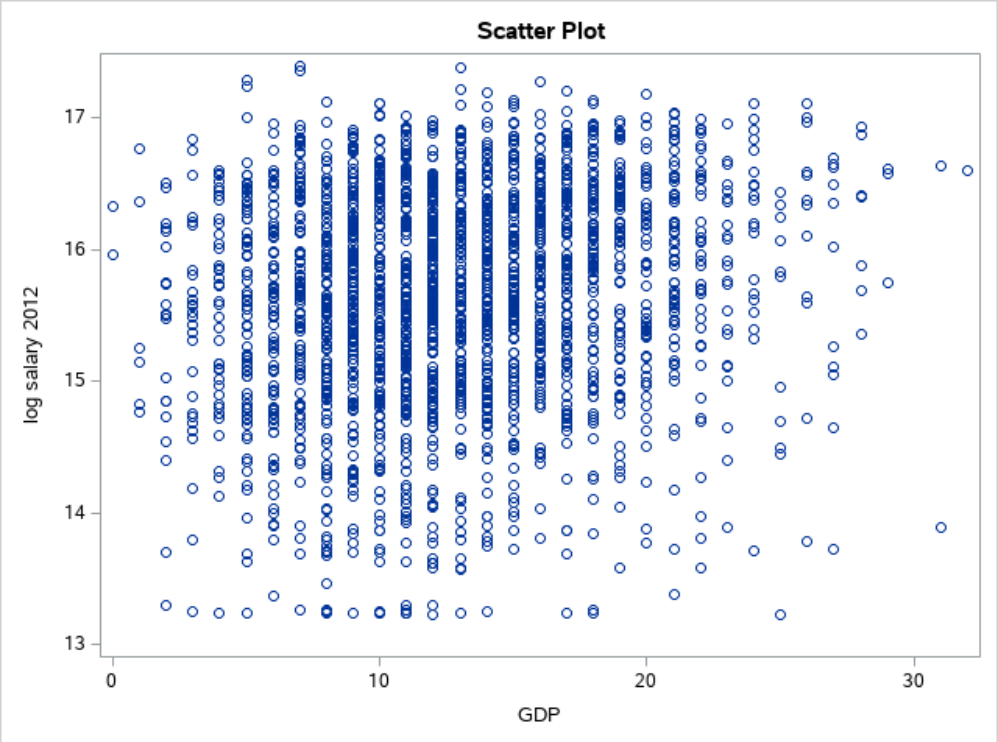
BA vs. Salario



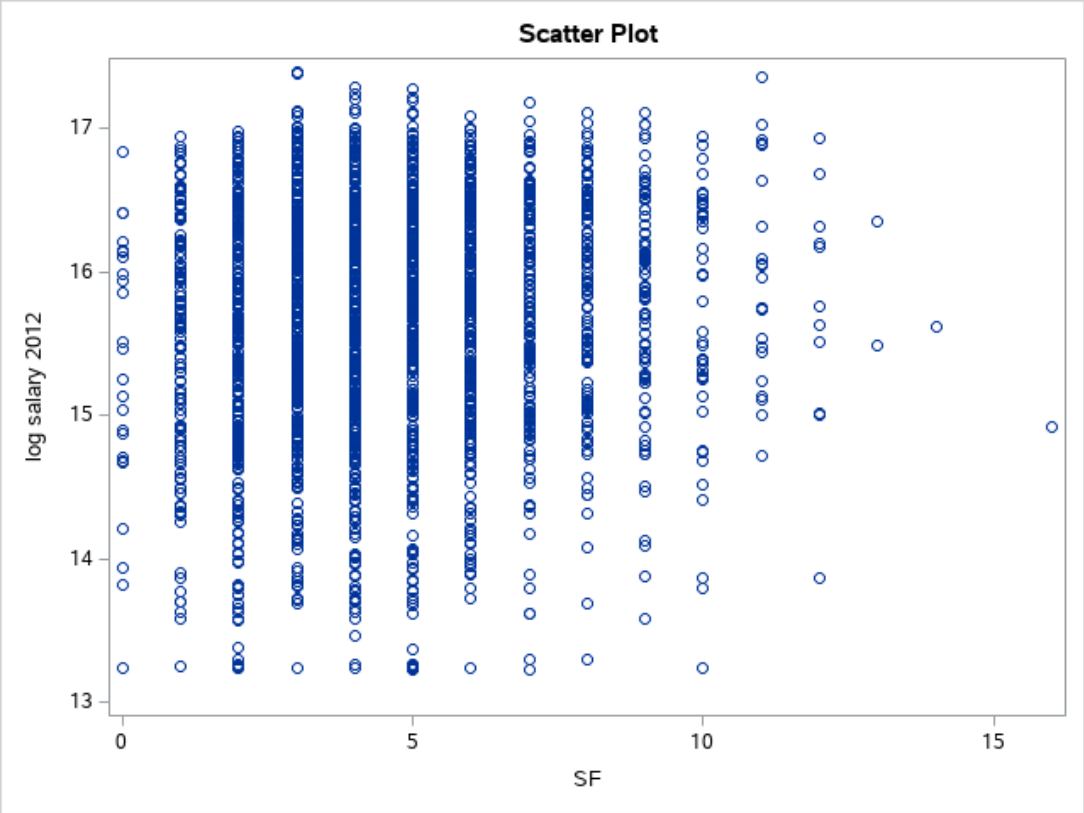
OPS vs. Salario



GDP vs. Salario



SF vs. Salario



Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=PROYFIN.DATOS2 nocorr nosimple noprob plots=scatter(noinset
        ellipse=none nvar=10 nwith=10);
    var G R '2B'n '3B'n HR RBI SO BA OPS GDP SF;
    with 'log salary 2012 usd'n;
run;
```

Anexo 8: Puntos de Influencia y Valores Atípicos

Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2
alpha=0.05
plots(only label)=(RStudentByLeverage CooksD);
    model 'log salary 2012'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF/r;
    run;
```

Anexo 9: Primer Modelo de Regresión

Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2 alpha=0.05 plots=none;
    model 'log salary 2012 usd'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF /;
    run;
quit;
```

Salida Completa:

Model: MODEL1
Dependent Variable: log salary 2012 usd log salary 2012 usd

Number of Observations Read	2093
Number of Observations Used	2093

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	34.63125	3.14830	26.52	<.0001
Error	2081	247.00231	0.11869		
Corrected Total	2092	281.63356			

Root MSE	0.34452	**R-Square**	0.1230
Dependent Mean	6.79005	**Adj R-Sq**	0.1183
Coeff Var	5.07389		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6.58437	0.17272	38.12	<.0001
G	G	1	-0.00225	0.00096051	-2.34	0.0192
R	R	1	0.00395	0.00071191	5.55	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
2B	2B	1	-0.00529	0.00129	-4.10	<.0001
3B	3B	1	-0.02369	0.00347	-6.83	<.0001
HR	HR	1	-0.00452	0.00235	-1.92	0.0549
RBI	RBI	1	0.00206	0.00082155	2.51	0.0121
SO	SO	1	-0.00029136	0.00031187	-0.93	0.3503
BA	BA	1	-1.74144	0.63358	-2.75	0.0060
OPS	OPS	1	0.93450	0.26024	3.59	0.0003
GDP	GDP	1	0.00871	0.00165	5.28	<.0001
SF	SF	1	0.00586	0.00359	1.63	0.1027

Anexo 10: Residuales Modelo Completo

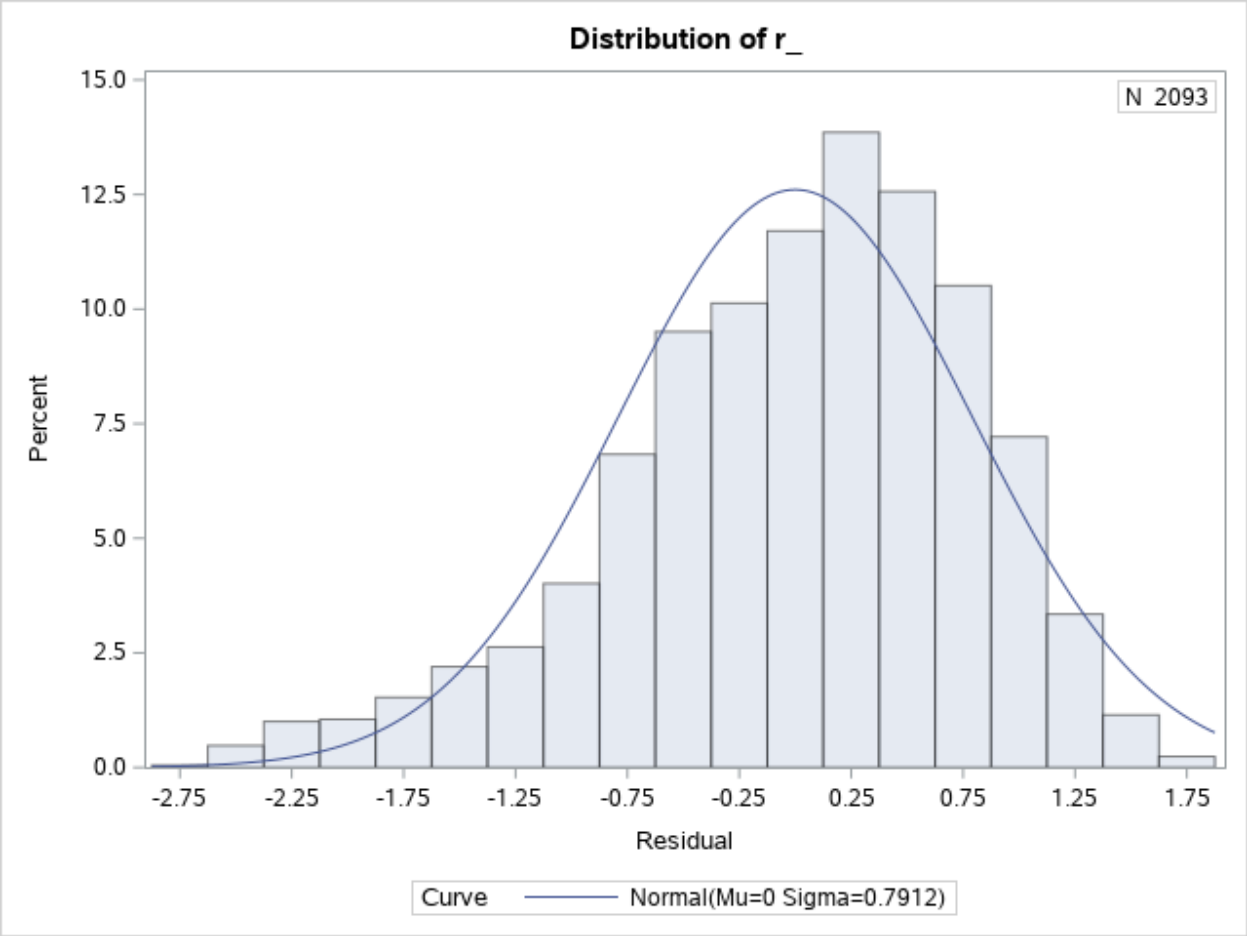
Código en SAS para pruebas de Normalidad:

```
ods noproctitle;
ods graphics / imagemap=on;

proc univariate data=WORK.REG_STATS;
    ods select Histogram GoodnessOfFit;
    var r_;
```

Código en SAS para generar residuales y gráficos:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2 alpha=0.05 plots(only)=(diagnostics residuals
        rstudentbypredicted);
    model 'log salary 2012'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF /;
    output out=work.Reg_stats r=r_;
    run;
quit;
```

Histograma Generado para la normalidad:



Anexo 11: Modelo de Regresión Completo Transformado

Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2 alpha=0.05 plots(only)=(rstudentbypredicted);
    model ' salary normalizado'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF /;
    output out=work.Reg_stats r=r_;
    run;
quit;
```

Salida Completa:

Model: MODEL1
Dependent Variable: salary normalizado salary normalizado

Number of Observations Read	2093
Number of Observations Used	2093

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	257.24409	23.38583	26.52	<.0001
Error	2081	1834.75591	0.88167		
Corrected Total	2092	2092.00000			

Root MSE	0.93897	**R-Square**	0.1230
Dependent Mean	-2.4672E-14	**Adj R-Sq**	0.1183
Coeff Var	-3.80581E15		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.56057	0.47073	-1.19	0.2338
G	G	1	-0.00613	0.00262	-2.34	0.0192
R	R	1	0.01076	0.00194	5.55	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
2B	2B	1	-0.01441	0.00352	-4.10	<.0001
3B	3B	1	-0.06457	0.00945	-6.83	<.0001
HR	HR	1	-0.01231	0.00641	-1.92	0.0549
RBI	RBI	1	0.00562	0.00224	2.51	0.0121
SO	SO	1	-0.00079409	0.00084998	-0.93	0.3503
BA	BA	1	-4.74620	1.72678	-2.75	0.0060
OPS	OPS	1	2.54695	0.70927	3.59	0.0003
GDP	GDP	1	0.02373	0.00450	5.28	<.0001
SF	SF	1	0.01598	0.00979	1.63	0.1027

Anexo 12: Residuales Modelo Completo Transformado

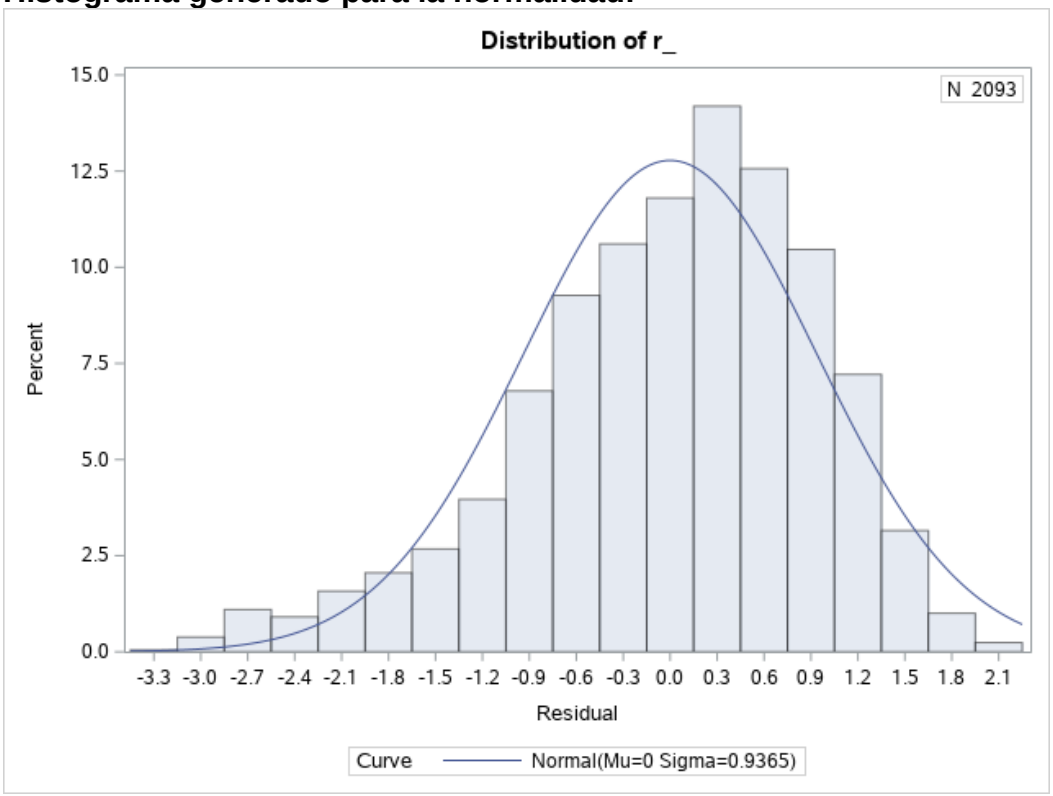
Código en SAS para pruebas de Normalidad:

```
ods noproctitle;
ods graphics / imagemap=on;

proc univariate data=WORK.REG_STATS;
    ods select Histogram GoodnessOfFit;
    var r_;

    /* Checking for Normality */
    histogram r_ / normal(mu=est sigma=est);
    inset n / position=ne;
run;
```

Código en SAS para generar residuales y gráficos:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2 alpha=0.05 plots(only)=(diagnostics residuals
        rstudentbypredicted);
    model ' salary normalizado'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF /;
    output out=work.Reg_stats r=r_;
    run;
quit;
```

Histograma generado para la normalidad:



Gráficos de Residuales para cada Variable Regresora

Anexo 13: VIFs Modelo Completo Transformado

Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=PROYFIN.DATOS2 alpha=0.05 plots=none;
    model ' salary normalizado'n=G R '2B'n '3B'n HR RBI SO BA OPS GDP SF / vif;
    run;
quit;
```

Salida Completa:

Model: MODEL1
Dependent Variable: salary normalizado salary normalizado

Number of Observations Read	2093
Number of Observations Used	2093

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	257.24409	23.38583	26.52	<.0001
Error	2081	1834.75591	0.88167		
Corrected Total	2092	2092.00000			

Root MSE	0.93897	**R-Square**	0.1230
Dependent Mean	-2.4672E-14	**Adj R-Sq**	0.1183
Coeff Var	-3.80581E15		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-0.56057	0.47073	-1.19	0.2338	0
G	G	1	-0.00613	0.00262	-2.34	0.0192	1.93009
R	R	1	0.01076	0.00194	5.55	<.0001	2.91161

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
2B	2B	1	-0.01441	0.00352	-4.10	<.0001	1.78809
3B	3B	1	-0.06457	0.00945	-6.83	<.0001	1.42695
HR	HR	1	-0.01231	0.00641	-1.92	0.0549	10.59717
RBI	RBI	1	0.00562	0.00224	2.51	0.0121	5.84071
SO	SO	1	-0.00079409	0.00084998	-0.93	0.3503	1.93716
BA	BA	1	-4.74620	1.72678	-2.75	0.0060	5.51756
OPS	OPS	1	2.54695	0.70927	3.59	0.0003	10.85189
GDP	GDP	1	0.02373	0.00450	5.28	<.0001	1.40000
S	SF	1	0.01598	0.00979	1.63	0.1027	1.31923

Anexo 14: Matriz de Correlaciones de Regresores Modelo Completo Transformado

Código en SAS:

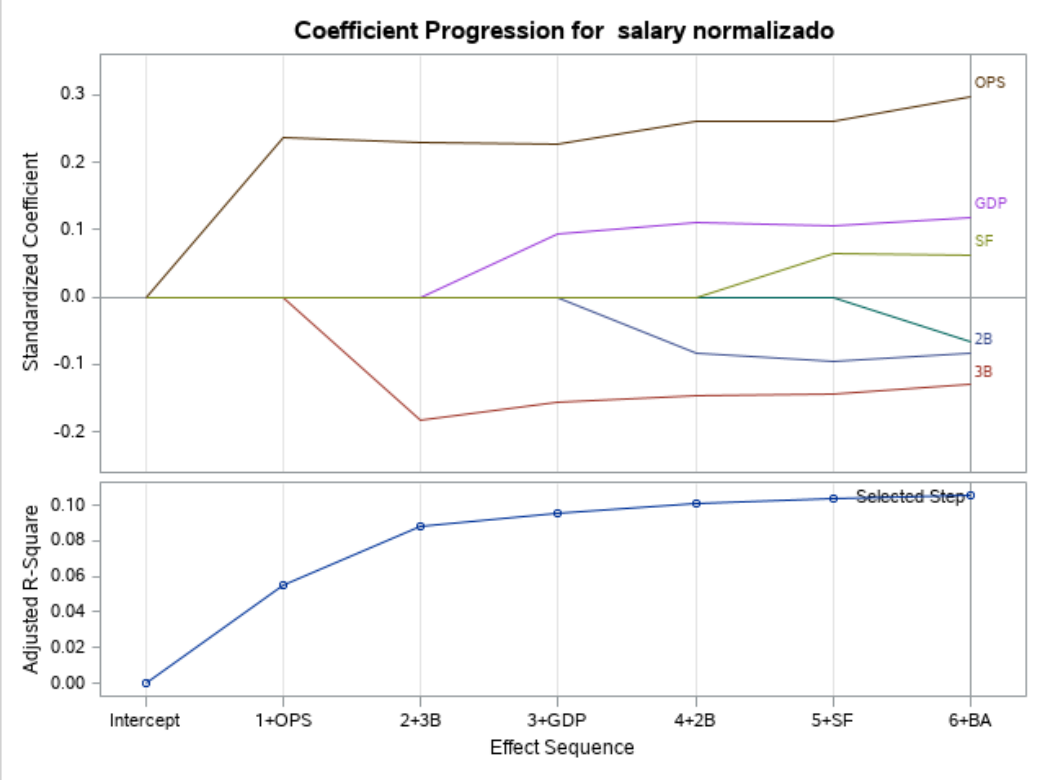
```
ods noproctitle;
ods graphics / imagemap=on;

proc corr data=PROYFIN.DATOS2 pearson nosimple noprob plots=none;
    var G R '2B'n '3B'n HR RBI SO BA OPS GDP SF;
run;
```

Anexo 15: Selección del Modelo Actualizado

Código en SAS:

```
ods noproctitle;
ods graphics / imagemap=on;

proc glmselect data=PROYFIN.DATOS2 outdesign(addinputvars)=Work.reg_design
        plots=(criterionpanel coefficientpanel);
    model ' salary normalizado'n=G '2B'n '3B'n SO BA OPS GDP SF / showpvalues
        selection=stepwise

   (select=adjrsq stop=adjrsq choose=adjrsq) stats=(adjrsq);
run;
```

Criterio de Selección Gráficamente:



Anexo 16: Pruebas de Normalidad Modelo Actualizado

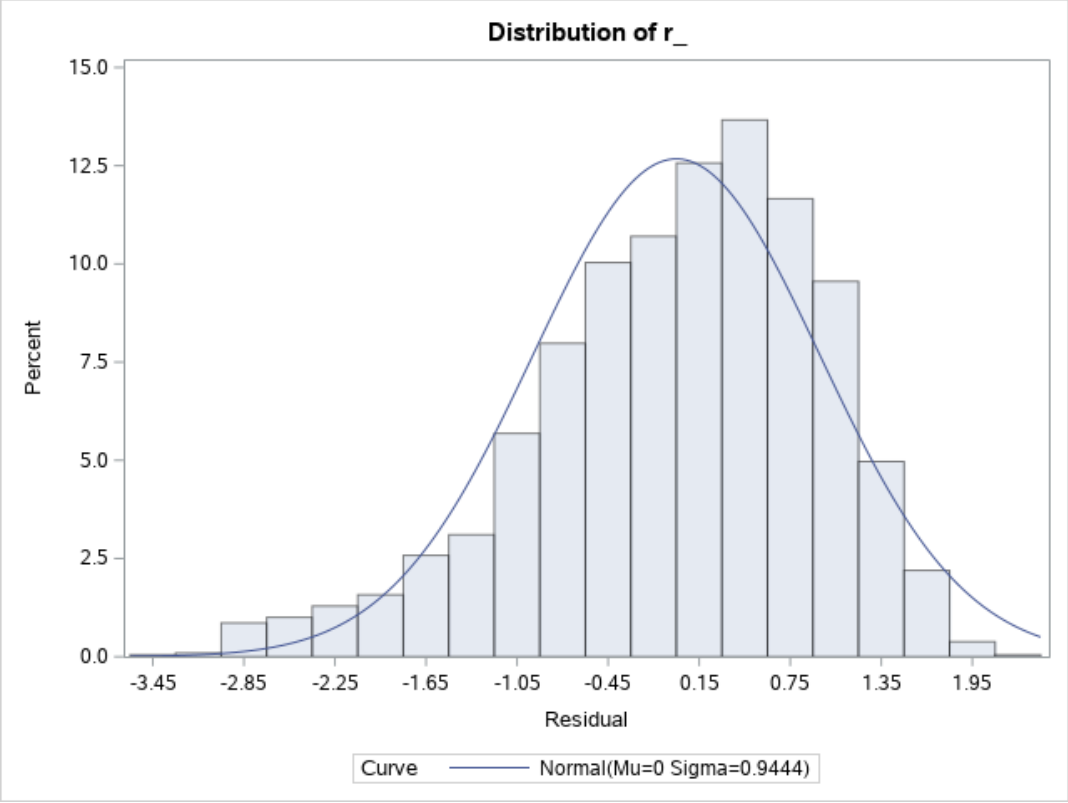
Código en SAS para generar residuales y diagnóstico:

```
proc reg data=Work.reg_design alpha=0.05 plots(only)=(rstudentbypredicted);
    ods select RStudentByPredicted;
    model ' salary normalizado'n=&_GLSMOD /;
    output out=work.Reg_stats r=r_;
    run;
quit;

proc delete data=Work.reg_design;
run;
```

Código en SAS para pruebas de normalidad:

```
ods noproctitle;
ods graphics / imagemap=on;

proc univariate data=WORK.REG_STATS;
    ods select Histogram GoodnessOfFit;
    var r_;

    /* Checking for Normality */
    histogram r_ / normal(mu=est sigma=est);
run;
```

Histograma generado:



Residuales de variables regresoras:

Anexo 17: VIFs Modelo Actualizado

Código en SAS:

```
proc reg data=Work.reg_design alpha=0.05 plots(only)=(rstudentbypredicted);
    ods select ParameterEstimates RStudentByPredicted;
    model ' salary normalizado'n=&_GLSMOD / vif;
    output out=work.Reg_stats r=r_;
    run;
quit;
```

Salida Completa:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1.78272	0.21633	-8.24	<.0001	0
2B	2B	1	-0.01057	0.00315	-3.35	0.0008	1.41886
3B	3B	1	-0.05024	0.00864	-5.81	<.0001	1.17629
BA	BA	1	-2.34934	1.06279	-2.21	0.0272	2.06039
OPS	OPS	1	3.10815	0.29794	10.43	<.0001	1.88770
GDP	GDP	1	0.02166	0.00420	5.16	<.0001	1.20237
SF	SF	1	0.02579	0.00886	2.91	0.0036	1.06541

Anexo 18: Factores de Conversión USD 2012

Año	Factor Conversión
2003	0.80226277025010700
2004	0.82907694370321000
2005	0.85675609049351100
2006	0.87838042392343300
2007	0.91447143641797200
2008	0.91426816768373100
2009	0.93999679959865200
2010	0.95351200799235400
2011	0.98270918298943500
2012	1.00000000000000000
2013	1.01512838366757000
2014	1.02175840429719000
2015	1.02828462812634000
2016	1.04937267808720000
2017	1.07172358912037000
2018	1.09224940641205000
2019	1.11695304492239000
2020	-
2021	1.21150760527807000
2022	1.27725422863840000

Fuente: FRED 2022