USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Métodos estadísticos aplicados al deporte

Daniela Gómez González

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO EN MATEMÁTICAS

Traballo Fin de Grao

Métodos estadísticos aplicados al deporte

Daniela Gómez González

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación de Operativa
Título: Métodos estadísticos aplicados al deporte
Breve descripción del contenido
En los últimos años aparece una gran variedad de medidas estadísticas asociadas a eventos deportivos. La aplicación de métodos estadísticos al deporte viene de la mano del deporte profesional en Estados Unidos, principalmente del béisbol (Sabermetrics) que después se ha extendido a otros ámbitos (NBA, NFL...). El objetivo de este trabajo es revisar y aplicar los principales procedimientos estadísticos y su utilidad en las disciplinas deportivas.
Recomendaciones
Otras observaciones

Índice

Resumen

La estadística está presente en todos los ámbitos de la vida, incluyendo deportes como el baloncesto. Gracias a esta disciplina, ligas como la NBA generan grandes cantidades de dinero al año y optimizan diferentes facetas del juego. Este trabajo se centra en el estudio del salario de los jugadores de la liga de baloncesto americana en la temporada 2020-2021, a través de ciertas características como la edad o características de su juego. El objetivo principal es el de, mediante la teoría de los modelos de regresión, aplicar los conocimientos de los métodos lineales, así como introducir métodos no paramétricos buscando una aplicación práctica e interpretable de los datos recogidos, con el propósito de ver cuales de las características son las más influyentes y las que más información aportan a la hora de predecir los sueldos. Además se estudia si existen casos que supongan irregularidades en el planteamiento de estos modelos y sus posibles interpretaciones analizando jugadores que están siendo pagados por encima o por debajo de su sueldo estimado.

Abstract

Statistics is present in all areas of life, including sports such as basketball. Thanks to this discipline, leagues like the NBA generate large amounts of money per year and enhance different facets of the game. This work focuses on the study of the salary of players in the American basketball league in the 2020-2021 season, through certain characteristics such as the age, the Box Plus/Minus or the value over replacement player. The aim of this study is to apply linear methods, as well as non-parametric methods. In this way, a practical and interpretable application of the collected data is carried out, with the purpose of seeing which of the players' characteristics are the most influential and the ones that provide the most information when predicting salaries. In addition, we study whether there are certain cases that involve irregularities in the approach of these models and their possible interpretations by analyzing players who are being paid above or below their estimated salary.

Introducción

El baloncesto es un deporte de equipo y de balón, que surge en 1881 de la mano de un profesor de educación física, James Naismith, que buscaba una alternativa a los deportes de exterior, cuando las condiciones meteorológicas no permitían realizar actividades al aire libre. En este deporte participan 10 jugadores, 5 por equipo, y consiste en introducir la pelota en el aro que defiende el equipo contrario. La liga de baloncesto americana, más conocida como NBA (National Basketball Association) surge en 1946 como fusión de otras dos asociaciones y desde entonces ha tenido mucho éxito. Con el tiempo ha ido ganando fama y riqueza, de forma que ha dejado de ser simple deporte para acabar convirtiéndose en espectáculo y todo lo que eso conlleva.

La estadística en el deporte empieza a ganar importancia en el momento en que se ve que la relación de la actuación de los jugadores influye en la victorias y derrotas de los equipos. Tiene su origen en el béisbol, donde se desarrolló una técnica que permitía estudiar a los jugadores (Sabermetrics) y posteriormente se ha ido extendiendo a otros deportes. Los datos de los jugadores se transforman en una herramienta de mucha ayuda para los entrenadores a la hora de configurar sus equipos así como conocer a sus oponentes.

Además las estadísticas en la NBA son utilizadas por todo tipo de usuarios del mundo de las apuestas. Dado que son medidas reales de la actuación de los jugadores, permiten obtener ideas claras de lo que puede suceder en un partido y así poder apostar a la opción más favorable.

Es por todo esto que las estadísticas han tomado un puesto muy importante en el mundo del baloncesto, generando mucho dinero en la actualidad, tanto por parte de los equipos como por parte de los consumidores de este espectáculo.

El objetivo de este trabajo es hacer uso de esta herramienta de las matemáticas para tratar de dar una explicación de la asignación de los salarios de los jugadores de la NBA según su intervención durante la temporada 2020-2021. Partiendo de un conjunto de datos asociados a 540 jugadores se trata a través de distintos modelos estadísticos, de dar una buena predicción de los salarios según la calidad de los deportistas. Es claro que los salarios se pactan previamente a la temporada, por lo que, lo que se analiza, es que el salario asignado sea acorde con lo

que se esperaba del jugador para la temporada 20-21. Dado que la gran cantidad de datos tanto asociados a la pura estadística como los asociados a la parte más económica son elevados y presentan diversos problemas a la hora de tratarlos, el punto de partida será realizar las modificaciones pertinentes que ayuden a la hora de utilizarlos en la práctica.

Se parte de los modelos más básicos conocidos, los lineales. En primer lugar, a lo largo del Capítulo 2, desde un punto de vista teórico se trata de entender el modelo y aplicarlo posteriormente de manera práctica a los datos con los que se trabaja. El inconveniente de que la mayoría de las situaciones de la vida cotidiana no siguen un comportamiento lineal induce en la búsqueda de otro tipo de métodos que permitan una mejor aproximación. Es por esto que a lo largo del trabajo se introducirán métodos no paramétricos que ayudan al ajuste de datos que no se distribuyen linealmente. Los métodos seleccionados son los modelos aditivos generalizados, introducidos por Hastie y Tibshirani que presentan la particularidad de que aproximan los datos mediante funciones de suavizado que permiten un ajuste mucho más coherente. Estos modelos son mayormente conocidos como GAM y se describen de la siguiente forma que se desarrollará en el Capítulo 3.

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon.$$

Sin embargo aun más importante es la interpretación del modelo. La herramienta de los *Marginal Effects* resulta de gran ayuda a la hora de exponer los resultados obtenidos a partir de los datos observados, además permite representaciones gráficas útiles a la hora de plasmar y comparar modelos.

Finalmente se plantean a lo largo del Capítulo 4 las relaciones que pueden existir entre distintos tipos de medidas estadísticas que engloban muchas cualidades de los jugadores, analizando aquellos datos más extraños que hacen que los modelos difieran del ajuste perfecto.

Capítulo 1

Preliminares

1.1. Datos

El objetivo principal de este trabajo es hacer un estudio entre los salarios de los jugadores de la NBA en función de sus estadísticas en la temporada 2020-2021 para ver si existe una relación entre la calidad de los jugadores y sus salarios. Los datos que reflejan la participación de los jugadores durante la liga regular que se usarán en este trabajo, fueron obtenidos de la web basketball-reference.com (Forman, 2000) que importa sus datos de la web oficial de la liga americana más conocida como NBA. Así mismo, los salarios se extrajeron de HoopsHype.com (Sierra, 2002). En cuanto a la actuación de los jugadores se recogieron un total de 17.650 datos asociados a 540 jugadores repartidos en 25 variables. Para la variable que se quiere explicar, se encontraron salarios asociados a 578 jugadores, por lo que posteriormente se filtrarán los datos para que concuerden. Las variables que se recogen en el fichero y que se estudian son las siguientes: la posición (Pos), la edad (Age), el equipo (Tm), el número de partidos jugados (G), el número de minutos jugados (MP), la calificación de eficiencia del jugador (PER), el porcentaje de tiro verdadero (TS) que combina los tiros de 2 y los de 3, tasa de intentos de tiros de 3 (3PAr), tasa de intentos de tiros libres (Fr), porcentaje de rebotes ofensivos (ORB), porcentaje de rebotes defensivos (DRB), porcentaje total de rebotes (TRB), porcentaje de asistencias (AST), porcentaje de robos (STL), porcentaje de bloqueos (BLK), porcentaje de pérdidas (TOV), el porcentaje de *usage* (USG) que no es más que una medida del porcentaje de jugadas que usa un jugador cuando está en pista, número de acciones ganadas por un jugador en ataque (OWS), en defensa (DWS) y totales (WS), así como las acciones ganadas en 48 minutos (WS48), la contribución de un jugador en ataque cuando está en pista (OBPM), en defensa (DBPM), y en total (BPM), por último el valor sobre jugador de reemplazo (VORP).

1.1.1. Depuración de los datos

Una vez escogidos los datos que se van a utilizar para el análisis, se procede a su depuración. Se tienen dos ficheros de datos. Uno de ellos contiene una lista de jugadores con sus correspondientes datos asociados a las variables que se introdujeron anteriormente. Como durante una misma temporada los jugadores pueden cambiar de equipo varias veces, se hace una media ponderada de cada jugador según los partidos jugados por equipo. De esta forma se obtienen una fila de datos para cada jugador. El otro fichero contiene los salarios asociados a los deportistas, si bien, como los datos no fueron extraídos de la misma fuente, presentan un orden distinto, e incluso contiene salarios de jugadores que no han sido estudiados. Para solventar este problema, se identifican y asocian los jugadores de ambos ficheros y se combinan en una misma matriz de datos, eliminando las no coincidencias. Estos dos procesos se pueden ver en el código de R del Anexo II. Con el comando `match` se buscan las coincidencias de dos columnas de dos ficheros de datos distintos. Así se pueden identificar los jugadores de los cuales se tienen estadísticas y salario. Una vez hecho esto, se reordenan los salarios de acuerdo al fichero de estadísticas y se eliminan las no coincidencias, para posteriormente, asociarlos en una misma matriz (con el comando `cbind`).

Además se eliminan los jugadores que no alcancen los 100 minutos jugados a lo largo de la temporada con el objetivo de obtener unos resultados más coherentes. Esto es porque en la NBA se permiten contratos de 10 días que garantizan un mínimo de 3 partidos. De esta forma al ser períodos de duración muy corta, no permiten obtener buenos datos y conviene eliminarlos.

Por último se hace una transformación logarítmica de los salarios, esto permite hacer una reducción de su escala, de forma que se hace más pequeño el rango donde varía el salario. Es útil en el sentido de que permite controlar la heterocedasticidad y la asimetría a la derecha, así la ventaja es que se reduce también la diferencia de las observaciones atípicas con respecto a las más razonables. Es importante destacar que todas las cantidades son positivas.

1.1.2. Análisis de los datos

El estudio comienza por hacer un análisis de las variables.

Se recoge en la Tabla 1.1, a modo de ejemplo, la salida del resumen de dos de las medidas que se están estudiando. Si uno se centra en la variable que representa la edad de los jugadores, a la vista de los resultados se tiene que el jugador más joven de la temporada 2020-2021 tenía 19 años, el 25 % de los jugadores tenían entre 19 y 23 años, la media era de 25.75 y el jugador más veterano tenía 37 años de edad.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	19.00	23.00	25.00	25.75	29.00	37.00
USG	7.90	14.90	18.05	18.79	21.90	36.00

Tabla 1.1: Ejemplo de resumen de variables explicativas

En cuanto a la variable USG, previamente se vio que es una medida que permite ver que cantidad del total de las jugadas de un equipo en un partido utiliza un jugador para finalizar un ataque cuando está en pista, luego permite ver la contribución de un jugador en el ataque de su equipo. Así, observando la salida del `summary` vemos que la media de los ataques del equipo que finaliza un jugador es de 18.05 por cada 100 y que el jugador que más ataques finaliza dentro de los que hace su equipo, lo hace en la cantidad de 36 de cada 100.

A partir de aquí y para el siguiente capítulo, los conceptos y definiciones básicas que se usarán se recogen en Draper et al. (1998), Ryan (2008) e Ibarrola (2004).

Definición 1.1. Dada una variable aleatoria X unidimensional, $X : \omega \to \mathbb{R}$ con función de densidad f. Se define su **media** como

$$E(X) = \int_{\mathbb{R}} xf(x)dx. \tag{1.1}$$

Para el caso discreto se tiene

$$E(X) = \sum_k x_k P(X = x_k), \tag{1.2}$$

con $x_1, \ldots, x_n$ los posibles valores de X.

Definición 1.2. Sea X una variable aleatoria unidimensional con media $\mu = E(X)$ se define su **varianza** como

$$Var(X) = E[(X - \mu)^2]$$

Definición 1.3. Sea $(X_1, X_2)'$ un vector aleatorio bidimensional tal que existen las medias $\mu_1 = E(X_1)$ y $\mu_2 = E(X_2)$ se define su **covarianza** como

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]. \tag{1.3}$$

Por último se define el coeficiente de correlación simple, y el múltiple.

Definición 1.4. Se define **coeficiente de correlación simple** de dos variables aleatorias X e Y como

$$\rho_{xy} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1.4}$$

Definición 1.5. Se define **coeficiente de correlación múltiple** de Y sobre $X_1, X_2, \ldots, X_{p-1}$ como el coeficiente de correlación simple entre Y y el ajuste $\beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$

$$Corr(Y; X_1, \ldots, X_{p-1}) = Corr(Y, \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}) \tag{1.5}$$

Si se observa la matriz de correlaciones de las variables explicativas recogida en la Tabla I del Anexo I, estudiando por ejemplo los minutos jugados, se puede ver que se obtienen valores pequeños, cerca del 0, esto sugiere que en este caso, esta variable no aporta información redundante que ya podría estar siendo dada por el resto de variables. Si se valora ahora la contribución en el ataque cuando un jugador está en pista (OBPM) se ve que las correlaciones están más próximas a uno, lo cual indica que podría ser ocasionado por una información ya dada por otra variable. Esto tiene sentido pues, esta variable incluye información sobre las estadísticas de los jugadores en posiciones de ataque.

Ahora bien, con el propósito de ver cuales de las características proporcionan una mejor descripción del salario, se realiza un nuevo cálculo de las matrices de covarianzas y correlaciones incluyendo esta vez el salario. En la Tabla 1.2 se recogen las correlaciones de las variables en función del salario.

Age	MP	PER	TS	X3Par	FTr
0.427	0.615	0.461	0.230	-0.078	0.054
ORB	DRB	TRB	AST	STL	BLK
-0.071	0.168	0.085	0.356	0.035	-0.005
TOV	USG	OWS	DWS	WS	WS48
0.047	0.436	0.498	0.557	0.572	0.308
OBPM	DBPM	BPM	VORP	Salario	
0.541	0.148	0.533	0.570	1.000	

Tabla 1.2: Correlación simple de los regresores con el salario

Surge entonces la duda de cual es el mejor modelo para empezar con el análisis. Si bien se podría pensar en escribir un modelo con todas las variables de las cuales se tienen datos, quizás, debido al alto número de variables, algo más razonable sería escoger aquellas que puedan describir mejor el salario. Esto es, aquellas que hagan que el salario cambie más, en función de su variación, o lo que es lo mismo, aquellas variables explicativas que presenten una mayor correlación con el salario. De esta forma, parece adecuado pensar en estudiar el salario en función de la edad, minutos jugados, la calificación de eficiencia del jugador, el porcentaje de pérdidas, el *usage*, el porcentaje de acciones ganadas en ataque, en defensa y totales de un jugador, la contribución en ataque y la total, y por último el VORP. En secciones posteriores a la hora de seleccionar

las variables que predigan mejor el salario, se tendrán en cuenta todas las variables iniciales. De esta forma no se descartan aquellas que por si mismas no presentan una buena correlación con el salario, pero que junto con otros regresores, podrían tener un papel importante en el modelo.

En la siguiente tabla se recoge la correlación entre las variables seleccionadas, así como el salario. De forma que se puede ver la relación que presentan entre ellas y con la variable respuesta.

	Age	MP	PER	USG	OWS	DWS	WS	OBPM	BPM	Salario
Age	1.00	0.16	0.17	0.01	0.22	0.20	0.24	0.25	0.31	0.43
MP	0.16	1.00	0.42	0.38	0.64	0.80	0.77	0.55	0.51	0.62
PER	0.17	0.42	1.00	0.60	0.75	0.49	0.74	0.87	0.87	0.46
USG	0.01	0.38	0.60	1.00	0.37	0.27	0.37	0.60	0.43	0.44
OWS	0.22	0.64	0.75	0.37	1.00	0.59	0.95	0.80	0.78	0.50
DWS	0.20	0.80	0.49	0.27	0.59	1.00	0.81	0.48	0.60	0.56
WS	0.24	0.77	0.74	0.37	0.95	0.81	1.00	0.77	0.80	0.57
OBPM	0.25	0.55	0.87	0.60	0.80	0.48	0.77	1.00	0.93	0.54
BPM	0.31	0.51	0.87	0.43	0.78	0.60	0.80	0.93	1.00	0.53
Salario	0.43	0.62	0.46	0.44	0.50	0.56	0.57	0.54	0.53	1.00

Una vez analizados los datos que se utilizan, el siguiente paso es la creación del modelo. Se verán distintas posibilidades, tratando de identificar cual es la que mejor se amolda a los datos.

Capítulo 2

Modelo múltiple

2.1. Definición del modelo

Un modelo de regresión lineal múltiple trata de representar como se comporta una variable respuesta Y en función de una serie de variables explicativas $X_1, X_2, \cdots, X_p$ suponiendo que existe una relación lineal entre la respuesta y el resto de variables. Siendo entonces $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ un vector de parámetros que acompaña a las variables, se define el modelo lineal múltiple como:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{2.1}$$

donde ε es el error.

Cabe destacar que β_0 es el intercepto, dicha cantidad representa el valor de la respuesta si todas las variables explicativas fuesen nulas. Los coeficientes asociados al resto de parámetros, representan como varía la respuesta, si la variable asociada a cada uno de ellos aumenta una unidad, y el resto permanecen constantes.

Bajo un diseño fijo, es decir, conociendo de antemano un conjunto de datos, se puede describir el modelo en función de cada individuo del cual se estudian una serie de factores, para el i-ésimo individuo se tendría:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i. \tag{2.2}$$

En los modelos de regresión lineal múltiple se suponen ciertas las hipótesis de homocedasticidad, normalidad e independencia. Estas se pueden resumir en la siguiente condición para los errores:

$$\varepsilon_1, \cdots, \varepsilon_n \in N(0, \sigma^2) \text{ independientes.} \tag{2.3}$$

Se puede formular (2.1) de forma matricial,

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}. \tag{2.4}$$

De forma abreviada sería

$$Y = \boldsymbol{X}\beta + \varepsilon, \tag{2.5}$$

Y es el vector de la respuesta, $\boldsymbol{X}$ la matriz de diseño donde cada fila es un individuo y cada columna un rasgo a estudiar, β es el vector de coeficientes asociado a cada variable, y ε el vector de errores verificando $\varepsilon \in N_n(0, \sigma^2 \boldsymbol{I_n})$

Observación 2.1. La matriz de diseño $\boldsymbol{X}$ contiene una primera columna de unos para que en el modelo se tenga en cuenta el término independiente.

En cuanto a los parámetros del modelo, se pueden estimar mediante el método de mínimos cuadrados. Para la estimación de β se procede de la siguiente forma: Se busca $\hat{\beta}$ tal que cumpla

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - x_i\beta)^2, \tag{2.6}$$

y así se llega a:

$$\hat{\beta} = \boldsymbol{(X'X)^{-1}X'Y}. \tag{2.7}$$

Es importante tener en cuenta que la matriz $\boldsymbol{X'X}$ sea no singular. Una vez visto esto, los valores estimados se calculan sin más que multiplicar el estimador de los parámetros por la matriz de diseño $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X(X'X)^{-1}X'Y} = \boldsymbol{HY}$, con $\boldsymbol{H} = \boldsymbol{X(X'X)^{-1}X'}$ conocida como matriz *hat*.

Además los residuos se pueden escribir como la diferencia entre la respuesta observada y el ajuste $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{(I_n - H)Y} = \boldsymbol{MY}$, denotando por $\boldsymbol{M}$ la matriz generadora de residuos. Dado que la matriz *hat* es de proyección, tiene la particularidad de ser idempotente, luego $\boldsymbol{M}$ también lo es. Para terminar, falta la estimación del otro parámetro del modelo, la varianza. Esto viene dado de la siguiente forma

$$\hat{\sigma}^2 = \frac{RSS}{n - p + 1}, \tag{2.8}$$

con RSS la suma residual de cuadrados ($RSS = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y'MY}$).

2.2. Formulación del modelo

Una vez seleccionadas las variables que mejor explican los salarios e introducido el modelo, se formula para los datos que se quieren estudiar. Para ello se puede escribir el modelo en R y una vez formulado, en la salida del `summary`, se obtienen los coeficientes de cada una de las variables, así como el intercepto y otros datos que se analizarán más adelante para comprobar si se tiene un modelo válido. La sintaxis empleada para el código se puede consultar en el Anexo II.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	10.6005922	0.4970030	21.329	<2e-16 ***
Age	0.1002132	0.0110409	9.077	<2e-16 ***
MP	0.0008665	0.0001635	5.298	1.93e-07 ***
PER	0.0063118	0.0272033	0.232	0.81664
USG	0.0548040	0.0166989	3.282	0.00112 **
OWS	1.0641905	0.9505746	1.120	0.26359
DWS	1.1035123	0.9529828	1.158	0.24757
WS	-1.1360142	0.9404665	-1.208	0.22779
OBPM	-0.0813631	0.0730540	-1.114	0.26606
BPM	0.0895188	0.0689791	1.298	0.19511
VORP	0.1472507	0.1444356	1.019	0.30858
R^2	0.5651			

Tabla 2.1: Estimaciones de los coeficientes, errores, estadísticos de contraste y p-valores asociados a la primera formulación del modelo lineal.

A la vista de la Tabla 2.1, el modelo quedaría formulado de la siguiente manera.

$$\begin{aligned} Y = {} & 10.6 + 0.100 \times \text{Age} + 0.001 \times \text{MP} + 0.006 \times \text{PER} + 0.055 \times \text{USG} \\ & + 1.064 \times \text{OWS} + 1.104 \times \text{DWS} - 1.136 \times \text{WS} - 0.081 \times \text{OBPM} \\ & + 0.090 \times \text{BPM} + 0.147 \times \text{VORP}. \end{aligned} \tag{2.9}$$

En cuanto a la información más relevante que aporta la Tabla 2.1, en la primera columna, se tienen los coeficientes asociados a cada variable. Estos valores permiten escribir la ecuación del modelo, en este caso, (2.9). Además se interpretan como sigue: el valor de dicho coeficiente indica cuanto varía la respuesta si se aumenta una unidad de la variable asociada, y el resto de variables permanecen constantes. También se debe destacar la última columna, esta proporciona

el p-valor del contraste

$$\begin{cases} \mathrm{H}_0\text{: } \beta_i = 0 \\ \mathrm{H}_\mathrm{a}\text{: } \beta_i \neq 0 \end{cases} \tag{2.10}$$

Es claro que el contraste se realiza para cada regresor y determina que parámetros estimados son significativamente distintos de 0. Además del p-valor asociado al contraste, se incluyen ciertos símbolos que permiten visualizar la significación de las variables. Esto es, valores entre el 0 y 0.001 se representan con "***", entre 0.001 y 0.01 con "**", para los p-valores entre 0.01 y 0.05 se usa un único "*", entre 0.05 y 0.1 se representa con un "." y los p-valores mayores que 0.1 no se corresponden con ningún símbolo. De esta forma las variables señaladas con "***" son aquellas que son significativamente distintas de 0 por lo que deberían ser utilizadas en el modelo. Las variables con " **" también muestran evidencias estadísticamente significativas por debajo del 1 % y también se deben considerar como distintas de 0.

2.3. Validación y diagnosis

Una vez planteado el modelo, el siguiente paso es proceder a la validación. Se trata de ver, en primer lugar, cuales de los coeficientes son significativos, es decir, significativamente no nulos. Así, viendo la salida del resumen del apartado anterior que se recogen en la Tabla 2.1, se ve claramente que los coeficientes significativos, son los asociados a las variables Age, MP y USG al nivel del 1 %.

Además se puede ver que el valor del coeficiente de determinación, R^2, es de 0.5651, lo cual indica que con este modelo queda explicada el 56.51 % de la varianza de la respuesta. Este valor coincide con el cuadrado del coeficiente de correlación múltiple definido en (1.5). Dicho valor no es muy elevado, por lo que de cumplirse las hipótesis básicas del modelo lineal, convendría buscar otro planteamiento que explique mejor la respuesta.

Ahora bien, para validar el modelo, se debe ver que cumple las cuatro hipótesis básicas de los modelos lineales que se habían introducido anteriormente. Se tienen múltiples librerías de R que permiten contrastarlas, sin embargo, resulta útil consultar los gráficos del modelo que se obtienen utilizando el comando `plot` e incluyendo como argumento el modelo.

La primera gráfica de la Figura 2.1 representa los residuos frente a los valores ajustados, que no son más que una combinación lineal de las variables explicativas que intervienen en el modelo. Permite visualizar como se distribuyen los residuos, si estos siguen algún tipo de tendencia (linea recta, parábola...) o si se distribuyen sin ningún tipo de patrón. Para el modelo que se está estudiando, parece que los residuos tienen una tendencia lineal, sin embargo en la parte izquierda del eje X parece haber una gran cantidad de valores atípicos que alejan a los

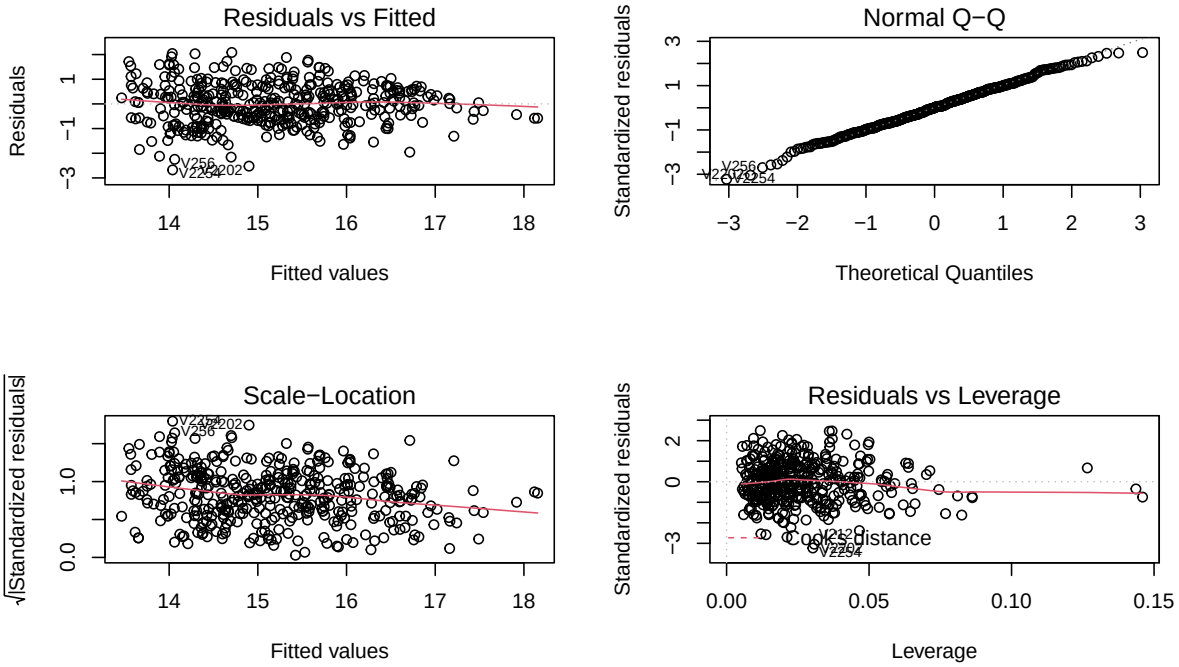

Figura 2.1: Gráficos para la validación y diagnosis del modelo (2.9).

residuos de esta tendencia.

La segunda gráfica permite hacer un estudio de la normalidad. Es un QQ-plot que compara los residuos estandarizados con la distribución normal teórica. Si estos residuos se sitúan sobre la recta de la normal, entonces se puede asumir que el modelo cumple la hipótesis de normalidad. En el caso que se está tratando, parece que se amoldan bastante bien a dicha recta, así que se puede intuir que hay normalidad en el modelo.

La tercera es un gráfico de localización y escala. Representa los valores ajustados frente a la raíz cuadrada de los residuos estandarizados y permite ver si hay homocedasticidad en el modelo. Se busca que la línea roja sea lo más horizontal posible, y que los residuos se distribuyan en torno a ella de manera aleatoria y de forma que todos varíen más o menos lo mismo. Lo cual sería un indicativo de que la varianza se mantiene constante. En este caso, podría intuirse algún problema con la homocedasticidad pues la recta parece tener cierta inclinación. Se aplicará posteriormente un test que permita una mejor interpretación.

La última se conoce como Residuals Vs Leverages, conviene introducir unos conceptos que entrarán en juego. Dichas definiciones se pueden encontrar en (Draper et al., 1998).

Definición 2.2. Se denomina **leverage** (o apalancamiento) en regresión lineal simple, al peso

que ejerce un individuo sobre su propia predicción y se representa por h_{ii}

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}, \tag{2.11}$$

donde n es el tamaño de la muestra, x_i la i-ésima observación y $\bar{x}$ su media. Para el caso de la regresión multiple, los h_{ii} no son más que los elementos de la diagonal de la matriz *Hat*, pues la varianza de los residuos es $\sigma^2(1 - h_i i)$.

Definición 2.3. Se denomina **distancia de Cook** a la siguiente cantidad:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\hat{\sigma}^2}, \tag{2.12}$$

donde $\hat{Y}_j$ son los ajustes incluyendo todos los datos, $\hat{Y}_{j(i)}$ los ajustes sin el dato i-ésimo y p el número de variables explicativas. Análogamente se puede definir en términos de los apalancamientos.

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}}. \tag{2.13}$$

Así un punto causará influencia sobre la recta de regresión si tiene un apalancamiento alto (mayor que $\frac{2p}{n}$) y si su distancia de Cook es grande.

En la gráfica de la Figura 2.1 se representan en las abscisas los apalancamientos y en las ordenadas los residuos estandarizados, así como dos rectas discontinuas para distancias de Cook mayores que 0.5 y 1. Los puntos situados en estas regiones conviene analizarlos y tratar de darles una explicación en cuanto a su distanciamiento del modelo.

Una vez observados los gráficos de apoyo a la validación y diagnosis del problema, se continúa con la realización de los contrastes de hipótesis para comprobar lo que se suponía previamente.

Para la hipótesis de linealidad, se aplica el **Ramsey RESET test**, propuesto por Ramsey y que se encuentra en (Ramsey, 1968). Este test supone que se tiene un modelo de la siguiente forma, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + \varepsilon$ que presenta como variables explicativas, funciones no lineales de los valores ajustados. Esto permite ver si en la primera definición del modelo, como un modelo lineal, falta información que debe ser explicada mediante modelos no lineales. Se realiza el siguiente test:

$$\begin{cases} \text{H}_0\text{: } \delta_1 = 0, \delta_2 = 0 \\ \text{H}_\text{a}\text{: } \delta_1 \neq 0, \delta_2 \neq 0 \end{cases} \tag{2.14}$$

Reset Test			
Data: modelo			
RESET = 4.3635,	df1 = 1	df2 = 401	p-value = 0.037

Tabla 2.2: Estadístico de contraste, grados de libertad y p-valor asociado al contraste de linealidad

En la Tabla 2.2, en la que se recogen los datos del test, se puede ver que el p-valor asociado al contraste es de 0.037, menor que los niveles de significación habituales del 5 y del 10 %, sin embargo, no es menor que el 1 %, aun así, hay evidencias estadísticamente significativas para rechazar la hipótesis de linealidad como ya se venía suponiendo en el análisis de la gráfica.

En cuanto a la normalidad, se utiliza un **test de Shapiro**, cuyo resultado es el p-valor asociado al siguiente contraste

$$\begin{cases} \mathrm{H}_0\colon \varepsilon \in N_n(0, \sigma^2 I_n) \\ \mathrm{H}_\mathrm{a}\colon \varepsilon \notin N_n(0, \sigma^2 I_n) \end{cases} \tag{2.15}$$

Este test se realiza sobre los residuos estandarizados de la regresión, esto es porque al estandarizarlos, los residuos están más cerca de tener una varianza común, lo cual, permite ver más fácilmente si se incumple alguna de las hipótesis de los modelos lineales.

Definición 2.4. Se definen los **residuos estandarizados** de un modelo de regresión como el cociente

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}. \tag{2.16}$$

Shapiro-Wilk normality test	
Data: res	
W = 0.99	p-value = 0.607

Tabla 2.3: Estadístico de contraste y p-valor asociado al contraste de normalidad

Se puede observar que el p-valor asociado al contraste es de 0.607, mayor que los niveles de significación habituales ya mencionados anteriormente, por lo que no hay evidencias para rechazar la hipótesis nula, aceptando así que se tiene normalidad en el modelo, es decir, los errores siguen una distribución normal, tal y como se intuía en el QQ-plot de la Figura 2.1.

En cuanto a la homocedasticidad, se quiere ver si la varianza de los errores permanece constante, para estudiarlo, se puede utilizar el **test de Harrison Mc-Cabe**.

$$\begin{cases} \mathrm{H}_0\colon \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2. \\ \mathrm{H}_\mathrm{a}\colon \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_n^2. \end{cases} \tag{2.17}$$

En la Tabla 2.4 se recoge la información de este contraste. De nuevo, el p-valor asociado a este contraste, 0.662, es mayor que los niveles de significación habituales, por lo que no hay pruebas estadísticamente significativas para rechazar la hipótesis nula, aceptando así que el modelo es homocedástico. A pesar de haber visto en la Figura 2.1 que podría haber heterocedasticidad, a la vista de este contraste esto no es cierto.

Harrison-McCabe test	
Data: Modelo	
HMC = 0.51	p-value = 0.662

Tabla 2.4: Contraste de Homocedasticidad

Por último se estudia la independencia de los errores. El **test de Durbin Watson** realiza el siguiente contraste:

$$\begin{cases} H_0\text{: Los errores son independientes} \\ H_a\text{: Los errores no son independientes} \end{cases} \tag{2.18}$$

lag	Autocorrelation	D-W Statistic	p-value
1	0.038	1.92	0.424
Alternative hypothesis: rho != 0			

Tabla 2.5: Contraste independencia de los errores

Una vez más, como el p-valor asociado al contraste es 0.43 aproximadamente, mayor que los niveles de significación habituales del 1, 5 y 10 %, no hay evidencias para rechazar la hipótesis nula y por tanto se tiene independencia de los errores.

Una vez hecha la validación y diagnosis de este modelo lineal, se puede plantear la búsqueda de otro modelo con menos variables que se ajuste mejor a los datos, ya que se podrían estar estimando parámetros innecesarios. Se puede aplicar entonces un criterio de selección de variables.

2.4. Selección de variables

Como se adelantaba anteriormente, el valor del coeficiente de correlación, 0.5651, indica que el modelo planteado anteriormente explica el 56,51 % de la variabilidad de la respuesta. Se quiere encontrar un modelo que explique mejor dicha variabilidad, pues es un indicativo de que es un mejor modelo. Un intento de mejoría del mismo, consiste en reducir las variables que se involucran en él.

Existen diversos modelos de selección de variables para un modelo múltiple, entre ellos destacan dos grandes tipos. Los métodos **Backward** y **Forward**. En los primeros, se parte de un modelo complejo, con todas las variables potenciales que se quieran incluir y se van eliminando términos hasta que no se puede mejorar más el modelo. En los *forward*, se actúa de forma inversa, se parte de un modelo simple, y se van añadiendo términos siguiendo alguna regla, hasta que se

tiene un modelo que no necesita más variables.

En cuanto a las normas a seguir para añadir o eliminar términos, se pueden utilizar criterios de significación, que buscan que los coeficientes del modelo sean los más significativos posibles; o criterios globales que buscan obtener la mejor medida global del modelo. Dos criterios globales son el **Akaike Information Criterion (AIC)** y el **Bayesian Information Criterion (BIC)** (Sheather, 2009). Las medidas globales en el caso del modelo múltiple para ambos criterios se calculan de la siguiente forma.

$$AIC = n \ln (RSS/n) + 2p \tag{2.19}$$

$$BIC = n \ln (RSS/n) + p \ln n, \tag{2.20}$$

donde RSS es la suma residual de cuadrados, p es el número de parámetros y n el número de observaciones. Estas medidas permiten valorar que modelo es mejor en función de que variables contiene. Por ejemplo, la función `step` de R, permite ir eliminado variables hasta que el AIC no se puede mejorar más (esto es, hasta obtener el menor AIC posible).

En cuanto a otro tipo de selección de variables más novedoso, se tiene el método LASSO. De acuerdo con Tibshirani (2011), el **LASSO** (*Least Absolute Shrinkage and Selection Operator*), como su propio nombre indica, es una técnica de selección de variables que busca evitar el sobreajuste de los datos. Para ello se añade un término de penalización para disminuir la varianza de los datos, aunque al añadir un elemento de penalización, aumente el sesgo.

Se trata de resolver un problema de minimización sobre un modelo de regresión penalizado,

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda P_\lambda(|\beta_{(0)}|). \tag{2.21}$$

Para el caso del LASSO, la penalización del vector de coeficientes sin el término independiente, $P_\lambda(|\beta_{(0)}|)$, se toma como $||\beta_{(0)}||_1$. Por tanto, no es mas que resolver el problema de programación (2.21) aplicándole este término.

$$\min_{\beta} (y - \boldsymbol{X}\beta)^T (y - \boldsymbol{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.22}$$

En la práctica se puede calcular haciendo uso del software R. Con la librería `glmnet` se define el modelo de regresión penalizado y se calculan los coeficientes. Dado que el LASSO busca quedarse con las variables más relevantes, fuerza a valer 0 a aquellos coeficientes de variables que se pueden evitar.

Variables	Coeficientes
Intercepto	12.097
Age	0.070
MP	0.001
PER	.
USG	0.029
OWS	.
DWS	0.015
WS	.
OBPM	.
BPM	0.019
VORP	0.092
TS	.
3Par	.
ORB	.
DRB	.
TRB	.
AST	.
BLK	.
TOV	.
WS48	.

Tabla 2.6: Coeficientes de la regresión LASSO

Los coeficientes obtenidos son los que se muestran en la Tabla 2.6. Además se introducen en este caso el resto de variables que se habían descartado inicialmente debido a la baja correlación con el objetivo de ver si cabría incluirlas en el modelo. Sin embargo, la regresión LASSO indica que ninguna de estas variables tiene coeficientes significativos.

Ahora bien, si se comparan con los resultados de la Tabla 2.1 se puede observar que el término asociado al intercepto incrementa una unidad y media sobre el logaritmo del salario. En cuanto a las variables explicativas todas disminuyen al hacer la selección. El hecho de que las variable WS y OBPM que presentaban coeficientes negativos en (2.9), hace que sus valores se repartan entre el resto disminuyendo su valor. También se eliminan las variables PER y WS, ambas positivas; sin embargo el efecto global es el que se ha mencionado.

Se representan a continuación unos gráficos propios del método LASSO. El la Figura 2.2 se puede ver entre que valores se obtiene el λ óptimo. Las rectas discontinuas delimitan el intervalo donde se puede encontrar dicho valor. Además los números que aparecen en la parte superior

son los coeficientes que son distintos de 0 según el λ seleccionado. En la figura 2.3 se representan los coeficientes en función del logaritmo de λ valores de lambda grandes llevan los coeficientes a cero, y pequeños llevan los valores de los coeficientes a los del modelo lineal.

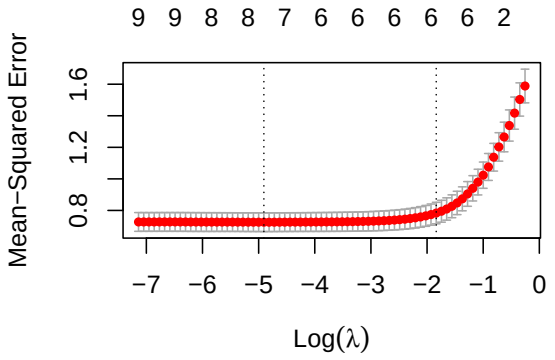

Figura 2.2: Intervalo y número de variables en el modelo según los distintos λ.

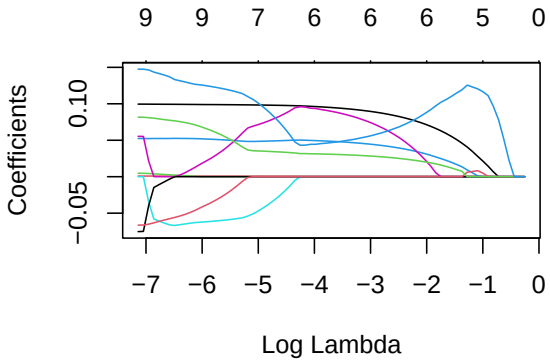

Figura 2.3: Coeficientes de la regresión LASSO en función del logaritmo de λ

Se define entonces un nuevo modelo lineal. El modelo debe incluir las variables explicativas cuyos coeficientes no han sido forzados a valer 0 en la regresión LASSO, entonces según la Tabla 2.6 se tiene

$$\begin{aligned} Y = {} & 12.097 + 0.077 \times \text{Age} + 0.001 \times \text{MP} + 0.029 \times \text{USG} + 0.014 \times \text{DWS} \\ & - 0.070 \times \text{OWS} + 0.019 \times \text{BPM} + 0.092 \times \text{VORP}. \end{aligned} \tag{2.23}$$

Antes de continuar con la validación del modelo con las variables seleccionadas, se puede ver si dicho modelo es mejor que el planteado anteriormente. Para ello se calcula el AIC de ambos, y se comprueba cual de ellos es más bajo. Se obtiene el siguiente resultado.

	AIC
Modelo 1	1042.53
Modelo 2	1039.56

Tabla 2.7: Valores de AIC para el primer ajuste (modelo 1) y para el modelo con variables seleccionadas por el método LASSO (modelo 2).

Sin más que ver la Tabla 2.7 es claro que el mejor modelo es el definido tras realizar la selección de variables por el método LASSO.

Se procede entonces igual que en el primer modelo para hacer su validación y diagnosis.

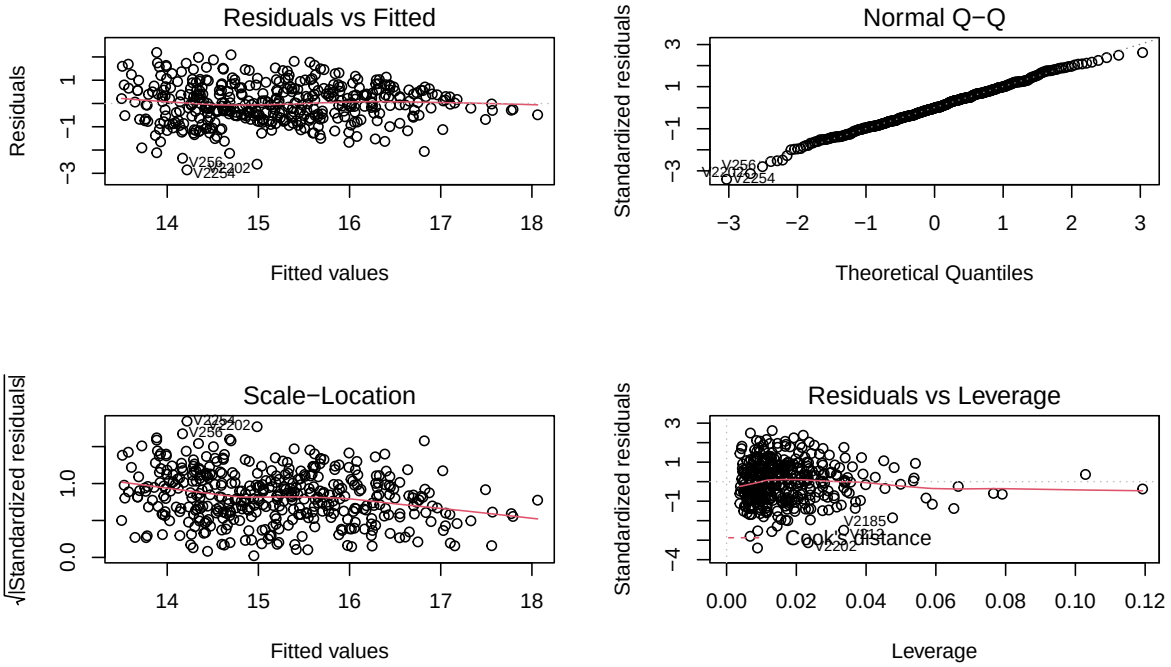

Figura 2.4: Gráficos para la validación y diagnosis

A la vista de los resultados se puede pensar que este modelo con menos variables seleccionadas por el método LASSO tampoco cumple todas las hipótesis del modelo de regresión lineal múltiple.

Parece que la linealidad todavía no se ha corregido, que si que se tiene normalidad en los errores, a pesar de que algunos puntos parezcan alejarse de la recta normal. En cuanto a la homocedasticidad, parece que los residuos se distribuyen de forma aleatoria y con variación parecida en torno a la linea roja, sin embargo parece estar un poco más inclinada que en el modelo anterior. Por último, observando el último gráfico se puede ver que hay algunos datos con distancias de Cook altos que podrían ser datos influyentes en el modelo.

Para comprobar lo que se ha intuido previamente se realizan los tests de contrastes de hipótesis (2.14), (2.15), (2.17) y (2.18).

Se puede ver en la Tabla 2.8 que efectivamente, se rechaza la linealidad, al nivel del 10 % aunque no a niveles del 1 y 5 %, y se aceptan las otras 3 hipótesis básicas del modelo de regresión lineal múltiple.

Test	P-valor
Linealidad	0.073
Normalidad	0.533
Homocedasticidad	0.593
Independencia	0.412

Tabla 2.8: P-valores asociados a los contrastes

Una vez estudiados estos dos modelos, se ha llegado a la conclusión de que ninguno de los dos se ajusta bien a los datos. Se ha visto que ambos incumplen la hipótesis de linealidad (aunque los p-valores no estén por debajo del 1 y 5 %, por lo que, conviene pensar en modelos no lineales que ajusten mejor el salario de los jugadores en función de sus números durante la liga regular 2020-2021.

Capítulo 3

Modelo no Lineal

Una vez se ha tratado de ajustar un modelo lineal y se ha visto que no se cumplían las hipótesis, cabe pensar en los modelos no lineales para tratar de encontrar un modelo que explique mejor la variable respuesta en función de las variables explicativas.

Con objeto de ver cual es la tendencia que siguen las variables explicativas, se realiza un gráfico de dispersión que permite ver que formas siguen los datos de cada variable, y que tipo de relación tienen con la variable respuesta. Se separan las 6 variables explicativas en dos gráficos, para tener una comparación más clara son el salario, que en la Figura 3.1 y la Figura 3.2 viene definido como V2.

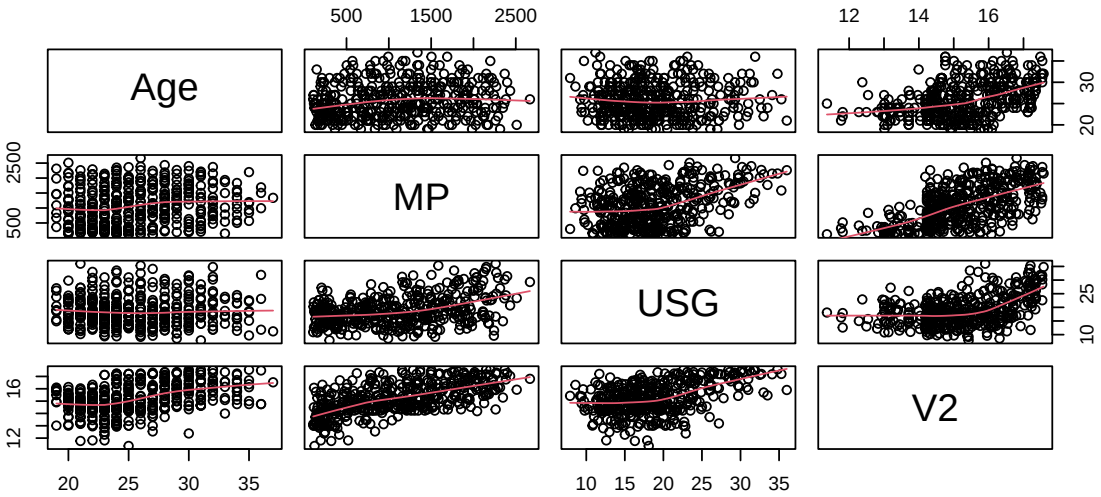

Figura 3.1: Gráficos de dispersión para las variables Age, MP, USG y el salario

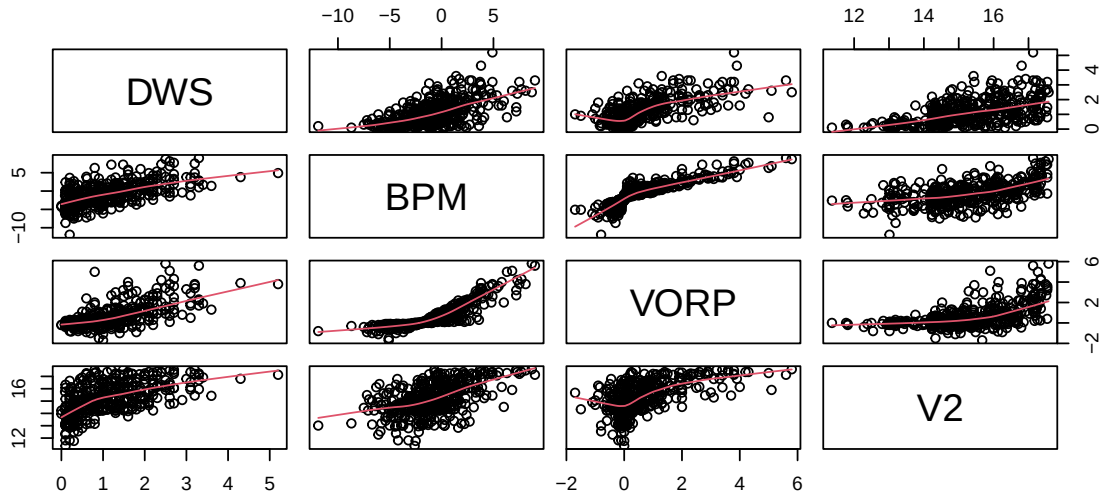

Figura 3.2: Gráficos de dispersión para las variables DWS, BPM, VORP y el salario

Con un simple golpe de vista de la Figura 3.1 y la Figura 3.2 , se ve claramente lo que se esperaba. Utilizando funciones de suavizado para describir la forma en la que se disponen los datos de una variable al relacionarla con la respuesta, se ve que la mayoría no tienen una tendencia lineal, si no que siguen distintos tipo de curvas.

3.1. Modelos aditivos generalizados

Dado que en la vida real, muchas situaciones no son explicables mediante modelos lineales, surgen métodos para explicar otro tipo de relaciones. Uno de ellos, sobre los que se desarrollarán las principales ideas a continuación, son los modelos aditivos generalizados.

Según el capítulo 9 de (Hastie et al., 2009), dado un conjunto de variables explicativas $X_1, X_2, \cdots, X_p$ y una respuesta Y, se define el **modelo aditivo generalizado (GAM)** de la siguiente manera.

$$E(Y|X_1, X_2, \cdots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \tag{3.1}$$

Las f_j, $j \in \{1, \cdots, p\}$ son funciones suaves que se estiman utilizando un algoritmo conocido como *Backfitting*, se recoge en el Algoritmo 3.1. Es un proceso iterativo que va suavizando los residuos.

El hecho de ser "generalizado" tiene que ver con que pueden existir distintas funciones *link* que relacionan la media condicionada de una variable respuesta Y con una función aditiva de las variables explicativas. Existen distintos tipos de funciones *link* y se define el modelo de acuerdo con la expresión (3.2).

$$g[\mu(X)] = \alpha + f_1(X_1) + \ldots + f_p(X_p). \tag{3.2}$$

Para los datos sobre los que se está trabajando, se usa la función identidad, luego en adelante se hará referencia a estos modelos simplemente como modelos aditivos. A la hora de introducir la *deviance* característica de estos modelos, cabe destacar que en el caso de tener normalidad en la respuesta (es el caso de la función identidad como función *link*)se corresponde con la suma residual de cuadrados, *RSS*, propia de los modelos lineales.

Una característica importante que los hace útiles, es que no todas las funciones deben ser no lineales, esto reduce la complejidad de un modelo totalmente no paramétrico.

3.1.1. Ajuste teórico de un modelo aditivo generalizado

Con la notación habitual que se está utilizando, se tiene que un modelo aditivo tiene la siguiente forma

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{3.3}$$

y además $\varepsilon \in N(0, \sigma^2)$. Este tipo de modelos presentan las mismas hipótesis que las del modelo lineal, exceptuando su forma. Además se impone que $E(f_j(X_j)) = 0$ de forma que no haya dos modelos diferentes que aporten la misma predicción.

Como se había mencionado anteriormente, las f_j son funciones de suavizado que se determinan a partir de los datos, mediante el algoritmo de **Backfitting**.

Algoritmo 3.1 (Backfitting)**.**

1. *Inicialización* $\hat{\alpha} = \frac{1}{N}\sum_1^N y_i, \;\; \hat{f}_j = 0, \;\; \forall i, j$
2. *Bucle* : $j = 1, 2, \cdots, p, 1, 2, \cdots, p.$

   $\hat{f}_j \leftarrow \mathcal{S}_j[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N],$

   $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^N \hat{f}_j(x_{ij})$

hasta que las $\hat{f}_j$ *varíen menos que un valor predeterminado.*

Es importante destacar que $\mathcal{S}$ es el operador de suavizado. Hay distintos tipos de operadores de suavizado, pero en este caso se usa el spline cúbico.

Un **spline** es una función polinomial por partes que trata de ajustarse a un conjunto de datos. Para conseguir una aproximación de los mismos, utiliza un proceso consistente en una interpolación entre cada par de datos que están cerca, de forma que la estimación de la función final se obtiene sumando todas las aproximaciones entre datos contiguos. El spline cúbico, que es el que utiliza este modelo, tiene la particularidad de que usa un polinomio cúbico para cada aproximación entre puntos. Así, el conjunto final, es decir, el spline, es una función de clase 2, esto es, una función continua cuyas derivadas primera y segunda existen y son continuas cumpliendo que la segunda derivada se hace 0 en los extremos del intervalo donde se define esta función. Esta información se recoge en el Capítulo 5 de Hastie et al. (2009) y en Wood (2006).

Así se tiene que las f_j son splines cúbicos, uno por cada variable explicativa que se tenga, y que se ajuste no linealmente. Como las funciones de suavizado son muy flexibles, no serán únicas. La constante α se ajusta dependiendo de las funciones suaves, además como se impone que $\sum_1^N f_j(x_{ij}) = 0 \ \forall j$, se tiene $\hat{\alpha} = \frac{1}{N}\sum_{i=1}^N y_i$.

Una vez fijada la condición anterior para $\hat{\alpha}$ se itera hasta conseguir todas las funciones que se necesitan. Se va aplicando el operador de suavizado a $[\{y_i - \hat{\alpha} - \sum_{k\neq j}\hat{f}_k(x_{ik})\}_1^N]$ como función de x_{ij} para estimar cada $\hat{f}_j$, hasta que las estimaciones se estabilicen.

Observación 3.2. El paso 2 del algoritmo 3.1 no es necesario realizarlo ya que se ha impuesto que la suma de las estimaciones de las funciones sea nula. Pero se define igualmente por posibles problemas de cálculo.

3.1.2. Formulación del modelo

Una vez introducidos los modelos aditivos generalizados también conocidos como GAM, se busca aplicarlos a los datos que se están tratando de estudiar en este trabajo. Si se recuerdan los pasos llevados a cabo en los anteriores apartados, se llegó a la conclusión de que las variables que mejor explicaban en salario eran la edad (Age), el número de minutos jugados (MP), el usage (USG), el número de acciones ganadas por un jugador en defensa (DWS), la contribución total de un jugador cuando está en pista (BPM), y el valor sobre jugador de reemplazo (VORP).

Teniendo esto en cuenta se formula el modelo aditivo a partir de la fórmula general vista anteriormente (3.1).

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon \tag{3.4}$$

Para los datos que se estudian:

$$Y = \alpha + f_1(Age) + f_2(MP) + f_3(USG) + f_4(DWS) + f_5(BPM) + f_6(VORP) + \varepsilon, \quad (3.5)$$

y se supone que $E(\varepsilon) = 0$

Observación 3.3. Los parámetros del modelo se aproximarán con el método **REML** (*Restricted Maximum Likelihood*), que no es más una forma de estimación de máxima verosimilitud que utiliza una función de probabilidad calculada de un conjunto de datos transformados y no de toda la información. Esto tiene una ventaja, y es que los parámetros que no intervienen directamente (pero que si son necesarios para estimar otros) no tienen ningún tipo de efecto.

Observación 3.4. Para la validación de este modelo se seguirán los pasos descritos en (Ross, 2019) y (Faraway, 2016).

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	15.19410	0.03937	385.9	<2e-16 ***
	edf	Ref.df	F	p-value
s(Age)	4.178	5.152	19.553	<2e-16 ***
s(MP)	3.534	4.377	9.779	<2e-16 ***
s(USG)	2.915	3.680	9.162	1.74e-06 ***
s(DWS)	1.005	1.009	0.539	0.462
s(BPM)	2.356	3.061	0.491	0.717
s(VORP)	2.609	3.269	1.470	0.247

Tabla 3.1: Estimaciomaciones y características del intercepto y de las funciones suaves.

En la Tabla 3.1 se recogen algunos de los datos de la salida del resumen del primer Modelo Aditivo planteado. En la tabla se distinguen dos partes, la primera de ellas contiene los términos que no necesitan una función de suavizado, en este caso, unicamente el intercepto, ya que, por construcción, todas las variables explicativas se han estimado usando la forma más general de un modelo aditivo generalizado, esto es aproximando funciones para cada variable.

A partir de la tercera fila de tabla se tiene la información asociada a las variables que si se han estimado con una función suave. La primera columna muestra los grados de libertad efectivos, cuyo valor representa la complejidad de dicha función. Es decir, para un grado de libertad efectivo, la suavización es lineal, para 2, cuadrática y así sucesivamente. Por un lado, sin más que ver dicha columna se puede ver que la variable DWS tiene un comportamiento lineal, siendo así la variable con función de suavizado más simples. Por otro lado, la variable Age tiene la función de suavizado más compleja ya que sus grados de libertad efectivos superan 4.

Además también se incluye un estadístico de contraste y un p-valor asociados al contraste de significación que indica que funciones de suavizado deben incluirse en el modelo. De acuerdo con esto, es claro que DWS,BPM y VORP presentan p-valores altos, lo cual indica que no son significativas. Dado que los modelos aditivos no presentan una elección automática de las variables, posteriormente se hará una selección manual de cuales son las variables que convienen ser introducidas en el modelo y eliminar aquellas que produzcan malos efectos en el ajuste.

Con ánimos de visualizar la información explicada hasta ahora sobre este modelo se hace uso de un gráfico. Con la función de R `plot` para un modelo aditivo, se obtiene un gráfico distinto para cada variable explicativa, donde viene representada su función de suavizado (Ross, 2019).

Efectivamente la Figura 3.3 muestra un resultado acorde al de la Tabla 3.1. Se intuye una linealidad sobre la función de suavizado de la variable DWS, tal y como se había visto antes, mientras que las funciones de las variables Age y MP oscila más.

A pesar de esto, se puede pensar que valores próximos a 1 en los grados de libertad efectivos, también podrían estar ajustándose de forma lineal. Esto se puede analizar desde el gráfico. Para la variable VORP, por ejemplo, cuyos grados de libertad efectivos son 2.609, si uno se fija en la representación de su función de suavizado, se ve que donde se acumulan la mayor parte de los datos (los datos vienen representados en el eje de abscisas por rayas negras), la gráfica sigue un comportamiento que se parece a una conducta lineal. Para la variable BPM, que tiene un valor de grados efectivos de 2.356, también podría seguir una tendencia que se aproximase a la linealidad cerca de donde hay una mayor concentración de datos. Debido a estos resultados, está claro que se debe definir un nuevo modelo que tome la variable DWS como lineal y plantearse si BPM y VORP se ajustan mejor al salario de forma lineal o con las funciones de suavizado.

Las otras tres columnas de la Tabla 3.1 son muy similares a las de un resumen de un modelo lineal y hacen referencia a los contrastes de significación de las funciones de suavizado.

Dado que los modelos aditivos no cuentan con una selección de variables, es necesario llevarla a cabo manualmente. Esto es, eliminar del modelo una a una, aquellas variables no significativas (con p-valores más grandes), hasta obtener un modelo con todas ellas significativas.

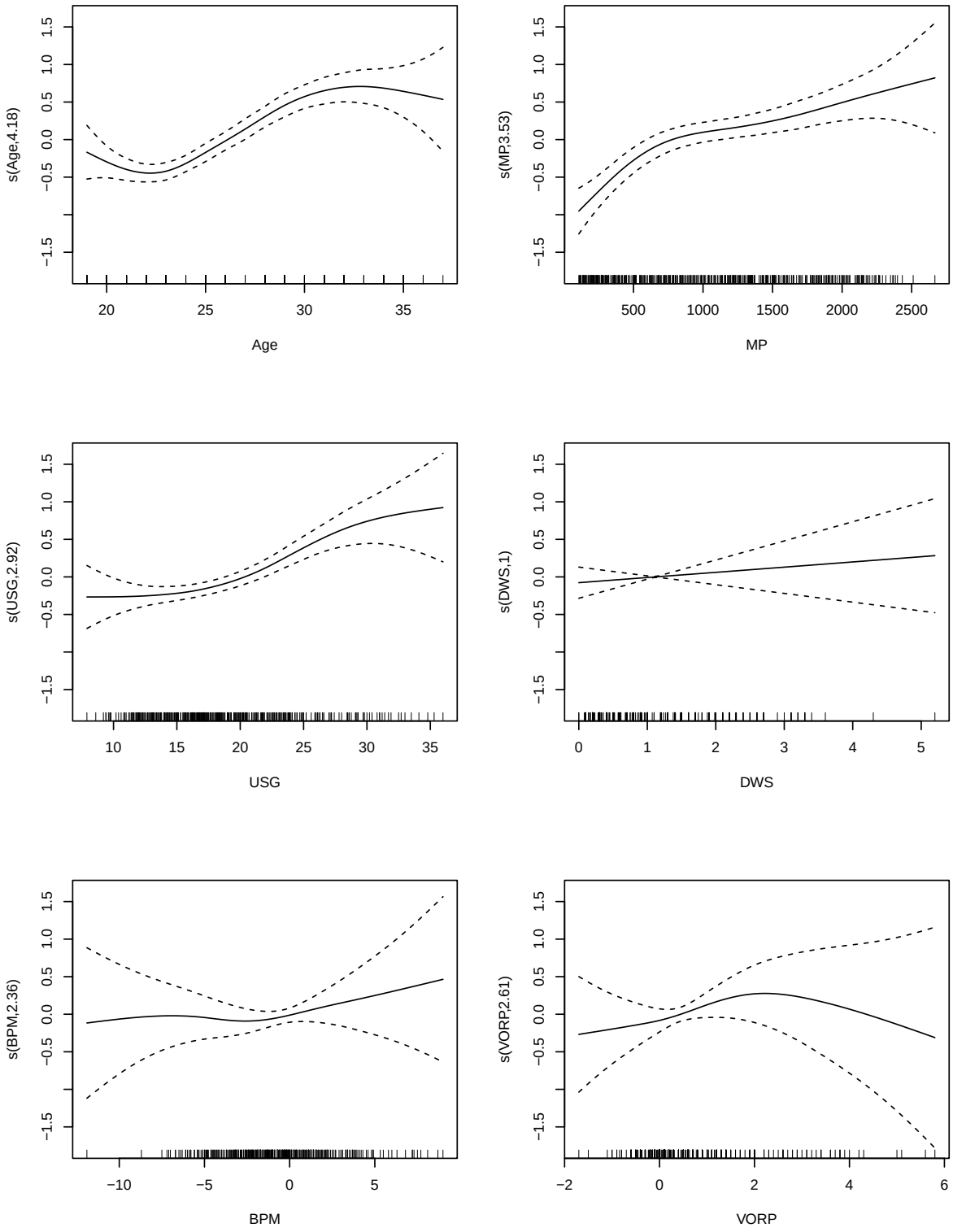

Figura 3.3: Funciones de suaves de las variables explicativas (línea negra) e intervalos de confianza (línea punteada).

De acuerdo con la Tabla 3.1, el mayor p-valor, 0.717, es el de BPM. Luego se define un modelo eliminando dicha variable.

$$Y = \alpha + f_1(\text{Age}) + f_2(\text{MP}) + f_3(\text{USG}) + f_4(\text{DWS}) + f_5(\text{VORP}) + \varepsilon, \tag{3.6}$$

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	15.194	0.039	385.5	<2e-16 ***
	edf	Ref.df	F	p-value
s(Age)	4.227	5.211	19.924	<2e-16 ***
s(MP)	3.460	4.286	9.433	<2e-16 ***
s(USG)	2.832	3.580	9.717	1.35e-06 ***
s(DWS)	1.004	1.007	0.616	0.43192
s(VORP)	3.109	3.899	3.645	0.00756 **

Tabla 3.2: Características del intercepto y de las funciones suaves para el modelo aditivo sin BPM.

Una vez formulado el nuevo modelo (3.6) en su resumen, que aparece recogido en la Tabla 3.2 se ve que el VORP tiene asociado un p-valor grande, 0.43192, lo cual indica que no es significativamente distinta de 0. Se procede eliminándolo y se obtiene:

$$Y = \alpha + f_1(\text{Age}) + f_2(\text{MP}) + f_3(\text{USG}) + f_4(\text{VORP}) + \varepsilon. \tag{3.7}$$

Los resultados obtenidos para el modelo (3.7) se recogen en la Tabla 3.3.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	15.194	0.039	385.7	<2e-16 ***
	edf	Ref.df	F	p-value
s(Age)	4.248	5.233	19.846	<2e-16 ***
s(MP)	3.466	4.293	17.784	<2e-16 ***
s(USG)	2.813	3.557	9.785	1.21e-06 ***
s(VORP)	3.167	3.967	5.029	0.000645 ***

Tabla 3.3: Características del intercepto y de las funciones suaves para el modelo aditivo (3.7).

Los resultados recogidos en la Tabla 3.3 indican que todas las funciones suaves del modelo son significativas, por lo que el proceso de selección finaliza y se prosigue con la validación del

modelo seleccionado, en este caso el modelo (3.7).

Se utilizará la función `gam.check()` para ver si se ha obtenido un modelo correcto.

Se recoge la información más relevante de la salida del `summary` en la Tabla 3.4. En primer lugar se debe observar la convergencia del modelo. Se necesita una convergencia total, lo cual significa que el Algoritmo 3.1 ha encontrado una solución para cada función de suavizado. En este caso, se puede ver que se ha alcanzado dicha convergencia en 5 pasos.

A continuación se tienen entradas para los términos no lineales. Se corresponde con contrastes asociados a las bases de las funciones, por lo que se tendrá cuatro columnas en las que se describen el número de bases que se han usado para estimar el parámetro de suavizado (la columna k'), los grados de libertad efectivos (edf), el estadístico de contraste(k-index) y un p-valor (p-value).

Dicho p-valor indica si los residuos están o no distribuidos aleatoriamente. Cuanto más pequeño sea ese valor indica que hay menos distribución aleatoria y es un indicador de que probablemente, no se hayan usado suficientes bases de estimación.

full convergence after 5 iterations.				
	k'	edf	k-index	p-value
s(Age)	9.00	4.25	0.94	0.120
s(MP)	9.00	3.47	1.03	0.730
s(USG)	9.00	2.81	1.02	0.710
s(VORP)	9.00	3.17	1.01	0.61

Tabla 3.4: Resumen comprobación modelo

A la vista de los valores de la última columna de la Tabla 3.4 el modelo que se está estudiando, no presenta p-valores bajos en los contrastes, por lo que se puede pensar en aleatoriedad de los residuos y consecuentemente, en que hay bases suficientes para la estimación de los parámetros.

Se analizan ahora estos resultados mediante un gráfico con la función `plot(gam.check())`.

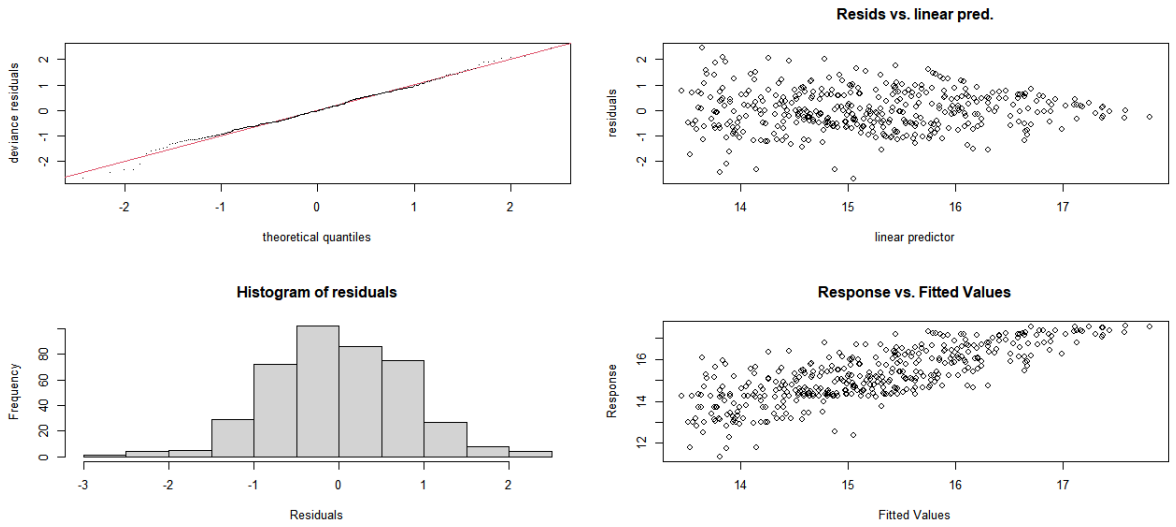

Figura 3.4: Gráficos asociados a la validación del modelo aditivo

Los 4 gráficos anteriores proporcionan información sobre los residuos del modelo. Se analizan a continuación uno a uno. El primer gráfico a la izquierda es un QQ-plot que permite comparar una distribución normal con la de los residuos del modelo planteado. Un buen resultado sería obtener una recta o curva que se asimile lo máximo posible a la recta de la normal. En la Figura 3.4 se ve que la mayor parte de los residuos se disponen sobre la línea roja. Se aprecian algunas irregularidades en las colas, sin embargo esto no supondría ningún problema en la normalidad en caso de aceptar dicho modelo.

En cuanto al gráfico situado arriba a la derecha, representa los valores de los residuos frente al predictor lineal. En el caso de que estuviesen distribuidos de forma aleatoria en torno a la recta $y = 0$, sería un indicador de que la varianza de los residuos se mantiene constante. Sin más que imaginar dicha línea sobre el 0, se ve que los residuos se distribuyen sin ningún patrón en torno a ella, a pesar de esto, la cantidad de residuos en la parte izquierda del gráfico es mayor que en la derecha, esto no implica que la varianza no sea constante, es consecuencia de los valores de la variable respuesta, no influye entonces la disposición en el eje X.

En la parte inferior izquierda de la Figura 3.4 se tiene un histograma de los residuos, este al igual que el primer gráfico que se analizó, permite ver su distribución. A diferencia del primero, este puede sugerir que tipo de distribución tiene. Ya que para un modelo bien ajustado se requiere una distribución normal, se busca que este histograma sea simétrico y con forma de campana. Los resultados son coherentes con el QQ-plot ya mencionado, se tiene una distribución casi normal, con alguna variación en la cola izquierda de la campana.

Para acabar con la comprobación se recurre a una representación de la variable respuesta

frente a los valores ajustados. Lo ideal sería tener una recta, si bien, no tenerla no tiene porqué indicar que el modelo no sea válido, simplemente no será perfecto; si que es necesario para tener un buen modelo, que la variable respuesta cambie con respecto a los valores ajustados en una proporción 1 a 1.

Ahora bien, dado que el mejor modelo es el que responde a la fórmula (3.7), se puede valorar si conviene eliminar alguna de las funciones suaves para introducir la variable correspondiente de forma lineal. A la vista de la Tabla 3.3, las variables con menos grados de libertad efectivos son el USG y el VORP. Sin embargo, ninguna de ellas está cerca de tener un único grado de libertad efectivo, lo cual se corresponde con una recta. Se tiene en la Figura 3.5 una representación gráfica de las funciones suaves. El objetivo de este gráfico es mostrar si siguen alguna tendencia lineal donde hay una mayor concentración de datos.

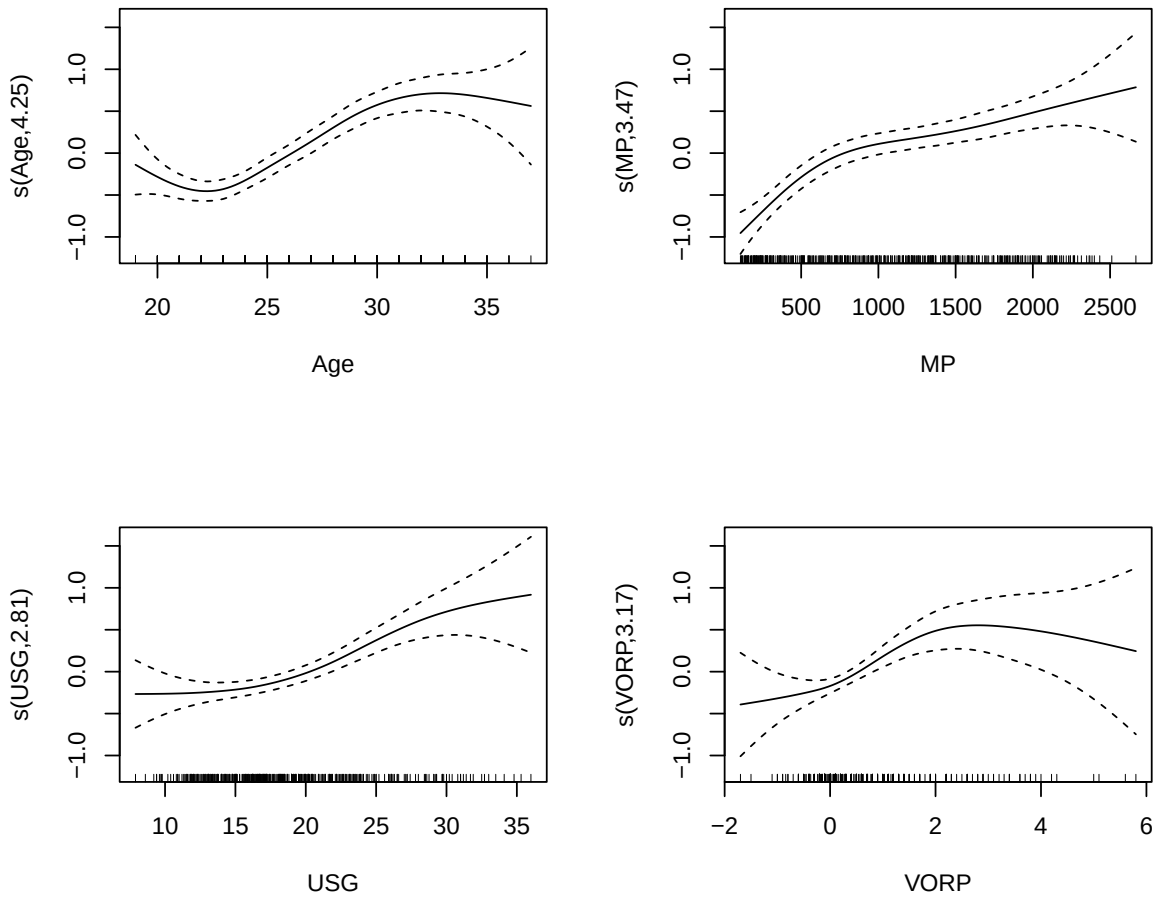

Figura 3.5: Funciones suaves para las variables del modelo (3.7)

En la Figura 3.5 no se aprecia ningún tipo de comportamiento lineal, por lo que no interesa introducir ninguna de estas variables como lineales.

De esta forma se ha obtenido un modelo que aparentemente es el que mejor explica el logaritmo del salario en función de las variables de partida. Dado que el atractivo de un modelo deportivo es sacar información de utilidad que pueda ser fácilmente interpretable, se trata de ver ahora una interpretación de los resultados.

3.1.3. Marginal Effects

Con objeto de conseguir una interpretación en cuanto a lo que estima el modelo, se usará la librería de `marginaleffects`, se seguirán los pasos descritos en Arel-Bundock (2022). En primer lugar se tienen los **predictores ajustados**, esto permite calcular la respuesta para algunos valores de las variables explicativas. De esta forma, usando la función `predictions()` se predicen, por defecto, los valores ajustados de la regresión para todos los datos que se estudian. También se pueden seleccionar variables, de forma que se obtengan predicciones en función de un conjunto de predictores. Utilizando la función `mean` dentro del argumento `newdata`, lo que se consigue obtener el valor de la predicción a partir de las variables seleccionadas y tomando como valor sus medias.

El resultado de dicha observación se recoge en la siguiente tabla.

type	predicted	std.error	conf.low	conf.high	Age	MP	USG	VORP
response	15.21	0.11	14.99	15.42	25.75	1085.23	18.79	0.59

Tabla 3.5: Predicción sobre el valor medio de las variables, intervalos de confianza y valor de las medias de cada variable.

Entonces, para un jugador promedio, sin más que fijarse en la Tabla 3.5, el valor de la respuesta será de aproximadamente 15.21. Es importante recordar que se ha hecho una transformación logarítmica de los salarios y la función `predictions` mantiene dicha transformación. De esta forma el valor del salario que se predice para un jugador cuyas estadísticas sean las medias de los jugadores estudiados, es de 4032915 dólares al año, se consigue sin más que aplicarle la función exponencial a la predicción.

Como se dijo anteriormente, con la función `predicticions` se obtiene un valor de predicción para cada dato del archivo de datos; si a esta función se le aplica un `summary` se obtendrá una media de todas las predicciones, se puede comprobar que efectivamente coincide con el resultado de la Tabla 3.5. Para ver el procedimiento llevado a cabo véase el código de R del Anexo II. Además haciendo uso de `plot_cap` se puede representar la respuesta en función de una o dos

variables. Por ejemplo, si se toma la variable que representa la edad (Age) y los minutos jugados (MP), en la Figura 3.6 se representa gráficamente como varía el salario en función de estas dos variables.

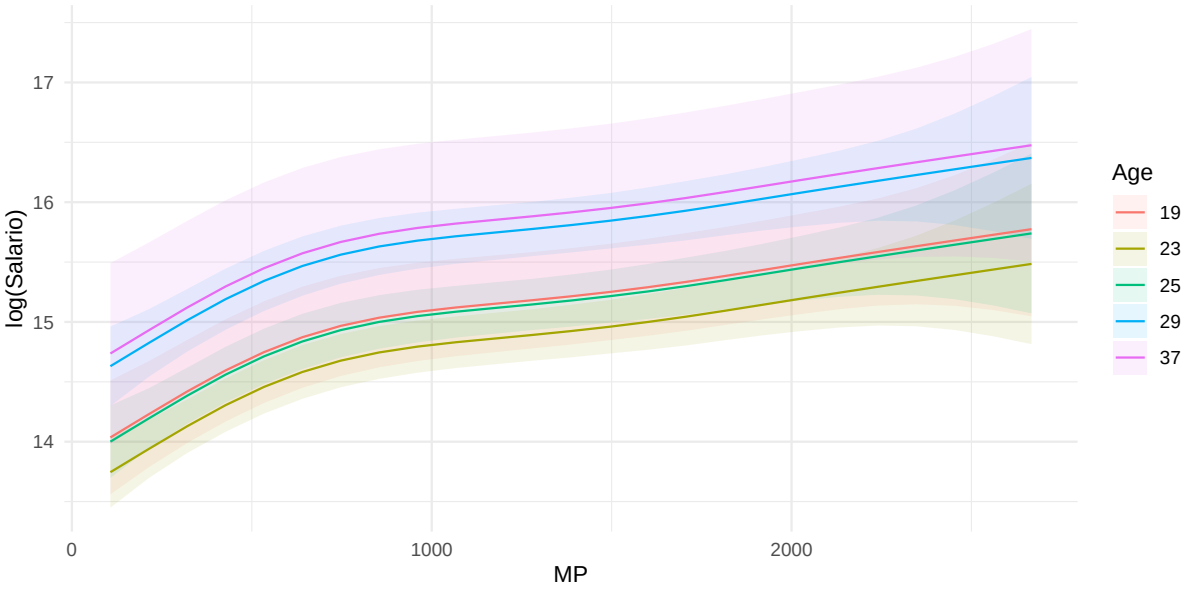

Figura 3.6: Predicción del logaritmo del salario según la edad y los minutos jugados

Por otro lado, también se pueden estudiar los **efectos marginales**. No son más que una medida de la relación que existe entre el cambio de la variable respuesta y el cambio en las explicativas. Formalmente, son derivadas parciales de la ecuación que define el modelo de la regresión respecto de cada variable y por cada unidad de los datos. El valor de los efectos marginales varía para cada observación, pues depende de los valores que tome cada variable explicativa en cada conjunto de datos. El hecho de que sean todas distintas, las hace un poco más complejas de interpretar, y es por eso que conviene utilizar un promedio de todas ellas para llevar a cabo el análisis.

Los resultados de promediar los efectos marginales se recogen en la Tabla 3.6.

type	term	estimate	std.error	statistic	p.value	conf.low	conf.high
response	Age	0.06	0.01	4.66	<0.001	0.04	0.09
response	MP	0.00	0.00	7.48	<0.001	0.001	0.001
response	USG	0.04	0.01	4.50	<0.001	0.02	0.06
response	VORP	0.24	0.07	3.24	0.001	0.10	0.39

Tabla 3.6: Promedio de efectos marginales (AME) para cada variable.

En cada fila de la Tabla 3.6 se tiene un promedio del cálculo de los efectos marginales sobre cada variable. La columna de las estimaciones indica como varía (aumenta o disminuye) el logaritmo del salario (ya que estamos considerando esta escala) en función del aumento de una unidad en las variables explicativas. Se hará una interpretación con detalle en el caso de los efectos marginales sobre la media.

	term	dydx	std.error	statistic	p.value	conf.low	conf.high	Age	MP	USG	VORP
response	Age	0.15	0.05	3.23	<0.001	0.06	0.25	25.75	1085.23	18.79	0.59
response	MP	0.00	0.00	1.06	0.29	-0.00	0.00	25.75	1085.23	18.79	0.59
response	USG	0.05	0.02	2.36	0.02	0.01	0.09	25.75	1085.23	18.79	0.59
response	VORP	0.39	0.12	3.27	<0.001	0.15	0.62	25.75	1085.23	18.79	0.59

Tabla 3.7: Efectos marginales sobre la media (MEM)

En cuanto a las interpretaciones que se pueden hacer sobre estos resultados, se debe destacar que los efectos marginales se calculan sobre un jugador promedio. Esto es, aquel cuyos datos para cada variable es una media de los valores de todos los jugadores de la NBA que participan en el estudio. Así, si se considera la edad, el logaritmo del salario aumenta en 0.15 unidades si el jugador cumple años.

Si se considera ahora la variable que se refiere a los minutos jugados, según los resultados recogidos en la Tabla 3.7 uno puede pensar que no produce ningún efecto sobre la respuesta. Sin embargo esto no es cierto, la tabla recoge unas cantidades aproximadas, en realidad el resultado que se ha obtenido en la salida de R indica que el logaritmo del salario aumenta 0.0008425083 por cada minuto jugado más.

El resto de estimaciones para las variables USG y VORP se interpretan de igual forma. Son todas positivas por lo que el aumento en las variables de regresión provocará un mayor salario para el jugador, en este caso, un deportista cuyos datos observados sean la media de todos los jugadores que se están considerando en el estudio.

Observación 3.5. En el caso del modelo lineal múltiple, el promedio de efectos marginales (AME) y los efectos marginales sobre la media (MEM) coinciden.

Por otra parte existe lo que se conoce por Promedio por Grupos de los Efectos Marginales (Group-Average Marginal Effect, G-AME). Dicha medida permite cuantificar cual es el efecto de una cierta variable sobre la respuesta, teniendo en cuenta distintos grupos de los datos estudiados. En el modelo planteado, todas las variables eran continuas. Para llevar a cabo un estudio a través de distintos grupos, se introduce la variable BPM categorizada, a modo de ejemplificar esta herramienta de los efectos marginales. Como ya se dijo, es una medida que involucra como

contribuye un jugador cuando está en la pista de forma que las contribuciones que aportan puntos al equipo se evalúan positivamente, mientras que las que hacen perder al equipo restan. Así se toma un jugador como jugador de alta efectividad si el valor del BMP es positivo, y baja efectividad si dicho valor es negativo.

Se estudia el efecto de la edad sobre el salario, teniendo en cuenta si el jugador pertenece al grupo de alta efectividad o al de baja. La sintaxis que se usa en R es la misma que el caso anterior añadiendo el argumento `by`.

type	term	BPM_cat	estimate	std.error	statistic	p.value	conf.low	conf.high
response	Age	Baja efectividad	0.05	0.02	3.20	<0.001	0.02	0.08
response	Age	Alta efectividad	0.09	0.01	7.84	<0.001	0.07	0.11

Tabla 3.8: Efecto de la edad sobre el logaritmo del salario según un BPM negativo (baja efectividad) y positivo (alta efectividad).

Es fácil interpretar los resultados de la Tabla 3.8. Suponiendo un posible jugador que contribuye a su equipo positivamente, un aumento en la edad supone tener un aumento de 1.09 en el salario. Mientras que para un jugador con baja efectividad, el salario incrementa unicamente 1.05. Para la obtención de estos resultados se ha aplicado la función exponencial a los valores de la columna *estimate*.

Para dar por finalizados los efectos marginales, se puede hacer una interpretación de como varía el salario para un jugador con ciertas características. A efectos prácticos es algo que puede resultar útil a un jugador a la hora de como podría aumentar o disminuir su salario según sus números de la temporada. Por ejemplo consideramos a Ricky Rubio. Se tiene en cuenta entonces que el modelo que se estudia sigue la ecuación (3.7) y que los valores obtenidos por Ricky en la temporada 2020-2021 son los que se recogen en la Tabla 3.9.

Age	MP	USG	VORP
30	1772	16	0

Tabla 3.9: Valores de Ricky Rubio para las variables del modelo en la temporada 20-21

Luego incluyendo un `datagrid` en la función `marginaleffects()` con los valores de Rubio, se obtienen los siguientes resultados.

type	term	estimate	std.error	statistic	p.value	conf.low	conf.high
response	Age	0.10	0.05	1.84	0.07	-0.01	0.21
response	MP	0.00	0.00	1.69	0.09	-0.0001	0.001
response	USG	0.03	0.02	1.33	0.18	-0.01	0.07
response	VORP	0.23	0.13	1.75	0.08	-0.03	0.49

Tabla 3.10: Variación del logaritmo del salario de Ricky Rubio según el incremento en las distintas variables explicativas.

La interpretación que se le puede dar a los resultados de la tabla anterior es la misma que se hace en los casos anteriores. No se debe olvidar que el salario, la variable respuesta del modelo (3.7), fue transformada mediante logaritmos, luego las cantidades de la columna *estimate* de la Tabla 3.10 indica cuanto aumenta el logaritmo del salario de dicho jugador en función de la variable sobre la que se aumente una unidad.

Por otro parte, la librería que se está utilizando, permite también hacer comparaciones entre jugadores "ficticios" que presentan ciertas características. Por ejemplo se puede plantear de que manera influye cumplir años en la NBA, con la función `comparisons` se obtiene el resultado de la Tabla 3.11.

type	term	contrast	estimate	std.error	statistic	p.value	conf.low	conf.high
response	Age	(x + 1) - x	0.06	0.01	4.64	<0.001	0.04	0.09

Tabla 3.11: Comparación en el salario según la edad

Según el resultado anterior, es claro que la diferencia entre tener un año menos o tener un año más para un jugador de la NBA, supone un aumento de 0.06 en el logaritmo del salario. El mismo tipo de contraste se puede plantear de otra forma. Por ejemplo, para un posible jugador de 26 años, se quiere ver cual es la diferencia de salario de uno de 31.

	type	term	contrast	estimate	std.error	statistic	p.value	conf.low	conf.high
1	response	Age	31 - 26	0.68	0.12	5.486	0.00	0.43	0.92

Tabla 3.12: Variación del salario en función de tener 26 o 31 años

Así, esta comparación se puede interpretar como sigue. Que un jugador tenga 31 años implica que cobrará 1.97 ($=e^{0.68}$) unidades más que un jugador de 26 años.

Ahora bien, lo que se acaba de hacer para la edad, se puede hacer para cada una de las variables que pertenecen al modelo, incluso se pueden comparar entre ellas.

Una medida interesante sobre la cual realizar las comparaciones es la media. Por ejemplo al pensar en el VORP, que es un valor sobre el cual no se tiene una idea tan intuitiva como la edad, las comparaciones se pueden realizar viendo cual es el efecto de la respuesta cuando la variable se distancia una o dos desviaciones típicas de la media. Para llevarlo a cabo en R basta indicar si se quiere una desviación (sd), dos ($2sd$) o las que se quiera.

type	term	contrast	estimate	std.error	statistic	p.value	conf.low	conf.high
response	VORP	(x + sd/2) - (x - sd/2)	0.40	0.10	3.86	0.00	0.20	0.61

Tabla 3.13: Efecto sobre la respuesta de una desviación típica sobre la media en el VORP

Luego para un posible jugador cuyo VORP esté una desviación típica por encima de la media, su salario será $1.49(=e^{0.40})$ unidades mayor que el salario de uno cuyo VORP esté una desviación típica por debajo de la media.

Durante estos capítulos, se ha tratado de buscar cual es el mejor modelo para explicar como varía el salario de los jugadores de la NBA en función de sus datos sobre la pista. Se ha analizado el modelo lineal múltiple, concluyendo que este tipo de ajuste no era el mejor ya que no se cumplía la hipótesis de linealidad. En este contexto, se introdujeron los modelos aditivos generalizados, que permitían combinar funciones lineales y no lineales que permitiesen una mejor aproximación de los datos. En el próximo capítulo, se tratará de seleccionar distintas variables y ver en que sentido están relacionadas y además se tratará de dar una explicación a aquellos datos que pueden resultar anómalos.

Capítulo 4

Comparaciones

Durante los capítulos 2 y 3 se describieron modelos que mediante un conjunto de variables respuestas apropiadas trataban de explicar como se podía estimar el salario de los jugadores de la liga de baloncesto americana. En primer lugar se escogió un conjunto de variables que mediante su correlación se vio que podían ser las adecuadas, y a partir de ahí fueron surgiendo los distintos modelo estudiados. Llegados a este punto uno puede plantearse como podría influir alguna de las variables por si misma sobre el salario o a su vez sobre otras variables respuesta. En este capítulo se tratará de dar una interpretación de estas interacciones, así como de los puntos que puedan resultar anómalos y como se pueden explicar dichas anomalías.

4.1. Salario vs Box Plus/Minus

El Box Plus/Minus es una medida interesante de la eficacia de un jugador, esto es porque engloba muchas características, prácticamente resume todo lo que hace un jugador los minutos que está en el campo. Se usa por primera vez en el hockey y posteriormente en la NFL y la NBA. Se trata de una medida que estima lo que contribuye un jugador por cada 100 posesiones, y es una forma de ver que cantidad de puntos aporta a su equipo. Para calcularlo, se utiliza la diferencia de puntuación desde que un jugador entra en la pista, hasta que sale. Es importante destacar que esta medida depende de la posición de cada jugador. Parece lógico pensar que una mayor contribución supondría un mayor salario, veamos si esto es cierto.

4.1.1. Formulación del modelo

De acuerdo con lo que se explica en Winston (2012) se tratará de ver como influye el Box Plus/Minus en el salario de los jugadores. Para este trabajo, se han sacado los datos de Forman

(2000) y Sierra (2002) como ya se ha mencionado anteriormente. En esta base de datos, ya se proporcionan las contribuciones de cada jugador en ataque, en defensa y totales a su equipo, esto es, las variables OBPM, DBPM y BPM respectivamente. El problema de los valores brutos para estas variables, es que dependen de cuales sean sus rivales y quienes sean sus compañeros en el campo. Para darle una solución a esto, se trabaja con el Box Plus/Minus ajustado. Los valores obtenidos tienen una representación. Por ejemplo, Luka Doncic, jugador de los Mavericks, tiene un BPM de 6.8, esto significa que un equipo de jugadores promedio de la NBA, mejoraría si jugase Luka, aumentando 6.8 puntos por cada 100 posesiones.

Ya que se tienen 3 medidas distintas en primer lugar se trabajará con la medida total y a partir de ahí se verá si está mejor pagado en la NBA ser bueno en ataque o en defensa; esto es, ver si tener un jugador con buen OBPM tiene mejor salario que uno con mejor DBPM.

Para llevar a cabo este estudio se partirá del modelo más simple posible, el modelo lineal simple. Dicho modelo no es más que un caso particular de (2.1) con un único regresor (Draper et al., 1998; Ryan, 2008).

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \tag{4.1}$$

donde Y es el salario, X_1 el BPM y ε el error. Aplicando el `summary` de R, se obtiene que la estimación del intercepto es de 15.36 y la del coeficiente β_1 es 0.224. Luego el modelo planteado es:

$$Y = 15.36 + 0.224 \times \text{BPM}. \tag{4.2}$$

Para poder utilizar este modelo, de acuerdo con lo descrito anteriormente, se deben cumplir las hipótesis del modelo lineal. Por tanto hay que realizar los contrastes (2.14), (2.15), (2.17) y (2.18).

Test	P-valor
Linealidad	0.053
Normalidad	0.008
Homocedasticidad	0.828
Independencia	0.302

Tabla 4.1: p-valores asociados a los contrastes

Los resultados de realizar los contrastes indican que no se tiene la hipótesis de normalidad. (Se puede ver una representación gráfica en la Figura I.1 en el Anexo I). Además de no tenerse normalidad en los errores lo cual influiría a la hora del cálculo de los intervalos de confianza,

al aplicar un test para contrastar la normalidad, el p-valor obtenido indica que existen pruebas estadísticamente significativas a un nivel de entre el 5 y el 10 % para rechazar la hipótesis nula de estar ante un modelo lineal. A pesar de no estar por debajo del 5 %, aplicando el RESET test ya introducido anteriormente, se obtiene una significancia mayor del 10 %, y si uno se apoya en la representación gráfica del ajuste, se puede pensar que el uso de un modelo lineal no es el más adecuado.

Sin embargo planteando un caso particular de (3.3), donde solo se tiene una única variable X_1 y una única función de suavizado f_1, se obtiene que donde se encuentra la mayor cantidad de datos, el ajuste sigue una línea recta como se puede ver en la Figura 4.1.

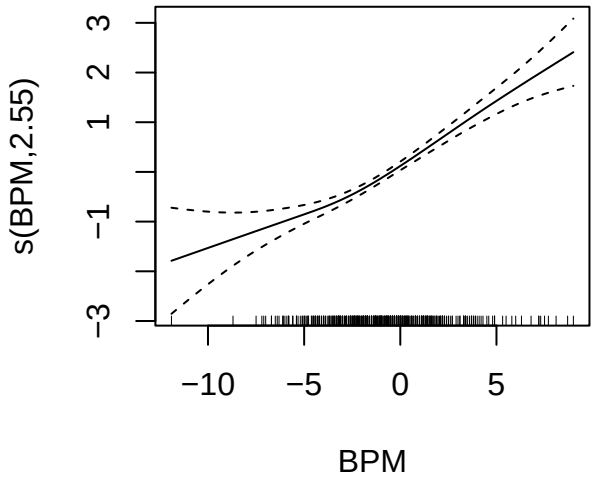

Figura 4.1: Función suave para el BPM (línea negra) e intervalos de confianza (línea discontinua).

De acuerdo con la Figura 4.1, es lógico utilizar el modelo formulado en (4.1).

El objetivo principal de este ajuste es poder interpretar los resultados de forma que resulten útiles y se pueden aplicar a la práctica. De nuevo para ayudarse en la interpretación, se hace uso del paquete `marginaleffects` de R (Arel-Bundock, 2022), y que se utilizó en el capítulo anterior para predecir los resultados.

Interpretación de los marginal effects

En primer lugar, es interesante predecir el salario según este modelo para un jugador con BPM medio, sirve como una estimación media del salario, lo cual puede resultar útil.

	rowid	type	predicted	std.error	conf.low	conf.high	BPM
1	1	response	15.19	0.05	15.09	15.30	-0.76

Tabla 4.2: Predicción del logaritmo del salario en función del BPM medio

Así, el salario para un posible jugador cuyo valor fuese el promedio de todos, sería de 3953058 dólares. Obsérvese que dicha predicción se puede calcular manualmente substituyendo en (4.2) el valor del BPM medio, esto es -0.76.

Haciendo uso de las funciones para calcular el máximo y el mínimo de una variable así como con que jugador se corresponde, se ha encontrado que el jugador con el máximo BPM de la temporada 2020-2021 fue Giannis Antetokoumpo, con un valor de 9, mientras que el jugador con el mínimo fue Josh Hall. Sus salarios reales y los estimados por el modelo son los que siguen.

Jugador	Salario estimado	Salario real
Giannis Antetokoumpo	35321415	27528088
Josh Hall	324486.8	449115

Tabla 4.3: Salarios para jugador con mejor y peor BPM

Es claro según los resultados de la Tabla 4.3 que existen datos que hacen que el modelo difiera de los resultados ideales. Por ejemplo, ya que estamos ante un modelo lineal, la estimación del salario de Antetokoumpo es la más alta, sin embargo su salario real está por debajo de dicho valor. Giannis ocupa el puesto 29 en el ranking de pagos más elevados (Sierra, 2002), lo cual no coincide con la formulación del modelo. Por otro lado Josh Hall tiene un salario mayor del que estima el modelo, esto es también debido a los outliers que se analizarán en la siguiente sección. Se debe tener en mente que en la NBA se firman contratos multianuales, esto supone que a la hora de firmarlo los números obtenidos por los jugadores no sean los mismos que los de temporadas posteriores.

Para ver cuanto varía el salario según la variación del BPM, se usan los *marginal effects*. En este caso dado que solo se está estudiando la respuesta en función de un grupo simplemente se calcula el promedio de lo efectos marginales para tener una idea general de cual es la variación global.

	type	term	estimate	std.error	statistic	p.value	conf.low	conf.high
1	response	BPM	0.22	0.02	12.78	<0.001	0.19	0.26

Tabla 4.4: Promedio de efectos marginales del BPM sobre el salario

En general, un aumento en una unidad en los valores del Box Plus/Minus supone un aumento de 1.246 en el salario, que no es más que la exponencial del valor que se especifica en la columna *estimate* de la Tabla 4.4.

Finalmente, en la Figura 4.2 se tiene una representación gráfica de la recta del ajuste del salario en función del BMP, se consigue utilizando la función `plot_cap`.

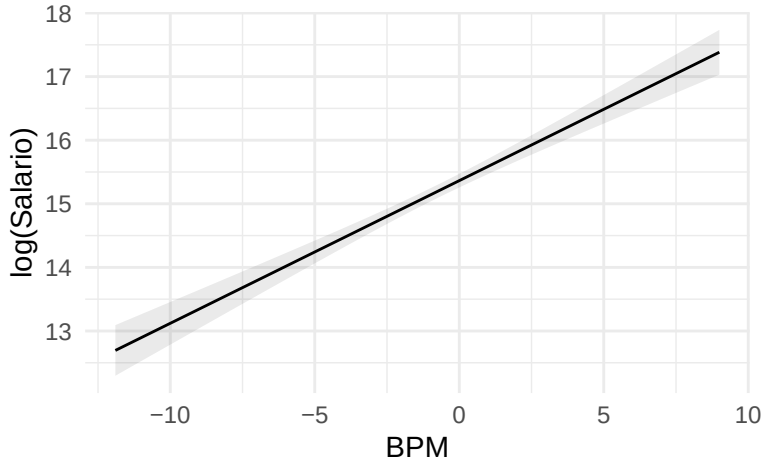

Figura 4.2: Predicciones para el logaritmo del salario según el BPM

Interpretación de los outliers

Como se ha argumentado anteriormente, el jugador con mejor BPM, Giannis Antetokoumpo no es el jugador que más cobra, y según la Figura 4.2 esto debería ser así. El problema puede venir de datos considerados como extraños que interfieran en el ajuste del modelo y provoquen este tipo de resultados. Se pretende entonces hacer un estudio de los datos que no se ajustan bien mediante el modelo,para tratar de darles una explicación.

Los siguientes conceptos y procedimientos se basan en la búsqueda de datos atípicos en modelo de regresión lineal simple. Para esta sección se hará uso de las definiciones y resultados en los trabajos de Rousseeuw and Leroy (2005), Cook and Weisberg (1982), Ryan (2008) y Draper et al. (1998).

La búsqueda de datos atípicos conocidos en inglés por *outliers* se centra en encontrar aquellos

valores que tienen una gran repercusión sobre el estimador de mínimos cuadrados, que es aquel que se usa para el cálculo de los parámetros del modelo, incluyendo los coeficientes.

En primer lugar se introducen los tipos de residuos de un modelo de regresión. Se sabe que dichos valores se definen como la diferencia de los valores de la respuesta observados y el ajuste. Esto es $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - x_i\hat{\beta} \;\; i \in \{1, \ldots, n\}$. Estos errores son conocidos como residuos brutos de la regresión, sin embargo, tienen la característica de que estos no varían en la misma escala, lo cual supone un problema a la hora de detectar irregularidades. Para solventar este inconveniente, se trabaja con los residuos estandarizados, definidos en (2.16) o los estudentizados (4.3).

Una particularidad que hace a estos residuos más útiles que los generados por el modelo sin ningún tipo de transformación es que estos están más próximos a tener varianza común 1, lo cual permite detectar los valores atípicos más fácilmente.

Definición 4.1. Los **residuos estudentizados** se definen a partir de los estandarizados de la siguiente forma:

$$t_i = \frac{r_i}{\sqrt{\frac{n-p-r_i^2}{n-p-1}}}, \tag{4.3}$$

donde r_i son los residuos estandarizados, p el número de parámetros del modelo y n el número de observaciones.

Los candidatos a ser *outliers* de la regresión son aquellos que tienen valores grandes de los residuos. Como valor a superar para ser considerado un valor atípico, se toma 1.96, dicho valor no es más que el de una distribución normal que contiene a más del 95 % de los individuos. A modo de ilustración se representan los residuos estandarizados frente a las predicciones de la respuesta. Además en color rojo se pintan aquellos residuos que superan el umbral determinado y que por tanto son candidatos a formar parte de los valores que influencian este ajuste. Toda esta información se recoge en la Figura 4.3.

Como ya se ha visto anteriormente, lo que se busca para tener un modelo correcto, es que los residuos se distribuyan en torno a una línea imaginaria sobre el 0 de forma constante. Los residuos más alejados de la línea mencionada, serían los valores aislados que se buscan. Efectivamente esta suposición coincide non la representación gráfica de la Figura 4.3.

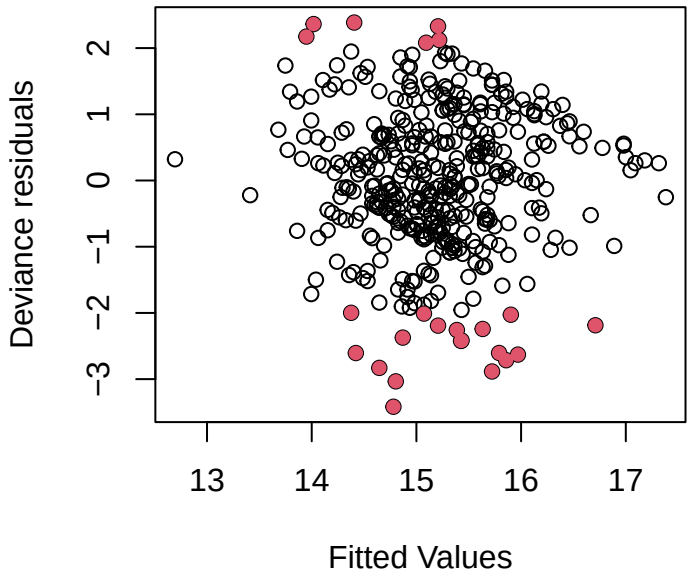

Figura 4.3: Residuos frente a valores ajustados

A continuación se recogen en la Tabla 4.5 los valores del BPM correspondientes a los puntos rojos junto con el salario estimado por el modelo y el salario real, así como los jugadores a los que están asociados.

Las entradas de la Tabla 4.5 muestran lo que cabía esperar. Las observaciones del ajuste realizado que presentan altos residuos vienen de jugadores que están pagados muy por encima de su valor o al contrario muy por debajo. Por ejemplo, Robert Williams presenta el BPM más alto de los jugadores recogidos en la tabla, es por esto que, según el modelo, debería ser el jugador con salario más alto en la Taba 4.5. Sin embargo se le paga 2029920 dólares, cuando la predicción es que debería cobrar aproximadamente 16 millones de dólares más. Si se fija uno ahora en Rodney Hood, el jugador con peor BPM y por tanto el que debería tener un salario más bajo, es claro que se le paga mucho más de lo que se debería. Según su actuación en la temporada 2020-2021, su salario debería ser unos 10 millones de dólares menos.

Es claro que los salarios están pactados previamente a la temporada, por lo que estos datos extraños pueden venir de un jugador que pase de ser un alguien poco conocido y con números malos en años anteriores, a un jugador promesa con muy buenos datos la siguiente temporada o viceversa, y de ahí a estas irregularidades en los salarios.

Jugador	Valor de BPM	Salario estimado	Salario Real
Jordan Bone	-4.20	1833135	135350
Jarrell Brantley	2.20	7704588	510586
Armoni Brooks	-3.20	2294166	135362
Chris Chiozza	-0.70	4019769	449115
Andre Drummond	-1.21	3588858	28751774
Tacko Fall	1.20	6156288	653924
Yogi Ferrell	0.28	5008215	444149
Trent Forrest	-1.3	3513523	470690
Blake Griffin	-0.65	4063131	33900241
Donta Hall	1.60	6734279	376276
Gary Harris	-4.27	1804884	19610714
Jalen Harris	0.30	5030736	449115
Jaylen Hoard	-2.50	2684270	129209
Rondae Hollis-Jefferson	0.10	4810004	502957
Rodney Hood	-6.3	1142801	10047450
Mike James	-4.40	1752703	237648
Alize Johnson	2.70	8619154	621587
Dakota Mathias	-2.60	2624721	86132
Skylar Mays	1.90	7203114	532218
Justin Patton	-2.20	2871147	267623
Khyri Thomas	2.40	8058153	1061283
John Wall	-0.70	4019769	41254920
Robert Williams	6.00	18071124	2029920
Justise Winslow	-6.00	1224108	13000000

Tabla 4.5: Valores para los datos atípicos

Jugadores como Griffin o Wall entre otros, sufrieron lesiones durante la temporada. En el caso de Griffin el jugador estuvo fuera de las pistas durante 6 meses. Múltiples lesiones y el paso del tiempo llevan a este jugador a no producir como solía hacerlo.La lesión de John Wall no le permitió jugar los 11 partidos que restaban de la temporada, además que el cambio de equipo (manteniendo el salario del equipo anterior) son condiciones que influyen en la eficacia de este jugador.

Dichos acontecimientos interfieren en el ajuste de forma que, la poca cantidad de datos no permite reflejar su actuación durante los partidos disputados, lo cual puede tener influencia sobre la estimación de su salario, que había sido previamente pactado. Luego la actuación que cabía

esperar de dichos jugadores, no pudo ser ejecutada debido a las lesiones, y de ahí la discrepancia entre el salario estimado y el real.

Por otro lado, se tienen jugadores como Andre Drummond, un jugador veterano el cual en dicha temporada sigue cobrando de su contrato multianual. Así se entiende la diferencia entre su salario real y su salario estimado.

Finalmente, un caso de un jugador veterano que acepta contratos básicos para poder continuar en las pistas por más tiempo es Jalen Harris, es claro que sus números son buenos y por eso se estima un salario mucho mayor que el real.

Por otra parte jugadores como Jarrell Brantley, cuyo salario real está por debajo de su estimación, son casos de deportistas que no suelen estar en la plantilla habitual, luego sus contratos no son elevados en comparación con las rotaciones habituales, y por tanto si en los pocos minutos que juegan consiguen buenos números, esto hace ue la estimación de su salario aumente con su actuación.

Es por esto que a pesar de que se hayan generado modelos capaces de explicar como influyen ciertas características del juego sobre lo que deberían cobrar, no siempre se obtienen estimaciones ajustadas, y es claro que se deben analizar diferentes factores a la hora de establecer los contratos.

4.1.2. Comparación de modelos Salario vs Ataque y Salario vs Defensa

Como ya se adelantó anteriormente, además del estudio de los datos atípicos se pretende ver que está mejor pagado en la NBA, ser bueno en ataque o ser bueno en defensa. En primer lugar se plantean ambos modelos de forma lineal, según la fórmula (4.1) y se contrastan las hipótesis como se ha hecho en todos los casos anteriores. La variable respuesta será la misma para ambos, el salario, mientras que las variables explicativas serán OBPM y DBPM, X_1, W_1 respectivamente.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{4.4}$$

$$Y = \gamma_0 + \gamma_1 W_1 + \epsilon, \tag{4.5}$$

con $\beta_0, \beta_1, \gamma_0, \gamma_1$ los coeficientes de la regresión y ε y ϵ los errores.

Cabe destacar que dado que el valor del BPM se consigue mediante la suma de los valores del OBPM y del DBPM, introducir ambas variables en un único modelo, sería realizar el estudio del apartado anterior.

Para evaluar la linealidad de los modelos se va a utilizar la función `sm.regression()`. El contraste que se realiza es H_0: Modelo lineal frente a H_a: Modelo no paramétrico.

Los p-valores obtenidos al realizar el contraste para los dos modelos (4.4) son los de la Tabla 4.6. Ambos están por debajo del nivel de significación del 5 % por lo que en los dos casos hay evidencias estadísticamente significativas para rechazar la hipótesis nula, por lo que se debe plantear un modelo no paramétrico, de nuevo se usan los modelos aditivos que se están usando a lo largo del trabajo.

Modelo	P-valor
Salario frente a OBPM	0.001
Salario frente a DBPM	0.043

Tabla 4.6: P-valores asociados a los contrastes de linealidad

A partir de aquí los modelos que se estudian son,

$$Y = \alpha + f_1(\text{OBPM}) + \varepsilon, \tag{4.6}$$

$$Y = \lambda + g_1(\text{DBPM}) + \epsilon, \tag{4.7}$$

donde Y es el salario en ambos casos, f_1, g_1 las funciones de suaves para los regresores y ε, ϵ son los errores.

El resumen de estos modelos se puede encontrar en el Anexo I, no obstante el objetivo es el de comparar los resultados de predicción del modelo.

Interpretación de los Marginal Effects

Se sigue el procedimiento utilizado en la sección anterior empezando por predecir los valores del ajuste para un jugador ficticio cuya puntuación sea la media del OBPM y DBPM, esto dará una primera idea de que da más dinero en la NBA.

Regresor	type	predicted	std.error	conf.low	conf.high	mean
OBPM	response	15.04	0.08	14.89	15.19	-0.75
DBPM	response	15.19	0.09	15.01	15.37	-0.002

Tabla 4.7: Predicciones para el valor medio de OBPM y DBPM

En la Tabla 4.7 se distinguen 6 columnas. La última de ellas, la media, indica sobre que valor se hace la predicción y dicha estimación se recoge en la columna *predicted*. Además se presentan intervalos de confianza y el error típico para cada grupo de la variable categórica. Con intención de deshacer el logaritmo se aplica la función exponencial sobre las estimaciones de la Tabla 4.7

y se obtiene que el salario estimado para un jugador con valor medio de OBPM será de 3402429 y para otro jugador cuyo valor de DBPM es la media de todos será 8040485.

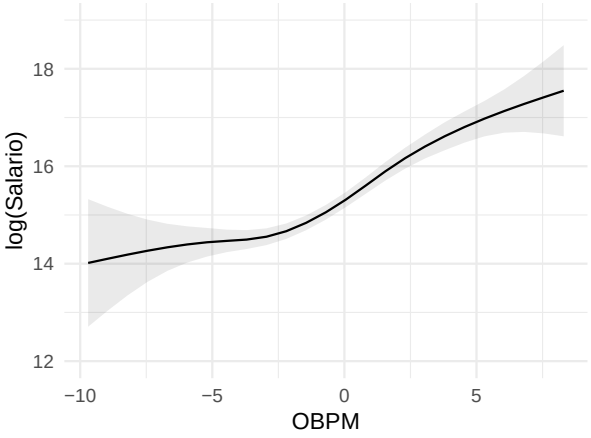

Figura 4.4: Predicciones para el logaritmo del salario según OBPM (línea negra). Intervalos de confianza para la predicción (sombreado gris).

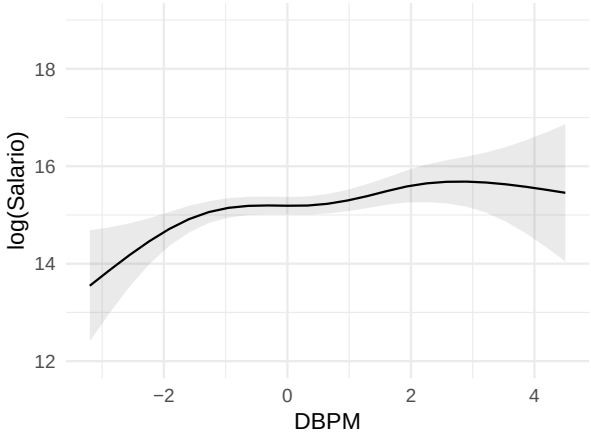

Figura 4.5: Predicciones para el logaritmo del salario según DBPM (línea negra). Intervalos de confianza para la predicción (sombreado gris).

Además se quiere ver cual es el salario estimado según estos modelos para los jugadores con mejor y peor OBPM y DBPM.

Jugador	Salario estimado	Salario real
Stephen Curry	41866645	43006362
Josh Hall	1226899	449115

Tabla 4.8: Salario estimado y real para jugador con mejor y peor OBPM

Jugador	Salario estimado	Salario real
Matisse Thybulle	5126840	2711280
Markus Howard	766814.3	535294

Tabla 4.9: Salario estimado y real para jugador con mejor y peor DBPM

Los resultados de las Tablas 4.8 y 4.9 son muy acordes a los gráficos de las Figuras 4.4 y 4.5. Si además se tiene en cuenta las predicciones para la media que se tienen en la Tabla 4.7, es fácil ver que partiendo de los peores valores de OBPM y DBPM, según los modelos planteados, está mejor pagado tener más nivel en ataque. Sin embargo, a medida que se avanza hacia los valores medios, los mejores salarios se los llevan aquellos que contribuyen más en la defensa. Esta tendencia cambia cuando los datos se acercan a su máximo de forma los mejores jugadores en ataque son los que tienen los sueldos más elevados, por encima de los sueldos de los mejores jugadores en defensa. Si esto se compara con los salarios reales de estos jugadores "extremos" en la temporada 2020-2021, se aprecia que se le da más importancia a ser bueno en el ataque, ya

que por ejemplo, a Stephen Curry se le paga por encima de lo que predice el modelo, mientras que al jugador con menos OBPM, Josh Hall, se le paga menos de lo que se debería. Por otro lado, el modelo ajustado según los datos de DBPM indica que Matisse Thybulle debería cobrar más de 2 millones de dólares más y Howard, el peor defensor, debería cobrar unos miles de dólares menos.

La conclusión a la que se puede llegar tras observar estos resultados es que para jugadores que no destacan por encima del resto, los entrenadores buscan que sean efectivos en defensa, aportando al equipo de esa forma, y dejando así que destaquen en el ataque aquellos jugadores más visuales y por tanto más exitosos con salarios desorbitados.

4.2. VORP vs Age

Uno se puede plantear si existe algún tipo de relación lineal entre el valor de un jugador y su edad, esto es si los jugadores más jóvenes presentan una mayor eficiencia por tener una mejor condición física, o menor eficacia por carecer de experiencia.

Una buena forma de resolver esta duda es planteando un modelo que relacione la edad de los jugadores de la NBA con su valor. Se introduce el VORP en Winston (2012) como una medida estudiada por primera vez en el béisbol. Esta cualidad, convierte el ya estudiado BMP en una contribución total al equipo por parte de cada jugador comparado con lo que un jugador de reemplazo aportaría al equipo. Se entiende por un jugador de reemplazo a aquel con el mínimo salario o aquel que no está en las rotaciones habituales. El nivel de reemplazamiento para la liga de baloncesto americana se estableció en -2.0 medido en puntos por encima o por debajo de la media por cada 100 posesiones. De acuerdo con este valor, el VORP de un jugador se calcula según la siguiente fórmula.

$$\text{VORP} = [\text{BPM} - (-2.0)] \cdot (\%\text{posesiones jugadas}) \cdot (\text{partidos jugados}/82) \tag{4.8}$$

Observación 4.2. El hecho de que se divida entre 82 es porque dicho número es el número de partidos de liga regular de la NBA.

Además sin más que multiplicar este valor por 2.7 se obtienen las victorias sobre reemplazo, lo cual relaciona la eficiencia de un jugador con las victorias de su equipo (Myers, 2022).

Una cualidad interesante es que debería estimar el salario linealmente. De hecho, es razonable pensar que los salarios de los jugadores más veteranos de la NBA son los más altos, ya que aquellos que siguen jugando pasada una cierta edad, quiere decir que son jugadores con muy buenos números, pues en otro caso se habrían retirado. Entonces lo que se pretende ver en esta

sección es si existe algún tipo de relación entre la edad y el valor de un jugador, de forma que se pueda aplicar posteriormente a un estudio sobre el salario.

4.2.1. Formulación del modelo

Aplicando el mismo procedimiento que se lleva realizando durante los distintos capítulos se parte del modelo más simple posible que se conoce. De esta forma, la variable respuesta de este modelo, Y, será el VORP y la variable explicativa X_1 es la edad. El resto de elementos de la ecuación (4.9), son β_0, β_1, los coeficientes de la regresión y ε los errores.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon. \tag{4.9}$$

Una vez formulado el modelo en R se deben contrastar las hipótesis habituales del modelo de regresión lineal. Dichos contrastes muestran que no se tiene un modelo lineal y que tampoco hay normalidad en los errores. Esto último es consecuencia de que los datos sean tomados sobre experiencias reales. En cuanto a intentar corregir la hipótesis de linealidad, se plantea un modelo aditivo que utiliza una función de suavizado sobre la edad. Si se representa dicha función se puede ver que se aprecia una curvatura, pero que sin embargo semeja mucho una línea recta. Además en el resumen de las características de este posible modelo, se ve que los grados de libertad efectivos de la función de suavizado no llegan a 2 (lo correspondiente a un polinomio cuadrático). Esto es un indicativo de que podría presentar un comportamiento casi lineal. Para ver dicha representación se puede consultar la Figura I.4

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	-1.13367	0.34248	-3.31	0.001**
Age	0.06701	0.01314	5.10	<0.001 ***

Tabla 4.10: Resumen del modelo lineal simple VORP vs edad

Ahora bien, se tiene en la Tabla 4.10 la salida del `summary` de R. De acuerdo con los resultados, el modelo seguiría la siguiente formulación.

$$Y = -1.134 + 0.067 \times \text{Age}. \tag{4.10}$$

Luego un aumento de una unidad en la variable edad supone incrementar el VORP en 0.067 unidades. Por ejemplo, si se toma la media de las edades de los jugadores que se están estudiando, es decir, 25.75, se tendrá un VORP de 0.5917.

Análisis de los outliers

Se pretende entonces hacer un estudio de los datos que no se ajustan bien mediante el modelo, con el objetivo de saber si eliminando dichos datos se consigue tener una buena aproximación y tratar de darle una explicación a dichas anomalías.

El procedimiento que se sigue es análogo al de la sección anterior.

Por un lado, en primer lugar se puede hacer una representación gráfica del modelo, del ajuste y de los puntos con capacidad de influencia, es decir, los apalancamientos. Dado que en el modelo lineal simple se presenta una única variable que pretende explicar la respuesta, no existe la opción de que la influencia de otras variables interfieran en el ajuste, de forma que no se pueda dar una explicación razonable de los valores de los residuos. Es por esto que no será necesario utilizar los apalancamientos, y se pueden estudiar directamente los residuos estandarizados de la regresión.

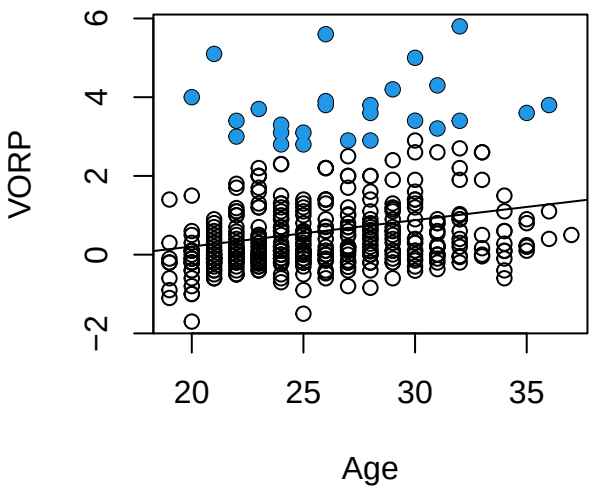

Figura 4.6: Ajuste del modelo, recta de regresión y residuos estandarizados (en azul).

En la Figura 4.6, se pintan en azul los residuos estandarizados que superan en valor absoluto el límite previamente definido, es decir, el valor de una normal estándar que contiene al 95 % de los individuos que se estudian.

Una vez se localizan los puntos con altos valores absolutos para los residuos estandarizados, se trata de detectar a que jugadores hacen referencia y ver que es lo que les hace convertirse en valores extraños.

Jugador	Edad	VORP estimado	VORP Real
Bam Adebayo	23.00	3.70	0.407
Giannis Antetokounmpo	26.00	5.60	0.608
Bradley Beal	27.00	2.90	0.676
Mikal Bridges	24.00	3.10	0.475
Jimmy Butler	31.00	4.30	0.944
Stephen Curry	32.00	5.80	1.011
Luka Doncic	21.00	5.10	0.274
Joel Embiid	26.00	3.80	0.609
Rudy Gobert	28.00	3.80	0.7243
James Harden	31.00	3.20	0.944
Tobias Harris	28.00	2.90	0.743
Kyrie Irving	28.00	3.60	0.743
LeBron James	36.00	3.80	1.279
Zach LaVine	25.00	3.10	0.542
Kawhi Leonard	29.00	4.20	0.810
Damian Lillard	30.00	5.00	0.877
Donovan Mitchell	24.00	2.80	0.475
Chris Paul	35.00	3.60	1.212
Julius Randle	26.00	3.90	0.609
Domantas Sabonis	24.00	3.30	0.475
Jayson Tatum	22.00	3.40	0.341
Karl-Anthony Towns	25.00	2.80	0.542
Nikola Vucevic	30.00	3.40	0.877
Russell Westbrook	32.00	3.40	1.010
Zion Williamson	20.00	4.00	0.207
Trae Young	22.00	3.00	0.341

Tabla 4.11: Jugadores que producen residuos grandes en el ajuste

El modelo que se ha definido para este análisis, es un modelo lineal simple con la característica de que, conforme aumenta la edad, también se tiene un aumento en el VORP del jugador. Dicho razonamiento se puede ver en la Figura 4.6, donde se ha representado la recta de regresión.

Ahora bien, parece razonable pensar que un aumento en la edad mejore el nivel de un jugador por el hecho de que ganar experiencia así como entrenar durante largos períodos de tiempo hacen que un jugador pueda aportar más. Esto puede parecer lógico hasta un punto, esto es, al llegar a ciertas edades, la condición física de los jugadores empeora, de forma que su valor comienza a

disminuir (Kalén et al., 2020).

En cuanto a los jugadores recogidos en la Tabla 4.11, es claro ver que el modelo predice valores de VORP pequeños. Todos ellos aumentan con la edad, sin embargo son todos más bajos que los valores reales. Si uno se centra en el jugador más joven de dicha tabla, Zion Williamson, su VORP tiene un valor de 4, mayor que 2.9 que es el VORP de Tobias Harris, 8 años mayor que el. Por otro lado, Lebron James, cuya edad se encuentra entre las más elevadas de la NBA (ya se vio anteriormente que el máximo estaba en 37 años), tiene un valor de 3.80 y el valor estimado por el modelo sería de 1.28. A pesar de superar este valor, es claro que no es acorde si se tienen en cuenta que a mayor edad, mayor VORP.

Todos estos valores anómalos llevan a uno a pensar que la explicación del valor sobre un jugador de reemplazo no es explicable unicamente mediante la edad. De hecho si se realiza el `summary` del modelo, se obtiene que el valor del coeficiente de determinación es de 0.05952, lo cual indica que dicho ajuste solo explica menos de un 6 % de la varianza de la respuesta.

Dado que la intención principal del presente trabajo era la estimación de los salarios a través de ciertas características del juego de los deportistas, conviene relacionar los resultados obtenidos en este apartado con dicho estudio. Así de acuerdo con la explicación de las variables sacadas de Forman (2000), el salario debe relacionarse linealmente con el VORP, y por tanto se concluye tras el estudio realizado en este apartado, que la edad, no sería una buena característica para tratar de explicar la asignación justa de los pagos de la NBA, como era lógico pensar.

A lo largo del trabajo se ha hecho un análisis de ciertas variables con el objetivo principal de buscar una buena aproximación que pueda ser usada por entrenadores, responsables de los equipos y directivos, a la hora de decidir cual es el sueldo justo de un jugador según su actuación en la liga regular. Además puede resultar también útil a los jugadores en ánimos de ver en que deberían enfocar sus entrenamientos para obtener un sueldo mayor. Los resultados no fueron óptimos a la hora de plantear modelos lineales, pero gracias a los modelos aditivos se obtuvieron mejores aproximaciones y consecuentemente mejores interpretaciones. Finalmente los pequeños modelos planteados que trataban de comparar distintas variables que parecían ser importantes por si solas, mostraron por un lado, que efectivamente el Box Plus/Minus es un tipo de estadística muy útil y que produce buenos resultados y por otro lado que la edad no basta para tener una predicción de los valores del VORP, y por tanto que la misma, tampoco influye por si sola sobre el salario.

Bibliografía

Vincent Arel-Bundock. *marginaleffects: Marginal Effects, Marginal Means, Predictions, and Contrasts*, 2022. URL `https://CRAN.R-project.org/package=marginaleffects`. R package version 0.5.0.

R.D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1982. ISBN 9780412242809.

N.R. Draper, N.R. Draper, H. Smith, and H.K. Smith. *Applied Regression Analysis*. Number v. 1 in Applied Regression Analysis. Wiley, 1998. ISBN 9780471170822.

J.J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. A Chapman & Hall Book. CRC Press, Taylor & Francis Group, 2016. ISBN 9781498720960.

Sean Forman. Basketball Reference sports reference, 2000. URL `https://www.basketball-reference.com/leagues/NBA_2021_advanced.html`.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387848587.

R.V. Ibarrola. *Cálculo de probabilidades 2*. Number v. 2 in Cálculo de probabilidades. EDICIONES ACADÉMICAS S.A., 2004. ISBN 9788496062412.

Anton Kalén, Alexandra Pérez-Ferreirós, Pablo B. Costa, and Ezequiel Rey. Effects of age on physical and technical performance in national basketball association (nba) players. *Research in Sports Medicine*, 29(3):277–288, 2020. doi: 10.1080/15438627.2020.1809411.

Daniel Myers. About box plus/minus (bpm) | basketball-reference.com, 2022. URL `https://www.basketball-reference.com/about/bpm2.html`.

J.B. Ramsey. *Tests for Specification Errors in Classical Linear Least Squares Regression Analysis*. University of Wisconsin–Madison, 1968.

Noam Ross. Gams in R by Noam Ross noam ross, 2019. URL https://noamross.github.io/gams-in-r-course/.

P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* Wiley Series in Probability and Statistics. Wiley, 2005. ISBN 9780471725374.

T.P. Ryan. *Modern Regression Methods.* Wiley Series in Probability and Statistics. Wiley, 2008. ISBN 9780470081860.

S. Sheather. *A Modern Approach to Regression with R.* Springer Texts in Statistics. Springer New York, 2009. ISBN 9780387096070.

Jorge Sierra. Hoops hype, 2002. URL https://hoopshype.com/salaries/players/2020-2021/.

Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011. doi: 10.1111/j.1467-9868.2011.00771.x.

W.L. Winston. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football.* Princeton University Press, 2012. ISBN 9781400842070.

S. Wood. *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2006. ISBN 9781584884743.

Anexo I

Tablas y Figuras complementarias

	Age	MP	PER	TS	X3Par	FTr	ORB	DRB	TRB	AST	STL	BLK	TOV	USG	OWS	DWS	WS	WS48	OBPM	DBPM	BPM	VORP	V2
Age	1.00	0.16	0.17	0.19	0.11	-0.04	-0.04	0.06	0.02	0.15	0.06	-0.01	0.00	0.01	0.22	0.20	0.24	0.25	0.25	0.23	0.31	0.24	0.43
MP	0.16	1.00	0.42	0.29	-0.01	0.03	-0.14	0.04	-0.03	0.32	-0.01	-0.10	-0.03	0.38	0.64	0.80	0.77	0.32	0.55	0.07	0.51	0.61	0.62
PER	0.17	0.42	1.00	0.64	-0.46	0.43	0.36	0.42	0.44	0.40	0.09	0.32	-0.05	0.60	0.75	0.49	0.74	0.84	0.87	0.25	0.87	0.78	0.46
TS	0.19	0.29	0.64	1.00	-0.18	0.30	0.28	0.23	0.28	-0.09	-0.10	0.25	-0.07	0.05	0.61	0.33	0.57	0.80	0.63	0.25	0.65	0.46	0.23
X3Par	0.11	-0.01	-0.46	-0.18	1.00	-0.02	-0.08	-0.47	-0.60	-0.08	-0.01	-0.45	-0.32	-0.14	-0.19	-0.18	-0.21	-0.32	-0.06	-0.19	-0.13	-0.13	-0.08
FTr	-0.04	0.03	0.43	0.30	-0.02	1.00	0.45	0.36	0.43	0.07	-0.02	0.35	0.26	0.15	0.28	0.15	0.26	0.37	0.17	0.18	0.22	0.22	0.05
ORB	-0.04	-0.14	0.36	0.28	-0.68	0.45	1.00	0.65	0.86	-0.32	-0.05	0.65	0.18	-0.18	0.12	0.13	0.13	0.44	-0.01	0.27	0.09	0.02	-0.07
DRB	0.06	0.04	0.42	0.23	-0.47	0.36	0.65	1.00	0.95	-0.12	-0.14	0.52	0.22	0.07	0.18	0.33	0.25	0.36	0.15	0.34	0.26	0.25	0.17
TRB	0.02	-0.03	0.44	0.28	-0.60	0.43	0.86	0.95	1.00	-0.22	-0.12	0.62	0.22	-0.03	0.18	0.28	0.23	0.44	0.10	0.35	0.22	0.18	0.08
AST	0.15	0.32	0.40	-0.09	-0.08	0.07	-0.32	-0.12	-0.22	1.00	0.28	-0.29	0.31	0.58	0.33	0.24	0.33	0.13	0.47	0.05	0.43	0.51	0.36
STL	0.06	-0.01	0.09	-0.10	-0.01	-0.02	-0.05	-0.14	-0.12	0.28	1.00	-0.06	0.22	0.00	-0.03	0.13	0.02	0.10	0.02	0.51	0.21	0.12	0.04
BLK	-0.01	-0.10	0.32	0.25	-0.45	0.35	0.65	0.52	0.62	-0.29	-0.06	1.00	0.11	-0.13	0.06	0.18	0.11	0.35	-0.00	0.46	0.17	0.06	-0.01
TOV	0.00	-0.03	-0.05	-0.07	-0.32	0.26	0.18	0.22	0.22	0.31	0.22	0.11	1.00	-0.03	-0.11	0.10	-0.04	-0.14	-0.22	0.26	-0.10	-0.01	0.05
USG	0.01	0.38	0.60	0.05	-0.14	0.15	-0.18	0.07	-0.03	0.58	0.00	-0.13	-0.03	1.00	0.37	0.27	0.37	0.15	0.60	-0.27	0.43	0.53	0.44
OWS	0.22	0.64	0.75	0.61	-0.19	0.28	0.12	0.18	0.18	0.33	-0.03	0.06	-0.11	0.37	1.00	0.59	0.95	0.77	0.80	0.20	0.78	0.88	0.50
DWS	0.20	0.80	0.49	0.33	-0.18	0.15	0.13	0.33	0.28	0.24	0.13	0.18	0.10	0.27	0.59	1.00	0.81	0.49	0.48	0.46	0.60	0.68	0.56
WS	0.24	0.77	0.74	0.57	-0.21	0.26	0.13	0.25	0.23	0.33	0.02	0.11	-0.04	0.37	0.95	0.81	1.00	0.75	0.77	0.32	0.80	0.90	0.57
WS48	0.25	0.32	0.84	0.80	-0.32	0.37	0.44	0.36	0.44	0.13	0.10	0.35	-0.14	0.15	0.77	0.49	0.75	1.00	0.76	0.49	0.85	0.69	0.31
OBPM	0.25	0.55	0.87	0.63	-0.06	0.17	-0.01	0.15	0.10	0.47	0.02	-0.00	-0.22	0.60	0.80	0.48	0.77	0.76	1.00	0.12	0.93	0.84	0.54
DBPM	0.23	0.07	0.25	0.25	-0.19	0.18	0.27	0.34	0.35	0.05	0.51	0.46	0.26	-0.27	0.20	0.46	0.32	0.49	0.12	1.00	0.48	0.34	0.15
BPM	0.31	0.51	0.87	0.65	-0.13	0.22	0.09	0.26	0.22	0.43	0.21	0.17	-0.10	0.43	0.78	0.60	0.80	0.85	0.93	0.48	1.00	0.87	0.53
VORP	0.24	0.61	0.78	0.46	-0.13	0.22	0.02	0.25	0.18	0.51	0.12	0.06	-0.01	0.53	0.88	0.68	0.90	0.69	0.84	0.34	0.87	1.00	0.57
V2	0.43	0.62	0.46	0.23	-0.08	0.05	-0.07	0.17	0.08	0.36	0.04	-0.01	0.05	0.44	0.50	0.56	0.57	0.31	0.54	0.15	0.53	0.57	1.00

Tabla I.1: Colinealidad entre las variables explicativas

Salario vs BPM

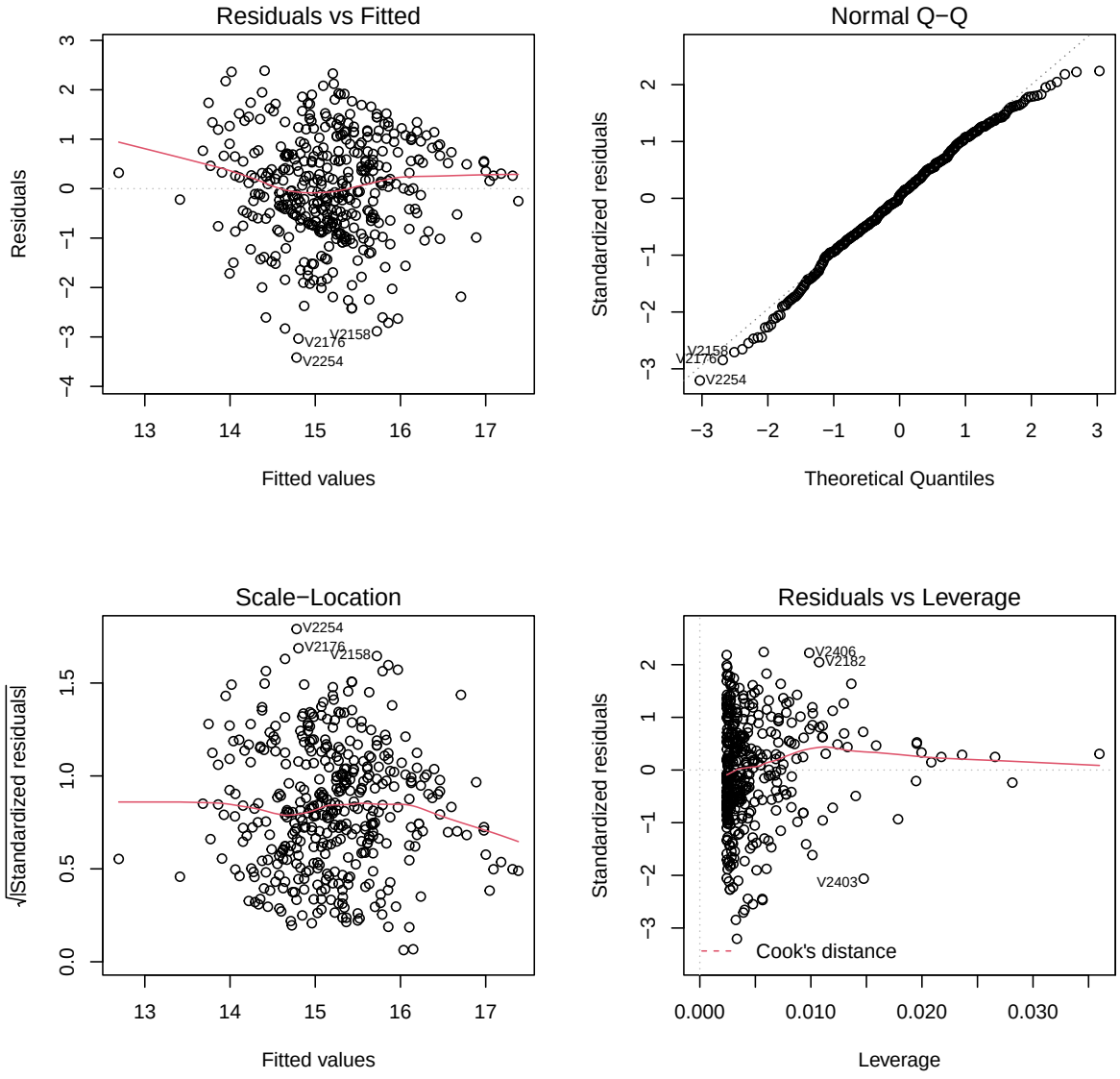

Figura I.1: Gráficos para la validación y diagnosis del Salario vs BPM

Salario Vs OBPM

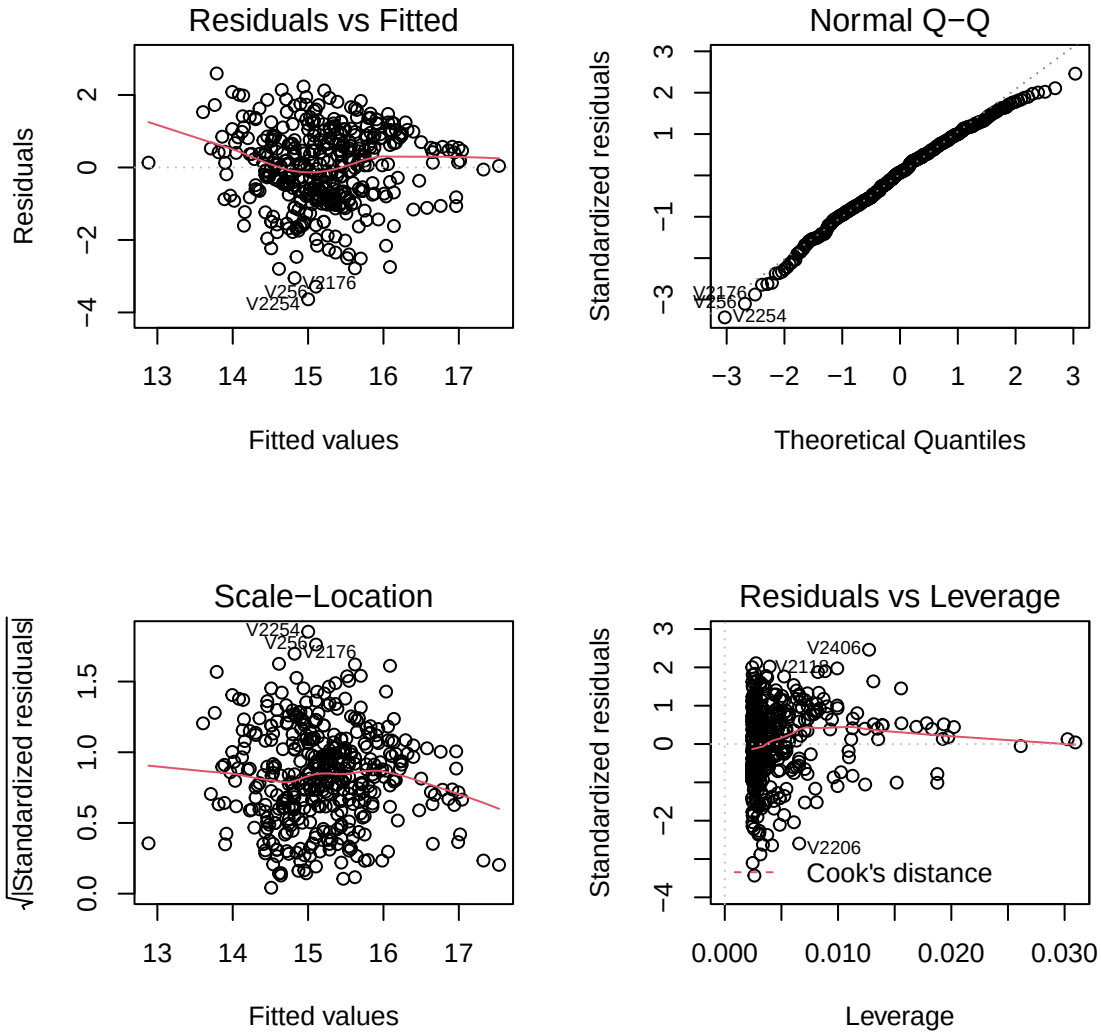

Figura I.2: Gráficos para la validación y diagnosis. Salario vs OBPM

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	15.19410	0.05125	296.5	<2e-16 ***
	edf	Ref.df	F	p-value
s(OBPM)	3.769	4.715	40.85	<2e-16 ***

Tabla I.2: Características modelo adtivos Salario vs OBPM.

Salario Vs DBPM

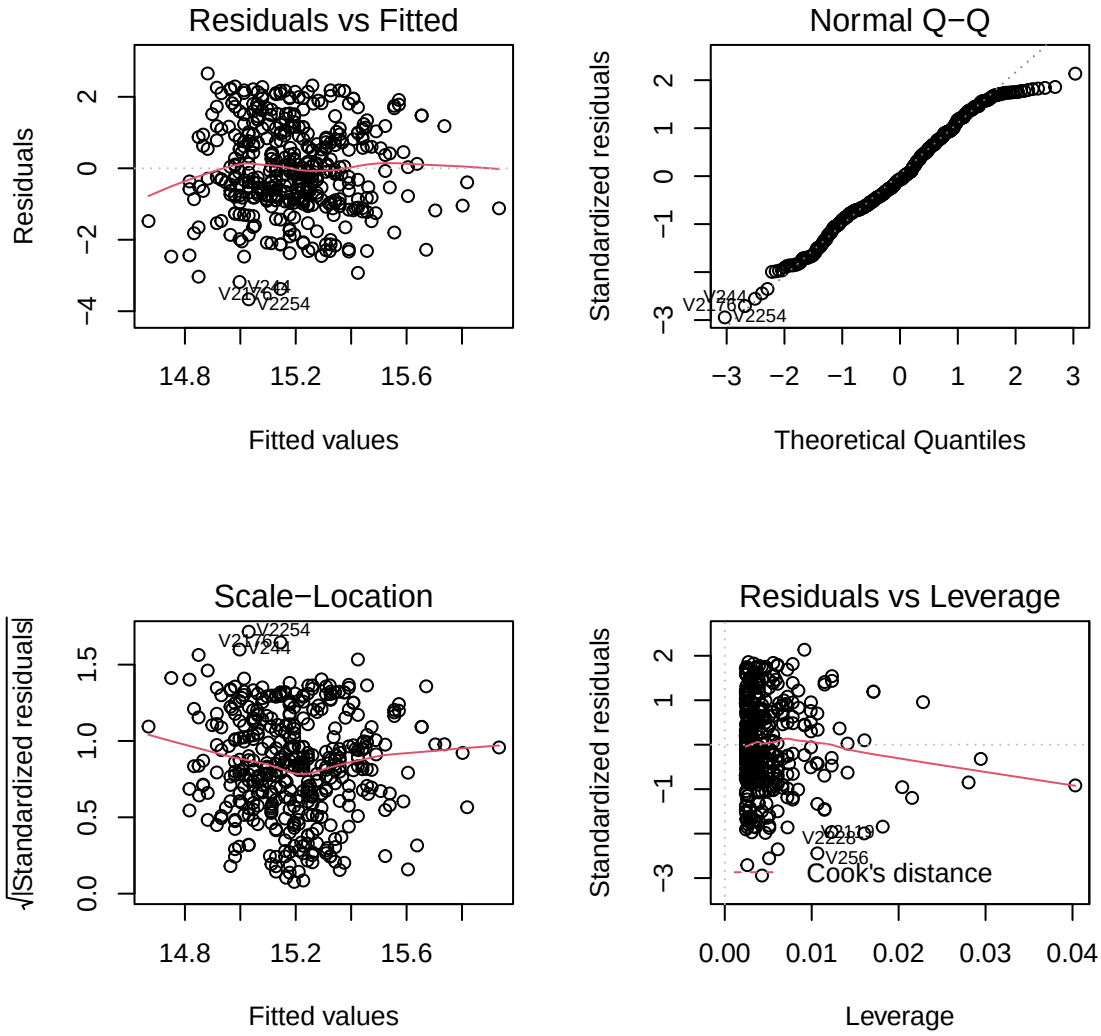

Figura I.3: Gráficos para la validación y diagnosis. Salario vs DBPM.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercepto	15.19410	0.06098	249.2	<2e-16 ***
	edf	Ref.df	F	p-value
s(DBPM)	3.695	4.617	3.266	0.00852 **

Tabla I.3: Características modelo aditivo Salario vs DBPM.

VORP vs Age

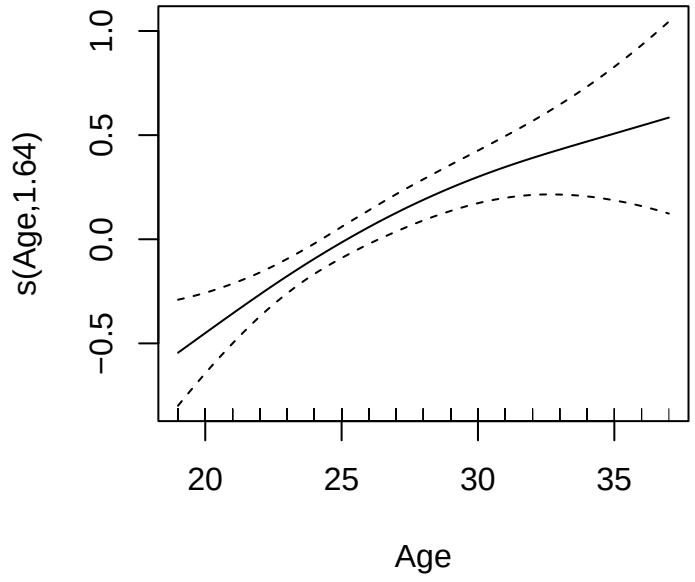

Figura I.4: Gráficos para la validación y diagnosis para VORP vs Age

Anexo II

Código de R

```
#-----------------------------------------------------------#
# Librerias y paquetes necesarios
#-----------------------------------------------------------#
library(data.table)
#install.packages("car")
library(car)
#install.packages("lmtest")
library(lmtest)
#install.packages("sm")
library(sm)
library(MASS)
library(glmnet)
library(mgcv)
library(visreg)
library(xtable)
#install.packages("marginaleffects")
library(marginaleffects)
library(ggplot2)

#-----------------------------------------------------------#
# Cargamos los datos
#-----------------------------------------------------------#
dataplayer=read.csv2("datanew.csv",dec=".",string=TRUE)
salarios=read.csv2("salarios2021.csv",dec=".",string=TRUE)
head(dataplayer)
```

```
head(salarios)
names(dataplayer)
names(salarios)
attach(dataplayer)
attach(salarios)
log(salarios)

#-------------------------------------------------------------#
#            ESTUDIO DE LAS VARIABLES
#-------------------------------------------------------------#

# Seleccionamos variables que queremos estudiar (las continuas)
data_cont=c("Age","G","MP","PER","TS","X3PAr","FTr","ORB","DRB",
 "TRB","AST","STL","BLK","TOV","USG","OWS","DWS","WS" ,
 "WS48","OBPM","DBPM","BPM","VORP")
data_cont2=as.matrix(na.omit(dataplayer[,data_cont]))
PF<- ifelse(Pos == 'PF', 1,0)
PG<- ifelse(Pos == 'PG', 1, 0)
C<- ifelse(Pos == 'C', 1, 0)
SG<- ifelse(Pos == 'SG', 1, 0)
SF<- ifelse(Pos == 'SF', 1, 0)


# Resumen de las variables
summary(dataplayer)
xtable(summary(dataplayer))

# Matriz de covarianzas y correlaciones de las variables
# explicativas
Sigma=cov(data_cont2); Sigma
Rho=cor(data_cont2); Rho



#-------------------------------------------------------------#
#           ESTUDIO DE LOS SALARIOS                           #
#-------------------------------------------------------------#
```

```
salariosreales2=as.matrix(as.numeric(na.omit(salarios[,"Real"])))
summary(salariosreales2)
logsalarios=as.matrix(as.numeric(na.omit(log(salarios[,"Real"]))))

plot(salariosreales2)

#-----------------------------------------------------------------#
# Medias ponderadas de los jugadores que juegan en mas de un equipo
# en la misma temporada.

Rk_new<-as.character(Rk)

DT <- data.table(data_cont2)
#DT[,list(wret = weighted.mean(MP,G)),by=Rk] solo para los minutos jugados
newdata<-DT[,list(Age=mean(Age), MP = weighted.mean(MP,G),
PER=weighted.mean(PER,G),TS=weighted.mean(TS,G),
X3Par=weighted.mean(X3PAr,G),FTr=weighted.mean(FTr,G),
ORB=weighted.mean(ORB,G),
DRB=weighted.mean(DRB,G),TRB=weighted.mean(TRB,G),
AST=weighted.mean(AST,G),STL=weighted.mean(STL,G),
BLK=weighted.mean(BLK,G),TOV=weighted.mean(TOV,G),
USG=weighted.mean(USG,G),OWS=weighted.mean(OWS,G),
DWS=weighted.mean(DWS,G),WS=weighted.mean(WS,G),
WS48=weighted.mean(WS48,G),OBPM=weighted.mean(OBPM,G),
DBPM=weighted.mean(DBPM,G),BPM=weighted.mean(BPM,G),
VORP=weighted.mean(VORP,G)),by=Player]

posiciones<-DT[,list(pos = unique(Pos)),by=Player]
# Pruebas para ver si calcula bien las medias.
newdata[1]
newdata[5]
newdata[6]



# Emparejamos jugadores con salarios
```

```
x<-match(newdata[,Player],Jugador)
sum(is.na(x))
salariosord<-salarios[x,]
length(salariosord[,"Jugador"])
sts_salar<-cbind(newdata,log(salariosord[,"Real"]))

sts_salar1<-na.omit(sts_salar)
length(x)-sum(is.na(x))



#------------------------------------------------------------------#
# Eliminamos los jugadores que hayan jugado menos de
#   x minutos

out<-which(sts_salar1[,MP]<100)
sts_salar_final<-sts_salar1[-out]

# Comprobación de que estamos eliminando los jugadores con menos de 100
# minutos.

length(sts_salar1[,Player])
length(sts_salar_final[,Player])
length(sts_salar1[,Player])-length(sts_salar_final[,Player])
length(out)

summary(sts_salar_final)
xtable(summary(sts_salar_final))
xtable(summary(Age))

#------------------------------------------------------------------#
#  Estudio de las variables que mejor me ajusten la
#  variable explicativa
    #------------------------------------------------------------------#

# Matriz de covarianzas y correlaciones
Sigma=cov(data_cont2)
Rho=cor(data_cont2)
```

```
xtable(Rho)

a<-sts_salar_final[,c("Age","MP","PER","TS","X3Par","FTr","ORB","DRB",
"TRB","AST","STL","BLK","TOV","USG","OWS",
"DWS","WS","WS48","OBPM", "DBPM","BPM","VORP","V2")]
Sigma2=cov(a); Sigma2[,"V2"]
Rho2=cor(a)
Rho2[,"V2"]
xtable(Rho2)
#-----------------------------------------------------------------------#
#     MODELO DE REGRESIÓN LINEAL MÚLTIPLE                               #
    #-----------------------------------------------------------------------#
salario<-unlist(sts_salar_final[,"V2"])
Age<-unlist(sts_salar_final[,"Age"])
MP<-unlist(sts_salar_final[,"MP"])
PER<-unlist(sts_salar_final[,"PER"])
STL<-unlist(sts_salar_final[,"STL"])
USG<-unlist(sts_salar_final[,"USG"])
OWS<-unlist(sts_salar_final[,"OWS"])
DWS<-unlist(sts_salar_final[,"DWS"])
WS<-unlist(sts_salar_final[,"WS"])
OBPM<-unlist(sts_salar_final[,"OBPM"])
BPM<-unlist(sts_salar_final[,"BPM"])
VORP<-unlist(sts_salar_final[,"VORP"])
DBPM<-unlist(sts_salar_final[,"DBPM"])

modelo=lm(salario~Age+MP+PER+STL+USG+OWS+DWS+WS+OBPM+BPM+VORP)

#-----------------------------------------------------------------------#
# Analisis del modelo
#-----------------------------------------------------------------------#
summary(modelo)
xtable(summary(modelo))

# Validación
windows()
par(mfrow=c(2,2))
```

```
plot(modelo)

# Contraste de hipótesis:


# LINEALIDAD
#x=c(Age,MP,PER,STL,USG,OWS,DWS,WS,OBPM,BPM,VORP)
#sm.regression(x,salario,model="linear")
resettest(modelo,power=3)

# NORMALIDAD
res<-rstandard(modelo)
shapiro.test(res)

# HOMOCEDASTICIDAD
hmctest(modelo)

# INDEPENDENCIA
durbinWatsonTest(modelo)

#-----------------------------------------------------------------#
# Nueva selección de las varibles
#-----------------------------------------------------------------#

X=matrix(c(Age,MP,PER,STL,USG,OWS,DWS,WS,OBPM,BPM,VORP),ncol=11)
res.lasso=cv.glmnet(X,salario,alhpa=1)
res.lasso
coef(res.lasso,s=res.lasso$lambda.1se)
coef(res.lasso,s=res.lasso$lambda.min)
windows()
par(new=TRUE)
plot(res.lasso)
plot(res.lasso$glmnet.fit,xvar="lambda")

# Seleeción de variables por AIC
step(modelo)
```

```
# Definicion del modelo con las nuevas variables
modelo2<-lm(salario~Age+MP+USG+DWS+BPM+VORP)

# mode1=glmnet(X,salario,alhpa=1)
# coef(mode1,s=res.lasso$lambda.1se)
summary(modelo2)
windows()
par(mfrow=c(2,2))
plot(modelo2)
# LINEALIDAD

#sm.regression(x,salario,model="linear")
resettest(modelo2,power=3)

# NORMALIDAD
res2<-rstandard(modelo2)
shapiro.test(res2)

# HOMOCEDASTICIDAD
hmctest(modelo2)

# INDEPENDENCIA
durbinWatsonTest(modelo2)

AIC(modelo)
AIC(modelo2)
#-----------------------------------------------------------------#
#                        MODELO NO LINEAL                         #
#-----------------------------------------------------------------#
b1<-sts_salar_final[,c("Age","MP","USG","V2")]
b2<-sts_salar_final[,c("DWS","BPM","VORP","V2")]
pairs(b1, panel = panel.smooth)
pairs(b2, panel = panel.smooth)
gam1<-gam(salario~s(Age,bs="cr")+s(MP,bs="cr")+s(USG,bs="cr")
+s(DWS,bs="cr")+s(BPM,bs="cr")+s(VORP,bs="cr"),method="REML")
plot(gam1)
summary(gam1)
```

```
par(mfrow=c(3,2))
plot(gam1,pages=1,se = TRUE)

gam2<-gam(salario~s(Age,bs="cr")+s(MP,bs="cr")+s(USG,bs="cr")
+DWS+s(BPM,bs="cr")+s(VORP,bs="cr"),method="REML")
summary(gam2)
gam.check(gam2)
plot(gam.check(gam2))

#------Seleccion manual del mejor modelo-----------#

gam1b<-gam(salario~s(Age,bs="cr")+s(MP,bs="cr")+s(USG,bs="cr")
+s(DWS,bs="cr")+s(VORP,bs="cr"),method="REML")
gam1c<-gam(salario~s(Age,bs="cr")+s(MP,bs="cr")+s(USG,bs="cr")
+s(VORP,bs="cr"),method="REML")
summary(gam1b)
summary(gam1c)

gam.check(gam1c)

#-----------------------------------------------------------------#
# Interpretación del gam1c (marginal effects)
#-----------------------------------------------------------------#
pred<-predictions(gam1c, newdata = "mean")
summary(pred)
xtable(predictions(gam2b, newdata = "mean"))
exp(15.21)
plot_cap(gam1c,condition=c("MP","Age"))+ggplot2::ylab("log(Salario)")

#------------------ AME----------------------#
mf <- marginaleffects(gam1c)
summary(mf)
xtable(summary(mf))

marginaleffects(gam1c, newdata = "mean")
xtable(marginaleffects(gam1c, newdata = "mean"))
```

```
#---------------------G AME --------------------#
BPM_cat<-cut(BPM,breaks=c(-12,0,10),
labels=c("Baja efectividad", "Alta efectividad"))
gam4<-gam(salario~s(Age,bs="cr")+s(MP,bs="cr")+s(USG,bs="cr")
+BPM_cat+s(VORP,bs="cr"),method="REML")

game <- marginaleffects(gam4, variables = "Age")
summary(game, by = "BPM_cat")
exp(0.05)
exp(0.09)
summary(gam4)

summary(marginaleffects(gam1c,
newdata = datagrid(
Age = 30,MP=1772,USG=16,VORP=0)))

#--------------- Comparisons-----------------------#

com<-comparisons(gam1c,variables = "Age")
summary(com)
summary(comparisons(gam1c, variables = list(Age = c(26, 31))))
summary(comparisons(gam1c, variables = list(VORP= "sd")))
exp(0.4)

#-------------------------------------------------------------------#
#           DISTINTAS VARIABLES VS DISTINTAS RESPUESTAS             #
#-------------------------------------------------------------------#
mod1<-lm(salario~BPM)
summary(mod1)
par(mfrow=c(2,2))
plot(mod1)
summary(mod1)
sm.regression(BPM,salario,model="linear")
resettest(mod1,power=3)
harvtest(mod1)
res1<-rstandard(mod1)
shapiro.test(res1)
```

```
hmctest(mod1)
durbinWatsonTest(mod1)

gam_mod1<-gam(salario~s(BPM,bs="cr"),method="REML")
summary(gam_mod1)
gam.check(gam_mod1)
plot(gam_mod1)
res1g<-residuals(gam_mod1)
mean(res1g)

predgam<-predictions(mod1, newdata = "mean")
summary(predgam)
xtable(predictions(mod1, newdata = "mean"))
exp(15.19)


# Estimación para peor y mejor BPM
which(a[,BPM]==min(BPM))
sts_salar_final[which(a[,BPM]==min(BPM))]
pred2<-predictions(mod1, newdata = datagrid(BPM=-11.9))
summary(pred2)
xtable(summary(pred2))
which(a[,BPM]==max(BPM))
sts_salar_final[which(a[,BPM]==max(BPM))]
pred3<-predictions(mod1, newdata = datagrid(BPM=9))
summary(pred3)

mfg <- marginaleffects(mod1)
summary(mfg)
xtable(summary(mfg))
plot_cap(mod1,condition="BPM")+ggplot2::ylab("log(Salario)")

#Analisis de los residuos

res<-residuals(mod1)
plot(residuals(mod1) ~ predict
(gam_mod1,type="response"),
```

```
xlab="Fitted Values",ylab="residuals")
inddev<-which(abs(res)>2)
points(residuals(mod1)[inddev]~predict
(mod1,type="response")[inddev],col=2,pch=16)

resp_outliers<-predict(mod1,type="response")
resp_outliers[inddev]
exp(resp_outliers[inddev])

#Comparación OBPM con DBPM
mod2<-lm(salario~OBPM)
mod3<-lm(salario~DBPM)
par(mfrow=c(2,2))
plot(mod2)
plot(mod3)
sm.regression(OBPM,salario,model="linear")
sm.regression(DBPM,salario,model="linear")

gam_mod2<-gam(salario~s(OBPM,bs="cr"),method="REML")
summary(gam_mod2)
gam.check(gam_mod2)

gam_mod3<-gam(salario~s(DBPM,bs="cr"),method="REML")
summary(gam_mod3)
gam.check(gam_mod3)
plot(gam_mod3)

predgam_mod2<-predictions(gam_mod2, newdata = "mean")
summary(predgam_mod2)
xtable(predictions(gam_mod2, newdata = "mean"))
predgam_mod3<-predictions(gam_mod3, newdata = "mean")
summary(predgam_mod3)
xtable(predictions(gam_mod3, newdata = "mean"))


windows()
par(mfrow=c(1,2))
```

```
plot_cap(gam_mod2,condition="OBPM")+ggplot2::ylim(12,19)
+ggplot2::ylab("log(Salario)")
plot_cap(gam_mod3,condition="DBPM")+ggplot2::ylim(12,19)
+ggplot2::ylab("log(Salario)")


sts_salar_final[which(a[,OBPM]==min(OBPM))]
predgam_mod2_1<-predictions(gam_mod2, newdata = datagrid(OBPM=-9.7))
summary(predgam_mod2_1)
xtable(summary(predgam_mod2_1))
sts_salar_final[which(a[,OBPM]==max(OBPM))]
predgam_mod2_2<-predictions(gam_mod2, newdata = datagrid(OBPM=8.3))
summary(predgam_mod2_2)
xtable(summary(predgam_mod2_2))

sts_salar_final[which(a[,DBPM]==min(DBPM))]
predgam_mod3_1<-predictions(gam_mod3, newdata = datagrid(DBPM=-3.2))
summary(predgam_mod3_1)
xtable(summary(predgam_mod3_1))
sts_salar_final[which(a[,DBPM]==max(DBPM))]
predgam_mod3_2<-predictions(gam_mod3, newdata = datagrid(DBPM=4.5))
summary(predgam_mod3_2)
xtable(summary(predgam_mod3_2))




#-----------------------------------------------------------------#
#                VALOR DE UN JUGADOR VS SU EDAD                   #
#-----------------------------------------------------------------#
mod4<-lm(VORP~Age)
plot(VORP~Age)
abline(mod4)
plot(mod4$residuals, pch=16, col=2)

abline(mod4);summary(mod4)
mod_gam5<-gam(VORP~s(Age,bs="cr"))
```

```
gam.check(mod_gam5)
windows()
par(mfrow=c(2,2))
plot(mod4)


resettest(mod4,power=2)
harvtest(mod4)
library(sm)
sm.regression(VORP,Age,model="linear")
res4<-rstandard(mod4)
shapiro.test(res4)
hmctest(mod4)
durbinWatsonTest(mod4)


r1<-residuals(mod4)
r2<-rstandard(mod4)
r3<-rstudent(mod4)
which(abs(r2)>1.96)
which(abs(r3)>1.96)
which(abs(r2)>qt(0.95,df=n-2))
which(abs(r3)>qt(0.95,df=n-2))
plot(VORP~Age)
indr<-which(abs(r2)>1.96)
points(VORP[indr]~Age[indr],col=4,pch=16)

summary(mod4)

vorp_outliers<-predict(mod4,type="response")
vorp_outliers[indr]
```