Un modelo de regresión lineal para caracterizar las victorias de un equipo en temporada regular de la NBA

Uriel Martínez, Guillaume Domenge, Sergio Arroyo

Mayo del 2020

Resumen

En este trabajo se propone un modelo de regresión lineal múltiple que caracteriza el porcentaje de victorias en temporada regular de los equipos de la *National Basketball Associaton* mediante las estadísticas generales tradicionales de los diversos equipos en la liga en los últimos 10 años.

1. Introducción

La *National Basketball Association* (NBA) es la liga profesional de baloncesto más popular y competitiva de del mundo. Anualmente, entre los meses de octubre y abril, cada uno de los 30 equipos afiliados a la liga disputan 82 juegos, 41 como locales y el resto como visitantes; en total, la temporada regular consta de 1230 partidos[1]. La alta cantidad de juegos por temporada que cada equipo disputa relativa a la cantidad de juegos que son usuales en otros deportes[2] y la captación de toda clase de datos y estadísticas alrededor de todas las facetas del juego, han facilitado una importante evolución en la manera en la que los equipos alrededor de la NBA toman decisiones dentro y fuera de la cancha, así como la producción de una cantidad considerable de artículos relacionados al tema producidos por profesionales del análisis de datos [5][6] . Sin ahondar demasiado en esto, en un sentido estadístico, se ha observado que la cantidad de triples intentados por partido y por temporada ha ido en ascenso cada año (figura 1). Gracias a esta revolución no es nada descabellado [7] el análisis de estadísticas concernientes a este aspecto del juego moderno pues es esencial para entenderlo globalmente.

En este marco y como ejercicio de las prácticas asociadas a la regresión lineal, es pertinente preguntarse cuáles son los factores que vuelven "bueno" a un equipo actual de baloncesto, más aún, ¿podemos encontrar un modelo para determinar la cantidad de juegos que un equipo ganará en la temporada regular basados en el conocimiento de algunas estadísticas y que este modelo no sea trivial, arrojando, de este modo, luz sobre cuáles son los aspectos importantes del juego? Intentaremos responder a esto mediante las estadísticas tradicionales generales [3] de los equipos de la NBA desde 2010 a la fecha [8].

[1] La actual crisis sanitaria interrumpió la temporada desde marzo, impidiendo este formato.

[2] Por ejemplo, las temporadas de la NFL y de *La Liga*, ligas comparables en popularidad y profesionalización, constan de 256 y 380 partidos respectivamente.

[3] Son tradicionales en contraposición a las estadísticas *avanzadas* diseñadas para representar mejor los

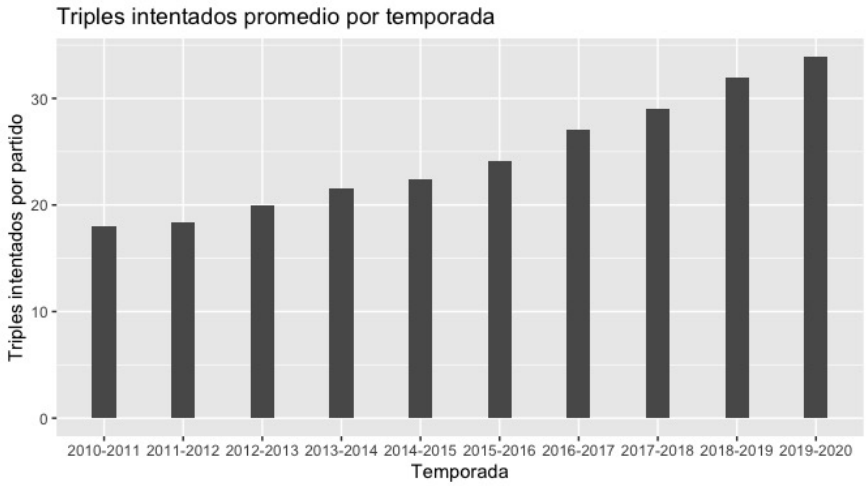

Figura 1: Los lanzamientos de triples han aumentado año con año.

2. Sobre el baloncesto y las estadísticas generales por equipo

La descripción de Wikipedia en español explica de manera bastante adecuada los aspectos más básicos del juego:

> El baloncesto [...] es un deporte de equipo, jugado entre dos conjuntos de cinco jugadores cada uno durante cuatro períodos [...] de doce minutos cada uno. El objetivo del equipo es anotar puntos introduciendo un balón por la canasta, un aro a 3,05 metros sobre la superficie de la pista de juego del que cuelga una red. La puntuación por cada canasta o cesta es de dos o tres puntos, dependiendo de la posición desde la que se efectúa el tiro a canasta, o de uno, si se trata de un tiro libre por una falta de un jugador contrario. El equipo ganador es el que obtiene el mayor número de puntos.[2]

Adicionalmente, los jugadores pueden hacerse del balón robando su posesión del equipo contrario o tras un rebote, acción de recuperar el balón tras un tiro fallado (ya sea a favor o en contra), y en muchas ocasiones las anotaciones se producen tras haber recibido la pelota de un compañero, es decir, una asistencia. Naturalmente, no todos los tiros que se hacen a la canasta son exitosos, así que se suele contabilizar el porcentaje de éxitos de los tiros en general (**FG %**), los triples y los tiros libres.

Los datos que utilizaremos en las siguientes páginas provienen de la página oficial de la **NBA** y pueden ser consultados aquí (figura 2). Entre las columnas cuya interpretación no es contextualmente clara a partir de la descripción básica del juego, podemos destacar a la última columna (+/-), la diferencia promedio entre los puntos totales anotados en un juego y los puntos recibidos, y **PFD**, la cantidad promedio de faltas que los jugadores de

aspectos relevantes del baloncesto. Estas pueden ser de gran utilidad para el estudio del basketball, pero en este trabajo nos decantamos por estadísticas más comunes y fácilmente interpretables[9].

	TEAM	GP	W	L	WIN%	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	BLKA	PF	PFD	+/-
1	Milwaukee Bucks	65	53	12	.815	48.2	118.6	43.5	91.2	47.7	13.7	38.6	35.6	17.8	24.0	74.2	9.5	42.2	51.7	25.9	14.9	7.4	6.0	4.6	19.2	21.3	11.3
2	Los Angeles Lakers	63	49	14	.778	48.2	114.3	42.9	88.6	48.5	11.2	31.4	35.5	17.3	23.7	73.0	10.6	35.5	46.1	25.9	15.1	8.6	6.8	3.7	20.6	21.4	7.4
3	Toronto Raptors	64	46	18	.719	48.3	113.0	40.6	88.5	45.8	13.8	37.0	37.1	18.1	22.6	80.0	9.7	35.5	45.2	25.4	14.4	8.8	4.9	5.3	21.5	20.0	6.5
4	LA Clippers	64	44	20	.688	48.2	116.2	41.6	89.7	46.4	12.2	33.2	36.6	20.8	26.2	79.2	11.0	37.0	48.0	23.8	14.8	7.1	5.0	4.9	22.0	22.8	6.5
5	Boston Celtics	64	43	21	.672	48.4	113.0	41.2	89.6	45.9	12.4	34.2	36.3	18.3	22.8	80.1	10.7	35.3	46.0	22.8	13.6	8.3	5.6	5.6	21.4	20.6	6.2
6	Denver Nuggets	65	43	22	.662	48.5	110.4	41.8	88.9	47.1	10.9	30.4	35.8	15.9	20.5	77.5	10.8	33.5	44.3	26.5	13.7	8.1	4.6	4.5	20.0	20.0	3.0

Figura 2: Las estadísticas tradicionales para algunos equipos esta temporada

determinado equipo reciben por juego. Gracias a que logramos recabar la información de las estadísticas generales de los últimos 10 años, nuestra base cuenta con 300 observaciones.

3. Análisis exploratorio de datos: variables relevantes

3.1. Reducción de la dimensión y colinealidad mediante scatterplots y la experiencia

Hay que recordar que la meta de este trabajo es saber qué es lo que hace que un equipo de la NBA sea ganador y que el modelo con el cual podemos responder esta pregunta no sea trivial. Evidentemente, el mejor equipo es aquel que gana más juegos, es decir, aquel cuyo porcentaje victorias (**WIN %**) es más alto, así que la información de las victorias (**W**) y las derrotas (**L**) se encuentra resumida perfectamente en **WIN %** y de ahora en adelante consideraremos esta como nuestra variable a explicar. De este modo, podemos hacer una consideración preliminar para reducir nuestro espacio de columnas.

Ya que deseamos que nuestro modelo no sea trivial o, dicho de otro modo, que no haya una variable que sea tan buena explicando el porcentaje de victorias que no vale ni siquiera la pena hacer un modelo, podemos eliminar a la columna (+/-). Es claro, intuitivamente hablando, que el equipo que tiene una relación favorable entre puntos recibidos y anotados sea un equipo ganador, entonces es igualmente claro que la relación entre el porcentaje de victorias y el +/- debe ser casi lineal. Para disuadir al escéptico de cualquier reclamo a esta afirmación, basta con observar la gráfica 3.

Podríamos establecer una narrativa similar a la arriba expuesta para ir eliminando columnas de nuestro análisis, sin embargo, esto sería demasiado tardado y poco interesante para el lector[4], a pesar de esto, podemos proseguir como se hizo en el párrafo pasado y eliminar variables que estén claramente relacionadas mediante una inspección sesuda de la figura 4 y consideraciones empíricas. Esto nos es útil por partida doble: en primer lugar, libera a nuestro modelo de variables explicativas innecesarias y, en segundo lugar, reduce la colinealidad. En general, podemos esperar un da un mejor modelo tras haber realizado esto.

[4]Sólo para que quede ilustrado, la cantidad de tiros libres que se anotan, los tiros libres que se intentan, y la cantidad de faltas que un equipo consigue están íntimamente relacionadas, al punto en que la información puede ser resumirse en sólo una de ellas.

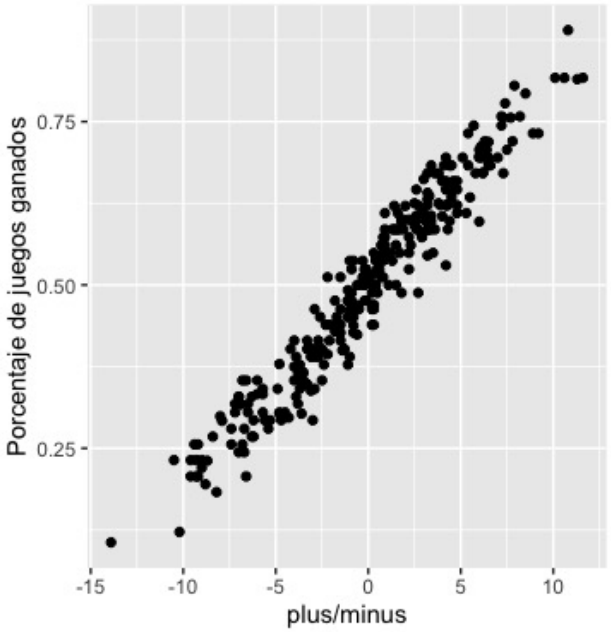

Figura 3: Plus/minus contra el porcentaje de victorias

Ahora consideremos a las correlaciones entre las variables que hemos conservado tras el análisis anterior. En la figura 5 podemos destacar que la correlación entre los puntos anotados y los triples intentados es muy alta y, por tanto, sería tentador colapsar una variable en otra, sin embargo, si consideramos la evolución reciente del juego[5] esto no sería aconsejable y nos perderíamos de una poderosa variable explicativa.

3.2. ¿La conferencia importa?

La NBA, así como otras ligas deportivas estadounidenses, se divide en un sistemas de conferencias y divisiones que determina en gran medida su calendario de juego; no es razonable que un equipo de Oregon juegue relativamente seguido contra un equipo de Florida. Por esto es pertinente preguntarse si la pertenencia a determinada conferencia, ya sea la Este o la Oeste, juega un papel determinante en el éxito de un equipo y, para no dar muchas vueltas al asunto, la respuesta es afirmativa, pero ¿qué tanto importa? La respuesta a esta pregunta es más complicada de lo que parece. En primera instancia y sin necesidad de una prueba estadística formal, al comparar las distribuciones del porcentaje de victorias por conferencia de los últimos 10 años podemos asegurar que en promedio los equipos de la conferencia Oeste son mejores que los de la conferencia Este (figura 6). Esto fundamenta la creencia popular de que, en general, los equipos de una costa son mejores que los de la otra.

[5]Nos referimos en particular a las últimas iteraciones de equipos como los Houston Rockets, Golden State Warriors y Portland Trail Blazers que han conseguido un gran éxito en la temporada regular gracias a sus estrategias ofensivas donde los triples son la herramienta principal.

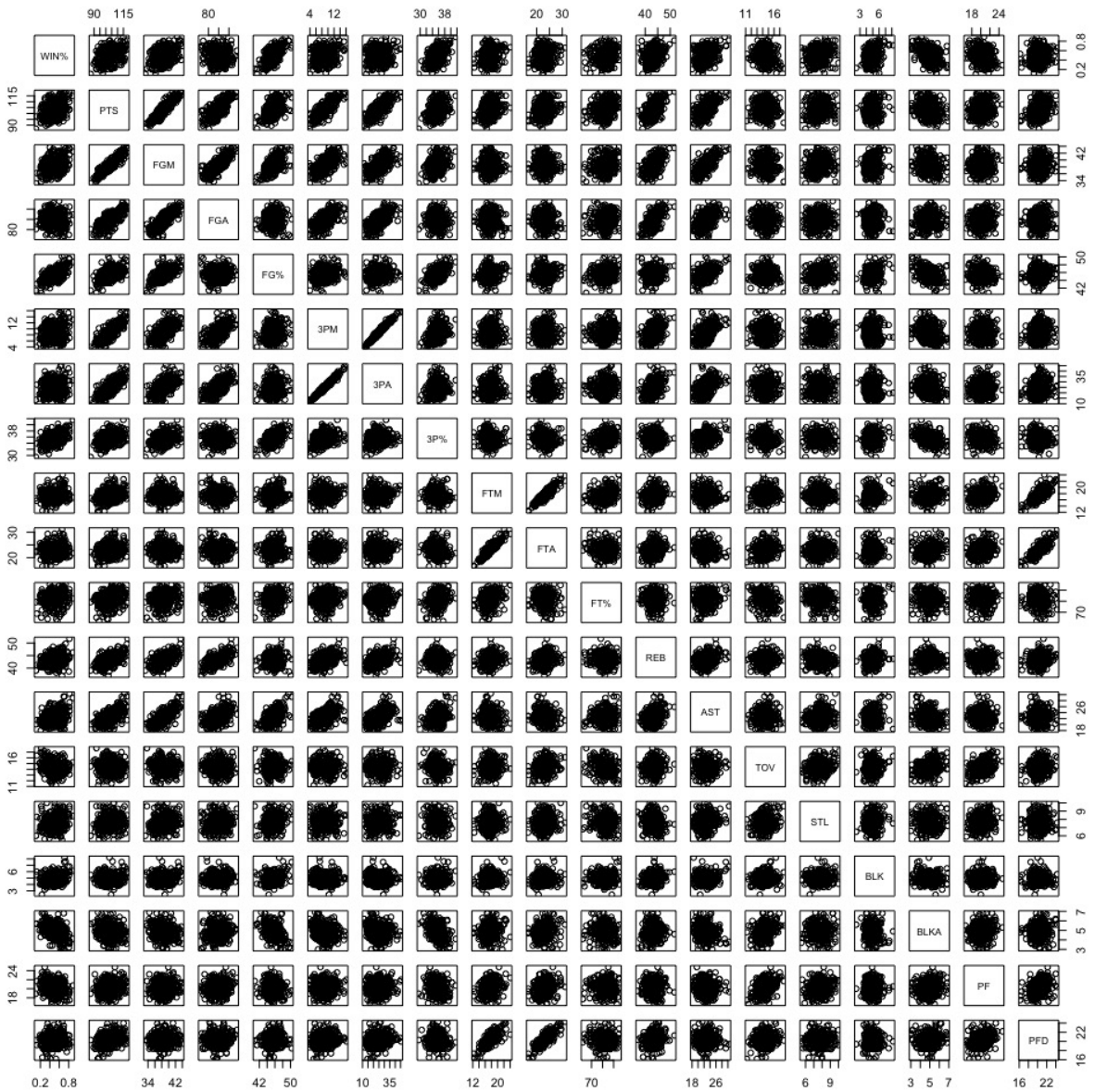

Figura 4: Scatterplots entre las distintas columnas de nuestros datos.

En segunda instancia y en función de nuestros fines, ¿agregar como columna categórica para nuestro modelo la pertenencia a determinada conferencia aumente significativamente el poder explicativo de nuestro modelo lineal? Una respuesta *ex post* nuestro modelaje nos dice que no. En este caso observamos que el tamaño del coeficiente asociado a la conferencia fue despreciable y que el tamaño de la R^2 marginalmente mayor[6]. Así que para responder a nuestra pregunta de trabajo proponemos el modelo que se presenta a continuación.

[6]Estos cálculos se pueden encontrar en el documento de R anexo, ya que por cuestiones de exposición decidimos excluirlos.

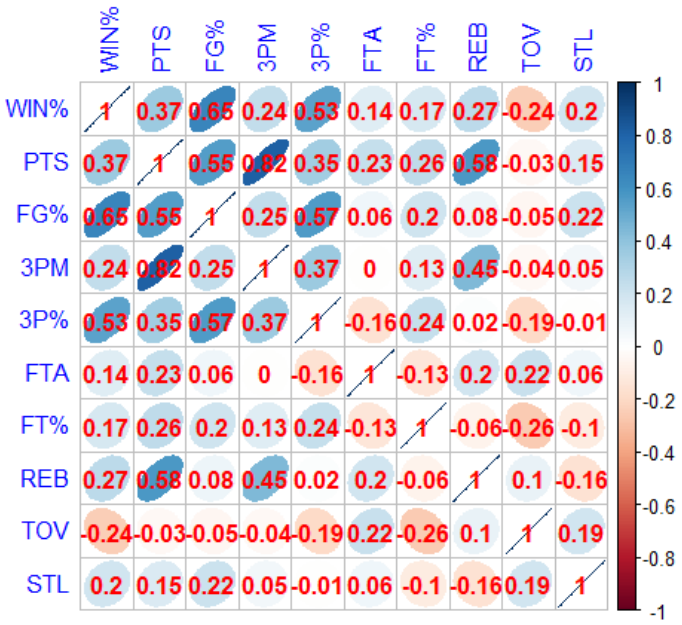

Figura 5: Correlaciones entre distintas variables a utilizar en el modelo

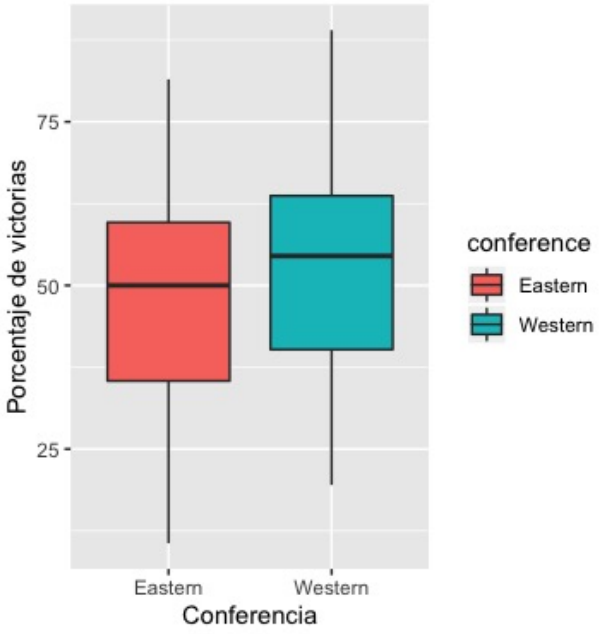

Figura 6: Boxplots de la distribución del porcentaje de victorias de los equipos de la NBA por conferencia

Acrónimo	Significado
FG %	Porcentage de tiros de campo anotados
PTS	Número de puntos por partida
3PM	Número de tiros de 3 puntos realizados
3P %	Porcentaje de tiros de 3 puntos anotados
FTA	Número tiros libres intentados
FT %	Porcentaje de tiros libres anotados
REB	Número de rebotes
TOV	Número de turnovers
STL	Número de balones robados

Cuadro 1: Glosario

4. El modelo

Primero recordemos la correcta pronunciación del modelo de regresión que utilizaremos para nuestro análisis:

Modelo: Regresión Lineal Múltiple

Supongamos que $\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ son vectores aleatorios, con n entradas, tales que satifacen

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \cdots + \beta_n\mathbf{X}_m + \varepsilon$$

donde ε es un vector aleatorio cuyas entradas $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ con σ^2 constante. En la notación de Wilkinson & Rogers, lo anterior se resume en:

$$\mathbf{Y} \sim \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_m$$

Tras haber realizado un exhaustivo análisis de las variables relevantes para determinar qué hace a un equipo actual de la NBA exitoso, llegamos a la realización del siguiente modelo lineal:

$$\mathbf{WIN\,\%} \sim \mathbf{FG\,\%} + \mathbf{PTS} + \mathbf{3PM} + \mathbf{3P\,\%} + \mathbf{FTA} + \mathbf{FT\,\%} + \mathbf{REB} + \mathbf{TOV} + \mathbf{STL}. \quad (1)$$

Es decir, buscamos explicar el porcentaje de victorias de los equipos de la NBA por medio de sus estadísticas particulares promedio por partido de efectividad a la hora de tirar, puntos anotados, triples intentados, efectividad al tirar triples, tiros libres intentados, efectividad al tirar tiros libres, rebotes, balones regalados y robos. Sin mucho más que decir a este respecto, pasemos a ver los resultados de la regresión por medio de mínimos cuadrados ordinarios[7].

El coeficiente de determinación R^2 obtenido nos indica que, aproximadamente, el 82,58 % de la variación de la variable respuesta **WIN %** está explicada por las variables explicativas descritas en el cuadro 1. Otro valor de importancia arrojado por el comando `lm` están los

[7]Los resultados numéricos exactos, así como el código implementado en el software estadístico `RStudio`, pueden encontrarse en el Anexo al final de este documento.

	Estimado	Error estándar	Valor de `t`	$Pr(>\|t\|)$
(Intercept)	-4.965759	0.205601	-24.152	< 2e-16
PTS	-0.040811	0.002217	-18.412	< 2e-16
FG %	0.106928	0.004805	22.256	< 2e-16
3PM	0.056933	0.004088	13.927	< 2e-16
3P %	0.008635	0.003019	2.860	0.00454
FTA	0.028275	0.002013	14.050	< 2e-16
FT %	0.015222	0.001547	9.837	< 2e-16
REB	0.055987	0.002741	20.429	< 2e-16
TOV	-0.052573	0.003873	-13.573	< 2e-16
STL	0.068824	0.005171	13.309	< 2e-16

Cuadro 2: Resumen de la regresión sobre nuestros datos

asociados a las pruebas de hipótesis de los parámetros de la regresión: la estadística de prueba `F` y el valor `p` asociado a ella. En particular, nos indica que para la prueba de significancia de la regresión:

$$\begin{cases} H_0: & \beta_i = 0 \text{ para todas las variables explicativas} \\ H_1: & \beta_i \neq 0 \text{ para alguna variables explicativa.} \end{cases}$$

el valor `p` asociado es extremadamente cercana a cero, lo cual nos indica que significativamente al menos una de las variables explicativas es distinta de cero.

Meditemos un poco sobre los resultados de la regresión. El modelo de regresión lineal tiene muchos coeficientes de regresión y, por tanto, muchas hipótesis sobre ellos son posibles, sin embargo, en el cuadro 2 se arroja poderosa información a favor de nuestra selección de variables. Por ejemplo, consideremos la prueba sobre β_1, coeficiente para el regresor **FG %** (efectividad en los tiros):

$$\begin{cases} H_0 & : \beta_1 = 0, \text{ el resto de las } \beta_i \text{ arbitrarias} \\ H_1 & : \beta_1 \neq 0, \text{ el resto de los coeficientes arbitrarios.} \end{cases}$$

Esta prueba podemos interpretarla como el efecto de agregar el regresor **FG %** al modelo[8]. La tabla nos comunica que para este coeficiente contamos con una estadística `t` con 46 grados de libertad y que para nuestros datos $t = 21{,}721$. Más aún, la probabilidad de encontrar una observación que sea mayor que $|t|$ es extremadamente baja, es decir, el valor de $Pr(>|t|)$ es prácticamente cero. En consecuencia, tenemos evidencia significativa para rechazar que, dado el resto de los predictores fijos, β_1 es idénticamente cero, es decir, la efectividad con la que un equipo tira a la canasta tiene un efecto sobre la cantidad de partidos que ganaran en la temporada regular de la NBA.

[8]Es decir, si consideramos el modelo **WIN % ~ PTS + 3PM + 3P % + FTA + FT % + REB + TOV + STL**., la prueba mide el efecto de agregar otra variable explicativa.

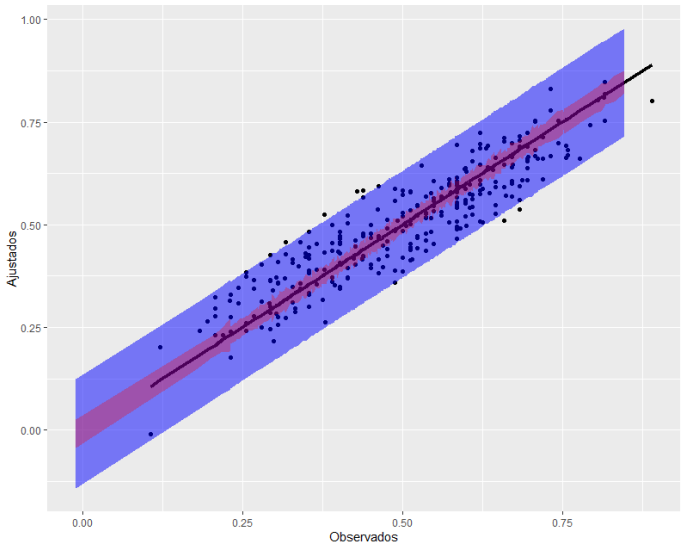

Figura 7: Intervalos de confianza y de predicción al 95 %.

Este argumento se puede hacer para cada uno de los coeficientes asociados a nuestras variables explicativas, así que, sin considerar como se relacionan unas entre otras, cada una de las variables juega un papel determinante en ganar los juegos. A pesar de estas buenas noticias, debemos ser cautos respecto del modelo. Debemos notar que el signo del coeficiente asociado a los puntos es negativo, lo cual es completamente contra intuitivo, ya que uno esperaría que, por lo general, si un equipo anota muchos puntos, entonces el equipo gane. Esto no es el caso, pero tampoco la situación contraria: un equipo que anota pocos puntos no va a ganar mucho. ¿Es esta una deficiencia insalvable del modelo? Primero validemos de los supuestos del modelo.

Por último, consideraremos el poder que tiene nuestro modelo para predecir un porcentaje de victorias basado en nuevas observaciones. En la gráfica 7 podemos observar los datos observados contra los ajustados por el modelo, donde se puede determinar el error de predicción midiendo la distancia en línea recta desde el punto a la línea recta colocada en medio. De igual forma, la franja roja es el intervalo de confianza para la predicción hecha por modelo. Es decir, para cada punto fijo y_F se tiene un intervalo de confianza para el estimador puntual $\hat{\mathbf{E}}[Y_F] = X_F'\underline{\hat{\beta}}$. Por otro lado, la franja azul es el intervalo de confianza para el predictor de y_F, es decir, para $\hat{Y_F} = X_F'\underline{\hat{\beta}}$. A pesar de contar con las mismas expresiones para estimar y predecir, el intervalo de confianza para predecir es más ancho que el correspondiente al del estimador.

5. Análisis de supuestos y de varianzas

Para corroborar que nuestro modelo se ajusta a los supuestos de la regresión lineal analizamos la distribución de los residuales arrojados por la regresión lineal. La gráfica observada

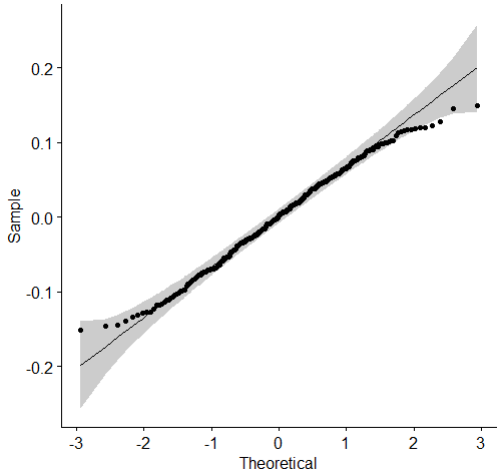

Figura 8: Gráfico Cuantil-Cuantil de los residuales del modelo contra los cuantiles teóricos de una distribución normal

en la figura 8 nos permite comparar la distribución de los residuos con la distribución normal teórica. Por lo tanto, si los residuos tienen una distribución normal deberíamos observar que siguen aproximadamente la línea recta diagonal en el gráfico. Es decir, podemos validar que los residuales se distribuyen de manera normal. También podemos podemos observar que los datos presentan *colas ligeramente pesadas*, sin embargo, no nos concentraremos en estos detalles y seguiremos con el análisis de varianzas.

Otro aspecto relevante de la validación de supuesto del modelo se basa en la relación de nuestros regresores con los residuales. Para ello graficamos los valores residuales contra los valores predichos por el modelo mediante un diagrama de dispersión. De esta manera, podemos observar en las gráficas de la figura 9 que los datos presentan un patrón de residuos al azar lo cual indica que no hay sesgos en los residuos (tendencias) ni una dispersión (varianza) no constante ni valores que desvíen el comportamiento observado (outliers). Por lo tanto, deducimos que el modelo que hemos presentado en la sección 4 resulta ser válido.

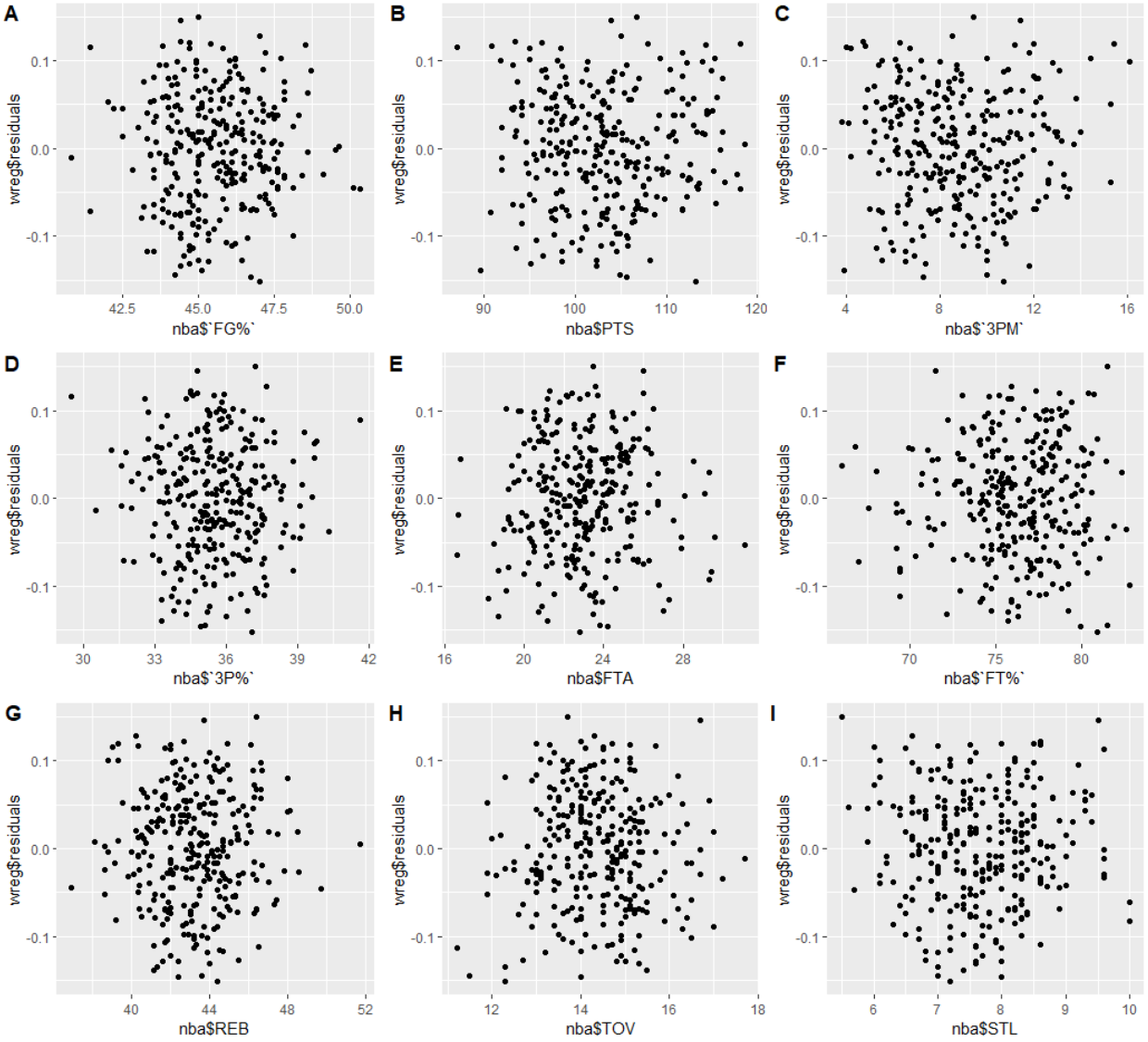

Figura 9: Relación entre los residuos y cada una de las variables explicativas.

	GL	SCR	SCR/GL	Valor de F	$Pr(>F)$	η^2
FG %	1	3.0178	3.0178	710.478	< 2e-16	0.2917
PTS	1	0.0006	0.0006	0.153	0.69619	0.1978
3PM	1	0.1321	0.1321	31.109	5.59e-08	0.1115
3P %	1	0.1545	0.1545	36.383	4.93e-09	0.0369
FTA	1	0.2399	0.2399	56.490	7.07e-13	0.1160
FT %	1	0.0402	0.0402	9.453	0.00231	0.0566
REB	1	1.0010	1.0010	235.662	< 2e-16	0.2475
TOV	1	0.4997	0.4997	117.654	< 2e-16	0.1103
STL	1	0.7523	0.7523	177.116	< 2e-16	0.1027
Residuals	290	1.2318	0.0042			

Cuadro 3: ANOVA y η^2 de los datos.

La tabla ANOVA en el **Cuadro 3** nos muestra que todas nuestras variables menos **PTS** presentan un valor de F grande y un valor p significativamente pequeño, es decir, para cada una de estas variables podemos rechazar nuestra hipótesis nula, $H_0 : \mu_{ij} = \bar{\mu}_i$. En otras palabras para las variables **FG %**, **3PM**, **3P %**, **FTA**, **FT %**, **REB**, **TOV**, **STL** podemos ver que nuestras muestras presentan diferencias en sus promedios estadísticamente significativas. Por otro lado, podemos notar que en el caso de la variable **PTS** se observa un valor p mayor de casi 0,7, lo cual, junto con una suma cuadrada de la regresión de 0,0006, nos muestra que la variable **PTS** contribuye poco de la suma total de cuadrados. Podríamos concluir de esta ANOVA que los puntos promedios por partida no contribuyen al **WIN %**, pero esto es contra intuitivo puesto que el juego se gana anotando canastas. Es por esto que agregamos a la tabla ANOVA una columna η^2 que mide la proporción de la variabilidad de la variable dependiente que se puede explicar gracias al i-ésimo regresor. Gracias a los valores en la columna η^2 podemos ver que la variable **PTS** contribuye en realidad al 19,8 % del a variabilidad de **WIN %**. El comportamiento de la variable **PTS** es problemático, y esto se puede deber a que su varianza está compuesta de la varianza natural entre los equipos en un dado año y una varianza artificial que se produce cuando juntamos los datos de varios años, que como ya mostramos anteriormente tienen medias distintas dado por un crecimiento en el promedio de puntos año tras año. Finalmente, podemos ver en la columna η^2 que las variables de nuestro modelo tienen un efecto mediano o grande sobre la variabilidad de **WIN %** según la regla de oro de Cohen[3], la unica variable con un efecto medio chico es **3P %** con $\eta^2 = 0{,}0369$, pero está más cerca de ser un efecto mediano que pequeño.

6. Conclusión

A pesar de que la validación de supuestos demostró que la elección de nuestras variables junto con la propuesta del modelo eran válidos, obtuvimos resultados inesperados. En primer lugar, el coeficiente asociado a la variable **PTS** se estimó con un valor negativo, lo cual podría ir contra toda intuición; es decir, se espera que uno de los "requisitos" para ser un ganador debe conseguir varios puntos. Sin embargo, una de las posibles explicaciones para este fenómeno es que encestar muchos puntos no asegura la victoria. Por ejemplo, sugonga que el equipo **A** encesta muchas canastas acumulando muchos puntos pero el equipo contrario, el equipo **B**, encesta tan solo 1 más que el **A**. Podemos intuir que la variable **PTS** contrarresta el efecto provocado por las variables **FT %**, **3PM** y **FTA**, pues éstas presentan efectos positivos al acumular o encestar puntos.

Otro manera de analizar el caso de la variable **PTS** es realizar una regresión lineal en función de las demás variables explicativas. Los resultados arrojados por el modelo indican que el 92,93 % de la variabilidad de los puntos anotados está descrita por el resto de las variables explicativas. De este modo, otra posible conclusión es que los efectos cruzados de las variables explicativas repercutan en la dinámica con **WIN %**. Sin embargo, para sustentar estas conclusiones falta validar los supuestos de regresión lineal. Es decir, errores normales con media cero y varianza constante.

Finalmente, el modelo propuesto en la sección 4 nos ayuda a caracterizar cuáles son las variables con mayor peso para ser ganador durante la temporada. Destacamos las variables **FG %**, **3PM** y **REB** cuya influencia en el porcentage de victorias es más alta que la del resto de las variables explicativas. Es decir, para que un equipo gané deberá enfocar sus esfuerzos en mantener una tasa alta de canastas; también tendrá que capacitar a sus jugadores en tiros efectivos de 3 puntos; por último, y en caso de que esto suceda, mantener una tasa elevada en la recuperación de tiros fallidos a la canasta (ofensivos y defensivos). No obstante, encontramos fallas basadas en conocimiento empírico y contra la intuición sobre la información recolectada durante varias temporadas. Por ejemplo, los datos se muestran de forma agregada por temporada, en primer nivel, y por partido jugado, en segundo nivel. Además, después de realizar este análisis, consideramos que un acercamiento adecuado para obtener resultados más precisos es obtener los mismos datos pero por partido para un equipo en especifico durante varias temporadas. De esta manera podremos validar los resultados descritos en este documento y dar instrucciones específicas para generar una estrategia fundamentada en el seguimiento particular de un equipo.

7. ANEXO: Código en RStudio

Regresión lineal mostrada en la sección 4.

```
nba_data <- read.csv(file = "nba_data1.csv", head = TRUE)
fit <- lm(WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB + TOV + STL), data = nba_data)
summary(fit)

##
## Call:
## lm(formula = WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB +
##     TOV + STL), data = nba_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151442 -0.044527  0.002307  0.046980  0.149996
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.965759   0.205601 -24.152  < 2e-16 ***
## FG.          0.106928   0.004805  22.256  < 2e-16 ***
## PTS         -0.040811   0.002217 -18.412  < 2e-16 ***
## X3PM         0.056933   0.004088  13.927  < 2e-16 ***
## X3P.         0.008635   0.003019   2.860  0.00454 **
## FTA          0.028275   0.002013  14.050  < 2e-16 ***
## FT.          0.015222   0.001547   9.837  < 2e-16 ***
## REB          0.055987   0.002741  20.429  < 2e-16 ***
## TOV         -0.052573   0.003873 -13.573  < 2e-16 ***
## STL          0.068824   0.005171  13.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06517 on 290 degrees of freedom
## Multiple R-squared:  0.8258,Adjusted R-squared:  0.8204
## F-statistic: 152.7 on 9 and 290 DF,  p-value: < 2.2e-16
```

Intervalos de confianza para los estimadores por mínimos cuadrados de los coeficientes.

```
confint(fit, level = 0.95)

##                     2.5 %      97.5 %
## (Intercept) -5.370418943 -4.56110003
## FG.          0.097471925  0.11638430
## PTS         -0.045173444 -0.03644823
## X3PM         0.048887035  0.06497823
## X3P.         0.002692974  0.01457623
## FTA          0.024314241  0.03223618
## FT.          0.012176559  0.01826749
## REB          0.050593108  0.06138106
## TOV         -0.060196814 -0.04494977
## STL          0.058645710  0.07900230
```

Resultados de la tabla ANOVA.

```
summary(aov(fit))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## FG.           1 3.0178  3.0178 710.478  < 2e-16 ***
## PTS           1 0.0006  0.0006   0.153  0.69619
## X3PM          1 0.1321  0.1321  31.109 5.59e-08 ***
## X3P.          1 0.1545  0.1545  36.383 4.93e-09 ***
## FTA           1 0.2399  0.2399  56.490 7.07e-13 ***
## FT.           1 0.0402  0.0402   9.453  0.00231 **
## REB           1 1.0010  1.0010 235.662  < 2e-16 ***
## TOV           1 0.4997  0.4997 117.654  < 2e-16 ***
## STL           1 0.7523  0.7523 177.116  < 2e-16 ***
## Residuals   290 1.2318  0.0042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regresión lineal considerando la división por conferencia.

```
fit2 <- lm(WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB + TOV + STL
    + conference_Eastern), data=nba_data)

summary(fit2)
```

```
##
## Call:
## lm(formula = WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB +
##     TOV + STL + conference_Eastern), data = nba_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.149717 -0.043177  0.002404  0.045838  0.154786
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.043661   0.212843 -23.697  < 2e-16 ***
## FG.                 0.107197   0.004801  22.329  < 2e-16 ***
## PTS                -0.040542   0.002222 -18.249  < 2e-16 ***
## X3PM                0.056290   0.004108  13.703  < 2e-16 ***
## X3P.                0.008826   0.003017   2.925  0.00371 **
## FTA                 0.028629   0.002026  14.134  < 2e-16 ***
## FT.                 0.015363   0.001548   9.923  < 2e-16 ***
## REB                 0.056124   0.002738  20.498  < 2e-16 ***
## TOV                -0.052424   0.003869 -13.550  < 2e-16 ***
## STL                 0.069360   0.005178  13.396  < 2e-16 ***
## conference_Eastern  0.011232   0.008110   1.385  0.16714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06507 on 289 degrees of freedom
## Multiple R-squared:  0.8269,Adjusted R-squared:  0.8209
## F-statistic: 138.1 on 10 and 289 DF,  p-value: < 2.2e-16
```

Prueba de significancia de la variable categórica.[9]

```
anova(fit, fit2)

## Analysis of Variance Table
##
## Model 1: WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB + TOV + STL)
## Model 2: WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB + TOV + STL +
##     conference_Eastern)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    290 1.2318
## 2    289 1.2237  1 0.0081213 1.9181 0.1671
```

Regresión lineal para explicar **PTS** a partir de las demás variables.

```
fit3 <- lm(PTS ~ (FG. + X3PM + X3P. + FTA + FT. + REB + TOV + STL),
    data=nba_data)
summary(fit3)

##
## Call:
## lm(formula = PTS ~ (FG. + X3PM + X3P. + FTA + FT. + REB + TOV +
##     STL), data = nba_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5214 -1.0699 -0.0744  1.1504  5.6060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.53069    4.91895  -8.036 2.33e-14 ***
## FG.           1.60392    0.08547  18.766  < 2e-16 ***
## X3PM          1.62700    0.05090  31.963  < 2e-16 ***
## X3P.         -0.44935    0.07537  -5.962 7.20e-09 ***
## FTA           0.45613    0.04602   9.911  < 2e-16 ***
## FT.           0.34064    0.03572   9.536  < 2e-16 ***
## REB           0.79516    0.05550  14.327  < 2e-16 ***
## TOV          -0.34517    0.10042  -3.437 0.000673 ***
## STL           0.72545    0.12999   5.581 5.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.724 on 291 degrees of freedom
## Multiple R-squared:  0.9293,Adjusted R-squared:  0.9273
## F-statistic: 477.9 on 8 and 291 DF,  p-value: < 2.2e-16
```

[9]Esta prueba nos indica que la presencia de la conferencia como variable categórica no es significativa para nuestro modelo. Es decir, la prueba `ANOVA` sugiere que no incluyamos la variable categórica que indique la conferencia en nuestro análisis.

Regresión lineal sin β_0.

```
fit4 <- lm(WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB + TOV + STL
    -1), data=nba_data)

summary(fit4)

##
## Call:
## lm(formula = WIN. ~ (FG. + PTS + X3PM + X3P. + FTA + FT. + REB +
##     TOV + STL - 1), data = nba_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.311041 -0.082836 -0.001706  0.079918  0.310346
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## FG.   0.052019   0.007333   7.094 9.93e-12 ***
## PTS  -0.017995   0.003474  -5.180 4.15e-07 ***
## X3PM  0.035722   0.006916   5.165 4.47e-07 ***
## X3P.  0.004446   0.005221   0.852 0.395160
## FTA   0.014540   0.003344   4.348 1.90e-05 ***
## FT.  -0.008831   0.002052  -4.305 2.29e-05 ***
## REB   0.015180   0.003738   4.061 6.29e-05 ***
## TOV  -0.072306   0.006559 -11.024  < 2e-16 ***
## STL   0.031153   0.008542   3.647 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1129 on 291 degrees of freedom
## Multiple R-squared:  0.9548,Adjusted R-squared:  0.9534
## F-statistic: 683.2 on 9 and 291 DF,  p-value: < 2.2e-16
```

Regresión lineal con parámetro cuadrático en **PTS**.[10]

```
fit5 <- lm(WIN. ~ (FG. + PTS + I(PTS^2) + X3PM + FTA + FT. + REB + TOV + STL),
data=nba_data)
summary(fit5)

##
## Call:
## lm(formula = WIN. ~ (FG. + PTS + I(PTS^2) + X3PM + FTA + FT. +
##     REB + TOV + STL), data = nba_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.185260 -0.048905  0.002865  0.046403  0.172599
```

[10]De acuerdo con los resultados arrojados consideramos que este modelo logra resolver el problema del coeficiente negativo [4], pues de esta manera se disminuye el efecto negativo de los puntos y se compensa con el coeficiente de su versión cuadrática. A simple vista, resulta ser un modelo prometedor debido a que el valor p de la prueba t asociada a `I(PTS^2)` indica que es significativo para el modelo.

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.885e+00  8.792e-01  -3.282  0.00116 **
## FG.          1.157e-01  3.875e-03  29.867  < 2e-16 ***
## PTS         -8.285e-02  1.681e-02  -4.927  1.4e-06 ***
## I(PTS^2)     1.914e-04  7.995e-05   2.395  0.01727 *
## X3PM         6.206e-02  3.691e-03  16.813  < 2e-16 ***
## FTA          2.886e-02  2.028e-03  14.233  < 2e-16 ***
## FT.          1.611e-02  1.524e-03  10.569  < 2e-16 ***
## REB          5.661e-02  2.740e-03  20.657  < 2e-16 ***
## TOV         -5.536e-02  3.868e-03 -14.311  < 2e-16 ***
## STL          6.947e-02  5.219e-03  13.312  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06544 on 290 degrees of freedom
## Multiple R-squared:  0.8243,Adjusted R-squared:  0.8189
## F-statistic: 151.2 on 9 and 290 DF,  p-value: < 2.2e-16
```

Referencias

[1] Alexander M. Petersen and Orion Penner. *Renormalizing individual performance metrics for cultural heritage management of sports records. https://arxiv.org/pdf/2004.08428.pdf.*

[2] *Baloncesto*. Wikipedia. *https://es.wikipedia.org/wiki/Baloncesto*

[3] *Eta-squared*. Wikiversity. *https://en.wikiversity.org/wiki/Eta-squared*

[4] Peter Dalgaard. *Introductory Statistics with R*. (2 ed). Springer: Statistics and Computing Series. **ISBN: 978-0-387-79053-4**

[5] Samuel Henry. *Improving upon NBA point-differential rankings. https://arxiv.org/pdf/1912.01574.pdf.*

[6] Samuel Henry. *Time-based analysis of the NBA hot hand fallacy. https://arxiv.org/pdf/1912.07442.pdf.*

[7] Shane Young. *The NBA's 3-Point Revolution Continues To Take Over*. Forbes Magazine.

[8] *Teams general traditional statistics*. NBA. *stats.nba.com/teams/traditional/*

[9] *Teams general advanced statistics*. NBA. *stats.nba.com/teams/advanced/*