Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Trabajo de Especialización

Impacto de los jugadores de la NBA según la métrica Plus Minus Ajustado

Autor: Lic. Franco Betteo

Supervisores:
Dr. Marcelo Soria
Dr. Ricardo Maronna

Índice

Resumen

No hay una manera única e inequívoca de comparar rendimientos de jugadores en los distintos deportes y eso da lugar a discusiones sin fin. Este trabajo propone un ranking de jugadores de la NBA para la temporada 2017-2018 basado en una regresión Ridge controlando por los otros jugadores en cancha. Se analizan los resultados de este enfoque y se construye un ranking de equipos basado en ellos. Posteriormente se testea el ordenamiento de equipos en las fases definitorias de la temporada.

Introducción

No hay una manera única e inequívoca de comparar rendimientos de jugadores en los distintos deportes y eso da lugar a discusiones sin fin. Más díficil aún en los deportes en equipo donde hay roles diferentes y contribuciones de distinta índole. A raíz de esto se han ido desarrollando métricas que intentan resumir el aporte al equipo de manera integral para hacer comparables los jugadores. En básquetbol una de las medidas más conocidas de este tipo es el "plus-minus" (originalmente implementado en el hockey sobre hielo), que calcula la diferencia de puntos de un equipo mientras cada jugador estuvo en cancha. Es decir que valores positivos (negativos) revelan que durante un partido el equipo hizo más (menos) puntos de los que recibió mientras el jugador estuvo en cancha. Es una métrica sencilla de calcular y resume el aspecto más importante de un partido de manera general para cada jugador. Es fácil de interpretar pero no está exenta de problemas. La idea de este trabajo es aplicar un método más robusto, "Plus-Minus ajustado", basado en los aportes de Justin Jacobs[1] y Joseph Sill[2], donde el principal agregado es controlar por los otros jugadores en cancha. El objetivo es actualizar la métrica para la temporada 2017-2018 y generar un ranking de jugadores. Posteriormente generar un ranking de equipos, basado en los rankings individuales de jugadores, y comparar contra los resultados de las rondas definitorias del certamen cuyos datos no son utilizados para calcular la métrica.

Datos

Para poder calcular la métrica "plus minus ajustado" necesitamos tener, para cada momento del partido, los jugadores que hay en cancha y el resultado ya que el objetivo es ver la performance del equipo en presencia y ausencia de cada jugador de la liga. A tales fines se decidió utilizar la información provista por la API de MySportsFeed[3]. En ella podemos encontrar datos a un nivel suficientemente granular. En particular, para cada partido tenemos información jugada a jugada marcada por ciertas situaciones particulares, entre ellas tiros al aro y si se convirtieron puntos o no, rebotes, faltas y sustituciones. Con la primera y la última de estas características podemos recolectar los datos necesarios para generar nuestro dataset. Al estar todas las sustituciones y tener la alineación inicial de cada equipo podemos obtener todos los segmentos del partido

donde hubo distintas combinaciones de jugadores en cancha (tanto compañeros como rivales) para cada jugador (denominados “stints” en la literatura). A su vez, al tener los puntos anotados podemos obtener el diferencial de puntos para cada uno de estos segmentos.

A los fines de utilizar toda esta información para entrenar el modelo de “plus minus ajustado” necesitamos armar un dataset con el formato siguiente:

- Cada observación es un segmento de un partido donde se mantuvo constante la alineación de ambos equipos.
- Las variables independientes son cada uno de los jugadores de la liga, con valor de 1 si estaban en cancha en ese segmento para el equipo local y -1 si estaban en cancha siendo del equipo visitante.
- La variable dependiente es el diferencial de puntos del equipo local. Valores positivos (negativos) es que anotó más puntos el equipo local (visitante).

Esta tabla esta conformada con todos los equipos de la liga y para cada partido de la temporada, de manera que para apendizar la información de cada encuentro hay que tener las variables independientes de cada equipo, es decir, todos sus jugadores. Dado que cada partido solo involucra a dos equipos, la gran mayoría de las columnas tendrán valor de 0 en cada observación, siendo el dataset una matriz dispersa (sparse matrix). Se toman valores de 1 y -1 según la localía para que el signo quede acorde a la medición del diferencial de puntos y por ende el signo de los coeficientes sea siempre positivo para aportes beneficiosos a un equipo y negativos para aportes perjudiciales. Ver Tabla 1

El dataset conformado cuenta con información de la temporada regular de la NBA 2017/2018, cuyo inicio es el 17 de Octubre de 2017 y finaliza el 11 de Abril de 2018. En este período se disputaron 1230 partidos, resultado de un calendario que cuenta con 82 fechas donde participan 30 equipos en cada una. Ver Cuadro 6. Dividiendo los partidos en stints obtenemos 35551 observaciones. A lo largo de estos 1230 partidos hubo incluidos oficialmente en las alineaciones 540 jugadores.

Tanto para la consulta de la API como para el manejo y procesamiento de los datos se utilizó el lenguaje R [4]

Cuadro 1: Primeras 20 observaciones del dataset (solo un subset de 10 jugadores (variables independientes

posesiones	dif_cada_100_posesiones	9082	9083	9084	9086	9087	9088	9089	9091
19	-5.263158	-1	0	0	0	0	0	0	0
11	9.090909	-1	0	0	0	0	0	0	0
8	100.000000	0	0	0	0	0	0	0	0
9	44.444444	0	0	0	0	0	0	0	0
8	0.000000	-1	0	0	0	0	0	0	0
7	57.142857	0	0	0	0	0	0	0	0
14	-14.285714	0	0	0	0	0	0	0	0
34	-26.470588	-1	0	0	0	0	0	0	0
9	-33.333333	0	0	0	0	0	0	0	0
11	-45.454546	-1	0	0	0	0	0	1	0
6	-16.666667	0	0	0	0	0	0	0	0
10	-10.000000	0	0	0	0	0	0	0	0
16	37.500000	-1	0	0	0	0	0	0	0
20	40.000000	0	0	0	0	0	0	0	0
6	83.333333	0	0	0	0	0	0	0	0
7	0.000000	0	0	0	0	0	0	0	0
21	47.619048	0	0	0	0	0	0	0	0
8	50.000000	0	0	0	0	0	0	0	0
14	-14.285714	0	0	0	0	0	0	0	0
8	-25.000000	0	0	0	0	0	0	0	0

Modelo

La metodología para calcular la métrica "plus minus ajustado" implica correr una regresión Ridge (regularización con *l2*) [5] donde la variable dependiente es el diferencial de puntos por segmento y las variables independientes son los jugadores en cancha. La idea de fondo es calcular el aporte de cada jugador al diferencial, controlando por sus compañeros y por los adversarios. Se intenta eliminar el factor que sobreestima el aporte de algún jugador solo por el hecho de compartir tiempo en cancha con compañeros de primer nivel, que son los que generan realmente los diferenciales positivos. De la misma manera se intenta no sobreestimar a los jugadores que anotan muchos puntos contra equipos de baja performance o que solo juegan en los minutos llamados "basura" que corresponden a los minutos finales de un partido cuando ya está todo definido y suelen haber jugadores de menor nivel. Dado que es una regresión lineal los coeficientes pueden interpretarse como un proxy del aporte neto de cada jugador, ya descontados los aportes del resto. Dada la codificación de variables - positiva para el local y negativa para visitantes - todo signo positivo de un coeficiente es aporte real en puntos y negativo es tendencia a recibir más puntos que los que se convierten con el jugador en cancha.
Los equipos suelen tener una plantilla de alrededor de 10 jugadores activos y los que más minutos disputan suelen ser menos aún por lo tanto es de esperar que haya una gran correlación entre los jugadores de cada equipo. Este es el principal motivo por el que se decide ir por una regresión regularizada y no una regresión lineal multivariada clásica.
Ridge lo que hace es reducir la varianza de los coeficientes estimados incluyendo una penalización *l2* a la función de pérdida que implica reducir mínimos cuadrados sumado a la diferencia al cuadrado de los coeficientes respecto de 0. Esto último ponderado por un parámetro λ a definir. A mayor λ la penalización es mayor y los coeficientes tienden a valores cercanos a 0. El procedimiento reduce la varianza de los coeficientes a costa de introducir un sesgo en su estimación, pero que de tener éxito, el tradeoff es tal que las predicciones son más certeras a pesar de no ser insesgado.

Formalización

Formalmente el modelo especificado es:

$$Y = \sum_{1}^{n} \beta_i X_i + \epsilon$$

Donde Y es el diferencial de puntos visto desde el equipo local, las X son las variables de presencia/ausencia de cada jugador del partido y ϵ es un término de error con distribución normal de media 0. Siendo estrictos, en la matriz estarán todos los jugadores de la liga por lo que las X incluyen a muchísimos jugadores que no son parte del partido pero obviamente tendrán valor 0 y no tendrán injerencia en la suma. Los β son los coeficientes de la regresión.

Dado nuestro modelo, lo siguiente es preguntarse cómo estimar los coeficientes con los datos que tenemos a disposición. Mínimos cuadrados ordinarios (MCO) es posiblemente la primera opción dado que sus coeficientes tienen propiedades interesantes: los estimadores son insesgados y si se cumplen ciertas condiciones sobre los errores del modelo también son los estimadores de mínima varianza dentro de los insesgados.

Los estimadores de MCO se obtienen mediante

$$\hat{\beta} = argmin\ S(\beta)$$

donde

$$S(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2$$

A pesar de mantener la propiedad de insesgadez, estos estimadores son bastante sensibles a la alta correlación entre variables ya que al existir en planos muy próximos existen mayor cantidad de combinaciones lineales entre ambas variables que dan resultados similares en cuanto al ajuste del modelo. Esto se traduce en que los

estimadores de MCO puedan variar mucho entre distintas muestras de la misma población, es decir que son estimadores con varianza elevada. Aunque en promedio los estimadores se centren en los verdaderos valores de β, lo van a hacer de manera muy errática y los valores que encontremos en nuestra muestra no van a ser confiables y por lo tanto nuestras predicciones tampoco.

En este momento es donde hay que tener en mente el tradeoff entre sesgo y varianza de un modelo. Este concepto nos dice que el error cuadrático medio (MSE por sus siglas en inglés) de un set de testeo puede descomponerse entre sesgo elevado al cuadrado más varianza - del modelo estimado y aplicada a la nueva observación - más la varianza del error irreducible del proceso generador de datos.
El último elemento está fijo pero los primeros dos varían según el modelo utilizado y estimado. En general modelos más flexibles tienden a tener menor sesgo ya que pueden ajustarse mejor a las no linealidades de los datos, pero a su vez suelen tener más varianza ya que cambios en los datos tienen impacto sobre cómo ajustan entre muestras distintas. Dijimos que MCO sufre de alta varianza en sus estimadores ante presencia de variables correlacionadas y como mencionamos, el dataset de este trabajo presenta tal característica ya que los jugadores de un equipo son pocos y suelen compartir minutos en cancha de manera reiterada. La propuesta de utilizar una regresión Ridge apunta a tratar de resolver la problemática de la alta varianza. La regularización por Ridge agrega una penalización a la minimización del desvío cuadrático respecto a cero y eso genera dos consecuencias:

- Los estimadores dejan de ser insesgados ya que se los restringe al penalizar su alejamiento de 0.
- Los estimadores ven su varianza reducida ya que se limita el espacio de búsqueda.

Formalmente la minimización es la siguiente:

$$S(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde λ es un hiperparámetro que regula el peso de la penalización en la función de costo. Mayores (menores) valores de λ hacen más (menos) costosos los desvíos de los estimadores respecto de cero. En los extremos, si $\lambda = 0$, los estimadores son los mismos que MCO y si $\lambda = +\infty$ todos los estimadores son 0 y queda solo el intercepto. El objetivo de incluir la penalización es agregar un poco de sesgo a cambio de reducir en mayor medida la componente de varianza del MSE mediante la restricción que se aplica sobre los coeficientes estimados. El éxito de este enfoque depende en gran medida del λ elegido. Para seleccionar el valor de λ la práctica habitual es hiperparametrizar el modelo mediante crossvalidation.

Aplicación

Mencionamos a continuación algunos aspectos metodológicos a tener en cuenta:

- Se eliminó el primer cuartil de jugadores medido en cantidad de posesiones en toda la temporada para reducir la dimensionalidad del problema. Son jugadores con pocos minutos en cancha y perjudican más de lo que aportan. Ver Figura 1.

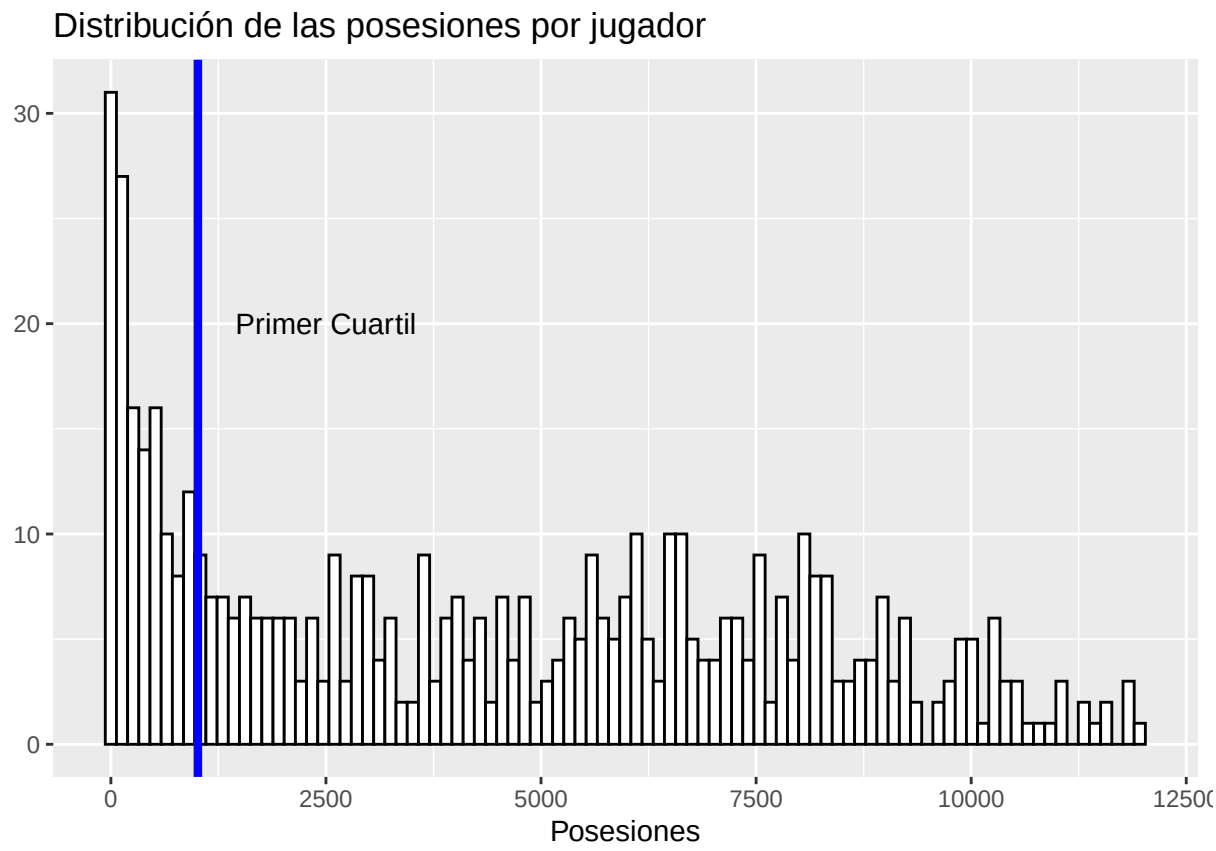

Figura 1: Distribución de las posesiones por jugador

- Se eliminaron las observaciones con menos de 5 posesiones. Este número corresponde a la mediana de la cantidad de posesiones por observación. Se considera que observaciones con menos de esa cantidad de posesiones no son representativas y aumentan el ruido dentro del modelo. Ver figura 2

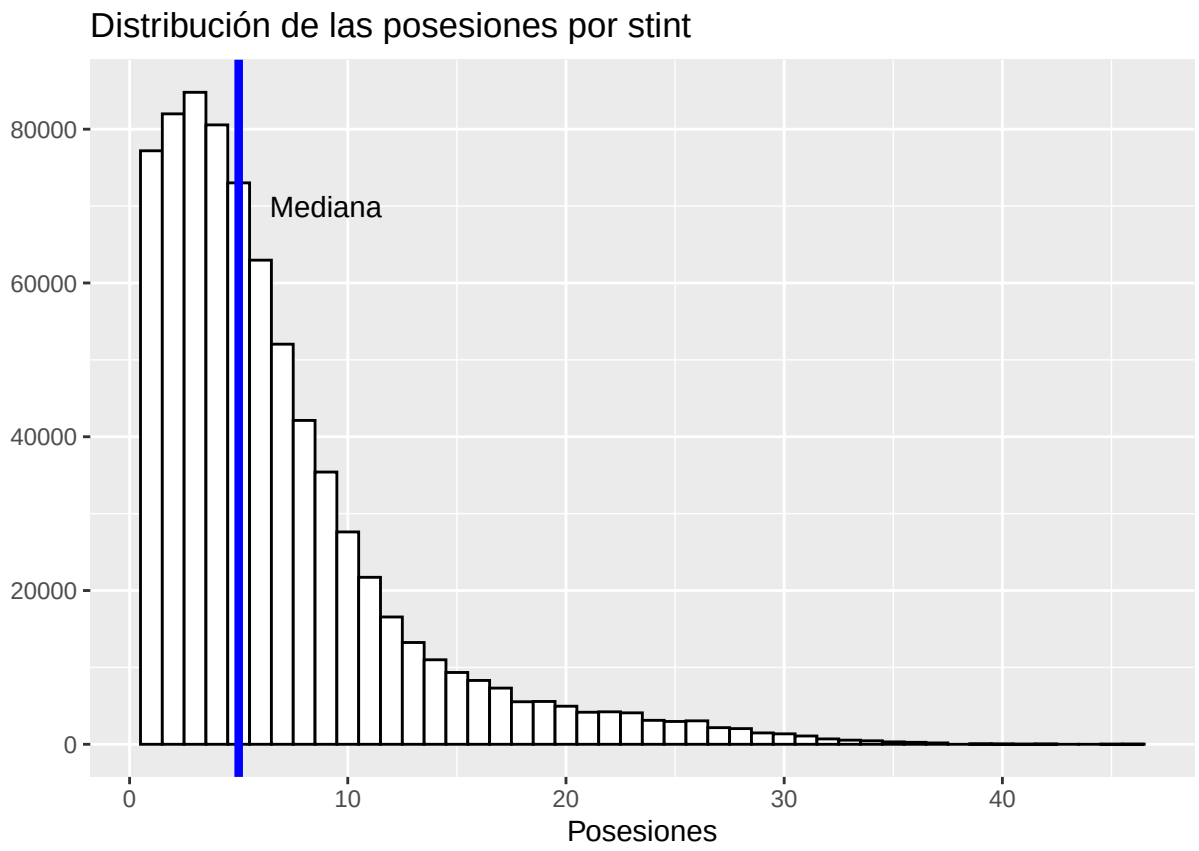

Figura 2: Disitribución de las posesiones por stint

- Se normalizó la variable dependiente, llevando la métrica a diferencial de puntos cada 100 posesiones para hacer comparables los segmentos de tiempo.
- Luego de estos pasos de preprocesamiento nuestro dataset cuenta con 16775 observaciones y 405 variables independientes.

Una vez preprocesados todos los datos y habiendo definido el modelo, procedemos a la etapa de correr la regresión. La librería que vamos a utilizar es *glmnet*[6].
Dijimos que uno de los hiperparámetros de la regresión Ridge es λ que define el peso de la penalización. Lo primero que hacemos es definir qué valor va a tomar para nuestro problema. La solución propuesta es el calcular el λ óptimo mediante Cross Validation con 10 folds. La librería provee la función *cv.glmnet* para tal fin, donde lo que hace es dividir el set de datos en 10 subsets y entrenar un modelo con 9 de los 10 subsets y validarlo con el décimo. Así para cada combinación de subsets. Esto lo realiza reiteradas veces para distintos valores de λ. Finalmente calcula el error cuadrático medio para cada λ en los sets de validación. Entre todos esos resultados buscamos el λ cuyo error promedio en validación sea el menor y será el que utilicemos para nuestro modelo con todos los datos.

Al aplicar este procedimiento obtenemos el óptimo para nuestro set de datos: $\lambda = 1.369778$

Tomando este valor de λ, lo siguiente es correr el modelo Ridge con todo el training set imputando ese valor para el hiperparámetro.

Resultados

A continuación mostramos los primeros diez jugadores según el coeficiente obtenido (de mayor a menor), donde un coeficiente más elevado está asociado a mayor diferencial de puntos a favor mientras se está en cancha, controlando por el resto de los jugadores presentes (compañeros como oponentes). Ver Tabla 2

Cuadro 2: Primeras 10 jugadores según coeficiente

playerid	nombre	apellido	equipo_abre	coef
9418	Joel	Embiid	PHI	2.173248
9420	Robert	Covington	PHI	2.159759
9402	Victor	Oladipo	IND	2.108312
9265	Chris	Paul	HOU	2.019688
9352	Eric	Gordon	HOU	1.932603
9218	Stephen	Curry	GSW	1.791738
9325	Giannis	Antetokounmpo	MIL	1.784740
9152	Jimmy	Butler	MIN	1.776535
9524	Otto	Porter Jr.	WAS	1.705704
9224	Shaun	Livingston	GSW	1.651169

Puede discutirse el orden y si falta algún jugador importante de la liga pero no es un primer ranking totalmente descabellado. Son jugadores de primer nivel y varios de ellos son "indiscutidos". Es destacable que en 10 jugadores haya 3 duplas que comparten equipo. Es posible que esté relacionado a cierta correlación en la presencia de estos jugadores en cancha y a su vez se de en equipos que han tenido muchas victorias en la temporada. Covington, Gordon y Livingston son de alguna manera una sorpresa en este ranking.

Viendo el ranking surge una nueva pregunta. ¿Qué tan precisos son esos coeficientes? ¿Qué tan estable es ese ranking?
La manera de cuantificar esto en una regresión lineal es calcular los desvíos estándar de los coeficientes estimados. Hay un punto muy importante a mencionar respecto a esto. La regresión Ridge es lineal en parámetros, pero como dijimos, incluye una penalización y eso lleva a que los coeficientes sean sesgados. Dicho esto, los desvíos estándar que calculemos serán sobre parámetros estimados que no sabemos en principio qué tan lejos están en promedio de los verdaderos parámetros si el modelo especificado fuera el correcto. Es posible un escenario donde gran parte del MSE de nuestra regresión se de por el sesgo y la varianza no contribuya casi nada, dando la impresión de gran precisión en la estimación ignorando el sesgo que introdujimos.[7]

Teniendo en cuenta esta limitación a la hora de interpretar los resultados pasamos al cálculo de los desvíos estándar. El paquete *glmnet* no tiene una función que devuelva los desvíos, posiblemente para evitar que se reporten sin recaudos. La manera de calcularlos es, de manera resumida:[8]

$$Var[\hat{\beta}(\lambda)] = \sigma^2((X^TX + \lambda I_{nxn}))^{-1}X^TX[(X^TX + \lambda I_{nxn})^{-1}]^T$$

Donde σ^2 es el error del modelo, X es la matriz de variables independientes de dimensión nxp, λ es el peso de la penalización e I es la matriz identidad de dimensión nxn.

Dado que nuestro modelo es sobre una muestra de la población utilizaremos una estimación de $\hat{\sigma}^2$. Implementamos en R la función para calcular los desvíos estándar de nuestros coeficientes y actualizamos el ranking de los primeros diez jugadores con esta nueva información. (Ver Anexo 1) Ver tabla 3

Cuadro 3: Primeras 10 jugadores según coeficiente y su desvío estándar

playerid	nombre	apellido	equipo_abre	coef	sd
9418	Joel	Embiid	PHI	2.173248	3.843356
9420	Robert	Covington	PHI	2.159759	2.741971
9402	Victor	Oladipo	IND	2.108312	3.295909
9265	Chris	Paul	HOU	2.019688	2.543393
9352	Eric	Gordon	HOU	1.932603	2.473872
9218	Stephen	Curry	GSW	1.791738	3.121359
9325	Giannis	Antetokounmpo	MIL	1.784740	2.602599
9152	Jimmy	Butler	MIN	1.776535	3.205942
9524	Otto	Porter Jr.	WAS	1.705704	3.048218
9224	Shaun	Livingston	GSW	1.651169	3.150640
9385	Andre	Roberson	OKL	1.637995	3.395044
9480	LaMarcus	Aldridge	SAS	1.613438	3.156169
10134	Jamal	Murray	DEN	1.551241	3.102419
9445	Damian	Lillard	POR	1.529906	3.126568
10124	Fred	VanVleet	TOR	1.524522	3.079476
9452	Al-Farouq	Aminu	POR	1.345331	3.229132
10087	Ben	Simmons	PHI	1.329176	3.164346
9490	Kyle	Lowry	TOR	1.323359	3.695157
9347	Tyus	Jones	MIN	1.303166	5.090437
13750	OG	Anunoby	TOR	1.256121	3.030928

A primera vista ya vemos que los desvíos estándar se sitúan alrededor de entre 2.5 y 3.5 puntos por cada 100 posesiones, mientras que el jugador con el mayor coeficiente aporta en promedio 2.17 puntos cada 100 posesiones. Los coeficientes están sesgados pero mantienen la propiedad de ser variables normales por lo que podemos aplicar los mismos tests que en una regresión lineal multivariada clásica.

Realizando el test t para ver diferencias estadísticamente significativas respecto a 0 (Anexo 2) de los coeficientes podemos ver que tanto el jugador rankeado número uno "Embiid" como el jugador rankeado número veinte "Anunoby" tienen p-valores de 0.29 y 0.34 respectivamente, no encontrando evidencia suficiente para decir que tienen impacto positivo en los puntos de su equipo teniendo un umbral de $\alpha = 0.05$. Más allá de las limitaciones de nuestros estimadores, no poder asegurar estadísticamente que el jugador mejor rankeado aporta positivamente a su equipo parece un obstáculo importante.

Yendo más allá, proponemos un test de Wald para comparar los coeficientes del rankeado número 1 contra el rankeado en la última posición (Anexo 3). La hipótesis nula es que ambos coeficientes son iguales. El test de Wald se basa en que la variable aleatoria generada por la resta de ambos coeficientes sea igual a 0 en la hipótesis nula. Se la divide por su desvío estándar generando un estadístico con distribución T de Student y luego se compara con las regiones críticas. En este caso el p-valor es de 0.21, lo cual nos dice que de ser cierta la hipótesis nula, encontrar un valor al menos tan extremo como este para el estadístico tiene 0.21 de probabilidad. Nuevamente no podemos rechazar la hipótesis nula con $\alpha = 0.05$ y podemos decir que no hay evidencia estadística para diferenciar al primero del último jugador del ranking por sus coeficientes.
Al margen de lo ya aclarado de que los desvíos estándar están sesgados, resultados como estos no parecen darle demasiada entidad al modelo. Pareciera que la reducción de la varianza mediante la regresión Ridge, tal como se hizo el experimento, no alcanza para generar coeficientes estables. Ver figura 3.

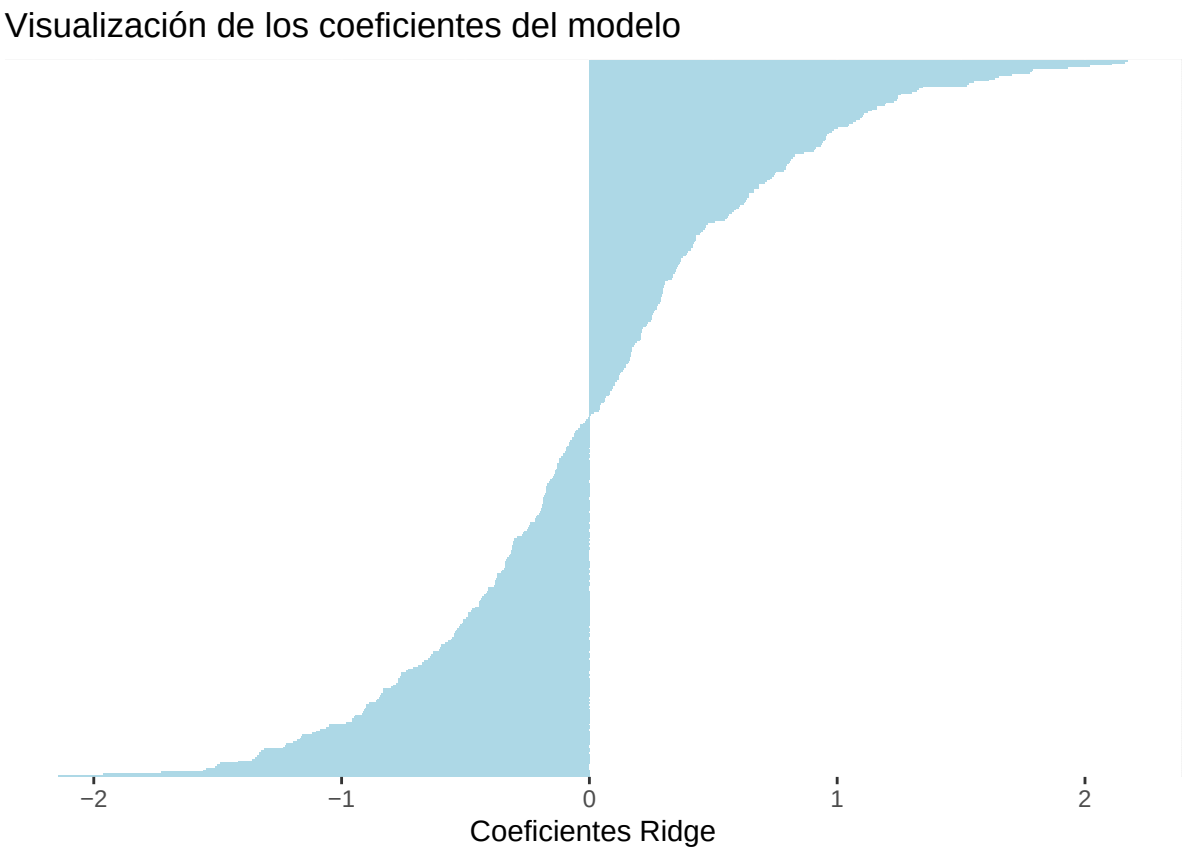

Figura 3: Coeficientes de los jugadores según el modelo Ridge

A pesar de los resultados del test que le quitan importancia a los coeficientes podemos darnos cuenta que aunque sea direccionalmente no es totalmente descabellado (aunque no sirva para obtener conclusiones sólidas). Los jugadores de los primeros puestos del ranking son en gran parte de elite y no sorprende verlos allí. De la misma manera, en la parte baja de la tabla vemos jugadores poco participativos de equipos que obtuvieron buenos resultados en la temporada y principalmente jugadores de equipos que han tenido una campaña pobre (Tabla 4)
Esta situación donde los resultados parecen coherentes, pero no permiten concluir estadísticamente acerca de ellos, hace pensar que quizás la muestra de un solo año no es suficientemente grande para discernir efectos entre jugadores y recolectar información de mayor cantidad de temporadas es necesario. Por otra parte, la regresión Ridge penaliza la diferencia de los coeficientes respecto a 0, lo cual lleva a acercar todos los coeficientes a ese valor. Puede ser interesante probar si otro enfoque permite diferenciar más a los jugadores, manteniendo el requisito de regularizar. Esto incluiría diferentes valores para los cuales la distancia es penalizada, es decir, no regularizar hacia 0 todos los jugadores sino hacia distintos valores según algún criterio o experiencia en el área. De la misma manera se podría probar algún enfoque bayesiano con distintos priors para los coeficientes que permitan mayor diferenciación entre ellos para compensar en parte la todavía elevada varianza que poseen.

Cuadro 4: Ultimos 20 jugadores según coeficiente

playerid	player.firstName	player.lastName	team.abbreviation	coef	sd
13731	Josh	Jackson	PHX	-1.196522	2.415776
9206	Jodie	Meeks	WAS	-1.224898	3.756302
13726	Bogdan	Bogdanovic	SAC	-1.226335	2.685511
9263	Jamal	Crawford	MIN	-1.234573	3.423356
9340	Gorgui	Dieng	MIN	-1.312150	3.672307
10158	Paul	Zipser	CHI	-1.326052	3.337279
13815	Kobi	Simmons	MEM	-1.331096	3.588547
9195	Jameer	Nelson	DET	-1.333314	3.208418
10104	DeAndre	Liggins	NOP	-1.342334	3.405218
10112	Patrick	McCaw	GSW	-1.350563	3.111211
9421	Jahlil	Okafor	BRO	-1.362062	4.798204
9127	Courtney	Lee	NYK	-1.417688	3.105015
9104	Evan	Turner	POR	-1.491482	2.834678
9407	Elfrid	Payton	PHX	-1.501153	2.812676
9194	Emmanuel	Mudiay	NYK	-1.508399	2.956592
9156	Cristiano	Felicio	CHI	-1.545031	3.916368
9497	Bismack	Biyombo	ORL	-1.559086	3.781430
13739	Malik	Monk	CHA	-1.729935	3.845401
9458	Ben	McLemore	MEM	-1.962150	2.871301
9423	TJ	McConnell	PHI	-2.143543	2.695342

Dicho esto podemos ir un paso más allá en el análisis y proponer un ranking de equipos basado en la suma de los coeficientes obtenidos para los jugadores que los componen multiplicado por el promedio de posesiones por partido de cada jugador.

$$ValorEquipo_i = \sum_{j=1}^{n} Coef_{ji} * AvgPos_{ji}$$

Donde j es cada jugador del equipo i.

Nuevamente los resultados son bastante creíbles, particularmente en la parte alta de la tabla donde los equipos que más ganaron en la temporada regular figuran en los primeros puestos. En la NBA los mejores 8 equipos de cada conferencia (este y oeste) clasifican a playoffs. De esos 16 equipos, 14 figuran en los primeros 16 puestos del ranking propuesto en este trabajo. Las únicas diferencias es que Denver y Detroit figuran entre los 16 primeros en nuestro ranking y Cleveland y Milwaukee no. Lo curioso es que Denver en realidad tuvo más victorias que Milwaukee pero quedo noveno en su conferencia, es decir que la diferencia entre ambas tablas es muy fina en términos de quienes clasifican a la instancia de eliminación directa. Por otra parte, Cleveland que en nuestro ranking queda fuera de los primeros ocho del este, en la realidad salió cuarto de su conferencia, dando lugar a una diferencia significativa entre su ranking y su posición en la tabla.
Ver tablas 5 y 6.
No es menor que el ranking se armó usando los partidos que dieron lugar a los resultados de la temporada regular y por lo tanto es como comparar con el set de entrenamiento, sin embargo, lo destacable es que el ranking se arma a partir de la suma de coeficientes de los distintos jugadores y no mirando a los equipos directamente.

Cuadro 5: Ranking de Equipos agregando coeficientes

equipo_abre	coef_total
TOR	780.40794
GSW	765.95924
HOU	730.41546
UTA	600.07384
PHI	491.22468
BOS	460.42148
OKL	432.35755
POR	360.03984
IND	335.03032
NOP	329.89205
SAS	324.15970
MIN	300.39813
DEN	230.85456
WAS	218.68353
DET	143.91906
MIA	125.47316
MIL	66.36723
CLE	53.82317
LAC	48.89531
ORL	-20.12870
CHA	-98.09310
LAL	-178.28902
NYK	-232.22049
BRO	-282.95437
MEM	-308.55069
DAL	-319.49053
ATL	-381.78566
PHX	-384.90047
SAC	-524.88377
CHI	-614.03632

Cuadro 6: Equipos ordenados según cantidad de victorias

equipo_abre	equipo	G	P	G_ratio
HOU	Houston Rockets* (1)	65	17	0.793
TOR	Toronto Raptors* (1)	59	23	0.720
GSW	Golden State Warriors* (2)	58	24	0.707
BOS	Boston Celtics* (2)	55	27	0.671
PHI	Philadelphia 76ers* (3)	52	30	0.634
CLE	Cleveland Cavaliers* (4)	50	32	0.610
POR	Portland Trail Blazers* (3)	49	33	0.598
IND	Indiana Pacers* (5)	48	34	0.585
OKC	Oklahoma City Thunder* (4)	48	34	0.585
UTA	Utah Jazz* (5)	48	34	0.585
NOP	New Orleans Pelicans* (6)	48	34	0.585
SAS	San Antonio Spurs* (7)	47	35	0.573
MIN	Minnesota Timberwolves* (8)	47	35	0.573
DEN	Denver Nuggets (9)	46	36	0.561
MIA	Miami Heat* (6)	44	38	0.537
MIL	Milwaukee Bucks* (7)	44	38	0.537
WAS	Washington Wizards* (8)	43	39	0.524
LAC	Los Angeles Clippers (10)	42	40	0.512
DET	Detroit Pistons (9)	39	43	0.476
CHA	Charlotte Hornets (10)	36	46	0.439
LAL	Los Angeles Lakers (11)	35	47	0.427
NYK	New York Knicks (11)	29	53	0.354
BRO	Brooklyn Nets (12)	28	54	0.341
CHI	Chicago Bulls (13)	27	55	0.329
SAC	Sacramento Kings (12)	27	55	0.329
ORL	Orlando Magic (14)	25	57	0.305
ATL	Atlanta Hawks (15)	24	58	0.293
DAL	Dallas Mavericks (13)	24	58	0.293
MEM	Memphis Grizzlies (14)	22	60	0.268
PHX	Phoenix Suns (15)	21	61	0.256

Test en Playoffs

Lo siguiente que vamos a analizar es cómo performa este ranking construido en la siguiente ronda de la competencia: los playoffs. Los partidos de esa instancia no se usaron para el entrenamiento del modelo.

El enfoque será el más simple posible y consiste en declarar como favorito para cada partido al equipo cuyo score sea mayor en el cruce, teniendo en cuenta los jugadores en el plantel para cada enfrentamiento, contemplando posibles lesiones, sanciones ,etc. Es decir que para cada partido se recalcula el score como la suma ponderada de coeficientes por posesiones promedio de cada jugador y se compara con el score del adversario.

$$\begin{cases} Gana_A & \text{si } ValorEquipo_A > ValorEquipo_B \\ Gana_B & \text{si } ValorEquipo_B > ValorEquipo_A \end{cases}$$

En la figura 4 podemos ver que los scores por equipo no varían demasiado salvo algunas excepciones, lo cual tiene sentido ya que los jugadores tienen un coeficiente fijo y quienes forman parte del partido no suele cambiar tan frecuentemente. Generalmente se debe a lesiones o imprevistos. Dicho esto, es de esperar que si comparamos scores de dos equipos para cierta cantidad de partidos mayor a uno, salvo que sean muy parejos, sea siempre el mismo el que tenga un score mayor. En los playoffs, los cruces son al mejor de siete partidos, por lo tanto aquel que logra ganar cuatro partidos pasa a la siguiente ronda y el otro queda eliminado. Dado

lo simple del modelo propuesto, donde no se toma en cuenta ningún otro factor más que el score, podemos ver cómo performa a un nivel agregado, es decir, no mirando partido a partido, donde cómo dijimos en la gran mayoría de los casos se va a predecir siempre al mismo ganador, si no a nivel cruce, donde se predice que el equipo que ganará la serie será aquel que tenga mayor score en mayor cantidad de partidos del cruce. Luego se compara esa predicción con el equipo que realmente pasó de ronda. Ver tablas 7 y 8

Cuadro 7: Predicción del ganador de la serie basado en score

equipo1	equipo2	ganador	predicho	fit
BOS	CLE	CLE	BOS	0
BOS	MIL	BOS	BOS	1
BOS	PHI	BOS	PHI	0
CLE	GSW	GSW	GSW	1
CLE	IND	CLE	IND	0
CLE	TOR	CLE	TOR	0
GSW	HOU	GSW	GSW	1
GSW	NOP	GSW	GSW	1
GSW	SAS	GSW	GSW	1
HOU	MIN	HOU	HOU	1
HOU	UTA	HOU	HOU	1
MIA	PHI	PHI	PHI	1
NOP	POR	NOP	NOP	1
OKC	UTA	UTA	UTA	1
TOR	WAS	TOR	TOR	1

Cuadro 8: Tabla de resultados

acierto	Freq
0	4
1	11

Vemos en la tabla 7 que utilizando los scores basados en el modelo Ridge se pudieron predecir correctamente 11 de los 15 cruces de los playoffs de la temporada 2017-2018. Vemos que la principal falencia se da con Cleveland, donde al tener un score tan bajo dados los coeficientes lo pronosticaba perdedor en todas las instancias cuando finalmente llegó a la final. Está claramente relacionado a lo mencionado anteriormente, que Cleveland quedaba fuera de los 16 primeros puestos según nuestro ranking cuando en realidad clasificó en cuarto lugar en su conferencia. A pesar de le gran varianza de los coeficientes de la regresión que da origen a este ranking y de lo simple del modelo para seleccionar ganadores los resultados no son decepcionantes por lo que aparentemente la regresión es atinada direccionalmente.

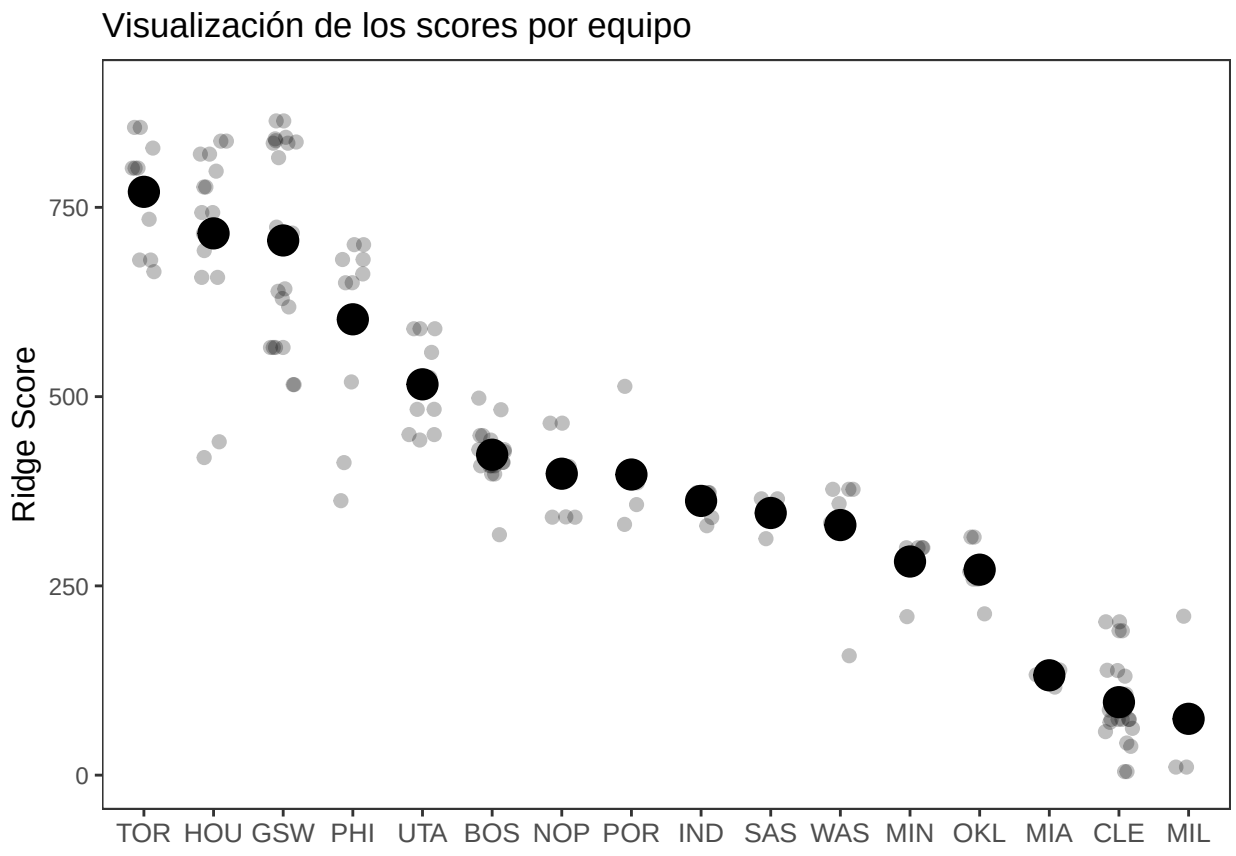

Figura 4: Scores por equipo para los distintos partidos de playoffs. El círculo grande representa el score promedio de cada equipo mientras que los círculos pequeños son los scores para cada partido de playoffs.

Conclusiones

El objetivo principal de este trabajo es tratar de dar una medida de la incidencia de un jugador en los puntos de su equipo controlando por sus compañeros y adversarios en cancha para no sobreestimar (subestimar) a quienes se ven beneficiados (perjudicados) por los otros jugadores presentes. Trata de dar un resultado superador al "Plus Minus" clásico que se suele ver en las estadísticas de los partidos de la NBA. La manera propuesta es realizando una regresión Ridge donde los las variables independientes son los jugadores en cancha en cada "stint". Se busca regresar de esta manera para lograr este control mencionado y, dada la alta correlación porque los jugadores no son muchos, se intenta regularizar los coeficientes para evitar resultados poco consistentes. Lo primero que se puede observar de los resultados obtenidos es que los jugadores con coeficientes más altos, y por lo tanto con mayor aporte positivo a sus equipos, son en gran medida jugadores de mucho renombre y que no sorprende verlos en los primeros puestos. Por otra parte, con los jugadores de coeficiente más negativo pasa algo similar. No hay ninguno que parezca imposible encontrarlo allí. Sin embargo, para los segmentos no tan extremos son más difíciles de distinguir y jugadores que en principio uno esperaría más cerca de la cima se los ve por la mitad de la tabla. Evaluando esta situación vimos que los desvíos estándar de los coeficientes son muy altos y estadísticamente no podemos diferenciar entre el jugador con el mayor coeficiente y el de menor coeficiente. A pesar de la reducción en varianza que implica una regresión Ridge comparado con regresión lineal clásica no podemos encontrar coeficientes con estos datos que sean robustos. Parece entonces que direccionalmente el modelo es acertado, pero no permite concluir con certeza acerca del orden de los coeficientes.
Dicho esto, para ver de manera más agregada que tan creíbles son los coeficientes se procedió a crear un

estadístico por equipo siendo este la suma ponderada por posesiones en cancha de los coeficientes de los jugadores de cada equipo. Esto dio lugar a un nuevo ranking, esta vez de equipos, donde nuevamente parece haber lógica. Los equipos mejor rankeados son los que terminaron con mayor cantidad de victorias en la temporada, y 14 de los 16 equipos que clasificaron a playoffs se encuentran en las primeras 16 posiciones. A pesar de la varianza de los coeficientes de los jugadores, la agregación da resultados que no parecen muy errados. Dada la aparente buena performance en el ranking a nivel equipos en la temporada regular, que sería el training set, se decidió probar en los cruces de playoff el ranking para ver si servían como predictores de quién ganaría. Se mencionó que partido a partido las alineaciones casi no cambian por lo que los estadísticos se mantienen bastante parejos. Dado que el modelo era muy simple (el mayor score se predice como ganador) se decidió hacerlo a nivel serie, es decir, predecir quién pasa de ronda que equivale a ganar 4 partidos de 7. De los 15 cruces que hay en playoffs se predijeron bien 11 de ellos, lo cual no parece nada mal. De los 4 fallados, 3 corresponden a Cleveland, quién llegó a las finales y consistentemente era predicho como perdedor según el modelo. Cleveland es uno de aquellos equipos que el ranking de equipos clasificaba fuera de los 16 primeros equipos y que finalmente tuvo una gran temporada. El modelo Ridge no logró posicionar correctamente a sus jugadores (quizás consecuencia de la varianza) o quizás el equipo elevó su rendimiento en playoffs.

Claramente el modelo no alcanza a ser robusto como para asegurar que los coeficientes representan el aporte de los jugadores, pero aparentemente es suficientemente bueno como para que en promedio los equipos tengan un ranking bastante acorde a su rendimiento durante la temporada y no es descabellado usarlo de parámetro para predecir resultados en playoffs.

Próximos pasos

Dado que la varianza parece ser el principal obstáculo de la regresión Ridge se sugiere que un próximo paso posible sea repetir el ejercicio incluyendo datos de más temporadas para intentar ampliar la muestra por jugador y reducir la correlación. La contraparte es que quizás los jugadores cambian su rendimiento a lo largo del tiempo y puede que sus aportes sean más dispersos. El desafío es ver si a pesar de esta nueva dimensión que se agrega se puede llegar a coeficientes más estables. Quizás se puede probar incluyendo año a año un prior basado en el coeficiente obtenido en el período anterior o alguna variante similar para reducir el espacio de búsqueda de los coeficientes.

Bibliografía

[**1**] Jacobs, J. (2017). Deep Dive on Regularized Adjusted Plus-Minus I: Introductory Example. Disponible en: https://squared2020.com/2017/09/18/deep-dive-on-regularized-adjusted-plus-minus-i-introductory-example/

[**2**] Sill, J. (2010) Improved NBA Adjusted +/- Using Regularizationand Out-of-Sample Testing. Disponible en: http://www.sloansportsconference.com/wp-content/uploads/2015/09/joeSillSloanSportsPaperWithLogo.pdf

[**3**] MySportsFeeds (2019). API Documentation. Disponible en: https://www.mysportsfeeds.com/data-feeds/api-docs

[**4**] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[**5**] Hoerl, A. and Kennard, R. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[**6**] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

[**7**] Goeman, J., Meijer, R., Chaturvedi, N. (2018). L1 and L2 Penalized Regression Models. Disponible en: https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

[8] van Wieringen, W. N. (2019). Lecture notes on ridge regression. Disponible en https://arxiv.org/pdf/1509.09169;Lecture

Anexo

1. Función para calcular desvíos estándar de la regresión Ridge

```
ridge_se <- function(xs,y,yhat,my_mod){
  x2 <- as.matrix(xs)
  n <- dim(x2)[1]
  k <- dim(x2)[2]
  sigma_sq <- sum((y-yhat)^2)/ (n-k)
  lam <- my_mod$lambda
  if(is.null(my_mod$lambda)==TRUE){lam <- 0}
  i_lams <- matrix(diag(x=1,nrow=k,ncol=k))# ,sparse=TRUE)
  xpx <- t(x2)%*%x2
  xpxinvplam <- solve(xpx+lam*as.vector(i_lams))
  var_cov <- sigma_sq * (xpxinvplam %*% xpx %*% xpxinvplam)
  se_bs <- sqrt(diag(var_cov))
  return(se_bs)
}
```

xs son los predictores, y la variable dependiente, yhat la predicción, my_mod la regresión Ridge.

2. Test T, diferencias de coeficientes respecto a 0

```
tscore_1 = (player_ranking$coef[1] - 0)/player_ranking$sd[1]
pvalue_1 = 1 - pt(q = tscore_1, df = model$nobs - model$df)

tscore_20 = (player_ranking$coef[20] - 0)/player_ranking$sd[20]
pvalue_20 = 1 - pt(q = tscore_20, df = model$nobs - model$df)
```

3. Test de Wald. Diferencias entre coeficientes del rankeado número uno y el último.

```
wald_1_405_init = player_ranking$coef[1] - player_ranking$coef[405]

sd_1_405 = sqrt(var_cov_ridge[paste0("x",player_ranking$playerid[1]),
              paste0("x",player_ranking$playerid[1])]+
              var_cov_ridge[paste0("x",player_ranking$playerid[405]),
              paste0("x",player_ranking$playerid[405])] -
              2*var_cov_ridge[paste0("x",player_ranking$playerid[1]),
              paste0("x",player_ranking$playerid[405])])

wald_1_405_t = wald_1_405_init / sd_1_405
pvalue_1_405 = 1 - pt(q = wald_1_405_t, df = model$nobs - model$df)
```