Selección de Modelos

Carlos Hernani Morales

20 de Diciembre, 2019

Abstract

Los modelos de regresión múltiple pueden presentar ciertas dificultades debidas a la presencia de más de una variable predictora. Al aumentar el número de predictores el modelo pierde explicabilidad, algunos predictores pueden estar correlacionados con otros, y por tanto aportan información redundante, y otros directamente pueden no aportar ninguna información relevante a la variable de interés de nuestro análisis. Para resolver estos problemas se plantean diversos modelos con diferentes predictores y compararemos para ver cuál de ellos es el mejor. Esto es lo que se conoce por Selección de Modelos. En el presente trabajo explicaré los diversos procedimientos aplicándolos sobre un banco de datos de la NBA.

Introducción

El banco de datos *NBA* presente en Kaggle contiene 835 observaciones con 20 columnas, donde nuestra variable respuesta de interés es **PTS**, los puntos marcados por un equipo en una temporada. Por tanto nuestro **objetivo** es seleccionar el mejor modelo que prediga **PTS**.

Table 1: Tabla de variables y estadísticos.

Variable	Percentil 0	Percentil 25	Percentil 50	Percentil 75	Percentil 100
Team	NA	NA	NA	NA	NA
SeasonEnd	1980	1989.0	1996	2005.0	2011
Playoffs	0	0.0	1	1.0	1
W	11	31.0	42	50.5	72
PTS	6901	7934.0	8312	8784.5	10371
oppPTS	6909	7934.0	8365	8768.5	10723
FG	2565	2974.0	3150	3434.5	3980
FGA	5972	6563.5	6831	7157.0	8868
X2P	1981	2510.0	2718	3296.0	3954
X2PA	4153	5269.0	5706	6753.5	7873
X3P	10	131.5	329	481.5	841
X3PA	75	413.0	942	1347.5	2284
FT	1189	1502.5	1628	1781.0	2388
FTA	1475	2008.0	2176	2352.0	3051
ORB	639	953.5	1055	1167.0	1520
DRB	2044	2346.5	2433	2516.5	2753
AST	1423	1735.0	1899	2077.5	2575
STL	455	599.0	658	729.0	1053
BLK	204	359.0	410	469.5	716
TOV	931	1192.0	1289	1395.5	1873

El conocimiento previo del problema nos permite eliminar ciertas variables predictoras como:

- **SeasonEnd,Playoffs,W,oppPTS** que se corresponden con el año de cierre de temporada, si jugaron o no *playoffs*, el número de partidos que ganaron y los puntos que han recibido de oponentes. Está justificado descartar estas variables puesto que no existe relación causal entre estas y los puntos marcados.
- **FG** y **FGA** son los *field goals* y *field goals attempts* que son la suma de los puntos dobles y triples, sin contar tiros libres. Si los tuviéramos en cuenta existirían problemas de colinealidad con las variables asociadas a dobles y triples.
- Además como nuestro interés es predecir los puntos que marcarán un equipo nos interesa saberlo en función de los intentos o *attempts*, ya que estos incluyen los puntos marcados al rival.

De modo que realizaremos el análisis sobre un conjunto inicial de variables que será:

Variables	Descripción
PTS	Variable respuesta
X2PA	Intentos tiro de campo 2 pts
X3PA	Intentos tiro de campo 3 pts
FTA	Intentos tiro libre
ORB	Rebotes ofensivos
DRB	Rebotes defensivos
AST	Asistencias
STL	Robos de balón
BLK	Bloqueos
TOV	Pérdida del balón

Selección de modelos

Para seleccionar el mejor modelo debemos definir primero lo que entendemos por *mejor*. En este caso es aquel modelo que generaliza mejor y esto lo consigueremos para aquel que tenga menor error en test. El cálculo de este error puede ser directo, utilizando otro conjunto de datos de validación o mediante *cross-validation* o bien, de forma indirecta utilizando estimadores como el AIC,BIC, etc. que nos permiten comparar modelos con diferente dimensionalidad.

El banco de datos que he usado ya tiene un conjunto de validación, así que podré comprobar todos los métodos de selección de modelos.

Una vez definido el criterio de elección del mejor[1] modelo hay que proponer los algoritmos que nos permitan explorar los diferentes modelos posibles. En el curso hemos visto:

- *Best Subset o Mejor Modelo*: exploramos todas las combinaciones posibles de predictores, i.e. búsqueda exhaustiva.
- *Selección por etapas*: exploramos un espacio de modelos más reducido, es más rápido pero no garantiza encontrar el *mejor* modelo. Hay tres tipos:
  - Hacia delante.
  - Hacia atrás.
  - Híbrido.

Todos estos procedimientos se encuentran implementados en el comando *regsubsets* de la librería *leaps* de R.

Mediante la ayuda de gráficas obtendremos el número de variables óptimo para cada método según los diferentes criterios que hemos mencionado con anterioridad.

[1] La nomenclatura puede resultar ambigua, pero por contexto se puede entender a que me refiero

Mejor Modelo:

El algoritmo de Mejor Modelo, recorre todas las posibles combinaciones de predictores. Esta búsqueda exhaustiva puede llegar a ser costosa computacionalmente o imposible a partir de un número de predictores grande, p.ej.: 40 predictores, en nuestro caso solo tenemos 9 por lo que es asequible.
Para los modelos del mismo número de predictores se elige el que mejor valor de RSS o R^2 tiene y de entre estos se elige como el "mejor" modelo aquel que mejor valor de C_p (AIC), BIC, R^2 ajustado, error de validación o de validación cruzada tenga.

Para regresión múltiple C_p y AIC son proporcionales.

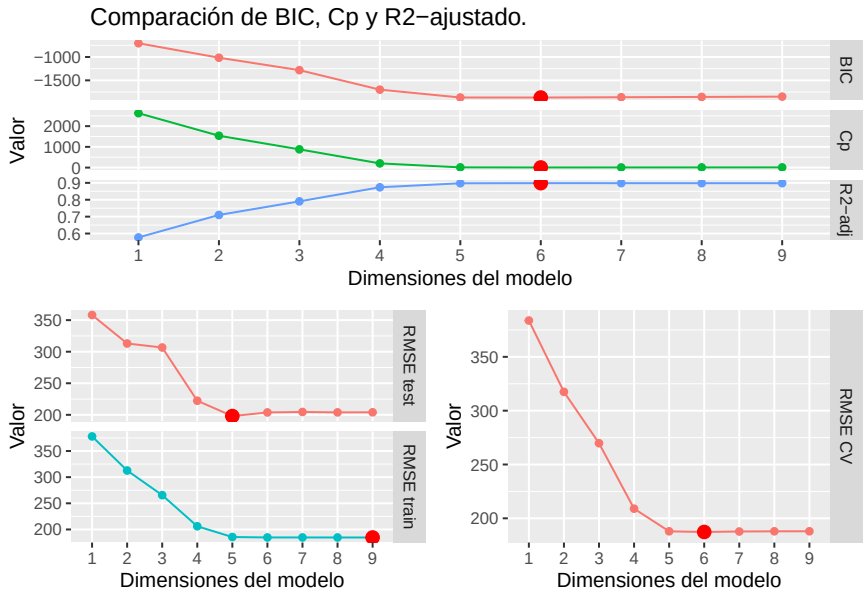

Figure 1: Algoritmo Mejor Modelo

Como ya he mencionado las gráficas nos ayudan a encontrar el mejor número de variables para nuestro modelo, en este caso el método es el "Mejor Modelo" y lo que obtenemos es que el modelo que minimiza el error en la muestra de validación es aquel cuyo número de variables es 5. También se puede apreciar el *overfitting* o sobreajuste en el RMSE de la muestra de entrenamiento.

La búsqueda exhaustiva del mejor modelo nos proporciona el siguiente resultado:

	Coeficientes
(Intercept)	-2039.2852012
X2PA	1.0454406
X3PA	1.2665846
FTA	1.1175738
AST	0.8650863
ORB	-1.0177804

Selección por etapas:

En la selección por etapas recorremos un espacio reducido de los modelos de modo que en vez de calcular 2^p modelos, donde p es el número de predictores, recorremos $1+p(p+1)/2$ que en nuestro caso sería pasar de 512 a 46 modelos.

Selección por etapas hacia delante.

En este método se parte del modelo nulo y se van añadiendo variables sucesivamente.

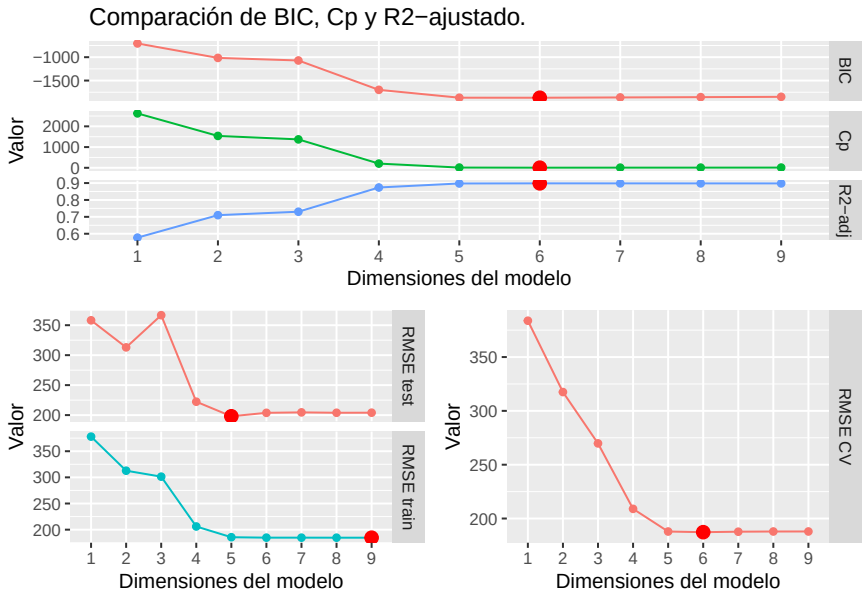

Figure 2: Algoritmo Selección por etapas hacia delante.

Fijémonos ahora que no todos los criterios nos dan el mismo número de variables. El cálculo indirecto del error mediante C_p (AIC), BIC, R^2 ajustado y el cálculo directo mediante validación cruzada nos da un modelo de 6 variables igual que antes pero en el caso del error de validación nos da un modelo de 5 variables. Como el objetivo propuesto en Kaggle es mejorar el error en la muestra de validación nos quedamos con el modelo de 5 variables para el método de selección hacia delante.

	Coeficientes
(Intercept)	-2039.2852012
X2PA	1.0454406
X3PA	1.2665846
FTA	1.1175738
AST	0.8650863
ORB	-1.0177804

Selección por etapas hacia atrás.

Un requisito previo de la selección por etapas hacia atrás es que el número de observaciones sea mayor que el número de variables del modelo. Como estamos tratando 835 observaciones frente a 9 variables no hay ningún problema.

En este método se procede a la inversa que en el anterior, partimos del modelo completo y vamos eliminando sucesivamente las variables que menos aporten información al modelo.

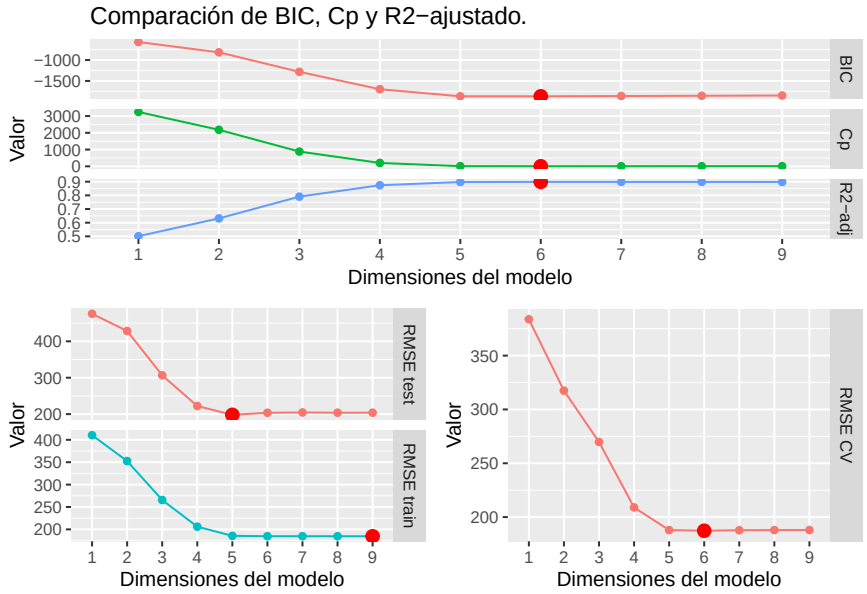

Figure 3: Algoritmo Selección por etapas hacia atrás.

Igual que en el anterior el modelo que mejor resultado da en la muestra de validación es el de 5 variables.

	Coeficientes
(Intercept)	-2039.2852012
X2PA	1.0454406
X3PA	1.2665846
FTA	1.1175738
AST	0.8650863
ORB	-1.0177804

Selección por etapas híbrida.

Como el nombre indica, la selección por etapas híbrida es una combinación de los dos métodos anteriores de modo que se añaden variables y luego se desechan otras que no supongan una mejora del modelo.

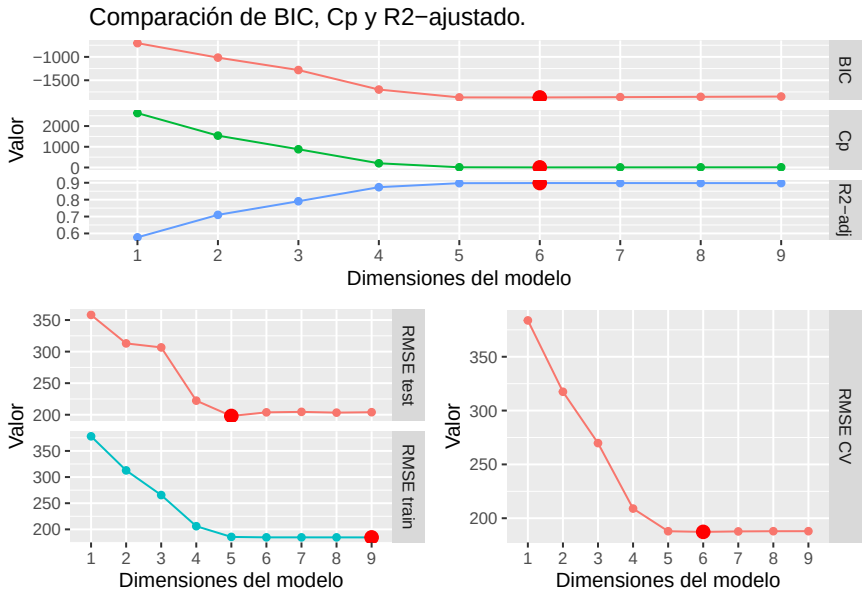

Figure 4: Algoritmo Selección por etapas híbrida.

Al igual que en los anteriores métodos de selección obtenemos el mismo resultado, un modelo de 5 variables.

	Coeficientes
(Intercept)	-2039.2852012
X2PA	1.0454406
X3PA	1.2665846
FTA	1.1175738
AST	0.8650863
ORB	-1.0177804

Comparación de métodos.

Todos los métodos nos han proporcionado el mismo resultado si nos fijamos en el error de validación como criterio de selección, esto no tiene porque ser así en general, nuestro problema tiene muy pocas variables como para que haya demasiada disonancia entre los resultados proporcionados por los diferentes métodos, en casos de muchas más variables, aparecerían diferencias entre estos.

El estadístico utilizado para el criterio de error de validación es el RMSE, una estimación de σ del modelo:

$$RMSE = \sqrt{\frac{RSS}{n-2}}$$

La diferencia entre el RMSE de los modelos con 5 y 6 variables:

$$RMSE_5 = 198 \quad RMSE_6 = 204$$

Y la media de los puntos es 8370 por temporada y equipo, a unos 82 partidos por temporada y equipo equivale a unos 102 puntos por partido y un RMSE por partido de:

$$RMSE_5 \approx 2 \quad RMSE_6 \approx 3$$

Es decir que por cada partido nos vamos en la mayoría de casos tres veces esta cantidad, unos 6 o 9 puntos.

Conclusiones

Concluimos por tanto que el modelo que mejor explica nuestros datos es una regresión lineal múltiple con 5 variables:

$$PTS \sim X2PA + X3PA + FTA + AST + ORB$$

Con los siguientes coeficientes asociados:

	Coeficientes
(Intercept)	-2039.2852012
X2PA	1.0454406
X3PA	1.2665846
FTA	1.1175738
AST	0.8650863
ORB	-1.0177804

Y que nos proporciona un RMSE en el conjunto de validación de: 198

Anexo: Código utilizado

Dataset: https://www.kaggle.com/amanajmera1/national-basketball-associationnba-dataset
Repositorio de GitHub: https://github.com/carhermo/TrabajoModelosLineales