Salarios NBA

Sergio Cañón

28 octubre, 2020

Table of Contents

Introducción

El objetivo del trabajo es el de crear un modelo con regresión lineal para predecir los salarios de los jugadores de la NBA.

Data set

El conjunto de datos desde la NBA y pertenece a un conjunto de datos de la misma página oficial. Es de tipo multivariante y contiene 485 muestras de jugadores con 28 variables, siendo la variable NBA$SALARY nuestro target.

Metodología

La metodología a seguir es la creación de una fórmula de regresión lineal e intentar ver sí con la normalidad y sin ella se puede obtener un modelo predictivo para un salario. Vamos a incluir los equipos de la NBA por qué el tope salarial de cada equipo varía en función de diferentes variables y también vamos a incluir la procedencia del jugador que, aunque en la vida real no se pueda poner un salario a un jugador por su procedencia queremos introducirlo para poder ver si nos da algún tipo de información.

Cargar librerías

```
LIBRARY(READR)
LIBRARY(PANDER)
LIBRARY(RMDFORMATS)
LIBRARY(DPLYR)
LIBRARY(SKIMR)
LIBRARY(TIDYR)
LIBRARY(GGPLOT2)
LIBRARY(GRIDEXTRA)
LIBRARY(PSYCH)

LIBRARY(FASTDUMMIES)
LIBRARY(GVLMA)
LIBRARY(CAR)
LIBRARY(CARDATA)

LIBRARY(GVLMA)
LIBRARY(MASS)
```

Importación del Data Set

Player	Salary	NBA_Country	NBA_DraftNumber	Age	Tm
Zhou Qi	815615	China	43	22	HOU
Zaza Pachulia	3477600	Georgia	42	33	GSW
Zach Randolph	12307692	USA	19	36	SAC
Zach LaVine	3202217	USA	13	22	CHI
Zach Collins	3057240	USA	10	20	POR
Yogi Ferrell	1312611	USA	62	24	DAL

Limpieza de datos

Los países de origen de los jugadores son los siguientes:

NBA_Country	n	NBA_Country	n
Argentina	2	Russia	1
Australia	8	Senegal	1
Austria	1	Serbia	5
Bahamas	1	Slovenia	1
Bosnia	1	South Sudan	1
Bosnia & Herz…	1	Spain	7
Brazil	5	Sweden	1
Cameroon	3	Switzerland	2
Canada	12	Tunisia	1
China	1	Turkey	5
Croatia	6	Ukraine	2
Czech Republic	1	United Kingdo…	2
Democratic Re…	2	USA	374
Democratic Re_	1		
Dominican Rep…	2		
Egypt	1		
Finland	1		
France	9		
Georgia	1		
Germany	5		
Greece	2		
Haiti	1		
Israel	1		
Italy	2		
Latvia	2		
Lithuania	3		
Mali	1		
Montenegro	2		
New Zealand	1		
Poland	1		
Puerto Rico	2		

China, *Georgia*, *USA*, *Canada*, *Spain*, *France*, *Czech Republic*, *Russia*, *South Sudan*, *Switzerland*, *New Zealand*, *Haiti*, _Democratic Re__, *Tunisia*, *Brazil*, *Germany*, *Australia*, *Cameroon*, *Israel*, *Turkey*, *United Kingdo…*, *Montenegro*, *Serbia*, *Argentina*, *Bosnia*, *Lithuania*, *Croatia*, *Italy*, *Poland*, *Dominican Rep…*, *Finland*, *Latvia*, *Bosnia & Herz…*, *Sweden*, *Ukraine*, *Austria*, *Puerto Rico*, *Senegal*, *Slovenia*, *Greece*, *Democratic Re…*, *Mali*, *Bahamas* and *Egypt*

Bosnia es el único pais que aparece dos veces escrito por lo que hay que corregirlo.

```
NBA$NBA_COUNTRY[NBA$NBA_COUNTRY == "BOSNIA & HERZ..."] <- "BOSNIA"
```

Ahora tenemos todos los datos de países limpios.

El siguiente paso es eliminar las variables inútiles en nuestro modelo como son: NBA$PLAYER

```
NBA$PLAYER <- NULL
```

Tratamiento de Dummy

Tenemos dos dummies posibles: el país de origen NBA$NBA_COUNTRY y el equipo del jugador NBA$TM. Creamos dos tablas con valores 1 y 0 para crear las dummies.

DUMMY PAIS (MUESTRA)

.data_Canada	.data_Cameroon	.data_Brazil	.data_Bosnia
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

DUMMY EQUIPO (MUESTRA)

.data_DEN	.data_DAL	.data_CLE	.data_CHO
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	1	0	0

Hechas las tablas dummies tenemos que hacer lo siguiente:

1. Eliminar `NBA$COUNTRY` porque está formada por caracteres.
2. Eliminar `NBA$.DATA`
3. Eliminar `NBA$TM` al ser también de tipo carácter.
4. Eliminar una variable por cada tabla dummy para evitar la dependencia: `NBA$.DATA_ARGENTINA` y `NBA$.DATA_BOS`

```
NBA$NBA_COUNTRY <- NULL
NBA$.DATA <- NULL
NBA$.DATA <- NULL
NBA$TM <- NULL
NBA$.DATA_ARGENTINA <- NULL
NBA$.DATA_BOS <- NULL
```

Crear "Train Set" y "Test Set"

En este paso vamos a dividir el data set `NBA` en dos para en una parte crear el modelo y en otra para probarlo.

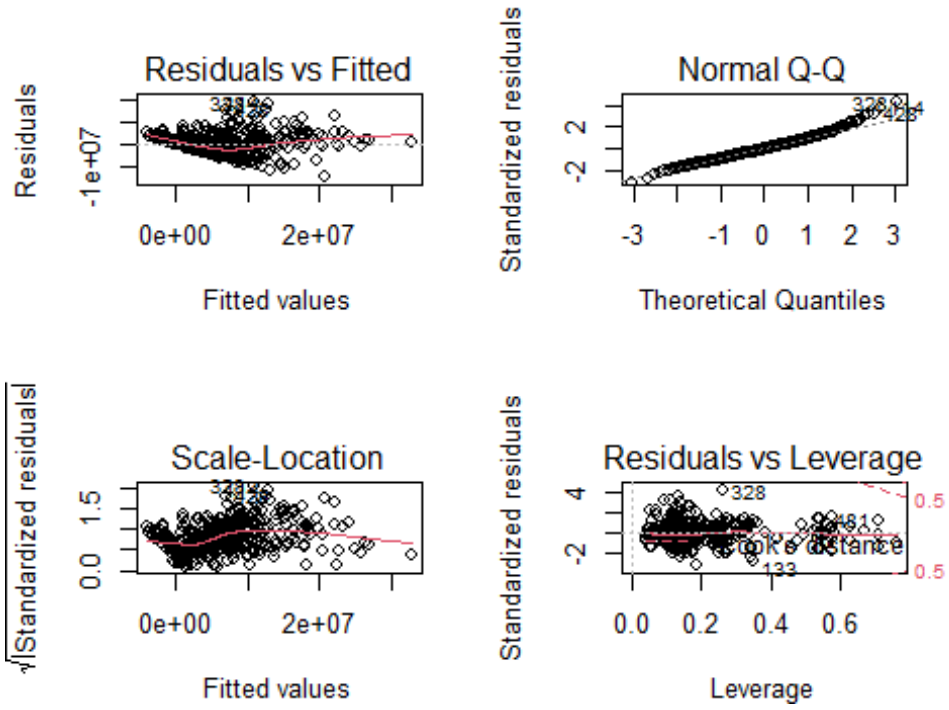
El *train set* tendrá 436 observaciones y el *test set* tendrá 50 observaciones.

```
SET.SEED(1234)
ROW <- SAMPLE(NROW(NBA))
TRAINSET <- NBA[ROW[1:436],]
VIEW(TRAINSET)

TESTSET <- NBA[ROW[437:485],]
VIEW(TESTSET)

VIEW(NBA)
```

Modelo sin normalidad

El primer modelo predictivo que vamos a crear contiene todas las variables del data set sin ninguna transformación

```
MODELOSSINNORMALIDAD <- LM(SALARY ~. ,DATA = TRAINSET)
```




```
GVMODEL <- GVLMA(MODELOSSINNORMALIDAD)
SUMMARY(GVMODEL)
```

	Value	p-value	Decision
Global Stat	123.073	0.000e+00	Assumptions NOT satisfied!
Skewness	27.210	1.826e-07	Assumptions NOT satisfied!
Kurtosis	47.528	5.421e-12	Assumptions NOT satisfied!
Link Function	46.040	1.159e-11	Assumptions NOT satisfied!
Heteroscedasticity	2.295	1.298e-01	Assumptions acceptable.

No hay normalidad, además la forma de curvilínea de los gráficos nos indica que la distribución del salario cambia en función de la cantidad (tiene dos comportamientos)

Transformación de variables y tratamiento de outliers

Para corregir nuestros datos necesitamos volver a las gráficas anteriores y ver lo que nos dice cada uno para poder tomar decisiones.

El gráfico de los Residuals vs Fitted nos dice que no hay una tendencia lineal entre las observaciones por la curva que hace la linea ajustada.

En el QQ-normal sí que podemos pensar en normalidad, pero hay outlier.

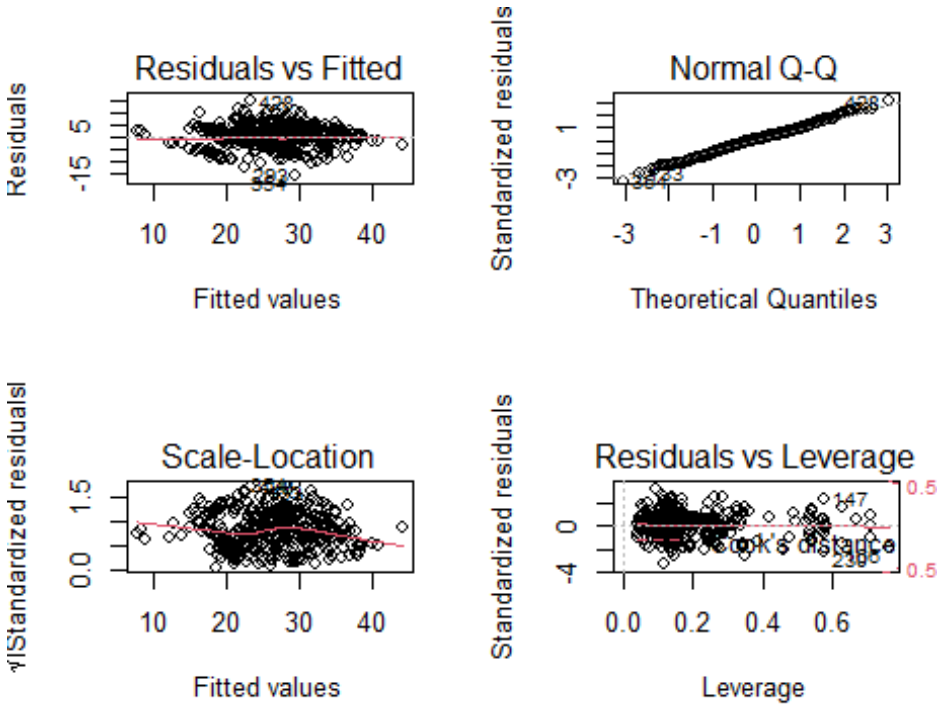
En el Scale-Location plot, que muestra si los residuos están igualmente distribuidos a lo largo de los predictores, se ve que en la parte la izquierda hay una concentración de puntos con una tendencia curvilínea por lo que nos indica que no hay homocedasticidad

Y en el Residuals vs Leverage, se utiliza para identificar los outliers que son influyentes en la regresión y que habría que eliminar. En nuestro caso ningún valor supera la distancia de 0.5 de Cook. Sin embargo, hay datos que están en extremos en los cuatro gráficos: 242,387, 114, 175 y 387.

```
powerTransform(NBA$Salary)

## Estimated transformation parameter
## NBA$Salary
##  0.2146177

summary(powerTransform(NBA$Salary))

## bcPower Transformation to Normality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## NBA$Salary    0.2146       0.21      0.1604      0.2689
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 66.83997  1 3.3307e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 528.8981  1 < 2.22e-16
```

Modelo con normalidad

Aplicando la transformación por la potencia (0.2146177) de la variable Salary, tenemos un nuevo modelo.



	Value	p-value	Decision
Global Stat	2.644402	0.6190	Assumptions acceptable.
Skewness	0.005965	0.9384	Assumptions acceptable.
Kurtosis	1.154845	0.2825	Assumptions acceptable.
Link Function	0.008319	0.9273	Assumptions acceptable.
Heteroscedasticity	1.475274	0.2245	Assumptions acceptable.

Una vez que hemos obtenido mediante una transformación la aceptación de linealidad del modelo, podemos mejorarlo en base al criterio Akaike que usa un trade-off entre la bondad de ajuste del modelo y la complejidad del modelo, por lo que podemos tener un modelo más resumido sin perder mucha información.

```
modeloConNormalidad <- lm(formula = Salary^0.2146177 ~ NBA_DraftNumber + Age + G +
 MP +
    PER + `TS%` + FTr + `ORB%` + `DRB%` + `TRB%` + `AST%` + `TOV%` +
    `USG%` + `WS/48` + OBPM + BPM + .data_Australia + .data_Cameroon +
    `.data_Democratic Re...` + .data_France + .data_Russia +
    .data_Senegal + .data_Turkey + `.data_United Kingdo...` +
```

```
    .DATA_ATL + .DATA_BRK + .DATA_CHI + .DATA_CHO + .DATA_CLE +
    .DATA_DEN + .DATA_DET + .DATA_IND + .DATA_LAC + .DATA_MEM +
    .DATA_MIA + .DATA_MIL + .DATA_MIN + .DATA_NYK + .DATA_OKC +
    .DATA_ORL + .DATA_PHO + .DATA_POR + .DATA_SAC + .DATA_TOR +
    .DATA_TOT + .DATA_WAS, DATA = TRAINSET)


GVMODEL <- GVLMA(MODELOCONNORMALIDAD)
SUMMARY(GVMODEL)
```

Ahora si aceptamos linealidad en nuestro modelo y podemos predecir salarios en función de un *input* dado.

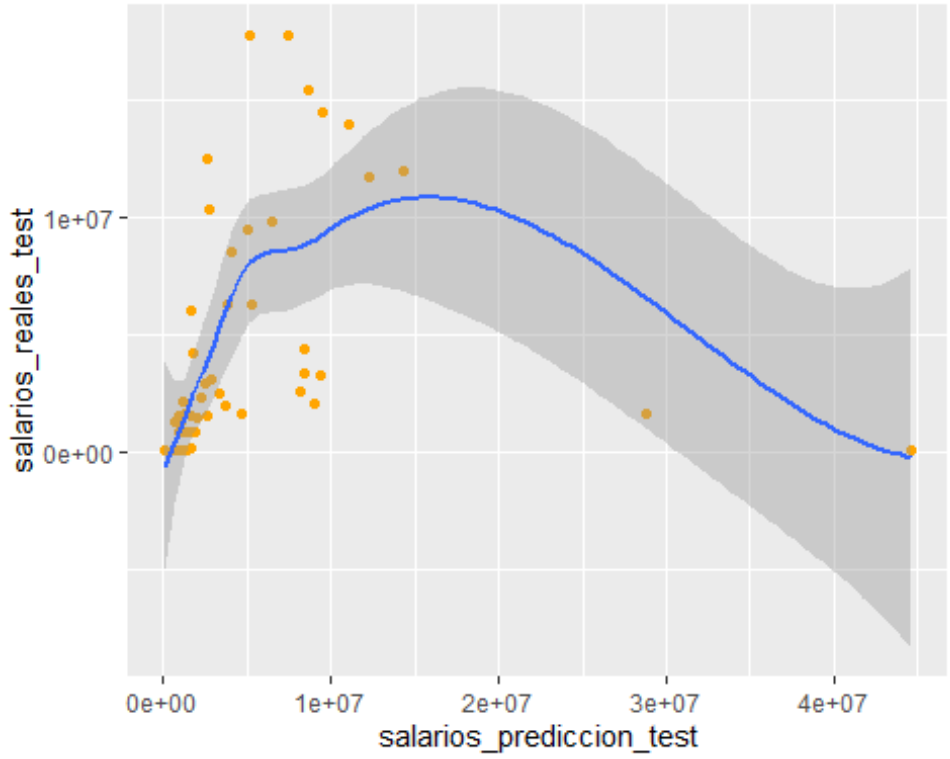
Cross validation Test

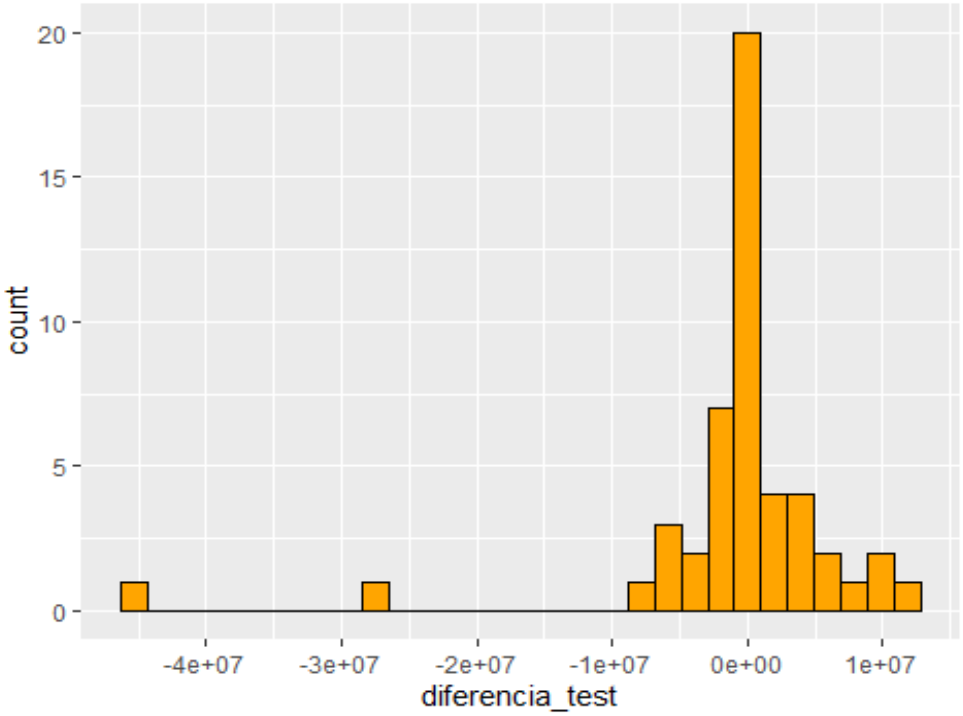
Modelo con normalidad

Antes de dar al modelo como válido, conviene probarlo en un conjuto de datos a los que no ha tenido acceso `TRAINSET`.

Para esto, aplicamos el modelo al `TRAINSET` y el resultado lo elevamos a la inversa de 0.21 debido a la anterior transformación de la varible Y. Luego le restamos los salarios reales para ver la precisión del modelo.

```
SALARIOS_PREDICCION_TEST <- PREDICT(MODELOCONNORMALIDAD,NEWDATA =TESTSET)^(1/0.2146)
```

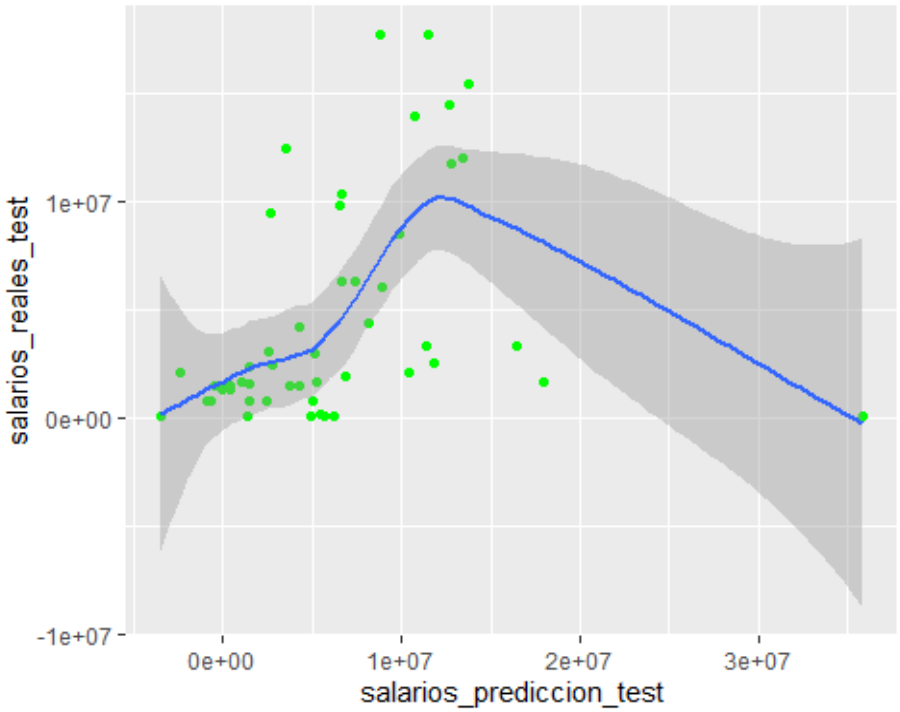


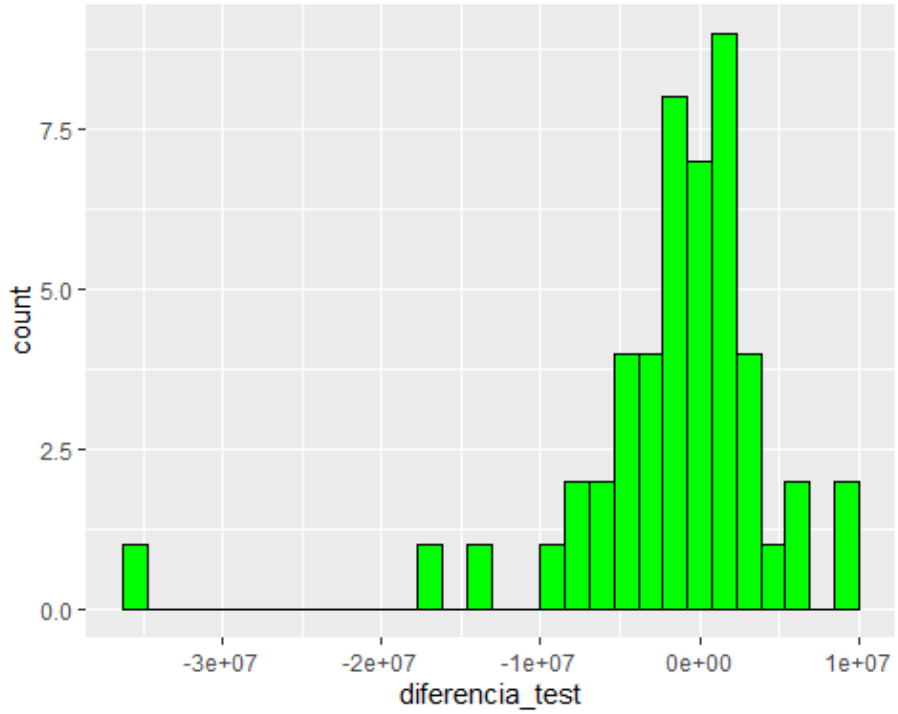


En este gráfico se muestra un histograma de la frecuencia de la diferencia de salario entre lo previsto por el modelo y el salario real.

Modelo sin normalidad

```
SALARIOS_PREDICCION_TEST <- PREDICT(MODELOSINNORMALIDAD,NEWDATA =TESTSET)
```





En este gráfico se muestra un histograma de la frecuencia de la diferencia de salario entre lo previsto por el modelo y el salario real.

El modelo corregido para la normalidad tiene más precisión incluso con menos variables debido a que la varianza de la distribución de diferencias es más baja. Ambos tienen la misma media.

Observamos una cierta relación, pero no suficiente para predecir y poner un salario a un jugador usando todas las variables del modelo.

```
AIC(modeloConNormalidad, modelosSinNormalidad)

##                       DF      AIC
## modeloConNormalidad  48  2636.143
## modelosSinNormalidad 98 14743.231

BIC(modeloConNormalidad, modelosSinNormalidad)

##                       DF      BIC
## modeloConNormalidad  48  2831.649
## modelosSinNormalidad 98 15142.389
```

Conclusión

Como conclusión decir que ningún modelo es suficientemente predictivo. Esto se debe a la complejidad que tiene la NBA en cuanto a salarios: por un lado, los equipos tienen límite salarial por lo que esto tiene bastante influencia en el tope salarial de un jugador que gana menos que lo que indica su rendimiento. Por otro lado, la situación de cada jugador es única: lesiones, jugadores retirados que siguen cobrando del equipo o que su anterior temporada haya sido la mejor y el actual salario no refleje la realidad.

Una variable para ver esto es en `NBA$MP` (minutos jugados), donde se aprecia que hay muchos jugadores con menos de 500 minutos jugados cobrando más dinero que la media (posiblemente por las razones anteriores).

La realidad que hay que aplicar al modelo de negocio es poner en contexto para qué quieres el jugador: un jugador que juegue muchos minutos y sea fundamental o no. O dividir los jugadores en defensivos u ofensivos y aplicar un modelo u oro en función del fichaje que quieras hacer. De aquí se haría un modelo clúster que explique mejor el salario para un tipo de jugador.