Una estimacion de los salarios de la NBA

Adrian Gonzalez Retamosa

10/28/2020

ESQUEMA

1. Descripcion del DataSet
2. Tratamiento del DataSet
3. Regresion global
4. Analisis de los regresores estadisticamente significativos
5. Metodo de seleccion de variables
6. Eleccion del modelo
7. Analisis de los residuos del modelo
8. Predicciones

1. Descripcion del DataSet

Las variables a trabjar son:

Player - Nombre jugador

Salary - Salario del jugador

NBA_Country - Pais origen jugador

NBA_DraftNumber - Numero en el Draft del jugador

Age - años

Tm - Equipo del jugador

G - Partidos jugados

MPMP = Minutos jugados

PER - Índice de eficiencia del jugador Una medida de la producción por minuto estandarizada de manera que el promedio de la liga es 15.

TSp - Porcentaje de tiros reales Una medida de la eficiencia de los tiros que tiene en cuenta tiros de campo de 2 puntos, tiros de campo de 3 puntos y tiros libres.

3PAr - Tasa de intentos de 3 puntos Porcentaje de intentos de FG desde el rango de 3 puntos

FTr - Tasa de intentos de tiros libres Número de intentos de FT por intento de tiro libre

ORBp - Porcentaje de rebote ofensivo Una estimación del porcentaje de rebotes ofensivos disponibles que un jugador agarró mientras estaba en la cancha.

DRBp - Porcentaje de rebote defensivo Una estimación del porcentaje de rebotes defensivos disponibles que un jugador agarró mientras estaba en la cancha.

TRBp - Porcentaje de rebote total Una estimación del porcentaje de rebotes disponibles que un jugador agarró mientras estaba en la cancha.

ASTp - Porcentaje de asistencia Una estimación del porcentaje de goles de campo de un compañero de equipo que un jugador ayudó mientras estaba en la cancha.

STLp - Porcentaje de robo Una estimación del porcentaje de posesiones del oponente que terminan con un robo del jugador mientras estaba en la cancha.

BLKp - Porcentaje de bloqueo Una estimación del porcentaje de intentos de gol de campo de dos puntos del oponente bloqueados por el jugador mientras estaba en el suelo.

TOVp - Porcentaje de rotación Una estimación de las pérdidas de balón cometidas por cada 100 jugadas.

USGp - Porcentaje de uso Una estimación del porcentaje de jugadas de equipo utilizadas por un jugador mientras estaba en la cancha.

OWS - Offensive Win Shares Una estimación del número de victorias aportadas por un jugador debido a su infracción.

DWS - Defensive Win Shares Una estimación del número de victorias aportadas por un jugador debido a su defensa.

WS - Win Shares Una estimación del número de victorias aportadas por un jugador.

WS / 48 - Acciones de ganancias por 48 minutos Una estimación del número de victorias aportadas por un jugador por 48 minutos (el promedio de la liga es aproximadamente .100)

OBPM - Offensive Box Plus / Minus Una estimación de la puntuación de la caja de los puntos ofensivos por cada 100 posesiones que un jugador contribuyó por encima de un jugador promedio de la liga, traducido a un equipo promedio.

DBPM - Defensive Box Plus / Minus Una estimación de la puntuación de caja de los puntos defensivos por cada 100 posesiones que un jugador contribuyó por encima de un jugador promedio de la liga, traducido a un equipo promedio.

BPM - Box Plus / Minus Una estimación de la puntuación de caja de los puntos por cada 100 posesiones que un jugador contribuyó por encima de un jugador promedio de la liga, traducido a un equipo promedio.

VORP - Valor sobre el jugador de reemplazo Una estimación de la puntuación de los puntos por cada 100 posesiones del EQUIPO que un jugador contribuyó por encima de un jugador de nivel de reemplazo (-2.0), traducido a un equipo promedio y prorrateado a una temporada de 82 juegos

El objetivo del trabajo es:

La construccion de un modelo a traves de los algoritmos de eleccion de variables con el fin de construir la prediccion mas acertada posible.

2. Carga y tratamiento del DataSet

Una vez que tenemos cargado el DataSet vamos a cambiar el nombre de sus columnas para verlo de una forma mas clara, omitir las observaciones con algun dato NaN y eliminar las observaciones repetidadas.

Por otro lado, eliminaremos de nuestro DataFrame las variales Player, NBA_Country, Tm ya que no son significativas estadisticamente

```
##  [1] "Salary"          "NBA_DraftNumber" "Age"             "G"
##  [5] "MP"              "PER"             "TSp"             "3PAr"
##  [9] "FTr"             "ORBp"            "DRBp"            "TRBp"
## [13] "ASTp"            "STLp"            "BLKp"            "TOVp"
## [17] "USGp"            "OWS"             "DWS"             "WS"
## [21] "WSd48"           "OBPM"            "DBPM"            "BPM"
## [25] "VORP"
```

3. Regresion global

Para realizar un primer acercamiento a nuestro datos vamos a correr dos regresion, ambas ellas incluyendo todas las variables, con la diferencia de poner nuestra variable dependiente en terminos de logaritmos o no.

3.1. Regresion en terminos NO logaritmicos

```
m1 <- lm(Salary ~ . , nba)
summary(m1)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = nba)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15354874  -3000738   -385019   2114291  21684587
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2257549    4909948  -0.460    0.646
## NBA_DraftNumber   -60665      12807  -4.737 2.90e-06 ***
## Age               517569      56727   9.124  < 2e-16 ***
## G                 -154422      25036  -6.168 1.53e-09 ***
## MP                   5670       1087   5.217 2.76e-07 ***
## PER               -353478     281293  -1.257    0.210
## TSp              -2184162    5160878  -0.423    0.672
## `3PAr`           -3441403    2364332  -1.456    0.146
## FTr               -151800     889330  -0.171    0.865
## ORBp             -1067597     906856  -1.177    0.240
## DRBp              -868491     897793  -0.967    0.334
## TRBp              2033598    1797209   1.132    0.258
## ASTp               -19939      47665  -0.418    0.676
## STLp              -194112     423061  -0.459    0.647
## BLKp               110027     318948   0.345    0.730
## TOVp                 4106      52808   0.078    0.938
## USGp               167442     105277   1.590    0.112
## OWS              -1323604    4518535  -0.293    0.770
## DWS              -1790690    4544902  -0.394    0.694
## WS                1877630    4523606   0.415    0.678
## WSd48             1893689   11822704   0.160    0.873
## OBPM              1935687    4772501   0.406    0.685
## DBPM              1494922    4687825   0.319    0.750
```

```
## BPM             -1353095    4705755  -0.288    0.774
## VORP              632085     635260   0.995    0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5114000 on 456 degrees of freedom
## Multiple R-squared:  0.5469, Adjusted R-squared:  0.5231
## F-statistic: 22.94 on 24 and 456 DF,  p-value: < 2.2e-16
```

Observamos que hay muchas variables que no son significativas estadisticamente

Los criterios de informacion de este modelo son:

```
AIC(m1)
```

```
## [1] 16251.85
```

```
BIC(m1)
```

```
## [1] 16360.43
```

3.2. Regresion en terminos logaritmicos

```
##
## Call:
## lm(formula = log(Salary, base = 10) ~ ., data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58376 -0.23901  0.01466  0.26193  1.45382
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.645e+00  4.231e-01  10.979  < 2e-16 ***
## NBA_DraftNumber -9.577e-03  1.104e-03  -8.678  < 2e-16 ***
## Age              4.317e-02  4.888e-03   8.832  < 2e-16 ***
## G               -2.206e-03  2.157e-03  -1.022 0.307144
## MP               4.562e-04  9.365e-05   4.872 1.53e-06 ***
## PER             -7.977e-02  2.424e-02  -3.291 0.001076 **
## TSp              1.378e+00  4.447e-01   3.098 0.002069 **
## `3PAr`          -1.772e-01  2.037e-01  -0.870 0.384961
## FTr             -1.257e-01  7.664e-02  -1.640 0.101649
## ORBp            -3.321e-02  7.815e-02  -0.425 0.671047
## DRBp            -2.743e-02  7.736e-02  -0.355 0.723123
## TRBp             8.961e-02  1.549e-01   0.579 0.563116
## ASTp             6.827e-03  4.107e-03   1.662 0.097192 .
## STLp             1.563e-02  3.646e-02   0.429 0.668281
## BLKp             4.643e-03  2.748e-02   0.169 0.865925
## TOVp            -9.124e-03  4.551e-03  -2.005 0.045551 *
## USGp             3.165e-02  9.072e-03   3.489 0.000532 ***
## OWS             -7.375e-02  3.894e-01  -0.189 0.849850
## DWS             -1.810e-01  3.916e-01  -0.462 0.644159
```

```
## WS               7.616e-02  3.898e-01   0.195 0.845182
## WSd48            2.679e+00  1.019e+00   2.630 0.008827 **
## OBPM            -2.433e-01  4.113e-01  -0.592 0.554450
## DBPM            -2.185e-01  4.040e-01  -0.541 0.588888
## BPM              2.530e-01  4.055e-01   0.624 0.533019
## VORP             1.191e-02  5.474e-02   0.218 0.827876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4407 on 456 degrees of freedom
## Multiple R-squared:  0.5586, Adjusted R-squared:  0.5354
## F-statistic: 24.04 on 24 and 456 DF,  p-value: < 2.2e-16
```

Tambien comprobamos que hay variables no significaticvas pero este modelo presenta un mayor R2 adj

Los criterios de informacion de este modelo son:

```
AIC(m1_ln)
```

```
## [1] 603.0783
```

```
BIC(m1_ln)
```

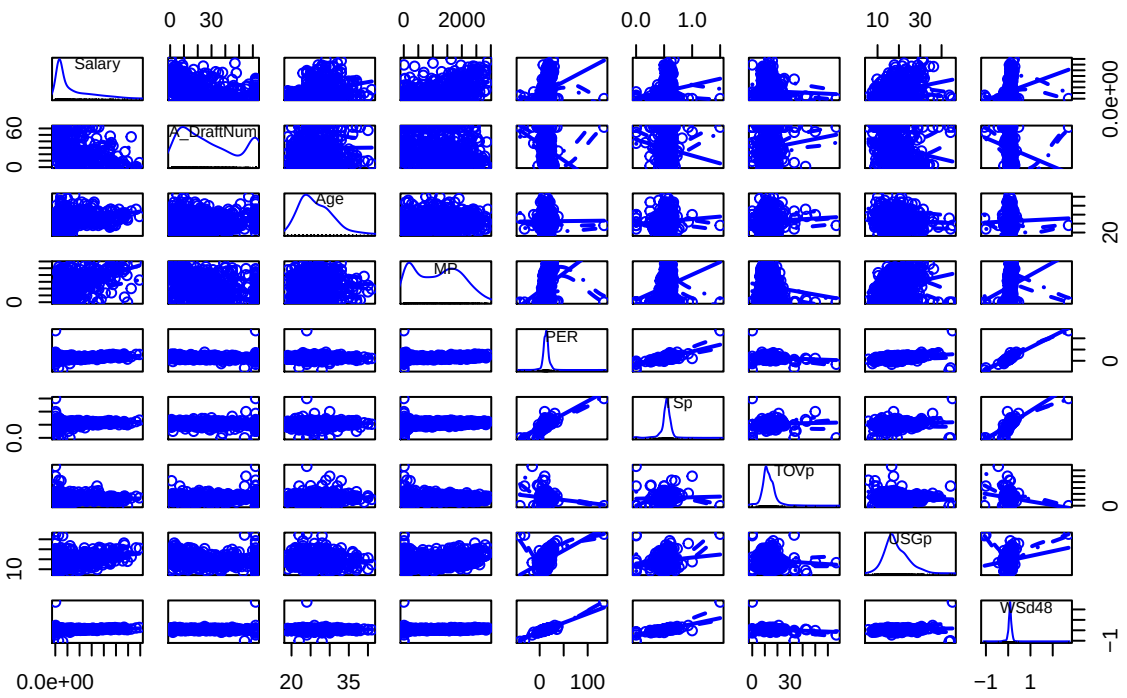
```
## [1] 711.6508
```

Analizando los dos modelos y sus criterios de informacion vamos a tomar la decision de poner la variable dependiente en terminos Logaritmicos.

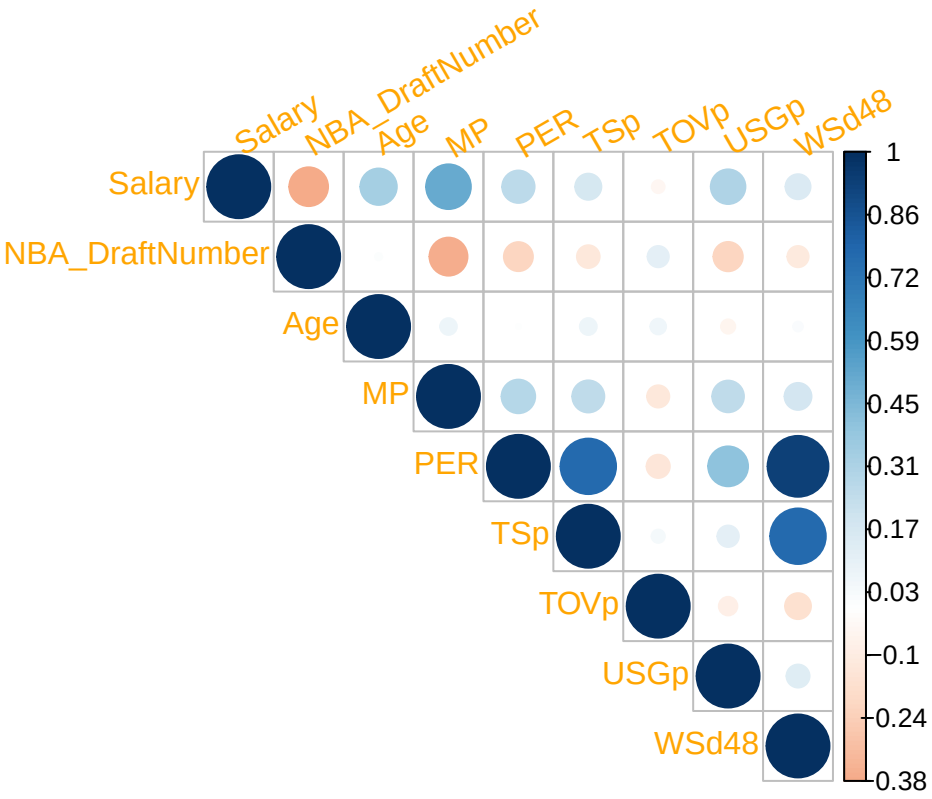
4. Analisis de los regresores estadisticamente significativos

En este apartado nos vamos a quedar solo con los regresores signioficativos para estudiar la relacion entre ellos y con la variable dependiente. Para ello construiremos un nuevo DataFramre que solo contenga estas variables.

A continuacion presentamos un grafico de distribucion de cada una de las variables, asi como la representacion de la dipersion entre cada una de ellas. En este grafico observamos como, por ejemplo, las variables NBA_DraftNumber y MP siguen una distribucion binomial.



El siguiente grafico a representar es un mapa de calor que nos indica las correlaciones que hay entre estas variables.En esta representacion observamos algo que era de esperar, ya que comprabamos una relacion negativa de la variable NBA_DraftNumber con todas las demas. Tambien nos podemos fijar en la alta correlacion que hay entre variables como TSp, PER y WSd48.



5. Metodo de seleccion de variables

En el siguiente apartado pondremos en practica tres algoritmos que nos indicaran que regresores deberemos incluir en nuestro modelo para que sea lo mas preciso posible. Esta eleccion la realizaremos a traves de sus criterios de informacion, el Cp y R2 Adj.

5.1. BEST SUBSET

Consiste en estimar todas las regresiones posibles con todas las combinaciones de los n regresores y asi encontrar aquel modelo.

```
## Subset selection object
## Call: regsubsets.formula(log(Salary, base = 10) ~ ., nba)
## 24 Variables  (and intercept)
##                  Forced in Forced out
## NBA_DraftNumber     FALSE      FALSE
## Age                 FALSE      FALSE
## G                   FALSE      FALSE
## MP                  FALSE      FALSE
## PER                 FALSE      FALSE
## TSp                 FALSE      FALSE
## `3PAr`              FALSE      FALSE
## FTr                 FALSE      FALSE
## ORBp                FALSE      FALSE
```

```
## DRBp                FALSE      FALSE
## TRBp                FALSE      FALSE
## ASTp                FALSE      FALSE
## STLp                FALSE      FALSE
## BLKp                FALSE      FALSE
## TOVp                FALSE      FALSE
## USGp                FALSE      FALSE
## OWS                 FALSE      FALSE
## DWS                 FALSE      FALSE
## WS                  FALSE      FALSE
## WSd48               FALSE      FALSE
## OBPM                FALSE      FALSE
## DBPM                FALSE      FALSE
## BPM                 FALSE      FALSE
## VORP                FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          NBA_DraftNumber Age G   MP  PER TSp `3PAr` FTr ORBp DRBp TRBp ASTp
## 1  ( 1 ) " "             " " " " "*" " " " " " " "    " " " "  " "  " "  " "
## 2  ( 1 ) "*"             " " " " "*" " " " " " " "    " " " "  " "  " "  " "
## 3  ( 1 ) "*"             "*" " " "*" " " " " " " "    " " " "  " "  " "  " "
## 4  ( 1 ) "*"             "*" " " "*" " " " " " " "    " " " "  "*"  " "  " "
## 5  ( 1 ) "*"             "*" " " "*" " " " " " " "    " " " "  "*"  " "  " "
## 6  ( 1 ) "*"             "*" " " "*" "*" "*" " " "    " " " "  "*"  " "  " "
## 7  ( 1 ) "*"             "*" " " "*" "*" " " "*" "    " " " "  "*"  " "  " "
## 8  ( 1 ) "*"             "*" " " "*" "*" "*" " " "    " " " "  "*"  " "  " "
##          STLp BLKp TOVp USGp OWS DWS WS  WSd48 OBPM DBPM BPM VORP
## 1  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  " "  " " " "
## 2  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  " "  " " " "
## 3  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  " "  " " " "
## 4  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  " "  " " " "
## 5  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  "*"  " " " "
## 6  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  " "  " "  " " " "
## 7  ( 1 ) " "  " "  " "  " "  " " " " " " " " "  "*"  " "  " " " "
## 8  ( 1 ) " "  " "  " "  "*"  " " " " " " " " "  " "  " "  "*" " "
```

```
reg_summary1$rss # Por la SRC el mejor modelo seria el (8)
```

```
## [1] 132.63823 113.04729  96.58245  94.07716  93.93373  93.25851  92.72742
## [8]  91.81947
```

```
reg_summary1$cp # Por el estadistico Cp el mejor modelo seria el (8)
```

```
## [1] 205.959135 107.084692  24.306590  13.406788  14.668251  13.191528  12.456919
## [8]   9.781848
```

```
reg_summary1$bic # Por el metodo bayesiano el mejor modelo seria el (4)
```

```
## [1] -186.7020 -257.3991 -326.9371 -333.4027 -327.9608 -325.2549 -321.8261
## [8] -320.3832
```

Observando los criterios de informacion el modelo sugerido por este metodo seria:

$\text{Log Salary} = \text{b1}\mathit{NBA\_DraftNumber} + \mathit{b2}\,\text{Age} + \text{b3}\mathit{MP} + \mathit{b4}\,\text{PER} + \text{b5}\,\mathit{TSp} + \mathit{b6}\,\text{DRBp} + \text{b7}\mathit{USGp} + \mathit{b8}\,\text{BPM} + \text{Ut}$

5.2. FORWARD STEPWISE

Este metodo consiste en empezar con un modelo que no incluye ningún regresor y se van añadiendo regresores de uno en uno. En cada etapa la variable que más mejora adicional aporta al modelo es incluida.

```
## Subset selection object
## Call: regsubsets.formula(log(Salary, base = 10) ~ ., nba, method = "forward")
## 24 Variables  (and intercept)
##                  Forced in Forced out
## NBA_DraftNumber     FALSE      FALSE
## Age                 FALSE      FALSE
## G                   FALSE      FALSE
## MP                  FALSE      FALSE
## PER                 FALSE      FALSE
## TSp                 FALSE      FALSE
## `3PAr`              FALSE      FALSE
## FTr                 FALSE      FALSE
## ORBp                FALSE      FALSE
## DRBp                FALSE      FALSE
## TRBp                FALSE      FALSE
## ASTp                FALSE      FALSE
## STLp                FALSE      FALSE
## BLKp                FALSE      FALSE
## TOVp                FALSE      FALSE
## USGp                FALSE      FALSE
## OWS                 FALSE      FALSE
## DWS                 FALSE      FALSE
## WS                  FALSE      FALSE
## WSd48               FALSE      FALSE
## OBPM                FALSE      FALSE
## DBPM                FALSE      FALSE
## BPM                 FALSE      FALSE
## VORP                FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          NBA_DraftNumber Age G   MP  PER TSp `3PAr` FTr ORBp DRBp TRBp ASTp
## 1  ( 1 ) " "             " " " " "*" " " " " " "    " " " "  " "  " "  " "
## 2  ( 1 ) "*"             " " " " "*" " " " " " "    " " " "  " "  " "  " "
## 3  ( 1 ) "*"             "*" " " "*" " " " " " "    " " " "  " "  " "  " "
## 4  ( 1 ) "*"             "*" " " "*" " " " " " "    " " " "  "*"  " "  " "
## 5  ( 1 ) "*"             "*" " " "*" " " " " " "    " " " "  "*"  " "  " "
## 6  ( 1 ) "*"             "*" " " "*" " " " " " "    " " " "  "*"  " "  " "
## 7  ( 1 ) "*"             "*" " " "*" " " " " " "    " " " "  "*"  " "  " "
## 8  ( 1 ) "*"             "*" " " "*" " " "*" " "    " " " "  "*"  " "  " "
##          STLp BLKp TOVp USGp OWS DWS WS  WSd48 OBPM DBPM BPM VORP
## 1  ( 1 ) " "  " "  " "  " "  " " " " " " " "  " "  " "  " " " "
## 2  ( 1 ) " "  " "  " "  " "  " " " " " " " "  " "  " "  " " " "
## 3  ( 1 ) " "  " "  " "  " "  " " " " " " " "  " "  " "  " " " "
## 4  ( 1 ) " "  " "  " "  " "  " " " " " " " "  " "  " "  " " " "
## 5  ( 1 ) " "  " "  " "  " "  " " " " " " " "  " "  "*"  " " " "
## 6  ( 1 ) " "  "*"  " "  " "  " " " " " " " "  " "  "*"  " " " "
## 7  ( 1 ) " "  "*"  "*"  " "  " " " " " " " "  " "  "*"  " " " "
## 8  ( 1 ) " "  "*"  "*"  " "  " " " " " " " "  " "  "*"  " " " "
```

```
reg_summary_fwd$rss # Por la SRC el mejor modelo seria el (8)
```

```
## [1] 132.63823 113.04729  96.58245  94.07716  93.93373  93.73599  93.56292
## [8]  93.37318
```

```
reg_summary_fwd$cp ## Por el estadistico Cp el mejor modelo seria el (8)
```

```
## [1] 205.95914 107.08469  24.30659  13.40679  14.66825  15.65007  16.75895
## [8]  17.78194
```

```
reg_summary_fwd$bic # Por el metodo bayesiano el mejor modelo seria el (4)
```

```
## [1] -186.7020 -257.3991 -326.9371 -333.4027 -327.9608 -322.7985 -317.5116
## [8] -312.3121
```

Con este algoritmo elegiriamos el mismo modelo que por el metodo Best Subset

5.3. BACKWARD STEPWISE

Empieza con un modelo que incluye todos los regresores y se van eliminando regresores de uno en uno. En cada etapa la variable que menos mejora adicional aporta al modelo es excluida.

Con este metodo el el modelo quedaria de la siguiente forma:

$\text{Log Salary} = \text{b1}\,NBA\_DraftNumber + b2\,\text{Age} + \text{b3}\,MP + b4\,\text{PER} + \text{b5}\,TSp + b6\,\text{DRBp} + \text{b7}\,USGp + b8\,\text{BPM} + \text{b9}\,TRBp + b10\,\text{ASTp} + \text{b11}\,TOVp + b12\,\text{DWS} + \text{b13}\,WSd48 + b14\,\text{OBPM} + \text{Ut}$

6. Eleccion del modelo

Una vez que hemos utilizado los 3 algoritmos para la seleccion de variables pasamos a comprobar cual de los 2 modelos propuestos presenta unos mejores criterios de informacion y un mejor R2 adj

6.1. Modelo obtenido por el metodo Best Subset

$\text{Log Salary} = \text{b1}\,NBA\_DraftNumber + b2\,\text{Age} + \text{b3}\,MP + b4\,\text{PER} + \text{b5}\,TSp + b6\,\text{DRBp} + \text{b7}\,USGp + b8\,\text{BPM} + \text{Ut}$

Este modelo presenta un R2 adj de 53,34% y sus criterios de informacion son :

```
AIC(model8)
```

```
## [1] 588.4622
```

```
BIC(model8)
```

```
## [1] 630.2209
```

6.2. Modelo obtenido por el metodo Backward Stepwise

Log Salary = b1$NBA_DraftNumber$ + $b2$Age + b3MP + $b4$PER + b5TSp + $b6$DRBp + b7$USGp$ + $b8$BPM + b9$TRBp$ + $b10$ASTp + b11$TOVp$ + $b12$DWS + b13$WSd48$ + $b14$OBPM + Ut

Este modelo presenta un R2 adj de 54,11% y sus criterios de informacion son :

```
AIC(m1_b)
```

```
## [1] 586.5935
```

```
BIC(m1_b)
```

```
## [1] 649.2315
```

Despues de analizar ambos modelos nos vamos a quedar con 'm1_b' ya que es el que mejor R2 adj presenta y tiene un mejor valor para el criterio de informacion de Akaike. Sera a este modelo al que le realizaremos una analisis a sus residuos para la posterior prediccion.

```
m1_b <- lm(formula = log(Salary, base = 10) ~ NBA_DraftNumber + Age +
     MP + PER + TSp + TRBp + ASTp + TOVp + USGp + DWS + WSd48 +
     OBPM + BPM, data = nba)
summary(m1_b)
```

```
##
## Call:
## lm(formula = log(Salary, base = 10) ~ NBA_DraftNumber + Age +
##     MP + PER + TSp + TRBp + ASTp + TOVp + USGp + DWS + WSd48 +
##     OBPM + BPM, data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55411 -0.23916  0.00956  0.27596  1.47282
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.529e+00  2.758e-01  16.419  < 2e-16 ***
## NBA_DraftNumber -9.866e-03  1.060e-03  -9.306  < 2e-16 ***
## Age              4.380e-02  4.782e-03   9.159  < 2e-16 ***
## MP               3.902e-04  5.816e-05   6.709 5.69e-11 ***
## PER             -6.533e-02  1.632e-02  -4.004 7.24e-05 ***
## TSp              1.202e+00  3.685e-01   3.263 0.001185 **
## TRBp             2.759e-02  6.976e-03   3.955 8.84e-05 ***
## ASTp             6.893e-03  3.435e-03   2.007 0.045368 *
## TOVp            -9.086e-03  4.128e-03  -2.201 0.028219 *
## USGp             3.030e-02  8.513e-03   3.559 0.000411 ***
## DWS             -8.839e-02  4.791e-02  -1.845 0.065690 .
## WSd48            2.271e+00  8.391e-01   2.706 0.007058 **
## OBPM            -4.025e-02  2.508e-02  -1.605 0.109175
## BPM              4.230e-02  1.449e-02   2.919 0.003681 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

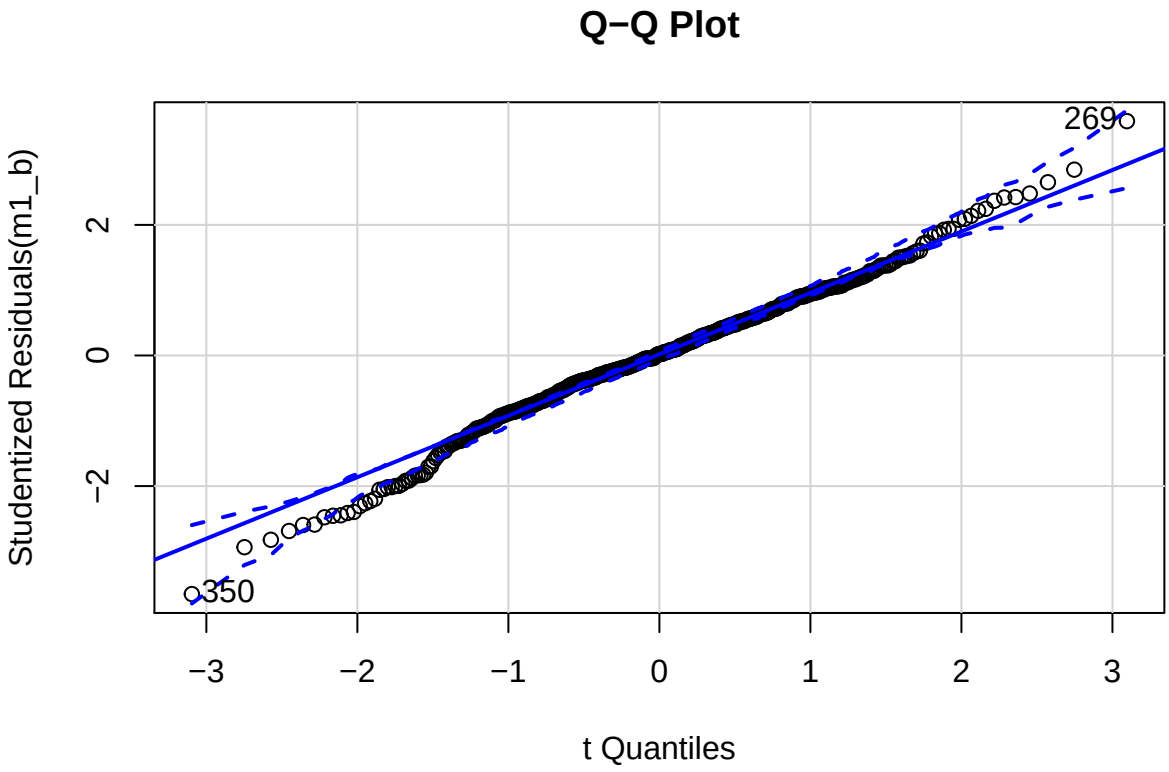
```
##
## Residual standard error: 0.438 on 467 degrees of freedom
## Multiple R-squared:  0.5535, Adjusted R-squared:  0.5411
## F-statistic: 44.53 on 13 and 467 DF,  p-value: < 2.2e-16
```

7. Analisis de los residuos del modelo

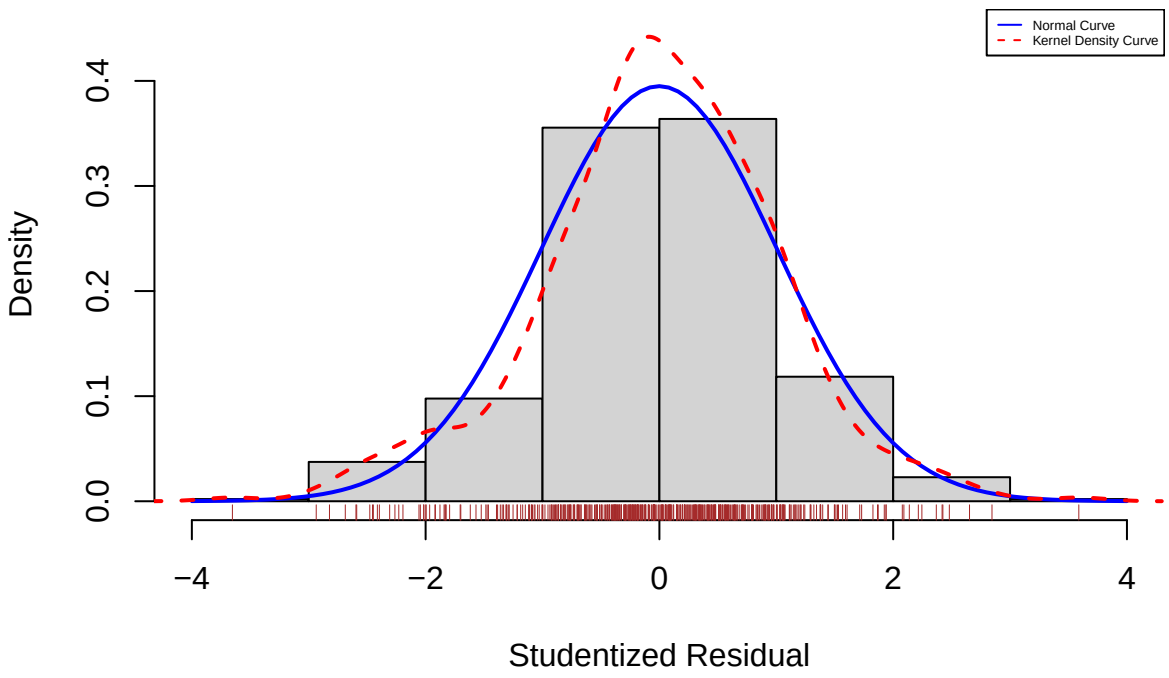
7.1. Normalidad

Con los siguientes dos graficos podemos observar cual el la distribucion de los residuos de nuestro modelo

```
## Warning in rlm.default(x, y, weights, method = method, wt.method = wt.method, :
## 'rlm' failed to converge in 20 steps
```




```
## [1] 269 350
```



Para confirma la normalidad de los residuos realizaremos el constraste de Shapiro-Wilk. Obtenemos un Pvalor = 0,004, inferior que el 5%, por lo que rechazamos la Hipotesis Nula de normalidad de los residuos.

```
shapiro.test(resid(m1_b))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(m1_b)
## W = 0.99091, p-value = 0.004684
```

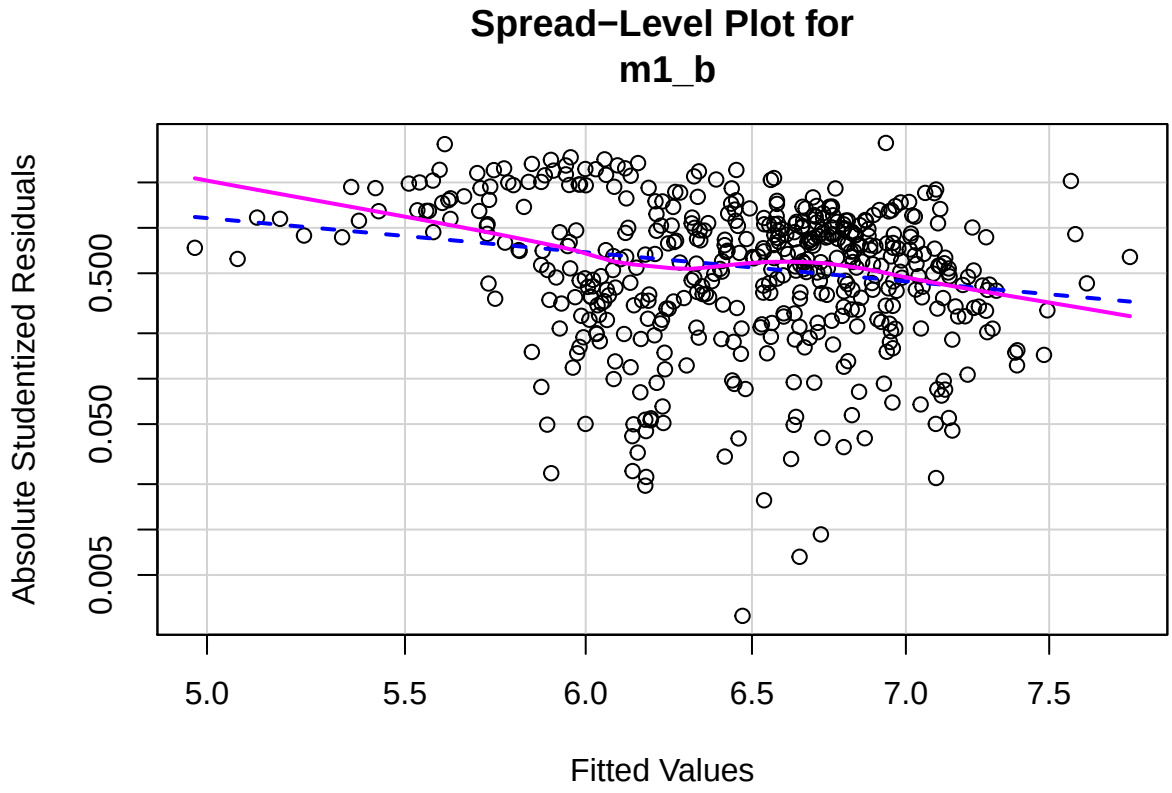
7.2. Homoscedasticidad

Con un Pvalor menor al 5% en el constraste de Breusch-Pagan rechazmos la hipotesis nula de que los resiudos tienen varinza constante

```
ncvTest(m1_b)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 52.5075, Df = 1, p = 4.2861e-13
```

```r
spreadLevelPlot(m1_b)
```




```
##
## Suggested power transformation:  3.867926
```

7.3. Contraste global

Con la realizacion del test global observamos que nuestro modelo pasa el test de simetria y el de homoscedasticidad, a pesar de lo que nos digo el constraste de Breusch-Pagan, pero en general es un modelo con problemas estadistico en su utilizacion para explicar la variable dependiente.

```r
gvlma(m1_b)
```

```
##
## Call:
## lm(formula = log(Salary, base = 10) ~ NBA_DraftNumber + Age +
##     MP + PER + TSp + TRBp + ASTp + TOVp + USGp + DWS + WSd48 +
##     OBPM + BPM, data = nba)
##
## Coefficients:
##     (Intercept)  NBA_DraftNumber              Age               MP
##       4.5286117       -0.0098659        0.0437968        0.0003902
##             PER              TSp             TRBp             ASTp
```

```
##      -0.0653310        1.2023677        0.0275924        0.0068926
##            TOVp             USGp              DWS             WSd48
##      -0.0090857        0.0302956       -0.0883894        2.2707356
##            OBPM              BPM
##      -0.0402456        0.0423008
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = m1_b)
##
##                       Value   p-value                   Decision
## Global Stat        32.72560 1.359e-06 Assumptions NOT satisfied!
## Skewness            3.57633 5.861e-02    Assumptions acceptable.
## Kurtosis            8.37409 3.806e-03 Assumptions NOT satisfied!
## Link Function      20.68987 5.400e-06 Assumptions NOT satisfied!
## Heteroscedasticity  0.08531 7.702e-01    Assumptions acceptable.
```

8. Preddiccion

En el ultimo apartado observaremos como de fiable es nuestro modelo a la hora de la prediccion.

Lo primero calcularemos la desviacion tipica de los residuos de nuestro modelo ya que es un buen indicador a la hora de predecir y observamos que es relativamente baja

```
sd(resid(m1_b))
```

```
## [1] 0.4320049
```

Acontinuacion construiremos una tabla con tres columnas, el valor real del salario, el valor que predice nuestro modelo y la diferencia de ambos en valor absoluto

```
## # A tibble: 481 x 3
##    LogSal predLogSalary difSalary
##     <dbl>         <dbl>     <dbl>
##  1   5.91          5.90    0.0102
##  2   6.54          6.82    0.277
##  3   7.09          7.20    0.112
##  4   6.51          6.35    0.158
##  5   6.49          6.45    0.0402
##  6   6.12          6.42    0.306
##  7   4.87          5.94    1.07
##  8   4.66          5.51    0.847
##  9   7.08          7.10    0.0217
## 10   6.16          6.18    0.0195
## # ... with 471 more rows
```

Para finalizar este trabajo calcularemos la media y la desviacion tipica de la diferencia del salario real y del salario estimado para comprobar en cuanto se equivoca nuestro modelo

```r
mean(pre_salary$difSalary)
```

```
## [1] 0.3308638
```

```r
sd(pre_salary$difSalary)
```

```
## [1] 0.2773613
```