Proyecto NBA

Alejandro Carrillo Vera

10/10/2019

```
#Cargamos la base de datos de los jugadores de la NBA
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(TeachingDemos)
library (leaps)
library(ISLR)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```r
nba <- read_csv("nba.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Player = col_character(),
##   NBA_Country = col_character(),
##   Tm = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```r
#Vamos a renombrar todas las variables para estudiarlas más fácil
nba<- rename(nba,Jugador=Player)
nba<- rename(nba,Salario=Salary)
nba<- rename(nba,Pais=NBA_Country)
nba<- rename(nba,Draft=NBA_DraftNumber)
nba<- rename(nba,Edad=Age)
nba<- rename(nba,Equipo=Tm)
nba<- rename(nba,Partidos=G)
nba<- rename(nba,Minutos_jugados=MP)
nba<- rename(nba,Eficiencia=PER)
nba<- rename(nba,Exito="TS%")
nba<- rename(nba,Triples="3PAr")
nba<- rename(nba,Tiros_Libres=FTr)
nba<- rename(nba,Reb_Atq="ORB%")
nba<- rename(nba,Reb_Def="DRB%")
nba<- rename(nba,Reb_Total="TRB%")
nba<- rename(nba,Asistencias="AST%")
nba<- rename(nba,Robos="STL%")
nba<- rename(nba,Bloqueos="BLK%")
nba<- rename(nba,Porc_perdidas="TOV%")
nba<- rename(nba,Porc_participacion="USG%")
nba<- rename(nba,Buen_Atq=OWS)
nba<- rename(nba,Buen_Def=DWS)
nba<- rename(nba,Buen_Tot=WS)
nba<- rename(nba,Buen_Tot_48="WS/48")
nba<- rename(nba,Calidad_Atq=OBPM)
nba<- rename(nba,Calidad_Def=DBPM)
#Omitimos los NAs
nba<- na.omit(nba)
#Tenemos que responder a la pregunta, ¿cuánta relación hay entre los datos de los jugadores y sus salar
#Para ello asignamos como variable dependiente al salario
vec_y<- nba$Salario
mat_x<- cbind(1,nba[,2:28])

head(vec_y)
```

```
## [1]   815615  3477600 12307692  3202217  3057240  1312611
```

```r
head.matrix(mat_x)
```

```
##   1  Salario    Pais Draft Edad Equipo Partidos Minutos_jugados Eficiencia
## 1 1   815615   China    43   22    HOU       16              87        0.6
## 2 1  3477600 Georgia    42   33    GSW       66             937       16.8
## 3 1 12307692     USA    19   36    SAC       59            1508       17.3
## 4 1  3202217     USA    13   22    CHI       24             656       14.6
## 5 1  3057240     USA    10   20    POR       62             979        8.2
## 6 1  1312611     USA    62   24    DAL       79            2238       11.5
##   Exito Triples Tiros_Libres Reb_Atq Reb_Def Reb_Total Asistencias Robos
## 1 0.303   0.593        0.370     6.5    16.8      11.7         1.5   1.1
## 2 0.608   0.004        0.337    11.0    25.0      18.5        15.4   1.9
## 3 0.529   0.193        0.140     7.0    23.8      15.0        14.9   1.4
## 4 0.499   0.346        0.301     1.4    14.4       7.7        18.6   1.8
## 5 0.487   0.387        0.146     4.9    18.3      11.7         7.3   0.8
## 6 0.543   0.489        0.141     1.3    11.3       6.1        13.3   1.4
##   Bloqueos Porc_perdidas Porc_participacion Buen_Atq Buen_Def Buen_Tot
## 1      6.8          18.2               19.5     -0.4      0.1     -0.2
## 2      1.3          19.3               17.2      1.7      1.4      3.1
## 3      0.6          12.5               27.6      0.3      1.1      1.4
## 4      0.5           9.7               29.5     -0.1      0.5      0.4
## 5      2.5          15.6               15.5     -0.4      1.2      0.8
## 6      0.3           9.1               17.0      1.6      1.6      3.1
##   Buen_Tot_48 Calidad_Atq Calidad_Def   BPM VORP
## 1      -0.121       -10.6         0.5 -10.1 -0.2
## 2       0.160        -0.6         1.3   0.8  0.7
## 3       0.046        -0.6        -1.3  -1.9  0.0
## 4       0.027        -0.7        -2.0  -2.6 -0.1
## 5       0.038        -3.7         0.9  -2.9 -0.2
## 6       0.067        -0.4        -0.5  -0.9  0.6
```

```r
#Vamos a realizar una selección a través del metodo de Forward Stepwise quitandole las variables Jugado
#ya que no las encuentro relevantes

regfit.full=regsubsets(Salario ~ .-Jugador-Pais,nba, method = "forward")
summary(regfit.full)
```

```
## Subset selection object
## Call: regsubsets.formula(Salario ~ . - Jugador - Pais, nba, method = "forward")
## 54 Variables  (and intercept)
##                    Forced in Forced out
## Draft                  FALSE      FALSE
## Edad                   FALSE      FALSE
## EquipoBOS              FALSE      FALSE
## EquipoBRK              FALSE      FALSE
## EquipoCHI              FALSE      FALSE
## EquipoCHO              FALSE      FALSE
## EquipoCLE              FALSE      FALSE
## EquipoDAL              FALSE      FALSE
## EquipoDEN              FALSE      FALSE
## EquipoDET              FALSE      FALSE
## EquipoGSW              FALSE      FALSE
## EquipoHOU              FALSE      FALSE
```

```
## EquipoIND              FALSE      FALSE
## EquipoLAC              FALSE      FALSE
## EquipoLAL              FALSE      FALSE
## EquipoMEM              FALSE      FALSE
## EquipoMIA              FALSE      FALSE
## EquipoMIL              FALSE      FALSE
## EquipoMIN              FALSE      FALSE
## EquipoNOP              FALSE      FALSE
## EquipoNYK              FALSE      FALSE
## EquipoOKC              FALSE      FALSE
## EquipoORL              FALSE      FALSE
## EquipoPHI              FALSE      FALSE
## EquipoPHO              FALSE      FALSE
## EquipoPOR              FALSE      FALSE
## EquipoSAC              FALSE      FALSE
## EquipoSAS              FALSE      FALSE
## EquipoTOR              FALSE      FALSE
## EquipoTOT              FALSE      FALSE
## EquipoUTA              FALSE      FALSE
## EquipoWAS              FALSE      FALSE
## Partidos               FALSE      FALSE
## Minutos_jugados        FALSE      FALSE
## Eficiencia             FALSE      FALSE
## Exito                  FALSE      FALSE
## Triples                FALSE      FALSE
## Tiros_Libres           FALSE      FALSE
## Reb_Atq                FALSE      FALSE
## Reb_Def                FALSE      FALSE
## Reb_Total              FALSE      FALSE
## Asistencias            FALSE      FALSE
## Robos                  FALSE      FALSE
## Bloqueos               FALSE      FALSE
## Porc_perdidas          FALSE      FALSE
## Porc_participacion     FALSE      FALSE
## Buen_Atq               FALSE      FALSE
## Buen_Def               FALSE      FALSE
## Buen_Tot               FALSE      FALSE
## Buen_Tot_48            FALSE      FALSE
## Calidad_Atq            FALSE      FALSE
## Calidad_Def            FALSE      FALSE
## BPM                    FALSE      FALSE
## VORP                   FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          Draft Edad EquipoBOS EquipoBRK EquipoCHI EquipoCHO EquipoCLE
## 1  ( 1 ) " "   " "  " "       " "       " "       " "       " "
## 2  ( 1 ) " "   "*"  " "       " "       " "       " "       " "
## 3  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
## 4  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
## 5  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
## 6  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
## 7  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
## 8  ( 1 ) "*"   "*"  " "       " "       " "       " "       " "
##          EquipoDAL EquipoDEN EquipoDET EquipoGSW EquipoHOU EquipoIND
```

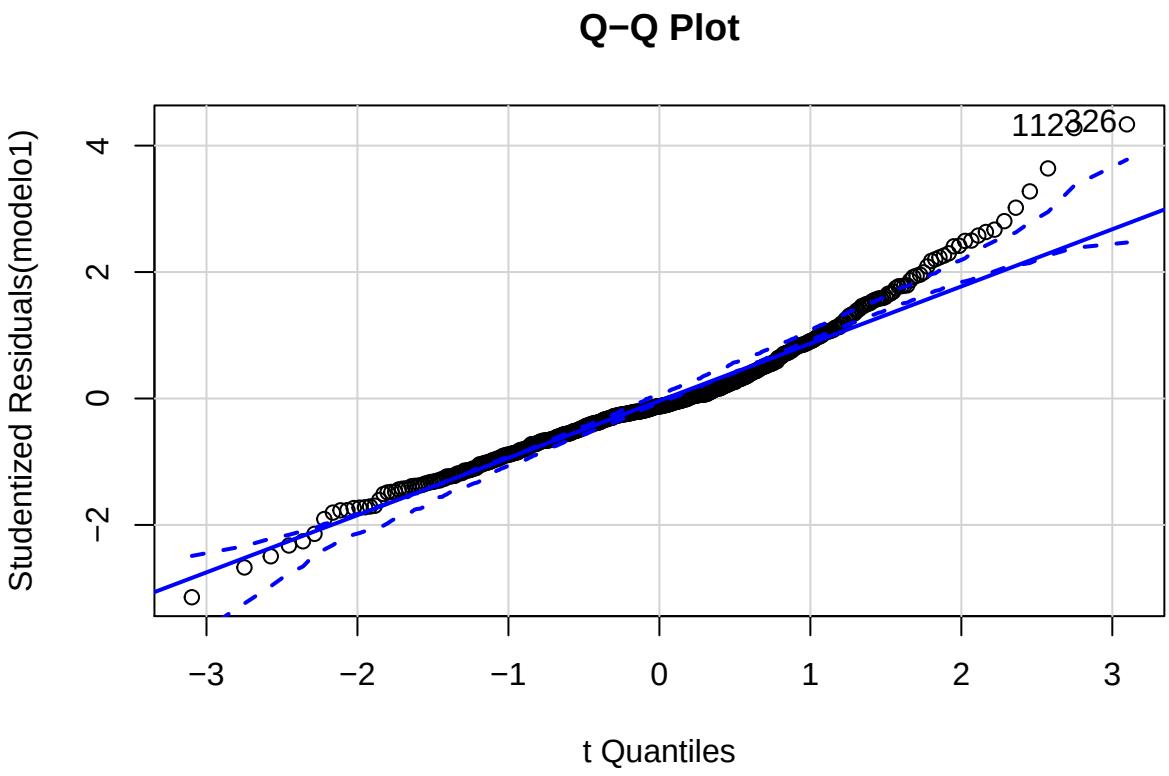
```
## 1  ( 1 ) " "       " "       " "       " "       " "       " "
## 2  ( 1 ) " "       " "       " "       " "       " "       " "
## 3  ( 1 ) " "       " "       " "       " "       " "       " "
## 4  ( 1 ) " "       " "       " "       " "       " "       " "
## 5  ( 1 ) " "       " "       " "       " "       " "       " "
## 6  ( 1 ) " "       " "       " "       " "       " "       " "
## 7  ( 1 ) " "       " "       " "       " "       " "       " "
## 8  ( 1 ) " "       " "       " "       " "       "*"       " "
##          EquipoLAC EquipoLAL EquipoMEM EquipoMIA EquipoMIL EquipoMIN
## 1  ( 1 ) " "       " "       " "       " "       " "       " "
## 2  ( 1 ) " "       " "       " "       " "       " "       " "
## 3  ( 1 ) " "       " "       " "       " "       " "       " "
## 4  ( 1 ) " "       " "       " "       " "       " "       " "
## 5  ( 1 ) " "       " "       " "       " "       " "       " "
## 6  ( 1 ) " "       " "       " "       " "       " "       " "
## 7  ( 1 ) " "       " "       " "       " "       " "       " "
## 8  ( 1 ) " "       " "       " "       " "       " "       " "
##          EquipoNOP EquipoNYK EquipoOKC EquipoORL EquipoPHI EquipoPHO
## 1  ( 1 ) " "       " "       " "       " "       " "       " "
## 2  ( 1 ) " "       " "       " "       " "       " "       " "
## 3  ( 1 ) " "       " "       " "       " "       " "       " "
## 4  ( 1 ) " "       " "       " "       " "       " "       " "
## 5  ( 1 ) " "       " "       " "       " "       " "       " "
## 6  ( 1 ) " "       " "       " "       " "       " "       " "
## 7  ( 1 ) " "       " "       " "       " "       " "       " "
## 8  ( 1 ) " "       " "       " "       " "       " "       " "
##          EquipoPOR EquipoSAC EquipoSAS EquipoTOR EquipoTOT EquipoUTA
## 1  ( 1 ) " "       " "       " "       " "       " "       " "
## 2  ( 1 ) " "       " "       " "       " "       " "       " "
## 3  ( 1 ) " "       " "       " "       " "       " "       " "
## 4  ( 1 ) " "       " "       " "       " "       " "       " "
## 5  ( 1 ) " "       " "       " "       " "       " "       " "
## 6  ( 1 ) " "       " "       " "       " "       " "       " "
## 7  ( 1 ) " "       " "       " "       " "       " "       " "
## 8  ( 1 ) " "       " "       " "       " "       " "       " "
##          EquipoWAS Partidos Minutos_jugados Eficiencia Exito Triples
## 1  ( 1 ) " "       " "      " "             " "        " "   " "
## 2  ( 1 ) " "       " "      " "             " "        " "   " "
## 3  ( 1 ) " "       " "      " "             " "        " "   " "
## 4  ( 1 ) " "       " "      " "             " "        " "   " "
## 5  ( 1 ) " "       "*"      " "             " "        " "   " "
## 6  ( 1 ) " "       "*"      "*"             " "        " "   " "
## 7  ( 1 ) " "       "*"      "*"             " "        " "   " "
## 8  ( 1 ) " "       "*"      "*"             " "        " "   " "
##          Tiros_Libres Reb_Atq Reb_Def Reb_Total Asistencias Robos Bloqueos
## 1  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 2  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 3  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 4  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 5  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 6  ( 1 ) " "          " "     " "     " "       " "         " "   " "
## 7  ( 1 ) " "          " "     "*"     " "       " "         " "   " "
## 8  ( 1 ) " "          " "     "*"     " "       " "         " "   " "
##          Porc_perdidas Porc_participacion Buen_Atq Buen_Def Buen_Tot
```

```
## 1  ( 1 ) " "           " "                " "      " "      "*"
## 2  ( 1 ) " "           " "                " "      " "      "*"
## 3  ( 1 ) " "           " "                " "      " "      "*"
## 4  ( 1 ) " "           "*"                " "      " "      "*"
## 5  ( 1 ) " "           "*"                " "      " "      "*"
## 6  ( 1 ) " "           "*"                " "      " "      "*"
## 7  ( 1 ) " "           "*"                " "      " "      "*"
## 8  ( 1 ) " "           "*"                " "      " "      "*"
##          Buen_Tot_48 Calidad_Atq Calidad_Def BPM VORP
## 1  ( 1 ) " "         " "         " "         " " " "
## 2  ( 1 ) " "         " "         " "         " " " "
## 3  ( 1 ) " "         " "         " "         " " " "
## 4  ( 1 ) " "         " "         " "         " " " "
## 5  ( 1 ) " "         " "         " "         " " " "
## 6  ( 1 ) " "         " "         " "         " " " "
## 7  ( 1 ) " "         " "         " "         " " " "
## 8  ( 1 ) " "         " "         " "         " " " "
```

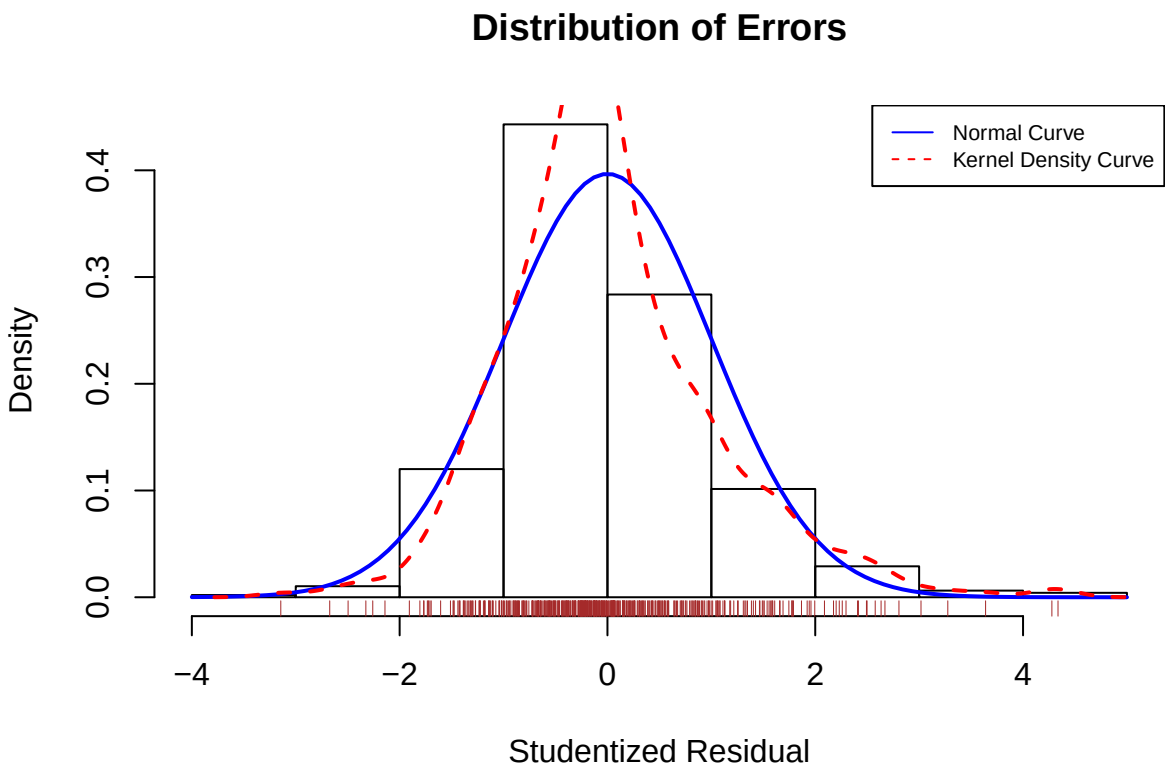
```
#Una vez hecho esto me salen 8 modelos con distintas variables en cada uno.
#Voy a comparar estos modelos para ver cuales de ellos tienen menos BIC
summary(regfit.full)$bic
```

```
## [1] -195.4434 -247.4542 -277.9638 -287.7754 -290.6157 -323.9518 -323.0798
## [8] -320.0409
```

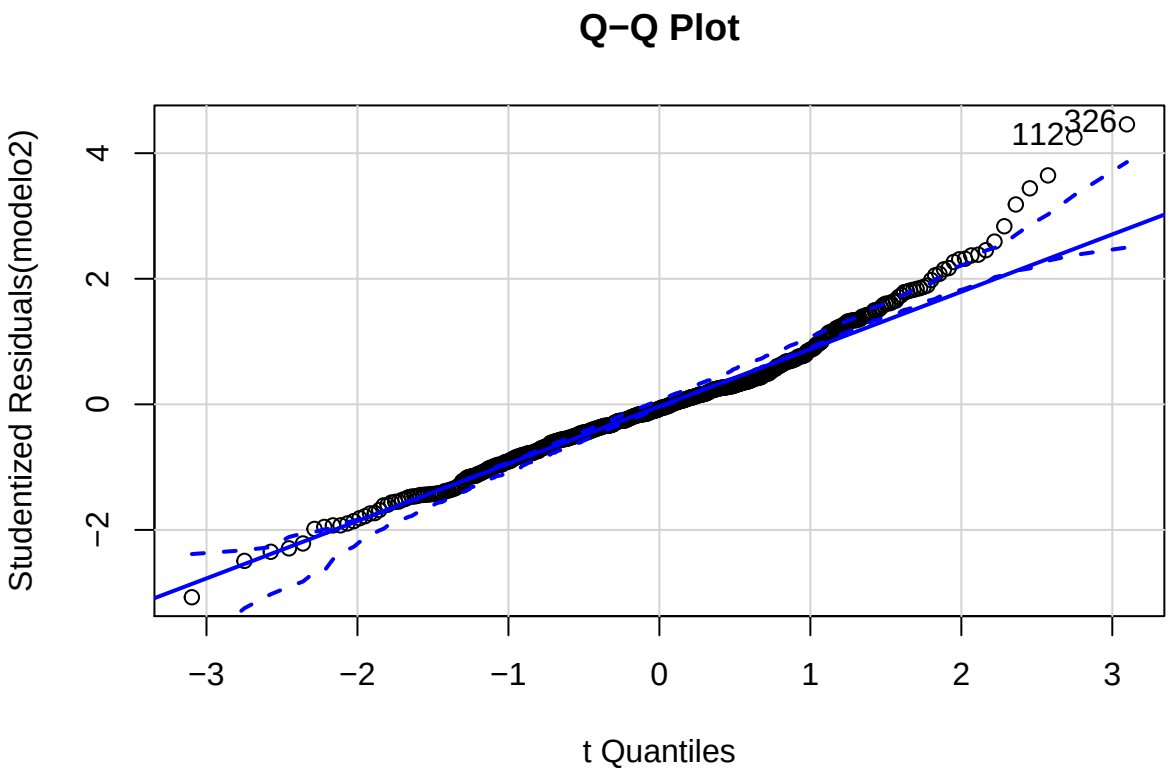
```
#Me quedo con 4 de los 8 modelos ya que estos tienen el BIC menor:

modelo1 <- lm(Salario~ Draft+Edad+Partidos+Buen_Tot,data=nba)
modelo2 <- lm(Salario~Draft+Edad+Partidos+Minutos_jugados+Buen_Tot,data=nba)
modelo3 <- lm(Salario~Draft+Edad+Partidos+Minutos_jugados+Reb_Def+Buen_Tot,data=nba)
modelo4 <- lm(Salario~Draft+Edad+Equipo+Partidos+Minutos_jugados+Reb_Def+Buen_Tot,data=nba)

#Estos son los 4 modelos que mejor explican la variable salario, para todos ellos es común las variable
#Draft,Edad,Partidos y Buen_Tot
#Vamos a realizar una serie de test para estudiar estos modelos

###Test de normalidad###
residplot <- function(fit, nbreaks=10) {
  z <- rstudent(fit)
  hist(z, breaks=nbreaks, freq=FALSE,
       xlab="Studentized Residual",
       main="Distribution of Errors")
  rug(jitter(z), col="brown")
  curve(dnorm(x, mean=mean(z), sd=sd(z)),
        add=TRUE, col="blue", lwd=2)
  lines(density(z)$x, density(z)$y,
        col="red", lwd=2, lty=2)
  legend("topright",
         legend = c( "Normal Curve", "Kernel Density Curve"),
         lty=1:2, col=c("blue","red"), cex=.7)
}
#Modelo1
n_modelo1 <- qqPlot(modelo1, labels=row.names(nba), id.method="identify",
       simulate=TRUE, main="Q-Q Plot")
```



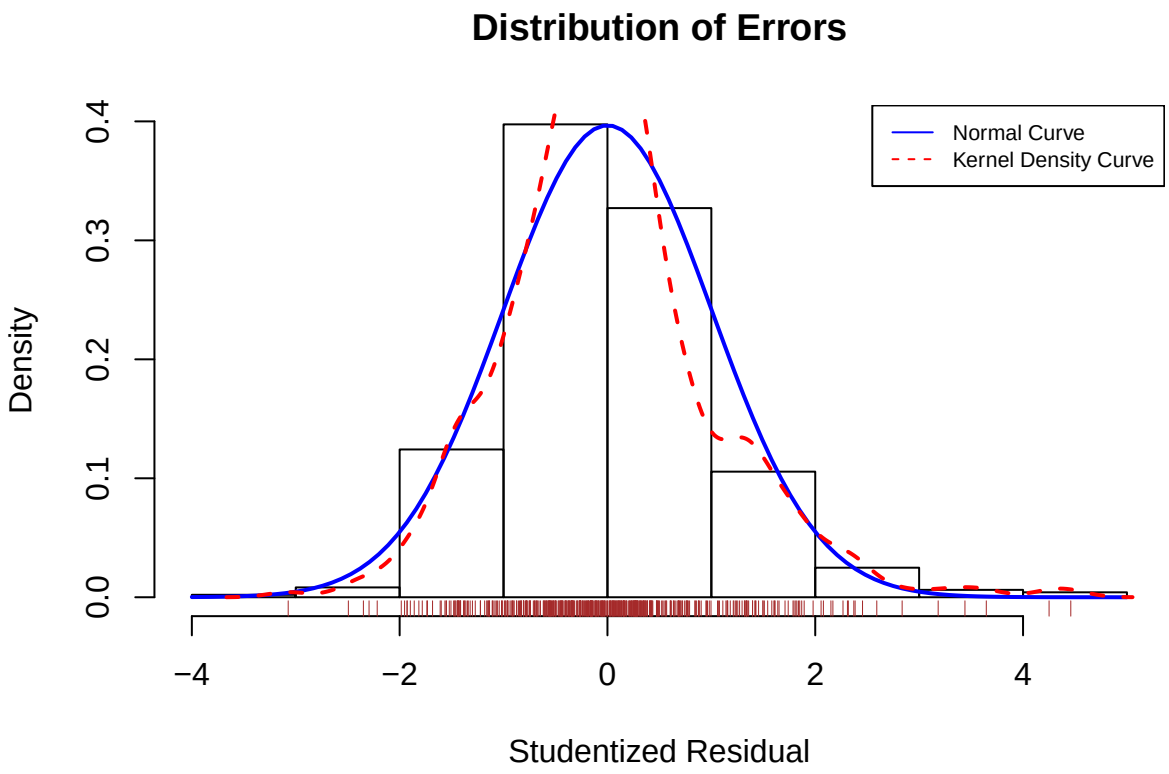

```
residplot1 <- residplot(modelo1)
```



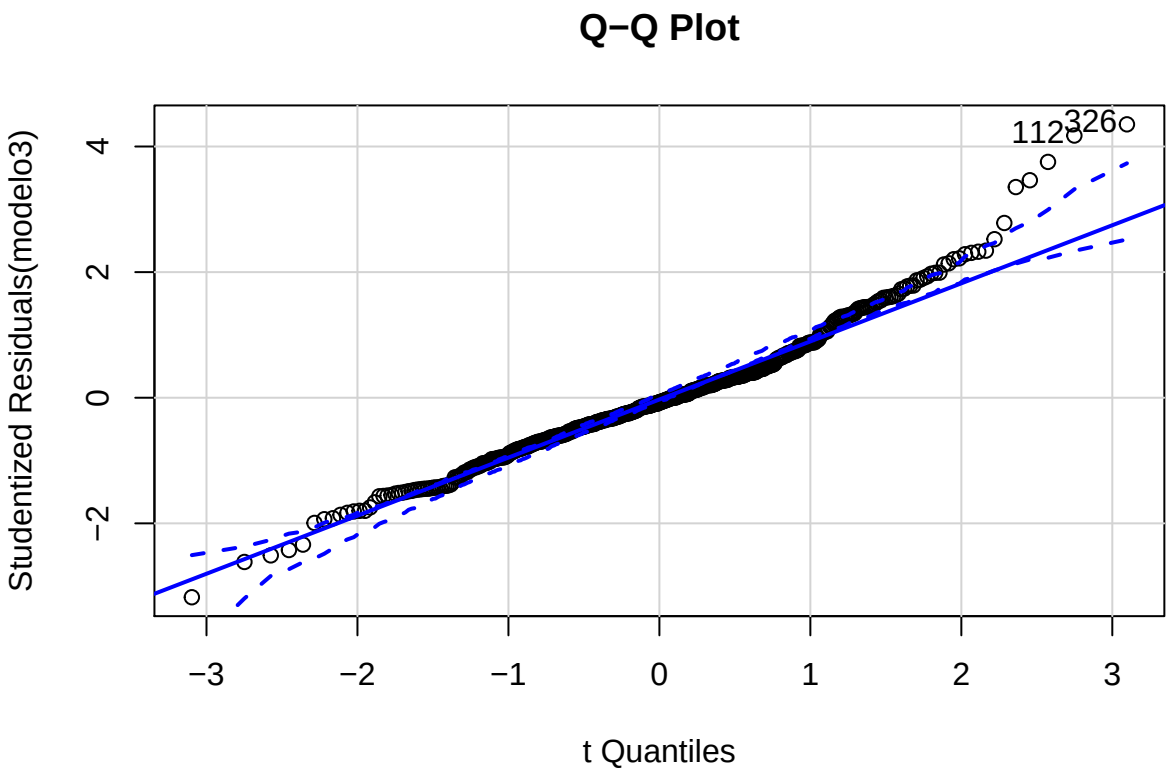

```
#Modelo2
n_modelo2 <- qqPlot(modelo2, labels=row.names(nba), id.method="identify",
                    simulate=TRUE, main="Q-Q Plot")
```



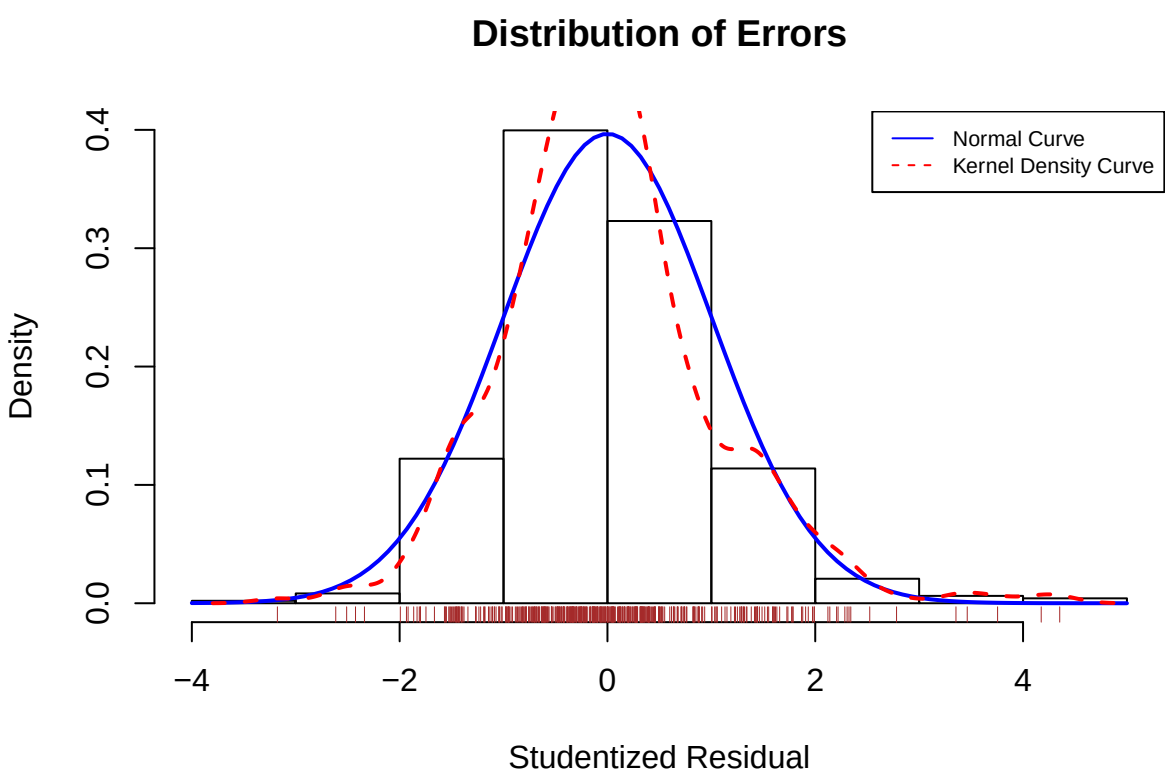

```
residplot2 <- residplot(modelo2)
```



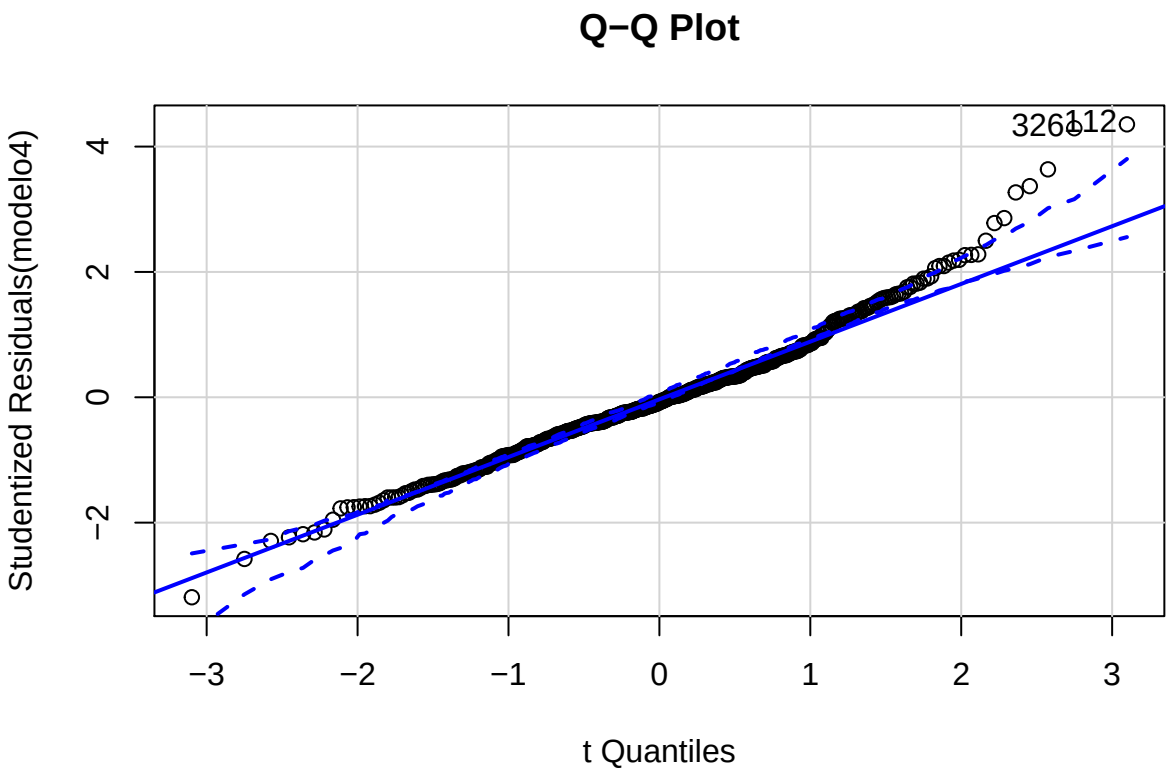

```
#Modelo3
n_modelo3 <- qqPlot(modelo3, labels=row.names(nba), id.method="identify",
                    simulate=TRUE, main="Q-Q Plot")
```



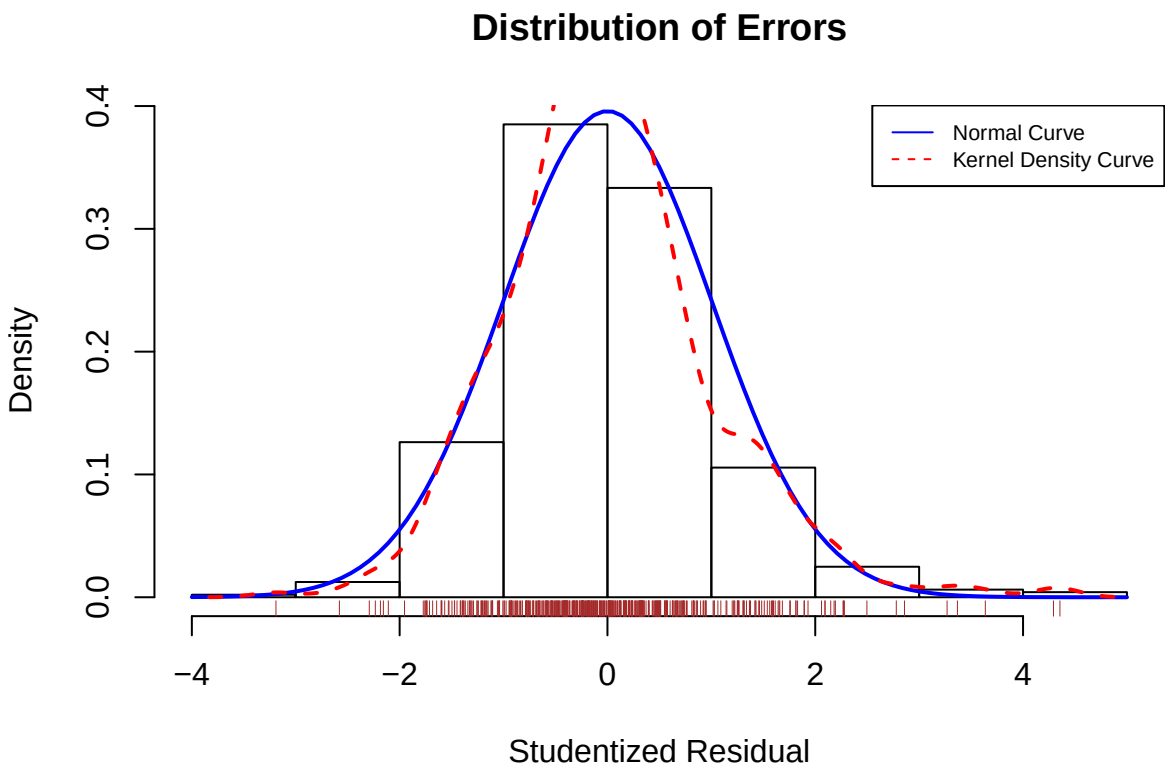

```
residplot3 <- residplot(modelo3)
```



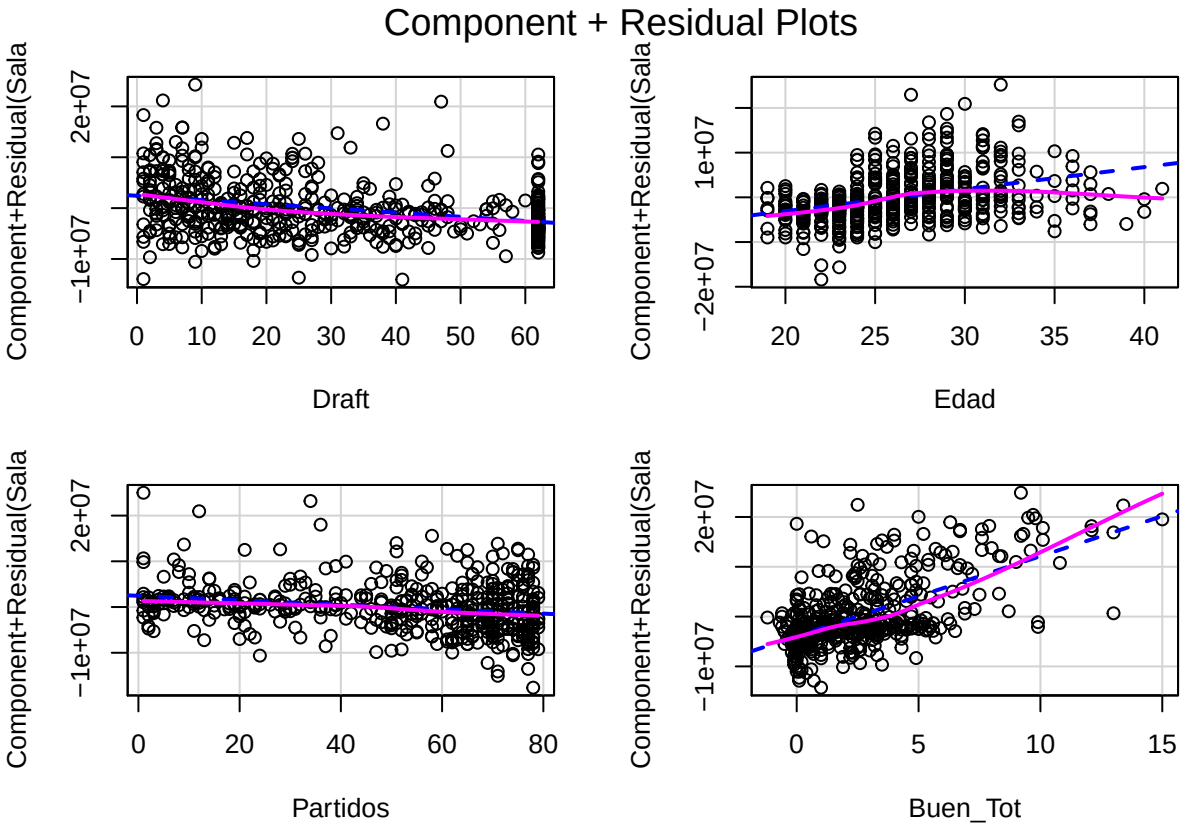

```
#Modelo4
n_modelo4 <- qqPlot(modelo4, labels=row.names(nba), id.method="identify",
                    simulate=TRUE, main="Q-Q Plot")
```



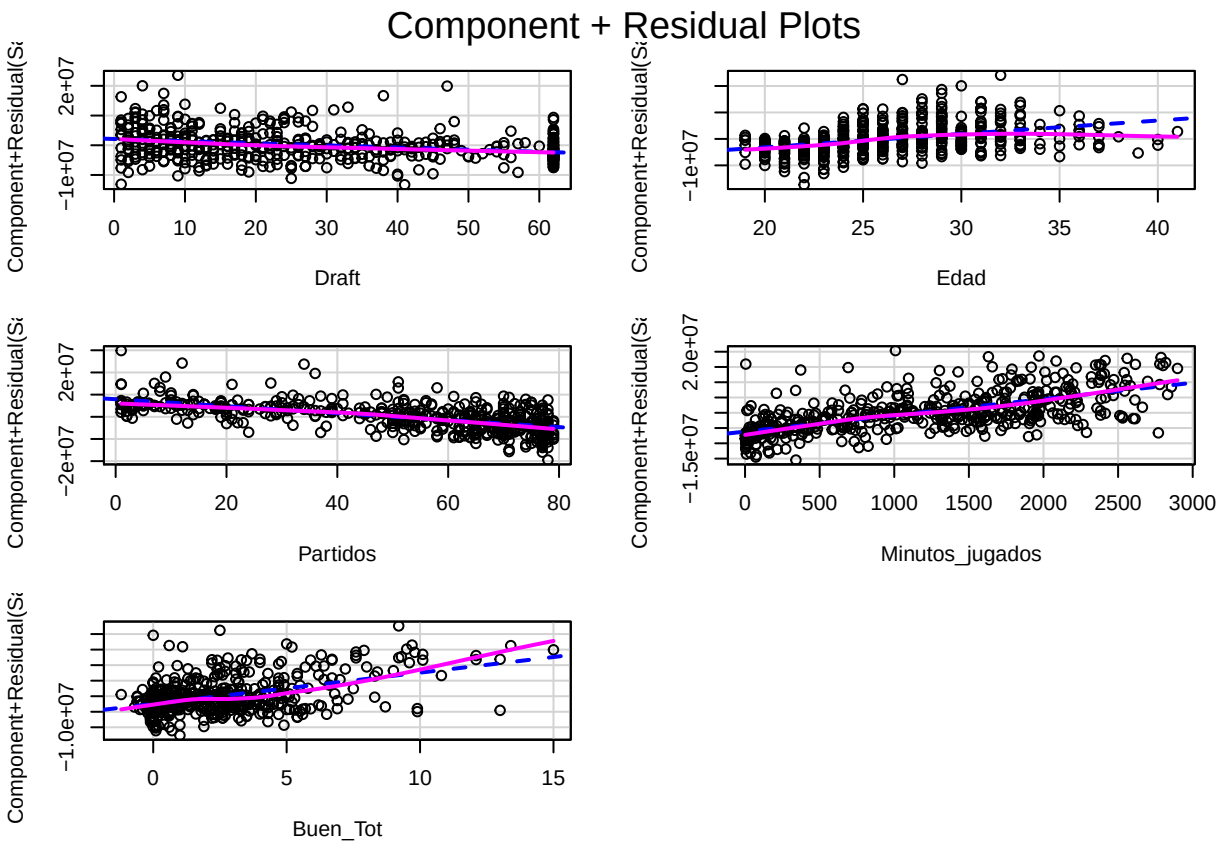

```
residplot4 <- residplot(modelo4)
```



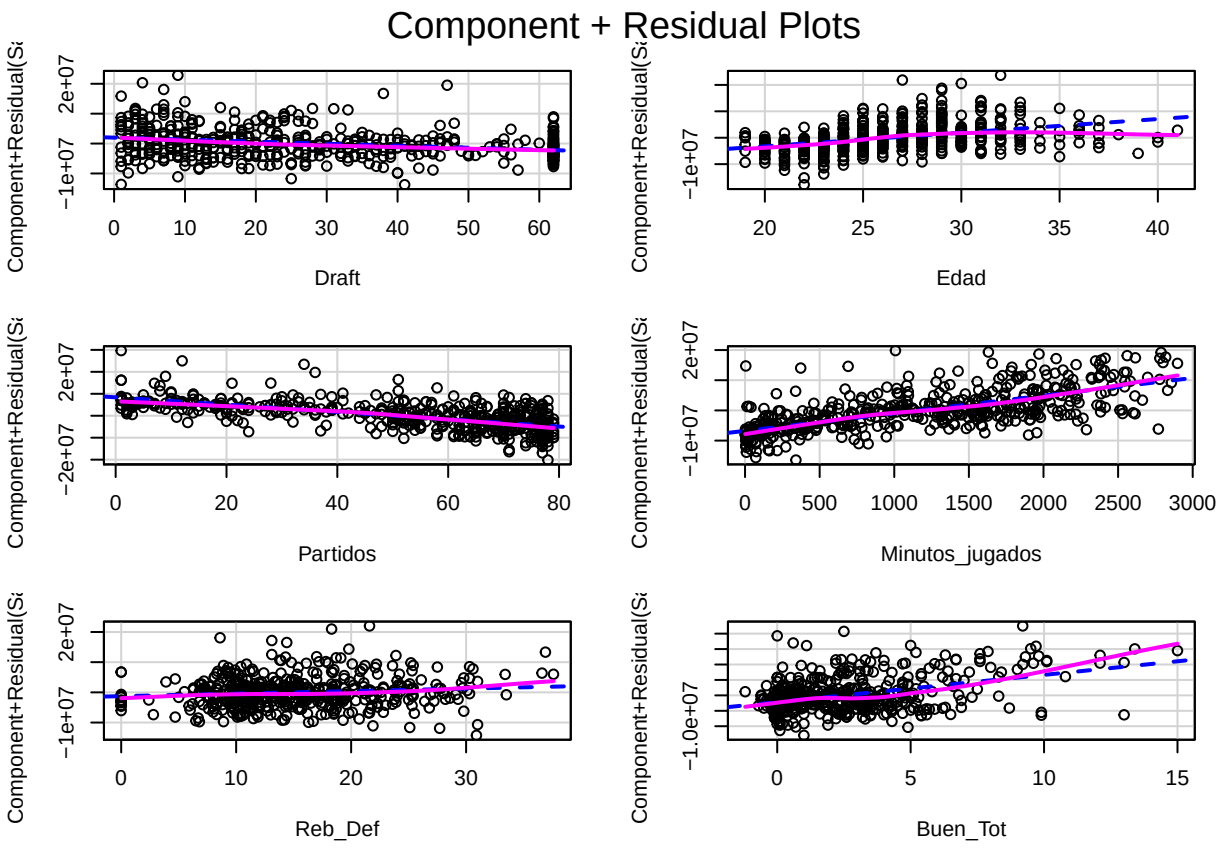

```
#Podemos concluir en que nuestros 4 modelos siguen una distribución normal
#al igual que sus errores

###Linealidad###
#Modelo1
lin_mod1 <- crPlots(modelo1)
```



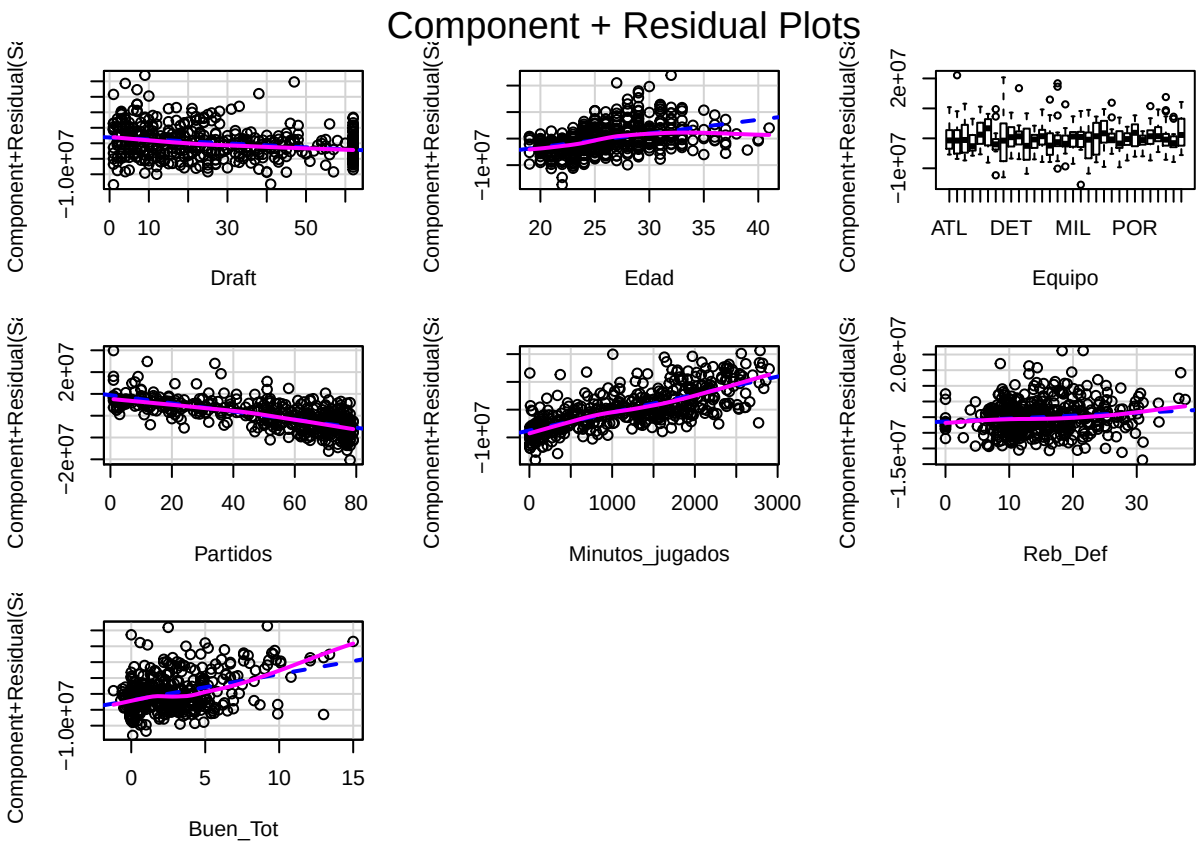

```
#Modelo2
lin_mod2 <- crPlots(modelo2)
```




```
#Modelo3
lin_mod3 <- crPlots(modelo3)
```

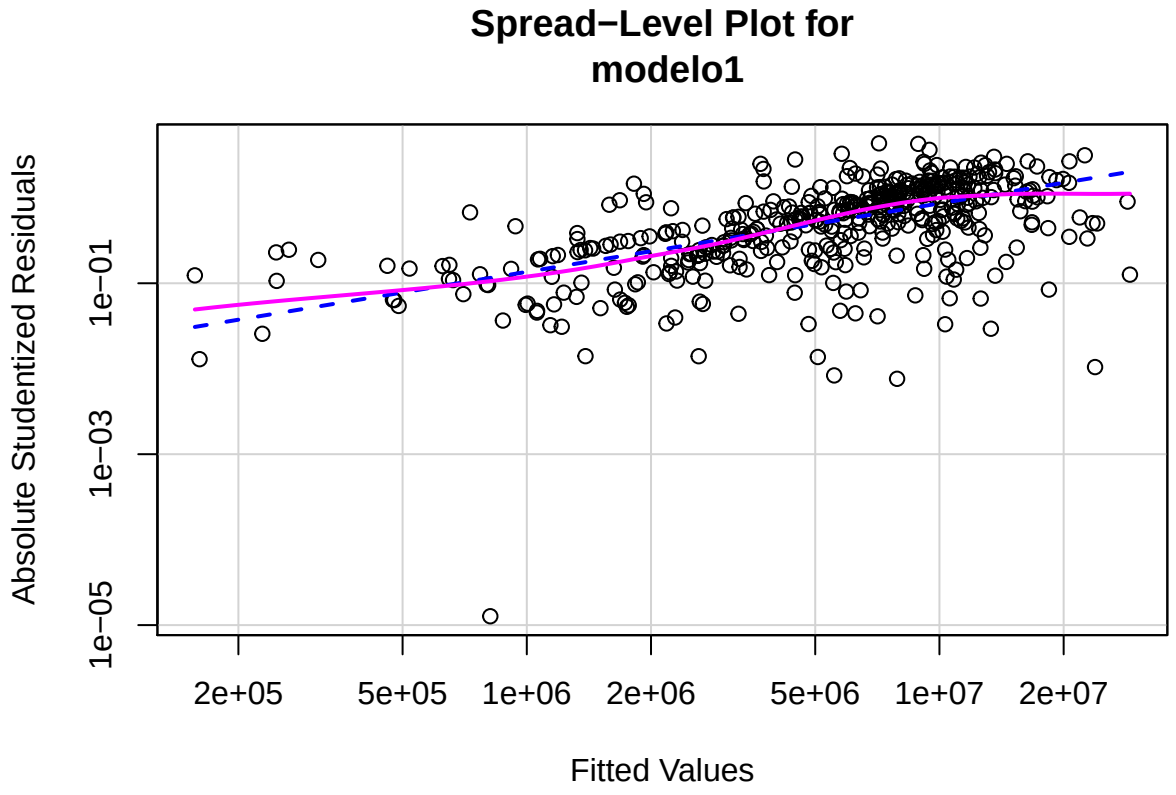



```
#Modelo3
lin_mod4 <- crPlots(modelo4)
```



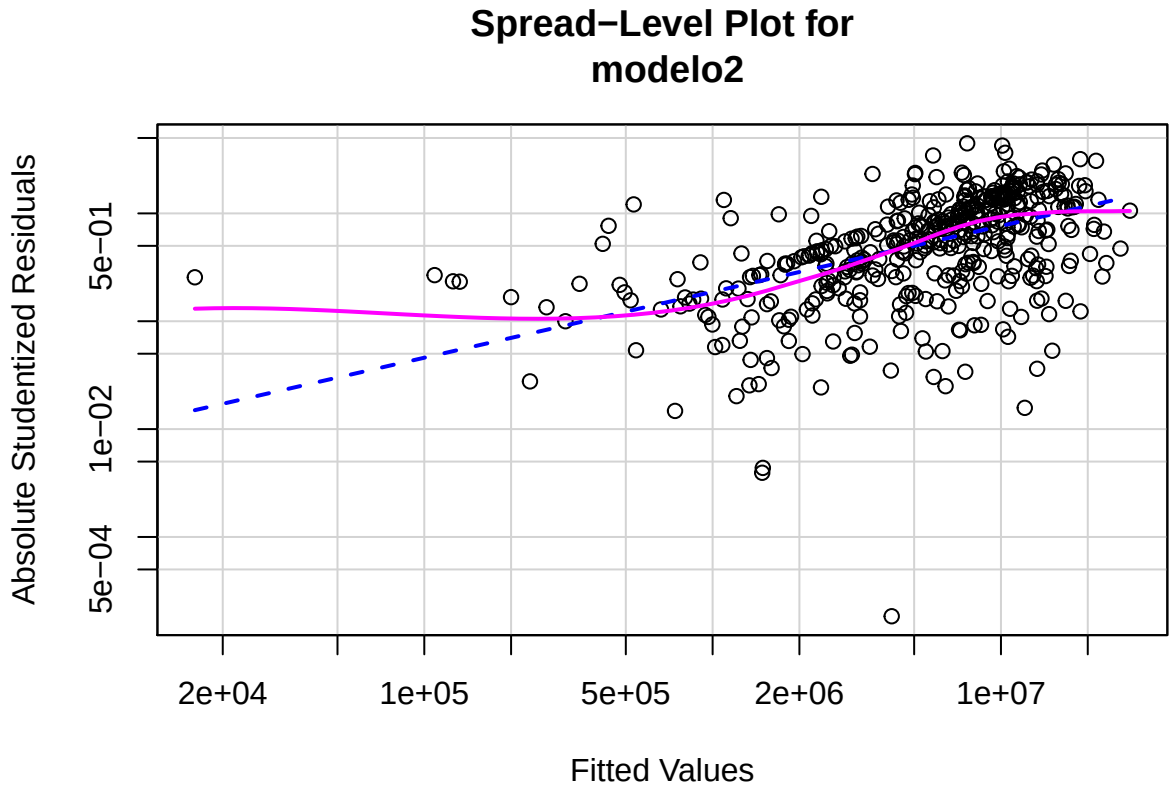

```
#En general las varibales son todas lineales menos la variable equipo del cuarto modelo ya que
#Son los nombres de los equipos y no tienen valores numericos

###Homocedasticidad###
#Modelo1
hom_mod1 <- spreadLevelPlot(modelo1)
```

```
## Warning in spreadLevelPlot.lm(modelo1):
## 24 negative fitted values removed
```



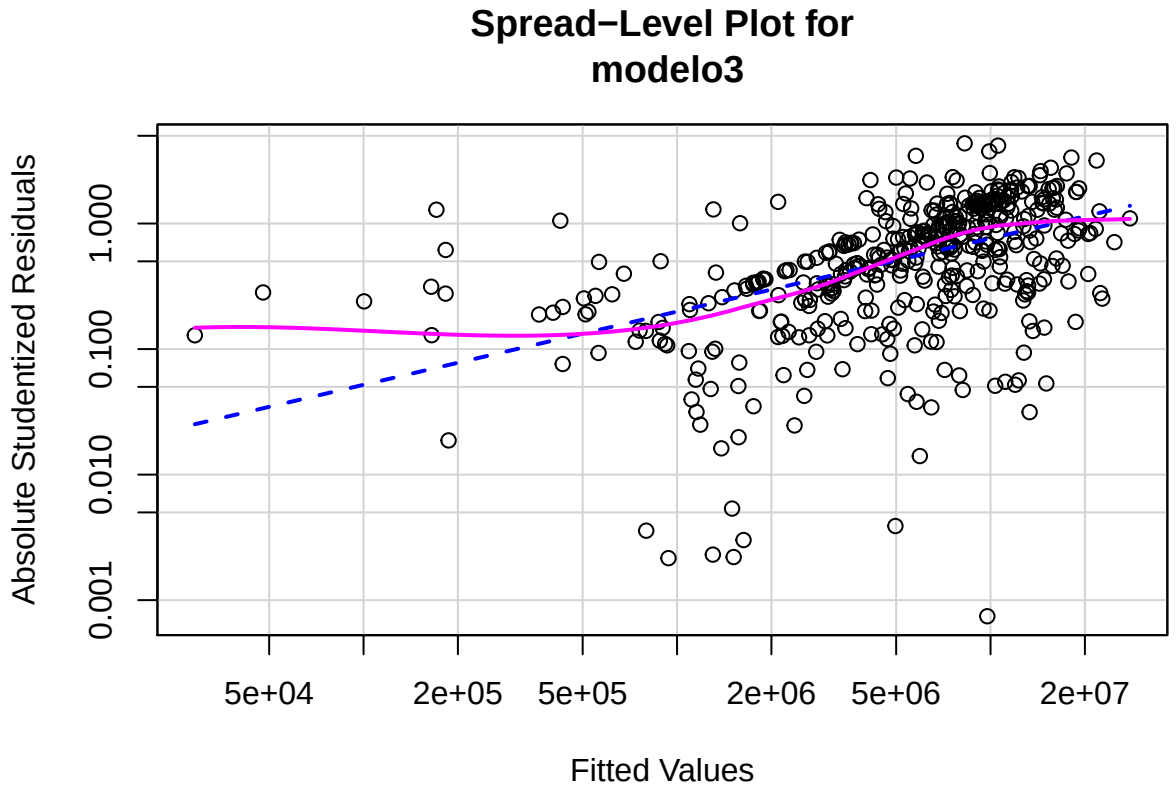

```
#Modelo2
hom_mod2 <- spreadLevelPlot(modelo2)
```

```
## Warning in spreadLevelPlot.lm(modelo2):
## 37 negative fitted values removed
```



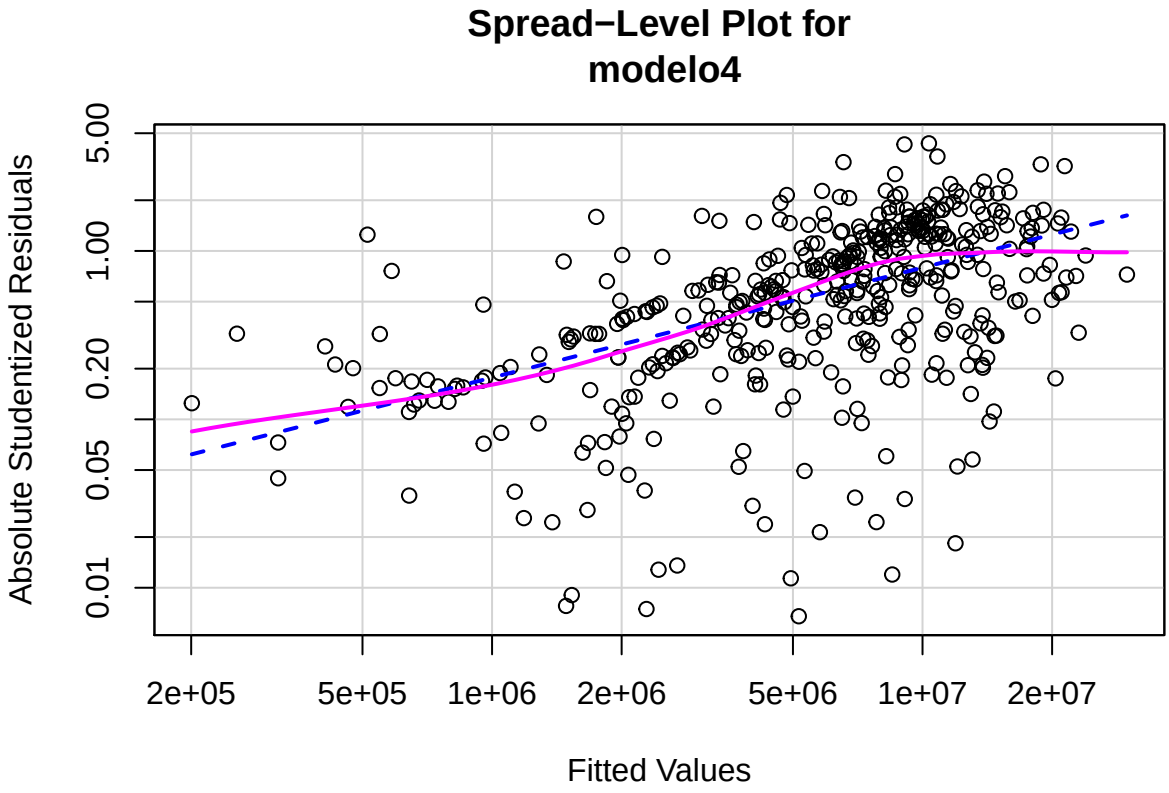

```
#Modelo3
hom_mod3 <- spreadLevelPlot(modelo3)
```

```
## Warning in spreadLevelPlot.lm(modelo3):
## 35 negative fitted values removed
```




```
#Modelo3
hom_mod4 <- spreadLevelPlot(modelo4)
```

```
## Warning in spreadLevelPlot.lm(modelo4):
## 43 negative fitted values removed
```




```
#Conforme recorremos los modelos, las gráficas tienen más pendiente, por lo que el
#primer modelo es más homocedástico que el cuarto esto tiene que ver con que desde el
#modelo 1 al 4 estamos añadiendo variables

##############################################################################################################

#Ahora vamos a hacer un Cross Validation para ver que modelo predice mejor
#Para el modelo 1
set.seed(250)
numData=nrow(nba)
train=sample(numData ,numData/2)

regres.train =lm(Salario~ Draft+Edad+Partidos+Buen_Tot,nba ,subset =train )
attach(nba)
mean((Salario-predict(regres.train ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 483 rows
```

```
## [1] 2.725836e+13
```

```
detach(nba)
sqrt(2.725836e+13)
```

```
## [1] 5220954
```

```
#El error cuadrático medio para el modelo 1 es de 2.725836e+13, el error es de 5.220.954 millones

#Para el modelo 2
regres.train2 =lm(Salario~Draft+Edad+Partidos+Minutos_jugados+Buen_Tot,nba ,subset =train )
attach(nba)
mean((Salario-predict(regres.train2 ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 483 rows
```

```
## [1] 2.430289e+13
```

```
detach(nba)
sqrt(2.430289e+13)
```

```
## [1] 4929796
```

```
#El error cuadrático medio para el modelo 2 es de 2.430289e+13, el error es de 4.929.796 millones

#Para el modelo 3
regres.train3 =lm(Salario~Draft+Edad+Partidos+Minutos_jugados+Reb_Def+Buen_Tot,nba,subset =train )
attach(nba)
mean((Salario-predict(regres.train3 ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 483 rows
```

```
## [1] 2.420056e+13
```

```
detach(nba)
sqrt(2.420056e+13)
```

```
## [1] 4919406
```

```
#El error cuadrático medio para el modelo 3 es de 2.420056e+13, el error es de 4.919.406 millones

#Para el modelo 4
regres.train4 =lm(Salario~Draft+Edad+Equipo+Partidos+Minutos_jugados+Reb_Def+Buen_Tot,nba,subset =train
attach(nba)
mean((Salario-predict(regres.train4 ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 483 rows
```

```
## [1] 2.767676e+13
```

```
detach(nba)
#El error cuadrático medio para el modelo 4 es de 2.767676e+13, el error es de 5.260.871 millones
sqrt(2.767676e+13)
```

```
## [1] 5260871
```

```
#Como podemos ver el modelo que mejor predice es el modelo 3 ya que tiene el error mas pequeño
#En este caso nuestro error es de 4.919.406 millones de dolares
#Las variables que mejor explican el salario serían la posición en el Draft, la edad el jugador,los par
#los minutos jugados, los rebotes en defensa y que haya contribuido de forma notable en total con el eq
```