Practica 5

Agustin Riquelme y Heriberto Espino

2023-10-11

Extracción de datos

Este conjunto de datos contiene información recopilada por el Servicio del Censo de EE. UU. sobre la vivienda en el área de Boston, Massachusetts, con 14 variables que van desde tasas de criminalidad per cápita hasta la proximidad a carreteras y la calidad del aire El conjunto de datos contiene un total de 506 casos. El nombre de este conjunto de datos es simplemente "boston".

Hay 14 variables, que son:

CRIM - tasa de criminalidad per cápita por ciudad

ZN - proporción de terreno residencial zonificado para lotes de más de 25,000 pies cuadrados.

INDUS - proporción de acres de negocios no minoristas por ciudad.

CHAS - variable ficticia del río Charles (1 si el lote limita con el río; 0 en caso contrario)

NOX - concentración de óxidos de nitrógeno (partes por 10 millones)

RM - número promedio de habitaciones por vivienda

AGE - proporción de unidades ocupadas por el propietario construidas antes de 1940

DIS - distancias ponderadas a cinco centros de empleo de Boston

RAD - índice de accesibilidad a carreteras radiales

TAX - tasa de impuesto a la propiedad de valor completo por cada $10,000

PTRATIO - ratio de alumnos por profesor por ciudad

BLACK - la proporción de personas de raza negra por ciudad

LSTAT - % de población de estatus socioeconómico bajo

MEDV - Valor mediano de las viviendas ocupadas por el propietario en miles de dólares (en unidades de $1000)

donde MEDV va a ser nuestra variable de respuesta para los modelos de regresión.

```r
library(MASS)
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'
```

```
## The following object is masked from 'package:MASS':
##
##     Boston
```

```r
library(PerformanceAnalytics)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

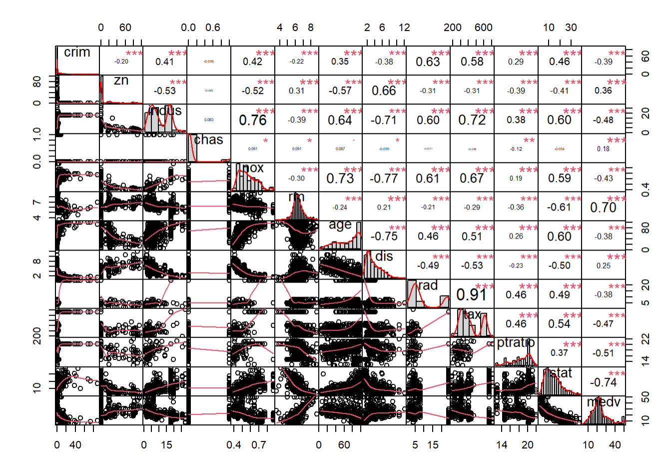
```
## The following object is masked from 'package:graphics':
##
##     legend
```

```r
library(lmtest)
library(nortest)

datos <- data.frame(Boston)

crim <- datos$crim
zn  <- datos$zn
indus <- datos$indus
chas <- datos$chas
nox <- datos$nox
rm <- datos$rm
age <- datos$age
dis <- datos$dis
rad <- datos$rad
tax <- datos$tax
ptratio <- datos$ptratio
black <- datos$black
lstat <- datos$lstat
medv <- datos$medv
```

Diagrama de pares



Descartaremos a las variables zn, indus, chas, tac y black por ser variables categoricas.

Las variables que ayudarán a predecir a medv pueden ser las variables stat, dis, rad y age.

Construcción del modelo 1

```r
m1 <- lm(medv ~ rm + lstat + ptratio + age)
summary(m1)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + ptratio + age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6689  -3.0889  -0.7632   1.8436  28.5351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.24370    3.93168   4.895 1.33e-06 ***
## rm           4.36784    0.43568  10.025  < 2e-16 ***
## lstat       -0.61815    0.05160 -11.980  < 2e-16 ***
## ptratio     -0.94678    0.11794  -8.028 7.12e-15 ***
## age          0.01653    0.01061   1.559     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.222 on 501 degrees of freedom
## Multiple R-squared:  0.6802, Adjusted R-squared:  0.6776
## F-statistic: 266.4 on 4 and 501 DF,  p-value: < 2.2e-16
```

La prueba general del modelo es buena, nos da un p-value: < 2.2e-16.

En la prueba individual del modelo podemos ver que la variable age no es significativa, por lo que el modelo queda descartado.

Construcción del modelo 2

La construcción del segundo modelo es sin la variable age.

```r
m2 <- lm(medv ~ rm + lstat + ptratio)
summary(m2)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + ptratio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4871  -3.1047  -0.7976   1.8129  29.6559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
## rm           4.51542    0.42587  10.603  < 2e-16 ***
## lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
## ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.229 on 502 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6767
## F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

La prueba general del modelo es buena, nos da un p-value: < 2.2e-16.

En la prueba individual todas las variablese son significativas.

Tiene un Adjusted R-squared: 0.6767, que indica que el modelo es bueno.

Ahora analizaremos los residuos:

```r
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.88804, p-value < 2.2e-16
```

El p-value es menor que 2.2e-16, concluimos que los errores no se distribuyen normal.

```r
dwtest(m2)
```

```
##
##  Durbin-Watson test
##
## data:  m2
## DW = 0.90124, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es menor a 2.2e-16, lo que indica que el test es significativo, y el DW es 0.90124, por lo que podemos concluir que hay una correlación positiva de entre los residuos.

Por último, obtendremos el AIC para compararlo con futuros modelos:

```r
AIC(m2)
```

```
## [1] 3116.097
```

Construcción del modelo 3

Como lstat sigue algo parecido a una exponencial en el diagrama de pares, vamos a agregar la variable elevada al cuadrado, para hacer un modelo cuadratico que tenga mejores aproximaciones

```r
m3 <- lm(medv ~ rm + lstat + I(lstat^2) + ptratio)
summary(m3)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + I(lstat^2) + ptratio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3166  -2.9506  -0.4864   2.2152  28.3357
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.784917   3.691814   6.984 9.14e-12 ***
## rm           3.875438   0.398850   9.717  < 2e-16 ***
## lstat       -1.649855   0.121054 -13.629  < 2e-16 ***
## I(lstat^2)   0.032019   0.003404   9.406  < 2e-16 ***
## ptratio     -0.730838   0.110633  -6.606 1.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.826 on 501 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7247
## F-statistic: 333.3 on 4 and 501 DF,  p-value: < 2.2e-16
```

La prueba general del modelo es buena, nos da un p-value: < 2.2e-16.

En la prueba individual todas las variablese son significativas.

Tiene un Adjusted R-squared: 0.7247, indicando que este modelo puede ser mejor que el anterior.

Ahora analizaremos los residuos:

```r
shapiro.test(m3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m3$residuals
## W = 0.91752, p-value = 5.361e-16
```

El p-value es 5.361e-16, concluimos que los errores no se distribuyen normal.

```r
dwtest(m3)
```

```
##
##  Durbin-Watson test
##
## data:  m3
## DW = 0.89721, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es menor que 2.2e-16, lo que indica que el test es significativo, y el DW es 0.89721, por lo que podemos concluir que hay una correlación positiva de entre los residuos.

Por último, obtendemos el AIC para compararlo con el modelo anterior:

```r
AIC(m3)
```

```
## [1] 3035.806
```

Como el AIC de este modelo es menor que el AIC del modelo pasado, concluimos que este modelo es mejor. 3035.806 es menor que 3116.097.

Construcción del modelo 4

Al igual que en el modelo pasado, ahora rm sigue algo parecido a una exponencial en el diagrama de pares, por lo que vamos a repetir el procedimiento.

```r
m4 <- lm(medv ~ rm +  I(rm^2) + lstat + I(lstat^2) + ptratio)
summary(m4)
```

```
##
## Call:
## lm(formula = medv ~ rm + I(rm^2) + lstat + I(lstat^2) + ptratio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4702  -2.5546  -0.5134   2.0633  28.4098
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 112.942417   9.644669  11.710  < 2e-16 ***
## rm          -24.797530   2.992621  -8.286 1.08e-15 ***
## I(rm^2)       2.237192   0.231739   9.654  < 2e-16 ***
## lstat        -1.260768   0.118325 -10.655  < 2e-16 ***
## I(lstat^2)    0.018729   0.003418   5.480 6.77e-08 ***
## ptratio      -0.651786   0.102002  -6.390 3.80e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.435 on 500 degrees of freedom
## Multiple R-squared:  0.7698, Adjusted R-squared:  0.7675
## F-statistic: 334.4 on 5 and 500 DF,  p-value: < 2.2e-16
```

La prueba general del modelo es buena, nos da un p-value: < 2.2e-16.

En la prueba individual todas las variablese son significativas.

Tiene un Adjusted R-squared: 0.7675 , indicando que este modelo puede ser mejor que el el modelo 3.

Ahora analizaremos los residuos:

```r
shapiro.test(m4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m4$residuals
## W = 0.87582, p-value < 2.2e-16
```

El p-value es menor que 2.2e-16, concluimos que los errores no se distribuyen normal.

```r
dwtest(m4)
```

```
##
##  Durbin-Watson test
##
## data:  m4
## DW = 0.93323, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es menor que 2.2e-16, lo que indica que el test es significativo, y el DW es 0.93323, por lo que podemos concluir que hay una correlación positiva de entre los residuos.

Por último, obtendemos el AIC para compararlo con el modelo 3:

```r
AIC(m4)
```

```
## [1] 2951.32
```

Como el AIC de este modelo es menor que el AIC del modelo 3, concluimos que este modelo es mejor.
2951.32 es menor que 3035.806.

Construcción del modelo 5

Estuve viendo diferentes combinaciones y esta tuvo un AIC menor junto con un R ajustado mayor.

```r
m5 <- lm(medv ~ rm +  I(rm^2) + lstat + I(lstat^2) + ptratio + I(crim^2) + nox + I(nox^2) )
summary(m5)
```

```
##
## Call:
## lm(formula = medv ~ rm + I(rm^2) + lstat + I(lstat^2) + ptratio +
##     I(crim^2) + nox + I(nox^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1824  -2.2942  -0.3592   1.9051  28.0071
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.037e+01  1.078e+01   8.384 5.30e-16 ***
## rm          -2.369e+01  2.963e+00  -7.995 9.15e-15 ***
## I(rm^2)      2.133e+00  2.311e-01   9.232  < 2e-16 ***
## lstat       -1.336e+00  1.281e-01 -10.424  < 2e-16 ***
## I(lstat^2)   2.153e-02  3.478e-03   6.192 1.24e-09 ***
## ptratio     -8.402e-01  1.071e-01  -7.842 2.72e-14 ***
## I(crim^2)   -1.592e-03  3.813e-04  -4.176 3.51e-05 ***
## nox          8.343e+01  1.680e+01   4.966 9.41e-07 ***
## I(nox^2)    -7.016e+01  1.366e+01  -5.137 4.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.264 on 497 degrees of freedom
## Multiple R-squared:  0.7885, Adjusted R-squared:  0.7851
## F-statistic: 231.6 on 8 and 497 DF,  p-value: < 2.2e-16
```

La prueba general del modelo es buena, nos da un p-value: < 2.2e-16.

En la prueba individual todas las variablese son significativas.

Tiene un Adjusted R-squared: 0.7851, indicando que este modelo puede ser mejor que el el modelo 3.

Ahora analizaremos los residuos:

```r
shapiro.test(m5$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m5$residuals
## W = 0.87128, p-value < 2.2e-16
```

El p-value es menor que 2.2e-16, concluimos que los errores no se distribuyen normal.

```r
dwtest(m5)
```

```
##
##  Durbin-Watson test
##
## data:  m5
## DW = 1.0055, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

El p-value es menor que 2.2e-16, lo que indica que el test es significativo, y el DW es 1.0055, conluimos que no hay una correlación entre los residuos.

Por último, obtendemos el AIC para compararlo con el modelo 4:

```r
AIC(m5)
```

```
## [1] 2914.472
```

Como el AIC de este modelo es menor que el AIC del modelo 4, concluimos que este modelo es mejor.
2914.472 es menor que 2951.32.