Modelos velocidad

IZASKUN LOPEZ-SAMANIEGO

19 de noviembre de 2017

Preparación del entorno

```
library(data.table)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
setwd(ruta)
source('./src/definitivos/funciones_opendata.R')
```

Cargar ficheros

```
a. Cargamos el fichero con la información normalizada y nos quedamos con los
datos necesarios para ejecutar el modelo.
```

```
dt.analisis   <- as.data.table(read.csv('F:/201711_dataton_opendata_madrid/d
at/PM16_dataset.csv'))

dt.analisis <- dt.analisis[,list(identif, ds,
                                 intensidad, ocupacion, carga,
                                 vmed, vel.med, carga.med,
                                 diaSemana, diaMes, Mes, fechaTrunc,
                                 prec, prec_norm, prec_zscore,
                                 Dia_semana, laborable...festivo...domingo.f
estivo,
                                 Tipo.de.Festivo, Festividad)]
```

```
dt.datos.prev   <- as.data.table(read.csv('F:/201711_dataton_opendata_madri
d/dat/trafico_outlier_datos_previos.csv', sep = ';'))
dt.datos.prev <- dt.datos.prev[,list(identif, ds,
                                     carga.1 = carga.45/100,
                                     vmed.1  = vmed.45/100,
                                     carga.2 = carga.60/100,
                                     vmed.2  = vmed.60/100,
                                     carga.3 = carga.75/100,
                                     vmed.3  = carga.75/100,
                                     carga.4 = carga.90/100,
                                     vmed.4  = vmed.90/100)]
```

```
dt.analisis   <- merge(dt.analisis, dt.datos.prev,
                       by.x = c('identif', 'ds'),
                       by.y = c('identif', 'ds'),
                       all.x = FALSE, all.y = FALSE)
dt.analisis <- Transformacion_variables(dt.analisis)
```

Dividimos la muestra en casos de test y casos de training

```
inTrain <- sample(1:nrow(dt.analisis),
                  nrow(dt.analisis)*0.3)

train.analisis  <- dt.analisis[-inTrain,]
test.analisis   <- dt.analisis[inTrain,]
```

Regresión Líneal Múltivariante

```
lm.M30 <- lm(vmed ~ vel.med   +
                    carga.med +
                    carga.1   +
                    vmed.1    +
                    carga.2   +
                    vmed.2    +
                    carga.3   +
             #       vmed.3    +
                    diaMes     +
                    Mes        +
                    prec_norm +
             #       var.carga.1 +
             #       var.carga.2 +
                    var.carga.3 +
                    var.vmed.1   +
                    var.vmed.2   +
                    var.vmed.3   +
                    diaLunes      +
                    diaMartes     +
                    diaMiercoles +
                    diaJueves     +
                    diaViernes     +
                    diaSabado     +
               #     diaDomingo    +
                    n.festivo   ,
            data = train.analisis)
print(lm.M30$coefficients)
```

```
##   (Intercept)       vel.med     carga.med       carga.1        vmed.1
## -2.7907696897  0.1603093947 -0.1020299900 -0.2130701979  3.7265228022
##       carga.2        vmed.2       carga.3        diaMes           Mes
##  0.2355567180  0.3941102483 -0.0631577107 -0.0001930377 -0.0034262834
##     prec_norm   var.carga.3    var.vmed.1    var.vmed.2    var.vmed.3
## -0.0402700429 -0.2668032583  0.0108414883 -0.0047742675 -0.1187681811
##      diaLunes     diaMartes  diaMiercoles     diaJueves    diaViernes
## -0.0678784553 -0.0755632197 -0.0764366222 -0.0832623718 -0.0746151249
##     diaSabado     n.festivo
##  0.0072834723  0.0276107423
```

```
summary(lm.M30)
```

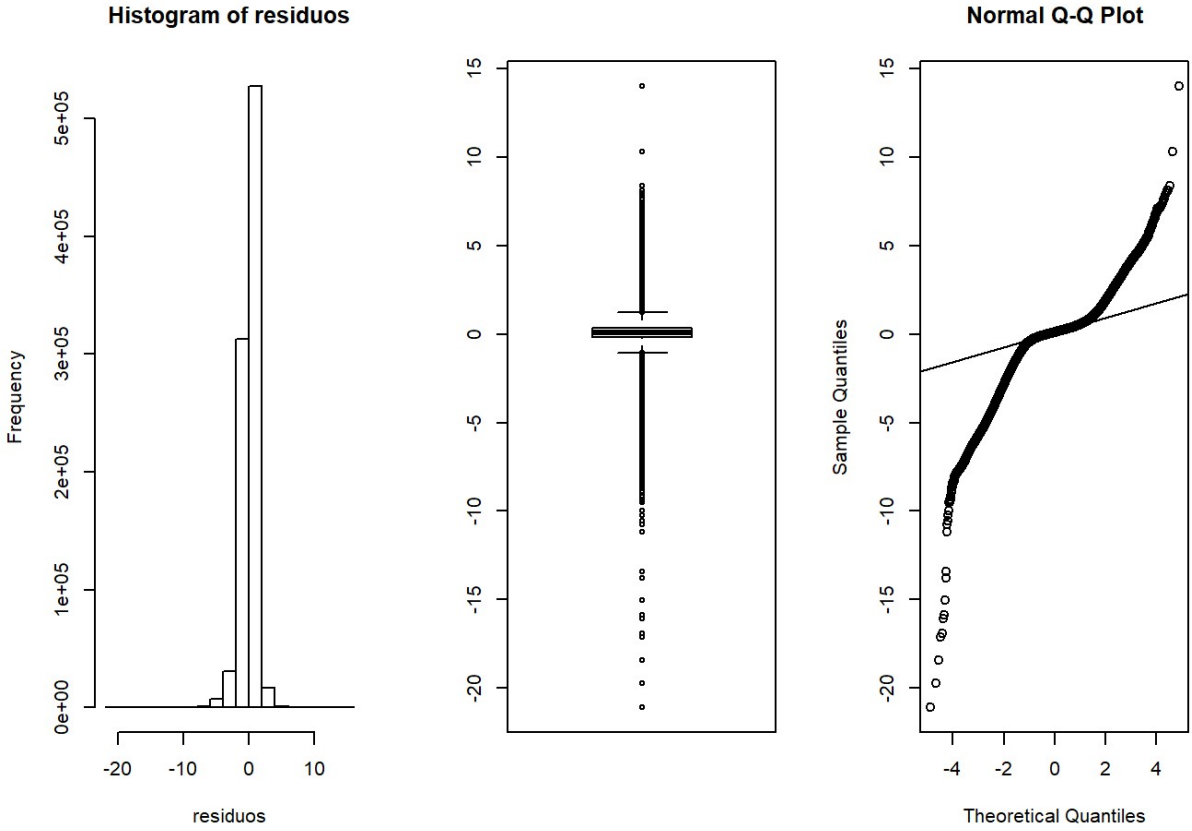
```
##
## Call:
## lm(formula = vmed ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + diaMes + Mes + prec_norm + var.carga.3
+
##     var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes + diaMartes +
##     diaMiercoles + diaJueves + diaViernes + diaSabado + n.festivo,
##     data = train.analisis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2896  -0.0936   0.0529   0.1871   6.9753
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -2.791e+00  4.849e-03 -575.553  < 2e-16 ***
## vel.med       1.603e-01  8.829e-04  181.582  < 2e-16 ***
## carga.med    -1.020e-01  6.946e-03  -14.688  < 2e-16 ***
## carga.1      -2.131e-01  1.091e-02  -19.535  < 2e-16 ***
## vmed.1        3.727e+00  1.045e-02  356.644  < 2e-16 ***
## carga.2       2.356e-01  1.604e-02   14.684  < 2e-16 ***
## vmed.2        3.941e-01  1.042e-02   37.817  < 2e-16 ***
## carga.3      -6.316e-02  1.113e-02   -5.673  1.4e-08 ***
## diaMes       -1.930e-04  6.035e-05   -3.199 0.001381 **
## Mes          -3.426e-03  1.543e-04  -22.199  < 2e-16 ***
## prec_norm    -4.027e-02  3.774e-03  -10.669  < 2e-16 ***
## var.carga.3  -2.668e-01  1.078e-02  -24.756  < 2e-16 ***
## var.vmed.1    1.084e-02  2.936e-03    3.693 0.000222 ***
## var.vmed.2   -4.774e-03  7.974e-05  -59.877  < 2e-16 ***
## var.vmed.3   -1.188e-01  1.102e-03 -107.793  < 2e-16 ***
## diaLunes     -6.788e-02  2.541e-03  -26.718  < 2e-16 ***
## diaMartes    -7.556e-02  2.539e-03  -29.765  < 2e-16 ***
## diaMiercoles -7.644e-02  2.555e-03  -29.913  < 2e-16 ***
## diaJueves    -8.326e-02  2.562e-03  -32.502  < 2e-16 ***
## diaViernes   -7.462e-02  2.510e-03  -29.722  < 2e-16 ***
## diaSabado     7.283e-03  2.145e-03    3.396 0.000684 ***
## n.festivo     2.761e-02  9.039e-04   30.547  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 897413 degrees of freedom
## Multiple R-squared:  0.7284, Adjusted R-squared:  0.7284
## F-statistic: 1.146e+05 on 21 and 897413 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_vel_45min.RData')
```

Análisis de los residuos

```
a. Supuesto 1: Normalidad
```

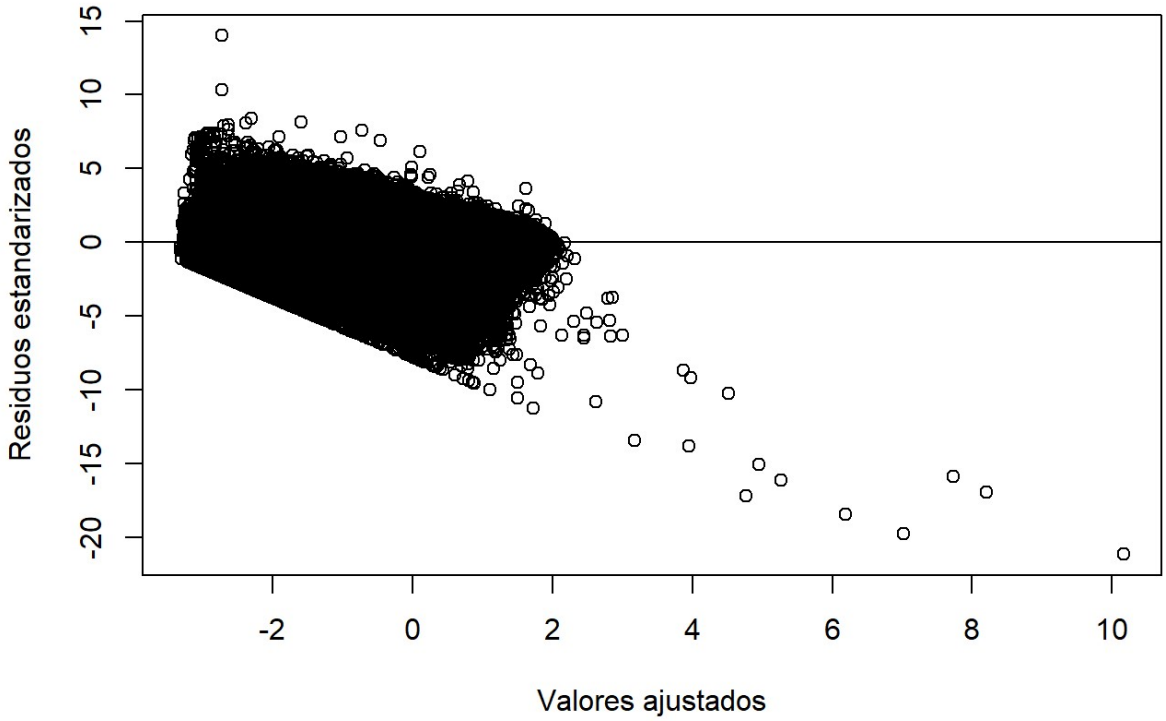
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
qqline(residuos)
```




```
b. Supuesto 2: Varianza de los errores es constante:
    - No es constante, tiene tendencia lo que indica que hay una variable de
sconocida que impacta en el tráfico pero no la hemos detectado.
```

```
par(mfrow=c(1,1))

# gráfico 2D de los valores ajustados vs. los residuos estandarizados
plot(fitted.values(lm.M30),rstandard(lm.M30),
     xlab="Valores ajustados",
     ylab="Residuos estandarizados")
 # dibuja la recta en cero
abline(h=0)
```



CALCULO RMSE

a. Training

```
predict.M30 <- predict(lm.M30, interval = "prediction")
```

```
## Warning in predict.lm(lm.M30, interval = "prediction"): predictions on cu
rrent data refer to _future_ responses
```

```
calculo_error(train.analisis, as.data.table(predict.M30))
```

```
##       error
## 1: 62885.53
```

b. Test

```
predict.M30 <- predict(lm.M30, test.analisis, interval = "prediction")
calculo_error(test.analisis, as.data.table(predict.M30))
```

```
##       error
## 1: 27186.51
```