Modelos

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(carga ~ 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
##  4.750755e-03 -2.023061e-04  1.550922e-01  1.178930e+00 -1.131048e-01
##       carga.2        vmed.2       carga.3        vmed.3        diaMes
##  6.493472e-02  1.246211e-01 -4.044939e-01            NA -1.515203e-04
##           Mes     prec_norm   var.carga.3    var.vmed.1    var.vmed.2
## -9.601845e-05  4.665605e-03 -3.053123e-01 -8.928743e-03  1.956225e-03
##    var.vmed.3      diaLunes     diaMartes  diaMiercoles     diaJueves
##  1.142694e-02  4.258230e-03  3.889016e-03  4.246557e-03  4.485114e-03
##    diaViernes     diaSabado     n.festivo
##  3.411747e-03 -8.089245e-03 -9.857227e-03
```

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

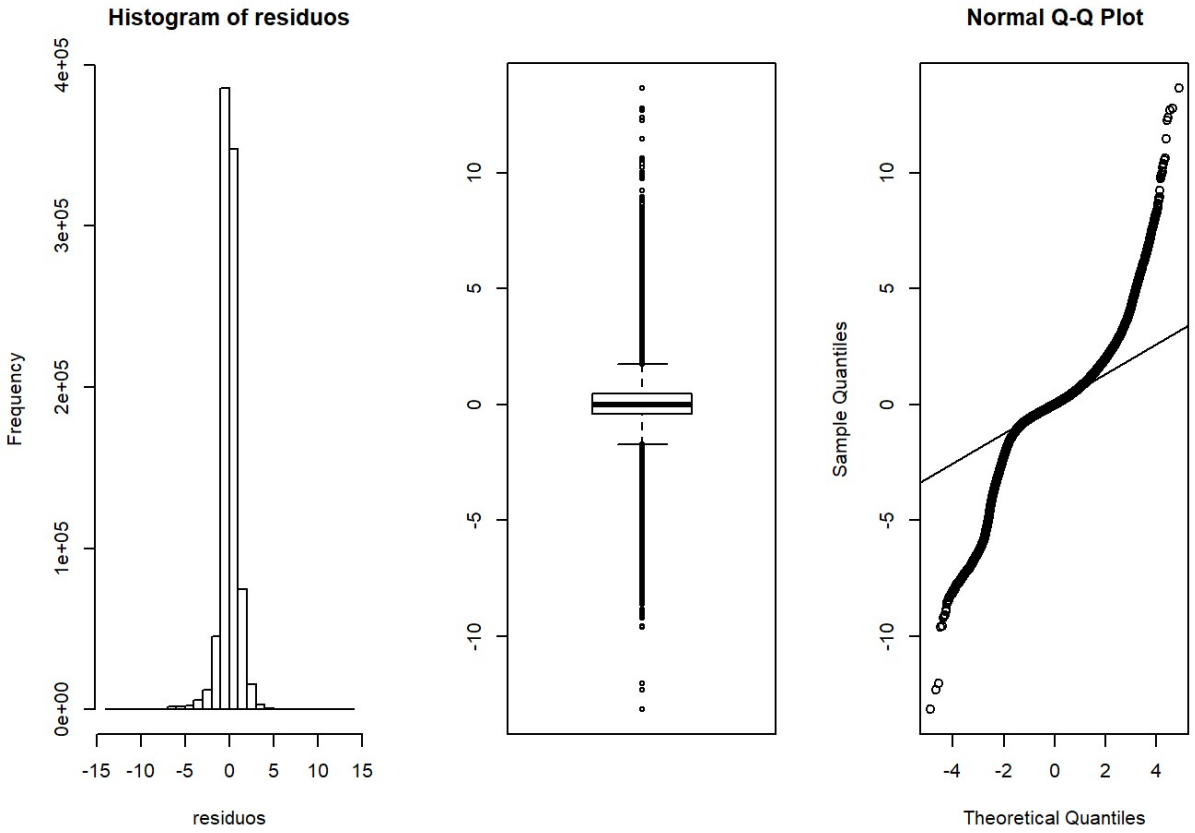
```
##
## Call:
## lm(formula = carga ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + vmed.3 + diaMes + Mes + prec_norm +
##     var.carga.3 + var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes +
##     diaMartes + diaMiercoles + diaJueves + diaViernes + diaSabado +
##     n.festivo, data = train.analisis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17217 -0.03682 -0.00098  0.04039  1.18693
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   4.751e-03  8.662e-04    5.485 4.15e-08 ***
## vel.med      -2.023e-04  1.578e-04   -1.282 0.199706
## carga.med     1.551e-01  1.240e-03  125.103  < 2e-16 ***
## carga.1       1.179e+00  1.951e-03  604.388  < 2e-16 ***
## vmed.1       -1.131e-01  1.886e-03  -59.957  < 2e-16 ***
## carga.2       6.493e-02  2.869e-03   22.635  < 2e-16 ***
## vmed.2        1.246e-01  1.884e-03   66.137  < 2e-16 ***
## carga.3      -4.045e-01  1.991e-03 -203.204  < 2e-16 ***
## vmed.3               NA         NA       NA       NA
## diaMes       -1.515e-04  1.079e-05  -14.046  < 2e-16 ***
## Mes          -9.602e-05  2.758e-05   -3.482 0.000498 ***
## prec_norm     4.666e-03  6.736e-04    6.926 4.33e-12 ***
## var.carga.3  -3.053e-01  1.925e-03 -158.629  < 2e-16 ***
## var.vmed.1   -8.929e-03  5.393e-04  -16.557  < 2e-16 ***
## var.vmed.2    1.956e-03  1.426e-05  137.208  < 2e-16 ***
## var.vmed.3    1.143e-02  2.016e-04   56.676  < 2e-16 ***
## diaLunes      4.258e-03  4.542e-04    9.375  < 2e-16 ***
## diaMartes     3.889e-03  4.537e-04    8.571  < 2e-16 ***
## diaMiercoles  4.247e-03  4.565e-04    9.302  < 2e-16 ***
## diaJueves     4.485e-03  4.577e-04    9.799  < 2e-16 ***
## diaViernes    3.412e-03  4.488e-04    7.601 2.94e-14 ***
## diaSabado    -8.089e-03  3.830e-04  -21.123  < 2e-16 ***
## n.festivo    -9.857e-03  1.616e-04  -61.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08902 on 897413 degrees of freedom
## Multiple R-squared:  0.8755, Adjusted R-squared:  0.8755
## F-statistic: 3.006e+05 on 21 and 897413 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_45min_noprec.RData')
```

Análisis de los residuos

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

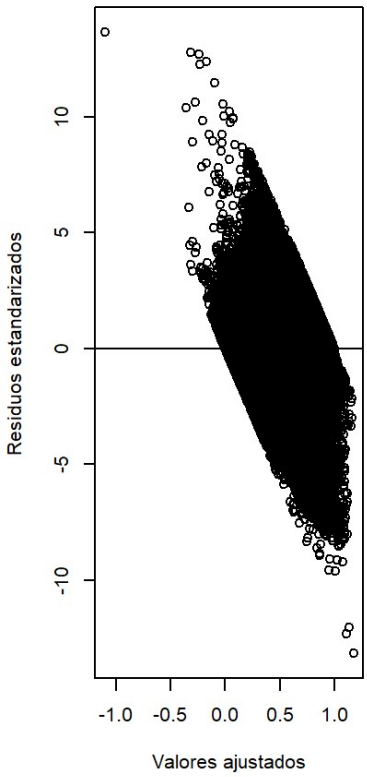
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados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

# 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: 3.600795e-23
```

b. Test

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

```
## Warning in predict.lm(lm.M30, test.analisis, interval = "prediction"):
## prediction from a rank-deficient fit may be misleading
```

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

```
##         error
## 1: 0.04235311
```