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
##  4.993267e-03 -2.278846e-04  1.555650e-01  1.176612e+00 -1.136409e-01
##       carga.2        vmed.2       carga.3        diaMes           Mes
##  6.663942e-02  1.255978e-01 -4.039960e-01 -1.466079e-04 -8.393701e-05
##   var.carga.3    var.vmed.1    var.vmed.2    var.vmed.3      diaLunes
## -3.068721e-01 -8.511144e-03  1.957845e-03  1.107859e-02  4.206273e-03
##     diaMartes  diaMiercoles     diaJueves    diaViernes     diaSabado
##  3.756091e-03  3.633489e-03  4.187883e-03  3.108988e-03 -8.309536e-03
##     n.festivo
## -9.952902e-03
```

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

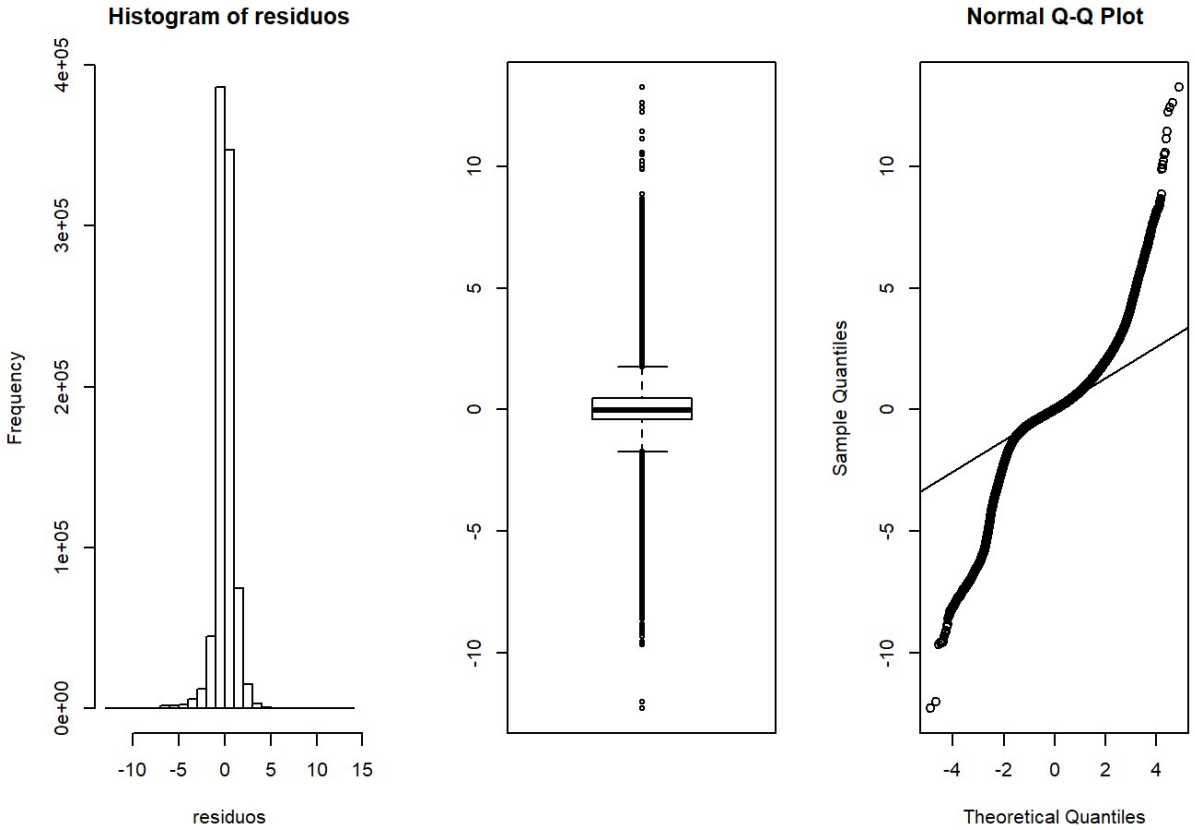
```
##
## Call:
## lm(formula = carga ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + diaMes + Mes + var.carga.3 +
##     var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes + diaMartes +
##     diaMiercoles + diaJueves + diaViernes + diaSabado + n.festivo,
##     data = train.analisis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09483 -0.03681 -0.00106  0.04035  1.15358
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   4.993e-03  8.652e-04    5.771 7.88e-09 ***
## vel.med      -2.279e-04  1.578e-04   -1.444  0.14879
## carga.med     1.556e-01  1.241e-03  125.317  < 2e-16 ***
## carga.1       1.177e+00  1.953e-03  602.473  < 2e-16 ***
## vmed.1       -1.136e-01  1.878e-03  -60.508  < 2e-16 ***
## carga.2       6.664e-02  2.869e-03   23.223  < 2e-16 ***
## vmed.2        1.256e-01  1.874e-03   67.010  < 2e-16 ***
## carga.3      -4.040e-01  1.989e-03 -203.084  < 2e-16 ***
## diaMes       -1.466e-04  1.078e-05  -13.599  < 2e-16 ***
## Mes          -8.394e-05  2.756e-05   -3.045  0.00233 **
## var.carga.3  -3.069e-01  1.928e-03 -159.147  < 2e-16 ***
## var.vmed.1   -8.511e-03  5.348e-04  -15.914  < 2e-16 ***
## var.vmed.2    1.958e-03  1.424e-05  137.531  < 2e-16 ***
## var.vmed.3    1.108e-02  2.001e-04   55.363  < 2e-16 ***
## diaLunes      4.206e-03  4.546e-04    9.254  < 2e-16 ***
## diaMartes     3.756e-03  4.543e-04    8.269  < 2e-16 ***
## diaMiercoles  3.633e-03  4.568e-04    7.955 1.79e-15 ***
## diaJueves     4.188e-03  4.577e-04    9.150  < 2e-16 ***
## diaViernes    3.109e-03  4.494e-04    6.918 4.57e-12 ***
## diaSabado    -8.310e-03  3.827e-04  -21.715  < 2e-16 ***
## n.festivo    -9.953e-03  1.618e-04  -61.505  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08902 on 897414 degrees of freedom
## Multiple R-squared:  0.8755, Adjusted R-squared:  0.8755
## F-statistic: 3.156e+05 on 20 and 897414 DF,  p-value: < 2.2e-16
```

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

Análisis de los residuos

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

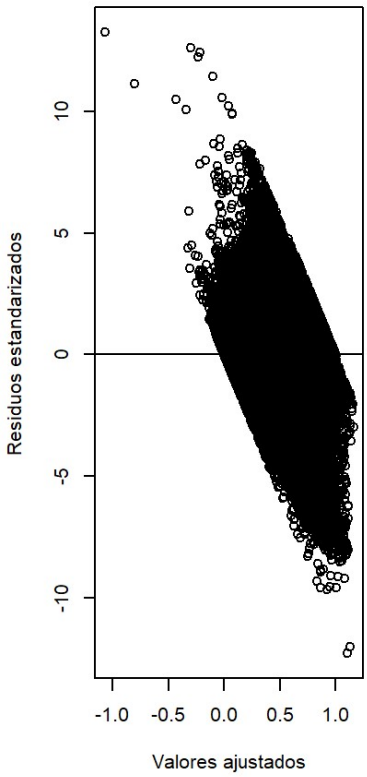
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
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: 3.529178e-23
```

b. Test

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

```
##          error
## 1: 0.008653125
```