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.30/100,
                                     vmed.1  = vmed.30/100,
                                     carga.2 = carga.45/100,
                                     vmed.2  = vmed.45/100,
                                     carga.3 = carga.60/100,
                                     vmed.3  = vmed.60/100,
                                     carga.4 = carga.75/100,
                                     vmed.4  = vmed.75/100)]
```

```
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
## -1.720075e-02 -3.552149e-03  8.111758e-02  1.149897e+00 -7.596341e-02
##       carga.2        vmed.2       carga.3        vmed.3        diaMes
## -2.676290e-02  2.829991e-02 -2.014679e-01  7.581761e-02 -8.298892e-05
##   var.carga.3    var.vmed.3      diaLunes     diaMartes  diaMiercoles
## -2.656433e-01  1.215987e-02  1.888950e-03  1.755724e-03  1.551237e-03
##     diaJueves    diaViernes     diaSabado     n.festivo
##  1.817479e-03  2.006558e-03 -3.919736e-03 -4.955152e-03
```

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

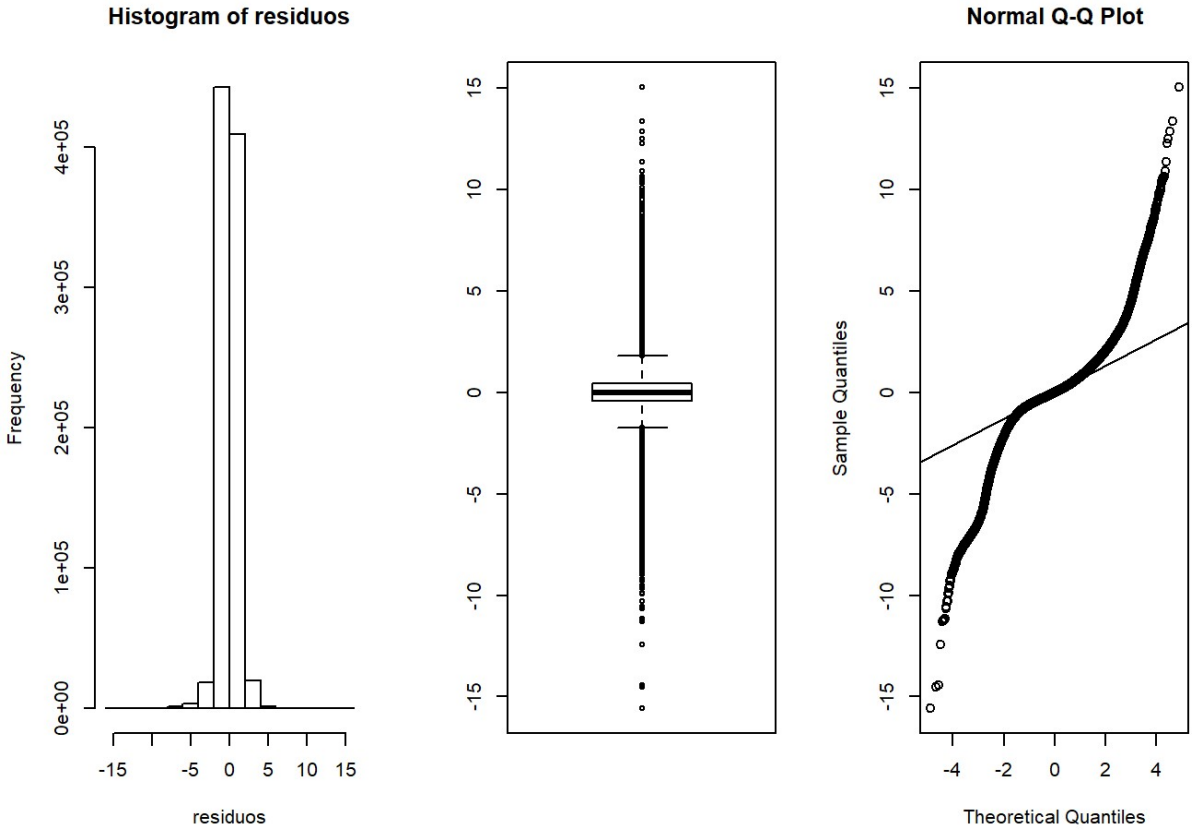
```
##
## Call:
## lm(formula = carga ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + vmed.3 + diaMes + var.carga.3 +
##     var.vmed.3 + diaLunes + diaMartes + diaMiercoles + diaJueves +
##     diaViernes + diaSabado + n.festivo, data = train.analisis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10681 -0.03033 -0.00226  0.03227  1.06623
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -1.720e-02  6.385e-04  -26.941  < 2e-16 ***
## vel.med      -3.552e-03  1.235e-04  -28.767  < 2e-16 ***
## carga.med     8.112e-02  9.806e-04   82.721  < 2e-16 ***
## carga.1       1.150e+00  1.552e-03  740.823  < 2e-16 ***
## vmed.1       -7.596e-02  1.174e-03  -64.704  < 2e-16 ***
## carga.2      -2.676e-02  2.289e-03  -11.692  < 2e-16 ***
## vmed.2        2.830e-02  1.441e-03   19.642  < 2e-16 ***
## carga.3      -2.015e-01  1.545e-03 -130.396  < 2e-16 ***
## vmed.3        7.582e-02  1.186e-03   63.940  < 2e-16 ***
## diaMes       -8.299e-05  8.602e-06   -9.648  < 2e-16 ***
## var.carga.3  -2.656e-01  1.538e-03 -172.668  < 2e-16 ***
## var.vmed.3    1.216e-02  3.266e-04   37.236  < 2e-16 ***
## diaLunes      1.889e-03  3.616e-04    5.224 1.75e-07 ***
## diaMartes     1.756e-03  3.614e-04    4.858 1.18e-06 ***
## diaMiercoles  1.551e-03  3.636e-04    4.267 1.98e-05 ***
## diaJueves     1.817e-03  3.641e-04    4.991 6.00e-07 ***
## diaViernes    2.007e-03  3.575e-04    5.613 1.99e-08 ***
## diaSabado    -3.920e-03  3.046e-04  -12.868  < 2e-16 ***
## n.festivo    -4.955e-03  1.276e-04  -38.844  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07103 on 897416 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9207
## F-statistic: 5.788e+05 on 18 and 897416 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_30min_noprec.RData')
```

Análisis de los residuos

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

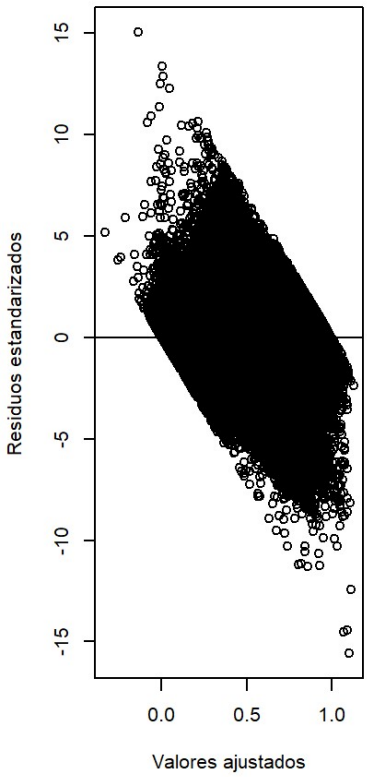
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados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



b. Supuesto 2: Varianza de los errores es constante:
    - No es constante, tiene tendencia lo que indica que hay una variable desconocida 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: 1.312595e-23
```

b. Test

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

```
##         error
## 1: 0.03269643
```