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.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(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
## -3.0340437503  0.1287930239 -0.0168229984 -0.2885525810  3.9363716831
##       carga.2        vmed.2       carga.3        vmed.3        diaMes
##  0.3601985871  0.8931995241 -0.0551626885 -0.4986248570 -0.0002017646
##           Mes     prec_norm    var.vmed.1    var.vmed.2    var.vmed.3
## -0.0028871629 -0.0309875018  0.0321202267  0.0316816479 -0.1197916634
##      diaLunes     diaMartes  diaMiercoles     diaJueves    diaViernes
## -0.0488924734 -0.0541916606 -0.0549868985 -0.0588716683 -0.0533993242
##     n.festivo
##  0.0162367786
```

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

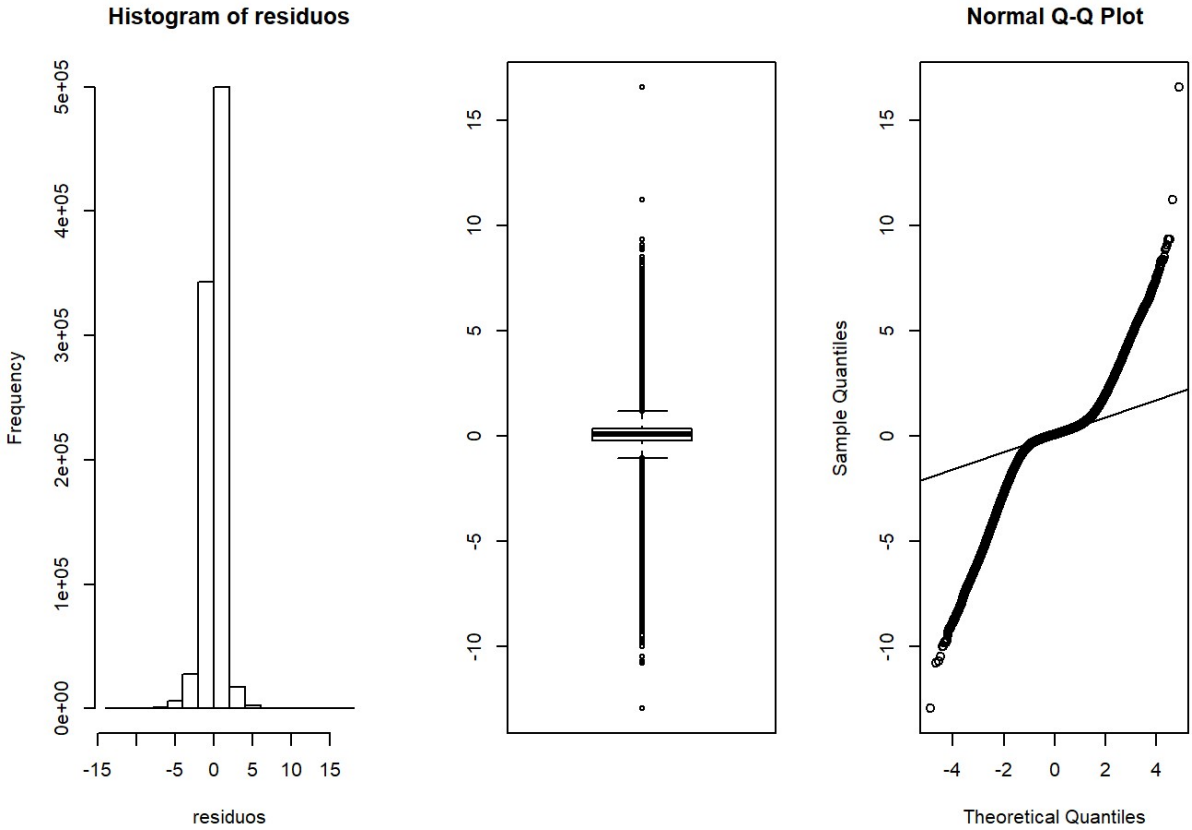
```
##
## Call:
## lm(formula = vmed ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + vmed.3 + diaMes + Mes + prec_norm +
##     var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes + diaMartes +
##     diaMiercoles + diaJueves + diaViernes + n.festivo, data = train.anali
sis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6199 -0.0909  0.0330  0.1519  7.1932
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -3.034e+00  3.902e-03 -777.627  < 2e-16 ***
## vel.med       1.288e-01  7.580e-04  169.915  < 2e-16 ***
## carga.med    -1.682e-02  5.986e-03   -2.811 0.004946 **
## carga.1      -2.886e-01  9.394e-03  -30.717  < 2e-16 ***
## vmed.1        3.936e+00  9.298e-03  423.357  < 2e-16 ***
## carga.2       3.602e-01  1.396e-02   25.797  < 2e-16 ***
## vmed.2        8.932e-01  1.251e-02   71.405  < 2e-16 ***
## carga.3      -5.516e-02  9.386e-03   -5.877 4.18e-09 ***
## vmed.3       -4.986e-01  1.007e-02  -49.500  < 2e-16 ***
## diaMes       -2.018e-04  5.261e-05   -3.835 0.000125 ***
## Mes          -2.887e-03  1.344e-04  -21.487  < 2e-16 ***
## prec_norm    -3.099e-02  3.273e-03   -9.469  < 2e-16 ***
## var.vmed.1    3.212e-02  2.649e-03   12.127  < 2e-16 ***
## var.vmed.2    3.168e-02  2.801e-03   11.311  < 2e-16 ***
## var.vmed.3   -1.198e-01  2.039e-03  -58.759  < 2e-16 ***
## diaLunes     -4.889e-02  1.790e-03  -27.310  < 2e-16 ***
## diaMartes    -5.419e-02  1.774e-03  -30.552  < 2e-16 ***
## diaMiercoles -5.499e-02  1.793e-03  -30.675  < 2e-16 ***
## diaJueves    -5.887e-02  1.802e-03  -32.676  < 2e-16 ***
## diaViernes   -5.340e-02  1.753e-03  -30.463  < 2e-16 ***
## n.festivo     1.624e-02  7.303e-04   22.234  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4341 on 897414 degrees of freedom
## Multiple R-squared:  0.7933, Adjusted R-squared:  0.7933
## F-statistic: 1.722e+05 on 20 and 897414 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_vel_30min.RData')
```

Análisis de los residuos

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

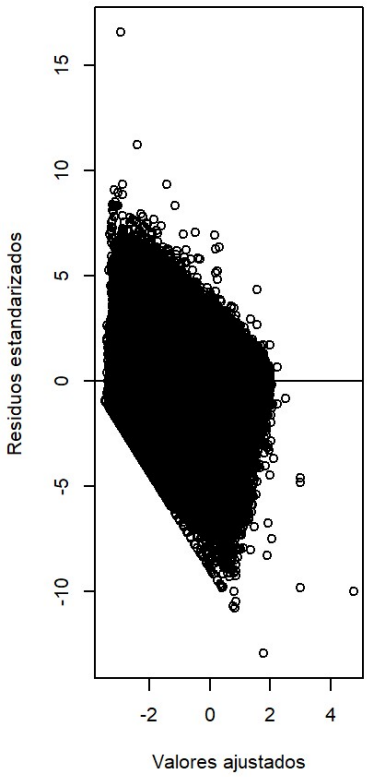
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
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: 62885.49
```

b. Test

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

```
##       error
## 1: 26829.01
```