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.15/100,
                                     vmed.1  = vmed.15/100,
                                     carga.2 = carga.30/100,
                                     vmed.2  = vmed.30/100,
                                     carga.3 = carga.45/100,
                                     vmed.3  = vmed.45/100,
                                     carga.4 = carga.60/100,
                                     vmed.4  = vmed.60/100)]
```

```
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
## -3.4010138310  0.0713555924 -0.0072786930 -0.3969904755  3.9679834995
##       carga.2        vmed.2       carga.3        vmed.3        diaMes
##  0.4020546381  0.9812275323 -0.0002096483 -0.1398396262 -0.0000899899
##           Mes   var.carga.3    var.vmed.1    var.vmed.2    var.vmed.3
## -0.0014301745 -0.0358820257  0.0105172304  0.0314808017 -0.0555012299
##      diaLunes     diaMartes  diaMiercoles     diaJueves    diaViernes
## -0.0258390879 -0.0305155824 -0.0305386742 -0.0323292450 -0.0293478299
##     n.festivo
##  0.0090577541
```

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

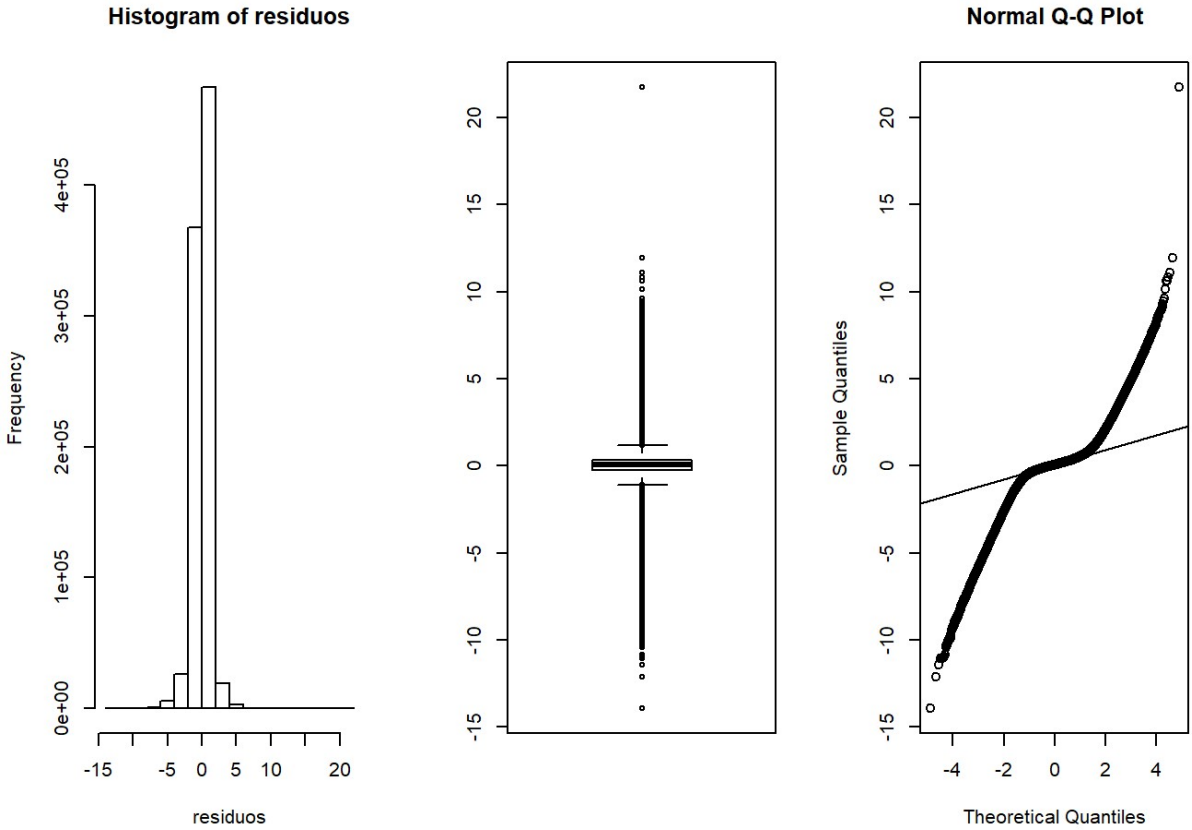
```
##
## Call:
## lm(formula = vmed ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + vmed.3 + diaMes + Mes + var.carga.3 +
##     var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes + diaMartes +
##     diaMiercoles + diaJueves + diaViernes + n.festivo, data = train.anali
sis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8793 -0.0822  0.0183  0.1172  7.6013
##
## Coefficients:
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  -3.401e+00  3.140e-03 -1083.183  < 2e-16 ***
## vel.med       7.136e-02  6.110e-04   116.778  < 2e-16 ***
## carga.med    -7.279e-03  4.828e-03    -1.508   0.1316
## carga.1      -3.970e-01  7.632e-03   -52.017  < 2e-16 ***
## vmed.1        3.968e+00  7.591e-03   522.707  < 2e-16 ***
## carga.2       4.021e-01  1.125e-02    35.749  < 2e-16 ***
## vmed.2        9.812e-01  1.020e-02    96.160  < 2e-16 ***
## carga.3      -2.096e-04  7.594e-03    -0.028   0.9780
## vmed.3       -1.398e-01  8.209e-03   -17.035  < 2e-16 ***
## diaMes       -8.999e-05  4.234e-05    -2.125   0.0336 *
## Mes          -1.430e-03  1.083e-04   -13.211  < 2e-16 ***
## var.carga.3  -3.588e-02  7.570e-03    -4.740 2.14e-06 ***
## var.vmed.1    1.052e-02  2.196e-03     4.789 1.68e-06 ***
## var.vmed.2    3.148e-02  2.315e-03    13.596  < 2e-16 ***
## var.vmed.3   -5.550e-02  1.611e-03   -34.453  < 2e-16 ***
## diaLunes     -2.584e-02  1.440e-03   -17.938  < 2e-16 ***
## diaMartes    -3.052e-02  1.428e-03   -21.367  < 2e-16 ***
## diaMiercoles -3.054e-02  1.439e-03   -21.221  < 2e-16 ***
## diaJueves    -3.233e-02  1.444e-03   -22.385  < 2e-16 ***
## diaViernes   -2.935e-02  1.413e-03   -20.774  < 2e-16 ***
## n.festivo     9.058e-03  5.890e-04    15.378  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3498 on 897414 degrees of freedom
## Multiple R-squared:  0.8657, Adjusted R-squared:  0.8657
## F-statistic: 2.893e+05 on 20 and 897414 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_vel_15min_noprec.RData')
```

Análisis de los residuos

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

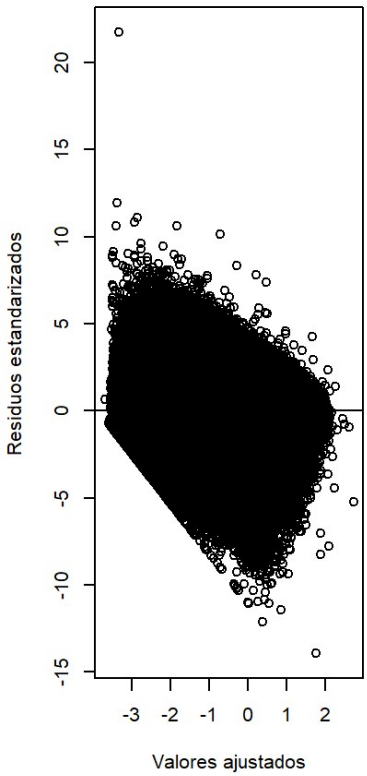
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados
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: 62622.68
```

b. Test

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

```
##       error
## 1: 27321.73
```