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
## -3.0436035797  0.1277091698 -0.0170350599 -0.2821951215  3.9132785227
##       carga.2        vmed.2       carga.3        vmed.3        diaMes
##  0.3360403731  0.9198929053 -0.0367366141 -0.4940389131 -0.0002009038
##           Mes    var.vmed.1    var.vmed.2    var.vmed.3      diaLunes
## -0.0028471409  0.0214274067  0.0317838699 -0.1222766509 -0.0454635948
##     diaMartes  diaMiercoles     diaJueves    diaViernes     n.festivo
## -0.0528723118 -0.0539224778 -0.0579389114 -0.0522358168  0.0165769202
```

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

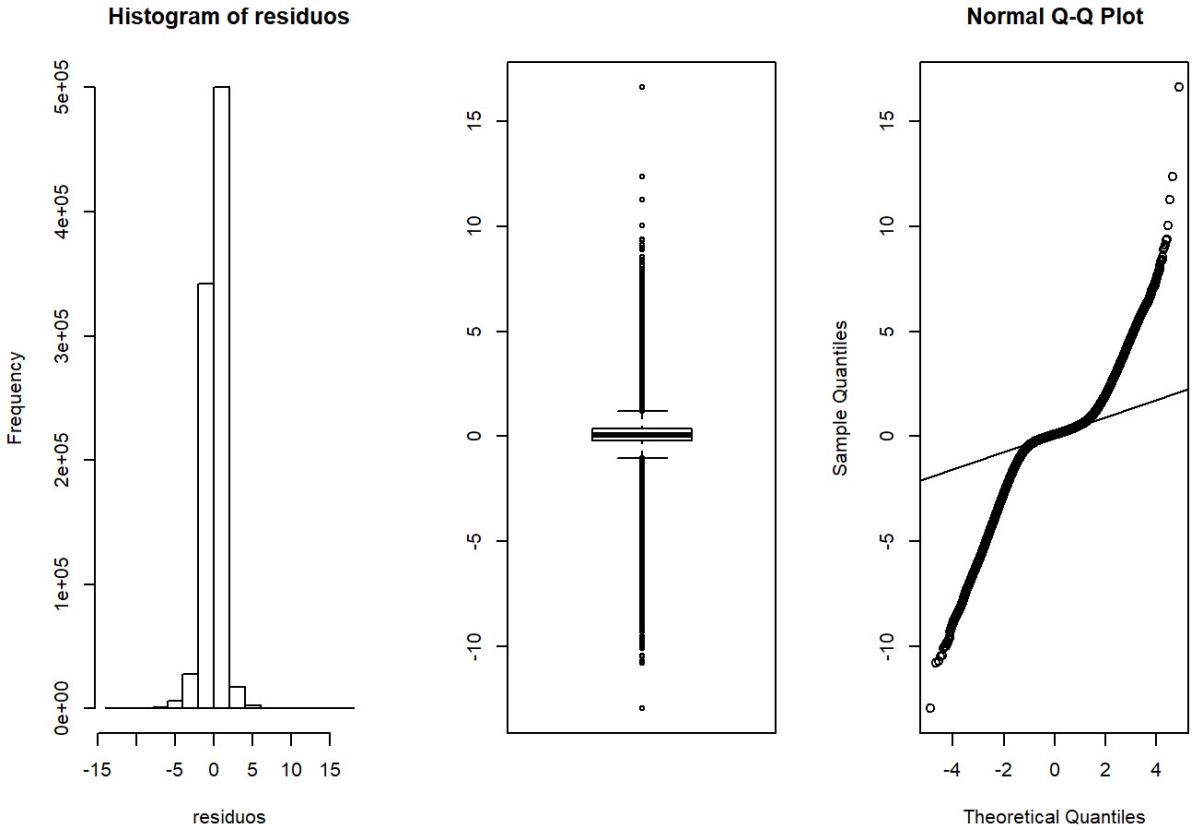
```
##
## Call:
## lm(formula = vmed ~ vel.med + carga.med + carga.1 + vmed.1 +
##     carga.2 + vmed.2 + carga.3 + vmed.3 + diaMes + Mes + var.vmed.1 +
##     var.vmed.2 + var.vmed.3 + diaLunes + diaMartes + diaMiercoles +
##     diaJueves + diaViernes + n.festivo, data = train.analisis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6155 -0.0906  0.0333  0.1518  7.1995
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -3.0436036  0.0038848 -783.469  < 2e-16 ***
## vel.med       0.1277092  0.0007570  168.711  < 2e-16 ***
## carga.med    -0.0170351  0.0059758   -2.851  0.00436 **
## carga.1      -0.2821951  0.0093748  -30.101  < 2e-16 ***
## vmed.1        3.9132785  0.0095875  408.163  < 2e-16 ***
## carga.2       0.3360404  0.0139519   24.086  < 2e-16 ***
## vmed.2        0.9198929  0.0128260   71.721  < 2e-16 ***
## carga.3      -0.0367366  0.0093735   -3.919 8.88e-05 ***
## vmed.3       -0.4940389  0.0101974  -48.448  < 2e-16 ***
## diaMes       -0.0002009  0.0000525   -3.827  0.00013 ***
## Mes          -0.0028471  0.0001343  -21.207  < 2e-16 ***
## var.vmed.1    0.0214274  0.0028415    7.541 4.67e-14 ***
## var.vmed.2    0.0317839  0.0028875   11.007  < 2e-16 ***
## var.vmed.3   -0.1222767  0.0020737  -58.966  < 2e-16 ***
## diaLunes     -0.0454636  0.0017863  -25.452  < 2e-16 ***
## diaMartes    -0.0528723  0.0017712  -29.851  < 2e-16 ***
## diaMiercoles -0.0539225  0.0017851  -30.208  < 2e-16 ***
## diaJueves    -0.0579389  0.0017914  -32.343  < 2e-16 ***
## diaViernes   -0.0522358  0.0017508  -29.836  < 2e-16 ***
## n.festivo     0.0165769  0.0007305   22.693  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4337 on 897415 degrees of freedom
## Multiple R-squared:  0.7936, Adjusted R-squared:  0.7936
## F-statistic: 1.816e+05 on 19 and 897415 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_vel_30min_noprec.RData')
```

Análisis de los residuos

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

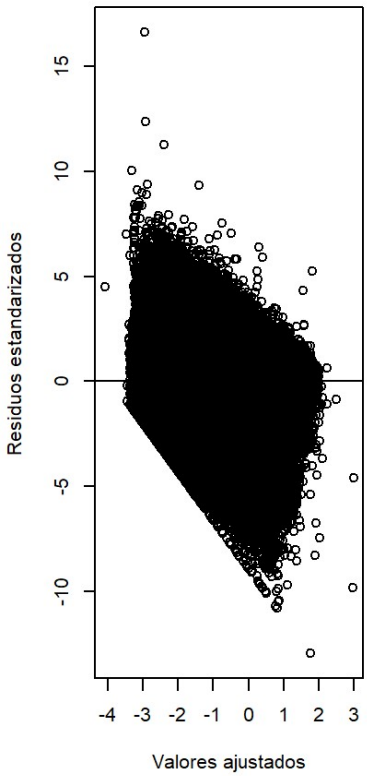
```
residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (compl
eto)
par(mfrow=c(1,3))
hist(residuos) # histograma de los residuos estandarizados
boxplot(residuos) # diagrama de cajas de los residuos estandarizados
qqnorm(residuos) # gráfico de cuantiles de los residuos estandarizados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




```
b. Supuesto 2: Varianza de los errores es constante:
    - No es constante, tiene tendencia lo que indica que hay una variable 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: 62937.32
```

b. Test

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

```
##       error
## 1: 26937.77
```