

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.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(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
## -3.567617e-03 -1.011868e-03  3.577770e-02  1.082067e+00 -5.027458e-02
##      carga.2      vmed.2      carga.3      vmed.3      diaMes
## -2.991414e-03  2.068452e-02 -1.142287e-01  3.768158e-02 -3.860199e-05
##   var.carga.3   var.vmed.3      diaLunes   diaViernes   diaSabado
## -1.003185e-01  5.562321e-03  3.002502e-04  1.371549e-04 -2.343963e-03
##      n.festivo
## -2.342737e-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 + diaViernes + diaSabado + n.festivo,
##      data = train.analisis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85400 -0.02258 -0.00139  0.02336  1.02987
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.568e-03  3.779e-04  -9.441 < 2e-16 ***
## vel.med      -1.012e-03  8.451e-05 -11.974 < 2e-16 ***
## carga.med     3.578e-02  6.717e-04  53.261 < 2e-16 ***
## carga.1       1.082e+00  1.064e-03 1016.922 < 2e-16 ***
## vmed.1       -5.027e-02  8.039e-04 -62.541 < 2e-16 ***
## carga.2      -2.991e-03  1.565e-03  -1.911  0.0560 .
## vmed.2        2.068e-02  9.882e-04  20.931 < 2e-16 ***
## carga.3      -1.142e-01  1.057e-03 -108.111 < 2e-16 ***
## vmed.3        3.768e-02  8.132e-04  46.339 < 2e-16 ***
## diaMes       -3.860e-05  5.889e-06  -6.555 5.58e-11 ***
## var.carga.3  -1.003e-01  1.053e-03 -95.254 < 2e-16 ***
## var.vmed.3    5.562e-03  2.258e-04  24.634 < 2e-16 ***
## diaLunes      3.003e-04  1.572e-04   1.909  0.0562 .
## diaViernes    1.372e-04  1.512e-04   0.907  0.3645
## diaSabado    -2.344e-03  1.561e-04 -15.020 < 2e-16 ***
## n.festivo    -2.343e-03  6.415e-05 -36.517 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04868 on 897419 degrees of freedom
## Multiple R-squared:  0.9628, Adjusted R-squared:  0.9628
## F-statistic: 1.55e+06 on 15 and 897419 DF, p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30, './modelos/lmM30_15min_noprec.RData')
```

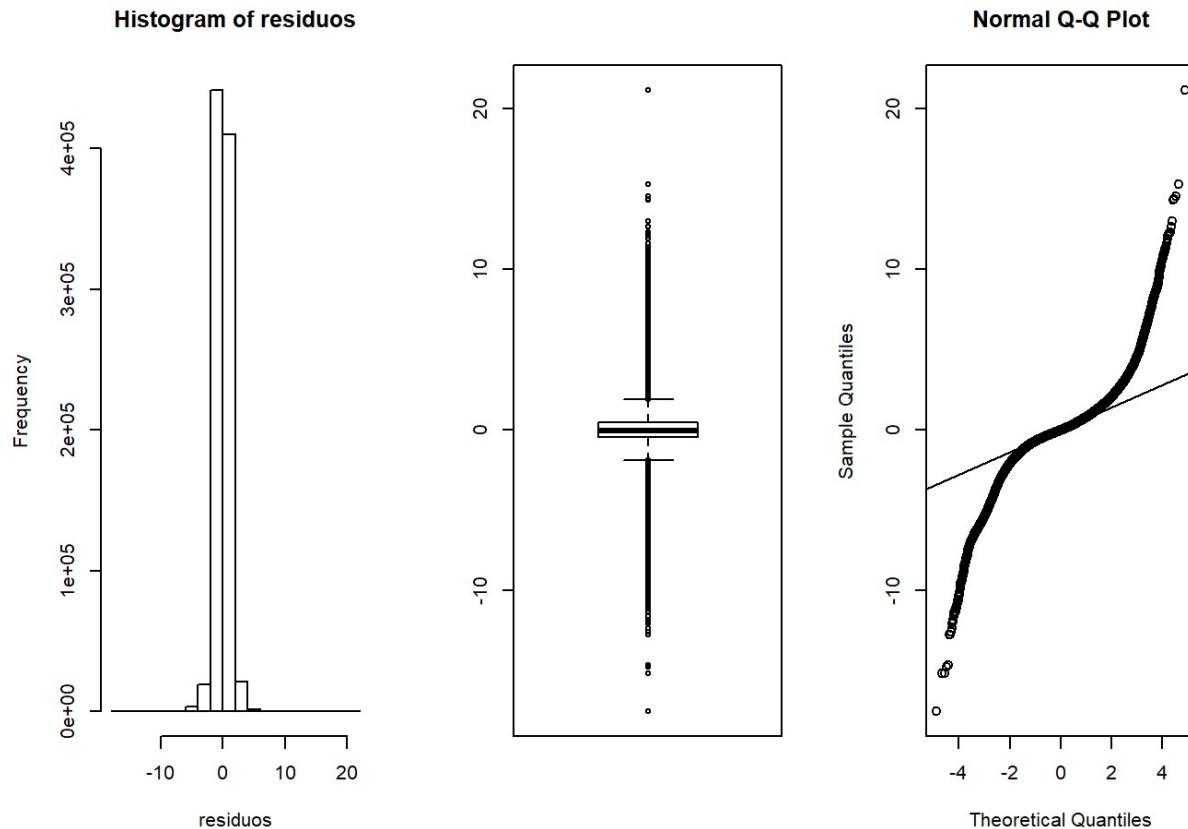
Análisis de los residuos

a. Supuesto 1: Normalidad

```

residuos<-rstandard(lm.M30) # residuos estándares del modelo ajustado (completo)
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:

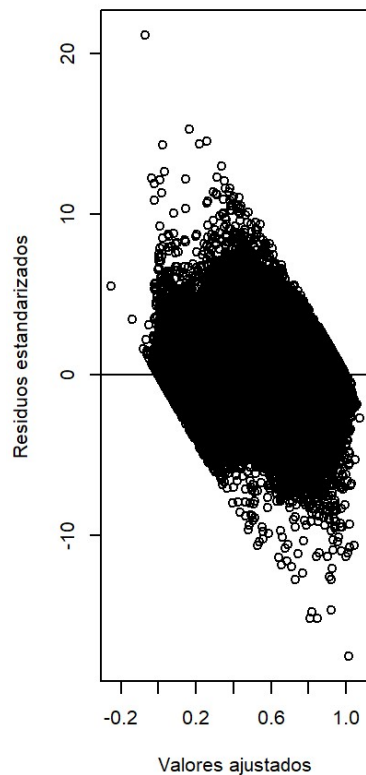
- No es constante, tiene tendencia lo que indica que hay una variable desconocida que impacta en el tráfico pero no la hemos detectado.

```

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 current data refer to _future_ responses
```

```
calculo_error(train.analysis, as.data.table(predict.M30))
```

```
##           error
## 1: 1.659433e-24
```

b. Test

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

```
##           error
## 1: 0.001970447
```