

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.45/100,
                                     vmed.1 = vmed.45/100,
                                     carga.2 = carga.60/100,
                                     vmed.2 = vmed.60/100,
                                     carga.3 = carga.75/100,
                                     vmed.3 = carga.75/100,
                                     carga.4 = carga.90/100,
                                     vmed.4 = vmed.90/100)]
```

```
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
## -2.7832742422  0.1598206517 -0.0938941740 -0.2138891975  3.7220265390
##      carga.2      vmed.2      carga.3      diaMes      Mes
##  0.2472373314  0.4126171502 -0.1154034094 -0.0001166685 -0.0034234844
##  var.carga.3  var.vmed.1  var.vmed.2  var.vmed.3  diaLunes
## -0.2737630625  0.0142725974 -0.0047033007 -0.1412171573 -0.0660238694
##  diaMartes  diaMiercoles  diaJueves  diaViernes  diaSabado
## -0.0747451987 -0.0753819256 -0.0796065681 -0.0707800389  0.0068626166
##      n.festivo
##  0.0282197338
```

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

```
##
## Call:
## lm(formula = vmed ~ vel.med + carga.med + carga.1 + vmed.1 +
##      carga.2 + vmed.2 + carga.3 + diaMes + Mes + var.carga.3 +
##      var.vmed.1 + var.vmed.2 + var.vmed.3 + diaLunes + diaMartes +
##      diaMiercoles + diaJueves + diaViernes + diaSabado + n.festivo,
##      data = train.analisis)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-10.6887	-0.0934	0.0534	0.1874	5.1428

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.783e+00	4.839e-03	-575.150	< 2e-16 ***
vel.med	1.598e-01	8.828e-04	181.032	< 2e-16 ***
carga.med	-9.389e-02	6.939e-03	-13.531	< 2e-16 ***
carga.1	-2.139e-01	1.091e-02	-19.596	< 2e-16 ***
vmed.1	3.722e+00	1.047e-02	355.503	< 2e-16 ***
carga.2	2.472e-01	1.604e-02	15.414	< 2e-16 ***
vmed.2	4.126e-01	1.046e-02	39.450	< 2e-16 ***
carga.3	-1.154e-01	1.116e-02	-10.341	< 2e-16 ***
diaMes	-1.167e-04	6.027e-05	-1.936	0.05290 .
Mes	-3.423e-03	1.541e-04	-22.216	< 2e-16 ***
var.carga.3	-2.738e-01	1.078e-02	-25.393	< 2e-16 ***
var.vmed.1	1.427e-02	2.968e-03	4.809	1.52e-06 ***
var.vmed.2	-4.703e-03	7.980e-05	-58.937	< 2e-16 ***
var.vmed.3	-1.412e-01	1.206e-03	-117.092	< 2e-16 ***
diaLunes	-6.602e-02	2.537e-03	-26.023	< 2e-16 ***
diaMartes	-7.475e-02	2.537e-03	-29.462	< 2e-16 ***
diaMiercoles	-7.538e-02	2.550e-03	-29.556	< 2e-16 ***
diaJueves	-7.961e-02	2.554e-03	-31.174	< 2e-16 ***
diaViernes	-7.078e-02	2.509e-03	-28.216	< 2e-16 ***
diaSabado	6.863e-03	2.139e-03	3.209	0.00133 **
n.festivo	2.822e-02	9.017e-04	31.297	< 2e-16 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4978 on 897414 degrees of freedom
## Multiple R-squared:  0.7283, Adjusted R-squared:  0.7283
## F-statistic: 1.203e+05 on 20 and 897414 DF,  p-value: < 2.2e-16
```

```
setwd(ruta)
saveRDS(lm.M30,'./modelos/lmM30_vel_45min_noprec.RData')
```

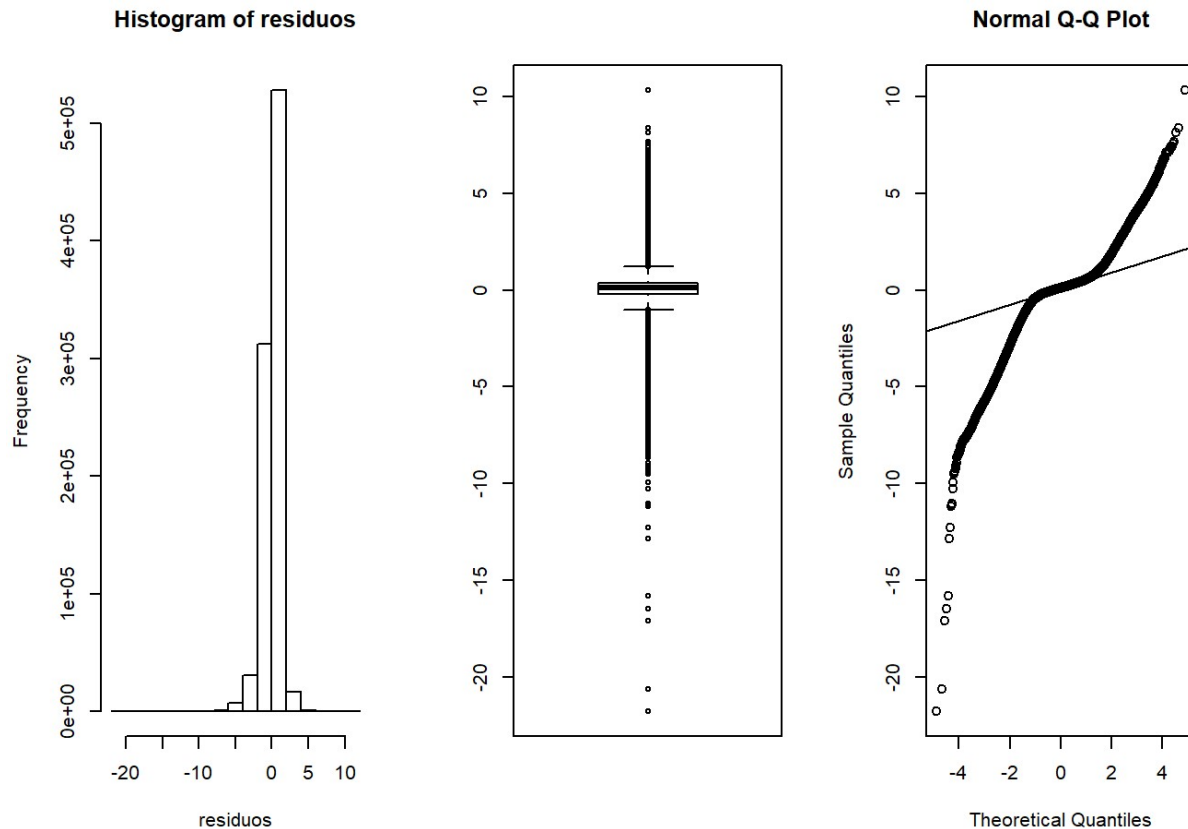
Análisis de los residuos

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:

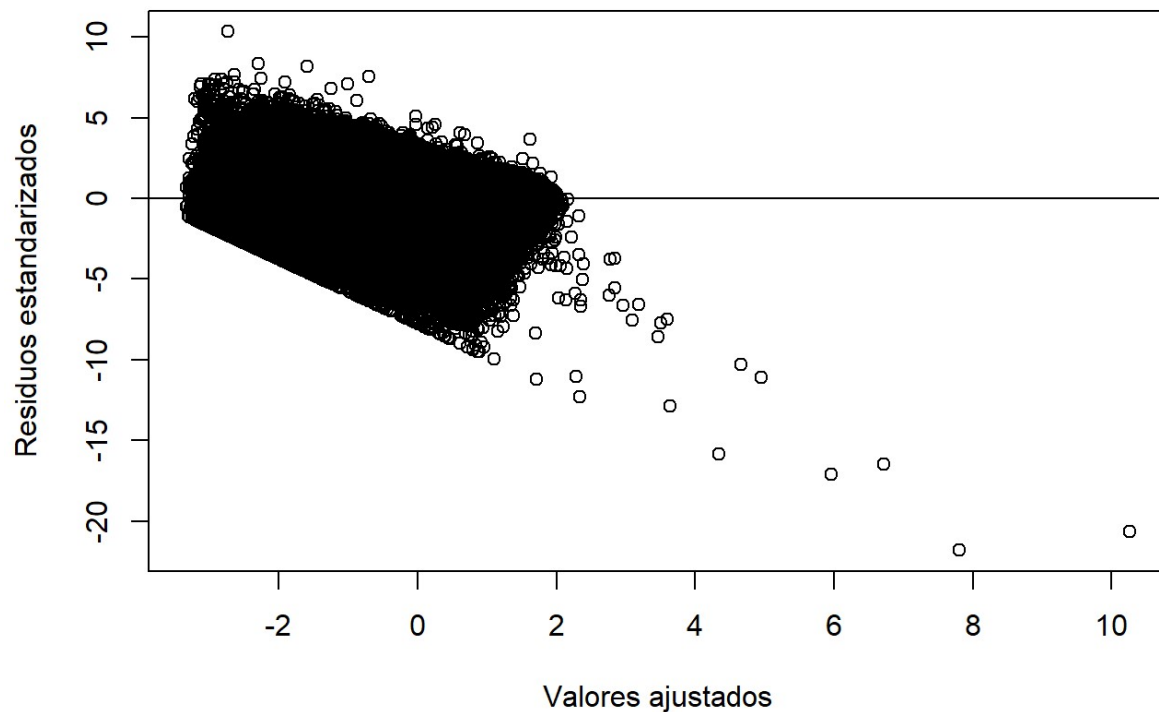
- No es constante, tiene tendencia lo que indica que hay una variable desconocida que impacta en el tráfico pero no la hemos detectado.

```

par(mfrow=c(1,1))

# 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: 63217.99
```

b. Test

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

```
##          error
## 1: 26673.72
```