Trabajo 2 TAE

Esteban Moreno Rodríguez-1152459914

6/1/2022

Planteamiento del problema.

Se busca construir un modelo de regresión que ayude a predecir el número de vehículos que se registrarán diariamente en el Registro Único Nacional de Tránsito (RUNT) para el 2018. Para esto se tienen los datos hisotóricos de registros de autos desde el 01/01/2012 hasta el 31/12/2017.

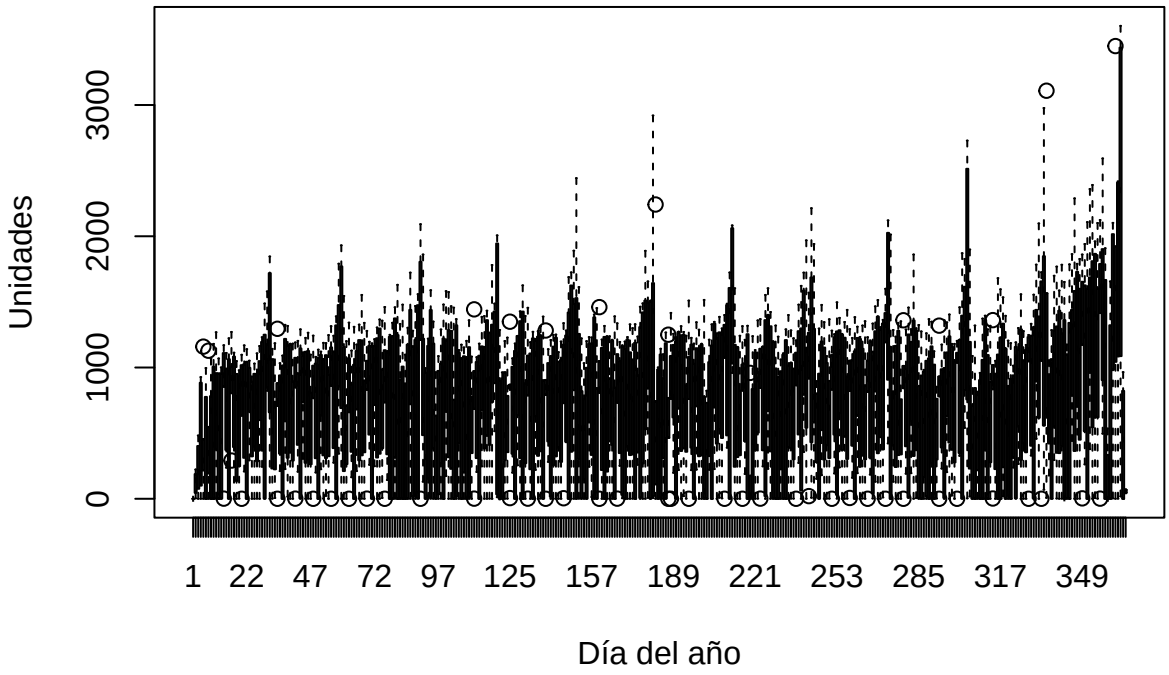
Datos

Fecha	Unidades	dia_semana	mes
2012-01-01	0	domingo	enero
2012-01-02	188	lunes	enero
2012-01-03	482	martes	enero
2012-01-04	927	miércoles	enero
2012-01-05	1159	jueves	enero
2012-01-06	996	viernes	enero

Los datos principales constan de las columnas **Fecha** y **Unidades**, adicionalmente se les agregó las columnas **dia_semana** (nombre del día de la semana correspondiente a la fecha) y **mes** (nombre del mes correspondiente a la fecha). esto con ayuda de *Excel*.Otras variables serán consideradas para ayudar a explicar el comportamiento de los datos, estás variables serán derivadas de la variable fecha.

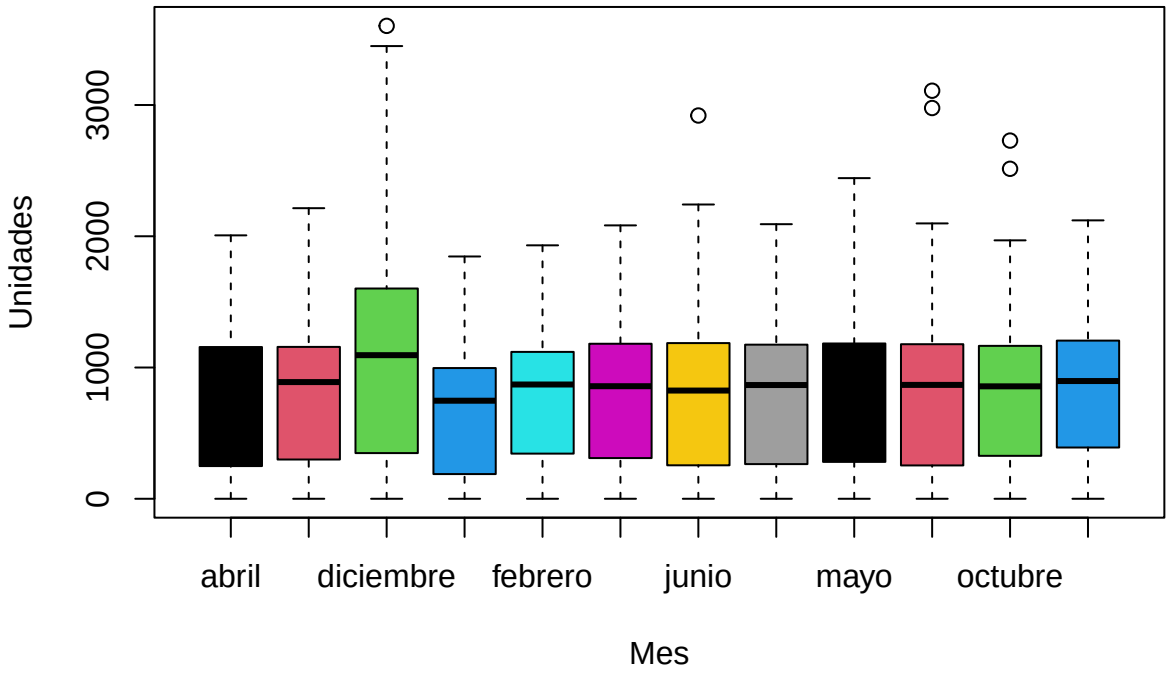
Veamos como se comportan las unidades registradas para cada una de las variables consideradas hata ahora.

Unidades registradas por día del año



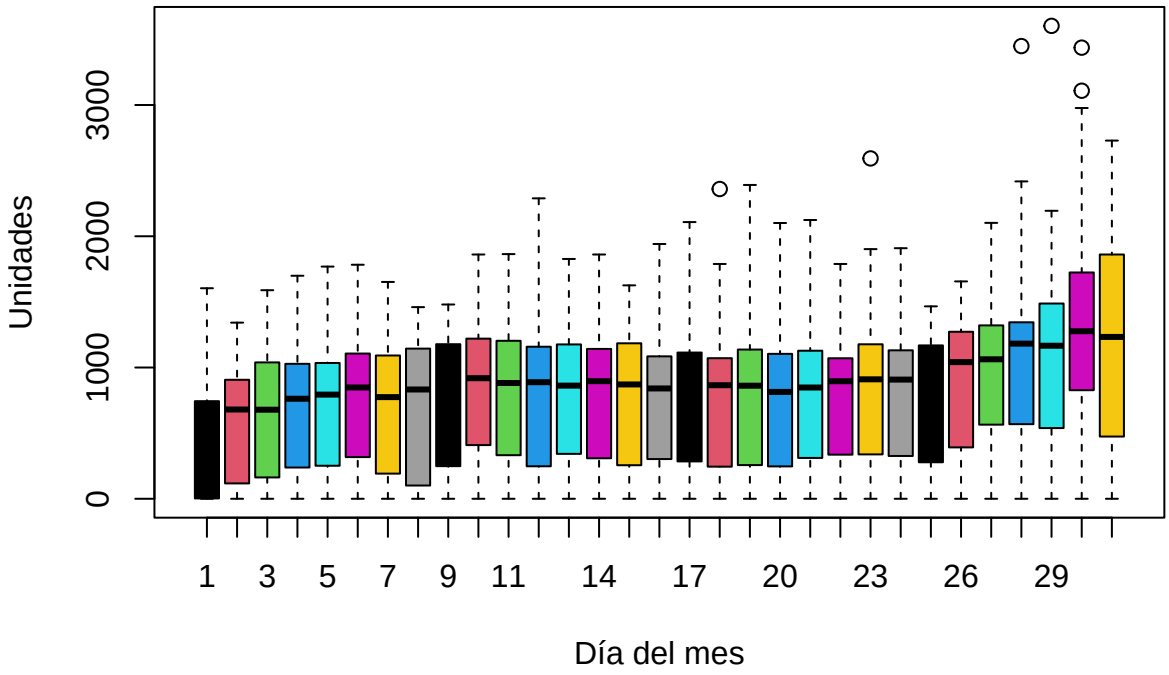
Se pueden ver unos picos que representan una gran cantidad de autos registrados a finales y comienzo de mes, sí parece haber una tendencia relacioanda con el tiempo.

Unidades registradas por mes



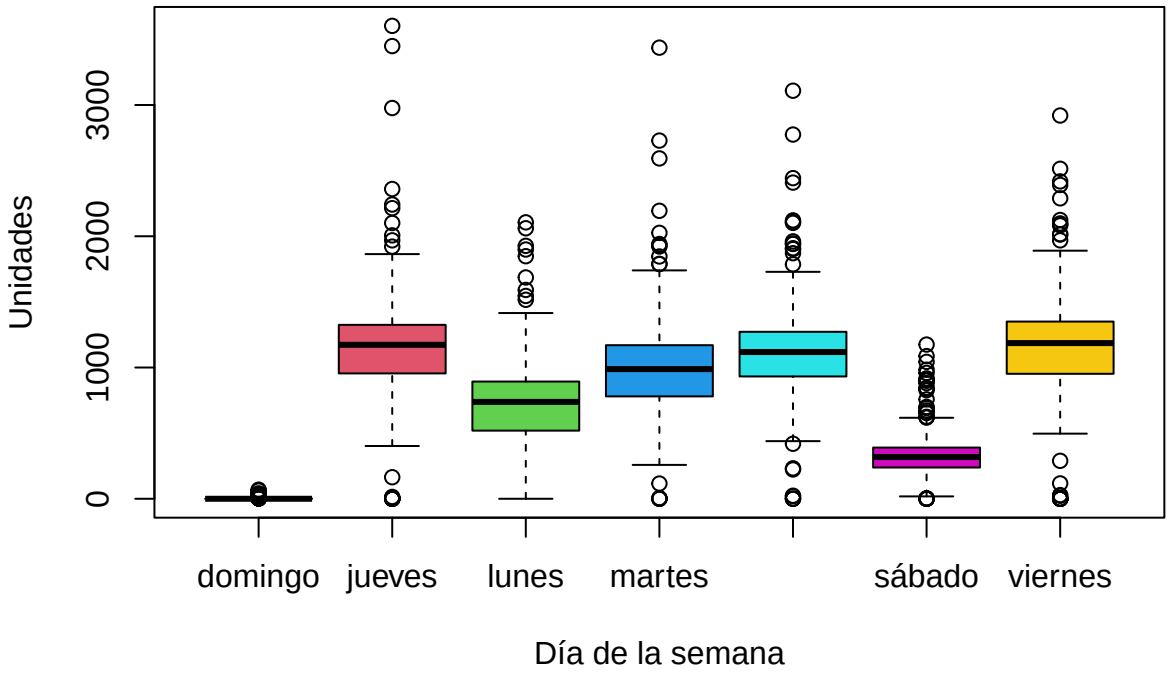
Al parecer no hay diferencias significativas de unidades registradas por mes para cada año ,pero se ve que en el mes de Diciembre se obtienen los valores más altos y la variabilidad más alta.

Unidades registradas por día del mes



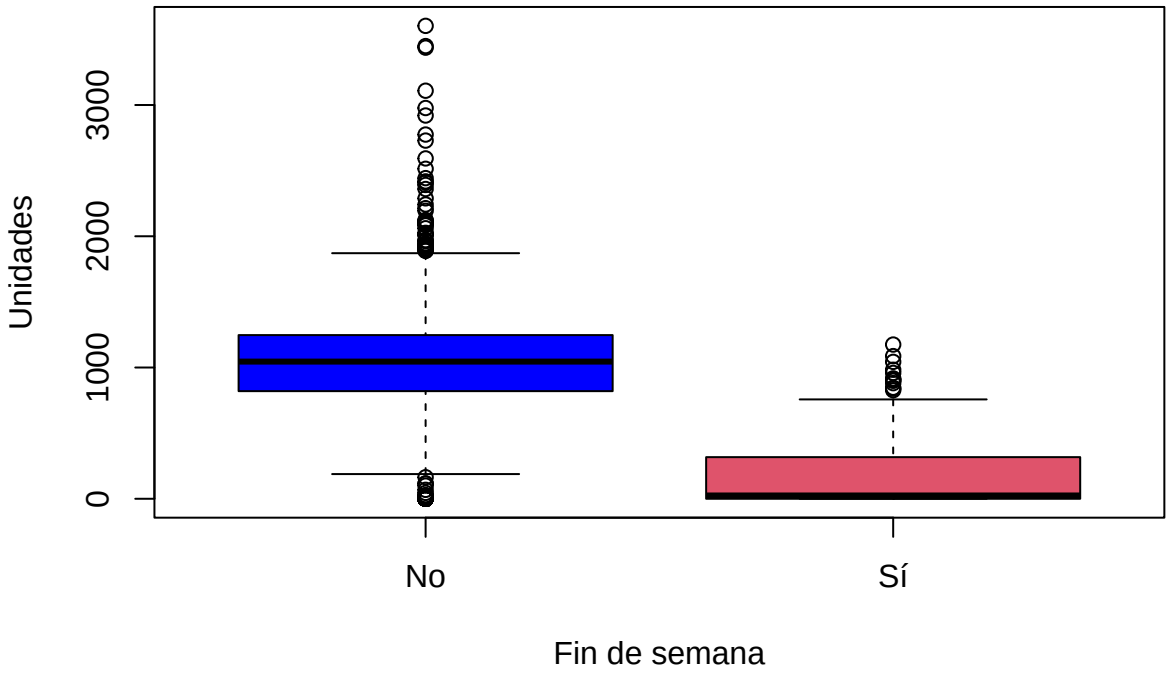
Se puede ver no hay diferencia entre las unidades registradas para los días de los meses para cada uno de los años, sinembargo se ve que en los últimos días de los meses se alcanzan los valores más altos.

Unidades registradas por semana



Se puede ver que los fines de semana son los días que menos autos se registran, eso puede deberse a que le Domingo es un día de descanso y que los sábados muchas personas trabajan hasta tempranas horas de la tarde. A partir de esto se crea una variable llamada **finde** (Fin de semana) que pudiera ayudar a aexplciar el comportamiento de los registros de las unidades.

Unidades registradas en semana y en fin de semana



Modelo de regresión

Con las variables anteriormente definidas se buscará crear un modelo de regresión que pueda explicar y posteriormente predecir las unidades de autos que se registran.

Para encontrar el mejor modelo de regresión se utiliza la regresión **Stepwise**, la cual nos permite escoger las mejores variables capaces de explicar el comportamiento de las unidades de autos registrados. El mejor modelo con el conjunto de variables indicado es escogido con el criterio **AIC**.

- Primero se intentará predecir y explicar las unidades registradas que corresponden a los años 2012 hasta 2016
- Segundo se intentará predecir las unidades registradas que corresponden al año 2017 para ver como predice el modelo
- Tercero se intentará predecir las unidades registradas para los días comprendidos entre el 01/01/2018 y el 30/06/2018.

Para esto se dividen los datos en dos grupos, datos desde los años 2012 hasta el 2016 (registros_1216) y los datos que corresponden al año 2017 (registros_17), para intentar predecir los datos del primer semestre de 2018 se utilizarán todos los datos.

```
registros_1216 <- registros[1:1827,]
registros_17 <- registros[1828:2192,]
```

Regresión Stepwise

Paso 1. Se crea un modelo de regresión lineal sin variables

$$\widehat{Y} = \bar{Y}$$

```
##
## Call:
## lm(formula = Unidades ~ 1, data = registros_1216)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -807.3 -493.8  109.7  393.7 2795.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   807.34      12.92    62.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 552.2 on 1826 degrees of freedom
```

El modelo vacío (sólo con el intercepto), tiene una desviación estandar de $\sigma = 552.2$. Como no tiene variables explicativas no se calcula el R^2.

Paso 2. Se crea un modelo de regresión lineal con todas las variables que se puedan considerar

En este caso todas las variables son indicadoras puesto que los datos carecen de variables numéricas.

```
## lm(formula = Unidades ~ dia_semana + mes + as.factor(dia_anual) +
##     finde + factor(dia_mes), data = registros_1216)
```

Componente	Valor
Sigma	259.30
R cuadrado	0.83
R cuadrado ajustado	0.78

La desviación estandar del modelo con todas las variables es de $\sigma = 259.30$, y un $R^2 = 0.83$, los que significa que las covariables del modelo explican un 83% de la variabilidad de las unidades de autos registradas, y con un $R^2_{Ajustado} = 0.78$ que está muy cerca del R^2 (5%), lo que apoya la veracidad de lo concluído con el R^2.

Paso 3. Regresión Stepwise

La regresión Stepwise o paso a paso genera el mejor modelo con las variables más significativas usando la regresión hacía atrás (o Backward) y la regresión hacía adelante (o Forward). Escoge el mejor modelo con el criterio AIC

```
## Start:  AIC=23071.82
## Unidades ~ 1
##
##                          Df Sum of Sq       RSS   AIC
## + dia_semana              6 352380980 204348556 21253
## + finde                   1 286967713 269761823 21750
## + factor(dia_mes)        30  35458632 521270904 23012
## + mes                    11  18176180 538553356 23033
## <none>                                556729536 23072
## + as.factor(dia_anual) 365 111564728 445164808 23393
##
## Step:  AIC=21252.71
## Unidades ~ dia_semana
##
##                          Df Sum of Sq       RSS   AIC
## + as.factor(dia_anual) 365  96271710 108076846 20819
## + factor(dia_mes)        30  34826239 169522316 20971
## + mes                    11  18560937 185787618 21101
## <none>                                204348556 21253
## - dia_semana              6 352380980 556729536 23072
##
## Step:  AIC=20818.93
## Unidades ~ dia_semana + as.factor(dia_anual)
##
##                          Df Sum of Sq       RSS   AIC
## + mes                    10   9067246  99009600 20679
## + factor(dia_mes)        30   9088936  98987910 20718
## <none>                                108076846 20819
## - as.factor(dia_anual) 365  96271710 204348556 21253
## - dia_semana              6 337087962 445164808 23393
##
## Step:  AIC=20678.84
## Unidades ~ dia_semana + as.factor(dia_anual) + mes
##
##                          Df Sum of Sq       RSS   AIC
## + factor(dia_mes)        30   3873144  95136457 20666
## <none>                                99009600 20679
## - mes                    10   9067246 108076846 20819
## - as.factor(dia_anual) 364  86778018 185787618 21101
## - dia_semana              6 333600171 432609772 23361
##
## Step:  AIC=20665.94
## Unidades ~ dia_semana + as.factor(dia_anual) + mes + factor(dia_mes)
##
##                          Df Sum of Sq       RSS   AIC
## <none>                                95136457 20666
## - factor(dia_mes)        30   3873144  99009600 20679
## - mes                    10   3851453  98987910 20718
```

```
## - as.factor(dia_anual) 364  55813931 150950388 20781
## - dia_semana             6 332468523 427604980 23400
```

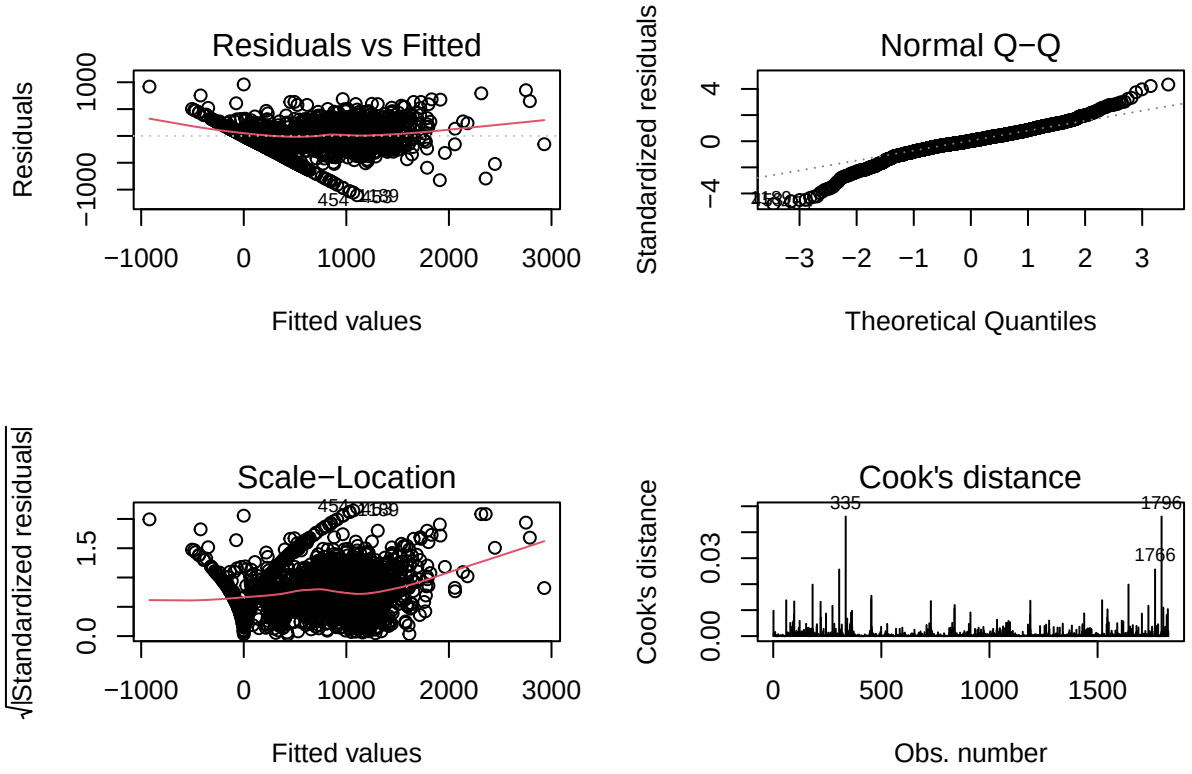
El modelo final obtenido es:

```
summary(modelo_step_final)$call
```

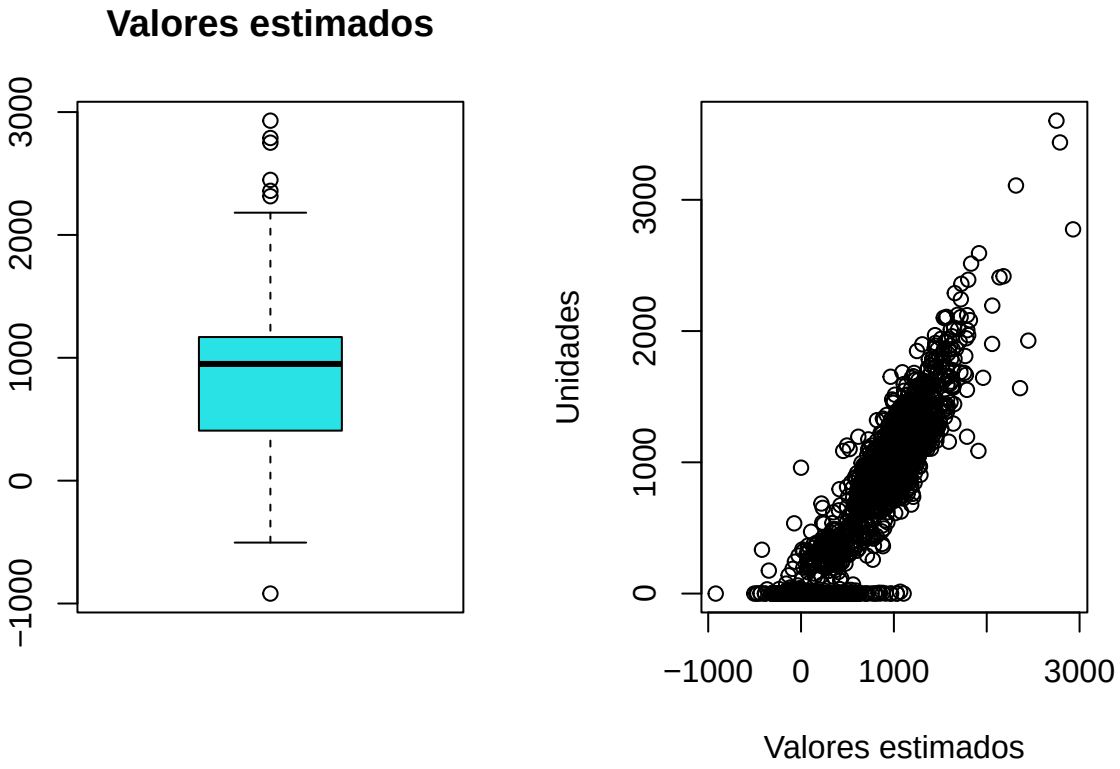
```
## lm(formula = Unidades ~ dia_semana + as.factor(dia_anual) + mes +
##     factor(dia_mes), data = registros_1216)
```

La desviación estandar del modelo con todas las variables es de $\sigma = 259.30$, y un $R^2 = 0.83$, los que significa que las covariables del modelo explican un 83% de la variabilidad de las unidades de autos registradas, y con un $R^2_{Ajustado} = 0.78$. Son los mismos resultados obtenidos que cuando utilizamos todas las variables, con la diferencia que este modelo no tiene encuenta la variable **finde**, la cual no aportaba información significativa para explicar las unidades registradas.

Supuestos del modelo



Estimación de los datos

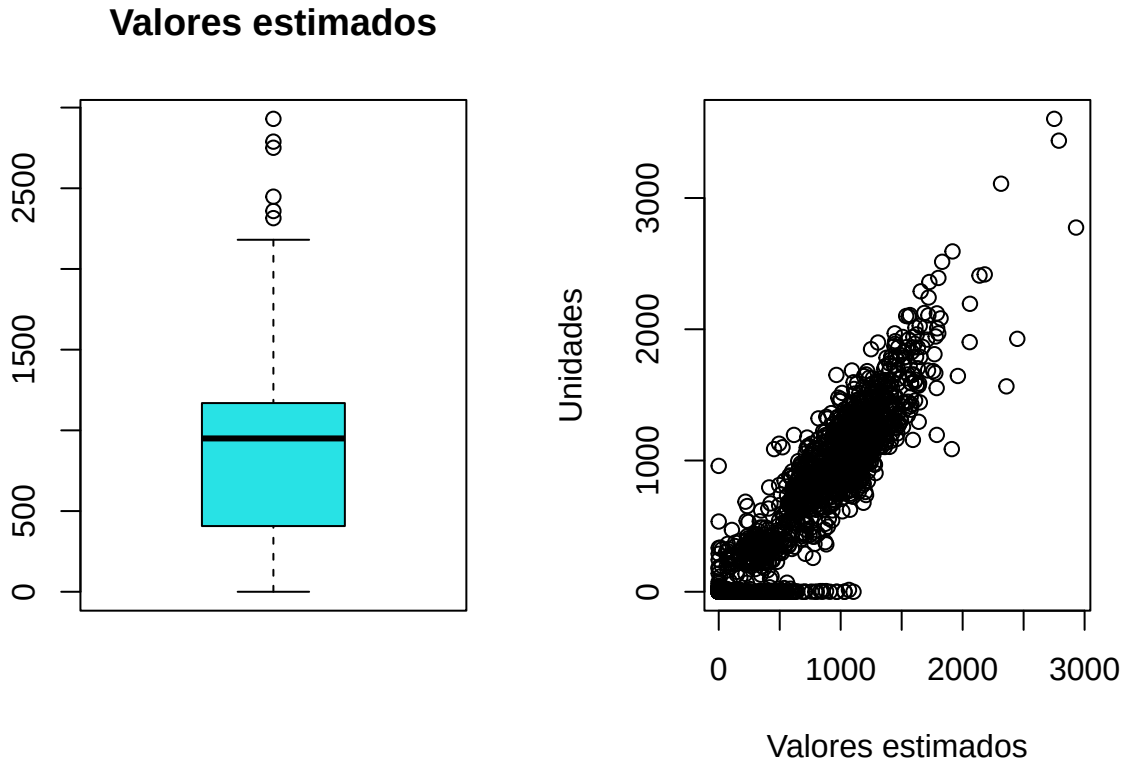


MSE
52070.35

Vemos que entre los datos existen valores negativos, lo cual no tienen ninguna interpretación pues la variable Unidad es una cantidad entera que no puede ser negativa, por lo tanto para un mejor ajuste se opta por cambiar los valores negativos a cero. En el otro gráfico podemos ver que hay cierta relación entre los valores estimados y las Unidades reales de la base de datos que contiene las unidades registradas desde el 2012 hasta el 2016.

El MSE es de: 52070.35.

Estimación de los datos con ajuste en los datos negativos

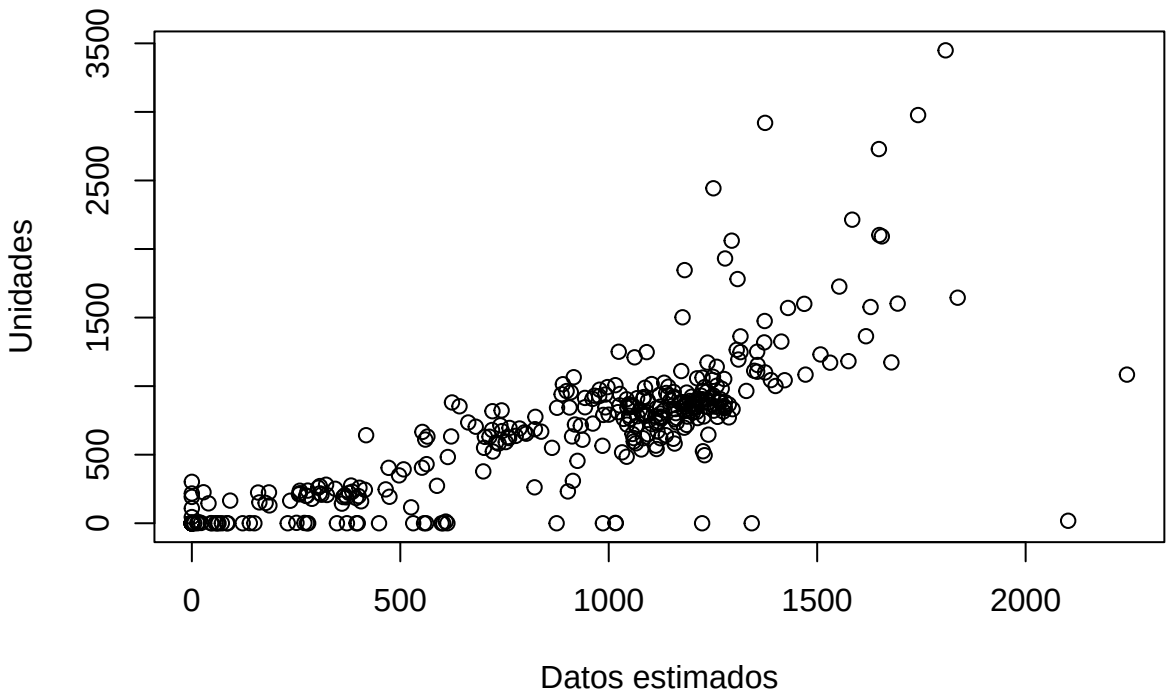


MSE
48900.54

El MSE es de: 48900.54, los que significa que el error cuadrático medio se reduce 3169.81 unidades cuadradas con respecto al MSE anterior de los datos estimados que contenían los negativos.

Predicción para 2017

Aquí se corrige de una vez el problema de los datos estimados negativos, haciéndolos igual a cero.



Los datos no parecen seguir una linea recta de 45° con respecto al eje X, al parecer las estimaciones subestiman los valores reales en su mayoría. Sinembargo fue le mejor modelo hallado.

MSE
140145.4

El MSE de los errores calculados para la base de datos de unidades registradas para el 2017 es de: 140145.4. Casi el doble que el MSE para los datos con los que se hicieron el modelo.

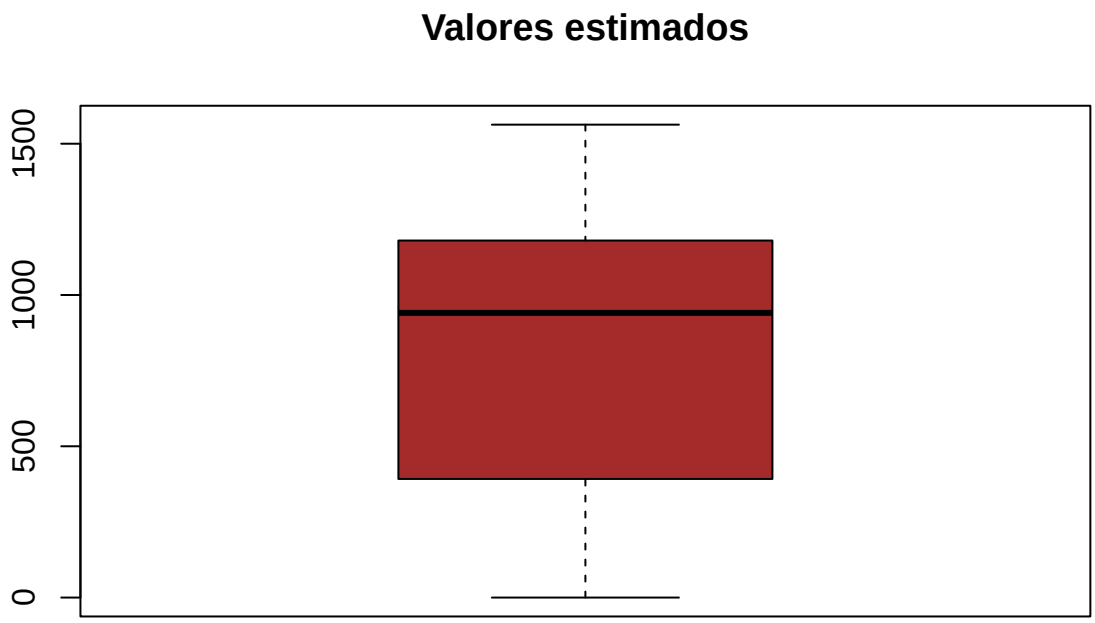
Predicciones para el primer semestre del 2018

Para esto se carga una nueva base de datos con las fechas correspondientes al primer semestre de 2018, además se le agregan las variables necesarias para que el modelo pueda hacer la predicción, la base se llama **registros_18**

Fecha	dia_semana	mes	dia_anual	dia_mes
2018-01-01	lunes	enero	1	1
2018-01-02	martes	enero	2	2
2018-01-03	miércoles	enero	3	3
2018-01-04	jueves	enero	4	4
2018-01-05	viernes	enero	5	5
2018-01-06	sábado	enero	6	6

Datos estimados

Para la estimación de los datos se corrige el problema de los datos negativos igualándolos a cero.



Al parecer los datos que se estimaron para el primer semestre del 2018 se encuentran entre 1 y 1500 unidades registradas, lo que puede ser un problema de subestimación de los datos reales como se vio en el caso para los datos del 2017.

Resultados y Conclusiones.

Algunas de las variables creadas para intentar explicar el comportamiento de las unidades de autos registradas resultaron ser útiles, pues se obtuvo un $R^2 = 0.83$ con una desviación estandar de $\sigma = 259.30$. Estás variables intentaron explciar los datos con los que se contruyó el modelo y se logró un MSE de 52070.35 y cuando se corrigió el problema de los valores estimados negativos se logró un MSE de 48900.54, En el caso de la predicción no fueron muy útiles pues cuando se intentó predecir los valores para la base de datos que contenía las unidades de autos registradas en 2017 se obtuvo un MSE de: 140145.4.

Con solo la fecha es complicado obtener variables que den buenos resultados a la hora de crear un modelo lineal que explique y haga predicciones sobre la cantidad de unidades de autos que se van a registrar en el Registro Único Nacional de Tránsito (RUNT). Se recomendaría otras técnicas como Series de tiempo.

Bibliografía

- Mendez J.(13 de octubre de 2019).*Stepwise Regresión*.Rpubs.https://rpubs.com/jorge_mendez/609253