Análisis del Dengue

Se tiene el conjunto de datos de dengue el cual contiene 5 series de tiempo (de longitud $t = 98$ dias) como variables predicadoras:

“Temp_Seca_Max","Precipitacion","Temp_Seca_Min","Hum_Rel_Min","Hum_Rel_Max"

Y como variable a predecir se tienen:

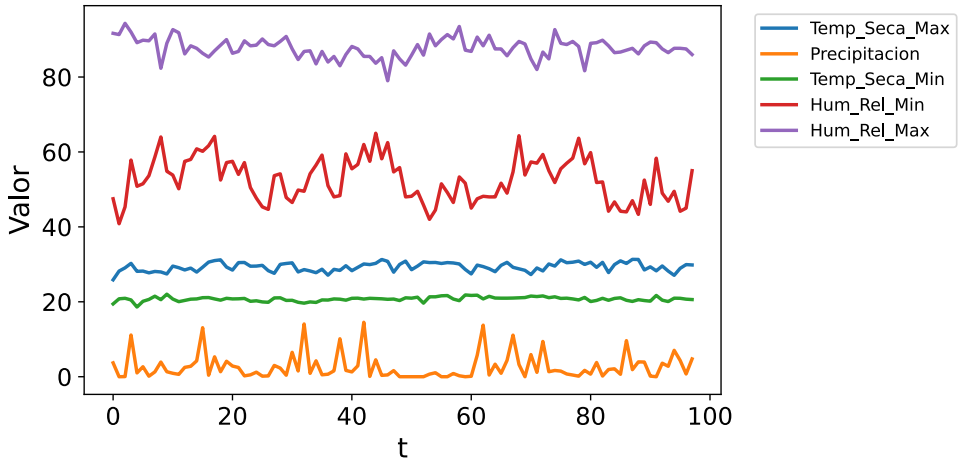
“brote” (Discreta, 0 = No, 1 = Si)
“Dengue” (Continua)

	Temp_Seca_Max	Precipitacion	Temp_Seca_Min	Hum_Rel_Min	Hum_Rel_Max	brote	Dengue
0	25.866667	3.700000	19.433333	47.500000	91.666667	0	2
1	28.200000	0.000000	20.800000	40.833333	91.333333	0	2
2	29.066667	0.033333	20.933333	45.333333	94.333333	0	1
3	30.266667	11.133333	20.500000	57.833333	92.000000	0	1
4	28.133333	1.016667	18.600000	50.833333	89.166667	0	2
...	...	...	...	...	...	...	...
93	28.200000	2.783333	20.066667	46.833333	86.500000	0	22
94	27.066667	7.050000	20.966667	49.500000	87.666667	0	18
95	28.933333	4.350000	20.933333	44.166667	87.666667	0	17
96	29.933333	0.733333	20.700000	45.000000	87.500000	0	14
97	29.833333	4.750000	20.600000	55.000000	86.000000	0	18

98 rows × 7 columns

Regresión Logística

En este experimento podemos ver a cada dato temporal $x_t \in R^5$ como variable predicadora y al brote como variable a predecir y_t.

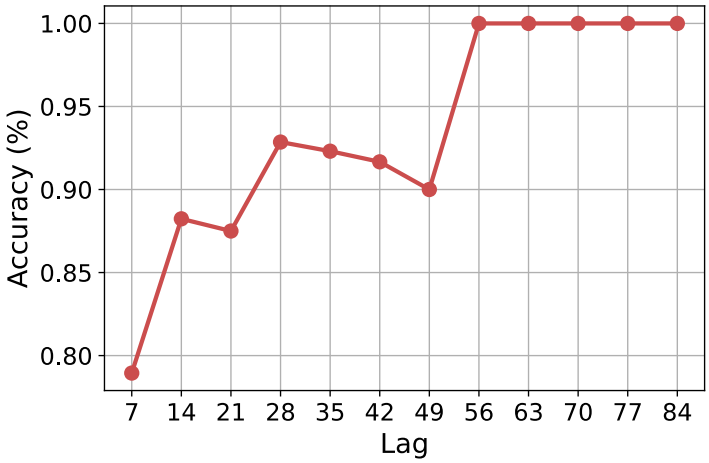


Para el análisis temporal se definió el operador rezago (LAG)

$$LAG(x_t) = x_{t-1}, LAG(LAG(x_t)) = x_{t-2}, \ldots$$

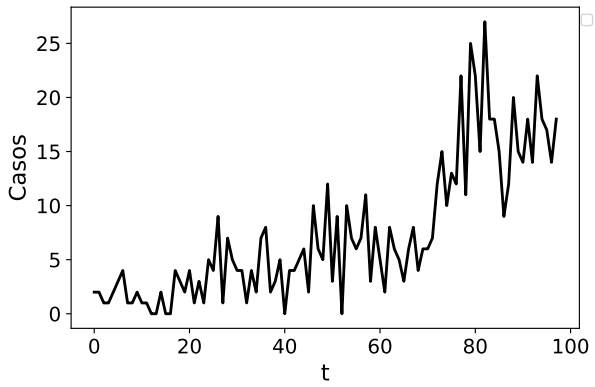
En particular se tomaron variables predictoras con diferente tipo de rezago. De forma que por ejemplo, si usamos los rezagos $LAG, LAG^2, \ldots, LAG^7$, tomaremos $[x_1 : x_2 : \ldots : x_7]$ para predecir y_7. De esta forma tenemos un conjunto de datos que tiene en cuenta un rezago de los 7 días anteriores, es decir, una matriz de 91x35. De manera similar se pueden definir problemas con una ventana temporal mas amplia (Mas *LAG*).

Para este problema se implementará un **regresor logístico,** ya que su variable dependiente es cualitativa (Hay brote o No hay brote). La evaluación del modelo se hizo tomando como entrenamiento los primeros 80% datos y se evaluó con el 20% restante. La siguiente figura ilustra el efecto de tomar mas *LAG* para predecir el brote de dengue. En ella se observa que cuando se toman como variables predictoras los 56 rezagos anteriores se alcanza un 100% de precisión, es decir, datos de 8 semanas antes de la fecha a predecir.



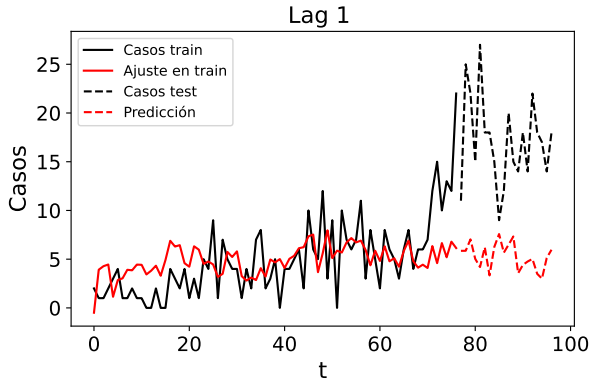
Regresión lineal múltiple

Del conjunto de datos, también se tiene la variable a predecir: “Dengue”. Esta representa el conteo de casos de dengue cada día. Es decir, es un conteo (valores enteros). A continuación de presenta la serie temporal de esta variable.

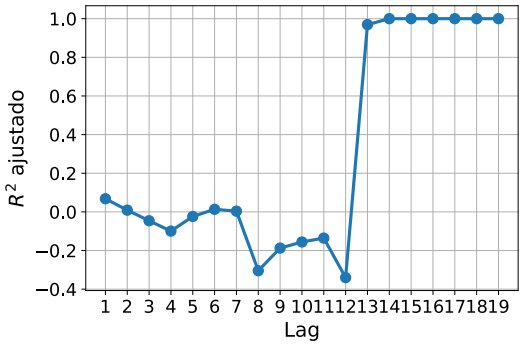


Para ajustar un modelo que logre estimar el número de casos de dengue utilizaremos como variables predictoras las mismas series de tiempo que el problema anterior. Para introducir el *LAG* se procederá también de la misma manera. En particular, utilizaremos ahora una **regresión lineal múltiple.**

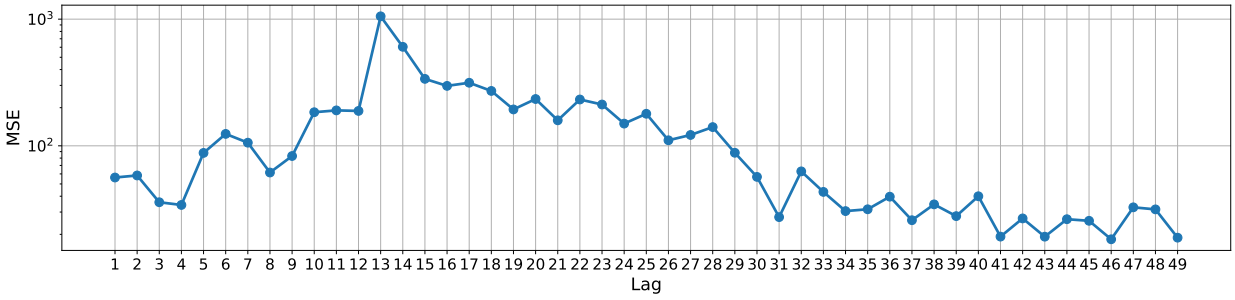
A continuación se presenta el ajuste alcanzado por el modelo sin utilizar *LAG* (i.e *LAG =1*) y dejando para validar el último 20% de los datos. El modelo logró ajustarse en el entrenamiento con un MSE = 13.88 y un $R^2(ajustado)$ = 0.73. Por el contrario en la validación obtuvo un MSE = 162.34.



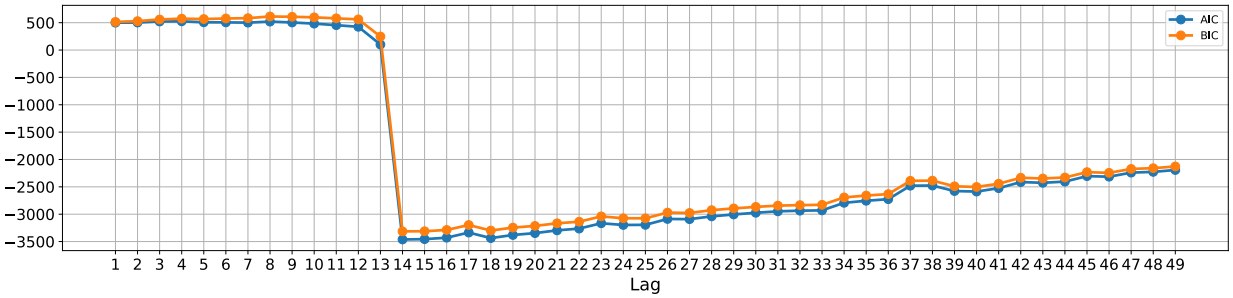
Debido a la poca explicabilidad de las variables, se varió el número de *LAG* como se muestra en la Figura. Se observa cómo mejora a partir de 12 días de *LAG.*



En general, como es de esperarse, al introducir mas variables el modelo mejora como se puede observar en el siguiente gráfico.

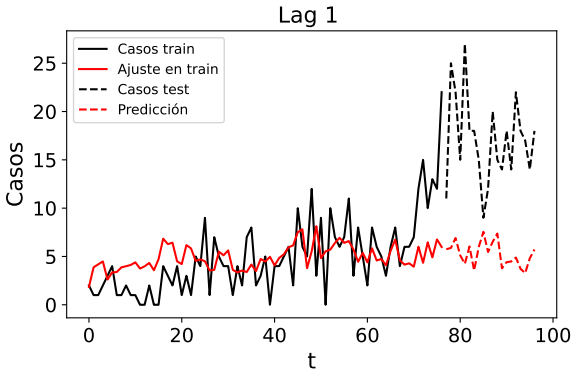


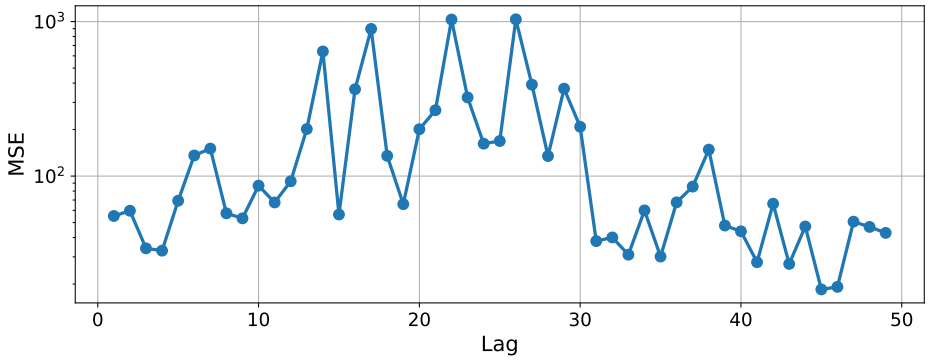
Adicionalmente, como estamos agregando complejidad a la vez que agregamos variables predictoras con el *LAG,* decidimos calcular los indicadores de información AIC y BIC. Donde se observa que el modelo utilizando 14 *LAG* es el mejor.



Regresión de Poisson

Para el mismo problema anterior, una regresión de Poisson fue diseñado para predecir la variable de conteo “Dengue” siguiendo el mismo esquema de evaluación que el anterior experimento. Sin usar LAG el modelo se comporta como se ve en la siguiente figura, teniendo un error MSE de ajuste de 162.7, muy similar a la regresión lineal.





Análisis de enfermedad cardiaca

Para el análisis de enfermedad cardíaca se cuenta con un conjunto de datos de 300 muestras (160 personas normales y 140 personas con enfermedad cardíaca). En total cada muestra consiste de 14 características médicas como presión sanguínea, electrocadiogramas, atributos de la personas, entre otros.

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	presence
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
297	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
298	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
299	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
300	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1
301	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1

La tarea de clasificación consiste en clasificar la presencia de enfermedad cardíaca usando como variables predictoras las anteriormente mencionadas (14 en total). Para esto se utilizará un **regresor logístico.** Bajo este esquema se obtuvieron los siguientes resultados.

- Utilizando una partición de datos 80%/20% para Train y Test respectivamente, se alcanzó una precisión perfecta del 100%.
- Se realizó una validación mas utilizando un K-Fold estratificado, donde se mantuvo este desempeño del 100% en cada uno de los Folds.