Aplicación de modelos lineales generalizados: Caso Regresión Gamma para analizar información generada de estaciones metereológicas

Centro de Investigación en Matemáticas. Unidad Monterrey
Marcelo Alberto Sanchez Zaragoza
Karla Mauritani Reyes Maya
marcelo.sanchez@cimat.mx
karla.reyes@cimat.mx

***Resumen*—El proyecto toma información de 54 estaciones metereológicas ubicadas dentro de la republica Mexicana, cada estación proporciono cerca de 6 años de datos de precipitación(2015-2020). El objetivo consiste en construir un modelo lineal generalizado para el caso Gamma para cada una de las estaciones. Finalmente porporcionar una estimación del valor de precipitación máxima del año 2022 con un intervalo de confianza máximo del 99 %.**

I. Pre-procesamiento

Cada una de las estaciones cuenta con información de 31 semanas aproximadamente, donde no se especificaba el mes al que pertenecian cada una de ellas. En la siguienta tabla se muestra como son los datos para cada una de las estaciones y en particular se ilustra la correspondiente a la estación de Acapulco.

Semana	Plaza	Precipitación_MM
201553	Acapulco	0.007536232
201601	Acapulco	2.184099353
201602	Acapulco	0.132298137
201603	Acapulco	0
201604	Acapulco	0.188819876
201605	Acapulco	0.643229808
201606	Acapulco	0
⋮	⋮	⋮
202136	Acapulco	19.75043478
202137	Acapulco	32.45565217
202138	Acapulco	9.164720497
202139	Acapulco	36.24149068
202140	Acapulco	0

Cuadro I
Visualización de los datos sin procesar para la Plaza meteorológica: Acapulco. El total de la semanas contempla de la semana 53 de 2015 a la semana 40 de 2021. Un aproximado de 6 años.

Como primer paso los datos se tenian que dividir en grupos de semanas correspondientes a los meses de cada año, es decir, encontrar el mes para cada uno de los registros de la estación.
Una vez que los datos ya tenian el mes al que pertenecian lo siguiente fue agregar variables que sirvieran de apoyo al momento de ocupar el modelo. La primer variable que se agrego fue una variable Dummy de estacionalidad por mes, esta variable dummy contaba con valores ceros y solo un uno. El valor uno se colocaba en el mes al que pertenecia el dato y las entradas restantes se dejaban con cero. Al agregar la variable dummy se observo a que mes pertenecia el dato con solo prestar atención a la posición del valor uno.
La siguiente variable que se agrego fue una variable de tendencia(consecutivo), es decir, comenzaba la variable con un valor uno y de forma consecutiva se iba sumando una unidad a este primer valor, al final teniamos una variable consecutiva para la tendencia. La siguiente variable que se agrego fue la media de precipitación, para encontrar dicho valor previamente nuestros datos ya contaban con el mes al que pertenecian y solo bastaba con realizar el promedio de todos los valores de precipitación.
La ultima variable que se agrego fue la correspondiente a la máxima precipitación para cada uno de los meses, al agregar esta variable los datos se redujeron ya que si el mes contaba con 4 o 5 semanas solo nos quedabamos con 1 semana, dicha semana llevaba consigo la variable dummy de estacionalidad y las antes mencionadas. Este procesamiento se realizo a cada una de las 54 estaciones, se guardaron para luego trabajar con ellas el modelo y realizar los ajustes correspondientes a cada una de ellas.
El lenguaje de programación con el que nos apoyamos fue R, al final se agregaran los códigos implementados para cada una de las 54 estaciones.

Nota: Los datos proporcionados corresponden a los años 2015, 2016, 2017, 2018, 2019, 2020 y 2021, fueron proporcionados y no se alteraron en ningun momento.

II. Regresión Gamma

La especificación de un modelo lineal generalizado se realiza en tres partes:

- La *componente aleatoria* correspondiente a la variable Y que sigue una distribución de la familia exponencial(normal, log-normal, Poisson, Gamma).

- La *componente sistematica*, también llamada predictor lineal y corresponde al vector de n componentes, siendo cada una de ellas igual a $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} = X^t\beta$.
- La *función de ligadura*(función link) relaciona la esperanza matemética de la variable dependiente con el predictor lineal $\eta_i = g(\mu_i)$, i=1,...,n. La función de ligadura debe ser monótona y diferenciable.

En la siguiente tabla se muestra algunos enlaces canonicos utilizados regularmente:

Distribución	Enlace Canonico
Normal	$\eta_i = \mu_i$(identity link)
Binomial	$\eta_i = ln(\frac{\pi_i}{1-\pi_i})$(logistic link)
Poisson	$\eta_i = ln(\mu_i)$(log link)
Exponencial	$\eta_i = \frac{1}{\mu_i}$(reciprocal link)
Gamma	$\eta_i = \frac{1}{\mu_i}$(reciprocal link)

Cuadro II
TABLA DE ENLACES CANONICOS.

Una vez que se explico brevemente las partes importantes de un modelo lineal generalizado y mostrado algunas funciones de enlace, nos concentraremos en el caso regresión Gamma(GLM) para realizar pronosticos sobre el siguiente año(2022) del valor maximo de precipitación para cada una de las estaciones.
La regresión Gamma(GLM) relaciona una variable de interés real y estrictamente positiva, a una o más variables(variables predictoras) que se espera que influya en la variable de destino.
La función de distribución Gamma es la siguiente:

$$f(y) = \frac{1}{\Gamma(r)} \left(\frac{1}{\lambda}\right)^r e^{-y/\lambda} y^{r-1}; \quad y \geq 0, r > 0, \lambda > 0 \quad (1)$$

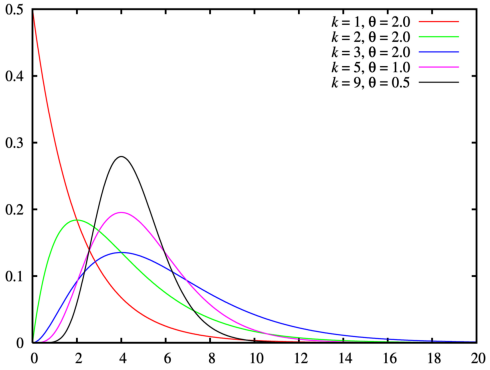

Figura 1. Ejemplo de la distribución Gamma

En la figura 1 se muestran algunos ejemplos de la distribución Gamma, en ellos puede llegar a ser exponencial o incluso tener una forma completamente simetrica(La media y la varianza no son independientes ya que existe una relación entre ellas, la varianza es proporcional a la media).

Como tarea importante para poder usar el modelo Gamma, fue la selección de la función de enlace ya que no siempre se obtuvieron resultados buenos o coherentes con la función canonica que anteriormente se mostra en la tabla 1.
La función de enlace que se tomo en cuenta fue la $log : \eta_i = log(\mu_i)$, ya que la función canonica no contemplaba los casos donde la variable de respuesta no fuera positiva y dado que se tienen valores mayores o iguales a cero no pueden existir valores negativos. Una vez que se definio la función de enlace con la que se iba a trabajar lo siguiente fue ajustar el modelo adecuado para cada una de las estaciones.

III. IMPLEMENTACIÓN

El resultado de agregar la variable de tendecia(continua), media de precipitación y estacionalidad(dummy), se muestra en la siguiente tabla:

Plaza	Max_Pre_MM	Media_Pre	Tendencia	mes_1	mes_2	mes_3
Colima	6.79334039	1.70365196	2	1	0	0
Colima	0.17518797	0.08308271	3	0	1	0
Colima	3.14672401	0.94355533	4	0	0	1
Colima	0	0	5	0	0	0
Colima	2.09999997	1.29057465	6	0	0	0
Colima	9.01815243	6.61458105	7	0	0	0
Colima	14.4611171	11.9402792	8	0	0	0
Colima	14.4792693	11.8405745	9	0	0	0
Colima	20.6278197	11.9795059	10	0	0	0
Colima	5.56015044	4.15185284	11	0	0	0
Colima	8.46390972	3.55385604	12	0	0	0
Colima	0.04092374	0.01023094	13	0	0	0

Cuadro III
VISUALIZACIÓN DE LAS VARIABLES DE REGRESIÓN, MEDIA MENSUAL, TENDENCIA(CONSECUTIVA) Y ESTACIONALIDAD(VARIABLE DUMMY).

El cuadro III tiene hasta el mes_3 debido a que no se pueden mostrar las doce columnas debido a las dimensiones pero al realizar el ajuste del modelo se tomaron en cuenta todas ellas. Una vez que cada estación contaba con las variables mencionadas lo siguiente fue la implementación del modelo, en donde el criterio para la selección del mejor modelo fue tomando el valor de AIC.
Cabe mencionar que para la gran mayoria de las estaciones la variable Media_precipitacion fue altamente significativa lo que implicaba que esta variable no podia estar fuera del modelo. Respecto a la variable de tendencia(consecutiva) casi en todas las estaciones no fue significativa por lo que no se omitia en el modelo. Las variable dummy en muchos modelos fue variando ya que al tener un total de 11(se omitio una para evitar la colinealidad) teniamos distintas combinaciones donde muchas de ellas se repetian y otras no.

III-A. Poryección de la Media

Una vez que se obtuvieron los modelos adecuados para cada plaza y en cada modelo fuera necesario encontrar el valor de la media para las futuras proyecciones, se realizo una regresión simple, en esta regresión se tomaron en cuenta unicamente la variable de tendencia(Dummy), es decir, las 11 variables respecto a los meses. Ya que se obtuvo el modelo de regresión lo que resto fue hacer el pronostico del valor de Media_precipitacion para cada estación que lo

necesitara. En el siguiente cuadro se muestra los resultados de las proyecciones de la media para la estación de colima:

Media_Precipitacion	Tendencia	mes_1	mes_2	mes_3
0.244787849	74	1	0	0
0.347882327	75	0	1	0
0.139590964	76	0	0	1
0.004249088	77	0	0	0
0.031190102	78	0	0	0
0.618933147	79	0	0	0
1.746537182	80	0	0	0
3.304952670	81	0	0	0
3.511007297	82	0	0	0
1.218307384	83	0	0	0
0.978424905	84	0	0	0
0.339959296	85	0	0	0

Nuevamente se omitieron las 12 columnas debido a las dimensiones pero para implementar el modelo se tomaron en cuenta todas ellas.

III-B. Estimación del intervalo de confianza al 99 %

Los intervalos de incertidumbre (intervalos de confianza o de predicción) para valores ajustados son bastante fáciles si solo se trabaja con modelos lineales, pero cuando los modelos se vuelven más complejos, por ejemplo, modelos lineales generalizados, las facilidades para cuantificar la incertidumbre para las predicciones son más complicadas o inexistente.

Para calcular los intervalos de confianza se hace un intervalo en la escala del predictor lineal y luego aplicando la función de liga inversa g^{-1} del ajuste del modelo para transformar los intervalos de confianza del nivel lineal en el nivel de respuesta. Donde los intervalos están dados por:

$$g^{-1}\left(x'\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}x'\left(X'X\right)^{-1}x}\right)$$

IV. Resultados

Finalmente para cada una de las plazas se realizo el pronostico del año 2022 para posteriormente seleccionar la máxima precipitación que se espera para ese año asi como el intervalo de confianza al 99 %, tenemos los siguientes resultados para once plazas:

Plaza	2015	2016	2017	2018	2019	2020	2021	2022	Intervalo C.	
Colima	0	20.6278	18.9683	24.7124	20.3172	17.431	45.2791	24.8179	13.1273	46.9018
Cuernavaca	0.2167	10.7571	16.0286	14.3289	15.9934	16.2607	12.0088	14.4555	8.5375	24.4625
Culiacan	0	11.7569	11.7435	17.7607	10.1676	13.9238	18.9024	26.3554	11.1664	62.1342
Oaxaca	0.0993	19.756	34.3143	19.0423	13.4268	23.7375	30.3191	17.5918	9.6294	32.1384
San Luis Potosi	0.206	6.2714	9.3265	12.0613	6.0824	7.9719	10.8801	7.6612	5.8311	10.0657
Puebla	0.8114	11.0413	10.8808	8.5227	14.1686	12.2633	17.1162	13.123	9.5791	17.9781
Queretaro	0.7008	10.5099	8.8434	13.6853	10.1803	9.7124	13.4652	10.4578	8.2395	13.2735
Reynosa	11.1771	8.9709	6.9045	12.1516	11.2895	29.2805	29.5235	11.8642	6.2751	22.4312
Saltillo	4.2	7.3857	6.2857	8.2	3.9844	26.2817	17.7667	5.1867	3.5714	7.5327
San Luis Potosi	0.206	6.2714	9.3265	12.0613	6.0824	7.9719	10.8801	7.6612	5.8311	10.0657
Zacatecas	0.4269	9.5436	15.4575	25.6081	10.8178	13.965	22.0742	10.7068	5.581	20.5403

Cuadro IV
HISTORICO DE LA PRECIPITACIÓN DE 2015-21 ASÍ COMO LA ESTIMACIÓN PARA 2022 Y E INTERVALO DE CONFIANZA.

Los resultados de la tabla muestran solamente los que corresponden a 11 estaciones,el resto de ellas se anexaran al final del documento. Debido a las dificultades de mostrar las 54 plazas y sus pronosticos en esta sección a continuación de agregan dos figuras más ilustrativas para observar los cambios de un año a otro.

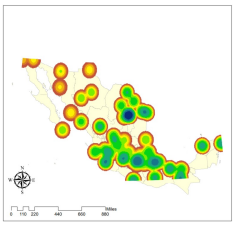

Figura 2. Precipitación Máx Historico 2021.

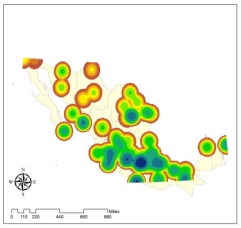

Figura 3. Precipitación Máx Estimada 2022.

Gracias a la figura 2 y 3 es más sencillo observar si hubo cambios entre los valores historicos y los estimados para el años 2022. El color rojo significa que no hay valores altos de precipitación mientras que para el color azul nos indica que altos valores de precipitación en esa región o estación.
En algunas regiones del sur se observa que los valores de precipitación van aumentar

V. Concluciones

Respecto a la estimación del valor máximo de precipitación anual para las estaciones meteorologicas,

- La información de la media de la precipitación mensual resulto una variable significativa en la mayoría de los modelos.
- Mientras que en casi todos los casos, aúnque haya un coeficiente de tendencia negativo o positivo, no resultó realmente significativo aportando menor información.
- Así, sí la precipitación se une a la Mediay a la estacionalidad para lograr una estimación del maximo mensual por año se obtuvieron mejores resultados

I. Resultados Finales

Plaza	2015	2016	2017	2018	2019	2020	2021	2022	Intervalo C.	
10VCZ Boca	0.16547	18.0857	21.7222	12.87163	22.0367	28.6973	30.5795	17.5737	12.8904	23.949
10VCZ Coatza	2.001	15.2142	20.2857	20.0821	24.5676	53.746	36.566	20.2879	14.346	28.686
10VCZ Cordoba	0.155	17.869	21.502	12.9176	21.8757	23.243	30.145	17.892	12.6245	24.221
10VCZ Poza	3.814	24.145	30.632	17.16	23.4688	19.93	22.0436	18.8265	13.4089	26.421
10VCZ Xalapa	0.3614	17.3933	20.395	12.7803	22.1281	20.1586	21.35	15.8825	11.4749	21.976
Acapulco	0.00753	17.1673	14.107	23.3381	20.486	20.0085	36.241	41.3070	27.8981	61.133
Aguascalientes	0.00	12.38	11.02	15.57	10.39	11.85	20.185	11.230	8.098	15.589
Cancun	0.633	16.900	30.442	27.199	16.9	46.514	35.5	21.708	11.6178	31.8015
Chihuahua	0.0332	10.590	6.73	6.29	5.754	6.58	12.97	5.401	3.241	8.995
Ciudad J.	1.506	6.178	4.461	3.741	3.64	4.35	6.997	3.06	2.102	4.45
Colima	0	20.6278	18.9683	24.7124	20.3172	17.431	45.2791	24.8179	13.1273	46.9018
Cuernavaca	0.2167	10.7571	16.0286	14.3289	15.9934	16.2607	12.0088	14.4555	8.5375	24.4625
Culiacan	0	11.7569	11.7435	17.7607	10.1676	13.9238	18.9024	26.3554	11.1664	62.1342
Durango	0.622	9.535	9.085	12.029	8.99	15.24	30.854	8.950	4.902	16.33
Guadalajara A.	0.002	9.209	15.089	12.40	13.184	11.391	15.06	11.435	10.491	12.382
Guadalajara C.	0.0	9.26	15.65	12.06	14.31	12.07	16.91	13.38	6.60	27.117
Guadalajara T.	0.0041	9.317	14.58	11.497	14.957	12.80	19.40	13.938	6.51	29.772
Hermosillo	0.134	12.544	4.730	6.758	8.96	4.602	5.3412	6.370	3.3108	12.22
La Paz	0.0	22.901	31.68	17.502	18.201	14.68	13.13	17.509	15.808	19.206
Laguna	4.328	7.248	5.29	9.81	3.96	11.52	13.36	4.35	2.48	7.649
Laredo	9.74	11.60	9.76	13.68	7.89	14.207	22.54	6.70	4.51	9.95
Leon	0.0027	9.87	11.14	17.33	7.87	15.139	14.44	9.75	4.51	16.60
Los Mochis	0.0040	9.28	11.24	14.54	11.47	8.39	7.33	10.01	5.10	19.60
Matamoros	6.89	14.29	6.92	18.42	10.94	25.84	29.753	8.91	6.225	12.76
Merida	1.772	7.35	15.49	15.92	10.04	28.34	34.14617	11.981	8.61	16.66
Mexicali	1.127	4.162	3.47	2.14	4.09	6.05	6.32	1.32	0.746	2.349
Mexico A.	0.445	10.66	14.07	13.38	8.99	12.79	9.44	11.11	7.16	17.226
Mexico O.	0.41	10.70	15.23	10.98	9.41	11.94	9.82	9.97	6.385	15.584
Mexico R.	0.4166	10.714	15.214	11.98	9.767	12.08	9.98	5.066	4.316	5.9470
Mexico S.	0.4156	10.68	15.18	11.07	9.48	11.52	10.665	4.66	3.941	5.521
Mexico V.	0.4156	10.692	13.85	12.144	9.486	12.27	10.39	11.71	7.799	17.605
Monclova	4.551	11.35	12.15	11.109	6.32	10.29	20.63	7.175	4.929	10.44
Monterrey C.	2.46	8.0180	6.99	8.012	12.74	44.805	26.803	8.41	5.877	12.054
Monterrey N.	2.4622	8.031	6.987	8.63	10.48	36.48	26.90	8.43	5.85	12.14
Monterrey O.	2.4654	8.79	8.607	11.701	11.017	33.21	32.790	9.64	6.453	14.422
Monterrey S.	2.4611	8.679	8.0607	10.92	12.57	36.88	31.89	9.40	6.419	13.787
Morelia	1.58	19.657	16.060	35.192	19.57	21.38	39.92	21.13	10.42	42.83
Nogales	3.558	10.6657	8.22	5.72	11.07	7.15	9.03	9.45	6.742	13.25
Oaxaca	0.0993	19.756	34.3143	19.0423	13.4268	23.7375	30.3191	17.5918	9.6294	32.1384
Obregon	0.0	31.001	7.043	9.7423	12.01	4.4735	3.951	18.34	9.0224	37.284
Pachuca	0.986	11.97	15.37	9.15	13.72	14.1122	10.57	11.54	8.512	15.656
Piedras N.	3.633	14.38	16.07	15.67	11.27	9.451	26.666	8.4704	5.572	12.874
Puebla	0.8114	11.0413	10.8808	8.5227	14.1686	12.2633	17.1162	13.123	9.5791	17.9781
Queretaro	0.7008	10.5099	8.8434	13.6853	10.1803	9.7124	13.4652	10.4578	8.2395	13.2735
Reynosa	11.1771	8.9709	6.9045	12.1516	11.2895	29.2805	29.5235	11.8642	6.2751	22.4312
Saltillo	4.2	7.3857	6.2857	8.2	3.9844	26.2817	17.7667	5.1867	3.5714	7.5327
San Luis Potosi	0.206	6.2714	9.3265	12.0613	6.0824	7.9719	10.8801	7.6612	5.8311	10.0657
Tampico	3.034	16.94	14.028	8.084	12.700	17.487	14.256	12.67	10.48	15.31
Tijuana	5.995	4.93	13.49	8.34	8.886	6.217	4.110	4.146	2.954	5.818
Toluca	0.3742	9.539	16.343	18.22	9.886	14.66	11.86	12.34	10.224	14.89
Tuxtla	0.0261	27.235	25.712	43.153	29.119	41.07	45.144	27.79	15.514	49.807
Vallarta	0.0	17.29	18.626	11.874	22.783	25.832	24.430	12.992	10.127	18.118
Villahermosa	4.9166	16.88	10.85	14.528	15.07	44.0608	23.256	12.901	9.8396	16.9162
Zacatecas	0.4269	9.5436	15.4575	25.6081	10.8178	13.965	22.0742	10.7068	5.581	20.5403

Cuadro I

Historico de la precipitación de 2015-21 así como la estimación para 2022 y e intervalo de confianza.

REFERENCIAS

[1] Matthew N., Carrie J., Michael J. Lyons, Miyuki K., Hanae I. y Jiro(2010). *Generalized Linear Models with Applications in Engineering and the Sciences*.

[2] Matthew N., Carrie J., Michael J. Lyons, Miyuki K., Hanae I. y Jiro(2010). *Bayesian and Frequentist Regression Methods*.

[3] Matthew N., Carrie J., Michael J. Lyons, Miyuki K., Hanae I. y Jiro(2010). *Linear Models With R*.