Presentación del proyecto de estadística

Rodrigo Castillo Camargo

May 2020

1 Introducción

El proyecto consiste en ajustar una regresión lineal a una función que sabemos que es sinosoidal y no lineal usando R, Todo el proyecto está basado en el paper "Studying seasonality by using sine and cosine functions in regression analysis" de A M Stolwijk, H Straatman, G A Zielhuis.

1.1 Descripción del método implementado

Un modelo de regresión lineal tiene la forma de $y = B_0 + B_1 + \epsilon + B_{G1}xG1 + ... + B_{Gn}$ x G_n en donde
B_0 indica el intercepto y y es una variable contiunua que indica la presencia de, en el caso del estudio, malformaciones congeniales. Lo que vamos a hacer es un ajuste a esa regresión para que se comporte como un modelo sinosoidal y no como un modelo lineal, pues es una variable que se comporta de esta manera con respecto a los meses del año, , para eso, vamos a definir la amplitud descrito como una función cosenosoidal $f(t) = \alpha \times cos[(\frac{2\pi t}{T}) - \theta]$
en donde :
T = numero de tiempo descrito entre 0 y 2 π (Ejemplo : T = 12 meses)
t = periodo de tiempo (EJemplo: Enero t =1 , Febrero t =2 .... Diciembre t=12)
α = Amplitud
θ = Desplazamiento horizontal de la función coseno
como tetha es desconocido vamos a ajustar el modelo a:
$f(t) = B_1 \times \sin(2\ \pi t_{\overline{T)}} + B_0 \times \cos(2\ \pi t_{\overline{T)}}$
Ahora vamos a ajustar esa fórmula a la regresión lineal, por lo que la regresión lineal nos quedará de esta manera:
$ln(\frac{P_t}{1-P_t}) = B_0 + B_1 \times sin(\frac{2\pi t}{T}) \times cos(\frac{2\pi t}{T}) + B_{G1} \times G_1 + ... + B_{Gn} \times G_n$
por lo tanto tendremos que la probabilidad de una malformación en cada periodo puede calcularse de la forma :

$$P_1 = \frac{e^(B_0+B_1)\times sin(\frac{2\pi t}{T})+B_2\times cos(\frac{2\pi t}{T})+B_{G1}\times G_1+...+B_{Gn}\times G_n}{1+e^(B_0+B_1)\times sin(\frac{2\pi t}{T})+B_2\times cos(\frac{2\pi t}{T})+B_{G1}\times G_1+...+B_{Gn}\times G_n}$$

Como queremos estimar un ajuste de los puntos a una función sinosoidal, lo que haremos será una aproximación usando un método de máxima verosimilitud para minimizar el error , con el cuál maximizaremos la aproximación de forma sinosoidal

1.2 Procedimiento

Se tiene una tabla en la cuál están los datos a estudiarse:

Table 1 Numbers of anencephalus cases and total births, fictitiously divided into data from boys and girls

	Data from Walter and Elwood[6]			*Fictitious data*					
	Total			*Boys*			*Girls*		
Month of birth	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*
January	468	340 797	137	463	252 695	183	5	88 102	5.68
February	399	318 319	125	392	215 431	182	7	102 888	6.80
March	471	363 626	130	459	205 341	224	12	158 285	7.58
April	437	359 689	121	417	156 571	266	20	203 118	9.85
May	376	373 878	101	347	120 846	287	29	253 032	11.46
June	410	361 290	113	375	93 400	401	35	267 890	13.07
July	399	368 867	108	364	95 359	382	35	273 508	12.80
August	472	358 531	132	444	115 886	383	28	242 645	11.54
September	418	363 551	115	398	158 252	251	20	205 299	9.74
October	448	352 173	127	436	198 874	219	12	153 299	7.83
November	409	331 964	123	402	224 665	179	7	107 299	6.52
December	397	336 894	118	392	249 801	157	5	87 093	5.74

leeremos ese dataset "Walter and Elwood 6" bajo el nombre de WyE en R , una vez tengamos el dataset leído podremos hacer un ajuste de como se comportan los datos de las diferentes columnas con respecto a los meses.
En este ejercicio, vamos a enfocarnos únicamente en la tabla de Prevalence con respecto a los meses del año , Prevalence es una tabla que indica un valor cada 100.000 pruebas.
Es necesario hacer una función que calcule el Error, pues queremos minimizarlo ; Esto lo podemos hacer, pues previamente definimos la probabilidad, que está dada por la ecuación :

$$P_1 = \frac{e^(B_0+B_1)\times sin(\frac{2\pi t}{T})+B_2\times cos(\frac{2\pi t}{T})+B_{G1}\times G_1+...+B_{Gn}\times G_n}{1+e^(B_0+B_1)\times sin(\frac{2\pi t}{T})+B_2\times cos(\frac{2\pi t}{T})+B_{G1}\times G_1+...+B_{Gn}\times G_n}$$

Así, el error se define como :
Sea k = P_1 , $H = Error$, se tiene que :

H = $\frac{Prevalence}{100000}$ - ($\frac{k}{k+1}$)

Ahora lo que queremos hacer es reducir H , por lo que haremos un método de máxima verosimilitud para minimizar H.

No sabía que método usaron en el paper, entonces, implementé un método llamado mle2(máximum likelihood estimation 2) de la librería "bbmle" de R . el cuál me retorna:
B_0 = -6.72415738
B_1 = 0.04289461
B_2 = 0.06018819

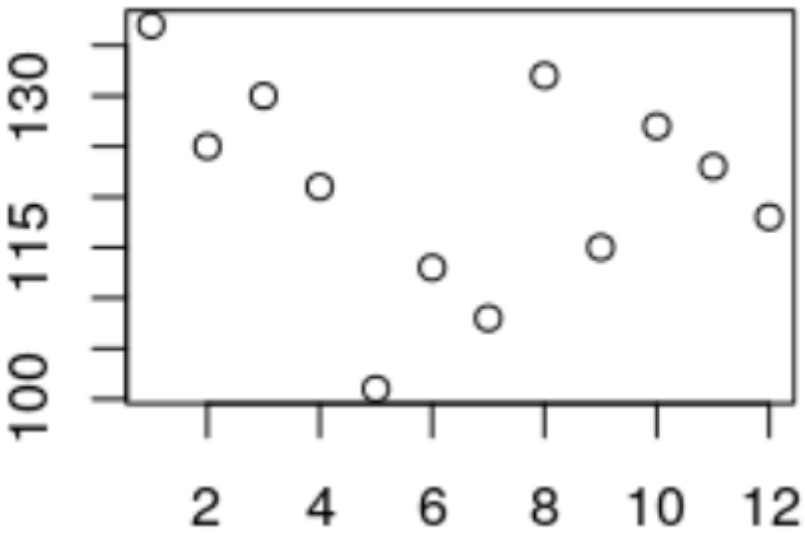
```
Coefficients:
      beta0       beta1       beta2
-6.72415738  0.04289461  0.06018819
```

que son unos coeficientes parecidos a los que obtuvieron en el paper:

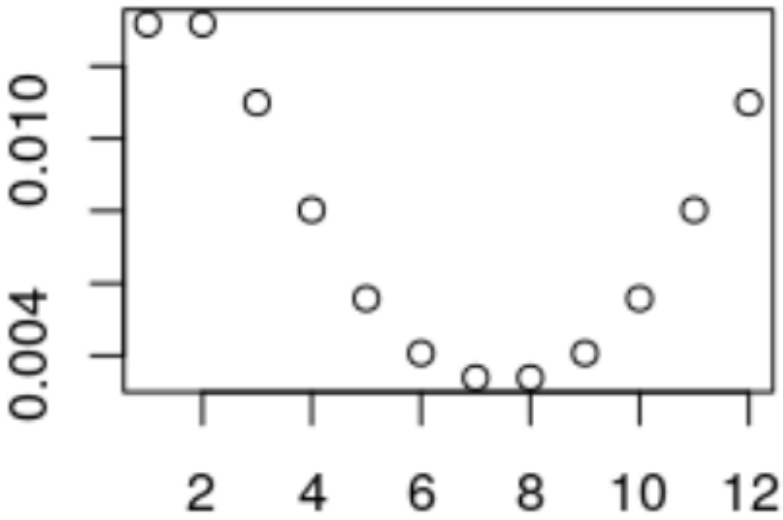
C Logistic, cosine function with 12 month period:
$\ln[P/(1-P)] = -6.718 + 0.009822 \times \sin(2\pi t/12) + 0.06929 \times \cos(2\pi t/12)$

1.3 Gráficas

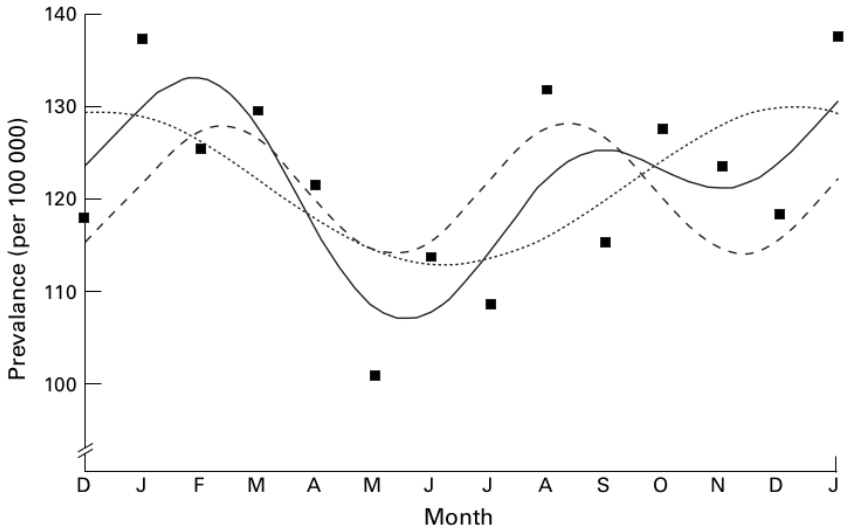
Los datos se ven de esta manera , en donde el eje y hace referencia a los datos de prevalencia y el eje x a los meses :



De esta manera, nuestro ajuste de la regresión lineal a una función sinosoidal quedará así :



así, nuestra estimación se verá de esta manera:



1.4 anotaciones y dudas

1 - Tuve que usar los valores iniciales de los beta igual a como los usaron en el paper, pues no sabía de donde los implementaron en el paper realmente.

2-No conozco el algoritmo del método MLE2 , soy consciente de que no tengo bagage suficiente para entenderlo, sin embargo, es el único método de máxima verosimilitud que me retornó los coeficientes parecidos a los del paper, por eso lo implementé.

1.5 Explorando otras maneras de hacer el ajuste en R

Existen funciones definidas en R diseñadas para hacer ajustes de regresiones de regresiones lineales a otro tipo de funciones, el método LM es un método muy poderoso que permite hacer estos ajustes de manera mas simple. Sin embargo, debido a que no conozco la función de máxima verosimilitud que se emplea , no entiendo la función de LM para este caso.
Como el propósito del trabajo es recrear el experimento del paper dado, me parece que tiene mas sentido exponer el algoritmo implementado por mi, así, los valores concuerdan con los del paper y no con los de un algoritmo el cuál desconozco.

1.6 Conclusión

Las regresiones lineales son una herramienta muy poderosa para estimar comportamientos de variables , sin embargo, conociendo que se pueden ajustar a otro tipo de funciones no lineales , mientras se sepa de que forma se comportan las variables, se pueden hacer estimaciones mucho mas precisas y elegantes , como la que está citada en el paper "Studying seasonality by using sine and cosine functions in regression analysis" y es conocer esos métodos de ajuste para , en un futuro, hacer estimaciones consistentes en caso de que se trabaje en análisis de datos o estadística.