Modelos log lineales en polling

Antonio Manguart Páez 175701 MCD

Mayo 2019

1 Introducción

Tenemos las encuestas de ciertas casas encuestadoras para las elecciones presidenciales de Estados Unidos con la siguiente información:
- Votos democratas
- Votos republicanos
- Indecisos
- Año de la elección
- Días para la elección
- Casa encurestadora
- Estado

Algunos artículos académicos sugieren que para hacer predicciones relevantes del tema no es buena idea asumir independencia de lo anterior, puede que haya casas encuestadoras que den preferencia a cierto partido o interacciones raras entre estado, casa encuestadora y partido. Para ello proponemos otra clase de modelos que pueden ser relevantes como los log-lineales.

2 Modelos log lineales

Llamémosle π a la frecuencia. Si queremos conocer la frecuencia de dos eventos independientes i, j

$$\pi_{i,j} = \pi_i * \pi_j$$

Si tenemos el total de veces que sucede un evento, el total de veces esperado es:

$$\mu_{i,j} = n\pi_{i,j}$$

Esto quiere decir que:

$$\mu_{i,j} = n * \pi_i * \pi_j$$

Sacandole logaritmo de ambos lados:

$$ln(\mu_{i,j}) = ln(n) * ln(pi_i) * ln(\pi_j)$$

$$ln(\mu_{i,j}) = ln(n) + ln(pi_i) + ln(\pi_j)$$

Renombrando algunas cosas:

$$ln(\mu_{i,j}) = \lambda + \lambda_i^x + \lambda_j^y$$

Es fácil notar que cuando no hay independencia:

$$ln(\mu_{i,j}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{i,j}^{x,y}$$

Esto explicaría todo el dataset pues hay más variables que filas, a este modelo se le conoce como *modelo saturado*. Estos parámetros se pueden encontrar con cualquier algoritmo de optimización pero con el fin de tener intervalos de confianza típicamente se sacan con simulación monte carlo.

2.1 Tipos de modelos log lineales

Suponga que se tienen 3 variables, x, y, z

2.1.1 Mutuamente independientes

$$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z$$

2.1.2 Conjuntamente independientes (ejemplo, y es conjuntamente independiente de x y z)

$$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,z}$$

Es decir, hay interacción entre x y z, pero no para y.

2.1.3 x, y son independientes dado Z

$$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,z} + \lambda^{y,z}$$

2.1.4 Asosiación homogenea

$$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,z} + \lambda^{y,z} + +\lambda^{x,y}$$

2.1.5 Modelo saturado

$$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,z} + \lambda^{y,z} + +\lambda^{x,y} + \lambda^{x,y,z}$$

2.2 Resumen y simbolos

$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z = (X, Y, Z)$
$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,y} = (XY, Z)$
$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,y} + \lambda^{y,z} = (XY, YZ)$
$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,y} + \lambda^{y,z} + \lambda^{x,z} = (XY, YZ, XZ)$
$\mu_{i,j,k} = \lambda + \lambda^x + \lambda^y + \lambda^z + \lambda^{x,y} + \lambda^{y,z} + \lambda^{x,z} + \lambda x, y, z = (XYZ)$

Estos son los modelos que vamos a utilizar para tratar de modelar las encuestas con los datos que tenemos.

3 Tratamiento de datos

Se utiliza el dataset *us-pres-state-polling-2004-2016.csv*. La variable mean days to election será dividida en 2 secciones (mayor que la mediana y menor que la mediana la cual es 16.72) y aquellos lugares donde tiene NA's serán sustituidos por el promedio.

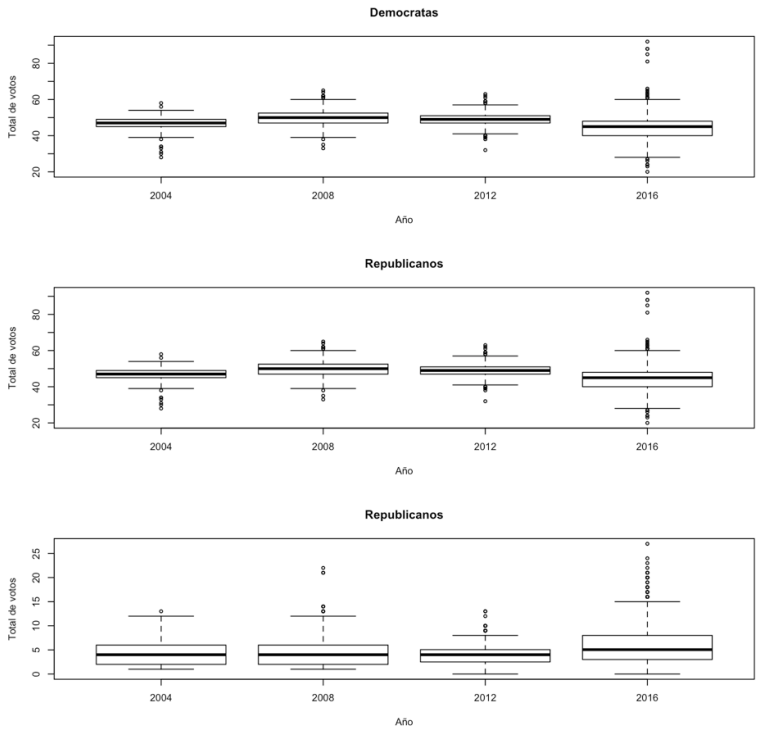
Al final nos quedamos con las siguientes variables:
- Votos
- Partido
- Estado
- Casa encuestadora
- Quintil de días para las elecciones
- Año

3.0.1 Código utilizado para tratar datos

```
library(bayesloglin)
library(tidyverse)

datos <- read.csv("us-pres-state-polling-2004-2016.csv")

NA2mean <- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))
replace(datos, TRUE, lapply(datos, NA2mean))
datos[] <- lapply(datos, NA2mean)
datos$year <- as.factor(datos$year)

datos$tile_days <- ntile(datos$mean_days_to_election, 2)

datos <- datos %>%
    select(Dem_poll, Rep_poll, Undecided, state, pollster_grp, tile_days
    gather(partido,Votos, -state, -pollster_grp, -tile_days, -year) %>%
    mutate(
        partido = as.factor(partido),
        tile_days = as.factor(tile_days)
    )
```

4 Exploración de datos

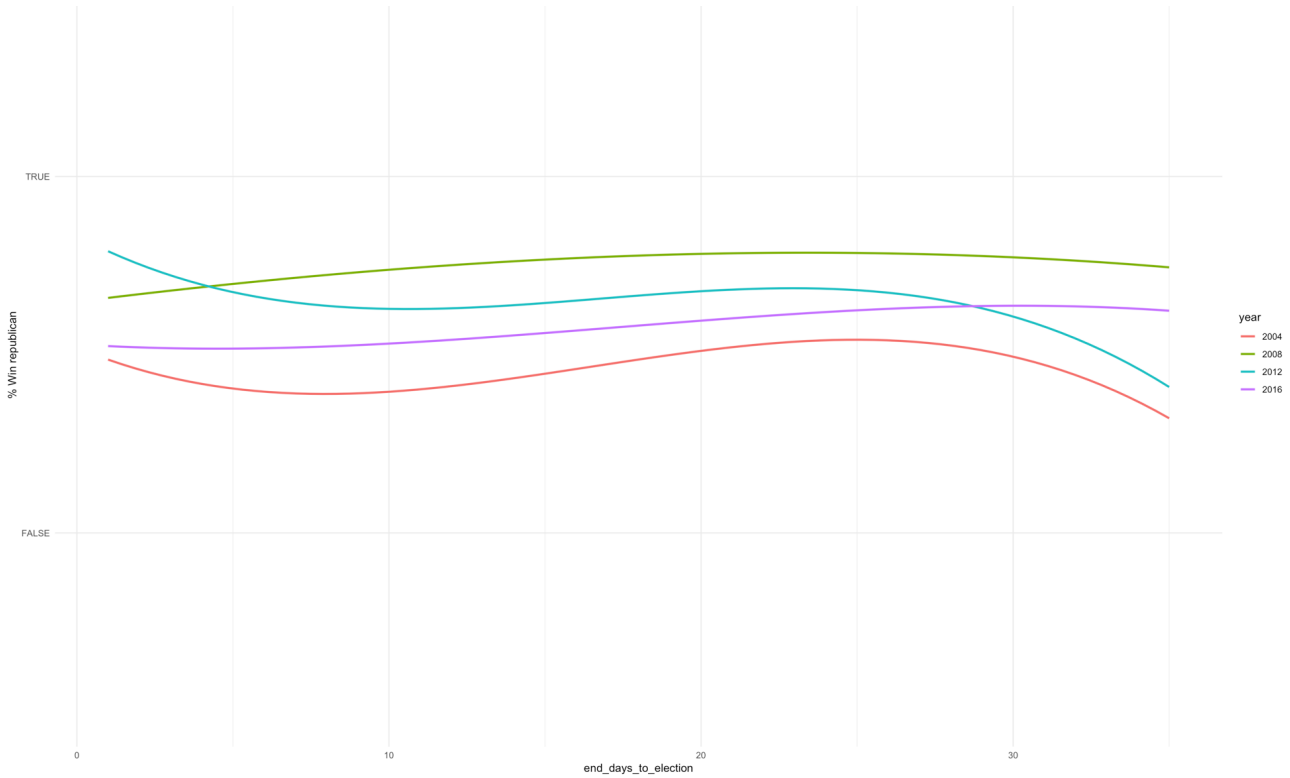
Si vemos cuantos votos demócratas, republicanos e indecisos hay en el dataset por año, vemos que esto está balanceado:



En todo el set de datos, el 61% de las encuestas dan por ganador a los democratas. Visto esto por año:
- 2004 44.5%
- 2008 57.7%
- 2012 60.37%
- 2018 76.67%

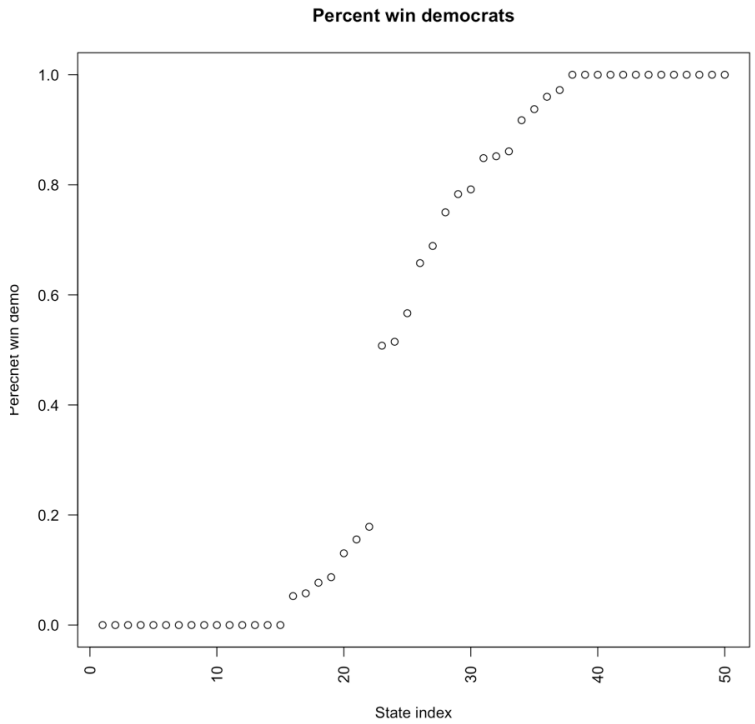
Por lo que claramente el año y el voto a X partido es un factor importante.

Si vemos como fue avanzando el porcentaje de voyos a los democratas respecto al día de la elección:

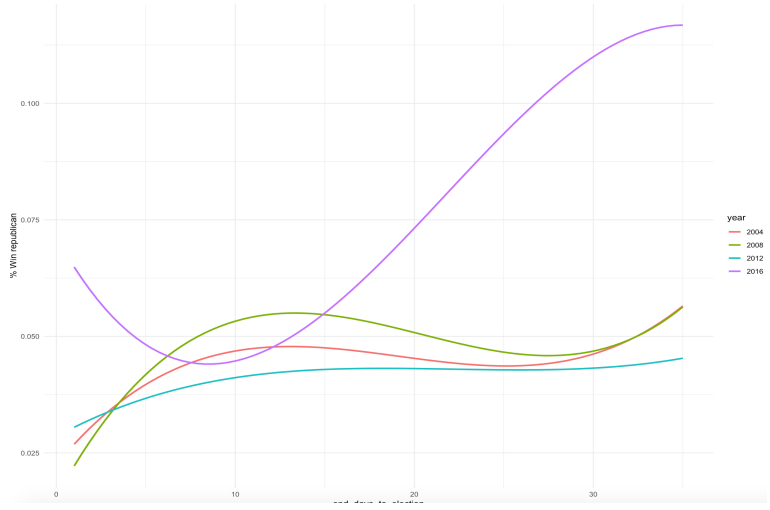


Vemos que no es la misma tendencia en cada año.

Existen estados donde las encuestas siempre dieron como ganadores a los demócratas como Hawai, Nueva York o Washingon mientras otros tienen siempre a republicanos como ganadores como Kansas, Texas o Alabama.



Viendo ahora los indecisos, al final hubo más en 2016, quizá por eso Trump pudo ganar.



La correlación entre los días para la elección y el voto democrata es negatia

pero pequeña

[1] −0.102956

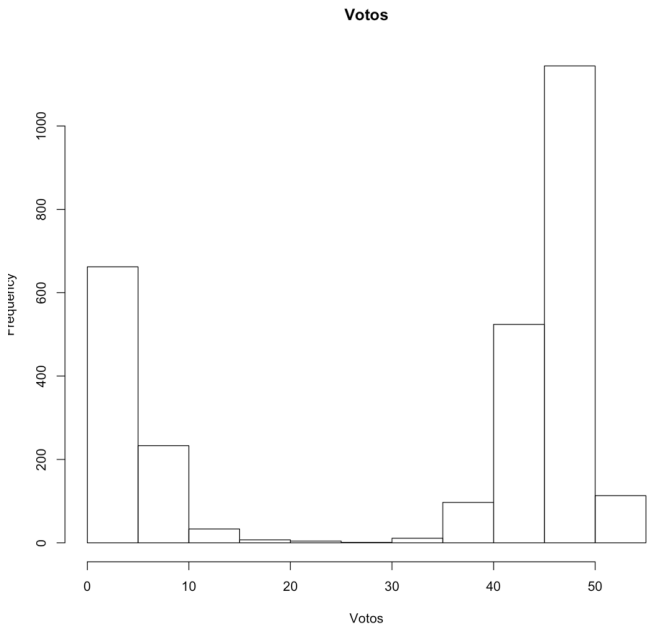
Contraintuitivamente también lo es para el voto republicano.

[1] −0.1883793

En realidad donde hay correlación alta es para el los indecidos, obviamente mientras más días faltan para la elección las personas están más indecisas.

[1] 0.3430182

En nuestro set de datos final, observamos el histograma del total de votos, vemos que es multimodal.



5 Resultados

5.1 Modelo con todos

Ahora utilizamos todos los datos. - Partido
- Estado
- Año

- Bucket de días promedio para elección
- Encuestadora
- Votos
Se usaron 2,000 iterciones.

5.1.1 Resultados

(state,pollster grp,tile days,year id)(state,partido) 609996.1
(state,pollster grp,tile days,year id)(partido) 609469.4
(state,pollster grp,tile days,year id)(state,tile days,partido) 609309.8

5.1.2 Interpretando el modelo con todas las variables

- Tenemos 2 secciones, estado, encuesta, días y año y por otro lado estado y partido
- Parece ser que estas dos secciones son independientes dado el año, es decir en distintos años hubo comportamientos distintos que no tienen que ver

5.1.3 Código para generar el modelo

```
model.graf <- MC3(iterations = 20000,
                  data = datos,
                  mode = "Decomposable")
```

6 Preguntas

6.1 Evaluar si la intención de voto en polling por estado es indpeendientemente inducida por el tiempo

Para evaluar esto tomamos las variables:
- Partido

- La variable de los días dividida en 2 cuantiles

- Estado

- Frecuencia

Utilizamos 10,000 iteraciones

6.1.1 Resultados

Los dos mejores modelos fueron:
- (partido,tile days)(state)
- (partido)(tile days)(state)
El mejor modelo, puede interpretarse como:

$$log\mu_{i,j,k} = \lambda_{partido} + \lambda_{dias} + \lambda_{estado} + \lambda_{partido,dias}$$

Es decir le estaríamos agregando la dependencia de partido y días, diferente al modelo donde todos son independientes.

Podemos concluir que el tiempo y el estado si son independientes.

6.1.2 Código para generar este modelo

```
datos1 <- datos %>% select(partido,tile_days, state, freq)
model.graf <- MC3(iterations = 10000,
                  data = as.data.frame(datos1),
                  mode = "Graphical")
head(model.graf, n = 5)
```

6.1.3 Generando modelos con MASS

Si asumimos independencia y generamos el siguiete modelo

```
model <- loglm(formula = freq ~ partido *
               tile_days + state, data = datos)
deviance(model)
[1] 7202.091
```

Ahora bien, si incluimos la interacción de año y partido:

```
model <- loglm(formula = freq ~ partido + state +
               year_id + partido*year_id, data = datos)
deviance(model)
[1] 6537.015
```

Con el año la devianza baja. Ahora bien incluyendo la interacción estado partido:

```
model <- loglm(formula = freq ~ partido + state +
               year_id + partido*year_id + partido*state, data = datos
```

```
deviance(model)
[1] 4069.136
```

Vemos como efectivamente existe interacción entre las variables.

6.2 Evaluar si la casa encuestadora es independientes de los resultados de la intención de voto.

Según el modelo con todas las variables, tomando como aproximación aquel con mejores resultados en la log verosimilitud, no hay relación.

6.3 Asociar a nivel estatal el resultado actual con los resultados anticipados agregados de los polling

Según el modelo con todas las variables y tomando como aproximación aquel con mejores resultados en la log verosimilitud si hay una relación entre el estado y el partido.

6.4 Identificar el efecto de la edición (año) de la elección presidencial en las preguntas anteriores

Según nuestra exploración de datos existe una diferencia marcada pero una vez que se mete al modelo solo hay relación con el estado.

7 Referencias

- Coppi, An Introduction to Multi-Way Data and Their Analysis.
- Agresti, Categorical Data Analysis
- Razia, Categorical Data Analysis for the behavioral and social sciences