Documento Ejecutivo - Estadística multivariada

Luis Federico Puente 103108

May 23, 2019

1. Introducción

Durante los últimos años los resultados de las elecciones en Estados Unidos de América (EUA), así como las encuestas de dichas elecciones han generado gran expectativa en el mundo académico. Por ejemplo, en cuantificar el efecto de dichas encuestas en la intención de voto de las personas.

En este sentido, el objetivo de este documento es analizar los siguientes puntos:

1) Evaluar si la intención de voto es independientemente inducida por la cercanía al día de la elección. En términos estadísticos se busca conocer la interacción/comovimiento/dependencia entre la ambas variables.
2) Evaluar si la edición de la elección tiene algún efecto en el punto previo.

2. Descripción de los modelos y relevancia para los datos

Tabla de contingencia

El primer paso que realiza es generar la tabla de contingencia asociada a los datos de las encuestas de EUA.

Para un conjunto de datos de J `variables/dimensiones`, donde cada dimension contiene K_j categorias, $\mathcal{X}_j = \{1, \ldots, K_j\}$, la `tabla de contingencia` J-dimensional se define como el arreglo J-dimensional

$$T_J = \left(n_{e(1),\ldots,e(J)}\right),$$

donde

$$e(j) \in \mathcal{X}_j,$$

para $j = 1, \ldots, J$.

Es decir, cada dimension j del arreglo contiene K_j entradas.

Además, se cumple que

$$n = \sum_{\{e(1),\ldots,e(J)\}} n_{e(1),\ldots,e(J)},$$

donde

$$n_{e(1),\ldots,e(J)} = \#\{x_i : x_{i1} = e(1), \ldots, x_{iJ} = e(J)\},$$

donde x_is corresponden a los vectores J-dimensionlaes que componen los n datos, con

$$x_i = (x_{i1}, \ldots, x_{iJ}).$$

Para el caso concreto de las elecciones en EUA se utilizan las sigientes dimensiones con sus respectivas entradas, esta fue la forma de agrupar los datos:

- Variable 1 (result_margin6). Toma 3 valores: Strong Rep, Strong Dem y Close margin
- Variable 2 (state). Toma 50 valores relacionados con los estados de EUA
- Variable 3 (tiempo). Toma 9 valores de acuerdo a la cercanía del día del levantamiento de la encuesta y el día de la elección, cada intervalo se seleccinó considerando aproximadamente 5 días

Cabe destacar que uno de lo principales propositos del uso de `tablas de contingencia` es analizar las `interacciones` o `comovimientos` o `dependencia` entre las X_js (en todos sus categorias). En el caso del ejemplo de las elecciones en EUA: evaluar si la intención de voto en polling por estado es independientemente inducida por el tiempo de duración para el día de la elección. Es decir, dos de las dimensiones del modelo.

Para el caso de las elecciones se sigue la siguiente notación:

- Denotaremos el conjunto de `vertices` como

  $$V = \{1, 2, 3\},$$

  correspondientes a las tres `dimnesiones/variables`

  $$\{Intenciónde voto, estado de EUA, días previos a la elección\}.$$

- El `vector aleatorio discreto`, X_i, es denotado por

  $$X_i = (X_{i,v} : v \in V),$$

  para $i = 1, \ldots, n = 490$. Este valor de 490 corresponde al número de combinaciones entre las tres dimensiones que tuvieron frecuencia positiva.

- La `configuracion` de X_i, que se refiere a las categorias especificas de la observacion i, se denota como

  $$x_i = (x_{i,v} : v \in V),$$

  donde cada

  $$x_j = e(j) \in \mathcal{X}_j,$$

  para $j = 1, \ldots, J = 3$.

- Los `conteos` son las entradas de la `tabla de contingencia`, y corresponden a los casos

  $$n_{e(1),e(2),e(3)} = \#\{x_{i,v} : x_{i,1} = e(1), x_{i,2} = e(2), x_{i,3} = e(3)\}.$$

- La `probabilidad` de que una observacion X_v sea igual a $e(v) = (e(1), \ldots, e(J))$ es denotada por

  $$\theta_v = p(x_v) = \mathbb{P}(X_v = x_v) = \mathbb{P}(X_1 = e(1), \ldots, X_J = e(J)).$$

La tabla de contingencia para el caso de las elecciones de EUA quedo definida de la siguiente forma

```
head(datos)
```

```
##   result_margin6          state   tiempo
## 1    Strong Rep. washington-d-c   Day0-5
## 2    Strong Rep. washington-d-c  Day5-10
## 3    Strong Rep. washington-d-c Day10-15
## 4    Strong Rep. washington-d-c Day10-15
## 5    Strong Rep. washington-d-c Day15-20
## 6    Strong Dem.  west-virginia   Day0-5
```

Especificación del modelo loglineal

La especificación del modelos sigue una forma loglineal para estimar las `probabilidades`:

$$\log \theta_v = \alpha^0 + \alpha^{12}_{e(1,2)} + \alpha^{23}_{e(2,3)} + \alpha^{13}_{e(1,3)},$$

donde $\alpha^{jk}_{e(j,k)}$ corresponde a un factor particular para las `dimensiones` $\{jk\}$, con $j, k = 1, \ldots, J = 3$, y sus correspondientes `categorias` $\{e(j), e(k)\}$.

3. Descripción del método de agregación

La forma en la que se agruparon los datos fue la siguiente:

- Variable 1 (result_margin6). Toma 3 valores: Strong Rep, Strong Dem y Close margin
- Variable 2 (state). Toma 50 valores relacionados con los estados de EUA
- Variable 3 (tiempo). que toma 9 valores de acuerdo a la cercanía del día del levantamiento de la encuesta y el día de la elección, cada intervalo se seleccinó considerando aproximadamente 5 días

Cabe destacar que en la base original de las elecciones de la elección existían valores perdidos (NA's) en la variable de dias medios para la elección. En estos casos, y con el objetivo de no perder información, se utilizó la información dispoble de el inicio o término del día del levantamiento de la encuesta.

4. Descripción del procedimiento de estimación

Para realizar las estimaciones se utilizó el algorítmo MC3 que utiliza la forma log-lineal en el modelo para encontrar la mejor probabilidad posterior.

Adicionalmente, con el objetivo de cuantificar el efecto de la edición (año) se calculo este modelo condicionando la estimación los años de la elección, con el objetivo de comparar/contrastar los resultados.

Para realizar el cálculo de las `verosimilitudes` y `probabilidades` se utilizó la libreria `bayesloglin` de R, de la siguiente forma:

```
suppressMessages(library("bayesloglin"))
model.graf <- MC3 (init = NULL,
                   alpha = 1,
                   iterations = 5000,
                   replicates = 1,
                   data = datos.tc,
                   mode = "Graphical")
```

5. Resultados

A continuación se muestra una tabla en la que es posible apreciar los generadores especificos mas importantes (i.e. con mayor probabilidad):

```
head(model.graf, n = 10)
```

```
##                                        formula logPostProb
## 1                [result_margin6,state][tiempo]    2676.006
## 2               [result_margin6][state][tiempo]    2675.948
## 3          [result_margin6,state][state,tiempo]    2672.647
## 4                [result_margin6][state,tiempo]    2672.472
## 5 [result_margin6,state][result_margin6,tiempo]    2672.316
## 6                [result_margin6,tiempo][state]    2672.254
## 7         [result_margin6,tiempo][state,tiempo]    2668.767
## 8                 [result_margin6,state,tiempo]    2666.321
```

Los resultados inferenciales para el `mejor modelo`, correspondiente al generador

$$\mathcal{C}^* = \{\{intenciónde voto, estado\}, \{cercaníaalaelección\}\}$$

Es decir, que existe una fuerte interacción entre las variables intensión de voto y estado de EUA, además de que cercanía al día la elección también es relevante pero independiente de las primeras dos variables.

Para calcular la media posterior se utilizó la función `findPostMean`, que para los modelos descomponibles, se conoce en forma cerrada.

```
library(bayesloglin)

colnames(datos.tc)[4] <- "freq"
sum(datos.tc$freq)
```

```
## [1] 1905
```

```
formula <- freq  ~ result_margin6 + state + tiempo

s <- findPostMean (formula, alpha = 1, data = datos.tc)
print(s)
```

```
##     (Intercept) result_margin62 result_margin63          state2
##     -1.01537850     -0.92719823     -0.47797314      0.38460392
##          state3          state4          state5          state6
##      1.65297058      0.53273551      1.53384458      2.35147830
##          state7          state8          state9         state10
##      0.82855992     -0.63245457      2.86444306      1.79899681
##         state11         state12         state13         state14
##     -0.63245457      0.00000000      0.53273551      1.67518293
##         state15         state16         state17         state18
##      2.02399630      0.30140925      0.72048979      0.38460392
##         state19         state20         state21         state22
##      0.77598368      0.38460392      1.24212684      2.00837618
##         state23         state24         state25         state26
##      1.83708100      0.38460392      1.87376787      0.97149697
##         state27         state28         state29         state30
##     -0.12468828      2.12699631      2.27005784      1.75940465
##         state31         state32         state33         state34
##      1.33890571      0.82855992      2.46737187      0.21066514
##         state35         state36         state37         state38
##      2.85773255      0.59931341      1.42713997      2.64882962
##         state39         state40         state41         state42
##     -0.63245457      0.38460392     -0.43325138      0.00000000
##         state43         state44         state45         state46
##      0.59931341      0.72048979     -0.43325138      2.41635125
##         state47         state48         state49         state50
##      1.50821670     -0.63245457      0.77598368      2.22037116
##         tiempo2         tiempo3         tiempo4         tiempo5
##      0.09902985     -0.39236826     -0.74785584     -0.27691050
##         tiempo6         tiempo7         tiempo8         tiempo9
##     -0.91654386     -1.94691247     -2.08470363      0.13444515
```

Adicionalmente, tambien se avaluó el mismo modelo pero condicionando para el año de la eleción. Esto con la intensión de analizar si a lo largo del tiempo ha variado la interacción entre la intensión de voto considerando la cercanía de la elección. A continuación, se muestran los resultados para cada año.

Resultados por año (2004, 2008, 2012 y 2016)

Para 2004:

```
##                                       formula logPostProb
## 1 [result_margin6][state][tiempo]    275.0721
```

```
## 2                [result_margin6][state,tiempo]    273.9466
## 3                [result_margin6,tiempo][state]    272.7969
## 4                [result_margin6,state][tiempo]    272.2703
## 5         [result_margin6,tiempo][state,tiempo]    271.5035
## 6          [result_margin6,state][state,tiempo]    271.1500
## 7 [result_margin6,state][result_margin6,tiempo]    270.0108
## 8                 [result_margin6,state,tiempo]    266.0026
```

Para 2008:

```
##                                       formula logPostProb
## 1                [result_margin6][state,tiempo]    596.3743
## 2               [result_margin6][state][tiempo]    595.8102
## 3          [result_margin6,state][state,tiempo]    593.3175
## 4         [result_margin6,tiempo][state,tiempo]    593.2270
## 5                [result_margin6,tiempo][state]    592.7947
## 6                [result_margin6,state][tiempo]    592.7828
## 7 [result_margin6,state][result_margin6,tiempo]    589.7712
## 8                 [result_margin6,state,tiempo]    587.0419
```

Para 2012:

```
##                                       formula logPostProb
## 1               [result_margin6][state][tiempo]    534.6916
## 2                [result_margin6,tiempo][state]    532.6227
## 3                [result_margin6,state][tiempo]    531.9709
## 4                [result_margin6][state,tiempo]    531.6739
## 5 [result_margin6,state][result_margin6,tiempo]    529.8854
## 6         [result_margin6,tiempo][state,tiempo]    529.6254
## 7          [result_margin6,state][state,tiempo]    528.9283
## 8                 [result_margin6,state,tiempo]    524.1042
```

Para 2016:

```
##                                       formula logPostProb
## 1               [result_margin6][state][tiempo]    493.2862
## 2                [result_margin6,tiempo][state]    491.0380
## 3                [result_margin6,state][tiempo]    490.2203
## 4                [result_margin6][state,tiempo]    490.1515
## 5 [result_margin6,state][result_margin6,tiempo]    487.9741
## 6         [result_margin6,tiempo][state,tiempo]    487.9255
## 7          [result_margin6,state][state,tiempo]    487.0984
## 8                 [result_margin6,state,tiempo]    481.8787
```

6. Comentarios finales

El principal resultado es que en caso de que se agregan los datos de todos los años, los resultados inferenciales para el `mejor modelo`, correspondiente al generador:

$$\mathcal{C}^* = \{\{intenciónd evoto, estado\}, \{cercaniaalaelección\}\}$$

Sin embargo, esta interacción cambia al estimar los modelos segmentando por año, ya que en la mayoría de los casos se obtuvo (salvo en 1 año):

$$\mathcal{C}^* = \{\{intenciónd evoto\}, \{estado\}\{cercaniaalaelección\}\}$$

Es decir, el año si tiene una impacto en la interación de variables.

7. Bibliografia

Bon, J., Joshua and Ballard, Timothy and Baffour, Bernard (2018). Polling bias and undecided voter allocations: US presidential elections, 2004-2016. Journal of the Royal Statistical Society: Series A (Statistics in Society).

Notas del curso de Estadística Multivariada. Profesor: Juan Carlos Martínez Ovando. ITAM. Primavera 2019.