Numérika®

Mayo 2018

Modelo de agregación de encuestas

Fernando Antonio Zepeda Herrera

Objetivo

A lo largo de un proceso electoral, las diferentes casas encuestadoras y campañas de los candidatos realizan encuestas con el objetivo de obtener información sobre el estado de las preferencias de los votantes. Debido a la volatilidad natural de las encuestas, en los últimos años han cobrado relevancia los modelos de agregación de encuestas, conocidos como *Poll of Polls*. Uno de los principales atractivos de estos modelos es que, al considerar varias estimaciones individuales, utilizan una mayor cantidad de información. Asimismo, permiten realizar estimaciones más estables pues tratan de separar el ruido estadístico de la tendencia real. De manera particular, el modelo aquí presentado busca reflejar la incertidumbre existente sobre la intención de voto presidencial en cada semana previa a la jornada electoral, agregando los resultados de las encuestas publicadas por distintos medios electrónicos y escritos desde enero de 2018.

Modelo

Considerando el objetivo de agregación de encuestas, podemos pensar que necesitamos un modelo que contemple los siguiente supuestos:

- En cada punto en el tiempo existen unas *preferencias subyacentes* entre los votantes que determinarían el resultado de la elección, *si esta se realizara en ese momento*.
- Cada encuesta busca estimar dichas preferencias pero, debido a errores muestrales y no muestrales, el dato obtenido es una desviación de las mismas.
- Dicha desviación depende, primordialmente, de la casa encuestadora que realizó el ejercicio demoscópico a través de dos fuentes independientes:
    - Las posibles propensiones, constantes en el tiempo, a sobre o subestimar a cada candidato. Es decir, pueden existir sesgos constantes de casa.
    - Asimismo, cada casa encuestadora tiene una volatilidad común entre sus encuestas, proveniente de todos los errores muestrales y no muestrales que no pueden clasificarse como parte de los sesgos constantes.

En este sentido, resulta atractiva la propuesta hecha por Cargnoni, Müller y West (1997) para modelar series de tiempo multinomiales. Con base en lo expuesto por los autores de dicho trabajo, cada encuesta se modelará como una realización de una distribución multinomial con parámetros propios:

$$Y_e \sim Mult(\theta_e, n_e)$$

Los autores proponen lo que ellos llaman un modelo condicionalmente normal que modele el vector de probabilidades θ_e mediante una transformación no lineal. En lugar de modelar directamente desde el simplex— como hacen los modelos basados en distribuciones Dirichlet, por ejemplo— se utiliza una transformación que permita "flexibilidad y libertad para describir patrones arbitrarios de correlación" (p.640, traducción propia).

Av. Colonia del Valle 511, piso 1
Colonia Del Valle, C. P. 03100
Del. Benito Juárez, Ciudad de México
Tel: (55) 63.63.62.97

Cada encuesta tiene asociada una *medición latente* η_e relacionada con las probabilidades multinomiales θ_e mediante la transformación logística aditiva:

$$\eta_{e,i} = ln\left(\frac{\theta_{e,i}}{\theta_{e,I}}\right) \quad \forall\, i = 1, \ldots, I-1,$$

donde I es el número de candidatos.[1]

Son estos parámetros transformados los que se modelarán de una manera jerárquica y dinámica que incorpora los supuestos antes enlistados. La medición, se asume, se distribuye de manera normal, centrada en una media que depende del nivel "real" del sistema en la semana de la encuesta— $\mu_{t[e]}$— y de los posibles sesgos constantes de la casa encuestadora— $\gamma_{c[e]}$—; su varianza depende exclusivamente de un parámetro de dispersión de la casa encuestadora, $V_{c[e]}$:

$$\eta_e \sim N(\mu_{t[e]} + \gamma_{c[e]}, V_{c[e]}).$$

La evolución del sistema depende del último punto en el tiempo, teniendo como origen una media inicial m:[2]

$$\mu_t \sim N(\mu_{t-1}, W),$$
$$\mu_0 \sim N(m, W).$$

Se requiere completar el modelo asignando distribuciones iniciales a los parámetros. Para no imputar, *a priori*, un sesgo determinado a ninguna casa encuestadora, se decidió que las distribuciones de los sesgos fueran normales centradas en 0:

$$\gamma_c \sim N(0, \Omega_c) \quad \forall\, c = 1, \ldots, C,$$

donde C es el número de casas encuestadoras en muestra. A todos los parámetros de varianza les fueron asignadas distribuciones iniciales siguiendo la sugerencia de Cargnoni, Müller y West y mediante pruebas con datos de encuestas para las elecciones presidenciales en México de 2006 y 2012. De manera concreta, estas fueron distribuciones Wishart inversas con 6 grados de libertad y matrices de escala diagonales con elementos iguales a 0.05.

Referencias

Cargnoni, Claudia, Peter Müller y Mike West. 1997. "Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models". *Journal of the American Statistical Association* 92 (438): 640-647.

[1] Esta es una transformación diferente a la de seno inverso que utilizan Cargnoni, Müller y West en su aplicación, principalmente porque garantiza que el resultado sea un vector de probabilidades válido, lo que permite eliminar un paso de aceptación y rechazo a la hora de hacer las simulaciones para obtener la distribución posterior.

[2] Cabe mencionar que la elección de este parámetro no influye de manera determinante en la estimación actual, pues los datos van dominando a la inicial. Sin embargo, por transparencia, reportamos que se utilizaron valores equivalentes a una distribución efectiva entre los candidatos con los siguientes porcentajes: Andrés Manuel López Obrador 35%, Ricardo Anaya Cortés 30%, José Antonio Meade 25%, Margarita Zavala 5% y Jaime Rodríguez "El Bronco" 5%.

Av. Colonia del Valle 511, piso 1
Colonia Del Valle, C. P. 03100
Del. Benito Juárez, Ciudad de México
Tel: (55) 63.63.62.97