Estadística Multivariada y Datos Categóricos

Proyecto Final

000113018 Paulina Gómez Mont Wiechers
000124433 Alejandra Lelo de Larrea Ibarra
000130901 Bernardo García Bulle Bueno
000144089 Laura López Santibañez Jácome

23 de mayo del 2018

Resumen

Este documento presenta un análisis de modelos loglineales para grafos probabilísticos con la finalidad de poder responder preguntas específicas sobre si la intención de voto ha tenido algún impacto en el proceso electoral de Estados Unidos para las elecciones de 2004, 2008, 2012 y 2016. En general, encontramos que existe evidencia de que la intención de voto es condicionalmente independiente del estado en que se levante la encuesta condicional a la venta de tiempo previa a las elecciones. Además, encontramos que bajo el grafo probabilístico asociado, el modelo loglineal tiene una tasa de acierto del 55 % de los votos; mientras que el modelo reducido tiene una tasa más alta (87 % aproximadamente) por lo que parece que algunas variables del modelo están generando ruido.

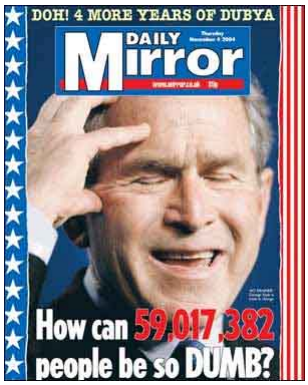
Índice

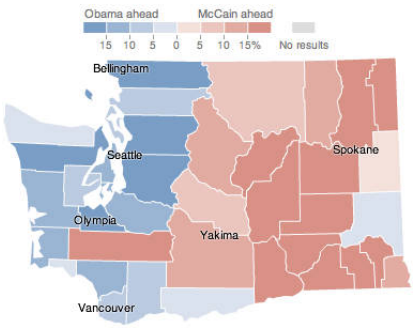
1. Introducción

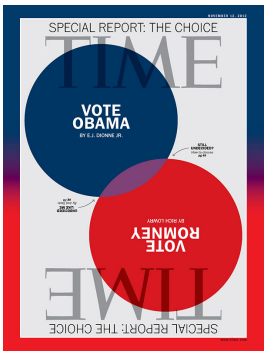
Contabilizar a los votantes indecisos e inciertos es una tarea difícil cuando se trata de predecir los resultados electorales a partir de las encuestas de opinión pública. Los votantes indecisos caracterizan la incertidumbre de los llamados "swing voters" en las encuestas, ya que éstos pueden cambiar por completo el resultado de una elección; sin embargo, a menudo éstos son ignorados o simplemente se les asigna un candidato al azar. Históricamente, esto puede haber sido adecuado ya que los votantes indecisos eran un grupo lo suficientemente pequeño como para suponer que no afectaba las proporciones relativas de los votantes decididos en el momento de la elección; no obstante, en las últimas elecciones la cantidad de votantes indecisos ha aumentado significativamente y ha mostrado ser causa de un cambio radical entre lo observado y lo esperado. [1]

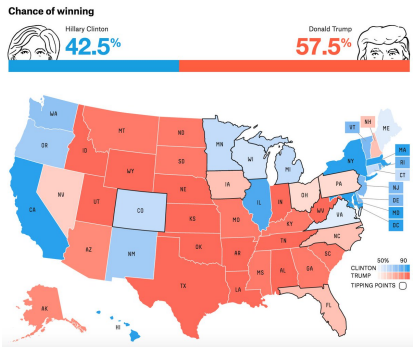
Ante dicho aumento, ignorar o simplemente asignar candidatos a los votantes indecisos puede sesgar las predicciones electorales y arrojar resultados incorrectos, como fue el caso de las últimas elecciones en Estados Unidos. Por ello, tomar en cuenta a los votantes indecisos como una categoría específica se ha vuelto un tema de gran importancia y es imperativo generar modelos estadísticos que nos ayuden a entender sus efectos potenciales.[1]

De esta manera, existen varias preguntas preguntas que serían interesantes de investigar. Por ejemplo, ¿cómo se ha distribuido la intención de voto por partido en cada año electoral?; para un determinado país, ¿ha sido distinta la intención de voto por partido a lo largo de las distintas elecciones?; ¿la intención de voto es distinta dependiendo de quién realice la encuesta?, ¿el periodo de anticipación a las elecciones en que se levanta la encuesta tienene algún impacto en la estimación de los votos indecisos? Responder algunas de estas preguntas ayudaría a tener claridad sobre cómo la intención de voto afecta el proceso electoral de un determinado país de tal forma que, tanto los políticos como los ciudadanos puedan tener estimaciones que sean un mejor reflejo de la situación electoral y puedan ejercer su derecho a voto de manera más informada.









Así, el objetivo de este proyecto, consiste en tratar de responder algunas preguntas relacionadas con estas interrogantes para las elecciones del 2004, 2008, 2012 y 2016 en Estados Unidos. Para ello, se utilizan los mismos datos y un enfoque similar al de Bon, J. J., Ballard, T., and Baffour, B. (2019). En particular, nos vamos a centrar en tratar de responder las siguientes dos preguntas: ¿la intención de voto por estado es independientemente inducida por el tiempo de duración promedio previo al día de la elección? y ¿existe una asociación a nivel estatal del resultado actual con los resultados anticipados agregados de los polling? La primer pregunta sería de gran utilidad, por ejemplo, para poder determinar el momento óptimo en que se deben reforzar las campañas de los candidatos en cada estado, dependiendo de cual se la intención de voto predominante. Adicionalmente, la segunda pregunta sería de gran utilidad para poder entender si los votos indecisos han sido determinantes en los resultados electorales a lo largo del tiempo o no.

Para poder repsonder las preguntas de interés, utilizamos un enfoque de estimación de modelos log-lineales a partir de grafos probabilísticos. En particular, utilizamos las variables relacionadas con la intención de voto por cada partido (e indecisos), el año electoral, la casa encuestadora, el número de días promedio antes de la elección en que se levantó la encuesta y el estado en que se levantó la encuesta. La única variable numérica en este análisis es el número de días promedio antes de la elección, por ello se categoriza esta variable tomando ventanas de tiempo de 5 días. Una vez realizadas las tranformaciones correspondientes, se estiman los posibles modelos de grafos generadores que maximizan la log-verosimilitud de la distribución posterior y se elige el mejor modelo que satisfaga la condición de ser triangular (*decomposable*). Lo anterior, con la finalidad de poder obtener distribuciones marginales y ahondar en la respuesta a nuestra preguntas de interés para buscar, entre otras cosas, independencia condicional de las variables.

En general, el modelo loglineal asociado al grafo probabilistico que mejor se ajusta a los datos muestrales es un grafo cíclico, por lo que no permite realizar un análisis más profundo. De esta manera, se utilizó el segundo gráfo con mayor logverosimilitud posterior como generado del modelo loglineal. A partir de dicho modelo, encontramos que existe evidencia

[1] Bon, J. J., Ballard, T., and Baffour, B. (2019).

de que la intención de voto es condicionalmente independiente del estado en que se levante la encuesta, condicional a la venta de tiempo previa a las elecciones. Además, encontramos que bajo el grafo probabilístico asociado, el modelo loglineal tiene una tasa de acierto del 55 % de los votos; mientras que el modelo reducido tiene una tasa más alta (87 % aproximadamente) por lo que parece que algunas variables del modelo están generando ruido.

El resto del documento está organizado del a siguiente manera. La sección 2 muestra un breve análisis exploratorio de los datos, con la finalidad de tener una idea de qué podemos esperar de la esitmación de los modelos. La descripción general del modelo a utilizar se encuentra en la sección 4. Posteriormente, la esitmación de los modelos junto con el método de agregación se desarrolla en la sección 3. Los resultados se presentan en la sección 5 y, en particular, las respuestas a las dos preguntas en las que nos enfocamos se encuentran en las secciones 5.1 y 5.2 respectivamente. Por último, la sección 6 presenta nuestras consideraciones finales. Cabe mencionar que todo el código utilizado para el desarrollo de este documento se incluye en el apéndice A.

2. Análisis Exploratorio de los Datos

En esta sección realizamos un breve análisis exploratorio de los datos en `pollster` y `voting` para entender un poco más a detalle su naturaleza y tener una mejor intuición sobre si los resultados arrojados por los modelos están alineados con lo observado. Los datos de `pollster` tienen 1905 observaciones y 20 variables. Entre estas variables destacan: la proporción de intención de voto por cada partido (Demócratas y Republicanos), la proporción de indecisos, el tamaño de la muestra, el número de días previos a la elección al inicio, al final y en promedio de la encuesta, el estado, el año de la eleccion y la casa encuestadora. Por su parte, los datos de `voting` tienen 129 observaciones y 10 variables. Entre estas variables se encuentran: el estado, el año electoral, así como el total de votos para demócratas y republicanos.

2.1. ¿Cómo se Distribuye la Intención de Voto por Partido y por Año electoral?

Como se puede obervar en la figura 1, la distribución de la intención de voto por partido y por año electoral ha sido muy distinta. Comparando dentro de un mismo partido, las elecciones de 2004 y 2012 tienen distribuciones muy similares que se destacan por están concentrdas alrededor del 50; mientras que la elecciones del 2008 y sobre todo de 2016 se caracterizan por tener mayor dispersión. Comparando entre partidos, la distribución de la intención de voto entre Demócratas y Republicanos ha sido similar a lo largo del tiempo. Ahora bien, la distribución de los indecisos se ha caracterizado por estár sesgada a la derecha y en las elecciones del 2016 tiene mayor dispersión que en el resto.

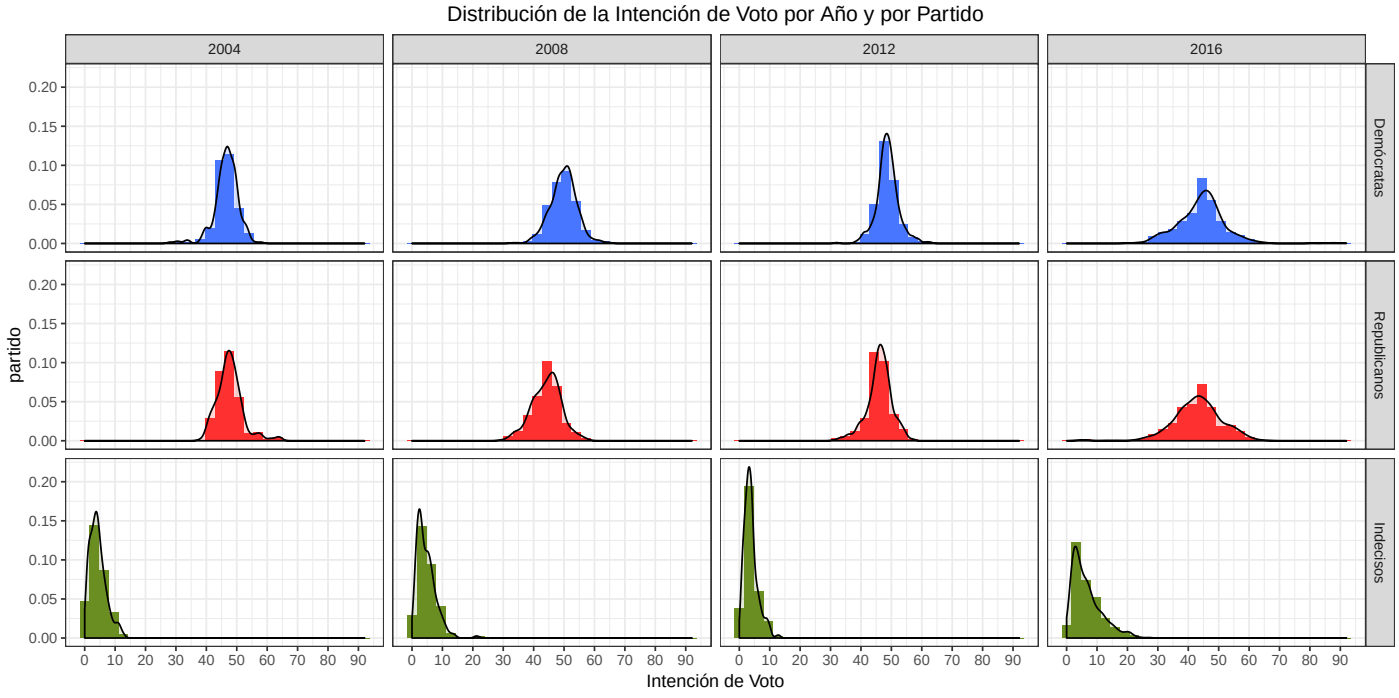


Figura 1: Intención de Voto por Partido y por Año Electoral.

2.2. ¿Cómo se Distribuye la Intención de Voto por Partido en Cada Estado Para Cada Año electoral?

La figura 2 muestra la intención de voto (promedio de todas las casas encuestadoras) por partido y por estado. Se puede notar que, la intención de voto por estado se mantiene realtivamente constante a lo largo del año; lo que nos hace pensar que es muy dificil que un estado cambie de ser Demócrata a Republicano o viceversa. En contraste, la proporción de indecisos en cada estado presenta mayor variabilidad. Comparando los estados, como era de esperarse, se puede notar que sí existen diferencias en la intención de voto de cada partido independientemente del año electoral.

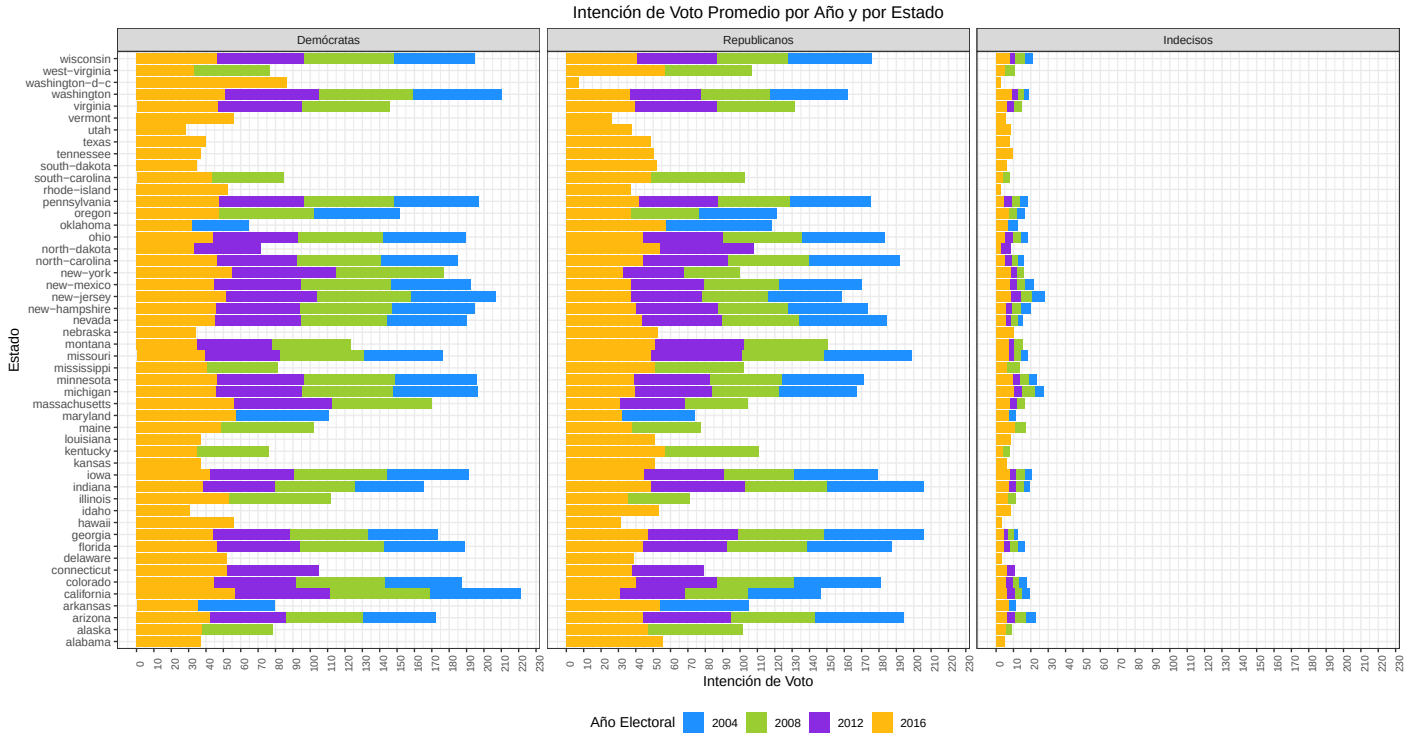


Figura 2: Intención de Voto Promedio por Estado y por Año Electoral.

2.3. ¿Cómo se Distribuye la Intención de Voto por partido en las Casas Encuestadoras?

De la figura 3 muestra la intención de voto para cada partido y por casa encuestadora (`pollster_id`). Se puede notar que, sí existen diferencias en la intención de voto de cada partido dependiendo de la casa encuestadora. Por ejemplo, con la casa encuestadora número 29 se tiene una mayor proporción de votos a los Demócratas mientras que con la casa encuestadora 32 se tiene una proorción similar de voto para republicanos y demócratas.

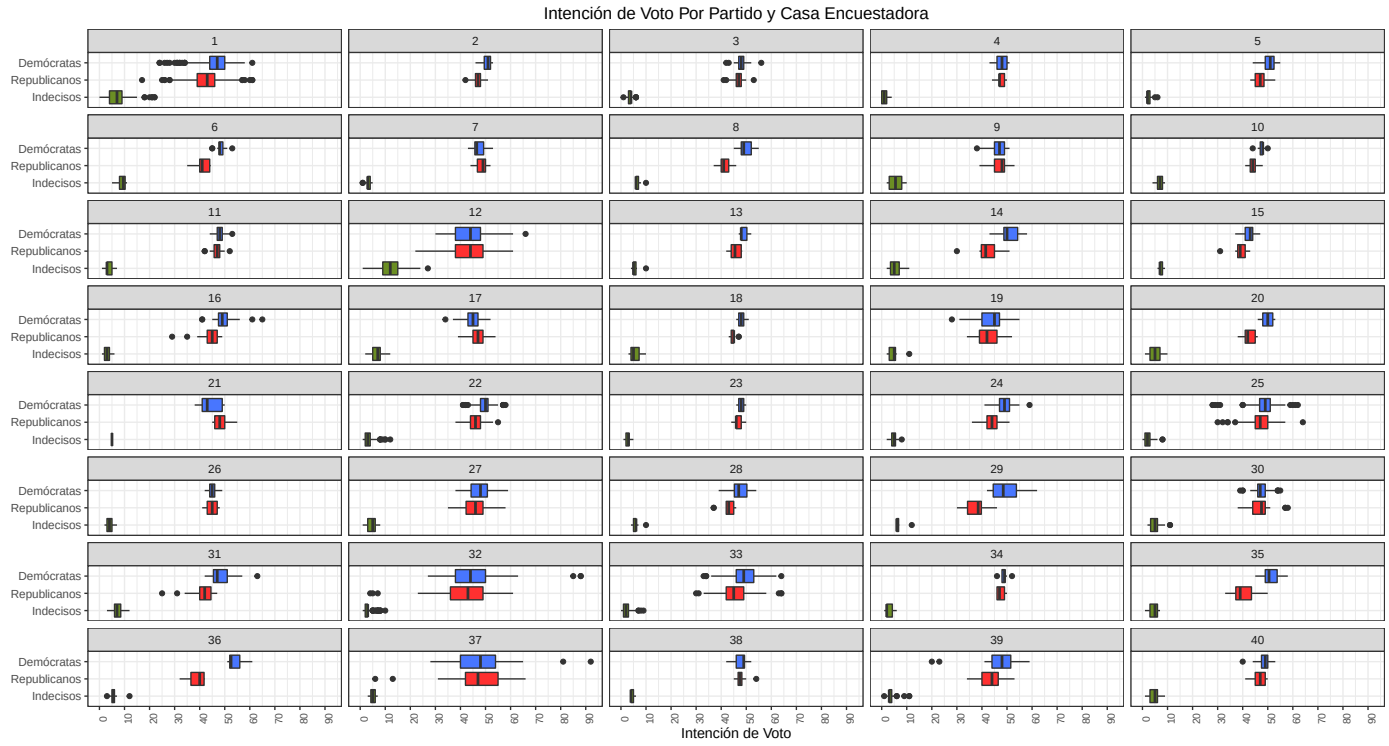


Figura 3: Intención de Voto por Estado y Casa Encuestadora.

2.4. ¿Hay diferencia entre la intención de votos y los votos observados?

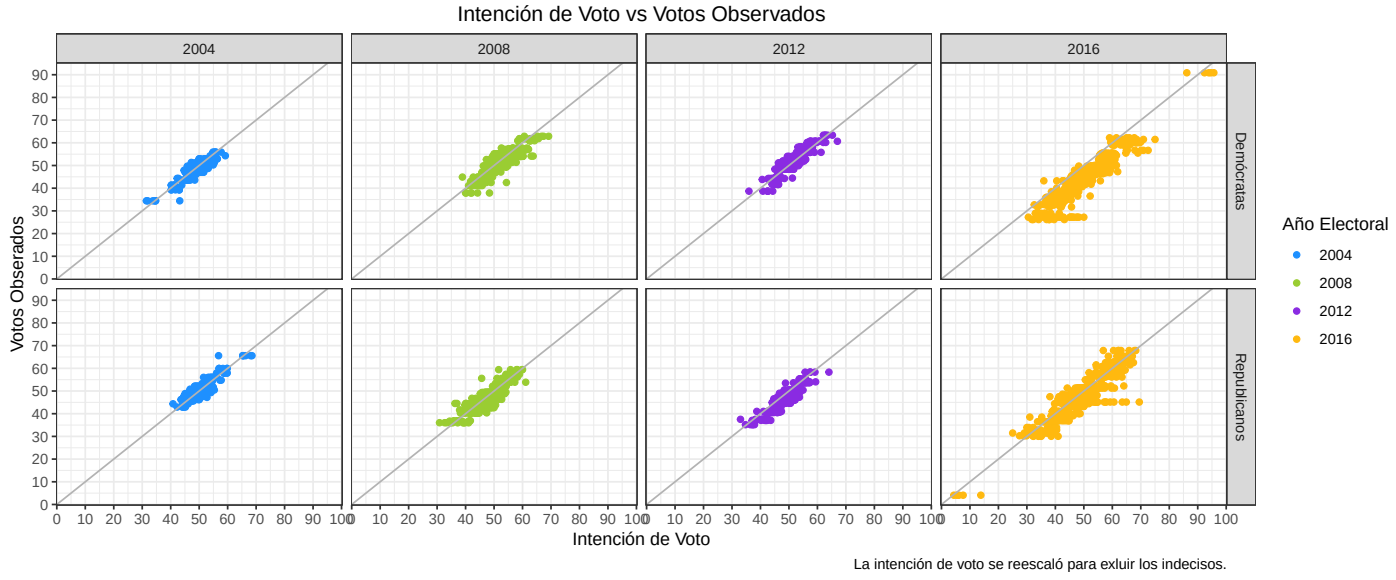


Figura 4: Intención de Voto por Estado y Casa Encuestadora.

Como se puede observar en la figura 4, la votación final fue muy similar a la intención de voto (ajustando para eliminar indecisos) tanto para Demócratas como para Republicanos en las elecciones de 2004, 2008 y 2012. En contraste, para el 2016, si bien los republicanos tienen una intención de voto muy similar a los votos observados, la intención de votos para los Demócrtas está evidentemente sesgada respecto a la votación observada; además, es donde se pueden observar más datos atípicos. Es ahí donde se puede ver el efecto de los indecisos en estas elecciones.

De este análisis exploratorio, surgen varias preguntas:

1. ¿La intención de voto por estado es independientemente inducida por el tiempo faltante para el día de la elección?
2. ¿La casa encuestadora es independiente de los resultados de la intención de voto?
3. ¿Los resultados son homogéneos entre las entidades de los Estados Unidos?
4. ¿Cuál es la asociación a nivel estatal del resultado actual de las elecciones con los resultados anticipados agregados de la intención de voto?
5. ¿Los rangos de indecisión en la intención de voto tienen un efecto sobre los resultados actuales?
6. ¿Cuál es el efecto de la edición electoral en las preguntas anteriores?

En las siguientes secciones trataremos de ahondar en la respuesta a las preguntas 1 y 4 utilizando modelos de grafos probabilísticos.

3. Descripción del Proceso de Estimación y Método de Agregación

3.1. Selección de Variables

Como ya se mencionó, para poder responder a algunas de las preguntas de interés, se plantea la estimación de grafos probabilísticos con los datos disponibles del set `pollster`. A continuación se describe el proceso de estimación de los modelos.

En primer lugar, de la base de datos se utilizaron las siguientes variables:

- `Dem_poll`: porcentaje de intención de voto por el partido Demócrata.
- `Rep_poll`: porcentaje de intención de voto por el partido Republicano.
- `Undecided`: porcentaje de indecisos.
- `sample_size`: tamaño de la muestra.
- `mean_days_to_election`: número de días promedio faltantes para las elecciones.
- `state`: estado en el que se levantó la encuesta.
- `year`: año electoral correspondiente.
- `pollster_id`: identificador de la casa encuestadora.

Una vez seleccionadas las variables de interés, las variables `Dem_poll`, `Rep_poll` y `Undecided` se convirtieron a frecuencia observada (en lugar de porcentaje de intención de voto) multiplicando por el tamaño observado de la muestra. [2]

Posteriormente, se crearon las variables `partido` e `int_voto` agrupando las columnas `Dem_poll`, `Rep_poll` y `Undecided` bajo la variable `Partido` y sus respectivas frecuencias bajo la variable `Int_Voto`. Además, la variable `mean_days_to_election` se categorizó utilizando cortes de 5 días.

3.2. Tabla de Frecuencias

Una vez realizadas todas las transformaciones necesarias, las variables categóricas (`partido`,`mean_days_to_election`,`state`,`year` y `pollster_id`) se convierten a factores con valores de 0 a $k_j - 1$ con k_j el número de categorías de la j-ésima variable y se eliminan aquellas observaciones con frecuencia cero en intención de voto. De esta manera, ya se tiene lista la tabla de contingencias (en formato largo) para poder hacer inferencia. De esta manera, los datos tienen la siguiente estructura:

[2] Cabe destacar que este procedimiento puede introducir sesgo en el sentido de un desbalanceo de la muestra. Sin embargo, se hizo un ajuste proporcional asumiendo que todas las encuestas tenían un tamaño de muestra fijo de 100 personas (e incluso un tampo de muestra fijo igual al tamaño de muestra promedio) pero, bajo esta especificación, los mejores modelos indicaban independencia incondicional de todas las variables. Por ello, y con fines académicos, se decidió utilizar el procedimiento antes mencionado a reservar de estar consciente de que las estimaciones podrían estar sesgadas.

	mean_days_to_election <dbl>	state <dbl>	year <dbl>	pollster_id <dbl>	partido <dbl>	int_voto <int>
1	0	47	3	31	0	264
2	1	47	3	36	0	287
3	2	47	3	31	0	292
4	2	47	3	36	0	251
5	3	47	3	31	0	444
6	0	48	3	31	0	137

3.3. Estimación de Modelos

Ahora bien, como vimos en clase, cuando las probabilidades del modelo son desconocidas, están referidas a cada una de las observaciones y, por lo tanto, a la tabla de contingencias muestrales; además, no podemos garantizar un único generador para los datos. Por ello, uitlizamos la función `MC3` del paquete `bayesloglin` en R para calcular la log-verosimilitud de aquellos posibles generadores, bajo el modelo log-lienal, que tengan evidencia de ser plausibles para los datos muestrales. En particular utilizamos los siguientes parámetros en la estimación:

- `init`= NULL
- `alpha`=1
- `iterations`=100,000
- `replicates`=1
- `mode`= Graphical

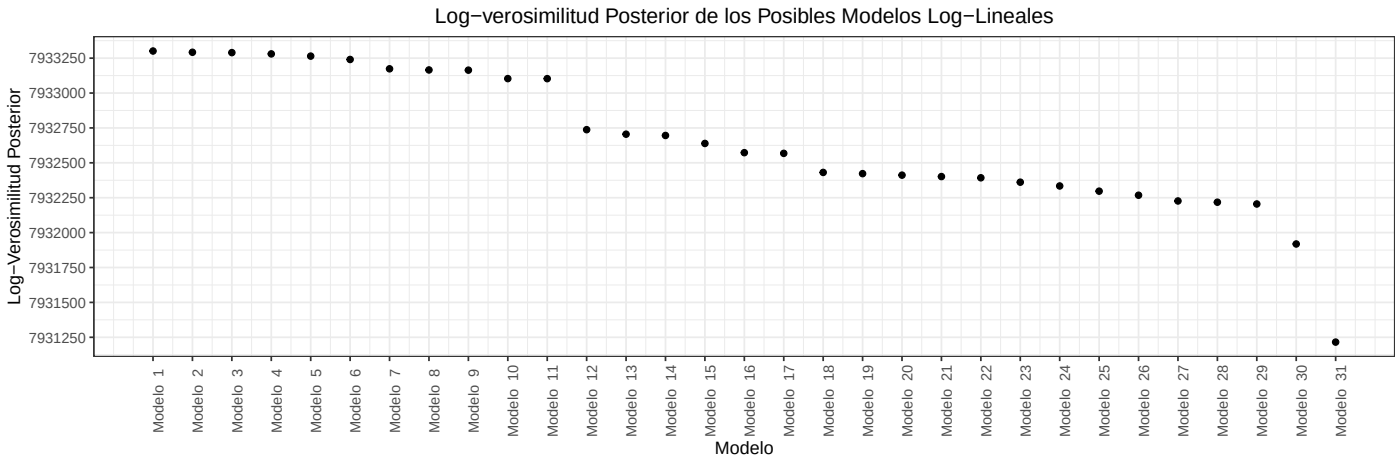


Figura 5: Log-Versimilitud Posterior de los Modelos Estimados.

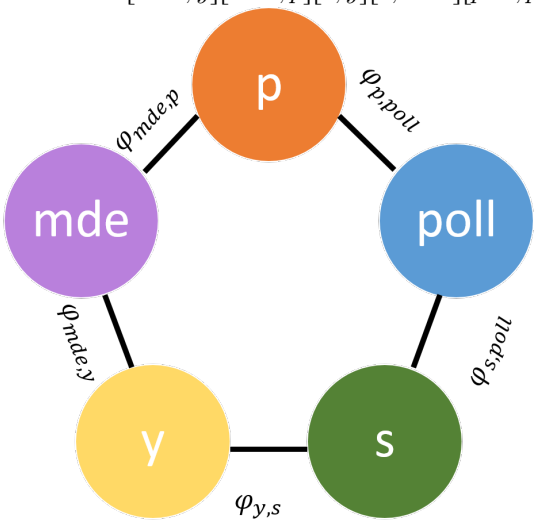
La figura 5 muestra la log-verosimilitud de los 31 distintos generadores obtenidos para los datos. Como se puede apreciar, los primeros 5 modelos obtienen prácticamente la misma log-verosimilitud, por lo que son los potenciales generadores para ajustar nuestros datos. Sea p :=`partido`, *mde* :=`mean_day_to_election`,s :=`state`,y :=`year` y *poll* :=`pollster_id`; la tabla 1 muestra el resultado y la configuración de estos 5 primeros modelos.

Tabla 1: Log-Verosimilitud Posterior de los Mejores Modelos

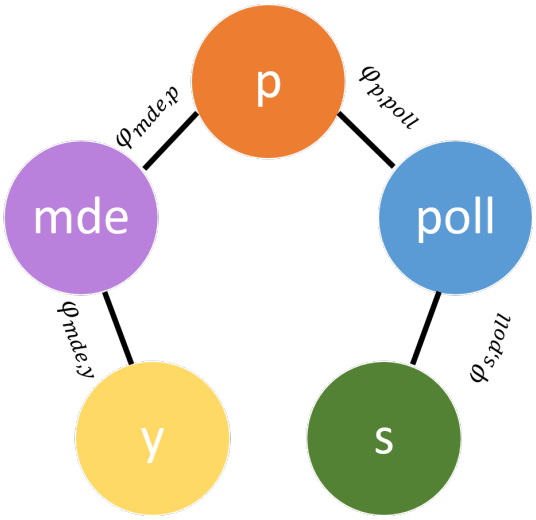
Modelo	Formula	LogVerosim
Mod1	[mde,y][mde,p][s,y][s,Poll][poll,p]	7933301.10
Mod2	[mde,y][mde,p][s,Poll][poll,p]	7933292.13
Mod3	[mde,s][mde,y][mde,p][s,poll][poll,p]	7933289.68
Mod4	[mde,s,y][mde,p][s,poll][poll,p]	7933280.42
Mod5	[mde,y][mde,p][s,y][s,poll,p]	7933264.63

Entonces, los 5 mejores modelos están representados gráficamente como

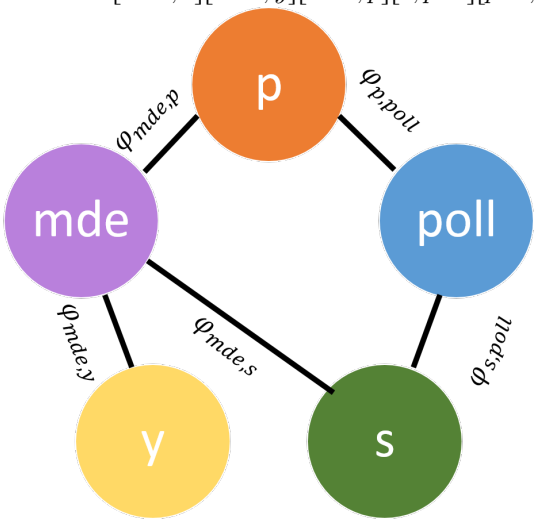
Modelo 1: $[mde, y][mde, p][s, y][s, Poll][poll, p]$



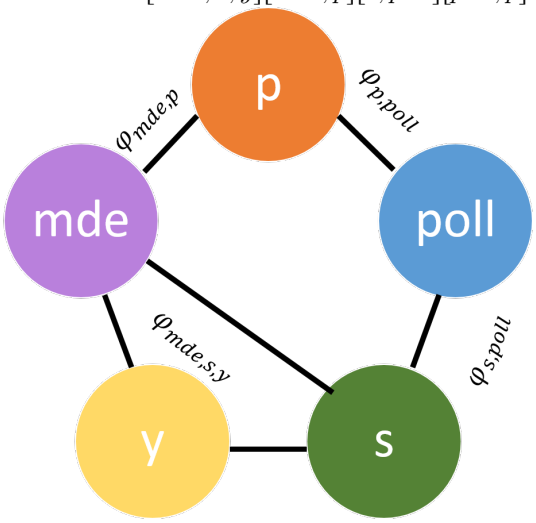
Modelo 2: $[mde, y][mde, p][s, Poll][poll, p]$



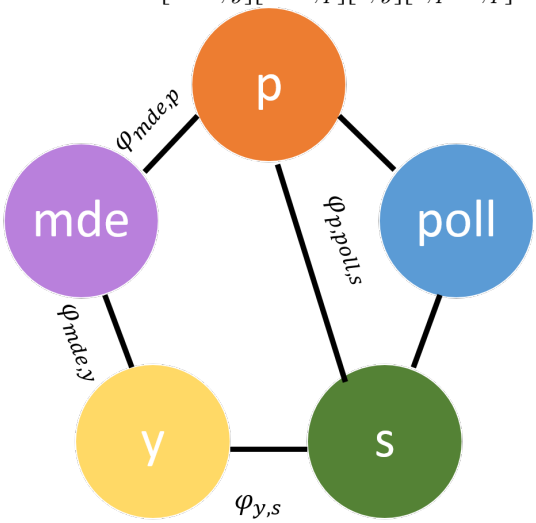
Modelo 3: $[mde, s][mde, y][mde, p][s, poll][poll, p]$



Modelo 4: $[mde, s, y][mde, p][s, poll][poll, p]$



Modelo 5: $[mde, y][mde, p][s, y][s, poll, p]$



De los grafos anteriores podemos notar lo siguiente:

- Sólo el grafo correspondiente al modelo 2 es triangular ("decomposable") pues es acíclico.
- Los grafos correspondientes a los modelos 1, 4 y 5 tiene un ciclo máximo de tamaño 5.
- EL grafo 3 tiene un ciclo de tamaño 4.
- Todos los modelos presentan una relación entre *mde* y *p*; es decir entre las variables `mean_days_to_election` y `partido`.
- Los modelos 1, 2 y 3 tienen relaciónes entre pares de elmentos. Los modelos 4 y 5 tienen relaciones entre pares de elementos y entre tercias de elementos.

Si bien, el modelo con la menor log-verosimilitud posterior es el modelo 1, para resolver las preguntas de interés se utilizará el modelo 2 pues el único modelo que es triangular, además de que la diferencia en log-verosimilitud es tan sólo de 8.97 unidades. De esta manera, el modelo queda definido como:

$$freq \sim mde \times y + mde \times p + s \times poll + poll \times p$$

4. Descripción de los Modelos

La selección del modelo se llevo acabo haciendo una evaluación de distintos grafos probabilísticos bajo el criterio de log-verosimilitud, utilizando los datos de las encuestas. Tomamos los primeros 5 modelos con mayor veoosimilitud y obtenemos estos modelos:

4.1. Modelo 1

$$log(\theta_v) = \alpha^0 + \sum_{a \in A} \alpha^a + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)} + \alpha^{y,s}_{e(y,s)}$$

Aunque, este modelo tiene mayor log-verosimilitud que el modelo 2, es un modelo cíclico, por lo tanto fue descartado.

4.2. Modelo 2

$$log(\theta_v) = \alpha^0 + \sum_{a \in A} \alpha^a + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)}$$

Este es el modelo elegimos ya que es el modelo ya que es el modelo con mayor log-verosimilitud y adicionalmente es acíclico.

4.3. Modelo 3

$$log(\theta_v) = \alpha^0 + \sum_{a \in A} \alpha^a + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)} + \alpha^{y,s}_{e(y,s)} + \alpha^{mde,s}_{e(mde,s)}$$

4.4. Modelo 4

$$log(\theta_v) = \alpha^0 + \sum_{a \in A} \alpha^a + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)} + \alpha^{p,s}_{e(p,s)} + \alpha^{mde,s}_{e(mde,s)}$$

4.5. Modelo 5

$$log(\theta_v) = \alpha^0 + \left(\sum_{a \in A} \alpha^a\right) + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)} + \alpha^{p,s}_{e(p,s)}$$

donde α^0 corresponde al factor común para las 5 dimensiones y $\alpha^{j,k}_{e(j,k)}$ corresponde al factor asociado a la combinación particular $e(j,k)$ de las dimensiones j y k con $j,k \in \{mde, y, p, poll, s\}$ representando las dimensiones `mean_days_to_election`, `year`, `partido`, `pollster_id` y `estado`, tal que $mde \in \{0,1,...,9\}$, $y \in \{0,1,2,3\}$, $p \in \{0,1,2\}$, $s \in \{0,1,...,49\}$ y $poll \in \{0,1,...,39\}$.

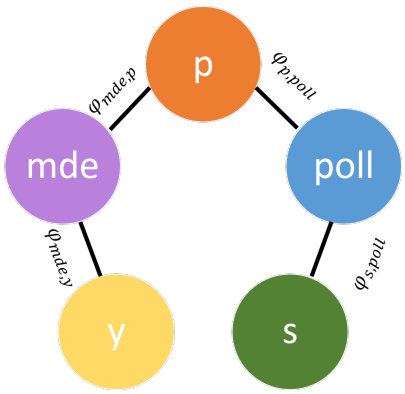
Podemos notar que en los 5 modelos con mayor log-verosimilitud se encuentra las aristas de [year,day],[day,partido],[partido, pollster_id] y [pollster_id, state], es decir, la relación entre estas variables es bastante importante ya que se encuentra en los grafos probabilísticos más probables. En los modelos 1, 3 y 4 el grafo considera la relación [year, state] para estos grafos esto se puede interpretar como que por cada estado el voto por cierto partido sí se puede afectar dependiendo el año. En los modelos 4 y 5 aparece la relación [state, partido] esta relación muestra que hay algunos Estados que tienen una afilación muy fuerte a ciertos partidos.

5. Resultados

En esta sección vamos a profundizar en dar respuesta a las preguntas de interés. Para ello, asumimos que el modelo generador de los datos es el definido en la sección anterior y está dado por

$$freq \sim mde \times y + mde \times p + s \times poll + poll \times p$$

y cuyo grafo asociado es



5.1. Pregunta 1: ¿La intención de voto en polling por estado es independientemente inducida por el tiempo de duración para el día de la elección?

5.1.1. ¿Porqué es importante esta pregunta?

Durante las elecciones es de suma importancia para los candidatos saber cuales los estados que apoyan a su partido, cuales al partido rival o en cuales estados predomina la gente está indecisa. Esta información es de especial relevandia sobre todo en los estados donde hay mayor número de indecisos ya que puede ser utilizada por los candidatos para cambiar sus estrategias de campaña e inveritir más recursos en dichos estados para asegurar su voto.

Ahora bien, es importante saber si la intención de voto es condicionalmente independiente o no al número de días previos a las elecciones en que se realice la encuenta ya que, si resulta que no son condicionalmente independientes, la decisión de inversión variaría dependiendo de los días que falten para las elecciones y podría llegar a no ser tan útil.

5.1.2. Análisis de Resultados del Modelo

Para poder responder esta pregunta, necesitamos es conocer si se cumple que las variables `partido` y `estado` son condicionalmente independientes dados los dias previos a la elección en que se levantó la encuesta. Para ello, y dado que el modelo de referencia es un modelo triangular, necesitamos realizar el análisis de independencia correspondiente.

5.1.2.1. Análisis de independencia

En esta sección aplicamos la Prueba de Independencia Condicional con la función `ciTest` del paquete `gRim` en R. Utilizando la notación definida en la sección anterior, esta prueba nos permite contrastar la hipótesis nula:

$$H_0 : (p \perp s) \mid mde \qquad \text{vs} \qquad H_a : (p \not\perp s) \mid mde$$

La prueba `ciTest` arroja un estadístico de 41.490978 con 708 grados de libertad y un valor-p de `ciTest_preg1$p.value`; por lo tanto, no se puedfe rechazar la hipótesis nula de independencia condicional entre la intención de voto por un partido y el estado dado la ventana de tiempo con que se lleva a cabo la encuesta antes de las elecciones.

Además, podemos analizar la prueba de independencia condicional para cada una de las categorías de la variable `mean_days_to_election` con la finalidad de comprobar si existe una diferencia dependiendo de la ventana de tiempo con que se lleva a cabo la encuesta. La tabla 2 muestra los estadísticos y sus correspondientes valores-p para cada una de las 10 posibles ventanas de tiempo.

Tabla 2: Prueba de Independencia Condicional Para la Pregunta 1

mde	Valor	Est.	Grados Lib.	Valor-p
0	0-4 días	1.08	98.00	1.00
1	5-9 días	28.23	98.00	1.00
2	10-14 días	3.22	98.00	1.00
3	15-19 días	2.72	98.00	1.00
4	20-24 días	3.92	56.00	1.00
5	25-29días	2.31	58.00	1.00
6	30-34 días	0.00	46.00	1.00
7	35-39 días	0.00	78.00	1.00
8	40-44 días	0.00	78.00	1.00
9	45+ días	0.00	0.00	1.00

Se puede observar que, para **todas** las ventanas de tiempo previas a las elecciones, no se puede rechazar la hipótesis nula (todos los valores-p son mayores a 0.05); por lo que concluimos que la intención de voto por el partido es independiente del estado condicional a cualquier ventana de tiempo antes de la elección.

5.2. Pregunta 4: ¿Existe una asociación a nivel estatal del resultado actual con los resultados anticipados agregados de los polling?

5.2.1. ¿Porqué es importante esta pregunta?

Por lo general, durante las campañas electorales en los distintos países se practican encuestas para intentar predecir quién será el ganador de la elecciones. Este tema es de especial relevancia pues puede influir, por ejemplo, en los distintos indicadores financieros dependiendo del riesgo que uno u otro ganador represente para el país, para los inversionistas y para el diseño de las políticas públicas. Sería de gran utilidad poder analizar si a partir de la información de las encuestas se puede predecir, de manera correcta, quien es el ganador de la elección aplicando un modelo loglineal a un grafo probabilístico o bien, si los votos indecisos causan efectos mixtos.

5.2.2. Análisis de Resultados del Modelo

Utilizando el modelo seleccionado, y las varaibles categorizadas, queremos estimar el modelo loglineal para obtener los valores estimados de θ_v de manera aditiva como sigue:

$$log(\theta_v) = \alpha^0 + \left(+ \sum_{a \in A} \alpha^a \right) + \alpha^{mde,y}_{e(mde,y)} + \alpha^{mde,p}_{e(mde,p)} + \alpha^{s,poll}_{e(s,poll)} + \alpha^{poll,p}_{e(poll,p)}$$

donde α^0 corresponde al factor común para las 5 dimensiones y $\alpha^{j,k}_{e(j,k)}$ corresponde al factor asociado a la combinación particular $e(j,k)$ de las dimensiones j y k con $j,k \in \{mde, y, p, poll, s\}$ representando las dimensiones `mean_days_to_election`, `year`, `partido`, `pollster_id` y `estado`, tal que $mde \in \{0,1,...,9\}$, $y \in \{0,1,2,3\}$, $p \in \{0,1,2\}$, $s \in \{0,1,...,49\}$ y $poll \in \{0,1,...,39\}$.

Dado que el modelo generador elegido es triangular, utilizamos los datos obtenidos en `polling` y la función `findPostMean` del paquete `bayesloglin` para calcular la media posterior de los coeficientes del modelo (α's). Se obtuvieron un total de 2,137 medias posteriores que corresponden a: * un intercepto,

- 102 posibles categorías para el coeficiente α_0 (0 de mde, 3 de y, 40 de s, 2 de p y 39 de $poll$),
- 27 posibles combinaciones de mde con y,
- 18 posibles combinaciones de mde con p,
- 1911 posibles combinacions de s con $poll$ y
- 78 posibles combinaciones de $poll$ con p.

Cabe mencionar que la tabla de medias posteriores no se reporta por su longiutd. Una vez calculados los valores de $\alpha_{i,j}$, podemos calcular las frecuencias estimadas $\hat{\theta}_v$ a partir de

$$\hat{p}(x_{mde,y,p,s,poll}) = \hat{\theta}_{mde,y,p,s,poll} = \frac{\hat{\theta}_{mde,y,p,s,\cdot}\hat{\theta}_{\cdot,y,p,s,poll}}{\hat{\theta}_{\cdot,y,p,s,\cdot}}$$

con $mde \in \{0,1,...,9\}$, $y \in \{0,1,2,3\}$, $p \in \{0,1,2\}$, $s \in \{0,1,...,49\}$ y $poll \in \{0,1,...,39\}$. La función utilizada para calcular estos estimadores se muestra en el apéndice A.

Como el grafo generador no es un DAG, la suma de las $\hat{\theta}_v$ no es 1. Por lo tanto, dividimos estos valores estimados entre su suma para convertirlo en probabilidades. Finalmente, una vez que se tiene el valor ajustado de Ya que tenemos las $\theta_{mde,y,s,p,poll}$ marginalizamos para obtener el valor de $\hat{\theta}_{y,s,p}$. De esta manera se obtiene una tabla de contingencias estimada con 600 entradas correspondientes a las combinaciones de los 4 años, para los 50 estados y los 3 partidos (Demócratas, Republicanos e indecisos). Nuevamente, la tabla completa de frecuencias estimada no se reporta por sus dimensiones. Sin embargo, la figura 6 muestra la distribución de los valores estimados para cada partido en cada estado.

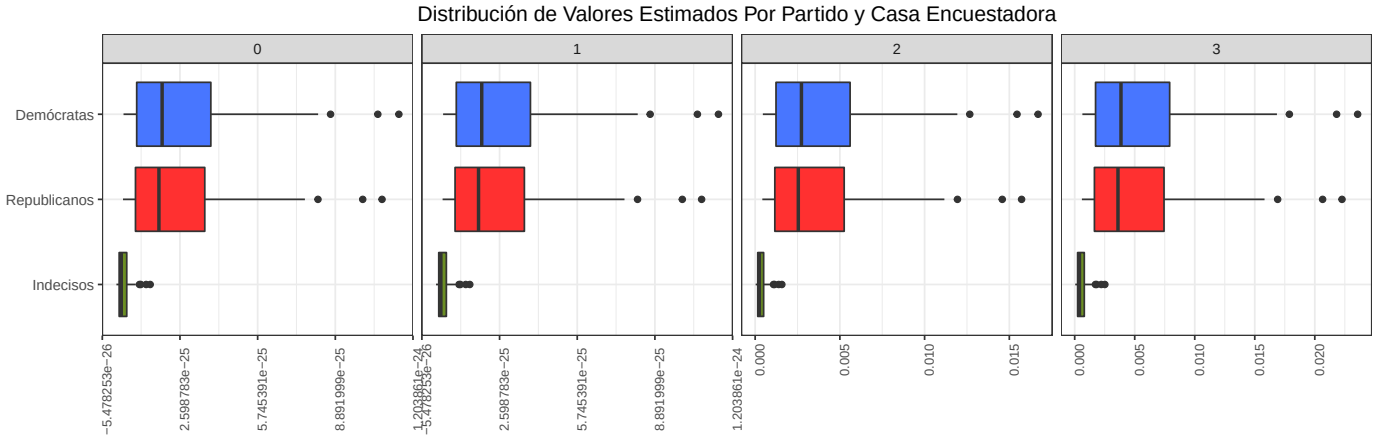


Figura 6: Distribución de Thetas Estimadas por Año Electoral.

Otra forma de ver los datos es por porcentaje de votos por Estado. Obtenemos el porcentaje de votos asumiendo que los votos Undecided se distribuyen proporcional a la distribución por estado y extraemos el partido ganador en cada estado.

Finalmente, tomamos los datos reales de las elecciones, los ajustamos igual que los datos del polling y comparamos. La figura 7 muestra una comparación entre las estimaciones y los valores ajustados.

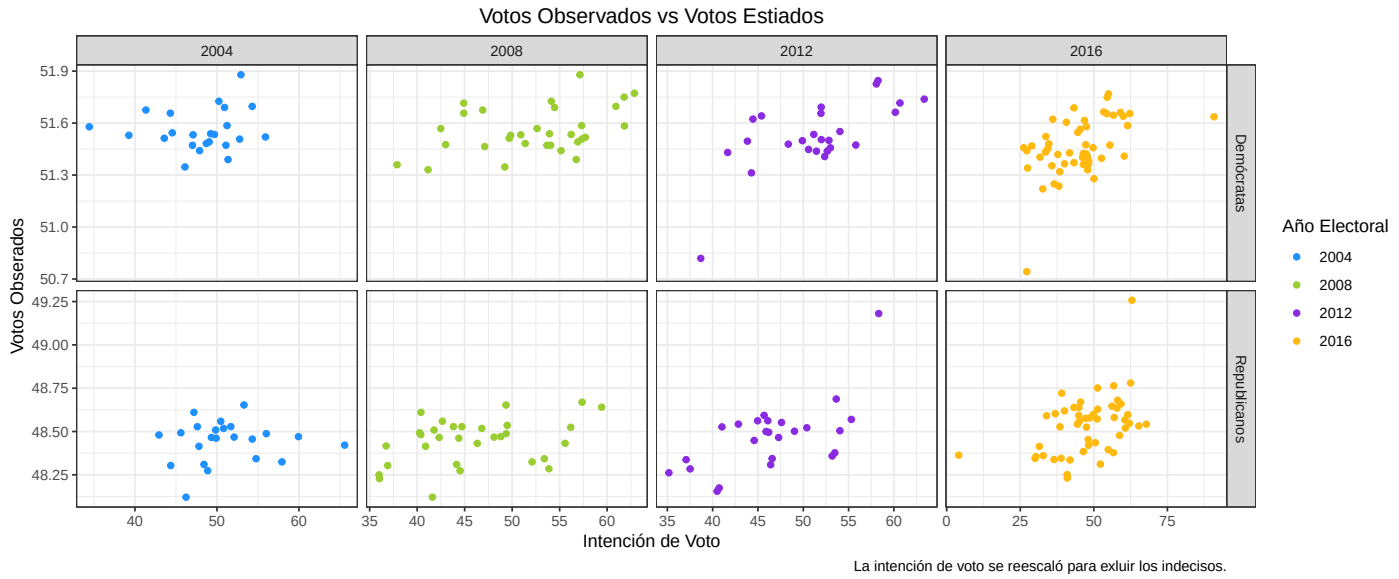


Figura 7: Comparación de Votos Estimados y Observados.

Haciendo el cálculo, le atinamos al 55 % de las observaciones. Las estimaciones que generamos están en un rango muy pequeño, lo cual quiere decir que el modelo no tiene información suficiente de qué determina que una votación vaya muy arriba o muy abajo de la media. Al correr un modelo con menos covariables, resulta que la estimación es muy precisa. Por lo tanto, creemos que eso puede ser parte del problema.

Nuestra conclusión al respecto es que agregar covariables puede arrojar más luz sobre relaciones complejas entre ésas, pero a costa de generar mejores estimaciones. Por otro lado, reducir covariables dice menos de la relación entre las variables, pero mejora las estimaciones.

6. Consideraciones Finales

En este trabajo exploramos la estructura subyaciente de los datos de encuestas en EEUU en distintos años. Las predicciones fueron muy similares a los resultados finales de las votaciones. Esto implica que conocer la dinámica entre ambos puede ser de ayuda en planeación y política pública.

Con base en los modelos podemos concluir que, en general, la relación que tiene un estado con un partido específico puede ser totalmente intermedidada por las casas encuestadoras. Esto significa que si el modelo que tomamos como modelo general es cierto, dada la casa encuestadora, el estado es independiente del partido. Esto implica si alguna casa encuestadora trabaja en varios estados, las posturas de esos estados deben ser similares. De otra forma, no se cumpliría la independencia. Este aprendizaje sugiere que fue un acierto tomar datos de muchas casas encuestadoras, pues quedarse con una sola (o pocas) traduciría sesgo a los resultados.

Con base en el éxito de las predicciones con modelos más simples, al comparar con los resultados de las votaciones, aprendemos que existe una disyuntiva entre desempeño predictivo e inferencia de la estructura subyaciente de los datos. Por un lado tenemos un modelo que sólo tomó en cuenta tres variables (año, estado y partido) y otro que toma en cuenta ésas más: días a la elección y casa encuestadora. Las dos últimas pueden arrojar información muy interesante para: entender cómo cambian las preferencias de los votantes conforme se acerca la elección, y para entender la relación que tienen las casas encuestadoras con los estados, y los posibles sesgos que puedan inducir. El primer modelo fue computacionalmente más rápido de correr, y mucho más preciso en términos del éxito de las predicciones (87 % vs. 55 %). El segundo modelo dio información interesante acerca de la posible independencia condicional de los estados con respecto a su postura política dada la casa encuestadora. Sin embargo, dado que las predicciones del primer modelo fueron tan excepcionalmente malas, estos resultados debemos tomarlos con escepticismo pues posiblemente hay demasiadas variables, con demasidos valores únicos, como para poder modelar correctamente la realidad con los datos que tenemos.

Concluímos que el análisis loglineal de Bayes puede ofrecer información interesante cuando hay datos suficientes. Que no siempre más variables es mejor. Finalmente que el propósito del análisis debe ser considerado para decidir qué hacer. Por ejemplo, si el análisis de hará en un cuartel de un partido político para poder decidir en qué estados invierten más recursos, vale la pena priorizar precisión de las estimaciones. Si el propósito es crear legislación regulatoria a casas encuestadoras, es de mayor interés entender la relación entre cada variable, que generar estimaciones precisas.

7. Fuentes de Consulta

- Bon, J. J., Ballard, T., & Baffour, B. (2019). *Polling bias and undecided voter allocations: US presidential elections, 2004–2016.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(2), 467-493.
- Martínez, J.C. (2019). *Sesión 22 Datos Categóricos y Modelo de Grafos Probabilísticos Parte 5/7* [Documento html]. Recuperado del repositorio privado https://drive.google.com/drive/folders/1RDVvjAaS77XbQHLrJvfjN95olZRylQmq?usp=sharing_eil&ts=5c3e47d9
- Martínez, J.C. (2019). *Sesión 23 Datos Categóricos y Modelo de Grafos Probabilísticos Parte 6/7* [Documento html]. Recuperado del repositorio privado https://drive.google.com/drive/folders/1RDVvjAaS77XbQHLrJvfjN95olZRylQmq?usp=sharing_eil&ts=5c3e47d9
- Martínez, J.C. (2019). *Sesión 24 Datos Categóricos y Modelo de Grafos Probabilísticos Parte 7/7* [Documento html]. Recuperado del repositorio privado https://drive.google.com/drive/folders/1RDVvjAaS77XbQHLrJvfjN95olZRylQmq?usp=sharing_eil&ts=5c3e47d9

A. Apéndice

A.1. Análsis Exploratorio de los Datos

```
# Instalar paquetes en caso necesario
if(!requireNamespace("tidyverse")){install.packages("tidyverse")}

# Cargar paquetes
suppressWarnings(suppressMessages(library(tidyverse)))
```

```
# Cargar los datos
data_pollster<-read.csv("../Data/us-pres-state-polling-2004-2016.csv")
data_voting<-read.csv("../Data/us-pres-state-voting-2004-2016.csv")
```

Grafica 1: ¿Cómo se Distribuye la Intención de Voto por Partido y por Año electoral?

```
aux <- data_pollster %>%
  gather("Partido","Int_Voto",Dem_poll:Undecided)

labs_partido <- c("Demócratas", "Republicanos", "Indecisos")
names(labs_partido) <- c("Dem_poll", "Rep_poll", "Undecided")

ggplot(aux,aes(x=Int_Voto,fill=Partido))+theme_bw()+
  geom_histogram(aes(y=..density..),position="identity")+
  geom_density(alpha=0.2)+
  facet_grid(Partido~as.factor(year),labeller = labeller(Partido = labs_partido))+
  ggtitle('Distribución de la Intención de Voto por Año y por Partido')+
  scale_x_continuous(breaks=seq(0,100,10))+
  scale_fill_manual(values=c("royalblue1","firebrick1","olivedrab"))+
  scale_color_manual(values=c("royalblue1","firebrick1","olivedrab"))+
  xlab('Intención de Voto')+
  ylab('Partido')+
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'none')
```

Grafica 2: ¿Cómo se Distribuye la Intención de Voto por Partido en Cada Estado Para Cada Año electoral?

```
aux <- data_pollster %>%
  gather("Partido","Int_Voto",Dem_poll:Undecided)

labs_partido <- c("Demócratas", "Republicanos", "Indecisos")
names(labs_partido) <- c("Dem_poll", "Rep_poll", "Undecided")

ggplot(aux,aes(x=as.factor(state),y=Int_Voto,fill=as.factor(year)))+theme_bw()+
  geom_bar(stat = "summary", fun.y = "mean",position = 'stack')+
  facet_grid(.~Partido,labeller = labeller(Partido = labs_partido))+
  ggtitle('Intención de Voto Promedio por Año y por Estado')+
  scale_fill_manual(values=c("dodgerblue1","yellowgreen","blueviolet","darkgoldenrod1"))+
  scale_color_manual(values=c("dodgerblue1","yellowgreen","blueviolet","darkgoldenrod1"))+
  scale_y_continuous(breaks=seq(0,300,10))+
  xlab('Estado')+
  ylab('Intención de Voto')+
  labs(fill='Año Electoral')+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90,size=8),
        legend.position = 'bottom')+
  coord_flip()
```

Grafica 3: ¿Cómo se Distribuye la Intención de Voto por partido en las Casas Encuestadoras?

```
aux <- data_pollster %>%
   gather("Partido","Int_Voto",Dem_poll:Undecided)%>%
  mutate(Partido=recode(Partido,
                        Dem_poll = "Demócratas",
                        Rep_poll = "Republicanos",
                        Undecided = "Indecisos"))

aux$Partido<-factor(as.factor(aux$Partido), levels = c("Indecisos","Republicanos","Demócratas"))

ggplot(aux,aes(x=Partido,y=Int_Voto,fill=Partido))+theme_bw()+
  geom_boxplot()+
  facet_wrap(as.factor(aux$pollster_id),nrow=8)+
  ggtitle('Intención de Voto Por Partido y Casa Encuestadora')+
  scale_fill_manual(values=c("olivedrab","firebrick1","royalblue1"))+
  scale_color_manual(values=c("olivedrab","firebrick1","royalblue1"))+
  scale_y_continuous(breaks=seq(0,300,10))+
  xlab('')+
  ylab('Intención de Voto')+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90,size=8),
        legend.position = 'none')+
  coord_flip()
```

Grarfica 4: ¿Hay diferencia entre la intención de votos y los votos observados?

```
aux1<-data_pollster %>%
  mutate(Dem_poll_aj=Dem_poll/(Dem_poll+Rep_poll)*100,
         Rep_poll_aj=Rep_poll/(Dem_poll+Rep_poll)*100)%>%
  gather("Partido","Int_Voto",Dem_poll_aj:Rep_poll_aj)%>%
  mutate(Partido=recode(Partido,
                            Dem_poll_aj = "Demócratas",
                            Rep_poll_aj = "Republicanos"))

aux2<-data_voting %>%
  select(state,year,Dem_vote, Rep_vote) %>%
  gather("Partido","Votos",Dem_vote:Rep_vote)%>%
  mutate(Partido=recode(Partido,
                           Dem_vote = "Demócratas",
                           Rep_vote = "Republicanos"))

aux1<-aux1 %>%
  left_join(aux2,by=c("year","state","Partido"))%>%
  select(year,state,Partido,Int_Voto,Votos)

ggplot(aux1,aes(x=Int_Voto,y=Votos,color=as.factor(year)))+theme_bw()+
  geom_jitter()+
  geom_abline(slope=1,intercept=0,color='gray70')+
  facet_grid(Partido~year)+
  ggtitle('Intención de Voto vs Votos Observados')+
  scale_color_manual(values=c("dodgerblue1","yellowgreen","blueviolet","darkgoldenrod1"))+
  scale_x_continuous(breaks=seq(0,100,10))+
  scale_y_continuous(breaks=seq(0,100,10))+
  xlab('Intención de Voto')+
  ylab('Votos Obserados')+
  labs(color='Año Electoral',
       caption="La intención de voto se reescaló para exluir los indecisos.")+
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'right')
```

A.2. Estimación de Modelos

```
# Descarga paquetes en caso necesario
if(!requireNamespace("BiocManager", quietly = TRUE)){install.packages("BiocManager")}
library(BiocManager)
BiocManager::install("graph", version = "3.8")
BiocManager::install("RBGL", version = "3.8")
BiocManager::install("Rgraphviz", version = "3.8")
if(!requireNamespace("gRbase", quietly = TRUE)){install.packages("gRbase", dependencies=TRUE)}
if(!requireNamespace("gRain", quietly = TRUE)){install.packages("gRain", dependencies=TRUE)}
if(!requireNamespace("gRim", quietly = TRUE)){install.packages("gRim", dependencies=TRUE)}
if(!requireNamespace("bayesloglin", quietly = TRUE)){install.packages("bayesloglin", dependencies=TRUE)}
if(!requireNamespace("xtable", quietly = TRUE)){install.packages("xtable", dependencies=TRUE)}

# Carga librerías
suppressWarnings(suppressMessages(library(graph)))
suppressWarnings(suppressMessages(library(RBGL)))
suppressWarnings(suppressMessages(library(Rgraphviz)))
suppressWarnings(suppressMessages(library(gRbase)))
suppressWarnings(suppressMessages(library(gRain)))
suppressWarnings(suppressMessages(library(gRim)))
suppressWarnings(suppressMessages(library(bayesloglin)))
suppressWarnings(suppressMessages(library(xtable)))
```

Selección y transformación de los datos

```
# Seleccionamos las variables
datos_agrupados <- data_pollster %>%
  dplyr::select(Dem_poll,Rep_poll,Undecided,sample_size,
         mean_days_to_election,state,year,pollster_id)


# Volvemos la variable mean_days_to_election categorica
datos_agrupados<-datos_agrupados %>%
  mutate(mean_days_to_election =cut(mean_days_to_election,
                                    breaks=seq(0,50,5),
                                    labels=FALSE) %>% as.numeric()-1)

# Sacamos el total de intención de votos por partido
datos_agrupados$Dem_poll <- datos_agrupados$Dem_poll/100*datos_agrupados$sample_size
datos_agrupados$Rep_poll <- datos_agrupados$Rep_poll/100*datos_agrupados$sample_size
datos_agrupados$Undecided <- datos_agrupados$Undecided/100*datos_agrupados$sample_size

# Agrupa las intenciones de voto por partido
datos_agrupados <- datos_agrupados %>%
  gather(key="partido","int_voto",Dem_poll:Undecided) %>%
  select(-sample_size)

# Trunca decimales en frecuencia
datos_agrupados$int_voto <- as.integer(round(datos_agrupados$int_voto),digits=0)

# Convierte las variables a factor (iniciando en 0)
datos_agrupados$year <- as.factor(datos_agrupados$year) %>%
  as.numeric() %>% as.integer()-1
datos_agrupados$pollster_id <- as.factor(datos_agrupados$pollster_id) %>%
  as.numeric()  %>% as.integer()-1
datos_agrupados$partido <- as.factor(datos_agrupados$partido) %>%
  as.numeric()  %>% as.integer()-1
datos_agrupados$mean_days_to_election <- as.factor(datos_agrupados$mean_days_to_election) %>%
```

```
  as.numeric()  %>% as.integer()-1
datos_agrupados$state <- as.factor(datos_agrupados$state) %>%
  as.numeric()  %>% as.integer()-1

# Quitamos observaciones tienen intención de voto cero o NA (quita 188 observaciones)
datos_agrupados <-datos_agrupados %>%
  filter(int_voto!=0) %>%
  filter(!is.na(int_voto)) %>%
  as.data.frame()
```

Estimación de Modelos

```
# Estimamos modelos
# Buscamos los generadores posibles para los datos
model_graf <- MC3(init = NULL,
                  alpha = 1,
                  iterations =100000 ,
                  replicates = 1,
                  data = datos_agrupados ,
                  mode = "Graphical")
```

Gráfica de log verosimilitudes

```
# Graficamos la loguerosimilitud de los modelos
# Hacemos un plot de la log-verosimilitud
aux<- modelo_graf %>%
  mutate(Id=1:nrow(modelo_graf))

ggplot(aux, aes(x=Id,y=logPostProb)) + theme_bw() +
  geom_point() +
  ggtitle('Log-verosimilitud Posterior de los Posibles Modelos Log-Lineales') +
  scale_x_continuous(breaks=1:nrow(modelo_graf),
                     labels=paste("Modelo ", 1:nrow(modelo_graf))) +
  scale_y_continuous(breaks=seq(7931000,7933500,250))+
  xlab("Modelo") +
  ylab("Log-Verosimilitud Posterior") +
  theme(plot.title = element_text(hjust=0.5),
        axis.text.x = element_text(angle=90))
```

Mejores modelos

```
# Vemos los primeros modelos
head(model_graf,5)
```

A.3. Resultados

A.3.1. Pregunta 1: ¿La intención de voto en polling por estado es independientemente inducida por el tiempo de duración para el día de la elección?

```
# Convertimos los datos a factor
datos_agrupados$mean_days_to_election<-datos_agrupados$mean_days_to_election%>% as.factor()
datos_agrupados$state<-datos_agrupados$state%>% as.factor()
datos_agrupados$partido<-datos_agrupados$partido%>% as.factor()

# Aplicamos ciTest para ver la independencia condicional
ciTest_p1<-ciTest(datos_agrupados, set=c("partido","state","mean_days_to_election"))

ciTest_p1
```

A.3.2. Pregunta 4: ¿Existe una asociación a nivel estatal del resultado actual con los resultados anticipados agregados de los polling?

```
# Extraemos las medias posteriores de la distribución
formula <- freq ~ mean_days_to_election*year + mean_days_to_election*partido + state*pollster_id + pollster_i

s <- findPostMean (formula, alpha = 1, data = datos_agrupados)
s
```

```
conseguir_teta <- function(s, datos_agrupados) {
total_filas=length(unique(datos_agrupados$mean_days_to_election))*length(unique(datos_agrupados$state))*lengt
 datos_cont<-data_frame(mean_days_to_election=rep(0,total_filas),
                        state=rep(0,total_filas),
                        year=rep(0,total_filas),
                        pollster_id=rep(0,total_filas),
                        partido=rep(0,total_filas),
                        theta=rep(0,total_filas))
 i=1
 for(md in unique(datos_agrupados$mean_days_to_election)){
    for(st in unique(datos_agrupados$state)){
    for(ja in unique(datos_agrupados$year)){
        for(po in unique(datos_agrupados$pollster_id)){
          for(pa in unique(datos_agrupados$partido)){

            th=s["(Intercept)"]

            sum1 = s[paste("mean_days_to_election",as.character(md),":year",as.character(ja), sep="")]

            sum2 = s[paste("mean_days_to_election",as.character(md),":partido",as.character(pa), sep="")]

            sum3 = s[paste("pollster_id",as.character(po),":state",as.character(st), sep="")]

            sum4 = s[paste("partido",as.character(pa),":pollster_id",as.character(po), sep="")]

            sum5 = s[paste("mean_days_to_election",as.character(md),sep="")]

            sum6 = s[paste("year",as.character(ja),sep="")]

            sum7 = s[paste("partido",as.character(pa),sep="")]

            sum8 = s[paste("pollster_id",as.character(po),sep="")]

            sum9 = s[paste("state",as.character(st),sep="")]


            for(p in c(sum1, sum2, sum3, sum4, sum5, sum6, sum7, sum8, sum9)){
              if(!is.na(p)){
              th=th+p
              }
            }
            th=as.numeric(exp(th))
            datos_cont$mean_days_to_election[i]=md
            datos_cont$state[i]=st
            datos_cont$year[i]=ja
            datos_cont$pollster_id[i]=po
            datos_cont$partido[i]=pa
            datos_cont$theta[i]=th
            i=i+1
          }
```

```
        }
      }
    }
  }
return(datos_cont)
}

# Estima las tetas
tabla_contingencias <- conseguir_teta( s , datos_agrupados1 )

# Ajusta para que sumen 1
tabla_contingencias$true_theta <- tabla_contingencias$theta / sum(tabla_contingencias$theta)
```

```
# Thetas marginalizando
tabla_theta <- tabla_contingencias %>% group_by(
  year,state,partido
) %>% summarise(
  theta=sum(true_theta)
)
tabla_theta$partido[tabla_theta$partido==0] <- "Dem_vote"
tabla_theta$partido[tabla_theta$partido==1] <- "Rep_vote"
tabla_theta$partido[tabla_theta$partido==2] <- "Undecided"

# Vemos los primeros valores
head(tabla_theta)


# Grarfica de los valores estimados
aux <- tabla_theta %>%
  mutate(partido=recode(partido,
                        Dem_vote = "Demócratas",
                        Rep_vote = "Republicanos",
                        Undecided = "Indecisos"))

aux$partido<-factor(as.factor(aux$partido), levels = c("Indecisos","Republicanos","Demócratas"))

labs_year <- c("2004", "2008", "2012","2016")
names(labs_year) <- as.character(0:3)


ggplot(aux,aes(x=partido,y=theta,fill=partido))+theme_bw()+
  geom_boxplot()+
  facet_wrap(as.factor(aux$year),nrow=1,scales="free_x",labeller = labeller(year = labs_year))+
  ggtitle('Distribución de Valores Estimados Por Partido y Casa Encuestadora')+
  scale_fill_manual(values=c("olivedrab","firebrick1","royalblue1"))+
  scale_color_manual(values=c("olivedrab","firebrick1","royalblue1"))+
  xlab('')+
  ylab('')+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90,size=8),
        legend.position = 'none')+
  coord_flip()


tabla_por_estado <- tabla_theta %>% spread(partido, theta)
tabla_por_estado <- tabla_por_estado %>% mutate(por_tot=sum(Dem_vote,Rep_vote,Undecided))
tabla_por_estado$Dem_vote <- tabla_por_estado$Dem_vote/tabla_por_estado$por_tot
tabla_por_estado$Rep_vote<- tabla_por_estado$Rep_vote/tabla_por_estado$por_tot
tabla_por_estado$Undecided <- tabla_por_estado$Undecided/tabla_por_estado$por_tot
```

```r
tabla_por_estado <- tabla_por_estado %>%  mutate(
  Dem_vote_adj=Dem_vote/(1-Undecided)*100,
  Rep_vote_adj=Rep_vote/(1-Undecided)*100,
  Ganador_poll="Rep"
) %>% select(-c(Undecided,Rep_vote,Dem_vote,por_tot))


for(i in 1:dim(tabla_por_estado)[1]){
  if(!is.na(tabla_por_estado$Dem_vote_adj[i])){
  if(tabla_por_estado$Dem_vote_adj[i]>tabla_por_estado$Rep_vote_adj[i]){
    tabla_por_estado$Ganador_poll[i] <- "Dem"
  }
  }
}

votos <- data_voting %>%
  select(year_id,state,Dem_vote,Rep_vote)

votos$year_id <- votos$year_id-1

votos$state <- as.factor(votos$state) %>% as.numeric()  %>% as.integer() -1
votos <- votos %>% mutate( Ganador="Rep")

for(i in 1:dim(votos)[1]){
  if(votos$Dem_vote[i]>votos$Rep_vote[i]){
    votos$Ganador[i] <- "Dem"
  }

}


comparacion <- inner_join(votos,tabla_por_estado,by=c("year_id"="year","state"))

comparacion$atinamos <- 0
for(i in 1:dim(comparacion)[1]){
  if(comparacion$Ganador_poll[i]==comparacion$Ganador[i]){
    comparacion$atinamos[i] <- 1
  }

}


aux1<-comparacion %>%
  select(year_id,state,Dem_vote_adj,Rep_vote_adj)%>%
  gather("partido","Estimado",Dem_vote_adj:Rep_vote_adj)%>%
  mutate(partido=recode(partido,
                        Dem_vote_adj = "Demócratas",
                        Rep_vote_adj = "Republicanos"))

aux2<-comparacion %>%
  select(year_id,state,Dem_vote,Rep_vote)%>%
  gather("partido","Observado",Dem_vote:Rep_vote)%>%
  mutate(partido=recode(partido,
                        Dem_vote = "Demócratas",
                        Rep_vote = "Republicanos"))

aux<-aux1 %>%
  left_join(aux2,by=c("year_id","state","partido"))%>%
```

```
  select(year_id,state,partido,Observado,Estimado)%>%
  mutate(year=recode(year_id,
                     "0"="2004",
                     "1"="2008",
                     "2"="2012",
                     "3"="2016"))

ggplot(aux,aes(x=Observado,y=Estimado,color=as.factor(year)))+theme_bw()+
  geom_jitter()+
  facet_grid(partido~year,scales="free")+
  ggtitle('Votos Observados vs Votos Estiados')+
  scale_color_manual(values=c("dodgerblue1","yellowgreen","blueviolet","darkgoldenrod1"))+
  xlab('Intención de Voto')+
  ylab('Votos Obserados')+
  labs(color='Año Electoral',
       caption="La intención de voto se reescaló para exluir los indecisos.")+
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'right')
```