Modelos estadísticos para sistemas electorales multipartidistas en Stata

Javier Márquez y Javier Aparicio

October 12, 2011

1 Introducción

¿Cuántos votos se requieren para conseguir mayoría en el Congreso? ¿Qué efecto tiene el gasto de campaña en los votos obtenidos por los partidos políticos? ¿Cuál es el impacto de las elecciones concurrentes en los resultados electorales y en la distribución de asientos en el Congreso? ¿Cuál es la probabilidad de que un candidato no experimentado gane una elección en un distrito reñido? Estas y otras muchas preguntas relacionadas con los sistemas electorales pueden analizarse a través de modelos estadísticos (Gelman & King, 1994). Sin embargo, los modelos estadísticos convencionales requieren algunas adecuaciones para adaptarse a las características más importantes de los regímenes políticos en América Latina.

El entramado institucional de los países de la región incluye tres elementos principales (Golder, 2005). En primer lugar, se tratan de regímenes presidenciales con separación de poderes entre el ejecutivo y el legislativo. En segundo lugar, la mayoría de esos países eligen a sus legisladores a través de sistemas electorales de representación proporcional o mixtos. Finalmente, el sistema de partidos tiende a ser fragmentado o multipartidista; es decir, existen más de dos partidos políticos relevantes.

Estas características tienen implicaciones importantes para los análisis estadísticos. Por una parte, la mayoría de los trabajos que intentan explicar o predecir la fuerza electoral de los partidos políticos utilizan técnicas que son adecuadas para sistemas bipartidistas –posblemente por la influencia de otros estudios que concentran su atención al caso de Estados Unidos. Sin embargo, este enfoque suele proporcionar resultados lógicamente inconsistentes cuando se utilizan en sistemas multipartidistas. Por otro lado, la asignación de asientos en la asamblea suele ser más compleja que la de un sistema mayoritario, en el que el candidato con más votos gana el asiento. Los sistemas proporcionales y mixtos utilizan fórmulas complicadas (por ejemplo, la fórmula de D'Hondt), y comúnmente establecen umbrales de votación para que los partidos tengan derecho al reparto de asientos o curules. Quizá por esta razón, las investigaciones sobre los sistemas electorales de América Latina tienden a menospreciar la utilidad de los modelos estadísticos para explicar o predecir la conformación de la asamblea –a diferencia de lo que ocurre en la literatura sobre Estados Unidos (véase por ejemplo, Gelman & King, 1994).

Nuestro propósito en este capítulo es presentar de manera práctica la implementación de una nueva versión (2.0) del módulo **camaradip** de Stata (Marquez & Aparicio, 2010), un paquete desarrollado por los autores que incorpora las técnicas de Katz & King (1999); Honaker et al. (2002); Tomz et al. (2002) para adecuar los modelos estadísticos convencionales al análisis de sistemas electorales multipartidistas. En este sentido, **camaradip** es una herramienta útil para analizar varias preguntas relacionadas con los sistemas electorales de un número considerable de países de América Latina. Puede decirse que **camaradip** es la contraparte del software *JudgeIt* (Gelman et al., 2007), que permite evaluar una gran variedad de características de los sistemas mayoritarios y bipartidistas Gelman & King (1994).

Este capítulo tiene como objetivo mostrar paso a paso los componentes de un modelo estadístico para sistemas de partidos multipartidistas y con sistemas electorales mixtos o proporcionales. La estructura del capítulo es como sigue. En la siguiente sección abordamos los problemas que surgen cuando se especifican modelos estadísticos para datos multipartidistas como si se trataran de datos bipartidistas, y presentamos algunas soluciones a estos problemas. La Sección 3 explica algunas técnicas para traducir estos modelos en información sustantiva relacionada con el proceso político, como por ejemplo, la relación entre elecciones legislativas y presidenciales. En la cuarta sección introducimos la técnica de simulación estadística para dotar a los análisis con medidas de incertidumbre tales como errores estándar e intervalos de confianza. Por último, en la quinta sección presentamos algunos comentarios finales y sugerimos diversas extensiones del modelo aquí analizado.

2 Estimación de modelos estadísticos con datos electorales multipartidistas

2.1 Mínimos Cuadrados Ordinarios

En ocasiones, los científicos sociales están interesados en explicar o predecir la votación de los partidos políticos de algún país. La técnica de regresión con Mínimos Cuadrados Ordinarios (MCO) es una de las herramientas más empleadas para analizar datos electorales de sistemas bipartidistas–como el de Estados Unidos. Los datos electorales bipartidistas pueden analizarse con una sola ecuación de regresión en la que la variable dependiente V_1 es la proporción de votos de uno de los partidos políticos, puesto que la proporción de votos del otro partido es simplemente $V_2 = 1 - V_1$.[1]

Algunos académicos utilizan dos variantes de este enfoque para analizar datos electorales multipartidistas. El primero consiste en plantear un modelo de regresión que contraste la proporción de votos de un partido de interés frente al de los demás, "bipartidizando" artificialmente el sistema de partidos. El segundo emplea el porcentaje

[1] En ocasiones, se aplica la transformación logística a la variable dependiente para que los resultados se encuentren en el intervalo de la unidad, de modo que $Y_1 = \ln\left[\frac{V_1}{1-V_1}\right]$ (véase por ejemplo Gelman & King, 1990). Sin embargo, en una investigación posterior, Gelman & King (1994) evitan esta complicación dado que la experiencia muestra que los resultados con y sin tranformación son muy similares.

de votación de cada partido político como una variable dependiente distinta, de modo que se plantean tantas ecuaciones de regresión como partidos existen en el sistema.

Por ejemplo, consideremos el fenómeno denominado "efecto de arrastre", que se define como la capacidad de algunos candidatos populares para atraer votos a otros candidatos de su mismo partido (Ferejohn & Calvert, 1984). En particular, analizaremos el efecto de arrastre que ejercieron los candidatos presidenciales en la votación de sus respectivos partidos en la elección de 2006 para la Cámara de Diputados en México. La base de datos `arrastre.dta` que se distribuye con el paquete **camaradip** contiene las siguientes variables para cada uno de los 300 distritos de mayoría relativa del país:

- la proporción de votos que obtuvo cada uno de los cinco partidos políticos o coaliciones para la elección legislativa (`pan`, `apm`, `pbt`, `asdc` y `otr`) por distrito,[2]
- la proporción de votos que obtuvieron los candidatos presidenciales de esos mismos partidos o coaliciones (`pres_pan`, `pres_apm`, `pres_pbt`, `pres_asdc`, `pres_otr`), y
- la proporción de votos que esos mismos partidos obtuvieron en la elección legislativa más reciente (2003) con el mismo nivel de agregación (`lag_pan`, `lag_apm`, `lag_pbt`, `lag_asdc`, `lag_otr`).[3]

Comencemos abriendo la base de datos en Stata y, siguiendo una de las variantes del modelo bipartidista, estimaremos cinco ecuaciones de regresión en las que el voto para la Cámara de Diputados está en función del voto para los candidatos presidenciales de ese mismo año. Dado que utilizaremos este ejemplo con fines meramente ilustrativos, mantendremos la especificación del modelo lo más sencilla posible introduciendo únicamente como variable de control el porcentaje de votos que cada partido obtuvo en la elección legislativa más reciente. La intención de incluir las variables rezagadas es controlar por la inercia o fuerza electoral del voto por los partidos. Así, para el caso del PAN, el modelo se puede especificar en Stata de la siguiente manera:

```
. use arrastre, clear
. reg pan pres_pan lag_pan
      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  2,   297) = 3641.53
       Model |  5.2221239      2  2.61106195            Prob > F      =  0.0000
    Residual |   .2129559    297  .000717023           R-squared     =  0.9608
-------------+------------------------------           Adj R-squared =  0.9606
       Total |  5.4350798    299  .018177524           Root MSE      =  .02678

------------------------------------------------------------------------------
         pan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

[2] La coalición Alianza por México (APM) integró al Partido Revolucionario Institucional (PRI) y al Partido Verde Ecologista de México (PVEM). La coalición Por el Bien de Todos (PBT) estuvo conformada por el Partido de la Revolución Democrática (PRD), el Partido del Trabajo (PT) y Convergencia. Los partidos Acción Nacional (PAN), Alternativa Socialdemócrata y Campesina (ASDC) y Nueva Alianza (denotado como "otr") compitieron de manera independiente.

[3] Para hacer compatibles los resultados, la variable `lag_apm` suma los votos del PRI, PVEM y los que obtuvo la coalición parcial Alianza para Todos (APT), `lag_pbt` suma los votos del PRD, PT y Convergencia. `lag_asdc` contiene los votos de México Posible (antecedente inmediato de ASDC), y `lag_otr` agrega los votos de los demás partidos que compitieron en esa elección. Los datos están ajustados para hacer compatibles los límites distritales de 2003 y 2006.

pres_pan	.8382251	.0199369	42.04	0.000	.7989896	.8774606
lag_pan	.1073357	.0220857	4.86	0.000	.0638714	.1507999
_cons	.0007425	.0042544	0.17	0.862	-.0076302	.0091152

Es decir, los resultados indican que por cada punto porcentual de votación que recibió el candidato presidencial del PAN, Felipe Calderón, los candidatos a diputado de su partido recibieron casi .84 puntos porcentuales, lo que representa un efecto de arrastre considerable. Este mismo ejercicio se puede repitir para el resto de los partidos políticos (por razones de espacio, omitimos los resultados de los siguientes comandos):

```
. reg apm  pres_apm  lag_apm
. reg pbt  pres_pbt  lag_pbt
. reg otr  pres_otr  lag_otr
. reg asdc pres_asdc lag_asdc
```

A simple vista es difícil observar que este ejercicio puede arrojar resultados lógicamente inconsistentes. Por ejemplo, supongamos que a partir del modelo de regresión deseamos predecir el resultado de la elección para la Cámara de Diputados en un distrito donde la competencia entre los principales candidatos presidenciales (PAN, APM y PBT) fuera sumamente reñida, de modo cada uno de ellos obtuviera 30% de los votos, y el resto de los votos se repartiera de manera simétrica entre los demás candidatos (5% cada uno). La predicción se puede expresar:

$$\hat{pan} = \_cons + (.838 * pres\_pan) + (.107 * lag\_pan) \tag{1}$$

Por simplicidad, supongamos que el porcentaje (rezagado) de votos que obtuvieron los partidos en ese distrito en la elección más reciente es igual al que recibieron en promedio en todos los distritos del país. Para el caso del PAN, esta predicción se puede calcular fácilmente con ayuda del comando `lincom`:

```
. sum lag_pan, meanonly
. lincom _cons + .3*pres_pan + `r(mean)´*lag_pan
 ( 1)  .3*pres_pan + .303111*lag_pan + _cons = 0
```

pan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	.2847447	.0019891	143.16	0.000	.2808302	.2886591

Es decir, en este distrito hipotético, el candidato del PAN para diputado obtendría poco más del 28% de los votos. Si extendemos el análisis a todos los partidos y sumamos las predicciones (véase columna 1 de la Tabla 1), observaríamos que la *votación total* sumaría 106%, lo cual es imposible. Pero los resultados pueden ser incluso más sensibles frente a otros escenarios hipotéticos; por ejemplo, si planteamos una situación extrema en que la votación de todos los candidatos presidenciales es idéntica (20% cada uno), la suma de los porcentajes sería de 133%!

Ahora consideremos el modelo de regresión del partido Nueva Alianza:

```
. reg otr pres_otr lag_otr
```

Source	SS	df	MS
Model	.061413334	2	.030706667
Residual	.074642257	297	.000251321

```
Number of obs =     300
F(  2,   297) =  122.18
Prob > F      =  0.0000
R-squared     =  0.4514
Adj R-squared =  0.4477
```

Total	.136055591	299	.000455035		Root MSE	=	.01585

otr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pres_otr	2.605222	.1669739	15.60	0.000	2.27662	2.933824
lag_otr	-.0242783	.0530429	-0.46	0.647	-.128666	.0801093
_cons	.0205118	.0021989	9.33	0.000	.0161845	.0248391

En realidad, el candidato presidencial de Nueva Alianza no transmitió ningún efecto de arrastre a sus correligionarios. De hecho, los resultados muestran que los diputados de Nueva Alianza recibieron 2.6 más votos que su candidato presidencial. Los resultados también indican que los votos que obtuvieron varios partidos pequeños en la elección de 2003 (y que perdieron su registro para competir en 2006) no tuvieron una relación significativa con el desempeño electoral del recién formado partido Nueva Alianza. Sin embargo, lo que queremos ilustrar es que si planteamos un escenario donde el candidato presidencial de Nueva Alianza obtuviera 40% de los votos, el partido obtendría *por sí solo* más del 100% de la votación para Diputados:

```
. sum lag_otr, meanonly
. lincom _cons + .4*pres_otr + `r(mean)´*lag_otr
 ( 1)  .4*pres_otr + .0197143*lag_otr + _cons = 0
```

otr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	1.062122	.0651253	16.31	0.000	.9339565	1.190287

Incluso, bajo otras especificaciones del modelo, un partido podría recibir una votación negativa, es decir, menos de cero votos. ¿Por qué MCO arroja estos resultados, y cómo pueden evitarse?

2.2 Datos composicionales

Estas inconsistencias ocurren porque los resultados electorales multipartidistas violan dos supuestos subyacentes del modelo estándar. En primer lugar, MCO asume que la variable dependiente (en este caso, la proporción de votos de un partido) es una variable continua irrestricta, es decir, que potencialmente puede tomar cualquier valor (Berry, 1993). Sin embargo, la proporción de votos de un partido se encuentra necesariamente acotada entre cero y uno. Denotemos a V_{ij} como la proporción de votos del partido j ($j = 1, 2, \ldots, J$) en el distrito electoral i ($i = 1, \ldots, n$). Formalmente,

$$V_{ij} \in [0, 1] \ \forall \ i, j \tag{2}$$

En segundo lugar, el modelo estándar presupone que la proporción de votos de un partido es independiente de la de los demás; naturalmente, en la composición porcentual de una contienda electoral, el porcentaje de votos de un partido está inversamente relacionado con el voto de los demás toda vez que el total de las proporciones de votos debe ser igual a uno. Puesto de manera formal,

$$\sum_{j=1}^{J} V_{ij} = 1 \ \forall \ i \tag{3}$$

Existen algunas soluciones que se pueden implementar fácilmente en Stata para que las predicciones del modelo se encuentren dentro del intervalo de la unidad (véase por ejemplo, Baum, 2008). Estas técnicas satisfacen la condición de la ecuación (2), pero ignoran la propiedad de la ecuación (3). Otra alternativa que satisface las dos propiedades está basada en la distribución Dirichlet y puede implementarse en Stata a través del paquete `dirifit` (Buis et al., 2006). Sin embargo, los resultados electorales multipartidistas tienen algunas características particulares, y por eso en este capítulo nos enfocamos en la aplicación de los modelos propuestos por Katz & King (1999); Honaker et al. (2002) y Tomz et al. (2002). Sugerimos al lector referirse a dichos trabajos para familiarizarse con los fundamentos de estos modelos con mayor detalle.

Katz & King (1999) proponen utilizar un modelo de Máxima Verosimilitud con Información Total (*Full information Maximum Likelihood* o FIML en inglés). Su modelo incluye varias adecuaciones al modelo estándar de regresión, entre las que destaca tratar a las proporciones de votos de los partidos como datos composicionales (Aitchison & J. Egozcue, 2005). Esta técnica se utiliza con frecuencia en otras disciplinas como Geología y Biología, y consiste aplicar una transformación logística multivariada a las variables dependientes. Específicamente, se transforma en *log ratios* la proporción de votos del partido $j = 1, \ldots, J-1$ respecto a un partido base J:

$$Y_i = \ln\left(V_{i1}/V_{iJ}\right) \;\; \forall \; j \neq J \tag{4}$$

La transformación logística cambia la proporción de votos a una escala irrestricta, satisfaciendo el supuesto de MCO. En nuestra aplicación, asumiremos que el partido base es ASDC, aunque en términos prácticos, no hace ninguna diferencia en los resultados el partido que se escoge para este fin. Para el caso del PAN, la transformación logística puede realizarse en Stata de la siguiente manera:

```
. generate t_pan = ln(pan/asdc)
```

Esta instrucción genera una nueva variable que transforma la variable `pan` en su *log ratio* de acuerdo con la ecuación (4). Ahora podemos especificar el mismo modelo que en el ejemplo anterior, pero con la transformación logística `t_pan` como variable dependiente.

```
. reg t_pan pres_pan lag_pan
```

Source	SS	df	MS		
Model	55.3268118	2	27.6634059		
Residual	99.9819877	297	.336639689		
Total	155.3088	299	.519427423		

Number of obs	=	300			
F(2, 297)	=	82.18			
Prob > F	=	0.0000			
R-squared	=	0.3562			
Adj R-squared	=	0.3519			
Root MSE	=	.58021			

t_pan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pres_pan	3.487068	.4319901	8.07	0.000	2.636919	4.337217
lag_pan	-.6673026	.478549	-1.39	0.164	-1.609079	.274474
_cons	1.818659	.0921847	19.73	0.000	1.637241	2.000077

Luego, en lugar de especificar J modelos de regresión–uno para cada partido–, se especifican $J-1$ ecuaciones, excluyendo al partido que se tomó como base:

```
. gen t_apm = ln(apm/asdc)
```

```
. reg t_apm  pres_apm  lag_apm
. gen t_pbt = ln(pbt/asdc)
. reg t_pbt  pres_pbt  lag_pbt
. gen t_otr = ln(otr/asdc)
. reg t_otr  pres_otr  lag_otr
```

Los coeficientes de las regresiones resultantes son muy distintos a los de MCO. La razón es que están expresados en *log ratios*, una escala difícil de interpretar directamente. Para interpretar estas cantidades en proporciones, debemos aplicar la tranformación inversa:

$$\hat{V}_{ij} = \frac{\exp\left(\hat{Y}_{ij}\right)}{1 + \sum_{j=1}^{J-1} \exp\left(\hat{Y}_{ij}\right)} \; \forall \; j \neq J, \tag{5}$$

mientras que la proporción del partido base J es simplemente la diferencia

$$\hat{V}_{iJ} = 1 - \sum_{j=1}^{J-1} \hat{V}_{ij} \tag{6}$$

Por ejemplo, podemos instruir a Stata a que realice la misma predicción con la que hemos trabajado hasta ahora con los coeficientes expresados en *log ratios*:

```
. sum lag_pan, meanonly
. lincom _cons + .3*pres_pan + `r(mean)´*lag_pan
 ( 1)  .3*pres_pan + .303111*lag_pan + _cons = 0
```

t_pan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	2.662513	.0430986	61.78	0.000	2.577695	2.74733

Para transformar el resultado en proporciones primero repetimos la misma instrucción para cada partido (por razones de espacio, omitimos los resultados). Luego, podemos obtener la predicción del PAN expresada en proporciones:

```
. display exp(2.6625)/(1+71.5757)
.19747759
```

donde 71.5757 es el resultado de la sumatoria $\sum_{j=1}^{J-1} \exp\left(\hat{Y}_{ij}\right)$. Al realizar el mismo ejercicio para el resto de los partidos políticos (con excepción de ASDC), se observa que la proporción de cada uno se encuentra dentro del intervalo de la unidad, y que por construcción, en conjunto suman 1. Las columnas 1 y 2 de la Tabla 1 contrastan los resultados de la predicción entre el modelo estándar y la técnica de *log ratios* recomendada por Katz & King (1999). Como se puede apreciar, existen diferencias importantes en la predicción de casi todos los partidos políticos con excepción de APM.

El modelo de Katz & King (1999) incorpora otros dos elementos que no consideraremos de manera detallada en este capítulo. Por una parte, los autores modelan los *log ratios* con una distribución t multivariada en lugar de la distribución Normal debido a que, en algunos casos, las proporciones de votos podrían ajustarse mejor a la primera que a la segunda. No obstante, los experimentos de Tomz et al. (2002) muestran que ambas distribuciones arrojan resultados muy parecidos, y por lo tanto, "al adoptar la

Table 1: Predicción de voto para Diputados 2006 en un distrito hipotético

Partido	MCO estándar	MCO con *Log ratios*	SURE(Tomz et al., 2002)
PAN	.285	.197	.2
APM (PRI PVEM)	.344	.358	.317
PBT (PRD PT Conv)	.248	.176	.172
NA ("otr")	.15	.255	.297
ASDC	.037	.014	.014
Total	1.064	1.0	1.0

Nota: Predicciones manteniendo los siguientes valores para el voto de candidatos presidenciales: PAN=.3; PRI=.3; PRD=.3, NA=.05; ASDC=.05, manteniendo el voto rezagado en su valor promedio.

[distribución] Normal se pierde poco de interés sustantivo pero se gana mucho en facilidad de implementación" (Tomz et al., 2002, p.71; también véase Breusch et al., 1997). Sin embargo, si el investigador observa que los datos se desvían signficativamente de la distribución normal, podría interesarle la regresión t implementada en Stata por Marchenko & Genton (2010).

En segundo lugar, Katz & King (1999) incorporan en su modelo un componente especial para lidiar con distritos con patrones de competencia parcial (en los que uno o varios partidos no presentan candidatos). En la práctica, dicho componente del modelo puede ser computacionalmente demandante y difícil de programar para el usuario. Afortunadamente, en un artículo más reciente (Honaker et al., 2002), los autores presentan una alternativa rápida y sencilla basada en imputación múltiple.[4] Para mayor referencia sobre cómo implementar el método de imputación múltiple en Stata, el usuario puede consultar el manual relacionado con ese tema.

2.3 SURE

Consideremos nuevamente la propiedad (3) de los datos electorales multipartidistas. Las variables dependientes del sistema de ecuaciones de la sección anterior están construidas a partir de proporciones de votos. Como mencionamos antes, estas proporciones están relacionadas en tanto un aumento en el *log ratio* de un partido significa un menor *log ratio* para los demás. Por lo tanto, es plausible suponer que los términos de error de las $J-1$ ecuaciones estimadas con MCO estarán correlacionados.

Cuando los términos de error están correlacionados, MCO produce estimadores consistentes, pero el método de Mínimos Cuadrados Generalizados Factibles (MCGF) proporciona estimadores más eficientes. Esta característica de MCGF es explotada por el método de Ecuaciones de Regresión Aparentemente Inconexas o SURE, por su abre-

[4]Por otro lado,Tomz et al. (2002) ofrecen un enfoque alternativo que consiste en en efectuar análisis separados para cada patrón de competencia.

viación en ingés (*Seemingly Unrelated Regression Equations*) (Zellner, 1962). Por esta razón, Tomz et al. (2002) argumentan que SURE es más conveniente y no menos eficiente que MCO para estimar sistemas de ecuaciones con datos electorales multipartidistas.

Existen dos circunstancias en las que los resultados que arrojan MCO y SURE son equivalentes. En primer lugar, cuando las variables independientes son las mismas en todas las ecuaciones de regresión. En segundo lugar, cuando la correlación entre los términos de error de las ecuaciones es muy pequeña o inexistente. Sin embargo, en términos prácticos, SURE es una herramienta valiosa aún en estas situaciones, pues permite al usuario especificar las $J-1$ ecuaciones partidistas en un solo paso.

Denotemos al vector de $J-1$ *log ratios* en el distrito i como:

$$\mathrm{Y}_i = \left[\ln\left(V_{i1}/V_{iJ}\right),\ \ln\left(V_{i2}/V_{iJ}\right), \ldots,\ \ln\left(V_{i(J-1)}/V_{iJ}\right)\right]$$

Y expresemos el modelo estadístico (King, 1998):

$$Y_i \sim N\left(\mu_i, \Sigma\right), \tag{7}$$

$$\mu_i = \left[\mathbf{X}_{i1}\beta_1, \ldots, \mathbf{X}_{i(J-1)}\beta_{(J-1)}\right], \tag{8}$$

donde Y_i representa el componente estocástico del modelo, y asumimos que se distribuye de manera normal multivariada con media μ_i y con matriz de varianza Σ. La ecuación (8) representa el componente sistemático del modelo, y se expresa como una función lineal del vector de variables explicativas $\mathbf{X}$ y el de parámetros o coeficientes $\boldsymbol{\beta}$.

El primer paso para estimar el modelo SURE es transformar las proporciones de votos en *log ratios* según la ecuación (4). Una vez hecho esto, la estimación del modelo estadístico en Stata es como sigue:

```
. sureg   (t_pan pres_pan lag_pan)        ///
>         (t_apm pres_apm lag_apm)        ///
>         (t_pbt pres_pbt lag_pbt)        ///
>         (t_otr pres_otr lag_otr)
Seemingly unrelated regression
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
t_pan	300	2	.5832537	0.3429	525.55	0.0000
t_apm	300	2	.5027267	0.6996	833.44	0.0000
t_pbt	300	2	.6395739	0.2108	423.42	0.0000
t_otr	300	2	.4249401	0.3425	277.75	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t_pan						
pres_pan	3.100458	.2537922	12.22	0.000	2.603035	3.597882
lag_pan	.2925297	.2587341	1.13	0.258	-.2145799	.7996393
_cons	1.667975	.0625521	26.67	0.000	1.545375	1.790575
t_apm						
pres_apm	5.60672	.3815648	14.69	0.000	4.858867	6.354573
lag_apm	1.023284	.2845275	3.60	0.000	.4656207	1.580948
_cons	1.028574	.0797296	12.90	0.000	.8723066	1.184841
t_pbt						

pres_pbt	2.974831	.286753	10.37	0.000	2.412805	3.536856
lag_pbt	.493751	.259587	1.90	0.057	-.0150302	1.002532
_cons	1.529438	.0738244	20.72	0.000	1.384745	1.674131
t_otr						
pres_otr	55.33552	3.462706	15.98	0.000	48.54875	62.1223
lag_otr	-4.642008	1.094761	-4.24	0.000	-6.7877	-2.496316
_cons	.4091474	.0480197	8.52	0.000	.3150305	.5032644

Como se aprecia, los resultados aparecen separados en ecuaciones, una para cada uno de los $J-1$ partidos políticos. Para verificar si efectivamente los residuales de las ecuaciones están correlacionados, implementamos la prueba de independencia de Breusch & Pagan (1979):

```
. sureg, notable noheader corr

Correlation matrix of residuals:

        t_pan   t_apm   t_pbt   t_otr
t_pan   1.0000
t_apm   0.6241  1.0000
t_pbt   0.8344  0.5341  1.0000
t_otr   0.6126  0.5006  0.5422  1.0000

Breusch-Pagan test of independence: chi2(6) =   687.211, Pr = 0.0000
```

Los resultados de la prueba indican que podemos rechazar la hipótesis nula de que los residuales son independientes u ortogonales, por lo que en este caso SURE resulta ser más eficiente que MCO.

Dado que el procedimiento estima varias ecuaciones a la vez, para calcular una predicción para un distrito hipotético debemos indicar la ecuación o partido con el que deseamos trabajar. Por ejemplo, para el caso del PAN:

```
. sum lag_pan, meanonly
. lincom [t_pan]_cons + [t_pan]pres_pan*.3 + [t_pan]lag_pan*`r(mean)´
 ( 1)  .3*[t_pan]pres_pan + .303111*[t_pan]lag_pan + [t_pan]_cons = 0
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	2.686781	.0369421	72.73	0.000	2.614376	2.759187

Y las instrucciones para realizar la misma predicción con el resto de las ecuaciones serían:

```
. lincom [t_apm]_cons + [t_apm]pres_apm*.3 + [t_apm]lag_apm*.43
. lincom [t_pbt]_cons + [t_pbt]pres_pbt*.3 + [t_pbt]lag_pbt*.24
. lincom [t_otr]_cons + [t_otr]pres_otr*.05+ [t_otr]lag_otr*.02
```

La tercera columna de la Tabla 1 compara los resultados del modelo SURE de Tomz et al. (2002) con los modelos estándar y de *log ratios* estimados anteriormente. Como se aprecia, los resultados de los modelos con tranformación logística son muy parecidos.

3 Cantidades de interés

Algunos académicos especializados utilizan el término *cantidades de interés* para referirse a funciones de los parámetros del modelo estadístico que contienen información sustantiva de los resultados (King et al., 2000). Muchas preguntas interesantes sobre los sistemas electorales –como las que se plantean al inicio de este capítulo– requieren algo

más que la interpretación del signo, magnitud y significancia de los coeficientes de un modelo de regresión. Por ejemplo, mientras que a un analista electoral le interesa conocer la magnitud del arrastre de los candidatos presidenciales en la votación distrital en la arena legislativa, a un estudioso de los sistemas presidenciales podría interesarle saber cuál es el porcentaje de votos que debe obtener un candidato presidencial ganador para que su partido consiga la mayoría en la Cámara de Diputados (véase por ejemplo, Campbell, 1986). Además, los efectos de arrastre de los candidatos presidenciales son relevantes para estudiar la gobernabilidad en la región, puesto que entre mayor sea el efecto de arrastre, la probabilidad que exista gobierno dividido es menor (situación que ocurre cuando un partido distinto al del Presidente mantiene la mayoría en la legislatura), y por lo tanto, es menos probable que exista *deadlock* o parálisis legislativa.

Para ilustrar esta idea a partir de un ejercicio sencillo, calcularemos la asignación o reparto de asientos en la asamblea a partir de los porcentajes de votos obtenidos por cada partido. Con este propósito, hemos escrito el comando `asignadip` que se documenta en el Apéndice. Por omisión, `asignadip` implementa las reglas vigentes en México para la asignación de asientos en la Cámara de Diputados: 300 asientos se reparten por regla de mayoría relativa en igual número de distritos uninominales, y 200 asientos de representación proporcional se distribuyen entre los partidos que obtuvieron más de 2% de los votos con la fórmula de D'Hont y el método de restos mayores en un distrito plurinominal nacional (Molinar Horcasitas & Weldon, 2001).[5]

```
. asignadip pan apm pbt otr asdc
(running asignadip)
Threshold:                .02
Max. num. of seats:       300
Max. overrepresentation:  .08
Iterations:                 0
Seat allocation
```

Party	SSD	List	Total	%Seats	%Votes	%Valid
pan	137	67	204	.408	.3373627	.3373627
apm	65	60	125	.25	.2982391	.2982391
pbt	98	60	158	.316	.2981795	.2981795
otr	0	9	9	.018	.0461007	.0461007
asdc	0	4	4	.008	.020118	.020118
Total	300	200	500	1	1	1

Los resultados indican que el PAN obtuvo 204 asientos o 40.8% de la Cámara de Diputados. Estos resultados difieren ligeramente de la asignación observada en la realidad, que puede calcularse con el comando `asignadipi` (la versión inmediata del comando `asignadip`) usando los porcentajes de votos de cada partido a nivel nacional:

```
. asignadipi, ssd(137 65 98 0 0) voteshare(.3341 .2818 .29 .0455 .0205)
(running asignadip)
Threshold:                .02
Max. num. of seats:       300
Max. overrepresentation:  .08
Iterations:                 0
```

[5]En México existen cinco circunscripciones plurinominales con listas regionales; sin embargo, los asientos plurinominales se asignan con base en la votación nacional. Para una explicación de las inconsistencias que surgen de estas reglas, véase Balinski & Ramírez González (1999)

Seat allocation

Party	SSD	List	Total	%Seats	%Votes	%Valid
party1	137	69	206	.412	.3341	.3437596
party2	65	58	123	.246	.2818	.2899475
party3	98	60	158	.316	.29	.2983846
party4	0	9	9	.018	.0455	.0468155
party5	0	4	4	.008	.0205	.0210927
Total	300	200	500	1	.9719	1

Las pequeñas diferencias se deben a que `asignadip` toma el porcentaje distrital promedio como una aproximación del porcentaje total de votos (este último es el que en realidad se considera al repartir los asientos de representación proporcional) y a que, por parsimonia, no incluimos el porcentaje de votos nulos y para candidatos no registrados.[6]

Ahora bien, una manera sencilla de calcular cantidades de interés relacionadas con la conformación de la Cámara de Diputados es a través del comando `predict`. Por ejemplo, supongamos que deseamos saber si un 10% de votación "extra" para el candidato presidencial del PAN hubieran sido suficientes para que su partido obtuviera la mayoría en la Cámara de Diputados. El primer paso consiste en estimar el modelo de la sección anterior, y luego plantear el escenario hipotético, modificando los valores de las variables independientes:[7]

```
. replace pres_pan  = pres_pan  +.1
(300 real changes made)
. replace pres_apm  = pres_apm  -.05
(300 real changes made)
. replace pres_pbt  = pres_pbt  -.05
(300 real changes made)
```

En nuestro escenario hipotético, asumimos que el aumento en la votación del candidato presidencial del PAN proviene de un *swing* uniforme en todos los distritos, y que las ganancias del PAN (+10%) provienen de pérdidas para los otros dos partidos más importantes (-5% para APM y PBT) (véase Borisyuk et al., 2008). Luego, el resultado de las predicciones, expresadas en *log ratios,* deben tranformarse a proporciones:

```
. local i = 0
. foreach partido in pan apm pbt otr {
  2.         predict pred_`partido´, xb equation(#`++i´)
  3.         replace pred_`partido´ = exp(pred_`partido´)
  4. }
(300 real changes made)
(300 real changes made)
(300 real changes made)
(300 real changes made)
. egen total = rowtotal(pred_*)
. foreach partido in pan apm pbt otr {
  2.         replace pred_`partido´ = pred_`partido´/total
```

[6]Normalmente, el porcentaje distrital promedio y el porcentaje total de votos es muy parecido. Sin embargo, entre mayor sea la diferencia de aporcionamiento y/o de tasa de participación entre los distritos electorales, mayor será la diferencia entre uno y otro. Es recomendable que el usuario verifique que ambas cantidades son parecidas antes de efectuar el análisis.

[7]Para no modificar permanentemente la base de datos en la memoria, el usuario podría usar el comando `preserve`.

```
  3. }
(300 real changes made)
(300 real changes made)
(300 real changes made)
(300 real changes made)
. egen total_prop = rowtotal(pred_*)
. generate pred_base = 1-total_prop
```

El primer bloque de comandos instruye a Stata a que calcule las predicciones de voto para cada distrito bajo el escenario hipotético, y las guarde en las variables `pred_pan`, `pred_apm`, `pred_pbt` y `pred_otr`. Luego, reemplaza cada una de estas variables con la función exponencial: `exp(pred_pan)`, `exp(pred_apm)`, etc. Para reducir el número de líneas de comando, utilizamos el comando `foreach` (Cox, 2002).

El segundo bloque de comandos genera una variable con la suma de las predicciones, y utiliza la ecuación (5) para tranformar las predicciones en proporciones. Finalmente, el último bloque calcula la predicción para el partido que utilizamos como base en la transformación logística multivariada (ecuación 6).

Una vez que contamos con nuestro escenario hipotético –el porcentaje de votos esperado por distrito para cada partido si el candidato presidencial del PAN obtuviera 10 puntos más de votación–, el siguiente paso consiste en asignar los asientos en la Cámara que le corresponden a cada partido político.

```
. asignadip pred_*
(running asignadip)
Threshold:                 .02
Max. num. of seats:        300
Max. overrepresentation:   .08
Iterations:                  1
Seat allocation
```

Party	SSD	List	Total	%Seats	%Votes	%Valid
pred_pan	189	75	264	.528	.4493737	.4493737
pred_apm	47	55	102	.204	.2444209	.2444209
pred_pbt	64	60	124	.248	.2631866	.2631866
pred_otr	0	10	10	.02	.0430187	.0430187
pred_base	0	0	0	0	9.93e-10	0
Total	300	200	500	1	1	1

Los resultados sugieren que efectivamente, el partido del Presidente hubiera gozadode una mayoría simple en la Cámara de Diputados bajo el escenario hipotético: el PAN obtendría 264 asientos (189 de mayoría relativa y 75 de representación proporcional), lo que representa 53% de la legislatura. Los resultados también indican que durante el procedimiento fue necesaria una iteración; es decir, el proceso de asignación se realizó en dos pasos debido a que algún partido se encontró en uno de los supuestos de la cláusula de sobrerrepresentación. En México, la cláusula de sobrerrepresentación ordena que ningún partido puede tener más de 300 asientos ó un porcentaje de asientos que represente 8 puntos porcentuales más que su votación válida.[8]

[8]Según el COFIPE, la votación válida (o votación nacional emitida) se obtiene al deducir de la votación total los votos nulos, para candidatos no registrados, y para partidos con menos de 2% de votación. El segundo supuesto de la cláusula no se aplica cuando la sobrerrepresentación se debe a los triunfos del partido en los distritos uninominales.

No obstante, las opciones del comando `asignadip` son lo suficientemente flexibles como para que pueda adaptarse a las reglas de asignación de otros países con sistemas mayoritarios, proporcionales, o mixtos. Por ejemplo, supongamos que deseamos analizar el mismo efecto del aumento en la votación del candidato presidencial del PAN en la configuración partidista de la Cámara de Diputados, pero esta vez ajustando las reglas de asignación de modo que no exista tope de sobrerrepresentación (o lo que es lo mismo, fijándola en 100%), y que el umbral mínimo de votación para tener derecho al reparto de asientos (*representation threshold*) sea de 4.5%:

```
. asignadip pred_*, maxoverr(1) threshold(.045)
(running asignadip)
Threshold:                .045
Max. num. of seats:       300
Max. overrepresentation:    1
Iterations:                 0
Seat allocation
```

Party	SSD	List	Total	%Seats	%Votes	%Valid
pred_pan	189	94	283	.566	.4493737	.4695742
pred_apm	47	51	98	.196	.2444209	.2554082
pred_pbt	64	55	119	.238	.2631866	.2750176
pred_otr	0	0	0	0	.0430187	0
pred_base	0	0	0	0	9.93e-10	0
Total	300	200	500	1	1	1

Los resultados indican que, si el candidato presidencial panista hubiera tenido 10% más de votos y no hubiera tope de sobrerrepresentación, el PAN obtendría 283 asientos o 57% de la legislatura. Como puede observarse, sus ganancias provienen enteramente del reparto de los asientos plurinominales, que pasaron de 75 a 94 curules. Esto se debe, por un lado, a que el aumento del umbral incrementó el número de asientos disponibles para los partidos con más de 4.5% de votación, y a la eliminación del tope de sobrerrepresentación que en un principio afectaba al PAN, por otro.

4 Medidas de incertidumbre

Una de las principales dificultades que surgen al calcular cantidades de interés como las de la sección anterior es estimar sus medidas de incertidumbre. Como mencionamos en la sección anterior, las cantidades de interés son funciones de los parámetros del modelo. En nuestra aplicación, dicha función consiste simplemente en las reglas de asignación de asientos. Puesto que nuestro conocimiento sobre los valores de los parámetros es incierto –lo cual se refleja en el hecho de que los investigadores usualmente reportan los coeficientes de regresión acompañados de errores estándar–, las cantidades de interés también son inciertas.

Ahora bien, ¿cuáles son las fuentes de incertidumbre que debemos considerar al elaborar cantidades de interés? En términos generales, los modelos estadísticos presentan dos fuentes de variabilidad (King, 1998). Primero, la variabilidad en la estimación de los parámetros –comúnmente denominada *error muestral*. Segundo, la que surge por la imposibilidad de hacer inferencias determinísticas a partir del modelo –también llamada incertidumbre fundamental– y que se expresa en la ecuación (7) del modelo.

En el caso concreto de los modelos estadísticos para datos multipartidistas, "la primera trata los votos observados como resultado de una muestra de votantes en cada distrito. El segundo elemento estocástico surge por el tratamiento usual del término de error en un modelo de regresión, es decir, la imposibilidad [...] de predecir un proceso manera perfecta" (Jackson, 2002).

Desafortunadamente, existen algunas cantidades de interés para las cuales es difícil o imposible calcular medidas de incertidumbre con los métodos tradicionales. Las reglas de asignación de asientos son muy variadas, suelen ser complejas y con varias salvaguardas, por lo que es difícil derivar una fórmula para calcular errores estándar e intervalos de confianza. Frente a esta situación, la técnica de simulación estadística es una manera práctica de propagar la incertidumbre del modelo a las cantidades de interés que se derivan del mismo(King et al., 2000).

4.1 Simulación post-estimación

La simulación estadística se basa en el principio de Monte Carlo, según el cual podemos conocer o describir cualquier variable aleatoria obteniendo una muestra de su distribución de probabilidad (Jackman, 2009). De acuerdo con el Teorema del Límite Central, la distribución de probabilidad de los parámetros del modelo es (asintóticamente) normal multivarada. Por lo tanto, podemos aproximarnos a la distribución de probabilidad de cualquier cantidad de interés seleccionando o *simulando* aleatoriamente T valores de la distribución de los parámetros, y calculando la cantidad de interés con cada uno de ellos. Normalmente, $T = 1,000$ es un número suficiente de simulaciones para la mayoría de las aplicaciones (Katz & King, 1999; King et al., 2000).

El paquete Clarify (Tomz et al., 2003) de Stata implementa el método de simulación post-estimación desarrollado por King et al. (2000). Además, Clarify incluye algunas adecuaciones para facilitar la estimación de modelos estadísticos con datos multipartidistas con SURE. Actualmente, Clarify puede calcular tres cantidades de interés que se describen con mayor detalle en Katz & King (1999): el voto predicho, el voto esperado, y efectos causales. El voto predicho es la distribución de probabilidad que describe la proporción de votos que obtiene cada partido en un distrito, manteniendo las variables explicativas en valores fijos. El resultado incorpora ambas fuentes de incertidumbre del modelo estadístico –variabilidad fundamental y muestral. En cambio, el voto esperado promedia la variabilidad fundamental, incorporando únicamente la variabilidad muestral. Por esta razón, los valores esperados tienen menor varianza que los valores predichos.[9] Por último, un efecto causal es la diferencia entre dos valores esperados dado un cambio en el valor de alguna variable explicativa.

El ejemplo que hemos utilizado a lo largo del capítulo puede estimarse con Clarify en cuatro sencillos pasos:[10]

```
. tlogit  pan t_pan       ///
>         apm t_apm       ///
>         pbt t_pbt       ///
```

[9] Sobre las situaciones en las que es más conveniente calcular un valor predicho o esperado, véase King et al. (2000).

[10] Para mayores detalles, véase Tomz et al. (2003)

```
>         otr t_otr,      ///
>         base(asdc)

. set seed 12345

. estsimp sureg ( t_pan pres_pan lag_pan) ///
>               ( t_apm pres_apm lag_apm) ///
>               ( t_pbt pres_pbt lag_pbt) ///
>               ( t_otr pres_otr lag_otr), sims(500)

Seemingly unrelated regression
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
t_pan	300	2	.5832537	0.3429	525.55	0.0000
t_apm	300	2	.5027267	0.6996	833.44	0.0000
t_pbt	300	2	.6395739	0.2108	423.42	0.0000
t_otr	300	2	.4249401	0.3425	277.75	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t_pan						
pres_pan	3.100458	.2537922	12.22	0.000	2.603035	3.597882
lag_pan	.2925297	.2587341	1.13	0.258	-.2145799	.7996393
_cons	1.667975	.0625521	26.67	0.000	1.545375	1.790575
t_apm						
pres_apm	5.60672	.3815648	14.69	0.000	4.858867	6.354573
lag_apm	1.023284	.2845275	3.60	0.000	.4656207	1.580948
_cons	1.028574	.0797296	12.90	0.000	.8723066	1.184841
t_pbt						
pres_pbt	2.974831	.286753	10.37	0.000	2.412805	3.536856
lag_pbt	.493751	.259587	1.90	0.057	-.0150302	1.002532
_cons	1.529438	.0738244	20.72	0.000	1.384745	1.674131
t_otr						
pres_otr	55.33552	3.462706	15.98	0.000	48.54875	62.1223
lag_otr	-4.642008	1.094761	-4.24	0.000	-6.7877	-2.496316
_cons	.4091474	.0480197	8.52	0.000	.3150305	.5032644

```
Simulating main parameters.  Please wait....

Note: Clarify is expanding your dataset from 300 observations to 500
observations in order to accommodate the simulations.  This will append
missing values to the bottom of your original dataset.

% of simulations completed: 8% 16% 25% 33% 41% 50% 58% 66% 75% 83% 91% 100%

Simulating Sigma matrix.  Please wait....
% of simulations completed: 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Number of simulations  : 500
Names of new variables : b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16
>  b17 b18 b19 b20 b21 b22

. setx mean

. setx pres_pan .3 pres_apm .3 pres_pbt .3 pres_otr .05

. simqi, ev tfunc(logiti)

Performing calculations.  Please wait...
10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

      Quantity of Interest |      Mean       Std. Err.     [95% Conf. Interval]
---------------------------+--------------------------------------------------
          E[logiti(t_pan)] |   .1989129     .0092293     .1804505     .2154857
          E[logiti(t_apm)] |   .3154078      .015141      .286759     .3445725
          E[logiti(t_pbt)] |    .175462     .0085039      .157693     .1921717
          E[logiti(t_otr)] |   .2959757     .0290625      .245878     .3526031
```

El primer comando (`tlogit`) genera un nuevo grupo de variables aplicando la tranformación logística multivariada a las variables dependientes. El comando `estsimp` *est*ima el modelo SUR (los resultados son idénticos a los de la sección anterior, y se

omiten por razones de espacio) y *sim*ula 500 valores de los *p*arámetros del modelo. Cada valor simulado representa un registro en las variables `b1`, `b2`, ..., `b18`. `setx` indica a Stata los valores de las variables explicativas con los que deseamos calcular la cantidad de interés (en este caso, fijamos los mismos valores que utilizamos en la sección anterior). Finalmente, el comando `simqi` *sim*ula la *c*antidad de *i*nterés (`ev`, el voto esperado para cada partido) y aplica la transformación logística inversa (`logiti`).

Los resultados varían ligeramente con respecto a los de la última columna de la Tabla 1. La razón es que los valores esperados se calculan con una muestra aleatoria de la distribución de los parámetros. Aumentar el número de simulaciones a través de la opción `sims()` contribuye a que los resultados sean más precisos.[11] Naturalmente los valores simulados difieren de una muestra a otra, por lo que para fines de replicabilidad, conviene fijar un arranque aleatorio con el comando `set seed`.

4.2 Medidas de incertidumbre para cantidades relacionadas con sistemas electorales

El voto esperado es una cantidad de interés que, como hemos visto en secciones anteriores, puede calcularse de manera analítica con los métodos tradicionales. La ventaja de hacerlo con simulación estadística es que los resultados pueden emplearse para calcular otras cantidades de interés más complejas y para las cuales es difícil o imposible derivar una solución analítica.

Consideremos nuevamente el problema de calcular la conformación de la Cámara de Diputados frente a un escenario hipotético. Las diferencias entre este problema y el ejemplo de la sección anterior pueden ilustarse fácilmente con la Tabla 2 (adaptada de Gelman & King, 1994). Cada fila representa un distrito, con el número del distrito en la primera columna y el porcentaje de votos observado del partido j en la segunda columna. Las demás columnas representan simulaciones de la distribución del porcentaje de votos. Por ejemplo, el voto observado del partido j en el distrito 1 es V_{1j}, y la primera simulación es $\tilde{V}_{1j}^{(hip)1}$, la segunda es $\tilde{V}_{1j}^{(hip)2}$, y así sucesivamente.

Normalmente, los usuarios utilizan Clarify para calcular T replicaciones hipotéticas de la proporción de votos de cada partido *en un distrito* bajo algún escenario específico, es decir, una fila de la Tabla 2. En cambio, para calcular la conformación de la Cámara requerimos T replicaciones hipotéticas del *resultado electoral*, lo cual implica trazar simulaciones para cada uno de los distritos. En la Tabla 2 se observa que cuando se simulan T valores de la distribución de votos para cada distrito (filas), se obtienen T resultados electorales hipotéticos para cada partido (columnas). Considerando el número de partidos y distritos en distintos países, se requerirían varios cientos o miles de líneas de código para obtener una matriz como la de la Tabla 2.

El paquete **camaradip** que documentamos en el Apéndice es un *wrapper* que automatiza y simplifica notablemente esta tarea y genera, con un par de comandos, una base datos con T elecciones simuladas para cada distrito y partido similar a la Tabla 2.

[11] Otra manera de lograr resultados similares a la solución analítica es utilizar la opción `antisim` del comando `estsimp`. Esta opción instruye a Stata a simular los parámetros con simulaciones antitéticas, de manera que las medias de las simulaciones sean prácticamente iguales a los coeficientes de regresión.

Cuadro 2: Estructura del modelo estadístico para la Cámara de Diputados

Distrito	Resultado observado	Replicaciones hipotéticas			
		1	2	$\cdots$	T
1	V_{1j}	$\tilde{V}_{1j}^{(hip)1}$	$\tilde{V}_{1j}^{(hip)2}$	$\cdots$	$\tilde{V}_{1j}^{(hip)T}$
2	V_{2j}	$\tilde{V}_{2j}^{(hip)1}$	$\tilde{V}_{2j}^{(hip)2}$	$\cdots$	$\tilde{V}_{2j}^{(hip)T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	V_{nj}	$\tilde{V}_{nj}^{(hip)1}$	$\tilde{V}_{nj}^{(hip)2}$	$\cdots$	$\tilde{V}_{nj}^{(hip)T}$
Cantidad de interés	ψ_j	$\tilde{\psi}_j^{(hip)1}$	$\tilde{\psi}_j^{(hip)2}$	$\cdots$	$\tilde{\psi}_j^{(hip)T}$

La etapa de estimación se ejecuta con el comando `estimadip`:

```
. estimadip (pan pres_pan lag_pan)        ///
>           (apm pres_apm lag_apm)        ///
>           (pbt pres_pbt lag_pbt)        ///
>           (otr pres_otr lag_otr),       ///
>           base(asdc) sims(500) seed(12345) replace
Multivariate Logistic Transformation
       variable     new variable       description
------------------------------------------------------------------------------
            pan     _tlog_pan          ln(pan/base)
            apm     _tlog_apm          ln(apm/base)
            pbt     _tlog_pbt          ln(pbt/base)
            otr     _tlog_otr          ln(otr/base)
           asdc     (base variable)
------------------------------------------------------------------------------

Seemingly unrelated regression
------------------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2         P
------------------------------------------------------------------------------
_tlog_pan         300      2    .5832537    0.3429     525.55    0.0000
_tlog_apm         300      2    .5027267    0.6996     833.44    0.0000
_tlog_pbt         300      2    .6395739    0.2108     423.42    0.0000
_tlog_otr         300      2    .4249401    0.3425     277.75    0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_tlog_pan    |
    pres_pan |   3.100458   .2537922    12.22   0.000     2.603035    3.597882
     lag_pan |   .2925297   .2587341     1.13   0.258    -.2145799    .7996393
       _cons |   1.667975   .0625521    26.67   0.000     1.545375    1.790575
-------------+----------------------------------------------------------------
_tlog_apm    |
    pres_apm |    5.60672   .3815648    14.69   0.000     4.858867    6.354573
     lag_apm |   1.023284   .2845275     3.60   0.000     .4656207    1.580948
       _cons |   1.028574   .0797296    12.90   0.000     .8723066    1.184841
-------------+----------------------------------------------------------------
_tlog_pbt    |
    pres_pbt |   2.974831    .286753    10.37   0.000     2.412805    3.536856
     lag_pbt |    .493751    .259587     1.90   0.057    -.0150302    1.002532
       _cons |   1.529438   .0738244    20.72   0.000     1.384745    1.674131
-------------+----------------------------------------------------------------
_tlog_otr    |
    pres_otr |   55.33552   3.462706    15.98   0.000     48.54875     62.1223
     lag_otr |  -4.642008   1.094761    -4.24   0.000      -6.7877   -2.496316
       _cons |   .4091474   .0480197     8.52   0.000     .3150305    .5032644
------------------------------------------------------------------------------

Simulating main parameters.  Please wait....
% of simulations completed: 8% 16% 25% 33% 41% 50% 58% 66% 75% 83% 91% 100%
```

```
Simulating Sigma matrix.  Please wait....
% of simulations completed: 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Number of simulations  : 500
Names of new variables : _b_1 _b_2 _b_3 _b_4 _b_5 _b_6 _b_7 _b_8 _b_9 _b_10 _b_
> 11 _b_12 _b_13 _b_14 _b_15 _b_16 _b_17 _b_18 _b_19 _b_20 _b_21 _b_22
```

`estimadip` utiliza como insumo las variables dependientes en su escala original (proporciones de votos) y genera un nuevo grupo de variables con la transformación logística multivariada. La información de este procedimiento se muestra en la parte superior de los resultados. Luego, el comando estima el modelo SUR y simula los parámetros del modelo (los resultados son idénticos a los de la sección anterior, y se omiten por razones de espacio). Además de facilitar la tarea de generar las variables dependientes transformadas, la sintaxis de `estimadip` es similar a la del comando `sureg`, por lo que es muy sencillo especificar el modelo cuando las variables explicativas son las mismas en todas las ecuaciones. Por ejemplo, la instrucción

```
estimadip (depvar1-depvar3 = indepvar1-indepvar5), base(depvar4)
```

es equivalente a:

```
estimadip (depvar1 indepvar1 indepvar2 indepvar3 indepvar4 indepvar5) ///
          (depvar2 indepvar1 indepvar2 indepvar3 indepvar4 indepvar5) ///
          (depvar3 indepvar1 indepvar2 indepvar3 indepvar4 indepvar5), base(depvar4)
```

Luego, las elecciones hipotéticas se guardan en un archivo con el comando `simuladip`:

```
. simuladip, ev seed(12345) saving(sims, replace)         ///
>         cmd(    replace pres_pan  = pres_pan  +.10,     ///
>                 replace pres_apm  = pres_apm  -.05,     ///
>                 replace pres_pbt  = pres_pbt  -.05,     ///
>         )
(running simuladip)
Districts (300)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................    100
..................................................    150
..................................................    200
..................................................    250
..................................................    300
(note: file sims.dta not found)
file sims.dta saved
Average district vote (Expected values)
```

Party	Sims	Mean	StdErr	[95%Conf	Interval]
party1	500	.4347215	.0065488	.4220469	.4479216
party2	500	.2440962	.0054142	.2333473	.2543551
party3	500	.2586628	.0040989	.2507176	.266755
party4	500	.0441501	.0011307	.0420591	.0463996
base	500	.0183694	.0005979	.0172515	.0195209

La opción `ev` indica a Stata que la distribución del porcentaje de votos se compone por valores esperados. Como es usual, la opción `seed` fija el arranque aleatorio del procedimiento de simulación. Por su parte, la opción `saving()` indica el nombre del archivo en el que se desea guardar los resultados (la subopción `replace` sobreescribe el archivo si éste ya existía previamente). Finalmente, la opción `cmd` sirve para fijar los valores de las variables dependientes de la misma manera que cuando utilizamos el comando `predict` para calcular la asignación de asientos bajo el escenario hipotético.

La opción `cmd` puede contener a su vez los comandos replace y recode. Cada uno de estos comandos debe estar separado por una coma, por lo que `cmd` no acepta subopciones. Sin embargo, `cmd` permite restringir las observaciones con los cualificadores `if` e `in` de la manera usual. Por ejemplo, la instrucción

```
simuladip in 1/100, cmd(replace indepvar=0 if indepvar==2)
```

calcula la distribución de los valores esperados para los primeros cien distritos, y reemplaza `indepvar` con el valor 0 cuando su valor original es 2. Si no se especifica la opción `cmd, simuladip` calcula los resultados con el valor real u observado de las variables explicativas.

El resultado que `simuladip` despliega consiste en el voto distrital promedio. Por ejemplo, los resultados indican que si el voto del candidato presidencial del PAN aumentara en 10 puntos, la votación distrital promedio de dicho partido para la Cámara de Diputados sería de entre 43.7% y 45.6% con un 95% de confianza estadística. Nótese que la media de las distribuciones de votos es muy similar a los puntos estimados por el comando `predict`, y que se muestra en los resultados del comando `asignadip` en la sección 3.

El siguiente paso es simular la distribución de asientos para cada resultado electoral hipotético. Una diferencia evidente entre Clarify y **camaradip** tiene que ver con la naturaleza de las cantidades de interés. La última fila de la Tabla 2 representa el número de asientos del partido j, denotado por ψ_j para el resultado observado, y por $\psi_j^{(hip)t}$ para las replicaciones hipotéticas. Las cantidades de interés se obtienen con la información de su respectiva columna. Para calcular la primera replicación $\tilde{\psi}_j^{(hip)1}$ debemos contar el número de distritos en que el partido j obtuvo más votos que los demás partidos en la primera columna de replicaciones hipotéticas (lo cual representa su número de asientos uninominales), y luego asignar las asientos plurinominales a las que tendría derecho dada su votación en todos los distritos de esa misma columna. Este ejercicio se repite sucesivamente para cada una de las elecciones hipotéticas de manera que la última fila de la tabla está formada por T simulaciones de la distribución de asientos $\tilde{\psi}_j^{(hip)}$.

Naturalmente, entre mayor sea el número de simulaciones T, la tarea de obtener $\tilde{\psi}^{(hip)}$ se torna sumamente difícil. Una de las virtudes de `asignadip` es que el usuario puede acceder a los resultados por medio de matrices y macros a través de `r()`. Esta utilería es una manera práctica de guardar las distribuciones del número de asientos en variables que pueden manipularse fácilmente con otros comandos de estadística descriptiva. Por ejemplo, para generar la distribución del número de asientos por partido utilizando los resultados de la base de datos `sims`, podemos combinar los comandos `asignadip` y `postfile`:

```
. use sims, clear
(simuladip: Expected values)
. tempname memhold
. postfile `memhold´ pan apm pbt na asdc using results, replace
(note: file results.dta not found)
. forvalues j = 1/500 {
  2.         qui asignadip party* base if _j == `j´
  3.         post `memhold´ `r(nTotal)´
  4. }
```

```
. postclose `memhold´
```

`postfile` indica a Stata que deseamos generar una base de datos llamada `results` que contiene las variables `pan`, `pri`, `prd`, `na` y `base`. El contenido de esas variables se llenará con los resultados de `asignadip` para cada una de las 500 elecciones hipotéticas (cuyo identificador en la base `sims` es la variable `_j`). `asignadip` calcula la distribución de asientos en cada ciclo creado por `forvalues`, y guarda los resultados en un macro llamado `r(nTotal)`. El comando `post` indica a Stata que las variables de la base `results` deben agregarse con el contenido de ese macro. De esta manera, la base `results` contiene 500 asignaciones de asientos al final del procedimiento, una para cada elección hipotética.

Para presentar los resultados del análisis, las distribuciones pueden describirse con medidas de tendencia central (e.g., media, moda, mediana) y medidas de dispersión (e.g., desviaciones estándar o percentiles que delimitan un intervalo de confianza). Con este propósito, el paquete **camaradip** incluye el comando `sumseats`, que proporciona estadística descriptiva básica (media, moda, error estándar e intervalos de confianza) de las distribuciones de asientos.

```
. use results, clear
. sumseats pan-asdc
```

Party	Sims	Mean	StdErr	Median	[95%Conf	Interval]
pan	500	260.914	3.337196	261	255	268
apm	500	104.9	4.629997	105	97	114
pbt	500	124.118	4.136944	124	116	131
na	500	10.05	.4096775	10	9	11
asdc	500	.018	.2860764	0	0	0

Así pues, los resultados sugieren que si el voto del candidato presidencial del PAN aumentara en 10 puntos, el número de asientos que su partido obtendría en la Cámara sería de entre 263 y 272 con un 95% de confianza estadística. Nuevamente, nótese que la media y mediana de las distribuciones de asientos es muy similar a los puntos estimados del comando `asignadip` en la sección 3.

5 Conclusiones

En este capítulo se han presentado diferentes formas de adecuar los modelos de regresión convencionales para estimar resultados electorales en sistemas multipartidistas y con sistemas electorales mixtos, como es el caso de México y otros países de América Latina. En primer lugar contrastamos la estimación por mínimos cuadrados ordinarios con la estimación basada en transformación de las proporciones de votos en *log ratios*, por un lado, y con los modelos SURE, por otro. Para ilustrar estas técnicas, se utilizó como ejemplo la elección para diputados federales de 2006 en México y su relación con los votos para presidente del mismo año.

En segundo lugar, se presentaron algunos ejemplos y técnicas para extraer *cantidades de interés* sustantivo a partir de los resultados de modelos de regresión múltiple. En nuestro ejemplo particular, se estimó la composición de la Cámara de Diputados a partir de los resultados electorales observados o bien de resultados hipotéticos. Por último,

discutimos las técnicas de simulación estadística para calcular medidas de incertidumbre para las cantidades de interés obtenidas a partir de modelos de regresión. A lo largo del capítulo, explicamos el uso del módulo **camaradip** de Stata para simplificar algunos de estos métodos(Marquez & Aparicio, 2010).

Algunas extensiones y aplicaciones de los métodos aquí discutidos son los siguientes. Una vez que se estima la composición del Congreso, también pueden realizarse pruebas de hipótesis para la probabilidad de que cierta bancada sea mayor o menor a una cifra cualquiera. Por ejemplo, para calcular la probabilidad de que un partido tenga mayoría simple en la Cámara, basta con sumar las veces que el número predicho de asientos es igual o mayor a 251, y dividirlo entre T. Del mismo modo, también se puede calcular la probablidad de que un partido tenga al menos un tercio de la cámara (lo que le daría poder de veto para reformas constitucionales) o bien si una coalición de partidos lograría tener mayoría calificada.

Además, la distribución del número de asientos puede tranformarse en otras distribuciones o cantidades de interés, como puede ser el número efectivo de partidos en la legislatura, índices de poder, las diferentes coaliciones mínimas ganadoras para una reforma legal o constitucional, etc. El marco en que se desarrolla el modelo también puede adaptarse para analizar otras consecuencias de los resultados electorales, más allá de la conformación propia del Congreso, como pueden ser la identificación de bancadas pivotales, o la asignación del financiamiento público y el acceso a medios para los partidos políticos. Esto, toda vez que desde un punto de vista estadístico, no existen grandes diferencias entre analizar cantidades de interés tales como el número de curules, el tamaño relativo de las bancadas, la probabilidad de ganar o perder un distrito, o incluso estimar las prerrogativas que corresponderían a cada partido de acuerdo los resultado electorales.

El el ámbito metodológico, el modelo estadístico también puede complementarse con otros métodos de estimación de los resultados electorales distintos al desarrollado en este artículo. Por ejemplo, las simulaciones pueden ser obtenidas con métodos bayesianos o de cadenas markovianas (MCMC), adaptarse para bases de datos longitudinales (TSCS, por sus siglas en inglés), o bien extenderse para la especificación de modelos jerárquicos que combinen datos con diferentes niveles de agregación.

References

Aitchison, J. & J. Egozcue, J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37(7), 829–850.

Balinski, M. & Ramírez González, V. (1999). Mexico's 1997 apportionment defies its electoral law. *Electoral Studies*, 18(1), 117–124.

Baum, C. (2008). Stata tip 63: Modeling proportions. *Stata Journal*, 8(2), 299–303.

Berry, W. (1993). *Understanding regression assumptions*, volume 92. Sage Publications, Inc.

Borisyuk, G., Johnston, R., Thrasher, M., & Rallings, C. (2008). Measuring bias: Moving from two-party to three-party elections. *Electoral Studies*, 27(2), 245–256.

Breusch, T. & Pagan, A. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, (pp. 1287–1294).

Breusch, T., Robertson, J., & Welsh, A. (1997). The emperor's new clothes: a critique of the multivariate t regression model. *Statistica Neerlandica*, 51(3), 269–286.

Buis, M., Cox, N., & Jenkins, S. (2006). Dirifit: Stata module to fit a dirichlet distribution. *Statistical Software Components*.

Campbell, J. (1986). Predicting seat gains from presidential coattails. *American Journal of Political Science*, (pp. 165–183).

Cox, N. (2002). Speaking stata: How to face lists with fortitude. *Stata Journal*, 2(2), 202–222.

Ferejohn, J. & Calvert, R. (1984). Presidential coattails in historical perspective. *American Journal of Political Science*, (pp. 127–146).

Gelman, A. & King, G. (1990). Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association*, (pp. 274–282).

Gelman, A. & King, G. (1994). A unified method of evaluating electoral systems and redistricting plans. *American Journal of Political Science*, (pp. 514–554).

Gelman, A., King, G., & Thomas, A. (2007). Judgeit ii: A program for evaluating electoral systems and redistricting plans. *URL http://gking. harvard. edu/judgeit*.

Golder, M. (2005). Democratic electoral systems around the world, 1946-2000* 1. *Electoral Studies*, 24(1), 103–121.

Honaker, J., Katz, J., & King, G. (2002). A fast, easy, and efficient estimator for multiparty electoral data. *Political Analysis*, 10(1), 84.

Jackman, S. (2009). *Bayesian analysis for the social sciences*, volume 846. Wiley.

Jackson, J. (2002). A seemingly unrelated regression model for analyzing multiparty elections. *Political Analysis*, 10(1), 49.

Katz, J. & King, G. (1999). A statistical model for multiparty electoral data. *American Political Science Review*, (pp. 15–32).

King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Univ of Michigan Pr.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, (pp. 347–361).

Marchenko, Y. & Genton, M. (2010). A suite of commands for fitting the skew-normal and skew-t models. *Stata Journal*, 10(4), 507–539.

Marquez, J. & Aparicio, F. (2010). Un modelo monte carlo para la camara de diputados en mexico. *Politica y Gobierno*, 17(2).

Molinar Horcasitas, J. & Weldon, J. (2001). Reforming electoral systems in mexico. *Mixed-member electoral systems: The best of both worlds*, (pp. 209–230).

Tomz, M., Tucker, J., & Wittenberg, J. (2002). An easy and accurate regression model for multiparty electoral data. *Political Analysis*, 10(1), 66.

Tomz, M., Wittenberg, J., & King, G. (2003). Clarify: Software for interpreting and presenting statistical results. *Journal of Statistical Software*, 8(1), 1–30.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, (pp. 348–368).