

Consulte los debates, las estadísticas y los perfiles de los autores de esta publicación en: <https://www.researchgate.net/publication/228213305>

## El "gráfico de separación": Un nuevo método visual para evaluar el poder predictivo de los modelos Logit/Probit

### Artículo

CITACIÓN

1

LECTURAS

1,684

3 autores, entre ellos:



[Brian Greenhill](#)

Universidad de Albany, Universidad Estatal de Nueva York

27 PUBLICACIONES 998 CITAS

VER PERFIL



[Michael Ward](#)

Universidad de Duke

120 PUBLICACIONES 4.967 CITAS

VER PERFIL

Todo el contenido que sigue a esta página fue subido por [Michael Ward](#) el 10 de enero de 2014.

El usuario ha solicitado la mejora del archivo descargado.

# UN NUEVO MÉTODO VISUAL PARA EVALUAR EL PODER PREDICTIVO DE LOS MODELOS BINARIOS

BRIAN D. GREENHILL, MICHAEL D. WARD Y AUDREY SACKS

Resumen. Presentamos un nuevo método visual para evaluar el poder predictivo de los modelos con resultados binarios. Esta técnica permite al analista elegir rápida y fácilmente entre las especificaciones de modelos alternativos, basándose en la capacidad de los modelos para hacer coincidir sistemáticamente las predicciones de alta probabilidad con los sucesos reales del evento de interés, y las predicciones de baja probabilidad con los no sucesos del evento de interés. A diferencia de los métodos existentes para evaluar el poder de predicción de los modelos logit y probit, como el uso de estadísticas de "porcentaje de predicción correcta", las puntuaciones de Brier y el gráfico ROC, nuestro "gráfico de separación" tiene la ventaja de producir una presentación visual que es más informativa y más fácil de explicar a un público general que un gráfico ROC, a la vez que sigue siendo insensible a la elección a menudo arbitraria del usuario del umbral para distinguir entre eventos y no eventos. Mostramos cómo implementar esta técnica en R y demostramos su eficacia en la construcción de modelos predictivos en cuatro áreas diferentes de la investigación política.

## 1. Estado del arte

Los datos binarios están muy extendidos en la ciencia política, y los politólogos han contribuido enormemente a los métodos de estudio sistemático de las variables dicotómicas. Aldrich y Nelson (1984) ofrecieron una introducción didáctica a la regresión logit y probit. La ciencia política se decantó por el logit, frente al probit, en gran medida por la disponibilidad de programas informáticos para el primero, y eso se ha convertido en la norma de facto. Sin embargo, hasta hace poco, los resultados de estos modelos de regresión binarios y discretos se presentaban principalmente como tablas de coeficientes y medidas de precisión asociadas. Dado que estos resultados numéricos son tan difíciles de interpretar y tan propensos a la mala interpretación, King, Wittenberg y Tomz (2000) introdujeron una forma más gráfica de presentar los resultados en términos de cálculo de los valores esperados condicionados por el modelo estimado. Este artículo se tituló acertadamente "¡Basta ya de coeficientes Logit!", pero ese título no sobrevivió, lamentablemente, al escrutinio del proceso editorial. Las ideas básicas presentadas en él sirven de base para el enfoque general de presentación de resultados que se encuentra en el paquete de software Omnibus conocido como *Zelig*.

A pesar de estas mejoras, se ha prestado muy poca atención a la noción de ajuste del modelo desde una perspectiva más moderna y visual. Hasta hace poco, la mayoría de los problemas de inferencia en ciencia política eran totalmente teóricos y retrospectivos: mirar las estadísticas

---

Preparado para su presentación en la Reunión Anual 2009 de la Asociación Americana de Ciencias Políticas, celebrada del 3 al 6 de septiembre en Toronto, Canadá. Agradecemos los comentarios de los colegas asociados al proyecto ICEWS, especialmente, Sean O'Brien, Philippe Loustanau y Laura Stuart. Los colegas de la Conferencia del 10º Aniversario de la Estadística y las Ciencias Sociales, celebrada los días 4 y 5 de junio de 2009 en la Universidad de Washington, Seattle, WA, también proporcionaron útiles aportaciones,

especialmente Andrew Gelman y Steve Fienberg. También presentamos una versión de estas ideas en la Conferencia de Verano sobre Metodología Política de 2009, celebrada del 23 al 25 de julio en la Universidad de Yale, y recibimos valiosas reacciones de diversos colegas. A pesar de todos estos buenos consejos, seguimos siendo responsables de los resultados aquí presentados.

significación y mirando a los datos observados. Sin embargo, hoy en día el interés por la validación cruzada y la predicción fuera de la muestra es mucho mayor que antes. Como resultado, a los incondicionales de la perspectiva clásica que querían "alguna medida de ajuste" se ha unido una amplia variedad de otros intereses que requieren mayor información sobre la calidad de los modelos empíricos que la que proporcionan las tablas de números que representan las medias y las varianzas de los parámetros estimados. Ahora hay un mayor interés por los gráficos ROC y las tablas de especificidad y sensibilidad como forma de calibrar la validez del modelo estimado, en términos empíricos. Cuando se trata de evaluar el poder predictivo de los modelos logit o probit, en última instancia se quiere poder distinguir los buenos modelos de los malos sobre la base de su capacidad para generar predicciones "correctas". Sin embargo, la dificultad reside en distinguir las predicciones "correctas" de las "incorrectas": los modelos logit y probit generan valores ajustados que se sitúan en algún punto de una escala continua de 0 - 1 (por ejemplo,  $\hat{y} = 0,68$ ) mientras que los valores reales de la variable dependiente son dicotómicos (para todas las observaciones,  $y = 1$  o  $y = 0$ ). Para ponerlo en términos más concretos, si uno tiene un modelo logit de los resultados de las elecciones presidenciales de EE.UU. que generó probabilidades predichas de 0,68 y 0,32 para las victorias de Obama y McCain en el 2008, ¿qué tan "correcto" fue el modelo? En este artículo presentamos un nuevo método visual -que denominamos "gráfico de separación"- para abordar este problema. Sin embargo, primero repasaremos brevemente las soluciones que se utilizan con mayor frecuencia para evaluar el poder predictivo de los modelos logit o probit.

1.1. **Opción 1: Dicotomizar todo.** Consideremos los datos hipotéticos sobre la guerra y la paz para una muestra de seis países que se muestran en la Tabla 1. Nuestra variable dependiente y se codifica de tal manera que a cada caso de guerra se le asigna el valor 1 y a cada caso de paz el valor 0. Nuestra variable dependiente,  $y$ , se codifica de forma que a cada caso de guerra se le asigna un valor de 1 y a cada caso de paz un valor de 0. Los valores ajustados,  $\hat{y}$ , obtenidos de nuestro modelo logit o probit se muestran en la tercera columna.

Tabla 1. Datos de la muestra

País	Resultado real ( $y$ )	Valor ajustado ( $\hat{y}$ )
A	0	0.774
B	0	0.364
C	1	0.997
D	0	0.728
E	1	0.961
F	1	0.422

Para determinar la frecuencia con la que el modelo hace las predicciones "correctas", podemos simplemente dicotomizar los valores ajustados. Por ejemplo, podemos establecer una regla según la cual cada valor ajustado inferior a, digamos, 0,5 se considera un caso de paz, mientras que cada valor superior o igual a 0,5 se considera un caso de guerra. En este caso, podemos generar la siguiente tabla de contingencia:

	Guerra prevista	Paz prevista
Guerra real {C, E}	{F}	
Paz real {A, D}	{B}	

Encontramos que tres países fueron predichos correctamente - C, E y B - mientras que otros tres fueron predichos incorrectamente - A, D y F. Pero estos resultados, por supuesto, dependen del

0,5 es un punto de corte totalmente arbitrario y si se eligiera un punto de corte de, por ejemplo, 0,3, los resultados serían bastante diferentes:

	Guerra prevista	Paz prevista
Guerra real{C , E, F}{}		
Paz real{A , B, D}{}		

En este caso somos capaces de predecir la guerra en tres de los casos en los que realmente hubo una guerra

- C, E y F - pero ahora no predecimos ningún caso de paz. En cambio, predecimos incorrectamente la guerra en los tres casos reales de paz: A, B y D. Al bajar el umbral de 0,5 a 0,3, hemos aumentado el número de verdaderos positivos a costa de un mayor número de falsos positivos.

Aunque estas tablas de 2x2 son fáciles de entender para los lectores, el problema es que los resultados dependen totalmente de la elección (a menudo arbitraria) del autor del umbral.<sup>1</sup>

**1.2. Opción 2: La curva ROC.** Las curvas ROC (Receiver Operating Characteristic) proporcionan un resumen visual del equilibrio entre falsos positivos y falsos negativos a medida que se varía el umbral. Normalmente se presentan en forma de un gráfico de la tasa de falsos positivos frente a la tasa de verdaderos positivos obtenida para cada umbral posible  $\tau$ . La tasa de falsos positivos (FPR) se define como el número de falsos positivos dividido por la suma de los falsos positivos y los verdaderos negativos (en otras palabras, todos los negativos incorrectamente identificados divididos por todos los negativos reales), mientras que la tasa de verdaderos positivos (TPR) se define como el número de verdaderos positivos dividido por la suma de los verdaderos positivos y los falsos negativos (de nuevo, todos los positivos correctamente identificados divididos por todos los positivos reales).<sup>2</sup>

La curva ROC para nuestro conjunto de datos de seis filas sobre la guerra y la paz se muestra en la Figura 1 a continuación. En este caso, los seis valores discretos de  $\tau$  dan lugar a siete combinaciones posibles de valores de FPR/TPR y, por tanto, a siete puntos de la curva. El cálculo de los siete puntos de la curva se muestra en la Tabla 2.

Tabla 2. Cálculo de los puntos de la curva ROC.

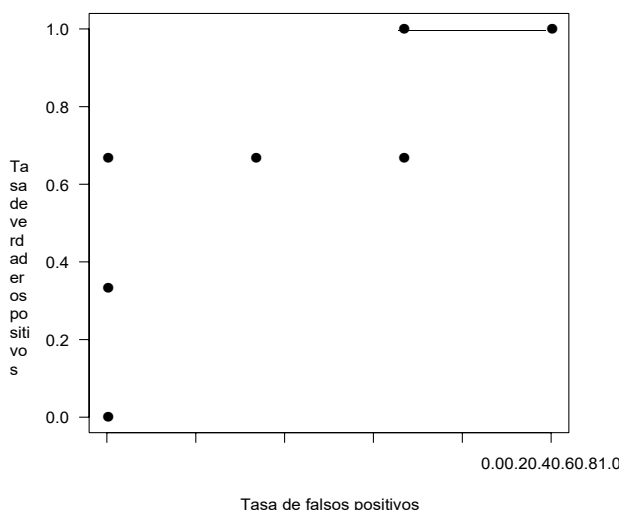
Umbral	TP	FP	FN	TN	FPR	TPR	$\frac{FP}{FP+TN}$	$\frac{TP}{TP+FN}$
$0 < \tau < 0,364$	{C , E, F}	{A, B, D}	{}	{}	$\frac{0}{0+0} = 0$	$\frac{3}{3+0} = 1$	0	1
$0,364 < \tau < 0,422$	{C, E, F}	{A, D}	{B}	{}	$\frac{2}{2+0} = 1$	$\frac{2}{2+0} = 1$	1	1
$0,422 < \tau < 0,728$	{C , E}	{A, D}	{F}	{B}	$\frac{2}{2+1} = 0,67$	$\frac{2}{2+0} = 1$	0,67	1
$0,728 < \tau < 0,774$	{C , E}	{A}	{F}	{B, D}	$\frac{1}{1+1} = 0,5$	$\frac{1}{1+0} = 1$	0,5	1
$0,774 < \tau < 0,961$	{C , E}	{}	{F}	{A, B, D}	$\frac{0}{0+3} = 0$	$\frac{0}{0+2} = 0$	0	0
$0,961 < \tau < 0,997$	{C}	{E, F}	{A, B, D}	{}	$\frac{3}{3+0} = 1$	$\frac{0}{0+0}$	1	0
$0,997 < \tau < 1$	{}	{C , E, F}	{A, B, D}	{}	$\frac{0}{0+3} = 0$	$\frac{0}{0+0}$	0	0

<sup>1</sup>Puede haber algunas aplicaciones en las que el umbral no sea arbitrario. Por ejemplo, si decidiéramos que cuando se trata de predecir guerras, un falso negativo es diez veces más costoso que un falso positivo, entonces un umbral de 0,1 podría considerarse apropiado.

La tasa de verdaderos positivos también se denomina *sensibilidad* de un clasificador. La tasa de falsos positivos equivale a  $1 - \text{specificidad}$ , donde la especificidad se define como  $\frac{TN}{TN + FP}$ .



Figura 1. Gráfico ROC para los datos mostrados en la Tabla 1.



Las curvas ROC tienen la ventaja de proporcionar una descripción visual del poder predictivo del modelo en todos los umbrales posibles. Los modelos con altos niveles de poder predictivo tenderán a tener tasas de verdaderos positivos que son consistentemente más altas que las correspondientes tasas de falsos positivos, dando lugar a curvas que tienen la apariencia de estar tiradas hacia la esquina superior izquierda del gráfico. En consecuencia, el poder predictivo global del modelo (en todos los umbrales posibles) puede resumirse en términos del área bajo la curva ROC, ya que ésta se define en el cuadrado de la unidad (la llamada "puntuación AUC").<sup>3</sup>

### 1.3. Opción 3: Otras estadísticas resumidas.

1.3.1. *Puntuaciones de Brier*. Una opción habitual es la puntuación de Brier, que es el valor medio de la diferencia al cuadrado entre los valores ajustados y reales de la variable dependiente. La puntuación de Brier se desarrolló a principios de la década de 1950 para proporcionar una forma de calificar las previsiones meteorológicas probabilísticas. La fórmula de esta métrica es bastante sencilla:

$$B = (p - X)^2$$

donde  $p$  es la previsión probabilística y  $X$  es la variable dicotómica y binaria que indica si el evento previsto se ha producido o no (1=sí; 0=no). Cuanto más se acerque a cero la puntuación de Brier, mejor será la previsión. Cuando se realizan muchas previsiones, como por ejemplo en un único modelo aplicado a muchos casos, se suele comunicar la puntuación media de Brier.

En este caso, el valor medio de la puntuación Brier en las seis observaciones es de 0,266.

<sup>3</sup>Las puntuaciones de AUC están limitadas entre 0 y 1. Un modelo que no es mejor que un simple lanzamiento

de moneda debería tener un AUC de 0,5; ocasionalmente se observan puntuaciones más bajas cuando, en promedio, la tasa de falsos positivos *supera* la tasa de verdaderos positivos.

Tabla 3. Cálculo de las puntuaciones Brier.

País	Resultado real ( $y$ )	Valor ajustado ( $\hat{p}$ )	Puntuación de Brier ( $(\hat{p} - p)^2$ )
A	0	0.774	0.599
B	0	0.364	0.132
C	1	0.997	0.000
D	0	0.728	0.530
E	1	0.961	0.002
F	1	0.422	0.334

1.3.2. *Pseudo  $R^2$* . A veces los académicos (y los programas informáticos) informan de una *Pseudo  $R^2$* , que suele ser uno menos la relación entre la probabilidad de un modelo nulo y la probabilidad del modelo estimado, de modo que si el modelo nulo y el modelo estimado tienen aproximadamente la misma probabilidad, la pseudo puntuación es cercana a cero. Al igual que la  $R^2$ , esta medida tiene muchos defectos como medida de ajuste.

Para ayudar a abordar la cuestión de proporcionar un criterio visual matizado para el rendimiento de tales modelos, desarrollamos en la siguiente sección un nuevo enfoque, denominado gráfico de separación.

## 2. La trama de la separación

2.1. **El concepto.** El método alternativo de evaluación del poder predictivo que presentamos aquí consiste en una simple reordenación de los datos presentados en la Tabla 1, de manera que los valores ajustados se presentan en orden ascendente. A continuación, anotamos si cada uno de ellos corresponde a un caso real del acontecimiento (la guerra) o a un no acontecimiento (la paz).

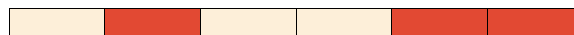
Tabla 4. Reorganización de los datos presentados en la Tabla 1 para su uso en el gráfico de separación.

País	Valor ajustado ( $\hat{y}$ )	Resultado real ( $y$ )
B	0.364	0
F	0.422	1
D	0.728	0
A	0.774	0
E	0.961	1
C	0.997	1

El poder de predicción puede evaluarse ahora simplemente midiendo hasta qué punto los casos reales del evento se concentran en el extremo inferior de la tabla, y los no eventos en el extremo superior de la misma. Un modelo sin capacidad de predicción, es decir, cuyos resultados pueden aproximarse a los de una moneda al azar, generaría una distribución uniforme de 0 y 1 en la columna de la derecha. En cambio, un modelo con un poder de predicción perfecto produciría una separación completa de los 0 y los 1 en la columna de la

derecha: los valores ajustados bajos siempre resultarían estar asociados a casos reales de

Figura 2. Diagrama de separación que representa los datos presentados en la Tabla 1.



paz (0s), mientras que los valores ajustados altos se asociarían siempre con casos reales de guerra (1s).

Resulta muy fácil discernir estas diferencias mediante una representación gráfica sencilla que denominamos "gráfico de separación" (véase la figura 2). En este gráfico, los paneles oscuros y claros corresponden a las instancias reales de los sucesos y los no sucesos, respectivamente, ordenados de forma que los valores  $\hat{y}$  correspondientes aumentan de izquierda a derecha.

Como muestra el gráfico, nuestro "modelo" describe razonablemente bien los datos. Evidentemente, un modelo perfecto produciría un gráfico en el que todos los sucesos se agrupan en el polo derecho y todos los no sucesos en el polo izquierdo, es decir, un modelo completamente ineficaz que no muestre tal separación podría tener un aspecto más parecido.<sup>4</sup>

**2.2. Ajuste del modelo.** La principal ventaja del gráfico de separación reside en su capacidad para proporcionar una descripción visual clara del ajuste de un modelo (ya sea dentro o fuera de la muestra). Esto puede ser especialmente útil en la fase de selección del modelo, en la que los resúmenes de un solo número del poder predictivo pueden carecer de matices y las curvas ROC son más difíciles de comparar.

Consideremos el siguiente ejemplo de un intento de construir un modelo que describa la incidencia de las insurgencias políticas entre un grupo de 29 países de la región de Asia-Pacífico durante el periodo 1998-2004 ( $n=812$ ). El proyecto y los datos se describen con más detalle en O'Brien (2010, de próxima publicación). El gráfico 3 muestra los resultados de la utilización de gráficos de separación en las sucesivas etapas del proceso de selección del modelo. En el primer gráfico, se muestran los resultados del ajuste de un modelo que tiene una intercepción y ninguna covariable. En este caso, los valores ajustados son idénticos para todas las observaciones del conjunto de datos y son simplemente la media de la variable insurgencia. El modelo hace claramente un mal trabajo al separar los eventos de los no eventos, y la línea negra que representa los valores correspondientes de  $\hat{y}$  permanece en un nivel bajo constante en todas las observaciones.

En el segundo gráfico añadimos una única covariable, el PIB per cápita (retrasado un año). Esto mejora claramente la capacidad del modelo para distinguir los eventos de los que no lo son. Ahora vemos tres grupos distintos de eventos dispuestos hacia el lado derecho del espacio del gráfico (correspondiente a los valores más altos de  $\hat{y}$ ). Mientras tanto, no se detectan eventos en las regiones correspondientes al 40% inferior de las observaciones.

El tercer gráfico muestra los resultados de añadir una segunda covariable, la anocracia, que puede considerarse un indicador del grado de inestabilidad de las instituciones políticas de un Estado (Marshall y Jaggers, 2003). La adición de esta variable mejora claramente la capacidad de predicción del modelo: los grupos de líneas rojas que indican casos reales del acontecimiento se desplazan más a la derecha, y no aparece ningún acontecimiento a la izquierda del acontecimiento de insurgencia más a la izquierda en el gráfico anterior.

Por último, el gráfico de la parte inferior de la Figura 3 muestra el efecto de la inclusión

de otras dos variables que representan datos detallados del flujo de eventos sobre el número de casos de hostilidad entre el gobierno y los grupos insurgentes en el periodo anterior. En este caso

---


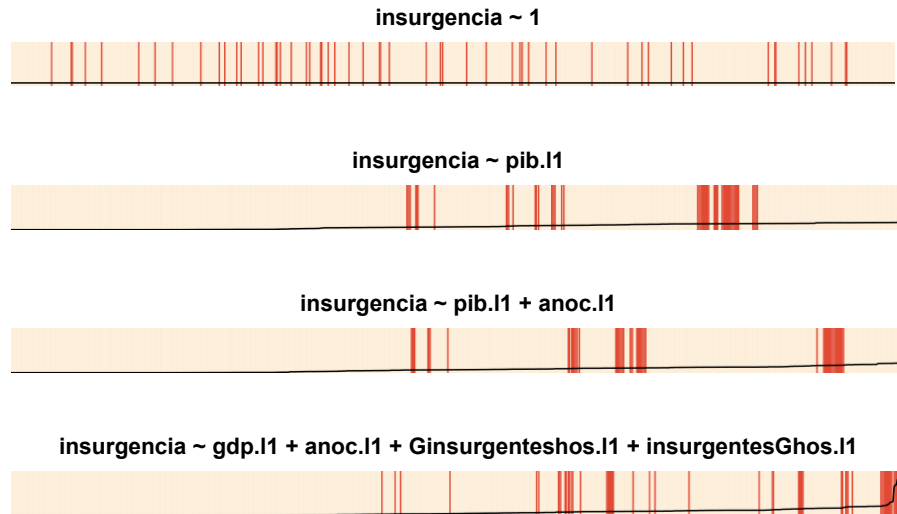
<sup>4</sup>Peor aún, un modelo que constantemente hace predicciones *erróneas* produciría un gráfico parecido a  - es decir, uno en el que los eventos y los no eventos se han ordenado en la dirección equivocada.

Figura 3. Gráficos de separación utilizados en el desarrollo de un modelo de insurgencia en la región de Asia-Pacífico, 1998-2004. El texto sobre cada gráfico de separación muestra las variables incluidas en el modelo.



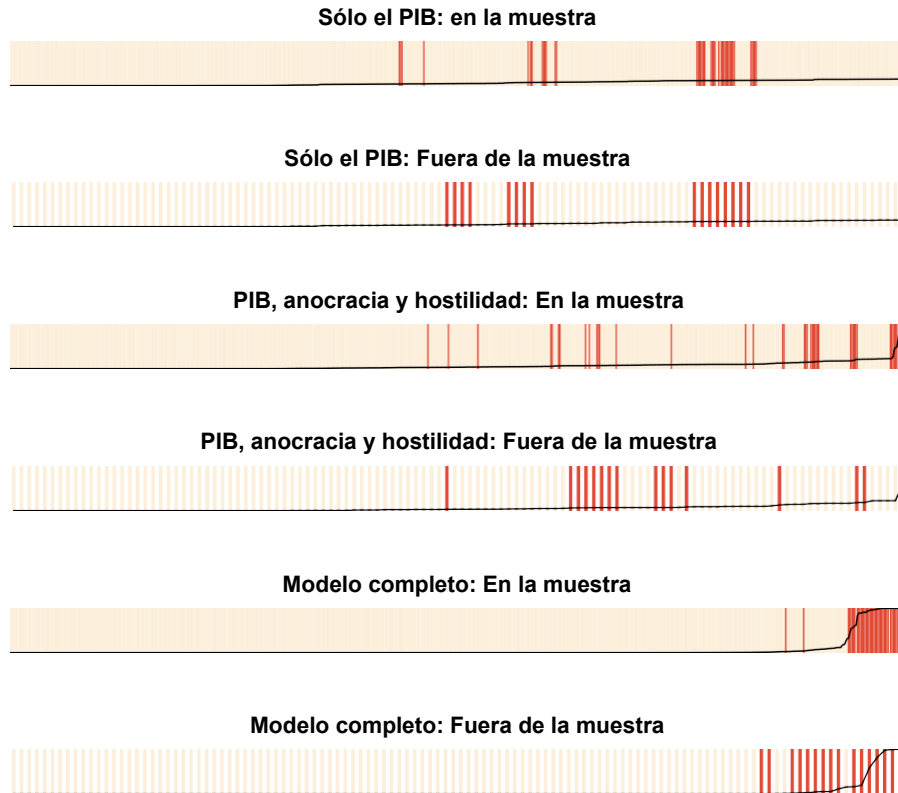
vemos una agrupación distintiva de observaciones correspondientes a eventos de insurgencia en el extremo derecho del gráfico, y también podemos ver que estos están asociados con valores significativamente más altos de  $\hat{y}$  (como se muestra por la posición vertical de la línea negra).

Estos resultados muestran cómo los gráficos de separación pueden utilizarse para obtener una comprensión más detallada de la forma en que el poder predictivo del modelo mejora a medida que se cambia la especificación del modelo. Su principal ventaja frente a una estadística única como el área bajo la curva ROC es que permite al usuario considerar las ganancias y pérdidas de poder predictivo en diferentes regiones de los datos. Por ejemplo, puede darse el caso de que el usuario esté especialmente interesado en desarrollar un modelo que maximice el número de eventos correctamente predichos entre los valores más altos de  $\hat{y}$ , en el que podría centrarse sólo en la región del extremo derecho ignorando otras partes del gráfico. Por otro lado, algunas aplicaciones pueden requerir que el usuario se centre en minimizar el número de eventos entre los valores más bajos de  $\hat{y}$ , en cuyo caso el tercer modelo de la figura 3 se consideraría realmente superior al cuarto.<sup>5</sup> Por supuesto, el gráfico de separación no sólo tiene que utilizarse para medir el ajuste en la muestra. También puede utilizarse para evaluar la capacidad de predicción fuera de la muestra en ejercicios de validación cruzada. Esto es especialmente útil cuando el usuario está interesado en maximizar el éxito predictivo dentro de una determinada fase de "prueba". La Figura 4 muestra tres pares de gráficos de separación obtenidos para la evaluación dentro y fuera de la muestra de los modelos de insurgencia en la región de Asia-Pacífico. El gráfico superior de cada par representa el ajuste dentro de la muestra de un modelo ajustado utilizando el periodo 1998-2003 como conjunto de entrenamiento, mientras que el gráfico inferior representa el poder predictivo fuera de la muestra en el periodo de 2004.

<sup>5</sup> En el paquete R para el gráfico de separación, también proporcionamos una facilidad para resaltar las observaciones individuales de interés. Esto permite al usuario ver cómo cambia la probabilidad predicha asignada a uno o más casos críticos bajo diferentes especificaciones del modelo.



Figura 4. Gráficos de separación utilizados en el desarrollo de un modelo de insurgencia en la región de Asia-Pacífico, 1998-2004. El texto sobre cada gráfico de separación muestra las variables incluidas en el modelo.



El primer par de gráficos de la Figura 4 muestra los resultados correspondientes dentro y fuera de la muestra para una regresión de la insurgencia sobre los niveles del PIB per cápita en el período anterior. (Obsérvese que el gráfico dentro de la muestra que se muestra aquí es ligeramente diferente del gráfico correspondiente de la Figura 3 porque éste muestra el rendimiento de un modelo ajustado utilizando sólo el conjunto de entrenamiento de 1998-2003, en lugar del período completo de 1998-2004).

El segundo par de gráficos muestra los resultados de un modelo ligeramente más complejo que incluye el PIB per cápita, la anocracia (una medida de la ausencia de un gobierno efectivo) y las puntuaciones de hostilidad del gobierno hacia los insurgentes y de los insurgentes hacia el gobierno. Aunque el resultado es una clara mejora del ajuste en la muestra, no es obvio que este modelo más complejo funcione mejor en el periodo de prueba de 2004 que el modelo simple de sólo PIB.

Por último, el tercer par de gráficos de la Figura 4 muestra los resultados de un modelo más complejo que incluye las siguientes covariables: PIB per cápita, anocracia, competitividad de la participación política, número de grupos minoritarios, puntuaciones

de hostilidad gobierno-insurgente e insurgente-gobierno, una variable de contador de años y un desfase espacial que refleja los niveles de insurgencia

entre los pares "similares" de cada país en el periodo anterior.<sup>6</sup> Como se desprende de los gráficos de separación, este modelo proporciona un excelente ajuste dentro de la muestra a los datos de 1998-2003 y predice casi perfectamente los casos de insurgencia en el periodo de prueba de 2004.

### 3. Ejemplos

Para ilustrar este planteamiento, veremos ahora cuatro ejemplos procedentes de diversos ámbitos de la disciplina.

#### 3.1. Campañas políticas: Hillygus y Jackman (2003).

Hillygus y Jackman (2003) desarrollan un modelo de intención de voto que muestra no sólo que los acontecimientos de la campaña, como las convenciones de los partidos y los debates presidenciales, provocan cambios en las preferencias de los votantes, sino que el efecto que estos acontecimientos tienen en las preferencias de los votantes varía en función de las preferencias que los votantes expresaron en una encuesta anterior. Interpretan este hallazgo como una prueba de que los votantes asimilan la nueva información sobre los candidatos de forma condicionada a sus preferencias anteriores (Hillygus y Jackman, 2003: 590).

Los autores emplean un modelo logit de preferencia de los votantes. Además de discutir las estimaciones de los coeficientes del modelo, también comentan el ajuste del modelo de la siguiente manera:

"En términos brutos, los modelos de transición estimados se ajustan muy bien a los datos, al igual que cualquier modelo con una variable dependiente retardada en la ecuación. El modelo de convención predice correctamente el 91% de las preferencias de voto, y el modelo de debate predice correctamente el 94% de las preferencias de voto, utilizando  $p = 0,5$  como umbral de clasificación. Además, cada uno de los modelos tiene un área bajo las curvas ROC que oscila entre 0,79 y 0,89. Estas cifras indican que cada uno de nuestros modelos hace un trabajo entre aceptable y excelente a la hora de discriminar a los que hicieron la transición de los que se mantuvieron estables". (Hillygus y Jackman, 2003: 592-593)

Creemos que el gráfico de separación proporciona una herramienta más informativa y compacta para transmitir el mismo punto. Como se muestra en la figura 5, los modelos, tal y como se ilustran en el gráfico de separación, parecen ajustarse perfectamente a los datos.

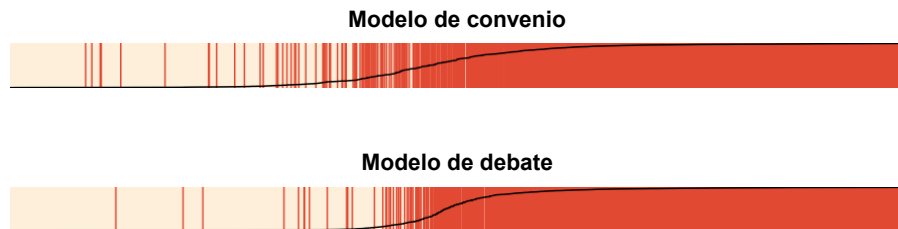
Además, se puede optar por simplificar aún más la presentación mostrando los gráficos de separación como un gráfico en línea similar al concepto de "sparkline" desarrollado por Tufte (2006). De este modo, podríamos decir que el modelo de Hillygus y Jackman sobre la preferencia de los votantes después de las convenciones de los partidos se ajusta a los datos así, mientras que el modelo correspondiente para el periodo posterior al debate se ajusta a los datos así. Lo que queda muy claro en este par de gráficos es que hay un número bastante grande de eventos en estos datos, es decir, declaraciones de intención de voto, y además que la mayoría de los Las probabilidades de predicción más altas son las de los individuos que declararon su intención de voto.

### 3.2. Guerra Civil: Fearon y Laitin (2003).

---

<sup>6</sup>La matriz de conectividad se basa en las distancias de Gower, calculadas a partir de diferentes variables diseñadas para captar la medida en que los pares de estados experimentan acontecimientos políticos similares.

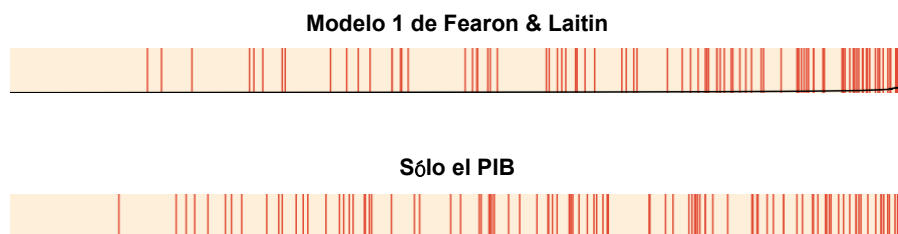
Figura 5. Gráficos de separación para los modelos de Hillygus y Jackman sobre la intención de voto en las elecciones presidenciales de 2000. El gráfico superior muestra los resultados de la encuesta realizada en el periodo posterior a las convenciones de los partidos, mientras que el gráfico inferior muestra los resultados de la encuesta realizada después de los debates presidenciales. Ambos modelos se ajustan perfectamente a los datos.



Fearon y Laitin (2003) desarrollaron un modelo muy influyente sobre el inicio de la guerra civil en el periodo posterior a la Segunda Guerra Mundial. Basándose en la significación estadística de las variables incluidas en el modelo de regresión logística, los autores afirman que diversos factores económicos y geográficos que favorecen las insurgencias (por ejemplo, la pobreza, las grandes poblaciones y el terreno montañoso) tienden a estar asociados con el inicio de la guerra civil. Y lo que es más importante, descubren que los factores culturales, como el grado de heterogeneidad étnica, no están asociados a una mayor probabilidad de inicio de la guerra civil.

Sin embargo, cuando se visualiza el poder predictivo dentro de la muestra mediante un gráfico de separación, el modelo parece ajustarse relativamente mal a los datos. De hecho, como mostramos en la Figura 6, este modelo se ajusta a los datos sólo marginalmente mejor que un modelo mucho más parsimonioso que incluye el PIB per cápita (registrado) como única covariable, a pesar de la presencia de un gran número de variables estadísticamente significativas. Este es un punto que desarrollamos con más detalle en Ward, Greenhill y Bakke (de próxima publicación).

Figura 6. Comparación de los gráficos de separación producidos al replicar el modelo 1 de Fearon y Laitin (2003), y al reestimar el modelo con el PIB per cápita registrado como única covariable.

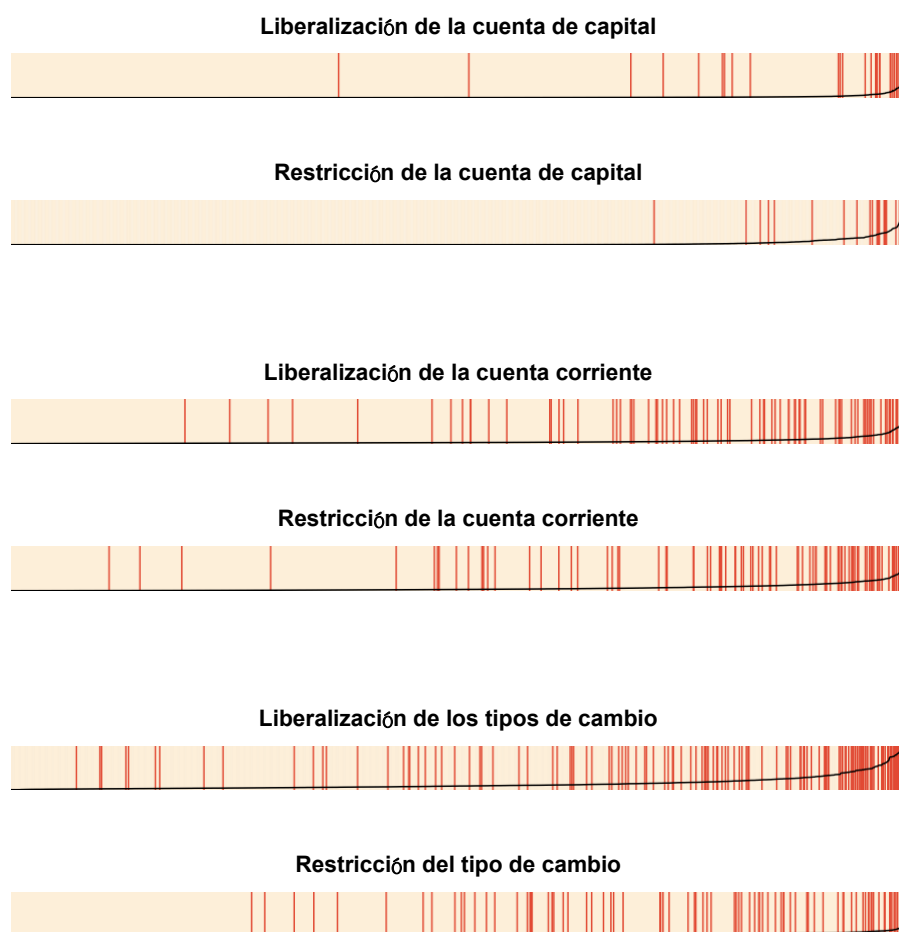


### 3.3. Difusión de políticas: Simmons y Elkins (2004).

Simmons y Elkins (2004) sostienen que la decisión de un país de adoptar políticas económicas liberales no puede explicarse únicamente por las condiciones internas, sino que depende en gran medida de las

de las decisiones tomadas por otros países influyentes. Uno de sus hallazgos más notables es que la decisión de un país de adoptar políticas liberales con respecto a su cuenta de capital, cuenta corriente y/o control de su tipo de cambio está estrechamente correlacionada con las decisiones tomadas por otros países que comparten una religión común con el país de interés. Los autores encuentran que una variable espacialmente retrasada que indica la naturaleza liberal/iliberal de las políticas económicas de los países correligionarios tiene un efecto estimado positivo y estadísticamente significativo ( $p < 0,05$ ) en todos los modelos. Esto es así en las tres dimensiones de la política económica que abarca su estudio, así como para las transiciones en ambas direcciones, es decir, las transiciones hacia una política más liberal y las transiciones hacia una política más restrictiva. Los autores sugieren que esta conclusión refleja el hecho de que, cuando se enfrentan a información incompleta sobre los costes y beneficios de adoptar una política concreta, los Estados tienden a imitar las políticas de otros culturalmente similares (Simmons y Elkins, 2004:187).

Figura 7. Gráficos de separación para los modelos de Simmons y Elkins.



Simmons y Elkins utilizan un modelo de supervivencia de Weibull para estimar el tiempo que tarda cada país en experimentar una transición hacia (o desde) una política económica liberal. En la Figura 7 mostramos los gráficos de separación obtenidos al reestimar el modelo utilizando una especificación logit más sencilla. Esta especificación es

una opción habitual para tratar una variable de duración que se produce dentro de



unidades discretas de tiempo (por ejemplo, cuando los datos sólo contienen información sobre el año de adopción de la política) y cuando los niveles de las covariables para cada país no son constantes a lo largo del tiempo (Box-Steffensmeier y Jones, 2004:108). Por lo tanto, el uso de un modelo logit nos permite generar una probabilidad de transición prevista para cada país-año hasta el año en el que se produjo la transición, inclusive. Estos valores se pueden comparar con los valores de la variable dicotómica que indica si la transición se produjo o no en cada caso de país-año. En este sentido, los gráficos de separación pueden interpretarse de la misma manera que antes: en la medida en que el modelo se ajuste bien a los datos, las probabilidades más altas de transición deberían corresponder a los sucesos reales de transición para cada país de la muestra.

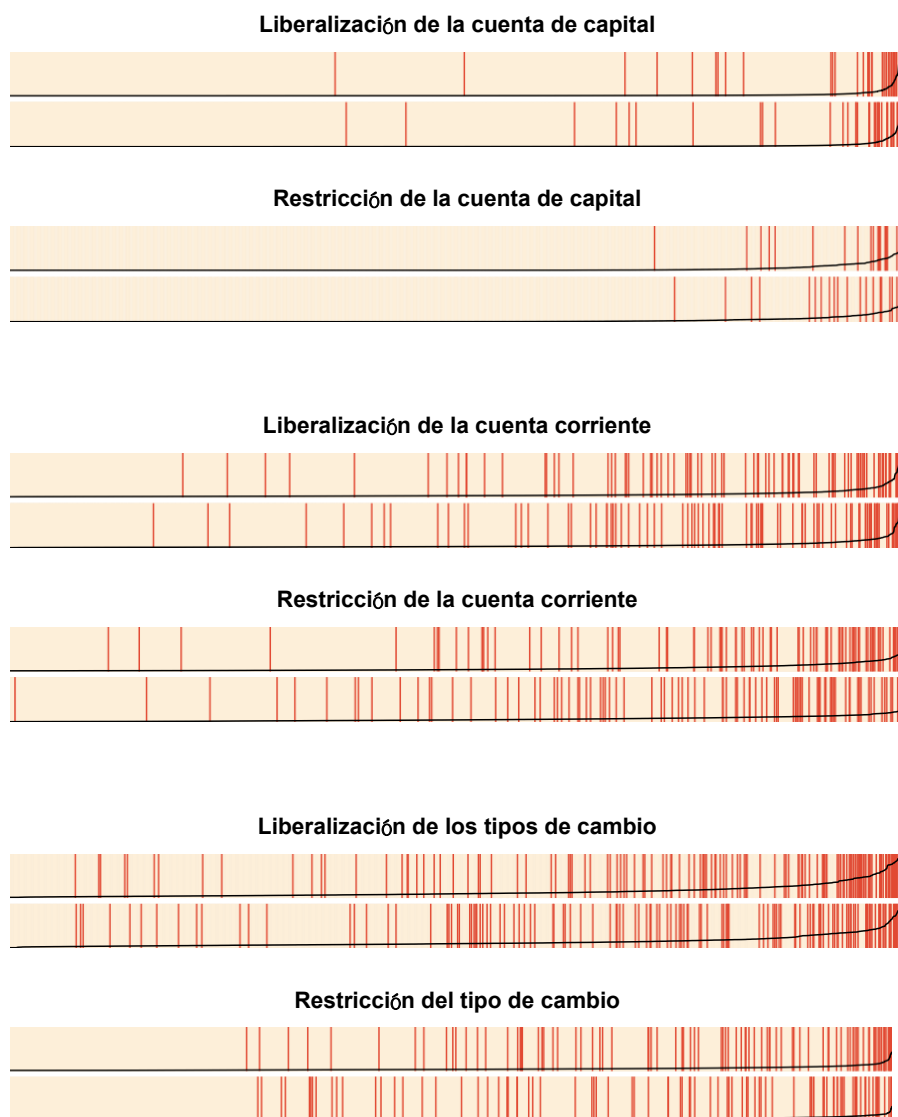
Los gráficos de separación muestran que el modelo de apertura de la cuenta de capital hace un trabajo razonablemente bueno en la predicción de los casos de países-año en los que se producen las transiciones. Esto es cierto para el modelo de adopción de políticas liberales (el gráfico superior de la Figura 7) y el correspondiente modelo de transición a políticas restrictivas (el segundo gráfico). Los resultados de los modelos de liberalización de la cuenta corriente y del tipo de cambio son, sin embargo, menos impresionantes; estos modelos asignan probabilidades relativamente bajas a muchos de los casos reales de transición.

En su artículo, Simmons y Elkins presentan los resultados de una serie de pruebas de razón de verosimilitud para demostrar que la inclusión conjunta de las tres medidas de difusión cultural utilizadas en su estudio (concretamente, los desfases espaciales calculados a partir de los datos sobre religiones comunes, lenguas comunes e historias coloniales comunes) mejora significativamente el ajuste de los modelos. Informan de que la inclusión de estas tres variables conduce a una mejora del ajuste que es estadísticamente significativa al nivel 0,05 para cinco de los seis modelos. El único que no lo es es el modelo de transiciones a políticas de cuenta de capital más restrictivas (Simmons y Elkins, 2004: Tabla 3).

En la figura 8 se muestra cómo se puede volver a realizar este análisis mediante una serie de gráficos de separación. La Figura 8 es esencialmente la misma que la Figura 7, salvo que ahora hemos añadido un gráfico de separación adicional que nos permite comparar los resultados de los modelos de la Figura 7 con modelos equivalentes estimados sin las tres variables de difusión cultural. En cada par de gráficos, el gráfico superior muestra los resultados del modelo totalmente especificado, mientras que el gráfico inferior representa el mismo modelo sin las variables de difusión cultural. Si la inclusión de estas variables conduce de hecho a una mejora sustancial del ajuste global del modelo, deberíamos esperar ver un mayor grado de separación en los gráficos superiores en comparación con los inferiores.

Sin embargo, los resultados de los pares de gráficos de separación mostrados en la Figura 8 no sugieren que la inclusión de las variables de difusión cultural conduzca a una mejora significativa del ajuste del modelo. En cambio, parece que el poder de predicción dentro de la muestra sólo mejora en dos de los seis modelos, concretamente en los modelos de restricción de la cuenta corriente y de liberalización del tipo de cambio. En ambos casos se observa una agrupación más estrecha de las líneas rojas (que representan los países-año en los que se adopta la política) en el lado derecho del gráfico. En los otros cuatro modelos hay poca o ninguna evidencia de una mejora en el poder de predicción, y en el caso de los modelos de cuenta de capital se podría argumentar que la inclusión de estas variables realmente *reduce* el poder de predicción del modelo.

Figura 8. Gráficos de separación para los modelos de Simmons y Elkins, estimados con y sin las variables de difusión cultural. Las cubiertas superiores de los pares de gráficos representan el ajuste dentro de la muestra de los modelos totalmente especificados (y, por tanto, son idénticos a los gráficos de la figura 7), mientras que las cubiertas inferiores representan el ajuste dentro de la muestra del modelo correspondiente reestimado sin las variables de difusión cultural.



#### 4. ¿Qué puede salir mal?

En los casos en que el tamaño de la muestra es muy grande ( $N > 10.000$ ), resulta difícil ver las líneas individuales que corresponden a los casos individuales en el gráfico de separación. Esto dificulta la interpretación cuando la cantidad de separación en el gráfico es baja y los valores predichos para los eventos y los no eventos caen en el rango de 0,0 a 1,0. En algunos casos, las líneas blancas son prácticamente imposibles de ver y el gráfico parece un sólido rectángulo rojo. Para solucionar este problema, hemos creado una versión alternativa del gráfico que separa los sucesos y los no sucesos en dos gráficos distintos y utiliza un espectro de colores en el que cada color corresponde a un rango de valores ajustados. Los tonos rojos más claros denotan probabilidades más pequeñas y los más oscuros, probabilidades más grandes. Además de facilitar la visualización de la cantidad de separación en el gráfico, nos permite comparar el número de sucesos y no sucesos que predecimos con probabilidades bajas y altas. A medida que el ajuste del modelo mejora, deberíamos observar menos tonos de rojo en cada uno de los gráficos.

##### 4.1. Ejemplo: Comportamiento del voto (Rosenstone y Hansen, 1993).

Como ejemplo, nos basamos en Rosenstone y Hansen (1993) y King, Tomz y Wittenberg (2000). Rosenstone y Hansen (1993) emplean un modelo logit para explicar por qué algunos individuos tienen más probabilidades que otros de votar en las elecciones presidenciales de Estados Unidos. King, Tomz y Wittenberg (2000) utilizaron entonces este modelo como medio para ilustrar cómo mejorar la interpretación estadística de los análisis de regresión. A efectos expositivos, King, Tomz y Wittenberg (2000) se centraron únicamente en las siguientes variables demográficas que Rosenstone y Hansen (1993) destacaron para explicar la participación de los votantes en las elecciones presidenciales de los años 1960 a 1996: Educación; Ingresos; Edad; y Raza (blancos y no blancos). King, Tomz y Wittenberg (2000) señalan que después de estimar un modelo logit, muchos estudiosos como Rosenstone y Hansen (1993) sólo presentan los coeficientes, los errores estándar y la significación estadística. En lugar de centrarse en las estimaciones de los parámetros y los estadísticos  $t$ , sugieren que los investigadores deberían presentar las cantidades de interés utilizando, por ejemplo, las primeras diferencias simuladas o los valores esperados.

Recomendamos que los investigadores vayan un paso más allá de calcular las cantidades de interés y el grado de certeza sobre estas cantidades para evaluar el ajuste del modelo. Ni King, Tomz y Wittenberg (2000) ni Rosenstone y Hansen (1993) comentan el ajuste del modelo. Reproducimos el modelo de participación electoral descrito anteriormente y utilizamos los valores predichos y reales para crear un gráfico de separación. Como se ilustra en la Figura 9, las líneas blancas y rojas individuales se vuelven imposibles de distinguir a esta escala debido al gran  $N$  ( $N=15.837$ ) de este estudio.

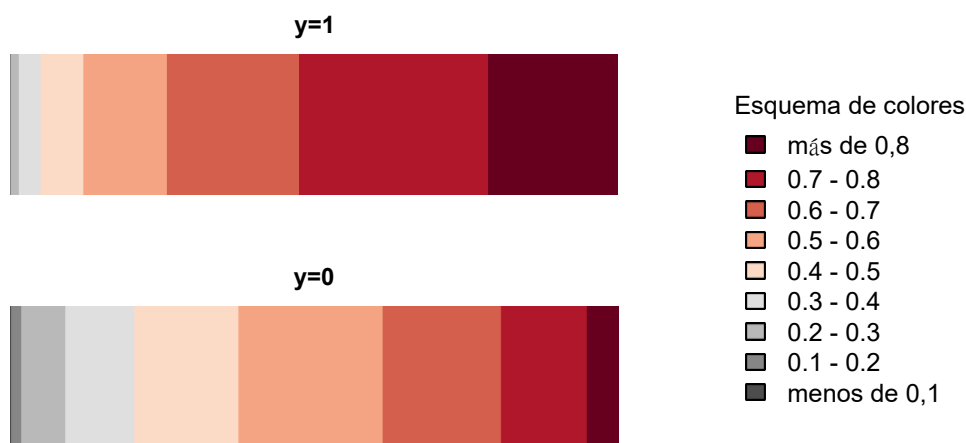
Figura 9. Gráficos de separación para el modelo de Rosenstone y Hansen (1993) y King, Tomz y Wittenberg (2000) de participación electoral en las elecciones presidenciales entre los años 1960 y 1996. Las líneas rojas y blancas individuales son imposibles de distinguir a esta escala cuando  $N$  es tan grande.



En lugar de ampliar el tamaño del gráfico, utilizamos la versión alternativa del gráfico de separación que agrupa las probabilidades en bandas discretas, como se ha descrito anteriormente. En la Figura 10, los colores corresponden a rangos de probabilidades; el tono más oscuro de rojo corresponde a una probabilidad de 0,8 y superior y el tono más claro de rojo corresponde a una probabilidad de 0,1 e inferior.

El gráfico de separación sugiere que el modelo funciona razonablemente bien a la hora de asignar altas probabilidades a los casos reales de participación electoral. Las bandas rojas más oscuras que representan las altas probabilidades de que se produzca el suceso son considerablemente más amplias en el piso superior del gráfico (que consiste en los sucesos reales) que en el piso inferior (que consiste en los no sucesos). Del mismo modo, las bandas grises correspondientes a las bajas probabilidades de que se produzca el acontecimiento son más amplias en el piso inferior que en el superior.

Figura 10. Una versión alternativa de los gráficos de separación para el modelo de Rosenstone y Hansen (1993) y King, Tomz y Wittenberg (2000) sobre la participación de los votantes en las elecciones presidenciales entre los años 1960 y 1996. El gráfico superior muestra los resultados para los casos de participación real, mientras que el gráfico inferior muestra los resultados para los casos de absentismo. El gráfico superior muestra un mayor ajuste a los datos que el gráfico inferior.



## 5. Futuros caminos

### 5.1. Variables dependientes politómicas.

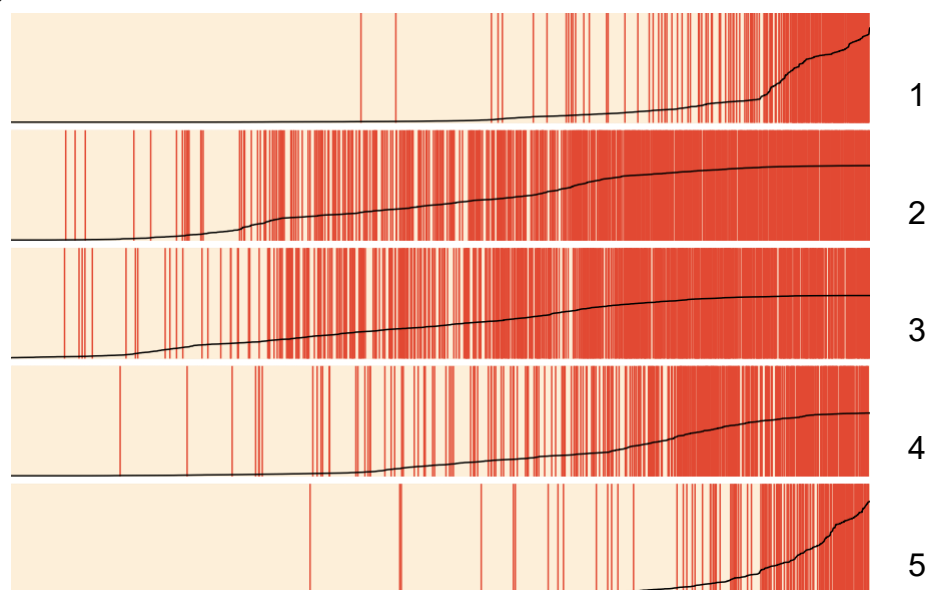
El concepto de gráfico de separación puede ampliarse para ayudar en el análisis de modelos probit/logit ordenados o multinomiales que tienen más de dos resultados categóricos. En estos casos, se genera un gráfico de separación individual para cada

categoría de resultado. Por lo tanto, cada gráfico compara las probabilidades de obtener un resultado concreto con un indicador dicotómico de si cada caso está o no asociado a esa categoría concreta

del resultado. El usuario puede entonces revisar una serie de gráficos de separación que representan el poder predictivo del modelo en todas las categorías posibles de resultados.

En la Figura 11 mostramos un ejemplo de cómo se pueden utilizar los gráficos de separación para evaluar el poder predictivo de un modelo probit ordenado. En esta figura reproducimos los resultados de Neumayer (2005: Tabla 2, Modelo 6), que examina el efecto que tiene la ratificación del Pacto Internacional de Derechos Civiles y Políticos en las prácticas de derechos humanos de los Estados. La variable dependiente es la "Escala de Terror Político" de cinco puntos, que mide hasta qué punto los gobiernos cometen violaciones de los derechos humanos destinadas a interrumpir las actividades políticas no violentas. Las puntuaciones más altas en la Escala de Terror Político representan mayores niveles de represión.

Figura 11. Gráficos de separación para cada uno de los cinco niveles de derechos de integridad personal estimados en Neumayer (2005: Tabla 2, Modelo 6).



Como muestran los gráficos, en general el modelo de Neumayer hace un buen trabajo al hacer coincidir las altas probabilidades de cada resultado con los sucesos reales de cada resultado. Por ejemplo, el primer gráfico de separación de la figura 11 muestra que las probabilidades más altas previstas de generar un resultado de 1 en la escala de terror político corresponden de hecho a resultados reales de 1. Para las categorías intermedias 2, 3 y 4, el modelo parece ser algo menos discriminatorio que para las categorías 1 y 5.

Cuando se evalúa el poder predictivo de los modelos para variables dependientes politómicas, es importante tener en cuenta que lo que importa no es la proporción de líneas oscuras y claras en cada gráfico (que, por supuesto, dependerá de cómo se distribuya la variable de resultado entre las posibles categorías), sino el grado general en que las líneas oscuras se separan de las claras. Por ejemplo, en el caso del modelo de Neumayer de la Figura 11, muchos más casos de países-año en ese conjunto de datos tienen valores

de la Escala de Terror Político de 2, 3 y 4 que los extremos de 1 y 5.

## 6. Conclusión

El gráfico de separación permite evaluar el ajuste de un modelo de regresión (logit o probit) para el que la variable dependiente es binaria. Añade información sobre el ajuste de los modelos de regresión discreta y puede mejorar nuestra capacidad para entender la incertidumbre de las predicciones que se hacen con esos modelos estadísticos. El gráfico de separación tiene las siguientes características:

- Proporciona un rápido resumen visual de la distribución de eventos y no eventos en los datos. Así, ¿los eventos son raros o frecuentes?
- Ilustra si las observaciones con altas probabilidades de predicción realmente experimentaron el evento. Si es así, el modelo tiene un alto grado de ajuste predictivo, que puede ser evaluado visualmente, sin predeterminedir el umbral para las predicciones de binning.
- Ilustra la existencia de grupos de falsos positivos y falsos negativos, en caso de que exista alguno o ambos.
- Permite una comparación bastante directa entre diferentes modelos, tanto para los mismos datos como para datos diferentes.
- Es relativamente fácil de aplicar, explicar y presentar. Al ser visual, se capta rápidamente.
- Para estar seguros, no sustituye el uso de la simulación para comparar los valores esperados en diferentes escenarios.

¿Cuál es el inconveniente de la trama de separación? La desventaja básica es su lado positivo. Es una presentación visual, que tiene un cierto elemento subjetivo. Hay muchos resúmenes de un solo número que podrían emplearse para evaluar el ajuste de este tipo de modelos. Algunos de ellos también son arbitrarios. ¿Cuál es el valor del área bajo la curva que indica que un modelo tiene un alto grado de ajuste? Otros, por ejemplo, los cocientes de probabilidad, tienen una base en la teoría estadística, pero generalmente requieren alguna comparación con otros modelos (a menudo imaginarios). El gráfico de separación no proporciona un resumen numérico único, sino que ofrece una presentación visual del ajuste. De este modo, se une a las presentaciones de las distribuciones de los valores esperados en diferentes escenarios, en las que rara vez se presentan pruebas numéricas específicas y se representa gráficamente el alcance de las diferencias. Sin embargo, sería posible calcular una variedad de estadísticas interesantes en el gráfico de separación, incluyendo pero no limitándose a la prueba U de Mann-Whitney. Preferimos dejar el desarrollo de esto a los estudiosos que lo consideren necesario. Más bien, pensamos que el uso de los gráficos de separación para presentar el ajuste de los modelos logísticos y probit servirá a las necesidades de muchos científicos sociales.

## Apéndice A. Paquete R

Un paquete de R que contiene las funciones necesarias para generar los distintos tipos de gráficos de separación que se discuten en este artículo estará pronto disponible a través de la Red de Archivos R. Mientras tanto, se pueden obtener copias de estas funciones poniéndose en contacto con Brian Greenhill (bdgreen@u.washington.edu).

La función principal del paquete se llama `separationplot`. Sólo requiere dos argumentos: un vector de probabilidades predichas (`pred`) y un vector de los resultados correspondientes (`actual`). Por lo tanto, la generación del gráfico de separación es



completamente independiente del procedimiento de estimación del modelo. Sin embargo, cuando se utiliza la función en los procedimientos de ajuste del modelo

como la descrita en la sección 2.2, recomendamos incorporar la función en una rutina de validación cruzada. Esto permite al usuario comparar fácilmente especificaciones alternativas de un modelo basándose en un rápido vistazo a sus gráficos de separación fuera de la muestra. Los argumentos opcionales de la función `separationplot` incluyen la posibilidad de identificar interactivamente uno o más casos individuales en el gráfico mediante clics del ratón (por ejemplo, `locate=2`), y de marcar una observación particular antes de generar el gráfico (por ejemplo, `flag=1234`, `flagcol="blue"`).

Otros argumentos proporcionan un control detallado sobre la forma de la salida gráfica.

Las dos funciones relacionadas incluidas en el paquete son `sp.categorical` - una envoltura para `separationplot` que genera matrices de gráficos para modelos con variables dependientes politómicas (véase el ejemplo de la Sección 5.1 anterior), y `sp.largeN` - una función para generar el gráfico de separación con probabilidades binadas como se demuestra en la Sección 9 anterior.

## Referencias

- Aldrich, John y Forrest Nelson. 1984. *Análisis con una variable dependiente limitada: Linear Probability, Logit, and Probit Models*. Beverly Hills, CA: Sage Publishers.
- Box-Steffensmeier, J.M. y B.S. Jones. 2004. *Event history modeling: A guide for social scientists*. Cambridge University Press.
- Brier, Glenn W. 1950. "Verificación de pronósticos expresados en términos de probabilidades". *Boletín de la Sociedad Meteorológica Americana* 78:1-3.
- Fearon, J.D. y D.D. Laitin. 2003. "Ethnicity, insurgency, and civil war". *American Political Science Review* 97(01):75-90.
- Hillygus, D.S. y S. Jackman. 2003. "Voter decision making in election 2000: campaign effects, partisan activation, and the Clinton legacy". *American Journal of Political Science* 47(4):583-596.
- Imai, Kosuke, Gary King y Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development". *Journal of Computational and Graphical Statistics* 17(4):892-913.
- King, G., M. Tomz y J. Wittenberg. 2000. "Sacar el máximo partido a los análisis estadísticos: Mejorando la interpretación y la presentación". *American Journal of Political Science* pp. 347- 361.
- Marshall, M. y K. Jaggers. 2003. "Political Regime Characteristics and Transitions 1800- 2003". *Proyecto Polity IV*.
- Neumayer, E. 2005. "¿Mejoran los tratados internacionales de derechos humanos el respeto de los derechos humanos ?" *Journal of conflict resolution* 49(6):925-953.
- O'Brien, Sean P. 2010 de próxima aparición. "Alerta temprana de crisis y apoyo a la toma de decisiones: Enfoques temporales y reflexiones sobre la investigación futura". *International Studies Review* 6(4):tba.
- Rosenstone, S.J. y J.M. Hansen. 1993. *Mobilization, participation, and democracy in America*. Macmillan Pub Co.
- Simmons, B.A. y Z. Elkins. 2004. "La globalización de la liberalización: Policy diffusion in the international political economy". *American Political Science Review* 98(1):171-189.
- Tufte, E.R. 2006. *Beautiful evidence*. Graphics Press Cheshire, Conn.
- Ward, M.D., B.D. Greenhill y K. Bakke. de próxima publicación. "The Perils of Policy by P-Value: Predicting Civil Conflict". *Journal of Peace Research*.

Departamento de Ciencias Políticas, Universidad de Washington, Seattle, WA, 98195  
*Dirección de correo electrónico:* bdgreen@u.washington.edu

Departamento de Ciencias Políticas, Universidad de Duke, Durham, NC, 27707  
*Dirección de correo electrónico:* mw160@duke.edu

Departamento de Sociología, Universidad de Washington, Seattle, WA, 98195  
*Dirección de correo electrónico:* sacks@u.washington.edu