Predicción de incumplimientos crediticios en PYMEs

Regresión Avanzada

David Edgardo Castillo Rodríguez
Miguel Ángel Ávila del Bosque
Jorge III Altamirano Astorga
Mario Alberto Cruz García

ITAM

Problema

Las PYMEs son muy importantes , y de acuerdo con la CONDUSEF:

- Aportan al **72%** de los empleos en el país.
- Aportan **52%** al PIB.

Aún así no se cuentan con suficientes mecanismos de financiamiento, debido a la escasa información financiera que se reporta en éste sector.

Con base en lo anterior, surge la necesidad de poder conocer los riesgos crediticios a los que está expuesto dicho sector empresarial.

Propuesta

Implementar un criterio de decisión que ayude a clasificar si una PYME va a poder cumplir con sus obligaciones contractuales crediticias.

Con el fin de fomentar el otorgamiento de créditos a este sector financiero, se propone crear una herramienta que permita clasificar con mayor precisión el incumplimiento de las PYMES.

Dicha herramienta toma información histórica de la institución acreditada con el fin de generar una probabilidad asociada al incumplimiento del crédito que permita reducir las reservas de las PYMES.

Variables

Se cuenta con las siguientes variables:

Abreviatura	Variable	Descripción
y	incumplido_prueba	Vale 1 si el crédito fue declarado en default y 0 e.o.c.
m3m	max_atraso_3m	Máximo número de atrasos en los 3 meses anteriores
psa6m	pje_sdo_atr0_6m	Porcentaje de saldo con atraso en los 6 meses anteriores
mofb	months_on_file_banking	Meses desde que la empresa fue reportada en buró por primera vez

Abreviatura	Variable	Descripción
nbkpromt	nbk12_pct_promt	Porcentaje de pagos en tiempo en los últimos 12 meses a instituciones financieras no bancarias
bkpromt	bk12_pct_promt	Porcentaje de pagos en tiempo en los últimos 12 meses a instituciones financieras bancarias
bkicq	bk12_ind_qcra	Indicador de quitas y castigos
bkipmor	bk_ind_pmor	Indicador de persona moral
pa0	prom_atr0	Promedio de atrasos a tiempo 0

Datos: Separación en 2 conjuntos

Separamos nuestros datos con una semilla estática, para hacerlo reproducible:

1. Observaciones utilizadas para el conjunto de entrenamiento (~70%): 10,595
2. Observaciones utilizadas para el conjunto de prueba (~30%): 4,542

Realizamos muestras con un promedio en la variable respuesta como se muestra a continuación al dividir el conjunto de datos original.

1. Media de la "y" en el conjunto original: 0.1936
2. Media de la "y" en el conjunto de entrenamiento: 0.1932
3. Media de la "y" en el conjunto de pruebas 0.194

ITAM

Datos: EDA

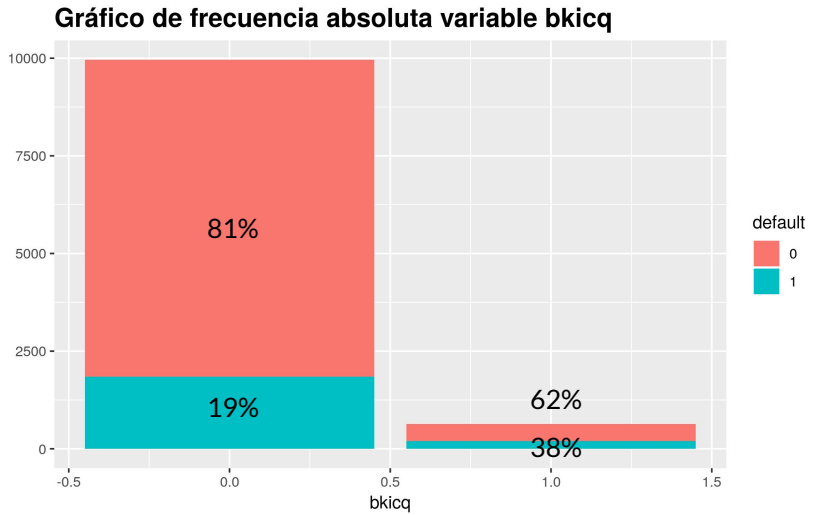


Gráfico de frecuencia absoluta variable bkicq
10000
7500
5000
2500
0
81%
19%
62%
38%
-0.5
0.0
0.5
1.0
1.5
bkicq
default
0
1

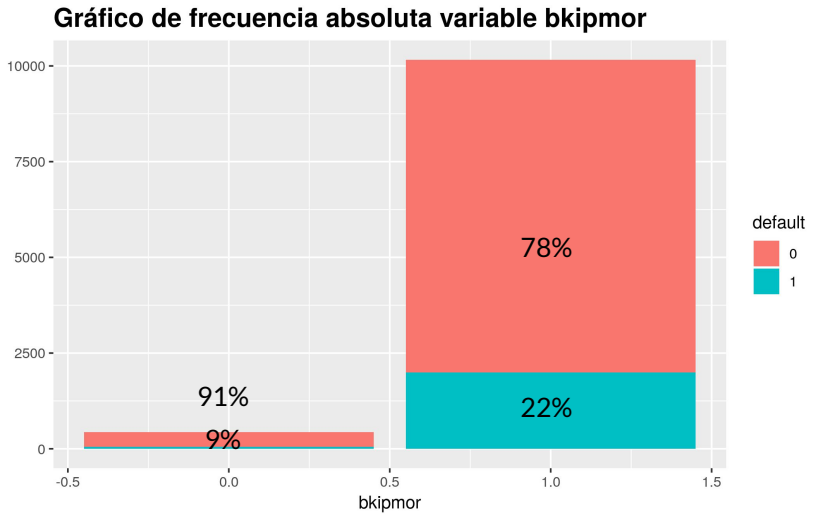


Gráfico de frecuencia absoluta variable bkipmor
10000
7500
5000
2500
0
91%
9%
78%
22%
-0.5
0.0
0.5
1.0
1.5
bkipmor
default
0
1

Datos: EDA

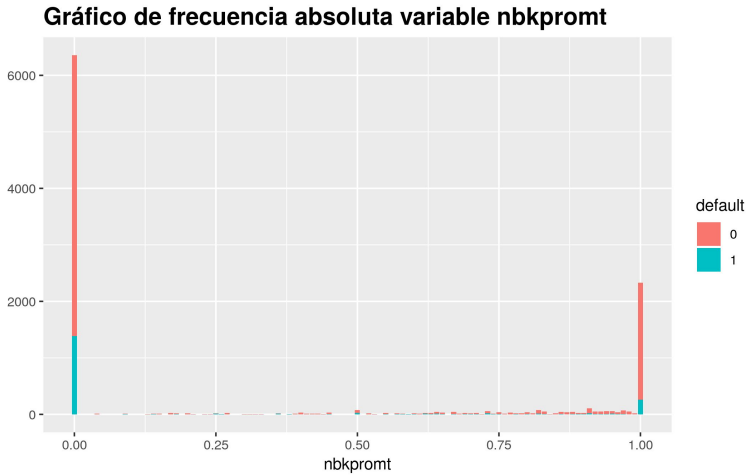


Gráfico de frecuencia absoluta variable nbkpromt
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
nbkpromt
default
0
1

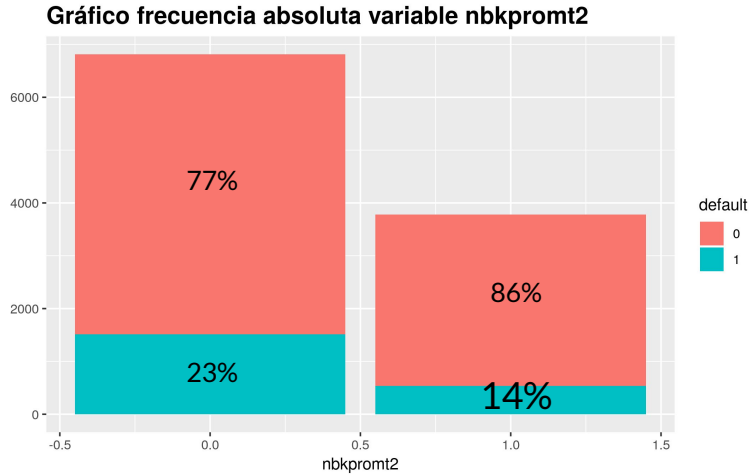


Gráfico frecuencia absoluta variable nbkpromt2
6000
4000
2000
0
77%
23%
86%
14%
-0.5
0.0
0.5
1.0
1.5
nbkpromt2
default
0
1

Datos: EDA

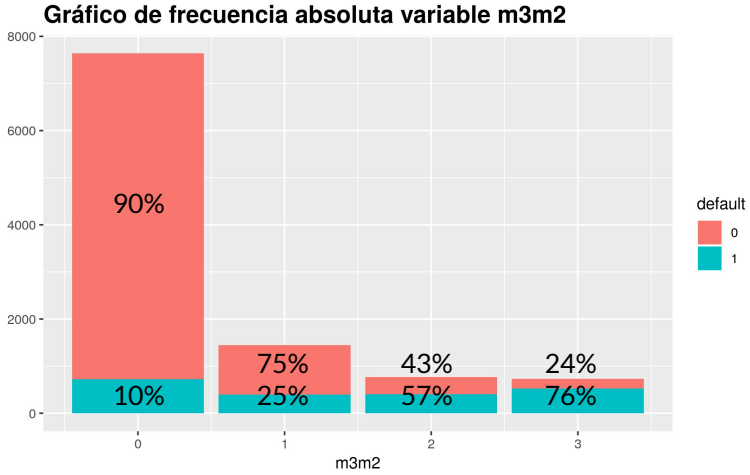


Gráfico de frecuencia absoluta variable m3m2
8000
6000
4000
2000
0
90%
10%
75%
25%
43%
57%
24%
76%
0
1
2
3
m3m2
default
0
1

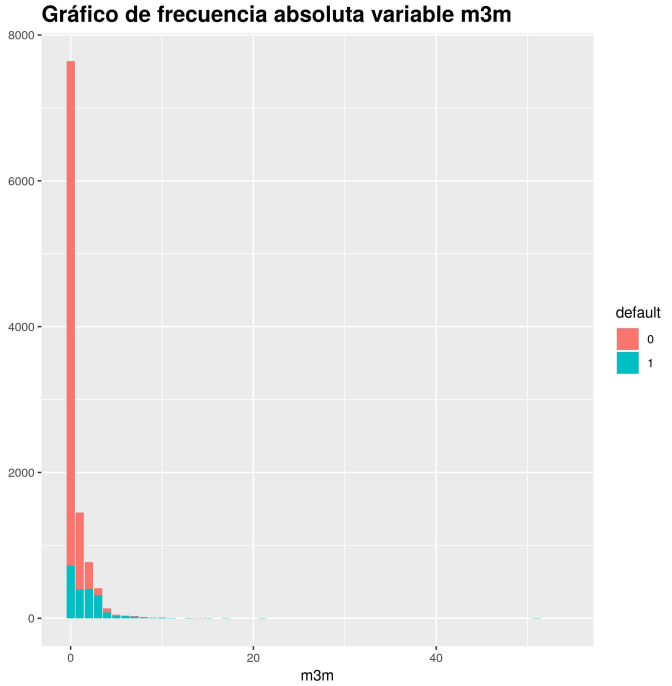


Gráfico de frecuencia absoluta variable m3m
8000
6000
4000
2000
0
0
20
40
m3m
default
0
1

Datos: EDA

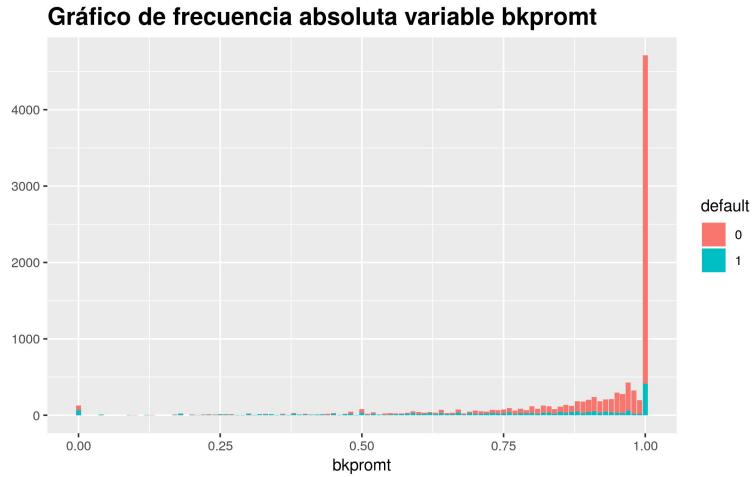


Gráfico de frecuencia absoluta variable bkpromt
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
bkpromt
default
0
1

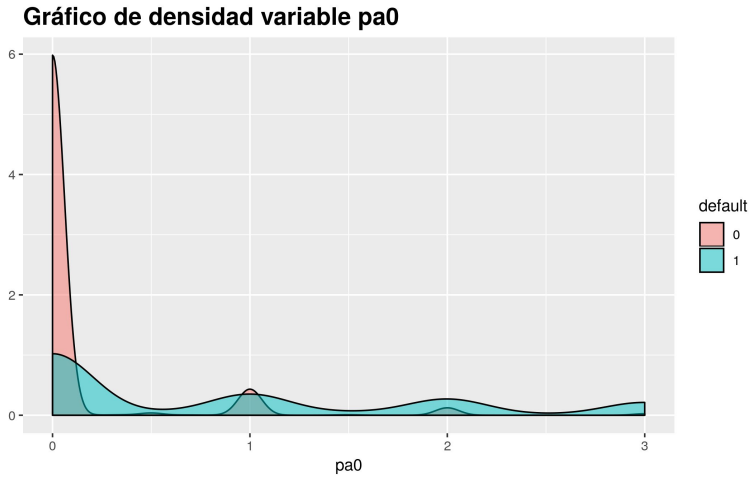
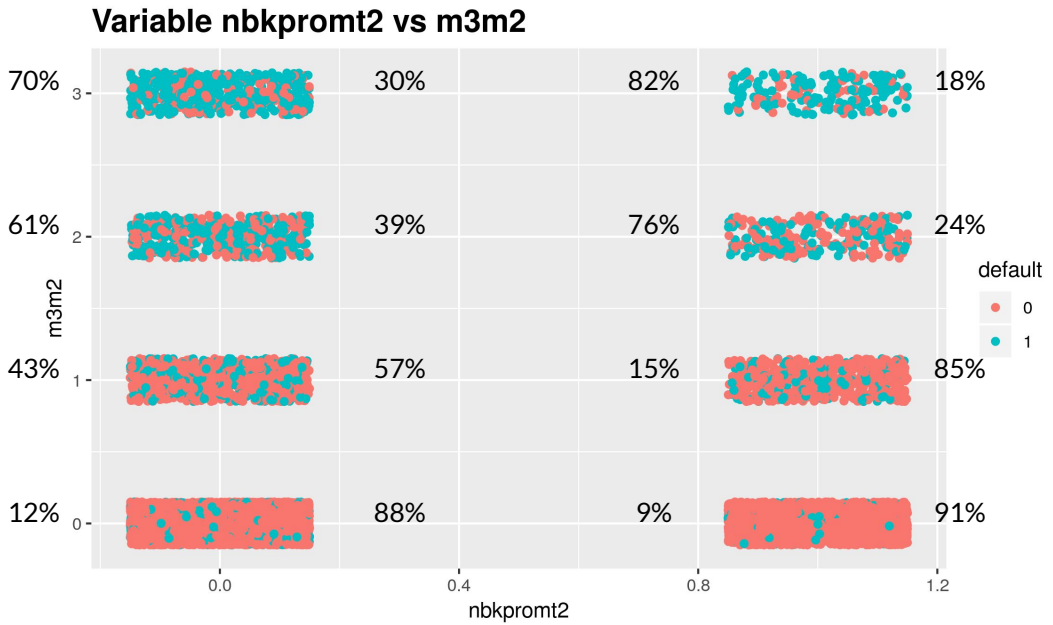
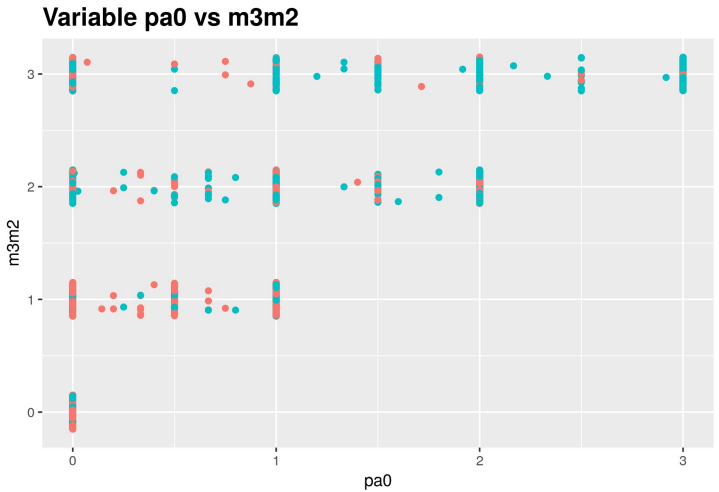


Gráfico de densidad variable pa0
6
4
2
0
0
1
2
3
pa0
default
0
1

Datos: EDA


Variable nbkpromt2 vs m3m2
70%
30%
82%
18%
61%
39%
76%
24%
43%
57%
15%
85%
12%
88%
9%
91%
m3m2
3
2
1
0
0.0
0.4
0.8
1.2
nbkpromt2
default
0
1

Datos: EDA


Variable pa0 vs m3m2
m3m2
pa0
default
0
1

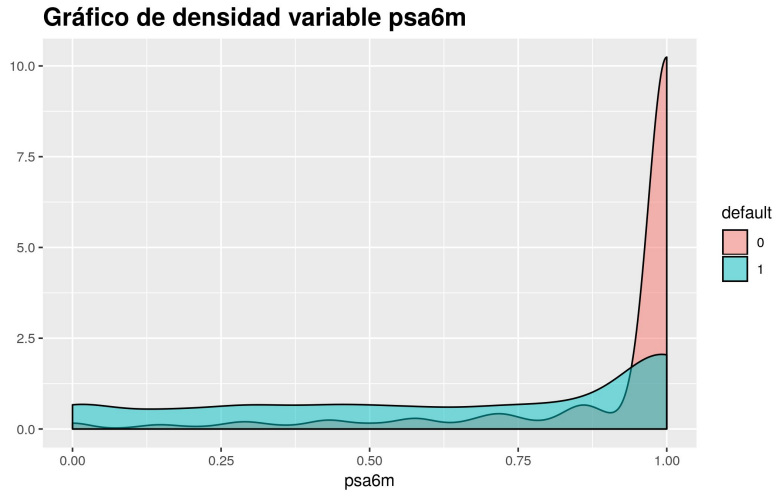


Gráfico de densidad variable psa6m
psa6m
default
0
1

Datos: PCA

Grupo 1	Grupo 2	Grupo 3	Grupo 4
pa0	m3m	mofb	bkpromt
	bkicq	nbkpromt	psa6m
		bkipmor	

Modelos

Se probaron modelos con verosimilitud Bernoulli y con cada una de las funciones liga mostradas en clase con el fin de comparar los mejores DICs.

$$y_i \mid \mu_i \sim Bernoulli(\mu_i = \theta)$$

$$\eta = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5$$

Liga Logística: $\theta = \frac{1}{1+e^{\eta}}.$

Liga C-Log Log: $\theta = log(-log(\eta)).$

Liga Probit: $\theta = \Phi(\eta).$

Liga Log Log: $\theta = log(-log(1-\eta)).$

Datos: Variables del Modelo

Variable	Descripción
y	Marca de incumplimiento (vale 1 si el crédito fue declarado en default y 0 e.o.c.)
m3m2	Máximo número de atrasos en los 3 meses anteriores
nbk_promt2	% pagos en tiempo en los últimos 12 meses a instituciones financieras no bancarias
bkpromt	% pagos en tiempo en los últimos 12 meses a instituciones financieras bancarias
pa0	Promedio de atrasos a tiempo 0
nbm3	Creada mediante ingeniería de variables: $(nbkpromt2 + .01) * ifelse(\text{m3m} \geq 3{,}3\ , \text{m3m})$

Predictores

$$\beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5$$

X_1: m3m2

X_2: nbkpromt2

X_3: bkpromt

X_4: pa0

X_5: nbm3

Comparación de Modelos

Aquí se pueden observar el desempeño de los mencionados modelos con las medias de los coeficientes lo cual se logró con 20,000 simulaciones cada uno en el muestreador de Gibbs para asegurar convergencia:

Modelo	DIC	β_1	β_2	β_2	β_3	β_4	β_5
LogLog	7958.21	0.56	-0.29	0.11	0.25	-0.28	-0.01
C-LogLog	8134.63	-1.83	0.44	-0.27	-0.21	0.24	0.11
Probit	8341.72	-1.01	0.3	-0.15	-0.22	0.24	0.05
Logit	8494.03	-1.72	0.52	-0.28	-0.38	0.41	0.09

Interpretación de Coeficientes

Se interpretarán a continuación los coeficientes. Para lo cual tomamos el modelo Bernoulli con liga log log, dado que obtuvo el mejor desempeño basándonos en su DIC, pudiéndose expresar como:

$$log(-log(\mu_i)) = \eta_i = x_i\beta \iff \mu_i = exp\{-exp[x_i\beta]\}$$

Tenemos que:

$$\frac{log(\mu_j)}{log(\mu_i)} = e^{\beta_i}$$

Cociente de logaritmos. Como el logaritmo transforma de [0,1] a valores negativos, los signos de esta liga son distintos a los de otras. Es decir, tienen el signo contrario.

Algunas Predicciones

Correctamente clasificados (umbral 62.5 % para clasificarse como default):

	y	y_hat	m3m2	nbkpromt2	bkpromt	pa0	nbm3
1	1	1.00	0.00	1.00	0.94	0.00	0.00
2	1	1.00	1.00	0.00	0.93	0.00	0.01
3	0	0.00	2.00	0.00	0.50	2.00	0.02
4	0	0.00	3.00	0.00	0.52	2.00	0.03

Incorrectamente clasificados (umbral 62.5 % para clasificarse como default):

	y	y_hat	m3m2	nbkpromt2	bkpromt	pa0	nbm3
1	1	0.00	1.00	0.00	0.21	0.00	0.01
2	1	0.00	2.00	0.00	0.77	1.00	0.02
3	1	0.00	3.00	0.00	0.54	0.00	0.03

Algunas Predicciones

Correctamente clasificados (umbral 12.5 % para clasificarse como default):

	y	y_hat	m3m2	nbkpromt2	bkpromt	pa0	nbm3
1	1	1.00	0.00	1.00	0.94	0.00	0.00
2	0	0.00	2.00	1.00	0.98	2.00	2.02
3	0	0.00	2.00	0.00	0.67	2.00	0.02
4	0	0.00	3.00	0.00	0.90	1.50	0.03

Incorrectamente clasificados (umbral 12.5 % para clasificarse como default):

	y	y_hat	m3m2	nbkpromt2	bkpromt	pa0	nbm3
1	0	1.00	0.00	0.00	0.78	0.00	0.00
2	0	1.00	1.00	1.00	1.00	0.50	1.01
3	0	1.00	2.00	1.00	0.87	1.00	2.02
4	0	1.00	3.00	0.00	0.83	0.00	0.03

ITAM

Conclusiones

Se cumplió el objetivo de obtener un (modelo de score) basado en la probabilidad de default con resultados razonables mediante un modelo lineal generalizado con enfoque bayesiano. Sin embargo, podemos decir que se requieren más pruebas en más créditos para poder tener mayor certidumbre de que nuestro modelo efectivamente fomentaría que los créditos sean dados a las PYMEs, manteniendo la certeza de que van a ser pagados aún sin un historial crediticio.

También mediante el análisis de los datos se logró el objetivo de identificar a un conjunto de variables que pudieran describir y predecir si cae en default.

ITAM

Referencias

- Luis E. Nieto-Barajas. Notas del curso de regresión avanzada. 2019.
- Nicky Best Dave Lunn David Spiegelhalter, Andrew Thomas. OpenBUGS User Manual. Cambdridge University, March 2014.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis. Chapman and Hall/CRC, 2013.
- David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. The BUGS book: A practical introduction to Bayesian analysis. Chapman and Hall/CRC, 2012.

ITAM

Anexos

Convergencia
de las Cadenas