Capítulo 8
Uso de paquetes para aplicar modelos avanzados

En los primeros siete capítulos de este libro, hemos tratado R como un programa de software estadístico tradicional y revisó cómo puede realizar la gestión de datos, informar estadísticas simples y estimar una variedad de modelos de regresión. En el resto de este libro, sin embargo, veremos la flexibilidad adicional queR ofrece, tanto en términos de capacidad de programación que está disponible para el usuario como en la provisión de herramientas adicionales aplicadas a través de *paquetes.* En este capítulo, nos enfocamos en cómo cargar lotes adicionales de código de paquetes escritos por el usuario puede agregar funcionalidad que muchos programas de software no permitirán. Aunque hemos usado paquetes para una variedad de propósitos en los siete capítulos anteriores (incluyendococho, gmodels, y enrejado, por nombrar algunos), aquí destacaremos los paquetes que permiten métodos únicos. Si bien el sitio web de CRAN enumera numerosos paquetes que los usuarios pueden instalar en un momento dado, nos centraremos en cuatro paquetes particulares para ilustrar los tipos de funcionalidad que se pueden agregar.

El primer paquete que discutiremos, lme4, permite a los usuarios estimar modelos multinivel, ofreciendo así una extensión a los modelos de regresión discutidos en los Capítulos. 6 y 7 (Bates et al. 2014). Los otros tres fueron desarrollados específicamente por científicos políticos para abordar los problemas de análisis de datos que encontraron en su investigación: MCMCpack permite a los usuarios estimar una variedad de modelos en un marco bayesiano usando la cadena de MarkovMonte Carlo (MCMC) (Martin et al. 2011). cem permite al usuario realizar un emparejamiento exacto aproximado, un método para la inferencia causal con datos de campo (Iacus et al. 2009, 2011). Finalmente,dominar permite al usuario escalar los datos de elección, como los datos de la lista legislativa, para estimar los puntos ideales ideológicos de los legisladores o encuestados (Poole y Rosenthal 1997; Poole y col.2011). Las siguientes cuatro secciones considerarán cada paquete por separado, por lo que cada sección presentará su ejemplo de datos a su vez. Estas secciones están diseñadas para ofrecer una breve descripción de la

Electrónico suplementario material: La en línea versión de esto capítulo (doi: 10.1007 / 978-3-319-23446-5_8) contiene usuarios autorizados material, que está disponible para suplementarios.

JE Monogan III, *Análisis político con R,* Utilice R !,
DOI 10.1007 / 978-3-319-23446-5_8

tipos de capacidades R ofrecen los paquetes, aunque algunos lectores pueden no estar familiarizados con los antecedentes de algunos de estos métodos. Se anima al lector interesado a consultar algunos de los recursos citados para aprender más sobre la teoría detrás de estos métodos.

8.1 Modelos multinivel con lme4

Habiendo discutido los modelos lineales en el Cap. 6 y varios ejemplos de modelos lineales generalizados en el Cap. 7, pasamos ahora a una extensión de este tipo de modelos: modelos multinivel. Los modelos multinivel, o modelos jerárquicos, son apropiados siempre que los datos de interés tengan una estructura anidada o longitudinal. Una estructura anidada ocurre cuando se puede pensar que las observaciones están dentro o como parte de una unidad de nivel superior: un ejemplo de política común es estudiar los resultados del aprendizaje de los estudiantes, pero los estudiantes están anidados dentro de las aulas. En tal caso, el investigador debería tener en cuenta el hecho de que los estudiantes de la muestra no son independientes entre sí, pero es probable que sean similares si se encuentran en la misma clase. De manera similar, siempre que un investigador estudia individuos que han repetido observaciones a lo largo del tiempo, es razonable pensar que las observaciones referenciadas en el tiempo están integradas dentro de las observaciones del individuo. Por ejemplo,1986) datos introducidos por primera vez en el Cap. 4, los ingresos de los participantes en su estudio se observan en 1974,
1975 y 1978. Algunos análisis de políticas pueden optar por considerar las tres observaciones temporales para cada individuo como anidadas dentro del caso de cada individuo.[1]

Se pueden encontrar explicaciones más completas de los modelos multinivel en Scott et al. (2013) y Gelman y Hill (2007). Continuamos ampliando dos de nuestros ejemplos anteriores para ilustrar un modelo lineal multinivel y un modelo logit multinivel.

8.1.1 Regresión lineal multinivel

En este ejemplo, volvemos a nuestro ejemplo del Cap. 6 sobre el número de horas que los docentes dedican a la docencia en el aula. Originalmente, ajustamos un modelo lineal usando mínimos cuadrados ordinarios (MCO) como nuestro estimador. Sin embargo, Berkman y Plutzer (2010) señalan que es probable que los profesores del mismo estado compartan características similares. Estas características podrían ser similitudes en la capacitación, en la cultura local o en la ley estatal. Para dar cuenta de estas similitudes no observadas, podemos pensar en los maestros como *anidado* dentro de los estados. Por esta razón, agregaremos un *efecto aleatorio* para cada estado. Los efectos aleatorios explican *correlación intraclase,* o correlación de errores entre

[1] Tenga en cuenta que este enfoque multinivel de datos de panel es más sensato para paneles cortos como estos, donde hay muchos individuos en relación con el número de puntos de tiempo. Para paneles largos en los que hay muchos puntos de tiempo en relación con el número de individuos, los modelos más apropiados se describen como métodos de sección transversal de series de tiempo agrupadas. Para obtener más información sobre el estudio de paneles cortos, consulte Monogan (2011) y Fitzmaurice et al. (2004).

observaciones dentro del mismo grupo. En presencia de correlación intraclase, las estimaciones de MCO son ineficaces porque los términos de perturbación no son independientes, por lo que un modelo de efectos aleatorios da cuenta de este problema.

Primero, recargamos los datos de la Encuesta Nacional de Maestros de Biología de Escuelas Secundarias de la siguiente manera:[2]

```
rm (lista = ls ())
biblioteca (extranjera)
evolución <-read.dta ("BPchap7.dta") evolución $ mujer
[evolución $ mujer == 9] <- NA evolución <-subconjunto
(evolución,! is.na (mujer))
```

Recuerde que tuvimos un puñado de observaciones de mujer que necesitaba ser recodificado como perdido. Como antes, subconjuntamos nuestros datos para omitir estas observaciones faltantes.

Para adaptarse a un modelo multinivel, hay algunos comandos disponibles. Optaremos por utilizar un comando dellme4**l**en el oido **metro**ijo **mi**ffect) biblioteca.[3] En nuestro primer uso, instalaremos el paquete, y luego, en cada uso, cargaremos la biblioteca:

```
biblioteca install.packages
("lme4") (lme4)
```

Una vez que hemos cargado la biblioteca, ajustamos nuestro modelo lineal multinivel usando el lmer (**l**en el oido **metro**ijo **mi**efectos **r**egresión) comando:

```
hours.ml <-lmer (hrs_allev ~ fase1 + senior_c + ph_senior + notest_p +
        ph_notest_p + female + biocred3 + degr3 + evol_course + certificado + idsci_trans
        + seguro + (1 | st_fip), datos = evolución)
```

La sintaxis de la lmer El comando es casi idéntico al código que usamos cuando ajustamos un modelo con OLS usando lm. De hecho, el único atributo que agregamos es el término adicional (1 | st_fip) en el lado derecho del modelo. Esto agrega una intersección aleatoria por estado. En cualquier ocasión en la que queramos incluir un efecto aleatorio, colocamos entre paréntesis el término para el que se incluye el efecto seguido de una barra vertical y la variable que identifica las unidades de nivel superior. Entonces, en este caso, queríamos una intersección aleatoria (de ahí el uso de1), y queríamos que estos fueran asignados por estado (de ahí el uso de st_fip).

Obtenemos nuestros resultados escribiendo:

```
resumen (horas.ml)
```

En nuestra salida de resultados, R imprime la correlación entre todos los *efectos fijos,* o parámetros de regresión estimados. Esta parte de la impresión se omite a continuación:

Ajuste de modelo lineal mixto por REML ['lmerMod'] Fórmula:

hrs_allev ~ fase1 + senior_c + ph_senior + notest_p + ph_notest _p +

[2] Si no tiene estos datos de antes, puede descargar el archivo BPchap7.dta del Dataverse en la página vii o el contenido del capítulo en la página 125.

[3] Ver también el nlme biblioteca, que fue un predecesor de lme4.

```
    female + biocred3 + degr3 + evol_course + certificada + idsci_
        trans +
    seguro + (1 | st_fip)
  Datos: evolución

Criterio REML en la convergencia: 5940

Residuos escalados:
      Mediana mínima 1T                 3T        Max
- 2,3478 -0,7142 -0,1754             0.5566     3.8846

Efectos aleatorios:
  Nombre de grupos          Variación Desv. Estándar
  st_fip (intercepción)       3.089       1,758
  Residual                   67.873       8.239
Número de obs: 841, grupos:          st_fip, 49

Efectos fijos:
                    Estimar Std. Valor t de error
(Interceptar)      10.5676          1.2138     8.706
fase 1              0,7577          0.4431     1.710
senior_c          - 0.5291          0.3098   - 1.708
ph_senior         - 0.5273          0.2699   - 1.953
notest_p            0.1134          0,7490     0,151
ph_notest_p       - 0,5274          0,6598   - 0,799
mujer             - 0,9702          0,6032   - 1,608
biocred3            0.5157          0.5044     1.022
degr3             - 0,4434          0.3887   - 1,141
curso_evolucionario 2.3894          0,6270     3.811
certificado       - 0.5335          0,7188   - 0,742
idsci_trans         1.7277          1.1161     1.548
seguro              2.6739          0,4468     5.984
```

La salida primero imprime una variedad de estadísticas de ajuste: AIC, BIC, log-verosimilitud, desviación, y desviación de máxima verosimilitud restringida. En segundo lugar, imprime la varianza y la desviación estándar de los efectos aleatorios. En este caso, la varianza de last_fip término es la varianza de nuestros efectos aleatorios a nivel estatal. La varianza residual corresponde a la varianza del error de regresión que normalmente calcularíamos para nuestros residuos. Por último, los efectos fijos que se informan son sinónimos de coeficientes de regresión lineal que normalmente nos interesan, aunque ahora nuestras estimaciones han tenido en cuenta la correlación intraclase entre profesores dentro del mismo estado. Mesa8.1 compara nuestras estimaciones OLS y multinivel una al lado de la otra. Como puede verse, el modelo multinivel ahora divide la varianza inexplicable en dos componentes (nivel estatal e individual), y las estimaciones de coeficientes han cambiado algo.

Cuadro 8.1 Dos modelos de horas de clase dedicadas a la enseñanza de la evolución de la biología de la escuela secundaria profesores

	OLS			Multi nivel		
Parámetro	Estimar	Std. error	Pr (>jzj)	Estimar	Std. error	Pr (>jzj)
Interceptar	10.2313	1.1905	0,0000	10,5675	1,2138	0,0000
Índice de estándares 2007	0,6285	0,3331	0.0596	0,7576	0,4431	0.0873
Antigüedad (centrada)	0,5813	0,3130	0.0636	0,5291	0,3098	0.0876
Antigüedad de los estándares	0.5112	0.2717	0.0603	0,5273	0,2699	0.0508
Cree que no hay exámenes Los	0,4852	0,7222	0.5019	0,1135	0,7490	0.8795
estándares creen que no hay exámenes	0,5362	0,6233	0.3899	0,5273	0,6598	0.4241
El maestro es mujer	1,3546	0,6016	0.0246	0,9703	0,6032	0.1077
Créditos obtenidos en biología (0-2)	0,5559	0,5072	0.2734	0,5157	0,5044	0.3067
Grados en ciencias (0-2)	0,4003	0,3922	0.3077	0,4434	0,3887	0.2540
Clase de evolución completada	2,5108	0,6300	0,0001	2.3894	0.6270	0,0001
Tiene certi fi cación normal Se	0,4446	0,7212	0.5377	0,5335	0,7188	0.4580
identifica como científico	1.8549	1.1255	0.0997	1.7277	1.1161	0.1216
Experiencia autoevaluada + 2)	2.6262	0.4501	0,0000	2,6738	0,4468	0,0000
(variación de 1 a nivel estatal	-			3.0892		
Varianza a nivel individual	69.5046			67,8732		

Notas: N D 841. Datos de Berkman y Plutzer (2010)

8.1.2 Regresión logística multinivel

Si bien es algo más complejo, la lógica del modelado multinivel también puede se aplicado al estudiar variables dependientes limitadas. Hay dos enfoques amplios para extender los GLM a un marco multinivel: modelos marginales, que tienen una interpretación de población promediada, y modelos lineales mixtos generalizados (GLMM), que tienen una interpretación a nivel individual (Laird y Fitzmaurice2013, págs. 149–156). Si bien se anima a los lectores a leer más sobre los tipos de modelos disponibles, su estimación y su interpretación, por ahora nos centramos en el proceso de estimación de un GLMM.

En este ejemplo, volvemos a nuestro ejemplo de Sect. 7.1.1 del último capítulo, sobre si un encuestado informó haber votado por el partido en el poder en función de la distancia ideológica del partido. Como Singh (2014a) observa, los votantes que hagan su elección en el mismo país-año se enfrentarán a muchas características de la elección que son exclusivas de esa elección. Por lo tanto, es probable que exista una correlación intraclase entre los votantes dentro de la misma elección. Además, el efecto de la ideología en sí puede ser más fuerte en algunas elecciones que en otras: los métodos multinivel, incluidos los GLMM, nos permiten evaluar si existe variación en el efecto de un predictor entre grupos, que es una característica que usaremos.

Volviendo a las especificaciones del código, si el lme4 la biblioteca no está cargada, la necesitamos de nuevo. Además, si los datos de no se cargan, entonces necesitamos cargar elextranjero biblioteca y el propio conjunto de datos. Todo esto se logra de la siguiente manera:[4]

```
biblioteca (lme4)
biblioteca (extranjera)
votando <-read.dta ("SinghJTP.dta")
```

Construyendo sobre el modelo de Table 7.1, primero simplemente agregamos una intersección aleatoria a nuestro modelo. La sintaxis para estimar el modelo e imprimir los resultados es:

```
inc.linear.ml <-glmer (votadainc ~ distanciainc + (1 | cntryyear),
        familia = binomio (enlace = "logit"), datos = votación)
resumen (incl.linear.ml)
```

Observe que ahora usamos el resplandecer comando**gramo**eneralizado **l**en el oido **metro**ijo **mi**efectos **r**egresión). Usando elfamilia opción, podemos utilizar cualquiera de las funciones de enlace comunes disponibles para la glm mando. Un vistazo al resultado muestra que, además de los efectos fijos tradicionales que reflejan los coeficientes de regresión logística, también se nos presenta la varianza de la intersección aleatoria para el país y el año. de la elección:

```
Ajuste de modelo lineal mixto generalizado por Laplace
    aproximación
Fórmula: voteinc ~ distanciainc + (1 | cntryyear)
      Datos: votación
         AIC          BIC       logLik desviación
 41998.96 42024.62 -20996.48 41992.96 Efectos
 aleatorios:
 Nombre de grupos                Variación Desv. Estándar
 cntryyear (intersección) 0.20663 0.45457 Número de
observaciones: 38211, grupos: cntryyear, 30

Efectos fijos:
                            Estimar       Std. Error          valor z Pr (> | z |)
(Intercepción) 0.161788717          0.085578393          1.89053 0.058687        .
distanciainc -0.          501250136       0,008875997 -56,47254 <2e-16                 * * *
- - -
Signif. códigos:          0 *** 0,001          * *   0,01     *   0,05     .    0,1     1

Correlación de efectos fijos:
                (Intr.)
distanciainc -0.185
```

Para replicar un modelo más acorde con Singh (2014a) resultados, ahora encajar en un modelo que incluye una intersección aleatoria y un coeficiente aleatorio de distancia ideológica, ambos condicionados por el país y el año de la elección. La sintaxis para estimar este modelo e imprimir el resultado es:

[4] Si no tiene estos datos de antes, puede descargar el archivo SinghJTP.dta del Dataverse en la página vii o el contenido del capítulo en la página 125.

```
inc.linear.ml.2 <-glmer (votadainc ~ distanciainc +
        (distanciainc | cntryyear), family = binomial (link = "logit"), data = vote)

resumen (incluido lineal.ml.2)
```

Observe que ahora hemos condicionado la variable distanciainc por cntryyear.
Esto agrega un coeficiente aleatorio para la distancia ideológica. Además, de forma predeterminada, agregar este efecto aleatorio también agrega una intersección aleatoria. Nuestro resultado en este caso es:

```
Ajuste de modelo lineal mixto generalizado por Laplace
    aproximación
Fórmula: voteinc ~ distanciainc + (distanciainc |
    cntryyear)
      Datos: votación
      Desviación de logLik de AIC BIC
  41074 41117-20532                41064
Efectos aleatorios:
  Nombre de grupos                 Variación Desv. Estándar   Corr
  cntryyear (intersección) 0,616658 0,78528
                      distanciainc 0.098081 0.31318           - 0,808
Número de obs: 38211, grupos: cntryyear,                      30

Efectos fijos:
                    Estimar Std. Valor z de error             Pr (> | z |)
(Interceptar)         0.26223          0.14531      1.805       0.0711.
distanciainc        - 0.53061          0.05816    - 9.124       <2e-16 ***
- - -
Signif. códigos: 0 ***                0,001 **       0,01     * 0,05. 0,1 1

Correlación de efectos fijos:
                    (Intr.)
distanciainc -0.808
```

Bajo efectos aleatorios, primero vemos la varianza para la intersección aleatoria con referencia a la elección, y luego la varianza para el coeficiente de distancia ideológica con referencia a la elección. Los efectos fijos de los coeficientes de regresión logística también se presentan de la forma habitual. El AIC indica que esta versión del modelo encaja mejor que el modelo con solo una intersección aleatoria o el modelo de la Tabla7.1 que no incluyó efectos aleatorios, ya que el puntaje de 41 074 es menor que el AIC de cualquiera de esos modelos. En resumen, esta discusión debería ofrecer una idea de los tipos de modelos jerárquicos queR puede estimar usando lme4Bates y col. 2014).

8.2 Métodos bayesianos utilizando MCMCpack

La MCMCpack El paquete permite a los usuarios realizar inferencias bayesianas en una variedad de modelos de regresión y modelos de medición comunes. El paquete incluso tiene un comando,MCMCmetrop1R, que construirá una muestra de MCMC a partir de una distribución definida por el usuario utilizando un algoritmo de Metropolis. Se anima a los lectores que deseen aprender más sobre los métodos bayesianos a consultar recursos como: Carlin y Louis (2009), Gelman et al. (2004), Branquias (2008) y Robert (2001).

Como una simple ilustración de cómo funciona el paquete, nos enfocamos en esta sección en algunos de los modelos de regresión comunes que están programados en el paquete. Esto es poderoso para elR usuario en el sentido de que los investigadores que prefieren informar modelos bayesianos pueden hacerlo fácilmente si su especificación de modelo se ajusta a una estructura común. Al ilustrar estas técnicas, revisaremos una vez más el modelo lineal de horas de evolución de Berkman y Plutzer (2010) y el modelo de regresión logística del apoyo del partido en el poder de Singh (2014a).

8.2.1 Regresión lineal bayesiana

Para estimar nuestro modelo de regresión lineal bayesiana, debemos volver a cargar los datos de la Encuesta Nacional de Maestros de Biología de Escuelas Secundarias, si aún no están cargados:

```
rm (lista = ls ())
biblioteca (extranjera)
evolución <-read.dta ("BPchap7.dta") evolución $ mujer
[evolución $ mujer == 9] <- NA evolución <-subconjunto
(evolución,! is.na (mujer))
```

Con los datos cargados, debemos instalar MCMCpack si este es el primer uso del paquete en la computadora. Una vez instalado el programa, debemos cargar la biblioteca:

```
biblioteca install.packages
("MCMCpack") (MCMCpack)
```

Ahora podemos usar **MCMC** para adaptarse a nuestro lineal bayesiano **regreso**modelo de iones con el MCMCregress mando:

```
mcmc.horas <-MCMCregress (hrs_allev ~ fase1 + senior_c + ph_senior +
      notest_p + ph_notest_p + female + biocred3 + degr3 + evol_course + certificado +
      idsci_trans + seguro, datos = evolución)
```

Esté preparado para que la estimación con MCMC generalmente toma más tiempo computacionalmente, aunque los modelos simples como este generalmente terminan con bastante rapidez. Además, debido a que MCMC es una técnica basada en simulación, es normal que los resúmenes de los resultados difieran ligeramente entre las repeticiones. Con este fin, si encuentra diferencias entre sus resultados y los impresos aquí después de usar el mismo código, no debe preocuparse a menos que los resultados sean marcadamente diferentes.

Si bien el código anterior se basa en los valores predeterminados del MCMCregress comando, algunas de las opciones de este comando son esenciales para resaltar: Una característica central de los métodos bayesianos es que el usuario debe especificar *previos* para todos los parámetros estimados. Los valores predeterminados deMCMCregress son priores conjugados vagos para los coeficientes y la varianza de las perturbaciones. Sin embargo, el usuario tiene la opción de especificar sus propios antecedentes sobre estas cantidades.[5] Se alienta a los usuarios a revisar estas opciones y otros recursos sobre cómo establecer antecedentes (Carlin y Louis 2009; Gelman y col.2004; Branquia2008; Robert2001). Los usuarios también tienen la opción de cambiar el número de iteraciones en la muestra de MCMC con elmcmc opción y el período de quemado (es decir, el número de iteraciones iniciales que se descartan) con la ardiendo opción. Los usuarios siempre deben evaluar el modelo*convergencia* después de estimar un modelo con MCMC (que se discutirá en breve) y considere si el quemado o el número de iteraciones deben cambiarse si hay evidencia de no convergencia.

Después de estimar el modelo, escribiendo resumen (mcmc.horas)ofrecerá un resumen rápido de la *posterior* muestra:

Iteraciones = 1001: 11000
Intervalo de dilución = 1 Número
de cadenas = 1
Tamaño de muestra por cadena = 10000

1. Media empírica y desviación estándar de cada variable, más el error estándar de la media:

	Significar	SD Naive SE	Serie temporal SE
(Interceptar)	10,2353	1,1922 0,011922	0.011922
fase 1	0,6346	0,3382 0,003382	0,003382
senior_c	- 0,5894	0,3203 0,003203	0,003266
ph_senior	- 0,5121	0,2713 0,002713	0,002713
notest_p	0,4828	0,7214 0,007214	0,007214
ph_notest_p	- 0,5483	0,6182 0,006182	0,006182
mujer	- 1,3613	0,5997 0,005997	0,006354
biocred3	0,5612	0,5100 0,005100	0,005100
degr3	- 0,4071	0,3973 0,003973	0,003973
curso_evolucionario	2,5014	0,6299 0,006299	0,005870
certificado	- 0,4525	0,7194 0,007194	0,007194
idsci_trans	1.8658	1.1230 0.011230	0.010938
seguro	2.6302	0.4523 0.004523	0,004590
sigma2	70,6874	3,5029 0,035029	0.035619

[5]Para los anteriores de los coeficientes, la opción b0 establece el vector de medias de un prior gaussiano multivariado, y B0 establece la matriz de varianza-covarianza del antes gaussiano multivariado. La distribución previa de la varianza del error de la regresión es Gamma inversa, y esta distribución se puede manipular estableciendo su parámetro de forma con la opciónc0 y parámetro de escala con opción d0. Alternativamente, la distribución Gamma inversa se puede manipular cambiando su media con la opción sigma.mu y su variación con la opción sigma.var.

2. Cuantiles para cada variable:

	2,5%	25%	50%	75%	97,5%
(Interceptar)	7,92359	9.438567	10.2273	11.03072	12.59214
fase 1	- 0.02787	0,405026	0,6384	0,86569	1,30085
senior_c	- 1.22527	- 0,808038	-0,5885	-0,37351	0,04247
ph_senior	- 1.04393	- 0,694228	-0,5105	-0,32981	0,03152
notest_p	- 0,92717	- 0,006441	0,4863	0,97734	1,88868
ph_notest_p	- 1.75051	- 0,972112	-0,5462	-0,13138	0,63228
mujer	- 2.52310	- 1,771210	-1,3595	-0,96109	-0,18044
biocred3	- 0,42823	0,212168	0,5558	0,90768	1,55887
degr3	- 1.19563	- 0,671725	-0,4048	-0,14536	0,38277
curso_evolucionario	1.26171	2.073478	2.5064	2.92601	3.73503
certificado	- 1.84830	- 0,942671	-0,4477	0,03113	0,95064
idsci_trans	- 0.33203	1.107771	1.8667	2.63507	4.09024
seguro	1.73568	2.324713	2.6338	2.94032	3.48944
sigma2	64.12749	68.277726	70.5889	72.95921	77.84095

Dado que MCMC produce una muestra de valores de parámetros simulados, toda la información reportada se basa en estadísticas descriptivas simples de la salida simulada (que son 10,000 conjuntos de parámetros, en este caso). La parte 1 del resumen anterior muestra la media de la muestra para cada parámetro, la desviación estándar de la muestra de cada parámetro y dos versiones del error estándar de la media. La parte 2 del resumen muestra los percentiles de la muestra de cada parámetro. Mesa8.2 muestra un formato común para presentar los resultados de un modelo como este: informar la media y el estándar

Cuadro 8.2 Modelo lineal de horas de clase dedicadas a la enseñanza de la evolución por alto profesores de biología de la escuela (estimaciones de MCMC)

Vaticinador	Significar	Std. Dev.	[95% Cred. En t.]
Interceptar	10.2353	1.1922	[7.9236: 12.5921]
Índice de estándares 2007	0,6346	0.3382	[0.0279: 1.3008]
Antigüedad (centrada)	0.5894	0.3203	[1.2253: 0.0425]
Antigüedad de los estándares	0.5121	0.2713	[1.0439: 0.0315]
Cree que no hay exámenes Los	0.4828	0,7214	[0.9272: 1.8887]
estándares creen que no hay exámenes	0.5483	0,6182	[1.7505: 0.6323]
El maestro es mujer	1.3613	0.5997	[2.5231: 0.1804]
Créditos obtenidos en biología (0-2)	0.5612	0.5100	[0.4282: 1.5589]
Grados en ciencias (0-2)	0.4071	0.3973	[1.1956: 0.3828]
Clase de evolución completada	2.5014	0,6299	[1.2617: 3.7350]
Tiene certi fi cación normal Se	0.4525	0,7194	[1.8483: 0.9506]
identifica como científico	1.8658	1.1230	[0.3320: 4.0902]
Experiencia autoevaluada (1 a C2)	2.6302	0.4523	[1.7357: 3.4894]
Varianza de error de regresión	70.6874	3.5029	[64.1275: 77.8410]

Notas: N D 841. Datos de Berkman y Plutzer (2010)

desviación de la distribución posterior marginal de cada parámetro, y un 95% *intervalo creíble* basado en los percentiles.[6]

Cuando un usuario carga MCMCpack en R, la coda la biblioteca también se cargará.[7] coda es particularmente útil porque permite al usuario evaluar la convergencia de modelos estimados con MCMC y reportar cantidades adicionales de interés. Como se mencionó anteriormente, cada vez que un investigador estima un modelo utilizando MCMC, debe evaluar si existe alguna evidencia de no convergencia. En el caso de que las cadenas de Markov no hayan convergido, el modelo debe ser muestreado para más iteraciones. MCMCregress estima el modelo usando una sola cadena. Por lo tanto, para nuestro modelo del número de horas de evolución enseñadas en las aulas de la escuela secundaria, podemos evaluar la convergencia utilizando**Geweke**la convergencia **diag**nostic, que simplemente pregunta si las medias de las primeras y últimas partes de la cadena son las mismas. Para calcular este diagnóstico, escribimos:

```
geweke.diag (mcmc.horas, frac1 = 0.1, frac2 = 0.5)
```

Aquí hemos especificado que queremos comparar la media del primer 10% de la cadena (frac1 = 0.1) a la media del último 50% de la cadena (frac2 = 0,5).
La salida resultante presenta una *z*-razón para esta diferencia de la prueba de medias para cada parámetro:

```
Fracción en la primera ventana = 0.1
Fracción en la segunda ventana = 0.5

(Interceptar)              fase 1          senior_c       ph_senior          notest_p
     - 1.34891        - 1.29015        - 1.10934         - 0.16417          0,95397
ph_notest_p                mujer         biocred3              degr3 curso_evolucion
     1.13720        - 0.57006          0.52718           1,25779 0,62082
   idsci_trans certificado              seguro                sigma2
     1,51121 - 0,87436                - 0.54549         - 0.06741
```

En este caso, ninguna de las estadísticas de prueba supera ningún umbral de significación común para una estadística de prueba distribuida normalmente, por lo que no encontramos evidencia de no convergencia. Con base en esto, podemos estar contentos con nuestra muestra original de MCMC de 10,000.

Una cosa más que quizás deseemos hacer con nuestra salida MCMC es **gráfico** el estimado general **guaridas**ity función de nuestras distribuciones marginales posteriores. Podemos graficar estos uno a la vez usando eldensplot función, aunque el analista deberá hacer referencia al parámetro de interés en función de su orden numérico de aparición en la tabla de resumen. Por ejemplo, si quisiéramos graficar el coeficiente para el indicador de si un maestro completó una clase de evolución (**evol_course),** ese es el décimo

[6] Para escribir una tabla similar a Table 8.2 en LaTeX, carga el xtable biblioteca en R y escriba lo siguiente en la consola:
xtable (cbind (resumen (mcmc.horas) $ estadísticas [, 1: 2], resumen (mcmc.horas) $ cuantiles [, c (1,5)]), dígitos = 4)

[7] Esto ocurre con frecuencia cuando un paquete depende del código de otro.

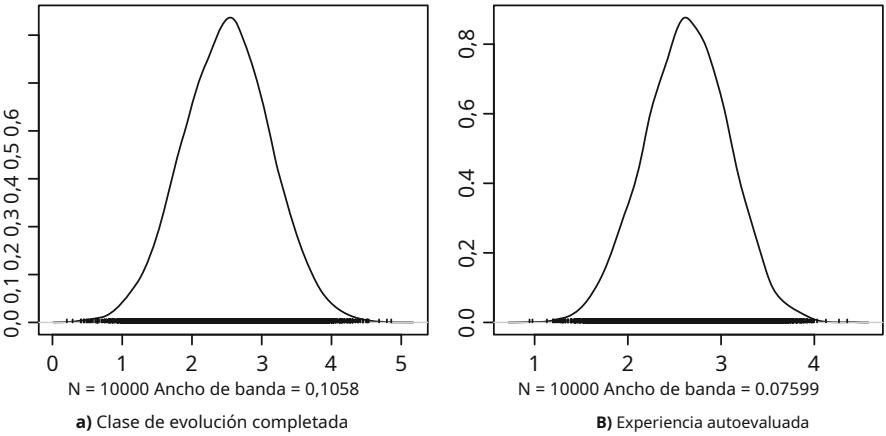

Figura 8.1 Gráficos de densidad de distribución marginal posterior de coeficientes para determinar si el maestro completó una clase de evolución y la pericia autoevaluada del maestro. Basado en una muestra de MCMC de 10,000 iteraciones (quemado de 1000 iteraciones). (**a)** Clase de evolución completada; (**B)** Experiencia autoevaluada

estimación de parámetros informada en la tabla. De manera similar, si quisiéramos informar la gráfica de densidad para el coeficiente de la pericia autoevaluada del maestro (**confiado),** ese es el decimotercer parámetro informado en la tabla resumen. Por lo tanto, podríamos trazar cada uno de estos escribiendo:

```
densplot (mcmc.horas [, 10])
densplot (mcmc.horas [, 13])
```

Las gráficas de densidad resultantes se presentan en la Fig. 8.1. Como muestran las figuras, ambas distribuciones posteriores marginales tienen una distribución normal aproximada, y la moda se encuentra cerca de la media y la mediana informadas en nuestro resultado resumido.

8.2.2 Regresión logística bayesiana

Como una ilustración adicional que MCMCpack estima una variedad de modelos, ilustramos la regresión logística bayesiana, utilizando Singh (2014a) datos sobre la votación de los partidos en el poder por última vez. Si no tiene estos datos o la biblioteca cargada, asegúrese de hacerlo:

```
biblioteca (MCMCpack)
biblioteca (extranjera)
votando <-read.dta ("SinghJTP.dta")
```

Para estimar el bayesiano **logit** modelo usando **MCMC,** escribimos:

```
inc.linear.mcmc <-MCMClogit (votadainc ~ distanciainc, datos = votando)
```

Al igual que con el MCMCregress comando, hemos optado por utilizar los valores predeterminados en este caso, pero se recomienda a los usuarios que consideren establecer sus propios valores a priori para satisfacer sus necesidades. De hecho, este es un caso en el que necesitaremos aumentar el número de iteraciones en nuestro modelo. Podemos verificar la convergencia de nuestro modelo usando el diagnóstico de Geweke:

```
geweke.diag (incluido linear.mcmc, frac1 = 0.1, frac2 = 0.5)
```

Nuestro resultado en este caso muestra una diferencia significativa entre las medias al principio y al final de la cadena para cada parámetro:

Fracción en la primera ventana = 0.1
Fracción en la segunda ventana = 0.5

(Intercepción) distanciainc
2.680 - 1.717

El valor absoluto de ambos *z*-las razones exceden 1.645, por lo que podemos decir que la media es significativamente diferente para cada parámetro en el nivel de confianza del 90%, lo cual es evidencia de no convergencia.

Como respuesta, podemos duplicar nuestro período de quemado y el número de iteraciones a 2,000 y 20,000, respectivamente. El código es:

```
inc.linear.mcmc.v2 <-MCMClogit (votadoinc ~ distanciainc,
    datos = votación, quemado = 2000, mcmc = 20000)
```

Ahora podemos verificar la convergencia de esta nueva muestra escribiendo:

```
geweke.diag (incluido linear.mcmc.v2, frac1 = 0.1, frac2 = 0.5)
```

Nuestro resultado ahora muestra no significativo *z*-ratios para cada parámetro, lo que indica que ya no hay evidencia de no convergencia:

Fracción en la primera ventana = 0.1
Fracción en la segunda ventana = 0.5

(Intercepción) distanciainc
- 1.0975 0.2128

Continuando con esta muestra de 20.000, entonces, si escribimos
resumen (incluido linear.mcmc.v2) en la consola, la salida es:

Iteraciones = 2001: 22000
Intervalo de dilución = 1 Número
de cadenas = 1
Tamaño de muestra por cadena = 20000

1. Media empírica y desviación estándar de cada variable, más el error estándar de la media:

	Significar	SD Naive SE	Serie temporal SE
(Interceptar)	0,1940	0,01846 1,305e-04	0.0003857
distanciainc	- 0,4946	0,00829 5,862e-05	0,0001715

2. Cuantiles para cada variable:

	2,5%	25%	50%	75%	97,5%
(Interceptar)	0,1573	0,1817	0,1944	0,2063	0.2298
distanciainc	- 0.5105	-0.5003	-0.4946	-0.4890	-0.4783

Estos resultados son similares a los que informamos en el último capítulo de la Tabla 7.1, aunque ahora tenemos la oportunidad de interpretar los hallazgos como un bayesiano. Al igual que con la regresión lineal bayesiana, si quisiéramos informar las gráficas de densidad de cada parámetro, podríamos aplicar ladensplot comando como antes. En general, esta breve ilustración debería mostrar a los investigadores lo fácil que es utilizar métodos bayesianos en R con MCMCpack. Se anima a los lectores que deseen utilizar métodos bayesianos más avanzados a consultar el MCMCpack manual de referencia y Martin et al. (2011).

8.3 Inferencia causal con cem

Una innovación destacada en la metodología política ha sido el desarrollo de varios métodos nuevos de emparejamiento. En resumen, el emparejamiento es una técnica diseñada para seleccionar un subconjunto de datos de campo para hacer una comparación justa de los individuos que reciben un tratamiento para controlar a los individuos que no recibieron el tratamiento. Con el emparejamiento, algunas observaciones se descartan para que las observaciones de control y tratamiento restantes sean similares en todas las covariables que se sabe que dan forma al resultado de interés. En ausencia de datos experimentales, los métodos de emparejamiento sirven para permitir al investigador aislar cómo la variable de tratamiento afecta las respuestas (ver Rubin2006 para un tratamiento completo de la inferencia causal con emparejamiento).

Los científicos políticos han desarrollado varios métodos nuevos de emparejamiento (Imai y van Dyk 2004; Sekhon y Grieve2012). Como ilustración de uno de estos, y cómo se implementa la nueva técnica enR, pasamos al método desarrollado por Iacus et al. (2009, 2011, 2012), **C**enfurecido **mi**xact **METRO**atching (CEM). En resumen, CEM procede recodificando temporalmente cada covariable en una variable ordenada que agrupa valores similares de la covariable. Luego, los datos se clasifican en estratos en función de sus perfiles de las variables aproximadas, y cualquier observación en un estrato que no contenga al menos un tratamiento y una unidad de control se desecha. La muestra resultante debe mostrar mucho más equilibrio en las variables de control entre las observaciones tratadas y de control. EnR, esta técnica se implementa con el cem comando dentro del cem paquete.

8.3.1 Desequilibrio de covariables, implementación de CEM y ATT

Como datos de nuestro ejemplo, volvemos a LaLonde (1986) estudio de la Demostración Nacional de Trabajo Apoyado que examinamos anteriormente en los Capítulos. 4 y 5. Nuestra variable de tratamiento (**tratado)** indica si la persona recibió el tratamiento de la Demostración Nacional de Trabajo Apoyado, lo que significa que la persona fue colocada en un trabajo del sector privado durante un año con fondos públicos que cubren los costos laborales. Nos gustaría conocer el efecto causal de este tratamiento sobre los ingresos del individuo en
1978, después de que se completó el tratamiento (**re78).** En la Secta. 5.1.1 Hicimos una prueba ingenua de esta hipótesis simplemente usando una prueba de diferencia de medias entre los individuos tratados y el grupo de control sin controlar ninguna de las otras variables que probablemente afecten el ingreso anual. En nuestra prueba ingenua, encontramos que los ingresos eran más altos para el grupo tratado, pero ahora podemos preguntar cuál es el efecto estimado cuando contabilizamos otras covariables importantes.

Como se señaló en capítulos anteriores, los datos de LaLonde ya están incluidos en el cem paquete, para que podamos cargar estos datos fácilmente si ya hemos cargado el cem Biblioteca. Las siguientes líneas de código limpian, instalancemen caso de que aún no lo hayas hecho), abre cem, y carga los datos de LaLonde (llamados LL):[8]

```
rm (lista = ls ())
biblioteca install.packages
("cem") (cem)
datos (LL)
```

Una tarea importante cuando se utilizan métodos de emparejamiento es evaluar el grado en que los datos son *equilibrado,* o el grado en que los casos tratados tienen una distribución similar de valores de covariables en relación con el grupo de control. Podemos evaluar el grado en que nuestros grupos de tratamiento y control tienen distribuciones diferentes con eldesequilibrio mando. En el código siguiente, primero aumentamos la penalización por la notación científica (una opción si prefiere la notación decimal). Luego, creamos un vector nombrando las variables para las que no queremos evaluar el equilibrio: la variable de tratamiento (**tratado)** y el resultado de interés (**re78).** Todas las demás variables del conjunto de datos son covariables que creemos que pueden dar forma al ingreso en 1978, por lo que nos gustaría tener un equilibrio sobre ellas. En la última línea, en realidad llamamos aldesequilibrio mando.

```
opciones (scipen = 8)
todrop <- c ("tratado", "re78") desequilibrio (grupo = LL $ tratado,
datos = LL, drop = todrop)
```

Dentro de desequilibrio comando, el grupo El argumento es nuestra variable de tratamiento que define los dos grupos para los que queremos comparar las distribuciones de covariables. Ladatos El argumento nombra el conjunto de datos que estamos usando, y el soltar La opción nos permite omitir ciertas variables del conjunto de datos al evaluar el equilibrio de covariables. Nuestro resultado de este comando es el siguiente:

[8] Los datos de LaLonde también están disponibles en el archivo LL.csv, disponible en el Dataverse (consulte la página vii) o el contenido del capítulo (consulte la página 125).