Capítulo 3

Visualización de datos

La presentación visual de datos y los resultados de los modelos se ha convertido en una pieza central del análisis político moderno. Muchas de las principales revistas de Ciencias Políticas, incluida la*Revista Estadounidense de Ciencias Políticas,* ahora solicite cifras en lugar de tablas siempre que ambos puedan transmitir la misma información. De hecho, Kastellec y Leoni (2007) argumentan que las cifras transmiten resultados empíricos mejor que las tablas. Cleveland (1993) y Tufte (2001) escribió dos de los volúmenes principales que describen los elementos de una buena visualización cuantitativa, y Yau (2011) ha producido una versión más reciente de la representación gráfica. Básicamente, estos trabajos sirven como manuales de estilo para gráficos.[1] Más allá de las sugerencias que estos académicos ofrecen por el bien de los lectores, ver los propios datos visualmente transmite información sustancial sobre las características univariadas, bivariadas y multivariadas de los datos: ¿Una variable parece sesgada? ¿Parecen correlacionarse sustancialmente dos variables? ¿Cuál es la relación funcional adecuada entre variables? ¿Cómo cambia una variable en el espacio o en el tiempo? Responder estas preguntas por uno mismo como analista y por el lector en general puede elevar la calidad del análisis presentado a la disciplina.

Al borde de este movimiento gráfico en el análisis cuantitativo, R ofrece visualización de modelos y datos de última generación. Muchos de los programas estadísticos comerciales han intentado durante años ponerse al díaRCapacidades gráficas. Este capítulo muestra estas capacidades, primero en elgráfico función que está disponible automáticamente como parte de la base paquete. En segundo lugar, analizamos algunos de los otros comandos de graficación que se ofrecen en elbase Biblioteca. Finalmente, pasamos alenrejado Biblioteca, que permite al usuario crear gráficos Trellis, una marco para la visualización

Electrónico suplementario material: La en línea versión de esto capítulo (doi: 10.1007 / 978-3-319-23446-5_3) contiene usuarios autorizados material, que está disponible para suplementarios.

[1]Otras cifras históricas particularmente clave en el desarrollo de medidas gráficas incluyen Halley (1686), Juega limpio (1786/2005) y Tukey (1977). Una historia más completa es presentada por Beniger y Robyn (1978).

© Springer International Publishing Suiza 2015

JE Monogan III, *Análisis político con R,* Utilice R !,
DOI 10.1007 / 978-3-319-23446-5_3

desarrollado por Becker, Cleveland y otros para poner el de Cleveland (1993) sugerencias en la práctica. Aunque el espacio no lo permite aquí, también se anima a los usuarios a buscar elggplot2 paquetes, que ofrece opciones de gráficos adicionales. Chang (2013), en particular, ofrece varios ejemplos de gráficos con ggplot2.

En este capítulo, trabajamos con dos conjuntos de datos de ejemplo. El primero es sobre el cabildeo en salud en los 50 estados estadounidenses, con un enfoque específico en la proporción de empresas de la industria financiera de la salud que están registradas para cabildear (Lowery et al.2008). Una variable de predicción clave es el número total de empresas de financiación de la salud abiertas al público, que incluye organizaciones que ofrecen planes de salud, servicios comerciales, coaliciones de empleadores de salud y seguros. El conjunto de datos también incluye la tasa de participación de los grupos de presión por estado, o el número de grupos de presión como una proporción del número de empresas, no solo en finanzas de salud sino para todas las empresas relacionadas con la salud y en otras seis subáreas. Estos son datos transversales del año 1997. La lista completa de variables es la siguiente:

stno: Índice numérico de 1 a 50 que ordena los estados alfabéticamente.

raneyfolded97: Índice de Ranney plegado de la competencia estatal bipartita en 1997.[2]

healthagenda97: Número de proyectos de ley relacionados con la salud considerados por la legislatura estatal.
tura en 1997.

negocio de suministro: Número de establecimientos de financiación sanitaria.

businesssupplysq: Número al cuadrado de establecimientos de financiación de la salud.

partratebusness: Tasa de participación en los grupos de presión para las finanzas sanitarias (número de inscritos)
ciones como porcentaje del número de establecimientos.

predecirbuspartrate: Predicción de la tasa de participación en el financiamiento de la salud como cuadrática
función del número de establecimientos de fi nanciación de la salud. (Sin variables de control en la
predicción).

partratetotalhealth: Tasa de participación en el lobby para toda la atención médica (incluidos siete
subáreas).

partratedpc: Tasa de participación en el lobby para la atención directa al paciente.

partratepharmprod: Tasa de participación en el lobby de medicamentos y productos sanitarios.

partrateprofessionals: Tasa de participación de los profesionales de la salud en el lobby.

partrateadvo: Tasa de participación del lobby para la promoción de la salud.

partrategov: Tasa de participación en el lobby del gobierno local.

rnmedschoolpartrate: Tasa de participación de los grupos de presión para la educación sanitaria.

En segundo lugar, analizamos las de Peake y Eshbaugh-Soha (2008) datos sobre el número de noticias de televisión relacionadas con la política energética en un mes determinado. En este marco de datos, las variables son:

Fecha: Vector de caracteres del mes y año observado.

Energía: Número de historias relacionadas con la energía transmitidas en los noticieros de televisión nocturnos por
mes.

Desempleo: La tasa de desempleo por mes.

Aprobación: Aprobación presidencial por mes.

oilc: Precio del petróleo por barril.

[2] Nebraska y Carolina del Norte son observaciones faltantes del índice de Ranney.

freeze1: Una variable indicadora codificada con 1 durante los meses de agosto a noviembre 1971, cuando se impuso la congelación de precios y salarios. Codificado 0 en caso contrario.

freeze2: Una variable indicadora codificada con 1 durante los meses de junio a julio de 1973, cuando se impusieron congelaciones de precios, salarios y precios. Codificado 0 en caso contrario.

embargo: Una variable indicadora codificada con 1 durante los meses de octubre de 1973 a marzo 1974, durante el embargo petrolero árabe. Codificado 0 en caso contrario.

rehenes: Una variable indicadora codificada 1 durante los meses de noviembre de 1979–Enero de 1981, durante la crisis de los rehenes en Irán. Codificado 0 en caso contrario.

Discursos presidenciales: Los indicadores adicionales se codifican como 1 durante el mes a El presidente pronunció un discurso importante sobre política energética, y 0 en caso contrario. Los indicadores de los respectivos discursos se denominan:**rmn1173, rmn1173a, grf0175, grf575, grf575a, jec477, jec1177, jec479, grf0175s, jec479s,** y **jec477s.**

3.1 Gráficos univariados en el base Paquete

Como primera mirada a nuestros datos, mostrar una sola variable gráficamente puede transmitir una idea de la distribución de los datos, incluyendo su modo, dispersión, sesgo y curtosis. La enrejado biblioteca en realidad ofrece algunos comandos más para visualización univariante que base lo hace, pero comenzamos con los principales comandos univariados incorporados. La mayoría de los comandos de gráficos enbase paquete llame al gráfico función, pero hist y diagrama de caja son excepciones notables.

La hist El comando es útil para simplemente tener una idea de la frecuencia relativa de varios valores comunes. Comenzamos cargando nuestros datos sobre la cobertura de noticias televisivas sobre política energética. Entonces creamos un**hist**El diagrama de esta serie temporal de historias mensuales cuenta con el hist mando. Primero, descargue los datos de Peake y Eshbaugh-Soha sobre la cobertura de la póliza energética, el archivo llamadoPESenergy.csv. El archivo está disponible en el Dataverse nombrado en la página vii o en el enlace de contenido del capítulo en la página 33. Es posible que deba usar setwd apuntar R a la carpeta donde ha guardado los datos. Después de esto, ejecute el siguiente código:

```
pres.energy <-read.csv ("PESenergy.csv") hist (pres.energyEnergy, xlab =
"Television Stories", main = "") abline (h = 0, col = 'gray60')

caja()
```

El resultado que produce este código se presenta en la Fig. 3.1. En este código, comenzamos leyendo Peake y Eshbaugh-Soha (2008) datos. El archivo de datos en sí es un archivo de valores separados por comas con una fila de encabezado de nombres de variables, por lo que los valores predeterminados de read.csv se adapte a nuestros propósitos. Una vez que se cargan los datos, trazamos un histograma de nuestra variable de interés usando elhist mando: pres.energy $ Energía llama a la variable de interés de su marco de datos. Usamos elxlab opción, que nos permite de fi nir la etiqueta R imprime en el eje horizontal. Dado que este eje nos muestra los valores de la variable, simplemente deseamos ver la frase “Historias de televisión”, describiendo brevemente lo que significan estos números. Laprincipal La opción de fi ne un título impreso en la parte superior de la fi gura. En este caso, la única forma de imponer un título en blanco es incluir citas sin contenido entre ellas. Una característica interesante de trazar en elbase el paquete es

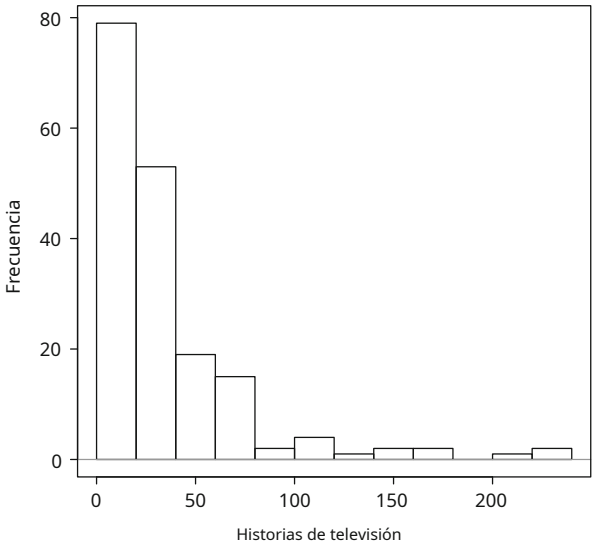


Figura 3.1 Histograma del recuento mensual de noticias de televisión relacionadas con la energía

que unos pocos comandos pueden agregar información adicional a un gráfico que ya se ha dibujado. Laabline El comando es una herramienta flexible y útil. (El nombre**línea ab** _ _ ____ se refiere a la fórmula lineal *y* D *a*C*bx.* Por lo tanto, este comando puede dibujar líneas con una pendiente e intersección, o puede dibujar una línea horizontal o vertical.) En este caso, abline agrega una línea horizontal a lo largo del punto 0 en el eje vertical, por lo tanto h = 0. Esto se agrega para aclarar dónde está la base de las barras en la figura. Finalmente, elcaja() El comando encierra la figura completa en una caja, a menudo útil en artículos impresos para aclarar dónde termina el espacio gráfico y comienza otro espacio en blanco. Como muestra el histograma, hay una fuerte concentración de observaciones en 0 y justo por encima de 0, y un claro sesgo positivo en la distribución. (De hecho, estos datos se vuelven a analizar en Fogarty y Monogan (2014) precisamente para abordar algunas de estas características de datos y discutir los medios útiles de analizar los recuentos de medios dependientes del tiempo).

Otro gráfico univariado es un diagrama de caja y bigotes. R nos permite obtener esto únicamente para la variable única, o para un subconjunto de la variable basado en alguna otra medida disponible. Primero dibujando esto para una sola variable:

```
boxplot (pres.energy $ Energy, ylab = "Historias de televisión")
```

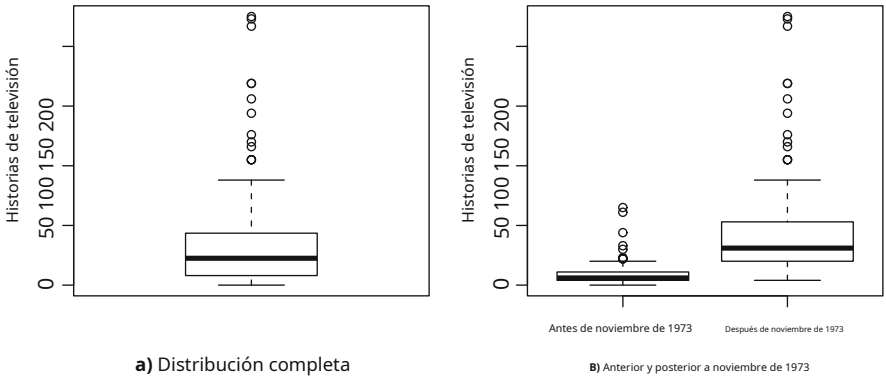
El resultado de esto se presenta en el panel (a) de la Fig. 3.2. En este caso, los valores de los recuentos mensuales están en el eje vertical; por lo tanto, usamos elylab opción para etiquetar el eje vertical (o **y-**eje **laboratorio**el) apropiadamente. En la figura, la parte inferior del cuadro representa el valor del primer cuartil (percentil 25), la línea sólida grande dentro del cuadro representa el valor mediano (segundo cuartil, percentil 50) y la parte superior del cuadro representa el valor del tercer cuartil (Percentil 75). Los bigotes, por defecto, se extienden a los valores más bajos y más altos de la variable que no son más de 1,5 veces el rango intercuartílico (o la diferencia entre el tercer y el primer cuartil) de distancia.

de la caja. El propósito de los bigotes es transmitir el rango sobre el que cae la mayor parte de los datos. Los datos que quedan fuera de este rango se representan como puntos en sus valores respectivos. Esta gráfica de caja se ajusta a nuestra conclusión del histograma: los valores pequeños que incluyen 0 son comunes y los datos tienen un sesgo positivo.

Los diagramas de caja y bigotes también pueden servir para ofrecer una idea de la distribución condicional de una variable. Para nuestra serie temporal de cobertura de la política energética, el primer evento importante que observamos es el discurso de Nixon de noviembre de 1973 sobre el tema. Por lo tanto, podríamos crear un indicador simple donde los primeros 58 meses de la serie (hasta octubre de 1973) se codifican con 0 y los 122 meses restantes de la serie (desde noviembre de 1973 en adelante) se codifican con 1. Una vez que hacemos esto, eldiagrama de caja El comando nos permite condicionar sobre una variable:

```
pres.energy $ post.nixon <-c (rep (0,58), rep (1,122)) boxplot
(pres.energy $ Energy ~ pres.energy $ post.nixon,
        ejes = F, ylab = "Historias de televisión") eje (1, en = c (1,2),
etiquetas = c ("Antes de noviembre de 1973",
        "Después de noviembre de 1973"))
eje (2)
caja()
```

Esta salida se presenta en el panel (b) de la Fig. 3.2. La primera línea de código define nuestra variable anterior a posterior a noviembre de 1973. Observe aquí que nuevamente definimos un vector conC. Dentro C, usamos el reps comando (para **reps**comer). Entoncesrep (0,58) produce 58 ceros, y rep (1,122) produce 122 unos. La segunda línea dibuja nuestros diagramas de caja, pero agregamos dos advertencias importantes en relación con nuestra última llamada adiagrama de caja: Primero, enumeramos pres.energy $ Energía~pres.energy $ post.nixon como nuestro argumento de datos. El argumento antes de la tilde (~) es la variable para la que queremos la distribución, y el argumento posterior es la variable condicionante. En segundo lugar, agregamos elejes = F mando. (También podríamos escribirejes = FALSO, pero R acepta F como abreviatura.) Esto nos da más control sobre cómo la horizontal y



a) Distribución completa

B) Anterior y posterior a noviembre de 1973

Figura 3.2 Tramas de caja y bigotes de la distribución del recuento mensual de nuevas historias televisivas relacionadas con la energía. Panel (**a**) muestra la distribución completa y el panel (**B**) muestra las distribuciones de los subconjuntos antes y después de noviembre de 1973

se presentan los ejes verticales. En el comando siguiente, agregamos el eje 1 (el eje horizontal inferior), agregando etiquetas de texto en las marcas de verificación 1 y 2 para describir los valores de la variable condicionante. Luego, agregamos el eje 2 (el eje vertical izquierdo) y un cuadro alrededor de toda la figura. El panel (b) de la Fig.3.2 muestra que la distribución antes y después de esta fecha es fundamentalmente diferente. Los valores mucho más pequeños persisten antes del discurso de Nixon, mientras que hay una media más grande y una mayor dispersión de valores después. Por supuesto, esto es solo una primera mirada y el efecto del discurso de Nixon se confunde con una variedad de factores, como el precio del petróleo, la aprobación presidencial y la tasa de desempleo, que contribuyen a esta diferencia.

3.1.1 Gráficos de barras

Los gráficos de barras pueden ser útiles siempre que queramos ilustrar el valor que toman algunas estadísticas para una variedad de grupos, así como para visualizar las proporciones relativas de datos medidos nominales u ordinalmente. Para ver un ejemplo de gráficos de barras, pasamos ahora al otro conjunto de datos de ejemplo de este capítulo, sobre el cabildeo por la salud en los 50 estados estadounidenses. Lowery y col. ofrecer un gráfico de barras de los medios en todos los estados de la tasa de participación en el cabildeo (o el número de cabilderos como porcentaje del número de empresas) para todos los cabilderos de salud y para siete subgrupos de cabilderos de salud (2008, Fig. 3). Podemos recrear esa figura enR tomando las medias de estas ocho variables y luego aplicando la gráfico de barras función al conjunto de medios. Primero debemos cargar los datos. Para hacer esto, descargue los datos de Lowery et al. Sobre cabildeo, el archivo llamado constructionData.dta. El archivo está disponible en el Dataverse nombrado en la página vii o en el enlace de contenido del capítulo en la página 33. Una vez más, es posible que deba usar setwd apuntar R a la carpeta donde ha guardado los datos. Dado que estos datos están en formato Stata, debemos utilizar elextranjero biblioteca y luego la read.dta mando:

```
biblioteca (extranjera)
health.fin <-read.dta ("constructionData.dta")
```

Para crear la figura real en sí, podemos crear un subconjunto de nuestros datos que solo incluya los ocho predictores de interés y luego usar el solicitar función para obtener la media de cada variable.

```
part.rates <-subset (health.fin, select = c (
      partratetotalhealth, partratedpc, partratepharmprod, partrateprofessionals,
    partrateadvo, partratebusness, partrategov, rnmedschoolpartrate))
    lobby.means <-apply (part.rates, 2, mean)

nombres (lobby.means) <- c ("Total Health Care",
      "Atención directa al paciente", "Medicamentos / productos de
      salud", "Profesionales de la salud", "Defensa de la salud", "
      Finanzas sanitarias "," Gobierno local "," Educación sanitaria ")
```

En este caso, tasas de part. es nuestro marco de datos subconjunto que solo incluye las ocho tasas de interés de participación del lobby. En la última línea, elsolicitar El comando nos permite tomar una matriz o un marco de datos (tasas parciales) y aplicar una función de interés (significar) a las filas o columnas del marco de datos. Queremos la media de

cada variable, y las columnas de nuestro conjunto de datos representan las variables. La2 ese es el segundo componente de este comando, por lo tanto, dice solicitar que queremos aplicar significar hacia *columnas* de nuestros datos. (Por el contrario, un argumento de1 se aplicaría a la *filas.* Los cálculos basados en filas serían útiles si tuviéramos que calcular alguna cantidad nueva para cada uno de los 50 estados) .Si simplemente escribimos lobby.medios en el R consola ahora, imprimirá los ocho medios de interés para nosotros. Para configurar nuestra cifra de antemano, podemos adjuntar un nombre en inglés a cada cantidad que se informará en el margen de nuestra cifra. Hacemos esto con elnombres comando, y luego asigne un vector con un nombre para cada cantidad.

Para dibujar realmente nuestro gráfico de barras, usamos el siguiente código:

```
par (mar = c (5.1, 10, 4.1, 2.1)) barplot (lobby.means, xlab = "Porcentaje de
registro en el lobby",
        xlim = c (0,26), horiz = T, cex.names = .8, las = 1) text (x =
lobby.means, y = c (.75,1.75,3,4.25,5.5,6.75,8 , 9),
        etiquetas = pegar (redondo (lobby.means, 2)), pos = 4)
caja()
```

Los resultados se representan en la Fig. 3.3. La primera línea llama alpar comando, que permite al usuario cambiar una amplia gama de valores predeterminados en el espacio gráfico. En nuestro

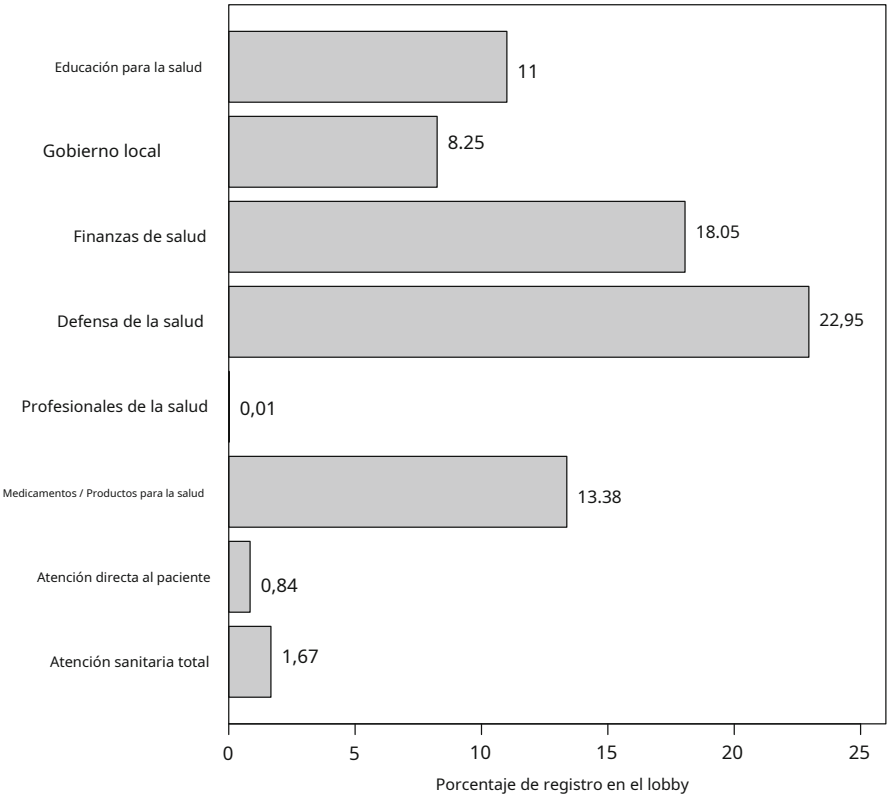


Figura 3.3 Gráfico de barras de la tasa media de participación de los grupos de presión en la atención médica y en siete subgremios en los 50 estados de EE. UU., 1997

caso, necesitamos un margen izquierdo más grande, así que usamos el mar para cambiar esto, estableciendo el segundo valor en el valor relativamente grande de 10. (En general, los márgenes se enumeran como inferior, izquierdo, superior y luego derecho). par se restablece a los valores predeterminados después de cerrar la ventana de trazado (o dispositivo, si se escribe directamente en un archivo). A continuación, usamos elgráfico de barras mando. El argumento principal es Lobby significa que es el vector de medias variables. El predeterminado paragráfico de barras consiste en dibujar un gráfico con líneas verticales. En este caso, sin embargo, configuramos la opciónhoriz = T para obtener barras horizontales. También usamos las opcionescex.names (**C**personaje **ex**pansion por eje **nombres)** y las = 1 (**l**Abel **a**xis **s**tyle) para encoger las etiquetas de nuestras barras al 80% de su tamaño predeterminado y obligarlas a imprimir horizontalmente, respectivamente.[3] La xlab El comando nos permite describir la variable para la que estamos mostrando las medias, y la xlim**X-**eje **lim**su) comando nos permite establecer el espacio de nuestro eje horizontal. Finalmente, usamos eltexto comando para imprimir la media de cada tasa de registro de lobby al final de la barra. Latexto El comando es útil cada vez que deseamos agregar texto a un gráfico, ya sean valores numéricos o etiquetas de texto. Este comando tomaX coordenadas para su posición a lo largo del eje horizontal, y coordenadas para su posición a lo largo del eje vertical, y etiquetas valores para que el texto se imprima en cada lugar. Lapos = 4 La opción especifica imprimir el texto a la derecha del punto dado (alternativamente, 1, 2 y 3 especificarían abajo, izquierda y arriba, respectivamente), para que nuestro texto no se superponga con la barra.

3.2 El gráfico Función

Pasamos ahora a gráfico, la función gráfica caballo de batalla en el base paquete. La gráfico El comando se presta naturalmente a parcelas bivariadas. Para ver la suma total de argumentos a los que se puede llamar usandográfico, tipo args (plot.default), que devuelve lo siguiente:

función (x, y = NULL, type = "p", xlim = NULL, ylim = NULL,
log = "", main = NULL, sub = NULL, xlab = NULL, ylab = NULL, ann = par ("ann"), axes = TRUE, frame.plot = axes, panel.first = NULL, panel.last = NULL, asp = NA, ...)

Obviamente, están sucediendo muchas cosas debajo del genérico gráfico función. Con el fin de comenzar con la creación de figuras enR queremos preguntarnos qué es lo esencial. La respuesta es sencilla: una variableX debe especificarse. Todo lo demás tiene un valor predeterminado o no es esencial. Para empezar a experimentar congráfico, Seguimos utilizando los datos de cabildeo de salud estatal de 1997 cargados en la Sect. 3.1.1. Con gráfico, podemos trazar las variables por separado con el comando plot (varname), aunque esto es definitivamente menos informativo que los tipos de

[3] El valor por defecto las El valor es 0, que imprime etiquetas paralelas al eje. 1, nuestra elección aquí, los imprime horizontalmente. 2 imprime perpendicularmente al eje y 3 imprime verticalmente.

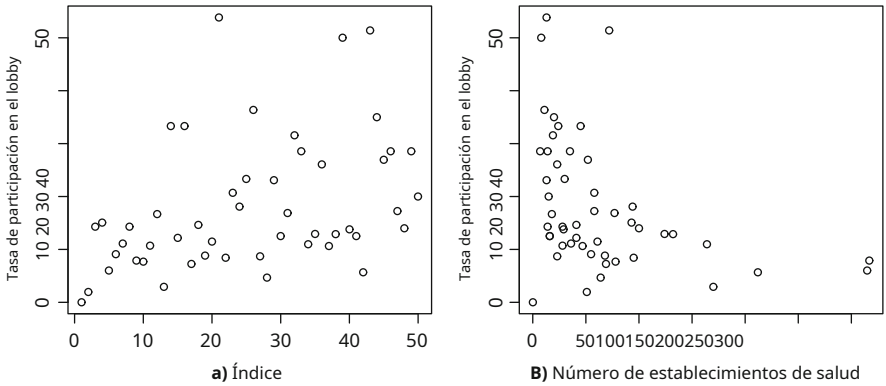


Figura 3.4 Tasa de participación del lobby de las finanzas de salud industria sola y contra el número de establecimientos comerciales de financiación de la salud. (**a)** Índice. (**B)** Número de establecimientos de salud

gráficos que se acaban de presentar en la secc. 3.1. Dicho esto, si simplemente quisiéramos ver todos los valores observados de la tasa de participación del lobby por empresas financieras del estado de salud (partratebusness), simplemente escribimos:

```
plot (health.fin $ partratebusness,
      ylab = "Tasa de participación en el lobby")
```

Figura 3.4a se devuelve en el R interfaz gráfica. Tenga en cuenta que esta figura traza la tasa de participación del lobby contra el número de fila en el marco de datos: con datos transversales, este índice es esencialmente insignificante. Por el contrario, si estuviéramos estudiando datos de series de tiempo y los datos se clasificaran a tiempo, podríamos observar cómo evoluciona la serie a lo largo del tiempo. Tenga en cuenta que usamos elylab opción porque, de lo contrario, el valor predeterminado etiquetará nuestro eje vertical con el aspecto pegajoso health.fin $ partratebusness. (Pruébelo y pregúntese qué pensaría el editor de una revista sobre cómo se ve el resultado).

Por supuesto, estamos más interesados en las relaciones bivariadas. Podemos explorarlos fácilmente incorporando una variableX en el eje horizontal (normalmente un *independiente* variable) y una variable y en el eje vertical (normalmente un *dependiente* variable) en la llamada a graficar:

```
plot (y = health.fin $ partratebusness, x = health.fin $ supplybusiness,
      ylab = "Tasa de participación en el lobby", xlab =
      "Número de establecimientos de salud")
```

Esto produce la Fig. 3.4b, donde nuestro eje horizontal se define por el número de empresas de financiación de la salud en un estado, y el eje vertical se define por la tasa de participación del lobby de estas empresas en el estado respectivo. Este gráfico muestra lo que parece ser una disminución en la tasa de participación a medida que aumenta el número de empresas, quizás en una relación curvilínea.

Una herramienta útil es trazar la forma funcional de un modelo bivariado en el diagrama de dispersión de las dos variables. En el caso de la Fig.3.4b, es posible que queramos comparar

cómo una función lineal versus una función cuadrática o al cuadrado del número de empresas se ajusta al resultado de la tasa de participación del lobby. Para hacer esto, podemos ajustar dos modelos de regresión lineal, uno que incluye una función lineal de número de empresas y el otro que incluye una función cuadrática. Los detalles adicionales sobre los modelos de regresión se analizan más adelante en el Cap.6. Nuestros dos modelos en este caso son:

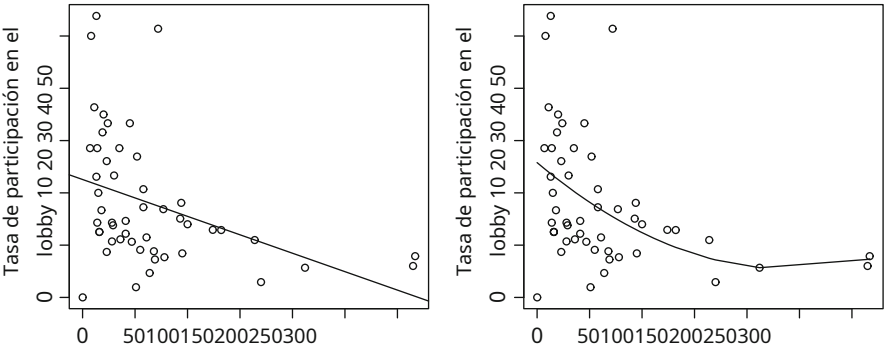
```
finance.linear <-lm (partratebusness ~ supplybusiness,
        datos = salud.fin)
resumen (finance.linear) finance.quadratic <-lm (partratebusness ~
supplybusiness +
        I (supplybusiness ^ 2), data = health.fin) resumen
        (finance.quadratic)
```

La lm**l**en el oido **metro**odel) se ajusta a nuestros modelos, y el resumen comando resume nuestros resultados. Nuevamente, detalles delm se discutirá en el Cap. 6. Con el modelo que es una función lineal del número de empresas, podemos simplemente introducir el nombre de nuestro modelo ajustado (Finance.linear) en el comando abline para agregar nuestra línea de regresión ajustada al gráfico:

```
plot (y = health.fin $ partratebusness, x = health.fin $ supplybusiness,
        ylab = "Tasa de participación en el lobby", xlab =
    "Número de establecimientos de salud") abline
    (finance.linear)
```

Como se mencionó antes, el abline El comando es particularmente fl exible. Un usuario puede especificara como la intersección de una línea y B como la pendiente. Un usuario puede especificarh como el valor del eje vertical donde se dibuja una línea horizontal, o v como el valor del eje horizontal donde se dibuja una línea vertical. O, en este caso, se puede insertar un modelo de regresión con un predictor para dibujar la línea de regresión que mejor se ajuste. Los resultados se presentan en la fig.3,5una.

Alternativamente, podríamos volver a dibujar este gráfico con la relación cuadrática esbozada en él. Desafortunadamente, a pesar deablineflexibilidad, no puede dibujar una cuadrática



a) Número de establecimientos de salud **B)** Número de establecimientos de salud

Figura 3.5 Tasa de participación del lobby de la industria financiera de la salud frente al número de establecimientos de salud, modelos lineales y cuadráticos. (**a)** Función lineal. (**B)** Función cuadrática

relación por defecto. La forma más fácil de trazar una forma funcional compleja es guardar los valores predichos del modelo, reordenar los datos según el predictor de interés y luego usar ellíneas función para agregar una línea conectada de todas las predicciones. Asegúrese de que los datos estén ordenados correctamente en el predictor; de lo contrario, la línea aparecerá como un desorden. El código en este caso es:

```
plot (y = health.fin $ partratebusness, x = health.fin $ supplybusiness,
      ylab = "Tasa de participación en el lobby", xlab = "Número de
    establecimientos de salud") finance.quadratic <-lm (partratebusness ~
    supplybusiness +
      I (supplybusiness ^ 2), data = health.fin) health.fin $ quad.fit <-
finance.quadratic $ fit.values   health.fin <-health.fin [order (health.fin $
supplybusiness),] líneas ( y = salud.fin $ quad.fit, x = salud.fin $ negocio de
suministros)
```

Este resultado se presenta en la Fig. 3,5B. Si bien no nos meteremos enlmlos detalles aún, tenga en cuenta que Yo (negocio de suministro ^ 2) se utiliza como predictor. I significa "como **I**s ", por lo que nos permite calcular una fórmula matemática sobre la marcha. Después de volver a dibujar nuestro diagrama de dispersión original, estimamos nuestro modelo cuadrático y guardamos los valores ajustados en nuestro marco de datos como la variablequad.fit. En la cuarta línea, reordenamos nuestro marco de datos health.fin según los valores de nuestra variable de entrada Supplybusiness. Esto se hace usando el pedido comando, que enumera los índices vectoriales en orden de valor creciente. Finalmente, ellíneas El comando toma nuestros valores predichos como coordenadas verticales (y) y nuestros valores del número de empresas como coordenadas horizontales (X). Esto agrega la línea al gráfico que muestra nuestra forma funcional cuadrática.

3.2.1 Gráficos de líneas con gráfico

Hasta ahora, nuestros análisis se han basado en gráfico predeterminado de dibujar un diagrama de dispersión. Sin embargo, en el análisis de series de tiempo, un gráfico de líneas a lo largo del tiempo suele ser útil para observar las propiedades de la serie y cómo cambia con el tiempo. (Más información sobre esto está disponible en el Cap.9.) Volviendo a los datos sobre la cobertura informativa televisiva de la política energética planteados por primera vez en el art. 3.1, visualicemos el resultado de la cobertura de la política energética y un insumo del precio del petróleo.

Comenzando con la cantidad de historias de energía por mes, creamos esta trama de la siguiente manera:

```
plot (x = pres.energy $ Energy, type = "l", axes = F,
      xlab = "Mes", ylab = "Historias de televisión sobre energía") eje (1, at = c
(1,37,73,109,145), labels = c ("Enero de 1969",
      "Ene. 1972", "Ene. 1975", "Ene. 1978", "Ene. 1981"), eje cex = .7)

eje (2)
abline (h = 0, col = "gray60")
cuadro ()
```

Esto produce la Fig. 3.6una. En este caso, nuestros datos ya están ordenados por mes, por lo que si solo especificamosX sin y, R mostrará todos los valores en tiempo correcto

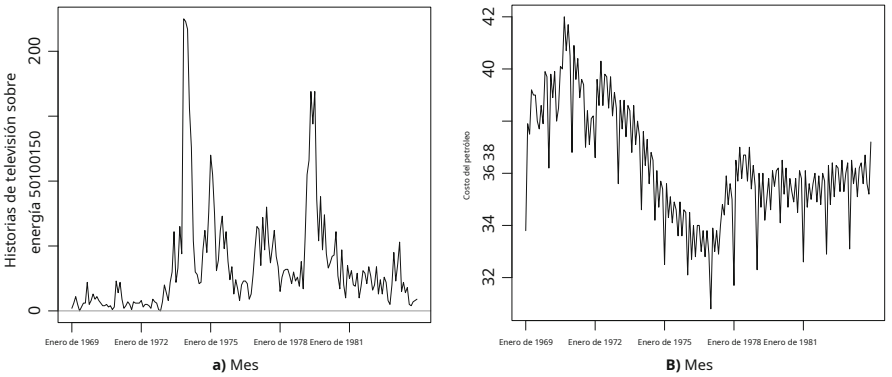


Figura 3.6 Número de reportajes televisivos sobre política energética y, el precio del petróleo por barril, respectivamente, por mes. (**a**) Cobertura de noticias. (**B**) Precio del petróleo

pedido.[4] Para designar que queremos un diagrama de líneas en lugar de un diagrama de dispersión de puntos, insertamos la letra l en el tipo = "l" opción. En este caso, hemos desactivado los ejes porque las marcas de graduación predeterminadas para el mes no son particularmente significativas. En su lugar, usamos eleje comando para insertar una etiqueta para el primer mes del año cada 3 años, ofreciendo una mejor sensación del tiempo real. Observe que en nuestra primera llamada aeje, usamos el eje cex. opción de encoger nuestras etiquetas al 70% de tamaño. Esto permite que las cinco etiquetas quepan en el gráfico. (Por ensayo y error, verá queR deja caer etiquetas de eje que no encajarán en lugar de sobreimprimir el texto.) Finalmente, usamos abline para mostrar el punto cero en el eje vertical, ya que se trata de un número significativo que re fl eja la ausencia total de cobertura de política energética en los informativos televisivos. Como demostraron nuestras cifras anteriores, vemos mucha más variabilidad y una media más alta después de los primeros 4 años. La cifra del precio del petróleo por barril se puede crear de manera similar:

```
plot (x = pres.energy $ oilc, type = "l", axes = F, xlab = "Month",
        ylab = "Costo del petróleo")
eje (1, en = c (1,37,73,109,145), etiquetas = c ("Enero de 1969",
        "Ene. 1972", "Ene. 1975", "Ene. 1978", "Ene. 1981"), eje cex = .7)

eje (2)
caja()
```

Nuevamente, los datos están ordenados, por lo que solo se necesita una variable. Figura3.6b presenta este gráfico.

[4]Alternativamente, sin embargo, si un usuario tuviera algún índice de tiempo en el marco de datos, se podría producir un gráfico similar escribiendo algo con el efecto de: pres.energy $ Time <-1: 180; plot (y = pres.energy $ Energy, x = pres.energy $ Time, type = "l").

3.2.2 Construcción de figuras con gráfico: Detalles adicionales

Habiendo probado nuestra suerte con las tramas del base paquete, ahora detallaremos en detalle las funciones y opciones básicas que aportan una flexibilidad considerable a la creación de figuras en R. Tener en cuenta que R en realidad ofrece la opción útil de comenzar con una pizarra en blanco y agregar elementos al gráfico bit a bit.

El sistema de coordenadas: En la Fig. 3.4, no estábamos preocupados por establecer el sistema de coordenadas porque los datos efectivamente lo hicieron por nosotros. Pero a menudo, querrá establecer las dimensiones de la figura antes de trazar cualquier cosa, especialmente si está construyendo a partir del lienzo en blanco. El punto más importante aquí es que suX y y debe tener la misma longitud. Esto es quizás obvio, pero los datos faltantes pueden crear dificultades que conduciránR resistirse.

Tipos de parcela: Ahora queremos trazar estas series, pero el gráfico función permite diferentes tipos de parcelas. Los diferentes tipos que se pueden incluir dentro del genéricográfico la función incluye:

- tipo = "p" Este es el valor predeterminado y traza el X y y coordenadas como *puntos.*
- tipo = "l" Esto traza el X y y coordenadas como *líneas.*
- tipo = "n" Esto traza el X y y coordenadas como *nada (*configura el espacio de coordenadas solamente).
- tipo = "o" Esto traza el X y y coordenadas como *puntos y lineas* superpuesto (es decir, se "superpone").
- tipo = "h" Esto traza el X y y coordenadas como *líneas verticales en forma de histograma.* (También llamado *trama de picos.)*
- type = "s" Esto traza el X y y coordenadas como *escalones como líneas.*

Ejes: Es posible apagar los ejes, ajustar el espacio de coordenadas usando el xlim y ylim opciones y para crear sus propias etiquetas para los ejes.

ejes = Le permite controlar si los ejes aparecen en la figura o no. Si tiene fuertes preferencias sobre cómo se crean sus ejes, puede desactivarlos seleccionando ejes = F dentro gráfico y luego cree sus propias etiquetas usando la etiqueta separada eje mando:

- eje (lado = 1, en = c (2, 4, 6, 8, 10, 12), etiquetas = c ("febrero", "abril", "junio", "agosto", "octubre", "diciembre")))

xlim =, ylim = Por ejemplo, si quisiéramos ampliar el espacio desde el R por defecto, podríamos ingresar:

- plot (x = ind.var, y = dep.var, type = "o", xlim = c (-5, 17), ylim = c (-5, 15))

xlab = "", ylab = "" Crea etiquetas para los ejes x e y.

Estilo: Hay varias opciones para ajustar el estilo en la figura, que incluyen cambios en el tipo de línea, grosor de línea, color, estilo de punto y más. Algunos comandos comunes incluyen:

asp = De fi ne el **áspid**ect relación de la parcela. Configuraciónasp = 1 es un poderoso y útil opción que permite al usuario declarar que los dos ejes se miden en la misma escala. Ver Fig.5.1 en la página 76 y Fig. 8.4 en la página 153 como dos ejemplos de esta opción.

lty = Selecciona el tipo de línea (sólida, discontinua, guión corto-largo, etc.).

lwd = Selecciona el ancho de la línea (líneas gruesas o delgadas).

pch = Selecciona el símbolo de trazado, puede ser un símbolo numerado (pch = 1) o una cartapch = "D"). col =Selecciona el color de las líneas o puntos de la figura.

cex = **C**personaje **ex**factor de expansión que ajusta el tamaño del texto y los símbolos en la figura. Similar,eje cex. ajusta el tamaño de la anotación del eje, cex.lab ajusta el tamaño de la fuente para las etiquetas de los ejes, cex.main ajusta el tamaño de fuente del título y cex.sub ajusta el tamaño de fuente de los subtítulos.

Parámetros gráficos: La par La función aporta funcionalidad adicional al trazado. en R dando al usuario control sobre los gráficos **par**ametros. Una característica notable depar es que le permite trazar múltiples llamadas a gráfico en un solo gráfico. Esto se logra seleccionando par (nuevo = T) mientras una ventana de trazado (o dispositivo) todavía está abierta y antes de la próxima llamada a gráfico. Ser *Cuidado,* aunque. Siempre que utilice esta estrategia, incluya la xlim y ylim comandos en cada llamada para asegurarse de que el espacio de la gráfica se mantenga igual. También tenga cuidado de que los márgenes del gráfico no cambien de una llamada a la siguiente.

3.2.3 Funciones complementarias

También hay una serie de funciones complementarias que se pueden utilizar una vez que se ha creado el sistema de coordenadas básico utilizando gráfico. Éstas incluyen:

flechas (x1, y1, x2, y2) Cree flechas dentro del gráfico (útil para etiquetas puntos de datos particulares, series, etc.).

texto (x1, x2, "texto") Cree texto dentro de la trama (modifique el tamaño del texto usando la opción de expansión de personajes cex).

lineas (x, y)Crea una gráfica que conecte líneas.

puntos (x, y) Crea una gráfica de puntos.

polígono() Crea un polígono de cualquier forma (rectángulos, triángulos, etc.).

leyenda (x, y, at = c ("", ""), etiquetas = c ("", "")) Crear a leyenda para identificar los componentes de la figura.

eje (lateral) Agregue un eje con etiquetas predeterminadas o personalizadas a uno de los lados de una parcela. Establezca el lado en 1 para la parte inferior, 2 para la izquierda, 3 para la parte superior y 4 para la derecha.

mtext (texto, lado) Comando para agregar **metro**argumentación **texto.** Esto le permite agregar un eje etiqueta en uno de los lados con más control sobre cómo se presenta la etiqueta. Vea el código que produce la Fig.7.1 en la página 108 para un ejemplo de esto.

3.3 Utilizando enrejado Gráficos en R

Como alternativa al base paquete de gráficos, es posible que desee considerar el enrejado paquete complementario. Estos producenconducción gráficos de la S lenguaje, que tienden a mostrar mejor los datos agrupados y numerosas observaciones. Algunas características interesantes delenrejado paquete es que los gráficos tienen valores predeterminados amigables para el espectador, y los comandos ofrecen una datos = opción que no requiere que el usuario enumere el marco de datos con cada llamada a una variable.

Para empezar, la primera vez que usamos el enrejado biblioteca, debemos instalarla. Luego, en cada reutilización del paquete, debemos llamarlo con elBiblioteca mando.

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

Para obtener una gráfica de dispersión similar a la que dibujamos con gráfico, esto se puede lograr en enrejado utilizando la xyplot mando:

```
xyplot (partratebusness ~ supplybusiness, data = health.fin,
        col = "black", ylab = "Tasa de participación en el lobby", xlab =
        "Número de establecimientos de salud")
```

Figura 3,7a muestra este gráfico. La sintaxis difiere de lagráfico funcionar de alguna manera: en este caso, podemos especificar una opción, datos = salud.fin, que nos permite escribir el nombre del marco de datos relevante una vez, en lugar de volver a escribirlo para cada variable. Además, ambas variables se enumeran juntas en un solo argumento utilizando el formulario,vertical.variable˜horizontal.variable. En este caso, también especificamos la opción, col = "negro" con el fin de producir una figura en blanco y negro. Por defectoenrejado los colores dan como resultado cian para permitir a los lectores separar fácilmente la información de datos de otros aspectos de la pantalla, como ejes y etiquetas (Becker et al. 1996, pag. 153). Además, de forma predeterminada,xyplot imprime marcas de graduación en el tercer y cuarto eje para proporcionar puntos de referencia adicionales para el espectador.

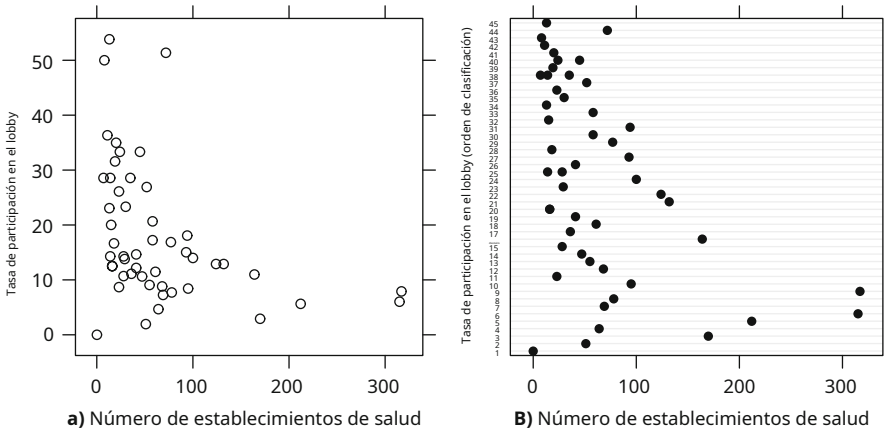
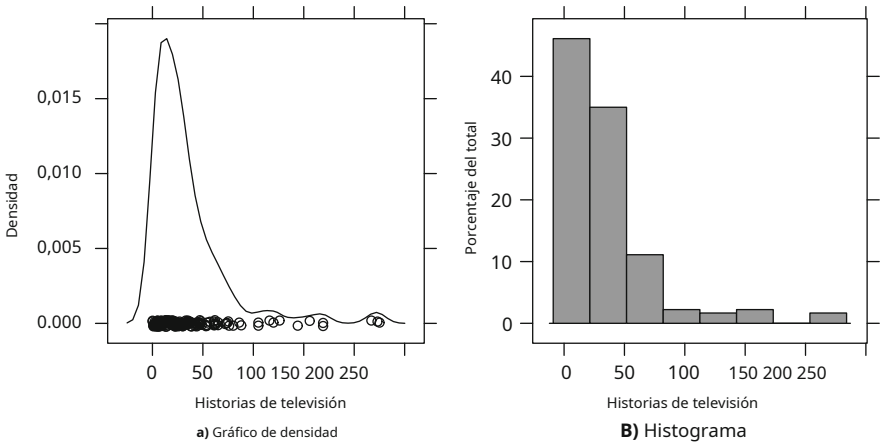


Figura 3.7 Tasa de participación del lobby de la industria financiera de la salud en comparación con el número de establecimientos de salud, (**a)** diagrama de dispersión y (**B)** Gráfica de puntos



a) Gráfico de densidad **B)** Histograma

Figura 3.8 (a) Gráfico de densidad y (**B)** histograma que muestra la distribución univariante del recuento mensual de noticias de televisión relacionadas con la energía

La enrejado El paquete también contiene funciones que dibujan gráficos que son similares a un diagrama de dispersión, pero en su lugar usan un orden de clasificación de la variable del eje vertical. Así es como eltrazar y Gráfica de puntos los comandos funcionan y ofrecen otra visión de una relación y su solidez. LaGráfica de puntos El comando puede ser algo más deseable ya que también muestra una línea para cada valor ordenado por rango, ofreciendo una sensación de que la escala es diferente. LaGráfica de puntos la sintaxis se ve así:

```
dotplot (partratebusness ~ supplybusiness,
        datos = salud.fin, col = "negro",
        ylab = "Tasa de participación en el lobby (orden de clasificación)",
        xlab = "Número de establecimientos de salud")
```

Figura 3,7b muestra este resultado. Latrazar La función utiliza una sintaxis similar.

Por último, el enrejado la biblioteca nuevamente nos da la opción de mirar la distribución de una sola variable trazando un histograma o un gráfico de densidad. Volviendo a los datos de la serie temporal presidencial que cargamos por primera vez en la Secta.3.1, ahora podemos dibujar una gráfica de densidad usando la siguiente línea de código:

```
diagrama de densidad (~ Energía, datos = energía pres.,
        xlab = "Historias de televisión", col = "negro")
```

Esto se presenta en la Fig. 3.8una. Esta salida muestra puntos dispersos a lo largo de la base, cada uno de los cuales representa el valor de una observación. La línea suavizada a lo largo del gráfico representa la densidad relativa estimada de los valores de la variable.

Alternativamente, un histograma en enrejado se puede dibujar con el histograma función:

```
histograma (~ Energía, datos = energía pres.,
        xlab = "Historias de televisión", col = "gray60")
```

Esto está impreso en la Fig. 3.8B. En este caso, el color se establece encol = "gray60". De nuevo, el valor predeterminado es para barras de color cian. Para una buena opción de escala de grises en

en este caso, un gris medio aún permite distinguir claramente cada barra. Una característica final interesante dehistograma Se deja al lector: la función dibujará distribuciones de histograma condicionales. Si todavía tienes el post.nixon variable disponible que creamos anteriormente, puede intentar escribir histograma (˜Energy | post.nixon, data = pres.energy), donde la tubería vertical (|) es seguida por la variable de acondicionamiento.

3.4 Salida gráfica

Un último punto esencial es una palabra sobre cómo los usuarios pueden exportar sus R gráficos en un procesador de texto o editor de escritorio deseado. La primera opción es guardar la salida de pantalla de una fi gura. En máquinas Mac, el usuario puede seleccionar la ventana de salida de la figura y luego usar el menú desplegable*Archivo! Guardar como. .* .para guardar la figura como un archivo PDF. En máquinas con Windows, un usuario puede simplemente hacer clic derecho en la ventana de salida de la figura y luego elegir guardar la figura como un metaarchivo (que se puede usar programas como Word) o como un archivo postscript (para usar en LATEXAS). También al hacer clic con el botón derecho en Windows, los usuarios pueden copiar la imagen y pegarla en Word, PowerPoint, o un programa de gráficos.

Una segunda opción permite a los usuarios una mayor precisión sobre el producto final. Específicamente, el usuario puede escribir el gráfico en un dispositivo gráfico, del cual hay varias opciones. Por ejemplo, al escribir este libro, exporté la Fig.3,5a escribiendo:

```
postscript ("lin.partrate.eps", horizontal = FALSE, ancho = 3,
        height = 3, onefile = FALSE, paper = "special", pointsize = 7) plot (y = health.fin $
partratebusness, x = health.fin $ supplybusiness,
        ylab = "Tasa de participación en el lobby", xlab =
    "Número de establecimientos de salud") abline
    (finance.linear)
dev.off ()
```

La primera línea llama al posdata comando, que creó un archivo llamado lin.partrate.eps que guardé el gráfico como. Entre las opciones clave de este comando se encuentranancho y altura, cada uno de los cuales puse a tres pulgadas. La tamaño de puntos El comando encogió el texto y los símbolos para encajar perfectamente en el espacio que asigné. Lahorizontal El comando cambia la orientación del gráfico de horizontal a vertical en la página. Cámbielo aCIERTO para que el gráfico adopte una orientación horizontal. Una vezposdata se llamó, todos los comandos de gráficos se escribieron en el archivo *y no a la ventana de gráficos.* Por lo tanto, suele ser una buena idea perfeccionar un gráfico antes de escribirlo en un dispositivo gráfico. Por lo tanto, lagráfico y abline los comandos servían para escribir toda la salida en el archivo. Una vez que terminé de escribir en el archivo, eldev.off () El comando cerró el archivo para que ningún otro comando gráfico pudiera escribir en él.

Por supuesto, los escritores que utilizan los gráficos postscript son los más utilizados lenguaje de autoedición de LATEXAS. Los escritores que utilizan procesadores de texto más tradicionales, como Word o Pages, querrán utilizar otros dispositivos gráficos. La

las opciones disponibles incluyen: jpeg, pdf, png, y pelea.[5] Para utilizar cualquiera de estos cuatro dispositivos gráficos, sustituya una llamada por la función correspondiente donde posdata está en el código anterior. ¿Asegúrate de escribir?png para tener una idea de la sintaxis de estos dispositivos alternativos, ya que cada uno de los cinco tiene una sintaxis ligeramente diferente.

Como circunstancia especial, los gráficos extraídos de la enrejado paquete utiliza un dispositivo gráfico diferente, llamado trellis.device. Es técnicamente posible utilizar los otros dispositivos gráficos para escribir en un archivo, pero no es aconsejable porque las opciones del dispositivo (por ejemplo, el tamaño del gráfico o el tamaño de la fuente) no se pasarán al gráfico. En el caso de la Fig. 3,7b, generé la salida usando el siguiente código:

```
trellis.device ("postscript", file = "dotplot.partrate.eps",
        tema = lista (tamaño de fuente = lista (texto = 7, puntos = 7)),
        horizontal = FALSO, ancho = 3, alto = 3,
        onefile = FALSE, paper = "especial") dotplot
(partratebusness ~ supplybusiness,
        datos = salud.fin, col = 'negro',
        ylab = "Tasa de participación en el lobby (orden de clasificación)",
    xlab = "Número de establecimientos de salud") dev.off ()
```

El primer argumento de la trellis.device El comando declara qué controlador desea utilizar el autor. ademásposdata, el autor puede usar jpeg, pdf, o png. El segundo argumento enumera el archivo en el que escribir. El tamaño de la fuente y del carácter debe establecerse a través deltema opción, y los argumentos restantes declaran las otras preferencias sobre la salida.

Este capítulo ha cubierto funciones gráficas univariadas y bivariadas en R. Varios comandos de ambos base y enrejado Se han abordado los paquetes. Esto está lejos de ser una lista exhaustiva deRcapacidades de creación de gráficos, y se anima a los usuarios a conocer más sobre las opciones disponibles. Sin embargo, este manual debe servir para presentar a los usuarios varios medios por los cuales los datos pueden visualizarse enR. Con un buen sentido de cómo tener una idea visual de los atributos de nuestros datos, el siguiente capítulo se centra en resúmenes numéricos de nuestros datos recopilados a través de estadísticas descriptivas.

3.5 Problemas de práctica

Además de su análisis de la cobertura de la política energética presentado en este capítulo, Peake y Eshbaugh-Soha (2008) también estudian la cobertura de la póliza de medicamentos. Estos datos cuentan de manera similar el número de noticias de televisión nocturnas en un mes centradas en las drogas, desde enero de 1977 hasta diciembre de 1992. Sus datos se guardan en formato separado por comas en el archivo denominadodrugCoverage.csv. Descargue sus datos del Dataverse mencionado en la página vii o del enlace de contenido del capítulo en la página 33. Las variables en este conjunto de datos son: un índice de tiempo basado en caracteres que muestra el mes y el año

[5] Mi experiencia personal indica que png a menudo se ve bastante claro y es versátil.

(**Año),** cobertura de noticias de drogas**drugmedia),** un indicador de un discurso sobre drogas que pronunció Ronald Reagan en septiembre de 1986 (**rwr86),** un indicador de un discurso pronunciado por George HW Bush en septiembre de 1989 (**ghwb89),** el índice de aprobación del presidente **aprobación),** y la tasa de desempleo (**desempleo).**

1. Dibuje un histograma del recuento mensual de historias relacionadas con las drogas. Puede utilizar cualquiera de los comandos de histograma descritos en el capítulo.
2. Dibuje dos diagramas de caja: uno de historias relacionadas con las drogas y otro de aprobación presidencial. ¿En qué se diferencian estas cifras y qué le dice eso sobre el contraste entre las variables?

3. Dibuja dos diagramas de dispersión:

   (a) En el primero, represente el número de historias relacionadas con las drogas en el eje vertical y coloque la tasa de desempleo en el eje horizontal.
   (b) En el segundo, represente el número de historias relacionadas con las drogas en el eje vertical y coloque la aprobación presidencial en el eje horizontal.
   (c) ¿En qué se diferencian las gráficas? ¿Qué te dicen de los datos?
   (d) Bonificación: agregue una línea de regresión lineal a cada uno de los diagramas de dispersión.

4. Dibuja dos gráficos de líneas:

   (a) En el primero, dibuje el número de historias relacionadas con las drogas por mes a lo largo del tiempo.
   (b) En el segundo, obtenga la aprobación presidencial por mes a lo largo del tiempo.
   (c) ¿Qué puedes aprender de estas gráficas?

5. Cargue el enrejado biblioteca y dibuje un diagrama de densidad del número de historias relacionadas con las drogas por mes.
6. Bonificación: Dibuje un gráfico de barras de la frecuencia de las tasas de desempleo observadas. (*Insinuación:* Intente usar el mesa comando para crear el objeto que graficará.) ¿Puede ir un paso más allá y dibujar un gráfico de barras del porcentaje de tiempo que se observa cada valor?