Capítulo 7
Modelos lineales generalizados

Si bien el modelo de regresión lineal es común a las ciencias políticas, muchas de las medidas de resultado que los investigadores desean estudiar son variables binarias, ordinales, nominales o de conteo. Cuando estudiamos estas variables dependientes limitadas, recurrimos a técnicas como regresión logística, regresión probit, regresión logit (y probit) ordenada, regresión logit (y probit) multinomial, regresión de Poisson y regresión binomial negativa. Se puede ver una revisión de estos y varios otros métodos en volúmenes como King (1989) y largo (1997).

De hecho, todas estas técnicas pueden considerarse como casos especiales de *modelo lineal generalizado,* o GLM (branquias 2001). El enfoque GLM en resumen es transformar la media de nuestro resultado de alguna manera para que podamos aplicar la lógica habitual del modelado de regresión lineal a la media transformada. De esta manera, podemos modelar una amplia clase de variables dependientes para las cuales la distribución de los términos de perturbación viola el supuesto de normalidad del teorema de Gauss-Markov. Además, en muchos casos, el resultado está acotado, por lo que *función de enlace* que usamos para transformar la media del resultado puede reflejar una forma funcional más realista (Gill 2001, págs. 31-32).

La glm comando en R es lo suficientemente flexible como para permitirle al usuario especificar muchos de los GLM más utilizados, como la logística y la regresión de Poisson. Un puñado de modelos que se usan con cierta regularidad, como el logit ordenado y la regresión binomial negativa, en realidad requieren comandos únicos que también cubriremos. En general, sin embargo, elglm El comando es un buen lugar para buscar primero cuando un investigador tiene una variable dependiente limitada. De hecho, elglm el comando toma un argumento llamado familia que permite al usuario especificar qué tipo de modelo desea para estimar. Escribiendo?familia en el R consola, el usuario puede obtener un rápido descripción general de qué modelos glm El comando puede estimar.

Electrónico suplementario material: La en línea versión de esto capítulo (doi: 10.1007 / 978-3-319-23446-5_7) 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_7

Este capítulo continúa ilustrando ejemplos con resultados binarios, resultados ordinales y resultados de conteo. Cada una de estas secciones utiliza un conjunto de datos de ejemplo diferente para considerar las variables dependientes de cada tipo. Cada sección presentará sus datos de ejemplo a su vez.

7.1 Resultados binarios

En primer lugar, consideramos las variables de resultado binarias o variables que toman sólo dos valores posibles. Por lo general, estos resultados se codifican con 0 o 1 para simplificar la interpretación. Como ejemplo en esta sección, usamos datos de encuestas del Estudio Comparativo de Sistemas Electorales (CSES). Singh2014a) estudia un subconjunto de estos datos que consta de 44.897 encuestados de 30 elecciones. Estas elecciones ocurrieron entre los años 1996-2006 en los países de Canadá, República Checa, Dinamarca, Finlandia, Alemania, Hungría, Islandia, Irlanda, Israel, Italia, Países Bajos, Nueva Zelanda, Noruega, Polonia, Portugal, Eslovenia, España, Suecia, Suiza y Reino Unido.

Singh utiliza estos datos para evaluar cómo la distancia ideológica determina la elección de voto de los individuos y su disposición a votar en primer lugar. Sobre la base del modelo espacial de la política propuesto por Hotelling (1929), Negro (1948), Downs (1957) y otros, el artículo muestra que las diferencias lineales en ideología explican mejor el comportamiento de los votantes que las diferencias cuadradas. Las variables del conjunto de datos son las siguientes:

votó: Indicador codificado 1 si el encuestado votó, 0 si no.
votadoinc: Indicador codificado con 1 si el encuestado votó por el partido en el poder, 0 si o votó por otro partido. (Faltan los no votantes).
cntryyear: Una variable de carácter que enumera el país y el año de la elección.
cntryyearnum: Un índice numérico que identifica el país y el año de la elección.
distanciainc: Distancia entre el encuestado y el titular en una escala de ideología de 0 a 10.
distanciaincsq: Distancia al cuadrado entre el votante y el partido en el poder.
ponderado a distancia: Distancia entre el encuestado y el más parecido partido político en una escala de ideología de 0 a 10, ponderado por la competitividad de la elección.
distanciasq ponderadas: Distancia ponderada al cuadrado entre el votante y la mayoría partido ideológico similar.

Los datos se guardan en formato Stata, por lo que necesitaremos cargar el extranjero Biblioteca. Descarga el archivoSinghJTP.dta desde el Dataverse vinculado en la página vii o el contenido del capítulo vinculado en la página 97. Luego, abra los datos de la siguiente manera:

```
biblioteca (extranjera)
votando <-read.dta ("SinghJTP.dta", convert.factors = FALSE)
```

Un buen paso inmediato aquí sería usar comandos como resumen así como gráficos para tener una idea de los atributos descriptivos de los datos. Esto queda para el lector.

7.1.1 Modelos Logit

Como primer modelo a partir de estos datos, modelaremos la probabilidad de que un encuestado haya votado por el partido en el poder, en lugar de por otro partido. Usaremos solo un predictor en este caso, y esa es la distancia entre el votante y el partido en el poder. Creamos esto como un*Regresión logística* modelo. La sintaxis de este modelo es:

```
inc. lineal <-glm (votadainc ~ distanciainc,
        familia = binomio (enlace = "logit"), datos = votación)
```

La sintaxis de glm**gramo**eneralizado **l**en el oido **metro**odel) es casi idéntica a lm: Todavía comenzamos con una especificación funcional que coloca la variable dependiente a la izquierda de la tilde (~) y los predictores a la derecha separados por signos más. Nuevamente, hacemos referencia a nuestro conjunto de datos con eldatos opción. Ahora, sin embargo, *deber* utilizar el familia opción para especificar qué GLM queremos estimar. Especificando binomial (enlace = "logit"), declaramos un resultado binario y que estamos estimando un modelo logit, en lugar de probit. Después de la estimación, escribiendo resumen (incluido lineal) obtenemos la salida de nuestro modelo de regresión logística, que es la siguiente:

```
Llamada:
glm (fórmula = voteinc ~ distanciainc,                familia = binomio
    (enlace = "logit"),
        datos = votación)

Desviación Derechos residuales de autor:
       Min          1T      Mediana          3T         Max
- 1.2608     - 0,8632     - 0.5570      1.0962      2.7519

Coeficientes:
                    Estimar Std. Error z valor Pr (> | z |)
(Interceptar)        0.19396    0.01880      10,32     <2e-16 ***
distanciainc      - 0,49469     0,00847    - 58,41     <2e-16 ***
- - -
Signif. códigos:          0    * * *      0,001     * *    0,01    * 0,05. 0,1
  1

(Parámetro de dispersión para          binomio      familia tomada como 1)

      Desviación nula: 47335          en 38210      grados de libertad
Desviación residual: 42910            en 38209      grados de libertad
    (6686 observaciones eliminadas por falta de información) AIC: 42914


Número de iteraciones de puntuación de Fisher: 4
```

Cuadro 7.1 Modelo logit de probabilidad de votar por el partido en el poder, 30 elecciones transnacionales

Vaticinador	Estimar	Std. error	z valor	Pr (> jzj)
Interceptar	0,1940	0,0188	10,32	0,0000
Distancia	0.4947	0,0085	58,41	0,0000

Notas: N D 38; 211. AIC D 42; 914. 69% predijo correctamente. Datos de Singh (2014a)

La impresión es similar a la impresión del modelo lineal que estimamos en el Cap. 6.[1] Puede encontrar una presentación más formal de nuestros resultados en la Tabla 7.1.[2]

Sin embargo, al comparar estos resultados con el modelo lineal, es importante tener en cuenta algunas diferencias. Primero, las estimaciones de coeficientes en sí mismas no son tan significativas como las del modelo lineal. Un modelo logit transforma nuestro resultado de interés, la probabilidad de votar por el partido en el poder, porque está acotado entre 0 y 1. La transformada logit vale la pena porque nos permite usar un marco de predicción lineal, pero requiere un paso adicional de esfuerzo por la interpretación. (Ver secc.7.1.3 para obtener más información sobre esto.) Una segunda diferencia en la salida es que informa z proporciones en lugar de t razones: al igual que antes, se calculan en torno a la hipótesis nula de que el coeficiente es cero, y la fórmula para la razón utiliza la estimación y el error estándar de la misma manera. Sin embargo, ahora debemos asumir que estas estadísticas siguen una distribución normal, en lugar de una t distribución.[3] En tercer lugar, se presentan diferentes estadísticas de ajuste: puntuaciones de desviación y el criterio de información de Akaike (AIC).[4]

En mesa 7.1, informamos los coeficientes, los errores estándar y la información inferencial mación. También reportamos el AIC, que es un buen índice de ajuste y tiene la característica de penalizar por el número de parámetros. a diferencia de R^2 sin embargo, en la regresión lineal, el AIC no tiene una métrica natural que dé un sentido absoluto de ajuste del modelo. Más bien, funciona mejor como medio de comparar modelos, con *más bajo* valores que indican un ajuste mejor penalizado. Para incluir una medida de ajuste que tenga una escala natural, también informamos qué porcentaje de respuestas predice correctamente nuestro modelo. Para calcular esto, todo lo que necesitamos hacer es determinar si el modelo predeciría un voto para el partido en el poder y compararlo con cómo votó realmente el encuestado. EnR, podemos rodar nuestro propio cálculo:

[1] En este caso, las estimaciones de coeficientes que obtenemos son similares a las reportadas por Singh (2014a). Sin embargo, nuestros errores estándar son más pequeños (y por lo tanto z y *pag* los valores son mayores) porque Singh agrupa los errores estándar. Ésta es una idea útil porque los encuestados están anidados dentro de las elecciones, aunque los modelos multinivel (que Singh también informa) también abordan este tema — ver Sect.8.1.

[2] LaLos usuarios de TEX pueden crear una tabla similar a esta rápidamente escribiendo: biblioteca (xtable); xtable (incluido lineal).

[3] Una explicación de cómo se derivan las propiedades inferenciales de este modelo se puede encontrar en Eliason (1993, págs. 26-27).

[4] La desviación se calcula como 2 veces la relación registrada entre la probabilidad ajustada y la saturada. probabilidad. Formalmente,2 Iniciar sesión $\frac{L_1}{L_2}$ dónde L_1 es la probabilidad ajustada y L_2 es la probabilidad saturada. R informa dos cantidades: el n2La desviación total calcula esto para un modelo de solo intercepto que siempre predice el valor modal, y la desviación residual calcula esto para el modelo informado.

```
predicho <-como numérico (
        predict.glm (inc.linear, type = "response")>. 5) verdadero <-
voting $ voteinc [vote $ vote == 1] correcto <-as.numeric (predicted
== true)
100 * tabla (correcta) / suma (tabla (correcta))
```

En la primera línea, creamos un vector de las predicciones del modelo. El código usa el predecir.glm comando, que puede pronosticar de manera útil a partir de cualquier modelo estimado con el glm mando. Especificandotype = "respuesta" aclaramos que queremos que nuestras predicciones estén en la escala de probabilidad (en lugar de la escala predeterminada de utilidad latente). Luego preguntamos si cada probabilidad es mayor que 0.5. Envolviendo todo esto en elas.numeric comando, contamos todas las probabilidades por encima de 0,5 como valores predichos de 1 (para el titular) y todos los que son inferiores a 0,5 como valores previstos de 0 (contra el titular). En la segunda línea, simplemente subconjuntamos el vector original del resultado de los datos originales con los que votaron y, por lo tanto, se incluyeron en el modelo. Este paso de subconjunto es esencial porque el glm El comando borra automáticamente los datos faltantes de la estimación. Por lo tanto, sin subconjuntos, nuestros valores predichos y verdaderos no se vincularían adecuadamente. En la tercera línea, creamos un vector codificado 1 si el valor predicho coincide con el valor verdadero, y en la cuarta línea creamos una tabla de este vector. La impresión es:

```
correcto
       0        1
30.99108 69.00892
```

Por lo tanto, sabemos que el modelo predice correctamente el 69% de los valores de resultado, que informamos en la Tabla 7.1.

Como un ejemplo más de regresión logística, Singh (2014a) compara un modelo con distancias ideológicas lineales a uno con distancias ideológicas al cuadrado. Para encajar en este modelo alternativo, escribimos:

```
inc. al cuadrado <-glm (votadainc ~ distanciaincsq,
        familia = binomio (enlace = "logit"), datos = votación)
resumen (incluido el cuadrado)
```

El resultado del comando de resumen en la segunda línea es:

```
Llamada:
glm (fórmula = votadainc ~ distanciaincsq, familia = binomi
  al (enlace = "logit"),
       datos = votación)

Desviación Derechos residuales de autor:
      Min        1T    Mediana        3T       Max
- 1,1020   - 0,9407   - 0.5519    1.2547    3.6552

Coeficientes:
                    Estimar Std. Error      valor z Pr (> | z |)
(Interceptar)     - 0,179971 0,014803       - 12,16 <2e-16 ***
```

distanciaincsq -0.101549 0,002075 - 48,94 <2e-16 ***
- - -
Signif. códigos: 0 * * * 0,001 * * 0,01 * 0,05.
0,1 1

(Parámetro de dispersión para familia binomial tomada como 1)

Desviación nula: 47335 en 38210 grados de libertad
Desviación residual: 43087 en 38209 grados de libertad
(6686 observaciones eliminadas debido a AIC: falta)
43091

Número de iteraciones de puntuación de Fisher:} 5

Con este segundo modelo podemos ver cómo el AIC puede ser útil: Con un valor mayor de 43.091 en el modelo cuadrático, concluimos que el modelo con distancia ideológica lineal encaja mejor con un AIC de 42.914. Esto corresponde a la conclusión del artículo original de que la forma lineal de la variable se ajusta mejor.

7.1.2 Modelos Probit

Los modelos Logit han ganado fuerza a lo largo de los años en aras de la simplicidad en el cálculo y la interpretación. (Por ejemplo, los modelos logit se pueden interpretar con *razones de probabilidades.)* Sin embargo, un supuesto clave de los modelos logit es que el término de error en el modelo de variable latente (o la utilidad latente) tiene una distribución logística. Podemos estar más contentos con suponer que el término de error de nuestro modelo se distribuye normalmente, dada la prevalencia de esta distribución en la naturaleza y en los resultados asintóticos.[5] *Regresión probit* nos permite ajustar un modelo con una variable de resultado binaria con un término de error distribuido normalmente en el modelo de variable latente.

Para mostrar cómo funciona este modelo alternativo de resultado binario, recurrimos a un modelo de probabilidad de que un encuestado haya votado. Singh2014a) modela esto en función de la proximidad ideológica al partido más cercano ponderada por la competitividad de la elección. La teoría aquí es que los individuos con una alternativa relativamente próxima en una elección competitiva tienen más probabilidades de considerar que vale la pena votar. Adaptamos este modelo de la siguiente manera:

```
turnout.linear <-glm (votado ~ ponderado a distancia,
        familia = binomio (enlace = "probit"), datos = votación)
resumen (participación.linear)
```

[5]Además, en entornos avanzados para los que necesitamos desarrollar una distribución multivariante para múltiples variables de resultado, es relativamente fácil trabajar con la distribución normal.

El resultado de nuestro comando de resumen es:

```
Llamada:
glm (fórmula = votado ~ ponderado a distancia, familia = binomi
  al (enlace    = "probit"),
        datos   = votación)

Desviación Derechos residuales de autor:
       Min              1T      Mediana              3T             Max
- 1,9732          0.5550          0.5550          0.5776          0,6644

Coeficientes:
                                  Estimar Std. Error              valor z Pr (> | z |)
(Interceptar)                     1.068134          0,009293      114,942 <2e-16
    * * *
ponderado a distancia -0,055074                 0.011724        - 4.698 2.63e-06
    * * *
- - -
Signif. códigos:              0    * * *    0,001      * *    0,01      *    0,05       .
0,1 1

(Parámetro de dispersión para             binomio       familia llevada a
    ser 1)

       Desviación nula: 37788             en 44896       grados de libertad
Desviación residual: 37766 AIC:           en 44895       grados de libertad
37770

Número de puntuación de Fisher       iteraciones: 4
```

El diseño de estos resultados del modelo probit es similar a los resultados del modelo logit. Sin embargo, tenga en cuenta que cambiar la distribución del término de error latente a una distribución normal cambia la escala de los coeficientes, por lo que los valores serán diferentes entre los modelos logit y probit. Las implicaciones sustantivas suelen ser similares entre los modelos, por lo que el usuario debe decidir qué modelo funciona mejor en términos de suposiciones e interpretación de los datos disponibles.

7.1.3 Interpretación de resultados logit y probit

Una característica importante de los GLM es que el uso de una función de enlace hace que los coeficientes sean más difíciles de interpretar. Con un modelo de regresión lineal, estimado en el Cap.6, podríamos simplemente interpretar el coeficiente en términos de cambio en el valor esperado del resultado mismo, manteniendo iguales las otras variables. Sin embargo, con un GLM, la media del resultado se ha transformado y el coeficiente habla del cambio en la media transformada. Por lo tanto, para análisis como modelos logit y probit,

necesitamos tomar pasos adicionales para interpretar el efecto que tiene una entrada en el resultado de interés.

Para un modelo de regresión logística, el analista puede calcular rápidamente la *razón de probabilidades* para cada coeficiente simplemente tomando el exponencial del coeficiente.[6] Recuerde que el *impares* de un evento es la razón entre la probabilidad de que ocurra el evento y la probabilidad no ocurre: $\frac{pag}{1\ pag}$. La razón de probabilidades nos dice el factor multiplicativo por el cual el las probabilidades cambiarán para un aumento unitario en el predictor. DentroR, si queremos la razón de posibilidades para nuestro coeficiente de distancia en la Tabla 7.1, simplemente escribimos:

```
exp (incluyendo coeficientes de $ lineales [-1])
```

Esta sintaxis tomará el exponencial de cada estimación de coeficiente de nuestro modelo, sin importar el número de covariables. La [-1] omite la intersección, por lo que una razón de probabilidades no tendría sentido. Teniendo solo un predictor, la impresión en este caso es:

```
distanciainc
  0.6097611
```

Debemos tener cuidado al interpretar el significado de las razones de probabilidades. En este caso, para un aumento de un punto en la distancia del partido en el poder en la escala de ideología, las probabilidades de que un encuestado vote por el partido en el poder disminuyen en un factor de 0,61. (Con múltiples predictores, necesitaríamos agregar el*ceteris paribus* advertencia.) Si, en lugar de interpretarlo como un factor multiplicativo, el analista prefirió discutir el cambio en términos porcentuales, escriba:

```
100 * (exp (con coeficientes de $ lineales [-1]) - 1)
```

En este caso, se devuelve un valor de 39.02389. Por lo tanto, podemos decir: para un aumento de un punto en la distancia del partido en el poder en la escala de ideología, las probabilidades de que un encuestado vote por el partido en el poder disminuyen en un 39%. Sin embargo, recuerde que todas estas declaraciones se relacionan específicamente con*impares,* por lo que en este caso nos referimos a una disminución del 39% en la relación entre la probabilidad de votar por el titular y la probabilidad de votar por cualquier otro partido.

Una interpretación alternativa que a menudo es más fácil de explicar en el texto es informar *probabilidades predichas* de un modelo. Para un modelo de regresión logística, ingresar las predicciones de la función lineal (las utilidades latentes) en la función de distribución acumulativa logística produce la probabilidad predicha de que el resultado tome un valor de 1. Un enfoque simple para ilustrar intuitivamente el efecto de un predictor es trazar las probabilidades predichas en cada valor que puede tomar un predictor, lo que muestra cómo cambia la probabilidad de forma no lineal a medida que cambia el predictor. Primero procedemos creando nuestras probabilidades predichas:

```
distancias <-seq (0,10, por = .1) entradas <-cbind (1, distancias)
nombres de columnas (entradas) <- c ("constante",
"distanciainc") entradas <-as.data.frame (entradas)
```

[6] Esto se debe a que la función de enlace logit es el logaritmo de las probabilidades del evento.

```
Forecast.linear <-predict (incluido lineal, nuevos datos = entradas,
        type = "respuesta")
```

En la primera línea, creamos un **seq**uencia de las posibles distancias del titular, que van desde el mínimo (0) al máximo (10) en pequeños incrementos (0,1). Luego creamos una matriz llamadaentradas que almacena valores de predictores de interés para todos los predictores en nuestro modelo (utilizando el **C**olumn **unir,** cbind, comando para combinar dos vectores como columnas en una matriz). Posteriormente, nombramos las columnas para que coincidan con los nombres de nuestras variables y recategorizamos esta matriz**como** a **marco de datos.** En la última línea, usamos el predecir comando, que guarda las probabilidades predichas en un vector. Observe el uso delnuevos datos opción para especificar nuestro marco de datos de valores predictores y la tipo opción para especificar que queremos nuestros valores predichos en el respuesta escala. Al establecer esto en la escala de respuesta, el comando devuelve probabilidades pronosticadas de votar por el partido en el poder en cada distancia hipotética.

Como alternativa al modelo en el que votar por el titular es una función de la distancia ideológica lineal entre el votante y el partido, también ajustamos un modelo utilizando la distancia al cuadrado. Podemos calcular fácilmente la probabilidad predicha de este modelo alternativo contra el valor de la distancia en su escala original. Nuevamente, las probabilidades predichas se calculan escribiendo:

```
entradas2 <-cbind (1, distancias ^ 2) nombres de columnas
(entradas2) <- c ("constante", "distanciaincsq") entradas2 <-
as.data.frame (entradas2)
Forecast.squared <-predict (incluido cuadrado, newdata = input2,
        type = "respuesta")
```

En este caso, usamos el vector original distancias que capturaron valores de predictores hipotéticos y los elevó al cuadrado. Al usar estos valores al cuadrado, guardamos nuestras probabilidades predichas del modelo alternativo en el vectorpronóstico al cuadrado.

Para graficar las probabilidades predichas de cada modelo en el mismo espacio, escribimos:

```
plot (y = previsión.lineal, x = distancias, ylim = c (0, .6), tipo = "l",
        lwd = 2, xlab = "", ylab = "")
líneas (y = pronóstico.cuadrado, x = distancias, lty = 2, col = "azul", lwd = 2) leyenda (x =
6, y = .5, leyenda = c ("lineal", "cuadrado"), lty = c (1,2),
        col = c ("negro", "azul"), lwd = 2)
mtext ("Distancia ideológica", lado = 1, línea = 2.75, cex = 1.2) mtext
("Probabilidad de votar por el titular", lado = 2,
        línea = 2.5, cex = 1.2)
```

En la primera línea, graficamos las probabilidades predichas del modelo con distancia lineal. En el eje vertical (y) son las probabilidades, y en el eje horizontal (X) son los valores de la distancia. Limitamos las probabilidades entre 0 y 0,6 para ver más de cerca los cambios, establezcatipo = "l" para producir un diagrama de línea, y use la opción lwd = 2 para aumentar el grosor de la línea. También configuramos las etiquetas de los ejes xey para que estén vacías (xlab = "", ylab = "") para que podamos completar las etiquetas más tarde con un comando más preciso. En la segunda línea, agregamos otra línea a la figura abierta del

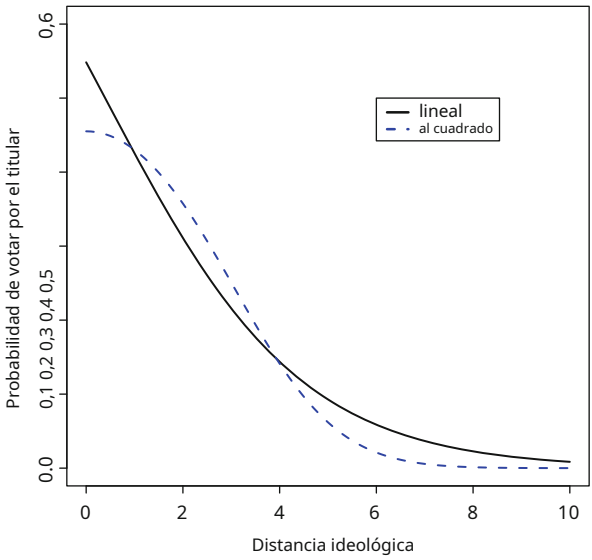

Figura 7.1 Predicho probabilidad de votar por el partido en el poder en función de la distancia ideológica de los gobernantes, basada en una línea lineal y una cuadrática forma funcional

probabilidades predichas a partir del modelo con distancia al cuadrado. Esta vez, coloreamos la línea de azul y la hacemos discontinua (lty = 2) para distinguirlo de las probabilidades predichas del otro modelo. En la tercera línea, agregamos una leyenda a la gráfica, ubicada en las coordenadas dondex = 6 y y = 0,5, que distingue las líneas en función de las distancias lineales y cuadradas. Finalmente, en las dos últimas líneas agregamos etiquetas de eje usando el
mtext comando: El lado La opción nos permite declarar en qué eje estamos escribiendo, la línea El comando determina qué tan lejos del eje se imprime la etiqueta, y el cex El comando nos permite ampliar el tamaño de la fuente (al 120% en este caso). Los resultados completos se presentan en la Fig.7.1. Como muestra la figura, el modelo con distancia al cuadrado responde mejor en valores medios, con una respuesta más plana en los extremos. Por lo tanto, Singh's (2014a) La conclusión de que la distancia lineal se ajusta mejor tiene implicaciones sustanciales para el comportamiento del votante.

Como ejemplo final de cómo informar las probabilidades predichas, recurrimos a un ejemplo del modelo probit que estimamos de participación. Las probabilidades pronosticadas se calculan de manera similar para los modelos probit, excepto que las predicciones lineales (o utilidades) ahora se ingresan en una función de distribución acumulativa normal. En este ejemplo, agregaremos a nuestra presentación de probabilidades predichas al incluir intervalos de confianza alrededor de nuestras predicciones, que transmiten al lector el nivel de incertidumbre en nuestro pronóstico. Comenzamos como lo hicimos en el último ejemplo, creando un marco de datos de valores de datos hipotéticos y produciendo probabilidades predichas con ellos:

```
wght.dist <-seq (0,4, by = .1) inputs.3 <-cbind (1, wght.dist) colnames
(input.3) <- c ("constante", "ponderada por distancia") entradas.3 <-
as.data.frame (entradas.3)

Forecast.probit <-predict (turnout.linear, newdata = inputs.3,
        type = "link", se.fit = TRUE)
```

En este caso, la distancia ideológica ponderada del partido ideológico más cercano es nuestro único predictor. Este predictor varía de aproximadamente 0 a 4, por lo que creamos un vector que abarca esos valores. En la última línea del código anterior, hemos cambiado dos características: Primero, hemos especificadotype = "enlace". Esto significa que nuestras predicciones son ahora predicciones lineales de la utilidad latente, y no las probabilidades en las que estamos interesados. (Esto se corregirá en un momento). En segundo lugar, hemos agregado la opciónse.fit = TRUE, lo que nos proporciona un error estándar de cada predicción lineal. Nuestro objeto de salida,Forecast.probit ahora contiene tanto los pronósticos lineales como los errores estándar.

La razón por la que guardamos las utilidades lineales en lugar de las probabilidades es que al hacerlo nos será más fácil calcular los intervalos de confianza que permanecen dentro de los límites de probabilidad de 0 y 1. Para hacer esto, primero calculamos los intervalos de confianza de las predicciones lineales. . Para el nivel de confianza del 95%, escribimos:

```
lower.ci <-forecast.probit $ fit-1.95996399 * Forecast.probit $ se.fit upper.ci <-
forecast.probit $ fit + 1.95996399 * Forecast.probit $ se.fit
```

Note que llamando Forecast.probit $ fit obtenemos las predicciones lineales de utilidades, y llamando Forecast.probit $ se.fit llamamos a los errores estándar de nuestro pronóstico. 1.95996399 es el valor crítico del 95% de dos colas de una distribución normal. Ahora que tenemos los vectores de los límites inferior y superior del intervalo de confianza, podemos insertarlos en la función de distribución acumulativa normal para poner los límites en la escala de probabilidad predicha.

Ahora podemos graficar las probabilidades predichas con intervalos de confianza de la siguiente manera:

```
plot (y = pnorm (pronóstico.probit $ ajuste), x = wght.dist, ylim = c (.7, .9),
      type = "l", lwd = 2, xlab = "Distancia ideológica ponderada", ylab =
      "Probabilidad de participación")
líneas (y = pnorm (lower.ci), x = wght.dist, lty = 3, col = "red", lwd = 2) líneas (y =
pnorm (upper.ci), x = wght.dist, lty = 3, col = "rojo", lwd = 2)
```

En la primera línea, graficamos las propias probabilidades predichas. Para obtener las probabilidades para el eje vertical, escribimosy = pnorm (pronóstico.probit $ ajuste). La pnorm La función es la función de distribución acumulativa normal, por lo que convierte nuestras predicciones de utilidad lineal en probabilidades reales. Entretanto,x = wght.dist coloca los posibles valores de la distancia ponderada a la parte más cercana en el eje horizontal. En la segunda línea, graficamos el límite inferior del intervalo de confianza del 95% de las probabilidades predichas. Aquí,pnorm (lower.ci) convierte el pronóstico del intervalo de confianza en la escala de probabilidad. Finalmente, repetimos el proceso en la línea tres para trazar el límite superior del intervalo de confianza. La salida completa se puede ver en la Fig. 7.2. Una característica notable de este gráfico es que el intervalo de confianza se vuelve notablemente amplio para los valores más grandes de la distancia ponderada. Esto se debe a que la media de la variable es baja y hay pocas observaciones en estos valores más altos.

Las probabilidades predichas en ambos casos fueron simples porque incluían solo un predictor. Para cualquier GLM, incluido un modelo logit o probit, las probabilidades predichas y su nivel de respuesta dependen del valor de todas las covariables. Siempre que un investigador tenga múltiples predictores en un modelo GLM,

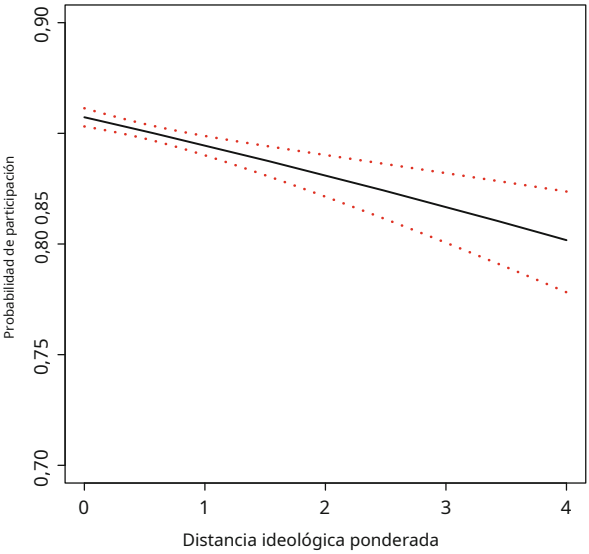

Figura 7.2 Probabilidad prevista de salir a votar en función de la distancia ideológica ponderada del partido más cercano, con intervalos de confianza del 95%

Los valores razonables de las variables de control deben incluirse en los pronósticos. Ver secc. 7.3.3 para un ejemplo del uso de la predecir función para un GLM que incluye múltiples predictores.

7.2 Resultados ordinales

Pasamos ahora a las medidas de resultado ordinales. Las variables ordinales tienen múltiples categorías como respuestas que se pueden clasificar de menor a mayor, pero aparte de las clasificaciones, los valores numéricos no tienen un significado inherente. Como ejemplo de una variable de resultado ordinal, nuevamente usamos datos de encuestas del CSES, esta vez de Singh (2014b) estudio de satisfacción con la democracia. En relación con el ejemplo anterior, estos datos tienen un alcance más amplio, incluidos 66.908 encuestados de 62 elecciones. Las variables de este conjunto de datos son las siguientes:

satisfacción: Nivel de satisfacción de la encuestada con la democracia. Escala ordinal codificado 1 (nada satisfecho), 2 (no muy satisfecho), 3 (bastante satisfecho) o 4 (muy satisfecho).
cntryyear: Una variable de carácter que enumera el país y el año de la elección.
cntryyearnum: Un índice numérico que identifica el país y el año de la elección.
libertad: Freedom House puntúa el nivel de libertad de un país. Rango de puntuaciones de 5.5 (menos gratis) a 1 (más gratis).
El crecimiento del PIB: Crecimiento porcentual del Producto Interno Bruto (PIB).

gdppercapPPP: PIB per cápita, calculado utilizando la paridad de precios de compra (PPA), encadenado a 2000 dólares internacionales, en miles de dólares.

IPC: Indice de Percepción de la corrupción. Las puntuaciones van de 0 (menos corrupto) a 7,6 (más corrupto).

eficacia: La Demandada cree que votar puede marcar la diferencia. Escala ordinal de 1 (en desacuerdo) a 5 (de acuerdo).

educ: Indicador codificado 1 si el encuestado se graduó de la universidad, 0 si no.

se abstuvo: Indicador codificado 1 si el encuestado se abstuvo de votar, 0 si el el encuestado votó.

prez: Indicador codificado 1 si el país tiene un sistema presidencial, 0 en caso contrario.

majoritarian_prez: Indicador codificado 1 si el país tiene un sistema mayoritario, 0 si no.

ganador: Indicador codificado 1 si el encuestado votó por el partido ganador, 0 si no.

vote_ID: Indicador codificado 1 si el encuestado votó por el partido que se identifica con, 0 si no.

vote_affect: Indicador codificado 1 si el encuestado votó por el partido que clasificado más alto, 0 si no.

vote_ideo: Indicador codificado 1 si el encuestado votó por el partido más similar en ideología, 0 si no.

Optimismo: Escala de optimización de votos que va de 0 a 3, codificada agregando **vot_ID, votó_afecto,** y **vote_ideo. ganadorXvoted_ID:**Término de interacción entre votar por el ganador y votar por identi fi cación del partido.

ganadorXvoted_affect: Término de interacción entre votar por el ganador y votar para la fiesta mejor valorada.

ganadorXvoted_ideo: Término de interacción entre votar por el ganador y votar por similitud ideológica.

ganadorXoptimality: Término de interacción entre la votación por el ganador y la votación escala de optimalidad.

Estos datos también están en formato Stata, por lo que si el extranjero la biblioteca aún no está cargada, será necesario llamarla. Para cargar nuestros datos, descargue el archivo SinghEJPR.dta del Dataverse vinculado en la página vii o el contenido del capítulo vinculado en la página 97. Luego escriba:

```
biblioteca (extranjera)
satisfacción <-read.dta ("SinghEJPR.dta")
```

En este ejemplo, deseamos modelar el nivel de satisfacción de cada encuestado con la democracia. Esta variable toma los valores 1, 2, 3 y 4, y solo podemos hacer declaraciones sobre qué valores reflejan niveles más altos de satisfacción. En otras palabras, podemos decir que un valor de 4 ("muy satisfecho") refleja una mayor satisfacción que un valor de 3 ("bastante satisfecho"), un valor de 3 es más de 2 ("no muy satisfecho"), y por lo tanto, un valor de 4 es mayor que un valor de 2. Sin embargo, no podemos decir *cuánto* más satisfacción refleja un valor en relación con otro, ya que los números no tienen un significado inherente más que proporcionar un orden de satisfacción. Hacerlo sería

requieren que cuantifiquemos adjetivos como "muy" y "bastante". Por lo tanto, un modelo logit ordenado o probit ordenado será apropiado para este análisis.

Como nuestro primer ejemplo, Singh (2014b, Tabla SM2) se ajusta a un modelo en el que la satisfacción en democracia se basa en si el encuestado votó por el candidato ideológicamente más próximo, si el candidato votó por el ganador y la interacción entre estas dos variables.[7] El más importante de estos términos es la interacción, ya que prueba la hipótesis de que los individuos que estaban del lado ganador *y* votado por el partido más similar ideológicamente expresará la mayor satisfacción.

Volviendo a los específicos, para los modelos de regresión ordinal debemos usar el comando especial polrcorto para **pag**proporcional **o**dds **l**ogístico **r**egresión), que es parte de la MASA paquete. La mayoríaR las distribuciones se instalan automáticamente MASA, aunque todavía tenemos que cargarlo con el Biblioteca mando.[8] Para cargar el MASA paquete y luego estimar un *logit ordenado* modelo con estos datos, escribiríamos:

```
biblioteca (MASA)
satisfacción $ satisfacción <-ordenados (as.factor (
        satisfacción $ satisfacción))
ideol.satisfaction <-polr (satisfacción ~ vote_ideo * ganador +
        abstencion + educ + eficacia + majoritarian_prez +
        freedom + gdppercapPPP + gdpgrowth + CPI + prez,
    method = "logistic", data = satisfacción) resumen
    (ideol.satisfaction)
```

Observe que recodificamos nuestra variable dependiente, satisfacción, utilizando la como.factor comando que se introdujo en la Secta. 2.4.2. Además, incorporamos esto en el ordenado comando para transmitir que el factor se puede ordenar numéricamente. Tuvimos que recodificar de esta manera porque el comando modelopolr requiere que el resultado se guarde como un vector de clase factor. Mientras tanto, el lado derecho del modelo se parece a todos los demás modelos que hemos estimado hasta ahora separando los nombres de las variables con signos más. Tenga en cuenta que usamos notación interactiva con votado_ideo * ganador, para incluir ambos términos más el producto.[9] En conjunto, hemos modelado la satisfacción en función de si el votante votó por el partido ideológicamente más similar, votó por el ganador, un término de interacción entre los dos y varias otras variables de control a nivel individual y nacional. Dentro del comando, la opciónmétodo = "logístico" especifica que deseamos estimar un ordenado *logit* modelo en lugar de utilizar otra función de enlace. Al final de la línea, especificamos nuestrodatos opción como de costumbre para señalar nuestro conjunto de datos de interés.

---

[7] Este y el siguiente ejemplo no replican exactamente los resultados originales, que también incluyen efectos aleatorios por país y año. Además, el siguiente ejemplo ilustra la regresión probit ordenada, en lugar del modelo logístico ordenado del artículo original. Ambos ejemplos se basan en modelos que se encuentran en el material de apoyo en línea en el *Revista europea de investigación política* sitio web.

[8] Si un usuario necesita instalar el paquete, install.packages ("MASA") hará el trabajo.

[9] Una especificación equivalente habría sido incluir vote_ideo + ganador + ganadorXvoted_ideo como tres términos separados de los datos.

Después de estimar este modelo, escribimos resumen (ideol. satisfacción) en la consola. La salida se ve así:

Llamada:
polr (fórmula = satisfacción ~ video_ votada * ganador + abstinencia +
educ + eficacia + majoritarian_prez + freedom + gdppercap
PPP +
gdpgrowth + CPI + prez, data = satisfacción, método =
"logístico")

Coeficientes:

	Valor	Std. Error	valor t
vote_ideo	- 0.02170	0.023596	- 0,9198
ganador	0.21813	0.020638	10.5694
se abstuvo	- 0.25425	0.020868	-12.1838
educ	0.08238	0.020180	4.0824
eficacia	0.16246	0,006211	26,1569
majoritarian_prez	0.05705	0.018049	3.1609
libertad	0.04770	0.014087	3.3863
gdppercapPPP	0.01975	0,001385	14,2578
El crecimiento del PIB	0.06653	0,003188	20,8673
IPC	- 0.23153	0,005810	-39,8537
prez	- 0.11503	0.026185	-4.3930
vote_ideo: ganador	0.19004	0.037294	5.0957

Intercepciones:

	Valor	Std.	Valor t de error
1 | 2	- 2.0501	0.0584	- 35.1284
2 | 3	- 0.0588	0.0575	- 1.0228
3 | 4	2.7315	0.0586	46.6423

Desviación residual: 146397.33 AIC:
146427.33

La salida muestra la estimación de cada coeficiente, el error estándar y la *z* valor. (AunqueR lo llama un *t* valor, los métodos de máxima verosimilitud suelen requerir *z* proporciones, como se mencionó anteriormente.) Después de la presentación de coeficientes, tres *puntos de corte* se presentan bajo la etiqueta de Intercepta. Estos puntos de corte identifican el modelo al encontrar en qué lugar de una escala de utilidad latente se encuentra el límite entre que el encuestado elija 1 frente a 2 como respuesta, 2 frente a 3 y 3 frente a 4. son importantes por el bien de pronosticar los resultados previstos. Las estadísticas de ajuste predeterminadas en la salida son la desviación residual y el AIC.

Los resultados se presentan de manera más formal en la Tabla 7.2. AunqueRLa salida base omite el *pag* valor, el usuario puede extraer fácilmente inferencias basadas en la información disponible: ya sea a través del cálculo de intervalos de confianza, comparando el *z* valor a un valor crítico, o computación *pag* se valora a uno mismo. Por ejemplo, la hipótesis clave

Cuadro 7.2 Elecciones nacionales logit ordenadas modelo de satisfacción con la democracia, 62

Vaticinador	Estimar	Std. error	*z* valor	Pr (> j*z*j)
Votado por el partido próximo	0: 0217	0.0236	0: 9198	0.3577
Votado por el ganador	0: 2181	0.0206	10: 5694	0,0000
Votado como ganador próximo	0: 1900	0.0373	5: 0957	0,0000
Se abstuvo	0: 2542	0.0209	12: 1838	0,0000
Graduado de la Universidad	0: 0824	0.0202	4: 0824	0,0000
Eficacia	0: 1625	0,0062	26: 1569	0,0000
Sistema mayoritario	0: 0571	0.0180	3: 1609	0,0016
Libertad	0: 0477	0.0141	3: 3863	0,0007
Desarrollo economico	0: 0197	0,0014	14: 2578	0,0000
Crecimiento económico	0: 0665	0,0032	20: 8673	0,0000
Corrupción	0: 2315	0,0058	39: 8537	0,0000
Sistema presidencial	0: 1150	0.0262	4: 3930	0,0000
1	2: 0501	0.0584	35: 1284	0,0000
2	0: 0588	0.0575	1: 0228	0.3064
3	2: 7315	0.0586	46: 6423	0,0000

Notas: N D 66; 908. AIC D 146; 427. Datos de Singh (2014b)

aquí es para que el término de interacción sea positivo. Por lo tanto, podemos obtener nuestro*pag* valor escribiendo:

```
1 pnorm (5.0957)
```

R imprime 1.737275e-07, que en notación científica significa *pag* D 0: 00000017. Por lo tanto, concluiremos con un 99,9% de confianza en que el coeficiente del término de interacción es perceptiblemente mayor que cero. En mesa7.2, hemos optado por informar de las dos colas *pag* valores.[10]

Una buena característica del uso de la función de enlace logit es que los resultados se pueden interpretar en términos de razones de probabilidades. Sin embargo, las razones de probabilidad deben calcularse e interpretarse de manera un poco diferente para un modelo ordinal. Ver largo (1997, págs. 138-140) para obtener una explicación completa. Para modelos logit ordenados, debemos exponenciar el*negativo* valor de un coeficiente e interpretar las probabilidades de estar en grupos más bajos en relación con los grupos más altos. A modo de ejemplo, las razones de probabilidades de nuestros coeficientes de la Tabla7.2, junto con los cambios porcentuales en las probabilidades, se pueden producir de una vez:

[10] Desafortunadamente, el xtable comando no produce L listoATablas TEX para resultados de polr. Sin embargo, al crear una matriz con los resultados relevantes, LALos usuarios de TEX pueden producir una tabla más rápido que la codificación manual, aunque son necesarias algunas revisiones del producto final. Intente lo siguiente:

```
coef <-c (ideol.satisfacción $ coeficientes, ideol.satisfacción $ zeta) se <-sqrt (diag (vcov
(ideol.satisfacción)))
z <-coef / se
p <-2 * (1-pnorm (abs (z)))
xtable (cbind (coef, se, z, p), digits = 4)
```

```
exp (-ideol.satisfaction $ coefficients) 100 * (exp (-
ideol.satisfaction $ coefficients) -1)
```

La impresión resultante de la segunda línea es:

```
      vote_ideo                  ganador              se abstuvo
       2.194186          - 19.597657                28.949139
           educ                 eficacia majoritarian_prez
     - 7.908003          - 14,994659 - 5,545347
        libertad        gdppercapPPP gdpgrowth
     - 4.658313          - 1.955471 - 6.436961
            IPC                      prez vot_ideo: ganador
      26.053177          12.190773 - 17.307376
```

Si quisiéramos interpretar el efecto de la eficacia, entonces, podríamos decir que para un aumento de un punto en una escala de eficacia de cinco puntos, las probabilidades de que un encuestado informe que no está "en absoluto satisfecho" con la democracia en relación con cualquiera de las tres categorías superiores disminuyen en un 15%, *ceteris paribus.* Además, las probabilidades de que un encuestado informe "nada satisfecho" o "no muy satisfecho" en relación con las dos categorías superiores también disminuyen en un 15%, todo lo demás igual. Además, las probabilidades de que un encuestado informe uno de los tres niveles inferiores de satisfacción en relación con la categoría más alta de "muy satisfecho" disminuyen en un 15%, manteniendo constantes los demás predictores. En general, entonces, podemos interpretar una razón de probabilidades para un logit ordenado como la configuración de las probabilidades de todas las opciones por debajo de un umbral en relación con todas las opciones por encima de un umbral.

Como segundo ejemplo, pasamos ahora a un modelo de satisfacción del votante que se centra no en el papel de la proximidad ideológica en la elección del voto, sino en qué partido evaluó mejor el votante cuando se le pidió que calificara a los partidos. Una vez más, la interacción entre votar por el partido con la calificación más alta y también votar por el partido ganador es el coeficiente de interés principal. Esta vez, también probaremos una función de enlace diferente y estimaremos un *probit ordenado* modelo en lugar de un modelo logit ordenado. En este caso teclearemos:

```
fect.satisfaction <-polr (satisfacción ~ vote_affect * ganador +
        abstencion + educ + eficacia + majoritarian_prez +
        freedom + gdppercapPPP + gdpgrowth + CPI + prez,
        method = "probit", data = satisfacción)
```

Además de cambiar una de las variables interactuadas, la única diferencia entre este código y el comando anterior para el comando logit ordenado es la especificación de método = "probit". Esto cambia un poco la escala de nuestros coeficientes, pero las implicaciones sustantivas de los resultados son generalmente similares independientemente de esta elección. Escribiendoresumen (afecta la satisfacción), obtenemos la salida:

```
Llamada:
polr (fórmula = satisfacción ~ efecto_ votada * ganador + abstinencia +

        educ + eficacia + majoritarian_prez + freedom + gdppercapPPP +
```

```
    gdpgrowth + CPI          + prez, datos = satisfacción, método
      = "probit")
```

Coeficientes:

	Valor	Std. Valor t de error	
vote_affect	0.03543	0.0158421	2.237
ganador	0.04531	0.0245471	1.846
se abstuvo	- 0.11307	0.0170080	- 6.648
educ	0.05168	0.0115189	4.487
eficacia	0.09014	0,0035177	25.625
majoritarian_prez	0.03359	0.0101787	3.300
libertad	0.03648	0,0082013	4.448
gdppercapPPP	0.01071	0.0007906	13.546
El crecimiento del PIB	0.04007	0,0018376	21.803
IPC	- 0,12897	0,0033005	-39,075
prez	- 0.03751	0.0147650	-2.540
vote_affect: ganador	0.14278	0,0267728	5,333

Intercepciones:

	Valor	Std. Valor t de error	
1 | 2	- 1,1559	0.0342	- 33,7515
2 | 3	- 0.0326	0.0340	- 0,9586
3 | 4	1.6041	0.0344	46.6565

Desviación residual: 146698.25 AIC: 146728.25

Una vez más, nuestra hipótesis de interés está respaldada por los resultados, con un efecto positivo y perceptible en la interacción entre votar por el partido ganador y votar por el partido mejor calificado.

7.3 Recuentos de eventos

Como tercer tipo de GLM, recurrimos a modelos de recuento de eventos. Siempre que nuestra variable dependiente sea el número de eventos que ocurren dentro de un período de tiempo definido, la variable tendrá la característica de que nunca puede ser negativa y debe tomar un valor discreto (por ejemplo, 0,1,2,3,4, ...). Por lo tanto, los resultados del recuento tienden a tener un fuerte sesgo hacia la derecha y una distribución de probabilidad discreta como la distribución binomial negativa o de Poisson.

Como ejemplo de datos de recuento, ahora volvemos a Peake y Eshbaugh-Soha (2008) datos que se discutieron previamente en el Cap. 3. Recuerde que la variable de resultado en este caso es el número de noticias de televisión relacionadas con la política energética en un mes determinado. (Ver Cap.3 para obtener detalles adicionales sobre los datos.) El número de noticias en un mes sin duda es un recuento de eventos. Sin embargo, tenga en cuenta que debido a

estos son datos mensuales, son *dependiente del tiempo,* que es una característica que ignoramos en este momento. En el Cap.9 revisamos este tema y consideramos modelos que dan cuenta del tiempo. Por ahora, sin embargo, esto ilustra cómo usualmente encajamos los modelos de conteo enR.

Primero, cargamos los datos nuevamente:[11]

```
pres.energy <-read.csv ("PESenergy.csv")
```

Después de ver las estadísticas descriptivas de nuestras variables y visualizar los datos como hicimos en el Cap. 3, ahora podemos pasar a ajustar un modelo.

7.3.1 Regresión de Poisson

El modelo de conteo más simple que podemos ajustar es un modelo de Poisson. Si tuviéramos que escribir:

```
energy.poisson <-glm (Energía ~ rmn1173 + grf0175 + grf575 + jec477 +
        jec1177 + jec479 + embargo + rehenes + petróleoc + Aprobación + Desempleo,
        familia = poisson (link = log), data = pres.energy)
```

Esto encajará en un modelo de regresión de Poisson en el que la cobertura televisiva de la política energética es una función de seis mandatos para los discursos presidenciales, un indicador del embargo petrolero árabe, un indicador de la crisis de rehenes de Irán, el precio del petróleo, la aprobación presidencial y el tasa de desempleo.[12] Tenga en cuenta que esta vez, establecemos familia = poisson (enlace = registro), declarando la distribución del resultado como Poisson y especificando nuestra función de enlace de registro. Si escribimos resumen (energía.poisson) en la consola, R devuelve la siguiente salida:

```
Llamada:
glm (fórmula = Energía ~ rmn1173 + grf0175 + grf575 + jec477 +

    jec1177 + jec479 + embargo + rehenes + oilc +
      Aprobación +
    Desempleo, familia = poisson (enlace = log), datos = pres.
      energía)

Residuos de desviación:
    Min       1T   Mediana       3T      Max
- 8.383 -2.994    - 1.054  1.536 11.399

Coeficientes:
                  Estimar Std.  Error z valor Pr (> | z |)
(Interceptar)      13.250093    0.329121 40.259    <2e-16    * * *
rmn1173            0,694714     0.077009 9.021     <2e-16    * * *
```

[11] La nota al pie debe decir: "Para los usuarios que no tienen el archivo a mano del Capítulo 3, descarguen el archivo del Dataverse vinculado en la página vii o el contenido del capítulo vinculado en la página 97.

[12] Tenga en cuenta que los términos del discurso presidencial se codifican con 1 solo en el mes del discurso y 0 en todos los demás meses. Los términos para el embargo de petróleo y la crisis de rehenes se codificaron con 1 mientras estos eventos estaban en curso y 0 en caso contrario.

grf0175	0,468294	0.096169	4.870	1.12e-06	* * *
grf575	- 0.130568	0.162191	- 0,805	0,420806	
jec477	1.108520	0.122211	9.071	<2e-16	* * *
jec1177	0.576779	0.155511	3.709	0,000208	* * *
jec479	1.076455	0.095066	11.323	<2e-16	* * *
embargo	0,937796	0.051110	18.349	<2e-16	* * *
rehenes	- 0.094507	0.046166	- 2.047	0.040647	*
oilc	- 0.213498	0,008052	- 26.515	<2e-16	* * *
Aprobación	- 0.034096	0,001386	- 24.599	<2e-16	* * *
Desempleo	- 0.090204	0,009678	- 9.321	<2e-16	* * *

- - -
Signif. códigos: 0 *** 0,001 * * 0,01 * 0,05. 0,1 1

(Parámetro de dispersión para poisson familia llevada a ser 1)

Desviación nula: 6009.0 en 179 grados de libertad
Desviación residual: 2598.8 AIC: 3488.3 en 168 grados de libertad

Número de iteraciones de puntuación de Fisher: 5

El formato de la salida: estimaciones de coeficientes, errores estándar, *z,* y *pag-*
Debería estar muy familiarizado a estas alturas, al igual que la desviación y las puntuaciones AIC. En este caso, la función de enlace es simplemente un logaritmo, por lo que, aunque los coeficientes en sí mismos no son muy significativos, la interpretación sigue siendo simple. Como una opción, si tomamos la exponencial de un coeficiente, esto nos ofrece una*relación de recuento,* lo que nos permite deducir el cambio porcentual en el recuento esperado para un cambio en la variable de entrada. Por ejemplo, si quisiéramos interpretar el efecto de la aprobación presidencial, podríamos escribir:

```
exp (-. 034096)
```

Aquí, simplemente insertamos el coeficiente estimado de la salida impresa. El resultado nos da una relación de recuento de 0,9664787. Podríamos interpretar esto como un significado para un aumento de un punto porcentual en el índice de aprobación del presidente, la cobertura de la política energética disminuye en un 3.4% en promedio y manteniendo todos los demás predictores iguales. Como una forma rápida de obtener la relación de recuento y el cambio porcentual para cada coeficiente, podríamos escribir:

```
exp (coeficientes de energía.poisson $ [-1]) 100 * (exp
(coeficientes de $ energía.poisson [-1]) - 1)
```

En ambas líneas el [-1] índice para el vector de coeficiente desecha la intersección término, por el cual no queremos un relación de recuento. La impresión de la segunda linea lee:

```
      rmn1173         grf0175          grf575          jec477         jec1177
100.313518       59.726654 -12.240295 202.987027 embargo       78.029428
       jec479         rehenes            oilc                      Aprobación
193.425875      155.434606 -9.017887 -19.224639                   - 3.352127
    Desempleo
 - 8.625516
```

De esta lista, podemos simplemente leer los cambios porcentuales para un aumento de una unidad en la entrada, manteniendo iguales las otras entradas. Para obtener un medio gráfico de interpretación, consulte la secc.7.3.3.

7.3.2 Regresión binomial negativa

Una característica intrigante de la distribución de Poisson es que la varianza es la misma que la media. Por lo tanto, cuando modelamos el logaritmo de la media, nuestro modelo modela simultáneamente la varianza. Sin embargo, a menudo encontramos que la varianza de nuestra variable de conteo es más amplia de lo que esperaríamos dadas las covariables, un fenómeno llamado*sobredispersión.* La regresión binomial negativa ofrece una solución a este problema al estimar un parámetro de dispersión adicional que permite que la varianza condicional difiera de la media condicional.

En R, **norte**egativo **B**Los modelos de regresión inomial en realidad requieren un comando especial del MASA biblioteca llamada glm.nb. Si el MASA la biblioteca no está cargada, asegúrese de escribir biblioteca (MASA) primero. Entonces, podemos ajustar el modelo binomial negativo escribiendo:

```
energy.nb <-glm.nb (Energía ~ rmn1173 + grf0175 + grf575 + jec477 +
        jec1177 + jec479 + embargo + rehenes + petróleoc + Aprobación + Desempleo,
        datos = energía pres.)
```

Observe que la sintaxis es similar a la glm comando, pero no hay familia opción ya que el comando en sí lo especifica. Escribiendoresumen (energy.nb) se imprimen los siguientes resultados:

```
Llamada:
glm.nb (fórmula = Energía ~ rmn1173 + grf0175 + grf575 +
    jec477 +
        jec1177 + jec479 + embargo + rehenes + oilc +
            Aprobación +
        Desempleo, datos = pres.energy, init.theta =
            2.149960724, enlace = registro)

Desviación Derechos residuales de autor:
      Min        1T    Mediana        3T       Max
- 2.7702  - 0,9635  - 0.2624    0.3569    2.2034

Coeficientes:
                    Estimar Std. Error  valor z  Pr (> | z |)
(Intercepción) 15.299318   1.291013   11.851    <2e-16 ***
rmn1173         0,722292   0,752005    0,960    0.33681
grf0175         0.288242   0,700429    0.412    0,68069
grf575        - 0.227584   0,707969  - 0.321    0,74786
jec477          0,965964   0,703611    1.373    0.16979
```

```
jec1177          0.573210   0,702534    0,816   0,41455
jec479           1.141528   0,694927    1.643   0.10045
embargo          1.140854   0.350077    3.259   0,00112 **
rehenes          0.089438   0.197520    0.453   0,65069
oilc           - 0.276592   0.030104  - 9.188   <2e-16 ***
Aprobación     - 0.032082   0,005796  - 5.536   3.1e-08 ***
Desempleo      - 0.077013   0.037630  - 2.047   0.04070 *
- - -
Signif. códigos: 0 ***          0,001 **        0,01 * 0,05. 0,1 1

(Parámetro de dispersión para      Negativo    Familia binomial (2.15)
    tomado
ser 1)

      Desviación nula: 393.02      en 179    grados de libertad
Desviación residual: 194,74 AIC:   en 168    grados de libertad
1526,4

Número de iteraciones de puntuación de Fisher: 1


                 Theta:  2.150
             Std. Errar.:  0,242

 2 x probabilidad logarítmica:  - 1500.427
```

Los coeficientes informados en esta salida se puede interpretar de la misma manera que Los coeficientes de un modelo de Poisson se interpretan porque ambos modelan el logaritmo de la media. La adición clave, informada al final de la impresión, es la dispersión parámetro. En este caso, nuestra estimación essobredosis 2:15, y con un error estándar de 0,242, el resultado es discernible. Esto indica que la sobredispersión está presente en este modelo. De hecho, muchas de las inferencias extraídas varían entre los modelos de Poisson y binomial negativo. Los dos modelos se presentan uno al lado del otro en la tabla.7.3. Como muestran los resultados, muchos de los resultados discernibles del modelo de Poisson no son discernibles en el modelo binomial negativo. Además, el AIC es sustancialmente más bajo para el modelo binomial negativo, lo que indica un mejor ajuste incluso cuando se penaliza por el parámetro de sobredispersión adicional.

7.3.3 Trazado de recuentos previstos

Si bien las razones de conteo son sin duda una forma sencilla de interpretar los coeficientes de los modelos de conteo, también tenemos la opción de graficar nuestros resultados. En este caso, modelamos el logaritmo de nuestro parámetro medio, por lo que debemos exponenciar nuestra predicción lineal para predecir el recuento esperado dadas nuestras covariables. Al igual que con los modelos logit y probit, para los resultados de recuento,predecir El comando facilita la previsión.

Supongamos que quisiéramos graficar el efecto de la aprobación presidencial en el número de noticias de televisión sobre la energía, según los dos modelos de Table 7.3. Esta situación contrasta un poco con los gráficos que creamos en la Secta.7.1.3. En todos los ejemplos logit y probit, solo teníamos un predictor. Por el contrario, en este caso tenemos varios otros predictores, por lo que tenemos que establecerlos en valores alternativos plausibles. Para este ejemplo, estableceremos el valor de todos los predictores de variables ficticias en su valor modal de cero, mientras que el precio del petróleo y el desempleo se establecen en su media. Si no insertamos valores razonables para las covariables, los recuentos predichos no se parecerán a la media real y el tamaño del efecto no será razonable.[13] En este ejemplo, la forma en que usamos el predecir El comando para pronosticar recuentos promedio con múltiples predictores se puede usar exactamente de la misma manera para un modelo logit o probit para pronosticar probabilidades predichas con múltiples predictores.

Cuadro 7.3 Dos modelos de conteo de nuevas historias de televisión mensuales sobre política energética, 1969-1983

	Poisson			Binomio negativo		
Parámetro	Estimar	Std. error	Pr (> jzj)	Estimar	Std. error	Pr (> jzj)
Interceptar	13.2501	0.3291	0,0000	15.2993	1.2910	0,0000
Nixon 11/73	0,6947	0.0770	0,0000	0,7223	0,7520	0.3368
Ford 1/75	0,4683	0.0962	0,0000	0.2882	0,7004	0,6807
Ford 5/75	0.1306	0.1622	0.4208	0.2276	0,7080	0,7479
Carter 4/77	1.1085	0.1222	0,0000	0.9660	0,7036	0.1698
Carter 11/77	0.5768	0.1555	0,0002	0.5732	0,7025	0.4145
Carter 4/79	1.0765	0.0951	0,0000	1,1415	0,6949	0.1005
Embargo de petróleo árabe	0,9378	0.0511	0,0000	1.1409	0.3501	0,0011
Crisis de rehenes en Irán	0.0945	0.0462	0.0406	0.0894	0,1975	0,6507
Precio del aceite	0.2135	0,0081	0,0000	0.2766	0.0301	0,0000
Aprobación presidencial	0.0341	0,0014	0,0000	0.0321	0,0058	0,0000
Desempleo	0.0902	0,0097	0,0000	0.0770	0.0376	0.0407
	-			2.1500	0.2419	0,0000
AIC	3488.2830			1526.4272		

Notas: N D 180. Datos de Peake y Eshbaugh-Soha (2008)

[13] Además de este enfoque de hacer predicciones utilizando valores centrales de variables de control, Hanmer y Kalkan (2013) argumentan que es preferible pronosticar los resultados basados en los valores observados de las variables de control en el conjunto de datos. Se anima a los lectores a consultar su artículo para obtener más consejos sobre este tema.

Volviendo a los específicos, en nuestros datos la variable aprobar oscila entre el 24% de aprobación y el 72,3%. Por lo tanto, construimos un vector que incluye el rango completo de aprobación, así como los valores plausibles de todos los demás predictores:

```
aprobación <-seq (24,72.3, por = .1) entradas 4 <-cbind (1,0,0,0,0,0,0,0,0, mean
(pres.energy $ oilc),
        aprobacion, media (pres.energy $ Unemploy)) colnames (input.4) <- c
("constante", "rmn1173", "grf0175",
        "grf575", "jec477", "jec1177", "jec479", "embargo", "rehenes", "oilc", "Aprobación",
        "Desempleo")
input.4 <-como.data.frame (input.4)
```

La primera línea de arriba crea el vector de valores hipotéticos de nuestro predictor de interés. La segunda línea crea una matriz de valores de datos hipotéticos: establece las variables indicadoras en cero, las variables continuas en sus medias y la aprobación en su rango de valores hipotéticos. La tercera línea nombra las columnas de la matriz después de las variables de nuestro modelo. En la última línea, la matriz de valores predictores se convierte en un marco de datos.

Una vez que tenemos el marco de datos de los predictores en su lugar, podemos usar el predecir comando para pronosticar los recuentos esperados para los modelos de Poisson y binomial negativo:

```
Forecast.poisson <-predict (energy.poisson, newdata = inputs.4,
        type = "respuesta")
Forecast.nb <-predict (energy.nb, newdata = inputs.4, type = "response")
```

Estas dos líneas solo difieren en el modelo del que extraen estimaciones de coeficientes para el pronóstico.[14] En ambos casos, especificamos type = "respuesta" para obtener predicciones en la escala de conteo.

Para graficar nuestros pronósticos de cada modelo, podemos escribir:

```
plot (y = previsión.poisson, x = aprobación, tipo = "l", lwd = 2,
        ylim = c (0,60), xlab = "Aprobación presidencial", ylab = "Recuento previsto
    de historias de política energética") líneas (y = previsión.nb, x = aprobación, lty =
    2, col = "azul", lwd = 2) leyenda (x = 50, y = 50, leyenda = c ("Poisson", "Binomio
    negativo"),
        lty = c (1,2), col = c ("negro", "azul"), lwd = 2)
```

La primera línea traza las predicciones de Poisson como una línea con el tipo = "l" opción. La segunda línea agrega las predicciones binomiales negativas, coloreando la línea de azul y punteándola conlty = 2. Finalmente, el leyenda El comando nos permite distinguir rápidamente qué línea representa qué modelo. La salida completa se presenta en la Fig.7.3. Las predicciones de los dos modelos son similares y muestran una similar efecto negativo de la aprobación. El modelo binomial negativo tiene un poco más bajo pronostica a valores bajos aprobación y un efecto ligeramente menos profundo se de aprobación, tal de los recuentos previstos superponen a valores altos de aprobación.

[14] Como nota al margen, en utilizando Rcomandos de álgebra matricial, descrito más adelante el Cap. 10, el usuario lata calcular predicho cuenta fácilmente con alterno sintaxis. Por ejemplo, para el modelo binomial negativo, podríamos haber escrito: Forecast.nb <-exp (as.matrix (input.4)% *% energy.nb $ coeficientes).

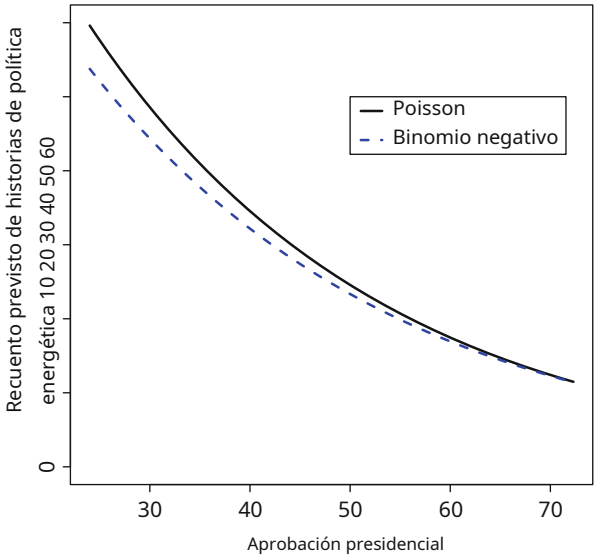

Figura 7.3 Recuento previsto de historias sobre política energética en los noticieros televisivos como función de presidencial aprobación, manteniendo los predictores continuos en su media y los predictores nominales en su modo. Predicciones basadas en Poisson y resultados del modelo binomial negativo

Después de los primeros siete capítulos de este volumen, los usuarios ahora deberían poder realizar la mayoría de las tareas básicas para las que está diseñado el software estadístico: administrar datos, calcular estadísticas simples y estimar modelos comunes. En los cuatro capítulos restantes de este libro, pasaremos ahora a las características únicas deR que permiten el usuario mayor fl exibilidad para aplicar métodos avanzados con paquetes desarrollados por usuarios y otro herramientas de programación en R.

7.4 Problemas de práctica

1. Regresión logística: cargue el extranjero biblioteca y descargue un subconjunto de Singh (2015) datos de encuestas transnacionales sobre participación electoral, el archivo stdSingh.dta, disponible en el Dataverse que se enumera en la página vii o en el contenido del capítulo que se enumera en la página 97. La variable de resultado es si el encuestado votó (**votado).** Un predictor clave, con el que interactúan varias variables, es el grado en que un ciudadano está sujeto a las reglas de votación obligatorias. Esto se mide con una escala de cuán severas son las reglas de voto obligatorio (**gravedad).** Se debe interactuar con cinco predictores **gravedad:** edad (**edad),** conocimiento político**polinfrel),** ingreso (**ingreso),** eficacia**eficacia),** y partidismo**partyID).** Se deben incluir cinco predictores más solo para efectos aditivos: magnitud del distrito (**dist_magnitude),** número de fiestas **enep),** margen de victoria

(**vicmarg_dist),** sistema parlamentario (**parlamentario),** y PIB per cápita (**desarrollo).** Todas las variables predictoras se han estandarizado.

una. Estime un modelo de regresión logística con estos datos, incluidos los cinco términos de interacción.

B. ¿Cuál es la razón de probabilidades para el número de partidos? ¿Cómo interpretaría este término?

C. Grafique el efecto del PIB per cápita sobre la probabilidad de salir. Mantenga todos los predictores distintos del desarrollo en su media.*Insinuación:* Construir sobre el código a partir de la página 107. Si usasteRla notación de interacción (p. ej., si edad * severidad es un término en el modelo), luego, cuando crea un nuevo conjunto de datos de valores de predictores, solo necesita definir sus valores para las variables originales, y no por los productos. En otras palabras, necesitaría una columna para**edad,** por **gravedad,** y para cualquier otro predictor, pero no para **gravedad de la edad.**

D. Bonificación: Trace el efecto de la edad sobre la probabilidad de participación en tres circunstancias: Cuando la severidad de las reglas de votación obligatoria es mínima, media y máxima. Mantenga todos los demás predictores, además de la edad y la gravedad, en su media. Su resultado final debe mostrar tres líneas de probabilidad predichas diferentes.

2. Logit ordenado: Los problemas de práctica del Capítulo 2 introdujeron los de Hanmer y Kalkan (2013) subconjunto del Estudio Electoral Nacional Estadounidense de 2004. Si aún no tiene estos datos, el archivohanmerKalkanANES.dta se puede descargar desde el Dataverse vinculado en la página vii o desde el contenido del capítulo vinculado en la página 97. Cargue el extranjero biblioteca y abra estos datos. (Nuevamente, asegúrese de especificar elconvert.factors = F opción.) Considere dos variables de resultado: evaluaciones económicas retrospectivas (**retecon** tomando valores ordinales codificados 1, 0.5, 0, 0.5 y 1) y evaluación del manejo de George W. Bush de la guerra en Irak (**bushiraq,** tomando valores ordinales codificados 0, 0.33, 0.67 y 1). Hay siete variables predictoras: partidismo en una escala de siete puntos (**partyid),** ideología en una escala de siete puntos (**ideol7b),** un indicador de si el encuestado es blanco (**blanco),** un indicador de si el encuestado es mujer (**mujer),** edad del encuestado**edad),** nivel de educación en una escala de siete puntos (**educ1_7),** e ingresos en una escala de 23 puntos (**ingreso).**

una. Estime un modelo logístico ordenado de evaluaciones económicas retrospectivas en función de los siete predictores.

B. ¿Cuál es la razón de posibilidades sobre el coeficiente para mujeres? ¿Cómo interpretaría este término?

C. Estime un modelo logístico ordenado de evaluación del manejo de Bush de la guerra en Irak en función de los siete predictores.

D. ¿Cuál es la razón de probabilidades en el coeficiente de la escala de partidismo de siete puntos? ¿Cómo interpretaría este término?

mi. Bonificación: los resultados de un modelo están sesgados si hay*causalidad recíproca,* lo que significa que una de las variables independientes no solo influye en la variable dependiente, sino que también la variable dependiente influye en la variable independiente

variable. Suponga que le preocupa el sesgo de causalidad recíproca en el modelo de evaluaciones económicas retrospectivas. ¿Qué variable o variables independientes serían más sospechosas de esta crítica?

3. Modelo de conteo: en los problemas de práctica de los capítulos 3 y 4, presentamos los de Peake y Eshbaugh-Soha (2008) análisis de la cobertura de la póliza de medicamentos. Si no tiene sus datos anteriores, descarguedrugCoverage.csv del Dataverse vinculado en la página vii o el contenido del capítulo vinculado en la página 97. La variable de resultado es la cobertura de noticias sobre drogas (**drugmedia),** y las cuatro entradas son 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).**[15]

una. Estime un modelo de regresión de Poisson de cobertura de pólizas de medicamentos en función de los cuatro predictores.

B. Estime un modelo de regresión binomial negativa de cobertura de pólizas de medicamentos en función de los cuatro predictores. Según los resultados de sus modelos, ¿qué modelo es más apropiado, Poisson o binomial negativo? ¿Por qué?

C. Calcule la proporción de recuento del predictor de aprobación presidencial para cada modelo. ¿Cómo interpretaría cada cantidad?

D. Grafique los recuentos previstos de cada modelo dependiendo del nivel de desempleo, que van desde el mínimo al máximo de los valores observados. Mantenga las dos variables del discurso presidencial en cero y mantenga la aprobación presidencial en su media. Con base en esta figura, ¿qué puede decir sobre el efecto del desempleo en cada modelo?

[15] Al igual que en el ejemplo del capítulo, estos son datos de series de tiempo, por lo que los métodos del Cap. 9 son más apropiados.