Capítulo 6
Modelos lineales y diagnósticos de regresión

El modelo de regresión lineal estimado con mínimos cuadrados ordinarios (MCO) es un modelo de caballo de batalla en Ciencias Políticas. Incluso cuando un académico utiliza un método más avanzado que puede hacer suposiciones más precisas sobre sus datos, como la regresión probit, un modelo de conteo o incluso un modelo bayesiano diseñado de manera única, el investigador a menudo se basa en la forma básica de un modelo que es lineal en los parámetros. De manera similar, muchos de losR Los comandos para estas técnicas más avanzadas utilizan una sintaxis funcional que se asemeja al código para estimar una regresión lineal. Por lo tanto, una comprensión de cómo usarR estimar, interpretar y diagnosticar las propiedades de un modelo lineal se presta al uso sofisticado de modelos con una estructura similar.

Este capítulo continúa describiendo el lm**l**en el oido **metro**odel) comando en R, que estima un modelo de regresión lineal con OLS y las diversas opciones del comando. Luego, el capítulo describe cómo realizar diagnósticos de regresión de un modelo lineal. Estos diagnósticos sirven para evaluar si los supuestos críticos de la estimación de MCO se mantienen, o si nuestros resultados pueden estar sujetos a sesgos o ineficiencia.

A lo largo del capítulo, el ejemplo práctico es un análisis del número de horas que los profesores de biología de la escuela secundaria dedican a enseñar la evolución. El modelo replica el trabajo de Berkman y Plutzer (2010, Tabla 7.2), quienes argumentan que este resultado de política se ve afectado por factores a nivel estatal (como los estándares del plan de estudios) y los atributos de los maestros (como la capacitación). Los datos provienen de la Encuesta Nacional de Maestros de Biología de Escuelas Secundarias y consisten en 854 observaciones de maestros de biología de escuelas secundarias que fueron encuestados en la primavera de 2007. El resultado de interés es el número de horas a docentes dedicados a la evolución humana y general en su escuela secundaria biología clase (**hrs_allev),** y las doce variables de entrada son como sigue:

Electrónico suplementario material: La en línea versión de esto capítulo (doi: 10.1007 / 978-3-319-23446-5_6) 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_6

fase 1: Un índice del rigor de los estándares de evolución de noveno y décimo grado en 2007 para el estado en el que trabaja el maestro. Esta variable está codificada en una escala estandarizada con media 0 y desviación estándar 1.

senior_c: Una variable ordinal para la antigüedad del profesor. Codificado3 por 1–2 Años de experiencia, 2 durante 3-5 años, 1 durante 6 a 10 años, 0 durante 11 a 20 años y 1 durante más de 21 años.

ph_senior: Una interacción entre estándares y antigüedad.

notest_p: Una variable indicadora codificada con 1 si el maestro informa que el estado no no tener una prueba de evaluación para la biología de la escuela secundaria, 0 si el estado tiene dicha prueba.

ph_notest_p: Una interacción entre estándares y ninguna prueba estatal.

mujer: Una variable indicadora codificada con 1 si el profesor es mujer, 0 si es hombre. Desaparecido los valores están codificados 9.

biocred3: Una variable ordinal para la cantidad de horas crédito de biología que tiene el maestro. (tanto de posgrado como de pregrado). Codificado 0 durante 24 horas o menos, 1 durante 25 a 40 horas y 2 durante más de 40 horas.

degr3: El número de títulos en ciencias que tiene el profesor, de 0 a 2.

evol_course: Una variable indicadora codificada con 1 si el instructor cursó una universidad específica. curso de nivel sobre evolución, 0 en caso contrario.

certi fi cado: Un indicador codificado con 1 si el maestro tiene una certi fi cación estatal normal, 0 de lo contrario.

idsci_trans: Una medida compuesta, que va de 0 a 1, del grado en que el el maestro piensa en sí mismo como un científico.

confiado: Experiencia autoevaluada en teoría evolutiva. Codificado1 por menos de muchos otros profesores, 0 para "típico" de la mayoría de los profesores, 1 para "muy bueno" en comparación con la mayoría de los profesores de biología de la escuela secundaria, y 2 para "excepcional" y a la par con los profesores de nivel universitario.

6.1 Estimación con mínimos cuadrados ordinarios

Para empezar, necesitamos cargar los datos de la encuesta, que nombraremos evolución. En este ejemplo, cargamos un conjunto de datos con formato Stata. Esto es fácilmente posible a través del extranjero biblioteca, que nos proporciona la read.dta mando:[1]

```
rm (lista = ls ())
biblioteca (extranjera)
evolución <-read.dta ("BPchap7.dta", convert.factors = FALSE)
```

[1]El archivo de datos de Berkman y Plutzer, llamado BPchap7.dta, está disponible en el Dataverse vinculado en la página vii o en el contenido del capítulo vinculado en la página 79. Recuerde que es posible que deba utilizar el setwd comando para señalar dónde ha guardado los datos.

Como regla general, queremos comenzar viendo las estadísticas descriptivas de nuestro conjunto de datos. Como mínimo, use elresumen comando, y quizás algunos de los otros comandos descritos en los Capítulos. 3 y 4:

```
resumen (evolución)
```

Además de las estadísticas descriptivas resumen nos da, también enumerará el número de observaciones faltantes que tenemos en una variable dada (bajo NA's), si falta alguno. La condición predeterminada para la mayoría de los comandos de modelado enR es eliminar cualquier caso en el que falte una observación sobre cualquier variable en un modelo. Por lo tanto, el investigador debe ser consciente no solo de la variación en las variables relevantes, sino también de cuántos casos carecen de observación.[2] Además, los investigadores deben tener cuidado de notar cualquier cosa en las estadísticas descriptivas que se desvíe de los valores de una variable que se enumeran en el libro de códigos. Por ejemplo, en este caso la variable**mujer** tiene un valor máximo de 9. Si sabemos por nuestro libro de códigos que 0 y 1 son los únicos valores observados válidos de esta variable, entonces sabemos que cualquier otra cosa es un código erróneo o (en este caso) un valor faltante.

Antes de continuar, necesitamos reclasificar las observaciones faltantes de **mujer:**

```
evolución $ mujer [evolución $ mujer == 9] <- Resumen de
NA (evolución)
evolución <-subconjunto (evolución,! is.na (mujer))
```

Este comando recodifica solo los valores de **mujer** codificado como 9 como faltante. Como la siguiente llamada aresumen muestra, los 13 valores codificados como 9 ahora se enumeran como faltantes, por lo que se omitirán automáticamente en nuestro análisis posterior. Para asegurarnos de que cualquier cálculo que hagamos se centre únicamente en las observaciones sobre las que ajustamos el modelo, subconjuntamos nuestros datos para excluir las observaciones que faltan. Como alternativa al usosubconjunto aquí, si tuviéramos valores perdidos en múltiples variables, en su lugar, podríamos haber querido escribir: evolución <-na.omit (evolución).

Habiendo limpiado nuestros datos, pasamos ahora al modelo de horas dedicadas a la enseñanza de la evolución descrito al comienzo del capítulo. Estimamos nuestro modelo lineal usando OLS:

```
mod.horas <-lm (hrs_allev ~ fase1 * senior_c + fase1 * notest_p +
        female + biocred3 + degr3 + evol_course + certificada + idsci_trans +
        seguro, datos = evolución) resumen
(horas mod.)
```

La sintaxis estándar para especificar la fórmula de un modelo es enumerar la variable de resultado a la izquierda de la tilde (~) y las variables de entrada en el lado derecho separadas por signos más. Tenga en cuenta que incluimos dos condiciones: fase1 * senior_c y fase1 * notest_p. Considerando el primero, fase1 * senior_c, esto *notación interactiva* agrega tres términos a nuestro modelo: **fase1, senior_c,** y el producto de los dos. Estos modelos interactivos permiten

[2]Una alternativa teóricamente atractiva a *eliminación de lista* como medio de manejar los datos faltantes es *imputación múltiple.* Ver Little y Rubin (1987), Frotar (1987) y King et al. (2001) para más detalles.

efectos condicionales de una variable.[3] La datos opción de lm nos permite llamar a variables del mismo conjunto de datos sin tener que hacer referencia al nombre del conjunto de datos con cada variable. Otras opciones destacadas para ellm comando incluye subconjunto, que permite al usuario analizar solo una parte de un conjunto de datos, y pesos que permite al usuario estimar un modelo lineal con mínimos cuadrados ponderados (WLS). Observe que tuvimos que nombrar nuestro modelo en la estimación, llamándolohoras mod. por elección, y para obtener los resultados de nuestra estimación, necesitamos llamar a nuestro modelo con el resumen
mando. La salida deresumen (horas mod.) Se ve como esto:

Llamada:
lm (fórmula = hrs_allev ~ fase1 * senior_c + fase1 * notest_p +
female + biocred3 + degr3 + evol_course + certificada + idsci
_trans +
seguro, datos = evolución)

Derechos residuales de autor:

Min	1T	Mediana	3T	Max
- 20.378	-6.148	- 1.314	4.744	32.148

Coeficientes:

	Estimar	Std.	Valor t	de error	Pr (> | t |)	
(Interceptar)	10,2313	1,1905	8,594	<2e-16	***	
fase 1	0,6285	0,3331	1,886	0,0596.		
senior_c	- 0,5813	0,3130	-1,857	0,0636.		
notest_p	0,4852	0,7222	0,672	0,5019		
mujer	- 1,3546	0,6016	-2,252	0,0246	*	
biocred3	0,5559	0,5072	1,096	0,2734		
degr3	- 0,4003	0,3922	-1,021	0,3077		
curso_evolucionario	2.5108	0.6300	3.985	7.33e-05	***	
certificado	- 0,4446	0,7212	-0,617	0,5377		
idsci_trans	1,8549	1,1255	1,648	0,0997.		
seguro	2.6262	0.4501	5.835	7.71e-09	***	
fase1: senior_c	- 0,5112	0,2717	-1,881	0,0603.		
fase1: notest_p	- 0,5362	0,6233	-0,860	0,3899		
- - -						
Signif. códigos:	0 *** 0,001 ** 0,01 * 0,05. 0,1					1

Error estándar residual: 8,397 en 828 grados de libertad
múltiple: 0,1226, Estadístico F de R ajustado: 9,641 en 12 - al cuadrado: 0.1099
y 828 DF, valor p: <2.2e-16

[3]Ver Brambor et al. (2006) para obtener más detalles sobre los términos de interacción. Además, tenga en cuenta que se podría lograr una especificación equivalente de este modelo reemplazando fase 1* senior_c y fase1 * notest_p con los términos fase1 + senior_c + ph_senior + notest_p + ph_notest_p. Simplemente estamos introduciendo cada uno de los términos por separado de esta manera.

Cuadro 6.1 Modelo lineal de horas de clase dedicadas a la enseñanza de la evolución por alto profesores de biología de la escuela (estimaciones OLS)

Vaticinador	Estimar	Std. Error	*t* valor	Pr (> j*t*j)
Interceptar	10: 2313	1.1905	8:59	0,0000
Índice de estándares 2007	0: 6285	0.3331	1:89	0.0596
Antigüedad (centrada)	0: 5813	0.3130	1:86	0.0636
Antigüedad de los estándares	0: 5112	0.2717	1:88	0.0603
Cree que no hay exámenes Los	0: 4852	0,7222	0:67	0.5019
estándares creen que no hay exámenes	0: 5362	0,6233	0:86	0.3899
El maestro es mujer	1: 3546	0,6016	2:25	0.0246
Créditos obtenidos en biología (0-2)	0: 5559	0.5072	1:10	0.2734
Grados en ciencias (0-2)	0: 4003	0.3922	1:02	0.3077
Clase de evolución completada	2: 5108	0,6300	3:99	0,0001
Tiene certi fi cación normal Se	0: 4446	0,7212	0:62	0.5377
identifica como científico	1: 8549	1.1255	1:65	0.0997
Experiencia autoevaluada (1 a +2)	2: 6262	0.4501	5:84	0,0000

Notas: N D 841. R^2 D 0: 1226. $F_{12;\,828}$ D 9: 641 (p <0: 001). Datos de Berkman y Plutzer (2010)

La parte superior de la impresión repite el comando de modelo especificado por el usuario y luego proporciona algunas estadísticas descriptivas de los residuos. La siguiente tabla presenta los resultados de interés principal: La primera columna enumera todos los predictores del modelo, incluida una intersección. La segunda columna presenta la estimación de MCO del coeficiente de regresión parcial. La tercera columna presenta el*t*-razón para una hipótesis nula de que el coeficiente de regresión parcial es cero, y la cuarta columna presenta un coeficiente de dos colas *pag*-valor para el *t*-proporción. Finalmente, la tabla imprime puntos y estrellas en función de los umbrales que las dos colas*pag*-cruces de valores.[4] Debajo de la tabla, se informan varios estadísticos de ajuste: el error estándar de regresión (o error estándar residual), el R^2 y ajustado R^2 valores, y el *F*-pruebe si el modelo en su conjunto explica una parte significativa de la varianza. Los resultados de este modelo también se presentan de manera más formal en la Tabla6.1.[5]

[4] Se recuerda a los usuarios que para las pruebas de una cola, en las que el usuario desea probar que el coeficiente parcial específicamente es mayor o menor que cero, el *pag*-el valor será diferente. Si el signo del coeficiente coincide con la hipótesis alternativa, entonces el correspondiente*pag*-el valor es la mitad de lo que se informa. (Naturalmente, si el signo del coeficiente es opuesto al signo de la hipótesis alternativa, los datos no concuerdan con la hipótesis del investigador). Además, los investigadores pueden querer probar una hipótesis en la que la hipótesis nula sea algo diferente de cero: En En este caso, el usuario puede construir el correcto*t*-relación utilizando la estimación informada y el error estándar.

[5] Investigadores que escriben sus documentos con LATEX puede transferir fácilmente los resultados de un modelo lineal desde R a una mesa usando el xtable Biblioteca. (HTML también es compatible conxtable.) En el primer uso, instale con: install.packages ("xtable"). Una vez instalado, simplemente ingresando biblioteca (xtable); xtable (horas mod.)produciría LACódigo listo para TEX para una tabla similar a Table 6.1. Como otra opción para generar resultados, consulte lartf paquete sobre cómo generar resultados en formato de texto enriquecido.

Muchos investigadores, en lugar de informar *t*-ratios y *pag*-valores presentados en la salida predeterminada de lm en su lugar informará *intervalos de confianza* de sus estimaciones. Se debe tener cuidado en la interpretación de los intervalos de confianza, por lo que se insta a los lectores que no estén familiarizados con estos a consultar un libro de texto de estadística o econometría para obtener más información (como Gujarati y Porter2009, págs. 108-109). Para construir tal**conf**idence **En t**erval en R, el usuario debe elegir un nivel de confianza y utilizar el confinar mando:

```
confint (mod. horas, nivel = 0,90)
```

La nivel La opción es donde el usuario especifica el nivel de confianza. 0,90 corresponde a un 90% de confianza, mientras que nivel = 0,99, por ejemplo, produciría un intervalo de confianza del 99%. Se informan los resultados de nuestro intervalo de confianza del 90% como sigue:

	5%	95%
(Interceptar)	8.27092375	12.19176909
fase 1	0.07987796	1.17702352
senior_c	- 1.09665413	- 0.06587642
notest_p	- 0,70400967	1.67437410
mujer	- 2.34534464	- 0.36388231
biocred3	- 0.27927088	1.39099719
degr3	- 1.04614354	0,24552777
curso_evolucionario	1.47336072	3.54819493
certificado	- 1.63229086	0,74299337
idsci_trans	0,00154974	3.70834835
seguro	1.88506881	3.36729476
fase1: senior_c	-0.95856134	fase1:
notest_p	-1.56260919	0,49020149

Entre otras características, un atributo útil de estos es que un lector puede examinar un intervalo de confianza del 90% (por ejemplo) y rechazar cualquier hipótesis nula que proponga un valor fuera del rango del intervalo para una prueba de dos colas. Por ejemplo, el intervalo de la variable**confiado** no incluye cero, por lo que podemos concluir con un 90% de con fi anza de que el coeficiente parcial de esta variable es diferente de cero.[6]

6.2 Diagnóstico de regresión

Solo nos contentamos con usar MCO para estimar un modelo lineal si es el mejor estimador lineal imparcial (AZUL). En otras palabras, queremos obtener estimaciones que, en promedio, produzcan el verdadero parámetro de población (insesgado), y entre insesgados

[6] De hecho, también podríamos concluir que el coeficiente es *mayor que* que cero en el nivel de confianza del 95%. Para obtener más información sobre cómo los intervalos de confianza también pueden ser útiles para las pruebas de una cola, consulte Gujarati y Porter (2009, pag. 115).

Estimadores Queremos el estimador que minimice la varianza del error de nuestras estimaciones (mejor o e fi ciente). Según el teorema de Gauss-Markov, MCO es AZUL y válido para inferencias si se cumplen cuatro supuestos:

1. Valores de entrada fijos o exógenos. En otras palabras, los predictores (*X)* debe ser independiente pendiente del término de error. Cov.$X_{2I;}$ $tu_{I/}$D Cov.$X_{3I;}$ $tu_{I/}$DD Cov.$X_{ki;}$ $tu_{I/}$D 0.
2. Forma funcional correcta. En otras palabras, la media condicional de la perturbación debe ser cero.
   $UEjX_{2I;}$ $X_{3I;}$ :::; $X_{ki\,/}$D 0.
3. Homoscedasticidad o varianza constante de las alteraciones ($tu_{I).}$ Var.$tu_{I/}$D 2.
4. No existe autocorrelación entre perturbaciones. Cov.$tu_{I;}$ $tu_{j\,/}$D 0 por I ¤ *j.*

Si bien nunca observamos los valores de las perturbaciones, ya que estos son términos de población, podemos predecir los residuos (*tu*O) después de estimar un modelo lineal. Por lo tanto, normalmente usaremos residuos para evaluar si estamos dispuestos a hacer los supuestos de Gauss-Markov. En las siguientes subsecciones, realizamos diagnósticos de regresión para evaluar las diversas suposiciones y describir cómo podríamos llevar a cabo medidas correctivas enR para corregir las aparentes violaciones de los supuestos de Gauss-Markov. La única excepción es que no probamos el supuesto de*sin autocorrelación* porque no podemos hacer referencia a nuestros datos de ejemplo por tiempo o espacio. Ver cap.9 para ver ejemplos de pruebas y correcciones de autocorrelación. Además, describimos cómo diagnosticar si los errores tienen una distribución normal, que es esencial para la inferencia estadística. Finalmente, consideramos la presencia de dos características de datos notables, multicolinealidad y observaciones atípicas, que no forman parte de los supuestos de Gauss-Markov pero que, sin embargo, vale la pena verificar.

6.2.1 Forma funcional

Es fundamental tener la forma funcional correcta en un modelo lineal; de lo contrario, sus resultados serán*tendencioso.* Por lo tanto, al estimar un modelo lineal, debemos evaluar si lo hemos especificado correctamente o si debemos incluir aspectos no lineales de nuestros predictores (como logaritmos, raíces cuadradas, cuadrados, cubos o splines). Como regla general, un diagnóstico esencial para cualquier modelo lineal es hacer un diagrama de dispersión de los residuos (*tu*O). Estos gráficos deben realizarse contra ambos valores ajustados (*Y*O) y contra los predictores (*X).* Para construir una gráfica de residuos contra valores ajustados, simplemente haríamos referencia a los atributos del modelo que estimamos en una llamada al gráfico mando:

```
plot (y = mod.horas $ residuales, x = mod.horas $ valores ajustados,
      xlab = "Valores ajustados", ylab = "Residuales")
```

Darse cuenta de horas mod. $ residuales nos permitió hacer referencia a los residuos del modelo (*tu*O), y mod.horas $ valores ajustados nos permitió llamar a los valores predichos (*Y* O). Podemos hacer referencia a muchas funciones con el signo de dólar ($). Tipo nombres (horas mod.) para ver todo lo que se guarda. Volviendo a nuestro gráfico de salida, se presenta en la Fig.6.1. Como analistas, deberíamos comprobar este gráfico en busca de algunas características: ¿El promedio local de los residuos tiende a permanecer alrededor de cero? Si el

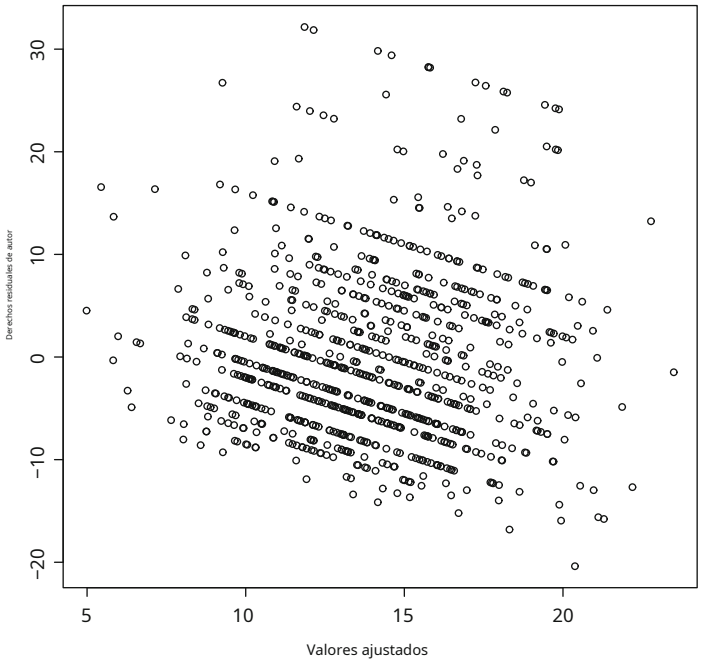
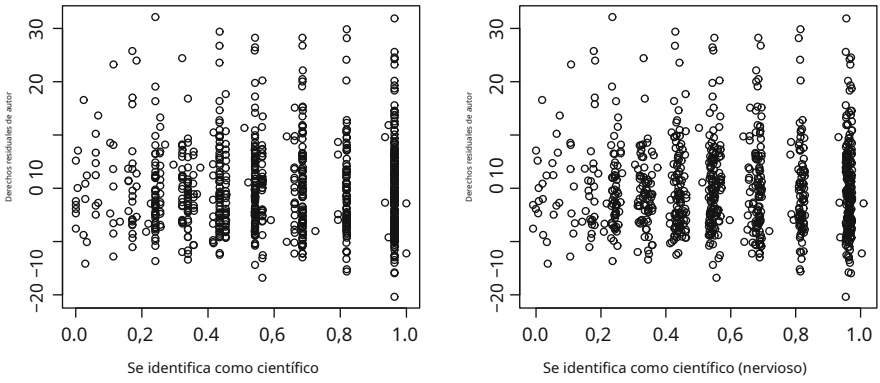


Figura 6.1 Diagrama de dispersión de residuos frente a valores ajustados del modelo de evolución de las horas de enseñanza

Los residuos muestran un patrón claro de aumento o disminución en cualquier rango, entonces la forma funcional de alguna variable puede ser incorrecta. ¿El margen de los residuos difiere en alguna parte del gráfico? Si es así, puede haber un problema de heterocedasticidad. Una característica aparente de la Fig.6.1 es que los residuos parecen golpear un "piso" diagonal cerca del fondo de la nube. Esto surge porque un profesor no puede dedicar menos de cero horas a la enseñanza de la evolución. Por tanto, este suelo natural refleja un límite en la variable dependiente. Una limitación de forma funcional como esta a menudo se aborda mejor dentro del marco del modelo lineal generalizado, que se considerará en el próximo capítulo.

Otra herramienta útil es dibujar cifras de los residuales contra uno o más predictores. Figura6.2 muestra dos gráficos de los residuos de nuestro modelo frente a la escala compuesta del grado en que el profesor se autoidentifica como científico. Figura6.2a muestra la gráfica básica utilizando los datos brutos, que un investigador siempre debe mirar. En este caso, el predictor de interés toma 82 valores únicos, pero muchas observaciones toman los mismos valores, particularmente en el extremo superior de la escala. En casos como este, muchos puntos de la trama se superpondrán entre sí. Por*tembloroso* los valores de **idsci_trans,** o agregando un pequeño número extraído al azar, es más fácil ver dónde está la preponderancia de los datos. Figura6.2b muestra una gráfica revisada que hace temblar al predictor. El riesgo de la figura nerviosa es que mover los datos puede distorsionar un patrón verdadero entre el predictor y los residuales. Sin embargo, en el caso de una variable de entrada ordinal (o quizás semi-ordinal), las dos subfiguras



a) Datos brutos **B)** Datos alterados

Figura 6.2 Diagrama de dispersión de residuos contra el grado a que un profesor identifica como científico. (**a)** Datos brutos. (**B)** Datos alterados

pueden complementarse para ofrecer la imagen más completa posible. Los dos diagramas de dispersión de la Fig. 6.2 se producen de la siguiente manera:

```
plot (y = mod.horas $ residuales, x = evolución $ idsci_trans,
      xlab = "Se identifica como científico", ylab = "Residuales") plot (y =
mod.hours $ residuales, x = jitter (evolution $ idsci_trans,
      cantidad = .01), xlab = "Se identifica como científico (nervioso)", ylab =
      "Residuales")
```

Al igual que el gráfico de valor residual a ajustado de la Fig. 6.1, examinamos los gráficos de predictores de residuos de la Fig. 6.2 para los cambios en la media local, así como las diferencias en la dispersión de los residuos, cada uno de ellos depende del valor del predictor. En la forma funcional, hay pocos indicios de que la media móvil esté cambiando marcadamente entre los valores. Por lo tanto, al igual que con el gráfico de residuo a ajustado, vemos poca necesidad de volver a especificar nuestro modelo con una versión no lineal de este predictor. Sin embargo, la propagación de los residuos parece un poco preocupante, por lo que volveremos a tratar este tema en la siguiente sección.

Además de los métodos gráficos, una estadística de prueba común para diagnosticar una forma funcional mal especificada es la prueba RESET de Ramsey (prueba de error de especificación de regresión). Esta prueba procede reestimando el modelo original, pero esta vez incluyendo los valores ajustados del modelo original en alguna forma no lineal (como una fórmula cuadrática o cúbica). Usando un *F*-La razón para evaluar si el nuevo modelo explica significativamente más varianza que el modelo antiguo sirve como prueba de si se debe incluir una forma diferente de uno o más predictores en el modelo. Podemos realizar esta prueba para una forma funcional cúbica potencial de la siguiente manera:

```
evolución $ ajuste <-mod.horas $ valores ajustados
reset.mod <-lm (hrs_allev ~ fase1 * senior_c + fase1 * notest_p +
      female + biocred3 + degr3 + evol_course + certificada + idsci_trans +
      seguro + I (ajuste ^ 2) + I (ajuste ^ 3), datos = evolución) anova
(mod.hours, reset.mod)
```

La primera línea de código guarda los valores ajustados del modelo original como una variable en el marco de datos. La segunda línea agrega formas al cuadrado y al cubo de los valores ajustados en el modelo de regresión. Al incorporar estos términos en elI función (nuevamente significando, "como **I** s "), podemos transformar algebraicamente la variable de entrada sobre la marcha a medida que estimamos el modelo. Tercero, elanova comando (para **un**análisis **o**F **Virginia**riance) presenta los resultados de un *F*-prueba que compara el modelo original con el modelo que incluye una forma cuadrática y cúbica de los valores ajustados. En este caso, obtenemos un resultado de F_{2826} D 2: 5626, con un *pag*-valor de *pag* D 0: 07772. Esto indica que el modelo con el polinomio cúbico de valores ajustados se ajusta significativamente mejor al 90% nivel, lo que implica que otra forma funcional sería mejor.

Para determinar qué predictor podría ser el culpable de la forma funcional mal especificada, podemos realizar pruebas de Durbin-Watson en los residuos, clasificando el predictor que puede ser problemático. (*Nota* que tradicionalmente las pruebas de Durbin-Watson se clasifican en *hora* para probar la autocorrelación temporal. Esta idea se revisa en el cap.9.) Un resultado discernible indica que los residuos toman valores similares en valores similares de la entrada, una señal de que el predictor necesita ser respecificado. Lalmtest biblioteca (los usuarios necesitarán instalar con install.packages la primera vez) proporciona comandos para varias pruebas de diagnóstico, incluida la prueba de Durbin-Watson. Clasificación de los residuos según el rigor de los estándares de evolución (**fase 1),** ejecutamos la prueba:

```
biblioteca install.packages ("lmtest")
(lmtest)
dwtest (mod.horas, order.by = evolución $ fase1)
```

Esto produce una estadística de *D* D 1: 8519 con un aproximado *pag*-valor de *pag* D 0: 01368, lo que indica que los residuos son similares en función del valor de la covariable. Por lo tanto, podríamos proceder a volver a cifrar nuestra forma funcional agregando términos polinomiales para**fase 1:**

```
mod.cubic <-lm (hrs_allev ~ fase1 * senior_c + fase1 * notest_p +
        female + biocred3 + degr3 + evol_course + certificada + idsci_trans +
        seguro + I (fase1 ^ 2) * senior_c + I (fase1 ^ 3) * senior_c + I (fase1 ^ 2) * notest_p +
        I (fase1 ^ 3) * notest_p, data = evolution)
```

Al igual que con la prueba RESET en sí, nuestro nuevo modelo (mod.cúbica) ilustra cómo podemos utilizar funciones adicionales del lm mando. Nuevamente, usando elI función, podemos realizar álgebra en cualquier variable de entrada dentro del comando modelo. Como antes, el signo de intercalación (^)ˆeleva una variable a una potencia, lo que permite nuestra función polinomial. De nuevo, por términos de interacción, simplemente multiplicar dos variables con un asterisco (*) asegura que se incluyan los efectos principales y los términos del producto de todas las variables en la interacción. Por lo tanto, permitimos que la antigüedad y si no hay una prueba de evaluación interactúen con la forma polinomial completa de los estándares de evolución.

6.2.2 Heteroscedasticidad

Cuando la varianza del error en los residuos no es uniforme en todas las observaciones, un modelo tiene una varianza del error heterocedástica, las estimaciones son ineficientes y los errores estándar están sesgados para ser demasiado pequeños. La primera herramienta que usamos para evaluar si la varianza del error es homocedástica (o constante para todas las observaciones) versus heterocedástica es una gráfica de dispersión simple de los residuos. Figura6.1 nos ofreció la gráfica de nuestros residuos contra los valores ajustados, y la Fig. 6.2 ofrece un gráfico de ejemplo de los residuos frente a un predictor. Además de estudiar la media móvil para evaluar la forma funcional, también evaluamos la dispersión de los residuos. Si la dispersión de los residuos es una banda constante alrededor de cero, entonces podemos usar esto como una con fi rmación visual de homocedasticidad. Sin embargo, en los dos paneles de la Fig.6.2, podemos ver que la preponderancia de residuos se concentra más estrechamente cerca de cero para los profesores que están menos inclinados a identificarse a sí mismos como científicos, mientras que los residuos están más dispersos entre aquellos que están más inclinados a identificarse como científicos. (Los residuos extremos son aproximadamente los mismos para todos los valores de*X,* lo que hace que sea un poco más difícil de detectar, pero la dispersión de puntos de datos concentrados en el medio se expande a valores más altos). Todo esto sugiere que la autoidentificación como científico se corresponde con la heterocedasticidad para este modelo.

Además de los métodos visuales, también tenemos la opción de utilizar una estadística de prueba en una prueba de Breusch-Pagan. Utilizando lalmtest biblioteca (que cargamos anteriormente), la sintaxis es la siguiente:

```
bptest (mod.horas, studentize = FALSE)
```

El defecto de bptest es utilizar la versión estudiantil de Koenker de esta prueba. Por lo tanto, la studentize = FALSO Esta opción le da al usuario la opción de utilizar la versión original de la prueba Breusch-Pagan. La hipótesis nula en esta prueba de chi-cuadrado es la homocedasticidad. En este caso, nuestra estadística de prueba es $_{12df}^{2}$ D 51: 7389 (p <0: 0001). Por lo tanto, rechazamos la hipótesis nula y concluimos que los residuos no son homocedásticos.

Sin homocedasticidad, nuestros resultados no son eficientes, entonces, ¿cómo podríamos corregir esto? Quizás la solución más común a este problema es utilizar errores estándar robustos de Huber-White o errores estándar tipo sándwich (Huber1967; blanco1980). La desventaja de este método es que ignora la ine fi ciencia de las estimaciones de MCO y continúa reportándolas como estimaciones de los parámetros. La ventaja, sin embargo, es que aunque las estimaciones de MCO son ineficientes bajo heterocedasticidad, son insesgadas. Dado que los errores estándar están sesgados, corregirlos soluciona el mayor problema que nos presenta la heterocedasticidad. El cálculo de los errores estándar de Huber-White se puede lograr utilizando elemparedado (necesita una primera instalación) y lmtest bibliotecas:

```
biblioteca install.packages ("sándwich")
(sándwich)
coeftest (horas mod., vcov = vcovHC)
```

La lmtest la biblioteca hace el coeftest comando disponible, y el emparedado biblioteca hace la matriz de varianza-covarianza vcovHC disponible dentro de este. (Ambas bibliotecas requieren instalación en el primer uso).coeftest El comando ahora presentará los resultados de horas mod. nuevamente, con las mismas estimaciones de MCO que antes, los nuevos errores estándar de Huber-White y los valores de *t* y *pag* que corresponden a los nuevos errores estándar.

Finalmente, también tenemos la opción de volver a estimar nuestro modelo usando WLS. Para hacer esto, el analista debe construir un modelo de los residuos cuadrados como una forma de pronosticar la varianza del error heterocedástico para cada observación. Si bien hay algunas formas de hacer esto de manera efectiva, aquí está el código para un plan. Primero, guardamos los residuos al cuadrado y ajustamos un modelo auxiliar del logaritmo de estos residuos al cuadrado:

```
evolución $ resid2 <-mod.horas $ residuales ^ 2
peso.reg <-lm (log (resid2) ~ fase1 * senior_c + fase1 * notest_p +
        female + biocred3 + degr3 + evol_course + certificada + idsci_trans +
        seguro, datos = evolución)
```

Una advertencia clave de WLS es que todos los pesos deben ser no negativos. Para garantizar esto, el código aquí modela el logaritmo de los residuos al cuadrado; por lo tanto, el exponencial de los valores ajustados de esta regresión auxiliar sirve como predicciones positivas de los residuos al cuadrado. (También existen otras soluciones a este problema). La regresión auxiliar simplemente incluye todos los predictores de la regresión original en su forma lineal, pero el usuario no está vinculado a esta suposición. De hecho, WLS ofrece el AZUL bajo heterocedasticidad, pero solo si el investigador modela correctamente la varianza del error. Por tanto, la especificación adecuada de la regresión auxiliar es esencial. En WLS, esencialmente queremos ponderar valores con una varianza de error baja y dar poca importancia a aquellos con una varianza de error alta. Por tanto, para nuestra regresión final de WLS, lapesos El comando toma el recíproco de los valores predichos (exponenciados para estar en la escala original de los residuos al cuadrado):

```
wls.mod <-lm (hrs_allev ~ fase1 * senior_c + fase1 * notest_p +
        female + biocred3 + degr3 + evol_course + certificada + idsci_trans +
        seguro, datos = evolución, pesos = I (1 / exp (peso.reg $
    valores ajustados))) resumen (wls.mod)
```

Esto nos presenta un conjunto de estimaciones que explica la heterocedasticidad en los residuos.

6.2.3 Normalidad

Si bien no forma parte del teorema de Gauss-Markov, una suposición importante que hacemos con los modelos de regresión lineal es que las perturbaciones se distribuyen normalmente. Si esta suposición no es cierta, entonces OLS sigue siendo AZUL. Sin embargo, el supuesto de normalidad es esencial para que nuestras estadísticas inferenciales habituales sean precisas. Por lo tanto, probamos este supuesto examinando la distribución empírica de la

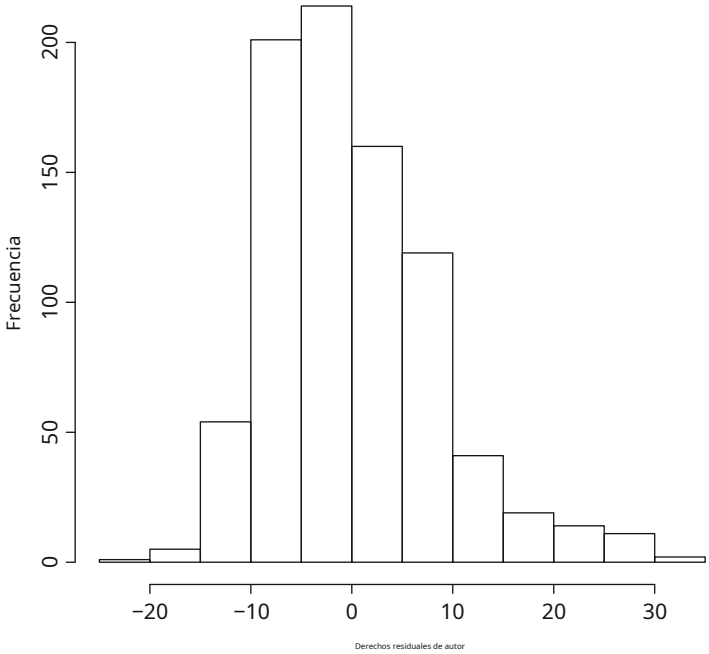


Figura 6.3 Histograma de residuos del modelo de horas de evolución docente

Residuos predichos. Un primer lugar fácil para comenzar es examinar los un histograma de residuos.

```
hist (mod.horas $ residuales, xlab = "Residuales", main = "")
```

Este histograma se muestra en la Fig. 6.3. Generalmente, nos gustaría una curva de campana simétrica que no sea ni excesivamente plana ni puntiaguda. Si ambos*sesgar*refiriéndose a si la distribución es simétrica o si las colas son pares) y *la curtosis*refiriéndose al pico de la distribución) son similares a una distribución normal, podemos usar esta figura a favor de nuestra suposición. En este caso, los residuos parecen estar sesgados a la derecha, lo que sugiere que la normalidad no es un supuesto seguro en este caso.

Una cifra un poco más compleja (aunque potencialmente más informativa) se denomina gráfico de cuantiles-cuantiles. En esta figura, los cuantiles de los valores empíricos de los residuos se grafican contra los cuantiles de una distribución normal teórica. Cuanto menos correspondan estas cantidades, menos razonable es suponer que los residuos se distribuyen normalmente. Tal figura se construye enR como sigue:

```
qqnorm (mod.horas $ residuales) qqline (mod.horas
$ residuales, col = "rojo")
```

La primera línea de código (qqnorm) en realidad crea el gráfico de cuantiles-cuantiles. La segunda línea (qqline) agrega una línea guía a la parcela existente. El gráfico completo se encuentra en la Fig. 6.4. Como puede verse, en cuantiles inferiores y superiores, los valores muestrales

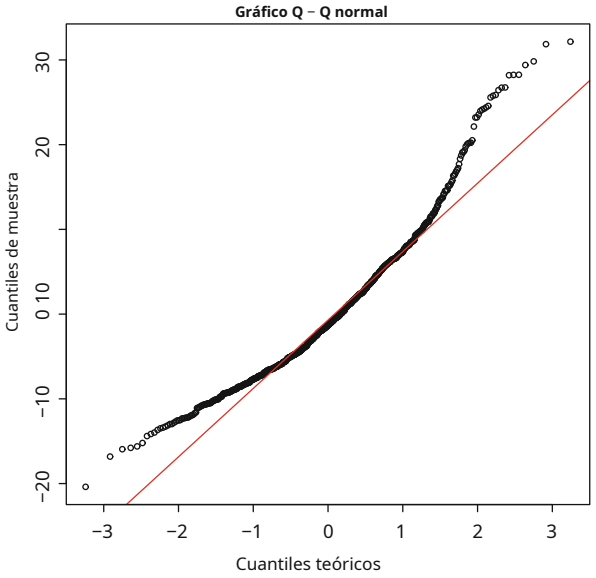


Figura 6.4 Gráfico de cuantiles-cuantiles normales para los residuos del modelo de evolución de las horas lectivas

desviarse sustancialmente de los valores teóricos. Nuevamente, esta figura cuestiona el supuesto de normalidad.

Además de estas evaluaciones sustancialmente enfocadas de la distribución empírica, los investigadores también pueden usar estadísticas de prueba. El estadístico de prueba más comúnmente utilizado en este caso es la prueba de Jarque-Bera, que se basa en el sesgo y la curtosis de la distribución empírica de los residuos. Esta prueba utiliza la hipótesis nula de que los residuos están distribuidos normalmente y la hipótesis alternativa de que no.[7] La tseries biblioteca puede calcular esta estadística, que instalamos en el primer uso:

```
biblioteca install.packages ("tseries")
(tseries)
jarque.bera.test (mod.horas $ residuales)
```

En nuestro caso, 2 D 191: 5709, por lo que rechazamos la hipótesis nula y concluimos que los residuos no se distribuyen normalmente. Al igual que los diagnósticos de heterocedasticidad, preferiríamos un resultado nulo ya que preferimos no rechazar la suposición.

Los tres diagnósticos indican una violación del supuesto de normalidad, entonces, ¿cómo podríamos responder a esta violación? En muchos casos, la mejor respuesta probablemente se encuentre en el próximo capítulo sobre modelos lineales generalizados (GLM). Bajo este marco, podemos asumir una gama más amplia de distribuciones para la variable de resultado, y también podemos transformar la variable de resultado a través de una función de enlace. Otro algo

[7] En otras palabras, si no rechazamos la hipótesis nula para una prueba de Jarque-Bera, entonces concluimos que no hay evidencia significativa de no normalidad. Tenga en cuenta que esto es diferente a concluir que tenemos normalidad. Sin embargo, esta es la conclusión más sólida que podemos sacar con esta estadística de prueba.

Una opción similar sería transformar la variable dependiente de alguna manera. En el caso de nuestro ejemplo de ejecución sobre las horas dedicadas a la evolución, nuestra variable de resultado no puede ser negativa, por lo que podríamos agregar 1 a la respuesta de cada maestro y tomar el logaritmo de nuestra variable dependiente. Sin embargo, tenga en cuenta que esto tiene un mayor impacto en la*forma funcional*er Gujarati y Porter 2009, págs. 162-164), y tenemos que suponer que las perturbaciones del modelo con una variable dependiente registrada se distribuyen normalmente con fines inferenciales.

6.2.4 Multicolinealidad

Aunque no es un supuesto estadístico del modelo lineal, ahora pasamos a diagnosticar la presencia de multicolinealidad entre predictores. Multicolinealidad significa que un predictor es una función de uno o más predictores. Si un predictor es una función exacta de otros predictores, entonces existe una multicolinealidad perfecta en el conjunto de regresores. En una multicolinealidad perfecta, el modelo no puede estimarse tal cual y debe respetarse. Por ejemplo, si un investigador incluye tanto el "año de nacimiento" como la "edad" de un encuestado en un análisis transversal, una variable sería una función perfecta de la otra y, por lo tanto, el modelo no sería estimable.

Una situación común es que un predictor tenga una multicolinealidad alta, pero no perfecta. El problema que surge es que los errores estándar de los coeficientes de regresión comenzarán a aumentar. Sin embargo, es importante destacar que OLS sigue siendo AZUL en el caso de multicolinealidad alta pero imperfecta. En otras palabras, los errores estándar grandes son precisos y aún reflejan el estimador más eficiente posible. Sin embargo, a menudo es una buena idea tener una idea de si la multicolinealidad está presente en un modelo de regresión.

El enfoque general para evaluar la multicolinealidad se basa en regresiones auxiliares de predictores. Entre las medidas de resumen de estos resultados se encuentra el factor de inflación de la varianza (VIF). Para cada predictor, el VIF nos da una idea del grado en que la varianza común entre los predictores aumenta el error estándar del coeficiente del predictor. Los VIF pueden tomar cualquier valor no negativo, y los valores más pequeños son más deseables. Una regla general es que siempre que un VIF excede
10, se puede concluir que la multicolinealidad está dando forma a los resultados.[8]

En R, Los VIF se pueden calcular para todos los coeficientes utilizando el carro biblioteca, instalada en el Cap. 2:

```
biblioteca (coche)
vif (horas mod.)
```

[8]Un VIF de 10 significa que el 90% de la varianza en un predictor puede explicarse por los otros predictores, lo que en la mayoría de los contextos puede considerarse como un gran grado de varianza común. Sin embargo, a diferencia de otras pruebas de diagnóstico, esta regla general no debe considerarse una estadística de prueba. En última instancia, el investigador debe sacar una conclusión sustancial de los resultados.

Cuadro 6.2 Factores de in fl ación de varianza para predictores de horas dedicadas a la enseñanza evolución

Vaticinador	VIF
Índice de estándares 2007	1,53
Antigüedad (centrada)	1.12
Antigüedad de los estándares	1,10
Cree que no hay exámenes Los	1.12
estándares creen que no hay exámenes	1,63
El maestro es mujer	1.08
Créditos obtenidos en biología (0-2)	1,15
Grados en ciencias (0-2)	1,11
Clase de evolución completada	1,17
Tiene certi fi cación normal Se	1.03
identifica como científico	1.12
Experiencia autoevaluada (1 a +2)	1,20

Los VIF calculados de esta manera se presentan en la Tabla 6.2. Como se puede ver en la tabla, todos los VIF son pequeños, lo que implica que la multicolinealidad no es un problema importante en este modelo. Sin embargo, en situaciones en las que surge la multicolinealidad, a veces el mejor consejo es no hacer nada. Para una discusión sobre cómo decidir si no hacer nada es el mejor enfoque o si otra solución funcionaría mejor, consulte Gujarati y Porter (2009, págs. 342-346).

6.2.5 Valores atípicos, apalancamiento y puntos de datos influyentes

Como diagnóstico final, es una buena idea determinar si alguna observación está ejerciendo una influencia excesiva en los resultados de un modelo lineal. Si una o dos observaciones generan un resultado completo que de otra manera no surgiría, entonces un modelo que incluya estas observaciones puede ser engañoso. Consideramos tres tipos de puntos de datos problemáticos: valores atípicos (para los cuales el residual es excesivamente grande), puntos de apalancamiento (que toman un valor de un predictor que está desproporcionadamente distante de otros valores) y puntos de influencia (valores atípicos con mucho apalancamiento). . Los más problemáticos son los puntos de influencia porque tienen la mayor capacidad para distorsionar los coeficientes de regresión parcial.

Una vez más, un diagnóstico simple para estas características de las observaciones es simplemente examinar diagramas de dispersión de residuos, como los que se muestran en las Figs. 6.1 y 6.2. Si una observación se destaca en la escala del predictor, entonces tiene apalancamiento. Si se destaca en la escala residual, entonces es un valor atípico. Si destaca en ambas dimensiones, entonces es un punto de influencia. Ninguna de las cifras de este modelo muestra señales de advertencia al respecto. Otra opción para evaluar estos atributos para las observaciones es calcular las cantidades de residuos estudentizados para detectar valores atípicos, valores de sombrero para detectar puntos de apalancamiento y distancias de Cook para detectar puntos de datos influyentes. Lacarro Una vez más, la biblioteca ofrece una forma sencilla de ver estas cantidades para todas las observaciones.

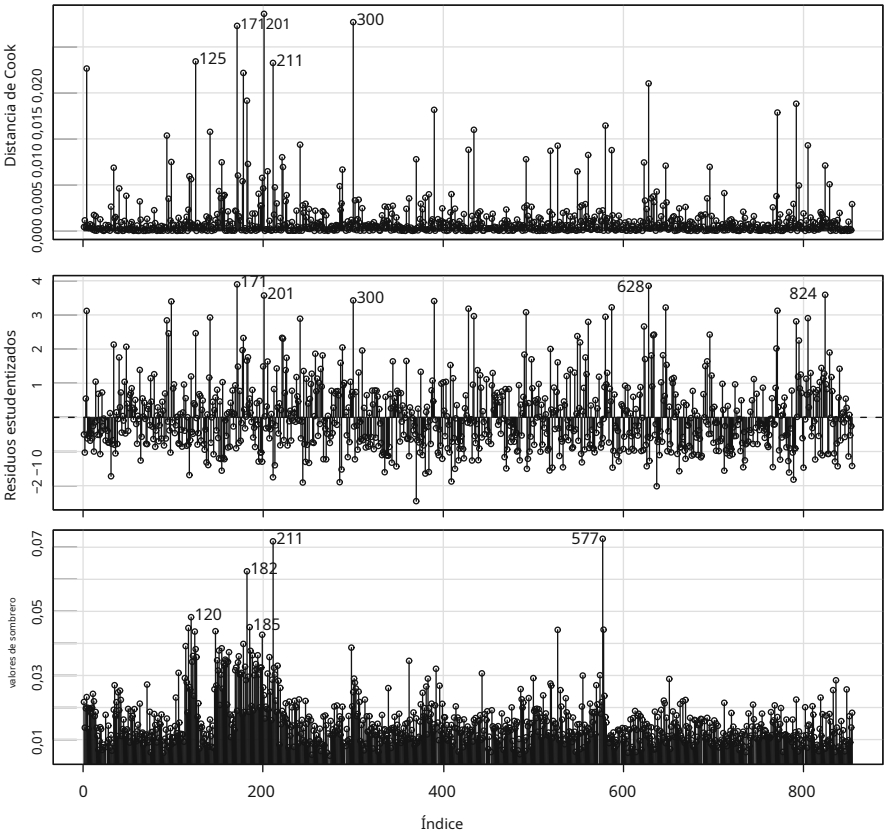


Figura 6.5 Distancias de Cook, residuos estudentizados y valores de sombrero del modelo de evolución de las horas de enseñanza

```
influenceIndexPlot (mod.horas,
        vars = c ("Cocinero", "Studentizado", "sombrero"), id.n = 5)
```

Los valores de estas tres cantidades se informan en la Fig. 6.5, que muestra las distancias de Cook, los residuos estudentizados y los valores de sombrero, respectivamente. En cualquiera de estos gráficos, un valor extremo en relación con los demás indica que una observación puede ser particularmente problemática. En esta figura, ninguna de las observaciones se destaca particularmente, y ninguno de los valores de la distancia de Cook está ni remotamente cerca de 1 (que es un umbral común de regla empírica para esta cantidad). Por lo tanto, ninguna de las observaciones parece ser particularmente problemática para este modelo. En un caso en el que algunas observaciones parecen ejercer influencia sobre los resultados, el investigador debe decidir si es razonable mantener las observaciones en el análisis o si alguna de ellas debe eliminarse. La eliminación de datos de un modelo lineal se puede lograr fácilmente con elsubconjunto opción de lm.

Ahora hemos considerado cómo ajustar modelos lineales en R y cómo realizar varios diagnósticos para determinar si OLS nos presenta el AZUL. Si bien este es un modelo común en Ciencias Políticas, los investigadores con frecuencia necesitan modelar variables dependientes limitadas en el estudio de la política. Para abordar las variables dependientes de esta naturaleza, pasamos al capítulo siguiente a GLM. Estos modelos se basan en el marco del modelo lineal, pero permiten variables de resultado de naturaleza limitada o categórica.

6.3 Problemas de práctica

Este conjunto de problemas de práctica se basará en la de Owsiak (2013) trabajo sobre democratización, en el que muestra que los estados que liquidan todas sus fronteras internacionales tienden a democratizarse. Cargue la biblioteca externa y luego descargue un subconjunto de los datos de Owsiak, guardados en el archivo con formato StataowsiakJOP2013.dta.
Este archivo se puede descargar desde el Dataverse vinculado en la página vii o el contenido del capítulo vinculado en la página 79. Estos son *panel de datos* que incluyen observaciones para 200 países desde 1918 hasta 2007, con un total de 10,434 países-año que forman los datos. Los países en estos datos cambian con el tiempo (tal como cambiaron en su libro de historia), lo que lo convierte en lo que llamamos un panel desequilibrado. Por lo tanto, nuestro modelo posterior incluye valores rezagados de varias variables o valores del año anterior. Ver cap.8 para obtener más información sobre los datos anidados y el Cap. 9 para obtener más información sobre los datos temporales. Para este ejercicio, nuestras herramientas OLS estándar funcionarán bien.

1. Empiece por utilizar el na.omit comando, descrito en la página 81, para eliminar las observaciones faltantes de estos datos. Luego calcule las estadísticas descriptivas para las variables en este conjunto de datos.
2. Para replicar el Modelo 2 de Owsiak (2013), estime una regresión lineal con MCO utilizando la siguiente especificación (con los nombres de las variables entre paréntesis): La variable dependiente es la puntuación Polity (**polity2),** y los predictores son un indicador de que se han establecido todas las fronteras (**allsettle),** PIB rezagado (**laggdpam),** cambio rezagado en el PIB (**laggdpchg),** apertura comercial rezagada**lagtradeopen),** personal militar rezagado**lagmilper),** población urbana rezagada**lagupop),** movimiento no democrático anterior rezagado (**lagsumdown),** y puntuación de Polity rezagada (**lagpolity).**
3. Grafique los residuales contra los valores ajustados.
4. ¿Existe heterocedasticidad en los residuos? Basado en diagramas de dispersión y una prueba de Breusch-Pagan, ¿qué concluye?

   una. Estime los errores estándar de Huber-White para este modelo con elemparedado biblioteca y coeftest mando.

   B. Para obtener un crédito de bonificación, puede reproducir el de Owsiak (2013) resultados exactamente calculando *errores estándar agrupados,* agrupación en el país (nombre de la variable: **ccode).** Puede hacer esto en tres pasos: Primero, instale el multiwayvcov Biblioteca. En segundo lugar, defina una matriz de varianza-covarianza de error utilizando la cluster.vcov mando. En tercer lugar, utilice esa matriz de varianza-covarianza de error como argumento en lacoeftest comando desde el lmtest Biblioteca.

5. Determine si la multicolinealidad es una preocupación calculando los VIF para los predictores en este modelo.
6. ¿Están distribuidos normalmente los residuos de este modelo? Utilice cualquiera de los métodos discutidos para sacar una conclusión.
7. Para el crédito de bonificación, puede evaluar si existe autocorrelación en los residuales, como se discutirá más adelante en el Cap. 9. Para ello, primero instale elplm Biblioteca. Segundo, reajuste su modelo usando elplm mando. (Asegúrese de especificar modelo = "agrupación" como una opción en el comando para estimar con MCO). En tercer lugar, utilice el pbgtest realizar una prueba de panel de Breusch-Godfrey para evaluar si existe una correlación serial en los residuos. ¿Qué conclusión sacas?