alimentarse de su corteza. En cada parcela, los investigadores determinaron el número de tocones resultantes de los árboles derribados por los castores y el número de larvas del coleóptero. He aquí los datos:[17]

Tocones	2	2	1	3	3	4	3	1	2	5	1	3
Larvas	10	30	12	24	36	40	43	11	27	56	18	40
Tocones	2	1	2	2	1	1	4	1	2	1	4	
Larvas	25	8	21	14	16	6	54	9	13	14	50	

(a) Haz un diagrama de dispersión que muestre cómo el número de tocones debidos a los castores influye sobre el de larvas. ¿Qué muestra tu diagrama? (Los ecólogos creen que los brotes que surgen de los tocones resultan más apetecibles para las larvas ya que son más tiernos que los de los árboles mayores.)

(b) Halla la recta de regresión mínimo-cuadrática y dibújala en tu diagrama.

(c) ¿Qué porcentaje de la variación observada en el número de larvas se puede explicar por la dependencia lineal con el número de tocones?

2.4.3 Residuos

Una recta de regresión es un modelo matemático que describe una relación lineal entre una variable explicativa y una variable respuesta. Las desviaciones de la relación lineal también son importantes. Cuando se dibuja una recta de regresión, se ven las desviaciones observando la dispersión de los puntos respecto a dicha recta. Las distancias verticales de los puntos a la recta de regresión mínimo-cuadrática son lo más pequeñas posible, en el sentido de que tienen la menor suma de cuadrados posible. A estas distancias les damos un nombre: *residuos*.

RESIDUOS

Un **residuo** es la diferencia entre el valor observado de la variable respuesta y el valor predicho por la recta de regresión. Es decir,

$$\begin{aligned} \text{residuo} &= y \text{ observada} - y \text{ predicha} \\ &= y - \hat{y} \end{aligned}$$

[17] G. D. Martinsen, E. M. Driebe y T. G. Whitham, "Indirect interactions mediated by changing plant chemistry: beaver browsing benefits beetles", *Ecology*, 79, 1998, págs. 192-200.

Tabla 2.7. Edad de la primera palabra y puntuación en la prueba Gesell.

Niño	Edad	Puntuación	Niño	Edad	Puntuación
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

EJEMPLO 2.12. Predicción de la inteligencia

¿Predice su inteligencia posterior la edad a la que un niño empieza a hablar? Un estudio del desarrollo de 21 niños, registró la edad, en meses, a la que cada niño pronunciaba la primera palabra y su puntuación en la prueba Gesell (*Gesell Adaptative Score*), una prueba de aptitud llevada a cabo mucho más tarde. Los datos aparecen en la tabla 2.7.[18]

La figura 2.14 es un diagrama de dispersión en el que se toma la edad en que se pronunció la primera palabra como variable explicativa x y la puntuación en la prueba Gesell como variable respuesta y. Los niños 3 y 13, y los niños 16 y 21 tienen valores idénticos para ambas variables, por lo que se utiliza un símbolo distinto para mostrar que estos puntos representan a dos individuos diferentes. El diagrama muestra una asociación negativa, es decir, los niños que empiezan a hablar más tarde tienden a tener puntuaciones más bajas en la prueba que los niños que hablan antes. El aspecto general de la relación es moderadamente lineal. La correlación describe la dirección y la fuerza de la relación lineal, $r = -0{,}640$.

La recta que se ha trazado en el diagrama es la recta de regresión mínimo-cuadrática de la puntuación Gesell con relación a la edad de la primera palabra. Su ecuación es

$$\hat{y} = 109{,}8738 - 1{,}1270x$$

[18]M. R. Mickey, O. J. Dunn y V. Clark, "Note on the use of stepwise regression in detecting outliers", *Computers and Biomedical Research*, 1, 1967, págs. 105-111. Estos datos han sido utilizados por varios autores; yo los he hallado en N. R. Draper y J. A. John, "Influential observations and outliers in regression", *Technometrics*, 23, 1981, págs. 21-26.

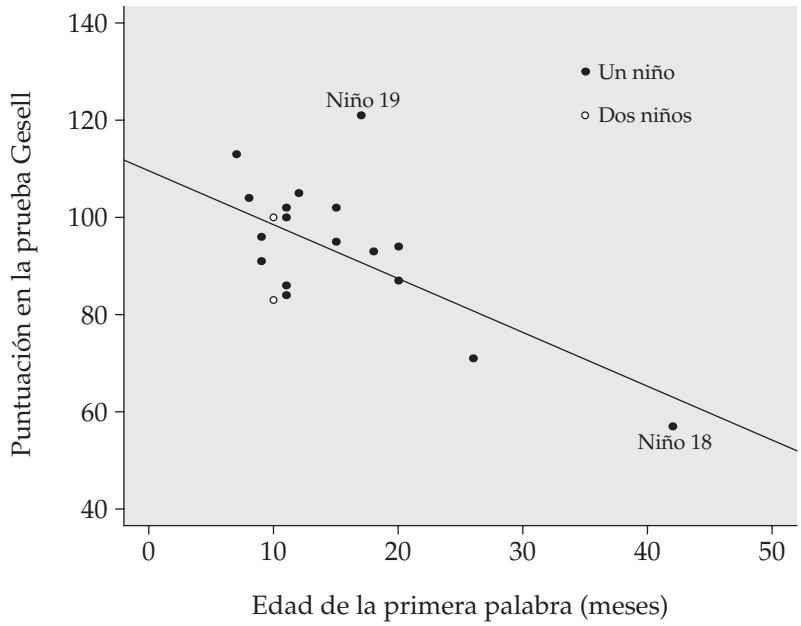

Figura 2.14. Diagrama de dispersión de las puntuaciones de la prueba Gesell con relación a la edad de la primera palabra de 21 niños. La recta es la recta de regresión mínimo-cuadrática para predecir la puntuación Gesell a partir de la primera palabra.

Para el primer niño, que empezó a hablar a los 15 meses, predecimos la puntuación

$$\hat{y} = 109{,}8738 - (1{,}1270)(15) = 92{,}97$$

La puntuación real de este niño fue de 95. El residuo es

$$\begin{aligned} \text{residuo} &= y \text{ observada} - y \text{ predicha} \\ &= 95 - 92{,}97 = 2{,}03 \end{aligned}$$

El residuo es positivo porque el punto se halla por encima de la recta. ■

Existe un valor residual para cada punto. Hallar los valores residuales con una calculadora es bastante laborioso, ya que primero tienes que hallar la respuesta predicha para cada x. Los programas estadísticos te dan todos los residuos

a la vez. He aquí los 21 residuos de los datos de la prueba Gesell obtenidos con un programa estadístico:

```
2.0310   -9.5721   -15.6040   -8.7309    9.0310    -0.3341     3.4120
2.5230    3.1421     6.6659   11.0151   -3.7309   -15.6040   -13.4770
4.5230    1.3960     8.6500   -5.5403   30.2850   -11.4770     1.3960
```

Debido a que los residuos muestran a qué distancia se hallan los datos de nuestra recta de regresión, el examen de los residuos nos ayuda a valorar en que medida la recta describe la distribución de los datos. A pesar de que los residuos se pueden calcular a partir de cualquier modelo que se haya ajustado a los datos, los de la recta de regresión mínimo-cuadrática tienen una propiedad especial: **la media de los residuos es siempre cero**.

Compara el diagrama de dispersión de la figura 2.14 con el *diagrama de residuos* correspondiente a los mismos datos de la figura 2.15. En dicha figura, la recta horizontal que pasa por cero nos ayuda a orientarnos. Esta recta corresponde a la recta de regresión de la figura 2.14.

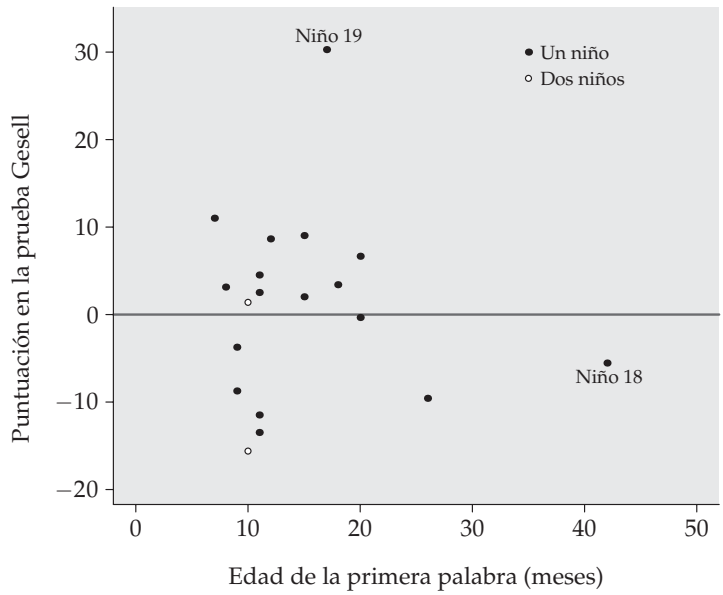

Figura 2.15. Diagrama de residuos para la regresión de las puntuaciones en la prueba Gesell en relación con la edad de la primera palabra. El niño 19 es una observación atípica. El niño 18 es una observación influyente que no tiene un residuo grande.

DIAGRAMA DE RESIDUOS

Un **diagrama de residuos** es un diagrama de dispersión de los residuos de la regresión con relación a la variable explicativa. Los diagramas de residuos nos ayudan a valorar el ajuste de la recta de regresión.

Si la recta de regresión se ajusta bien a la relación entre x e y, los residuos no tienen que tener ninguna distribución especial. En dicho caso, la distribución de residuos se parecerá a la distribución que de forma simplificada se muestra en la figura 2.16(a). Este diagrama muestra que la distribución de los residuos es uniforme a lo largo de la recta, no se detectan observaciones atípicas. Cuando examines los residuos en el diagrama de dispersión o en el diagrama de residuos, has de fijarte en algunos detalles:

- **Una forma curva** de la distribución de los residuos indica que la relación no es lineal. La figura 2.16(b) es un ejemplo ilustrativo. La recta no es una buena descripción para estos datos.
- **Un crecimiento o decrecimiento de la dispersión de los residuos** a medida que aumentan las x. La figura 2.16(c) es un ejemplo. En él, la predicción de y será menos precisa para valores de x mayores.
- **Los puntos individuales con residuos grandes**, como el del niño 19 de las figuras 2.14 y 2.15. Estos puntos son observaciones atípicas, ya que no encajan en el aspecto lineal de la nube de puntos.
- **Los puntos individuales que son extremos en el eje de las abscisas**, como el niño 18 de las figuras 2.14 y 2.15. Estos puntos pueden no tener grandes residuos, pero pueden ser muy importantes. Más adelante estudiaremos este tipo de puntos.

2.4.4 Observaciones influyentes

Los niños 18 y 19 del ejemplo Gesell son poco frecuentes, pero por motivos distintos. El niño 19 está lejos de la recta de regresión. Este niño empezó a hablar mucho más tarde que los demás. Su valor Gesell es tan alto que tendríamos que comprobar que no se trata de un error de transcripción de los datos. De todas

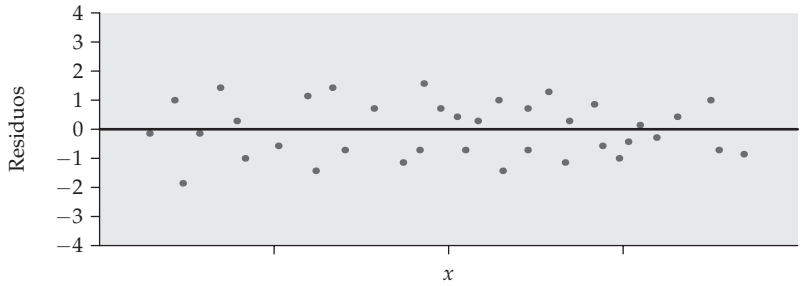

Figura 2.16(a). Distribuciones idealizadas de diagramas de residuos de la recta de regresión mínimo-cuadrática. El gráfico (a) indica un buen ajuste de la recta de regresión.

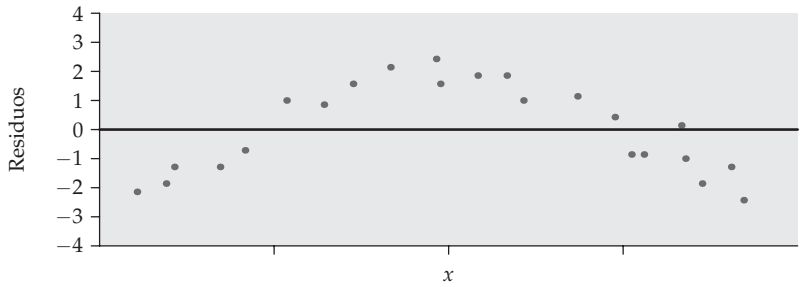

Figura 2.16(b). El gráfico (b) muestra una forma curva, por tanto, la recta se ajusta mal.

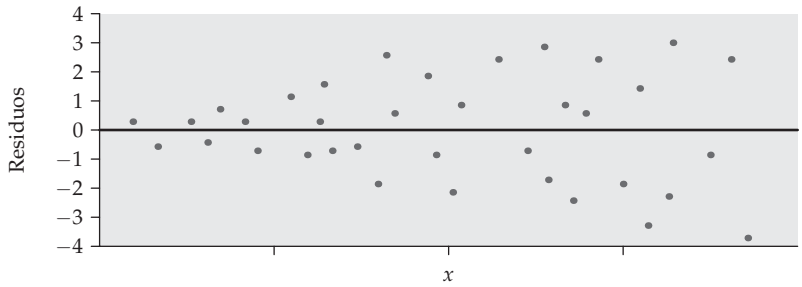

Figura 2.16(c). La variable respuesta y del gráfico (c) presenta más dispersión para los valores mayores de la variable explicativa x. Por tanto, la predicción será menos precisa cuanto mayor sea x.

formas, el valor Gesell es correcto. El punto correspondiente al niño 18 se halla cerca de la recta, sin embargo se encuentra alejado en la dirección de las abscisas. El niño 18 fue el que empezó a hablar más tarde. *Debido a su posición extrema en el eje de las abscisas tiene una gran influencia sobre la posición de la recta de regresión*. La figura 2.17 añade una segunda recta de regresión, calculada tras excluir al niño 18. Puedes ver que sin esta observación la posición de la recta se ha modificado. A estos puntos los llamamos *influyentes*.

OBSERVACIONES ATÍPICAS Y OBSERVACIONES INFLUYENTES EN REGRESIÓN

Una **observación atípica** es aquélla que queda separada de las restantes observaciones.

Una observación es **influyente** con relación a un cálculo estadístico si al eliminarla cambia el resultado del cálculo. En regresión mínimo-cuadrática, las observaciones atípicas en la dirección del eje de las abscisas son, en general, observaciones influyentes.

Los niños 18 y 19 son ambos observaciones atípicas de la figura 2.17. El niño 18 es una observación atípica en la dirección del eje de las abscisas y es también una observación influyente para la recta de regresión mínimo-cuadrática. El niño 19, en cambio, es una observación atípica en la dirección del eje de las ordenadas. Tiene menos influencia en la posición de la recta de regresión porque hay muchos puntos con valores de x similares que retienen la recta por debajo de la observación atípica. Los puntos influyentes suelen tener residuos pequeños, ya que tiran de la recta hacia su posición. Si sólo te fijas en los residuos, pasarás por alto los puntos influyentes. Las observaciones influyentes pueden modificar en gran manera la interpretación de unos datos.

EJEMPLO 2.13. Una observación influyente

La fuerte influencia del niño 18 hace que la recta de regresión de la puntuación Gesell con relación a la primera palabra sea engañosa. Los datos originales tienen $r^2 = 0{,}41$. Es decir, un 41% de la variación total observada en la prueba Gesell

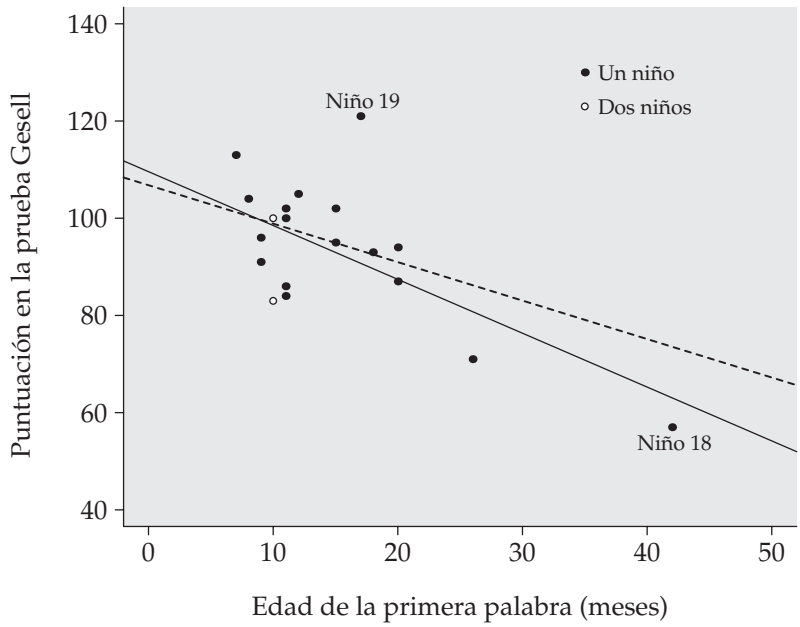

Figura 2.17. Dos rectas de regresión mínimo-cuadráticas de las puntuaciones Gesell en relación con la edad de la primera palabra. La recta de trazo continuo se ha calculado a partir de todos los datos. La de trazo discontinuo se ha calculado excluyendo al niño 18. El niño 18 es una observación influyente, ya que cuando se elimina este punto, la posición de la recta cambia.

se puede explicar a partir de la edad a la que los niños empiezan a hablar. Esta relación es suficientemente fuerte para que sea de interés para los padres. Pero si dejamos fuera al niño 18, r^2 cae al 11%. La fuerza aparente de la asociación se debía en gran medida a una sola observación influyente.

¿Qué debe hacer un investigador? Debe decidir si el desarrollo del niño 18 fue tan lento que no se debería permitir que influyera en el resultado del análisis. Si excluye al niño 18, desaparece en gran parte la evidencia de la relación entre la edad a la que un niño empieza a hablar y su posterior puntuación en la prueba. Si mantiene esta observación, necesitará datos adicionales de niños que hayan empezado a hablar tardíamente, de manera que el análisis no dependa tanto de una sola observación. ■

APLICA TUS CONOCIMIENTOS

2.36. Consumo de gasolina y velocidad. El ejercicio 2.6 proporciona datos sobre el consumo de gasolina y de un automóvil a distintas velocidades x. El consumo de carburante se ha medido en litros de gasolina por 100 kilómetros y la velocidad en kilómetros por hora. Con la ayuda de un programa estadístico hemos obtenido la recta de regresión mínimo-cuadrática y también los residuos. La recta de regresión es

$$\hat{y} = 11{,}058 - 0{,}01466x$$

Los residuos, en el mismo orden que las observaciones, son

10,09	2,24	−0,62	−2,47	−3,33	−4,28	−3,73	−2,94
−2,17	−1,32	−0,42	0,57	1,64	2,76	3,97	

(a) Dibuja un diagrama de dispersión con las observaciones y traza la recta de regresión en tu diagrama.

(b) ¿Utilizarías la recta de regresión para predecir y a partir de x? Justifica tu respuesta.

(c) Comprueba que la suma de los residuos es 0 (o muy cercana a 0, teniendo en cuenta los errores de redondeo).

(d) Dibuja un diagrama de residuos con relación a los valores de x. Traza una recta horizontal a la altura del valor 0 del eje de las ordenadas. Comprueba que la distribución de los residuos a lo largo de esta recta es similar a la distribución de los puntos a lo largo de la recta de regresión del diagrama de dispersión en (a).

2.37. ¿Cuántas calorías? La tabla 2.5 proporciona datos sobre el contenido real en calorías de diez alimentos y la media de los contenidos estimados por un numeroso grupo de personas. El ejercicio 2.23 explora la influencia de dos observaciones atípicas sobre la correlación.

(a) Dibuja un diagrama de dispersión adecuado para predecir la estimación de las calorías a partir de los valores reales. Señala los puntos correspondientes a los espaguetis y a los pasteles en tu diagrama. Estos dos puntos quedan fuera de la relación lineal de los ocho puntos restantes.

(b) Utiliza tu calculadora para hallar la recta de regresión de las calorías estimadas con relación a las calorías reales. Hazlo dos veces, primero, con todos los puntos y luego, dejando fuera los espaguetis y los pasteles.

(c) Dibuja las dos rectas de regresión en tu diagrama (una de trazo continuo y la otra con trazo discontinuo). Los espaguetis y los pasteles, tomados conjuntamente, ¿son observaciones influyentes? Justifica tu respuesta.

2.38. ¿Influyentes o no? Hemos visto que el niño 18 de los datos Gesell de la tabla 2.7 es una observación influyente. Ahora vamos a examinar el efecto del niño 19, que también es una observación atípica en la figura 2.14.

(a) Halla la recta de regresión mínimo-cuadrática de la puntuación en la prueba Gesell respecto a la edad a la cual un niño empieza a hablar, dejando fuera al niño 19. El ejemplo 2.12 da la recta de regresión con todos los niños. Dibuja ambas rectas en el mismo gráfico (no es necesario que lo hagas sobre un diagrama de dispersión; tan sólo dibuja las rectas). ¿Calificarías al niño 19 como muy influyente? ¿Por qué?

(b) La exclusión del niño 19, ¿qué efecto tiene sobre el valor r^2 de esta regresión? Explica por qué cambia r^2 al excluir al niño 19.

RESUMEN DE LA SECCIÓN 2.4

Una **recta de regresión** es una recta que describe cómo cambia una variable respuesta y al cambiar una variable explicativa x.

El método más común para ajustar una recta en un diagrama de dispersión es el método de mínimos cuadrados. La **recta de regresión mínimo-cuadrática** es la recta de la ecuación $\hat{y} = a + bx$ que minimiza la suma de cuadrados de las distancias verticales de los valores observados de y a la recta de regresión.

Puedes utilizar una recta de regresión para **predecir** el valor de y a partir de cualquier valor de x, sustituyendo esta x en la ecuación de la recta.

La **pendiente** b de una recta de regresión $\hat{y} = a + bx$ indica el cambio de la variable respuesta predicha $\hat{y}$ a lo largo de la recta de regresión, al cambiar la variable explicativa x. En concreto, b es el cambio de $\hat{y}$ al aumentar x en una unidad.

La **ordenada en el origen** a de una recta de regresión $\hat{y} = a + bx$ es la respuesta predicha $\hat{y}$ cuando la variable explicativa es $x = 0$. Esta predicción no tiene significado estadístico a no ser que x pueda tomar valores cercanos a 0.

La recta de regresión mínimo-cuadrática de y con relación a x es la recta de pendiente $r\frac{s_y}{s_x}$ y ordenada en el origen $a = \bar{y} - b\bar{x}$. Esta recta siempre pasa por el punto $(\bar{x}, \bar{y})$.

La **correlación y la regresión** están íntimamente relacionadas. Cuando las variables x e y se miden en unidades estandarizadas, la correlación r es la pendiente de la recta de regresión mínimo-cuadrática. El cuadrado de la correlación r^2 es la proporción de la variación de la variable respuesta explicada por la regresión mínimo-cuadrática.

Puedes examinar el ajuste de una recta de regresión estudiando los **residuos**, que son las diferencias entre los valores observados y los valores predichos de y.

Vigila los puntos atípicos con residuos anormalmente grandes, y también las distribuciones no lineales y desiguales de los residuos.

Fíjate también en las **observaciones influyentes**, que son los puntos aislados que cambian de forma sustancial la posición de la recta de regresión. Las observaciones influyentes suelen ser observaciones atípicas en la dirección de las abscisas y no tienen por qué tener residuos grandes.

EJERCICIOS DE LA SECCIÓN 2.4

2.39. Repaso sobre relación lineal. Antonio guarda sus ahorros en un colchón. Empezó con 500 € que le dio su madre y cada año fue añadiendo 100 €. Sus ahorros totales y después de x años vienen dados por la ecuación

$$y = 500 + 100x$$

(a) Representa gráficamente esta ecuación. (Escoge dos valores de x, tales como 0 y 10. Calcula los valores correspondientes de y a partir de la ecuación. Dibuja estos dos puntos en el gráfico y dibuja la recta uniéndolos.)

(b) Después de 20 años, ¿cuánto tendrá Antonio en su colchón?

(c) Si Antonio hubiera añadido cada año 200 € a sus 500 € iniciales, en vez de 100, ¿cuál sería la ecuación que describiría sus ahorros después de x años?

2.40. Repaso sobre relación lineal. En el periodo posterior a su nacimiento, una rata blanca macho gana exactamente 40 gramos (g) por semana. (Esta rata es extrañamente regular en su crecimiento, pero un crecimiento de 40 g por semana es un valor razonable.)

(a) Si la rata pesaba 100 gramos al nacer, da una ecuación para predecir su peso después de x semanas. ¿Cuál es la pendiente de esta recta?

(b) Dibuja un gráfico de esta recta para valores de x entre el nacimiento y las 10 semanas de edad.

(c) ¿Utilizarías esta recta para predecir el peso de la rata a los 2 años? Haz la predicción y medita sobre si el resultado es razonable.

2.41. Coeficiente de inteligencia y notas escolares. La figura 2.5 muestra las notas escolares medias y los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato. La media y la desviación típica de los coeficientes de inteligencia son

$$\bar{x} = 108{,}9 \qquad\qquad s_x = 13{,}17$$