2. ANÁLISIS DE RELACIONES

JOHN W. TUKEY

John W. Tukey (1915-) empezó como químico, siguió como matemático y finalmente se convirtió en estadístico debido a lo que él mismo denominó "la experiencia de los problemas reales y de los datos reales" que adquirió durante la II Guerra Mundial. En 1937, John W. Tukey fue a la Princeton University a estudiar química pero se doctoró en matemáticas en 1939. Durante la guerra trabajó en temas de precisión de tiro. Después de la guerra, simultaneó su labor en la Princeton University con su trabajo en los Laboratorios Bell, quizá el grupo de investigación industrial más importante del mundo.

Tukey dedicó la mayor parte de su atención al estudio estadístico de problemas especialmente difíciles de resolver, como son la seguridad de las anestésicos, el comportamiento sexual de los seres humanos, la comprobación del cumplimiento de la prohibición de las pruebas nucleares, y la determinación de la calidad del aire y la contaminación ambiental.

Basándose en "la experiencia de los problemas reales y de los datos reales", John Tukey desarrolló el análisis exploratorio de datos. Inventó algunas de las herramientas estadísticas que hemos visto en el capítulo 1 como, por ejemplo, los diagramas de tallos y los diagramas de caja. Tukey cambió el enfoque del análisis de datos, defendiendo un análisis de datos mucho más flexible, más exploratorio, cuyo objetivo no consiste simplemente en dar respuesta a preguntas concretas. El primer propósito es contestar a la pregunta: "¿Qué dicen los datos?". Este capítulo, igual que el capítulo 1, sigue el camino que marcó Tukey, y para ello presentamos más ideas y herramientas para examinar datos.

2.1 Introducción

Un estudio médico halló que las mujeres bajas son más propensas a sufrir ataques al corazón que las mujeres de altura media. Además, las mujeres altas sufren menos ataques al corazón que las de altura media. Por otro lado, un grupo asegurador informa que con coches grandes se producen menos muertes por cada 10.000 vehículos que con coches pequeños. Estos y muchos otros estudios estadísticos buscan relaciones entre dos variables. Para entender este tipo de relaciones, a menudo también tenemos que examinar otras variables. Para llegar a la conclusión de que las mujeres bajas sufren más ataques al corazón, los investigadores tuvieron que eliminar el efecto de otras variables como el peso y los hábitos deportivos. En este capítulo examinaremos la relación entre variables. También veremos que la relación entre dos variables puede verse afectada de forma importante por variables latentes de entorno.

Como la variación está siempre presente, las relaciones estadísticas son tendencias generales, no reglas blindadas. Nuestras relaciones admiten excepciones individuales. A pesar de que como media los fumadores mueren antes que los no fumadores, algunos fumadores que consumen más de tres paquetes diarios llegan a los noventa años. Para estudiar la relación entre dos variables, las medimos en los mismos individuos. A menudo, creemos que una de las variables puede explicar o influye sobre la otra.

VARIABLE RESPUESTA Y VARIABLE EXPLICATIVA

Una **variable respuesta** mide el resultado de un estudio. Una **variable explicativa** influye o explica cambios en la variable respuesta.

Variable independiente

Variable dependiente

A menudo, encontrarás que a las variables explicativas se les llama **variables independientes** y a las variables respuesta, **variables dependientes**. La idea es que el valor de la variable respuesta depende del de la variable explicativa. Como en estadística las palabras "independiente" y "dependiente" tienen otros significados que no están relacionados con lo que acabamos de ver, no utilizaremos esta terminología.

La manera más fácil de distinguir entre variables explicativas y variables respuesta es dar valores a una de ellas y ver lo que ocurre en la otra.

EJEMPLO 2.1. Los efectos del alcohol

El alcohol produce muchos efectos sobre el cuerpo humano. Uno de ellos es la bajada de la temperatura corporal. Para estudiar este efecto, unos investigadores suministraron distintas dosis de alcohol a unos ratones y al cabo de 15 minutos midieron la variación de temperatura de su cuerpo. La cantidad de alcohol es la variable explicativa y el cambio de temperatura corporal es la variable respuesta. ■

Cuando no asignamos valores a ninguna variable, sino que simplemente observamos los valores que adquieren, éstas pueden ser o no variables explicativas y variables respuesta. El que lo sean depende de cómo pensemos utilizar los datos.

EJEMPLO 2.2. Calificaciones en la prueba SAT

Alberto quiere saber qué relación existe entre la media de las calificaciones de Matemáticas y la media de las calificaciones de Lengua obtenidas por estudiantes de los 51 Estados de EE UU (incluyendo el Distrito de Columbia) en la prueba SAT. Inicialmente, Alberto no cree que una variable dependa de los valores que tome la otra. Tiene dos variables relacionadas y ninguna de ellas es una variable explicativa.

Julia, con los mismos datos, se plantea la siguiente pregunta: ¿puedo predecir la calificación de Matemáticas de un Estado si conozco su calificación de Lengua? En este caso, Julia trata la calificación de Lengua como variable explicativa y la de Matemáticas como variable respuesta. ■

En el ejemplo 2.1, el alcohol realmente causa un cambio en la temperatura corporal. No existe ninguna relación causa-efecto entre las calificaciones de Matemáticas y las de Lengua del ejemplo 2.2. De todas formas, como existe una estrecha relación entre las calificaciones de Matemáticas y de Lengua, podemos utilizar la de Lengua para predecir la de Matemáticas. En la sección 2.4 aprenderemos a hacer dicha predicción. Ésta requiere que identifiquemos una variable explicativa y una variable respuesta. Otras técnicas estadísticas ignoran esta distinción. Recuerda que llamar a una variable explicativa y a otra variable respuesta no significa necesariamente que los cambios en una de ellas causen cambios en la otra.

Muchos estudios estadísticos examinan datos de más de una variable. Afortunadamente, los estudios estadísticos de datos de varias variables se basan en las herramientas que hemos utilizado para examinar una sola variable. Los principios en los que se basa nuestro trabajo también son los mismos:

- Empieza con un gráfico; luego, añade resúmenes numéricos.
- Identifica el aspecto general y las desviaciones.
- Cuando el aspecto general sea bastante regular, utiliza un modelo matemático para describirlo.

APLICA TUS CONOCIMIENTOS

2.1. En cada una de las situaciones siguientes, ¿qué es más razonable, simplemente explorar la relación entre dos variables o contemplar una de las variables como variable explicativa y la otra como variable respuesta?

(a) La cantidad de tiempo que un alumno pasa estudiando para un examen de Estadística y la calificación obtenida en el examen.

(b) El peso y la altura de una persona.

(c) La lluvia caída durante un año y el rendimiento de un cultivo.

(d) Las calificaciones de Estadística y de Francés de los estudiantes.

(e) El tipo de trabajo de un padre y el de su hijo.

2.2. ¿Es posible predecir la altura que tiene un niño de 16 años a partir de la altura que tenía a los 6? Una manera de descubrirlo consistiría en medir la altura de un grupo suficientemente numeroso de niños de 6 años, esperar hasta que cumplieran los 16 años y entonces volver a medirlos. En este caso, ¿cuál es la variable explicativa y cuál es la variable respuesta? ¿Estas variables son categóricas o cuantitativas?

2.3. Tratamiento del cáncer de mama. El tratamiento más común para combatir el cáncer de mama consistía en la extirpación completa del pecho. Hoy en día, se suele extirpar únicamente el tumor y los ganglios linfáticos circundantes, aplicando después radioterapia en la zona afectada. El cambio de tratamiento se produjo tras un amplio experimento médico que comparó ambas técnicas: se seleccionó al azar dos grupos de enfermas; cada uno siguió un tratamiento distinto y se realizó un minucioso seguimiento de las pacientes para comprobar el periodo de supervivencia. ¿Cuál es la variable explicativa y cuál es la variable respuesta? ¿Son variables categóricas o cuantitativas?

2.2 Diagramas de dispersión

La manera más común de mostrar gráficamente la relación entre dos variables cuantitativas es un *diagrama de dispersión*. He aquí un ejemplo de diagrama de dispersión.

EJEMPLO 2.3. Notas en la prueba SAT

La tabla 2.1 proporciona datos sobre la educación en los diversos Estados de EE UU. La primera columna identifica los Estados, la segunda indica a qué región censal pertenece cada uno: *East North Central* (ENC), *East South Central* (ESC), *Middle Atlantic* (MA), *Mountain* (MTN), *New England* (NE), *Pacific* (PAC), *South Atlantic* (SA), *West North Central* (WNC) y *West South Central* (WSC). La tercera columna contiene la población de cada Estado en miles de habitantes. Las cinco variables restantes son la media de Lengua y de Matemáticas en la prueba SAT, el porcentaje de alumnos que se presentan a la prueba, el porcentaje de residentes que no se graduaron en secundaria y la media de los salarios de los profesores expresado en miles de dólares.

En EE UU se usan las medias de las calificaciones obtenidas en las pruebas SAT para evaluar los sistemas educativos, tanto estatales como locales. Este sistema de evaluación no es un buen procedimiento, ya que el porcentaje de alumnos de enseñanza media que se presenta a estas pruebas varía mucho según el Estado. Vamos a examinar la relación entre el porcentaje de alumnos que se presentan a estas pruebas en cada Estado y la media de las calificaciones de matemáticas.

Creemos que "el porcentaje de alumnos que se presentan" nos ayudará a entender "la media de los resultados". Por tanto, la variable "el porcentaje de alumnos que se presentan" es la variable explicativa y la variable "la media de las calificaciones de Matemáticas" es la variable respuesta. Queremos ver cómo varía la media de las calificaciones cuando cambia el porcentaje de alumnos que se presentan al examen. Es por este motivo que situaremos el porcentaje de alumnos que se presentan (la variable explicativa) en el eje de las abscisas. La figura 2.1 es el diagrama de dispersión en el que cada punto representa a un Estado. Por ejemplo, en Alabama el 8% de los alumnos se presentó al examen y la media de las calificaciones de Matemáticas fue de 558. Busca el 8 en el eje de las x (eje de las abscisas) y el 558 en el eje de las y (eje de las ordenadas). El Estado de Alabama aparece como el punto (8, 558), encima del 8 y a la derecha de 558. La figura 2.1 muestra cómo localizar el punto correspondiente a Alabama en el diagrama. ■

Tabla 2.1. Datos sobre la educación en EE UU.

Estado*	Región**	Población (1.000)	SAT Lengua	SAT Matemáticas	Porcentaje de alumnos presentados	Porcentaje sin estudios de secundaria	Salario de profesores ($1.000)
AL	ESC	4.273	565	558	8	33,1	31,3
AK	PAC	607	521	513	47	13,4	49,6
AZ	MTN	4.428	525	521	28	21,3	32,5
AR	WSC	2.510	566	550	6	33,7	29,3
CA	PAC	31.878	495	511	45	23,8	43,1
CO	MTN	3.823	536	538	30	15,6	35,4
CT	NE	3.274	507	504	79	20,8	50,3
DE	SA	725	508	495	66	22,5	40,5
DC	SA	543	489	473	50	26,9	43,7
FL	SA	14.400	498	496	48	25,6	33,3
GA	SA	7.353	484	477	63	29,1	34,1
HI	PAC	1.184	485	510	54	19,9	35,8
ID	MTN	1.189	543	536	15	20,3	30,9
IL	ENC	11.847	564	575	14	23,8	40,9
IN	ENC	5.841	494	494	57	24,4	37,7
IA	WNC	2.852	590	600	5	19,9	32,4
KS	WNC	2.572	579	571	9	18,7	35,1
KY	ESC	3.884	549	544	12	35,4	33,1
LA	WSC	4.351	559	550	9	31,7	26,8
ME	NE	1.243	504	498	68	21,2	32,9
MD	SA	5.072	507	504	64	21,6	41,2
MA	NE	6.092	507	504	80	20,0	42,9
MI	ENC	9.594	557	565	11	23,2	44,8
MN	WNC	4.658	582	593	9	17,6	36,9
MS	ESC	2.716	569	557	4	35,7	27,7
MO	WNC	5.359	570	569	9	26,1	33,3
MT	MTN	879	546	547	21	19,0	29,4
NE	WNC	1.652	567	568	9	18,2	31,5
NV	MTN	1.603	508	507	31	21,2	36,2
NH	NE	1.162	520	514	70	17,8	35,8
NJ	MA	7.988	498	505	69	23,3	47,9
NM	MTN	1.713	554	548	12	24,9	29,6
NY	MA	18.185	497	499	73	25,2	48,1
NC	SA	7.323	490	486	59	30,0	30,4
ND	WNC	644	596	599	5	23,3	27,0
OH	ENC	11.173	536	535	24	24,3	37,8
OK	WSC	3.301	566	557	8	25,4	28,4
OR	PAC	3.204	523	521	50	18,5	39,6
PA	MA	12.056	498	492	71	25,3	46,1
RI	NE	990	501	491	69	28,0	42,2
SC	SA	3.699	480	474	57	31,7	31,6
SD	WNC	732	574	566	5	22,9	26,3
TN	ESC	5.320	563	552	14	32,9	33,1
TX	WSC	19.128	495	500	48	27,9	32,0
UT	MTN	2.000	583	575	4	14,9	30,6

Tabla 2.1 *(continuación).*

Estado*	Región**	Población (1.000)	SAT Lengua	SAT Matemáticas	Porcentaje de alumnos presentados	Porcentaje sin estudios de secundaria	Salario de profesores ($1.000)
VT	NE	589	506	500	70	19,2	36,3
VA	SA	6.675	507	496	68	24,8	35,0
WA	PAC	5.533	519	519	47	16,2	38,0
WV	SA	1.826	526	506	17	34,0	32,2
WI	ENC	5.160	577	586	8	21,4	38,2
WY	MTN	481	544	544	11	17,0	31,6

* Para identificar los Estados véase la tabla 1.1.
** Las regiones censadas son East North Central, East South Central, Middle Atlantic, Mountain, New England Pacific, South Atlantic, West North Central y West South Central.
Fuente: *Statistical Abstract of the United States*, 1992.

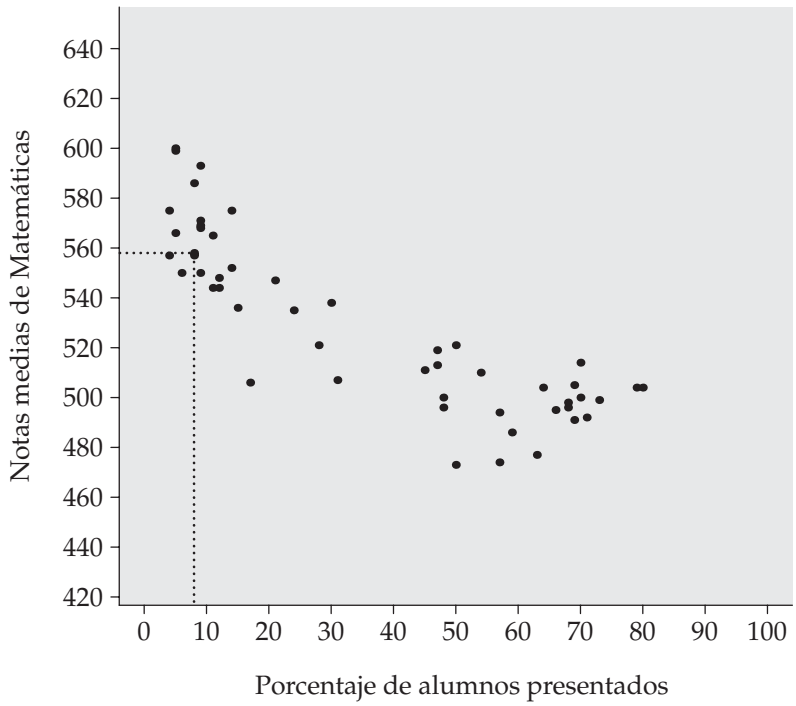

Figura 2.1. Diagrama de dispersión correspondiente a las notas medias de Matemáticas en la prueba SAT en relación con el porcentaje de alumnos que se presentan a dicho examen. La intersección de las líneas discontinuas corresponde al punto (8, 558), el dato del Estado de Alabama.

DIAGRAMA DE DISPERSIÓN

Un **diagrama de dispersión** muestra la relación entre dos variables cuantitativas medidas en los mismos individuos. Los valores de una variable aparecen en el eje de las abscisas y los de la otra en el eje de las ordenadas. Cada individuo aparece como un punto del diagrama. Su posición depende de los valores que toman las dos variables en cada individuo.

Sitúa siempre a la variable explicativa, si una de ellas lo es, en el eje de las abscisas del diagrama de dispersión. En general, llamamos a la variable explicativa x y a la variable respuesta y. Si no distinguimos entre variable explicativa y variable respuesta, cualquiera de las dos se puede situar en el eje de las abscisas.

APLICA TUS CONOCIMIENTOS

2.4. Manatís en peligro. Los manatís son unos animales grandes y dóciles que viven a lo largo de la costa de Florida. Cada año, lanchas motoras hieren o matan muchos manatís. A continuación, se presenta una tabla que contiene el número de licencias para lanchas motoras (expresado en miles de licencias por año) expedidas en Florida y el número de manatís muertos entre los años 1977 y 1990.

Año	Licencias expedidas (1.000)	Manatís muertos	Año	Licencias expedidas (1.000)	Manatís muertos
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

(a) Queremos analizar la relación entre el número de licencias anualmente expedidas en Florida y el número de manatís muertos cada año. ¿Cuál es la variable explicativa?

(b) Dibuja un diagrama de dispersión con estos datos. (Indica en los ejes los nombres de las variables, no te limites a indicar x e y.) ¿Qué nos dice el diagrama de dispersión sobre la relación entre estas dos variables?

2.2.1 Interpretación de los diagramas de dispersión

Para interpretar un diagrama de dispersión, aplica las estrategias de análisis de datos aprendidas en el capítulo 1.

EXAMEN DE UN DIAGRAMA DE DISPERSIÓN

En cualquier **gráfico de datos**, identifica el aspecto general y las **desviaciones** sorprendentes del mismo.

Puedes describir el aspecto general de un diagrama de dispersión mediante la **forma,** la **dirección** y la **fuerza** de la relación.

Un tipo importante de desviación son las **observaciones atípicas,** valores individuales que quedan fuera del aspecto general de la relación.

Grupos

La figura 2.1 muestra una *forma* clara: hay dos **grupos** distintos de Estados. En el grupo situado más a la derecha, el 45% o más de los alumnos se presentó a la prueba y las medias de los resultados estatales son bajas. Los Estados situados en el grupo de la izquierda tienen calificaciones más altas y porcentajes menores de alumnos presentados. No hay observaciones atípicas claras, es decir, no hay puntos situados de forma clara fuera de los grupos.

¿Qué puede explicar la existencia de dos grupos? En EE UU existen dos pruebas principales de acceso a la universidad: la prueba SAT (*Scholastic Assessment Test*) y la prueba ACT (*American College Testing*). En cada Estado predomina una de las dos pruebas. El grupo que aparece a la izquierda en el diagrama de dispersión de la figura 2.1 está constituido por Estados donde predomina la prueba ACT. El grupo de la derecha está formado por Estados en los que predomina la prueba SAT. En los Estados ACT, los alumnos que se presentan a la prueba SAT lo hacen porque quieren acceder a universidades más selectivas, que exigen una nota elevada en la prueba SAT. Este grupo selecto de estudiantes suele obtener unas notas en la prueba SAT superiores a las que obtienen los estudiantes de los Estados donde predomina dicha prueba.

La relación de la figura 2.1 tiene una *dirección* clara: los Estados donde el porcentaje de alumnos que se presentan a la prueba SAT es elevado tienden a tener notas medias más bajas. Tenemos una *asociación negativa* entre dos variables.

ASOCIACIÓN POSITIVA Y ASOCIACIÓN NEGATIVA

Dos variables están **asociadas positivamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores también situados por encima de la media de la otra variable, y cuando valores inferiores a la media también tienden a ocurrir conjuntamente.

Dos variables están **asociadas negativamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores inferiores a la media de la otra variable, y viceversa.

La *fuerza* de la relación del diagrama de dispersión está determinada por lo cerca que quedan los puntos de una determinada curva imaginaria. En general, la relación de la figura 2.1 no es fuerte —Estados con porcentajes similares de alumnos que se presentan a la prueba SAT muestran bastante variación en sus notas medias—. He aquí un ejemplo de una relación fuerte con una forma clara.

Tabla 2.2. Medias de grados-día y consumo de gas de la familia Sánchez.

Mes	Grados-día	Gas (m^3)
Noviembre	13,3	17,6
Diciembre	28,3	30,5
Enero	23,9	24,9
Febrero	18,3	21,0
Marzo	14,4	14,8
Abril	7,2	11,2
Mayo	2,2	4,8
Junio	0	3,4
Julio	0	3,4
Agosto	0,5	3,4
Septiembre	3,3	5,9
Octubre	6,7	8,7
Noviembre	16,7	17,9
Diciembre	17,8	20,2
Enero	28,9	30,8
Febrero	16,7	19,3

EJEMPLO 2.4. Calefacción del hogar

La familia Sánchez está a punto de instalar paneles solares en su casa para reducir el gasto en calefacción. Para conocer mejor el ahorro que puede significar, antes de instalar los paneles los Sánchez han ido registrando su consumo de gas en los últimos meses. El consumo de gas es más elevado cuando hace frío, por lo que debe existir una relación clara entre el consumo de gas y la temperatura exterior.

La tabla 2.2 muestra los datos de 16 meses.[1] La variable respuesta y es la media de los consumos de gas diarios durante el mes, en metros cúbicos (m^3). La variable explicativa x es la media de los grados-día de calefacción diarios durante el mes. (Los grados-día de calefacción son la medida habitual de la demanda de calefacción. Se acumula un grado-día por cada grado que la temperatura media diaria está por debajo de 18,5 ºC. Una temperatura media de 1 ºC, por ejemplo, corresponde a 17,5 grados-día de calefacción.

El diagrama de dispersión de la figura 2.2 muestra una asociación positiva fuerte. Más grados-día indican más frío y, por tanto, más gas consumido. La forma de la relación es **lineal**. Es decir, los puntos se sitúan a lo largo de una recta imaginaria. Es una relación fuerte porque los puntos se apartan poco de dicha recta. Si conocemos las temperaturas de un mes podemos predecir con bastante exactitud el consumo de gas. ■

Relación lineal

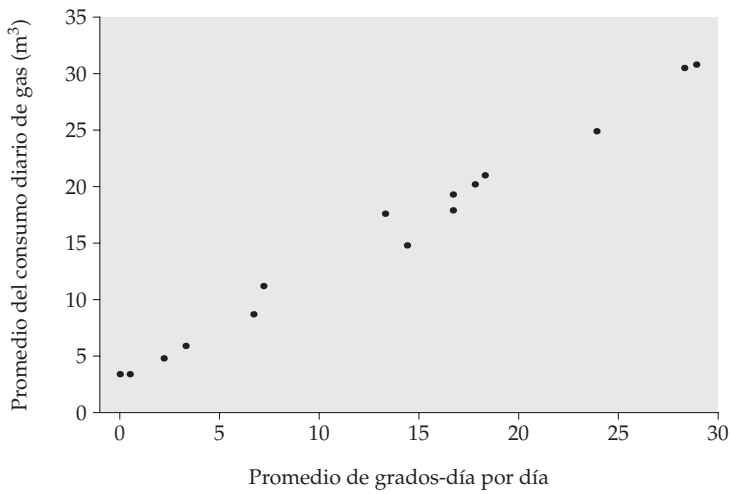

Figura 2.2. Diagrama de dispersión del consumo diario medio de gas de la familia Sánchez durante 16 meses en relación con la media diaria de grados-día en esos meses. Datos de la tabla 2.2.

[1]Datos de Robert Dale, Purdue University.

Por supuesto, no todas las relaciones son de tipo lineal. Es más, no todas las relaciones tienen una dirección clara que podamos describir como una asociación positiva o negativa. El ejercicio 2.6 da un ejemplo de una relación que no es lineal y que no tiene una dirección clara.

APLICA TUS CONOCIMIENTOS

2.5. Más sobre manatís en peligro. En el ejercicio 2.4 dibujaste un diagrama de dispersión del número de licencias para lanchas motoras registradas anualmente en Florida y del número de manatís que matan las lanchas cada año.

(a) Describe la dirección de la relación. Las variables, ¿están asociadas positiva o negativamente?

(b) Describe la forma de la relación. ¿Es lineal?

(c) Describe la fuerza de la relación. ¿Se puede predecir con precisión el número de manatís muertos cada año conociendo el número de licencias expedidas en ese año? Si Florida decidiera congelar el número de licencias en 716.000, ¿cuántos manatís matarían, aproximadamente, las lanchas motoras cada año?

2.6. El consumo, ¿aumenta con la velocidad? ¿Cómo varía el consumo de gasolina de un coche a medida que aumenta su velocidad? Aquí se presentan los datos correspondientes al modelo británico del Ford Escort. La velocidad se ha medido en kilómetros por hora y el consumo de carburante en litros de gasolina por 100 kilómetros.[2]

Velocidad (km/h)	Consumo (litros/100 km)	Velocidad (km/h)	Consumo (litros/100 km)
10	21,00	90	7,57
20	13,00	100	8,27
30	10,00	110	9,03
40	8,00	120	9,87
50	7,00	130	10,79
60	5,90	140	11,77
70	6,30	150	12,83
80	6,95		

[2]T. N. Lam, "Estimating fuel consumption from engine size", *Journal of Transportation Engineering*, 111, 1985, págs. 339-357.

(a) Dibuja un diagrama de dispersión. ¿Cuál es la variable explicativa?

(b) Describe la forma de la relación. ¿Por qué no es lineal? Explica lo que indica la forma de la relación.

(c) ¿Por qué no tiene sentido decir que las variables están asociadas positiva o negativamente?

(d) La relación, ¿es razonablemente fuerte o, por el contrario, es más bien débil? Justifica tu respuesta.

2.2.2 Inclusión de variables categóricas en los diagramas de dispersión

Desde hace tiempo, los resultados de los alumnos de las escuelas del sur de EE UU están por debajo del resto de escuelas del país. De todas formas, los esfuerzos para mejorar la educación han reducido la diferencia. En nuestro estudio sobre las pruebas de acceso a la universidad, los Estados del sur, ¿están por debajo de la media?

EJEMPLO 2.5. ¿El Sur es diferente?

Se han señalado en la figura 2.3 los Estados del Sur del diagrama de dispersión de la figura 2.1 con un símbolo diferente al resto de los Estados. (Consideramos como Estados del Sur los Estados de las regiones East South Central y South Atlantic.) En el diagrama, la mayoría de los Estados del Sur aparecen mezclados con los demás. De todas formas, algunos Estados del Sur se hallan en los bordes inferiores de sus grupos, junto con el Distrito de Columbia, que es más una ciudad que un Estado. Georgia, Carolina del Sur y Virginia Occidental tienen notas SAT inferiores a las que cabría esperar de acuerdo con el porcentaje de alumnos de secundaria que se presentan al examen. ■

Al clasificar los Estados en "Estados del Sur" y "resto de los Estados", hemos introducido una tercera variable en el diagrama de dispersión, una variable categórica que sólo tiene dos valores. Los dos valores se muestran con dos símbolos distintos. **Cuando quieras añadir una variable categórica a un diagrama de dispersión, utiliza colores o símbolos distintos para representar los puntos.**[3]

[3]W. S. Cleveland y R. McGill. "The many faces of a scatterplot", *Journal of the American Statistical Association*, 79, 1984, págs. 807-822.

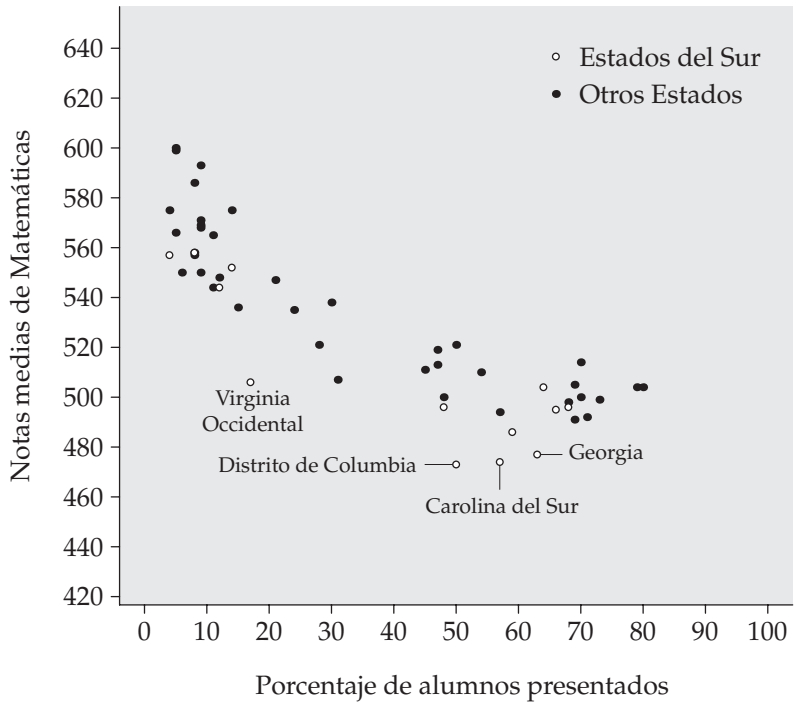

Figura 2.3. Nota media de Matemáticas en la prueba SAT y porcentaje de alumnos que se presenta a la prueba en cada Estado.

EJEMPLO 2.6. Los paneles solares, ¿reducen el consumo de gas?

Al poco tiempo de recopilar los datos que aparecen en la tabla 2.2 y en la figura 2.2, la familia Sánchez decidió instalar paneles solares en su casa. Para determinar el ahorro de gas que podía representar la instalación de estos paneles, los Sánchez registraron su consumo de gas durante 23 meses más. Para ver este efecto, añadimos los nuevos grados-día y el consumo de gas de estos meses en el diagrama de dispersión. La figura 2.4 es el resultado. Utilizamos símbolos distintos para distinguir los datos de "antes" de los de "después". En los meses poco fríos no hay mucha diferencia entre los dos grupos de datos. En cambio, en los meses más fríos el consumo de gas es claramente menor después de instalar los paneles solares. El diagrama de dispersión muestra que se ahorra energía después de instalar los paneles. ■

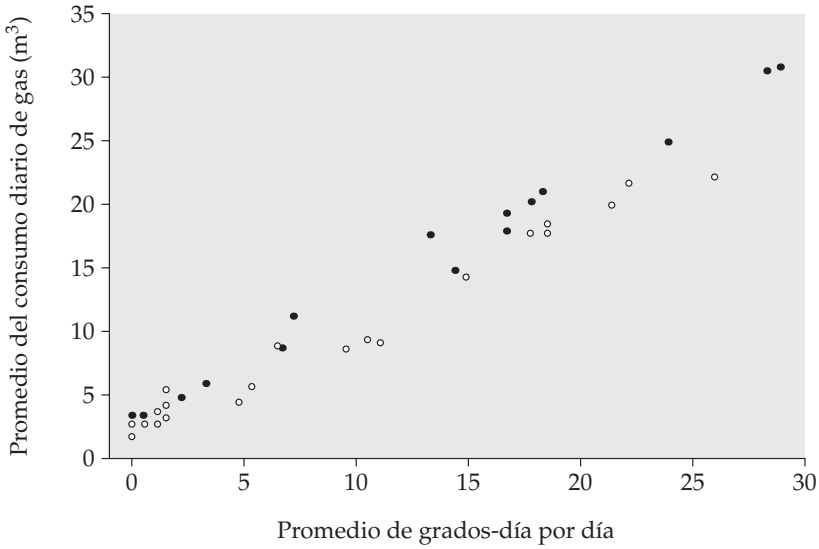

Figura 2.4. Consumo de gas con relación a los grados-día. Los puntos negros corresponden a los 16 meses sin paneles solares, mientras que los blancos corresponden a los meses con paneles solares.

Nuestro ejemplo sobre el consumo de gas tiene un problema que se presenta frecuentemente al dibujar diagramas de dispersión. Puede ser que no te des cuenta del problema cuando hagas el dibujo con un ordenador. Cuando algunos individuos tienen exactamente los mismos valores, ocupan el mismo punto del diagrama de dispersión. Fíjate en los valores de junio y julio de la tabla 2.2. La tabla 2.2 contiene datos de 16 meses, en cambio en la figura 2.2 sólo aparecen 15 puntos: junio y julio ocupan el mismo punto. Puedes utilizar símbolos distintos para señalar los puntos que representan más de una observación. Algunos programas estadísticos lo hacen automáticamente, pero otros no. Te recomendamos que utilices símbolos distintos para las observaciones repetidas cuando representes a mano un número pequeño de observaciones.

APLICA TUS CONOCIMIENTOS

2.7. La gente obesa, ¿consume más energía? El nivel metabólico de una persona, es decir, el ritmo al que su cuerpo consume energía, es un factor importante a

tener en cuenta en estudios de dietética. La tabla 2.3 proporciona datos sobre el sexo, el peso magro (peso total descontando su contenido en grasa) y el nivel metabólico en reposo de 12 mujeres y 7 hombres que eran los sujetos de un estudio de dietética. El nivel metabólico se expresa en calorías consumidas en 24 horas, la misma unidad utilizada para expresar el valor energético de los alimentos. Los investigadores creen que el peso magro corporal tiene una importante influencia en el nivel metabólico.

Tabla 2.3. Peso magro corporal y nivel metabólico.

Sujeto	Sexo	Peso (kg)	Nivel metabólico	Sujeto	Sexo	Peso (kg)	Nivel metabólico
1	H	62,0	1.792	11	M	40,3	1.189
2	H	62,9	1.666	12	M	33,1	913
3	M	36,1	995	13	H	51,9	1.460
4	M	54,6	1.425	14	M	42,4	1.124
5	M	48,5	1.396	15	M	34,5	1.052
6	M	42,0	1.418	16	M	51,1	1.347
7	H	47,4	1.362	17	M	41,2	1.204
8	M	50,6	1.502	18	H	51,9	1.867
9	M	42,0	1.256	19	H	46,9	1.439
10	H	48,7	1.614				

(a) Dibuja un diagrama de dispersión sólo con los datos de las mujeres. ¿Cuál sería la variable explicativa?

(b) La asociación entre estas dos variables, ¿es positiva o negativa? ¿Cuál es la forma de la relación? ¿Cuál es la fuerza de la relación?

(c) Ahora, añade en el diagrama de dispersión los datos de los hombres utilizando un color o un símbolo distinto al utilizado para las mujeres. La relación entre el nivel metabólico y el peso magro de los hombres, ¿es igual al de las mujeres? ¿En qué se distinguen el grupo de hombres y el grupo de mujeres?

RESUMEN DE LA SECCIÓN 2.2

Para estudiar la relación entre variables, tenemos que medir las variables sobre el mismo grupos de individuos.

Si creemos que los cambios de una variable x explican o que incluso son la causa de los cambios de una segunda variable y, a la variable x la **llamaremos variable explicativa** y a la variable y **variable respuesta**.

Un **diagrama de dispersión** muestra la relación entre dos variables cuantitativas, referidas a un mismo grupo de individuos. Los valores de una variable se

sitúan en el eje de las abscisas y los valores de la otra en el de las ordenadas. Cada observación viene representada en el diagrama por un punto.

Si una de las dos variables se puede considerar una variable explicativa, sus valores se sitúan siempre en el eje de las abscisas del diagrama de dispersión. Sitúa la variable respuesta en el eje de las ordenadas.

Para mostrar el efecto de las variables categóricas, representa los puntos de un diagrama de dispersión con colores o símbolos distintos.

Cuando analices un diagrama de dispersión, identifica su aspecto general describiendo la **dirección**, la **forma** y la **fuerza** de la relación, y luego identifica las **observaciones atípicas** y otras desviaciones.

Forma: relaciones lineales cuando los puntos del diagrama de dispersión se sitúan aproximadamente a lo largo de una recta, son una forma importante de relación entre dos variables. Las relaciones curvilíneas y las **agrupaciones** son otras formas en las que también tienes que fijarte.

Dirección: si la relación entre las dos variables tiene una dirección clara, decimos que existe una **asociación positiva** (si valores altos de las dos variables tienden a ocurrir simultáneamente) o una **asociación negativa** (si valores altos de una variable tienden a coincidir con valores bajos de la otra).

Fuerza: la **fuerza** de la relación entre variables viene determinada por la proximidad de los puntos del diagrama a alguna forma simple como, por ejemplo, una recta.

EJERCICIOS DE LA SECCIÓN 2.2

2.8. Inteligencia y calificaciones escolares. Los estudiantes que tienen coeficientes de inteligencia mayores, ¿tienden a ser mejores en la escuela? La figura 2.5 es un diagrama de dispersión correspondiente a las calificaciones medias escolares y a los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato en una escuela rural.[4]

(a) Explica en palabras qué significaría una asociación positiva entre el coeficiente de inteligencia y la nota media escolar. El diagrama, ¿muestra una asociación positiva?

(b) ¿Cuál es la forma de la relación? ¿Es aproximadamente lineal? ¿Es una relación muy fuerte? Justifica tus respuestas.

[4]Datos de Darlene Gordon, Purdue University.

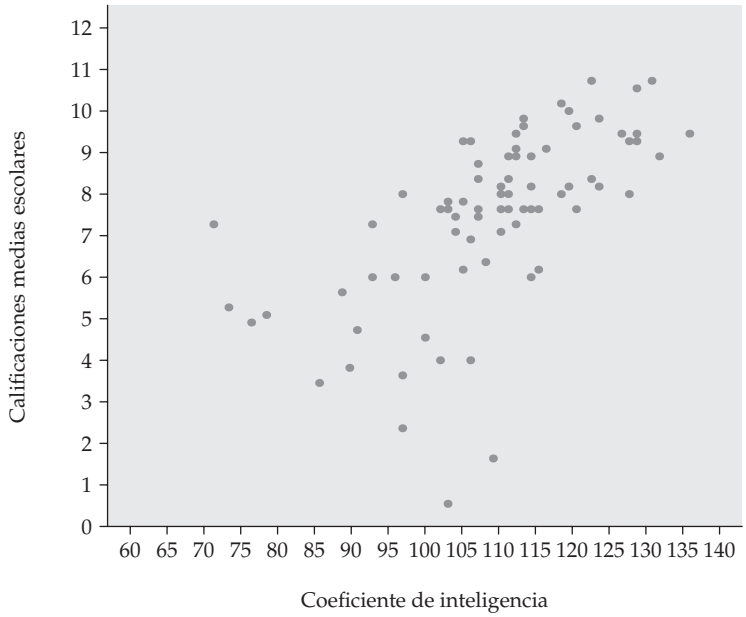

Figura 2.5. Diagrama de dispersión de las calificaciones medias escolares *versus* el coeficiente de inteligencia. Para el ejercicio 2.8.

(c) En la parte baja del diagrama aparecen algunos puntos que podríamos llamar observaciones atípicas. En concreto, un estudiante tiene una nota escolar muy baja, a pesar de tener un coeficiente de inteligencia medio. ¿Cuáles son, de forma aproximada, el coeficiente de inteligencia y la nota media escolar de este estudiante?

2.9. Calorías y sal en salchichas. Las salchichas con un contenido alto en calorías, ¿tienen también un contenido alto en sal? La figura 2.6 es un diagrama de dispersión que relaciona las calorías con el contenido en sal (expresado en miligramos de sodio) de 17 marcas distintas de salchichas elaboradas con carne de ternera.[5]

(a) Di de manera aproximada cuáles son los valores máximo y mínimo del contenido en calorías de las distintas marcas. De forma aproximada, ¿cuáles son los contenidos de sal de las marcas con más y con menos calorías?

[5] *Consumer Reports*, junio 1986, págs. 366-367.

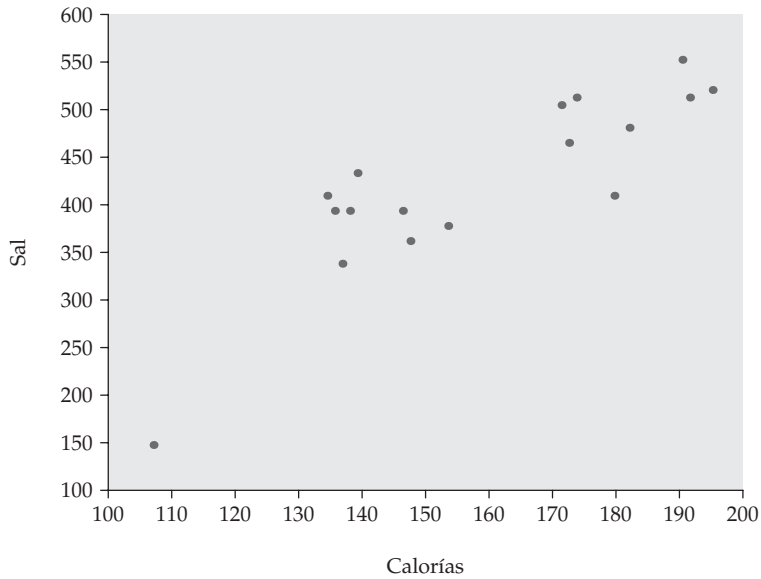

Figura 2.6. Diagrama de dispersión que relaciona las calorías y el contenido de sal de 17 marcas de salchichas. Para el ejercicio 2.9.

(b) El diagrama de dispersión, ¿muestra alguna asociación positiva o negativa clara? Explica con palabras el significado de esta asociación.

(c) ¿Has identificado alguna observación atípica? Prescindiendo de las posibles observaciones atípicas, ¿existe una relación lineal entre estas variables? Si ignoras las observaciones atípicas, ¿crees que existe una asociación fuerte entre ambas variables?

2.10. Estados ricos y Estados pobres. Una medida de la riqueza de un Estado es la mediana de ingresos por hogar. Otra medida de riqueza es la media de ingresos por persona. La figura 2.7 es un diagrama de dispersión que relaciona estas dos variables en EE UU. Ambas variables se expresan en miles de dólares. Debido a que las dos variables se expresan en las mismas unidades, la separación entre unidades es la misma en ambos ejes.[6]

[6]1997 *Statistical Abstract of the United States.*

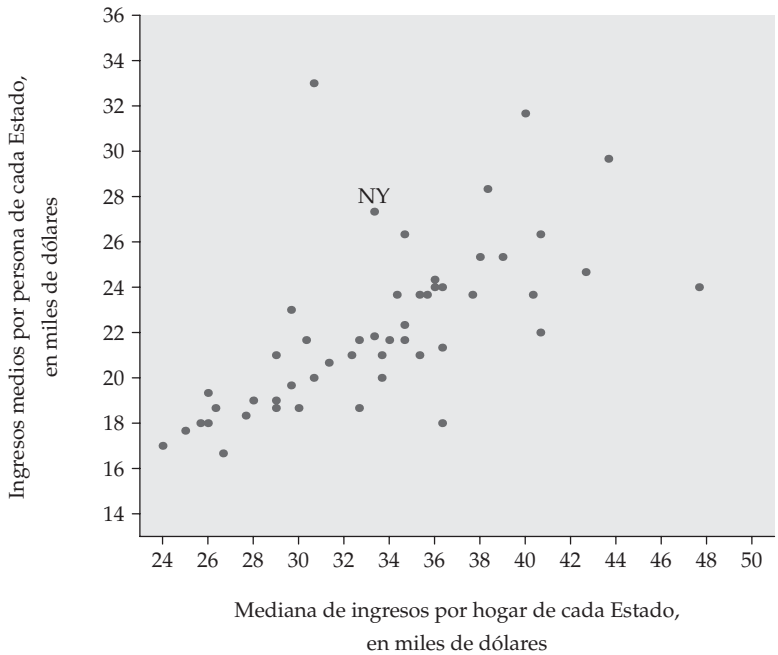

Figura 2.7. Diagrama de dispersión que relaciona los ingresos medios por persona con la mediana de ingresos por hogar. Para el ejercicio 2.10.

(a) En el diagrama de dispersión, hemos señalado el punto correspondiente a Nueva York. ¿Cuáles son, aproximadamente, los valores de la mediana de ingresos por hogar y la media de ingresos por persona?

(b) Explica por qué esperamos que haya una asociación positiva entre estas variables. Explica también, por qué esperamos que los ingresos por hogar sean mayores que los ingresos por persona.

(c) Sin embargo, en un determinado Estado, la media de los ingresos por persona puede ser mayor que la mediana de ingresos por hogar. De hecho, el Distrito de Columbia tiene una mediana de ingresos por hogar de 30.748 $ y una media de ingresos por persona de 33.435 $. Explica por qué esto puede ocurrir.

(d) Alaska es el Estado con la mediana de ingresos por hogar mayor. ¿Cuál es aproximadamente su mediana de ingresos por hogar? Podemos considerar Alaska y el Distrito de Columbia observaciones atípicas.

(e) Obviando las observaciones atípicas, describe la forma, la dirección y la fuerza de la relación.

2.11. El vino, ¿es bueno para tu corazón? Existe alguna evidencia de que tomar vino con moderación ayuda a prevenir los ataques al corazón. La tabla 2.4 proporciona datos sobre el consumo de vino (en litros de alcohol, procedente del vino, por cada 100.000 personas) y sobre las muertes anuales por ataques al corazón (muertos por cada 100.000 personas) en 19 países desarrollados.[7]

Tabla 2.4. Consumo de vino y enfermedades del corazón.

País	Consumo de alcohol*	Tasa de muertes por ataques al corazón**	País	Consumo de alcohol	Tasa de muertes por ataques al corazón
Alemania	2,7	172	Holanda	1,8	167
Australia	2,5	211	Irlanda	0,7	300
Austria	3,9	167	Islandia	0,8	211
Bélgica/Lux.	2,9	131	Italia	7,9	107
Canadá	2,4	191	Noruega	0,8	227
Dinamarca	2,9	220	N. Zelanda	1,9	266
España	6,5	86	Reino Unido	1,3	285
EE UU	1,2	199	Suecia	1,6	207
Finlandia	0,8	297	Suiza	5,8	115
Francia	9,1	71			

* Procedente del vino. ** Por 100.000 personas

(a) Dibuja un diagrama de dispersión que muestre cómo el consumo nacional de vino ayuda a explicar las muertes por ataques al corazón.

(b) Describe la forma de la relación. ¿Existe una relación lineal? ¿Es una relación fuerte?

(c) La dirección de la asociación, ¿es positiva o negativa? Explica de forma llana qué dice la relación sobre el consumo de vino y los ataques al corazón. Estos datos, ¿proporcionan una clara evidencia de que tomar vino causa una reducción de las muertes por ataques al corazón? ¿Por qué?

2.12. El profesor Moore y la natación. El profesor Moore nada 1.800 metros de forma regular. Un intento inútil de contrarrestar el paso de los años. He aquí los tiempos (en minutos) y su ritmo cardíaco después de nadar (en pulsaciones por minuto) en 23 sesiones de natación.

[7] M. H. Criqui, University of California, San Diego. Apareció en el *New York Times*, el 28 de diciembre de 1994.

Minutos	34,12	35,72	34,72	34,05	34,13	35,72	36,17	35,57
Pulsaciones	152	124	140	152	146	128	136	144
Minutos	35,37	35,57	35,43	36,05	34,85	34,70	34,75	33,93
Pulsaciones	148	144	136	124	148	144	140	156
Minutos	34,60	34,00	34,35	35,62	35,68	35,28	35,97	
Pulsaciones	136	148	148	132	124	132	139	

(a) Dibuja un diagrama de dispersión. (¿Cuál es la variable explicativa?)

(b) La asociación entre estas variables, ¿es positiva o negativa? Explica por qué crees que la relación va en este sentido.

(c) Describe la forma y la fuerza de la relación.

2.13. ¿Qué densidad de siembra es excesiva? ¿Cuál debe ser la densidad de siembra del maíz para que un agricultor obtenga el máximo rendimiento? Si se siembran pocas plantas, el suelo estará poco aprovechado y el rendimiento será bajo. Si se siembra muy denso, las plantas competirán por el agua y los nutrientes del suelo, por lo que el rendimiento tampoco será el deseado. Para determinar la densidad de siembra óptima, se hace un experimento que consiste en sembrar plantas de maíz a distintas densidades de siembra en parcelas de fertilidad similar. Los rendimientos obtenidos son los siguientes:[8]

Densidad de siembra (plantas por hectárea)	**Rendimiento (toneladas por hectárea)**			
30.000	10,1	7,6	7,9	9,6
40.000	11,2	8,1	9,1	10,1
50.000	11,1	8,7	9,4	10,1
60.000	9,1	9,3	10,5	
70.000	8,0	10,1		

(a) ¿Cuál es la variable explicativa: el rendimiento o la densidad de siembra?

(b) Dibuja un diagrama de dispersión con los datos del rendimiento y de la densidad de siembra.

(c) Describe el aspecto general de la relación. ¿Es una relación lineal? ¿Existe una asociación positiva, negativa o ninguna de las dos?

(d) Calcula los rendimientos medios de cada una de las densidades de siembra. Dibuja un diagrama de dispersión que relacione estas medias con la densidad de siembra. Une las medias con segmentos para facilitar la interpretación del

[8]W. L. Colville y D. P. McGill, "Effect of rate and method of planting on several plant characters and yield of irrigated corn", *Agronomy Journal*, 54, 1962, págs. 235-238.

diagrama. ¿Qué densidad de siembra recomendarías a un agricultor que quisiera sembrar maíz en un campo de fertilidad similar a la del experimento?

2.14. Salario de profesores. La tabla 2.1 muestra datos sobre la educación en EE UU. Es posible que los Estados con un nivel educativo menor paguen menos a sus profesores. Esto se podría explicar por el hecho de que son más pobres.

(a) Dibuja un diagrama de dispersión que relacione la media de los salarios de los profesores y el porcentaje de residentes que no tienen una carrera universitaria. Considera esta última variable como explicativa.

(b) El diagrama muestra una asociación negativa débil entre las dos variables. ¿Por qué decimos que la relación es negativa? ¿Por qué decimos que es débil?

(c) En la parte superior izquierda de tu diagrama hay una observación atípica. ¿A qué Estado corresponde?

(d) Existe un grupo bastante claro formado por nueve Estados en la parte inferior derecha del diagrama. Estos Estados tienen muchos residentes que no se graduaron en una escuela secundaria y además los salarios de los profesores son bajos. ¿Qué Estados son? ¿Se sitúan en alguna parte concreta del país?

2.15. Transformación de datos. Al analizar datos, a veces conviene hacer una **transformación de datos** que simplifique el aspecto general de la relación. A continuación se presenta un ejemplo de cómo transformando la variable respuesta se puede simplificar el aspecto del diagrama de dispersión. La población europea entre los años 1750 y 1950 creció de la siguiente manera:

Transformación de datos

Año	1750	1800	1850	1900	1950
Población (millones)	125	187	274	423	594

(a) Dibuja el diagrama de dispersión correspondiente a estos datos. Describe brevemente el tipo de crecimiento en el periodo señalado.

(b) Calcula los logaritmos de la población de cada uno de los años (puedes utilizar tu calculadora). Dibuja un nuevo diagrama de dispersión con la variable población transformada. ¿Qué tipo de crecimiento observas ahora?

2.16. Variable categórica explicativa. Un diagrama de dispersión muestra la relación entre dos variables cuantitativas. Vamos a ver un gráfico similar en el que la variable explicativa será una variable categórica en vez de una cuantitativa.

La presencia de plagas (insectos nocivos) en los cultivos se puede determinar con la ayuda de trampas. Una de ellas consiste en una lámina de plástico de distintos colores que contiene en su superficie un material pegajoso. ¿Qué colores

atraen más a los insectos? Para responder a esta pregunta un grupo de investigadores llevó a cabo un experimento que consistió en situar en un campo de avena 24 trampas de las cuales había 6 de color amarillo, 6 blancas, 6 verdes y 6 azules.[9]

Color de la trampa	Insectos capturados					
Amarillo	45	59	48	46	38	47
Blanco	21	12	14	17	13	17
Verde	37	32	15	25	39	41
Azul	16	11	20	21	14	7

(a) Dibuja un gráfico que relacione los recuentos de insectos capturados con el color de la trampa (sitúa el color de las trampas a distancias iguales en el eje de las abscisas). Calcula las medias de insectos atrapados en cada tipo de trampa, añádelas al gráfico y únelas con segmentos.

(b) ¿Qué conclusión puedes obtener de este gráfico sobre la atracción de estos colores sobre los insectos?

(c) ¿Tiene sentido hablar de una asociación positiva o negativa entre el color de la trampa y el número de insectos capturados?

2.3 Correlación

Un diagrama de dispersión muestra la forma, la dirección y la fuerza de la relación entre dos variables cuantitativas. Las relaciones lineales son especialmente importantes, ya que una recta es una figura sencilla bastante común. Decimos que una relación lineal es fuerte si los puntos del diagrama de dispersión se sitúan cerca de la recta, y débil si los puntos se hallan muy esparcidos respecto de la recta. De todas maneras, a simple vista, es difícil determinar la fuerza de una relación lineal. Los dos diagramas de dispersión de la figura 2.8 representan exactamente los mismos datos, con la única diferencia de la escala de los ejes. El diagrama de dispersión inferior da la impresión de que la asociación entre las dos variables es más fuerte. Es fácil engañar a la vista cambiando la escala.[10] Por ello, necesitamos seguir nuestra estrategia para el análisis de datos y utilizar una medida numérica que complemente el gráfico. La *correlación* es la medida que necesitamos.

[9] Adaptado de M. C. Wilson y R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug", *Journal of Economic Entomology*, 60, 1967, págs. 578-580.

[10] W. S. Cleveland, P. Diaconis y R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased", *Science*, 216, 1982, págs. 1138-1141.

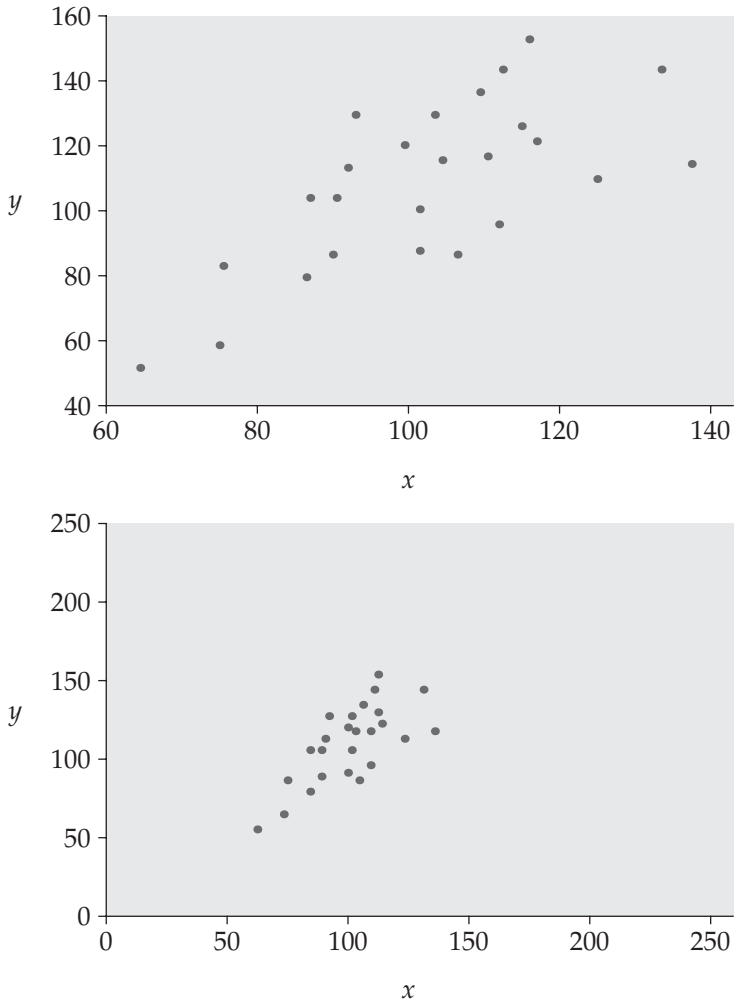

Figura 2.8. Dos diagramas de dispersión con los mismos datos. Debido a las diferentes escalas utilizadas, la fuerza de la relación lineal parece mayor en el gráfico inferior.

2.3.1 Correlación *r*

CORRELACIÓN

La **correlación** mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. La correlación se simboliza con la letra r.

Supón que tenemos datos de dos variables x e y para n individuos. Los valores para el primer individuo son x_1 e y_1, para el segundo son x_2 e y_2, etc. Las medias y las desviaciones típicas de las dos variables son $\bar{x}$ y s_x para los valores de x, e $\bar{y}$ y s_y para los valores de y. La correlación r entre x e y es

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Como siempre, $\sum$ (la letra sigma mayúscula del alfabeto griego) indica "suma estos términos para todos los individuos". La fórmula de la correlación r es algo complicada. Nos ayuda a entender qué es la correlación pero, en la práctica, conviene utilizar un programa estadístico o una calculadora para hallar r a partir de los valores de las dos variables x e y. Con el objetivo de consolidar tu comprensión del significado de la correlación, en el ejercicio 2.17 tienes que calcular la correlación paso a paso a partir de la definición.

La fórmula de r empieza estandarizando las observaciones. Supón, por ejemplo, que x es la altura en centímetros e y el peso en kilogramos y que tenemos las alturas y los pesos de n personas. Por tanto, $\bar{x}$ y s_x son la media y la desviación típica de las n alturas, ambas expresadas en centímetros. El valor

$$\frac{x_i - \bar{x}}{s_x}$$

es la altura estandarizada de la i-ésima persona, tal como vimos en el capítulo 1. La altura estandarizada nos indica a cuántas desviaciones típicas se halla la altura de un individuo con respecto a la media. Los valores estandarizados no tienen unidades de medida —en este ejemplo, las alturas estandarizadas ya no se expresan en centímetros—. Estandariza también los pesos. La correlación r es como una media de los productos de las alturas estandarizadas y de los pesos estandarizados para las n personas.

APLICA TUS CONOCIMIENTOS

2.17. Clasificación de fósiles. El *Archaeopteryx* es una especie extinguida que tenía plumas como un pájaro, pero que también tenía dientes y cola como un reptil. Sólo se conocen seis fósiles de estas características. Como estos especímenes difieren mucho en su tamaño, algunos científicos creen que pertenecen a especies distintas. Vamos a examinar algunos datos. Si los fósiles pertenecen a la misma especie y son de tamaños distintos porque unos son más jóvenes que otros, tiene que haber una relación lineal positiva entre las longitudes de algunos de los huesos en todos los individuos. Una observación atípica en esta relación sugeriría una especie distinta. He aquí los datos de las longitudes en centímetros del fémur y del húmero de cinco fósiles que conservan ambos huesos.[11]

Fémur	38	56	59	64	74
Húmero	41	63	70	72	84

(a) Dibuja un diagrama de dispersión. ¿Crees que los 5 fósiles pertenecen a la misma especie?

(b) Halla la correlación r, paso a paso. Es decir, halla la media y la desviación típica de las longitudes de los fémures y de los húmeros. (Utiliza tu calculadora para calcular las medias y las desviaciones típicas.) Halla los valores estandarizados de cada valor. Calcula r a partir de su fórmula.

(c) Ahora entra los datos en tu calculadora y utiliza la función que permite calcular directamente r. Comprueba que obtienes el mismo valor que en (b).

2.3.2 Características de la correlación

La fórmula de la correlación ayuda a ver que r es positivo cuando existe una asociación positiva entre las variables. Por ejemplo, el peso y la altura están asociados positivamente. La gente que tiene una altura superior a la media tiende también a tener un peso superior a la media. Para esta gente los valores estandarizados de altura y peso son positivos. La gente que tiene una altura inferior a la media

[11] M. A. Houck *et al*. "Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*", *Science*, 247, 1900, págs. 195-198. Los autores llegan a la conclusión de que todos los especímenes corresponden a la misma especie.

también tiende a tener un peso inferior a la media. Los dos valores estandarizados son negativos. En ambos casos los productos de la fórmula de r son en su mayor parte positivos y, por tanto, también lo es r. De la misma manera, podemos ver que r es negativa cuando la asociación entre x e y es negativa. Un estudio más detallado de la fórmula proporciona más propiedades de r. A continuación tienes las siete ideas que necesitas conocer para poder interpretar correctamente la correlación.

1. La correlación no hace ninguna distinción entre variables explicativas y variables respuesta. Da lo mismo llamar x o y a una variable o a otra.
2. La correlación exige que las dos variables sean cuantitativas para que tenga sentido hacer los cálculos de la fórmula de r. No podemos calcular la correlación entre los ingresos de un grupo de personas y la ciudad en la que viven, ya que la ciudad es una variable categórica.
3. Como r utiliza los valores estandarizados de las observaciones, no varía cuando cambiamos las unidades de medida de x, de y o de ambas. Si en vez de medir la altura en centímetros lo hubiéramos hecho en pulgadas, o si en lugar de medir el peso en kilogramos lo hubiéramos hecho en libras, el valor de r sería el mismo. La correlación no tiene unidad de medida. Es sólo un número.
4. Una r positiva indica una asociación positiva entre las variables. Una r negativa indica una asociación negativa.
5. La correlación r siempre toma valores entre -1 y 1. Valores de r cercanos a 0 indican una relación lineal muy débil. La fuerza de la relación lineal aumenta a medida que r se aleja de 0 y se acerca a 1 o a -1. Los valores de r cercanos a -1 o a 1 indican que los puntos se hallan cercanos a una recta. Los valores extremos $r = -1$ o $r = 1$ sólo se dan cuando existe una relación lineal perfecta y los puntos del diagrama de dispersión están exactamente sobre una recta.
6. La correlación sólo mide la fuerza de una relación lineal entre dos variables. La correlación no describe las relaciones curvilíneas entre variables aunque sean muy fuertes.
7. Al igual que ocurre con la media y la desviación típica, la correlación se ve fuertemente afectada por unas pocas observaciones atípicas. La correlación de la figura 2.7 es $r = 0{,}634$ cuando se incluyen todas las observaciones, de todas formas aumenta hasta $r = 0{,}783$ cuando obviamos Alaska y el Distrito de Columbia. Cuando detectes la presencia de observaciones atípicas en el diagrama de dispersión, utiliza r con precaución.

Los diagramas de dispersión de la figura 2.9 ilustran cómo los valores de r cercanos a 1 o a -1 corresponden a relaciones lineales fuertes. Para dejar más claro el significado de r, las desviaciones típicas de ambas variables en estos diagramas son iguales, y también son iguales las escalas en los ejes de las abscisas y de las ordenadas. No es fácil, en general, estimar el valor de r a partir de la observación del diagrama de dispersión. Recuerda que un cambio de escala puede engañar tu vista, pero no modifica la correlación.

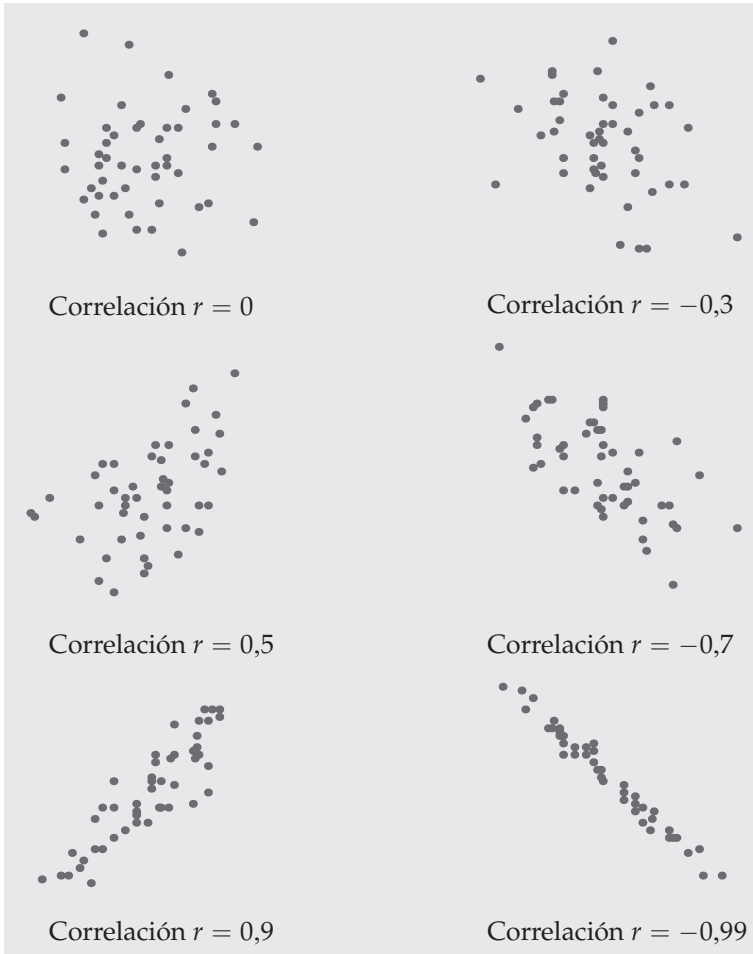

Figura 2.9. El coeficiente de correlación mide la fuerza de la asociación lineal. Cuando los puntos están muy cerca de la recta, los valores de r están más próximos a 1 o a -1.

Los datos reales que hemos examinado también ilustran cómo la correlación mide la fuerza y la dirección de relaciones lineales. La figura 2.2 muestra una relación lineal positiva muy fuerte entre los grados-día y el consumo de gas. La correlación es $r = 0{,}9953$. Compruébalo con tu calculadora utilizando los datos de la tabla 2.2. La figura 2.1 muestra una asociación negativa clara, aunque más débil, entre el porcentaje de estudiantes que se presentan a la prueba SAT y la nota media de Matemáticas en la prueba SAT de cada Estado de EE UU. En este caso, la correlación es $r = -0{,}8581$.

Recuerda que **la correlación no es una descripción completa de los datos de dos variables**, incluso cuando la relación entre las variables es lineal. Junto con la correlación tienes que dar las medias y las desviaciones típicas de x e y. (Debido a que la fórmula de la correlación utiliza las medias y las desviaciones típicas, estas medidas son las adecuadas para acompañar la correlación.) Conclusiones basadas sólo en las correlaciones puede que tengan que ser revisadas a la luz de una descripción más completa de los datos.

EJEMPLO 2.7. Puntuaciones para submarinistas

La condición física de los submarinistas profesionales se determina mediante las puntuaciones dadas por un grupo de jueces que utilizan una escala que va de 0 a 10. Existe alguna controversia sobre la objetividad de este método.

Tenemos las puntuaciones dadas por dos jueces, los señores Hernández y Fernández, a un grupo numeroso de submarinistas. ¿Concuerdan las puntuaciones de los dos jueces? Calculamos r y vemos que su valor es 0,9. Pero la media de las puntuaciones del Sr. Hernández está 3 puntos por debajo de la media del Sr. Fernández.

Estos hechos no se contradicen. Simplemente, son dos tipos diferentes de información. Las puntuaciones medias muestran que el Sr. Hernández puntúa más bajo que el Sr. Fernández. De todas formas, como el Sr. Hernández puntúa a todos los submarinistas con 3 puntos menos que el Sr. Fernández, la correlación permanece alta. Sumar o restar un mismo valor a todos los valores de x o de y no modifica la correlación. Aunque las puntuaciones de los jueces Hernández y Fernández son distintas, los submarinistas mejor puntuados por el juez Hernández son también los mejor puntuados por el juez Fernández. La r alta muestra la concordancia. Pero si el Sr. Hernández puntúa a un submarinista y el Sr. Fernández a otro, tenemos que añadir tres puntos a las puntuaciones del Sr. Hernández para que la comparación sea justa. ■

APLICA TUS CONOCIMIENTOS

2.18. Reflexiones sobre la correlación. La figura 2.5 es un diagrama de dispersión que relaciona las notas medias escolares y los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato.

(a) La correlación r de estos datos, ¿es próxima a -1, claramente negativa aunque no próxima a -1, próxima a 0, próxima a 1, claramente positiva pero no próxima a 1? Justifica tu respuesta.

(b) La figura 2.6, muestra las calorías y los contenidos de sodio de 17 marcas de salchichas. En esta ocasión, la correlación ¿es más próxima a 1 que la correlación de la figura 2.5? ¿es más próxima a 0? Justifica tu respuesta.

(c) Tanto la figura 2.5 como la figura 2.6 contienen observaciones atípicas. La eliminación de estas observaciones, ¿*aumentará* el coeficiente de correlación de una figura y lo *disminuirá* en la otra? ¿Qué ocurre en cada figura? ¿Por qué?

2.19. Si las mujeres siempre se casaran con hombres que fueran 2 años mayores que ellas, ¿cuál sería la correlación entre las edades de las esposas y las edades de sus maridos? (Sugerencia: dibuja un diagrama de dispersión con varias edades.)

2.20. Falta de correlación, pero asociación fuerte. A medida que aumenta la velocidad, el consumo de un automóvil disminuye al principio y luego aumenta. Supón que esta relación es muy regular, tal como muestran los siguientes datos de la velocidad (kilómetros por hora) y el consumo (litros por 100 km).

Velocidad (km/h)	30	45	55	70	85
Consumo (litros/100 km)	9,8	8,4	7,8	8,4	9,8

Dibuja un diagrama de dispersión del consumo con relación a la velocidad. Muestra que la correlación es $r = 0$. Explica por qué r es 0, a pesar de que existe una fuerte relación entre la velocidad y el consumo.

RESUMEN DE LA SECCIÓN 2.3

La **correlación** r mide la fuerza y la dirección de la asociación lineal entre dos variables cuantitativas x e y. Aunque puedes calcular r para cualquier diagrama de dispersión, r sólo mide la relación lineal.

La correlación indica la dirección de una relación lineal con su signo: $r > 0$ para asociaciones positivas y $r < 0$ para asociaciones negativas.

La correlación siempre cumple que $-1 \leq r \leq 1$. Valores de r cercanos a -1 o a 1 indican una fuerte asociación. Cuando los puntos de un diagrama de dispersión se sitúan exactamente a lo largo de una recta $r = \pm 1$.

La correlación ignora la distinción entre variables explicativas y variables respuesta. El valor de r no se ve afectado por cambios en las unidades de medida de cada una de las variables. De todas formas, r se puede ver muy afectada por las observaciones atípicas.

EJERCICIOS DE LA SECCIÓN 2.3

2.21. El profesor Moore y la natación. El ejercicio 2.12 proporciona datos sobre el tiempo que el profesor Moore, un hombre de mediana edad, tarda en nadar 1.800 metros y su ritmo cardíaco posterior.

(a) Si no lo hiciste en el ejercicio 2.12, calcula el coeficiente de correlación r. Explica, después de analizar el diagrama de dispersión, por qué el valor de r es razonable.

(b) Supón que los tiempos se hubieran medido en segundos. Por ejemplo, 34,12 minutos serían 2.047 segundos. ¿Cambiaría el valor de r?

2.22. Peso corporal y nivel metabólico. La tabla 2.3 proporciona datos sobre el nivel metabólico y el peso magro de 12 mujeres y 7 hombres.

(a) Dibuja un diagrama de dispersión si no lo hiciste en el ejercicio 2.7. Utiliza colores o símbolos distintos para las mujeres y para los hombres. ¿Crees que la correlación será aproximadamente igual para los hombres y las mujeres, o bastante distinta para los dos grupos? ¿Por qué?

(b) Calcula r para el grupo de las mujeres y también para el grupo de los hombres. (Utiliza la calculadora.)

(c) Calcula el peso magro medio de las mujeres y de los hombres. El hecho de que, como media, los hombres sean más pesados que las mujeres, ¿influye en las correlaciones? Si es así, ¿por qué?

(d) El peso magro se midió en kilogramos. ¿Cuál sería la correlación si lo hubiéramos medido en libras? (2,2 libras equivalen a 1 kilogramo.)

2.23. ¿Cuantas calorías? Una industria agroalimentaria solicita a un grupo de 3.368 personas que estimen el contenido en calorías de algunos alimentos. La tabla 2.5 muestra las medias de sus estimaciones y el contenido real en calorías.[12]

[12] De una encuesta de la Wheat Industry Council aparecida el 20 de octubre de 1983 en *USA Today*.

Tabla 2.5. Calorías estimadas y reales de 10 alimentos.

Alimento	Calorías estimadas	Calorías reales
225 g de leche entera	196	159
142 g de espaguetis con salsa de tomate	394	163
142 g de de macarrones con queso	350	269
Una rebanada de pan de trigo	117	61
Una rebanada de pan blanco	136	76
57 g de caramelos	364	260
Una galleta salada	74	12
Una manzana de tamaño medio	107	80
Una patata de tamaño medio	160	88
Una porción de pastel de crema	419	160

(a) Creemos que el contenido real en calorías de los alimentos, puede ayudar a explicar las estimaciones de la gente. Teniendo esto presente, dibuja un diagrama de dispersión con estos datos.

(b) Calcula la correlación r (utiliza tu calculadora). Explica, basándote en el diagrama de dispersión, por qué r es razonable.

(c) Las estimaciones son todas mayores que los valores reales. Este hecho, ¿influye de alguna manera en la correlación? ¿Cómo cambiaría r si todos los valores estimados fuesen 100 calorías más altos?

(d) Las estimaciones son demasiado altas para los espaguetis y los pasteles. Señala estos puntos en el diagrama de dispersión. Calcula r para los ocho alimentos restantes. Explica por qué r cambia en el sentido en que lo hace.

2.24. Peso del cerebro y coeficiente de inteligencia. La gente que tiene un cerebro mayor, ¿tiene también un coeficiente de inteligencia mayor? Un estudio realizado con 40 sujetos voluntarios, 20 hombres y 20 mujeres, proporciona una explicación. El peso del cerebro se determinó mediante una imagen obtenida por resonancia magnética (IRM). (En la tabla 2.6 aparecen estos datos. IRM es el recuento de "pixels" que el cerebro genera en la imagen. El coeficiente de inteligencia (CI) se midió mediante la prueba Wechsler.[13])

(a) Haz un diagrama de dispersión para mostrar la relación entre el coeficiente de inteligencia y el recuento de IRM. Utiliza símbolos distintos para hombres y mujeres. Además, halla la correlación entre ambas variables para los 40 sujetos, para los hombres y para las mujeres.

[13]L. Willerman, R. Schultz, J. N. Rutledge y E. Bigler, "In vivo brain size and intelligence", *Intelligence*, 15, 1991, págs. 223-228.

Tabla 2.6. Tamaño del cerebro y coeficiente de inteligencia.

Hombres				Mujeres			
IRM	**CI**	**IRM**	**CI**	**IRM**	**CI**	**IRM**	**CI**
1.001.121	140	1.038.437	139	816.932	133	951.545	137
965.353	133	904.858	89	928.799	99	991.305	138
955.466	133	1.079.549	141	854.258	92	833.868	132
924.059	135	945.088	100	856.472	140	878.897	96
889.083	80	892.420	83	865.363	83	852.244	132
905.940	97	955.003	139	808.020	101	790.619	135
935.494	141	1.062.462	103	831.772	91	798.612	85
949.589	144	997.925	103	793.549	77	866.662	130
879.987	90	949.395	140	857.782	133	834.344	83
930.016	81	935.863	89	948.066	133	893.983	88

(b) En general, los hombres son más corpulentos que los mujeres, por tanto sus cerebros suelen ser más grandes. ¿Cómo se muestra este efecto en tu diagrama? Halla la media del recuento de IRM para hombres y para mujeres para comprobar si existe diferencia.

(c) Tus resultados en (b) sugieren que para analizar la relación entre el coeficiente de inteligencia y el peso del cerebro, es mejor separar a hombres y mujeres. Utiliza tus resultados en (a) para comentar la naturaleza y la fuerza de esta relación para hombres y mujeres de forma separada.

2.25. Un cambio en las unidades de medida puede alterar drásticamente el aspecto de un diagrama de dispersión. Considera los siguientes datos:

x	-4	-4	-3	3	4	4
y	0,5	$-0,6$	$-0,5$	0,5	0,5	$-0,6$

(a) Dibuja un diagrama de dispersión con los datos anteriores en el que la escala de las ordenadas y la de las abscisas vayan de -6 a 6.

(b) Calcula, a partir de x e y, los valores de las nuevas variables: $x^* = \frac{x}{10}$ e $y^* = 10y$. Dibuja y^* en relación con x^* en el mismo diagrama de dispersión utilizando otros símbolos. El aspecto de los dos diagramas es muy diferente.

(c) Utiliza una calculadora para hallar la correlación entre x e y. Luego, halla la correlación entre x^* e y^*. ¿Cuál es la relación entre las dos correlaciones? Explica por qué este resultado no es sorprendente.

2.26. Docencia e investigación. Un periódico universitario entrevista a un psicólogo a propósito de las evaluaciones que hacen los estudiantes de sus profesores. El psicólogo afirma: "La evidencia demuestra que la correlación entre la

capacidad investigadora de los profesores y la evaluación docente que hacen los estudiantes es próxima a cero". El titular del periódico dice: "El profesor Cruz dice que los buenos investigadores tienden a ser malos profesores y viceversa". Explica por qué el titular del periódico no refleja el sentido de las palabras del profesor Cruz. Escribe en un lenguaje sencillo (no utilices la palabra "correlación") lo que quería decir el profesor Cruz.

2.27. Diversificación de inversiones. Un artículo en una revista de una asociación dice: "Una cartera bien diversificada incluye asientos con correlaciones bajas". El artículo incluye una tabla de correlaciones entre los rendimientos de varios tipos de inversiones. Por ejemplo, la correlación entre unos bonos municipales y acciones de grandes empresas es 0,50 y la correlación entre los bonos municipales y acciones de pequeñas empresas es 0,21.[14]

(a) María invierte mucho en bonos municipales y quiere diversificar sus inversiones añadiendo unas acciones que tengan unos rendimientos que no sigan la misma tendencia que los rendimientos de sus bonos. Para conseguir su propósito, ¿qué tipo de acciones debe escoger María, las acciones de grandes empresas o acciones de pequeñas empresas? Justifica tu respuesta.

(b) Si María quiere una inversión que tienda a aumentar cuando los rendimientos de sus bonos tiendan a disminuir, ¿qué tipo de correlación debe buscar?

2.28. Velocidad y consumo de gasolina. Los datos del ejercicio 2.20 se presentaron para mostrar un ejemplo de una relación curvilínea fuerte para la cual, sin embargo, $r = 0$. El ejercicio 2.6 proporciona datos sobre el consumo del Ford Escort con relación a la velocidad. Dibuja un diagrama de dispersión si no lo hiciste en el ejercicio 2.6. Calcula la correlación y explica por qué r está cerca de 0 a pesar de la fuerte relación entre la velocidad y el consumo.

2.29. ¿Dónde está el error? Cada una de las siguiente afirmaciones contiene un error. Explica en cada caso dónde está la incorrección.

(a) "Hay una correlación alta entre el sexo de los trabajadores y sus ingresos."

(b) "Hallamos una correlación alta ($r = 1{,}09$) entre las evaluaciones de los profesores hechas por los estudiantes y las hechas por otros profesores."

(c) "La correlación hallada entre la densidad de siembra y el rendimiento del maíz fue de $r = 0{,}23$ hectolitros."

[14] *T. Rowe Price Report*, invierno 1997, pág. 4.

2.4 Regresión mínimo-cuadrática

La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Si un diagrama de dispersión muestra una relación lineal, nos gustaría resumirla dibujando una recta a través de la nube de puntos. La regresión mínimo-cuadrática es un método para hallar una recta que resuma la relación entre dos variables, aunque sólo en una situación muy concreta: una de las variables ayuda a explicar o a predecir la otra. Es decir, la regresión describe una relación entre una variable explicativa y una variable respuesta.

RECTA DE REGRESIÓN

La **recta de regresión** es una recta que describe cómo cambia una variable respuesta y a medida que cambia una variable explicativa x.

A menudo, utilizamos una recta de regresión para predecir el valor de y a partir de un valor dado de x.

EJEMPLO 2.8. Predicción del consumo de gas

El diagrama de dispersión de la figura 2.10 muestra que existe una fuerte relación lineal entre la temperatura exterior media de un mes (medida en grados-día de calefacción diarios) y el consumo medio diario de gas de ese mes en casa de los Sánchez. La correlación es $r = 0{,}9953$, cerca de $r = 1$ que corresponde a los puntos situados sobre la recta. La recta de regresión trazada a través de los puntos de la figura 2.10 describe los datos muy bien.

La familia Sánchez quiere utilizar dicha relación para predecir su consumo de gas. "Si un mes tiene una media diaria de 10 grados-día, ¿cuánto gas utilizaremos?"

Predicción

Para **predecir** el consumo de gas de los días de un mes con una media de 10 grados-día, en primer lugar localiza el valor 10 en el eje de las abscisas. Luego, ves "hacia arriba y hacia la izquierda", como en la figura, para hallar el consumo de gas y que corresponde a $x = 10$. Predecimos que los Sánchez consumirán aproximadamente 12,5 metros cúbicos de gas cada día de ese mes. ■

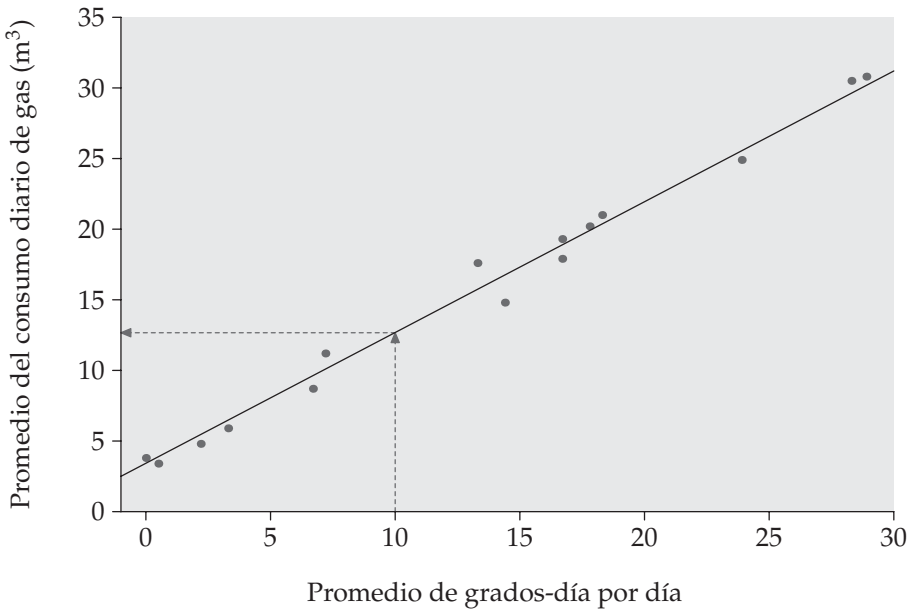

Figura 2.10. Datos del consumo de gas de la familia Sánchez, con una recta de regresión para predecir el consumo de gas a partir de los grados-día. Las líneas discontinuas muestran cómo predecir el consumo de gas en un mes con una media diaria de 10 grados-día.

2.4.1 Recta de regresión mínimo-cuadrática

Diferentes personas dibujarían, a simple vista, diferentes rectas en un diagrama de dispersión. Esto es especialmente cierto cuando los puntos están más dispersos que los de la figura 2.10. Necesitamos una manera de dibujar la recta de regresión que no dependa de nuestra intuición de por dónde tendría que pasar dicha recta. Utilizaremos la recta para predecir y a partir de x; en consecuencia, los errores de predicción estarán en y, el eje de las ordenadas del diagrama de dispersión. Si predecimos un consumo de 12,5 m^3 para un mes con 10 grados-día y el consumo real resulta ser de 13,45 m^3, nuestro error es

$$\begin{aligned} \text{error} &= \text{valor observado} - \text{valor predicho} \\ &= 13{,}45 - 12{,}5 = 0{,}95 \end{aligned}$$

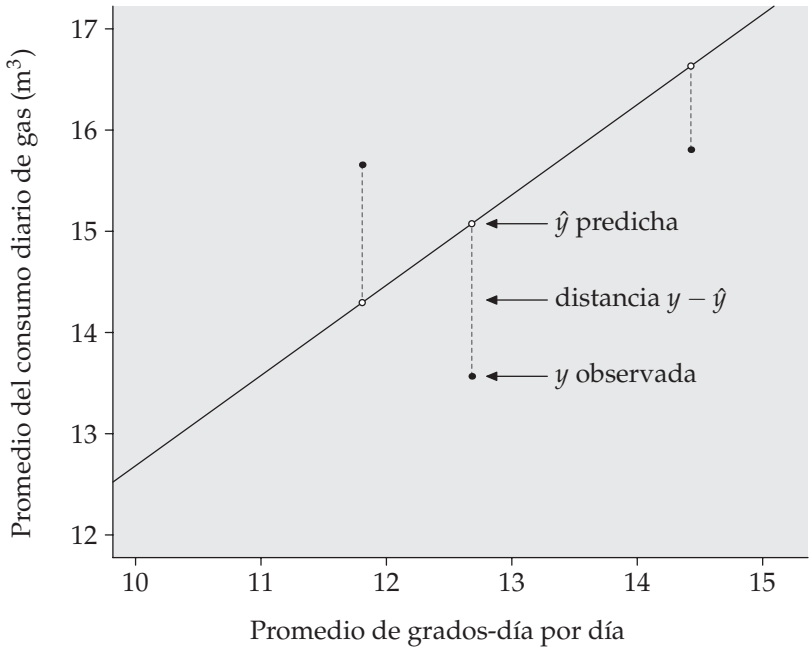

Figura 2.11. La idea de los mínimos cuadrados. Para cada observación, halla la distancia vertical de cada punto del diagrama de dispersión a la recta. La regresión mínimo-cuadrática hace que la suma de los cuadrados de estas distancias sea lo más pequeña posible.

Ninguna recta podrá pasar exactamente por todos los puntos del diagrama de dispersión. Queremos que las distancias *verticales* de los puntos a la recta sean lo más pequeñas posible. La figura 2.11 ilustra esta idea. El diagrama muestra sólo 3 puntos de la figura 2.10 conjuntamente con la recta y en una escala ampliada. La recta pasa por encima de dos de los puntos y por debajo de uno de ellos. Las distancias verticales de los puntos a la recta de regresión se han señalado con líneas discontinuas. Existen muchos procedimientos para conseguir que las distancias verticales "sean lo más pequeñas posible". El más común es el método de *mínimos cuadrados*.

RECTA DE REGRESIÓN MÍNIMO-CUADRÁTICA

La **recta de regresión mínimo-cuadrática** de y con relación a x es la recta que hace que la suma de los cuadrados de las distancias verticales de los puntos observados a la recta sea lo más pequeña posible.

Una de las razones de la popularidad de la recta de regresión mínimo-cuadrática es que el procedimiento para encontrar dicha recta es sencillo: se calcula a partir de las medias, las desviaciones típicas de las dos variables y su correlación.

ECUACIÓN DE LA RECTA DE REGRESIÓN MÍNIMO-CUADRÁTICA

Tenemos datos de la variable explicativa x y de la variable respuesta y para n individuos. A partir de los datos, calcula $\bar{x}$ e $\bar{y}$, las desviaciones típicas s_x y s_y de las dos variables y su correlación. La recta de regresión mínimo-cuadrática es

$$\hat{y} = a + bx$$

con **pendiente**

$$b = r\frac{s_y}{s_x}$$

y **ordenada en el origen**

$$a = \bar{y} - b\bar{x}$$

Escribimos $\hat{y}$ en la ecuación de la recta de regresión para subrayar que la recta *predice* una respuesta $\hat{y}$ para cada x. Debido a la dispersión de los puntos a lo largo de la recta, la respuesta predicha no coincidirá, por regla general, con la respuesta realmente *observada* y. En la práctica, no necesitas calcular primero las medias, las desviaciones típicas y la correlación. Cualquier programa estadístico, o tu calculadora, te dará la pendiente b y la ordenada en el origen a de la recta de regresión mínimo-cuadrática a partir de los valores de las variables x e y. Por tanto, puedes concentrarte en comprender y utilizar la recta de regresión.

EJEMPLO 2.9. Utilización de la recta de regresión

La recta de la figura 2.10 es de hecho la recta de regresión mínimo-cuadrática del consumo de gas con relación a los grados-día. Introduce los datos de la tabla 2.2 en tu calculadora y comprueba que la recta de regresión es

$$\hat{y} = 3{,}0949 + 0{,}94996x$$

Pendiente

La **pendiente** de una recta de regresión es importante para interpretar los datos. Esta pendiente es la tasa de cambio, la cantidad en que varía $\hat{y}$ cuando x aumenta en una unidad. La pendiente $b = 0{,}94996$ de este ejemplo dice que, como media, cada grado-día adicional predice un aumento diario del consumo de 0,94996 m^3 de gas.

Ordenada en el origen

La **ordenada en el origen** de la recta de regresión es el valor de $\hat{y}$ cuando $x = 0$. Aunque necesitamos el valor de la ordenada en el origen para dibujar la recta de regresión, sólo tiene significado estadístico cuando x toma valores cercanos a 0. En nuestro ejemplo, $x = 0$ ocurre cuando la temperatura exterior media es de al menos 18,5 °C. Predecimos que los Sánchez utilizarán una media de $a = 3{,}0949$ m^3 de gas diarios con 0 grados-día. Utilizan este gas para cocinar y para calentar el agua, y este consumo se mantiene incluso cuando no hace frío.

Predicción

La ecuación de la recta de regresión facilita la **predicción**. Tan sólo sustituye x por un valor concreto en la ecuación. Para predecir el consumo de gas a 10 grados-día, sustituye x por 10.

$$\begin{aligned} \hat{y} &= 3{,}0949 + (0{,}94996)(10) \\ &= 3{,}0949 + 9{,}4996 = 12{,}5945 \end{aligned}$$

Trazado de la recta

Para **trazar la recta** en el diagrama de dispersión, utiliza la ecuación para hallar $\hat{y}$ de dos valores de x que se encuentren en los extremos del intervalo determinado por los valores de x de los datos. Sitúa cada $\hat{y}$ sobre su respectiva x y traza la recta que pase por los dos puntos. ■

La figura 2.12 muestra los resultados de la regresión de los datos de consumo de gas obtenidos con una calculadora con funciones estadísticas y con dos programas estadísticos. Cada resultado da la pendiente y la ordenada en el origen de la recta mínimo-cuadrática, calculadas con más decimales de los que necesitamos. Los programas también proporcionan información que no necesitamos —la gracia de utilizar programas es saber prescindir de la información extra que siempre se proporciona—. En el capítulo 10, utilizaremos la información adicional de estos resultados.

```
LinReg
y = ax + b
a = .94996152
b = 3.09485166
r² = .99041538
r = .99519615
```

(a)

The regression equation is
Consumo-Gas = 3.09 + 0.95 G-dia

Predictor	Coef	Stdev	t-ratio	p
Constant	3.0949	0.3906	7.92	0.000
G-dia	0.94996	0.0250	38.04	0.000

s = 0.9539 R-sq = 99.0% R-sq(adj) = 99.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	1316.26	1316.26	1446.67	0.000
Error	14	12.74	0.91		
Total	15	1329.00			

(b)

Dependent variable is: Consumo-Gas
No Selector
R squared = 99.0% R squared (adjusted) = 99.0%
s = 0.9539 with 16-2 = 14 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1316.260	1	1316.260	1447
Residual	12.738	14	0.910	

Variable	Coefficient	s.e of Coeff	t-ratio	prob
Constant	3.094852	0.3906	7.92	$\leq$0.0001
G-dia	0.949962	0.0250	38.04	$\leq$0.0001

(c)

Figura 2.12. Resultados de la regresión mínimo-cuadrática del consumo de gas obtenidos con una calculadora y con dos programas estadísticos. (a) Calculadora TI-83. (b) Minitab. (c) Data Desk.

APLICA TUS CONOCIMIENTOS

2.30. El ejemplo 2.9 da la ecuación de la recta de regresión del consumo de gas y con relación a los grados-día x de los datos de la tabla 2.2 como

$$\hat{y} = 3{,}0949 + 0{,}94966x$$

Entra los datos de la tabla 2.2 en tu calculadora.

(a) Utiliza la función de regresión de la calculadora para hallar la ecuación de la recta de regresión mínimo-cuadrática.

(b) Utiliza tu calculadora para hallar la media y la desviación típica de x e y, y su correlación r. Halla la pendiente b y la ordenada en el origen a de la recta de regresión a partir de esos valores, utilizando las ecuaciones del recuadro *Ecuación de la recta de regresión mínimo-cuadrática*. Comprueba que en (a) y en (b) obtienes la ecuación del ejemplo 2.9. (Los resultados pueden ser algo distintos debido a los errores de redondeo.)

2.31. Lluvia ácida. Unos investigadores determinaron, durante 150 semanas consecutivas, la acidez de la lluvia en una zona rural de Colorado, EE UU. La acidez se determina mediante el pH. Valores de pH bajos indican una acidez alta. Los investigadores observaron una relación lineal entre el pH y el paso del tiempo e indicaron que la recta de regresión mínimo-cuadrática

$$\text{pH} = 5{,}43 - (0{,}0053 \times \text{semanas})$$

se ajustaba bien a los datos.[15]

(a) Dibuja esta recta. ¿La asociación es positiva o negativa? Explica de una manera sencilla el significado de esta asociación.

(b) De acuerdo con la recta de regresión, ¿cuál era el pH al comienzo del estudio (semana $= 1$)? ¿Y al final (semana $= 150$)?

(c) ¿Cuál es la pendiente de la recta de regresión? Explica claramente qué indica la pendiente respecto del cambio del pH del agua de lluvia en esta zona rural.

2.32. Manatís en peligro. El ejercicio 2.4 proporciona datos sobre el número de lanchas registradas en Florida y el número de manatís muertos por las lanchas motoras entre 1977 y 1990. La recta de regresión para predecir los manatís muertos a partir del número de lanchas motoras registradas es

$$\text{muertos} = -41{,}4 + (0{,}125 \times \text{lanchas})$$

[15] W. M. Lewis y M. C. Grant, "Acid precipitation in the western United States", *Science*, 207, 1980, págs. 176-177.

(a) Dibuja un diagrama de dispersión y añádele la recta de regresión. Predice el número de manatís que matarán las lanchas en un año en que se registraron 716.000 lanchas.

(b) He aquí nuevos datos sobre los manatís muertos durante cuatro años más.

Año	Licencias expedidas (1.000)	Manatís muertos
1991	716	53
1992	716	38
1993	716	35
1994	735	49

Añade estos puntos al diagrama de dispersión. Durante estos cuatro años, Florida tomó fuertes medidas para proteger a los manatís. ¿Observas alguna evidencia de que estas medidas tuvieron éxito?

(c) En el apartado (a) predijiste el número de manatís muertos en un año con 716.000 lanchas registradas. En realidad, el número de lanchas registradas se mantuvo en 716.000 durante los siguientes tres años. Compara las medias de manatís muertos en estos años con tu predicción en (a). ¿Qué nivel de exactitud has alcanzado?

2.4.2 Características de la regresión mínimo-cuadrática

La regresión mínimo-cuadrática tiene en cuenta las distancias de los puntos a la recta sólo en la dirección de y. Por tanto, en una regresión las variables x e y juegan papeles distintos.

Característica 1. La distinción entre variable explicativa y variable respuesta es básica en regresión. La regresión mínimo-cuadrática considera sólo las distancias verticales de los puntos a la recta. Si cambiamos los papeles de las dos variables, obtenemos una recta de regresión-mínimo cuadrática distinta.

EJEMPLO 2.10. El universo se expande

La figura 2.13 es un diagrama de dispersión dibujado con los datos que sirvieron de base para descubrir que el Universo se está expandiendo. Son las distancias a la Tierra de 24 galaxias y las velocidades con que éstas se alejan de nosotros, proporcionadas por el astrónomo Edwin Hubble en 1929.[16] Existe una relación

[16]E. P. Hubble, "A relation between distance and radial velocity among extra-galactic nebulae", *Proceedings of the National Academy of Sciences*, 15, 1929, págs. 168-173.

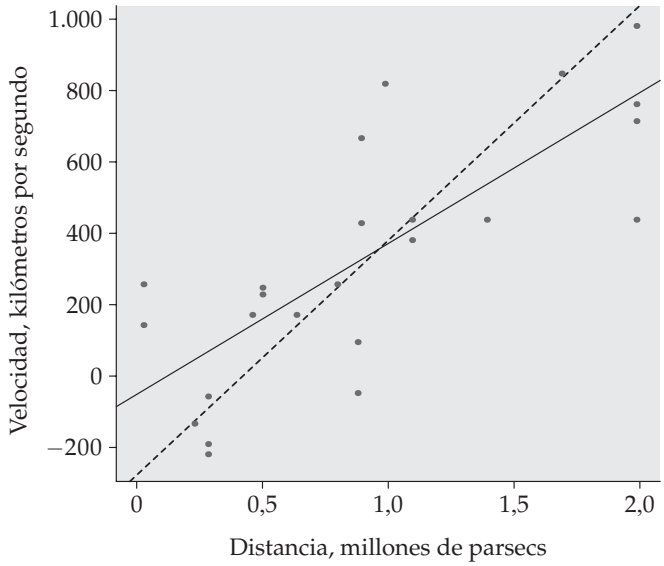

Figura 2.13. Diagrama de dispersión de los datos de Hubble sobre la distancia a la Tierra de 24 galaxias y la velocidad con la que éstas se alejan de nosotros. Las dos rectas son de regresión mínimo-cuadrática: la de la velocidad con relación a la distancia (línea continua) y la de la distancia con relación a la velocidad (línea discontinua).

lineal positiva, $r = 0{,}7842$, de manera que las galaxias que se hallan más lejos se alejan más rápidamente. De hecho, los astrónomos creen que la relación es perfectamente lineal y que la dispersión se debe a errores de medición.

Las dos rectas del dibujo son rectas de regresión mínimo-cuadrática. La recta de trazado continuo es la regresión de la velocidad con relación a la distancia, mientras que la de trazado discontinuo es la regresión de la distancia con relación a la velocidad. *La regresión de la velocidad con relación a la distancia y la regresión de la distancia con relación a la velocidad dan rectas distintas.* Al determinar la recta de regresión, debes saber cuál es la variable explicativa. ■

Característica 2. Existe una estrecha conexión entre la correlación y la regresión. La pendiente de la recta de regresión mínimo-cuadrática es

$$b = r\frac{s_y}{s_x}$$

Esta ecuación indica que, a lo largo de la recta de regresión, a **un cambio de una desviación típica de x le corresponde un cambio de r desviaciones típicas de y.** Cuando las variables están perfectamente correlacionadas ($r = 1$ o $r = -1$), el cambio en la respuesta predicha $\hat{y}$ es igual al cambio de x (expresado en desviaciones típicas). En los restantes casos, como $-1 \leq r \leq 1$, el cambio de $\hat{y}$ es menor que el cambio de x. A medida que la correlación es menos fuerte, la predicción $\hat{y}$ se mueve menos en respuesta a los cambios de x.

Carecterística 3. **La recta de regresión mínimo-cuadrática siempre pasa por el punto $(\bar{x}, \bar{y})$** del diagrama de dispersión de y con relación a x. Por tanto, la recta de regresión mínimo-cuadrática de y con relación a x es la recta de pendiente $r\frac{s_y}{s_x}$ que pasa a través del punto $(\bar{x}, \bar{y})$. Podemos describir completamente la regresión con $\bar{x}$, s_x, $\bar{y}$, s_y y r.

Característica 4. La correlación r describe la fuerza de la relación lineal. En este contexto se expresa de la siguiente manera: **el cuadrado de la correlación, r^2, es la fracción de la variación de las y que explica la recta de regresión mínimo-cuadrática de y con relación a x.**

La idea de la regresión es la siguiente: cuando existe una relación lineal, parte de la variación de y se explica por el hecho de que cuando x cambia, arrastra consigo a y. Mira otra vez la figura 2.10. Hay mucha variación en los valores observados de y, los datos de consumo de gas. Los valores de y toman valores que van de 3 a 31. El diagrama de dispersión muestra que la mayor parte de la variación de y se explica por la variación de la temperatura exterior (medida en grados-día x) que arrastra consigo el consumo de gas. Sólo existe una pequeña variación residual de y que aparece en la dispersión de los puntos a lo largo de la recta. Por otro lado, los puntos de la figura 2.13 están mucho más dispersos. La dependencia lineal de la velocidad con relación a la distancia explica sólo una parte de la variación observada en la velocidad. Podrías adivinar, por ejemplo, que cuando $x = 2$ el valor de y será mayor que cuando $x = 0$. De todas formas, existe todavía una variación considerable de y cuando x se mantiene fija —mira los cuatro puntos de la figura 2.13 cuando $x = 2$—. Esta idea se puede expresar algebraicamente, aunque no lo haremos. Es posible dividir la variación total de los valores observados de y en dos partes. Una de ellas es la variación que esperamos obtener de $\hat{y}$ a medida que x se mueve a lo largo de la recta de regresión. La otra mide la variación de los datos con relación a la recta. El cuadrado de la correlación r^2 es el primero de estos dos componentes expresado como fracción de la variación total.

$$r^2 = \frac{\text{variación de } \hat{y} \text{ junto con } x}{\text{variación total de las } y \text{ observadas}}$$

EJEMPLO 2.11. Utilización de r^2

En la figura 2.10, $r = 0{,}9953$ y $r^2 = 0{,}9906$. Más del 99% de la variación del consumo de gas se explica por la relación lineal con los grados-día. En la figura 2.13, $r = 0{,}7842$ y $r^2 = 0{,}6150$. La relación lineal entre la distancia y la velocidad explica el 61,5% de la variación de las dos variables. Hay dos rectas de regresión, pero existe sólo una correlación y r^2 ayuda a interpretar ambas regresiones. ■

Cuando presentes los resultados de una regresión, da el valor de r^2 como una medida de lo buena que es la respuesta que proporciona la regresión. Todos los resultados de programas estadísticos de la figura 2.12 incluyen r^2, en tanto por uno o en tanto por ciento. Cuando tengas una correlación, elévala al cuadrado para tener una idea más precisa de la fuerza de la asociación. Una correlación perfecta ($r = -1$ o $r = 1$) significa que los puntos se hallan perfectamente alineados a lo largo de una recta. En este caso $r^2 = 1$, es decir, toda la variación de una variable se explica por la relación lineal con la otra variable. Si $r = -0{,}7$ o $r = 0{,}7$, entonces $r^2 = 0{,}49$. Es decir, aproximadamente la mitad de la variación se explica con la relación lineal. En la escala de la r^2, una correlación de $r = \pm 0{,}7$ se halla a medio camino entre 0 y ± 1.

Las características anteriores son propiedades especiales de la regresión mínimo-cuadrática. No son ciertas para otros métodos de ajuste de una recta a unos datos. Otra razón por la cual el método de los mínimos cuadrados es el más común para ajustar una recta de regresión a unos datos es que tiene muchas propiedades interesantes.

APLICA TUS CONOCIMIENTOS

2.33. El profesor Moore y la natación. He aquí los tiempos (en minutos) que tarda el profesor Moore en nadar 1.800 metros y su ritmo cardíaco después de bracear (en pulsaciones por minuto) en 23 sesiones de natación.

Minutos	34,12	35,72	34,72	34,05	34,13	35,72	36,17	35,57
Pulsaciones	152	124	140	152	146	128	136	144
Minutos	35,37	35,57	35,43	36,05	34,85	34,70	34,75	33,93
Pulsaciones	148	144	136	124	148	144	140	156
Minutos	34,60	34,00	34,35	35,62	35,68	35,28	35,97	
Pulsaciones	136	148	148	132	124	132	139	

(a) Un diagrama de dispersión muestra una relación lineal negativa relativamente fuerte. Utiliza tu calculadora o un programa informático para comprobar que la recta de regresión mínimo-cuadrática es

$$\text{pulsaciones} = 479{,}9 - (9{,}695 \times \text{minutos})$$

(b) Al siguiente día el profesor tardó 34,30 minutos. Predice su ritmo cardíaco. En realidad su pulso fue 152. ¿Cómo de exacta es tu predicción?

(c) Supón que sólo conociéramos que las pulsaciones fueron 152. Ahora quieres predecir el tiempo que el profesor estuvo nadando. Halla la recta de regresión mínimo-cuadrática apropiada para la ocasión. ¿Cuál es tu predicción? ¿Es muy exacta?

(d) Explica de forma clara, a alguien que no sepa estadística, por qué las dos rectas de regresión son distintas.

2.34. Predicción del comportamiento de mercados de valores. Algunas personas creen que el comportamiento de un mercado de valores en enero permite predecir el comportamiento del mercado durante el resto del año. Toma como variable explicativa x el porcentaje de cambio en el índice del mercado de valores en enero y como variable respuesta y la variación del índice a lo largo de todo el año. Creemos que existe una correlación positiva entre x e y, ya que el cambio de enero contribuye al cambio anual. Cálculos a partir de datos del periodo 1960-1997 dan

$$\bar{x} = 1{,}75\% \qquad s_x = 5{,}36\% \qquad r = 0{,}596$$
$$\bar{y} = 9{,}07\% \qquad s_y = 15{,}35\%$$

(a) ¿Qué porcentaje de la variación observada en los cambios anuales del índice se explica a partir de la relación lineal con el cambio del índice en enero?

(b) ¿Cuál es la ecuación de la recta mínimo-cuadrática para la predicción del cambio en todo el año a partir del cambio en enero?

(c) En enero el cambio medio es $\bar{x} = 1{,}75\%$. Utiliza tu recta de regresión para predecir el cambio del índice en un año para el cual en enero sube un 1,75%. ¿Por qué podías haber conocido este resultado (hasta donde te permite el error de redondeo) sin necesidad de hacer ningún cálculo?

2.35. Castores y larvas de coleóptero. A menudo los ecólogos hallan relaciones sorprendentes en nuestro entorno. Un estudio parece mostrar que los castores pueden ser beneficiosos para una determinada especie de coleóptero. Los investigadores establecieron 23 parcelas circulares, cada una de ellas de 4 metros de diámetro, en una zona en la que los castores provocaban la caída de álamos al

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.