Análisis de los residuos

Se ha visto anteriormente que la correlación entre dos variables puede ser alta a pesar de que la relación entre las dos sea fuertemente **no lineal**.

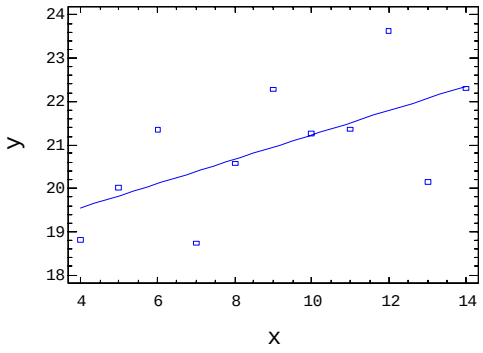
Se pueden utilizar los residuos para ver si el modelo de regresión lineal es adecuado.

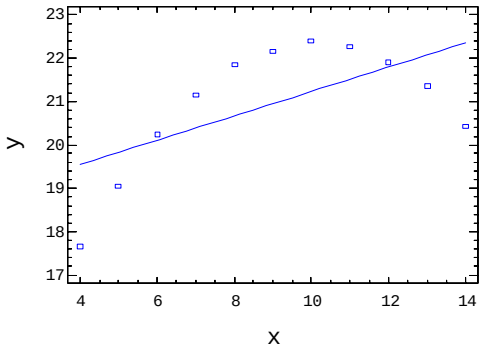
Casi siempre es útil hacer gráficos de los residuos (frente x, y o $\widehat{y}$) para ver si los supuestos del modelo lineal de regresión son justificados o no.

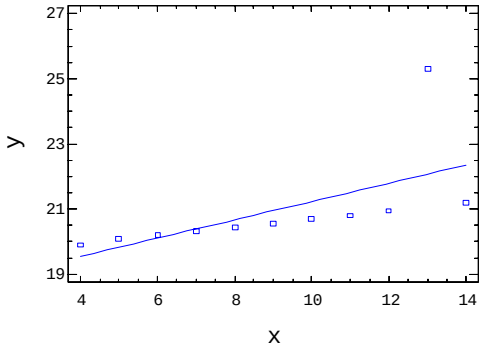
Ejemplo 83 *La recta de regresión para los cinco siguientes conjuntos de datos es la misma:*

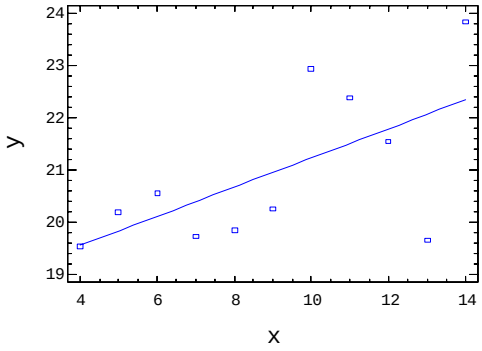
$$y = 18{,}43 + 0{,}28 * x$$

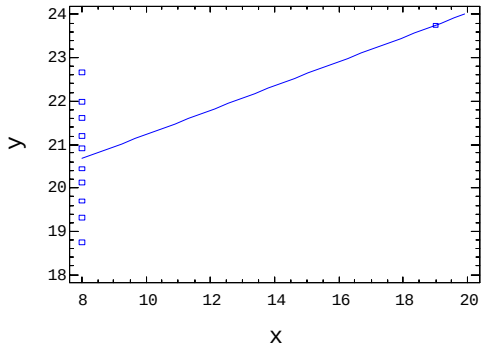
Bassett, E. et al (1986). *Statistics: Problems and Solutions*. London: Edward Arnold


y
x
24
23
22
21
20
19
18
4
6
8
10
12
14


y
x
23
22
21
20
19
18
17
4
6
8
10
12
14


y
x
27
25
23
21
19
4
6
8
10
12
14


y
x
24
23
22
21
20
19
4
6
8
10
12
14


y
x
24
23
22
21
20
19
18
8
10
12
14
16
18
20

- *El primer caso parece una regresión normal.*

- *En el segundo caso, hay una relación no lineal.*

- *En el tercer gráfico, se ve la influencia de un dato atípico.*

- *En el cuarto gráfico parece que la recta está más cerca a los datos cuando x es más pequeño.*

- *En el últimо caso, se ve el efecto de un punto influyente.*

Ahora hacemos gráficos de los residuos frente a las predicciones.

Gráfico de predicciones frente a residuos

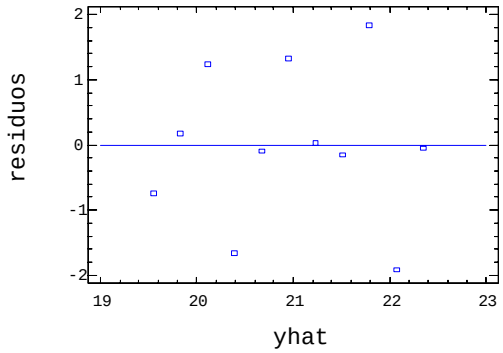


Gráfico de predicciones frente a residuos

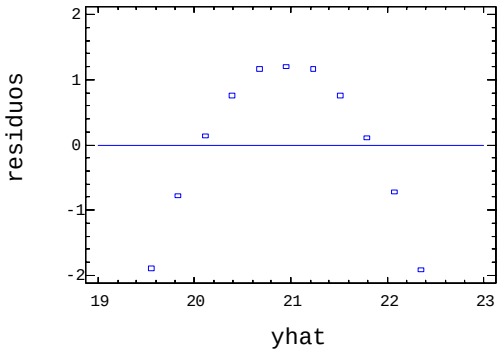


Gráfico de predicciones frente a residuos

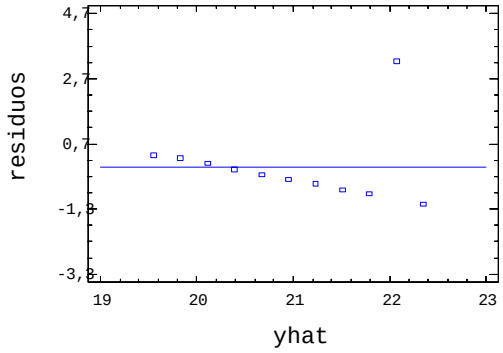


Gráfico de predicciones frente a residuos

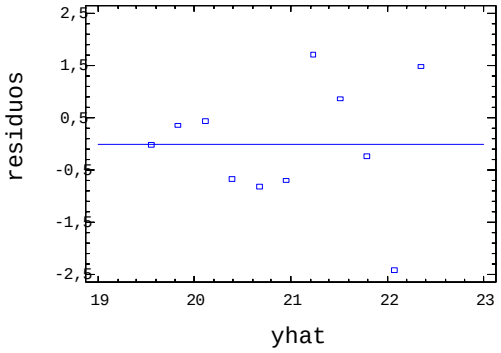
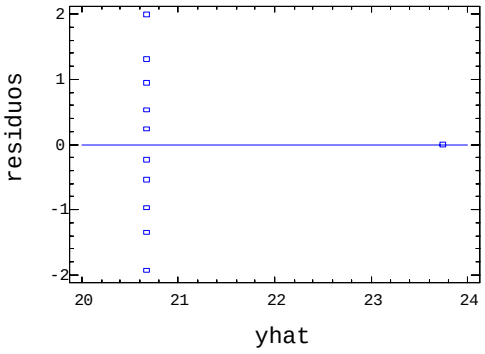


Gráfico de predicciones frente a residuos



- *En el primer caso, los residuos parecen aleatorios. Es una buena indicación que el modelo de regresión se ajusta bien.*

- *En el segundo caso, se ve una relación entre $\widehat{y}$ y los residuos. El modelo lineal no se ajusta bien.*

- *Cuando haya un dato atípico, se ve un residuo muy alto.*

- *Los residuos son más pequeños cuando $\widehat{y}$ es pequeño.*

- *Se ve el efecto del dato influyente.*

Dos rectas de regresión

Hasta ahora, hemos pensado en un modelo

$$y = \alpha + \beta x + \epsilon$$

y dada la muestra, hemos usado mínimos cuadrados para ajustar la rectas

$$y = a + bx$$

con $b = \frac{s_{xy}}{s_x^2}$ y $a = \bar{y} - b\bar{x}$.

Podríamos escribir el modelo de otra manera:

$$x = \gamma + \delta y + \nu$$

donde $\delta = \frac{1}{\beta}$, $\gamma = -\frac{\alpha}{\beta}$ y $\nu = -\frac{\epsilon}{\beta}$.

No obstante, si usamos mínimos cuadrados para ajustar la recta $x = c + dy$ a los datos muestrales tendrémos

$$d = \frac{s_{xy}}{s_y^2} \quad \text{y} \quad c = \bar{x} - d\bar{y}.$$

Observamos que $d \neq \frac{1}{b}$. ¡Las rectas no son iguales!

Ejemplo 84 *Volvemos al Ejemplo 72 sobre extensión (y) relativa a la fuerza (x) aplicada al muelle.*

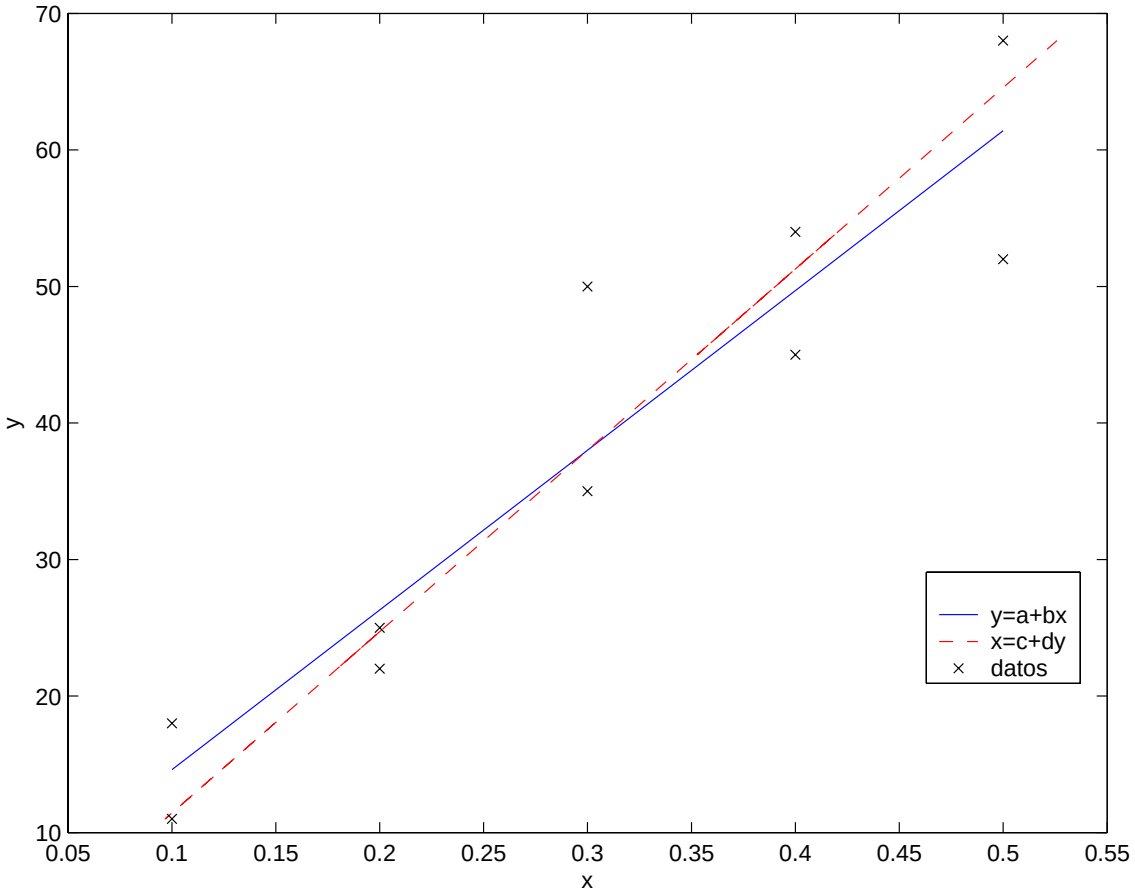
Antes hemos visto que ajustando la recta $y = a + bx$ por mínimos cuadrados, se tiene

$$y = 2{,}9 + 117x.$$

Ahora supongamos que queremos predecir la fuerza x que causaría una extensión de y. Ajustando la recta por mínimos cuadrados, tenemos

$$x = {,}0139 + {,}0075y.$$

El ajuste de ambas rectas aparece en el siguiente gráfico.



Para hacer regresión es importante saber cuales son las variables dependientes y independientes.

Correlación espuria

Si el coeficiente de correlación entre dos variables es alta, indica que están relacionadas entre si. No obstante, no permite concluir una relación *causal*.

Ejemplo 85 *Se ha descubierto que por más coches de bomberos que van al fuego, más es el daño causado. Pero el trabajo de los bomberos es extinguir los fuegos y reducir el daño y entonces el resultado es sorprendente.*

?Cómo podemos explicar el resultado?

Si el fuego es más intenso, entonces van más coches de bomberos y también el fuego causa más daño.

La alta correlación entre número de coches y daño es espuria, o sea debida al efecto de otra variable (intensidad del fuego) que influye a ambas.

Ejemplo 86 *De Los Simpsons*©.

Homer: No veo ningún oso. La patrulla de búsqueda de osos debe estar funcionando de maravilla.

Lisa: Papa eres idiota.

Homer: Gracias Lisa

Lisa: Siguiendo tu lógica, esta roca impede la presencia de tigres.

Homer: ?Cómo funciona?

Lisa: No funciona.

Homer: Uh-huh.

Lisa: Sólo es una estupida roca pero no veo ningún tigre. ?Y tu?

Homer: Lisa, quiero comprar tu roca.

La paradoja de Simpson

El siguiente ejemplo ilustra la paradoja.

Ejemplo 87 *La enfermedad de Grot es muy peligroso y puede llegar a ser fatal. De momento, no existe ningún tratamiento reconocido pero a muchas personas, les gusta usar la panacea de Blogg, un remedio natural. Se hace un estudio de un grupo de sufridores de la enfermedad con los siguientes resultados.*

	Nada	*Blogg*
Sobrevive	108	153
Muere	123	120

Parece que el tratamiento funciona, ya que $\frac{153}{153+120} = 56\,\%$ *de los pacientes tomando la panacea han sobrevivido mientras sólo* $\frac{108}{108+123} = 47\,\%$ *de las pacientes sin tratamiento se han recuperado.*

Ejemplo tomado de: `http://www.cawtech.freeserve.co.uk/simpsons.2.html`

Entonces, parece ser buena idea recetar la panacea a los pacientes.

No obstante, cuando se informa el colectivo de mujeres Grot sobre la decisión a recetar la panacea, ellas no están nada contentas porque ellas han visto los resultados sólo para las mujeres en la muestra.

	Nada	*Blogg*
Sobrevive	57	32
Muere	100	57

Sólo un $\frac{32}{32+57} = 36{,}0\,\%$ *de las mujeres tomando Blogg han sobrevivido mientras un* $36{,}3\,\%$ *de las que no toman Blogg han sobrevivido. Parece perjudicar (un poquito) a las mujeres tomar Blogg.*

?El Blogg debe ser un tratamiento machista que favorece la salud de los hombres?

Sacamos la tabla con respeto a los hombres.

	Nada	*Blogg*
Sobrevive	51	121
Muere	23	63

Un $\frac{121}{121+63} = 65{,}7\,\%$ *de los hombres bajo tratamiento han sobrevivido mientras un* $\frac{51}{51+23} = 68{,}9\,\%$ *de los hombres sin tratamiento han sobrevivido. Parece que el tratamiento perjudica a los hombres también.*

El resultado es paradojico. El tratamiento parece favorecer la población entera de pacientes pero no favorece ni a las mujeres ni a los hombres.

La paradoja de Simpson demuestra que si mezclamos datos de dos subpoblaciones bastante distíntos, podemos llegar a conclusiones opuestas a las obtenidas tratando los grupos por separados.

Pregunta de examen sobre el tema 3

Ejemplo 88 *(Examen de junio 2004)*

Una heladería quiere hacer un estudio sobre la cantidad de helados que vende. Para ello se han tomado al azar 10 semanas del año y se ha observado la temperatura media correspondiente a cada una de ellas, así como la cantidad de helados vendidos en cada uno de estos periodos, obteniendo los siguientes resultados:

temp ^oC	10	28	12	31	30	19	24	5	9	15
# *helados vendidos*	21	65	19	72	75	39	67	11	12	24

a) *Hallar la ecuación de regresión. Interpretar sus coeficientes, es decir la pendiente y la ordenada en el origen.* *(1 punto)*

b) *Supongamos que perdemos los datos de la variable dependiente y que tenemos todos los residuos menos el correspondiente a la primera semana.¿Cuál sera su valor?*

Hallar la desviación típica residual.

residuos	r_1	$-1{,}91$	$-4{,}34$	$-3{,}08$	$2{,}64$	$-3{,}40$	$10{,}98$	$6{,}71$	$-3{,}17$	$-7{,}51$

(1 punto)

- *¿Cuál será la cantidad de helados vendidos en una semana cuya temperatura sea de 23 grados?* *(0,5 puntos)*