

Ejercicio resuelto 8.1

La siguiente tabla muestra la distribución del gasto mensual en libros y el gasto mensual en audiovisual en euros en los últimos 8 años de una biblioteca.

Gasto en audiovisual	Gasto en libros		
	0–200	200–400	400–600
0–200	1	15	3
200–400	3	37	7
400–600	1	13	16

- ¿Cómo es la distribución del gasto en libros?
- ¿En qué apartado fue mayor el desembolso de la biblioteca en estos años, en libros o en audiovisual?
- ¿Cómo se distribuye el gasto en libros si se tienen en cuenta únicamente los meses en los que se gastan más de 200€ en audiovisual?, compara los resultados con los del apartado a).

Solución:

El **objetivo del Apartado a)** es describir la distribución del gasto en libros.

Planteamiento: el experimento consiste en seleccionar meses (individuos) y observar el gasto en libros (variable 1) y el gasto en audiovisual (variable 2). La población son todos los meses y la muestra son los meses de los últimos 8 años. Los datos son numéricos, el 0 significa que no hay gasto y, en principio, cualquier gasto (145, 420,5, etc.) es posible, luego la variable es cardinal, de razón y continua, aunque no se tienen datos aislados, sino datos agrupados y frecuencias.

Método y justificación: para describir la distribución se necesita saber la frecuencia de cada clase del gasto en libros, independientemente de cuál haya sido el gasto en audiovisual. En la tabla del enunciado se observa que con un gasto aproximado de 100€ (es decir, en la primera clase), hay 1 mes en el que el gasto en audio fue de 100 aproximadamente, más 3 meses en los que fue de 300, más 1 mes que fue de aproximadamente 500, es decir, un total de 5 meses (la suma de las frecuencias de la columna). De la misma forma se pueden calcular las frecuencias del resto de modalidades de la tabla. Estas frecuencias se pueden anotar en el margen de la tabla, o bien se pueden extraer a una tabla aparte para manejarlas igual que se hacía anteriormente.

Cálculos: la Tabla 8.1 contiene la distribución conjunta de ambas variables y las marginales de cada una por separado. La Tabla 8.2 representa la distribución de frecuencias unidimensional de la variable gasto en libros.

Gasto en audiovisual	Gasto en libros			Total
	0-200	200-400	400-600	
0-200	1	15	3	19
200-400	3	37	7	47
400-600	1	13	16	30
Total	5	65	26	96

Tabla 8.1: Tabla de doble entrada con marginales.

clases	n_i	f_i
0-200	5	0,0521
200-400	65	0,6771
400-600	26	0,2708
Total	96	1

Tabla 8.2: Tabla de frecuencias del gasto en libros.

Conclusión: el gasto más frecuente de los 96 meses de la muestra se sitúa alrededor de 300€ y hay más meses con gastos altos (alrededor de 500) que con gastos bajos (alrededor de 100).

El **objetivo del Apartado b)** es saber en qué apartado hubo un desembolso mayor. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: se necesita calcular el gasto total (total o suma de la variable) de libros y en audio de esos 8 años. Para calcular el total, se calculan las marcas de clase, se multiplican por las frecuencias marginales y se realiza la suma. Los gastos totales se pueden comparar mediante el cociente.

Cálculos: para el gasto en libros (ver Tabla 8.2) se tiene

$$x_T = 100 \times 5 + 300 \times 65 + 500 \times 26 = 33000,$$

análogamente para el gasto en audiovisual se obtiene que $y_T = 31000$. Comparando con el cociente se tiene que $33000/31000 = 1,065$.

Conclusión: el desembolso total en libros fue aproximadamente de 33000€ y en audiovisual de 31000€, lo que significa que el gasto en libros fue un 6,5 % mayor (aprox.) que en audiovisual.

El **objetivo del Apartado c)** es conocer la distribución del gasto en libros pero considerando *únicamente* los meses en los que el gasto en audio superó los 200€. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: hay que *seleccionar o filtrar* los meses en los que se cumple esa condición y calcular la distribución marginal de los datos seleccionados.

Cálculos: la Tabla 8.3 contiene la distribución conjunta de ambas variables eliminando los datos que no cumplen la condición y las marginales condicionadas de cada una por separado. La Tabla 8.4 representa la distribución de frecuencias unidimensional de la variable gasto en libros condicionada a que el gasto en audio sea mayor de 200.

Gasto en audiovisual	Gasto en libros			Total
	0–200	200–400	400–600	
200–400	3	37	7	47
400–600	1	13	16	30
Total	4	50	23	77

Tabla 8.3: Tabla de doble entrada con condicionadas.

clases	n_i	f_i
0–200	4	0,0519
200–400	50	0,6494
400–600	23	0,2987
Total	77	

Tabla 8.4: Tabla de frecuencias condicionada del gasto en libros.

Conclusión: 77 de los 96 meses tienen un gasto en audio de más de 200€. Si se consideran sólo esos 77 meses, se observa que la frecuencia (relativa) de gastos bajos (aproximadamente de 100) se mantiene prácticamente igual (ver Tabla 8.2). Los gastos moderados (alrededor de 300) son ligeramente menos frecuentes, mientras que los altos (alrededor de 500) son también algo más frecuentes.

Ejercicio resuelto 8.2

La siguiente tabla muestra la cuantía de ciertas subvenciones concedidas en las comunidades de Galicia y Asturias

Comunidades	Cuantía de las subvenciones	
	6000	12000
Galicia	21	14
Asturias	9	6

Calcula las distribuciones de frecuencias relativas de la cuantía de las subvenciones para cada comunidad, ¿qué deduces de ellas?

Solución:

El **objetivo** es calcular e interpretar las distribuciones de frecuencias relativas de la cuantía de las subvenciones para cada comunidad.

Planteamiento: el experimento consiste en seleccionar subvenciones (individuos) y observar su cuantía (variable 1) y la comunidad (variable 2) en la que se conceden. La población son todas las subvenciones y se tiene una muestra de 50 subvenciones. La variable 1 es cardinal, de razón y continua porque sus posibles valores son números, en principio cualquier valor sería posible, y el 0 representa ausencia de subvención. La segunda variable es nominal, porque sus posibles valores son nombres que no se pueden ordenar. Los datos se muestran en una tabla de doble entrada.

Método y justificación: se necesitan las distribuciones de las cuantías *condicionadas* a cada comunidad. Las distribuciones absolutas se observan directamente en la tabla de doble entrada, ya que cada una se corresponde con la fila correspondiente a cada comunidad, pero como hay distinto número de subvenciones en Asturias y en Galicia, para valorar la distribución es mejor calcular las frecuencias relativas, que es el objetivo.

Cálculos: las Tablas 8.5 y 8.6 contienen las distribuciones de frecuencias unidimensionales de la cuantía de las subvenciones para Asturias y Galicia, respectivamente.

Conclusión: en Asturias se concedieron 15 subvenciones y en Galicia 35. Tanto en Galicia como en Asturias el 60 % de las subvenciones fueron de 6000€ y el 40 % de 12000. La distribución relativa de las cuantías es la misma para Galicia que para Asturias, en proporción de 60/40. Es decir, *independientemente* de la comunidad la distribución de las cuantías es la misma. Saber si una subvención fue de 6000 o de 12000 no ayuda para

Cuantía	n_i	f_i
6000	9	0,6
12000	6	0,4
Total	15	

Tabla 8.5: Tabla de frecuencias de la cuantía para Asturias.

Cuantía	n_i	f_i
6000	21	0,6
12000	14	0,4
Total	35	

Tabla 8.6: Tabla de frecuencias de la cuantía para Galicia.

saber si es más probable que fuera de Asturias o de Galicia. Eso significa que la cuantía de la subvención es *estadísticamente independiente* de la comunidad a la que se le asigne. Si, por ejemplo, la distribución de subvenciones en Galicia fuera de 90 % de 12000 frente a 10 % de 6000, saber la cuantía de la subvención sí que ayudaría a saber si es más probable que fuese de Asturias o de Galicia (sería más probable que fuese de Galicia si fuese de 12000).

Ejercicio resuelto 8.3

Los siguientes datos muestran la cantidad de papel y plástico (en kg) que reciclaron varias familias de cierta ciudad a lo largo de un año.

Papel	134	45	15	85	103	94	27	57	71	46	89	94	54	67
Plástico	175	47	0	96	121	111	29	47	66	47	115	115	61	75
Papel	36	136	112	59	104	81	42	93	64	102	31	51	128	137
Plástico	39	168	143	65	121	89	47	109	75	121	21	55	161	171

- Gráficamente, ¿se observa algún tipo de relación entre la cantidad de plástico y de papel que recicla cada familia?
- ¿Se podría aproximar la relación de ambas variables mediante alguna fórmula?
- ¿Se podría predecir cuánto plástico reciclará aproximadamente una familia que recicló 55 kg de papel?, ¿y de una familia que no recicló nada de papel?, ¿es fiable la predicción?

Solución:

El **objetivo del Apartado a)** es determinar si gráficamente, se observa algún tipo de relación entre la cantidad de plástico y de papel que recicla cada familia.

Planteamiento: el experimento consiste en seleccionar familias (individuos) y observar la cantidad de papel (variable 1) y la cantidad de plástico (variable 2) que reciclan en un año. La población la constituyen todas las familias de esa ciudad y se tiene una muestra de 28 familias. Las variables son cardinales, continuas porque son números y, en principio, podrían tomar cualquier valor. El 0 significa ausencia de cantidad reciclada, luego son de razón. Se conocen los datos apareados para cada familia de la muestra.

Método y justificación: para ver gráficamente si existe relación entre la cantidad de plástico y papel reciclada hay que representar los pares de datos. De esta forma se podrá comprobar si al aumentar el valor de una de las variables tiende a aumentar o disminuir el valor de la otra, lo que indicaría que hay algún tipo de relación. Para decidir qué variable se va representar en el eje de las equis y cuál en el eje de las íes, hay que mirar si existe alguna relación causa-efecto lógica (en cuyo caso X sería la causa e Y el efecto) o bien si el objetivo posterior es predecir o aproximar una de las variables (Y) en función de la otra (X). En este

caso como posteriormente se trata de aproximar la cantidad de plástico en función de la cantidad de papel, Y será la primera y X la segunda de dichas variables.

Cálculos: en la Figura 8.3 se representa la nube de puntos que se buscaba. Cada punto representa el dato en X y en Y de una familia. Se muestra además una recta a la que más se parecen los datos.

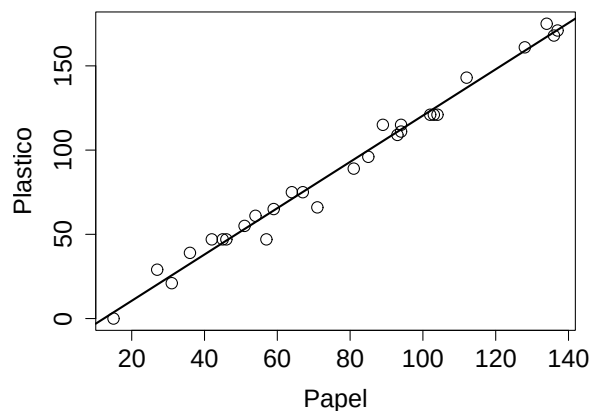


Figura 8.3: Nube de puntos.

Conclusión: en la Figura 8.3 se observa que si la cantidad de papel reciclada es pequeña, también lo es la de plástico y según aumenta la cantidad de papel, también aumenta la de plástico, luego existe relación entre ambas cantidades. Aunque la nube de puntos no determina una recta exacta, sí que se parece mucho. Este va a ser el origen de la regresión lineal, donde se buscarán las fórmulas de las funciones matemáticas que más se parecen a la nube de puntos.

El **objetivo del Apartado b)** es aproximar la relación de la cantidad de papel y plástico reciclada por cada familia mediante alguna fórmula. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: en el Apartado a) se comprobó que la nube de puntos seguía la forma de una recta, por lo que se buscará la fórmula de dicha recta, que es precisamente la de regresión. Esa recta será la que aproxime la relación entre X e Y , ya que los valores de Y son prácticamente iguales a los que proporciona la fórmula de la línea de regresión.

Cálculos: aplicando las fórmulas anteriores o con ayuda de un programa estadístico se obtiene que $\hat{b}_1 = 1,374$ y $\hat{b}_0 = -16,942$, por lo que la recta de regresión es

$$\hat{y}(x) = 1,374x - 16,942,$$

que es precisamente la recta representada en la Figura 8.3.

Conclusión: la recta $\hat{y}(x) = 1,374x - 16,942$ (ver Figura 8.3) es la que da valores más próximos a los y_i muestrales (en media) para cada valor de x . Es decir $\hat{b}_1 = 1,374$ y $\hat{b}_0 = -16,942$ son los valores que hacen que los pares $(x_i, \hat{b}_1x_i + \hat{b}_0)$ estén tan cerca a los verdaderos valores (x_i, y_i) como es posible (dentro de todos los pares que se podrían obtener con rectas). Así, dado un x , la fórmula $\hat{y}(x)$ da la mejor relación lineal para aproximar y . Como $\hat{b}_1 = 1,374$, eso significa que una familia que recicla 1 kg más de papel que otra, recicla en media aproximadamente 1,374 kg. más de plástico.

El **objetivo del Apartado c)** es precedir cuánto plástico reciclará aproximadamente una familia que recicló 55 kg de papel y otra familia que no recicló nada de papel. También se pide determinar si es fiable la predicción. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: para realizar la predicción se necesita la fórmula de la mejor aproximación de Y a partir de X . Ya se ha visto que la relación era lineal (ver Figura 8.3) y se ha hallado la fórmula de la regresión lineal (ver Apartado b)). Así, sólo hay que sustituir los valores de x en la recta de regresión, siempre y cuando el valor de x esté en el rango de valores (mínimo-máximo) para el que se calculó dicha recta. Esto es así porque fuera de ese rango no hay evidencias de que la relación sea la misma y se podrían hacer predicciones que no tengan sentido.

El valor de papel reciclado de la primera familia, 55, sí está en el rango 15-137, pero no el de la segunda. De hecho, si se intentase aplicar la regresión con la segunda, nos saldría un valor predicho de $\hat{y}(0) = -16,942$, es decir, una cantidad negativa, lo que no tiene sentido.

Cálculos: sustituyendo 55 en la recta de regresión queda una predicción de

$$\hat{y}(55) = 1,374 \cdot 55 - 16,942 = 58,63 \text{ kg.}$$

Conclusión: se predice que las familias que reciclan 55 kg de papel reciclan en media unos 58,63 de papel, así que lo que recicla una familia concreta será un valor que esté alrededor de 58,63 kg.

La fiabilidad de la predicción depende de lo cerca que estén en general los datos reales de los predichos por la recta. En este caso se ve que los datos

están muy cerca de la recta, por lo que se espera que la aproximación sea bastante buena. Para cuantificar objetivamente cómo de fiable es una recta de regresión se necesita analizar la correlación.

Parte del **objetivo del Apartado c)** era determinar si las predicciones eran fiables. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: anteriormente se indicó que gráficamente se veía que la regresión era fiable para hacer predicciones, porque los datos reales estaban cerca de la recta. Para cuantificar esta percepción visual hay que calcular el coeficiente de determinación, que indica que porcentaje de la variabilidad de Y se explica por su relación lineal con la variable X .

Cálculos: utilizando la fórmula anterior o con ayuda de un programa estadístico se obtiene que $R^2 = 0,9849$.

Conclusión: el 98,49 % de la variabilidad de la cantidad de plástico reciclada se explica por la cantidad de papel que se recicla mediante la recta de regresión. Es decir, casi toda la variabilidad de Y se produce sobre la recta y hay poquita variación que se desvíe de dicha recta. Como los errores se cometerían si la cantidad de papel se alejase mucho de la recta (que es lo que se predice), se puede decir que la predicción es bastante fiable.

Ejercicio resuelto 8.4

Se ha estudiado el tiempo en minutos que dedica a consultas telefónicas y el tiempo que dedica a consultas personales cada uno de los empleados de una Gestoría. Los datos obtenidos se encuentran en las siguientes tabla:

Telefónicas	99	157	141	135	35	41	138	269	147	138	137
Personales	200	123	142	132	441	373	136	88	143	159	164
Telefónicas	40	57	227	209	199	160	101	261	142	159	251
Personales	389	309	104	106	120	154	167	118	163	165	108
Telefónicas	82	255	101	130	145	138	206	94			
Personales	215	81	199	169	135	162	148	171			

- a) ¿Crees que hay una buena relación funcional entre el tiempo empleado en consultas telefónicas y el tiempo empleado en consultas personales?, ¿de qué tipo?
- b) Si un empleado dedica a consultas telefónicas 2 horas, ¿cuánto tiempo se espera que dedique a consultas personales?

Solución:

El **objetivo del Apartado a)** es determinar si hay una buena relación funcional entre el tiempo empleado en consultas telefónicas y en personales y de qué tipo es la relación.

Planteamiento: el experimento consiste en seleccionar empleados (individuos) y observar el tiempo que dedicaron cierto día a consultas telefónicas (variable 1) y a consultas personales (variable 2). La población la constituyen todos los empleados y hay una muestra de 30. Las variables son cardinales, de razón y continuas porque los posibles valores son números, el 0 significa que no dedican nada de tiempo y, en principio, cualquier valor es posible. Se tienen los datos apareados de toda la muestra.

Método y justificación: se trata de verificar si hay algún modelo (lineal o curvilíneo) bueno. Se considerarán los más comunes y se observará si alguno tiene un R^2 alto. En principio, se podría elegir como independiente cualquiera de las variables, pero en el Apartado b) se pretende hacer una predicción del tiempo dedicado a consultas personales **en función** del tiempo dedicado a consultas telefónicas, por lo que la variable independiente X será el tiempo dedicado a consultas telefónicas y la variable dependiente Y será el tiempo dedicado a consultas personales.

Cálculos: con ayuda de un programa estadístico se obtienen los valores $\hat{b}_0$ y $\hat{b}_1$ para cada modelo. En la Tabla 8.7 se recogen las fórmulas de cada modelo, los valores $\hat{b}_0$ y $\hat{b}_1$ para los datos de este problema y el R^2 . En las Figuras 8.4 y 8.5 se representan todas las regresiones calculadas junto con la nube de puntos.

Modelo	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	R^2
Lineal: $y(x) = b_0 + b_1x$	337,827	337,827			0,676
Logarítmica: $y(x) = b_0 + b_1 \ln(x)$	919,356	-152,783			0,896
Inversa: $y(x) = b_0 + b_1/x$	52,840	52,840			0,970
Cuadrático: $y(x) = b_0 + b_1x + b_2x^2$	505,475	3,770	3,770		0,911
Cúbico: $y(x) = b_0 + b_1x + b_2x^2 + b_3x^3$	654,529	-8,269	0,044	-0,00008	0,964
Potencial: $y(x) = b_0x^{b_1}$	5397,411	5397,411			0,928
Exponencial: $y(x) = b_0e^{b_1x}$	366,911	-0,006			0,815

Tabla 8.7: Regresiones curvilíneas.

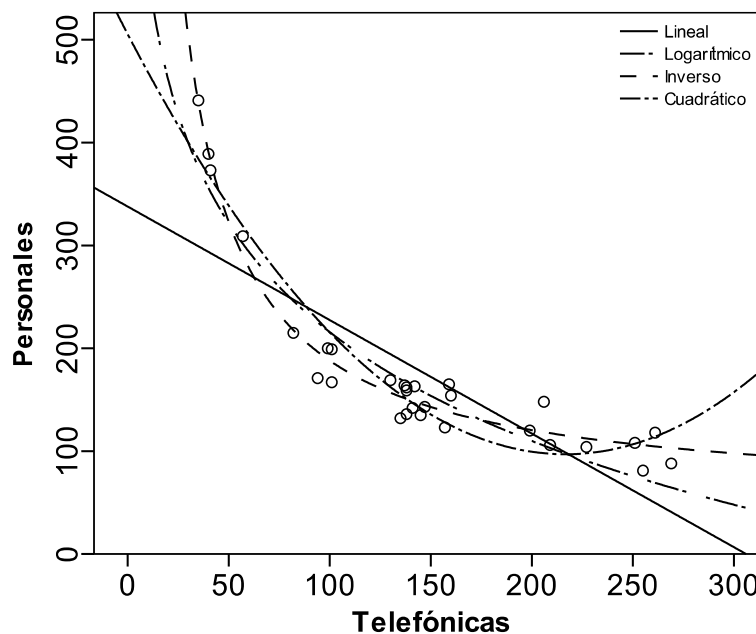


Figura 8.4: Regresiones curvilíneas (I).

Conclusión: según la Tabla 8.7 el mejor modelo es el inverso, que tiene dos parámetros y un coeficiente de determinación de 0,97, lo que significa que los datos se ajustan muy bien a ese modelo y que, por lo tanto, hay una buena relación de tipo inverso entre el tiempo dedicado a consultas telefónicas y el dedicado a personales.

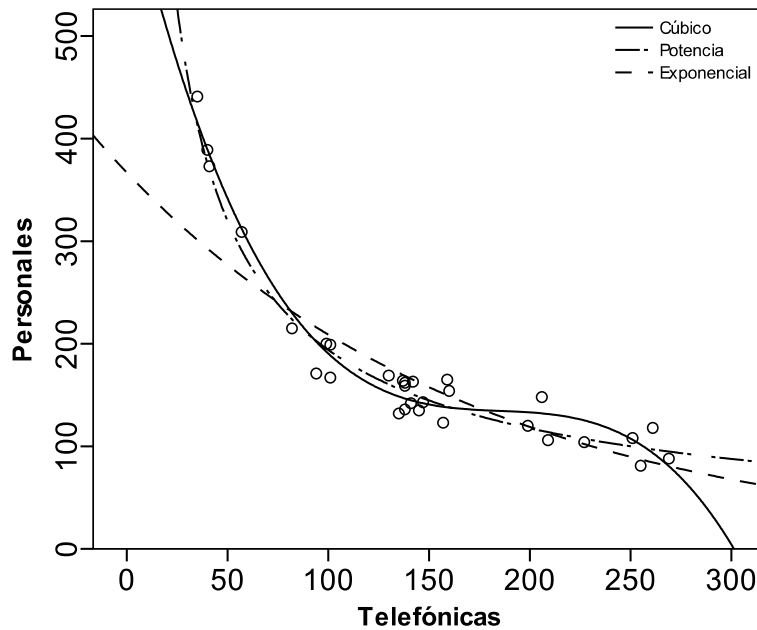


Figura 8.5: Regresiones curvilíneas (II).

La fórmula de la curva de regresión inversa es $\hat{y}(x) = 52,840 + 13516,735/x$. En la Figura 8.4 se aprecia que la nube de puntos varía poco alrededor de esa curva. En concreto puede decirse que el 97 % de la variabilidad del tiempo dedicado a consultas personales se explica por la relación o regresión inversa con el tiempo dedicado a consultas telefónicas.

El **objetivo del Apartado b)** es predecir cuánto tiempo se espera que dedique a consultas personales un empleado que dedica 2 horas a consultas telefónicas. El **planteamiento** es el mismo que el del Apartado a).

Método y justificación: para realizar la predicción se necesita la fórmula de la mejor aproximación de Y a partir de X . Ya se ha visto que la mejor relación era inversa y se ha hallado la fórmula de la regresión (ver Apartado a)). Sólo hay que sustituir los valores de x en la curva de regresión, si el valor de x está en el rango de valores (mínimo-máximo) para el que se calculó dicha curva (fuera de ese rango no hay evidencias de que la relación sea la misma y se podrían hacer predicciones que no tengan sentido).

El tiempo de consultas personales en minutos es de 120, que se encuentra entre el mínimo (que son 35 minutos) y el máximo (que son 269 minutos), luego se puede hacer la predicción.

Para cuantificar el grado de fiabilidad se necesita el coeficiente de determinación, que indica que porcentaje de la variabilidad de Y se explica por su relación con la variable X (ya calculado en el Apartado a)).

Cálculos: sustituyendo 120 en la curva de regresión queda una predicción de

$$\hat{y}(120) = 52,8398 + 13516,7/120 = 165,48 \text{ minutos.}$$

Conclusión: la previsión del tiempo de consultas personales que se espera que tenga un empleado que dedica 2 horas a consultas telefónicas es de unos 165,48 minutos. Esta previsión es una aproximación del valor real, no es exacta, pero es bastante fiable, ya que como se vió en el Apartado a), el 97% de la variabilidad del tiempo dedicado a consultas personales se explica por la relación o regresión inversa con el tiempo dedicado a consultas telefónicas. Hay muy poca variabilidad fuera de la curva, es decir, los valores reales oscilan muy cerca de los predichos por la curva.