TEMA 4
Modelo de regresión múltiple

José R. Berrendero
Departamento de Matemáticas
Universidad Autónoma de Madrid

Análisis de Datos - Grado en Biología

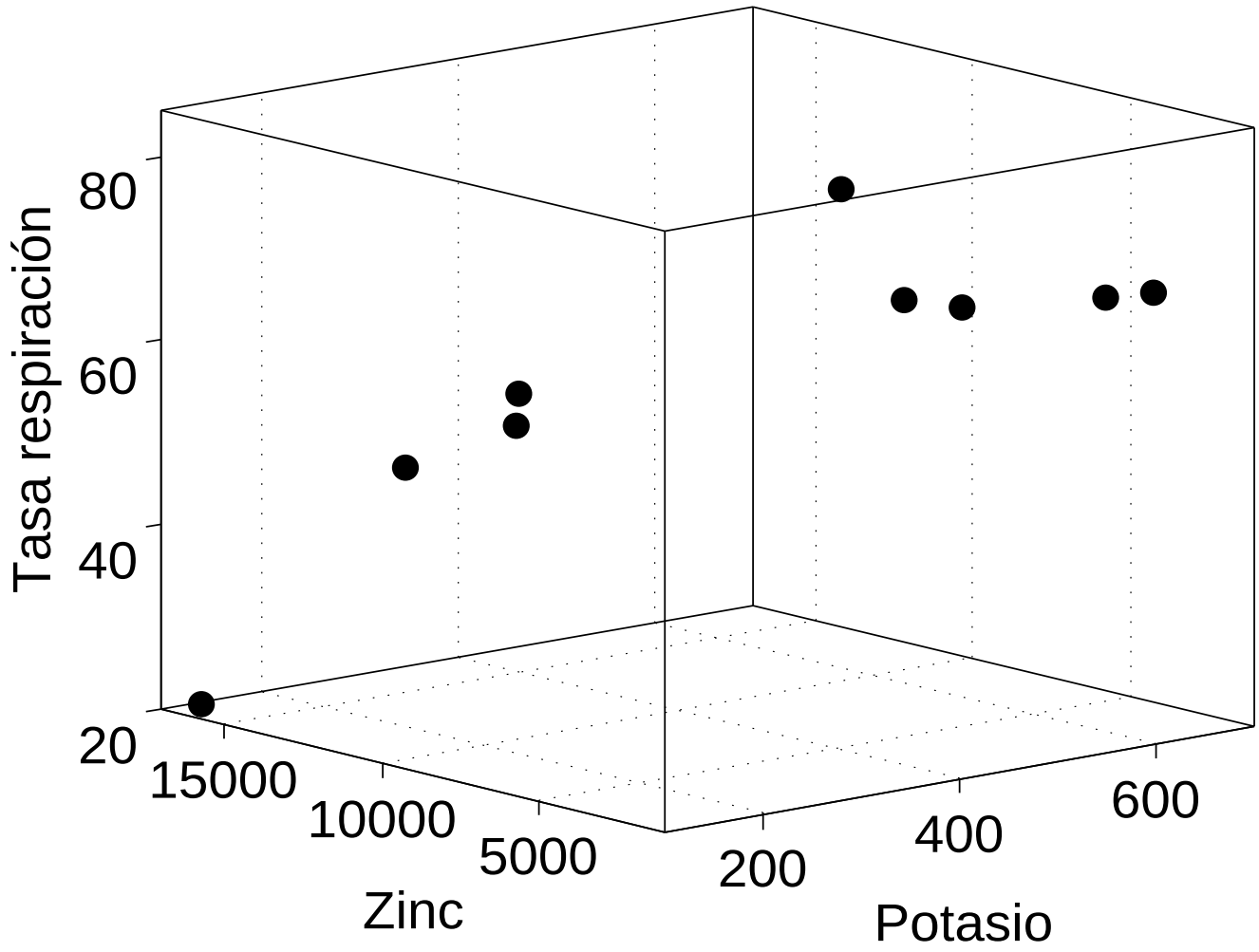
Estructura de este tema

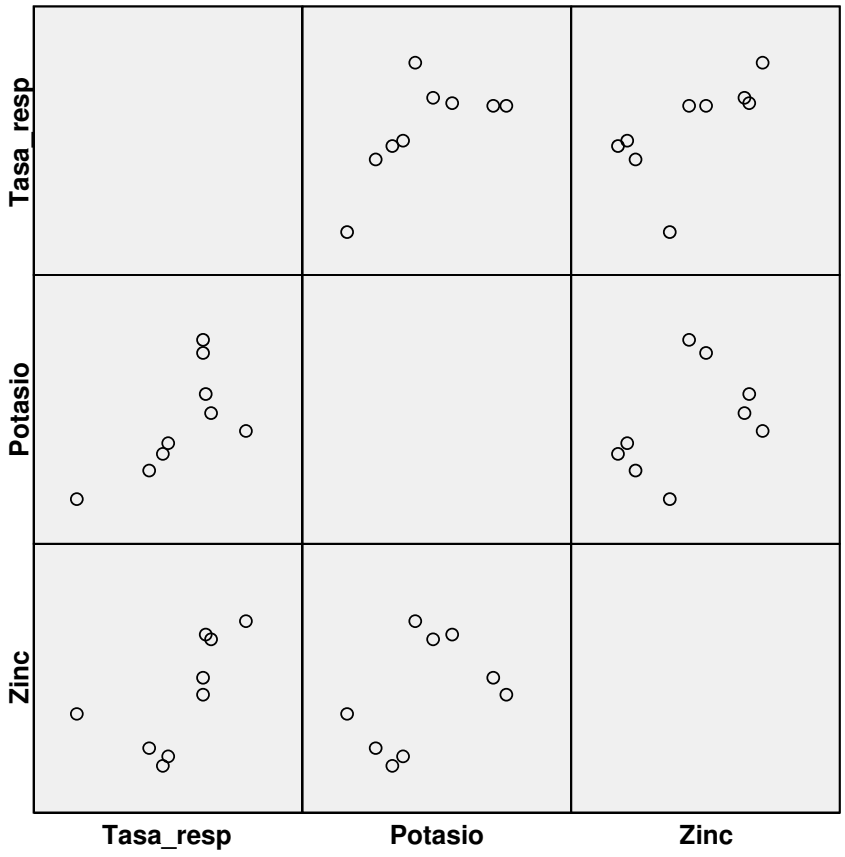
- Modelo de regresión múltiple. Ejemplos.
- Estimación e inferencia sobre los parámetros del modelo.
- Tabla ANOVA y contraste de la regresión.
- Regresión polinómica.
- Variables regresoras dicotómicas.
- Multicolinealidad.
- Diagnóstico del modelo.

Ejemplo

Se estudia Y = la tasa de respiración (moles $O_2/(g{\cdot}min)$) del liquen *Parmelia saxatilis* bajo puntos de goteo con un recubrimiento galvanizado. El agua que cae sobre el liquen contiene zinc y potasio, que utilizamos como variables explicativas. (Fuente de datos: Wainwright (1993), *J. Biol. Educ.*.)

Tasa de respiración	Potasio (ppm)	Zinc (ppm)
71	388	2414
53	258	10693
55	292	11682
48	205	12560
69	449	2464
84	331	2607
21	114	16205
68	580	2005
68	622	1825


Tasa respiración
80
60
40
20
15000
10000
5000
Zinc
200
400
600
Potasio



Correlaciones

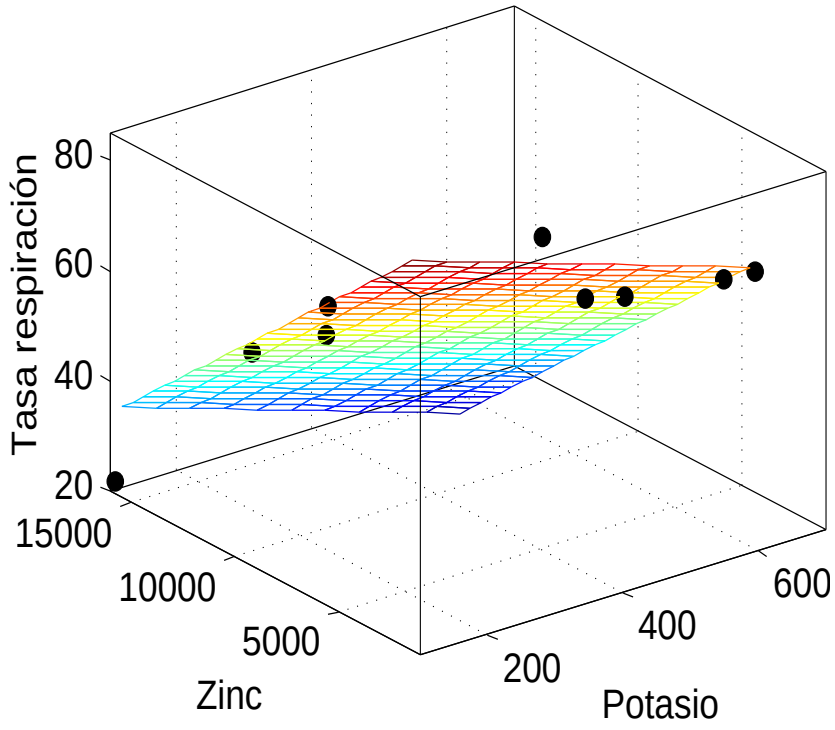
		Tasa_resp	Potasio	Zinc
Tasa_resp	Correlación de Pearson	1	,686	,653
	Sig. (bilateral)		,041	,057
	N	9	9	9
Potasio	Correlación de Pearson	,686	1	,443
	Sig. (bilateral)	,041		,232
	N	9	9	9
Zinc	Correlación de Pearson	,653	,443	1
	Sig. (bilateral)	,057	,232	
	N	9	9	9

Modelo de regresión lineal múltiple

En la regresión lineal múltiple de Y sobre $X_1, \ldots, X_k$ se supone que la función de regresión tiene la expresión

$$Y \approx \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Cuando $k = 2$ la función de regresión es un plano



Modelo de regresión lineal múltiple

Tenemos una muestra de n individuos en los que observamos las variables Y y $X_1, \ldots, X_k$. Para el individuo i, tenemos el vector de datos $(Y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$.

El modelo de regresión lineal múltiple supone que

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{ik} + u_i, \quad i = 1, \ldots, n,$$

donde las variables de error U_i verifican

a) u_i tiene media cero, para todo i.

b) $\mathrm{Var}(u_i) = \sigma^2$, para todo i (homocedasticidad).

c) Los errores son variables independientes.

d) u_i tiene distribución normal, para todo i.

e) $n \geq k + 2$ (hay más observaciones que parámetros).

f) Las variables X_i son linealmente independientes entre sí (no hay *colinealidad*).

Modelo de regresión lineal múltiple

Las hipótesis (a)-(d) se pueden reexpresar así: las observaciones Y_i son independientes entre con distribución normal:

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \sigma).$$

El modelo admite una expresión equivalente en forma matricial:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Estimación de los parámetros del modelo

Parámetros desconocidos: $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$.

Estimamos $\beta_0, \beta_1, \ldots, \beta_K$ por el método de mínimos cuadrados, es decir, los estimadores son los valores para los que se minimiza la suma:

$$\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2.$$

Cada coeficiente β_i mide el efecto que tiene sobre la respuesta un aumento de una unidad de la variable regresora x_i **cuando el resto de las variables permanece constante**.

Estimación de los parámetros del modelo

Al derivar la suma anterior respecto a $\beta_0, \beta_1, \ldots, \beta_k$ e igualar las derivadas a 0 obtenemos $k+1$ restricciones sobre los residuos:

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} e_i x_{i1} = 0, \quad \ldots, \quad \sum_{i=1}^{n} e_i x_{ik} = 0.$$

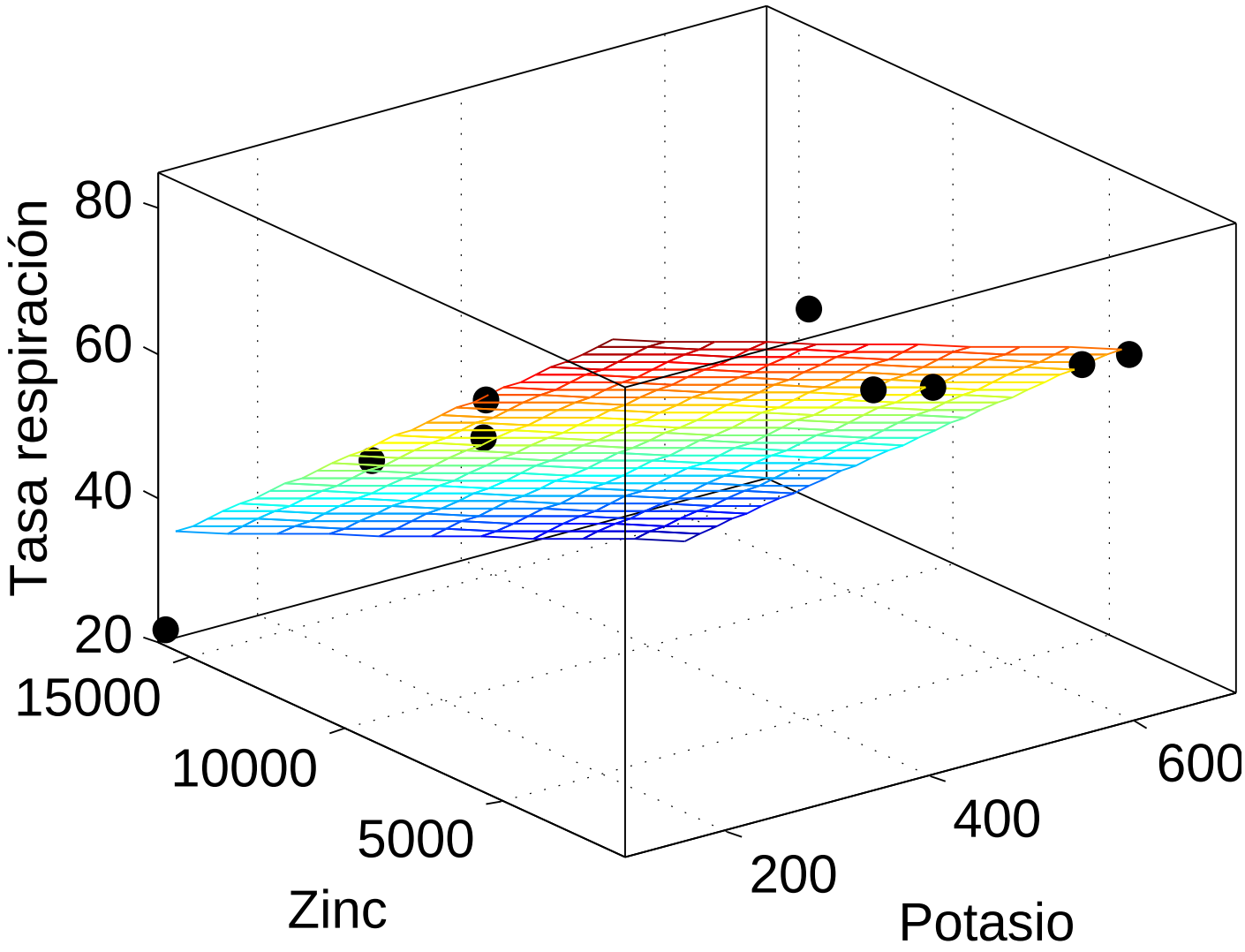
A partir de este sistema de $k+1$ ecuaciones es posible despejar los estimadores de mínimos cuadrados de $\beta_0, \beta_1, \ldots, \beta_k$.

Las hipótesis (e) y (f) hacen falta para que el sistema tenga una solución única. Llamamos $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ a los estimadores.

Le media de los residuos es cero. La correlación entre los residuos y cada una de las k variables regresoras es cero.

Los residuos tienen $n-k-1$ grados de libertad.

Estimación de los parámetros del modelo



Estimación de la varianza

Un estimador insesgado de σ^2 es la varianza residual S_R^2.

Como en los modelos anteriores, S_R^2 se define como la suma de los residuos al cuadrado, corregida por los gl apropiados:

$$S_R^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2.$$

Siempre se verifica $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \ldots + \hat{\beta}_k \bar{x}_k$, siendo

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1}, \quad \ldots, \quad \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}.$$

Por ejemplo, si $k = 2$, el plano de regresión pasa por el punto de medias muestrales $(\bar{x}_1, \bar{x}_2, \bar{y})$.

Inferencia sobre los parámetros del modelo

Distribución de los estimadores de los coeficientes:

Todos los estimadores $\hat{\beta}_j$ verifican:

$$\frac{\hat{\beta}_j - \beta_j}{\text{error típico de } \hat{\beta}_j} \equiv t_{n-k-1},$$

donde el error típico de $\hat{\beta}_j$ es un valor que se calcula con SPSS.

Intervalos de confianza para los coeficientes:

Para cualquier $j = 0, 1, \ldots, k$,

$$\text{IC}_{1-\alpha}(\beta_j) = \left(\hat{\beta}_j \mp t_{n-k-1;\alpha/2} \times \text{error típico de } \hat{\beta}_j\right).$$

Contrastes de hipótesis individuales sobre los coeficientes

Estamos interesados en determinar qué variables X_j son significativas para explicar Y.

$$H_0 : \beta_j = 0 \qquad (X_j \text{ no influye sobre } Y)$$
$$H_1 : \beta_j \neq 0 \qquad (X_j \text{ influye sobre } Y)$$

La región crítica de cada H_0 al nivel de significación α es

$$R = \left\{ \frac{|\beta_j|}{\text{error típico de } \hat{\beta}_j} > t_{n-k-1;\alpha/2} \right\}.$$

El cociente $\hat{\beta}_j/(\text{error típico de } \hat{\beta}_j)$ se llama estadístico t asociado a β_j.

Salida SPSS

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,789[a]	,622	,496	12,907

a. Variables predictoras: (Constante), Zinc, Potasio

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1644,390	2	822,195	4,935	,054[a]
	Residual	999,610	6	166,602		
	Total	2644,000	8			

a. Variables predictoras: (Constante), Zinc, Potasio

b. Variable dependiente: Tasa_resp

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	15,978	15,304		1,044	,337
	Potasio	,053	,030	,494	1,763	,128
	Zinc	,013	,009	,434	1,549	,172

a. Variable dependiente: Tasa_resp

Descomposición de la variabilidad

Como en modelos anteriores:

$$
\begin{aligned}
Y_i &= \hat{Y}_i + e_i \\
Y_i - \bar{Y} &= (\hat{Y}_i - \bar{Y}) + e_i \\
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2 \\
\text{SCT} &= \text{SCE} + \text{SCR}
\end{aligned}
$$

SCT mide la variabilidad total (tiene $n-1$ gl)

SCE mide la variabilidad explicada por el modelo (tiene k gl)

SCR mide la variabilidad no explicada o residual (tiene $n-k-1$ gl)

El contraste de la regresión

$H_0 : \beta_1 = \ldots = \beta_k = 0$ (el modelo no es explicativo: **ninguna de las variables explicativas influye en la respuesta**)

$H_1 : \beta_j \neq 0$ para algún $j = 1, \ldots, k$ (el modelo es explicativo: **al menos una de las variables X_j influye en la respuesta**)

Comparamos la variabilidad explicada con la no explicada mediante el estadístico F:

$$F = \frac{\text{SCE}/k}{\text{SCR}/(n-k-1)}.$$

Bajo H_0 el estadístico F sigue una distribución $F_{k,n-k-1}$.

La región de rechazo de H_0 al nivel de significación α es

$$R = \{F > F_{k,n-k-1;\alpha}\}$$

El coeficiente de determinación

Es una medida de la bondad del ajuste en el modelo de regresión múltiple

$$R^2 = \frac{\text{SCE}}{\text{SCT}}.$$

Propiedades:

- $0 \leq R^2 \leq 1$.
- Cuando $R^2 = 1$ existe una relación exacta entre la respuesta y las k variables regresoras.
- Cuando $R^2 = 0$, sucede que $\hat{\beta}_0 = \bar{y}$ y $\hat{\beta}_1 = \ldots = \hat{\beta}_k = 0$. No existe relación lineal entre Y y las X_i.
- Podemos interpretar R^2 o como un **coeficiente de correlación múltiple** entre Y y las k variables regresoras.
- Se verifica que $F = \dfrac{R^2}{1 - R^2} \dfrac{n - k - 1}{k}$.

El coeficiente de determinación ajustado

El coeficiente de determinación para comparar distintos modelos de regresión entre sí tiene el siguiente inconveniente:

Siempre que se añade una nueva variable regresora al modelo, R^2 aumenta, aunque el efecto de la variable regresora sobre la respuesta no sea significativo.

Por ello se define el *coeficiente de determinación ajustado o corregido por grados de libertad*

$$\bar{R}^2 = 1 - \frac{\text{SCE}/(n-k-1)}{\text{SCT}/(n-1)} = 1 - \frac{S_R^2}{\text{SCT}/(n-1)}$$

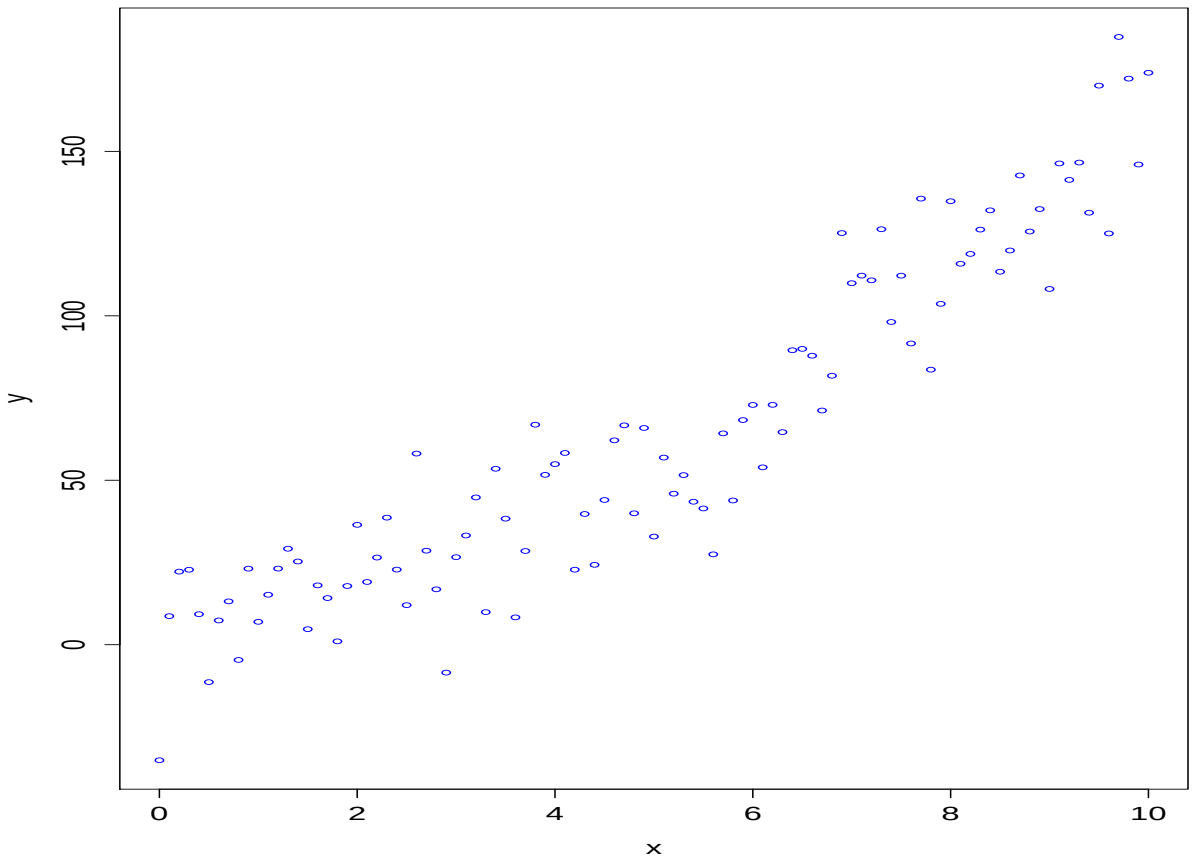
$\bar{R}^2$ sólo disminuye al introducir una nueva variable en el modelo si la varianza residual disminuye.

Regresión polinómica

Podemos utilizar el modelo de regresión múltiple para ajustar un polinomio:

$$Y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k.$$

Basta considerar las k variables regresoras $x, x^2, \ldots, x^k$.



Regresión polinómica

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,926[a]	,858	,857	19,04222

a. Variables predictoras: (Constante), x

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-14,376	3,762		-3,822	,000
	x	15,904	,650	,926	24,472	,000

a. Variable dependiente: y

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,947[a]	,896	,894	16,36427

a. Variables predictoras: (Constante), x2, x

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	6,846	4,790		1,429	,156
	x	3,042	2,214	,177	1,374	,172
	x2	1,286	,214	,774	6,004	,000

a. Variable dependiente: y

Regresión polinómica

Estimación curvilínea

Resumen del modelo y estimaciones de los parámetros

Variable dependiente:y

Ecuación	Resumen del modelo				
	R cuadrado	F	gl1	gl2	Sig.
Lineal	,858	598,866	1	99	,000
Cuadrático	,896	423,481	2	98	,000

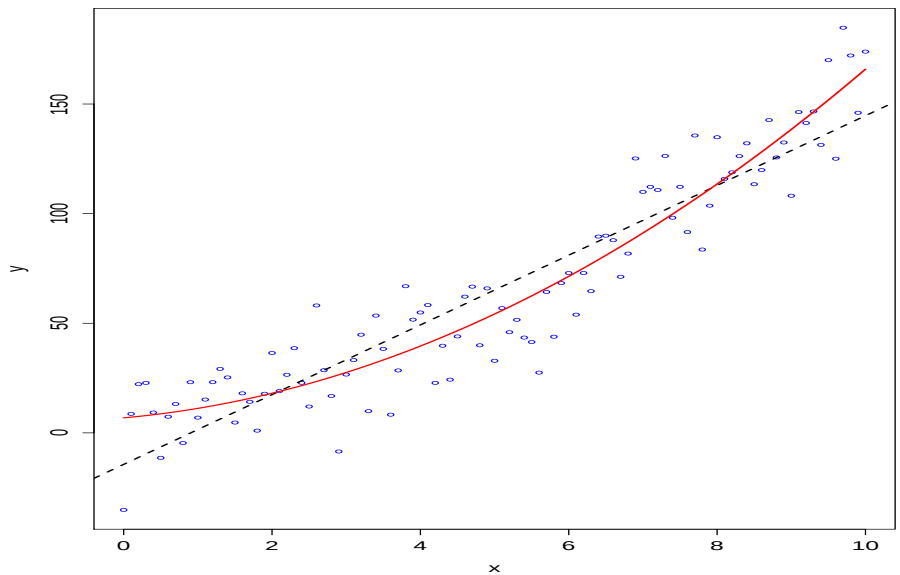
La variable independiente esx.

Resumen del modelo y estimaciones de los parámetros

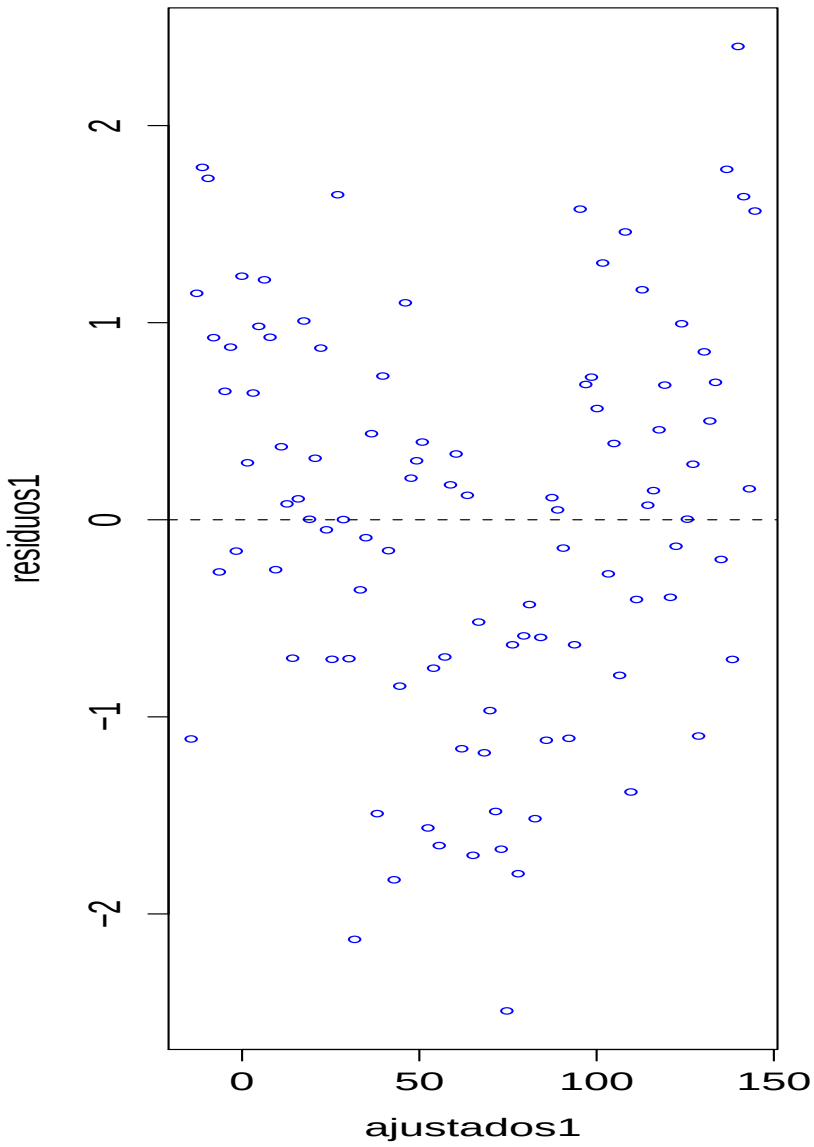
Variable dependiente:y

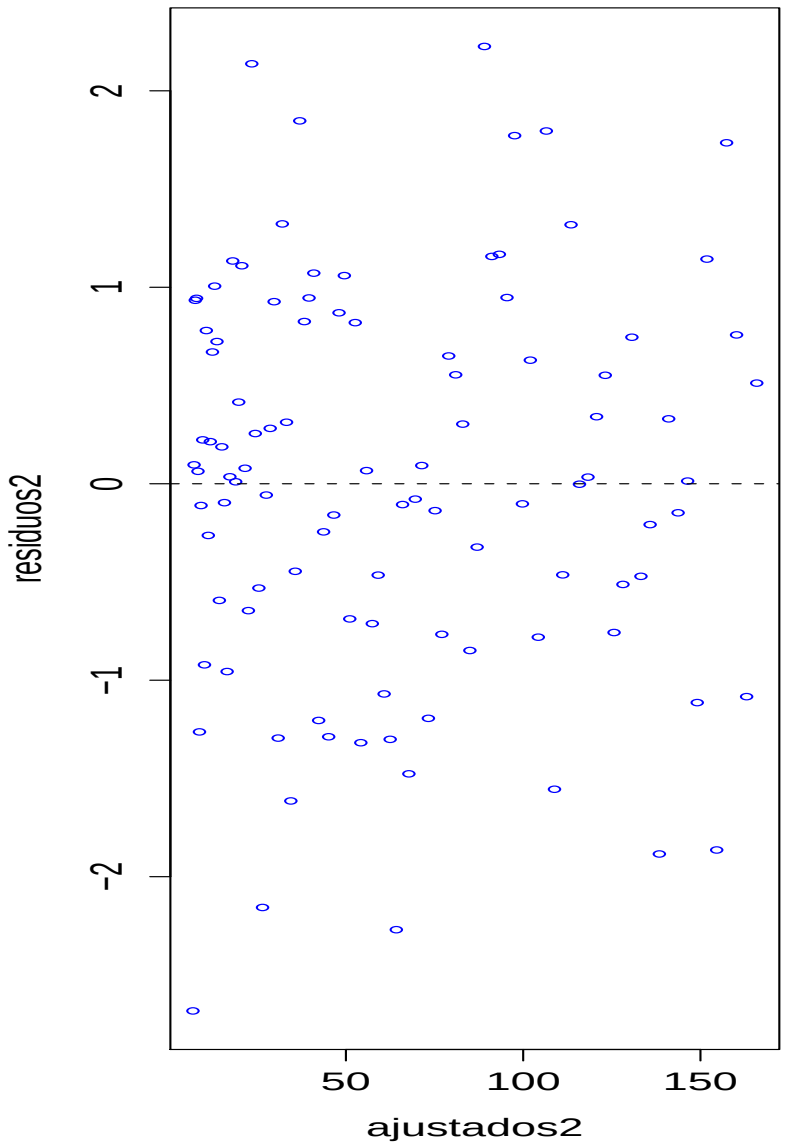
Ecuación	Estimaciones de los parámetros		
	Constante	b1	b2
Lineal	-14,376	15,904	
Cuadrático	6,846	3,042	1,286

La variable independiente esx.



Regresión polinómica





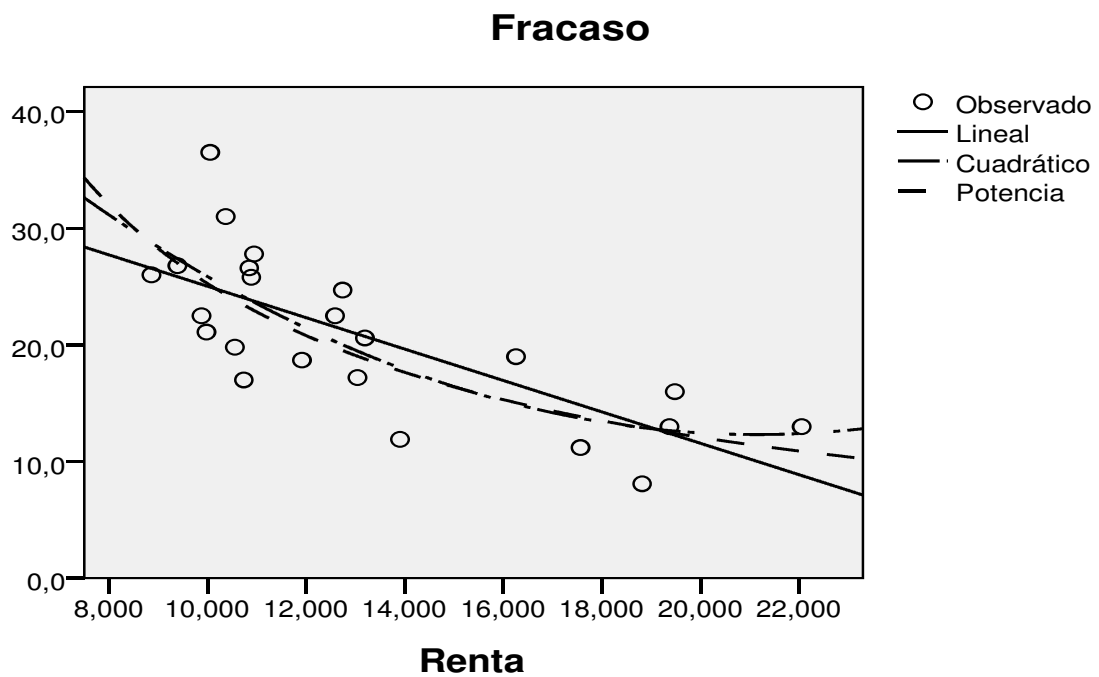
Regresión polinómica: rentas y fracaso escolar

Resumen del modelo y estimaciones de los parámetros

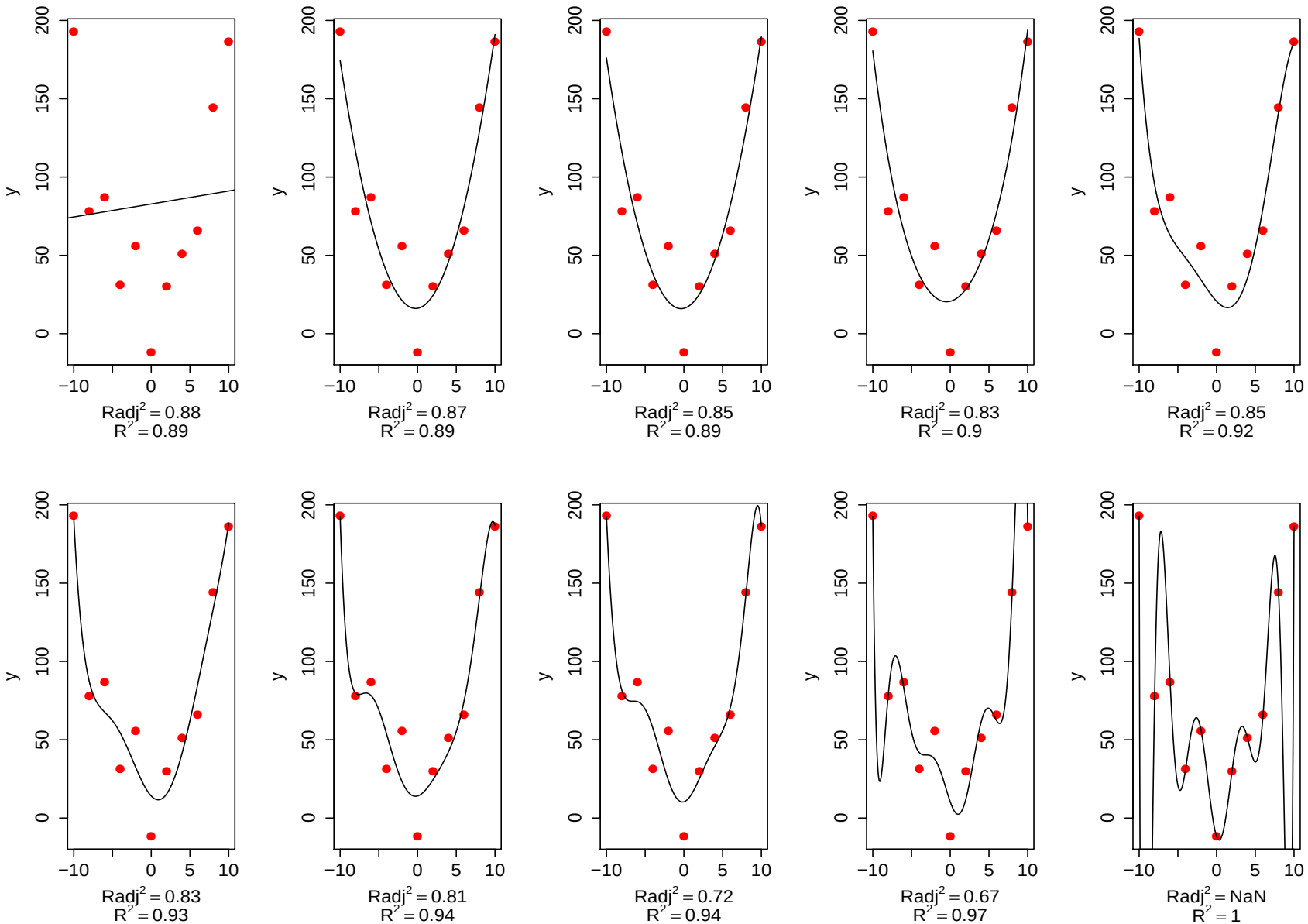
Variable dependiente:Fracaso

Ecuación	Resumen del modelo					Estimaciones de los parámetros		
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1	b2
Lineal	,550	25,658	1	21	,000	38,494	-1,347	
Cuadrático	,586	14,183	2	20	,000	61,088	-4,614	,109
Potencia	,610	32,809	1	21	,000	293,923	-1,066	

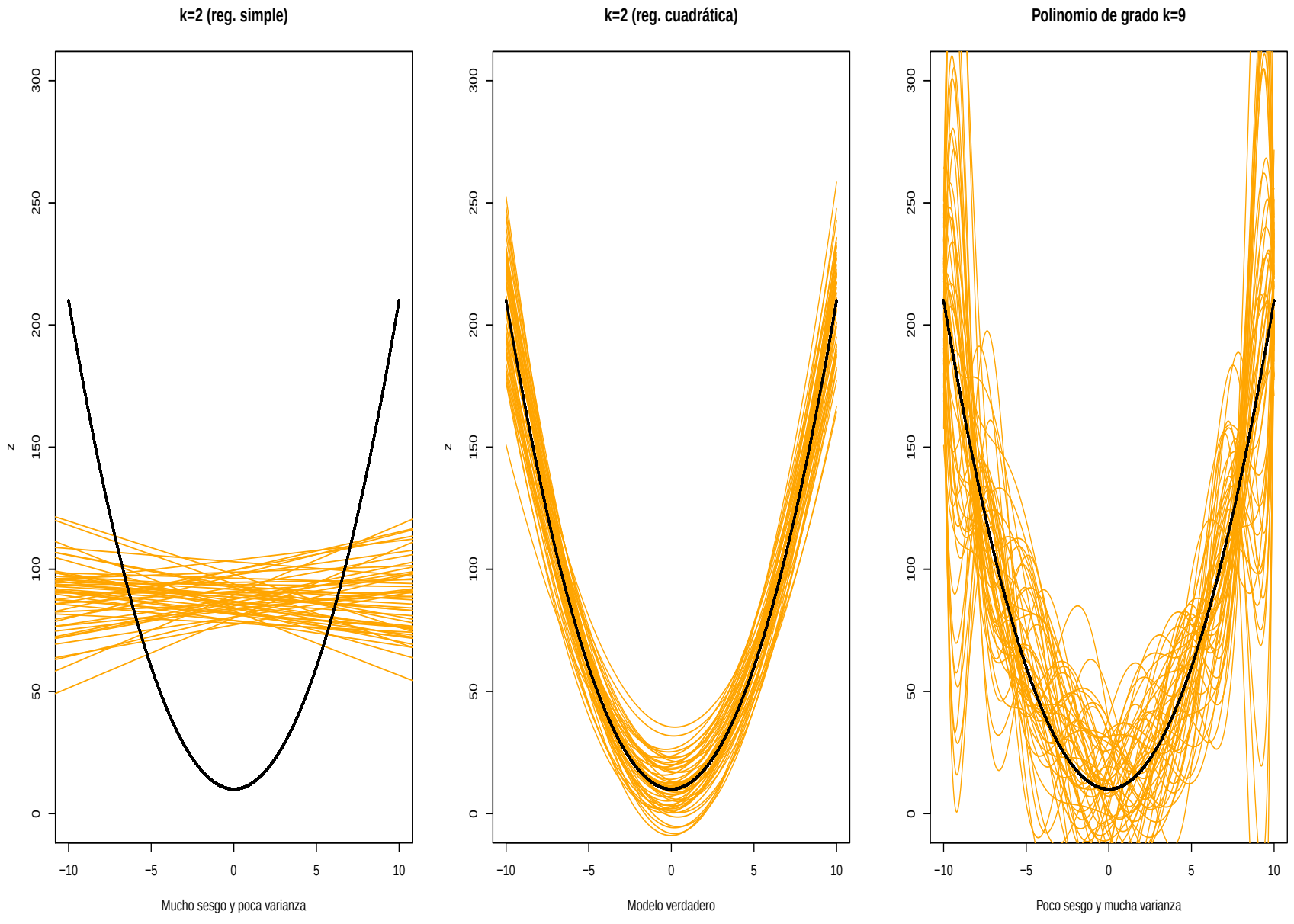
La variable independiente esRenta.



Regresión polinómica y sobreajuste

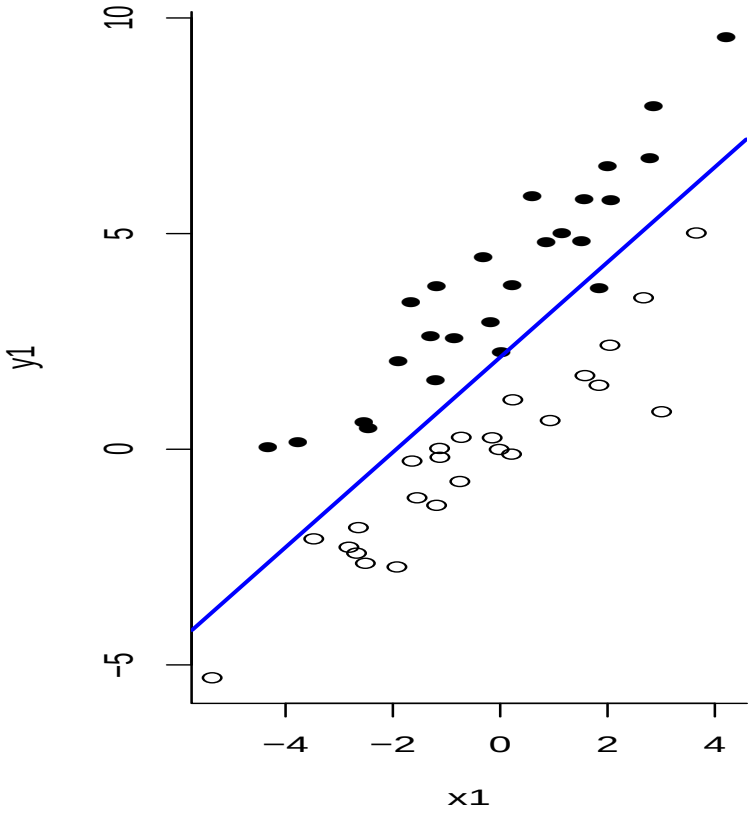


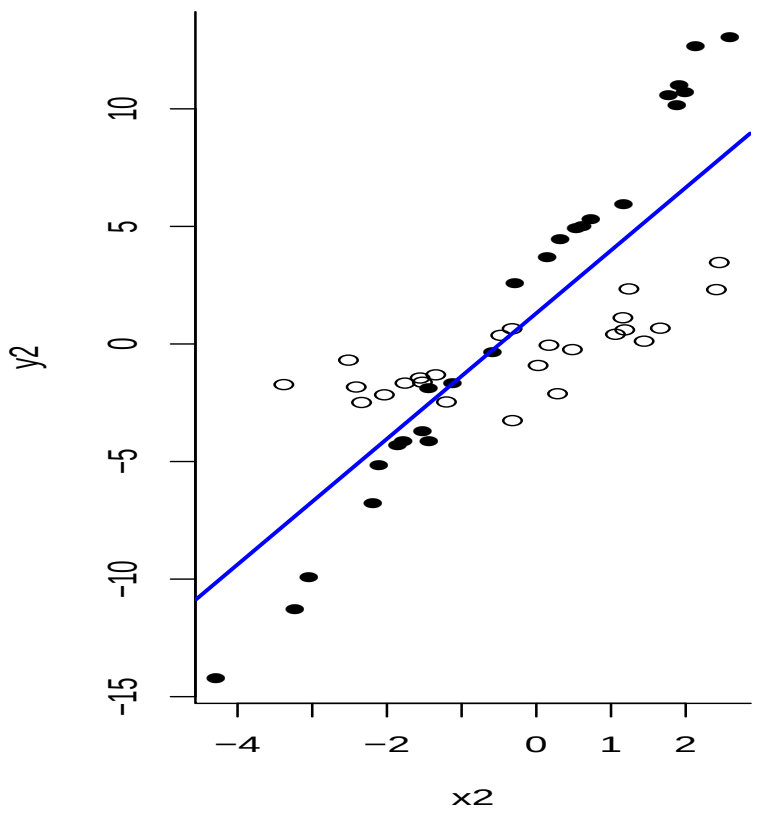
Curvas estimadas a partir de 50 muestras de 10 datos



Variables regresoras dicotómicas

Mezclar subpoblaciones en regresión no es adecuado.





¿En qué se diferencian los dos ejemplos anteriores?

Modelo aditivo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,963[a]	,928	,923	…

a. Variables predictoras: (Constante), x1z1, z1, x1

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	438,063	3	146,021	197,319	,000[a]
	Residual	34,041	46	,740		
	Total	472,104	49			

a. Variables predictoras: (Constante), x1z1, z1, x1

b. Variable dependiente: y1

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	,277	,177		1,560	,126
	x1	,927	,080	,647	11,632	,000
	z1	3,620	,247	,589	14,649	,000
	x1z1	,142	,114	,068	1,241	,221

a. Variable dependiente: y1

Modelo con interacciones

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,987[a]	,975	,973	…

a. Variables predictoras: (Constante), x2z2, z2, x2

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1533,096	3	511,032	593,559	,000[a]
	Residual	39,604	46	,861		
	Total	1572,700	49			

a. Variables predictoras: (Constante), x2z2, z2, x2

b. Variable dependiente: y2

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-,235	,189		-1,243	,220
	x2	,796	,115	,247	6,902	,000
	z2	3,025	,267	,270	11,320	,000
	x2z2	3,288	,152	,781	21,599	,000

a. Variable dependiente: y2

Multicolinealidad

El cálculo de los estimadores de los parámetros en regresión múltiple requiere resolver un sistema de $k+1$ ecuaciones con $k+1$ incógnitas.

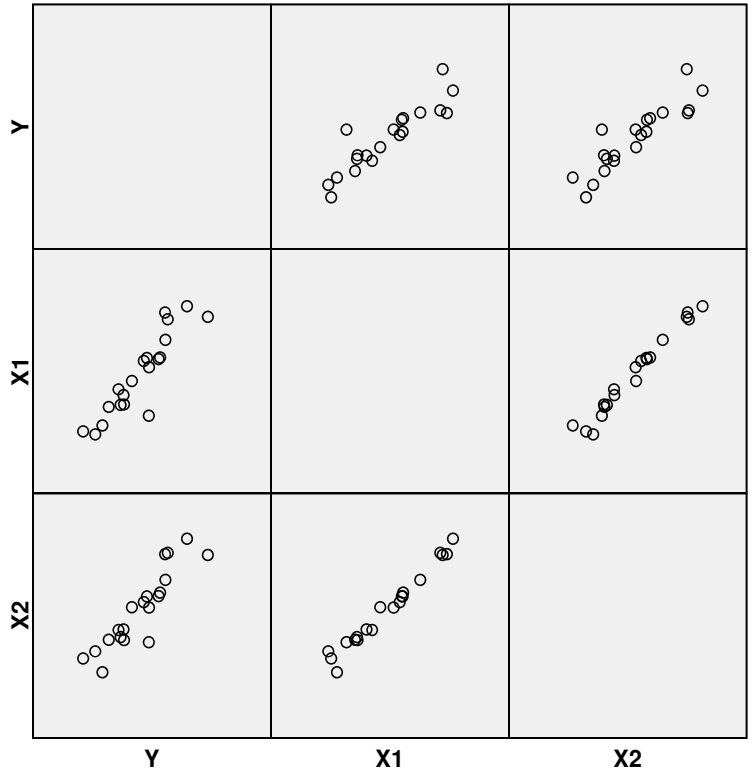
Cuando una de las X_j es combinación lineal de las restantes variables regresoras, el sistema es indeterminado. Entonces diremos que las variables explicativas son *colineales*.

En la práctica esto nunca pasa de manera exacta, aunque sí es posible que en un conjunto de datos algunas de las variables regresoras se puedan describir muy bien como función lineal de las restantes variables.

Este problema, llamado *multicolinealidad*, hace que los estimadores de los parámetros $\hat{\beta}_i$ tengan alta variabilidad (errores típicos muy grandes) y sean muy dependientes entre sí.

Multicolinealidad

y	x1	x2
-0.67	-0.43	-0.57
4.36	1.36	1.42
0.70	0.52	0.45
-1.00	-0.12	-0.33
-1.59	-0.48	-0.56
-3.13	-0.98	-1.00
-2.40	-1.04	-0.83
1.79	1.45	1.44
1.95	1.31	1.47
-0.70	-0.24	-0.32
-1.97	-0.86	-1.32
1.82	0.89	0.84
1.49	0.53	0.54
-0.88	-0.44	-0.50
1.40	0.50	0.46
0.82	-0.66	-0.62
0.51	0.46	0.32
0.83	0.33	0.19
3.11	1.58	1.80
-0.20	0.05	0.20



Correlaciones

		Y	X1	X2
Y	Correlación de Pearson	1	,906	,902
	Sig. (bilateral)		,000	,000
	N	20	20	20
X1	Correlación de Pearson	,906	1	,987
	Sig. (bilateral)	,000		,000
	N	20	20	20
X2	Correlación de Pearson	,902	,987	1
	Sig. (bilateral)	,000	,000	
	N	20	20	20

Multicolinealidad

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,907[a]	,823	,803	,84071

a. Variables predictoras: (Constante), X2, X1

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	56,049	2	28,025	39,651	,000[a]
	Residual	12,015	17	,707		
	Total	68,065	19			

a. Variables predictoras: (Constante), X2, X1

b. Variable dependiente: Y

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-,041	,202		-,205	,840
	X1	1,360	1,426	,601	,954	,354
	X2	,648	1,319	,309	,491	,630

a. Variable dependiente: Y