Técnicas de análisis cuantitativas y cualitativas

Sesión 2

Eduardo Corbelle Rico

Máster Universitario en Xestión Sustentable da Terra e o Territorio
Universidade de Santiago de Compostela

Curso 2015–2016

Contenidos

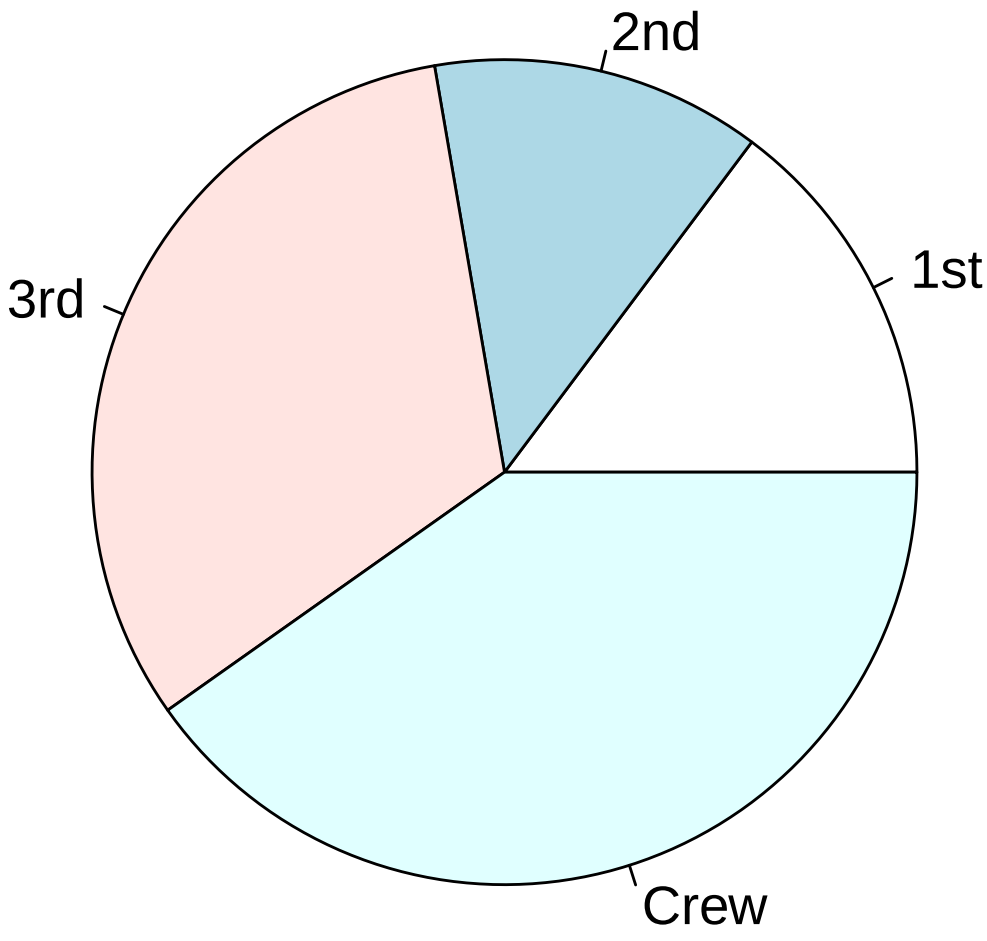
Variables categóricas

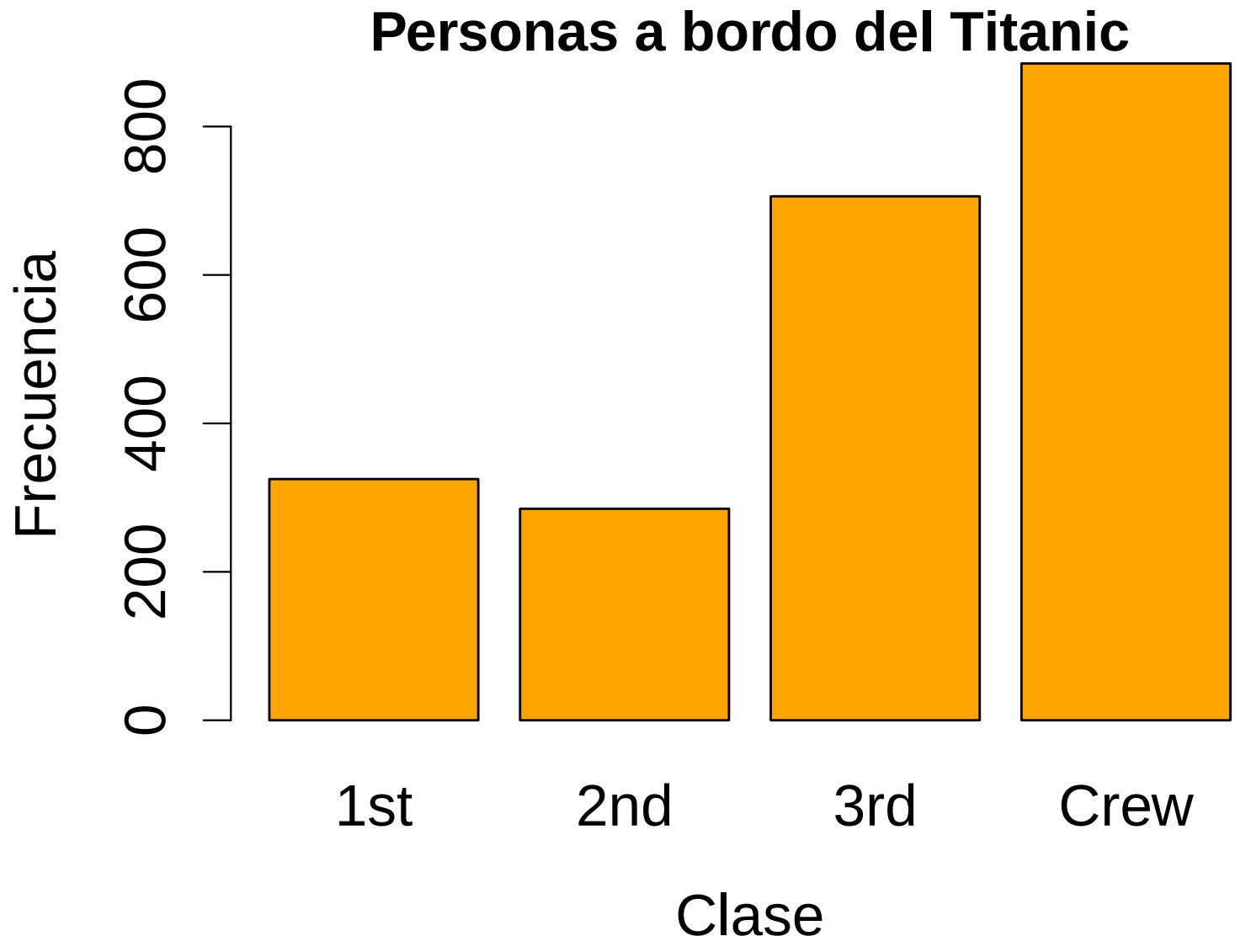
Tablas de frecuencias

Ejemplo: Personas a bordo del Titanic

Clase			
Primera	Segunda	Tercera	Tripulación
325	285	706	885

(N= 2201)


2nd
1st
3rd
Crew


Personas a bordo del Titanic
Frecuencia
0
200
400
600
800
1st
2nd
3rd
Crew
Clase

Tablas de contingencia

Ejemplo: Supervivencia a bordo del Titanic

Sobrevivió	Clase			
	1ª	2ª	3ª	Tripulación
No	122	167	528	673
Sí	203	118	178	212

$(N = 2201)$

Tablas de contingencia (% por filas)

Ejemplo: Supervivencia a bordo del Titanic

Sobrevivió	Clase				Total
	1ª	2ª	3ª	Tripulación	
No	8	11	35	45	100 %
Sí	29	17	25	30	100 %

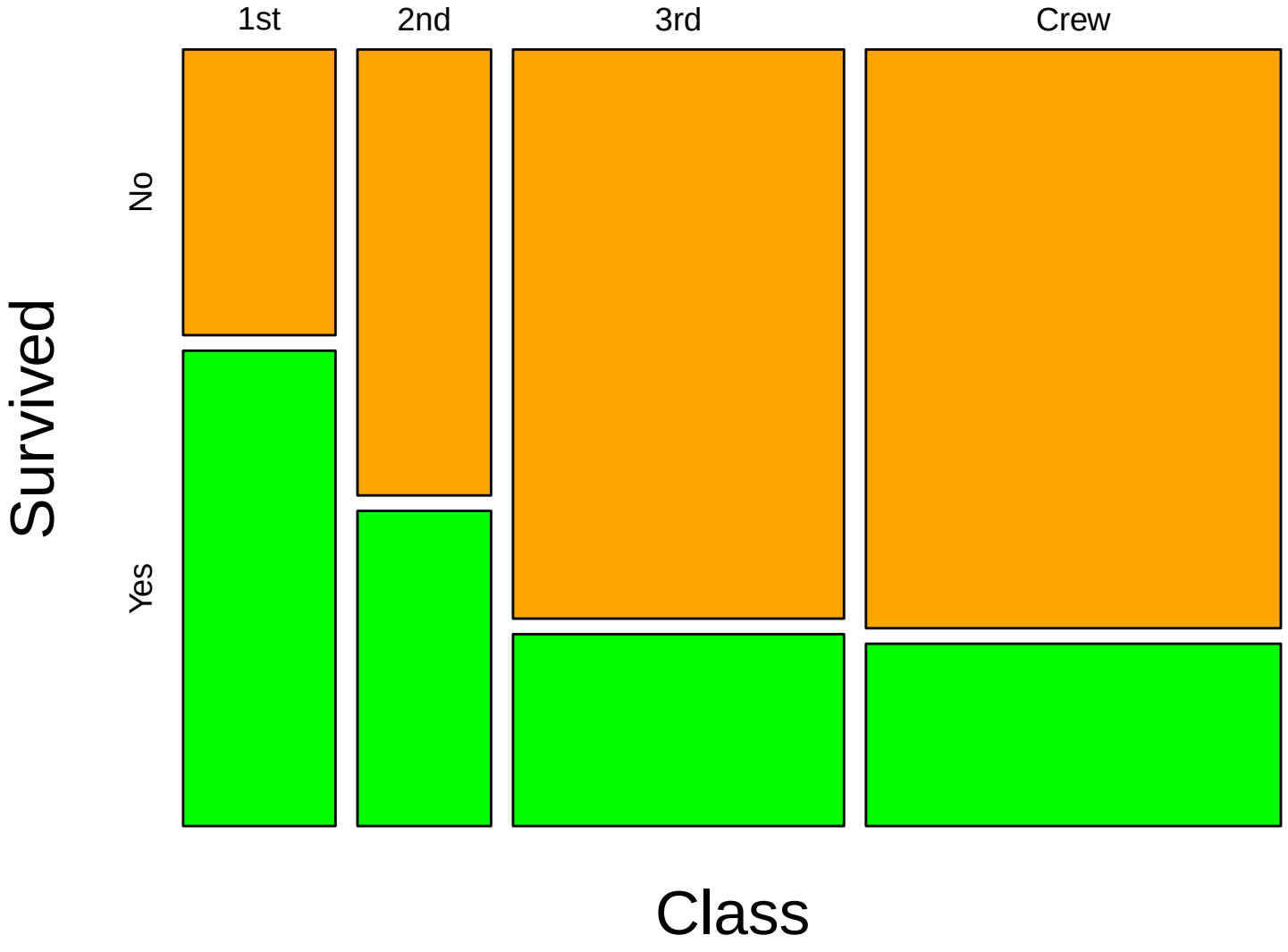
$(N = 2201)$

Tablas de contingencia (% por columnas)

Ejemplo: Supervivencia a bordo del Titanic

Sobrevivió	Clase			
	1ª	2ª	3ª	Tripulación
No	38	59	75	76
Sí	62	41	25	24
Total	100 %	100 %	100 %	100 %

($N = 2201$)


1st
2nd
3rd
Crew
No
Yes
Survived
Class

Pruebas χ^2

Contrastes de...

- bondad de ajuste
- homogeneidad de muestras
- independencia de caracteres

Contraste de bondad de ajuste

H_0 La variable observada procede de una determinada distribución modelo

Contraste de bondad de ajuste

H_0 La variable observada procede de una determinada distribución modelo

Ejemplo: determinar si un dado está trucado

Resultado	1	2	3	4	5	6
Observaciones	4	3	4	3	1	5

($N = 20$)

Contraste de bondad de ajuste

H_0 La variable observada procede de una determinada distribución modelo

Ejemplo: determinar si un dado está trucado

Resultado	1	2	3	4	5	6
Observaciones	4	3	4	3	1	5

($N = 20$)

```
> chisq.test(table(lanz), p = rep(1/6, 6))
Chi-squared test for given probabilities
data: table(lanz)
X-squared = 2.8, df = 5, p-value = 0.7308
```

Contraste de homogeneidad de muestras

H_0 Las muestras proceden de poblaciones con iguales características

Ejemplo: partidarios de la independencia, 3 CCAA

	Si	No	Indiferente
A	40	45	15
B	37	39	24
C	43	51	6

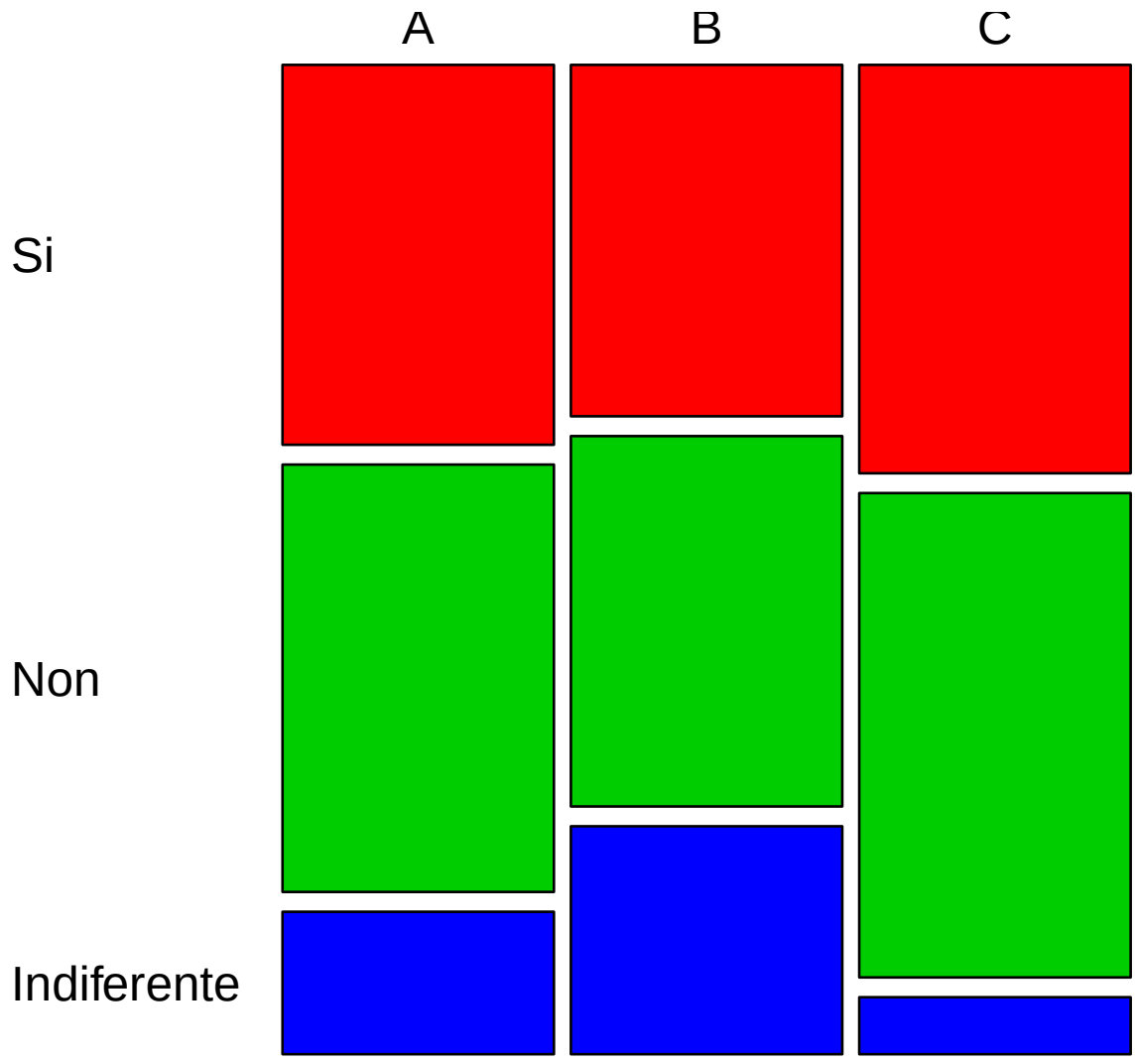
Contraste de homogeneidad de muestras

H_0 Las muestras proceden de poblaciones con iguales características

Ejemplo: partidarios de la independencia, 3 CCAA

	Si	No	Indiferente
A	40	45	15
B	37	39	24
C	43	51	6

```
> chisq.test(tabla)
Pearson's Chi-squared test
data: tabla
X-squared = 12.85, df = 4, p-value = 0.01203
```


A
B
C
Si
Non
Indiferente

Contraste de independencia de caracteres

H_0 Las dos variables (caracteres) son independientes

Contraste de independencia de caracteres

H_0 Las dos variables (caracteres) son independientes

Ejemplo: Nivel de denuncias de consumidores y sector de actividad

	Nulo	Bajo	Medio	Alto
G.almacenes	12	6	4	2
Bancos	6	12	16	6
Agencias de viajes	12	20	36	16
Telefonía	0	0	2	10

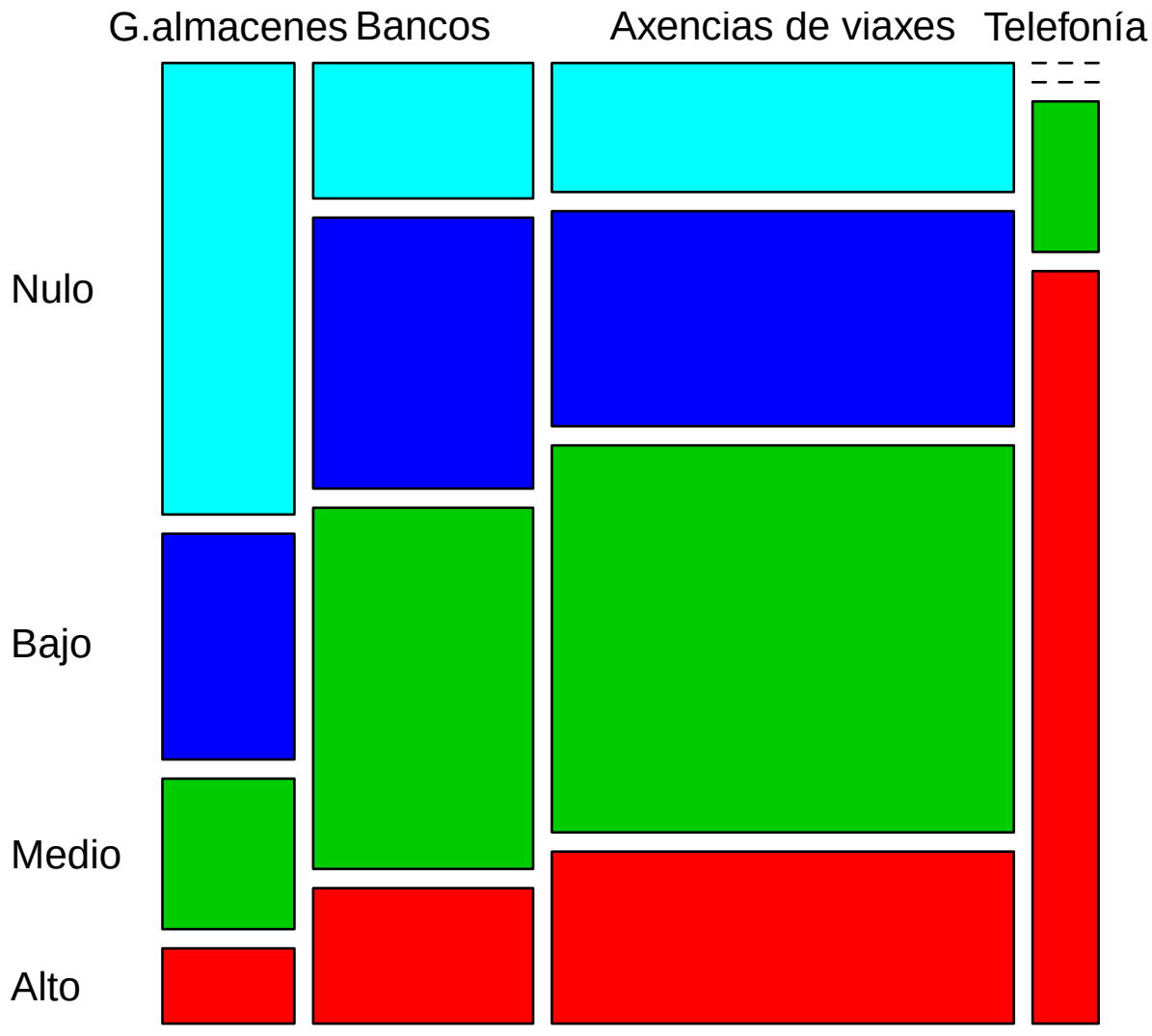
Contraste de independencia de caracteres

H_0 Las dos variables (caracteres) son independientes

Ejemplo: Nivel de denuncias de consumidores y sector de actividad

	Nulo	Bajo	Medio	Alto
G.almacenes	12	6	4	2
Bancos	6	12	16	6
Agencias de viajes	12	20	36	16
Telefonía	0	0	2	10

```
> chisq.test(taboa2)
Pearson's Chi-squared test
data: taboa2
X-squared = 49.0191, df = 9, p-value = 1.646e-07
```


G.almacenes
Bancos
Axencias de viaxes
Telefonía
Nulo
Bajo
Medio
Alto

Análisis de correspondencias

Método exploratorio para representar la asociación entre los niveles de dos variables categóricas
Normalmente asociado a un contraste de independencia de caracteres

- Simple (2 variables)
- Múltiple (máis de 2)

Práctica 3

- Tablas de contingencia
- Contraste de independencias de caracteres
- Análisis de correspondencias

Correlación

Correlación (definición: RAE)

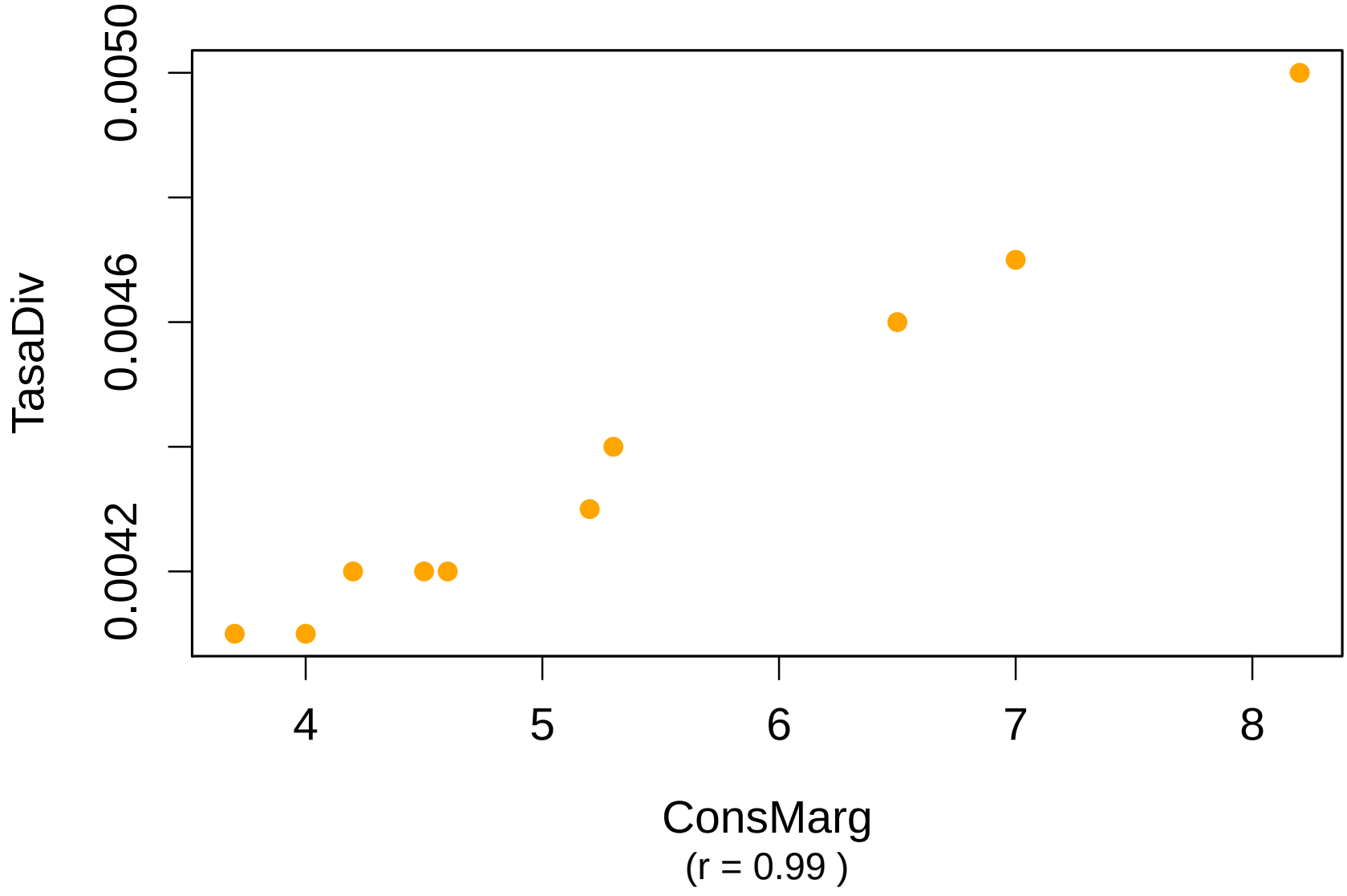
1. f. Correspondencia o relación recíproca entre dos o más cosas o series de cosas
2. f. *Fon.* Conjunto de dos series de fonemas opuestas por los mismos rasgos distintivos
3. f. *Fon.* Relación que se establece entre estas series
4. f. *Mat.* Medida de la dependencia existente entre variantes aleatorias

Correlación

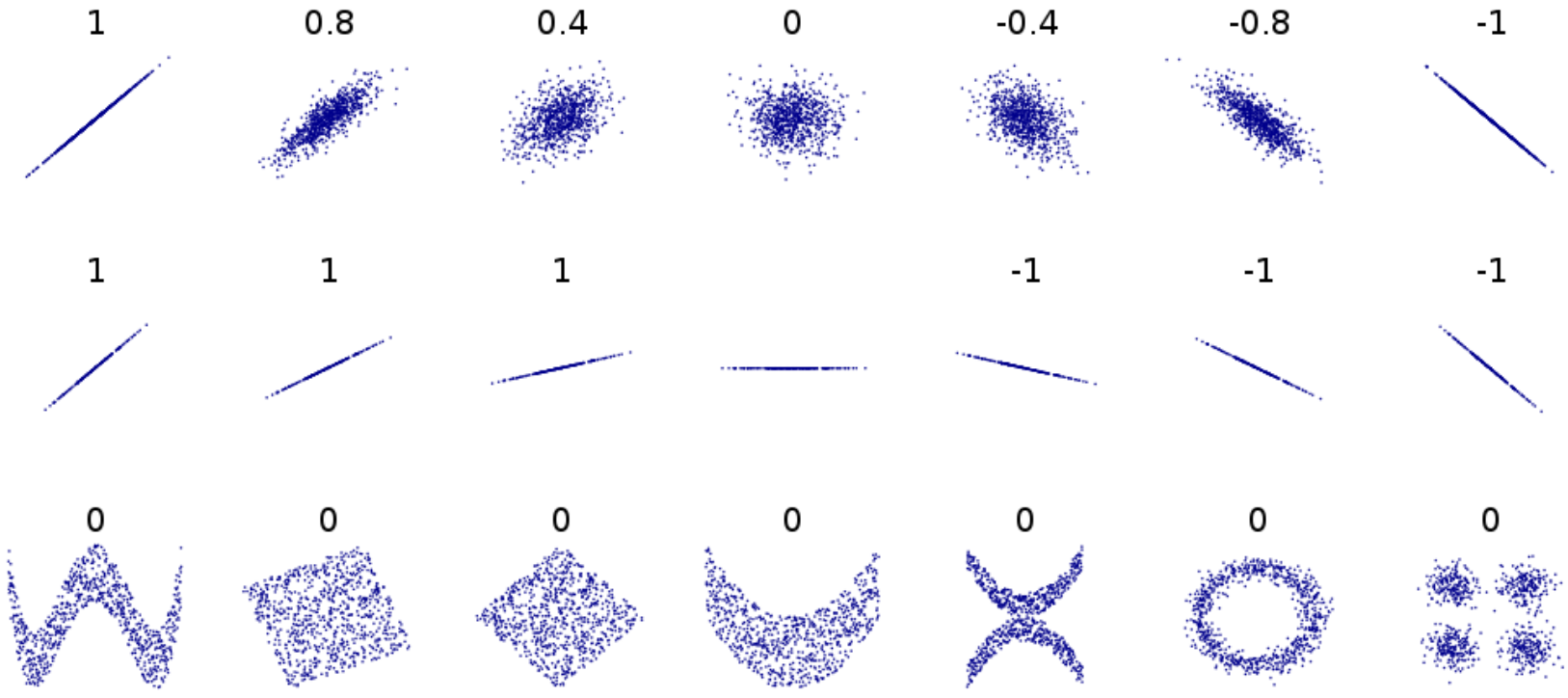
Correlación (definición: RAE)

1. f. Correspondencia o relación recíproca entre dos o más cosas o series de cosas
2. f. *Fon.* Conjunto de dos series de fonemas opuestas por los mismos rasgos distintivos
3. f. *Fon.* Relación que se establece entre estas series
4. f. *Mat.* Medida de la dependencia existente entre variantes aleatorias

Exploración visual



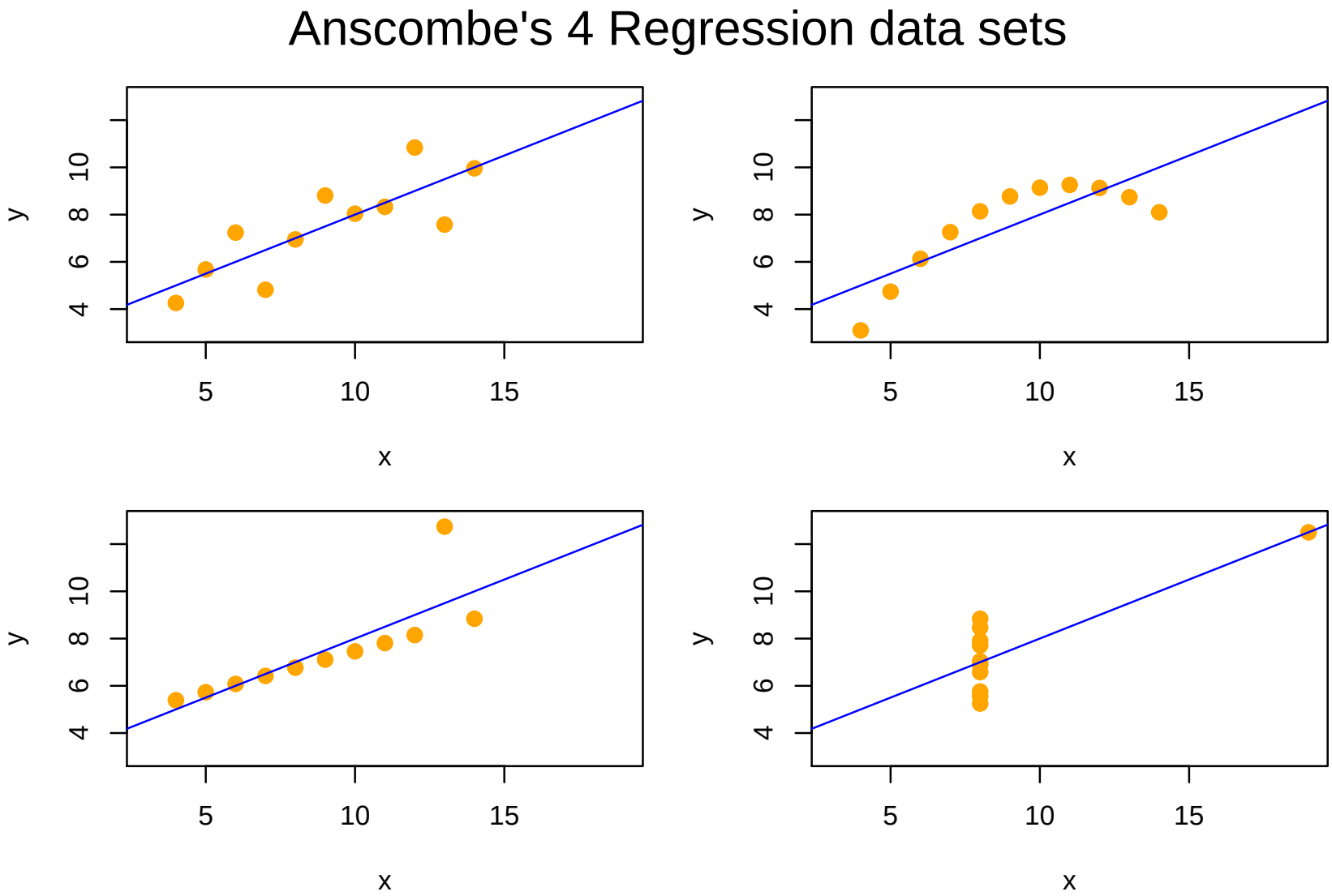
Coef. de correlación de Pearson



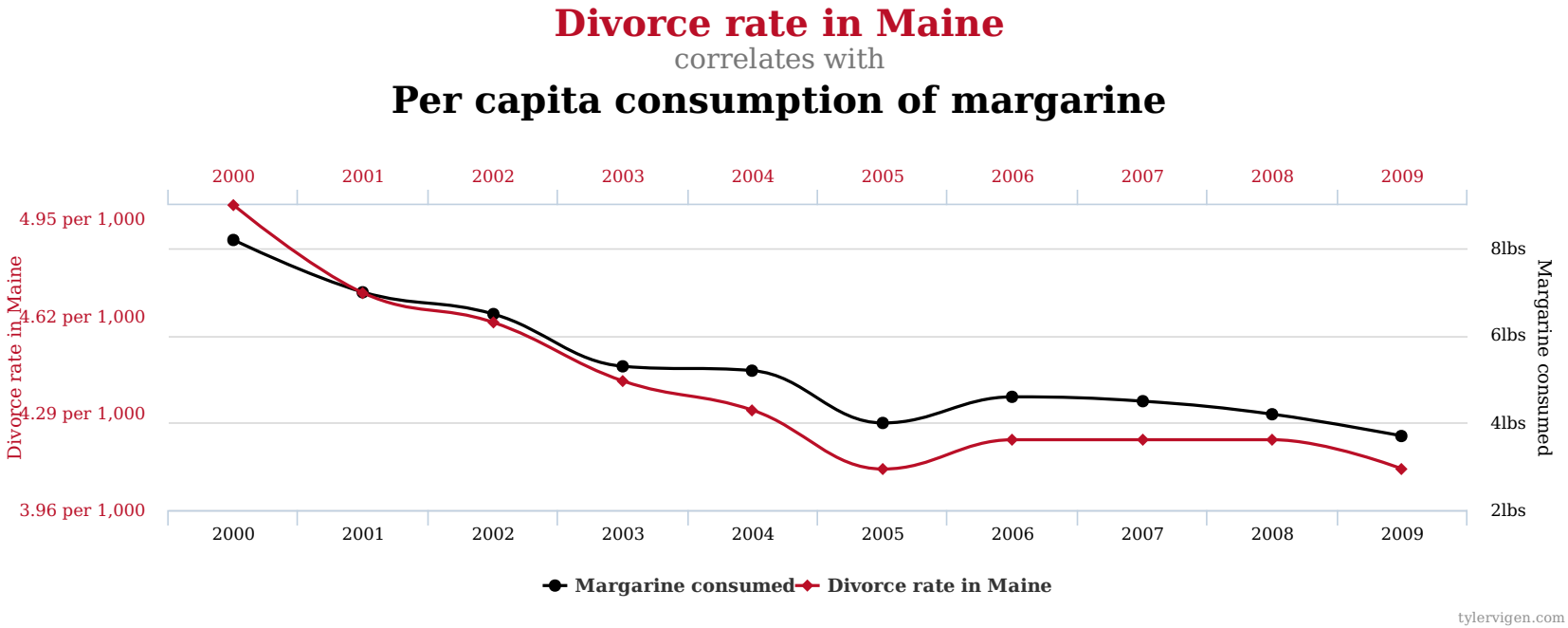
(Imagen: en.wikipedia.org)

Cuarteto de Anscombe

$(r = 0,816)$



Correlación no implica causa



Fuente: Spurious Correlations, `http://www.tylervigen.com/`

Regresión lineal simple

Objetivo: inferir una relación lineal entre dos variables

$$Y = \alpha + \beta X + e$$

Finalidad: explicativa / predictiva.
Habitualmente mediante mínimos cadrados: $Min(\sum e_i^2)$

Regresión lineal simple

Objetivo: inferir una relación lineal entre dos variables

$$Y = \alpha + \beta X + e$$

Finalidad: explicativa / predictiva.
Habitualmente mediante mínimos cadrados: $Min(\sum e_i^2)$

Supuestos de partida

- La relación entre las variables es lineal
- Varianza de e independiente de x (homocedasticidad)
- El residuo e sigue una distribución normal

Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + e$$

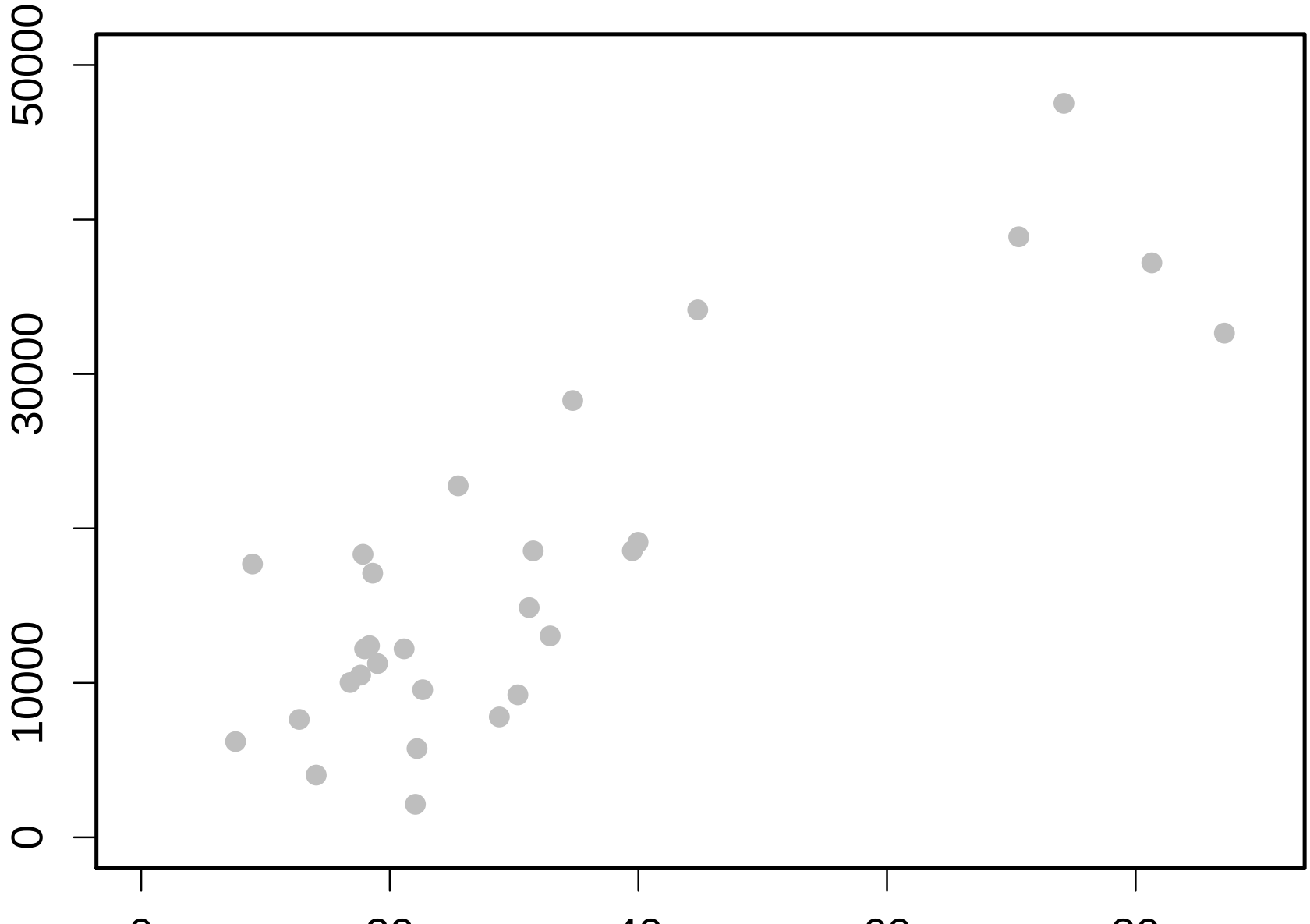
Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + e$$

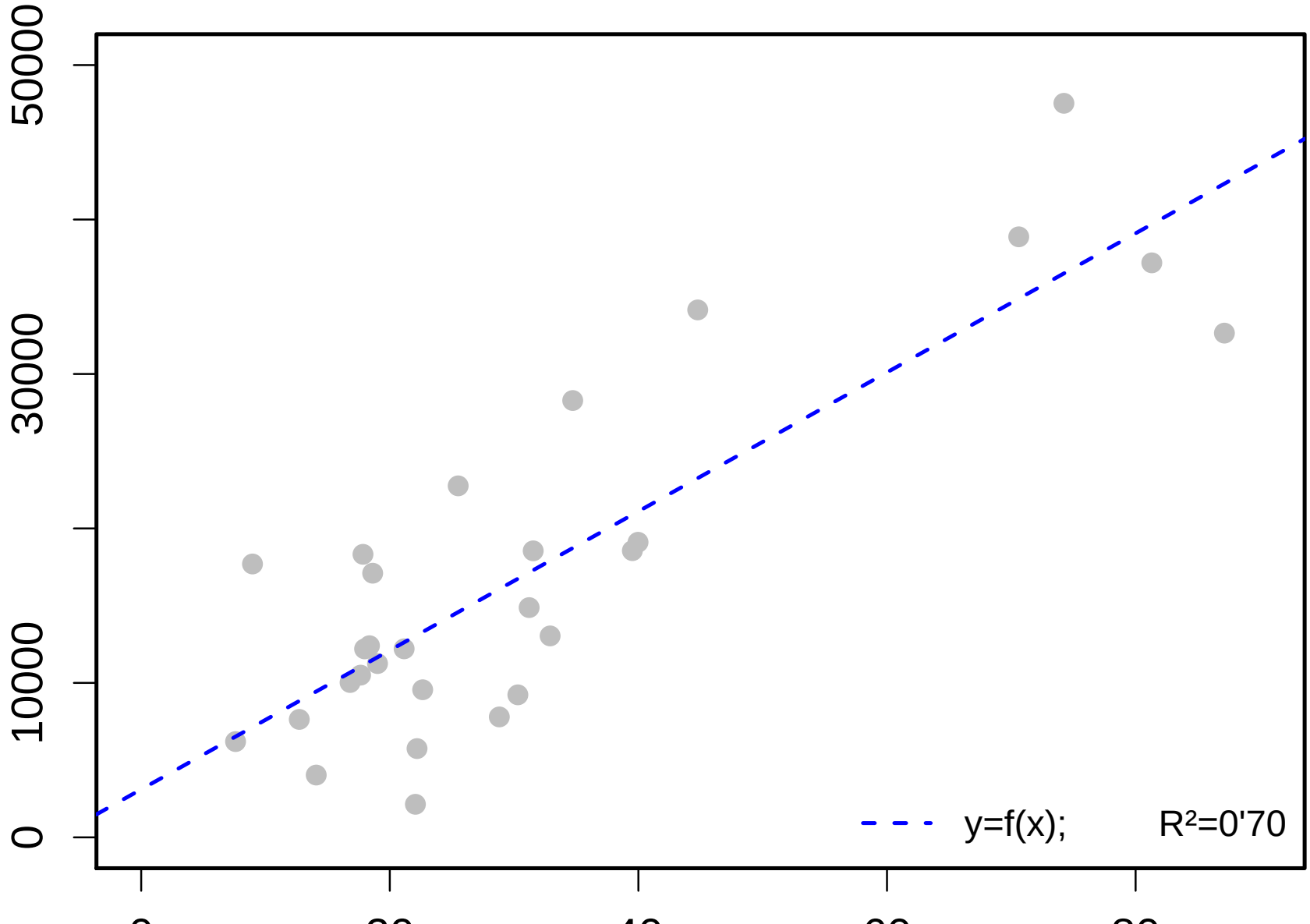
Tamaño muestral deseable

- > 20 y $< 1{,}000$ observaciones
- 15–20 observaciones por cada variable independiente (< 5 obs. por variable pueden causar sobreajuste)

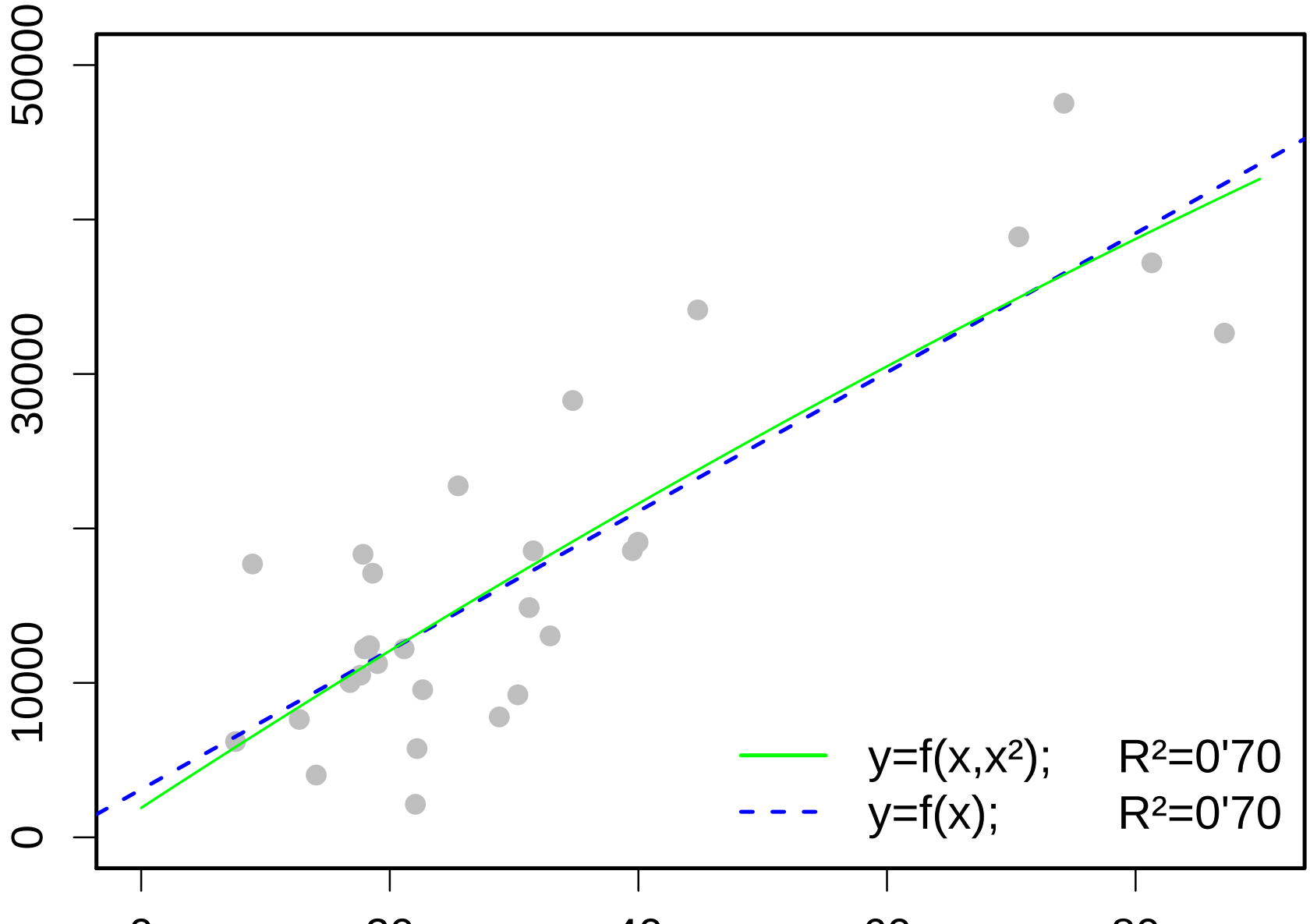
Sobreajuste (*overfitting*)



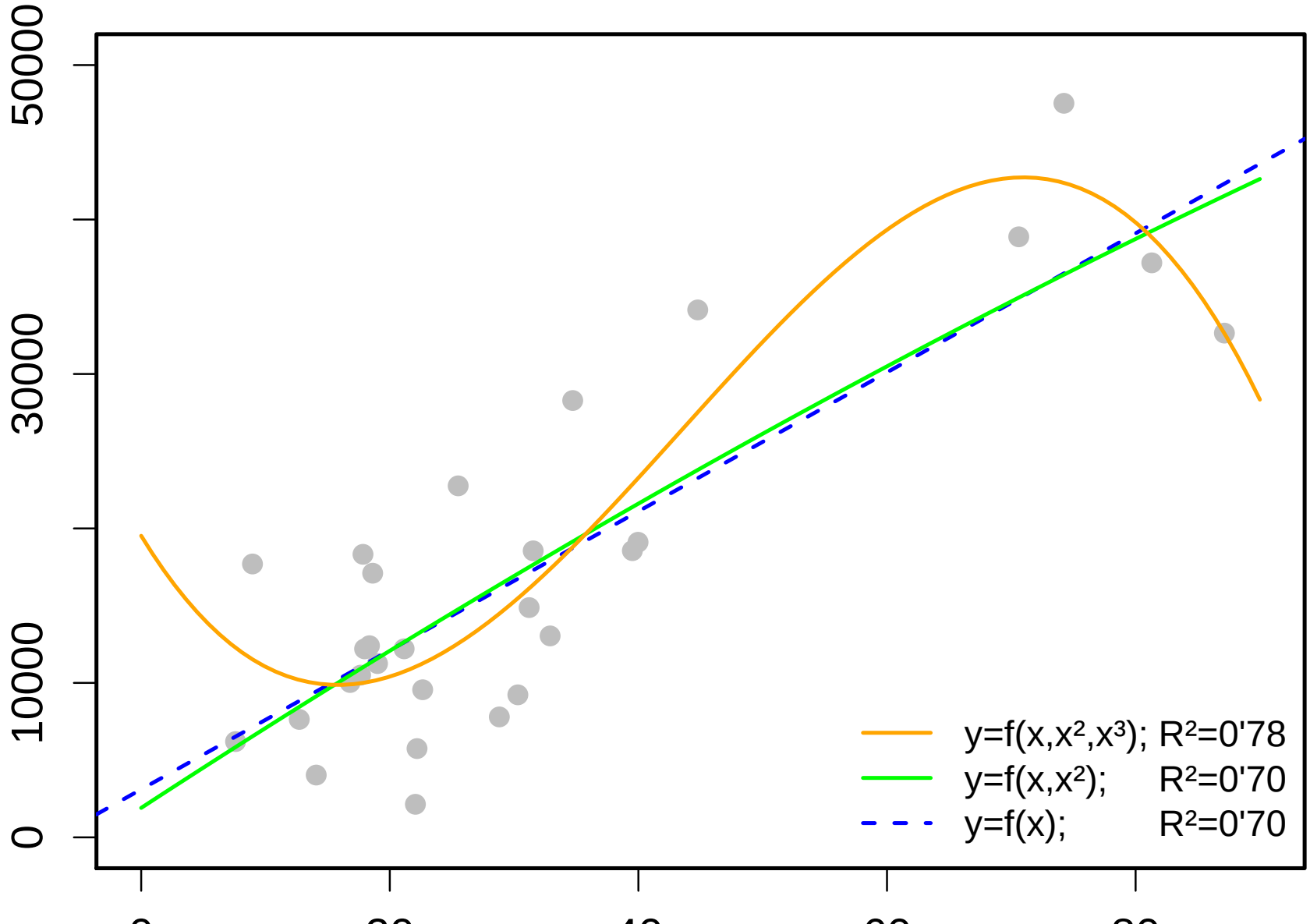
Sobreajuste (*overfitting*)



Sobreajuste (*overfitting*)



Sobreajuste (*overfitting*)



Variables de entrada

Transformación de las variables

- Para introducir variables no numéricas
- Para mejorar (linearizar) la relación
- Para hacer más simétrica su distribución
- Para estandarizar los coeficientes de la regresión

Variables de entrada

Transformación de las variables

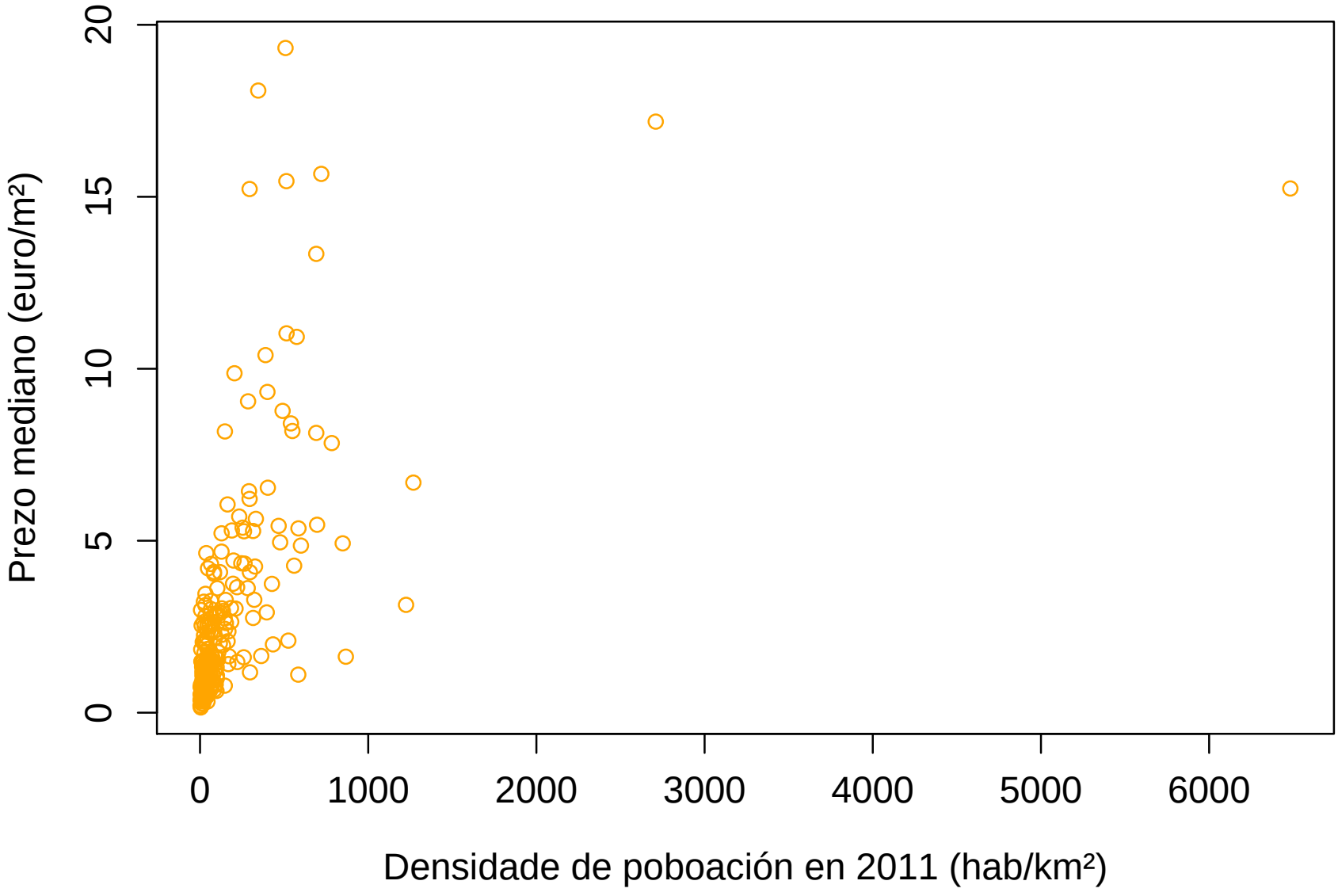
- Para introducir variables no numéricas
- Para mejorar (linearizar) la relación
- Para hacer más simétrica su distribución
- Para estandarizar los coeficientes de la regresión

Multicolinearidad

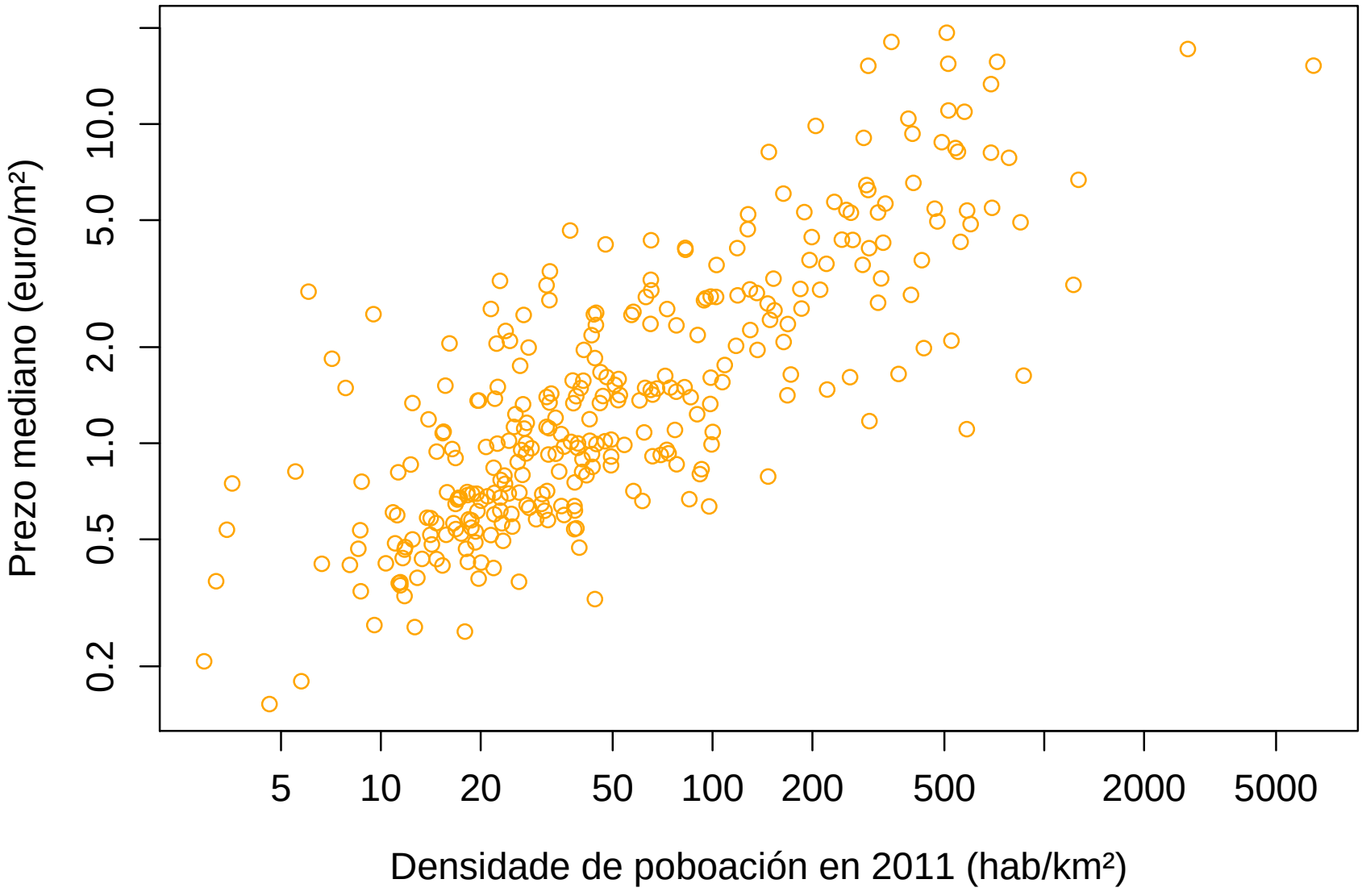
Correlación entre variables independientes

- Omitir una o varias
- Emplear sólo para predicción

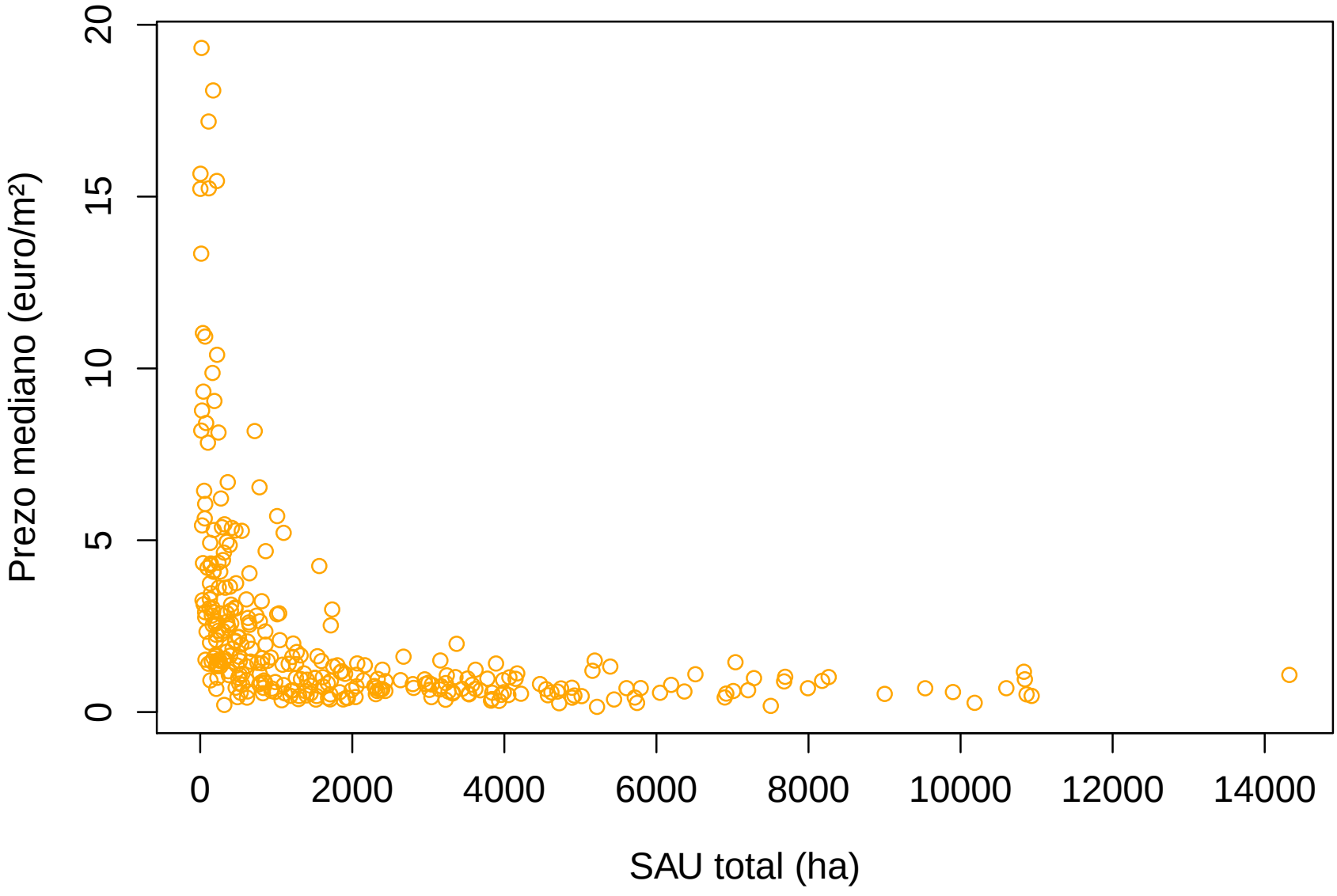
Transformación de variables



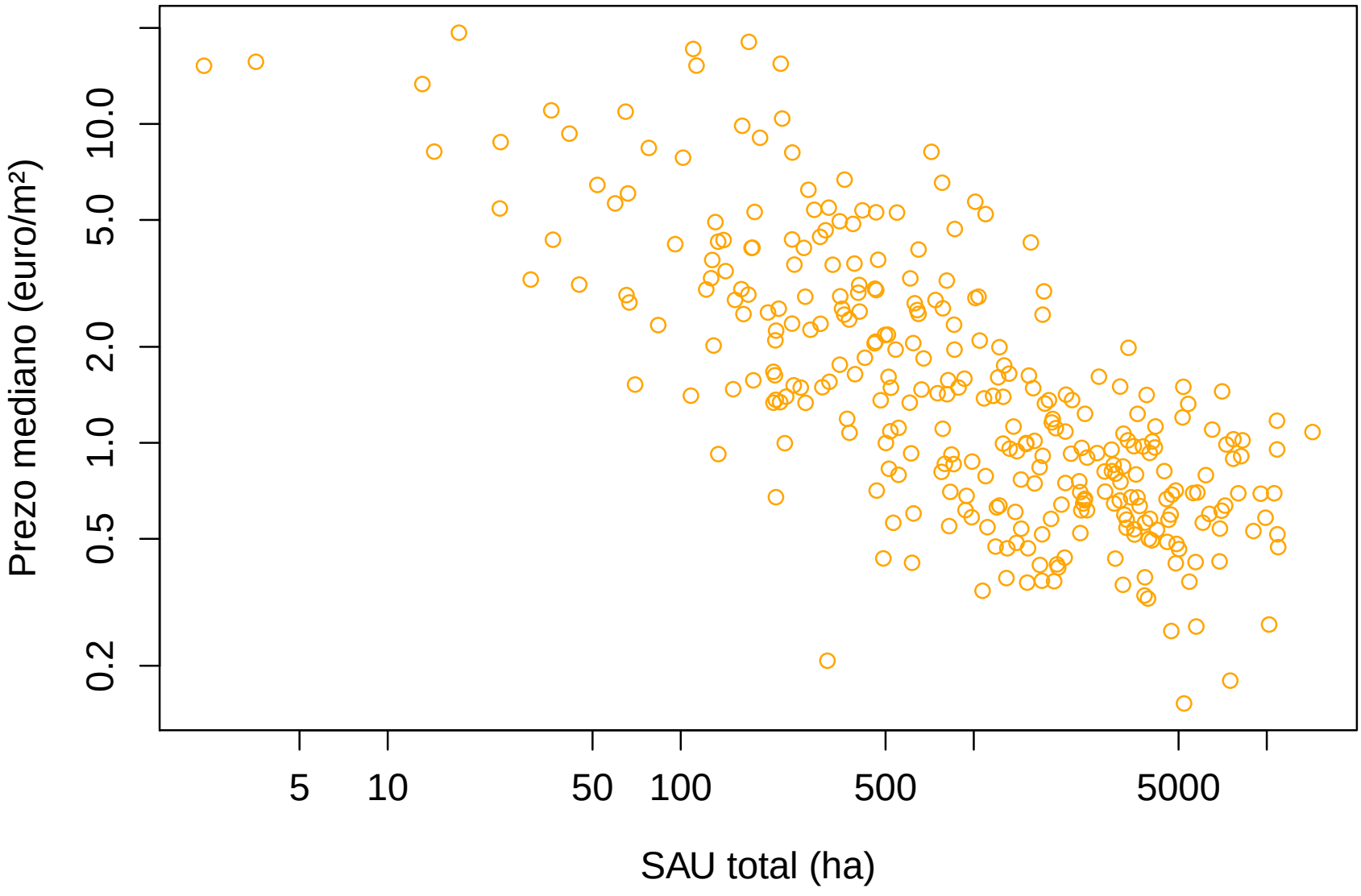
Transformación de variables



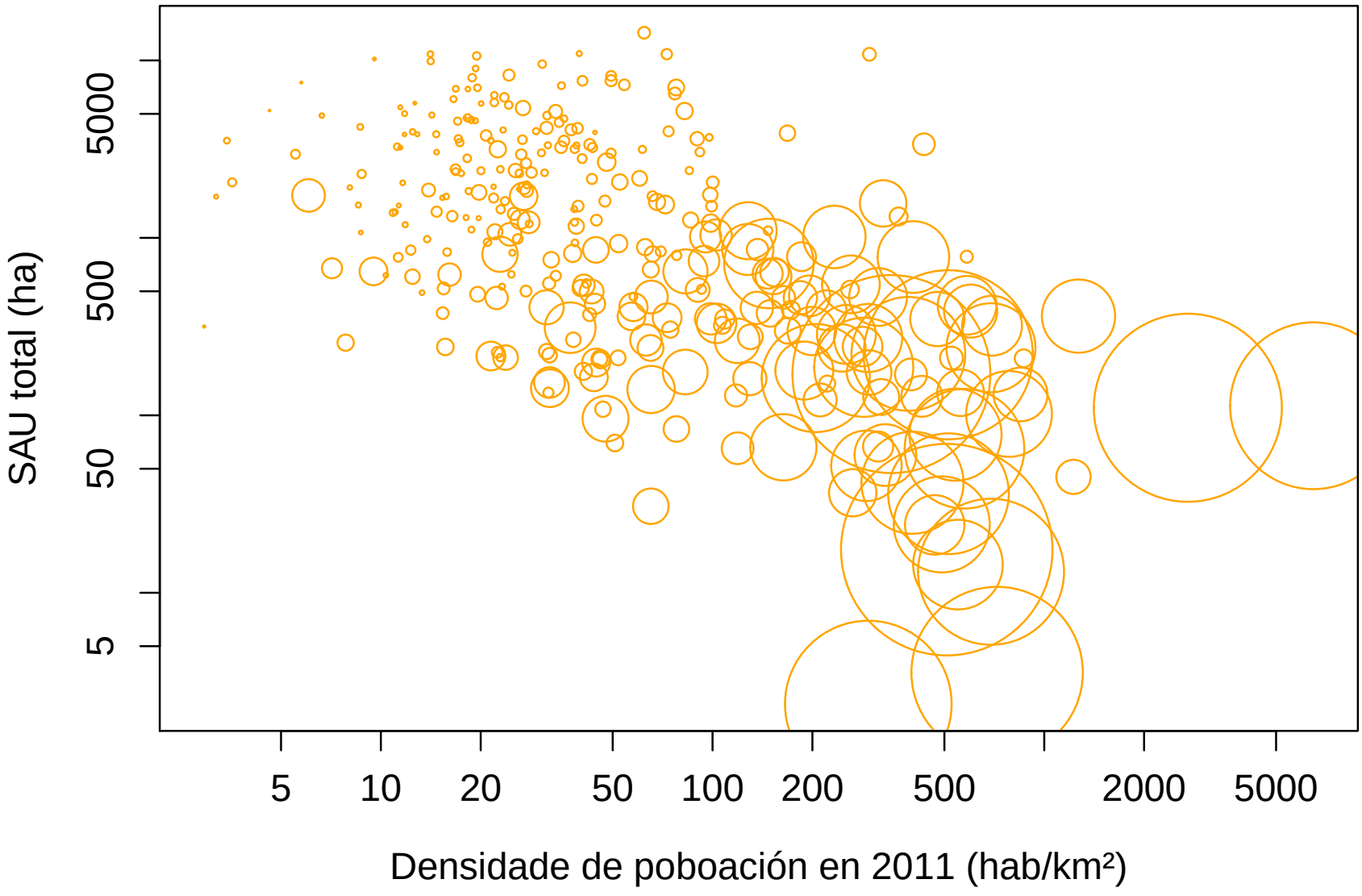
Transformación de variables



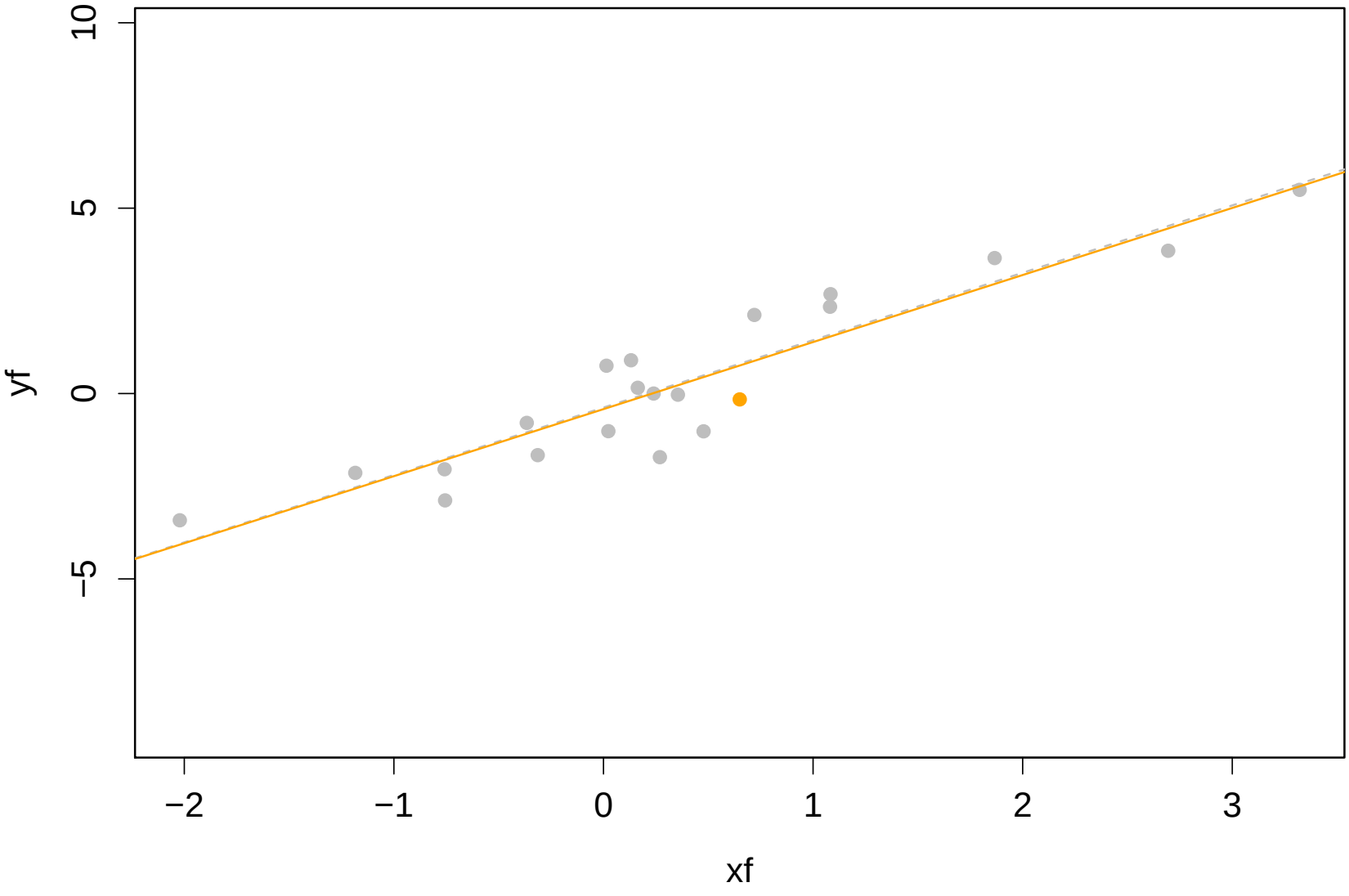
Transformación de variables



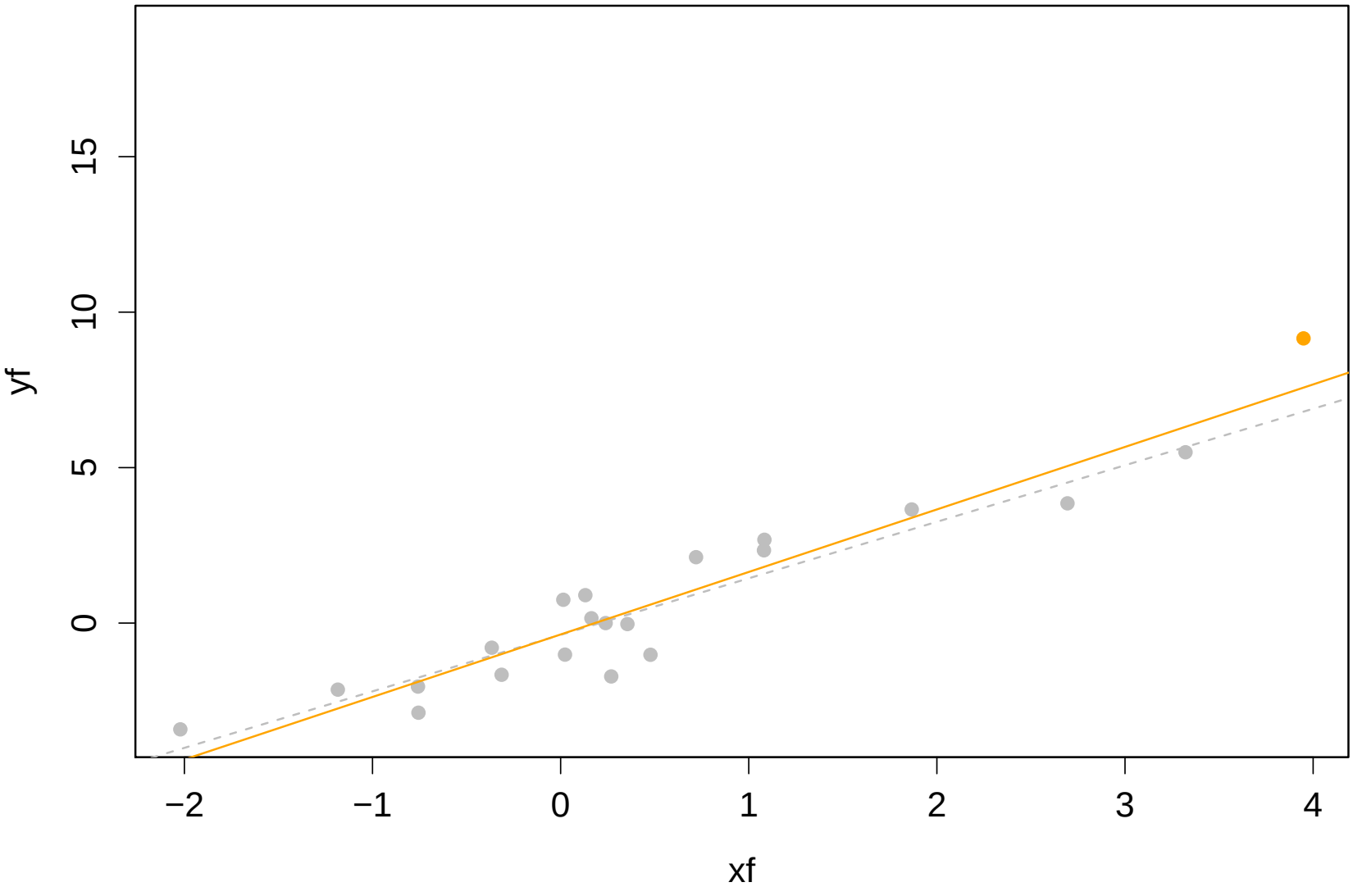
Transformación de variables



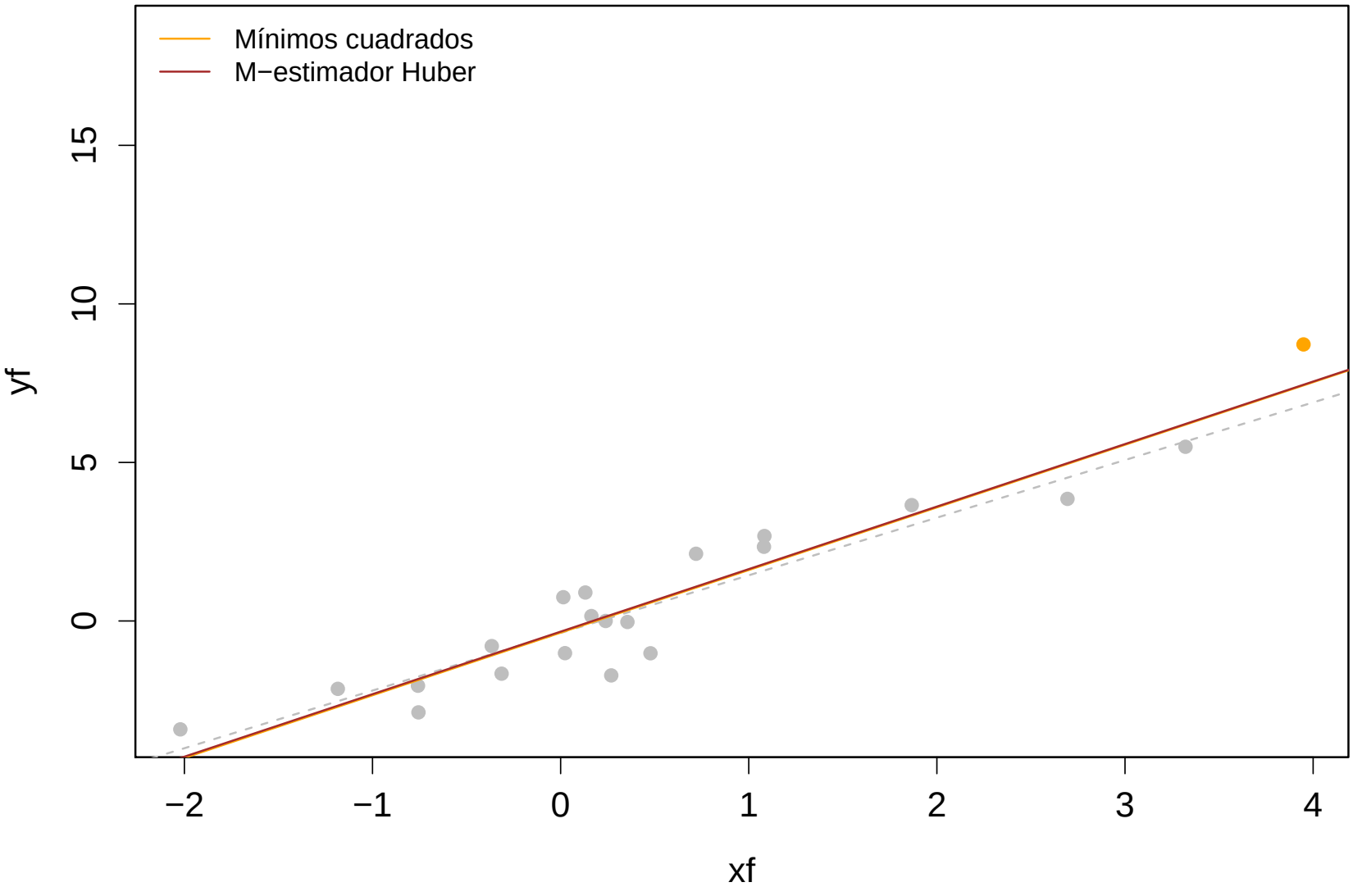
Datos anómalos (*outliers*) y apalancamiento



Datos anómalos (*outliers*) y apalancamiento



Datos anómalos (*outliers*) y apalancamiento



Datos anómalos: orígenes posibles

- Error en la toma o manipulación de datos
- Observación excepcional explicable por una situación extraordinaria
- Observación excepcional sin explicación plausible (preferible no eliminarla)

Regresión para fines predictivos

Residuo de ajuste y error de predicción

- Validación con una submuestra
- Validación cruzada

Práctica 4

- Regresión lineal simple
- Regresión lineal múltiple