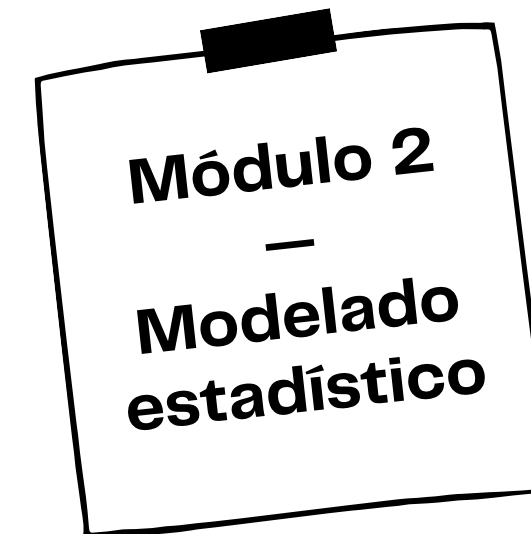


Regresión Logística

“All Models are wrong, but some are useful.” George Box



Q Agenda de hoy

- | | | | |
|---|--|---|----------------------|
| 1 | Regresión Logística | 5 | Regresión Polinómica |
| 2 | Estimación de la ecuación de regresión | 6 | Modelos no lineales |
| 3 | Prueba de significancia | | |
| 4 | Interpretación de la ecuación de regresión | | |

Regresión Logística (1/9)

Hasta ahora hemos utilizado modelos de regresión en los que la variable dependiente es continua: ventas mensuales, número de homicidios, ...



¿Qué pasa en el caso en el que la variable dependiente es discreta?



Ejemplos: Género de una persona, Cliente paga o no paga el próximo mes, Email es spam o no, Banco aprueba tarjeta de crédito o no, Voto de Sara va a votar en segunda vuelta)

Regresión Logística (2/9)

La Regresión Logística nos permite, dado un conjunto particular de valores de las variables independientes, estimar la probabilidad de pertenencia a cada categoría de la variable dependiente

Regresión Lineal

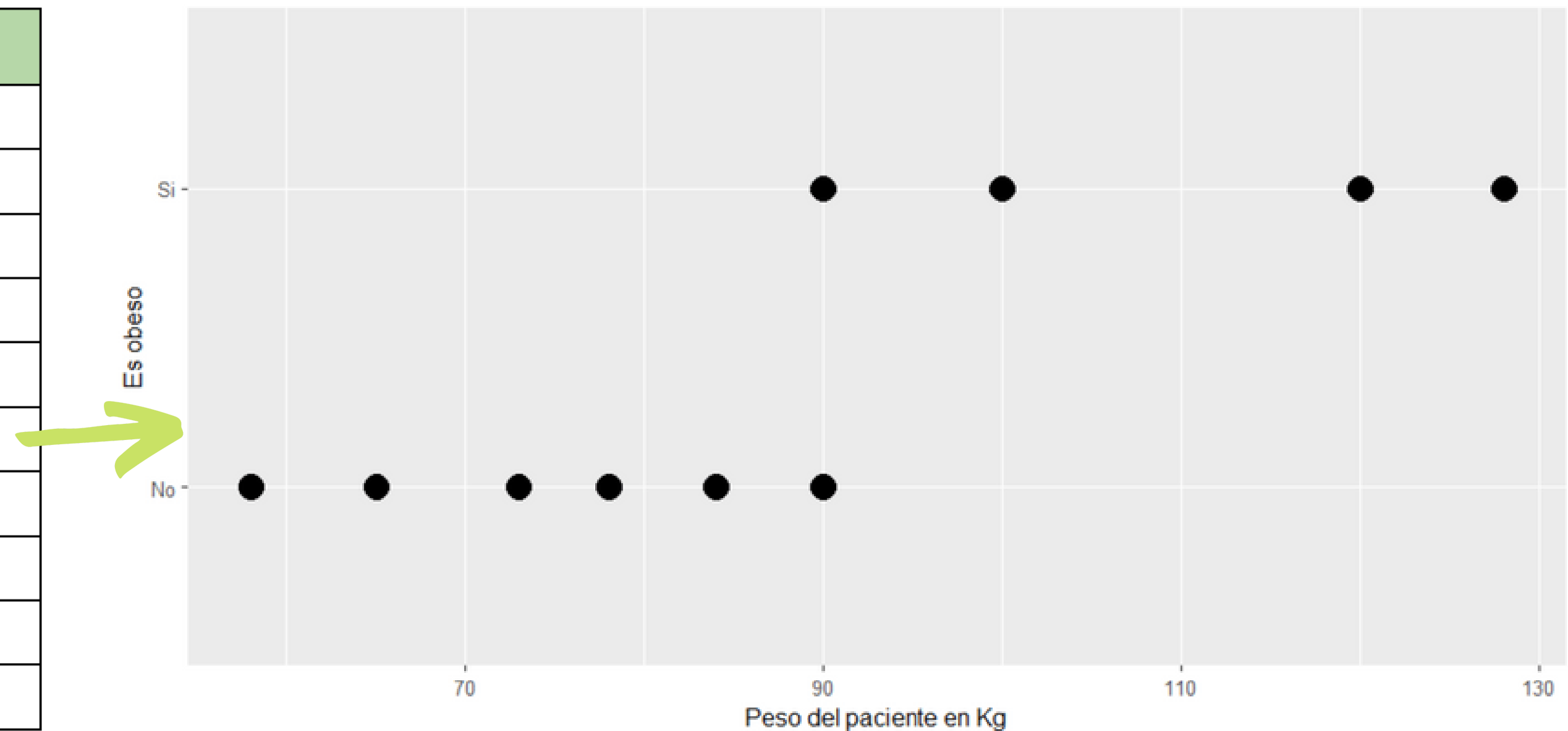
Los valores estimados son el promedio de la variable dependiente dados determinados valores de las variables independientes

Regresión Logística

Los valores estimados son la probabilidad de un nivel particular de la variable dependiente dados determinados valores de las variables independientes

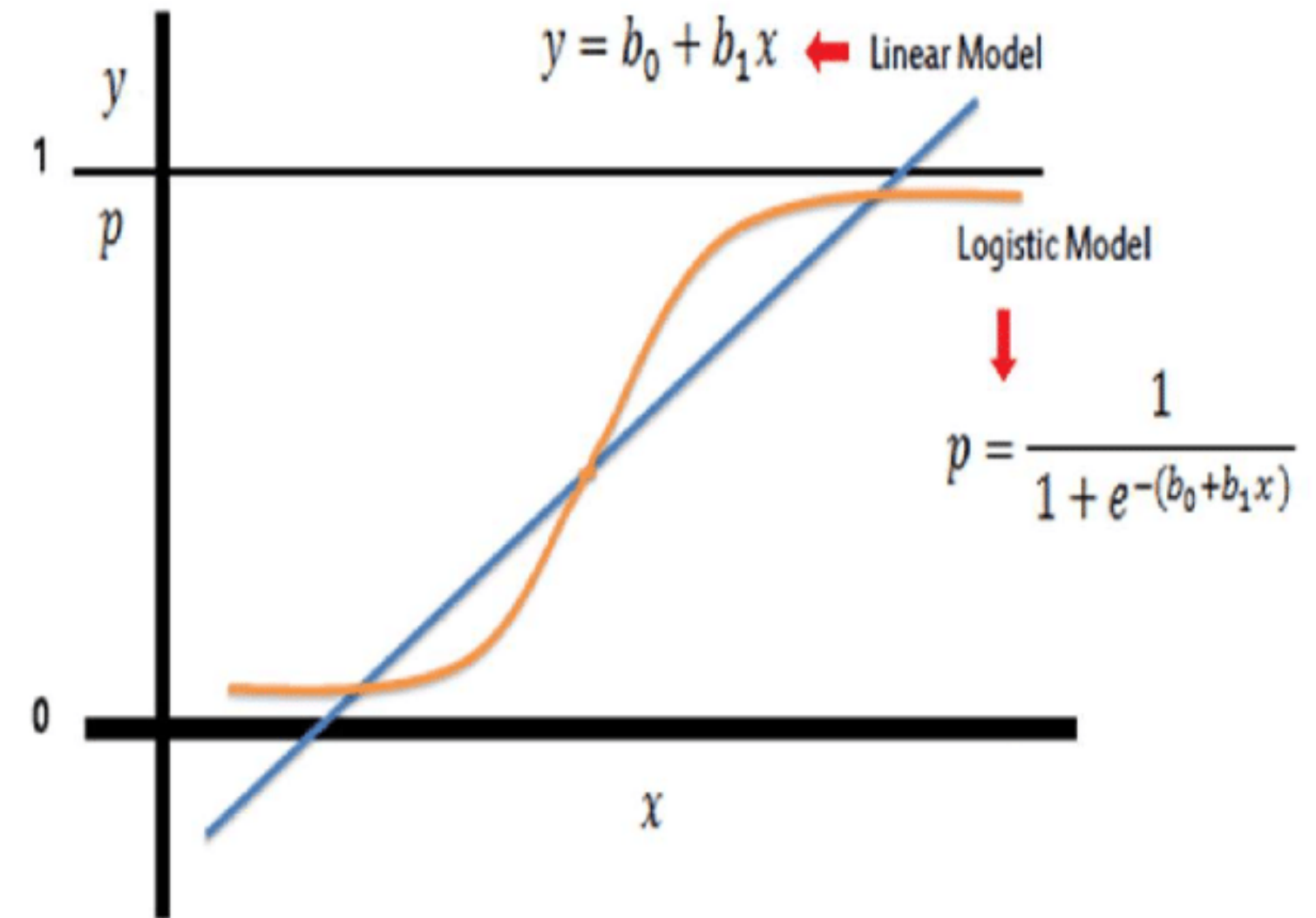
Regresión Logística (3/9)

Paciente i	Peso (Kg)	Es Obeso
1	100	Si
2	73	No
3	90	No
4	90	Si
5	120	Si
6	128	Si
7	65	No
8	58	No
9	78	No
10	84	No



Regresión Logística (4/9)

La Regresión Logística NO ajusta una línea a los datos (como la Regresión Lineal) sino una función con forma de **S** conocida como la **Función Logística**



Regresión Logística (5/9)

Ecuación de Regresión Logística




$$E(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \quad (1)$$

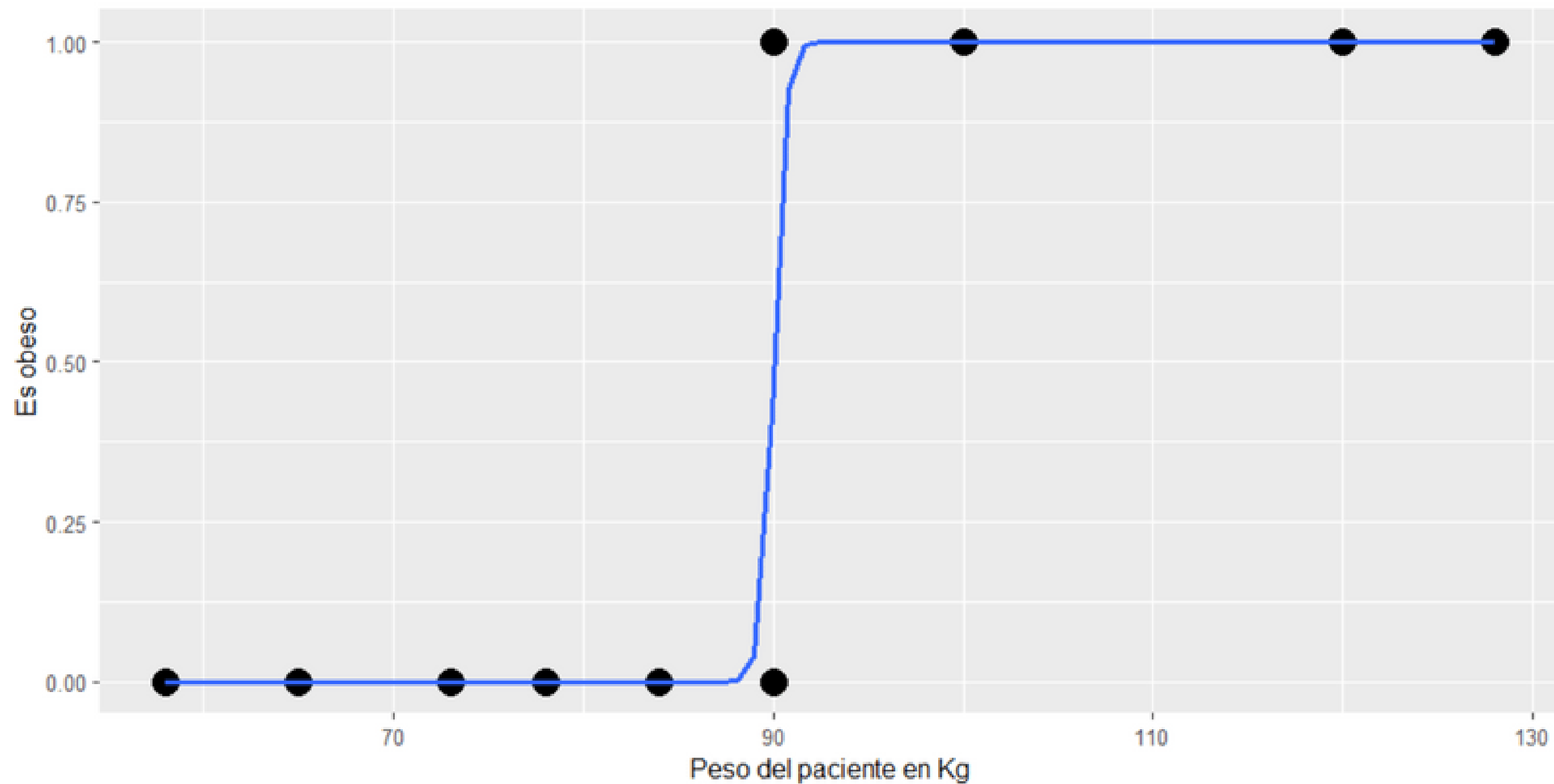
Regresión Logística (6/9)

En la práctica, los valores de los parámetros del modelo no se conocen y es necesario estimarlos usando **datos muestrales**

Ecuación de Regresión Logística Estimada

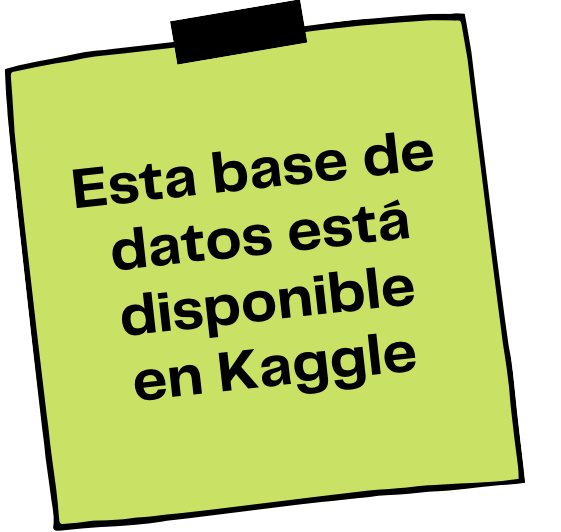

$$\hat{y} = \text{estimacion de } P(y = 1|x_1, x_2, \dots, x_p) = \frac{e^{b_0+b_1x_1+\dots+b_px_p}}{1 + e^{b_0+b_1x_1+\dots+b_px_p}} \quad (2)$$

Regresión Logística (7/9)



Regresión Logística (8/9)

Utilicemos una base de datos que contiene información de empleados de una compañía para predecir cuál es la probabilidad de que estos empleados renuncien (**IBM HR Analytics Employee Attrition & Performance**)



Esta base de
datos está
disponible
en Kaggle

Regresión Logística (9/9)

Employee	Resignation	Age	Marital Status	Gender	MonthlyIncome	YearsAtCompany	YearsInCurrentRole	YearsSinceLast Promotion
1	Yes	41	Single	Female	5993	6	4	0
2	No	49	Married	Male	5130	10	7	1
3	Yes	37	Single	Male	2090	0	0	0
4	No	33	Married	Female	2909	8	7	3
5	No	27	Married	Male	3468	2	2	2
6	No	32	Single	Male	3068	7	7	3
7	No	59	Married	Female	2670	1	0	0
8	No	30	Divorced	Male	2693	1	0	0
9	No	38	Single	Male	9526	9	7	1
10	Yes	41	Single	Female	5993	6	4	0

Ajustemos un modelo de Regresión Logística en R...

Prueba de Significancia

Prueba G (Chi-Squared Test)

La prueba G se usa para determinar si existe una relación de significancia entre la variable dependiente y las variables independientes

(Prueba de significancia global)

$$\chi^2 = \text{Null deviance} - \text{Residual deviance}$$
$$\chi^2 \sim \chi_p^2$$

Prueba z (Wald Test)

Si la prueba G indica que hay significancia global, posteriormente se usa la prueba z para determinar si cada una de las variables independientes es significativa

(Prueba de significancia individual)

Interpretación de la ecuación de regresión

Interpretación de parámetros: Variables cuantitativas



$$b_1 = -0.0264$$

- Un aumento en la edad de la persona está asociado con una disminución en la probabilidad de renunciar
- Se espera que la probabilidad de renunciar disminuya cuando la edad aumenta en una unidad

Interpretación de la ecuación de regresión

Interpretación de parámetros: Variables cualitativas



$$b_2 = 0.582$$

- La probabilidad de renuncia de una persona casada es mayor a la probabilidad de renuncia de una persona divorciada

Criterio AIC

AIC (Akaike Information Criterion) es una métrica usada para comparar el ajuste de diferentes modelos de regresión. **Entre menor sea el valor de AIC, mejor es el ajuste del modelo a los datos.**

$$AIC = 2k - 2\ln(L)$$



Número de parámetros
del modelo



Log-verosimilitud
del modelo

Medida de la
información
perdida al usar
un modelo
para aproximar
la realidad

Criterio BIC

BIC (Bayesian information criterion) es una métrica usada para comparar el ajuste de diferentes modelos de regresión. **Entre menor sea el valor de BIC, mejor es el ajuste del modelo a los datos.**

$$BIC = k \ln(n) - 2 \ln(L)$$



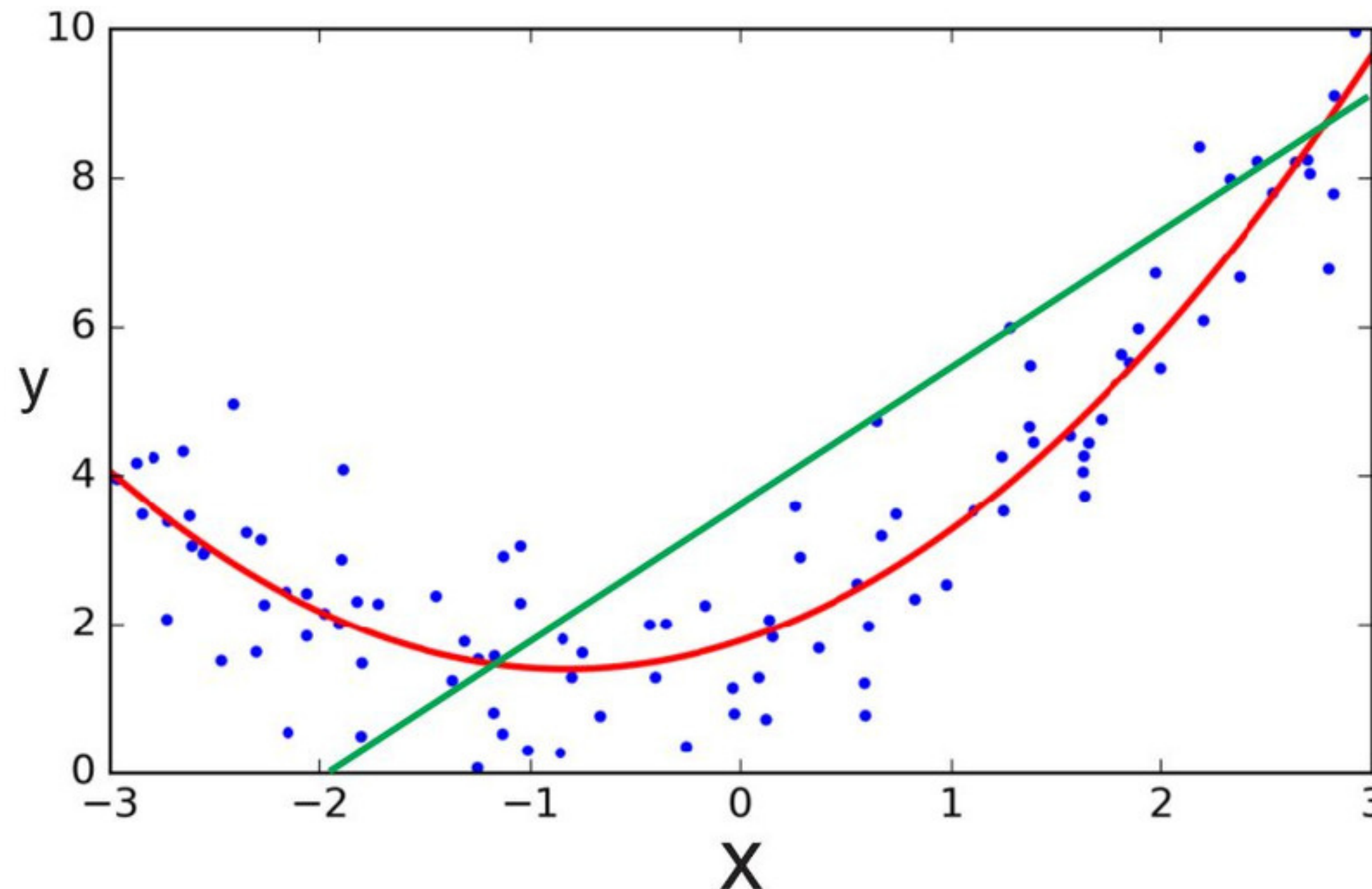
Número de parámetros
del modelo



Log-verosimilitud
del modelo

El término de
penalización
de BIC es
mayor al de
AIC

Regresión Polinomial (1/9)



¿La Regresión Lineal Simple es la mejor forma de aproximar la relación entre las variables X y Y?

Regresión Polinómica (2/9)

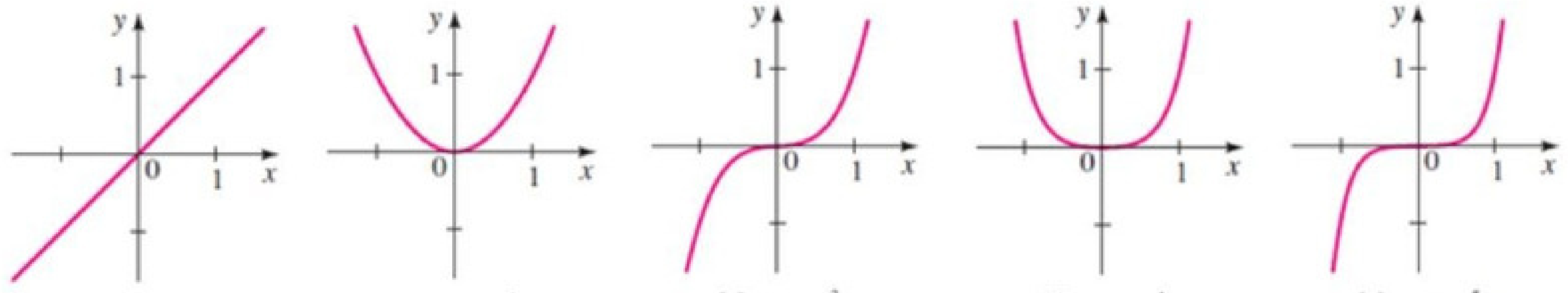
La Regresión Polinómica es un tipo de Análisis de Regresión en el que la relación entre la variable independiente X y la variable dependiente Y se modela como una función polinomial de grado n .

Modelo de Regresión Polinómico



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_n x^n + \epsilon$$

Regresión Polinómica (3/9)



Regresión Polinómica (4/9)

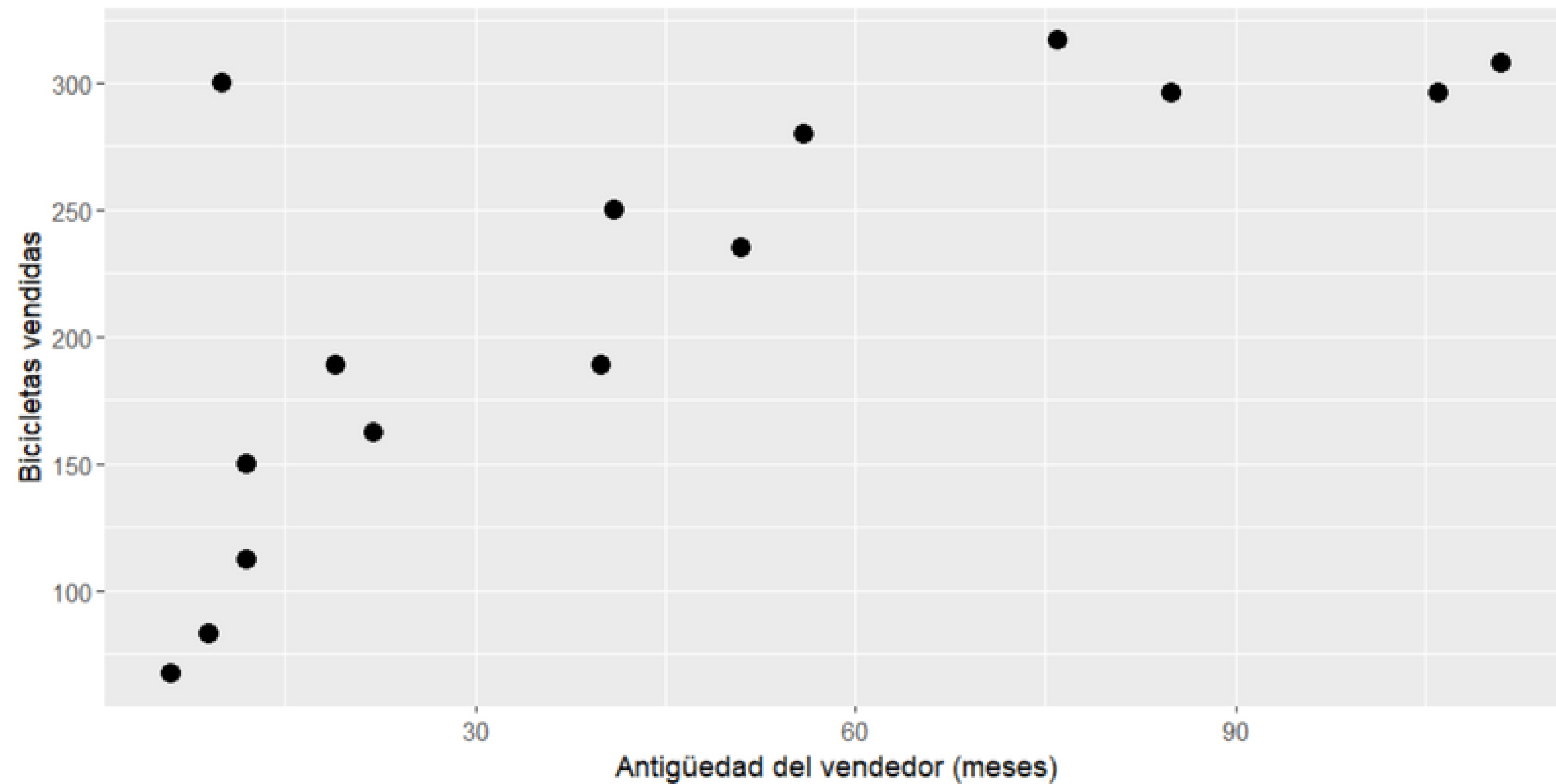
Vendedor i	Antigüedad (meses)	Bicicletas vendidas
1	41	375
2	106	296
3	76	317
4	10	300
5	22	162
6	12	150
7	85	296
8	111	308
9	40	189
10	51	235
11	9	83
12	12	112
13	6	67
14	56	280
15	19	189



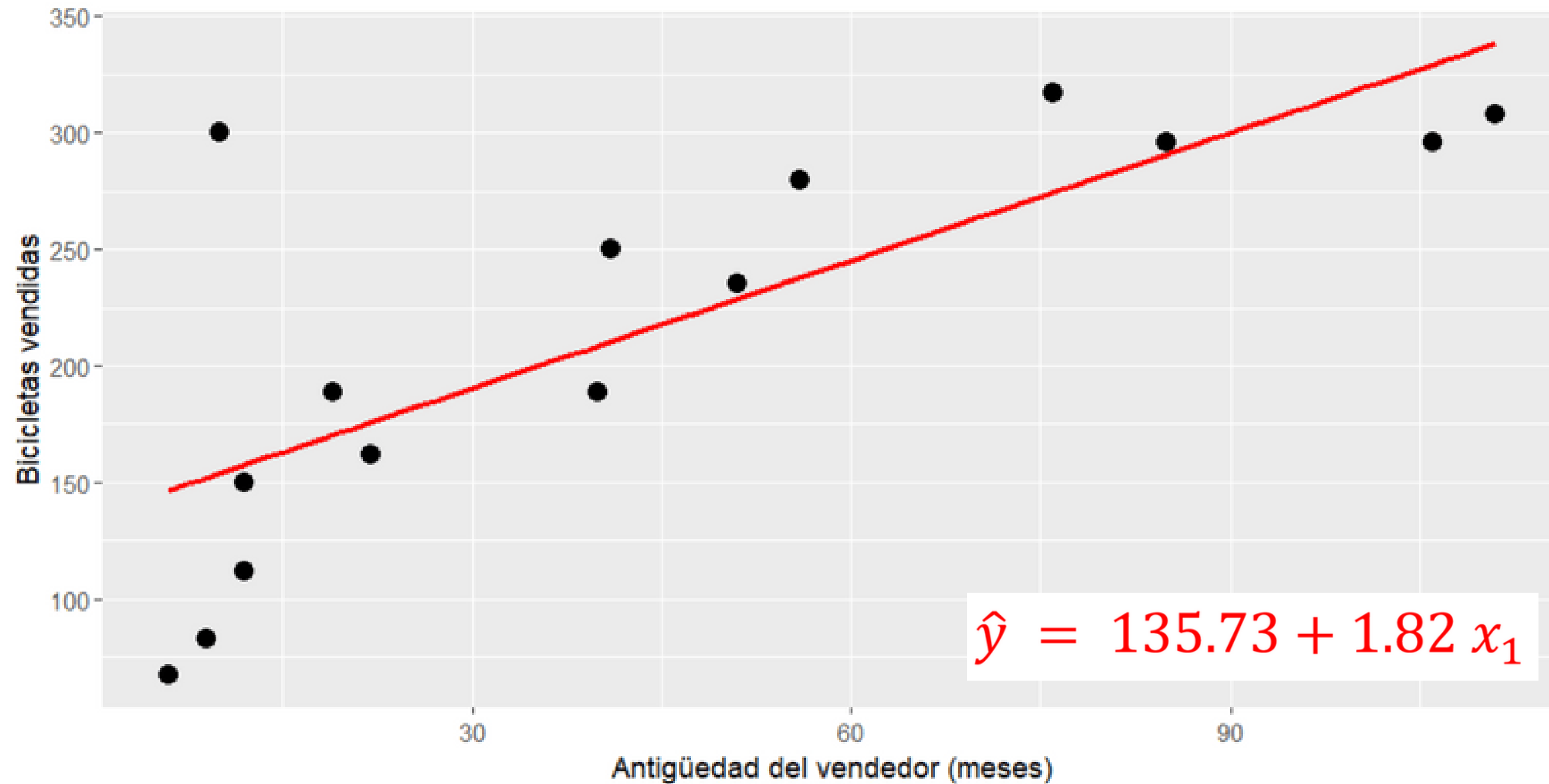
Creemos un modelo de regresión que nos permita modelar el número de bicicletas vendidas a partir de la antigüedad de los vendedores



Regresión Polinomial (5/9)



Regresión Polinomial (6/9)

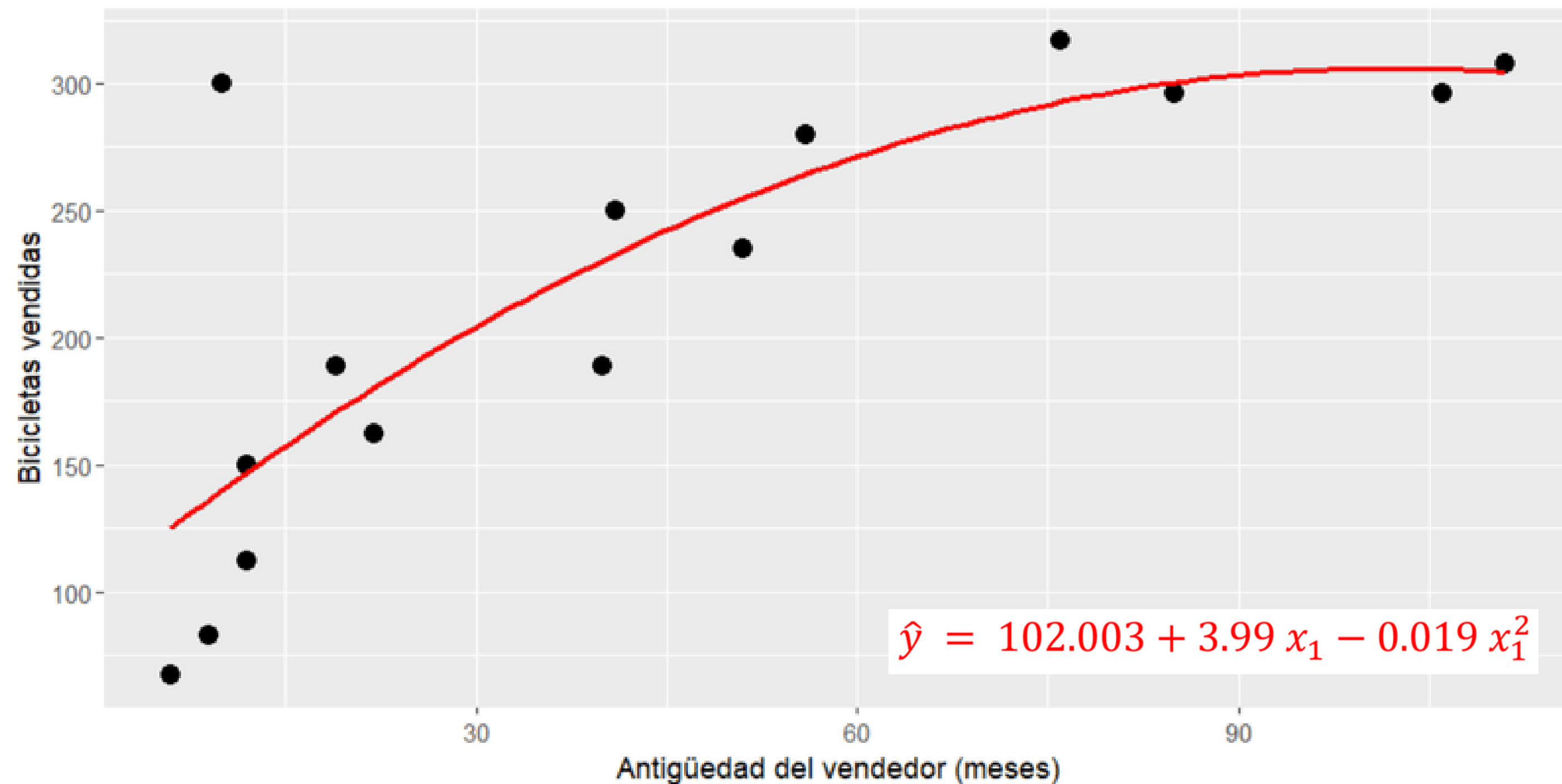


Regresión Polinomial (7/9)

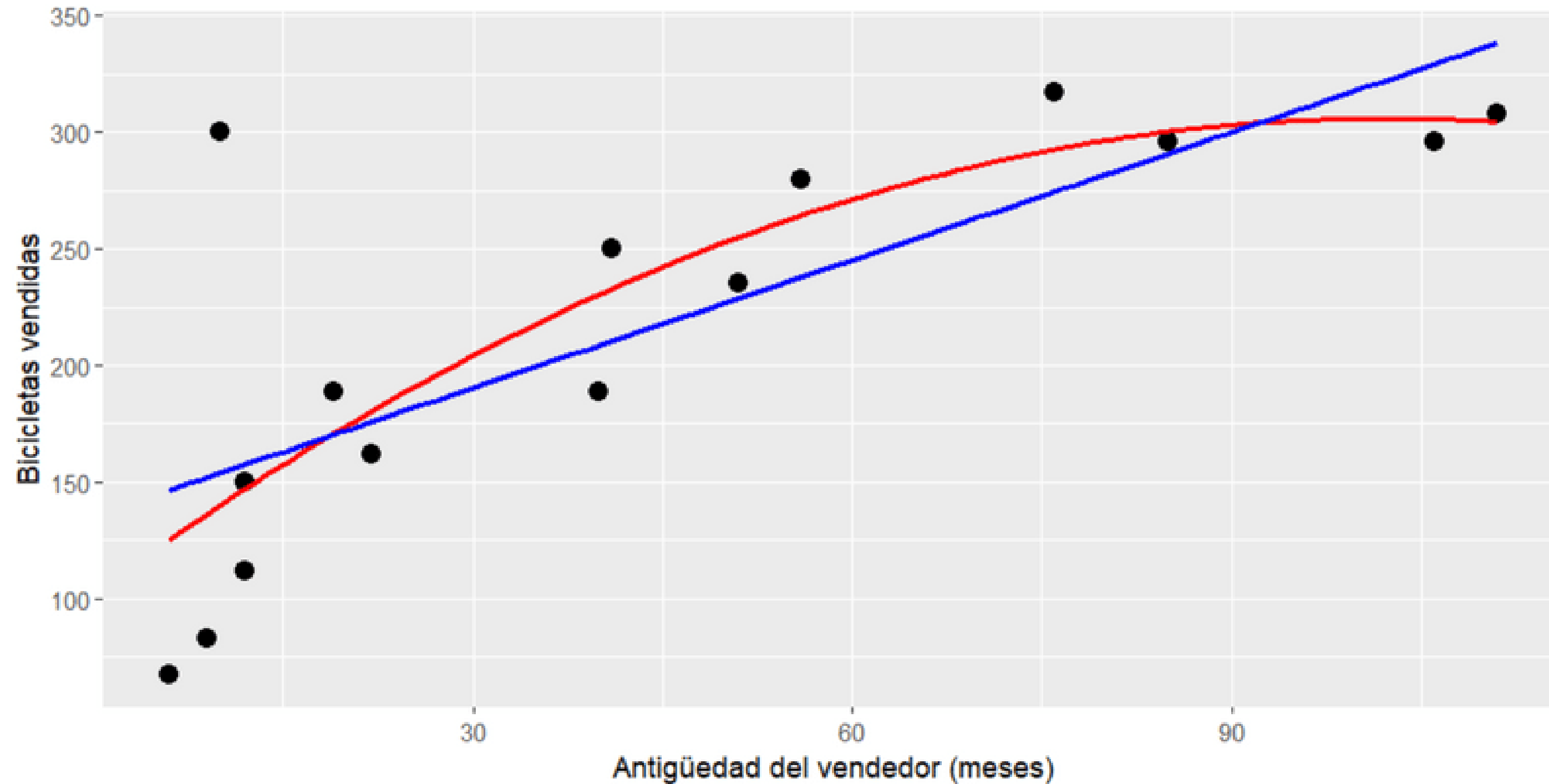
Usemos un modelo polinomial de grado 2 o de segundo orden para tratar de capturar la relación curvilínea que existe entre las dos variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

Regresión Polinomial (8/9)



Regresión Polinomial (9/9)



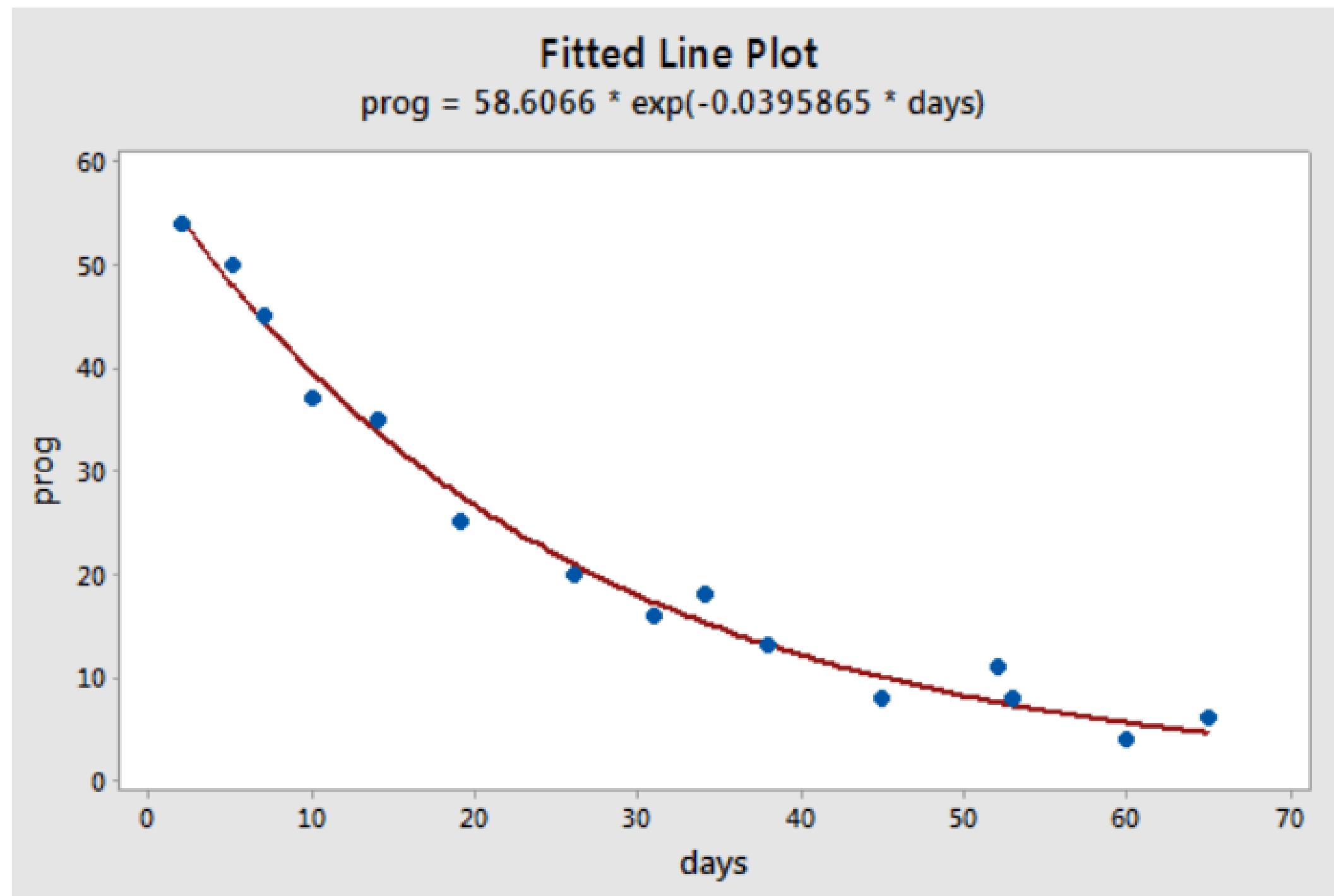
Modelos no lineales (1/9)

Los modelos no lineales son modelos en los que los parámetros $\beta_0, \beta_1, \dots, \beta_p$ tienen exponentes distintos de 1

Modelo Exponencial

$$y = \beta_0 e^{\beta_1 x}$$

Modelos no lineales (2/9)

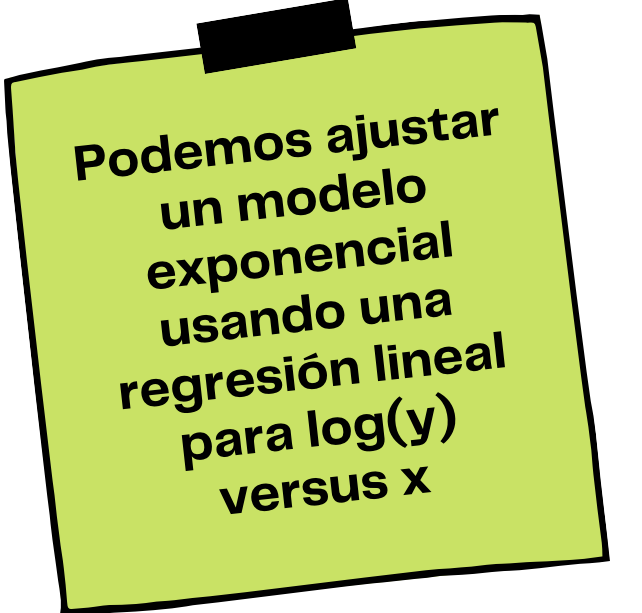


Modelos no lineales (3/9)

Muchos modelos no lineales pueden ser transformados a un modelo lineal equivalente

$$\log y = \log \beta_0 + \beta_1 x$$

$$y' = \beta'_0 + \beta_1 x$$

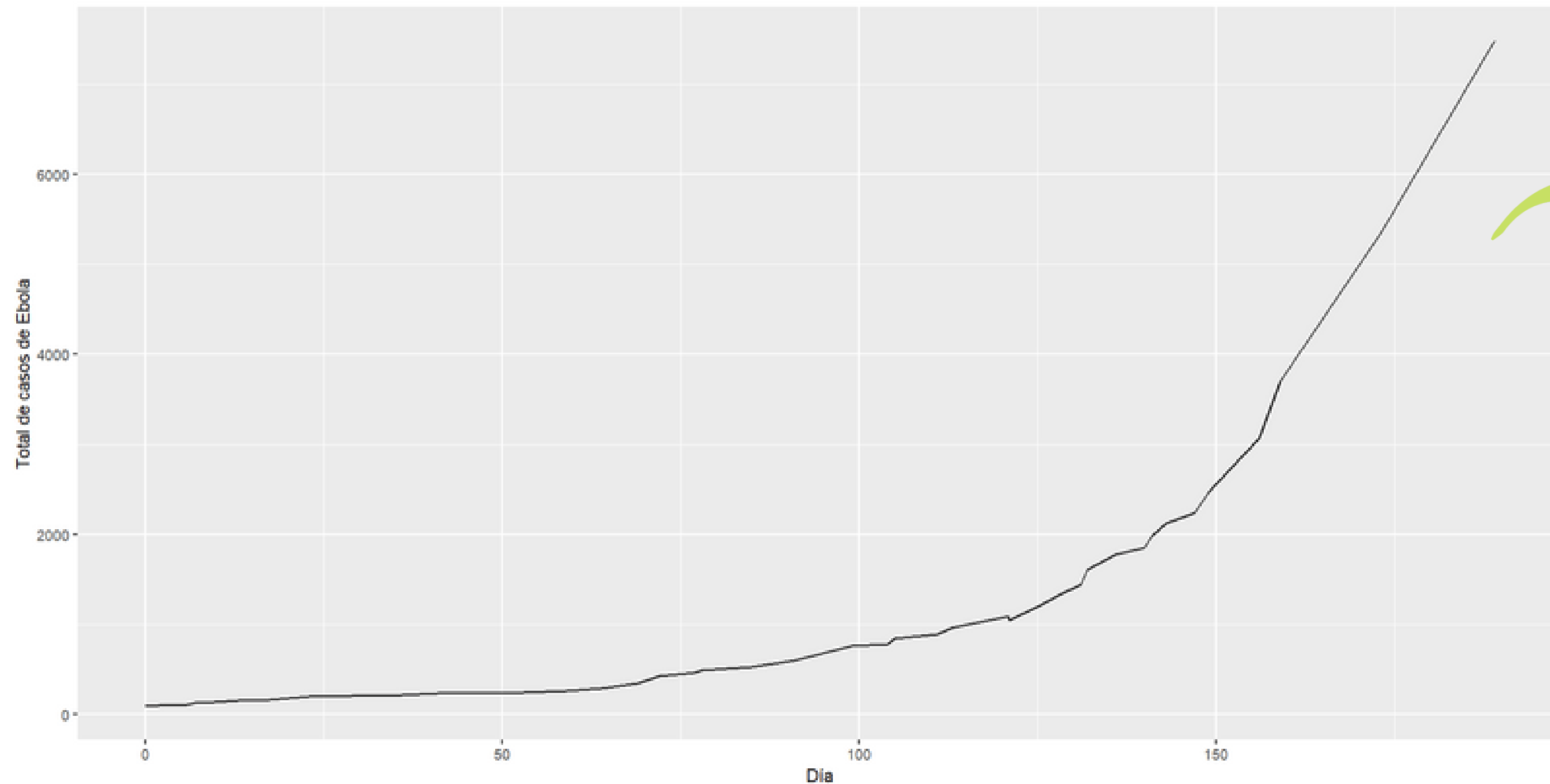


Podemos ajustar un modelo exponencial usando una regresión lineal para $\log(y)$ versus x

Modelos no lineales (4/9)

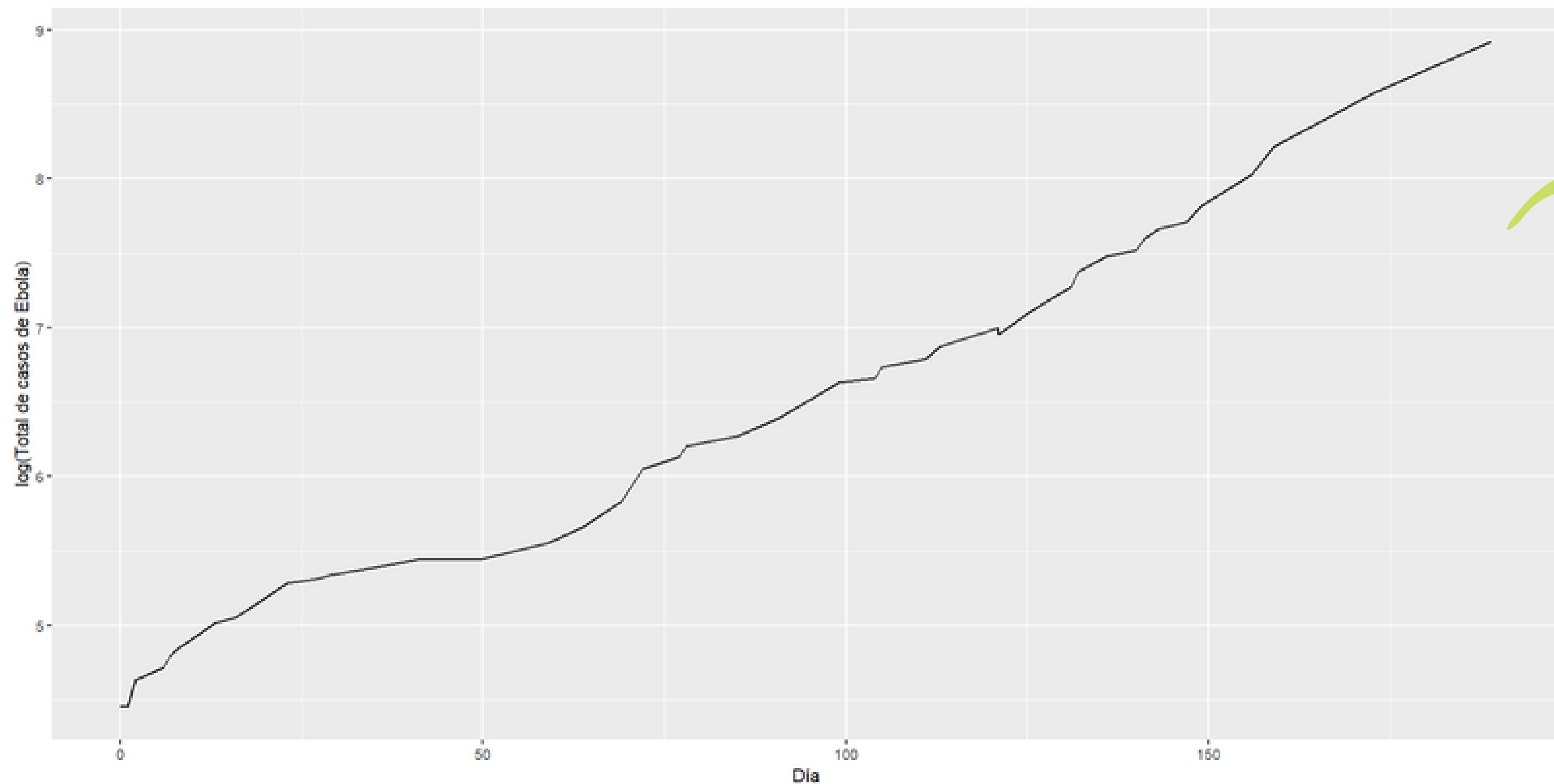
Utilicemos una base de datos con el número de casos de ébola en Guinea, Liberia, Níger, Sierra Leona y Senegal a lo largo del tiempo, a partir del 25 de marzo de 2014 mayor brote del virus del Ébola de la historia

Modelos no lineales (5/9)



**Esta relación
claramente NO es
lineal**

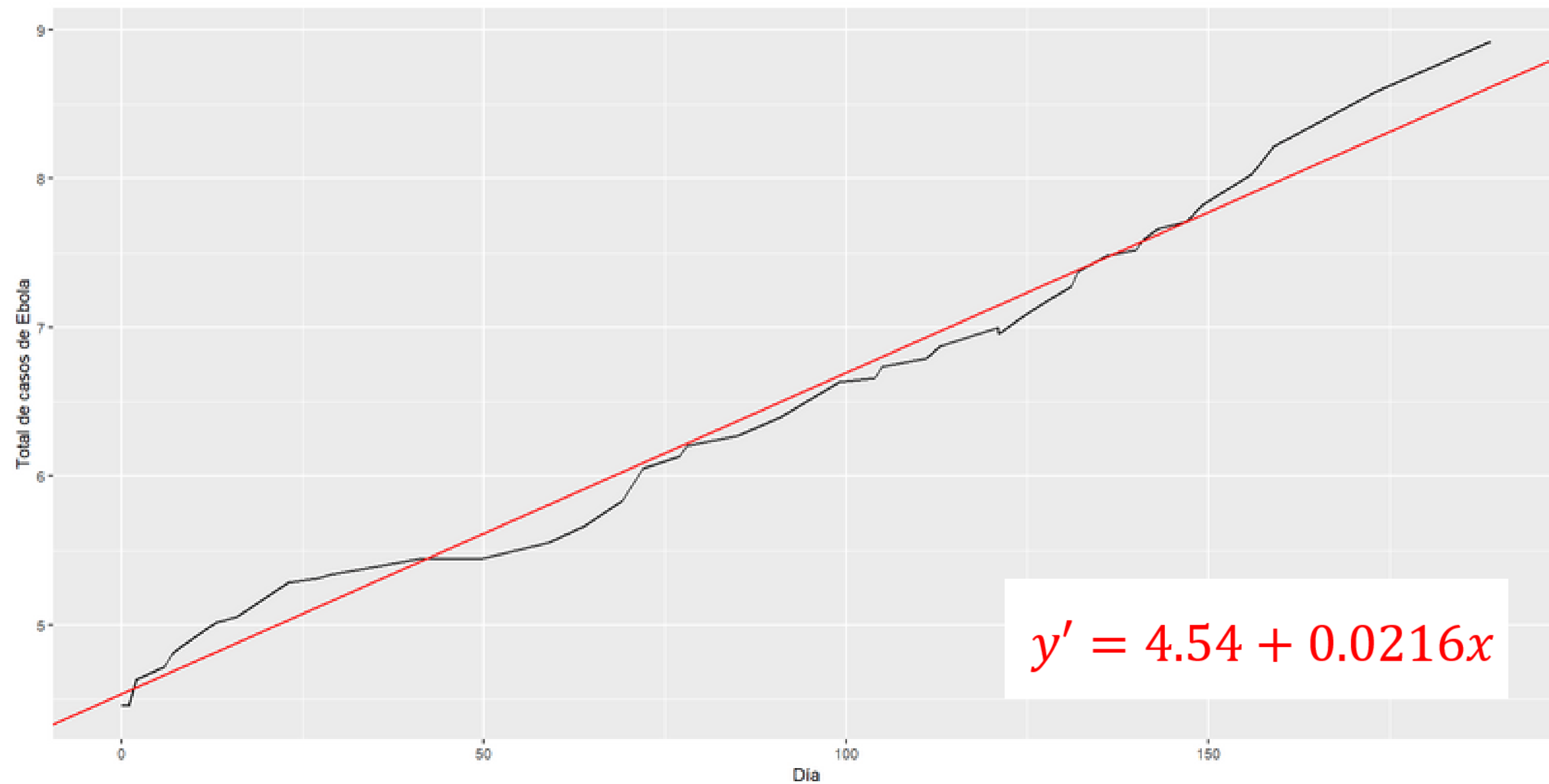
Modelos no lineales (6/9)



Al transformar la variable Y, se puede observar que la relación se aproxima a un comportamiento lineal

Ajustemos en R un modelo de regresión lineal cuya variable dependiente será la variable Y transformada...

Modelos no lineales (8/9)



Modelos no lineales (9/9)

$$y = \beta_0 e^{\beta_1 x} = 93.7 e^{0.0216x}$$



Intercepto
con el eje y



Tasa de crecimiento:
la tasa de crecimiento
diario en casos de ébola es
de aproximadamente 2,1%