

El modelo lineal

Regresión: modelos y métodos

Francesc Carmona Pontaque

PID_00298358



Universitat
Oberta
de Catalunya

Francesc Carmona Pontaque

Cómo citar este recurso de aprendizaje con el estilo Harvard:

Carmona Pontaque, F. (2024) *El modelo lineal. Regresión: modelos y métodos*. [Recurso de aprendizaje textual]. 1.^a ed. Barcelona: Fundació Universitat Oberta de Catalunya (FUOC).

Primera edición: febrero 2024

© de esta edición, Fundació Universitat Oberta de Catalunya (FUOC)

Av. Tibidabo, 39-43, 08035 Barcelona

Autoría: Francesc Carmona Pontaque

Producción: FUOC

Todos los derechos reservados

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Planteamiento y objetivos

Este primer módulo ejerce una función introductoria del curso. En él se presenta la noción de modelo lineal y las razones por las que resulta relevante su estudio.

El principal objetivo del módulo es plantear el concepto de modelo lineal como base de un gran número de modelos estadísticos, tales como la regresión o el análisis de la varianza. También se presentan las condiciones en que resulta adecuado utilizar modelos lineales y un ejemplo inicial desarrollado con el lenguaje de programación **R**.

En este documento los cálculos más complejos se han dejado en forma de apéndices. Cada estudiante puede decidir si profundiza en esos cálculos en función de sus conocimientos. Asimismo es interesante repasar o estudiar desde cero los principios básicos de álgebra matricial, imprescindible en el desarrollo de los siguientes módulos.

1. Un ejemplo

En el artículo de Prinzinger *et al.* (2002) se investigó la relación entre los latidos del corazón de ciertas aves en vuelo y su tasa metabólica, la cual resulta importante para cuestiones ecológicas. Los latidos del corazón por minuto se pueden medir en vuelo gracias al desarrollo de un sistema telemático. En cambio, la tasa metabólica se debe medir en el laboratorio. Por eso es interesante saber si es posible inferir esa tasa a partir del número de latidos por minuto. De los muchos datos del trabajo, vamos a utilizar como ejemplo únicamente los que se refieren a algunos buitres leonados (*griffon vultures*) que se tomaron el día 1999-05-17. Los datos son:

```
heartbpm <- c(47.53, 48.27, 49.51, 51.09, 52.57, 54.30,
             54.25, 54.45, 57.95, 60.92, 61.91, 77.92,
             82.07, 82.95, 83.94, 86.96, 90.42, 92.93, 100.05)
metabol <- c(6.15, 6.31, 6.43, 6.78, 6.86, 6.90, 7.37, 7.41,
            8.24, 9.22, 8.16, 12.61, 15.26, 13.09, 14.59,
            17.35, 18.57, 19.00, 20.70)
vulture <- data.frame(heartbpm, metabol)
rm(heartbpm, metabol)
attach(vulture)
```

Los datos en R

Con estas instrucciones creamos un *data.frame* que contiene los datos de las dos variables. La instrucción `attach()` permite acceder a las variables directamente (Verzani, 2002, p. 24).

Con estos datos, dibujamos un gráfico de dispersión como el de la figura 1.

```
library(ggplot2)
p <- ggplot(vulture, aes(x=heartbpm, y=metabol)) +
  geom_point() + labs(x = "heart beats (per minute)",
                    y = "metabolic rate [J/(g*h)]",
                    title = "Griffon vulture, 1999-05-17") +
  theme_light()
p
```

Una decisión importante

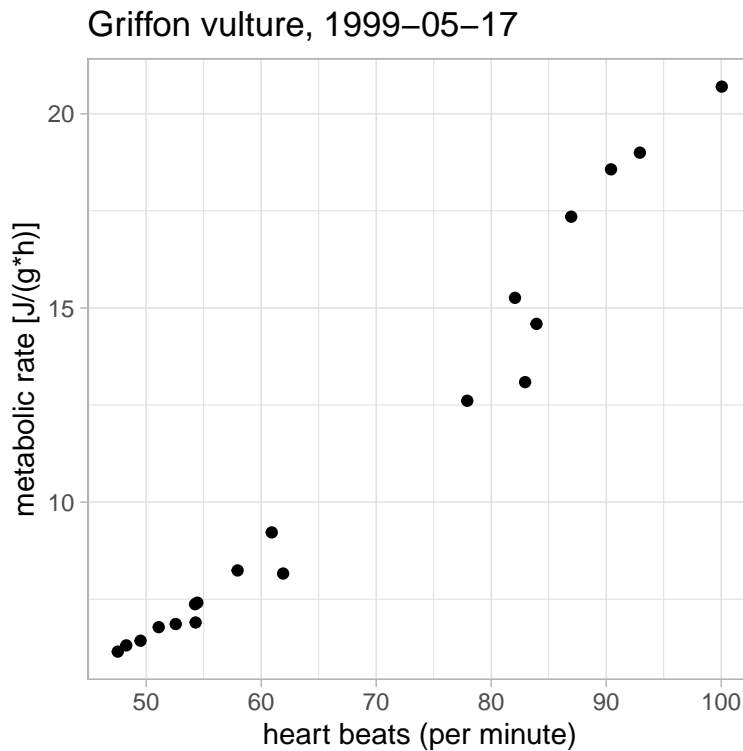
En cualquier regresión hay que decidir previamente cuál es la variable explicativa, también llamada regresora o independiente, y la variable respuesta, también llamada dependiente. Normalmente no son intercambiables.

El objetivo es hallar una recta que “se ajuste” a estos datos. Así, podremos explicar la relación entre las dos medidas con una función lineal muy simple, es decir, con una recta.

Para empezar y como primera aproximación, podríamos tomar la recta que une dos puntos representativos. Por ejemplo y ya que los datos están casi ordenados, podemos elegir los puntos cuarto y dieciseisavo de la muestra. La recta que pasa por esos dos puntos es $y - y_1 = m \cdot (x - x_1)$, que en este caso es $y = 0.295x - 8.275$.

Recta que pasa por dos puntos

La recta $y = mx + b$ que pasa por los puntos (x_1, y_1) y (x_2, y_2) tiene pendiente $m = (y_2 - y_1)/(x_2 - x_1)$.

**Figura 1**

Nube de puntos o diagrama de dispersión (*scatter plot*) con la variable independiente (latidos del corazón por minuto) en el eje horizontal y la variable dependiente (tasa metabólica) en el eje vertical.

```
p1 <- as.numeric(vulture[4,])
p2 <- as.numeric(vulture[16,])
dif <- p2 - p1
m <- dif[2]/dif[1]
b <- p1[2] - m * p1[1]
```

Con la siguiente instrucción podemos incorporar esta recta al gráfico de dispersión anterior.

```
p + geom_abline(intercept = b, slope = m)
```

Sin embargo, el objetivo que nos proponemos es hallar la mejor de las rectas por algún criterio.

Como veremos, el método de los mínimos cuadrados proporciona una recta que goza de muy buenas propiedades estadísticas, siempre que se cumplan algunas condiciones. Este método consiste en hallar los coeficientes m y b tales que hagan mínima la suma de los errores al cuadrado

$$e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - mx_i - b)^2$$

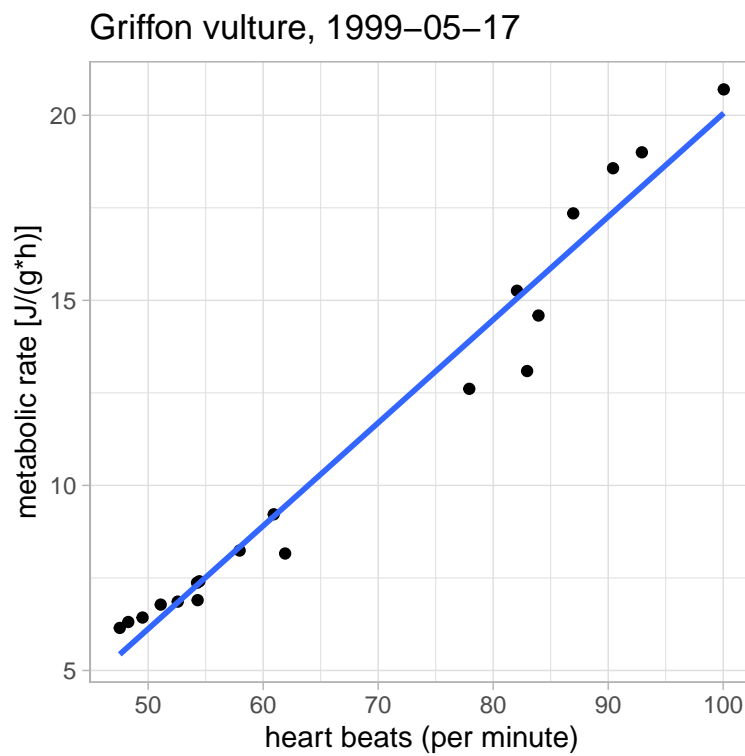
donde cada error e_i es la diferencia entre el valor observado de la respuesta y_i y el valor que nos da la recta para x_i , es decir, $\hat{y}_i = mx_i + b$.

Otras rectas

Es interesante conocer otros métodos para ajustar una recta a la nube de puntos. Métodos como la regresión lineal ortogonal, la regresión Deming o la recta resistente de los tres grupos. En **R** disponemos de funciones que calculan estas rectas.

Podemos añadir esta recta al gráfico de dispersión con la siguiente instrucción:

```
p + stat_smooth(method="lm", se=FALSE)
```



La función `stat_smooth()`

Con esta función añadimos la recta con el método "lm" (*linear model*). El parámetro `se = FALSE` se añade para que no se dibujen los intervalos de confianza.

Más adelante veremos cómo se calculan exactamente los coeficientes m y b , pero una forma de hacerlo con **R** es:

```
ajuste.mc <- lsfit(heartbpm, metabol)
coef(ajuste.mc)
```

```
Intercept      X
-7.7966734  0.2784028
```

Ahora vamos a comprobar que efectivamente la suma de errores al cuadrado es menor para la recta mínimo cuadrática o MC (o LS, siglas de *least squares*).

```
rmse <- function(y, y.ast) sqrt(sum((y - y.ast)^2))
rmse(metabol, m * heartbpm + b)      # recta entre dos puntos
```

```
[1] 4.814855
```

```
sqrt(sum(residuals(ajuste.mc)^2))    # recta MC
```

```
[1] 3.770082
```

La función `lsfit()`

Esta función calcula el ajuste LS para unos datos.

La función `rmse()`

La función *root mean square error* calcula la raíz cuadrada de la suma de los errores al cuadrado. El resultado tiene las mismas unidades que la variable respuesta.

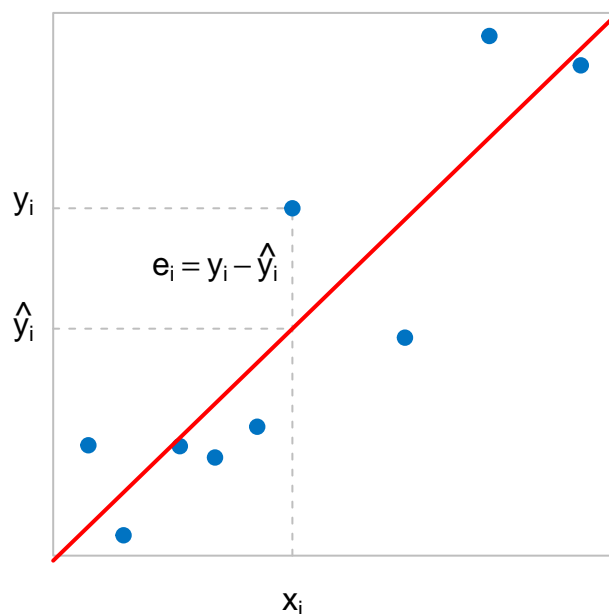
La función `residuals()`

Calcula los residuos o errores entre los valores respuesta observados y los valores de la recta ajustada por la función `lsfit()`.

2. El método de los mínimos cuadrados

La paternidad de este método se reparte entre Legendre que lo publicó en 1805 y Gauss que lo utilizó en 1795 y lo publicó en 1809.

Vamos a considerar los residuos entre los valores observados y los valores calculados según la recta, tal como se ve en el siguiente gráfico.



Obviamente, cuanto menores son los residuos, mejor es el ajuste. De todos los posibles valores de β_0 y β_1 , el método de los mínimos cuadrados selecciona aquellos que minimizan la expresión

$$S(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Gracias al análisis matemático sabemos que para hallar el mínimo de una función derivable, debemos derivar e igualar a cero las dos derivadas parciales. Los cálculos se pueden ver en el apéndice al final de este documento.

Lectura complementaria

En el apartado 1.4 del libro de Faraway (2014) podemos ver dos ejemplos históricos ilustrativos.

Notación

Para extender la regresión al caso múltiple con varias variables regresoras, es mejor utilizar letras griegas para los parámetros de pendiente $\beta_1 = m$ y de intercepto $\beta_0 = b$.

Estas derivadas igualadas a cero se llaman *ecuaciones normales* y son

$$\begin{aligned}\beta_0 n + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i\end{aligned}$$

La solución de este sistema de ecuaciones es (ver Apéndice A)

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

Así pues, la ecuación de la recta de regresión MC o LS es


$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

Con las estimaciones de los parámetros, podemos proceder al cálculo de predicciones \hat{y}_i y residuos e_i .

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ e_i &= y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

Como consecuencia resulta que

$$\sum_{i=1}^n e_i = 0$$

lo que no ocurre en un modelo sin β_0 . 

Finalmente, si queremos una medida del ajuste de la regresión podemos pensar en la suma de cuadrados $\sum_{i=1}^n e_i^2$, pero es una medida que depende de las unidades de y_i al cuadrado. En general, la medida que se utiliza es el coeficiente de determinación

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Más notación

Como es habitual, las medias de los valores x_i y y_i son \bar{x} y \bar{y} respectivamente.

Además, s_x^2 representa la varianza muestral de las x_i sin corregir, es decir, la suma de cuadrados dividida por n . Del mismo modo, el numerador de $\hat{\beta}_1$ también va dividido por n y, por ello, simplifican.

Resultado

La recta de regresión se puede escribir en la forma

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x})$$

lo que indica claramente que esta recta pasa por el punto (\bar{x}, \bar{y}) .

¡Atención!

Esta propiedad de los residuos no se verifica en una recta que obligamos a pasar por el origen de coordenadas, es decir, si no tiene β_0 .

Coeficiente de correlación

Aunque el coeficiente de determinación coincide con el coeficiente de correlación al cuadrado, ambos son conceptualmente distintos.

Sabemos que $0 \leq R^2 \leq 1$ y cuando $R^2 \approx 1$ el ajuste es bueno.

En el caso de considerar una recta sin β_0 , el coeficiente de determinación apropiado es

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n y_i^2}$$

Con los datos del ejemplo de los buitres leonados, los cálculos con **R** se pueden hacer con la función `lm()`.

```
mod <- lm(metabol ~ heartbpm, data=vulture)
coef(mod)

(Intercept)    heartbpm
-7.7966734    0.2784028

summary(mod)$r.squared

[1] 0.9688697
```

Los mismos resultados se pueden obtener con la matriz de diseño (ver Apéndice B) y un sencillo cálculo matricial. La función `solve()` resuelve sistemas de ecuaciones lineales.

```
X <- model.matrix(mod)
Y <- metabol
XtX <- crossprod(X)      # == t(X) %*% X
XtY <- crossprod(X,Y)    # == t(X) %*% Y
solve(XtX, XtY)

           [,1]
(Intercept) -7.7966734
heartbpm     0.2784028
```

¡Atención!

Los modelos que carecen de término β_0 no se pueden comparar con los que sí lo tienen, ya que los coeficientes de determinación son distintos.

La función `lm()`

Esta función se escribe con una fórmula de **R** del tipo `y ~ x`. Otras fórmulas se pueden ver en Verzani (2002, p. 35). Por defecto, la función `lm()` incorpora siempre el intercepto. Para eliminarlo hay que escribir una fórmula del tipo `y ~ 0 + x` o `y ~ -1 + x`.

El algoritmo

En realidad la función `lm()` utiliza la descomposición QR para calcular los coeficientes. Los detalles se pueden ver en el apartado 2.7 de Faraway (2014).

3. El modelo lineal

3.1. Definición

Hasta aquí, el ajuste a una recta y el método de los mínimos cuadrados parece un problema matemático. ¿Donde está la estadística?

A lo largo de esta asignatura vamos a ver que, si planteamos el mismo problema como un modelo estadístico y le imponemos algunas condiciones, podremos realizar estimaciones óptimas, contrastes de hipótesis y estudiar la bondad del ajuste desde esta óptica.

Cuando en el ejemplo anterior ajustamos los datos a una recta, implícitamente estamos asumiendo la hipótesis de que los datos siguen un patrón lineal subyacente del tipo $y = \beta_0 + \beta_1 x$. Pero el ajuste no es perfecto y contiene errores. La ecuación que define el modelo es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

donde ϵ_i son los errores aleatorios. Este es el modelo de **regresión simple** o con una sola variable independiente.

En el mismo ejemplo con los buitres leonados, podemos pensar en ajustar los datos a un modelo parabólico

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \quad i = 1, \dots, n$$

que continúa siendo un *modelo lineal*.

Un modelo es lineal si lo es para los parámetros. Por ejemplo, el modelo $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$ es lineal, mientras que $y_i = \beta_0 \exp(-\beta_1 x_i) + \epsilon_i$ no.

En general, suponemos que una cierta variable aleatoria Y es igual a un valor fijo η más una desviación aleatoria ϵ

$$Y = \eta + \epsilon$$

η representa la verdadera medida de la variable, es decir, la parte *determinista* de un experimento, que depende de ciertos factores cualitativos y variables cuantitativas que son controlables por la persona que experimenta.

Modelo lineal

Un modelo es lineal cuando cada parámetro multiplica a una variable o función de variables y se suman a otros de la misma forma. Además, en el modelo lineal el error es aditivo.

El término ϵ representa el *error*. Es la parte del modelo no controlable por la persona que experimenta debido a múltiples causas aleatorias, inevitables en los datos que proceden de la Biología, Psicología, Economía, Medicina,... El error ϵ convierte la relación matemática $Y = \eta$ en la relación estadística $Y = \eta + \epsilon$, obligando a tratar el modelo desde la perspectiva del análisis estadístico.

En particular, los modelos de la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \dots, n$$

con $k > 1$ variables independientes, predictoras o regresoras, se llaman modelos de **regresión múltiple**. La variable cuyos datos observados son y_i es la llamada variable dependiente o respuesta.

Los parámetros β_j son desconocidos y nuestro objetivo principal es su estimación. En cuanto a los errores ϵ_i , su cálculo explícito en forma de residuos nos permitirá, como veremos extensamente, la evaluación del modelo.

3.2. Las condiciones de Gauss-Markov

Ahora ya hemos planteado el ajuste de unos datos a una función lineal de unos parámetros como un modelo estadístico. ¿Cómo estimamos esos parámetros?

Fácil: con el método de los mínimos cuadrados. Llegado este punto, además, nos podemos hacer otra importante pregunta. ¿Qué tan bueno es el método de los mínimos cuadrados para estimar los parámetros? La respuesta es que este método proporciona un buen ajuste y buenas predicciones si se verifican las condiciones que llamamos de Gauss-Markov.

En el modelo lineal que hemos definido anteriormente, se supone que los errores ϵ_i son desviaciones que se comportan como variables aleatorias. Vamos a exigir que estos errores aleatorios verifiquen las siguientes condiciones:

1. Los errores tienen media teórica (esperanza) cero.
2. La varianza σ^2 de todos los errores es la misma.
3. Los errores son incorrelacionados entre sí.

Observación

Consideraremos los valores x_i como constantes y no como observaciones de una variable aleatoria, aunque lo fueran. Para más detalles, consúltese Carmona (2005, p. 18).

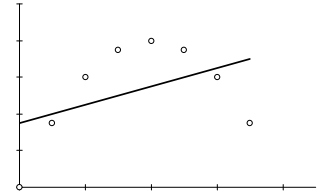
Veamos con detalle estas condiciones:

Primera condición

$$E(\epsilon_i) = 0 \quad i = 1, \dots, n$$

Se trata de una condición natural sobre un error.

De este modo nos aseguramos que $E(y_i) = \beta_0 + \beta_1 x_i$, de forma que el modelo lineal es correcto y la situación que representa el gráfico no se puede dar.

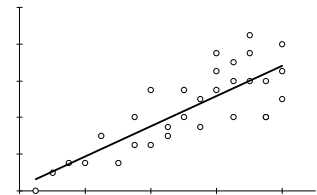


Segunda condición

$$\text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2 \quad i = 1, \dots, n$$

Es la propiedad de la *homocedasticidad*.

En el gráfico se representa una situación anómala llamada *heterocedasticidad*, en la que la $\text{var}(\epsilon_i)$ crece con x_i .



El parámetro desconocido σ^2 es la llamada varianza del modelo.

Tercera condición

$$E(\epsilon_i \epsilon_j) = 0 \quad \text{para todo } i \neq j$$

Las observaciones deben ser incorrelacionadas. Con dos puntos tenemos una recta de regresión. Con 20 copias de esos dos puntos, tenemos 40 puntos y la misma recta, poco representativa.

Las tres condiciones se pueden resumir en forma matricial como

$$E(\epsilon) = \mathbf{0} \quad \text{var}(\epsilon) = \sigma^2 \mathbf{I}_n$$

donde $E(\epsilon)$ es el vector de esperanzas matemáticas y $\text{var}(\epsilon)$ es la matriz de covarianzas de $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$. La matriz \mathbf{I}_n es la llamada matriz identidad, con unos en la diagonal y ceros en las otras posiciones. Así pues, la matriz de covarianzas de los errores es una matriz diagonal con σ^2 en la diagonal y ceros en las otras posiciones.

Como veremos en los siguientes módulos, la adopción de estas condiciones debería evitar teóricamente situaciones anómalas como la presencia de observaciones influyentes o atípicas y, sobre todo, garantizar la mejor estimación de los parámetros.

Errores independientes

Si los errores son estocásticamente independientes, también son incorrelacionados. Al revés, sin embargo, no siempre es cierto.

3.3. Otros tipos de modelos lineales

Como veremos ampliamente, con el mismo tratamiento podremos resolver otros modelos lineales, que aunque tienen diferentes objetivos, gozan de las

mismas bases teóricas.

Por ejemplo, el **análisis de la varianza** con un factor (*one-way Analysis of Variance*), representado por el modelo lineal

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{con } \epsilon_{ij} \sim N(0, \sigma^2) \text{ indep.,}$$

se resuelve de forma similar al modelo de regresión.

El **análisis de la covarianza**, que utiliza como variables independientes tanto variables cuantitativas como factores, generaliza la aplicación de los modelos lineales que vamos a estudiar.

También la **regresión logística**, que considera una variable respuesta dicotómica, se puede ver como un caso especial de modelo lineal.

4. Algunas preguntas

Un típico problema de estadística consiste en estudiar la relación que existe, si existe, entre dos variables aleatorias X e Y . Por ejemplo, altura y peso de una persona, altura de la madre y altura de la hija, longitud y anchura de unas hojas, temperatura y presión de un determinado volumen de gas, etc.

Si tenemos n pares de observaciones (x_i, y_i) $i = 1, 2, \dots, n$, podemos dibujar estos puntos en un gráfico o *scatter diagram* y tratar de ajustar una curva a los puntos, de forma que los puntos se hallen lo más *cerca* posible de la curva. No podemos esperar un ajuste perfecto porque ambas variables están expuestas a fluctuaciones al azar debido a factores incontrolables. Incluso, aunque en algunos casos pudiera existir una relación exacta entre variables físicas como temperatura y presión, también aparecerían fluctuaciones debidas a errores de medida. En muchos casos, la solución es un modelo lineal.

Algunas cuestiones que podemos plantearnos en este tipo de investigaciones son:

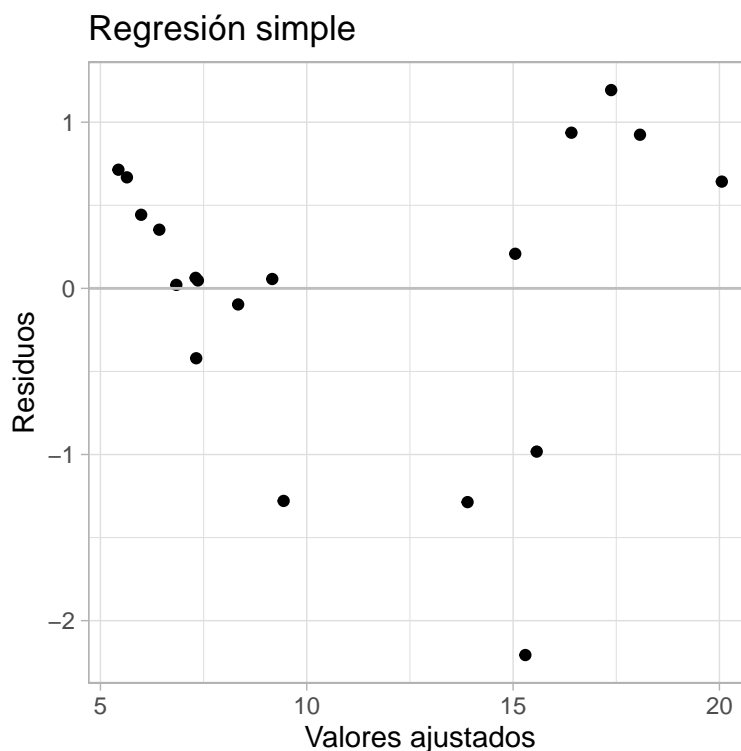
- Si existe un modelo físico teórico y lineal, podemos utilizar la regresión para estimar los parámetros.
- Si el modelo teórico no es lineal, podemos probar una transformación. Por ejemplo: $PV^\gamma = c \rightarrow \log P = \log c - \gamma \log V$
- Si no es una recta, se puede estudiar un modelo de regresión polinómico. ¿De qué grado?
- En el modelo múltiple intervienen varias variables “predictoras”. ¿Son todas necesarias? ¿Son linealmente independientes las llamadas “variables independientes”?
- ¿Se verifican realmente las condiciones de Gauss-Markov?
- ¿Qué ocurre si las variables predictoras son discretas?
- ¿Qué ocurre si la variable dependiente es discreta o una proporción?
- ¿Y si faltan algunos datos?
- ¿Qué hacemos con los llamados puntos atípicos (residuo muy grande) y los puntos influyentes?

Algunas de estas preguntas las iremos trabajando y resolviendo en los siguientes módulos, otras pueden quedar para una posterior profundización.

Regresión parabólica

En el ejemplo de los buitres leonados con el que empezamos este módulo, nos podemos preguntar si el ajuste a una recta es suficiente o no. Una primera aproximación para responder a la pregunta consiste en realizar un análisis de los residuos. Más adelante trataremos este análisis en profundidad. Por ahora, vamos a dibujar un gráfico de dispersión con los residuos vistos desde los valores de predicción.

```
ggplot(mod, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, col="gray") +
  labs(title='Regresión simple',
       x='Valores ajustados', y='Residuos') +
  theme_light()
```



En el gráfico observamos que los residuos no aparecen alrededor del cero de forma “aleatoria” y visualmente dibujamos una parábola. Tal vez, podemos pensar en ajustar una parábola o polinomio de grado dos.

Para ello consideramos el modelo lineal con dos variables regresoras x y x^2 .

```
mod2 <- lm(metabol ~ heartbpm + I(heartbpm^2), data=vulture)
```

Para ver si el modelo parabólico se ajusta mejor que la regresión simple podemos, entre otras cosas, comparar el gráfico de los residuos.

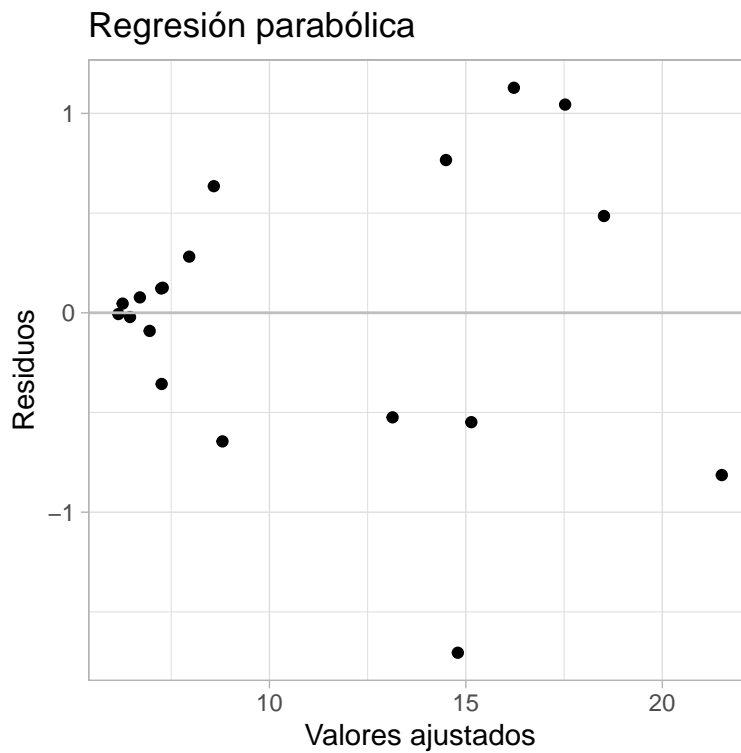
Nota

La comparación de los coeficientes de determinación no es la mejor forma para decidir el modelo. Como veremos, cuantas más variables regresoras, mejor es el R^2 .

La función `I()`

Esta función permite realizar un cálculo e incorporar el resultado como una nueva variable a una función `lm()`.

```
ggplot(mod2, aes(x = .fitted, y = .resid)) +  
  geom_point() +  
  geom_hline(yintercept = 0, col="gray") +  
  labs(title='Regresión parabólica',  
        x='Valores ajustados', y='Residuos') +  
  theme_light()
```



Ahora los residuos están mejor repartidos alrededor del cero, pero aparece otro problema. La variabilidad de los residuos para valores ajustados pequeños es menor que para los grandes. Esto es un indicio de heterocedasticidad. Así pues, habrá que seguir mejorando el modelo, tal vez con una transformación de la variable respuesta.

Bibliografía

Carmona, F. (2005) *Modelos lineales*. e-UMAB, Universitat de Barcelona.

Faraway, J.J. (2014) *Linear Models with R*. 2.^a ed. Chapman and Hall/CRC.

Prinzinger, R., Nagel, B., Bahat, O., Bögel, R., Karl, E., Weihs, D. y Walzer, C. (2002) *Energy metabolism and body temperature in the Griffon Vulture (Gyps fulvus) with comparative data on the Hooded Vulture (Necrosyrtes monachus) and the White-backed Vulture (Gyps africanus)*. Journal für Ornithologie, 143(4), 456-467.

Verzani, J. (2002) *simpleR - Using R for Introductory Statistics*. College of Staten Island. Disponible en: <https://biostat.jhsph.edu/~iruczins/teaching/Rresources/simpleR.pdf>

Verzani, J. (2014) *Using R for Introductory Statistics*. 2.^a ed. Chapman and Hall/CRC.

Apéndice A: Estimaciones mínimo cuadráticas

Vamos a hallar las estimaciones de los parámetros en un modelo de regresión lineal simple, minimizando la suma de los cuadrados de los errores.

$$S = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

La función S depende de dos incógnitas (β_0, β_1) , mientras que las (x_i, y_i) son las observaciones que se consideran conocidas. Entonces, para hallar el mínimo de esa función habrá que derivar e igualar a cero:

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-1) \\ \frac{\partial S}{\partial \beta_1} &= 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-x_i) \end{aligned}$$

y al igualar a cero obtenemos

$$\begin{aligned} - \sum_{i=1}^n y_i + n\beta_0 + \beta_1 \sum_{i=1}^n x_i &= 0 \\ - \sum_{i=1}^n x_i y_i + \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= 0 \end{aligned}$$

que son las llamadas *ecuaciones normales*:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Para resolver este sistema, dividimos la primera ecuación por n :

$$\beta_0 + \bar{x}\beta_1 = \bar{y} \quad \Rightarrow \quad \beta_0 = \bar{y} - \bar{x}\beta_1$$

y sustituimos en la segunda

$$(\bar{y} - \bar{x}\beta_1) \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

de donde

$$\beta_1 \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \quad (1)$$

Por otra parte,

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - n \bar{x} \bar{y} + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \end{aligned} \quad (2)$$

y también

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - 2n \bar{x}^2 + n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n \bar{x}^2 = \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \end{aligned} \quad (3)$$

De modo que podemos sustituir (2) y (3) en (1) y tenemos

$$\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

es decir

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

donde $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ y $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

Además $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$ y el sistema queda solucionado.

Esta solución única proporciona el mínimo de la función S , tal como se puede comprobar con las derivadas parciales de segundo orden.

Las predicciones son

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \bar{x}\hat{\beta}_1 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

y los residuos

$$e_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$$

Apéndice B: la matriz de diseño

Aunque hemos visto funciones de **R** que calculan directamente los parámetros de la recta y otros elementos como el coeficiente de determinación, vamos a comprobar con nuestros propios cálculos la solución a las ecuaciones normales.

Para ello, observamos que en las ecuaciones normales aparecen unas sumas como

$$\sum_{i=1}^n 1 = n, \quad \sum_{i=1}^n x_i, \quad \sum_{i=1}^n x_i^2, \quad \sum_{i=1}^n y_i, \quad \sum_{i=1}^n x_i y_i$$

y este sistema se puede escribir matricialmente así

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Estas sumas se pueden obtener con un sencillo cálculo matricial a partir de la llamada *matriz de diseño*, la cual se muestra a continuación:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

La matriz de diseño

El nombre tiene pleno sentido en los llamados diseños experimentales de las *ciencias naturales*, donde las variables son cualitativas (factores) y sus valores son ceros y unos.

de modo que $\mathbf{X}'\mathbf{X}$ es

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

que es la matriz del sistema de ecuaciones. Si además consideramos el vector de respuestas $\mathbf{Y} = (y_1, y_2, \dots, y_n)'$, los términos independientes de las ecuaciones son $\mathbf{X}'\mathbf{Y}$

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

En resumen, el sistema de ecuaciones normales es

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.