Tema 4
Modelos de regresión

Siguiendo con el esquema de clasificación de pruebas estadísticas presentado en el tema anterior, se tratará ahora el problema de analizar respuestas biológicas frente a variables ambientales cuantitativas.

1. Tipos de respuesta frente a variables ambientales cuantitativas

La forma en que la población de una determinada especie responde a los cambios de una variable ambiental cuantitativa, puede ser representada por una función matemática, más o menos compleja según la naturaleza de la relación. Una simple inspección visual de una gráfica que represente, por ejemplo, los valores de abundancia de una especie frente a los de una variable ambiental, puede proporcionar indicios sobre el tipo de relación existente entre ambos. En la naturaleza, el cambio gradual en los valores de una variable ambiental origina lo que se denomina **gradiente**. Generalmente, la presencia o abundancia de una especie depende de un gradiente ambiental, de manera que la probabilidad de encontrar dicha especie, o su abundancia, están estrechamente relacionadas con los cambios que experimentan una o más variables del medio. El estudio de estas relaciones es lo que se conoce en Ecología como **análisis directo del gradiente**, y que se pueden investigar por medio de las técnicas de regresión que se estudiarán en este apartado.

En el análisis de regresión, las relaciones entre dos variables se fundamentan en modelos de respuesta, es decir, en la forma cómo una variable (y), que se puede denominar **dependiente**, responde a los cambios que experimenta la otra variable (x), denominada **independiente**. Fácilmente se puede deducir que en nuestro caso las variables independientes son, por lo general, variables ambientales. Por tanto, el objetivo de un análisis de regresión es conocer la respuesta estimada (Ey) de una especie para un valor determinado de la variable o variables ambientales consideradas.

Un modelo de regresión sencillo es el siguiente

$$y_i = b_0 + b_1 x_i + \varepsilon_i \qquad (1)$$

donde y es la variable dependiente, x es la variable independiente, b_0 y b_1 son los coeficientes que deben ser estimados por la regresión, y ε es el término de error.

En este ejemplo, el modelo es el correspondiente a una línea recta, y por tanto, los coeficientes b_0 y b_1 son respectivamente la ordenada en el origen y la pendiente de la recta. Cada modelo de regresión consta de dos partes:

- una parte sistemática, que describe la relación entre x e y, y que en este caso sería

$$\mathrm{E}y_i = b_0 + b_1 x_i \qquad (2)$$

- una parte de error, que asume que los errores (y - Ey) se distribuyen normalmente, con una media igual a 0 y una varianza que no depende de la variable x.

En la Figura 1 se representan diversos modelos de respuesta de una especie frente a una variable ambiental cuantitativa.

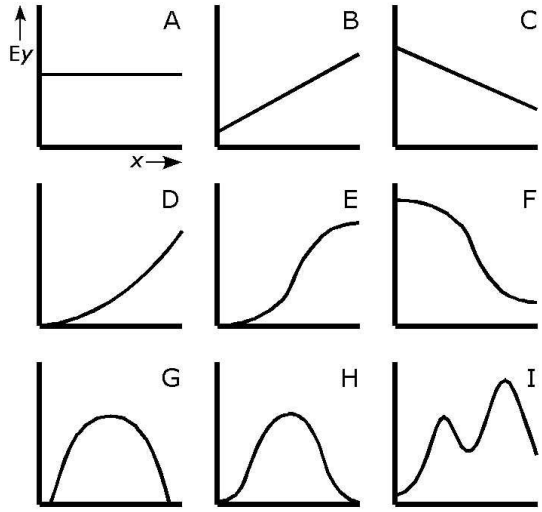


Figura 1. Diversos tipos de respuesta de una especie (E*y*) frente a una variable ambiental (*x*). Para explicación véase el texto.

En el caso A la respuesta de la especie (E*y*) es constante, y por tanto la especie es independiente de la variable ambiental considerada (X). B y C representan relaciones lineales, bien creciente (B) o decreciente (C). El caso D es un ejemplo de relación exponencial, y los casos E y F representan curvas sigmoides. Por su parte, las gráficas G y H representan a una parábola y a una curva gaussiana respectivamente. Ambas son ejemplo de modelos de respuesta unimodales (con un único máximo), mientras que el caso I la respuesta es bimodal. La expresión matemática de algunos de estos modelos se estudiará a continuación.

2. Regresión lineal

El modelo de regresión más sencillo es el que ajusta los datos observados a una línea recta, y en el que la parte sistemática del modelo es, como ya sabemos, Ey = b_0 + $b_1 \cdot x$. El cálculo de los coeficientes de regresión se realiza mediante el método de **mínimos cuadrados**, es decir mediante la búsqueda de la recta que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los estimados:

$$\sum \left(y_i - \mathrm{E}y_i \right)^2 \tag{3}$$

Para una línea recta, el valor del parámetro b_1 (o lo que es lo mismo, la pendiente de la recta) que minimiza esta suma de cuadrados, se calcula mediante la expresión:

$$b_1 = \frac{\sum \left(y_i - \bar{y} \right) \cdot \left(x_i - \bar{x} \right)}{\sum \left(x_i - \bar{x} \right)^2} \tag{4}$$

y la ordenada en el origen se obtiene a partir de:

$$b_0 = \bar{y} - b_1 x_i \tag{5}$$

Una vez calculados los valores de los coeficientes b_0 y b_1 se puede representar gráficamente la línea recta que mejor se ajusta a los datos. Si la pendiente es igual a 0, la expresión (2) se reduce a Ey = b_0. En este caso, la respuesta sería constante (Figura 1A) y podría concluirse que dicha especie es independiente de la variable ambiental considerada.

De igual manera, si b_1 es muy pequeño, la relación entre especie y variable ambiental no será significativa. Por tanto, hemos de utilizar un test que nos permita decidir sobre la significación estadística del parámetro b_1. En concreto se utiliza un test de la t con n-2 grados de libertad:

$$t = \frac{b_1 - 0}{error\ típico\ \ de\ \ b_1} \qquad (6)$$

El error típico de b_1 se calcula mediante una expresión compleja que no interesa estudiar aquí, pero que se puede encontrar en cualquier texto avanzado de estadística. Por otra parte, un análisis de regresión se puede completar con una tabla de ANOVA. La relación entre el análisis de la varianza y la regresión se puede deducir fácilmente de la expresión: ***valor observado = valor estimado + error*** (Tema 3), que ahora puede rescribirse como

$$y = \mathrm{E}y + error \qquad (7)$$

A continuación se comentarán detenidamente los resultados de un análisis de regresión con un ejemplo. Los datos que se utilizarán forman parte de un estudio sobre las relaciones entre la vegetación y la profundidad de la capa freática, llevado a cabo en arenales del Parque Nacional de Doñana. En cada unidad de muestreo se midió la cobertura lineal de diferentes especies, así como la profundidad de la capa freática. Los datos de cobertura fueron transformados logarítmicamente. En la Tabla 1 se presentan los resultados correspondientes a un análisis de regresión para la especie *Cistus libanotis*.

Tabla 1. Resultados del análisis de regresión lineal de los valores transformados de cobertura de *Cistus libanotis* sobre la variable Profundidad.

Término	coeficiente	error	t	P
Constante	2,5869	0,4248	6,089	3.85e-05
Profundidad	0,0130	0,0018	7,101	8.04e-06

TABLA DE ANOVA					
Fuente de variación	g.l.	s.c.	m.c.	F	P
Regresión	1	11,891	11,891	50,49	8.04e-06
Residual	13	3,066	0,236		
Total	14	14,957	1,068		

R^2 = 0,7950 R^2_{aj} = 0,7793

La parte sistemática del modelo de regresión es en este caso:

- $\mathrm{E}y = b_0 + b_1 \cdot$ Profundidad

Como se aprecia a partir de la Tabla 1, el coeficiente b_1 es significativamente diferente de 0, por lo que puede concluirse que existe una relación entre la cobertura de la especie y la profundidad del nivel freático. Como el valor es positivo, a mayor profundidad de la capa freática, mayor cobertura de *Cistus libanotis*. Esta relación se aprecia claramente en la Figura 2. El test de la t para la constante (b_0), no tiene mayor interés que

contrastar si la recta pasa o no por el origen de coordenadas, es decir, si el coeficiente b_0 es o no es significativamente distinto de 0.

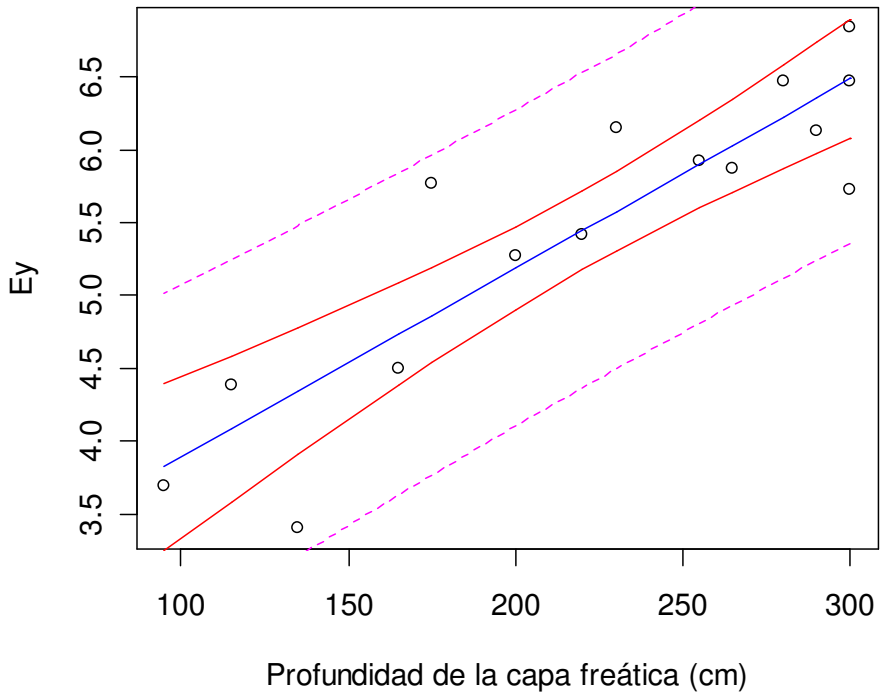


Figura 2. Recta de regresión de los datos transformados de cobertura de *Cistus libanotis* sobre la profundidad de la capa freática. Se representan también los intervalos de confianza y de predicción al 95 %. Los círculos representan los valores transformados de cobertura.

En la Figura 2 se observan los valores transformados de cobertura y la recta de regresión ajustada a los mismos. También se presentan el intervalo de confianza al 95 % de la recta de regresión y el intervalo de predicción al 95 %. El primero determina los límites de los valores estimados de la recta, mientras que el segundo (mucho más amplio por lo general) hace referencia a los valores de cobertura predichos por el modelo de regresión para nuevas muestras con una determinada profundidad de la capa freática. Así, por ejemplo, el valor estimado por la ecuación de regresión para una profundidad de 175 cm es:

- Ey = 2,6859 + 0.0130 · 175 = 4,96

El intervalo de confianza de esta estimación es (4,53, 5,19), mientras que el intervalo de predicción es (3,76, 5,96). Puesto que los datos fueron previamente transformados mediante la expresión ln(x), para calcular el valor real de cobertura estimada hay que calcular $e^{4,96}$ = 142,6 cm. Esta transformación, aplicada sobre los valores estimados de la recta de regresión origina una curva de tipo exponencial (véase la Figura 3), reflejando la relación real entre los valores originales de cobertura y la profundidad del nivel freático.

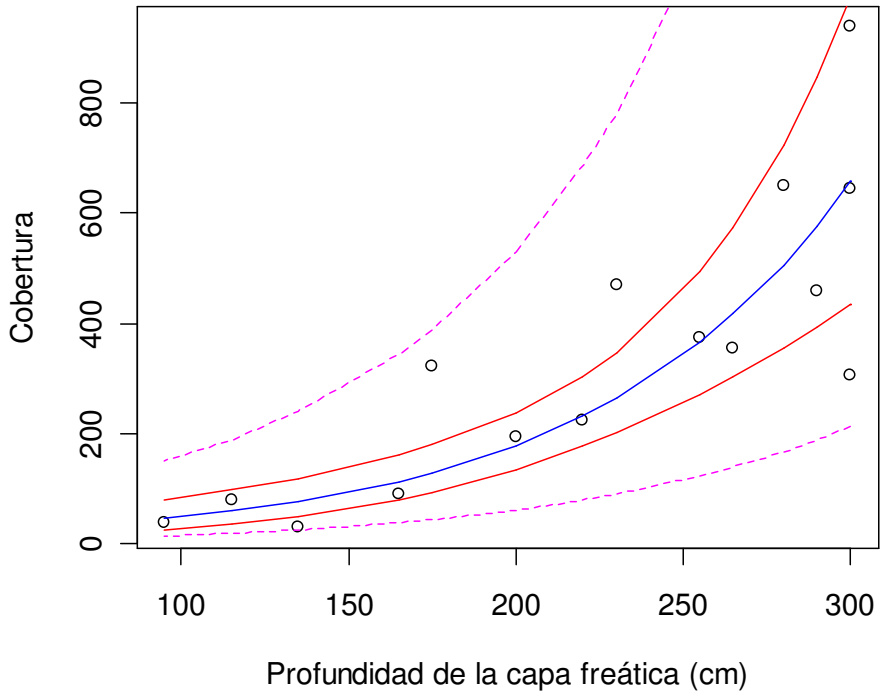


Figura 3. Relación entre la cobertura de *Cistus libanotis* y la profundidad de la capa freática. Las curvas se han obtenido de la transformación exponencial de la recta y los intervalos de confianza y predicción de la Figura 2. Los círculos corresponden a los valores originales de cobertura.

En la Figura 3 puede observarse cómo las asunciones de la parte de error del modelo no se cumplen con los valores originales, ya que la varianza de los datos depende del valor de la variable ambiental, siendo mucho mayor conforme aumentan los valores de ésta. La transformación logarítmica (Figura 2) sí permite asumir este requisito.

La tabla de ANOVA que se presenta en la Tabla 1 es similar a las que estudiadas en el Tema 3, si bien ahora sustituimos los términos "*entre grupos*" y "*dentro de grupos*" por los términos "*regresión*" y "*residual*". El test de la *F* sirve también para contrastar la significación estadística del análisis efectuado, y las conclusiones son las mismas que con los tests de la *t* para los coeficientes. Sin embargo, en modelos más complejos, con más de una variable ambiental, la significación estadística de cada una de ellas debe realizarse mediante los tests de la *t* para los coeficientes. También aparecen ahora dos nuevos valores que no habíamos calculado en las tablas de ANOVA anteriores. Uno de ellos es el **coeficiente de determinación** (R^2), que se calcula como

$$R^2 = 1 - (\text{s.c. residual} / \text{s.c. total}) \quad (8)$$

y que es el cuadrado del **coeficiente de correlación lineal,** *r* (o del **coeficiente de correlación múltiple**, *R*, en el caso de que el modelo de regresión cuente con dos o más variables ambientales).

El otro, más interesante en los análisis de regresión, es el **coeficiente de determinación ajustado** (R^2_{aj}) que se calcula como

$$R^2_{aj} = 1 - (\text{m.c. residual} / \text{m.c. total}) \quad (9)$$

y expresa el porcentaje de la varianza de los datos explicada por la variable o variables independientes del modelo. Así, para el ejemplo, la profundidad de la capa freática explicaría el 77,93 % de la varianza total de los datos de cobertura.

Puesto que para la interpretación de la regresión no es imprescindible la tabla de ANOVA, a partir de ahora las tablas de resultados se presentarán reducidas, sólo con aquellos parámetros de mayor interés.

3. Parábolas y curvas gaussianas

Las especies presentan a menudo modelos no lineales de respuesta frente a variables ambientales. Estos modelos requieren por tanto la introducción de términos no lineales, más o menos complejos, en las ecuaciones de regresión. Como se verá más adelante, el ajuste de los datos a curvas unimodales (con un único máximo) es especialmente interesante en Ecología. Así, si extendemos el modelo de la ecuación (2) con un término cuadrático obtendremos

$$Ey = b_0 + b_1 x + b_2 x^2 \qquad (10)$$

que es la ecuación de una parábola. De esta forma podemos ajustar los datos a una curva de tipo parabólico, siempre que el coeficiente b_2 sea significativamente distinto de 0.

Continuando con el ejemplo del estudio sobre las relaciones entre la vegetación y la profundidad de la capa freática, utilizaremos ahora los datos transformados de cobertura de otra especie: *Lavandula stoechas*. Puede observarse en la Tabla 2 y en la Figura 4 que el ajuste a un modelo lineal no es significativo (la recta no tiene pendiente), por lo que podríamos concluir que la abundancia de la especie es independiente de la profundidad de nivel freático.

Tabla 2. Tabla de regresión lineal para los datos transformados de cobertura de *Lavandula stoechas* sobre la variable Profundidad.

Término	coeficiente	error	*t*	*P*
Constante	5,338	0,4248	8,750	1,49e-06
Profundidad	3,528e-05	0,0025	0,014	0,99

R^2 = 1,59e-05 R^2_{aj} = -0,0833

Varianza residual = 0,3469 ; g.l. = 12

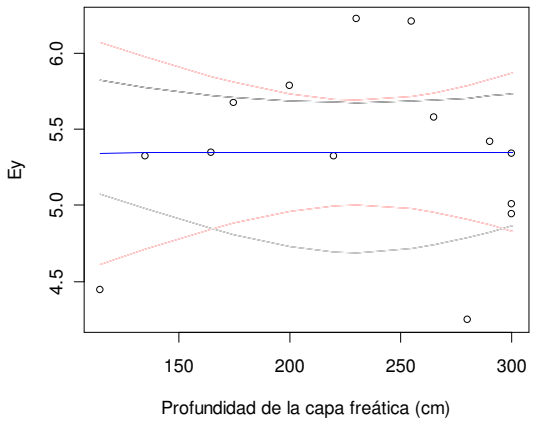


Figura 4. Ajuste de una recta por regresión de los datos transformados de cobertura de *Lavandula stoechas* sobre la profundidad de la capa freática.

Sin embargo, la inclusión del término (profundidad)² sí produce una regresión significativa (Tabla 3, Figura 5).

Tabla 3. Tabla de regresión lineal para los datos transformados de cobertura de *Lavandula stoechas* sobre la variable Profundidad.

Término	coeficiente	error	*t*	*P*
Constante	0,2607	1,727	0,151	0.8828
Profundidad	0,0514	1,692e-02	3,036	0.0113
Profundidad ^{2}	-1,187e-04	3,886e-05	-3,054	0.0110

R^2 = 0.4589 R^2_{aj} = 0.3606

Varianza residual = 0,2048 ; g.l. = 11

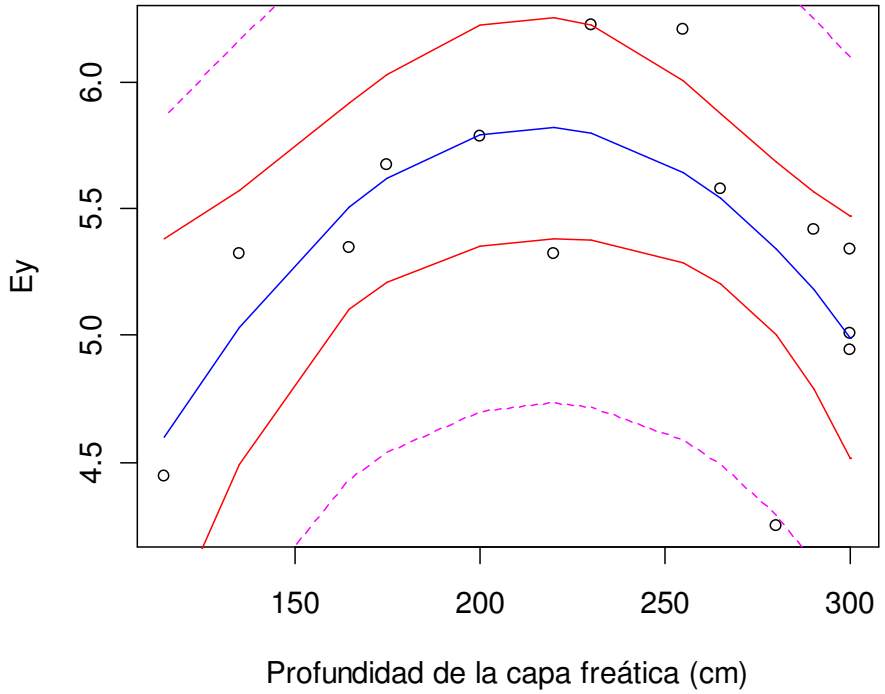


Figura 5. Ajuste de una parábola por regresión de los datos transformados de cobertura de *Lavandula stoechas* sobre la profundidad de la capa freática. Se representan también los intervalos de confianza y predicción al 95 %.

En la Tabla 3 puede observarse que el número de grados de libertad para el test de la t de los coeficientes, es ahora $n - k - 1 = 11$, siendo k el número de términos incluidos en el modelo (dos en este caso). Como b_2 es significativamente distinto de 0 (P = 0,011), la inclusión del término cuadrático en el modelo mejora notablemente el ajuste de los datos. Si b_2 es negativo la curva presenta un máximo, y en caso contrario, un mínimo (cosa poco habitual en respuestas ecológicas). La significación del coeficiente b_1 no tiene ahora importancia.

La transformación exponencial de la parábola para obtener los valores estimados para los datos originales de cobertura da lugar lo que se denomina **curva de Gauss**. Por

tanto, al ajustar una parábola a datos transformados logarítmicamente estamos ajustando una modelo de respuesta gaussiana a los datos originales. Las curvas gaussianas son de gran interés en Ecología, ya que permiten estimar el **óptimo**, la **tolerancia**, y el **máximo** (de abundancia, cobertura, crecimiento, etc.) de una especie a lo largo de un gradiente ambiental (Figura 6).

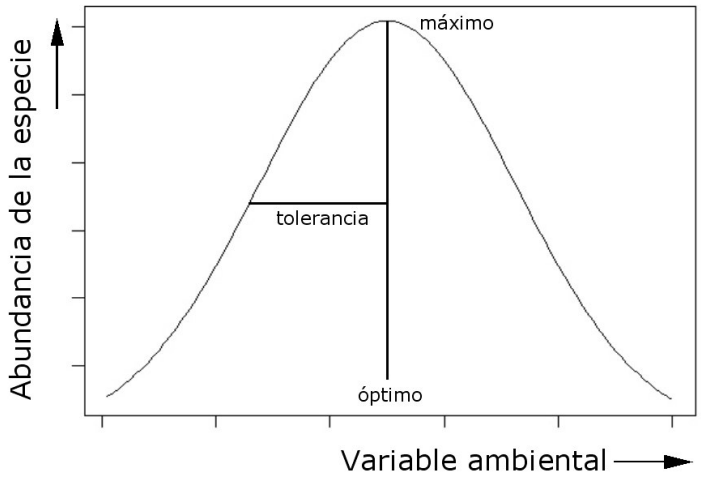


Figura 6. Representación de una curva gaussiana con los parámetros **óptimo**, **tolerancia** y **máximo**.

El óptimo es el valor de la variable ambiental en el que la especie alcanza su máxima abundancia. La tolerancia es una medida de la amplitud ecológica de la especie. Un valor de tolerancia pequeño es debido a que la distribución de la especie a lo largo del gradiente ambiental se reduce a una estrecha franja alrededor del óptimo; un valor de tolerancia elevado indica que la especie tiene distribución mucho más amplia.

La expresión matemática de la curva de Gauss es similar a la función de densidad de la distribución normal, y se expresa como:

$$y = c \cdot e^{-(x-u)^2/2t^2} \tag{11}$$

donde c es el máximo, u el óptimo y t la tolerancia. (Nótese que la tolerancia es equivalente a la desviación típica.)

A partir de los coeficientes de un modelo de regresión cuadrática podemos calcular el óptimo y la tolerancia mediante las ecuaciones:

$$u = \frac{b_1}{-2\,b_2} \tag{12}$$

$$t = \frac{1}{\sqrt{-2\,b_2}} \tag{13}$$

El máximo se obtiene simplemente calculando el valor estimado para el óptimo, es decir:

$$c = e^{\left(b_0 + b_1 u + b_2 u^2\right)} \tag{14}$$

Volviendo al ejemplo de *Lavandula stoechas*, encontramos que para esta especie el óptimo de profundidad de la capa freática es $u \approx 216$ cm, la tolerancia es $t \approx 65$ cm, y el máximo de cobertura estimado es $c \approx 308$ cm. Estos valores se pueden contrastar en la representación gráfica de la curva de Gauss ajustada a los datos originales de cobertura (Figura 7).

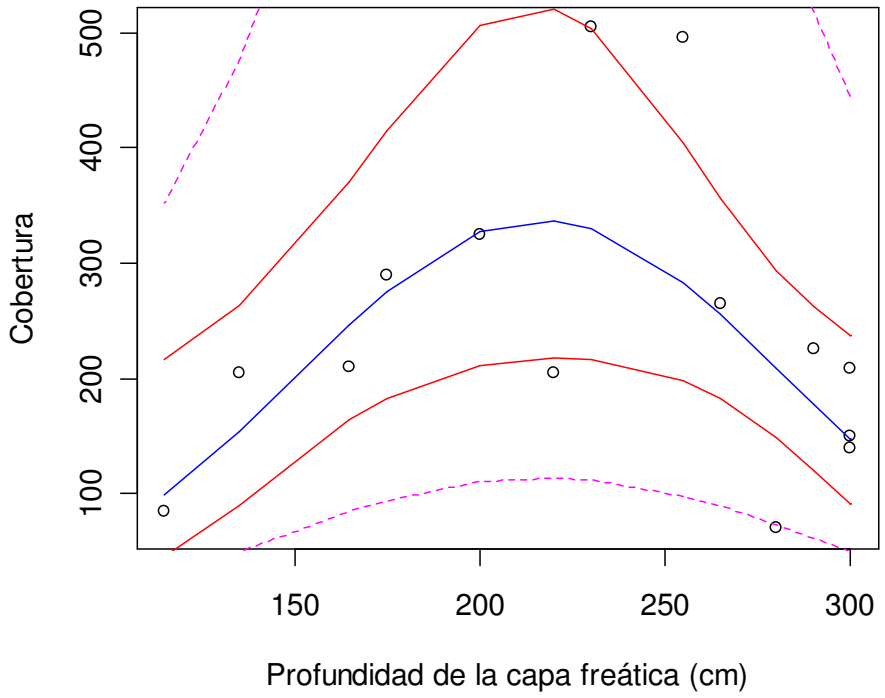


Figura 7. Relación entre la cobertura de *Lavandula stoechas* y la profundidad de la capa freática. Las curvas se han obtenido de la transformación exponencial de la parábola y los intervalos de confianza y predicción de la Figura 5.

4. Regresión discreta

Los datos procedentes de muestreos en los que se cuentan individuos contienen por lo general muchos ceros. En estos casos las transformaciones no suelen producir resultados satisfactorios por lo que respecta a la normalización de los datos y, además, la obligatoriedad de añadir un valor a los ceros antes de transformarlos (por ejemplo: $\log(x+1)$ ó $\sqrt{x+0{,}5}$), implica la necesidad de restarlo al retransformar los datos a la escala original y, como consecuencia, para determinados rangos de valores ambientales puede darse la circunstancia, poco elegante y no deseable, de que el modelo de regresión estime "valores negativos" de abundancia.

Estos problemas se solventan con la aplicación de un método de regresión específicamente apropiado para datos de esta naturaleza (valores discretos, obtenidos mediante conteos): la **regresión discreta** o **regresión de Poisson**. Este método, junto con la **regresión logística** que se presentará en el siguiente apartado son generalmente conocidos como **Modelos Lineales Generalizados** (**GLM**). En este tipo de técnicas los cálculos necesarios para estimar los coeficientes y errores se realizan de manera iterativa, de forma que su aplicación práctica requiere el uso de ordenadores. El parámetro fundamental de los GLM es la *deviance* (término que se traduce aquí como **desvianza**), equivalente a la varianza de la regresión lineal, y que representa, por tanto, una estima de la variabilidad de los datos antes (**desvianza nula**) y después de ajustar el modelo

(**desvianza residual**). De esta forma, la proporción de desvianza explicada por un modelo:

$$1 - (\text{desvianza residual} / \text{desvianza nula}) \qquad (15)$$

es una medida equivalente al coeficiente de determinación R^2 de la regresión lineal.

En los GLM, la expresión de los modelos es similar a la de los modelos de regresión lineal. Por ejemplo, un modelo sencillo de regresión discreta es:

$$\ln y = b_0 + b_1 x \qquad (16)$$

Con respecto a la expresión (2) cambia el término a la izquierda de la igualdad, denominado **función de enlace**. En este caso la función de enlace es logarítmica y la curva ajustada es exponencial ($y = e^{b_0 + b_1 x}$), y el efecto es similar al de transformar los datos logarítmicamente (pero sin necesidad de hacerlo, con lo que se evitan los problemas anteriormente mencionados).

La aplicación de un análisis de regresión discreta se presenta con el siguiente ejemplo. Se utilizan los datos de un estudio llevado a cabo en diversas sierras la Región de Murcia, en el que se investigaba las relaciones entre la abundancia de diferentes especies de aves (número de individuos contados en un total de 180 transectos lineales) y diversas variables ambientales. El modelo de regresión discreta que se presenta en la Tabla 4 corresponde a la Collalba Negra (*Oenanthe leucura*), y muestra las relaciones entre la abundancia de esta especie y la altitud sobre el nivel del mar.

Tabla 4. Tabla de regresión discreta para los datos de abundancia de *Oenanthe leucura* sobre la variable Altitud.

Término	coeficiente	error	*z*	*P*
Constante	1,2158	0,1833	6,632	3,31e-11
Altitud	-0,0039	4,361e-04	-9,020	<2e-16

Desvianza nula = 269,41 ; g.l. 179
Desvianza residual = 166,24 ; g.l. 178
AIC = 283,30

Autor: Carlos González Revelles

La Tabla 4 revela algunas diferencias más entre la regresión discreta y la regresión lineal. Así, el estadístico *t*, utilizado para contrastar la significación estadística de los coeficientes, se sustituye ahora por una *z* (equivalente a la *t*, pero con infinitos grados de libertad). Además se proporciona un nuevo parámetro: el denominado **criterio de información de Akaike** (**AIC**), que representa una medida comparativa del grado de ajuste del modelo, y resulta especialmente útil cuando se realizan modelos de regresión con múltiples variables ambientales (apartado 6).

La representación gráfica de la curva ajustada se muestra en la Figura 8. La respuesta es exponencial decreciente, lo que refleja la preferencia de esta especie por áreas de baja altitud. La curva corresponde a la expresión:

- $y = e^{(1{,}2158 - 0{,}0039 \cdot Altitud)}$

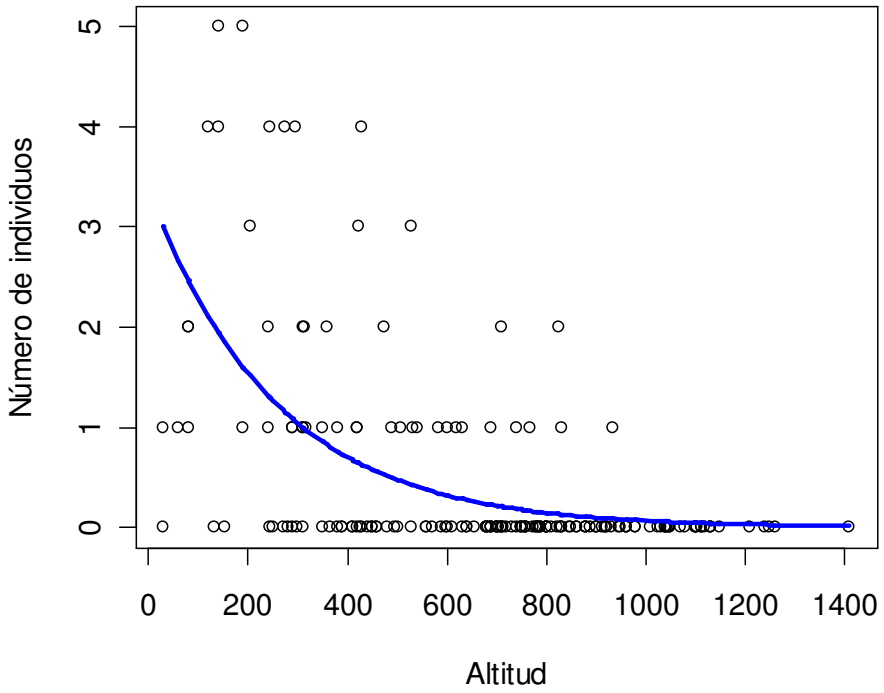


Figura 8. Modelo de regresión discreta de los datos de abundancia (número de individuos) de *Oenanthe leucura* sobre la altitud. Nótese la gran cantidad de ceros en los datos (representados mediante círculos).

5. Regresión logística

La regresión logística es un tipo de GLM especialmente apropiado para el análisis de proporciones y, en general, de datos biológicos cualitativos. Por tanto los modelos de regresión logística constituyen un método útil para el análisis de datos de presencia/ausencia (unos y ceros) de una especie en relación con variables ambientales de carácter cuantitativo. En este caso, la función de enlace se corresponde con la denominada transformación logística; los modelos logísticos tienen la forma:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \qquad (17)$$

donde p es el valor estimado de proporción o probabilidad. Las tablas de regresión logística, con sus coeficientes y demás parámetros, son iguales que las que se obtienen utilizando regresiones discretas. Sin embargo, las curvas estimadas son sigmoides y, dado que se estiman proporciones, el rango de valores del eje y, se encuentra entre el mínimo 0 y el máximo 1.

Como se ha comentado, la regresión logística permite analizar datos de presencia/ausencia. En estos casos los modelos estiman la probabilidad de encontrar una especie en función de los valores de una o más variables ambientales. Sin embargo, la regresión logística permite analizar de forma general datos de proporciones. El ejemplo que se presenta a continuación servirá para ilustrar las características de este tipo de análisis.

La Tabla 5 muestra los resultados de un análisis de regresión logística aplicado sobre los datos procedentes de un experimento realizado con semillas de *Halocnemum strobilaceum*; se trata de investigar la capacidad de germinación en función de la presión osmótica del medio, Ψ, medida en megapascales (MPa).

Tabla 5. Resultados del análisis de regresión logística aplicado sobre datos de germinación de semillas de *Halocnemum strobilaceum* en relación con la presión osmótica (Ψ).

Término	coeficiente	error	*z*	*P*
Constante	6.7613	0,6060	11,16	<2e-16
Ψ	2,0228	0,1885	10,73	<2e-16

Desvianza nula = 335,70 ; g.l. 23
Desvianza residual = 59,51 ; g.l. 22
AIC = 103,77

De la Tabla 5 se deduce que la respuesta la relación entre la proporción de semillas germinadas y la presión osmótica es significativa. La expresión de la curva ajustada es:

- $$p = \frac{e^{(6,7613+2,0228\cdot\Psi)}}{1+e^{(6,7613+2,0228\cdot\Psi)}}$$

Como puede observarse en la representación gráfica del modelo (Figura 9), a mayor presión osmótica (valores más negativos), menor proporción de semillas germinadas, o expresado de otra forma, menor probabilidad de que una semilla germine.

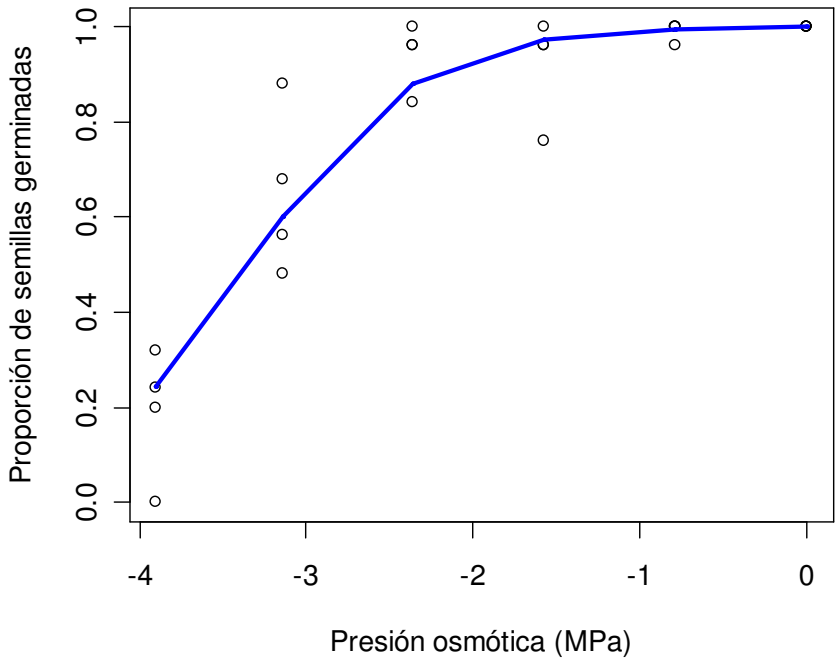


Figura 9. Modelo de regresión logística de los datos de proporción de semillas germinadas de *Halocnemum strobilaceum* en función de la presión osmótica.

6. Modelos de regresión múltiple

Los modelos estudiados hasta ahora pueden ampliarse para analizar el efecto de dos o más variables ambientales sobre la respuesta de una especie. Estas variables ambientales pueden ser tanto cuantitativas como cualitativas, y pueden incluirse de forma conjunta en un modelo de regresión múltiple. Para la construcción de estos modelos pueden seguirse varias estrategias. Una de ellas, denominada *forward stepwise* ("paso a paso hacia delante") se ilustra con el siguiente ejemplo de regresión discretea, en el que además se utilizarán conjuntamente variables ambientales cualitativas y cuantitativas.

Los datos analizados corresponden al mismo estudio sobre aves forestales comentado anteriormente. Aquí se trata de analizar la respuesta del Mito (*Aegithalus caudatus*) a la altitud sobre el nivel del mar y la temperatura media anual. También se incluye la variable *Hábitat*, de carácter cualitativo (*dummy*), que informa sobre el tipo de hábitat donde se realizó el transecto (la unidad de muestreo); en este caso *hábitat* tiene dos modalidades (pinar y matorral) y toma valor 1 para pinar, y valor 0 para matorral. El procedimiento de análisis (ilustrado en la Tabla 6) es el siguiente. En un primer paso realizamos regresiones con cada una de las variables ambientales disponibles; es conveniente probar el ajuste tanto a modelos lineales como cuadráticos, anotando en cada caso los valores de AIC.

A continuación, una vez realizados los modelos con todas las variables ambientales, se selecciona la variable con la que se obtenga el menor valor de AIC. (En el caso de regresiones lineales se seleccionaría aquella variable cuyo modelo proporcione el mayor valor de R^2 ajustada.) A partir de ese momento la variable seleccionada se mantiene "fija y, en los siguientes modelos, se van incorporando una a una todas las demás. De nuevo se repite el procedimiento seleccionando la variable de mejor ajuste, que se une a la primera. Este procedimiento de selección finaliza cuando ya no se pueden incorporar más variables significativas al modelo.

Tabla 6. Procedimiento de elaboración de un modelo de regresión múltiple mediante el método *forward stepwise*, para los datos de abundancia de Mito (*Aegithalus caudatus*) en sistemas forestales de la Región de Murcia. En cada paso se selecciona la variable para la que se obtenga un menor valor del criterio de información de Akaike (AIC). En este caso, la "mejor" variable es *Hábitat*, que en el siguiente paso se incorpora "fija" en los modelos. Una vez seleccionadas las variables significativas se completa el modelo con las interacciones. [***ns***: modelo no significativo.]

Modelo		AIC
Nulo		
$\ln(y) = b_0 + b_1$ Hábitat		932,3
$\ln(y) = b_0 + b_1$ Altitud		1056,4
$\ln(y) = b_0 + b_1$ Altitud + b_2 Altitud^2		1016,7
$\ln(y) = b_0 + b_1$ Temperatura		1078,8
$\ln(y) = b_0 + b_1$ Temperatura + b_2 Temperatura^2		1071.6
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Altitud		919,2
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Altitud + b_3 Altitud^2		891,8
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Temperatura		929,8
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Temperatura + b_3 Temperatura^2	***ns***	930,2
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Altitud + b_3 Altitud^2 + b_4 Temperatura	***ns***	891,9
$\ln(y) = b_0 + b_1$ Hábitat + b_2 Altitud + b_3 Altitud^2 + b_4 Temperatura + b_5 Temperatura^2	***ns***	893.9

Modelo final con interacciones:

$\ln(y) = -48{,}58 + 47{,}32 \cdot$Hábitat $+ 0{,}10613 \cdot$Altitud $- 5{,}732\text{e-}05 \cdot$Altitud$^2 - 0{,}1001 \cdot$Hábitat$\cdot$Altitud $+ 5{,}363\text{e-}05 \cdot$Hábitat$\cdot$Altitud^2

Desvianza = 665,12 ; g.l. = 174 ; AIC = 863,54

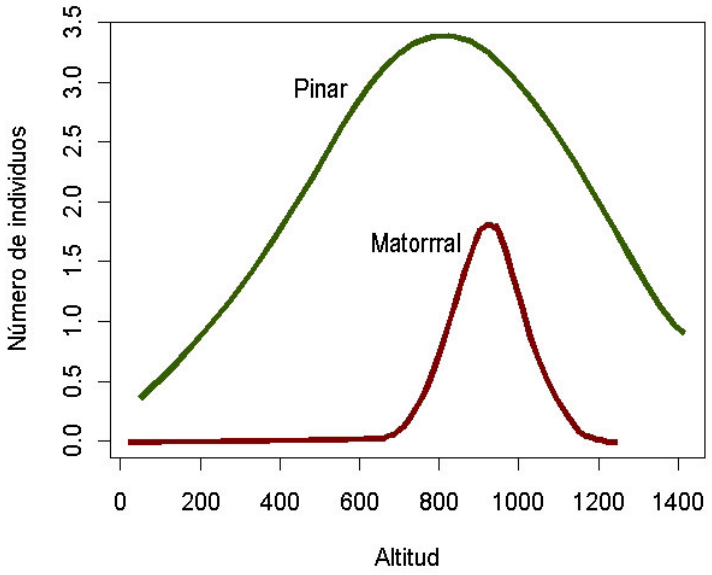
Finalmente, puede completarse el modelo testando la significación de las interacciones (es decir, el producto de dos variables incluidas en el modelo). Estas interacciones pueden ser especialmente interesantes en el caso de contar con variables *dummy*, como en el caso de la Tabla 6, ya que la consideración de las interacciones entre *hábitat* y *altitud*, proporciona en realidad el ajuste a dos curvas de respuesta distintas: una para matorral y otra para pinar (Figura 10). Dado que la variable *Hábitat* toma valor 1 ó 0, los diferentes coeficientes se anulan o se suman en función del tipo de hábitat considerado. Como resultado se obtienen dos curvas distintas:

Para pinar:

$$\ln(y) = (-48{,}58+47{,}32) + (0{,}10613 - 0{,}1001)\cdot\text{Altitud} + (-5{,}732\text{e-}05+5{,}363\text{e-}05)\cdot\text{Altitud}^2$$

Para matorral:

$$\ln(y) = -48{,}58 + 0{,}10613\cdot\text{Altitud} - 5{,}732\text{e-}05\cdot\text{Altitud}^2$$



Autor: Carlos González Revelles

Figura 10. Representación gráfica del modelo de respuesta de la Tabla 6 correspondiente a los datos de abundancia (número de individuos) de *Aegithalus caudatus*. El modelo estima diferentes respuestas unimodales a la altitud según el tipo de hábitat.

La consideración de dos variables cuantitativas en un modelo de regresión produce el ajuste a **superficies de respuesta**. La representación gráfica de las mismas adoptará diferentes formas en función de las respuestas individuales a cada variable. Como ejemplo, en la Figura 11 se representan tres tipos de superficies de respuesta correspondientes a modelos de regresión en los que el ajuste es lineal para ambas variables (caso del Carbonero Garrapinos, *Parus ater*), lineal para una y unimodal para la otra (caso del Piquituerto, *Loxia curvirostra*), o unimodal para ambas (caso del Verdecillo, *Serinus serinus*).

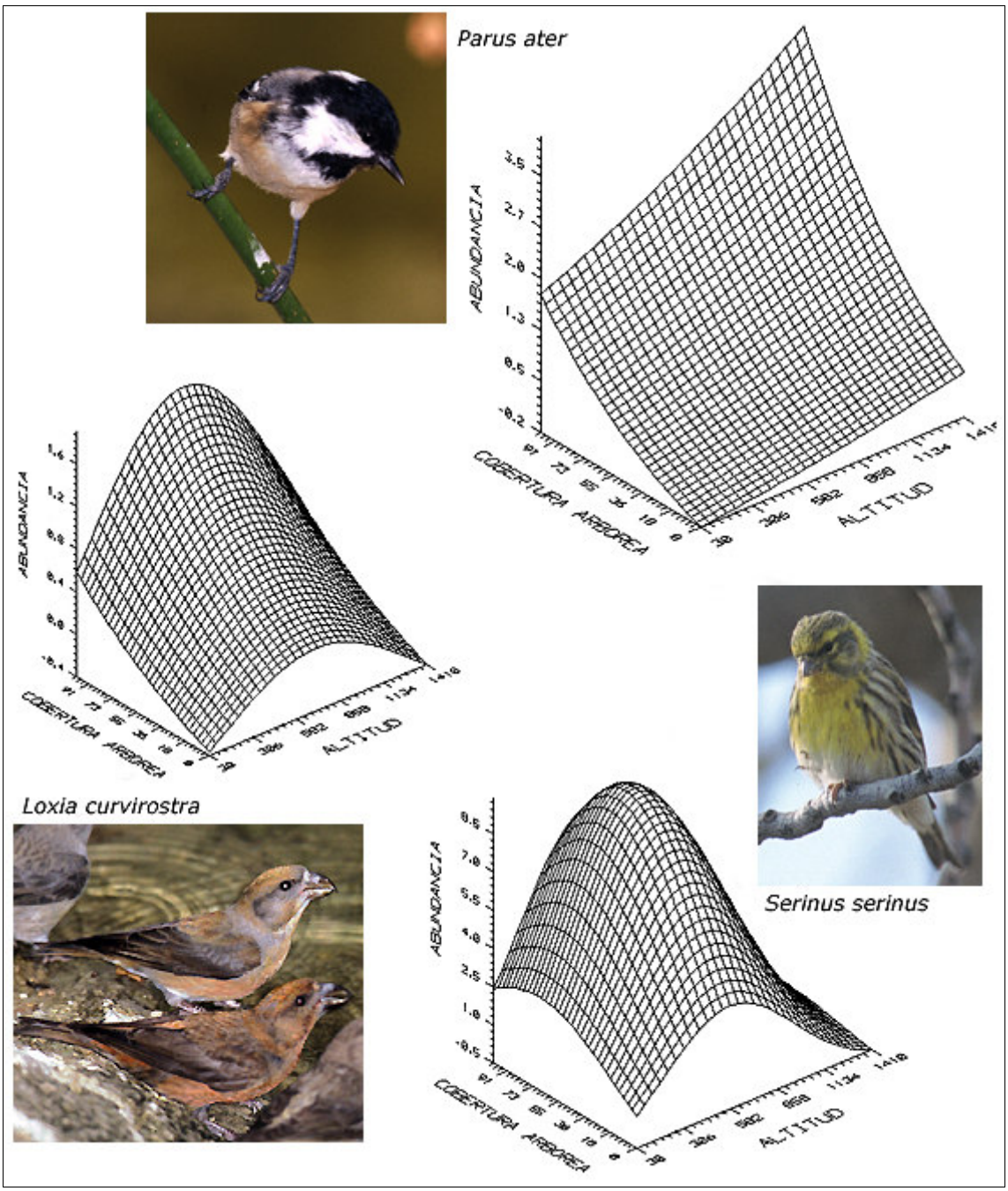


Figura 11. Representación gráfica de diferentes modelos de respuesta a dos variables ambientales cuantitativas. Las superficies pueden adoptar diferente forma en función del tipo de ajuste a cada variable individual. (Autor de las fotografías: Carlos González Revelles)