

Modelos lineales generalizados (GLM)

Luis Cayuela

9 de noviembre de 2009

EcoLab, Centro Andaluz de Medio Ambiente, Universidad de Granada –
Junta de Andalucía, Avenida del Mediterráneo s/n, E-18006, Granada.
E-mail: lcayuela@ugr.es.

Índice

1. ¿Qué son los GLM?	3
1.1. La estructura de los errores	4
1.2. La función de vínculo	6
2. Construcción y evaluación de un GLM	12
3. Criterios de evaluación de modelos	15
4. La función glm()	15
5. Modelos binomiales	16
5.1. Respuestas binarias (regresión logística)	16
5.1.1. Un ejemplo: Prediciendo la distribución de especies . .	17
5.2. Proporciones	22
5.2.1. Análisis de proporciones para factores con uno y dos niveles	23
5.2.2. Un ejemplo: ¿Son eficientes los pesticidas en el control de plagas?	23
6. Modelos Poisson	28
7. Ejercicios	29

1. ¿Qué son los GLM?

Los modelos lineales (regresión, ANOVA, ANCOVA), se basan en los siguientes supuestos:

1. Los errores se distribuyen normalmente.
2. La varianza es constante.
3. La variable respuesta se relaciona linealmente con la(s) variable(s) independiente(s).

En muchas ocasiones, sin embargo, nos encontramos con que uno o varios de estos supuestos no se cumplen. Por ejemplo, es muy común en ecología que a medida que aumenta la media de la muestra, aumente también su varianza. Estos problemas se pueden llegar a solucionar mediante la transformación de la variable respuesta (por ejemplo tomando logaritmos). Sin embargo estas transformaciones no siempre consiguen corregir la falta de normalidad, la heterocedasticidad (varianza no constante) o la no linealidad de nuestros datos. Además resulta muchas veces difícil interpretar los resultados obtenidos. Si decimos que la abundancia de pino silvestre es función de la elevación tenemos una idea más o menos clara de lo que esto puede significar. Si la relación es positiva, un aumento de la elevación aumentaría la abundancia de esta especie. Pero ¿qué quiere decir que el logaritmo de la abundancia de pino silvestre es función de la elevación? Esto ya no es tan intuitivo. La cosa se complica aún más cuando utilizamos otro tipo de transformaciones, como las exponenciales, las potencias, etc. Una alternativa a la transformación de la variable respuesta y a la falta de normalidad es el uso de los modelos lineales generalizados. Los modelos lineales generalizados (GLM de las siglas en inglés de *Generalized Linear Models*) son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc) y varianzas no constantes.

Ciertos tipos de variables respuesta sufren invariablemente la violación de estos dos supuestos de los modelos normales y los GLM ofrecen una buena alternativa para tratarlos. Específicamente, podemos considerar utilizar GLM cuando la variable respuesta es:

- un conteo de casos (p.e. abundancia de una especie);

- un conteo de casos expresados como proporciones (p.e. porcentaje de plántulas muertas en un experimento de vivero);
- una respuesta binaria (p.e. vivo o muerto, infectado o no infectado);

El supuesto central que se ha hecho hasta el momento con los modelos lineales es que la varianza es constante (**Figura 1a**). En el caso de los conteos, sin embargo, donde la variable respuesta está expresada en números enteros y en donde hay a menudo muchos ceros en los datos, la varianza podría incrementar linealmente con la media (**Figura 1b**). Con proporciones, donde hay un conteo del número de fallos de un evento, así como del número de éxitos, la varianza tendrá una forma de U invertida en relación a la media (**Figura 1c**). Cuando la variable respuesta siga una distribución Gamma, entonces la varianza incrementa de una manera no lineal con la media (**Figura 1d**).

Muchos de los métodos estadísticos más comunes, como la t de Student o la regresión, asumen que la varianza es constante, pero en muchas aplicaciones este supuesto no es aplicable. Y es precisamente en estos casos cuando los GLM pueden ser de gran utilidad. Los GLM tienen dos propiedades importantes:

1. La estructura de los errores.
2. La función de vínculo.

1.1. La estructura de los errores

Muchos datos tienen una estructura no normal. En el pasado, las únicas herramientas disponibles para tratar la ausencia de normalidad eran la transformación de la variable respuesta o la adopción de métodos no paramétricos. Hoy en día, existe otra alternativa, que son los modelos lineales generalizados o GLM. Los GLM permiten especificar distintos tipos de distribución de errores:

Poisson, muy útiles para conteos (p.e. número de muertos por accidentes de tráfico; número de días con heladas en el mes de enero; número de colonias de bacterias en una placa de agar; número de especies de plantas leñosas en un cuadrado de muestreo de 10 m^2).

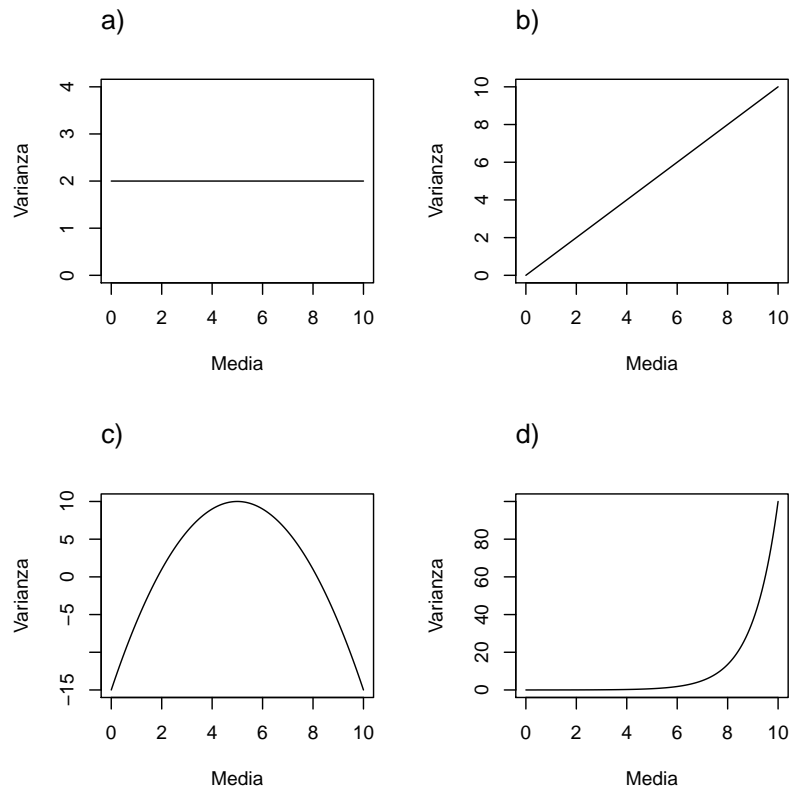


Figura 1: Relación entre la media y la varianza de los datos bajo distintos supuestos; (a) la varianza es constante; (b) la varianza incrementa con la media; (c) la varianza tiene forma de U invertida en relación a la media; (d) la varianza incrementa de manera no lineal con la media.

Binomiales, de gran utilidad para proporciones y datos de presencia/ausencia (p.e. tasas de mortalidad; tasas de infección; porcentaje de parasitismo; porcentaje de éxito reproductivo; presencia o ausencia de una determinada especie).

Gamma, muy útiles con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante (p.e. número de presas comidas por un predador en función del número de presas disponibles).

Exponenciales, muy útiles para los análisis de supervivencia, aunque no se verán de manera específica en este curso.

Además, los modelos lineales, con los que estamos más familiarizados, asumen que tanto la variable respuesta como los errores del modelo siguen una distribución normal. Una distribución normal es, por definición, continua. En ocasiones, sin embargo, la variable dependiente sigue una distribución que no es continua y, por tanto, los valores estimados por el modelo han de seguir el mismo tipo de distribución que los datos de partida. Cualquier otro tipo de valor estimado por el modelo no debería ser válido desde un punto de vista lógico, aunque en la práctica no se presta mucha atención a esto. Por ejemplo, un investigador está interesado en predecir cuántos niños tendrá una familia en función del ingreso neto anual y otros indicadores socio-económicos. La variable respuesta –número de niños– es discreta (es decir, una familia podrá tener 1, 2, 3 hijos y así sucesivamente, pero no 2.4 hijos) y además está muy sesgada (la mayoría de las familias tendrán 1, 2 o 3 hijos, algunas menos tendrán 4 o 5, y muy pocas familias tendrán 6 o 7 hijos) (**Figura 2**). En este caso, es razonable asumir que la variable dependiente seguirá una distribución de tipo Poisson y no una normal.

Para detectar si nuestros datos son o no normales es conveniente: (1) conocer el tipo de variable respuesta y su naturaleza; y (2) el análisis de los residuos del modelo una vez ajustado el modelo (ya sea un modelo lineal o un GLM con una distribución de errores no normal). Esto nos va a permitir observar alejamientos de la normalidad y saber cuándo es conveniente utilizar uno u otro tipo de distribuciones de errores.

1.2. La función de vínculo

Otra razón por la que un modelo lineal puede no ser adecuado para describir un fenómeno determinado es que la relación entre la variable respuesta y la(s) variable(s) independiente(s) no es siempre lineal. Un ejemplo lo tenemos en la relación entre la edad de una persona y su estado de salud (**Figura 3**). La salud de la gente de 30 años no es muy distinta de la de la gente de 40. Sin embargo, las diferencias sí son más marcadas entre la gente de 60 y 70 años. Por tanto, la relación entre edad y salud no es lineal. Tal vez una función de tipo exponencial sería más adecuada para describir la relación entre la edad de una persona y su salud. Este tipo de funciones indicaría que un incremento en años en edades más avanzadas tendría un mayor impacto sobre la salud de las personas que un incremento en años en edades más tempranas. Con otras palabras, el vínculo entre la edad y la salud se describe mejor con una función de tipo exponencial que con una relación lineal.

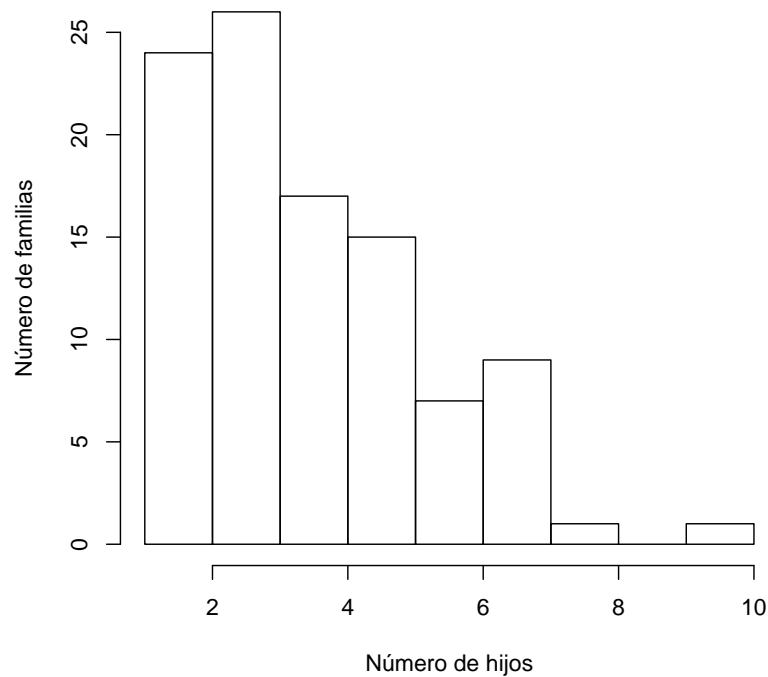


Figura 2: Gráfica de distribución del número de familias con un determinado número de hijos.

La función de vínculo, por tanto, se encarga de linealizar la relación entre la variable respuesta y la(s) variable(s) independiente(s) mediante la transformación de la variable respuesta. Tomemos por ejemplo la relación entre el número de ovejas muertas y el número de parásitos (**Figura 4**). Está relación como podemos ver no es del todo lineal (izquierda). Pero podemos linealizarla tomando logaritmos en la variable respuesta (derecha).

En este ejemplo, el modelo quedaría formulado de la siguiente forma:

$$\text{Log}(y_i) = \beta_0 + \beta_1 \cdot x_i$$

En dónde:

y = número de ovejas muertas,

x = número de parásitos,

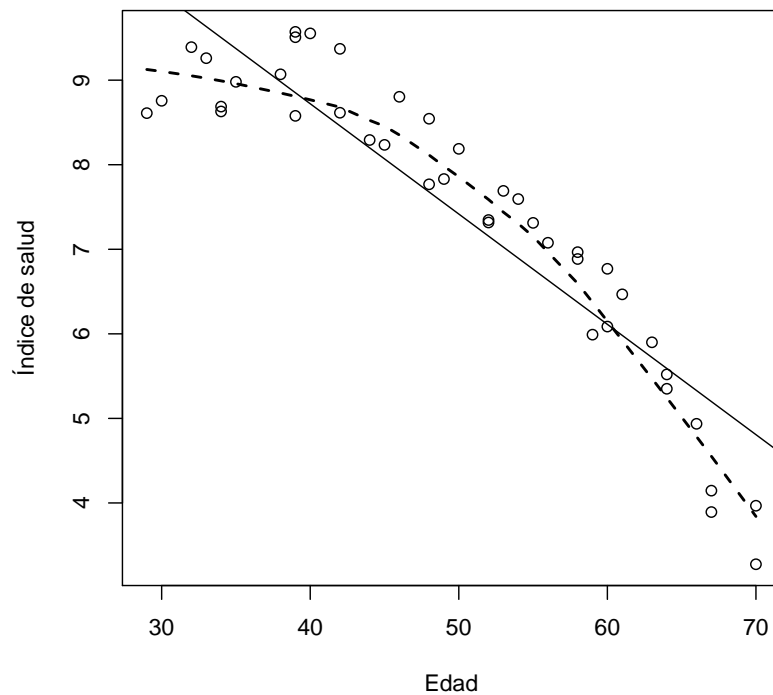


Figura 3: Relación entre el estado de salud, medido por medio de un índice, y la edad de las personas.

β_0 y β_1 = coeficientes del modelo.

Ahora bien, los valores estimados por este modelo no son los valores de y , sino los del $\text{Log}(y)$. Para obtener los valores estimados de y , lo que se debe de hacer es aplicar la función inversa a la función de vínculo utilizada, en este caso, la función exponencial. Por tanto:

$$\exp(\text{Log}(y_i)) = \exp(\beta_0 + \beta_1 \cdot x_i)$$

$$y_i = \exp(\beta_0 + \beta_1 \cdot x_i)$$

En realidad, aunque parezca muy complicado, lo que estamos haciendo es básicamente transformar la variable respuesta de modo similar a cómo haríamos en una regresión cuando tenemos problemas de linealidad, pero teniendo en cuenta los valores estimados por el modelo mediante la transformación inversa de la función de vínculo.

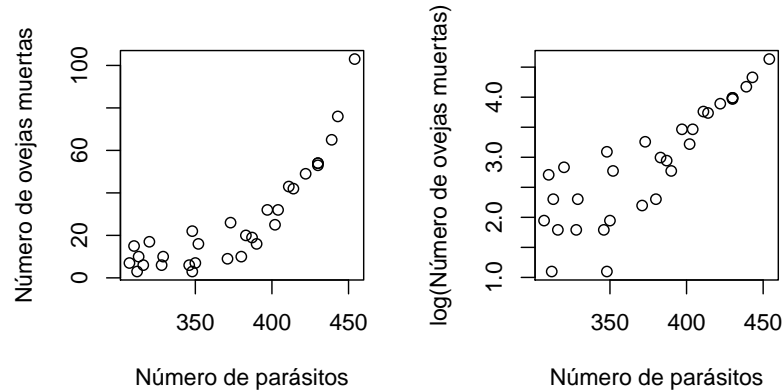


Figura 4: Relación entre el número de parásitos y: (izqda) el número de ovejas muertas; (dcha) el logaritmo del número de ovejas muertas.

Otra de las utilidades de la función de vínculo, es la de conseguir que las predicciones de nuestro modelo queden acotadas. Por ejemplo, si tenemos datos de conteo, no tiene sentido que nuestras predicciones arrojen resultados negativos, como en el caso del número de ovejas muertas o la abundancia de una determinada especie. En este caso, una función de vínculo de tipo logarítmica resolverá el problema (**Figura 5**). Otro ejemplo, si la variable respuesta es una proporción, entonces los valores estimados tienen que estar entre 0 y 1 o 0 y 100 (valores por debajo de 0 o por encima de 1 o 100 no tienen ningún sentido). En este otro caso, una función de vínculo de tipo 'logit' será más apropiada.

En la **Tabla 1** se resumen las funciones de vínculo más utilizadas.

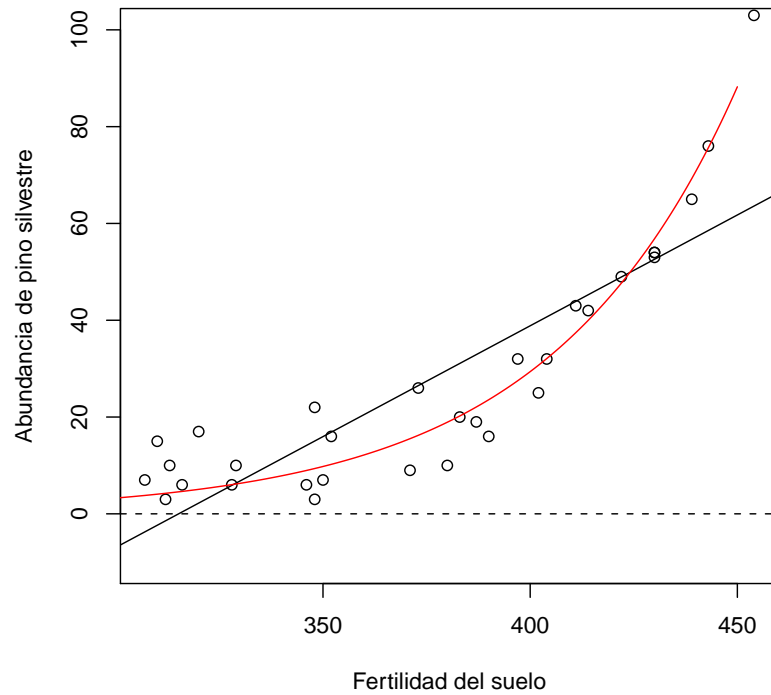


Figura 5: Relación entre la abundancia de una especie (p.e. número de pies de pino silvestre en una parcela forestal) y la fertilidad del suelo. Los valores quedan acotados por encima de 0 cuando utilizamos una función de vínculo de tipo logarítmica.

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$\text{Log}(\mu)$	Conteos con errores de tipo Poisson
Logit	$\text{Log}(\frac{\mu}{n-\mu})$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

Cuadro 1: Las funciones de vínculo más comunes utilizadas por los GLM

Se denominan funciones de vínculo canónicas a las funciones que se aplican por defecto a cada una de las distribuciones de errores. Esto no significa que siempre se deba usar una única función de vínculo para una determinada distribución. De hecho, puede ser recomendable comparar diferentes funciones de vínculo para un mismo modelo y ver con cuál se obtiene un mejor ajuste del modelo a los datos. En la **Tabla 2** se pueden ver las funciones de vínculo canónicas para cada una de las distribuciones de errores, así como otras posibles funciones de vínculo que pueden ser usadas.

Distribución de errores	Función de vínculo canónica	Otras funciones de vínculo posibles
Normal	Identidad	Logarítmica
Poisson	Logarítmica	Identidad, Raíz cuadrada
Binomial	Logit	Logarítmica
Gamma	Recíproca	Identidad, Logarítmica

Cuadro 2: Las funciones de vínculo canónicas y otras funciones de vínculo posibles usadas para distintas distribuciones de errores en GLM.

En la **Tabla 4** se muestran algunas de las combinaciones más comunes de variables respuestas y variables explicativas con distintos tipos de funciones de vínculo y distribuciones de errores.

Tipo de análisis	Variable respuesta	Variable explicativa	Función de vínculo	Distribución de errores
Regresión	Continua	Continua	Identidad	Normal
ANOVA	Continua	Factor	Identidad	Normal
Regresión	Continua	Continua	Recíproca	Gamma
Regresión	Conteo	Continua	Logarítmica	Poisson
Tabla de contingencia	Conteo	Factor	Logarítmica	Poisson
Proporciones	Proporción	Continua	Logit	Binomial
Regresión logística	Binaria	Continua	Logarítmica	Binomial
Análisis de supervivencia	Tiempo	Continua	Recíproca	Exponencial

Cuadro 4: Algunas de las combinaciones más comunes de variables respuestas y variables explicativas con distintos tipos de funciones de vínculo y distribuciones de errores.

2. Construcción y evaluación de un GLM

En la construcción de modelos lineales generalizados es importante tener en cuenta una cosa: no existe un único modelo que sea válido. Éste es uno de los errores más comunes implícitos en el uso de regresión o ANOVA, en donde el mismo modelo se usa una y otra vez, muchas veces sin una perspectiva crítica. En la mayoría de los casos, habrá un número variable de modelos plausibles que puedan ajustarse a un conjunto determinado de datos. Parte del trabajo de construcción y evaluación del modelo es determinar cuál de todos estos modelos son adecuados, y entre todos los modelos adecuados, cuál es el que explica la mayor proporción de la varianza sujeto a la restricción de que todos los parámetros del modelo deberían ser estadísticamente significativos. Esto es lo que se conoce como el modelo adecuado mínimo. En algunos casos habrá más de un modelo que describan los datos igual de bien. En estos casos queda a nuestro criterio elegir uno u otro, aunque puede ser recomendable utilizarlos todos y discutir las limitaciones que esto presenta desde el punto de vista inferencial.

Los pasos que hay que seguir en la construcción y evaluación de un GLM son muy similares a los de cualquier modelo estadístico. No obstante los detallamos a continuación:

Exploración de los datos. Conviene conocer nuestros datos. Puede

resultar interesante obtener gráficos que nos muestren la relación entre la variable respuesta y cada una de las variables explicativas, gráficos de caja (box-plot) para variables categóricas, o matrices de correlación entre las variables explicativas. El objetivo de este ejercicio es: a) Buscar posibles relaciones de la variable respuesta con la(s) variable(s) explicativa(s); b) Considerar la necesidad de aplicar transformaciones de las variables; c) Eliminar variables explicativas que estén altamente correlacionadas.

Elección de la estructura de errores y función de vínculo. A veces resultará fácil elegir estas propiedades del modelo. Otras resultará tremendamente difícil. No hay que preocuparse por esto, sin embargo, ya que con posterioridad analizaremos los residuos del modelo para ver la idoneidad de la distribución de errores elegida. Por otro lado, puede ser una práctica recomendable el comparar modelos con distintas funciones de vínculo para ver cuál se ajusta mejor a nuestros datos.

Ajuste del modelo a los datos. Debemos prestar particular atención a: a) Los tests de significación para los estimadores del modelo; b) La cantidad de varianza explicada por el modelo. Esto en GLM se conoce como devianza D^2 . La devianza nos da una idea de la variabilidad de los datos. Por ello, para obtener una medida de la variabilidad explicada por el modelo, hemos de comparar la devianza del modelo nulo (*Null deviance*) con la devianza residual (*Residual deviance*), esto es, una medida de cuánto de la variabilidad de la variable respuesta no es explicado por el modelo, o lo que es lo mismo:

$$D^2 = \frac{\text{Devianza.modelo.nulo} - \text{Devianza.residual}}{\text{Devianza.modelo.nulo}} \cdot 100$$

Análisis de los residuos. Los residuos son las diferencias entre los valores estimados por el modelo y los valores observados. Sin embargo, muchas veces se utilizan los residuos estandarizados, que tienen que seguir una distribución normal. Conviene analizar los siguientes gráficos:

1. Histograma de los residuos.
2. Gráfico de residuos frente a valores estimados. Estos gráficos pueden indicar falta de linealidad, heterocedasticidad (varianza no constante) y valores atípicos.

3. El gráfico probabilístico de normalidad (*q-q plot*), que permite contrastar la normalidad (simetría) de la distribución de los residuos.

Y, opcionalmente, pueden ser también de gran utilidad los siguientes gráficos:

1. Gráficos de residuos frente a variables explicativas. Pueden ayudar a identificar si la falta de linealidad o la heterocedasticidad es debida a alguna variable explicativa.
2. Gráfico de los residuos frente al tiempo (u orden de medida). Permiten detectar cambios sistemáticos en el muestreo (como cuando el experimentador adquiere mayor experiencia en el proceso de medición de un determinado fenómeno, o por el contrario, se vuelve menos cuidadoso a medida que aumenta el esfuerzo muestral).
3. Gráfico de valores atípicos. Existen tests que permiten detectar valores atípicos. Los índices más comunes son el índice de Cook y el apalancamiento o *leverage*.

Todos estos gráficos (y opcionalmente algunos tests estadísticos complementarios) nos pueden ayudar en la evaluación del modelo utilizado. En caso necesario, será preciso volver a plantear el modelo (paso 2), tal vez utilizando una estructura de errores más adecuada, otra función de vínculo o incluso eliminando ciertos datos que pueden estar ‘sobre-influenciando’ nuestro análisis.

Simplificación del modelo. El principio de parsimonia requiere que el modelo sea tan simple como sea posible. Esto significa que no debe contener parámetros o niveles de un factor que sean redundantes. La simplificación del modelo implica por tanto:

1. La eliminación de las variables explicativas que no sean significativas.
2. La agrupación de los niveles de factores (variables categóricas) que no difieran entre sí. Esto significa que cada vez que simplificamos el modelo debemos repetir los pasos 3 y 4. La simplificación del modelo tiene que tener, además, una cierta lógica para el analista y no debe incrementar de manera significativa la devianza residual. Por ello y para llegar a entender bien los datos y las relaciones existentes entre las variables conviene evitar, en la medida de lo posible, los procedimientos automatizados (p.e. *backward/forward stepwise regression procedures*).

3. Criterios de evaluación de modelos

Podemos utilizar la reducción de la devianza como una medida del ajuste del modelo a los datos. Los tests de significación para los parámetros del modelo son también útiles para ayudarnos a simplificar el modelo. Sin embargo, un criterio comunmente utilizado es el llamado **Criterio de Información de Akaike** (AIC del inglés *Akaike Information Criterion*). Aunque no vamos a explicar aquí los fundamentos matemáticos de este índice, sí diremos que es un índice que evalúa tanto el ajuste del modelo a los datos como la complejidad del modelo. Cuanto más pequeño es el AIC mejor es el ajuste. El AIC es muy útil para comparar modelos similares con distintos grados de complejidad o modelos iguales (mismas variables) pero con funciones de vínculo distintas. Las funciones `stepAIC()`, `addterm()` y `dropterm()` del paquete **MASS** permiten comparar modelos con distintos grados de complejidad en función del AIC. Todo esto se aplica igualmente a los modelos lineales que se vieron en la sesión anterior.

4. La función `glm()`

La función `glm()` viene especificada por los siguientes argumentos

```
> args(glm)
```

```
function (formula, family = gaussian, data, weights, subset,
  na.action, start = NULL, etastart, mustart, offset, control = glm.control(..
  model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL,
  ...)
NULL
```

dónde `formula` es una fórmula que especifica el modelo siguiendo la lógica de los modelos lineales especificados por la función `lm()` y `family` es la familia de errores de distribución, especificada de la siguiente forma:

- `binomial(link = "logit")`
- `gaussian(link = "identity")`
- `Gamma(link = "inverse")`

- `inverse.gaussian(link = "1/mu^2")`
- `poisson(link = "log")`
- `quasi(link = "identity", variance = "constant")`
- `quasibinomial(link = "logit")`
- `quasipoisson(link = "log")`

Si la función de vínculo (`link`) no se especifica, la primera opción de la lista es tomada como opción predeterminada en cada caso. Como en el caso de las funciones `lm()`, podemos acceder fácilmente al resultado de un modelo `glm()` con las funciones `summary()` y `anova()`.

5. Modelos binomiales

5.1. Respuestas binarias (regresión logística)

Muchas variables respuesta son del tipo:

- vivo o muerto,
- hombre o mujer,
- infectado o saludable,
- ocupado o vacío.

En estos casos podemos investigar qué variables están relacionados con la asignación de un individuo a una clase u otra mediante modelos GLM con una distribución de errores de tipo binaria, siempre y cuando exista al menos una variable explicativa que sea continua. La variable respuesta debe de contener sólo 0s ò 1s, en dónde un 0 representaría por ejemplo a un individuo muerto y un 1 a un individuo vivo. La manera en la que R trata datos binarios es asumiendo que los 0s y los 1s provienen de una distribución binomial de tamaño 1. Si la probabilidad de que un individuo esté muerto es p , entonces la probabilidad de obtener y (donde y es vivo o muerto, 0 ò 1) vendría dado por la forma abreviada de la distribución binomial con $n = 1$, conocida como la distribución de Bernoulli:

$$P(y) = p^y \cdot (1 - p)^{(1-y)}$$

El objetivo en este caso sería determinar cómo las variables explicativas influyen el valor de p . Para ajustar un modelo binomial en R hay que usar la función `glm()` especificando el argumento `family = binomial`.

5.1.1. Un ejemplo: Prediciendo la distribución de especies

Los modelos de distribución de especies son una de las herramientas más ampliamente usadas en ecología y biología de la conservación. Estos modelos asocian la presencia o ausencia de una especie a una serie de variables ambientales. Dichos modelos pueden extrapolarse posteriormente a la totalidad del territorio para predecir el rango de distribución “potencial” de dicha especie. La Junta de Andalucía ha impulsado recientemente un proyecto para predecir la distribución de todos los peces nativos y exóticos en los ríos andaluces.

En este ejemplo vamos a predecir si la presencia o ausencia de peces (variable 'Presencia') en ríos vadeables a lo largo de la cuenca del Guadalquivir depende del orden del tramo fluvial (variable 'Orden') y de la precipitación (variable 'Precipitacion') utilizando regresión logística¹. La base de datos (GLM_peces.txt)² está accesible en la siguiente dirección <http://tinyurl.com/mf2vx9>. Puedes descargar los datos directamente a tu ordenador y leerlos usando la función `read.table()` o leerlos directamente de la dirección web con la función `url()`.

```
> peces <- read.table(url("http://tinyurl.com/mf2vx9"), header = T,  
+ sep = "\t", dec = ",")
```

Visualizamos la base de datos con la función `str()` y `edit()`.

```
> str(peces)  
> edit(peces)
```

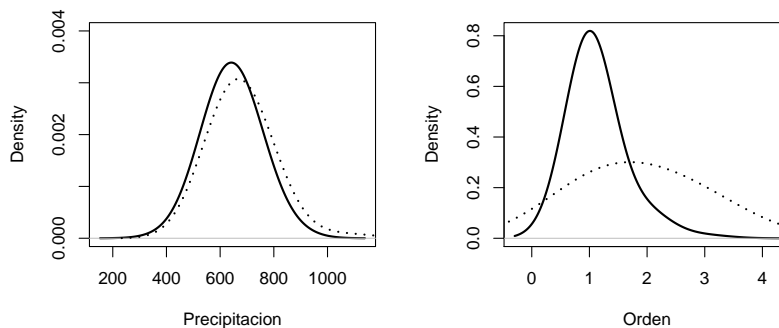
¹Hay muchos otros modelos que pueden utilizarse con este fin, como los GAM, modelos de sobre climáticos, índices de disponibilidad de hábitat, algoritmos genéticos, árboles de clasificación y regresión (CART), etc.

²Datos cedidos por Lucía Gálvez, Centro de Investigaciones de Recursos Cinegéticos (IREC). Estos datos no pueden ser usados para otros fines que no sean docentes sin permiso de la autora.

Para llevar a cabo el análisis, vamos a seguir los pasos descritos en la sección 2.

Exploración de los datos. No parece haber una diferencia muy clara en los valores de precipitación entre tramos con y sin peces. Sin embargo, la presencia de peces parece estar asociada, a primera vista, a tramos de río de orden superior.

```
> par(mfcol = c(1, 2))
> plot(density(peces$Precipitacion[peces$Presencia == 0], adjust = 3),
+      main = "", xlab = "Precipitacion", ylim = c(0, 0.004), lwd = 2)
> lines(density(peces$Precipitacion[peces$Presencia == 1], adjust = 3),
+       main = "", xlab = "", ylab = "", lty = 3, lwd = 2)
> plot(density(peces$Orden[peces$Presencia == 0], adjust = 3),
+      main = "", xlab = "Orden", lwd = 2)
> lines(density(peces$Orden[peces$Presencia == 1], adjust = 3),
+       main = "", xlab = "", ylab = "", lty = 3, lwd = 2)
```



Elección de la estructura de errores y función de vínculo. Como la variable respuesta es binomial (0-1) la familia de distribución de errores que elegiremos será la binomial. En este caso, es muy sencillo saber cómo analizar los datos. En principio, utilizaremos la función de vínculo canónica (logit), pero podríamos proponer un modelo alternativo utilizando una función de vínculo de tipo logarítmica para ver cuál ajusta mejor los datos al modelo.

Ajuste del modelo a los datos. Para ajustar el modelo a los datos, usaremos la función `glm()`. Es conveniente asignar el resultado del

ajuste del modelo a un nuevo objeto (p.e. `glm1` o `modelo1`). Este objeto será del tipo `glm` y contendrá información sobre los coeficientes del modelo, los residuos, etc. Para acceder a los resultados podemos utilizar las funciones `anova()` y `summary()`.

```
> peces$Orden <- as.factor(peces$Orden)
> glm1 <- glm(Presencia ~ Precipitacion + Orden, data = peces,
+           family = binomial)
> anova(glm1, test = "Chi")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: Presencia

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi)
NULL			149	207.944	
Precipitacion	1	6.288	148	201.656	0.01215 *
Orden	3	44.646	145	157.009	1.1e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> summary(glm1)
```

Call:

```
glm(formula = Presencia ~ Precipitacion + Orden, family = binomial,
    data = peces)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0596	-0.8621	-0.2077	0.8227	1.9165

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.214e+00	1.663e+00	-3.135	0.001717 **
Precipitacion	6.674e-03	2.448e-03	2.727	0.006393 **

```

Orden2      1.870e+00  4.795e-01   3.900  9.6e-05 ***
Orden3      3.950e+00  1.067e+00   3.700 0.000215 ***
Orden4      1.699e+01  1.455e+03   0.012 0.990687
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 207.94 on 149 degrees of freedom
Residual deviance: 157.01 on 145 degrees of freedom
AIC: 167.01

```

Number of Fisher Scoring iterations: 14

Como podemos ver, tanto la variable **Precipitacion** como el factor **Orden** son significativos ($P(>|\text{Chi}|) < 0.05$). Ahora bien, no todos los coeficientes del factor **Orden** son significativos. En principio, el **Intercept**, que resume el efecto del nivel de **Orden1** sobre la presencia de peces, es significativo y negativo. Esto indicaría que en niveles de **Orden1** la probabilidad de presencia de peces es menor que en el resto de niveles. Los niveles **Orden2** y **Orden3** también son significativos pero positivos, lo que indicaría que estos niveles incrementan la presencia de peces en los tramos fluviales. Por último, el nivel de **Orden4** no es significativo, lo que indicaría que el valor positivo del coeficiente no es significativamente distinto de cero y, por tanto, tiene un efecto nulo sobre la variable respuesta.

Por último, es interesante saber qué proporción de la varianza explica el modelo (es decir, la devianza). La función `summary()` nos da la *Null deviance* y la *Residual deviance*. Como vimos anteriormente, la devianza o varianza explicada por el modelo sería:

$$D^2 = \frac{\text{Devianza.modelo.nulo} - \text{Devianza.residual}}{\text{Devianza.modelo.nulo}} \cdot 100$$

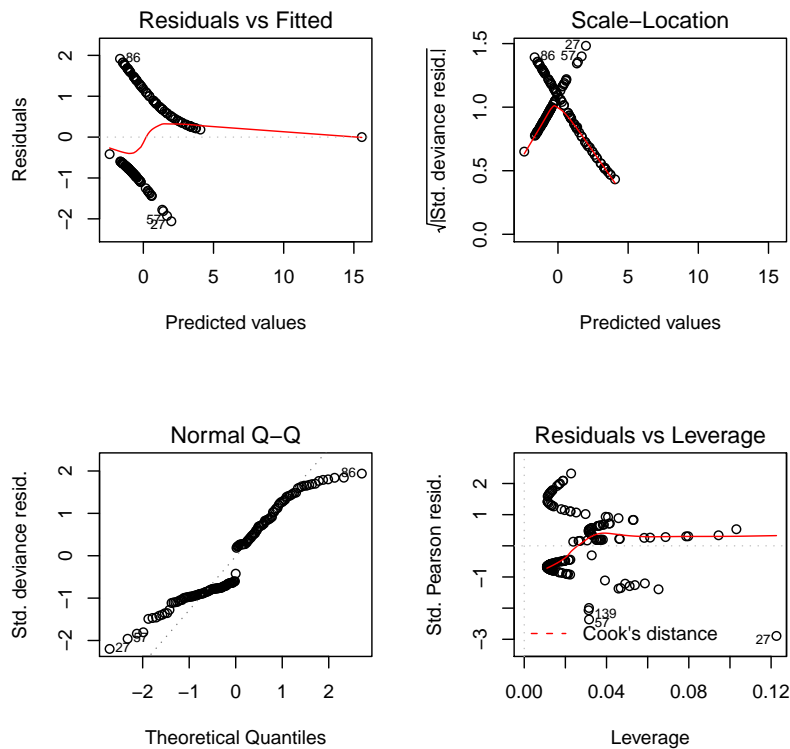
Si sustituimos estos valores obtenemos:

$$D^2 = \frac{207,97 - 157,01}{207,97} \cdot 100 = 24,49\%$$

Por lo que el modelo explicaría aproximadamente un 25 % de la presencia de peces en tramos fluviales.

Análisis de los residuos. La función genérica `plot()` muestra los principales gráficos de los residuos cuando su argumento es un objeto del tipo `lm` o `glm`.

```
> par(mfcol = c(2, 2))
> plot(glm1)
```



La función `plot()` genera un gráfico de residuos estandarizados frente a valores predichos (arriba izquierda), el gráfico probabilístico de normalidad (*q-q plot*, abajo izquierda) y el gráfico de valores atípicos (abajo derecha). El cuarto gráfico (arriba derecha) no ofrece ninguna información relevante para el análisis de los residuos. En el caso de los modelos binomiales, los gráficos de los residuos generalmente tienen formas poco “normales” dado que la respuesta siempre toma valores 0-1 y los valores predichos se mueven en el rango comprendido entre estos dos valores, por lo que el grado de discrepancia entre los valores observados y predichos por el modelo es generalmente grande. Sería también posible obtener un gráfico de los residuos con la función `hist()` y, de igual modo, hacer un test de normalidad de dichos residuos con la función `shapiro.test()`. La homogeneidad de varianzas entre grupos en el caso de variables explicativas categóricas podría realizarse con la función `levene.test()` del paquete `car`. Más importante sería, no obstante, investigar los datos atípicos y eliminar aquellos datos que estén

sobre-influyendo nuestro análisis. Estos datos se pueden detectar a primera vista en el **q-q plot** y el gráfico de valores atípicos.

Simplificación del modelo. En principio las dos variables son significativas. Sin embargo, parece que uno de los niveles del factor (**Orden4**) no es significativamente distinto de cero. Podríamos por tanto proponer un modelo alternativo con tres niveles del factor (juntando el nivel de **Orden1** y **Orden4**) en vez de cuatro y comparar la parsimonia de ambos modelos. Sin embargo, juntar estos dos niveles no parece tener mucho sentido biológico dado que los niveles del factor están ordenados. Por tanto, dejaremos el modelo como está.

5.2. Proporciones

En otras ocasiones la variable respuesta va a ser una proporción como:

- porcentaje de mortalidad,
- tasas de infección de enfermedades,
- proporción de pacientes que responden a un ensayo clínico,
- ratios de sexo en una población,
- porcentaje de germinación de plántulas en un estudio de vivero.

Todos estos datos tienen en común el hecho de que conocemos cuantos de los objetos experimentales están en una categoría (vivo, solvente, mujer, infectado) y cuantos están en la otra (muerto, insolvente, hombre, sano). En el caso de los datos de conteo de tipo Poisson, conoceríamos cuantas veces ocurre un evento, pero no cuantas veces no ocurre. Podemos modelar procesos que impliquen una respuesta proporcional en R especificando un GLM con una familia de tipo binomial, al igual que en el caso anterior. La única diferencia es que, mientras que en el caso de datos binomiales bastaba con indicar `family=binomial`, con los errores binomiales de tipo proporcional debemos además especificar el número de fallos así como el número de sucesos de un evento en una única variable respuesta. Para ello, debemos juntar dos vectores utilizando el comando `cbind()` en un único objeto, *y*, que constituirá la variable respuesta con el número de fallos y sucesos del evento en cuestión. El denominador binomial, *n*, es la muestra total, *y*:

```
numero.fallos = denominador.binomial - numero.sucesos
```

```
y = cbind(numero.sucesos, numero.fallos)
```

La forma tradicional de modelar este tipo de datos era usar el porcentaje de mortalidad (o la variable que fuera) como la variable respuesta. Hay cuatro problemas con esto:

1. Los errores no están normalmente distribuidos,
2. La varianza no es constante,
3. La respuesta está acotada a valores entre 0 y 1 (o 0 y 100),
4. Calculando el porcentaje, perdemos información sobre el tamaño de la muestra, n , a partir de la cual se estima la proporción.

5.2.1. Análisis de proporciones para factores con uno y dos niveles

Para comparaciones de una proporción binomial con una constante, podemos utilizar la función `binom.test()`. Para comparaciones de dos muestras de datos proporcionales (como en el test de la t), podemos utilizar la función `prop.test()`. El método que se explica a continuación permite comparar datos proporcionales entre varios niveles de un factor o la influencia de variables continuas (regresión binomial) sobre la variable binomial respuesta.

5.2.2. Un ejemplo: ¿Son eficientes los pesticidas en el control de plagas?

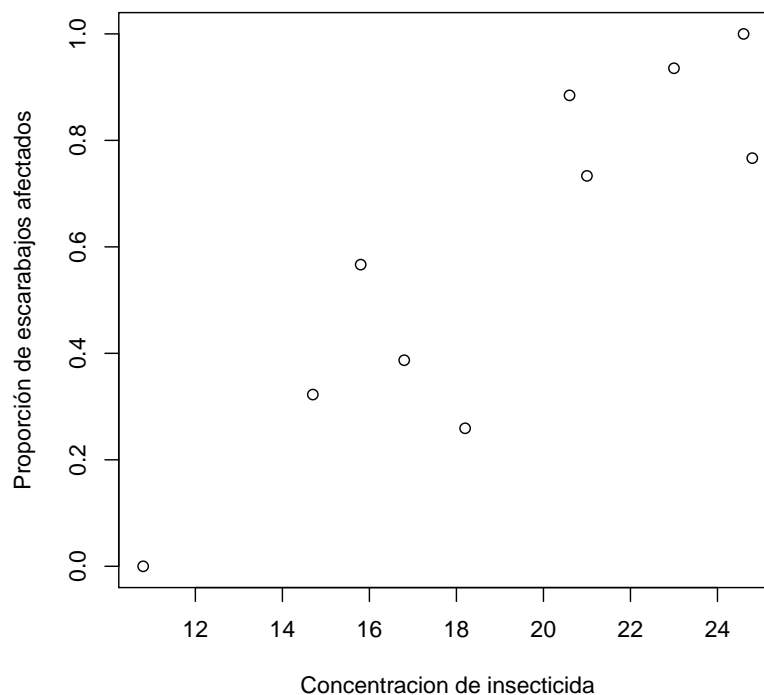
En este ejemplo vamos a ver cómo afectan los pesticidas a la proporción de escarabajos en cultivos de maíz. En cada cultivo se mide la concentración de pesticidas aplicada (`conc`), el número total de escarabajos capturados en trampas de feromonas (`exposed`) y, de éstos, los que se ven afectados por los pesticidas, que acaban muriendo en las trampas (`affected`). Vamos a seguir la misma secuencia de pasos definida en el capítulo 2. Los datos están disponibles en el archivo de datos `beetle` del paquete `faraway`.

```
> library(faraway)
> data(beetle)
> str(beetle)
```

```
'data.frame':      10 obs. of  3 variables:
 $ conc      : num  24.8 24.6 23 21 20.6 18.2 16.8 15.8 14.7 10.8
 $ affected: num  23 30 29 22 23 7 12 17 10 0
 $ exposed : num  30 30 31 30 26 27 31 30 31 24
```

El gráfico de dispersión de la variable respuesta (**yprop**) y la variable explicativa (**conc**) parece indicar que, efectivamente, hay una relación positiva entre la concentración de insecticida aplicada y la mortalidad de escarabajos.

```
> attach(beetle)
> yprop <- affected/exposed
> plot(yprop ~ conc, ylab = "Proporción de escarabajos afectados",
+      xlab = "Concentracion de insecticida")
```



Vamos a modelar esta relación con un GLM.

```
> y <- cbind(affected, exposed - affected)
> glm.beetle1 <- glm(y ~ conc, family = binomial)
```



```
> summary(glm.beetle1)
```

Call:

```
glm(formula = y ~ conc, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.0670	-1.7117	0.1495	1.7340	2.4150

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-6.01047	0.77937	-7.712	1.24e-14 ***
conc	0.34127	0.04153	8.217	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

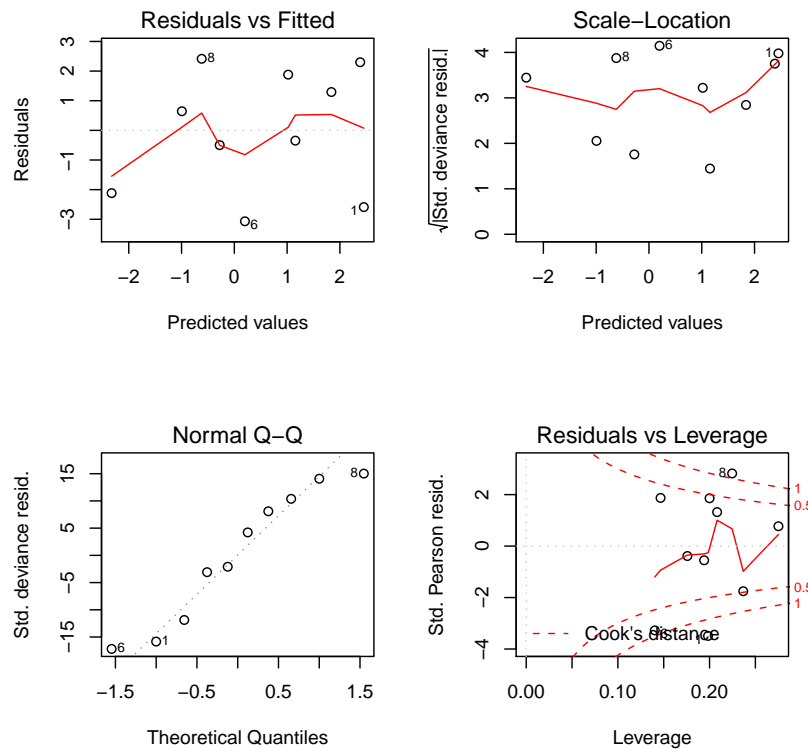
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 138.001 on 9 degrees of freedom
 Residual deviance: 37.697 on 8 degrees of freedom
 AIC: 69.27

Number of Fisher Scoring iterations: 5

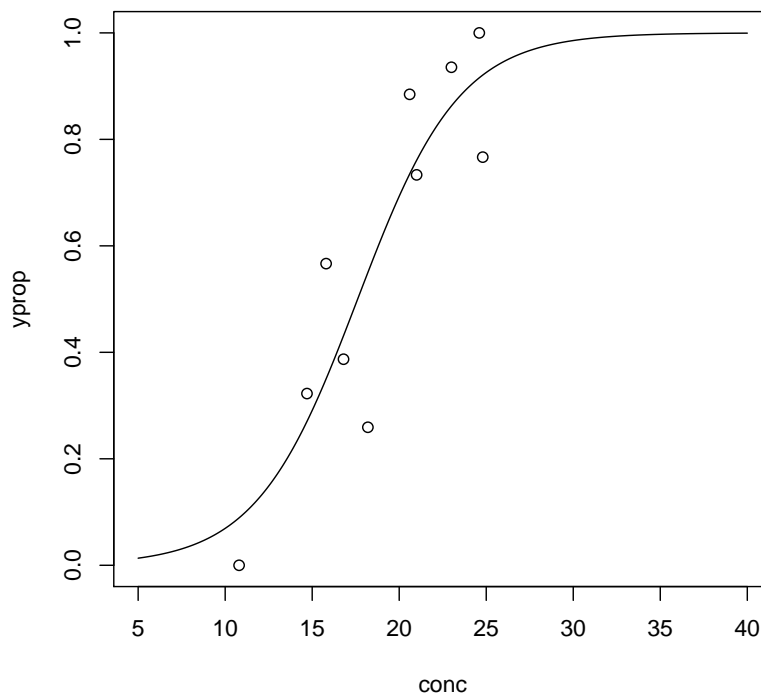
Lo que confirma nuestro supuesto: hay una relación positiva entre la mortalidad de escarabajos y la concentración de insecticidas aplicada a los cultivos. Por tanto el uso de insecticida parece una medida eficaz para reducir las plagas de escarabajo en los cultivos de maíz. ¿Qué variabilidad es explicada por el modelo? Debemos fijarnos, como en el caso anterior, en la relación entre la devianza nula y la devianza residual. Analicemos ahora los residuos del modelo.

```
> par(mfcol = c(2, 2))
> plot(glm.beetle1)
```



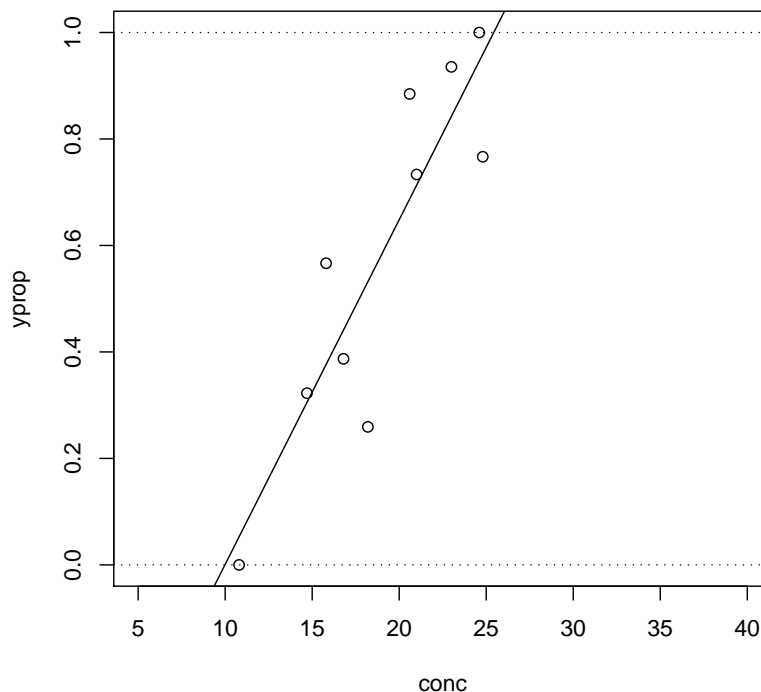
Ahora sí que los gráficos de los residuos son más explicativos de la normalidad (o falta de normalidad) de los mismos. Vemos que el gráfico de los residuos estandarizados frente a los valores predichos representa más o menos una nube de puntos, lo cual demuestra normalidad. También vemos que no hay desvíos muy grandes con respecto a la diagonal en el **q-q plot**, lo cual parece confirmar la linealidad de los datos. Por último, no hay datos atípicos ni sobre-influyentes. Por tanto el modelo parece bastante adecuado. Vamos a ver ahora la representación gráfica de los valores predichos por el modelo, para demostrar que este modelo permite acotar los valores entre 0 y 1.

```
> xv <- seq(5, 40, 0.1)
> yv <- predict(glm.beetle1, list(conc = xv), type = "response")
> plot(yprop ~ conc, xlim = c(5, 40))
> lines(xv, yv)
```



¿Qué ocurriría si utilizáramos un modelo de regresión (estructura de errores de tipo gaussiana o normal) para estos mismos datos? Lo que ocurriría es que, si bien el modelo sería adecuado desde el punto de vista de la normalidad y homocedasticidad de los residuos (resultados no mostrados), los valores predichos no estarían acotados entre 0 y 1 y, por tanto, el modelo no sería el más adecuado para representar una respuesta proporcional.

```
> glm.beetle2 <- glm(yprop ~ conc, family = gaussian)
> xv <- seq(5, 35, 0.1)
> yv <- predict(glm.beetle2, list(conc = xv), type = "response")
> plot(yprop ~ conc, xlim = c(5, 40))
> lines(xv, yv)
> abline(h = 0, lty = 3)
> abline(h = 1, lty = 3)
```



6. Modelos Poisson

Los modelos Poisson se utilizan generalmente para representar datos de conteos, es decir, datos enteros positivos, como por ejemplo el número de individuos que mueren (pero ¡ojó! no la proporción de individuos que mueren), el número de empresas que van a bancarrota, el número de días que hiela o la abundancia de una determinada especie. Con datos de conteos, el 0 aparece como un valor más de la variable respuesta, pero valores negativos no tienen lugar. En los conteos por tanto vamos a estar

interesados en modelar la frecuencia de un determinado suceso, pero sin tener información sobre el número de veces que dicho suceso NO tiene lugar. En el caso de tener información sobre el número de veces que dicho suceso NO tiene lugar, estaríamos ante datos proporcionales y, por tanto, un modelo con distribución de errores de tipo binomial sería mucho más apropiado.

El uso de modelos lineales (es decir, asumiendo varianza constante y errores normales) no sería adecuado ante datos de conteo por las siguientes razones:

1. El modelo lineal podría predecir valores negativos de la variable respuesta.
2. La varianza de la variable respuesta aumentará probablemente a medida que aumenta la media (varianza no constante).
3. Los errores no están normalmente distribuidos.
4. Los ceros son difícil de manejar en transformaciones de la variable respuesta.

En R, los datos de tipo conteo se pueden modelar de manera muy elegante mediante el uso de GLM con una distribución de errores de tipo Poisson (`family = poisson`) y `link = log`. La función de vínculo de tipo logarítmico asegura que todos los valores predichos sean positivos, mientras que la distribución de errores de tipo Poisson tiene en cuenta el hecho de que los datos son enteros y que la varianza aumenta proporcionalmente a la media.

7. Ejercicios

A continuación se proponen tres ejercicios. Para resolverlos, la clase se dividirá en tres grandes grupos y, dentro de cada grupo, los alumnos se dispondrán por parejas para resolver uno de los tres casos de estudio. Para ello hay 45 minutos. El código que se genere deberá guardarse en un archivo de texto. Una vez resuelto el ejercicio, todas las parejas dentro del mismo grupo pondrán los resultados en común y lo resolverán de manera conjunta en la wiki del curso (<http://curso-r-ceama2009.wikispaces.com/>).

1. El archivo `deforestacion.txt` contiene datos de cambio de cobertura forestal en Chile central calculados a partir de dos clasificaciones

obtenidas a partir de imágenes satelitales Landsat para el periodo 1999-2008. Se quiere modelar la deforestación (variable **defores**, dónde 1 = deforestado y 0 = no deforestado) en función de una serie de variables físicas: elevación (**Elev**), pendiente (**Pendiente**), distancia a ríos (**Dist.rio**), distancia a ciudades con más de 20.000 habts (**Dist.ciudad**), distancia a ciudades con menos de 20.000 habts (**Dist.pueblo**), distancia a carreteras (**Dist.carP**), distancia a caminos y pistas forestales (**Dist.carS**), distancia a suelo agrícola (**Dist.agri**) e insolación (**Insolac**). El archivo puede descargarse de la siguiente dirección web (<http://tinyurl.com/yklbde4>)³. Hay que buscar el modelo más parsimonioso (es decir, que tenga el menor número de variables significativas). Para ello, elimina las variables que no son significativas y vuelve a ajustar el modelo. Luego compara los resultados con los que obtendrías utilizando la función **stepAIC()**. ¿Se llega al mismo modelo final reducido por ambas vías?

2. El archivo **ailanto.txt** contiene datos de germinación de semillas de ailanto en un experimento de vivero en el que 48 bandejas, con cerca de 40 semillas de ailanto cada una, fueron sometidas a distintas exposiciones de luz (0-100 %) y agua (0-100 %). El archivo de datos puede descargarse de la siguiente dirección web (<http://tinyurl.com/lzco8s>)⁴. El objetivo es construir un modelo que explique la germinación de semillas en función de la luz y el agua (y su interacción). Para poder trabajar con los datos hay que transponer el arreglo de datos con la función **t()**. Pistas, para ver gráficamente la interacción entre dos factores se puede utilizar la función **interaction.plot()**.
3. McArthur y Wilson (1967) propusieron que el número de especies en una isla representa un equilibrio dinámico entre tasas opuestas de migración y extinción, dos procesos recurrentes que mantienen la riqueza de especies relativamente constante pese a cambios en su composición. Estos procesos se ven afectados por el tamaño y aislamiento de la isla. El primero afecta a la tasa de extinción de manera negativa (islas más grandes tendrán menor tasa de extinción y, por tanto, mayor número de especies), mientras que el segundo afecta de manera negativa a la tasa de migración (islas más

³Datos cedidos por Jennifer J. Schulz, Universidad de Alcalá. Estos datos no pueden ser usados para otros fines que no sean docentes sin permiso de la autora.

⁴Datos cedidos por Oscar Godoy y Pilar Castro, Universidad de Alcalá. Estos datos no pueden ser usados para otros fines que no sean docentes sin permiso de los autores.

aisladas tendrán mayor aislamiento). La relación entre el aislamiento y el número de especies no es, sin embargo, tan obvia, puesto que menor migración supondría en principio una menor riqueza de especies pero, por otro lado, favorecería el proceso de especiación (formación de nuevas especies) en las islas aisladas, aumentando por tanto la riqueza de especies. En este ejercicio, vamos a utilizar los datos de Johnson & Raven (1973)⁵, accesibles en el archivo de datos **gala** del paquete **faraway**. Estos datos contienen información sobre el número de plantas (**Species**) en 30 islas del archipiélago de las Galápagos, así como información de distintas variables, incluyendo el área de la isla (**Area**) y su proximidad a otras islas (**Nearest**). Vamos a ajustar un modelo GLM para ver si existe un efecto del área y el aislamiento de las islas sobre la riqueza de especies.

⁵M.P. Johnson and P.H. Raven. 1973. Species number and endemism: The Galapagos Archipelago revisited. *Science* 179: 893-895.