GLM-Introducción

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 dependiente se relaciona linealmente con la(s) variable(s) independiente(s).

de manera analítica tendríamos:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + U_i \\ E(U_i) &= 0 \quad i = 1, \ldots, n \end{aligned}$$

tomado la esperanza

$$E(Y_i) = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Además, suponemos que existe un comportamiento normal $U_i \backsim N(0, \sigma^2)$ o bien, equivalentemente $Y_i \backsim N(\mu, \sigma^2)$, siendo $\mu = E(Y_i) = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$

En el modelo de regresión lineal múltiple:

- El **predictor lineal**: $\beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- La relación entre:

$$E(Y_i)$$

y

$$\beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

es la **identidad**

$$E(Y_i) = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

- La distribución de probabilidad para Y_i es la **normal**.

En muchas ocasiones, sin embargo, nos encontramos con que uno o varios de estos supuestos no se cumplen por la naturaleza de la información.

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 interpretar los resultados obtenidos, si utilizamos transformaciones de la variable.

Una alternativa a la transformación de la variable dependiente/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) .

McCullagh, Peter & Nelder, John A. (1983, 1989) Generalized Linear Models, Chapman & Hall

Los GLM son, por tanto, 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 dependientes sufren invariablemente la violación de estos dos supuestos de los modelos normales y los GLM ofrecen una buena alternativa para tratarlos.

Estaríamos en el supuesto GLM

Cuando la variable dependiente/respuesta/endógena es

►Un variable de conteo, en concreto, casos (ejemplo: número de colisiones, accidentes, viviendas destruidas…)

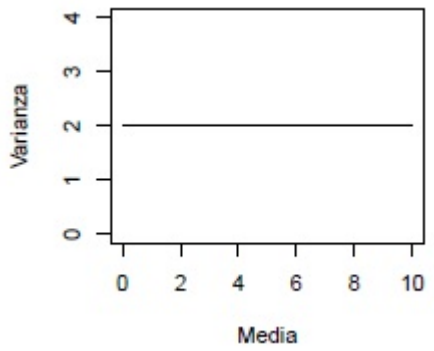
►Un variable de contero de casos expresados éstos como proporciones (ejemplo; porcentaje de heridos graves en accidentes, porcentaje de no carnet…)

►Una variable establecida como binaria (ejemplo: vivo o muerto, hombre mujer, carnet o no , joven o mayor.)

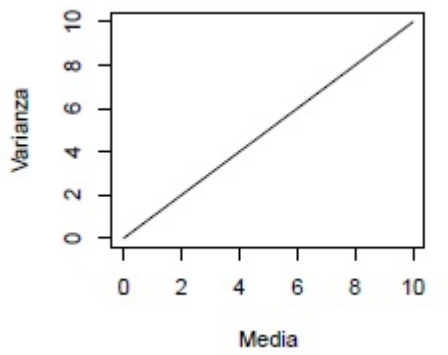
Varianza no Constante

Al analizar submuestras de datos de los que disponemos podemos encontrarnos con diversas realidades .

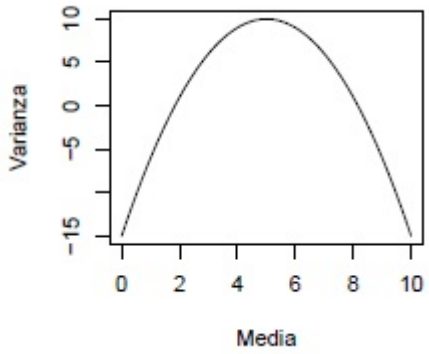
■ El supuesto central que se hace en los modelos lineales es que la varianza es constante así tendríamos que al variar la media se ésta(la varianza) se mantiene constante:



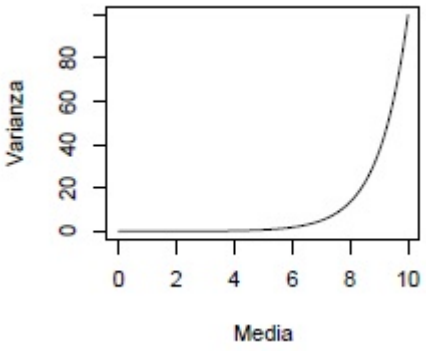
■ En el caso de variables de conteo como variable dependiente , sin embargo, donde ésta se expresa en números enteros y en dónde puede haber muchos ceros en los datos, la varianza se suele incrementar linealmente con la media:



■ Con proporciones de eventos como variable explicada es muy posible que la varianza se comporte en forma de U invertida :



■ Cuando la variable respuesta/dependiente se aproxime a una distribución Gamma, entonces la varianza se incrementa de una manera no lineal con la respecto a la media :



Son en estos casos de varianza no constante donde será necesario aplicar los GLM

Normalidad

Muchos datos tienen una estructura no normal. Las herramientas habituales para tratar la ausencia de normalidad eran la transformación de la variable respuesta o la adopción de métodos no paramétricos. Otra alternativa, son los modelos lineales generalizados o GLM. Los GLM permiten especificar distintos tipos de distribución de errores así:

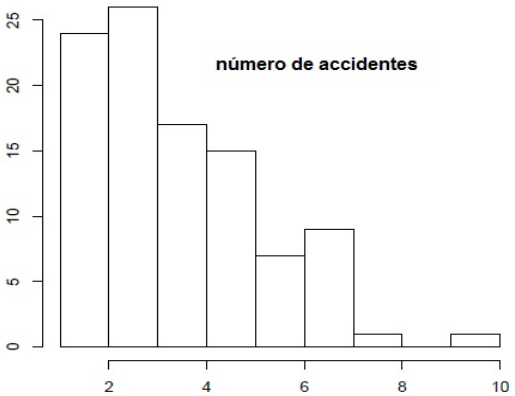
► Poisson, muy útiles para conteos de acontecimientos .Ejemplo: número de heridos por accidentes de tráfico ; número de hogares asegurados que dan parte de siniestro al día .

► Binomiales, de gran utilidad para proporciones y datos de presencia/ausencia Ejemplo : tasas de mortalidad; tasas de infección ; porcentaje de siniestros mortales .

► 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 Ejemplo : número de heridos en función del número de siniestros

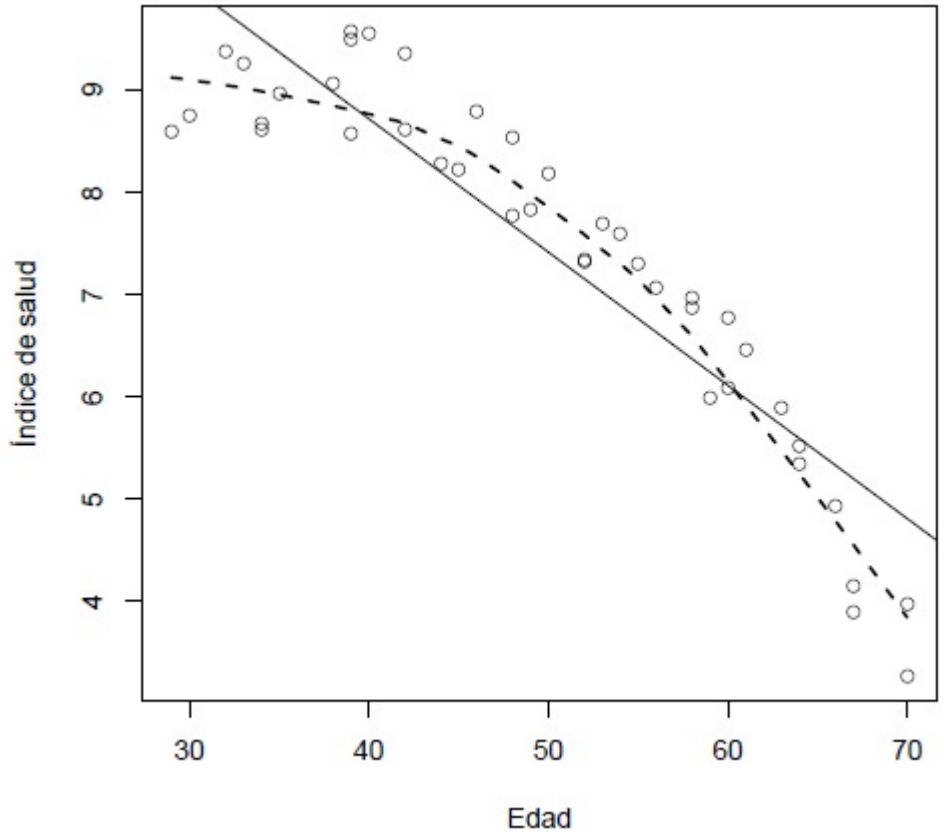
► Exponenciales, muy útiles para los análisis de supervivencia

Además, los modelos lineales, habituales ,asumen que tanto la variable respuesta como los errores del modelo siguen una distribución normal. Una distribución normal que es, como es sabido 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 deberá 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 cuantos accidentes e producen al día en un determinado municipio en base a datos de días con determinado número de accidentes como los de la figura siguiente En este caso, es razonable asumir que la variable dependiente seguirá una distribución de tipo Poisson y no una normal como en muchas ocasiones se utiliza por “comodidad”

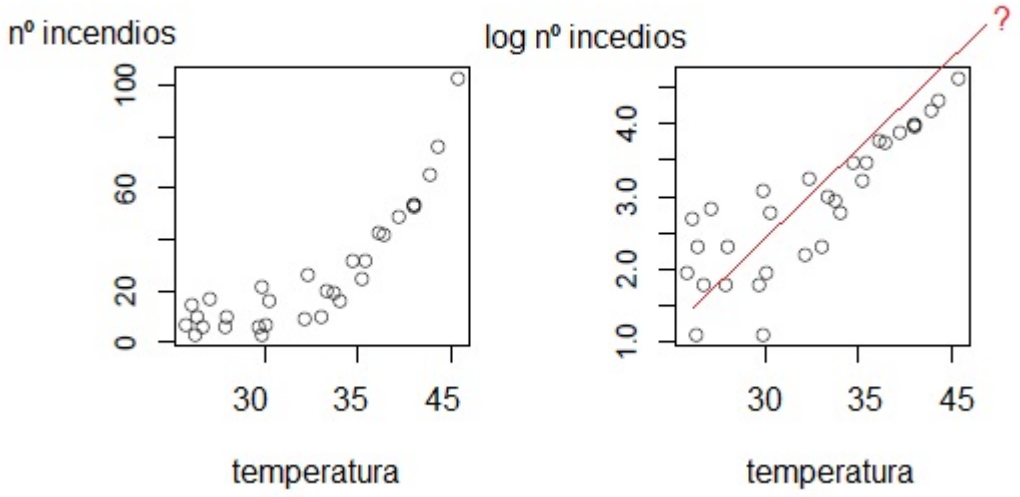


Función vínculo/ligadura

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 . La salud de la gente de 30 años no es muy distinta de la de la gente de 40. Sin embargo, las diferencias son más marcadas entre la gente de 60 y 70 años. Por tanto, la relación entre edad y salud no es de tipo lineal. Tal vez una función de tipo exponencial parece ser más adecuada .



La función de vínculo se encarga de linealizar la relación entre la variable dependiente y la(s) variable(s) independiente(s) mediante la transformación de la variable respuesta. Tomemos por ejemplo la relación entre el número de incendios forestales y la temperatura exterior .
Esta relación como podemos ver en el gráfico siguientes 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:

$$log(y_i) = \beta_0 + \beta_i x_i$$

$$donde:$$

$$log(y_i) = logaritmos\ de\ n^o\ de\ incendios$$

$$x_i = temperaturas$$

Ahora bien, los valores estimados por este modelo no son los valores de y, sino los del 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:

$\exp(log(y_i)) = \exp(\beta_0 + \beta_i x_i)$ por tanto :

$$y_i = \exp(\beta_0 + \beta_i x_i)$$

es evidente que es más intuitivo trabajar con el número de incendios que con el "logaritmo" del número de incendios

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 incendios. En este caso, una función de vínculo de tipo logarítmica resolverá el problema de la acotación En otras situaciones la variable respuesta/dependiente 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 mas apropiada.

Especificación del modelo

La especificación de un modelo lineal generalizado se realiza en tres partes:

- La **componente aleatoria** corresponde a la variable $\mathbf{Y}$ que sigue una distribución de la familia exponencial (normal, log-normal, Poisson, gamma,...). Además denotaremos por $\boldsymbol{\mu}$ a su esperanza matemática.
- La **componente sistemática**, también llamada predictor lineal, se denota por $\boldsymbol{\eta}$ y corresponde al vector de n componentes, siendo cada una de ellas igual a $\eta_i = \sum_{j=1}^{p} \beta_j x_{ji} = \mathbf{x}_i' \boldsymbol{\beta}$
- La **función de ligadura** (o función link, $g(\cdot)$) relaciona la esperanza matemática de la variable dependiente con el predictor lineal $\eta_i = g(\mu_i)$, $i = 1, \ldots, n$. La función de ligadura debe ser monótona y diferenciable.

Una varible aleatoria **Y** sigue una distribución de la **familia exponencial** si su densidad puede escribirse como:

$$f_Y(y;\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\}$$

siendo $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ funciones conocidas.

- Si ϕ es conocido, se cumplirá que $E(Y) = \mu = b'(\theta)$, donde $'$ indica la diferenciación respecto a θ. Además $var(Y) = b''(\theta)a(\phi)$.
- Cuando escribamos la varianza como $var(\mu)$ es porque estará en función del parámetro μ.

Resumen las funciones de ligadura/vínculo mas utilizadas:

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$Log(\mu)$	Conteos con errores de tipo Poisson
Logit	$Log(\frac{\mu}{n-\mu})$	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

Se denominan funciones de ligadura/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 siguiente tabla se plasma las funciones de vínculo canónicas para cada una de las distribuciones de errores, así como otras posibles funciones de vínculo habitualmente 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

En la siguiente tabla 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

Especificación de la estimación

La estimación de los parámetros $\beta_1, ..., \beta_p$ se realiza por el **método de máxima verosimilitud.**

Para valorar el ajuste de los modelos lineales generalizados podemos utilizar el **estadístico de Chi-cuadrado**. Se define como el doble de la diferencia entre el máximo del logaritmo de la verosimilitud que se podría conseguir con la mínima (o máxima) parametrización y el valor del máximo del logaritmo de la verosimilitud que se consigue con el modelo que se quiere evaluar.

Los **ajustes** de $\hat{\mu}_i$ que se calculan como $g^{-1}(\sum_{j=1}^{p} \widehat{\beta}_j x_{ji})$, una vez estimados los parámetros del vector $\boldsymbol{\beta}$.

Los **residuos** de Pearson son los más utilizados y se definen como:

$$r_P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}.$$

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 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 deberán 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 se detallan:

■ Exploración de los datos.(E.D.A. exploratory data analysis ,J.Tukey) Conviene conocer nuestros datos. Puede resultar interesante obtener gráficos que nos muestren la relación entre la variable explicada 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 análisis exploratorio es: a) Buscar posibles relaciones de la variable respuesta/dependiente 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, y será a posteriori cuando comprobemos , analizando los residuos, 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 del 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:

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

■ 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 comúnmente 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.

■ 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 de normalidad (q-q plot), que permite contrastar la normalidad (simetría) de la distribución de los residuos.

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. Gráficos de los residuos frente al tiempo (u orden de medida). Permiten detectar cambios sistema ticos 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. Gráficos de valores atípicos. Existen tests que permiten detectar valores atípicos. Los índices más comunes son el índice de Cook y el de apalancamiento o leverage.
Todos estos gráficos (y opcionalmente algunos tests estadísticos complementarios) nos pueden ayudar en la evaluación del modelo utilizado.

En caso necesario, sería 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 desviando 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).*

Los modelos de 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), el número de empresas que van a bancarrota…. 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 modelizar 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 aumentaría 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.

Desarrollo en ejemplo:

- Para cada asegurado i, $i = 1, ..., n$, el comportamiento de Y_i (el número de siniestros sufridos durante un año) sigue una distribución de Poisson de parámetro λ_i.
- El parámetro es distinto para cada individuo y depende de sus características de riesgo.
- Suponiendo la función de ligadura se establece que $\lambda_i = \exp(\sum_{j=1}^{p} \beta_j x_{ji})$.
- Una vez estimados los coeficientes o parámetros del modelo $\beta_1, ..., \beta_p$, se puede predecir λ_i en función de las características individuales.

La probabilidad de que el asegurado i sufra y_i siniestros, bajo el anterior modelo de Poisson es:

$$\Pr(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Por lo tanto:

$$\Pr(Y_i = y_i) = \frac{\exp(-\exp(\sum_{j=1}^{p} \beta_j x_{ji}))[\exp(\sum_{j=1}^{p} \beta_j x_{ji})]^{y_i}}{y_i!}$$

La función de verosimilitud se escribe:

$$L(\beta_1, ..., \beta_p) = \prod_{i=1}^{n} \Pr(Y_i = y_i)$$

y su logaritmo:

$$\begin{aligned} \ell(\beta_1, ..., \beta_p) &= \sum_{i=1}^{n} \ln \Pr(Y_i = y_i) = \\ &= \sum_{i=1}^{n} y_i (\sum_{j=1}^{p} \beta_j x_{ji}) - \sum_{i=1}^{n} \exp(\sum_{j=1}^{p} \beta_j x_{ji}) - \ln(y_i!) \end{aligned}$$

ejemplo en R

Supongamos una muestra de 80 994 asegurados de una entidad española. La siniestralidad que se observa en este grupo de conductores aparece en la siguiente tabla:

Número de siniestros	Frecuencia
0	75 428
1	5 127
2	405
3	31
4	3

El promedio es 0.075

■ los datos del ejemplo se encuentran en el archivo:

http://www.uv.es/lejarza/eaa/tareas/t4r/pois.csv

ingresar datos en R

sentencia/código

datos<-read.table("http://www.uv.es/lejarza/actu/glm/pois.csv",header=TRUE, sep=";")

variables utilizadas:

SIN2	Número de siniestros	
V9	sexo	1 mujer 0 hombre (dicotómica)
V10	Zona conducción	1 urbana 0 resto
V13	Experiencia media	1 entre 4-14 años, 0 resto
V14	Experiencia max	1 más de 15, 0 resto
V17	Edad	1 30 o mas , 0 resto
V18	Coberturas adicionales	1 si , 0 no
V19	Coberturas todo riesgo	1 si , 0 no
V20	potencia	1 más 120cv , 0 no

sentencia/código

```
pois<-
glm(datos$SIN2~datos$V9+datos$V10+datos$V13+datos$V14+datos$V17+datos$V18+datos$V19
+datos$V20,family=poisson(link="log"))
```

sentencia/código

```
summary(pois)
```

resultado:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.47468    0.16552 -14.951  < 2e-16 ***
datos$V9    -0.03090    0.06481  -0.477 0.633553
datos$V10   -0.04861    0.04924  -0.987 0.323528
datos$V13   -0.22010    0.14129  -1.558 0.119279
datos$V14   -0.35123    0.15267  -2.301 0.021413 *
datos$V17    0.11895    0.09924   1.199 0.230666
datos$V18    0.23387    0.06572   3.559 0.000373 ***
datos$V19    0.07946    0.05342   1.488 0.136840
datos$V20    0.16924    0.06734   2.513 0.011961 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9781.9  on 24207  degrees of freedom
Residual deviance: 9737.3  on 24199  degrees of freedom
AIC: 13174
```

consecuencias

dada que hemos tratado un modelo Poisson la función de vínculo/ligadura sería:

$$\lambda_i = \exp\left(\sum_{j=1}^{p} \beta_j x_{ji}\right).$$

en base a lo obtenido:
la predicción para (por ejemplo) un hombre de 40 años, que conduce en zona urbana ,con 20 de experiencia,que no tiene coberturas adicionales pero si a todo riesgo y con vehículo de 135cv. Es decir:

V9=0,V10=1,V13=0,V14=1,V17=0,V18=0,V19=1 , V20=1

El modelo sería:

Exp[-2,47468-0,0309(0)-0,04861(1)-0,22010(0)-
0,35123(1)+0,11895(0)+0,23387(0)+0,07956(1)+0,16924(1)]
Exp[-2,47468-0,04861-0,35123+0,07956+0,16924]=
Exp[-2,6257] =0,07238 siniestros , un poco por debajo de la media.

Evaluación del modelo

Según el Criterio de Información de Akaike (AIC del inglés Akaike Information Criterion) el resultado es 13174 , ni alto ni bajo , serviría para comparar con otros AIC

En base a la devianza

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

tendríamos: $D^2 = \frac{9781,9 - 9737,3}{9781,9} \cdot 100 = 0,4559$

luego el modelo explica el 0,4559 % de la variabilidad ,luego……

Los Modelos binomiales

Respuestas binarias (regresión logística)

Muchas variables respuesta son del tipo:
vivo o muerto,
hombre o mujer,
infectado o saludable,
En estos casos podemos investigar que 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 ceros y unos. La manera en la que los GLM tratan datos binarios es asumiendo que los ceros y los unos 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 o 1) vendrá dado por la forma abreviada de la distribución binomial con n = 1, conocida como la distribución de Bernoulli/dicotómica:

Ejemplo en R

En este ejemplo vamos a predecir si la presencia o ausencia de accidentes de automóvil (variable “presencia”) en diversos tramos de carretera depende de las características de tramo (variable “orden”, cuatro órdenes) y de la precipitación de lluvia (variable 'Precipitacion') utilizando regresión logística.

■La base de datos se encuentra en:

http://www.uv.es/lejarza/eaa/tareas/t4r/trese.txt

ingresar datos en R
sentencia/código

pece <- read.table(url("http://www.uv.es/lejarza/actu/glm/trese.txt"), header = T, sep = "\t", dec = ",")

donde se ha creado una base denominada “pece”

con la sentencia/código

str(pece)

obtenemos información sobre los datos:

```
> str(pece)
'data.frame':   150 obs. of  7 variables:
 $ Presencia    : int  1 1 0 0 1 0 0 0 0 1 ...
 $ montana      : num  -0.7 -0.7 -0.7 -0.69 0.08 0.08 0.08 0.(
 $ Densidad     : num  0 0 0.05 0.22 0.19 0 0.17 0.11 0 0.21
 $ usos         : num  1.55 2.28 0.18 -4.65 -0.23 -2.02 -0.52
 $ Orden        : num  1 2 1 1 4 1 2 1 1 3 ...
 $ Precipitacion: num  808 1110 754 421 568 535 570 690 668 5!
 $ Sup.Tramo    : num  140 953.9 41.2 109.4 2009.9 ...
```

Exploración de los datos (EDA). Observando los datos .No parece haber una diferencia muy clara en los valores de precipitación entre tramos con y sin presencia de accidentes. Sin embargo, la presencia de accidentes parece estar asociada, a primera vista, a tramos con número (orden) mayor

Con el código siguiente podemos llevar a cabo los gráficos que explicitan mejor lo anteriormente dicho

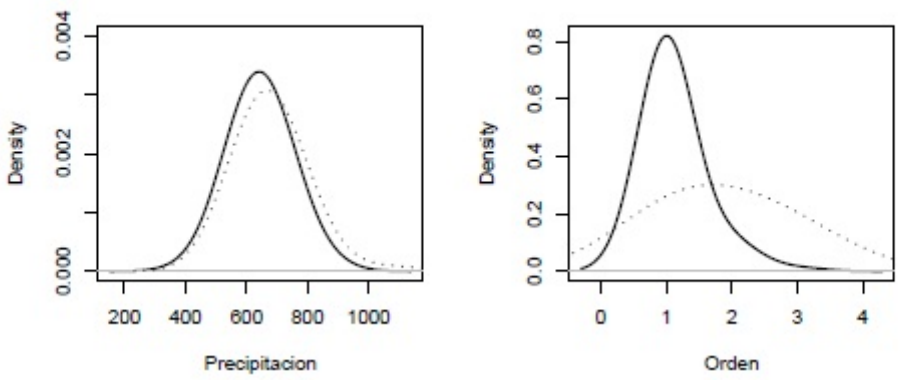
```
par(mfcol = c(1, 2))

plot(density(pece$Precipitacion[pece$Presencia == 0], adjust = 3), main = "", xlab = "Precipitacion", ylim = c(0, 0.004),lwd = 2)

lines(density(pece$Precipitacion[pece$Presencia == 1], adjust = 3),main = "", xlab = "", ylab = "", lty = 3, lwd = 2)

plot(density(pece$Orden[pece$Presencia == 0], adjust = 3), main = "", xlab = "Orden", lwd = 2)
```

lines(density(pece$Orden[pece$Presencia == 1], adjust = 3),main = "", xlab = "", ylab = "", lty = 3, lwd = 2)



en estos gráficos así creados lo punteado se refiere a valores 1(si) de presencia de accidentes. En precipitación casi no hay diferencias , en tramos si

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, (en R) 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().

Así la sentencia/ código será

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

el resultado obtenido sería:

```
> pece$Orden <- as.factor(pece$Orden)
> glm1 <- glm(Presencia ~ Precipitacion + Orden, data = pece, 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     207.94
Precipitacion  1    6.288       148     201.66   0.01215 *
Orden          3   44.646       145     157.01   1.1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Donde se observa que ambas variables son significativas

Mediante el comando

Summary(glm1)

Obtenemos:

```
> summary(glm1)

Call:
glm(formula = Presencia ~ Precipitacion + Orden, family = binomial,
    data = pece)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0596  -0.8620  -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
```

Donde se observan los valores estimados de los parámetros así como las devianzas y el AIC

Como podemos ver, tanto la variable Precipitación como el factor Tramo(orden) son significativos (P(>|Chi |<0.05). Ahora bien, no todos los coeficientes del factor tramo son significativos. En principio, el Intercept, que resume el efecto del nivel de Orden1 sobre la presencia de accidentes , es significativo y negativo. Esto indicaría que en niveles de Orden1 la probabilidad de presencia de accidentes es menor que en el resto de niveles. Los niveles Orden2 y Orden3 también son significativos pero positivos, lo que indicaría que en estos tramos/niveles se incrementa presencia de accidentes. 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).

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

tendríamos: $D^2 = \frac{207,94 - 157,01}{207,94} \cdot 100 = 24,49$

luego el modelo explica el 24,49 % de la variabilidad

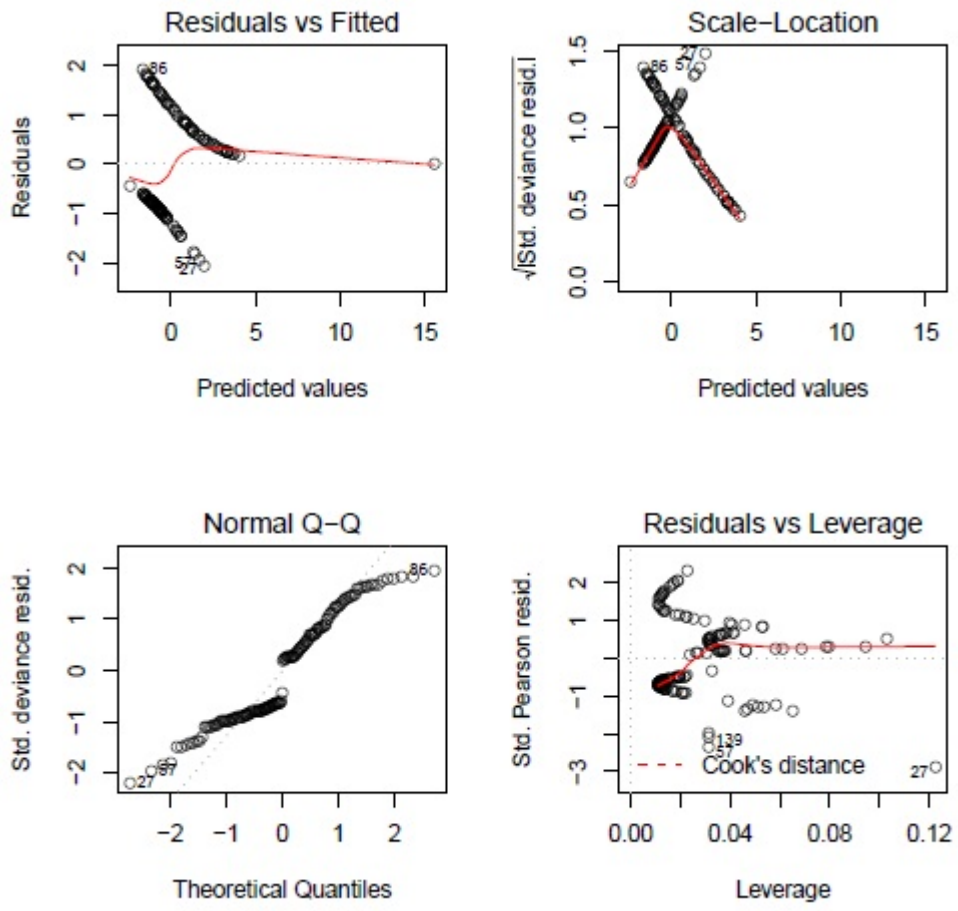
el AIC es de 167

Análisis de los residuos.

Mediante los comandos

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

obtendríamos



La función plot() de R genera un gráfico de residuos estandarizados frente avalores 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. Más importante será, no obstante, investigar los datos atípicos y eliminar aquellos datos que estén sobre-inuyendo nuestro análisis. Estos datos se pueden detectar a primera vista en el q-q plot y el gráfico de valores atípicos (abajo derecha. Ej 27).

Simplificación del modelo. En principio las dos variables son significativas. Sin embargo, parece que uno de los niveles del factor tramo (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 si está fisicamente lejanos. Por tanto, dejaremos el modelo como está.

Apéndice R

La función glm()

La funcióon glm() viene especifcada 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 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().

Todo el material basado en :

-Modelos lineales generalizados (GLM) – Luis Cayuela –Universidad de Granada
-Modelos lineales generalizados , en seguros – Monserrat Guillen y Catalina Bolancé – Universidad de Barcelona