Modelos Lineales Generalizados

María Durbán

Índice general

Capítulo 1

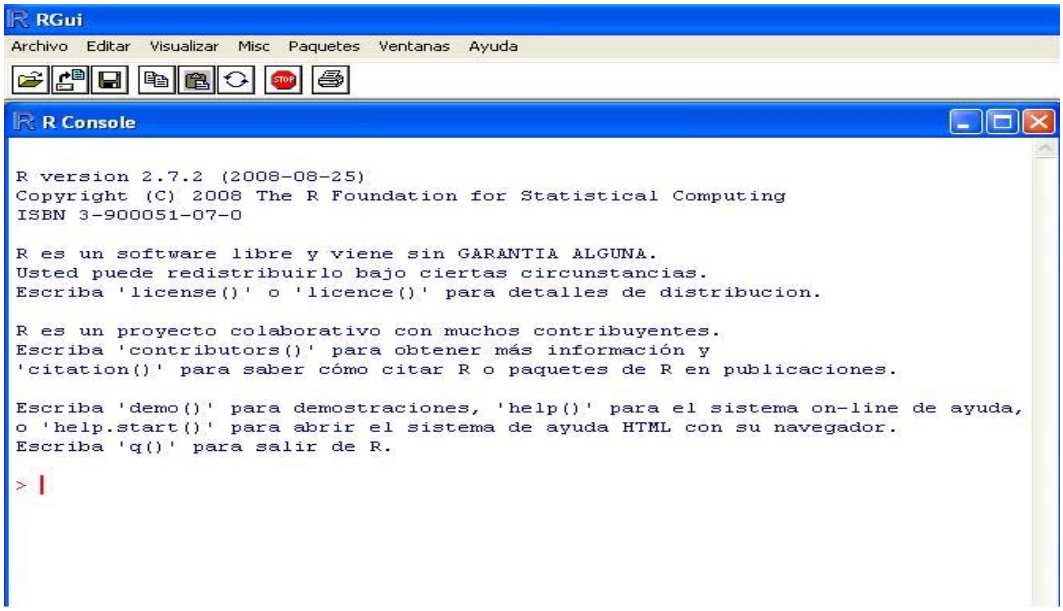
Introducción

1. Introducción a R

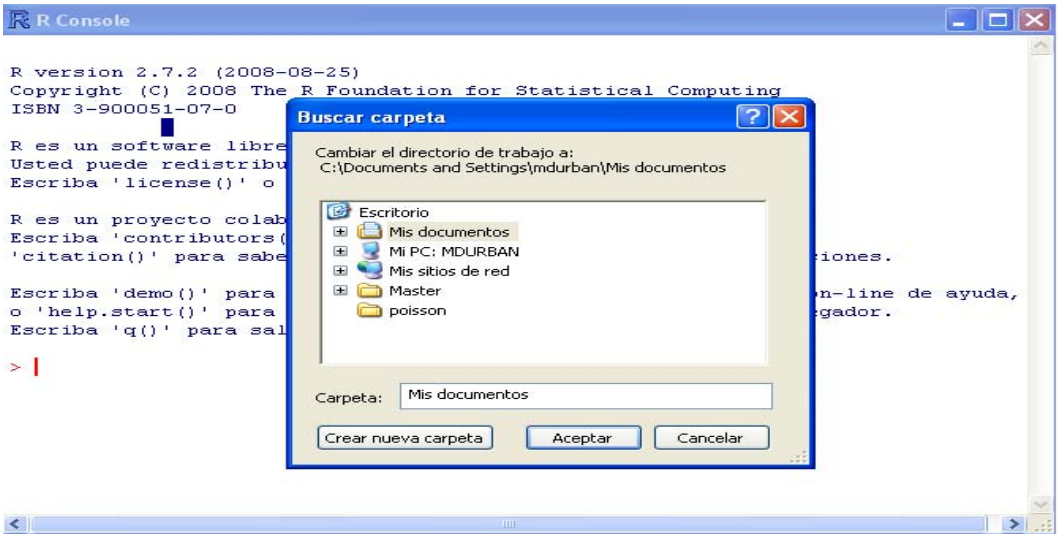
1.1. Instalalar y empezar con el programa

La versión actual de R es la 2.7.2. Para instalarla es necesario seguir los siguientes pasos:

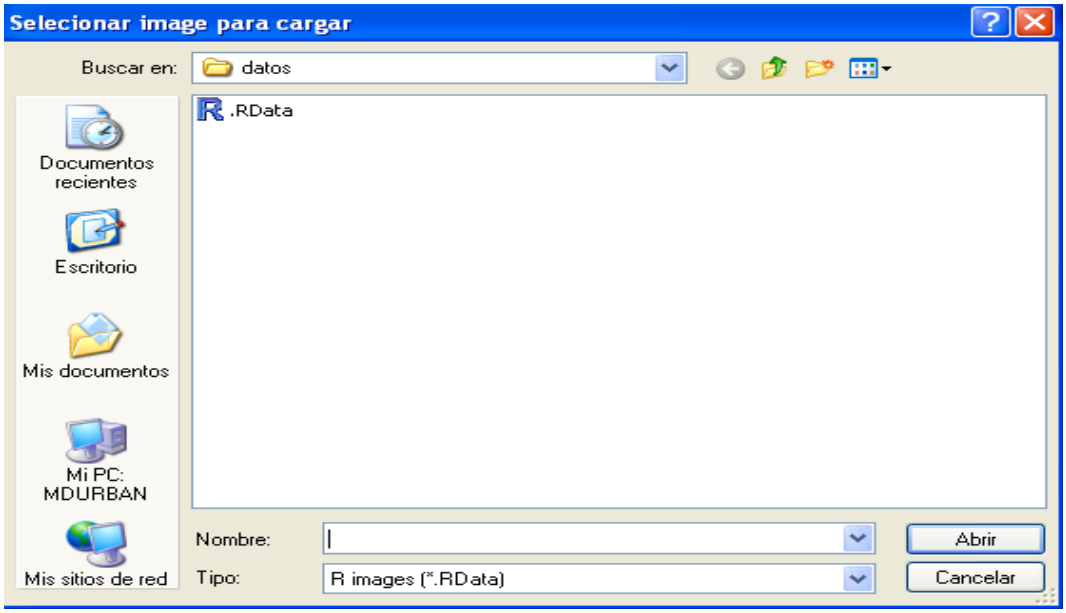
1. Ir a la página web `http://www.r-project.org`
2. Descargar el archivo `SetupR.exe` y una vez el el ordenador, pinchar en su icono para instalar el programa
3. Para empezar el programa sólo hay que pinchar en el icono, al hacerlo apartecerá la siguiente ventana



Cada vez que empezamos el programa hemos de decidir en qué directorio queremos trabajar e indicárselo a R. Para ello pinchamos en `Archivo` $\Longrightarrow$ `Cambiar dir` y aparecerá una ventana en la que podremos elegir la carpeta en la que queremos trabajar. Si ésta es la primera vez que utilizamos esta carpera no hemos de hacer nada más.



Cada vez que utilizamos un nuevo directorio, `R` crea un fichero llamado `.RData` donde se guardarán todas las variables, modelos, etc., que creemos en cada sesión, de modo que si hemos utilizados ya `R` en esa misma carpeta, hemos de indicarle al programa que queremos utilizar todos los objetos que habíamos creado en la sesión anterior, para ello pinchamos en `File` $\Longrightarrow$ `Cargar área de trabajo` y aparecerá una ventana en la que elegiremos el fichero `.RData`.



Algunas reglas en `R`

- Para nombrar a una variable se pueden usar letras, números, puntos y guión bajo. No puede empezar con mayúsculas

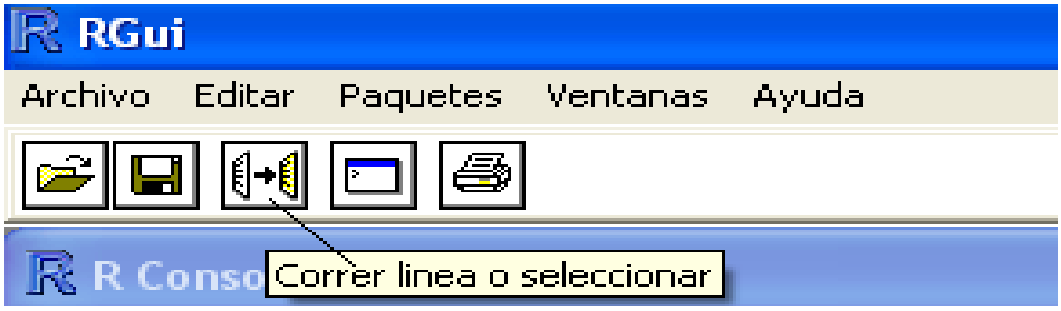
- `R` distingue mayúsculas y minúsculas
- `R` sobreescribirá las variables/objetos que tengan el mismo nombre
- Hay nombre que están reservados y no se pueden utilizar, y otros que aunque no lo están es mejor no utilizar (sobre todo aquellos que corresponden a funciones)

Ficheros .R

Un fichero .R contiene una secuencia de comandos que se pueden ejecutar de una vez, o sólo una parte de ellos, marcándolos con el rantón. Es una buena práctica el utilizar este tipo de ficheros ya que sirve de registro de las transformaciones y los pasos realizados y permite comentar dichos pasos. Para ejecutar un fichero .R hemos de abrirlo primero, para ello pinchamos `File` $\Longrightarrow$ `Abrir script`, esto abrirá el directorio en el que estemos trabajando y mostratá los ficheros .R que haya allí. Una vez abierto, tenemos dos opciones, ejecutar todos los comandos del fichero o marcar los que nos interesen y ejecutar sólo esos. Para esto pinchamos en el tercer icono que hay en la parte superior o:

- `Editar` $\Longrightarrow$ `Ejecutar todo`
- Marcamos y `Edit` $\Longrightarrow$ `Correr línea o selecionar`

IMPORTANTE: Siempre salvar el fichero antes de cerrarlo, si no quereis salvar las modificaciones, cerrar sin salvar.



En un fichero .R podemos añadir comentarios, para eso empezaremos la línea donde con el símbolo %

1.2. Librerías

Cuando se instala `R` se instalan automáticamente una serie de librerías o paquetes. Las librerías suelen ser bastante especializadas y permiten hacer gran variedad de cosas y todo tipo de análisis. Para ver las librerías que hay instaladas escribimos:

```
library()
```

para utilizar alguna de las que hay instaladas tenemos dos opcionesqueremos utilizar una librería llamada MASS:

1. Escribimos

```
library(MASS)
```

2. En el panel superior pinchamos en `Paquete` $\Longrightarrow$ `Cargar paquete` y seleccionamos MASS.

Hay cientos de librerías escritas por los usuarios de `R`, algunas de ellas están dedicadas a análisis estadísticos especializados (por ejemplo, mlogit es una librería que utilizaremos para datos multinomiales) y otras dan acceso a datos. Para poder instalarlas es necesario disponer de conexión a Internet.

2. Importar y editar datos

Los datos suelen leerse desde archivos. Normalmente se editarán los datos con un procesador de textos para ajustarlos a las necesidades de `R`. Hay dos funciones para leer datos: `read.table()` y `scan()`, aquí nos centraremos en la primera.

2.1. La función `read.table()`

Es el método más recomendable si los datos se van a almacenar el hojas de datos o `data.frame` (que es lo más habitual y que veremos más adelante. Para poder leer un fichero de datos utilizando esta función, la primera línea del archivo ha de contener el nombre de las variables. Lo mejor es tenet los datos como un fichero de texto `.txt` en el que las columnas están separadas por espacios. Esta función permite especificar varias opciones, las más importantes son:

- `file=` debe ir acompañada del nombre del fichero entre comillas, si le fichero de datos no está en el directorio en el que estamos trabajando hemos de darle la ubicación exacta del mismo
- `header=TRUE/FALSE`, en general escribiremosentenderá que en el fichero de datos, los decimales están separados por un `.`, si estubiera separados por una coma, escribiríamos `dec=’’,’’`

Por ejemplo:

```
> salud=read.table(‘‘salud.txt’’,header=TRUE)
```

El objeto `salud` es una matrix que tiene tantas columnas como variables y tantas filas como datos. Esta función almacena los valores numéricos tal y como aparecen y los caracteres como variables categóricas (que en `R` se llama factor).

3. Tipos de datos y su manipulación

Un objeto (ya sea una variable, el resultados de un análisis, etc.) puede ser creado mediante el operador `‘‘=’’`, por ejemplo:

```
> n=10
> n
[1] 10
```

como digimos antes, si el objeto ya existe, su valor anterior es borrado y sustituido por el nuevo. El valor asignado puede ser el resultado de una operación matemática o de una función

```
> n=3*4
> n
[1] 12
> x=sqrt(25)
> x
[1] 5
```

La función `ls()` lista todos los objetos que se han creado y que están en la memoria (y por tanto quedarán guardados en el archivo .RData)

```
> ls()
[1] "Escuelas" "n"        "x"
```

Los objetos se pueden borrar utilizando la función `rm()`, por ejemplo

```
> rm(n)
```

borra el objeto `n` de la memoria,

```
> n
Error: objeto "n" no encontrado
```

Los objetos con los que trabaja `R` tienen nombre y contenido, pero también tienen *atributos* que indican el tipo de datos representados por ese objeto. Por ejemplo, una variable que toma valores 1, 2 y 3, puede ser una variable que representa el número de hijos, puede ser una variable categórica con tres categorías. Toda variable tiene dos atributos básicos:

1. Longitud: Número de elementos que contiene
2. Tipo: Numérico, carácter o lógico

Para ver estos atributos utilizamos `length` y `mode`:

```
> x=2
> length(x)
[1] 1
> mode(x)
[1] "numeric"
> y="Maria"
> mode(y)
[1] "character"
```

Los datos que no están disponibles se representan por `NA`

3.1. Tipos de objetos en `R`

Hay varios tipos de objetos, nos vamos a centrar en los más utilizados.

Vectores

Un vector es una colección ordenada de objetos del mismo tipo (números, caracteres, etc.)

```
> x=c(1,2,3)
> y=seq(3,1) #esto creará una secuencia del 3 al 1
> y
[1] 3 2 1
```

`R` puede operar con vectores de una sola vez:

```
> y^2
[1] 9 4 1
> x+10
[1] 11 12 13
```

Entre las funciones más útiles se encuentran:

- `sort, order, max, min, mean, var, sd`
- `sum, prod, range`

Otras funciones importantes son los operadores comparativos:
$<$, $>$, $<=$, $>=$.

Podemos extraer elementos de un vector así:

```
> x=c(2,1,5,65,28,4,55)
> x[1:3]
[1] 2 1 5
> x[x>10]
[1] 65 28 55
> x[x>=55]
[1] 65 55
```

data.frame

Un objeto de este tipo es una colección de vectores (todos de la misma longitud) pero que pueden ser de distinto tipo. Cuando leemos datos utilizando `read.table` el objeto que creamos es un `data.frame`. Lo podemos comprobar:

```
> class(salud)
[1] "data.frame"
> names(salud)
 [1] "sexo"    "g01"     "g02"     "peso"    "altura"  "con_tab" "anio"
 [8] "educa"   "imc"     "bebedor" "edad"
```

El objeto `salud` corresponde a datos percepción de salud de personas en distintos distritos de Madrid. Contiene 11 variables:

- `sexo` Toma valores 1 si es hombre y 2 si es mujer
- `g01` En general, ¿cómo considera usted que es su salud?, es variable de interés con 5 categorías

    - Muy buena=1
    - Buena=2
    - Regular=3
    - Mala=4
    - Muy mala=5
- `g02` Una recodificación de la anterior con dos categorías: 1 = Buena salud, 0 = No buena salud
- `peso` en Kg
- `altura` en cm
- `con.tab` Consumo de tabaco, con 2 categorías:
    - No fumador o fumador ocasional=1
    - Fumador diario o ex-fumador=2
- `anio` año en el que se recogieron los datos
- `educa` nivel de estudios, con 4 categorías: Bajo, Medio-Bajo, Medio-Alto, Alto
- `imc` Índice de masa corporal
- `bebedor` con tres categorías: Poco/Nada, Ocasionalmente, Frecuentemente
- `edad` en años

Podemos tener un resumen si hacemos:

```
> summary(salud)
```

Todas estas variables están dentro del objeto `salud` pero `R` no las identifica por sí mismas:

```
> g01
Error: objeto "g01" no encontrado
```

Sin embargo, si hacemos:

```
> salud$g01
```

obtenemos los valores de la variable. El símbolo $ detrás del nombre del `data.frame` y seguido del nombre de una variable, extraerá la variable que queramos.

Todas las variables están identificadas como variables numéricas (esto es lo que `R` hace por defecto). Hay cuatro variables categóricas: `sexo, con_tab, educa y bebedor`:

```
> salud$sexo=factor(salud$sexo)
> salud$con_tab=factor(salud$con_tab)
> salud$educa=factor(salud$educa)
> salud$bebedor=factor(salud$bebedor)
```

si volvemos a hacer

```
> summary(salud)
```

vemos que ahora están identificadas como variables categóricas (los factores son otro tipo de objetos en R). Algunas funciones útiles para explorar las variables categóricas son: `table` y `tapply`:

```
> table(salud$bebedor)

   0    1    2
2977 4054  326

> tapply(salud$peso,salud$bebedor,mean)
       0        1        2
67.30971 70.86384 72.75767
```

Cuando estemos trabajando con un `data.frame` podemos decirle a R que identifique automáticamente las variables que hay en él mediante el comando `attach`

```
> attach(salud)
g01
```

cuando acabemos de trabajar con el haremos:

```
> detach(``salud'')
```

4. Gráficos

R tiene muchas funciones para crear gráficos y da gran libertad al usuario para personalizar el los gráficos. Aquí vamos a ver sólo las funciones básicas, pero las posibilidades son casi infinitas.

4.1. La función `plot()`

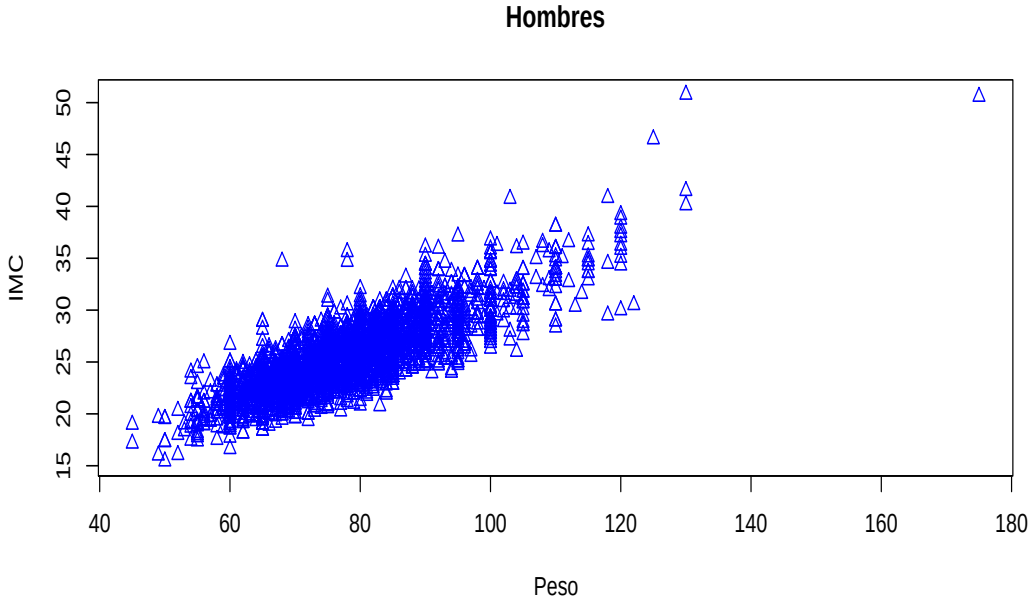
Esta es la función básica para dibujar. Tiene los siguientes argumentos:

- `x` e `y` que son las variables que van en el eje x e y
- `type=` Por defecto, la función dibuja puntos pero las opciones pueden ser:
  - "l" para líneas
  - "b" para líneas y puntos a la vez
  - "n" para no dibujar nada
- `col=` para dibujar con color en vez de en negro, los colores se especifican con números
- `main=` para poner un título en el gráfico
- `xlim, ylim` para cambiar los límites del gráfico
- `xlab, ylab` para cambiar las etiquetas del eje x e y
- `pch=` para cambiar el símbolo que dibujamos, se especifica con un número
- `lty=` para cambiar el tipo de líne (en caso de elegir líneas en vez de puntos)

- `lwd=` Para aumentar o disminuir la anchura de la línea

Por ejemplo:

```
x=peso[sexo==1]
y=imc[sexo==1]
plot(x,y,col=4,xlab="Altura", ylab="IMC",main="Hombres",pch=2)
```



4.2. Añadir elementos a un gráfico

Una vez que hemos hecho un gráfico es posible añadir otros elementos:

- `points()` permite añadir puntos
- `lines()` añade líneas
- `text()` añade texto

Si queremos añadir una línea que represente donde está el imc medio de los hombres:

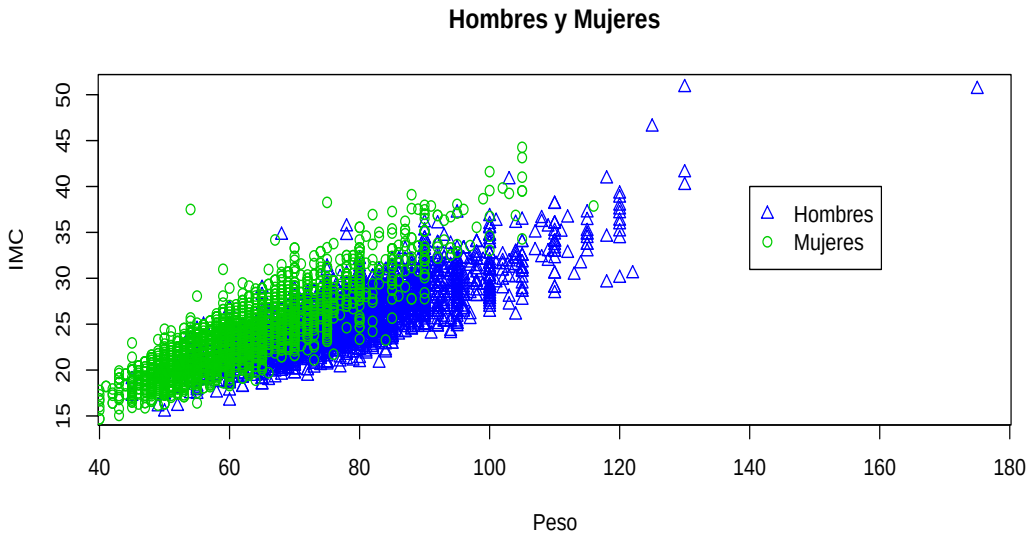
```
> mean(y)
[1] 25.4505
> length(x)
[1] 3666
> m=rep(25.4505,3666)
> lines(x,m,lwd=2,col=6)
> text(130,28,"IMC medio")
```

Otra manera de añadir el texto es utilizar `legend()`

```
legend(140,40,col=c(4,6),pch=c(2,-1),lty=c(-1,1),lwd=c(-1,2),c("Hombres","IMC
Medio"))
```

Otro ejemplo:

```
> xx=peso[sexo==2]
> yy=imc[sexo==2]
>plot(x,y,col=4,xlab="Peso", ylab="IMC",main="Hombres y Mujeres",pch=2)
> points(xx,yy,col=3)
> legend(140,40,col=c(4,3),pch=c(2,1),c("Hombres","Mujeres"))
```



4.3. Múltiples gráficos por página

En algunas ocasiones será conveniente tener más de un gráfico en la misma página, esto se puede conseguir con la función `par()`, esta función tiene muchos argumentos que permiten cambiar el aspecto de la ventana de gráficos, nossotros nos vamos a centrar solamente en el argumento `mfrow=` que es el que permite dibujar varios gráficos a la ver. Por ejemplo:

- `par(mfrow=c(1,2))` permite dibujar dos gráficos uno junto al otro
- `par(mfrow=c(2,1))` permite dibujar dos gráficos uno debajo del otro
- `par(mfrow=c(2,2))` permite dibujar 4 gráficos en dos filas y dos columnas

Por ejemplo

```
> par(mfrow=c(1,2))
> plot(x,y,col=4,pch=2,main="Hombres")
> plot(xx,yy,col=3,main="Mujeres")
> par(mfrow=c(2,1))
> plot(x,y,col=4,pch=2,main="Hombres")
> plot(xx,yy,col=3,main="Mujeres")
```

Los gráfico pueden parecer un poco deformados, podemos arreglarlo con otro argumento de la función `par()` que hace que los gráficos sean cuadrados:

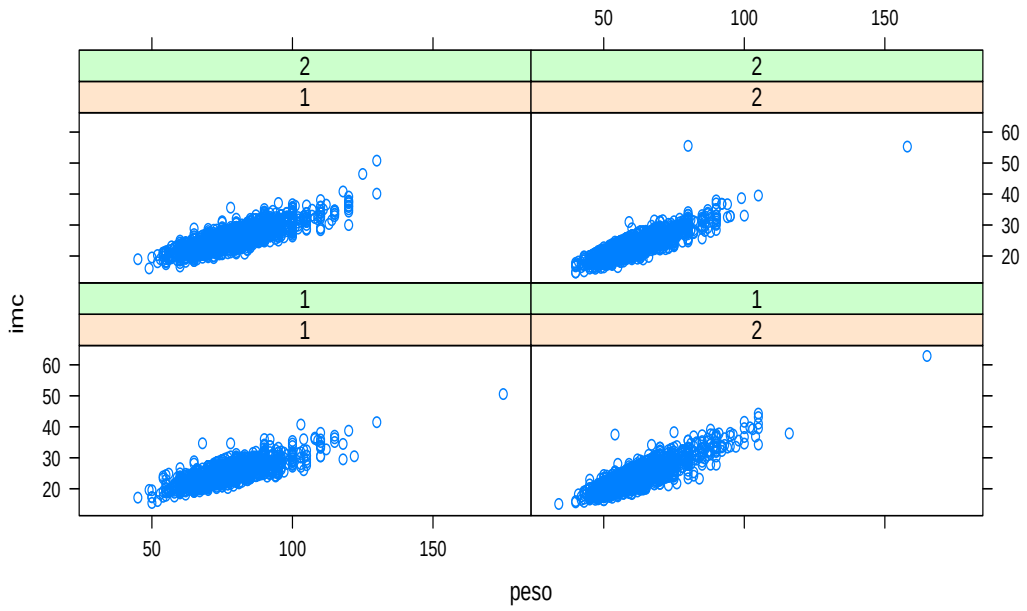
```
> par(mfrow=c(1,2),pty="s")
> plot(x,y,col=4,pch=2,main="Hombres")
> plot(xx,yy,col=3,main="Mujeres")
```

La función `xyplot`

Una de las librarías de `R` que viene instalada por defecto es `lattice`, esta contiene la función `xyplot()`, que permite hacer gráficos de una variable continua respecto a otra, pero condicionadas por los valores de una o más variables categóricas.

Por ejemplo si estamos interesados en dibujar el peso con respecto al imc, pero diferenciando por sexo:

```
> library(lattice)
> xyplot(imc~peso|sexo)
#o por sexo y tipo de fumador
> xyplot(imc~peso|sexo+con_tab)
```



4.4. Otros gráficos en `R`

Otros gráficos útiles en `R` son:

Histograma

```
hist(imc,nclass=6)
```

Produce un histograma (sólo tiene sentido para variables numéricas). La función `hist()` nos permite elegir el número de barras del histograma, si añadimos la opción `probability=TRUE` se representarán las frecuencias relativas en vez de las absolutas

```
hist(imc,nclass=6, probability=TRUE)
```

La función `identify()`

En ocasiones es interesante poder identificar puntos en un gráfico, por ejemplo, podemos estar interesados en identificar el tipo de fumador entre los hombres con mayor imc, para eso hacemos

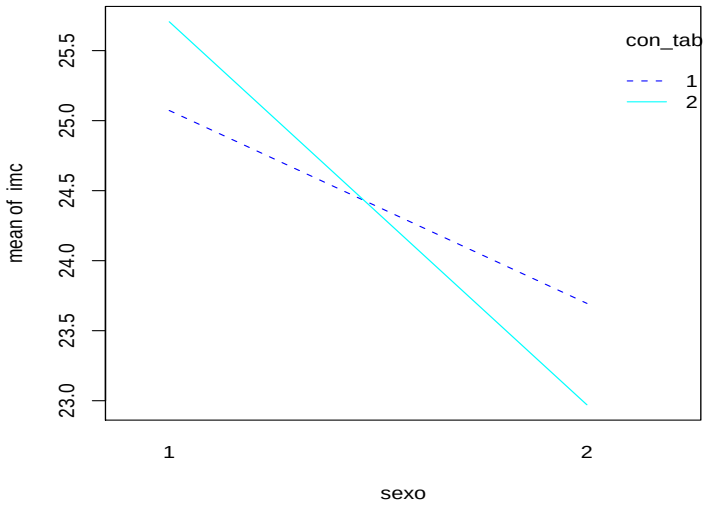
```
b=con_tab[sexo==1]
plot(x,y)
identify(x,y,b)
```

y pinchamos con el botón derecho del ratón en el punto que nos interesa. Aparecerá un 1 ó un 2 (ya que esa es la codificación para el tipo de fumador)

Gráficos de medias

En ocasiones, cuando tenemos varias variables explicativas categóricas es informativo hacer un gráfico del valor medio de la variable respuesta para cada categoría de las variables categóricas. Esto lo hacemos mediante la función `interaction.plot`:

```
interaction.plot(sexo, con_tab, imc,col=4:5)
```



¿Qué nos está indicando este gráfico?

Hay otras muchas funciones en `R` para hacer gráficos, algunas las iremos viendo a lo largo del curso.

5. Modelos estadísticos en `R`

`R` ofrece muchísimas posibilidades a la hora de analizar datos, aquí nos vamos a centrar en aquellos modelos que son relevantes para este curso: Modelos lineales.

5.1. La fórmula

La sintaxis de los modelos es en general común para todos ellos y depende de una fórmula. En un modelo, sea del tipo que sea siempre tenemos una variable respuesta (`y`), y una o más variables explicativas (continuas y/o categóricas $X_1, X_2, ...$, estos son algunos ejemplos de posibles fórmulas:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

```
y~X1+X2
```

$y = \beta_1 X_1 + \beta_2 X_2$ (sin la ordenada en el origen)

```
y~X1+X2-1
```

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

```
y~X1+I(X1^2)
```

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ (donde X_2 es una varaible categórica con dos niveles)

```
y~X1+X2+X1:X2
```

Cuando estamos trabajando con variables categóricas, hemos de especificar que queremos que el parámetro que corresponde al nivel más bajo sea 0, es decir que queremos comparar con el nivel más baja, en `R` lo hacemos de la siguiente forma:

```
options(contrasts=c("contr.treatment", "contr.poly"))
```

5.2. Modelos lineales

Cuando la variable respuesta es continua y queremos ajustar un modelo de regresión utilizamos la función `lm()` (linear model). El resultado de utilizar esta función es un objeto que guarda toda la información necesaria: valores ajustados, parámetros, errores estándar, contrastes, intervalos de confianza, resiuos, etc.
Supongamos que queremos ver la relación entre el peso y el índice de masa corporal:

```
modelo1=lm(imc~peso)
```

se ha creado el objeto `modelo1`.

¿Cómo obtenemos información a partir de este objeto?. La tabla que se muestra a continuación resume los métodos para obtener información:

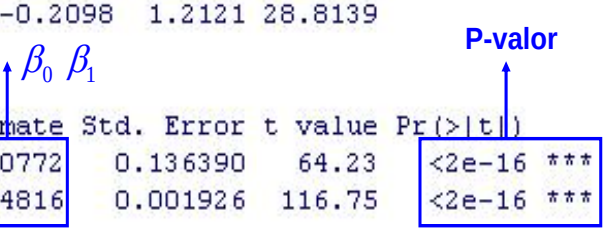
`print()`	resumen corto
`summary`	resumen completo
`coef()`	coeficientes
`confint()`	intervalos de confianza para los parámetros
`fitted.values()`	valores ajustados por el modelo
`residuals()`	residuos
`deviance()`	devianza
`logLik()`	Logaritmo de la verosimilitud y número de parámetros

El resultado de

```
summary(modelo1)
```

¿Cuáles son los intervalos de confianza para los parámetros?

```
Call:
lm(formula = imc ~ peso)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6705 -1.4898 -0.2098  1.2121 28.8139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.760772   0.136390   64.23   <2e-16 ***
peso        0.224816   0.001926  116.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.249 on 7355 degrees of freedom
Multiple R-squared: 0.6495,     Adjusted R-squared: 0.6495
F-statistic: 1.363e+04 on 1 and 7355 DF,  p-value: < 2.2e-16
```



Cómo dibujar los resultados

Para dibujar recta ajustada utilizamos una función llamada `abline`:

```
plot(peso,imc,xlab="Peso",ylab="IMC")
abline(modelo1,lwd=3,col=2)
```

Si hacemos simplemente

```
plot(modelo1)
```

lo que obtenemos son gráficos que analizan los residuos y que nos ayudan a ver si los datos cumplen las hipótesis en las que se basa el modelo: normalidad, varianza constante, etc.

Modelos con variables categóricas e interacción

Si ajustamos el modelo:

```
modelo2=lm(imc3~peso + sexo)
summary(modelo2)
```

¿Cómo interpretarías que el parámetro de `sexo` sea 2.626?

Ahora ajustamos el modelo con interacción:

```
modelo3=lm(imc~peso+sexo+peso:sexo)
summary(modelo3)
```

Ejercicio

Escribe el modelo con los coeficientes. ¿Cómo interpretamos ahora los parámetros?

Ahora lo que estamos ajustando son dos rectas, una para cada clase social. Par hacer el dibujo:

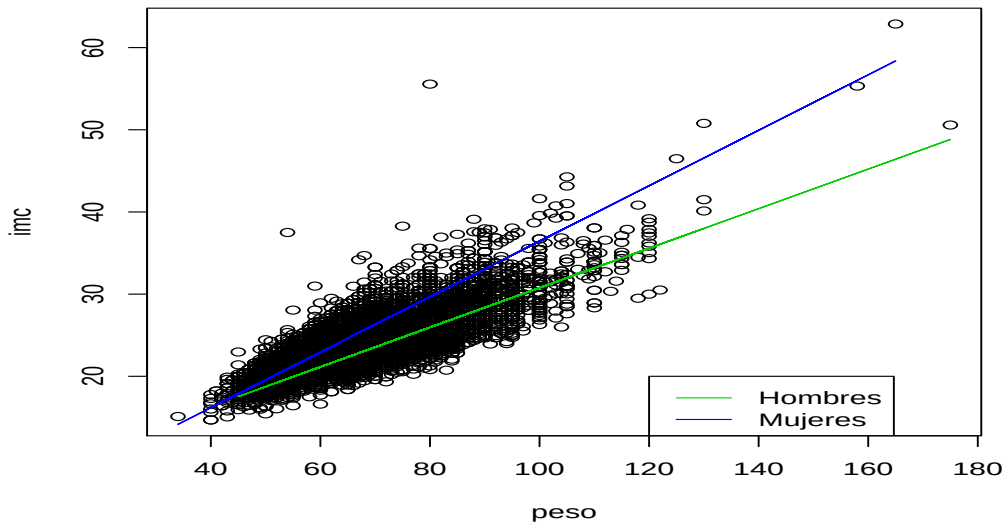
```
y.ajustados=fitted.values(modelo3)
plot(peso,y.ajustados)
```

¿qué ocurre si queremos dibujar líneas?

```
plot(peso,y.ajustados,type="l")
```

El gráfico no tiene sentido ya que los valores de `peso` no están ordenados y `R` une puntos consecutivos. Para solucionarlo vamos a dibujar las rectas por separado:

```
 peso.1=peso[sexo==1]
 peso.2=peso[sexo==2]
 y.ajustados.1=modelo3.coef[1]+modelo3.coef[2]*peso.1
 y.ajustados.2=modelo3.coef[1]+modelo3.coef[2]*peso.2+modelo3.coef[3]+
                modelo3.coef[4]*peso.2
 plot(peso,imc)
 lines(peso.1,y.ajustados.1,col=3)
 lines(peso.2,y.ajustados.2,col=4)
legend(120,20,col=c(3,4),lty=c(1,1),c("Hombres","Mujeres"))
```



Seleción de variables

Hay varios métodos para la seleción de variables en un modelo de regresión lineal con variable respuesta continua. Mediante la comparación de la devianza (en este caso mediante la comparación de la varianza residual), que en este caso corresponde al uso del test F, y mediante el estadístico AIC (Akaike information criteria).
Para mostrar como hacer selcción de variables, vamos a ajustar el siguiente:

```
modelo4=lm(imc~peso+sexo+sexo:peso+bebedor+bebedor:peso+sexo:bebedor)
anova(modelo4)

Analysis of Variance Table

Response: mat.3
                    Df  Sum Sq Mean Sq  F value  Pr(>F)
mat.1                1 14889.4 14889.4 642.1642 < 2e-16 ***
sexo                 1    23.3    23.3   1.0058 0.31624
clase.social         1    71.9    71.9   3.0994 0.07875 .
mat.1:sexo           1     1.8     1.8   0.0768 0.78175
mat.1:clase.social   1    32.0    32.0   1.3807 0.24037
sexo:clase.social    1    99.1    99.1   4.2750 0.03903 *
Residuals          721 16717.3    23.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A la vista de estos resultados eliminamos la interacción entre las dos variables categóricas:

```
modelo5=lm(imc~peso+sexo+sexo:peso+bebedor+bebedor:peso)
anova(modelo5)
```

Comapramos los dos modelos, para ello:

```
anova(modelo5,modelo4)
Analysis of Variance Table

Model 1: imc ~ peso + sexo + sexo:peso + bebedor + bebedor:peso
Model 2: imc ~ peso + sexo + sexo:peso + bebedor + bebedor:peso + sexo:bebedor
  Res.Df     RSS   Df Sum of Sq      F Pr(>F)
1   7349 27200.3
2   7347 27183.4    2      16.9 2.2802 0.1023
```

Es importante advertir que el primer modelo que hay que poner es el más sencillo (y por tanto la hipótesis nula). En este caso el p-valor es $> 0{,}05$ por lo tanto no rechazamos la hipótesis nula, es decir que nos quedamos con el modelo más sencillo. ¿Cuál es el modelo final?

Diagnosis del modelo: Análisis de residuos

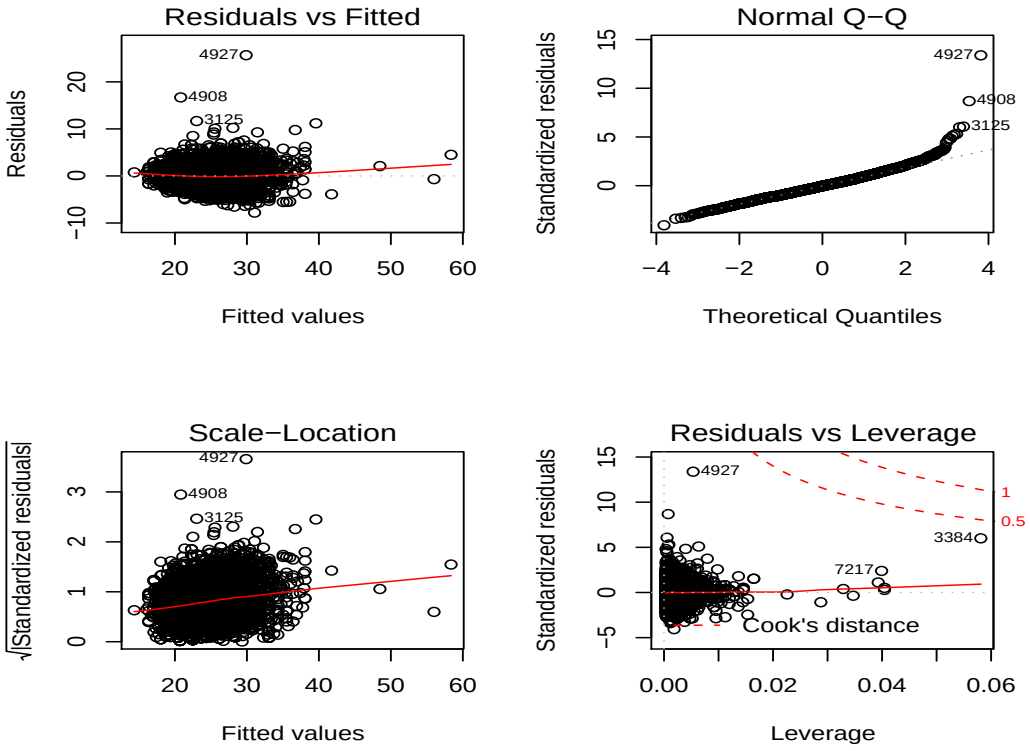
Hemos de verificar que se cumplen las hipótesis del modelo, ya que los contrastes de hipótesis se basan en ellas. En el caso de modelos de regresión lineal con variables continua, hay dos condiciones básicas que se han de satisfacer:

- Los datos vienen de una distribución normal
- La varianza del lod errores es constante

Para comprobar que estas condiciones se cumplen utilizamos los residuos del modelo (los residuos son los valores observados menos los ajustados), y hacemos dos gráficos; un gráfico de los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y el gráfico cuantil-cuantil que comparará los residuos del modelo con los valores de una variable normal. En `R` esto

es sencillo de hacer, si tenemos un objeto que sea un modelo lineal (como es el caso de `modelo5`), podemos utilizar la función `plot()`.

```
par(mfrow=c(2,2))
plot(modelo5)
```



El gráfico de la parte superior derecha indica los datos no corresponde a los de una variable Normal.

6. Tipos de datos

Todos los modelos que vamos a considerar en este curso intentan explicar la relación entre una **variable dependiente/variable respuesta** y un conjunto de **variables independiente/variables explicativas/predictores**. Estas variables están medidas en diferentes escalas y la metodología que hemos de utilizar depende de con qué tipo de variables estemos trabajando.

6.1. Clasificación de variables

Las variables con las que trabajamos pueden ser de dos tipos: **cuantitativas** o **cualitativas**.

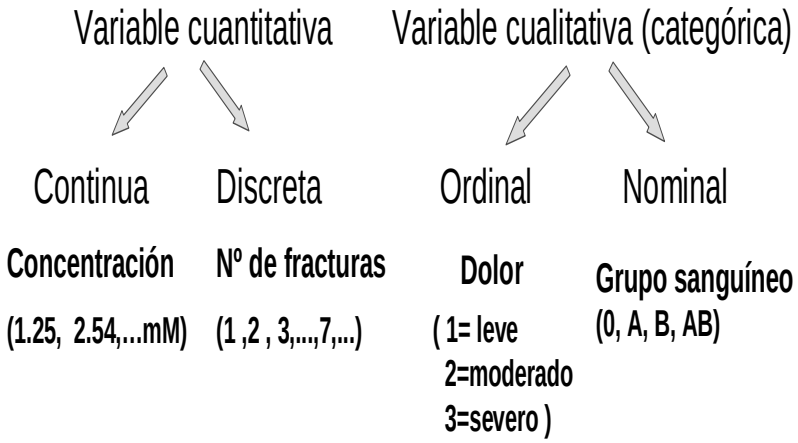
1. **Cuantitativas**: Miden una magnitud, por ejemplo, peso, altura, número de hijos, etc. Estas variables a su vez se clasifican en:

   - **Continuas**: Cuando puede adoptar cualquier valor numérico dentro de un intervalo, por ejemplo, edad, tensión arterial, etc.

- **Discretas**: Cuando sólo puede tomar ciertos valores enteros, por ejemplo, número de fallecidos en un año, número de hijos, etc.

2. **Cualitativas/Categóricas**: Expresan una graduación o una cualidad, pueden clasificarse en:

   - **Nominales**: Clasifican de forma no ordenada
     - Dicotómicas, binarias o binomiales: cuando presentan dos categorías o niveles, por ejemplo, respuestas sí/no
     - Politómicas o multinomiales: cuando presenta más de dos categorías, por ejemplo, color de los ojos
   - **Ordina** dad del dolor.



Una variable explicativa cualitativa se llama **factor** y a cada una de sus categorías se le llama **nivel**. Una variable explicativa cuantitativa se llama **covariable**. La naturaleza de estas variables va a determinar los métodos estadísticos que se han de utilizar para su análisis. En la siguiente tabla se muestran los principales métodos estadísticos para distintas combinaciones de variables respuesta y predictoras.

Variable Respuesta	Variables Predictoras	Métodos
Continua	Binaria	t-test
	Nominal,>2 categorías	Análisis de la Varianza
	Ordinal	Análisis de la Varianza
	Cualitativas y Continuas	Regresión Múltiple
Nominal/Binaria	Cualitativas	Tablas de Conting./Logística/Otros glm
	Continuas	Logística/Otros glm
	Categóricas y Continuas	Logística/Otros glm
Nominal >2 Categorías	Nominal	Tablas de Conting.
	Categóricas y Continuas	Logística Multinomial
Ordinal	Categóricas y Continuas	Logística Ordinal
Discretas	Categóricas	Modelos log-lineales
	Categóricas y Continuas	Regresión de Poisson

7. ¿Qué son y por qué hay que utilizar GLMs?

El objetivo de cualquier análisis es responder a la pregunta: ¿puede ser la variable de interés predicha por un conjunto de variables explicativas?, esta es la misma pregunta que nos hacemos cuando utilizamos un modelo de regresión lineal, entonces, ¿por qué es necesario utilizar otro tipo de modelos?. La razón fundamental es que para poder utilizar regresión lineal es necesario que la variable respuesta sea continua, y cumpla las hipótesis estándar del modelo lineal (datos Normales, varianza constante, etc.) Si la variable de interés es, por ejemplo binaria e ignoramos este hecho, lo que hacemos es ajustar este modelo:

$$p = Pr(\text{ocurra algo}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

si estimamos los parámetros utilizando el procedimiento `lm()` de R podríamos estar cometiendo dos graves errores:

1. Los valores predichos de la probabilidad podrían estar fuera del intervalo $(0, 1)$.
2. Los intervalos de confianza y los test para ver que variables son significativas están basados en la hipótesis de que los datos viene de una distribución Normal, cosa que no es cierta con datos binarios

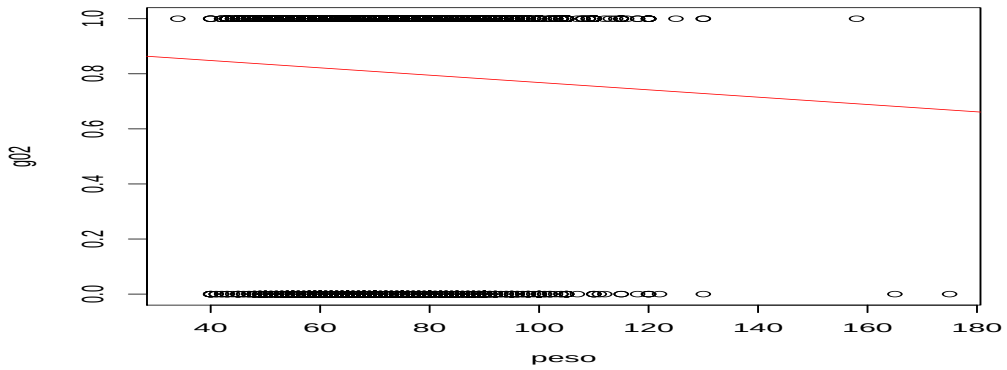
7.1. Ejemplos de motivación

Motivaremos la necesidad de los GLMs usando dos conjuntos de datos donde la distribución de la variable respuesta tiene la propiedad de que la respuesta media y la varianza están relacionadas.

Percepción de salud

Con los datos sobre percepción del estado de salud vamos a ajustar un modelo de regresión entre la variable dicotómica que representa el estado de salud y el peso:

```
ejemplo=lm(g02~peso)
y.ajustado=ejemplo$fitted
plot(peso,g02)
abline(ejemplo,col=2)
```



Vemos como no tiene sentido ajustar una recta, y además, a veces lugar a valores predichos menores que 0.

En 1944, Berkson utilizó por primera vez la regresión logística como una forma de solucionar el problema, ya que la función logit hace que en vez de estar trabajando con valores de la variable respuesta entre $(0, 1)$ estemos trabajando con una variable respuesta que puede tomar cualquier valor. No fue hasta 1972 cuando John Nelder introdujo los modelos lineales generalizados (GLM), de ahí que en general se considere la regresión logística como algo distinto a los GLM, cuando lo que ocurre es que tanto la regresión múltiple como la logística, de Poisson, ordinal, etc., son casos particulares de un GLM.

Experimento de turbinas

Se realiza un experimento para determinar la relación entre el tiepo de uso de unas turbinas y el número de fisuras que aparecen.

```
      horas   fisuras
 [1,]   400         0
 [2,]  1000       212
 [3,]  1400        66
 [4,]  1800       511
 [5,]  2200       150
 [6,]  2600       351
 [7,]  3000       378
 [8,]  3400        78
 [9,]  3800       748
[10,]  4200       840
[11,]  4600       756
```

Los datos siguen una distribución de Poisson, por lo tanto, la media es igual a la varianza y la restricción ahora es que los datos ajustados han de ser positivos. Si ajustamos `lm(fisuras ~ horas)`, obtenemos que el valor ajustado para la primera observación es negativo.

```
 lm(fissures~hours)$fitted
      1         2         3         4         5         6
-1.47929 101.17751 169.61538 238.05325 306.49112 374.92899
        7         8         9        10        11
443.36686 511.80473 580.24260 648.68047 717.11834
```

Para entender lo que es un GLM, volvamos al modelo de regresión múltiple, en este modelos suponemos que:

$$E[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

es decir que existe una relación lineal entre las X y $E[Y]$ (el valor medio de Y dado un cierto valor de las variables explicativas). Si las observaciones son binarias, entonces:

$$\begin{aligned} Pr(Y = 1) &= p \\ Pr(Y = 0) &= 1 - p \; E[Y] = 0 \times (Y = 0) + 1 \times (Y = 1) = p \end{aligned}$$

por lo tanto un modelo de regresión múltiple relaciona directamente la probabilidad de que ocurra un suceso con las variables explicativas y hemos visto cómo eso lleva a errores graves. Lo que hacen

los GLM es establecer esa relación lineal no entre la media de la variable respuesta y los predictores, sino entre una función de la media de variable respuesta y los predictores, es decir

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

según de qué tipo sea la varible Y, así será la función g. Es decir, un GLM tiene 3 componentes:

1. **Componente aleatorio**: La variable respuesta Y. Para poder utilizar un GLM, la distribución de Y ha de pertenecer a la **Familia Exponencial**, es decir, su función de densidad ha de poder escribirse como:

$$f(y; \boldsymbol{\theta}, \phi) = exp\left\{ \frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y, \phi) \right\} \tag{1.1}$$

   donde, en cada caso, $a(.)$, $b(.)$ y $c(.)$ serán funciones específicas. El parámetro $\boldsymbol{\theta}$ es lo que se llama *parámetro canónico de localización* y ϕ es un *parámetro de dispersión*. Pertenecen a la familia exponecial la distribución Normal, Bernouilli, Binomial, Poisson, Exponecial, Gamma, entre otras.

2. **Componente sistemático**: Las variables predictoras X_i $i = 1 \ldots k$

3. **Función *link***: La función que relaciona la media, $E[Y]$, con las variables predictoras X. En el caso del modelo de regresión ordinaria, $\boldsymbol{\mu} = \boldsymbol{\eta}$, por lo tanto la función link es la identidad. Hay muchas opciones par la función link. La función **link canonica** es una función que transforma la media en el parámetro canónico $\boldsymbol{\theta}$

$$g(E[Y]) = \boldsymbol{\theta} \Rightarrow g \quad \text{es una función link canónica}$$

   La siguiente tabla muestra las funciones link canónicas para las distribuciones más comunes usadas en los GLMs:

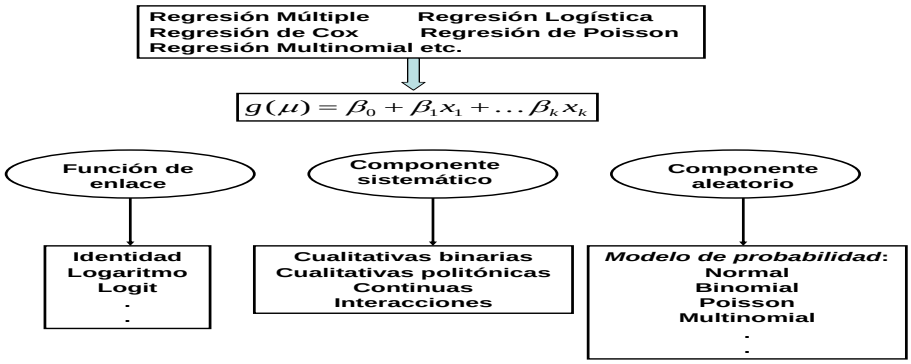
Distribución	Link canónica
Normal	$X\beta = E[Y]$ (identidad)
Binomial	$X\beta = \ln\left(\frac{P}{1-P}\right)$ (logistística)
Poisson	$X\beta = \ln(E[Y])$ (logarítmica)
Exponential	$X\beta = \frac{1}{E[Y]}$ (recíproca)
Gamma	$X\beta = \frac{1}{E[Y]}$ (recíproca)

Cuadro 1: Link canónicas usadas en los GLMs

La diferencia que hay entre usar la función link y usar una transformación, es que la función link transforma la media, $E[Y]$, y no los datos, Y.

Las etapas a la hora de utilizar GLMs (o cualquier otro modelo estadístico) son:

1. Especificación de modelos
2. Estimación de los parámetros
3. Selección de un modelo



4. Evaluación del modelo

5. Interpretación de modelo

Todos los modelos que vamos a ver en los capítulos siguientes son casos particulares de un GLM, por ejemplo, la regresión logística es un GLM en la que la variable respuesta es binaria (sigue una distribución de Bernoulli), y la función link es el logit ya que

$$g(E[y] = p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

La inferencia, selección de variables, contrastes de hipótesis, etc., en los GLM ha sido desarrollada a lo largo de los años y la metodología para llevarla a cabo está implementada en el software estadístico más común.

Capítulo 2

Regresión Logística

Este capítulo se centra en los distintos modelos para datos con respuesta binaria. Vamos a trabajar con los datos sobre percepción de salud y el objetivo va a ser el describir como cambia la percepción del estado de salud por sexo, edad y tipo de bebedor. Podemos definir

$$y_i = \begin{cases} 1 & \text{si buena salud} \\ 0 & \text{si no buena salud} \end{cases}$$

donde la variable toma el valor 1 con probabilidad p_i y el valor 0 con probabilidad $1 - p_i$. Esa variable aleatoria sigue una distribución de Bernoulli

$$Pr[Y_i = y_i] = p_i^{y_i}(1 - p_i)^{1-y_i} \quad y_i = 0, 1.$$

Supongamos ahora que los sujetos estudiados se pueden clasificar atendiendo a los factores de interés en k grupos. En nuestro ejemplo, los individuos pueden ser clasificados en 12 grupos atendiendo al sexo y al tipo de bebedor

```
ftable(list(g02,sexo,bebedor))
        x.3    0    1    2
x.1 x.2
0   1        223  335   36
    2        516  281   15
1   1        792 2090  190
    2       1446 1348   85
```

Sea n_i el número de observaciones en el grupo i, ahora y_i es el número de personas que consideran que tienen buena salud en el grupo i, por lo tanto ahora tenemos una variable binomial con parámetros n_i y p_i y

$$Pr[Y_i = y_i] = \binom{n_i}{r_i} p_i^{y_i}(1 - p_i)^{n-y_i} \quad y_i = 0, 1, \ldots, n_i.$$

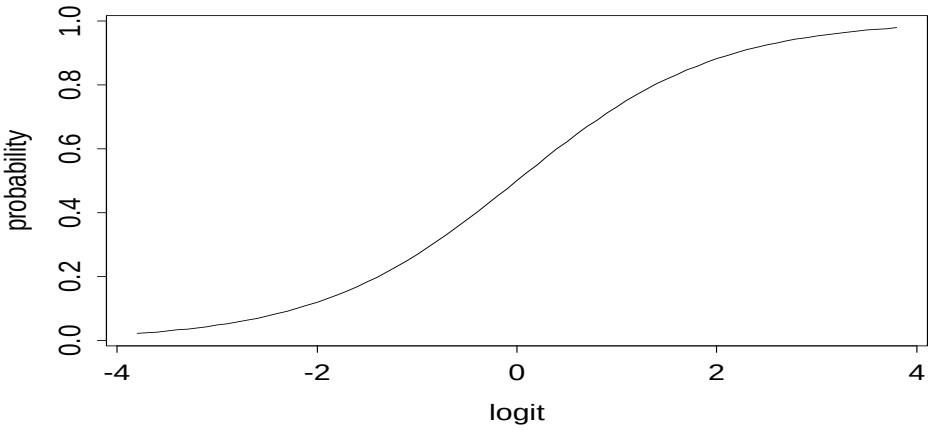
Desde el punto de vista matemático esta formulación para datos agrupados es la más general, ya que la que dimos en primer lugar es una caso particular de esta última con $k = n$ y $n_i = 1$. Desde el punto de vista práctico, hay que resaltar que si los predictores son categóricos y las observaciones son independientes, ambas maneras de analizar los datos son equivalentes. Una ventaja de trabajar con datos agrupados es que, dependiendo del tamaño de los grupos es posible hacer test sobre la bondad de ajuste del modelo.

Como ya comentamos en el capítulo anterior la regresión logística apareció en los años 50 y no fue hasta los 70 cuando aparecieron los GLM, por eso es bastante común pensar que son cosas diferentes, sin embargo, un modelo de regresión logística es un GLM como veremos más adelante.

Cuando trabajamos con probabilidades, el primer problema al que nos enfrentamos es que la probabilidad sólo toma valores entre 0 y 1, y si pretendemos relacionarla directamente con los predictores puede que nos salgamos del intervalo $[0, 1]$. Una solución sencilla es transformar la probabilidad para evitar este tipo de restricciones y relacionar esta probabilidad modificada con los predictores, una posible transformación es el *logit* or log-odds

$$\beta_0 + \beta_1 X_{1i} + \ldots + \beta_r X_{ri} = \eta_i = \ln \frac{p_i}{1 - pi}$$

esta transformación lo que consigue es que cuando la probabilidad se aproxima a 0 el logit lo hace a $-\infty$ y cuando la probabilidad se acerca a 1, el logit lo hace a $+\infty$. Si la probabilidad es 0.5, el odds es 1 y el logit 0. Logits negativos corresponden a probabilidades inferiores a 1/2 y viceversa.



La transformación logit es uno-a-uno, a la inversa se le suele llamar *antilogit* y permite calcular la probabilidad a partir del logit,

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Ahora estamos en disposición de definir el modelo de regresión logística al asumir que el logit de la probabilidad sigue un modelo lineal:

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_r X_{ri}$$

por lo tanto

$$p_i = \frac{e^{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_r X_{ri}}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_r X_{ri}}}$$

la probabilidad es una función no lineal de los predictores, por lo tanto es difícil expresar el efecto que tiene en la probabilidad un cambio en las variables predictoras, esta es una de las razones por las que se trabaja directamente con el logit.

1. El modelo de regresión logística como un GLM

Habíamos visto que un GLM tiene tres componentes:

1. Una distribución de probabilidad para la variable respuesta que pertenezca a la familia exponencial
2. Un predictor lineal $\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$
3. Una función link que relaciona a la media con el predictor lineal $g(\mu) = \eta$

El modelo de regresión logistica es un GLM donde la distribución de probabilidad es Bernoulli o Binomial, y la función link es el logit (ya que relaciona a la media, que en una Bernouilli es la probabilidad con el predictor lineal). Por lo tanto la estimación de los parámetros y los contrastes de hipótesis utilizan la teoría desarrollada para los GLMs.

Aunque los estimadores se van a calcular utilizando R, a continuación se da una breve descripción de los resultados en los que se basan.
Los parámetros se estiman maximizando la función de verosimilitud que representa la verosimilitud de que los datos observados sean una muestra de una variable con una determinada distribución, con lo cual lo que estamos haciendo es calcular qué valores de los parámetros hacen más verosímiles nuestros datos.

$$log(L(\beta)) = \sum (y_i \ln(p_i) + (n_i - y_i) \ln(1 - p_i))$$

escribimos p_i en función de β y maximizamos esa función, para ello hay que resolver una serie de ecuaciones que no tienen solución analítica y se utiliza un método iterativo basado en el algoritmo de *Newton-Raphson*.

1.1. Procedimientos para ajustar el modelo en R

En `R` la función que se utiliza para este tipo de modelos es la función `glm()`. Los argumentos de la función son:

- fórmula: similar a la que describimos anteriormente
- familia: binomial, poisson, etc.
- link: establece la relación entre la variable respuesta y los predictores: logit,probit,log, etc.

Vamos a describir la relación entre la percepción de salud, el sexo y el tipo de bebedor utilizando esta función:

```
logistica1=glm(g02~sexo+bebedor,family=binomial(link=logit))
summary(logistica1)
```

Esto da lugar al siguiente output:

```
Call:
glm(formula = g02 ~ sexo + bebedor, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9851   0.5480   0.5629   0.7040   0.7814
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.26858    0.06027  21.049  < 2e-16 ***
sexo2       -0.23868    0.06230  -3.831 0.000128 ***
bebedor1     0.55151    0.06288   8.771  < 2e-16 ***
bebedor2     0.49377    0.15984   3.089 0.002007 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7177.9  on 7356  degrees of freedom
Residual deviance: 7059.2  on 7353  degrees of freedom
AIC: 7067.2

Number of Fisher Scoring iterations: 4
```

Habíamos comentado que los parámetros del modelo (los β) se estiman maximizando la función de verosimilitud, las iteraciones que aparecen son debidas a que es necesario resolver las ecuaciones de forma iterativa, en este caso han sido necesarias 4 iteraciones para estimar los parámetros. El valor del *Null-deviance=-2× log-likelihood* en la primera iteración corresponde a lo que se llama el *null-model* un modelo en el que todos los parámetros son cero. Si el algoritmo no converge, los estimadores que aparecen no se deben utilizar. El valor final, `Residual deviance=7059.2` no tiene significado en sí mismo, pero se utilizará para comparar modelos anidados.

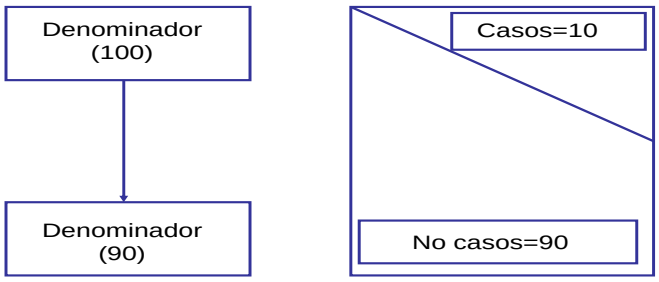
A continuación aparecen los valores estimados de los coeficientes del modelo:

$$log\left(\frac{p}{1-p}\right) = 1{,}27 - 0{,}238 * sexo + 0{,}55 * bebed\_1 + 0{,}49 * bebe\_2$$

También aparecen los errores estándar asociados con los coeficientes, y son utilizados para contrastar si el parámetro es significativamente distinto de cero y para calcular los intervalos de confianza. Los intervalos son muy útiles ya que nos orientan sobre los posibles valores que puede tomar el verdadero valor del parámetro, además, cuanto más ancho sea el intervalo, más ineficiente es la estimación del parámetro. Para calcular los intervalos de confianza utilizamos

```
> confint(logistica1)

                2.5 %     97.5 %
(Intercept)  1.1512601  1.3875492
sexo2       -0.3609674 -0.1167008
bebedor1     0.4283962  0.6749056
bebedor2     0.1891390  0.8169365
```

Para poder interpretar los parámetros, primero hemos de definir **Odds**:
Es la razón entre la probabilidad de experimentar un evento en relación con la probabilidad de no experimentarlo. La diferencia entre el concepto de tasa de incidencia y odds se muestra en la siguiente figura:



Tasa de Incidencia: 10/100=0.1(10 %)
Odds: 10/100/90/100=10/90=0.11(11 %)

Ahora el modelo está expresado en términos del log-odds, para obtenerlo en términos de los odds de la siguiente forma:

```
exp(coefficients(logistica1))
(Intercept)       sexo2    bebedor1    bebedor2
   3.555815    0.787666    1.735871    1.638488

exp(confint(logistica1))

                2.5 %    97.5 %
(Intercept) 3.1621750 4.0050226
sexo2       0.6970017 0.8898514
bebedor1    1.5347941 1.9638475
bebedor2    1.2082089 2.2635549
```

Si los datos los estuvieran agrupados como aparecen en la tabla aparece al principio del capítulo, los comandos en serían:

```
 salud sex drink count n
 1 1 0 792 1015
 0 1 0 223 1015
 1 2 0 1446 1962
 0 2 0 516 1962
 1 1 1 2090 2425
 0 1 1 335 2425
 1 2 1 1348 1629
 0 2 1 281 1629
 1 1 2 190 226
 0 1 2 36 226
 1 2 2 85  100
 0 2 2 15 100

salud2=read.table("salud2.txt",header=TRUE)
salud2$sex=factor(salud2$sex)
salud2$drink=factor(salud2$drink)
```

```
salud2=salud2[salud2$health==1,]
attach(salud2)

glm(cbind(count,n-count)~sex+drink,family=binomial(link=logit))
```

2. Interpretación de los parámetros

Los coeficientes estimados para cada variable independiente, representan la tasa de cambio de una función de la variable dependiente (en este caso el logit) por unidad de cambio de la variable independiente. Por lo tanto, a la hora de interpretar los parámetros hay que tener en cuenta dos cosas: la relación funcional entre la variable dependiente y las independientes y la unidad de cambio para las variables independientes. Además la interpretación de los coeficientes dependerá también de qué tipo de variables independientes tengamos: dicotómicas, politómicas o continuas.

2.1. Variable independiente dicotómica

Comenzamos con este caso ya que es la base para los demás tipos de variables independientes. Asumimos que la variable independiente está codificada como un 0 o un 1, como el modelo es:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X,$$

la diferencia en el logit para un individuo con $X = 0$ y $X = 1$ es β_1. Si en esta ecuación despejamos p obtenemos:

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad 1-p = \frac{1}{1+e^{\beta_0+\beta_1 X}}$$

En la siguiente tabla se dan los distintos valores de las probabilidades para las 4 combinaciones entre la variable dependiente y la independiente:

Y	$X=1$	$X=0$
$y=1$	$p(y=1\|x=1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(y=1\|x=0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$p(y=0\|x=1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$p(y=0\|x=0)=\frac{1}{1+e^{\beta_0}}$
$Total$	1	1

Habíamos definido el odds de que la variable respuesta esté presente entre los individuos con $X = 1$ como

$$\frac{p(y=1|x=1)}{p(y=0|x=1)} = \frac{p(y=1|x=1)}{1-p(y=1|x=1)}$$

y se define de forma similar para los que tienen $X = 0$. El odds-ratio es la razón entre los odds (razón de posibilidades o razón de ventajas):

$$OR = \frac{p(y=1|x=1)/1-p(y=1|x=1)}{p(y=1|x=0)/1-p(y=1|x=0)}$$

El valor nulo para la OR es el 1. Un OR = 1 implica que las dos categorías comparadas son iguales. El valor mínimo posible es 0 y el máximo teóricamente posible es infinito. Un OR inferior a la unidad se interpreta como que el desenlace es menos frecuente en la categoría o grupo que se ha elegido como de interés con respecto al otro grupo o categoría de referencia. Un OR = 3 se interpreta como una ventaja 3 veces superior de una de las categorías (X = 1) relativamente a la otra categoría (X=0).

Utilizando la tabla anterior,

$$\begin{aligned} OR &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{\beta_1} \end{aligned}$$

El hecho de que el odds-ratio sea la exponencial del coeficiente ha hecho que la regresión logística se haya hecho tan popular. Es bastante común encontrarse con el caso en el que el odd-ratio se interpreta como el riesgo relativo, es decir como

$$RR = \frac{p(y=1|x=1)}{p(y=1|x=0)}$$

es decir, se interpreta como una medida de cuanto más probable es encontrar el outcome entre los individuos con $X = 1$ que entre aquellos que tienen $X = 0$. **Para poder hacer esta aproximación es necesario que** $p(y = 1|x = 1)$ **y** $p(y = 1|x = 0)$ **sean pequeñas**.

Por ejemplo, supongamos que con los datos sobre percepción de salud queremos ver la relación entre esta variable y el sexo:

```
table(g02,sexo)
   sexo
g02    1    2
  0  594  812
  1 3072 2879
```

$$\begin{aligned} p(y=1|x=1) &= 2879/(2879+812) = 0{,}78 \qquad & p(y=0|x=1) &= 0{,}22 \\ p(y=1|x=0) &= 3072/(3072+594) = 0{,}838 \qquad & p(y=0|x=0) &= 0{,}162 \\ RR &= \frac{0{,}78}{0{,}838} = 0{,}93 \qquad & OR &= \frac{0{,}78/0{,}22}{0{,}838/0{,}162} = 0{,}68 \end{aligned}$$

tanto para $X = 0$ como para $X = 1$ la prevalencia de buena salud es alta, esta es la razón por la que hay diferencia entre RR y OR, pero de este tema nos ocuparemos más tarde. Suponiendo que pudieramos aproximar el RR mediante el OR, un OR estimado de 0.68 querría decir que la ocurrencia de buena salud en las mujeres es 0.68 veces menor.

Junto con el estimador del coeficiente o del OR , es importante utilizar los intervalos de confianza (I.C.) para obtener información adicional acerca del parámetro. Los I.C. para los coeficientes

del modelo (los β) se calculan asumiendo que la distribución de los parámetros estimados es aproximadamente normal, entonces un I.C. para el odds-ratio estimado se calcula tomando exponenciales en el I.C. para los parámetros:

$$exp\left[\hat{\beta}_1 \pm z_{\alpha/2} \times \widehat{s.e.}(\hat{\beta}_1)\right]$$

Que el I.C. para el OR sea $(0{,}61, 0{,}77)$ quiere decir que la buena salud entre las mujeres en la población de estudio es entre 0,61 y 0,77 veces menos probable que en los hombres.

2.2. Variable independiente politómica

En este caso la variable independiente tiene más de 2 categorías. En nuestro ejemplo agrupamos la edad en tres grupos: 18-29, 30-44 y 45-64 años

```
logistica2=glm(g02~edad2,family=binomial(link=logit))
exp(coefficients(logistica2))
(Intercept)      edad22      edad23
  9.0186914   0.6407686   0.2403349
```

R crea tres variables dicotómicas que indican si la persona está o no en el rango de edad correspondiente, al ajustar el modelo, R está suponiendo que el valor de la primera variable es 0 y por lo tanto está comparando los otros dos grupos de edad con ese. El odds ratio para `edad2_2` es odd de las personas entre 30 y 44 años dividido por el odds de las personas entre 18 y 29 años, 0,64, esto significa que estimamos que el odds de tener buena salud para personas entre 30 y 44 años es 0.64 veces el de las personas entre 18 y 29 años. Si estamos en el caso en el que podemos aproximar el RR mediante el OR diríamos que la prevalencia de buena salud en las personas de mediana edad es 0.64 veces menor que la prevalencia en personas jóvenes. Los I.C. se obtienen siguiendo el mismo procedimiento que en el caso de las variables dicotómicas.

2.3. Variable independiente continua

Cuando en un modelo de regresión logística utilizamos una variable predictora continua, la interpretación del parámetro estimado depende de cómo entra la variable en el modelo y de las unidades en que se mide. El parámetro β_1 representa el cambio en el log-odds cuando la variable X cambia en una unidad, o equivalentemente, e^{β_1} es el cambio en el odds cuando la variable X cambia en una unidad. Si en vez de cambiar en una unidad, cambia en c unidades, entonces el cambio vendría dado por $e^{c\beta_1}$.

Vamos a establecer la relación entre el índice de masa corporal (variable continua con rango de valores entre 14,69 y 62,87) y la percepción de salud.

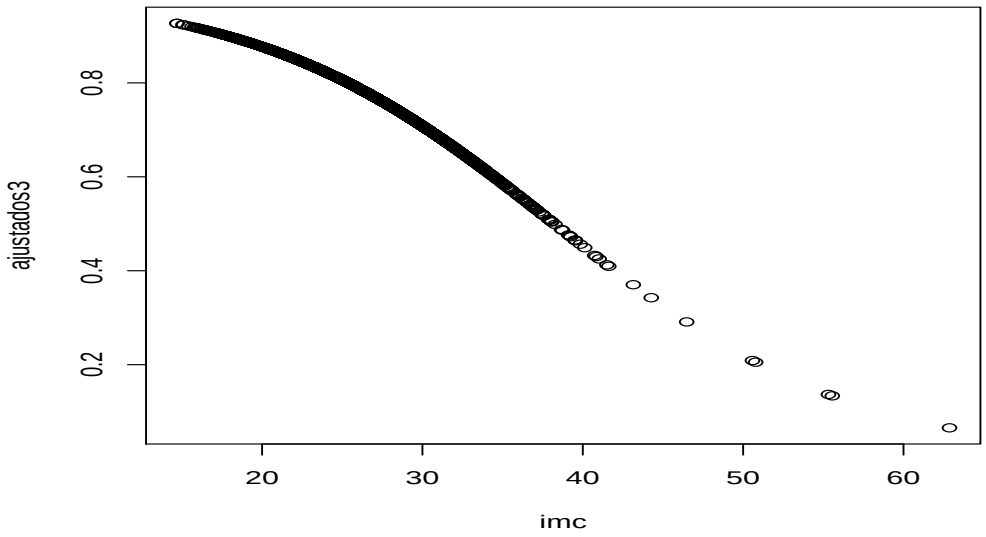
```
logistica3=glm(g02~imc,family=binomial(link=logit))
summary(logistica3)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.120383   0.195110   21.12   <2e-16 ***
imc         -0.107795   0.007637  -14.12   <2e-16 ***
---
    Null deviance: 7177.9  on 7356  degrees of freedom
```

```
Residual deviance: 6973.7  on 7355  degrees of freedom
AIC: 6977.7

Number of Fisher Scoring iterations: 4

 exp(coefficients(logistica3))
(Intercept)         imc
 61.5828259   0.8978114
```

Por lo tanto, por cada unidad que aumenta el índice de masa corporal el odds de tener buena salud es 0.89 veces menor, si aumenta en 10 unidades el odds sería $0{,}89^{10} = 0{,}31$ veces menor. Podemos ver los resultados en un gráfico:

```
ajustados3=predict(logistica3,type="response")
plot(imc,ajustados3)
```

Al poner type=response, le decimos que nos de las probabilidades, si type=link nos da en función del logit



La curva es decreciente ya que el odds ratio es < 1. Hay que tener cuidado al interpretar los resultados para imc altos ya que están basados en muy pocas observaciones.

2.4. Variables independientes categóricas y continuas

Hasta ahora hemos visto como ajustar modelos de regresión logística con una sola variable independiente, a estos modelos se les llama univariantes. Este tipo de modelos son válidos en pocas ocasiones, ya que en general, una variable independiente está asociada a otras variables independientes. Por eso es necesario considerar un modelo multivariante para poder compreneder mejor lo que ocurre con los datos. El objetivo es *ajustar* el efecto de cada variable en el modelo por las otras variables independientes. Por lo tanto, en el modelo de regresión logística, cada coeficiente β_j del modelo proporciona una estimación del logaritmo del odds ajustando por las otras variables incluídas en el modelo.

Para poder entender adecuadamente qué significan los coeficientes de un modelo de regresión logística multivariante es necesario tener claro lo que entendemos por *ajustar por otras variables*. Para ello, comenzamos por examinar el concepto de ajuste en el caso de regresión lineal, y luego lo extenderemos al caso de regresión logística.

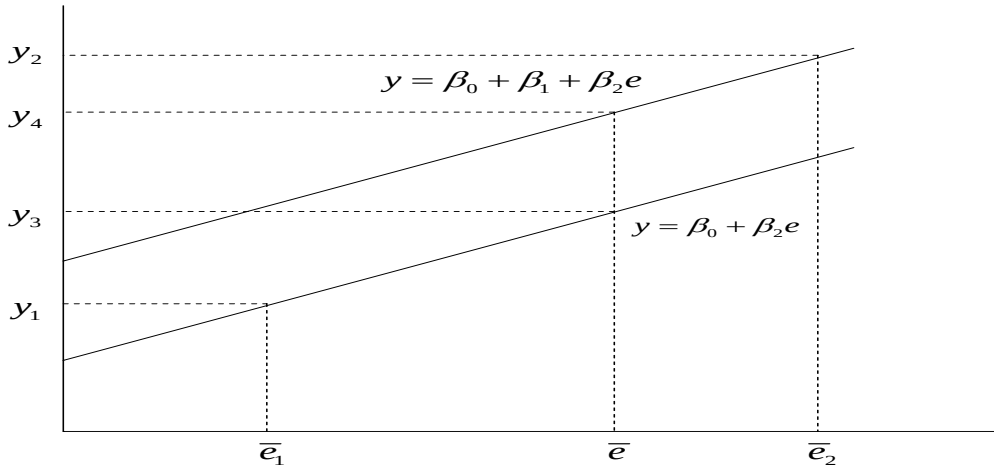
Supongamos que tenemos un modelo de regresión lineal con dos variables independientes, una dicotómica y otra continua, pero es la dicotómica la que es de interés, esta situación es frecuente en estudios epidemiológicos cuando se estudia la exposición a un riesgo y queremos ajustar por ejemplo por la variable edad.
Supongamos que queremos comparar la altura media de dos grupos de plantas jóvenes (unas sometidos a un cierto tratamiento y otros no), sabemos que la altura está asociada con la edad. Si la distribución de la edad es la misma en los dos grupos de plantas, entonces podríamos comparar directamente la altura media de los dos grupos, esta comparación nos daría una estimación de la diferencia en altura entre los dos grupos. Sin embargo, si un grupo fuera más joven que otro, entonces la comparación entre los dos grupos carecería de sentido, ya que una proporción de la diferencia en altura observada sería debida a la diferencia en edad, y por lo tanto sería imposible determinar el efecto del tratamiento en la altura si no eliminamos primero el efecto debido a diferencia en edad entre los grupos. Por simplicidad, vamos a suponer que no hay interacción entre ambas variables (más adelante veremos como comprobar si efectivamente la hay o no). El modelo estadístico en este caso sería:

$$y = \beta_0 + \beta_1 x + \beta_2 e + \epsilon$$

donde y representa la altura, x es una variable que toma el valor 0 si las plantas se han sometido a un tratamiento y valor 1 si las plantas no se han sometido ese tratamiento, y e representa la edad. En este modelo el parámetro β_1 representa la diferencia en altura entre los dos grupos, y el parámetro β_2 es la tasa de cambio en altura por semana de edad. Cada recta pasa por $(\overline{e}_1, y_1)$ y $(\overline{e}_2, y_2)$ ((y_1,y_2) representan la altura media de cada grupo), por lo tanto:

$$(y_2 - y_1) = \beta_1 + \beta_2(\overline{e}_2 - \overline{e}_1)$$

Por lo tanto la comparación no sólo implica la diferencia entre las medias de los grupos (β_1), sino también $\beta_2(\overline{e}_2 - \overline{e}_1)$ que representa la diferencia entre la edades de los grupos.
El proceso estadístico de ajustar por edad supone comparar los dos grupos para un mismo valor de la edad, el valor que se utiliza normalmente es la altura media de los dos grupos, $\overline{e}$, eso equivaldría



a comparar y_4 e y_3, es decir $(y_4 - y_3) = \beta_1 + \beta_2(\overline{e} - \overline{e}) = \beta_1$. La elección de la media de edad tiene sentido por dos razones: es un valor lógico y está dentro del rango de valores para los que creemos que la relación entre las variables es lineal.

Consideremos ahora una situación similar, pero en vez de la variable altura, tenemos la percepción de salud, por lo tanto la ecuación del modelo será:

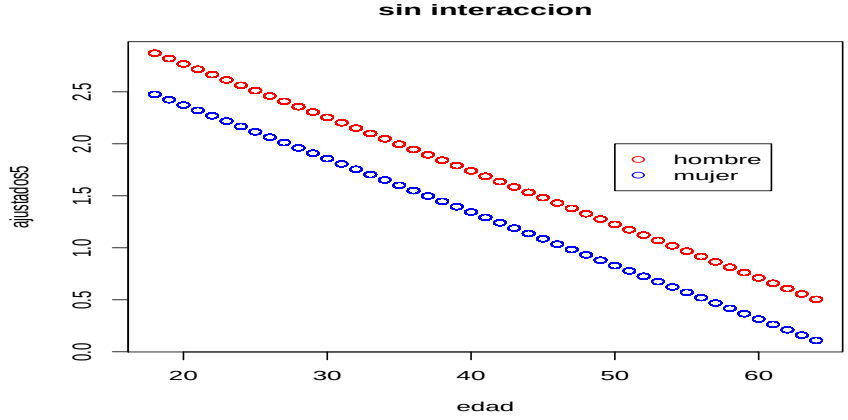
$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \beta_2 e$$

donde x es la variable sexo y e_1^{β} es odds ratio para un mismo valor de la edad, es decir el odds ratio que esperaríamos encontrar si la edad media en hombres y mujeres fuera similar.

```
logistica4=glm(g02~sexo,family=binomial(link=logit))
logistica5=glm(g02~sexo+edad,family=binomial(link=logit))
summary(logistica4)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.64320    0.04482  36.661  < 2e-16 ***
summary(logistica5)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.797233   0.120809  31.432  < 2e-16 ***
sexo2       -0.395887   0.061899  -6.396  1.6e-10 ***
edad        -0.051446   0.002512 -20.478  < 2e-16 ***
```

El parámetro β_1 es muy similar en ambos casos, esto es debido a que la edad media en hombre y en mujeres es muy similar (39.15 y 39.32 respectivamente), por lo tanto ajustar por edad no cambia el odds de la percepción de salud por sexos.

```
ajustados5=predict(logistica5,type="link")
plot(edad,ajustados5,type="n",main="sin interaccion")
points(edad[sexo==1],ajustados5[sexo==1],col=2)
points(edad[sexo==2],ajustados5[sexo==2],col=4)
legend(50,2, col=c(2,4),pch=c(1,1),c("hombre","mujer"))
```



El método de ajuste cuando las variables independientes son todas dicotómicas, politómicas, continuas o mezclas de ellas es idéntico al que acabamos de describir, por ejemplo si utilizamos la variable edad agrupada en tres categorías

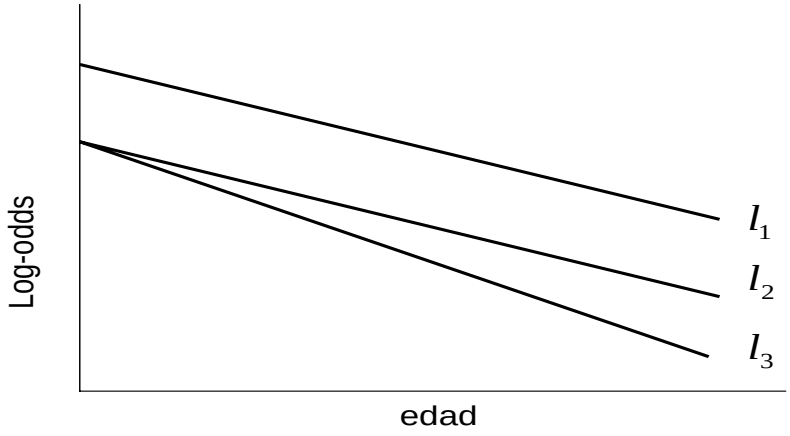
Un punto que hay que tener en cuenta a la hora de interpretar los parámetros del modelo es si se cumplen las hipótesis en las que se basa: que la relación es lineal y que las rectas son paralelas.

2.5. Interacción y confusión

El término confusión se utiliza para describir una covariable que está asociada al mismo tiempo a la variable respuesta y al factor de interés. Cuando estas dos asociaciones están presentes la relación entre la variable respuesta y el factor están *confundidos*. Es importante recordar que un indicio de confusión es el cambio en el valor de los parámetros estimados del factor cuando se introduce la covariable.

Si la asociación entre la covariable (edad) y la variable respuesta es la misma para cada nivel del factor (sexo) entonces no hay interacción entre la covariable y el factor. Gráficamente, la ausencia de interacción da lugar a un modelo con dos rectas paralelas, una para cada nivel del factor. Cuando la interacción está presente, la asociación entre el factor y la variable respuesta depende del nivel de la covariable, es decir la covariable modifica el efecto del factor.

En el ejemplo anterior, si hubiera interacción implicaría que el logit en el grupo de mujeres sigue



una línea con diferente pendiente. Por lo tanto, un aspecto importante a la hora de buscar el modelo más adecuado es determinar si hay evidencia de interacción entre las variables, este aspecto será tratado más adelante.

En la figura siguiente se muestran 3 logits diferentes. Supongamos que la línea l_1 corresponde al logit para mujeres en función de la edad, y l_2 lo es para los hombres. Estas dos líneas son paralelas, indicando que la relación entre la edad y la percepción de salud es la misma para hombres que para mujeres. En esta situación no hay interacción y el logaritmo del odds-ratio para sexo controlado por edad viene dado por la distancia entre las rectas, $l_1 - l_2$, esta diferencia es la misma para cualquier edad. Supongamos ahora que el logit para mujeres viene dado por la línea l_3, ésta indica que la relación entre la percepción de salud y la edad es distinta en hombres y en mujeres. Cuando esto ocurre diremos que hay *interacción*. El logaritmo del odds-ratio estimado para sexo controlado por edad viene dado por la distancia vertical $l_1 - l_3$, pero ahora la diferencia depende de la edad en la que se está haciendo la comparación. Por lo tanto, no podemos estimar el odds-ratio para sexo sin

especificar para qué valor de la edad estamos haciendo esta comparación. En este caso la edad es un *efecto modificador*.

Una manera de detectar una variable de confusión es ver si los parámetros estimados para el factor de interés cambiaban sustancialmente al introducir la covariable, este criterio no se puede aplicar al caso de la interacción ya que la inclusión de un término de interacción, especialmente si una de las variables es continua, normalmente da lugar a cambios en los parámetros estimados, aunque la interacción no sea significativa.

En el caso de los datos de percepción de salud, para saber si edad es un variable de confusión con respecto al sexo, vemos si el valor del parámetro del seco cambian mucho al introducir edad:

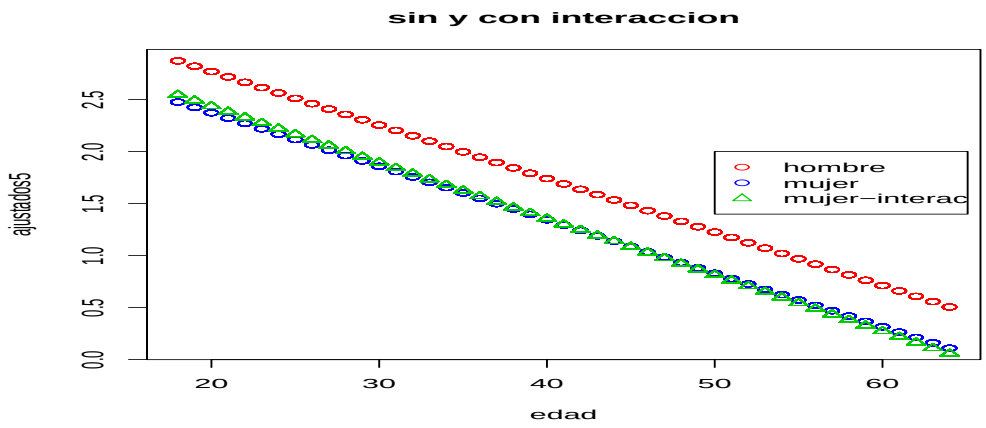
```
logistica4

Call:  glm(formula = g02 ~ sexo, family = binomial(link = logit))

Coefficients:
(Intercept)        sexo2
     1.6432      -0.3775

logistica5

Call:  glm(formula = g02 ~ sexo + edad, family = binomial(link = logit))

Coefficients:
(Intercept)        sexo2         edad
    3.79723     -0.39589     -0.05145
```

En este caso, hay muy poca diferencia, por lo que podemos considerar que no hay confusión. Ahora vemos si existe interacción entre ambas variables

```
logistica7=glm(g02~sexo+edad+sexo:edad,family=binomial(link=logit))
ajustados7=predict(logistica7,type="link")
plot(edad,ajustados5,type="n",main="sin y con interaccion")
points(edad[sexo==1],ajustados5[sexo==1],col=2)
points(edad[sexo==2],ajustados5[sexo==2],col=4)
points(edad[sexo==2],ajustados7[sexo==2],col=3,pch=2)
legend(50,2, col=c(2,4,3),pch=c(1,1,2),c("hombre","mujer", "mujer-interac"))
```

El gráfico nos muestra que el hecho de introducir un término de interacción casi no modifica la pendiente de la recta, lo cual es un indicio de que que no hay interacción. Por lo tanto edad no es una variable de confusión (como vimos antes) ni es un efecto modificador, estas conclusiones hay que comprobarlas mediante tests que comparen los modelos (lo veremos más adelante).
Si en vez de utilizar la variable edad, utilizamos el peso:

```
logistica8=glm(g02~sexo+peso,family=binomial(link=logit))
logistica8

Call:  glm(formula = g02 ~ sexo + peso, family = binomial(link = logit))
```




```
Coefficients:
(Intercept)        sexo2         peso
    3.71254     -0.82594     -0.02619

logistica4

Call:  glm(formula = g02 ~ sexo, family = binomial(link = logit))

Coefficients:
(Intercept)        sexo2
     1.6432      -0.3775
```

Vemos que al introducir el peso el parámetro de sexo pasa de $-0{,}38$ a $-0{,}82$ un descenso de más del 50 % lo que da indicios de que el peso pueda ser una variable de confusión.

```
logistica9=glm(g02~sexo+peso+sexo:peso,family=binomial(link=logit))
ajustados9=predict(logistica9,type="link")
plot(peso,ajustados9,type="n",main="con interaccion")
points(peso[sexo==1],ajustados9[sexo==1],col=2)
points(peso[sexo==2],ajustados9[sexo==2],col=4)
legend(140,2, col=c(2,4),pch=c(1,1),c("hombre","mujer"))
```

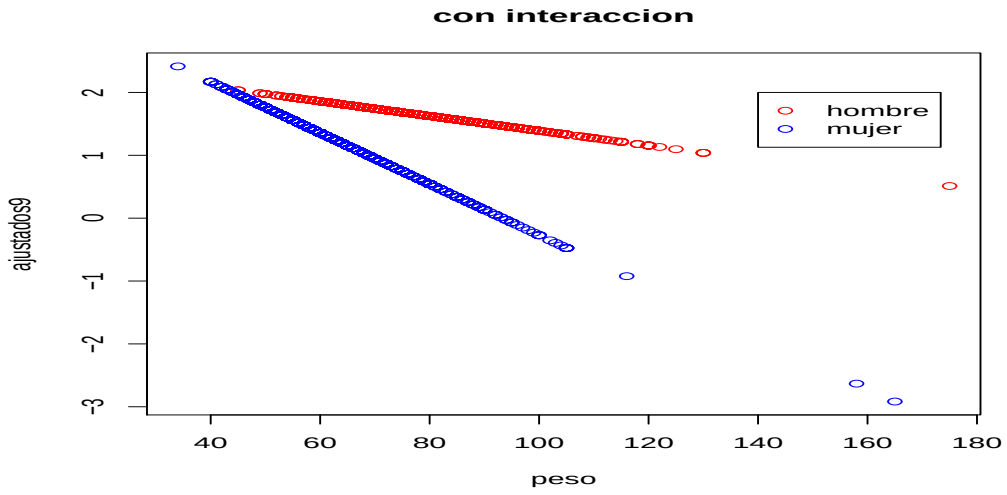
En el siguiente gráfico vemos que las rectas a las que da lugar el modelo con interacción no son paralelas y que la percepción de salud en función de la edad es distinta para hombres que para mujeres, éstas se ven más afectas por el peso.

En resumen, determinar si una variable es un efecto modificador y/o de confusión depende de varios aspectos. Para que sea una variable de confusión se han de cumplir dos condiciones:

1. La variable ha de estar asociada con la variable respuesta, es decir que el parámetro estimado para esta variable ha se ser distinto de 0.

2. La variable ha de estar asociada al factor de riesgo

Mientras que para detectar un efecto modificador hemos de mirar a la estructura paramétrica del logit.



En la práctica, un método para evaluar si una variable es de confusión, es comparar el coeficiente estimado para el factor de riesgo de los modelos que contienen y no contienen a la covariable. Cualquier cambio científicamente importante en el coeficiente del factor de riesgo, sugiere que la variable es de confusión. Si esto ocurre y la interacción no es estadísticamente significativa, la variable ha de salur del modelo. Por otra parte, una variable es un efecto modificador, sólo cuando el término de interacción añadido al modelo es científica y estadísticamente significativo.

2.6. Interpretación del OR en presencia de interacción

Antes hemos visto que cuando existe interacción entre un factor de riesgo y otra variable, el parámetro estimado para el factor de riesgo depende del valor de la variable que interactúa con este, por lo tanto no podemos obtener el OR simplemente exponenciando el valor del parámetro. Solución:

1. Escribir la ecuación del logit para los dos niveles del factor de riesgo.
2. Calcular la diferencia entre los logit
3. Exponenciar el valor obtenido.

Por ejemplo, consideremos un modelo que contiene sólo dos variables y su interacción. Llamamos F al factor, X a la covariable y $F \times X$ a la interacción. El logit para $F = f$ y $X = x$ es:

$$log\left(\frac{p(f,x)}{1-p(f,x)}\right) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x,$$

Supongamos que queremos el OR para los dos niveles de F, f_1 versus f_0:

$$\begin{aligned} log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) &= \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x \\ log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) &= \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x \\ log\left(\frac{p(f_1,x)}{1-p(f_1,x)}\right) - log\left(\frac{p(f_0,x)}{1-p(f_0,x)}\right) &= \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0) \\ OR &= exp\left[\beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0)\right] \end{aligned}$$

En el ejemplo de los datos sobre percepción de salud, supongamos que queremos saber el OR del sexo:

$$OR = exp[1{,}23 - 0{,}029x]$$

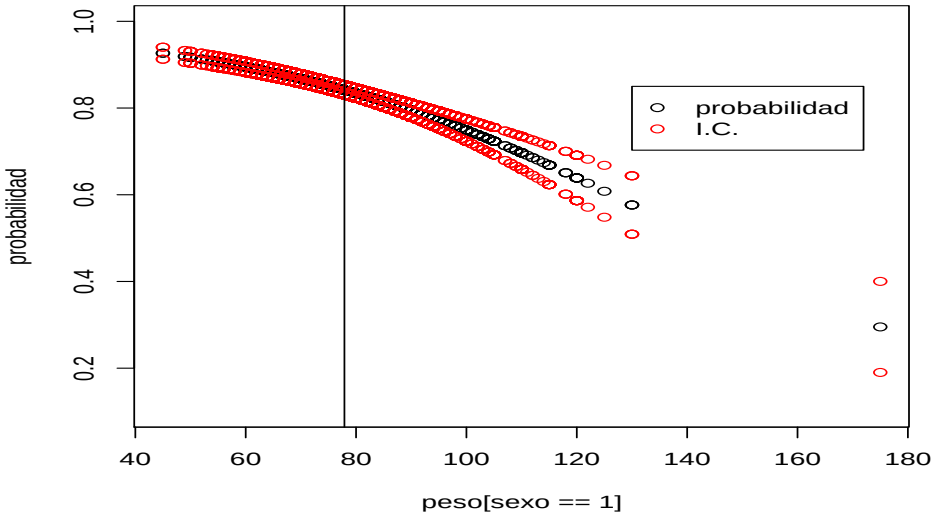
2.7. Interpretación de los valores ajustados

Aunque, en general, en regresión logística, los parámetros y los OR son el centro de interés, hay en ocasiones en las que los valores ajustados son también importantes. Calcular los intervalos de confianza para el logit es sencillo, ya que R da los errores estándar. Para calcular el intervalo de confianza para las probabilidades sólo hay que utilizar la relación entre el logit y la probabilidad:

$$\frac{e^{I.C.logit}}{1 + e^{I.C.logit}}$$

Por ejemplo supongamos que estamos interesados en el efecto del peso sobre la percepción de salud en los hombres:

```
ajustados8=predict(logistica8, se.fit=TRUE)
names(ajustados8)
L.inf=with(ajustados8,exp(fit-1.96*se.fit)/(1+exp(fit-1.96*se.fit)))
L.sup=with(ajustados8,exp(fit+1.96*se.fit)/(1+exp(fit+1.96*se.fit)))
with(ajustados8,plot(peso[sexo==1],(exp(fit)/(1+exp(fit)))[sexo==1],
                                  ylim=c(0.1,1),ylab="probabilidad"))
points(peso[sexo==1],L.inf[sexo==1],col=2)
points(peso[sexo==1],L.sup[sexo==1],col=2)
legend(130,0.85, c("probabilidad","I.C."),col=c(1,2),pch=c(1,1))
```



El intervalo de confianza es más ancho en los extremos ya que hay menos observaciones y por lo tanto las estimaciones son menos precisas. La línea vertical corresponde al peso medio de los hombres, por eso es por lo que el intervalo de confianza es más estrecho. Cada punto y su correspondiente intervalo de confianza está estimando la media de la variable respuesta, percepción de salud, entre los hombres para un valor específico del peso. Por ejemplo, a los 80 kilos, el valor estimado es 0,835 y el intervalo de confianza es $(0{,}823, 0{,}847)$. La interpretación es que la proporción estimada de hombres con buena salud que pesan 80 kilos es 0,835 y puede ser tan baja como 0,823

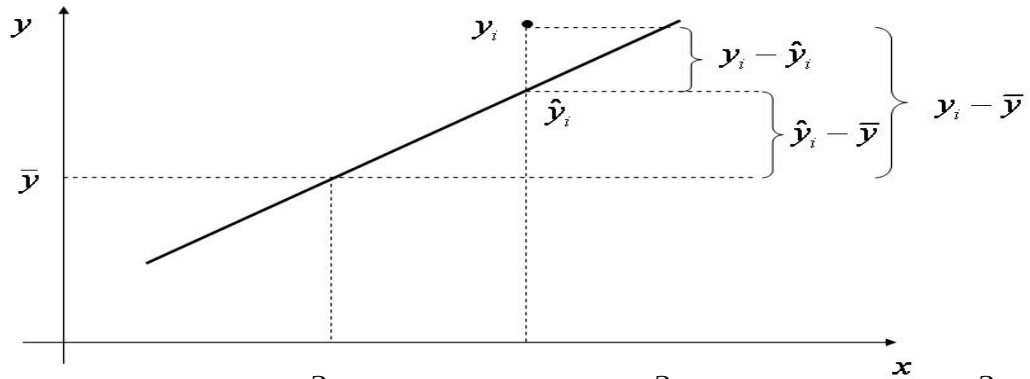
y tan alta como 0,847 con un 95 % de confianza. Si nos vamos al caso extremos de hombres que pesen 165 kilos, la estimación de la proporción es 0,65 y el intervalo de confianza $(0{,}494, 0{,}783)$. Un error que se comete a veces es aplicar las estimaciones de la probabilidad a los sujetos individuales, sin embargo, estas probabilidades representan la proporción de sujetos.

3. Selección de variables

Una vez que hemos estimado los parámetros, hemos de determinar si las variables del modelo son significativas o no. Esto supone la formulación de tests de hipótesis para determinar si las variables independientes del modelo están relacionadas *significativamente* con la variable respuesta. El método para realizar estos tests es bastante general y difiere sólo de un modelo a otro en pequeños detalles. Comenzamos con un enfoque general para el caso en el que hay una sola variable independiente en el modelo. Esta metodología se basa en la siguiente pregunta: *¿El modelo que incluye la variable en cuestión nos da más información sobre la variable respuesta que un modelo que no la incluya?*. La respuesta a esta pregunta se hace comparando los valores observados con los valores predichos por ambos modelos. Si los valores predichos por el modelo que incluye la variable en cuestión son mejores (en algún sentido) que los predichos cuando la variable no está en el modelo, diremos que la variable es *significativa*. Es importante que distingamos esta cuestión de la que nos plantearemos cuando se discuta la bondad de ajuste del modelo (ya que en ese caso nos fijaremos en la adecuación de los valores ajustados en un sentido absoluto).

Ilustramos la metodología partiendo del método de regresión lineal simple. En este caso el test se realiza mediante la tabla del análisis de la varianza. La tabla se basa en la división de la suma de cuadrados de la distancia entre las observaciones y su media en dos partes: (1) $SSE =$ suma de cuadrados de las desviaciones entre las observaciones y la recta, también se llama suma de cuadrados residual, y (2) $SSR =$ suma de cuadrados de las desviaciones entre los valores ajustados por el modelo y la media de las observaciones, también llamada suma de cuadrados debida a la regresión.

Esto nos ofrece una forma muy simple de comparar los valores observados con los ajustados por ambos modelos. En el modelo de regresión, comparamos los valores ajustados y los observados



$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

mediante SSE:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde $\hat{y}_i$ representan los valores ajustados por el modelo.

Si el modelo no contiene a la variable independiente, sólo hay un parámetro, β_0 y $\hat{\beta}_0 = \overline{y}$ por lo tanto en este caso SSE es igual a $\sum_{i=1}^{n}(y_i - \overline{y})^2$. Cuando introducimos la variable independiente en el modelo, cualquier reducción en SSE será debido a que el coeficiente de la variable independiente (la pendiente de la recta) no es cero. El cambio en SSE sería:

$$\left[\sum_{i=1}^{n}(y_i - \overline{y})^2\right] - \left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]$$

Un valor alto indica que la variable independiente es importante.
En el caso de un GLM y en particular en la regresión logística la filosofía es la misma: *Comparar los valores observados y predichos por el modelo con y sin la variable explicativa.* En un GLM la comparación se hace utilizando la función de verosimilitud que describimos al principio del capítulo

$$\ln L(\beta) = \sum (y_i \ln(p(x_i)) + (n_i - y_i)\ln(1 - p(x_i)).$$

Si definimos un modelo saturado como aquel en que hay tantos parámetros como observaciones, entonces en dicho modelo $\hat{y}_i = y_i$, y la comparación entre valores observados y ajustados de hace mediante:

$$D = -2\ln\left[\frac{\text{verosimilitud del modelo ajustado}}{\text{verosimilitud del modelo saturado}}\right]$$

la cantidad dentro de los corchetes se llama **razón de verosimilitud**, a esta cantidad se le toma logaritmo y se multiplica por -2 porque de esta manera obtenemos una cantidad que sigue una distribución conocida χ^2 y por lo tanto se pueden realizar contrastes de hipótesis, a este test se le llama **test de la razón de verosimilitud**. En el caso de un modelo de regresión logística viene dado por

$$-2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}(x_i)}{y_i}\right) + (1 - y_i)\ln\left(\frac{1 - \hat{p}(x_i)}{1 - y_i}\right)\right]$$

Al estadístico D se le llama *devicance*. Cuando la variable respuesta es dicotónica, la verosimilitud del modelo saturado es 1 por lo tanto, en un modelo de regresión logística la deviance es:

$$D = -2\ln[\text{verosimilitud del modelo ajustado}]$$

```
logistica4

Call:  glm(formula = g02 ~ sexo, family = binomial(link = logit))

Coefficients:
(Intercept)        sexo2
     1.6432      -0.3775

Degrees of Freedom: 7356 Total (i.e. Null);  7355 Residual
Null Deviance:      7178
Residual Deviance: 7138
```

A la hora de comprobar si una variable en el modelo es significativa lo que hacemos es comparar el valor de D con y sin la variable independiente en el modelo:

$$\begin{aligned} G &= -2\ln\left[\frac{\text{verosimilitud del modelo sin la variable}}{\text{verosimilitud del modelo con la variable}}\right] \\ &= D(\text{modelo sin la variable}) - D(\text{modelo con la variable}). \end{aligned}$$

En el ejemplo anterior:

```
  anova(logistica4,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: g02

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    7356     7177.9
sexo    1     40.1      7355     7137.8 2.397e-10
```

A la vista de estos resultados lo único que podemos decir es que la variable es significativa, sin embargo para determinar que la variable debe estar en el modelo hay que comprobar que el modelo es el adecuado, así como incluir otras variables.

Otro test alternativo es el **Wald test**, este test compara el estimador máximo verosímil del parámetro con su error estándar, bajo ha hipótesis nula $H_0 : \beta_1 = 0$ la razón entre estas dos cantidades sigue una distribución $N(0, 1)$, pero este test lleva a no rechazar H_0 cuando el parámetro es significativo, por lo tanto se debe utilizar el test de la razón de verosimilitud.
Si en vez de tener una sola variable tenemos más, este test se hace de la misma forma con la única diferencia que los grados de libertad de la distribución χ^2 serán igual a la diferencia en número de parámetros entre los dos modelos que estamos comparando. La decisión de incluir o no de un conjunto de variables en un modelo no puede basarse sólo en los test de significación. Un complemento a los test de hipótesis para los parámetros son los intervalos de confianza, que se calculan como:

$$\hat{\beta}_i \pm z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_i\right),$$

como hemos visto antes, a partir de esta expresión, es sencillo calcular los intervalos de confianza para el logit y para la probabilidad estimada.

Hasta ahora nos hemos centrado en la estimación, contraste e interpretación de los coeficeintes del modelo. Los ejemplos que hemos utilizado incluían pocas variables explicativas y, por lo tanto, es bastante sencillo seleccionar el mejor modelo. Sin embargo, hay situaciones en las que hay muchas variables explicativas que, en principio, podrían estar en el modelo. El objetivo de cualquier método de selección de variables es elegir aquellas variables que dan lugar al mejor modelo dentro del contexto científico del problema que se trata. Para conseguir esto necesitamos:

1. Un plan para seleccionar variables.

2. Un conjunto de métodos para comprobar el ajuste del modelo, tanto para cada variable de forma individual, como para todas en conjunto.

Encontrar un modelo adecuado es en parte ciencia, parte métodos estadísticos, parte experiencia y sobre todo sentido común.

3.1. ¿Cómo seleccionar las variables en la práctica?

Tradicionalmente buscamos el modelo más sencillo que ajuste bien los datos. La razón por la que buscamos minimizar el número de variables en el modelo es que si el número de variables es grande comparado con el número de observaciones o la proporción de ceros y unos es extrema, entonces aparecen problemas numéricos y los estimadores son inestables ya que dependen mucho de los datos observados.

Pasos a seguir en la seleción del modelo

1. **Análisis individual de cada variable**
   Cuando las variables explicativas son nominales, ordinales, o continuas pero toman pocos valores enteros, podemos utilizar tablas de contingencia o regresión logística univariante, y utilizar el test de la razón de verosimilitud o el de Pearson para determinar si hay asociación entre la variable respuesta y cada una de las explicativas. Hay que prestar atención a los casos en que hay celdas en la tabla de frecuencia que son 0. Si introducimos esta variable en el modelo dará lugar a errores, ya que alguno de los parámetros estimados estará muy próximo a cero o a infinito. Una posible solución es recodificar la variable y unir categorías, o si variable es ordinal, tratarla como si fuera continua. Para variables continuas hemos de utilizar un modelo de regresión logística para obtener los coeficientes y el test de la razón de verosimilitud.

   ```
   logistica10=glm(g02~peso,family=binomial(link=logit))
   anova(logistica10,test="Chisq")
   ```

2. **Selección de variables para el análisis multivariante**

   Cualquier variable que en el análisis individual tenga un p-valor en el test de la razón de verosimilitud menor que 0.25 es candidata para el análisis multivariante. La razón por la que incluímos variables con un p-valor tan alto es porque puede que si tomamos un valor más pequeño estemos eliminando variables importantes, ya que hay veces en que las variables por separado tienen escasa asociación con la variable respuesta, pero que al estar juntas en el modelo se convierten en un predictor importante.
   Cuando el número de observaciones es grande y el tamaño de cada valor de la respuesta es grande en comparación con el número de variables, no hay problema si queremos empezar introduciendo todas las variables en el modelo. Sin embargo, cuando éste no es el caso, introducir muchas variables da lugar a problemas numéricos, por lo tanto es necesario seleccionar un subconjunto de variables.
   Otra posibilidad es utilizar un método *stepwise* que consiste en la selección de variables por introducción o exclusión del modelo de forma secuencial, se puede hacer hacia delante o hacia atrás. Por último, está el *método de selección del mejor subconjunto*, con este método se ajustan modelos con una, dos, tres,..., variables, y se examinan para determinar cual es el mejor con respecto un cierto criterio específico. En cualquier caso, es responsabilidad del analista el que las variables incluídas en el modelo de partida tengan sentido desde el punto de vista

científico.
De entre las siguientes variables: educa, edad2, sexo, con.tab, peso ¿Cuáles incluirías en el análisis multivariante?.

```
anova(glm(g02~educa,family=binomial(link=logit)),test="Chisq")
.
.
.
```

3. **Verificar la importancia de cada variable**
   Una vez que se ha ajustado el modelo multivariante hay que verificar la importancia de las variables que hemos incluído en el modelo. Esto debe hacerse mediante el Wald test, y comparando el valor del coeficiente en el modelo univariante con el que tiene en el modelo multivariante, esto también nos va a indicar si alguna de la variables es de confusión. Las variables que no contribuyan al modelo basándose en estos criterios deben ser eliminadas y el modelo debe ser ajustado otra vez. Cada vez que eliminamos una variable hemos de comparar el nuevo modelo con el anterior mediante un test de la razón de verosimilitud. El proceso sigue hasta que todas las variables importantes estén en el modelo y todas las excluídas sean científica y/o estadísticamente no significativas.

```
logistica12=glm(g02~educa+edad2+sexo+peso,family=binomial(link=logit))
summary(logistica12)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.745780   0.250075  10.980  < 2e-16 ***
educa2       0.458386   0.104607   4.382 1.18e-05 ***
educa3       1.031458   0.111568   9.245  < 2e-16 ***
educa4       1.438206   0.112947  12.733  < 2e-16 ***
edad22      -0.314390   0.092066  -3.415 0.000638 ***
edad23      -1.072312   0.089383 -11.997  < 2e-16 ***
sexo2       -0.631798   0.077985  -8.102 5.43e-16 ***
peso        -0.018155   0.002796  -6.493 8.39e-11 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7177.9  on 7356  degrees of freedom
Residual deviance: 6464.0  on 7349  degrees of freedom
```

¿Cambian los coeficientes con respecto al análisis univariante?

```
glm(g02~educa,family=binomial(link=logit))
.
.
.
```

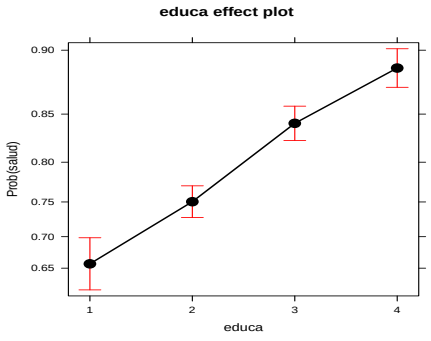
4. **Comprobar las interacciones**
   Una vez que hemos seleccionado las variables hay que comprobar las interacciones. Recordemos que la interacción entre dos variables significa que el efecto de una de las dos variables no es constante para todos los niveles de la otra. La decisión final sobre si un término de interacción debe ser incluído en el modelo debe basarse tanto en consideraciones estadísticas como en prácticas.
   Empezaremos por incluir las interacciones entre pares de variables que tengan alguna base científica para interaccionar. Introducimos una interacción cada vez y vemos si es significativa utilizando el test de la razón de verosimilitud. La inclusión de una interacción que no sea significativa tiene el efecto de no cambiar el valor de los parámetros estimados, pero aumenta el error estándar.

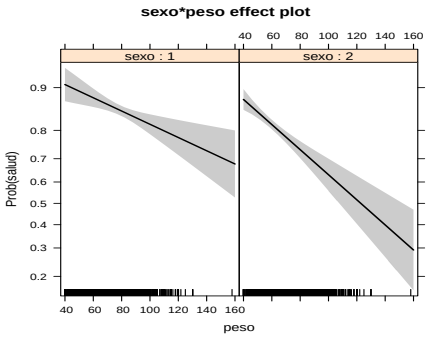
```
logistica13=glm(g02~educa+edad2+sexo*peso,family=binomial(link=logit))
anova(logistica12,logistica13, test="Chisq")
Analysis of Deviance Table

Model 1: g02 ~ educa + edad2 + sexo + peso
Model 2: g02 ~ educa + edad2 + sexo * peso
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      7349       6464
2      7348       6460    1        4   0.04509
```

Por lo tanto el modelo selecionado incluye la interacción.

Podemos realizar gráficos para representar el efecto de las variables explicativas e interacciones. Para ello utilizamos la librería `effects`:

```
library(effects)
plot(effect("educa",logistica13),ylab="Prob(salud)")
plot(effect("sexo*peso",logistica13),ylab="Prob(salud)")
```





En algunas ocasiones no sabemos qué variables pueden ser relevantes en el modelo, en estos casos podemos introducir todas y utilizar *stepwise* para elegir entre las variables de forma rápida y sencilla. Cualquier método stepwise está basado en un algoritmo que comprueba la importancia de las variables en el modelo y las incluye o excluye de acuerdo a una regla fija. En el caso de regresión logística (y de cualquier otro glm) se considera la variable más importante, aquella que da lugar al cambio más importante en un criterio llamado AIC, el cual está basado en el deviance y el número

de parámetros del modelo. El proceso empieza ajustando el modelo con sólo la constante y comparándolo con el modelo que resulta de introducir cada una de las variables, aquella que da lugar a un valor del AIC más pequeño es considerada la más importante y se introduce en el modelo. A continuación se compara el modelo con una variable con los modelos que resultan de introducir otra variables más, y así sucesivamente. Además en cada paso se comprueba que la variable que se ha introducido en paso anterior sigue siendo significativa. Una vez que se han elegido así las variables y sus interacciones, habría que comprobar una a una con el test de la razón de verosimilitud.

¿Coinciden tus conclusiones con las obtenidas utilzando un procedimiento stepwise?

```
logistica11=glm(g02~educa+edad2+sexo+con_tab+peso,family=binomial(link=logit))
step(logistica11,trace=0)
```

3.2. Bondad de ajuste del modelo

Cuando llegamos a este punto del análisis, estamos presuponiendo que el modelo contiene todas las variables importantes y que están presentes en la forma funcional correcta. Ahora el siguiente paso es saber lo eficiente que es es modelo a la hora de predecir la variable respuesta. A esto se le llama *bondad de ajuste*. Diremos que un modelo ajusta bien si la distancia entre los valores observados, y y los ajustados, $\hat{y}$, es pequeña. Esto implica:

1. Cálculo de medidas de bondad de ajuste
2. Diagnosis del modelo

Medidas de bondad de ajuste

X^2 de Pearson y Deviance

Son medidas de la diferencia entre los valores ajustados y observados basadas en los residuos del modelo, los *Pearson residuals* son:

$$r(y_j, \hat{p}_j) = \frac{y_j - m_j\hat{p}_j}{\sqrt{m_j\hat{p}_j(1-\hat{p}_j)}}$$

donde m_j es el número de individuos que comparten el mismo patrón de covariables, es decir la misma combinación de las variables explicativas. El estadístico X^2 de Pearson es

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{p}_j)^2,$$

donde J es el número de patrones diferentes. Los *Pearson residuals* estandarizados siguen aproximadamente una distribución N(0,1) por lo que deberían estar en el intervalo $(-3, 3)$

El otro estadístico está basado en los *Deviance residuals* que se calculan a partir de la definición de la deviance que vimos anteriormente. En ambos casos, se considerará que un valor del estadístico por encima de 4 podría ser un indicio de que el modelo no ajusta bien. Este test sólo es posible utilizarlo cuando el número de datos es bastante más grande que el número de patrones. En el caso en el que una de las variables es continua no deben usarse estos tests. Si este es el caso hay que utilizar el test de Hosmer- Lemeshow.

En la librería `ResourceSelection` hay una función que ajusta el test de Hosmer- Lemeshow.

```
library(ResourceSelection)
hoslem.test(g02,fitted(logistica13))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  g02, fitted(logistica13)
X-squared = 6.1108, df = 8, p-value = 0.6348
```

La probabilidad es 0.63, lo que indicaría que el modelo ajusta correctamente los datos.

4. Predicciones con el modelo: clasificación de sujetos

Tablas de clasificación

Esta tabla resulta de cruzar la variable respuesta, y, con una variable dicotómica cuyos valores resultan de las probabilidades estimadas en el modelo. Para definir esta variable hemos de decidir un punto de corte c y comparar las probabilidades estimadas con c, si son mayores, la variable vale 1 y si son menores 0. Lo más normal es tomar $c = 0{,}5$ o un valor mayor. Sin embargo, la tabla de clasificación por si sola no es una buena medida de ajuste, aunque puede ser útil junto a otras medidas basadas en los residuos.

Las tablas no deben usarse para comparar modelos ya que depende en gran parte de la distribu-

	$y = 1$	$y = 0$	Total
$\hat{y}_i = 1$	a	b	$a + b$
$\hat{y}_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	n

Cuadro 1: Tabla de clasificación

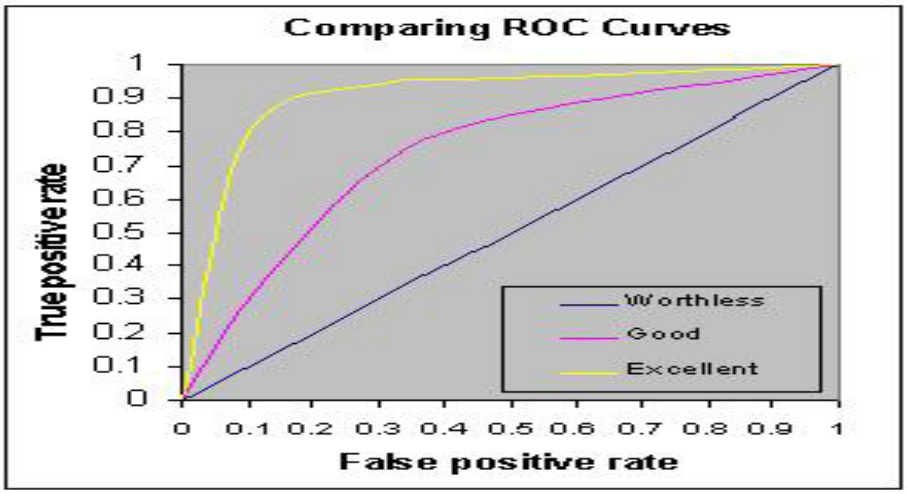
ción de las probabilidades en la muestra en la que están basadas, el mismo modelo evaluado en dos muestras diferentes puede dar lugar a clasificaciones distintas. Si el objetivo del análisis no es la clasificación de sujetos, entonces las tablas de clasificación son sólo el complemento de otras medidas de ajuste.

4.1. Área bajo la curva ROC

Definimos:

- La ***sensitividad*** es la proporción de verdaderos 1s estimados como 1s (la probabilidad de predecir correctamente un 1): $Ss = a/(a + c)$.
- La ***especificidad*** es la proporción de verdaderos 0s estimados como 0s (la probabilidad de predecir correctamente un 0): $Sp = d/(b + d)$.
- La ***tasa de falsos positivos*** es la proporción de verdaderos 0s estimados como 1s (la probabilidad de predecir incorrectamente un 0): $F+ = b/(b + d)$.
- La ***tasa de falsos negativos*** es la proporción de verdaderos 1s estimados como 0s (la probabilidad de predecir incorrectamente un 1): $F- = c/(a + c)$.

Cada vez que elegimos un punto de corte la sensitividad y especificidad cambian, si nuestro objetivo es buscar el punto de corte óptimo desde el punto de vista de la clasificación, entonces buscaremos aquel que maximiza la sensitividad y especificidad. La curva ROC (Receiver Operating Characteristic) es un gráfico de sensibilidad frente a 1−especificidad (tasa de falsos positivos) para todos los puntos de corte. El área por debajo de esa curva puede tomar valores entre 0 y 1, y nos proporciona una medida de la habilidad del modelo para discriminar entre aquellos individuos que presentan la respuesta de interés frente a los que no la presentan.



En general:

- Si $ROC \leq 0{,}5$ el modelo no ayuda a discriminar
- Si $0{,}6 \leq ROC \leq 0{,}8$ el modelo discrimina de forma adecuada
- Si $0{,}8 \leq ROC \leq 0{,}9$ el modelo discrimina de forma excelente
- Si $ROC \geq 0{,}9$ el modelo discrimina de forma excepcional

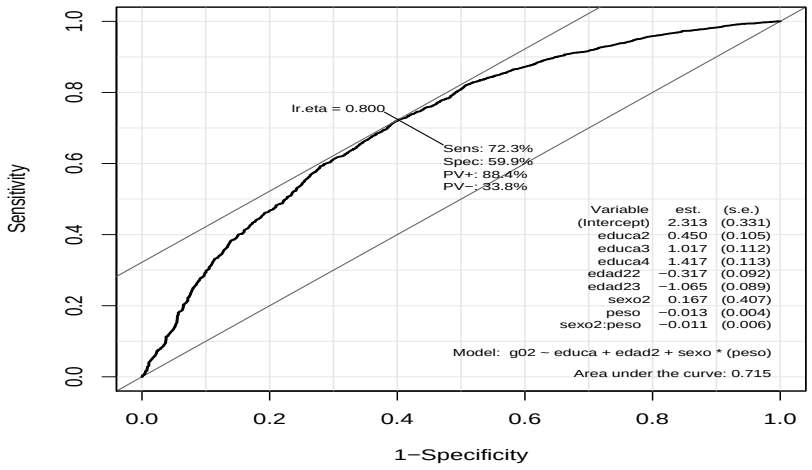
```
library(Epi)
ROC( form=g02~educa+edad2+sexo*(peso),plot="ROC")
```

5. Diagnosis en regresión logística

Cuando se violan las hipótesis en las que se basa el modelo de regresión logística, el modelo ajustado puede dar lugar a los siguientes errores:

- Coeficientes sesgados
- Estimadores ineficientes
- Inferencia estadística no válida

El sesgo se refiere a una tendencia sistemática a sobrestimar o subestimar los coeficientes del modelo. La ineficiencia se refiere a los errores estándar elevados para los coeficientes, esto hace que supongamos que el parámetro es cero cuando no lo es. La inferencia no válida se refiere a los casos



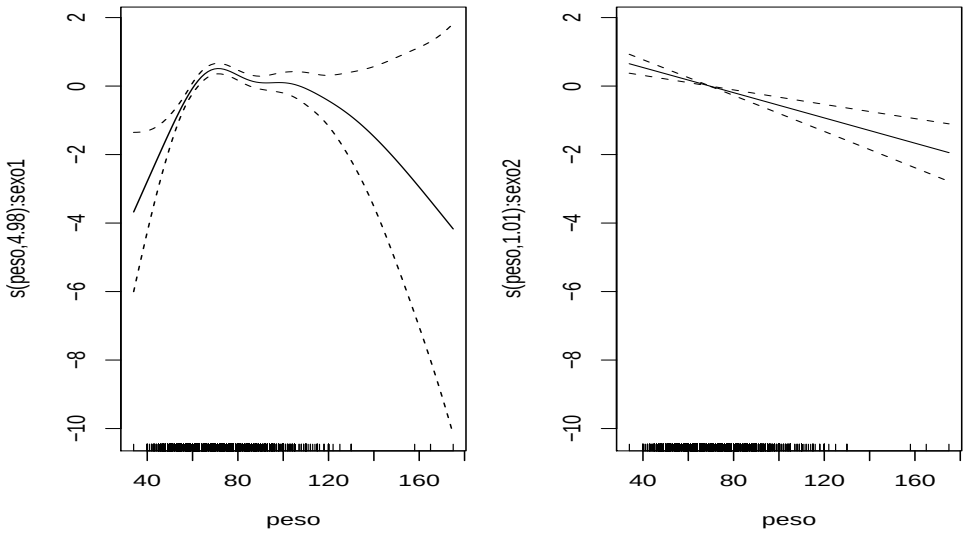
en los que la significación de los coeficientes es inexacta. Además, valores inusualmente altos o bajos de las variables independientes (*high leverage points*) o de la variable respuesta (*outliers*) pueden ser puntos influyentes que distorsionen la influencia de los parámetros. Nos vamos a centrar en las consecuencias de no satisfacer las hipótesis y en cómo detectarlas.

5.1. Error de especificación

Lo más importante es ver si el modelo está especificado correctamente. Esto tiene dos componentes: la relación funcional del modelo y la presencia de variables irrelevantes o la ausencia de variables importantes. Si la especificación del modelo no es correcta, esto da lugar a coeficientes sesgados. Puede ocurrir que la relación entre el logit y las variables explicativas no sea lineal o que la relación entre las variable explicativas sea multiplicativa y no aditiva, es decir, que haya interacción.
Si incluímos más variables de las necesarias, incrementarán los errores estándar de los coeficientes estimados, lo que reduce la eficiencia de las estimaciones, aunque no da lugar a sesgos. El incremento del error estándar depende de la correlación entre la variable innecesaria y las demás variables del modelo. Omitir variables relevantes dará lugar a sesgos en los coeficientes, lo cual es más importante que la ineficiencia.
Si la relación entre el logit y una variable explicativa no es lineal, entonces el incremento en una unidad de X no es constante, sino que depende del valor de X. Una manera de detectarlo es utilizar una técnica de suavizado para ver la relación entre el logit y la variable dependiente. Para ilustrarlo vamos a utilizar una función llamada `gam()`

```
library(mgcv)
logistica13.3=gam(g02~educa+edad2+s(peso,by=sexo),family=binomial,method="REML")
par(mfrow=c(1,2))
plot(logistica13.3)
```

La detección de la no aditividad hay que hacerla contrastando si las interacciones entre las variables son significativas o no. La correlación entre las variables explicativas (multicolinealidad) puede afectar también al sesgo de los coeficientes, un valor de la correlación mayor de 0.8 se considera alto y dará lugar a coeficientes elevados pero que son estadísticamente no significativos.



5.2. Análisis de residuos

Anteriormente vimos que en regresión logística podemos definir los *pearson y deviance residuals*. Cuando el tamaño de la muestra es grande, los *pearson residuals* se distribuyen como una $N(0,1)$, por lo tanto si se encuentran en $(-3,3)$ indican que el modelo ajusta bien los datos.

Las observaciones que a priori tienen influencia en los parámetros del modelo pueden ser identificadas por un alto valor del *leverage* h_i (en regresión simple $\hat{y}_j = \sum h_{ij} y_j$), este coeficiente captura la influencia de la observación y_j en el valor predicho $\hat{y}_j$. En regresión logística hay varias medidas de la influencia, las principales son:

1. ΔD de Hosmer-Lemeshow, mide la infuencia de cada observación en el ajuste del modelo.
2. $\Delta\beta$ de Pregibon que da una aproximación de lo que cambian los coeficientes al omitir cada observación.

Es conveniente hacer los siguientes gráficos:

- Residuos frente valores ajustados (o predictor lineal)
- ΔD_j frente a $\hat{p}_j$
- $\Delta\beta$ frente a $\hat{p}_j$

En el primer gráfico, los patrones de covariables que ajustan mal son aquellos con puntos en la parte superior derecha o izquierda del gráfico. Se considera un valor elevado aquel que está por encima de 3. En el gráfico de $\Delta\beta$, para que un patrón de variable tenga efecto en los coeficientes estimados, el valor ha de ser mayor que 1.
La función `influence.measures()`, nos da esta información.

6. Interpretación y presentación de resultados

Una vez que estamos conformes con que el modelo es adecuado, es posible extraer conclusiones. En general, vamos a utilizar los coeficientes del modelo para obtener odds-ratios que utilizaremos para estimar el riesgo relativo si los datos provienen de un estudio transversal.

El efecto principal es la variable continua peso, hay además una variable nominal dicotómica, sexo, dos variables politónicas, edad (edad2) y nivel de estudios (educa), además se incluye la interacción entre sexo y peso.

La siguiente tabla muestra los coeficientes del modelo:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.313414   0.330517   6.999 2.57e-12 ***
educa2       0.449794   0.104774   4.293 1.76e-05 ***
educa3       1.017172   0.111848   9.094  < 2e-16 ***
educa4       1.417463   0.113447  12.495  < 2e-16 ***
edad22      -0.316745   0.092114  -3.439 0.000585 ***
edad23      -1.064558   0.089488 -11.896  < 2e-16 ***
sexo2        0.166893   0.406530   0.411 0.681417
peso        -0.012573   0.003967  -3.170 0.001527 **
sexo2:peso  -0.011252   0.005626  -2.000 0.045488 *
```

Como vimos anteriormente los odds-ratios estimados se obtienen exponenciando los coeficientes y los extremos de los intervalos de confianza. Es importante recordar que cuando las variables explicativas son dicotómicas o politómicas, el nivel más bajo se toma como referencia.

```
 exp(logistica13$coeff)
(Intercept)      educa2      educa3      educa4      edad22      edad23
 10.1088729   1.5679890   2.7653644   4.1266397   0.7285166   0.3448804
      sexo2        peso  sexo2:peso
  1.1816273   0.9875053   0.9888107

exp(confint(logistica13))
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) 5.2863677 19.3239718
educa2      1.2765170  1.9250822
educa3      2.2208226  3.4433198
educa4      3.3046966  5.1561349
edad22      0.6074493  0.8717621
edad23      0.2889436  0.4104118
sexo2       0.5330294  2.6248014
peso        0.9798893  0.9952553
sexo2:peso  0.9779375  0.9997543
```

Comenzamos por comentar los resultados para la variable edad y educación:

En ambos casos el nivel de referencia es al que le corresponde el $OR = 1$. El OR estimado para personas de mediana edad es 0.73. La interpretación correcta es que el odd (o ventaja) a la hora de tener buena salud para una persona entre 30 y 44 años es 0.73 veces menor que el de un individuo similar (mismo sexo, educación y peso) pero más joven. Muchas veces se cae en la tentación de interpretar el OR como el riesgo relativo. Esto es cierto cuando el outcome es raro, en general se

Valor de la variable	Odds Ratio	I.C. 95 %
Edad		
Menor de 29	1.00	
30-44	0.73	0.61 0.87
45-64	0.34	0.29 0.41
Educación		
Bajo	1.00	
Medio-Bajo	1.56	1.27 1.92
Medio-Alto	2.76	2.22 3.44
Alto	4.12	3.3 5.15

considera raro cuando es menor del 15 %. En nuestro ejemplo significaría que la posibilidad de que un individuo tenga buena salud es pequeña, y en realidad es del 80 %. El intervalo de confianza es (0,61, 0,87) sugiere que el odds de buena salud para personas de mediana edad puede ser tan pequeño como 0.61 y tan grande como 0.87. Para personas de más edad el odds es menos de la mitad. En el caso de la educación todos los valores del odd están por encima del uno, lo que indica que tener más nivel de estudios es una ventaja a la hora de tener buena salud.
El caso para la interacción entre sexo y peso es diferente, para empezar no podemos interpretar las variables por separado. Empezamos por describir el efecto de un incremento del peso en la salud, ajustando por sexo; supongamos que queremos calcular el OR que corresponde a un incremento del peso en 30 kilos para los hombres o las mujeres cuando las demás variables se mantienen constantes (por esta razón no las incluímos al calcular el logit):

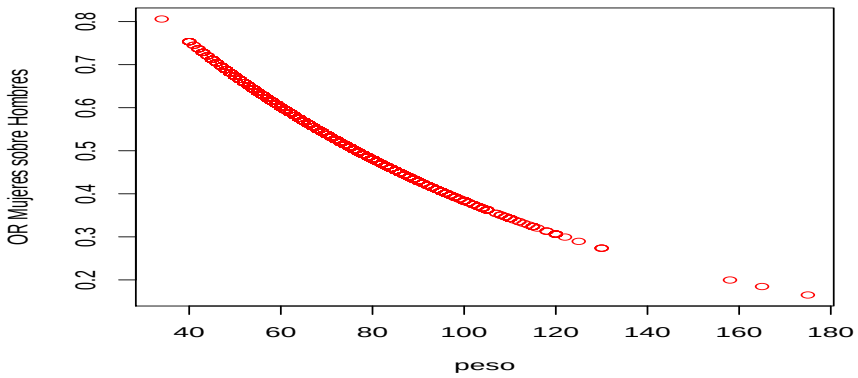
$$\begin{aligned} logit(sexo, peso) &= \hat{\beta}_0 + \hat{\beta}_1 sexo + \hat{\beta}_2 peso + \hat{\beta}_3 sexo \times peso \\ logit(sexo, peso + 30) &= \hat{\beta}_0 + \hat{\beta}_1 sexo + \hat{\beta}_2 (peso + 30) + \hat{\beta}_3 sexo \times (peso + 30) \\ logit(sexo, peso + 30) - logit(sexo, peso) &= \hat{\beta}_2 30 + \hat{\beta}_3 sexo \times 30 \\ OR &= exp\left(\hat{\beta}_2 30 + \hat{\beta}_3 sexo \times 30\right) \end{aligned}$$

```
exp(30*-0.0125)
[1] 0.6872893
exp(30*-0.0125-0.01257*30)
[1] 0.4893878
```

El odds ratio es mayor en los hombres que en las mujeres cuando el peso incrementa 30Kg. Si lo hacemos para distintos pesos veremos que cuanto mayor el el incremento del peso más distancia hay entre OR para hombres y mujeres. Queda describir el efecto del sexo ajustando por peso, en este caso

$$OR = exp\left(\hat{\beta}_1 + \hat{\beta}_3 peso\right)$$

```
OR_M=exp(0.16689-0.01125*peso)
plot(peso,OR_M,col=2,ylab="OR Mujeres sobre Hombres")
```



El gráfico muestra el OR de mujeres sobre hombres para distintos valores del peso, se ve claramente que las mujeres están en desventaja con respecto a los hombres (ya que los valores son menores que 1), y esa desventaja se acentúa con el peso.

7. Otros GLMs para datos con respuesta binaria

En los estudios tranversales las medidas de asociación más comunes son el OR y la razón de prevalencias PR (o RR). Recordemos que la PR se define en términos de cuantas veces es más probable que los individuos expuestos presenten la condición de interés que los no expuestos, mientras que el OR se define como el exceso o defecto de ventaja que tienen los individuos expuestos de presentar la condición de interés frente a no presentarla respecto a la ventaja de los individuos no expuestos de presentar la condición frente a no presentarla. Algunos estudios e expresan sus resultados en forma de OR. No hay nada intrínsicamente erróneo al hacer esto. Pero en los casos en los que la prevalencia es alta el OR sobrestima la PR. En muchas ocasiones el investigador olvida qué medida de asociación es el OR y hace interpretaciones de este tipo: *las plantas que reciben un cierto tratamiento tiene un riesgo de contraer la enfermedad cuatro veces mayor que el grupo no lo recibe*. La interpretación del OR como PR puede dar lugar a malos entendidos, tanto a nivel teorico como a nivel práctico. Además, en los casos en los que hay confusión y/o interacción, éstas dependen de la medida de efecto, de modo que controlar por una variable mediante OR no es lo mismo que hacerlo usando PR, lo que llevaría en algunos casos a que no estuvieramos controlando de forma adecuada.

En la literatura aparecen varias alternativas para el análisis de datos binarios en estudios transversales estimando directamente la PR:

1. **Transformar OR en RP**:
   Los resultados de ajustar un modelo de regresión logística, además de proporcionar los OR, proporcionan las PR, simplemente dividiendo las las probabilidades estimadas. Sin embargo, esto presenta dos dificultades: 1) Es necesario hacer un análisis estratificado, lo cual se complica bastante cuando hay varias variables politómicas o continuas y 2) el cálculo de los errores estándar no es inmediato, con lo cual es difícil obtener intervalos de confianza.

2. **Regresión de Breslow-Cox**:
   El modelo de Cox estima el RR de presentar una condición en aquellos individuos que han sufrido una exposición respecto a aquellos que no la han sufrido, ajustando por unas ciertas

variables teniendo en cuenta también el tiempo de exposición a partir de la función de riesgo acumulado:

$$h(t) = h_0(t)exp\left(\beta_1 X_1 + \ldots \beta_k X_k\right)$$

El principal inconveniente de este modelo es que asume que los errores del modelo siguen una distribución de Poisson, cuando en realidad son binomiales, esto dará lugar a que la varianza de los coeficientes sea sobrestimada

3. **Regression de Poisson**:
La regresión de Poisson se utiliza fundamentalmente para analizar estudios longitudinales donde la variable respuesta es el número de individuos con una característica que ocurren en un periodo de tiempo,

$$\log\left(\frac{n}{t}\right) = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k$$

En este contexto, la regresión de Poisson es equivalente a la regresión de Cox, sin embargo la regresión de Poisson da lugar en ciertas ocasiones a estimaciones de la prevalencia mayor que uno.

4. **GLM con vínculo binomial y familia logarítmica**:
Vimos que el modelo de regresión logística es un GLM con vínculo binomial y familia logit, necesitábamos transformar la probabilidad para que el predictor lineal pudiera tomar valores fuera del intervalo $(0, 1)$, otra posibilidad es tomar logaritmo a la probabilidad

$$log(p) = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k$$

por lo tanto,

$$\begin{aligned} log(p(X_1 = 1)) &= \beta_0 + \beta_1 + \ldots \beta_k X_k \\ log(p(X_1 = 0)) &= \beta_0 + \ldots \beta_k X_k \\ log(p(X_1 = 1)) - log(p(X_1 = 0)) &= \beta_1 \\ PR &= \frac{p(X_1 = 1)}{p(X_1 = 0)} \end{aligned}$$

con lo cual estamos estimando directamente la PR. Un inconveniente de esta metodología es que $\beta_0 + \beta_1 + \ldots \beta_k X_k$ tiene que ser menor que cero, esto da lugar a problemas de convergencia del algoritmo que estima los parámetros, especialmente si hay alguna covariable continua.

```
logistica14=glm(g02~educa+edad2+sexo*peso,family=binomial(link=log))
```

8. Ejemplo: Bajo peso al nacer

El bajo peso en los recién nacidos es una cuestión que ha preocupado desde hace tiempo. Esto es debido a que la tasa de mortalidad infaltil y la tasa de niños con discapacidad es muy alta para los nacidos con poco peso. El comportamiento de una mujer durante el embarazo (dieta, hábitos de tabaco, cuidados prenatales) puede alterar de forma importante la posibilidad de llevar al término el embarazo y, consecuentemente, la posibilidad de tener un bebé con peso normal o no. En la siguiente tabla se muestran algunas de las variables que se asocian con el bajo peso al nacer. El objetivo es identificar los factores de riesgo. En este estudio participaron 189 mujeres, 59 de las cuales tubieron bebés con bajo peso.

```
-----------------------------------------------------------------
low     Peso al nacer    (0 = Peso $>$ 2500g,
                          1 = Peso $<$ 2500g)
age     Edad de la madre en a\~nos
lwt     Peso de la madre en el \'ultimo ciclo menstrual
race    Raza (1 = Blance, 2 = Negra, 3 = Otra)
smoke   Fumadora durante el embarazo (1 = S\'i, 0 = No)
-----------------------------------------------------------------
```

El modelo

1. Empezamos haciendo una regresión logística univariante, es decir, los modelos continen una sola de las variables explicativas:

   ```
   glm1=glm(low~lwtfamily=binomial)
   glm2=glm(low~race,family=binomial)
   glm3=glm(low~smoke,family=binomial)
   glm4=glm(low~age,family=binomial)
   ```

   Comprueba que obtienes los siguientes resultados, donde la información de cada columna corresponde a: (1) coeficiente estimado, (2) error estándar estimado, (3) odds ratio estimado, (4) intervalo de confianza al 95 % para el odds ration, (5) el deviance del modelo, (6) el cambio en el deviance cuando contrastamos la hipótesis de que el modelo es igual a cero, (7) el p-valor para este test (usa la función anova()).

Variable	$\hat{\beta}$	$s\hat{e}.(\hat{\beta})$	Odds-ratio	95 % C.I.	Deviance	ΔD	p
Constant	-0.798	0.157			234.7		
age	-0.051	0.031	0.6	(0.32,1.11)	231.9	2.76	0.1
lwt	-0.014	0.006	0.87	(0.77,0.98)	228.7	5.98	0.02
race_2	0.845	0.463	2.33	(0.94,5.77)	229.7	5.01	0.08
race_3	0.636	0.347	1.89	(0.96,3.73)			
smoke	0.704	0.319	2.02	(1.08,3.78)	229.8	4.87	0.03

   ¿Qué variables incluirías en el modelo multivariante?

2. Basándonos en estos los modelos con una sola variables, ajustamos un modelos con todas las variables excepto `age`. Comprueba que los resultados que obtienes en este caso coinciden con los de la siguiente tabla:

Variable	$\hat{\beta}$	$s\hat{e}.(\hat{\beta})$	$\hat{\beta}/s\hat{e}.(\hat{\beta})$
Constant	-0.109	0.882	-0.124
lwt	-0.013	0.006	-2.101
race_2	1.29	0.511	2.525
race_3	0.97	0.412	2.354
smoke	1.06	0.378	2.802

3. Hemos de comprobar también las interacciones entre las variables del modelo anterior. Como hay sólo tres variables, es sencillo hacerlo. En los casos en los que haya un alto número de

variables es aconsejable comprobar sólo aquellas que tengan sentido científico o sobre las que tengamos alguna información.
Ajusta todas las posibles interacciones y usa la función anova() para comprobar si son significativas o no, debes obtener los siguientes resultados:

Interaction	Deviance	ΔD	df	p
Main effects only	215.01			
lwt×race	213.51	1.498	2	0.473
lwt ×smoke	214.09	0.918	1	0.338
race×smoke	212.61	2.403	2	0.301

4. Comprueba que los odds-ratio estimados para el modelo final son:

Variable	Odds-ratio	95 % C.I.
lwt	0.987	(0.974,0.999)
race_2	3.633	(1.335,9.888)
race_3	2.639	(1.176,5.921)
smoke	2.886	(1.375,6.058)

   ¿Cómo interpretarías esos resultados?

5. ¿Cuál sería el modelo seleccionado por el criterio AIC.?

6. Usa el test de Hosman-Lemeshow para ver la bondad de ajuste del modelo final (usando 10 grupos).

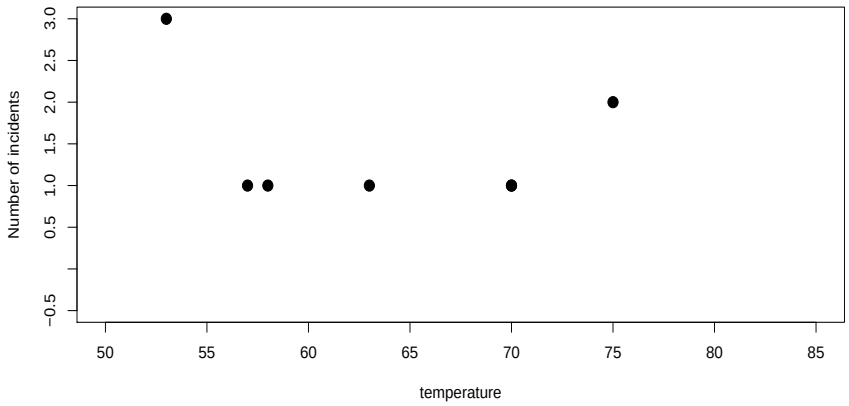
7. Usa los residuos deviance para realizar los gráficos de diagnóstico. Calcula el Delta Deviance, del Delta de Pearson. Comprueba que el punto de corte optimo es 0.339. ¿es un buen modelo predictor?

```
A=influence.measures(birth.glm)$infmat

dr=residuals(birth.glm,"deviance")
pr=predict(birth.glm,type="response")
plot(pr,dr)
plot(birth.glm,1)
plot(dr,A[,6])
plot(dr,A[,1]) #..............
```

9. Ejercicios

Desastre del transbordador Challenger

El 7 de Enero de 1986, la noche antes del despegue del transbordador, hubo una reunión en la que se discutio sobre la la temperatura mínima predicha para del día siguiente, 31^oF, y el efecto de la misma sobre el sello en las juntas de los O-rings. En la discusión utilizaron el siguiente gráfico en el que se muestra la relación entre la temperatura y el número de O-rings que sufrían problemas



El primer error que cometieron, fue el no dibujar los casos en los que no había incidentes, para saber cuál eran las temperaturas más propicias. La decisión final fue permitir el despegue en el que murieron 7 astronautas debido a la combustión de gas a través de un O-ring.

En el fichero de datos `challenger.txt` hay dos variables: temperatura y una variable dicotómica con valor 0 ó 1 que indica si hubo fallo del O-ring o no. Utlizando ese fichero responde a las siguientes preguntas:

1. Existe relación entre la temperatura y el fallo del O-ring?.
2. Haz un gráfico en el que vea la relación entre la temperatura y la probabilidad de fallo.
3. Interpreta el parámetro de la variable temperatura
4. Por debajo de qué temperatura la probabilidad de fallo es mayor de 0.5? (para responder a la pregunta utiliza el valor del parámetro estimado).
5. Si tú hubieras estado en esa reunión, cuál habría sido tu decisión?, en qué te basas para tomarla?

Ensayo toxicológico

El ensayo tiene como objetivo comparar tres tipos de insecticidas, aplicándolos a grupos de insectos en 17 dosis distintas (en escala logarítmica),pasadas 24 horas se contó el número de insectos muertos. Los datos están en el fichero `insectos.txt`:

- Codifica la variable `insecticida` como factor
- Ajusta un modelo con interacción. ¿Son significativos los coeficientes?
- A la vista de los resultados anteriores, recodifica `insecticida` para que tenga sólo dos niveles y vuleve a ajustar el modelo.
- ¿Cómo interpretas los parámetros?

Mortalidad por esfuerzo reproductivo

El objetivo de este estudio fue el analizar el coste reproductivo en plantas perennes. Se recogio información sobre el número de semillas producidas por un grupo de plantas, el tamaño de su raíz, y si al llegar al final de la estación estaban vivas o muertas. Los datos están en el ficheros `flores.txt`. Analiza si la probabilidad de muerte se ve influenciada por el número de semillas y el tamaño de la raíz.

Capítulo 3

Regresión Multinomial

Hasta ahora nos hemos centrado en el caso en el que la variable respuesta era dicotómica. Ahora nos centramos en el caso en el que la variable de interés tiene más de dos categorías, por simplicidad, vamos a ilustrar la metodología para el caso de tres categorías, ya que la generalización a más de tres es inmediata.

Supongamos que codificamos las tres categorías de la variable respuesta como 0, 1 y 2. En el caso de regresión logística, el logit es

$$log\left(\frac{Pr(Y=1)}{Pr(Y=0)}\right)$$

Ahora el modelo necesita dos funciones logit ya que tenemos tres categorías, y necesitamos decidir que categorías queremos comparar. Lo más general es utilizar $Y=0$ como referencia y formar logits comparándola con $Y=1$ y $Y=2$. Supongamos que tenemos k variables explicativas, entonces:

$$\begin{aligned} log\left(\frac{Pr(Y=1)}{Pr(Y=0)}\right) &= \beta_{10}+\beta_{11}X_1+\beta_{12}X_2+\ldots+\beta_{1k}X_k \\ log\left(\frac{Pr(Y=2)}{Pr(Y=0)}\right) &= \beta_{20}+\beta_{21}X_1+\beta_{22}X_2+\ldots+\beta_{2k}X_k \end{aligned}$$

con lo cual ahora tenemos el doble de coeficientes que en el caso de regresión logística.

Las probabilidades se calcularían como:

$$\begin{aligned} Pr(Y=0|X) &= \frac{1}{1+e^{g_1(X)}+e^{+g_2(X)}} \\ Pr(Y=1|X) &= \frac{e^{g_1(X)}}{1+e^{g_1(X)}+e^{+g_2(X)}} \\ Pr(Y=2|X) &= \frac{e^{g_2(X)}}{1+e^{g_1(X)}+e^{+g_2(X)}} \\ g_1(X) &= \beta_{10}+\beta_{11}X_1+\beta_{12}X_2+\ldots+\beta_{1k}X_k \\ g_2(X) &= \beta_{20}+\beta_{21}X_1+\beta_{22}X_2+\ldots+\beta_{2k}X_k \end{aligned}$$

La estimación de los parámetros se hace nuevamente mediante el método de máxima verosimilitud.

Los datos que vamos a utilzar en este caso provienen de un estudio cuyo objetivo es estudiar los factores que influyen en el conocimiento, actitud y comportamiento de mujeres hacia las mamografías. Son datos sobre 412 mujeres y las variables son:

Variable	Descripción	Valores
me	Frecuencia del chequeo	0=Nunca
		1=El último año
		2= Más de un año
symp	La mamografía no es	1=Muy de acuerdo
	necesaria a menos que	2=De acuerdo
	tengas algún síntoma	3=En desacuerdo
		4=Muy en desacuerdo
pb	Beneficio percibido	5-20
hist	Madre y hermana con	0=No
	historia de cancer de mama	1=Sí
bse	¿Te han enseñado a explorarte?	0=No
		1=Sí
detc	¿Es probable que una	0=No
	mamografía detecte un	1= Algo
	nuevo caso de cancer?	2= Muy

```
mamexp=read.table( "mamexp.txt",header=TRUE)
names(mamexp)
mamexp$symp=factor(mamexp$symp)
mamexp$hist=factor(mamexp$hist)
mamexp$bse=factor(mamexp$bse)
mamexp$dect=factor(mamexp$dect)

attach(mamexp)
table(me)

me
  0   1   2
234 104  74
```

1. El procedimiento multinom() en R

La sintaxis del comando `multinom` es similar a la de los comandos para regresión logística. Para utilizar esta función necesitamos cargar una nueva librería

```
library(nnet)
multi1=multinom(me~hist)
multi1
Coefficients:
  (Intercept)    hist1
1  -0.9509794 1.256531
2  -1.2504803 1.009384
```

$$
\begin{aligned}
logit(\text{Último año/Nunca}|\text{hist}) &= -0{,}95 + 1{,}25 \times \text{hist} \\
logit(\text{Más de un año/Nunca}|\text{hist}) &= -1{,}25 + 1{,}01 \times \text{hist}
\end{aligned}
$$

Por defecto, el nivel estamos tomando como referencia es el primer nivel. El output es similar al de regresión logística, sin embargo, tenemos dos grupos de parámetros, uno para cada logit.
En este caso, no tenemos un p-valor para los coeficientes, si queremos saber si la variable explicativa es significativa, utilizamos el test de la razón de verosimilitud

```
multi0=multinom(me~1)
anova(multi1,multi0)
```

2. Interpretación y significación de los parámetros

El OR en este caso sería

$$
OR_j(a,b) = \frac{Pr(Y=j|X=a)/Pr(Y=0|X=a)}{Pr(Y=j|X=b)/Pr(Y=0|X=b)} \quad j=0,1
$$

Donde j indica el nivel de la variable respuesta con el que estamos comparando y a y b son los valores que toman las variable explicativas. Si sólo tenemos una variable explicativa dicotómica que toma valores 0 y 1, entonces tendríamos $R_j(0,1)$. En el ejemplo:

```
table(me,hist)
   hist
me    0   1
  0 220  14
  1  85  19
  2  63  11
```

$$
\begin{aligned}
\widehat{OR}_1 &= \frac{19 \times 220}{85 \times 14} = 3{,}51 = exp(1{,}25) \\
\widehat{OR}_2 &= \frac{11 \times 220}{63 \times 14} = 2{,}74 = exp(1{,}009)
\end{aligned}
$$

La interpretación de los parámetros, los OR es similar al caso de regresión logística. La interpretación del efecto de la historia familiar en la frecuencia de mamografías es la siguiente:

- Mujeres con historia familiar de cancer de pecho tienen un odds de haberse hecho una mamografía en el último año que es 3.51 veces mayor que el de las mujeres sin historia familiar.
- Mujeres con historia familiar de cancer de pecho tienen un odds de haberse hecho una mamografía hace más de un año que es 2.7 veces mayor que el de las mujeres sin historia familiar.

En otras palabras, tener un familiar directo con cancer de pecho es un factor significativo en el uso de mamografías.

Si en vez de `hist` introducimos `dect` (la opinión de las mujeres sobre la habilidad de la mamografía para detectar el cancer):

```
 multi2=multinom(me~dect)
summary(multi2)
Call:
multinom(formula = me ~ dect)

Coefficients:
  (Intercept)      dect1     dect2
1   -2.565201  0.7062589 2.1062453
2   -1.178617 -0.3926042 0.1977986

Std. Errors:
  (Intercept)    dect1     dect2
1    1.037867 1.083278 1.0464712
2    0.571762 0.634349 0.5936115

Residual Deviance: 778.4011
AIC: 790.4011

exp(coef(multi2))
  (Intercept)     dect1    dect2
1  0.07690374 2.0263961 8.217330
2  0.30770391 0.6752959 1.218717
```

Si nos fijamos en los errores estándar, vemos que todos los intervalos de confianza contendrían al 0 excepto uno, ese OR se interpretaría de la siguiente forma: El odds de hacerse una mamografía dentro del último año entre las mujeres que piensan que es muy probable que la mamografía detecte un nuevo caso de cáncer es 8.22 veces mayor que el el odds entre las mujeres que piensan que no es probable que la mamografía lo detecte.

Si la variable explicativa es continua, aparece un parámetro para cada logit y representa el OR para un cambio en una unidad de la variable continua.

$$log\left(\frac{Pr(Y=1|X=x_1)}{Pr(Y=0|X=x_1)}\right) - log\left(\frac{Pt(Y=1|X=x_0)}{Pt(Y=0|X=x_0)}\right) = (\alpha_{11}+\beta_{11}x_1)-(\alpha_{11}+\beta_{11}x_0) = \beta_{11}(x_1-x_0)$$

3. Selección de variables

Los pasos a seguir serían los mismos que para regresión logística, comenzaríamos viendo si cada una de las variables por separado son significativas. Si tenemos los modelos

```
multi3=multinom(me~symp)
multi4=multinom(me~bse)
multi5=multinom(me~pb)
```

¿Cuáles son significativas?

```
multi6=multinom( me~hist+dect+symp+bse+pb)
summary(multi6)
Call:
multinom(formula = me ~ hist + dect + symp + bse + pb)
```

```
Coefficients:
  (Intercept)    hist1       dect1      dect2      symp2     symp3    symp4
1  -2.9990607 1.366274  0.01700706  0.9041717  0.1101408 1.9249158 2.457230
2  -0.9860258 1.065446 -0.92448094 -0.6906153 -0.2901346 0.8172916 1.132224
      bse1         pb
1 1.291772 -0.2194421
2 1.052183 -0.1482079

Std. Errors:
  (Intercept)     hist1     dect1     dect2     symp2     symp3    symp4
1    1.539293 0.4375239 1.1619394 1.1268643 0.9228324 0.7776651 0.775400
2    1.111829 0.4593986 0.7137332 0.6871025 0.6440622 0.5397872 0.547666
       bse1         pb
1 0.5299080 0.07551477
2 0.5149934 0.07636886

Residual Deviance: 693.9019
AIC: 729.902
```

Los dos coeficientes estimados para `symp` que estiman el log-odds de las que están de acuerdo) frente al grupo de referencia (que son las que están muy de acuerdo no parecen ser significativos, por lo que podríamos simplificar la codificación de la variable agrupando los dos primeros niveles. Vemos además que ninguno de los cuatro coeficientes de `detc` parece ser significativo, por lo que decidimos ajustar el modelo sin esta variable y nos fijamos en dos cosas: 1) si los parámetros de las demás variables cambian (con lo cual estamos comprobando si es una variable de confusión) y 2) qué nos dice el test de la razón de verosimilitud.

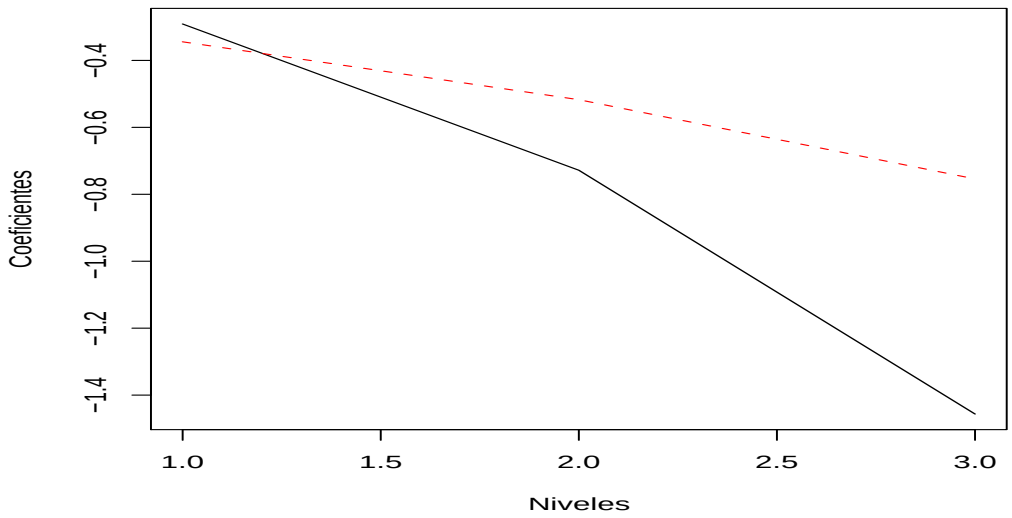
```
multi7=multinom( me~hist+symp+bse+pb)
coef(multi6)
coef(multi7)
anova(multi6,multi7)
```

En cuanto a la variable `pb` vamos a codificarla para que sea más sencillo comprobar si la relación es lineal (ya que la función `gam` que usamos en el capítulo anterior no permite usar la familia multinomial):

```
benef=pb
benef[benef<=5]=0
benef[benef>5 & benef<=7]=1
benef[benef>7 & benef<=9]=2
benef[benef>9]=3
benef=factor(benef)
multi8=multinom( me~hist+symp+bse+benef)
```

hacemos un gráfico de los coeficientes para los dos logits:

```
coefi=coef(multi8)
coefi=coefi[,7:9]
matplot(t(coefi),type="l",xlab="Niveles",ylab="Coeficientes")
```



El gráfico confirma que la relación es lineal, con lo cual tendríamos la opción de dejar la variable continua o utilizar la categorizada.

Ahora tenemos que comprobar si las interacciones son significativas, introduciendolas una a una en el modelo. En este caso, ninguna de las diez posibles intercacciones es significtiva.

Los OR estimados:

```
exp(coef(multi7))
  (Intercept)    hist1     symp2    symp3     symp4     bse1        pb
1   0.1128529 3.950369 1.1653385 8.049259 13.721009 3.548813 0.7785954
2   0.1766115 3.051151 0.7353411 2.320669  3.135382 2.668927 0.8710831
```

Los OR estimados muestran que hay un aumento en el odds de hacerse una mamografía para ambas categorías de frecuencia frente a la categoría "nunca" (excepto para la variable beneficio, ya que valores más altos de esta variable corresponden a una menor creencia en el beneficio de este tipo de diagnóstico). Las estimaciones de los OR son mayores para reciente"frente a "nunca", lo que indica que los dos grupos perciben de forma diferente el valor de la mamografía.

- En el caso de la variable `síntoma`, las mujeres que no están de acuerdo con la frase: "*La mamografía no es necesaria a menos que aparezca algún síntoma*"tienen un odds 13.65 veces mayor de haberse hecho una mamografía recientemente, y un odds 3.2 veces mayor de habérsela hacho no tan recientemente, comparado con mujeres que no están en desacuerdo con esa frase.

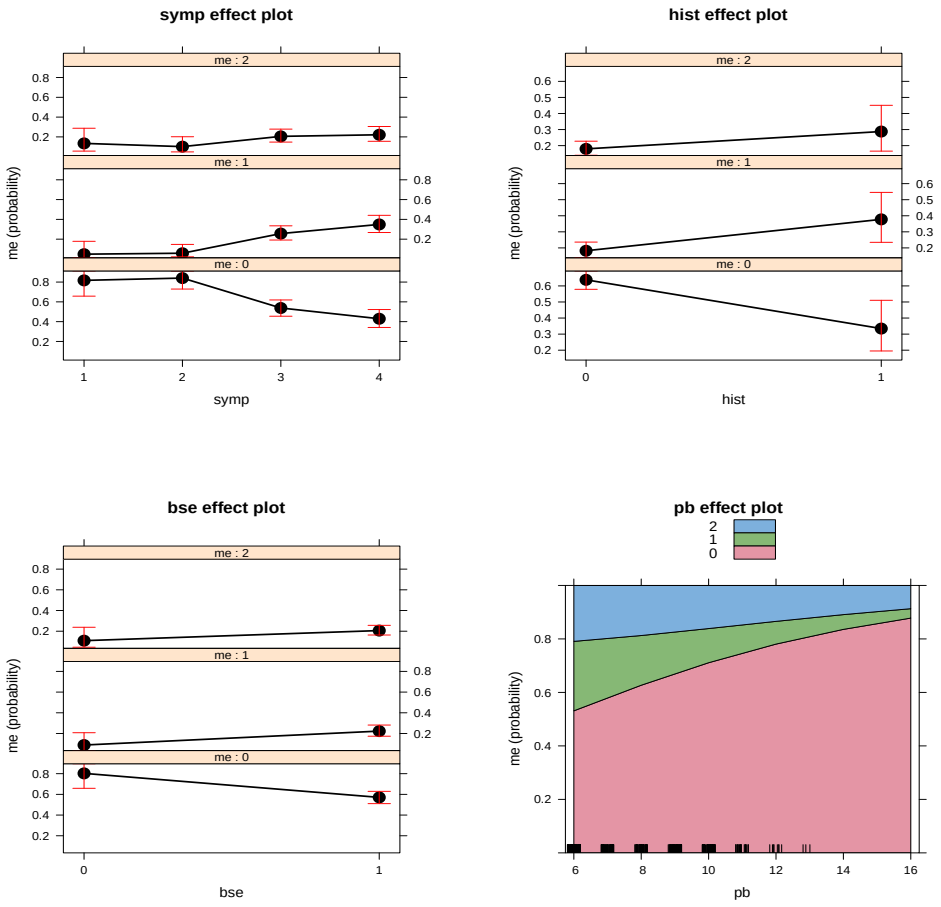
- Las mujeres con historia familiar de cáncer tienen un odds 3.9 veces mayor de haber hecho uso de la mamografía recientemente y 3 veces mayor en el caso no tan reciente comparado con mujeres que no tienen un familiar próximo con cáncer.

- Haber aprendido a explorarse el pecho es un factor significtivo a la hora de haberse hecho una mamografía en el último año, pero no lo es a la hora de hacerlo no tan recientemente (ver los errores estándar).

- La variable `pb` tiene un $OR < 1$ ya que valores más altos indica creencia de menor beneficio, además, cuanto más aumenta la desconfianza más disminuye el odds de someterse a una mamografía ya sea reciente o no recientemente.

Podemos ver como cambia la probabilidad de frecuencia del chequeo para las distintas variables de forma gráfica:

```
plot(effect("symp",multi7))
plot(effect("hist",multi7))
plot(effect("bse",multi7))
plot(effect("pb",multi7),style="stacked")
```



Como ya mencionamos anteriormente, cuando utilizamos un modelo de regresión multinomial, la interpretación de los coeficientes es más complicada, ya que tenemos más de un OR para cada variable. Sin embargo, utilizar una respuesta multinomial da una información más completa sobre el proceso que estamos estudiando. Por ejemplo, si hubiéramos codificado la variable respuesta como *alguna vez* frente a *nunca* habríamos perdido la información que indica que los odds son mayores para el uso frecuente.
Desde el punto de vista estadístico, no es conveniente pasar a una variable dicotómica a menos que los coeficientes en los distintos logits no se puedan considerar significativamente distintos.

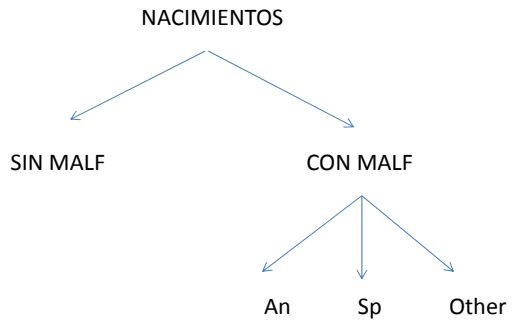
4. Ejemplo

Los datos que vamos a analizar fueron recogidos por Lowe, Roberts y Lloyd (1971) y corresponden a nacimientos de niños vivos con malformaciones del sistema nervioso central en Gales:

```
malf=read.table("malformaciones.txt",header=TRUE)
malf
```

	Area	NoCNS	An	Sp	Other	Water	Work
1	Cardiff	4091	5	9	5	110	NonManual
2	Newport	1515	1	7	0	100	NonManual
3	Swansea	2394	9	5	0	95	NonManual
4	GlamorganE	3163	9	14	3	42	NonManual
5	GlamorganW	1979	5	10	1	39	NonManual
6	GlamorganC	4838	11	12	2	161	NonManual
7	MonmouthV	2362	6	8	4	83	NonManual
8	MonmouthOther	1604	3	6	0	122	NonManual
9	Cardiff	9424	31	33	14	110	Manual
10	Newport	4610	3	15	6	100	Manual
11	Swansea	5526	19	30	4	95	Manual
12	GlamorganE	13217	55	71	19	42	Manual
13	GlamorganW	8195	30	44	10	39	Manual
14	GlamorganC	7803	25	28	12	161	Manual
15	MonmouthV	9962	36	37	13	83	Manual
16	MonmouthOther	3172	8	13	3	122	Manual

`NoCNS` indica el número de casos sin malformaciones, `An`, casos de anencefalia, `Sp`, espina bífida, y `Other`, otro tipo de malformaciones. `Water` es la dureza del agua, `Area` indica los diferentes municipios, y `Work` indica el tipo de trabajo realizado por los padres. Podríamos analizar los datos mediante un modelo multinomial con cuatro categorías, sin embargo, casi todos los casos corresponden a niños sin malformaciones. En este tipo de casos es mejor considerar un modelo de respuesta jerárquica:



De modo que primero ajustamos un modelo binomial agrupando todos los tipos de malformaciones, y después, condicionado ha que hay una malformación, ajustamos un modelo multinomial con tres categorías.

```
malf$CNS <- malf$An+malf$Sp+malf$Other
attach(malf)
```

Vemos si las variables son significativas:

```
bino1=glm(cbind(CNS,NoCNS) ~ Water,family=binomial)
anova(bino1,test="Chi")

      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     15     41.047
Water  1   15.689        14     25.359 7.466e-05 ***

bino2=glm(cbind(CNS,NoCNS) ~ Area,family=binomial)
anova(bino2,test="Chi")
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    15     41.047
Area  7   24.509         8     16.539 0.0009269 ***

bino3=glm(cbind(CNS,NoCNS) ~ Work,family=binomial)
anova(bino3,test="Chi")


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    15     41.047
Work  1   17.139        14     23.909 3.475e-05 ***
```

Ahora, las incluímos todas en el modelo y comparamos los coeficientes con los que se obtienen con los modelos con una sola variables:

```
bino4=glm(cbind(CNS,NoCNS) ~ Water+Area+Work,family=binomial)
bino4
bino1
bino2
```

Vemos que los coefficientes de `Area` y `Water` cambian mucho por lo que esas variable están confundidas, para ver con cual de las dos nos quedamos vemos cual es significativa al ponerlas juntas:

```
bino5=glm(cbind(CNS,NoCNS) ~ Area+Work,family=binomial)
bino6=glm(cbind(CNS,NoCNS) ~ Water+Work,family=binomial)
anova(bino5,bino4,test="Chi")
anova(bino6,bino4,test="Chi")
```

A continuación comprobamos si hay interacción entre `Water` y `Work`:

```
bino7=glm(cbind(CNS,NoCNS) ~ Water*Work,family=binomial)
anova(bino6,bino7,test="Chi")
summary((bino6)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.4325803  0.0897889 -49.367  < 2e-16 ***
Water         -0.0032644  0.0009684  -3.371 0.000749 ***
WorkNonManual -0.3390577  0.0970943  -3.492 0.000479 ***
```

Dado que $\exp(-0{,}3390577) = 0{,}71$, los nacidos de trabajadores no manuales tienen un 29 % menos de posibilidad de nacer con una malformación del sistema nervioso central. Un incremento de la dureza del agua de 80 unidades: $exp(-0{,}0032644 * 80) = 0{,}77$, supondría un descenso del 23 % de la posibilidad de nacer con malformaciones.

Ahora consideramos el modelo multinomial para los tres tipos de malformaciones:

```
mult1=multinom(cbind(An,Sp,Other) ~ Water)
mult2=multinom(cbind(An,Sp,Other) ~ Work)
mult0=multinom(cbind(An,Sp,Other) ~ 1)
anova(mult1,mult0)
anova(mult2,mult0)
```

Por lo que concluimos que la dureza del agua y la profesión de los padres tiene relación con la probabilidad de una malformación al nacer, pero no tienen efecto en el tipo de malformación.

Si hubiéramos usado un modelo multinomial desde el principio:

```
multi=multinom(cbind(NoCNS,An,Sp,Other) ~ Water + Work)
```

ambas variables son significativas, pero no hubiérmos descubierto que no afectan al tipo de malformación.

5. Ejercicios

Intención de voto en USA

Los datos del archivo `elections.txt` corresponden a un subconjunto de datos pertenecientes a un estudio de intención de voto en la elecciones americanas de 1996 (Faraway, 2006). Por simplicidad consideramos sólo las variables `age` (edad, variable continua), `educa` (variable categórica, con dos categorías *Low*, si no tienen estudios universitarios, y *High*, si tienen estudios universitarios), *income* (ingresos en miles de dólares, variable continua). La variable respuesta es `party`, intención de voto, con tres categorías: *Democrat* (que es la categoría base), *Independent* y *Republican*. Bus el modelo que mejor predice la intención de voto.

Caimanes

Los datos del archivo `gator.txt` corresponden a 219 caimanes capturados en cuatro lagos de Florida. La variable respuesta es el tipo de comida encontrada en el el estómago de los caimanes y tiene 5 categorías:

1. Pescado
2. invertebrados
3. Reptiles
4. Pájaros
5. Otros

Las variables explicativas son:

1. El lago donde se capturaron: Hancock,Oklawaha, Trafford, George
2. Sexo
3. Tamaño: $\leq 2{,}3m$ ó $\geq 2{,}3m$.

Contesta a las siguientes preguntas:

1. ¿Influye el sexo, el lago de captura y el tamano en el contenido del estómago?
2. ¿Hay interacción entre las variables explicativas?
3. Obten $P(food = fish|gender = Male, size \geq 2{,}3m, lake = Trafford)$

Capítulo 4

Regresión para datos ordinales

¿Cuándo unos datos son ordinales? Son datos que toman valores que tienen un orden natural, para los cuales los intervalos entre valores no tiene por qué tener un significado. Un ejemplo de datos ordinales: estado de salud: Excelente, muy bueno, bueno, regular, malo; encuestas de satisfacción: Mucho, Poco, Nada, etc.

Vamos a trabajar con datos de un estudio cuyo interés es buscar los factores que influyen en un graduado a la hora de decidir va a realizar estudios de postgrado o no. A 400 estudiantes de grado se les preguntó si era: nada probable, algo problable, o muy probable que realizaran estudios de postgrado. Se recogieron datos sobre el nivel de educación de sus padres (0=bajo, 1=alto), si el centro en el que estaban estudiando era público o privado, y su nota de expediente académico.
Cuando nos encontramos con este tipo de datos podemos tomar varias decisiones:

1. Recodificar los datos, pasar a una variable dicotómica y utilizar los modelos que vimos en el capítulo 2.
2. Analizar cada unas de las categorías por separado:Muy probable/Algo probable/Nada probable

Estos modelos son fáciles de interpretar, pero son estadísticamente ineficientes ya que no utilizan toda la información que proporcionan los datos. La regresión logístistica ordinal combina los modelos anteriores en uno solo.

1. Modelo de odds proporcionales

Lo que este modelo hace es trasformar la escala ordinal a varios puntos de corte binarios, el número de puntos de corte es siempre menor que el número de categrorías. Por ejemplo, si hay tres categorías, habría dos puntos de corte. Es lógico pensar estos puntos de corte como umbrales que necesitamos cruzar para pasar una categoría a la siguiente (más alta) categoría. Los dos umbrales en nuestro caso nos haría pasar de nada probable a algo probable, y de algo probable a muy probable. En este modelo no estimamos la probabilidad de algo, sino la probabilidad de observar un resultado o algo menor.

En general, supongamos que la variable respuesta Y toma valores de $1, \ldots, K$ (en el ejemplo, de 1 a 3), definimos

$$\tilde{p}_k = Pt(Y \leq k|X), \quad \text{para } k = 1 \ldots K-1$$

Las p_k son las **probabilidades acumuladas** y k los **puntos de corte**.

En el modelo de odds-proporcionales se comparan la probabilidad de obtener una respuesta $<= k$ con la de tener una respuesta $> k$

$$\begin{aligned} logit(\tilde{p}_1) &= log\left(\frac{p_1}{1-p_1}\right) = \alpha_1 - \beta_1 X_1 + \ldots - \beta_m X_m \\ logit(\tilde{p}_2) &= log\left(\frac{p_1+p_2}{1-p_1-p_2}\right) = \alpha_2 - \beta_1 X_1 + \ldots - \beta_m X_m \\ logit(\tilde{p}_k) &= log\left(\frac{p_1+p_2+\ldots+p_k}{1-p_1-p_2-\ldots-p_k}\right) = \alpha_k - \beta_1 X_1 + \ldots - \beta_m X_m \\ 1 &= p_1+p_2+\ldots+p_k+p_{k+1} \end{aligned}$$

El modelo asume que los coeficientes que acompañan a las covariables son iguales, y la única diferencia es en la ordenada en el origen, α_i.

¿Por qué ponemos $-\beta_i$ en ver de $+\beta_i$?: por analogía con regresión logística. En regresión logistica tenemos una variable W que toma valores 0 ó 1:

$$p = Pr(W = 1|x) \quad logit(p) = \beta_1 x$$

Si colapsamos las categorías de Y (en nuestro caso 1,2,3) en dos:

$$\underbrace{1,2}_{0} \quad \underbrace{3}_{1}$$

Entonces:

$$Pr(Y \leq 2|x) = Pr(W = 0|x) = 1 - p \Rightarrow logit(Pr(Y \leq 2|x) = log\left(\frac{1-p}{p}\right) = -\beta_1 x$$

Por eso, los β_i se interpretan como el cambio en el log-odds de categoría más alta frente a otra más baja, asociado con un incremento en una unidad de X_i, manteniendo las demás variables constantes. Este cambio es el mismo para cualquiera de las categorías ya que, por ejemplo:

$$\begin{aligned} logit(\tilde{p}_k|X_i = a) &= \alpha_k - \beta_1 X_1 - \ldots - \beta_i a + \ldots + -\beta_m X_m \\ logit(\tilde{p}_k|X_i = a+1) &= \alpha_2 - \beta_1 X_1 - \ldots - \beta_i (a+1) + \ldots + -\beta_m X_m \\ logit(\tilde{p}_k|X_i = a+1) - logit(\tilde{p}_k|X_i = a) &= -\beta_i \end{aligned}$$

y esto es cierto para cualquier categoría k.

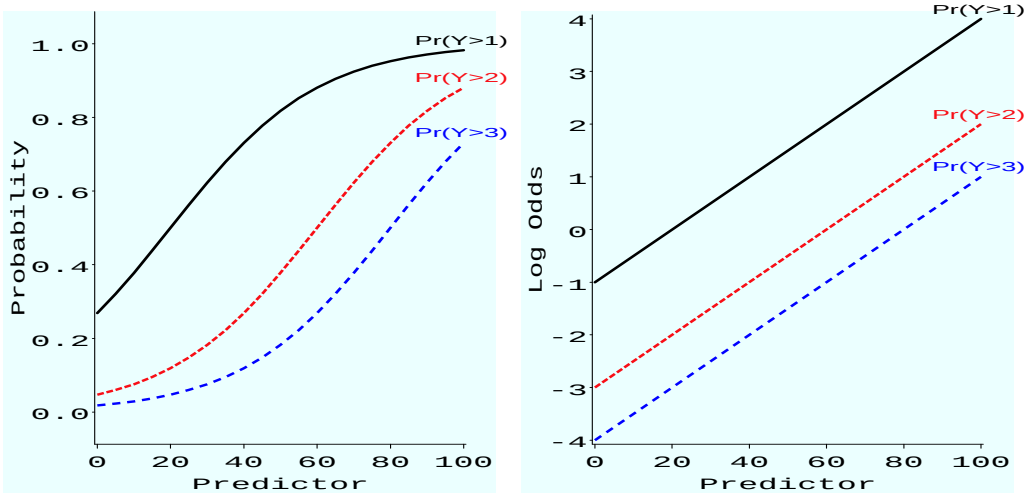
El modelo se llama de odds-proporcionales ya que el odds para cualquier k es proporcional a

$$= e^{-\beta_1 X_1 + \ldots - \beta_m X_p}$$

Si una cierta combinación de variables explicativas duplica el odds de estar en la categoría 1, también dobla el odds de estar en la categoría 2, etc. Los odds solo se diferencian en el punto de partida (cuando todas las covariables son cero), es decir, en e^{α_i}.

Los modelos de regresión ordinal se puede interpretar a partir de variables latentes. Supongamos que la variable que observamos, Y, proviene de una variable no observada Z, la cual hemos discretizado:

$$c_{j-1} < Z \leq c_j \Rightarrow Y = j \quad j = 1, \ldots, k+1$$



La variable Z, depende de covariables:

$$Z = \beta_1 X_1 + \ldots + \beta_m X_m + \epsilon$$

donde ϵ tiene una determinada función de distribución F, de modo que:

$$Pr(Y \leq j) = Pr(Z \leq \alpha_j) = F(\alpha_j - \beta_1 X_1 + \ldots - \beta_m X_m)$$

y

$$F^{-1}(p_i + \ldots p_j) = \alpha_j - \beta_1 X_1 + \ldots - \beta_m X_m$$

Si $F^{-1} = logit$, tenemos el modelo de odds-proporcionales.

Es importante darse cuenta de que para $k = 1$ estamos en el caso de una multinomial.
Los OR se calcularían:

$$\begin{aligned} log\left(\frac{Pt(Y \leq k|X = x_1)}{Pt(Y > k|X = x_1)}\right) - \left(\frac{Pt(Y \leq k|X = x_0)}{Pt(Y > k|X = x_0)}\right) &= (\alpha_k - \beta x_1) - (\alpha_k - \beta x_0) \\ &= \beta_{11}(x_1 - x_0) \Rightarrow OR = e^{-\beta_{11}(x_1 - x_0)} \end{aligned}$$

En general un coeficiente positivo indica un incremento en la posibilidad de que un individuo con valores más altos en la variable explicativa esté en una categoría superior de intención de estudios de postgrado, y un coeficiente negativo indica un descenso la en la posibilidad de que un individuo con valores más altos en la variable explicativa esté en una categoría superior intención de estudios de postgrado.

2. La función polr en R

Esta función forma parte de la librería MASSS. Es **importante** tener en cuenta que esta función usa una parametrización que implica que los odds ratio son de estar en una categoría superior

Vemos si las variables individuales son significativas en el modelo univariante:

```
library(MASS)
ord1=polr(apply~pared,postgrado)
summary(ord1)
Coefficients:
       Value Std. Error t value
pared1 1.127     0.2634    4.28

Intercepts:
                             Value   Std. Error t value
unlikely|somewhat likely     0.3768  0.1103     3.4152
somewhat likely|very likely  2.4519  0.1826    13.4302
```

Los parámetros se pueden interpretar de manera similar al caso de regresión logística, pero ahora hay dos transiciones en vez de una (que sería el caso de una variable dicotómica). El parámetro para `pared` es positivo lo que indica que el hecho de que los padres tengan estudios aumenta la posibilidad de hacer estudios de postgrado

```
ord2=polr(apply~public,postgrado)
summary(ord2)

ord3=polr(apply~gpa,postgrado)
summary(ord3)


ord0=polr(apply~1,postgrado)

anova(ord0,ord1)
anova(ord0,ord2)
anova(ord0,ord3)
```

El parámetro de `gpa` también es positivo, lo que indica que la nota media de expediente predispone a elegir los estudios, y el hecho de que el alumno esté en una Universidad pública o privada no afecta a la elección de estudios de postgrado. Si introducimos las variables juntas:

```
ord4=polr(apply~pared+gpa,postgrado)
```

Antes de seguir interpretando el modelo, hemos de comprobar que se satisface las condiciones del modelo de odds proporcionales, es decir, que si creáramos dos variable dicotómicas correspodientes a las dos transiciones y a justáramos las variables, los coeficientes en ambos casos sería muy similares. Para comprobarlo podemos utilizar la librería `VGAM`:

```
apply2=ordered(postgrado$apply)
m0 <- vglm(apply2~pared+gpa,family=cumulative(parallel=T),
data=postgrado)
m1 <- vglm(apply2~pared+gpa,family=cumulative(parallel=F),
data=postgrado)

test.po <- 2*logLik(m1)-2*logLik(m0)
df.po <- length(coef(m1))-length(coef(m0))
test.po
df.po
1-pchisq(test.po,df=df.po)
```

Por lo que podemos aceptar la hipótesis de odds proporcionales.

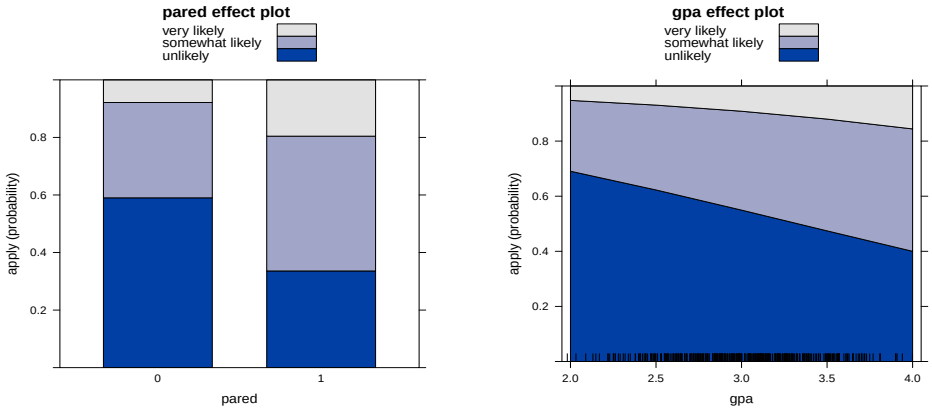
Calculamos los odds y sus intervalos de confianza:

```
summary(ord4)
Coefficients:
        Value Std. Error t value
pared1 1.0457     0.2656   3.937
gpa    0.6042     0.2539   2.379

Intercepts:
                              Value  Std. Error t value
unlikely|somewhat likely    2.1763 0.7671     2.8370
somewhat likely|very likely 4.2716 0.7922     5.3924

ci=confint(ord4)
exp(cbind(OR=coef(ord4),ci))
             OR    2.5 %   97.5 %
pared1 2.845412 1.693056 4.806454
gpa    1.829873 1.115180 3.022320
```

Entonces, la posibilidad de pasar de *nada probable* a *algo o muy probable*, es 2.8 veces mayor para los alumnos cuyos padres tienen estudios que para los que no los tienen (también sería 2.8 veces más probable para de *nada o algo probable* a *muy probable*). Algo similar ocurre con cada unidad que aumenta la nota del expediente. Las ordenadas en el origen correspondería a los umbrales de corte de la variable latente.

Podemos obtener y dibujar las probabilidades ajustadas:

```
plot(effect("pared", ord4, style='stack'))
plot(effect("gpa", ord4, style='stack'))

pr=predict(ord4,postgrado,type="p")
```



No hay métodos equivalentes a los vistos anteriormente para la diagnosis del modelo, con lo cual tendríamos que hacer las regresiones logísticas por separado para comprobar si se cumplen las hipótesis.

3. Ejercicios

3.1. Artritis

Los datos corresponden a un estudio sobre el efecto de un determinado tratamiento para la artritis. La variable respuesta `Improved` tiene tres categorías: None, Some, Marked, y hay tres variable explicativas: `Treatment` (Placebo, Tratado), `age`, `sex` (Mujer Hombre). Para obtener los datos:

```
library(vcd)
data("Arthritis")
```

¿Qué variables son significativas?, ¿hay interacciones?, ¿cómo interpretas los coeficientes?

Capítulo 5

Regresión de Poisson

Gran cantidad de datos son recogidos por científicos, médicos, o empresas como datos *de conteo*. En general, este tipo de datos aparecen en cuantro formatos distintos:

- Datos como *frecuencias*, donde contamos cuantas veces ocurre algo, pero no sabemos cuantas veces no ocurre.
- Datos como *proporciones*, donde ambos, el número de veces que ocurre algo, y el tamaño total del grupo es conocido.
- Datos por *categorías*, donde la varible cuenta cuántos individuos hay en cada nivel de una variable categórica.
- Datos *binarios*, donde recogemos la presencia o ausencia de una caracteríastica.

Los datos como frecuencias y datos binarios han sido tratados en el capítulo de regresión logística.

Hay cuatro razones por las que **es erroneo utlizar un modelo de regresión normal para datos de conteo** :

1. Puede dar lugar a predicciones negativas.
2. La varianza de la variable respuesta no es independiente de la media.
3. Los errores no siguen una distribución Normal.
4. Los ceros que aparecen en la variable respuesta dan problemas a la hora de transformar la variables.

Sin embargo, si los datos son elevados, es posible utilizar la distribución Normal.

1. La distribución de Poisson

La principal diferencia entre la distribución de Poisson y la Binomial, es que, aunque ambas cuentan el número de veces que ocurre algo, en la distribución de Poisson no sabesmos cuantas veces no ocurrio, y en la Binomial sí lo sabemos.

Supongamos que estamos haciendo un estudio sobre cuantas larvas de insectos hay en ciertos árboles, los datos de los que disponemos corresponden al número de larvas por hoja (y). Habrá hojas

que no tengan ninguna, y otras que tenga hasta 5 ó 6. Si el número medio de larvas por hoja es μ, la probabilidad de observar x larvas por hoja viene dada por:

$$P(y) = \frac{e^{\mu}\mu^{y}}{y!}$$

y μ sería el número de larvas, por la probabilidad de que una hoja tenga una larva. Implícitamente, lo que estamos haciendo es una aproximación:

$$\mu = np$$

para n grande y p pequeño. Es decir, que una distribución de Poisson se obtiene a partir de una Binomial con p pequeño y n grande y donde $\mu = np$. Veremos más adelante que si utilizamos una regresión logística con datos agrupados obtenemos los mismo coeficientes que si utilizamos una regresión de Poisson, pero la interpretación de ambas sólo es similar cuando la prevalencia es baja.

En general, estamos interesados en relación la media de nuestra variable respuesta con las covariables, es decir:

$$\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_k$$

sin embargo, la parte izquierda de la ecuación puede tomar cualquier valor, mientras que la derecha sólo toma valores positivos. Una solución inmediata es tomar logaritmos:

$$log(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_k$$

de modo que es una transformación de la media la que tiene una relación lineal con las covariables, y e^{β_i} representa un efector multiplicador del i-ésimo predictor sobre la media, ya que al incrementar X_i en una unidad, la media queda multiplicada por un factor e^{β_i}. Una de las ventajas de utilizar este tipo de modelo es que, en general, en el caso de datos de conteo, los predictores son multiplicativos y no aditivos, es decir, observamos conteos pequeños para efectos pequeños y conteos grandes para efectos grandes (o viceversa).

De nuevo estamos en el entorno de un glm con función link=logaritmo, y utilizaremos máxima verosimilitud para estimar los parámetros. La función soporte viene dada por:

$$logL(\beta_0, \ldots \beta_k) = \sum (y_i log(\mu_i) - \mu_i) = \sum (y_i(\beta_0 + \ldots + X_k\beta_k)_i - exp(\beta_0 + \ldots + X_k\beta_k)$$

Si tomamos derivadas con respeto a los β_i e igualando a cero obtenemos:

$$\boldsymbol{X'y} = \boldsymbol{X'}exp(\boldsymbol{X\beta})$$

donde $\boldsymbol{X}$ es la matriz del modelo. Estas ecuaciones sólo se pueden resolver de forma iterativa mediante el algoritmo llamado *iterative reweighted least squales* (ver anexo a estos apuntes).

En cuanto a las medidas de bondad de ajuste:

1. Test de la razón de verosimilitud. Sin embargo, hay que tener cuidado con este test de bondad de ajuste, ya que sólo es válido cuando el número de casos en la mayoría de los estratos es > 5, cosa que a veces no ocurre.

2. Para este tipo de datos utilizamos los residuos estandarizados:

   $$r_i = \frac{\mu_{iobs} - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

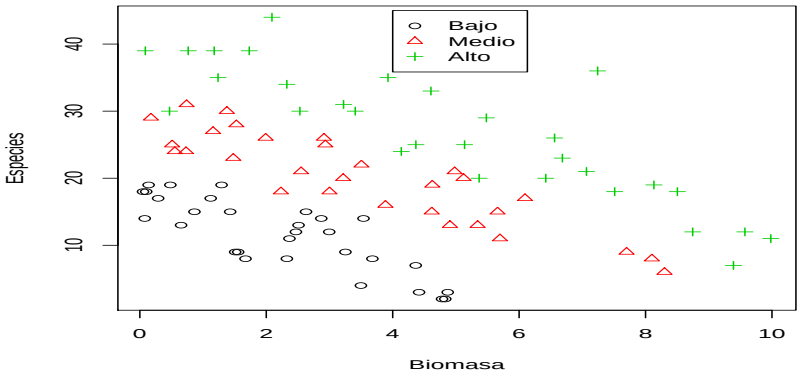
   estos residuos siguen una distribución $N(0, 1)$ por lo tanto la mayoría de ellos deben estar en el intervalo $(-2{,}5, 2{,}5)$.

2. Ejemplo

Un experimento agrícola consta de 90 parcelas de pradera, de $25m^2$ cada una, que se diferencian entre sí, en biomasa, pH del suelo, y la riqueza en especies. Es sabido que la riqueza de especies decrece cuando aumenta la biomasa, pero la cuestión de interés aquí es si esa disminución cambia con el pH del suelo. Las parcelas se clasificaron de acuerdo a tres niveles de pH: alto (=2), medio (=1) y bajo (=0). La variable respuesta es el número de especies por parcela, y las variable predictora son la biomasa media medida en el mes de Junio, y la variable categórica correspondiente al pH del suelo.

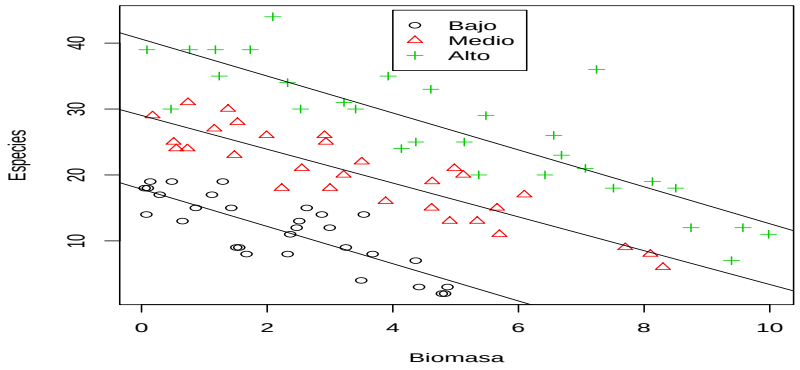
Comenzamos por dibujar los datos:

```
points(Biomasa[pH==0],Especies[pH==0])
points(Biomasa[pH==1],Especies[pH==1],pch=2,col=2)
points(Biomasa[pH==2],Especies[pH==2],pch=3,col=3)
legend(4,45,legend=c("Bajo","Medio","Alto"),pch=c(1,2,3),col=c(1,2,3))
```



Viendo el dibujo, podemos tener la tentación de ajustar un modelos de regresión lineal, pero sería erroneo ya que predeciría valores negativos para el número de especies si el valor para la variable Biomasa es alto. Sin embargo, podemos dibujar las líneas a modo de análisis exploratorio:

```
abline(lm(Especies[pH==0]~Biomasa[pH==0]))
abline(lm(Especies[pH==1]~Biomasa[pH==1]))
abline(lm(Especies[pH==2]~Biomasa[pH==2]))
```



Se puede observar claramente la diferencia en biomasa media para distinto valor del pH (hay distancia entre las rectas), pero no se aprecia que la pendiente varíe significativamente (las rectas son aproximadamente paralelas). Si hubiéramos ajustado un modelo de regresión lineal, hubiéramos obtenido que la interacción entre Biomasa y pH no era significativa. Veamos qué ocurre cuando analizamos los datos utilizando un modelo de regresión de poisson.

Primero ajustamos cada variable por separado, ¿son significativas?.

A continuación las incluímos las dos juntas en el modelo:

```
especies0=glm(Especies~Biomasa+pH,family=poisson)
```

¿son variables de confusión?. Incluímos la interacción y contastamos si es significativa:

```
especies1=glm(Especies~Biomasa*pH,family=poisson)
especies2=update(especies1,~.-Biomasa:pH)
anova(especies2,especies1,test="Chi")
Analysis of Deviance Table

Model 1: Especies ~ Biomasa + pH
Model 2: Especies ~ Biomasa * pH
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        86     99.242
2        84     83.201  2   16.040 0.0003288
```

Lo que nos indica que asumir que el descenso en número de especies cuando aumenta la biomasa es similar sea cual sea el pH, es erroneo.

```
 summary(especies1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.95255    0.08240  35.833  < 2e-16 ***
Biomasa     -0.26216    0.03803  -6.893 5.47e-12 ***
pH1          0.48411    0.10723   4.515 6.34e-06 ***
pH2          0.81557    0.10284   7.931 2.18e-15 ***
Biomasa:pH1  0.12314    0.04270   2.884 0.003927 **
Biomasa:pH2  0.15503    0.04003   3.873 0.000108 ***
```

El modelo ajustado es el siguiente:

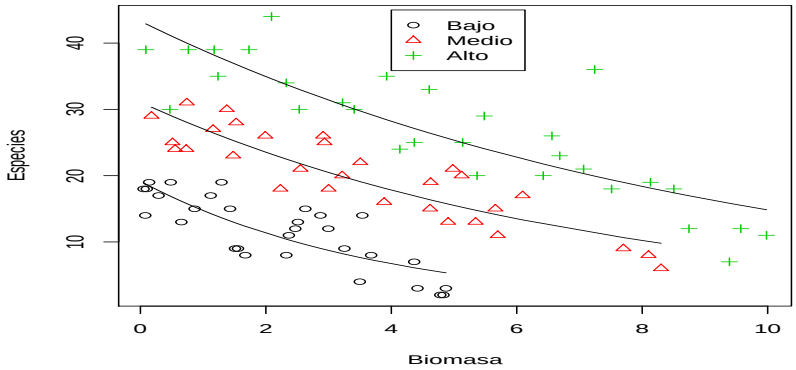
$$\begin{aligned} log(Especies) &= 2{,}95 - 0{,}261 \times Biomasa + 0{,}484 pH_{medio} + 0{,}815 \times pH_{alto} \\ &\quad +0{,}123 \times pH_{medio} \times Biomasa + 0{,}155 \times pH_{alto} \times Biomasa \end{aligned}$$

La interpretación de los parámetros sería la siguiente:

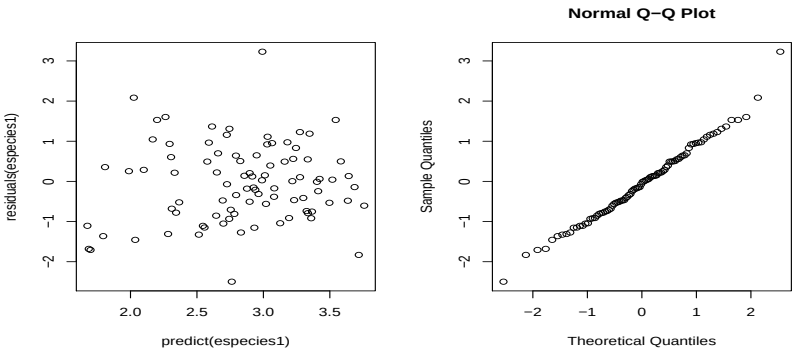
- Un cambio en una unidad de Biomasa supodría un número de especies $exp(-0{,}262) = 0{,}769$ menor en suelos con pH bajo
- Un cambio en una unidad de Biomasa supodría un número de especies $exp(-0{,}262 + 0{,}123) = 0{,}87$ menor en suelos con pH medio

- Un cambio en una unidad de Biomasa supodría un número de especies $exp(-0{,}262 + 0{,}155) = 0{,}898$ menor en suelos con pH alto

En el siguiente gráfico vemos las curvas ajustadas, y se aprecia cómo el utilizar un modelo de poisson resuelve el problema de los valores positivos, al ajustar una curva exponencial en vez de una recta. Podemos hacer un análisis de residuos:




```
par(mfrow=c(1,2))
plot(predict(especies1),residuals(especies1))
qqnorm(residuals(especies1))
```



En el caso de datos de Poisson, la media es igual a la varianza, si los datos no cumplen esta hipótesis los errores estándar de los estimadores de los parámetros serán incorrectos. En estas ocasiones podemos introducir un **parámetro de dispersión**, ϕ, de modo que $Var(Y) = \phi\mu$ (si $\phi = 1$ estamos en el caso Poisson). Dicho parámetro se puede estimar como:

$$\hat{\phi} = \frac{Deviance}{n - p}$$

En el ejemplo anterior, $Deaviance = 83{,}2$ y $n-p = 84$, por lo que el parámetro es muy próximo a 1. Podemos utilizar en la función `glm`, la familia *quasipoisson*, que automáticamente estima el parámetros de dispersión. dará lugar a los mismos valores de los parámetros pero cambiarán los errores estándar.

```
especies2=glm(Especies~Biomasa*pH,family=quasipoisson)
especies2
```

3. Regresión de Poisson para tasas de incidencia

Podemos usar la regresión logística para estimar la prevalencia (proporción), pero no podemos usarala para estimar la incidencia, ya que en este caso necesitamos utilizar el tiempo de exposición al riesgo, ya que en la tasas de incidencia medimos el número de sucesos en la unidad de tiempo.

Supongamos que la variable Y representa el número de eventos que aparecen con una tasa λ por unidad de tiempo de exposición y que la exposición es E (E se suele dar en términos de personas por año o 1000×personas por año) y va a determinar las unidades en las que se expresa la tasa. Dado que nuestro interés es la tasa, λ, y esta puede depender de distintas covariables, podríamos ajustar el modelo:

$$log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + X_k \beta_k$$

Pero es Y quien sigue una distribución de Poisson y no λ. Dado que , $y \sim P(\mu)$ y el número medio de eventos $E[Y] = \mu = \lambda \times E$:

$$log(\mu) = log(\lambda) + \log(E) = \beta_0 + \beta_1 X_1 + \ldots + X_k \beta_k + \log(E)$$

Al $log(E)$ se le llama *offset* y es una cantidad fija y no hay que estimar ningún parámetro para ella. La interpretación de los parámetros va unida al concepto de riesgo relativo. Por ejemplo, supomgamos que tenemos una sola covariable X

$$log(\lambda) = \beta_0 + X\beta_1$$

Si llamamos λ_0 a la tasa de incidencia cuando la covariable toma un valor x_o, y λ_1 a la tasa cuando $X = x_0 + 1$, entonces:

$$\begin{aligned} log\left(\frac{\lambda_1}{\lambda_0}\right) &= log(\lambda_1) - log(\lambda_0) = \beta_1 \\ RR &= \frac{\lambda_1}{\lambda_0} = e^{\beta_1} \end{aligned}$$

es decir, e^{β_1} es el riesgo relativo entre la población donde $X = x_0$ y $X = x_1$, y e_0^{β} es la tasa de incidencia cuando todas las variables son cero.

3.1. Ejemplo

Los datos corresponden al registro de incidencias de diabetes mellitus en niños de 0 a 14 años de la Comunidad de Madrid. Las variables son:

Variable	Valores
sexo	1=Varón
	2=Mujer
edad	1=0-4
	2=5-9
	3=10-15
mes	(numérica del 1 al 12)
periodo	1997-2003
casos	número de casos por estrato
poblacio	número de personas en el estrato
estacion	1=fría (Octubre a Marzo)
	2= caliente (Abril a Septiembre)

El objetivo del estudio era:

a) Ver si hay cambios en la incidencia en los 7 años de registro globalmente y por grupos de edad y sexo

b) Explorar la posibilidad de que la incidencia sea diferente en las diferentes estaciones del año, así como explicar si hay variaciones significativas en la incidencia según el mes.

```
diabetes=read.table("diabetes.txt",header=TRUE)
names(diabetes)
diabetes$periodo=factor(diabetes$periodo)
diabetes$estacion=factor(diabetes$estacion)
diabetes$sexo=factor(diabetes$sexo)
diabetes$mes=factor(diabetes$mes)
diabetes[1:10,]
   edad sexo mes periodo casos poblacion estacion
1     1    2  12    1997     2    110324        1
2     1    2   8    1997     1    110324        2
3     2    2   9    1997     0    120910        2
4     3    2   1    1997     1    146712        1
5     1    1  11    1997     1    116134        1
6     2    2  10    1997     3    120910        2
7     3    1   5    1997     3    154741        2
8     1    2   6    1997     0    110324        2
9     1    2  10    1997     0    110324        2
10    3    2   3    1997     3    146712        1
```

$\lambda = y/E$
por ejemplo para niños, entre 0 y 4 años, en el mes de Diciembre del año 1997, $y = 2$ y $E = 110324$, por lo tanto:

$$\lambda = \frac{2}{110324} = 0{,}0000181 \quad \text{o 1.81 cada 100000 personas}$$

El uso de la función glm() es similar a los ejemplos anteriores, aunque con la peculiaridad de que hemos de especificar E, para especificarlo utilizamos `offset()`:

```
logpobla=log(poblacion)
diabetes1=glm(casos~periodo+offset(logpobla),family=poisson)
summary(diabetes1)
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -11.21951    0.08944 -125.438   <2e-16 ***
periodo1998   0.02386    0.12649    0.189    0.850
periodo1999   0.04955    0.12599    0.393    0.694
periodo2000   0.13387    0.12391    1.080    0.280
periodo2001  -0.15710    0.13238   -1.187    0.235
periodo2002   0.08106    0.12369    0.655    0.512
periodo2003   0.03106    0.12527    0.248    0.804
---
Null deviance: 686.27  on 503  degrees of freedom
Residual deviance: 680.44  on 497  degrees of freedom
```

Podemos ver que el año no es una variable significativa. ¿Cómo usarías la función `anova()` para comprobarlo?.

¿Qué estaríamos calculando si hacemos:

```
exp(0.04955-0.02386)
```

Si queremos obtener los resultados como RR escribimos:

```
> exp(diabetes1$coeff)
 (Intercept)  periodo1998  periodo1999  periodo2000  periodo2001  periodo2002
1.340994e-05 1.024145e+00 1.050797e+00 1.143247e+00 8.546228e-01 1.084431e+00
 periodo2003
1.031550e+00
> exp(confint(diabetes1))
Waiting for profiling to be done...
                   2.5 %    97.5 %
(Intercept) 1.119437e-05 0.0000159
periodo1998 7.990119e-01 1.3127114
periodo1999 8.206769e-01 1.3456605
periodo2000 8.968513e-01 1.4585479
periodo2001 6.584309e-01 1.1071044
periodo2002 8.511035e-01 1.3829685
periodo2003 8.068865e-01 1.3192888

diabetes1.resid=rstandard(diabetes1)
summary(diabetes1.resid)
```

La tasa de incidencia ajustada

```
diabetes1.ajustados=predict(diabetes1,type="response")
tasa1.ajustada=diabetes1.ajustados/poblacion
```

Ahora estamos listos para responder a las cuestiones que se planteaban al principio
¡¡MANOS A LA OBRA!!

Capítulo 6

Regresión de Poisson con variables categóricas:el peligro de las tablas de contingencia

Cuando se lleva a cabo un estudio, se mide un número limitado de variables, por lo que es inevitable que pasemos por alto variables que tienen gran influencia en nuestro estudio. Sin embargo, este problema ocurre frecuentemente porque *agregamos datos sobre una variable predictora importante*.

Supongamos que se ha llevado a cabo un estudio sobre las defensas de los árboles frente a insectos que comen hojas. Un estudio preliminar había sugerido que la presencia de áfidos cuando la hoja es pequeña, puede causar cambios químicos en la hoja que reduzcan la probabilidad de que éstas sean atacadas por insectos más adelante. En el estudio se marcaron un gran número de hojas y se recogio información sobre cuantas eran infectadas por áfidos y cuantas aparecieron agujereadas al final de la estación. El estudio se llevó a cabo en dos tipos de árboles y los resultados (que están en el fichero `arboles.txt`) son:

Árbol	Áfido	Agujereado	Intacto	Total hojas	Proporción
Tipo 1	Sin	35	1750	1785	0.0196
	Con	23	1146	1169	0.0197
Tipo 2	Sin	146	1642	1778	0.0817
	Con	30	333	363	0.0826

Hay 4 variables,

- **Hojas**: Número de hojas, con 8 valores
- **Oruga**: Toma dos valores 0 =No, 1 = Sí
- **Áfido**: Toma dos valores 0 =No, 1 = Sí
- **Árbol**: Toma dos valores 0 y 1.

Vamos a ajustar en primer lugar el modelo **saturado**, es decir, el que tiene tantos parámetros como datos. El ajuste es perfecto, pero no hay grados de libertad disponibles.

```
arboles=read.table("arboles.txt", header=TRUE)
attach(arboles)
poisson1=glm(Hojas~Arbol*Afido*Oruga,family=poisson)

Coefficients:
      (Intercept)             Arbol             Afido             Oruga
         7.467371          -0.063701          -0.423338          -3.912023
      Arbol:Afido        Arbol:Oruga        Afido:Oruga  Arbol:Afido:Oruga
        -1.172190           1.491959           0.003484           0.009634

Degrees of Freedom: 7 Total (i.e. Null);  0 Residual
Null Deviance:      6573
```

El asterisco asegura que todas los efectos principales e interacciones se están ajustando en el modelo. Este modelo no tiene sentido, por lo que comenzamos por eliminar la interacción de orden 3, para eso podemo utilizar la función `update`:

```
poisson2=update(poisson1,~.-Arbol:Afido:Oruga)
```

La puntuación utilizada es importante, estamos diciendo que al modelo anterior hay que quitarle sólo la interacción de orden 3 (por eso utilizamos `:` en vez de `*`). Para saber si este término es significativo utilizamos la función `anova()`:

```
anova(poisson2,poisson1,test="Chi")
Analysis of Deviance Table

Model 1: Hojas ~ Arbol + Afido + Oruga + Arbol:Afido + Arbol:Oruga + Afido:Oruga
Model 2: Hojas ~ Arbol * Afido * Oruga
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         1    0.00079
2         0  1.572e-13  1  0.00079   0.97756
```

Esto muestra claramente que la interacción entre oruga y áfido no difiere de árbol a árbol. Si esta interacción hubiera sido significativa ya hubiéramos terminado la selección del modelo. En este caso podemos continuar y responder a la pregunta de interés para este estudio, ¿hay interacción entre la presencia de áfidos y los agujeros en las hojas?. Para comprobar esto eliminamos la interacción `Afido:Oruga`:

```
poisson3=update(poisson2,~.-Afido:Oruga)
anova(poisson3,poisson2,test="Chi")
Analysis of Deviance Table

Model 1: Hojas ~ Arbol + Afido + Oruga + Arbol:Afido + Arbol:Oruga
Model 2: Hojas ~ Arbol + Afido + Oruga + Arbol:Afido + Arbol:Oruga + Afido:Oruga
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         2    0.00409
2         1    0.00079  1  0.00329   0.95423
```

Por lo que podemos concluir que no hay evidencia de que la presencia de áfidos al principio de la estación mejore la defensa del árbol frente a las orugas.

Ahora, vamos a realizar el mismo análisis de forma errónea, para mostrar el peligro que tiene colapsar tablas de contingencia sobre variables importantes. Supongamos que no tenemos en cuenta el efecto del árbol, de modo que ajustamos el modelo:

```
poisson4=glm(Hojas~Afido*Oruga,family=poisson)
poisson5=update(poisson4,~.-Afido:Oruga)
anova(poisson5,poisson5,test="Chi")
Analysis of Deviance Table

Model 1: Hojas ~ Afido + Oruga
Model 2: Hojas ~ Afido + Oruga
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         5     556.85
2         5     556.85  0     0.00
```

La conclusión hubiera sido que el número de hojas agujereados disminuye (ya que el coeficiente es negativo) cuando se infectan los árboles con áfidos. El error es debido a la proporción de hojas agujereadas en un tipo de árbol es 4 veces mayor que en la del otro

Capítulo 7

Resultados teóricos en GLMs

1. La familia exponencial

Un concepto importante que unifica todos los GLms es la **familia exponencial de distribuciones** . Este concepto fue presentado por primera vez en Fisher (1934).
Todas las distribuciones pertenecientes a la familia exponencial tiene una función de densidad (o de probabilidad) que se puede expresar de la siguiente forma:

$$f(y;\boldsymbol{\theta},\phi) = exp\left\{\frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi)\right\} \tag{7.1}$$

donde, en cada caso, $a(.)$, $b(.)$ y $c(.)$ serán funciones específicas. El parámetro $\boldsymbol{\theta}$ es lo que se llama *parámetro canónico de localización* y ϕ es un *parámetro de dispersión*. La distribucion Binomial, Poisson y Normal (entre otras) son miembros de la familia exponencial. El caso más importante es el de la distribución Normal, cuya función de densidad es:

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-[y-\mu]^2/2\sigma^2\right\} \quad \text{que podemos reescribir como} \\ &= exp\left\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}\left[y^2/\sigma^2 + \ln(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Por lo tanto, $\boldsymbol{\theta} = \mu$, $b(\boldsymbol{\theta}) = \mu^2/2$, $a(\phi) = \phi$, $\phi = \sigma^2$ y

$$c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$$

1.1. La familia exponencial y la máxima verosimilitud

Para obtener los estimadores del vector de parámetros desconocidos, $\boldsymbol{\theta}$, podemos usar el método de máxima verosimilitud. Es más conveniente trabajar con el ln de la función de verosimilitud, y por supuesto, esto no cambia los estimadores obtenidos. Usando (7.1), el logaritmo de la función de verosimilitud para una distribución perteneciente a la familia exponencial es:

$$l(\boldsymbol{\theta},\phi|\boldsymbol{y}) = \ln\left(exp\left\{\frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi)\right\}\right) = \frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi) \tag{7.2}$$

La función *score* es la primdera derivada la función anterior con respecto a los parámetros de interés. Por simplicidad, trataremos ϕ como un *parámetro de ruido* (no de interés primordial), y

calcularemos las derivadas con respecto a $\boldsymbol{\theta}$, la función score resultante es:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{y - \frac{\partial}{\partial \boldsymbol{\theta}} b(\theta)}{a(\phi)} = \frac{y - b'(\boldsymbol{\theta})}{a(\phi)} \tag{7.3}$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = -\frac{b''(\boldsymbol{\theta})}{a(\phi)} \tag{7.4}$$

2. Componentes de un modelo lineal generalizado

Comencemos con el modelo de regresión estándar:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$$

donde $\boldsymbol{X\beta}$ es una combinación lineal de las variables predictoras llamada *predictor lineal* (el cual se representa como $\boldsymbol{\eta}$), en este caso la media $\boldsymbol{\mu}$ está directamente relacionada con el predictor lineal, ya que en este caso $\boldsymbol{\mu} = \boldsymbol{\eta}$. Usando este modelo sencillo, podemos ver que hay dos componentes en el modelo: la función de probabilidad de la variable respuesta y la estructura lineal del modelo. En general, un modelo lineal generalizado tendra los siguientes componentes:

1. **Componente aleatorio**: $\boldsymbol{y}$ es un vector aleatorio procedente de una distribución que pertenece a la familia exponencial y cuya media es $\boldsymbol{\mu}$.
2. **Componente sistemático**: es el predictor lineal $\boldsymbol{\eta} = \boldsymbol{X\beta}$. D
3. **La función *link***: es una función monótona, derivable que establece la relación entre la media y el predictor lineal

   $$\boldsymbol{\eta} = g(\boldsymbol{\mu}) \quad E(\boldsymbol{y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) \tag{7.5}$$

   En el caso del modelo de regresión ordinaria, $\boldsymbol{\mu} = \boldsymbol{\eta}$, por lo tanto la función link es la identidad.
   Hay muchas opciones par la función link. La función **link canonica** es una función que transforma la media en el parámetro canónico $\boldsymbol{\theta}$

   $$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\theta} \Rightarrow g \quad \text{es una función link canónica}$$

La siguiente tabla muestra las funcionen link canónicas para las distribuciones más comunes usadas en los GLMs:

Distribución	Link canónica
Normal	$\boldsymbol{\eta} = \boldsymbol{\mu}$ (identidad)
Binomial	$\boldsymbol{\eta} = \ln\left(\frac{P}{1-P}\right)$ (logistística)
Poisson	$\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$ (logarítmica)
Exponential	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (recíproca)
Gamma	$\boldsymbol{\eta} = \frac{1}{\boldsymbol{\mu}}$ (recíproca)

Cuadro 1: Link canónicas usadas en los GLMs

Podemos ver la elección de función link como algo sismilar a la elección de transformaciones de la variable respuesta. Sin embargo, la función link es una transformación de la *media poblacional*, no de los datos. Se pueden encontrar más detalles sobre funciones link en McCullagh and Nelder (1989, chap. 2)

3. Estimación de Modelos Lineales Generalizados

Anteriormente vimos como estimar el parámetro canónico, $\boldsymbol{\theta}$, mediante máxima verosimilitud. Sin embargo, esto no es útil en la práctica, ya que $\boldsymbol{\theta}$ dependerá de variables explicativas, y cuando usamos la link canónica, $\boldsymbol{\theta} = \boldsymbol{\eta} = \boldsymbol{X\beta}$, por lo tanto,$\boldsymbol{\beta}$ y no $\boldsymbol{\theta}$ son nuestros parámetros de interés.

3.1. Caso general

Dado un vector de observaciones $\boldsymbol{y} = (y_i, y_2, \ldots, y_n)'$. El logaritmo de la verosimilitud es

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} ((y_i\theta_i - b(\theta_i)/a(\phi) + c(y_i, \phi)) \tag{7.6}$$

Por lo tanto, la función score:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \tag{7.7}$$

$$= \sum_{i=1}^{n} \frac{(y_i - b'(\theta_i))}{a(\phi)}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \tag{7.8}$$

pero $\eta_i = g(\mu_i) = \boldsymbol{\beta}'\boldsymbol{x}_i$ y debido a que

$$E(y) = b'(\boldsymbol{\theta}) \quad Var(y) = b''(\boldsymbol{\theta})a(\phi)$$

tenemos que

$$\begin{aligned} g(b'(\theta_i)) &= \boldsymbol{\beta}'\boldsymbol{x}_i \\ g'(b'(\theta_i))b''(\theta_i)\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i \\ \text{por tanto } g'(\mu_i)b''(\theta_i)\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i \\ \text{y } \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{a(\phi)g'(\mu_i)b''(\theta_i)}\boldsymbol{x}_i \\ \text{i.e. } \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)V_i}\boldsymbol{x}_i \end{aligned}$$

donde $V_i = Var(y_i) = a(\phi)b''(\theta_i)$.

$\hat{\boldsymbol{\beta}}$ es la solución de $\frac{\partial l}{\partial \boldsymbol{\beta}} = 0$. Generalmente, este sistema de ecuaciones necesita resolverse de forma iterativa, usando el algoritmo de Newton-Raphson:

$$\left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_1} \approx \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} + \left.\frac{\partial^2 l}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_0} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \tag{7.9}$$

$$0 = \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} + \left.\frac{\partial^2 l}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}_0} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \tag{7.10}$$

calculamos $\hat{\boldsymbol{\beta}}_1$ a partir de $\boldsymbol{\beta}_0$ y así sucesivamente; este proceso da lugar a $\boldsymbol{\beta}_\gamma \to \hat{\boldsymbol{\beta}}$.

En la práctica, el valor de la segunda derivada se sustituye por su valor esperado, a esto se le llama **Mínimos cuadrados iterativamente reponderados** . Usando

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -E\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^2$$

Tenemos que

$$\begin{aligned} E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right) &= -E\left(\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{(g'(\mu_i)V_i)^2}\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\ &= -\sum_{i=1}^{n} \frac{V_i}{(g'(\mu_i))^2 V_i^2}\boldsymbol{x}_i\boldsymbol{x}_i' \\ &= -\sum_{i=1}^{n} w_i\boldsymbol{x}_i\boldsymbol{x}_i' \end{aligned}$$

donde $w_i = 1/V_i\,(g'(\mu_i))^2$. Podemos escribir la expresión anterior en forma matricial:

$$E\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right) = -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

donde $\boldsymbol{W}$ es una matriz diagonal con elementos w_i.Por lo tanto, podemos decir que si $\hat{\boldsymbol{\beta}}$ es solución de $\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, entonces $\hat{\boldsymbol{\beta}}$ es asintóticamente Normal con media $\boldsymbol{\beta}$ y matriz de covarianzas$(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$.

La ecuación (7.10) es ahora,

$$\begin{aligned} \boldsymbol{\beta}_{new} &= \boldsymbol{\beta}_{old} - \left(E\left[\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right]\right)^{-1}_{\boldsymbol{\beta}_{old}} \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_{old}} \\ &= \boldsymbol{\beta}_{old} + (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z} \end{aligned}$$

donde $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta}_{old} + (\boldsymbol{y}-\boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) = \boldsymbol{\eta}_{old} + (\boldsymbol{y}-\boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old})$, y $\boldsymbol{z}$ se llama *working vector*.

Por lo tamto, el IRLS basado en el método de Newton-Raphson se puede resumir así:

1. Obtener valores iniciales para $\boldsymbol{\beta}$, por ejemplo, $\hat{\boldsymbol{\beta}}_{old}$.
2. Usar $\hat{\boldsymbol{\beta}}_{old}$ para estimar $\boldsymbol{W}$ y $\boldsymbol{\mu}_{old}$.
3. Entonces $\hat{\boldsymbol{\eta}}_{old} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{old}$. Calcular $\boldsymbol{z}_{new}$.
4. Calcular nuevos estimadores $\hat{\boldsymbol{\beta}}_{new}$, y repetir los pasos 2 al 4 hasta que converja.

El caso del link canónico

En general, en GLMs el valor de la segunda derivada de la verosimilitud y su valor esperado son distintos. Sin embargo, si se usa el link canónico, ambas matrice son iguales, y $w_i = 1/g'(\mu_i)$. La demostración de este resultado se deja al lector

¿Qué ocurre cuando la variable respuesta no proviene de una familia exponencial?

Wedderburn (1974) desarrolló el concepto de quasi-verosimilitud que usa el hecho de que en la función score la distribución de la variable respuesta sólo está presente a través de los momentos de primer y segundo orden, es decir, para definir la quasi-verosimilitud sólo necesitamos esoecificar la relación entre la media y la varianza de las observaciones.
Para una familia exponencial de un sólo parámetro, la verosiimilitud es la misma que la quasi-verosimilitud.

3.2. Estimación del parámetro de dispersión

Con la excepción de la variable Binomial y Poisson, el parámetro de disperdión no tiene por qué ser conocido, y tendrá que ser estimado. Cuando estimamos $\boldsymbol{\beta}$, no es necesario conocer ϕ, ya que es sistema de ecuaciones de la función score, es independiente de ϕ. En el caso de la distribución Normal, el parámetro de dispersión $\phi = \sigma^2$ es estimado de modo que la varianza residual escalada es igual a los grados de libertad, es decir, $\hat{\sigma}^2 = RSS/(n-p) = D/(n-p)$, y por lo tanto $d/\hat{\phi} = n-p$. La extensión de esto al caso de los GLMs sería estimar ϕ como la media de los residuos de Pearson al cuadrado,

$$\hat{\phi} = \frac{\sum_{i=1}^{n} r_{iP}^2}{n-p}$$

3.3. Ejemplo 1

Usaremos una regresión lineal simple a través del origen como ejemplo. Estamos en el caso

$$y_i \sim N(\underbrace{\beta x_i}_{\mu_i}, \sigma^2).$$

En la sección 2 vimos que la función link es la identidad, y $\theta_i = \mu_i = x_i\beta$, $b(\theta) = \mu^2/2$, y $\phi = \sigma^2$. Por lo tanto, $V_i = \sigma^2$, $g(\mu) = \mu \Rightarrow g'(\mu) = 1$, y $w_i = 1/V_i(g'(\mu))^2 = 1/\sigma^2$. Entonces,

$$\begin{aligned}
\frac{dl}{d\beta} &= \sum_{i=1}^{n} \frac{y_i - \mu_i)x_i}{g'(\mu_i)V_i} = \sum_{i=1}^{n} \frac{(y_i - x_i\beta)x_i}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{x}'(\boldsymbol{y} - \boldsymbol{x}\beta) \\
E\left(\frac{d^2l}{d\beta^2}\right) &= -\sum_{i=1}^{n} w_i x_i^2 = -\sum_{i=1}^{n} x_i^2/\sigma^2 = -\boldsymbol{x}'\boldsymbol{x}/\sigma^2 \\
\boldsymbol{z} &= \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old})\boldsymbol{g}'(\boldsymbol{\mu}_{old}) = \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{\mu}_{old}) = \boldsymbol{x}\beta_{old} + (\boldsymbol{y} - \boldsymbol{x}\beta_{old}) = \boldsymbol{y}.
\end{aligned}$$

Por lo tanto, sólo una iteración es necesaria, y

$$\hat{\beta} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{y}$$

que es el estimador de mínimos cuadrados de β.

4. Inferencia

Volviendo al modelo GLM original, cuya función de verosimilitid viene dada por (7.6), con $E(\boldsymbol{y}) = \boldsymbol{\mu}$, $g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$. Hay varias maneras de hacer contrastes de hipótesis sobre las componentes de $\boldsymbol{\beta}$.

1. Si queremos contrastar por ejemplo, $\beta_1 = 0$ (el primer componente de $\boldsymbol{\beta}$), el procedimiento sería calcular $\hat{\beta}_1$ y $se(\hat{\beta}_1)$ y comparar $|\hat{\beta}_1|/se(\hat{\beta}_1)$ con $N(0,1)$, y rechazar $\beta_1 = 0$ si esa cantidad es demasiado grande.
   recuerda que $se(\hat{\beta}_1)$ es la raíz cuadrada del elemento $(1,1)$ de la inversa de la matriz

$$\left.\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right|_{\hat{\boldsymbol{\beta}}}$$

   De modo que aquí estamo usando la normalidad asintótica del estimador máximo versímil $\hat{\boldsymbol{\beta}}$, y la fórmula de su varianza asintótica. ¡Por lo tanto necesitamos una muestra grande!

2. Si la muestra de datos es pequeña, la aproximación a la Normal, puede no funcionar bien. Una alternativa es utilizar el *Análisis de la Varianza*. La base de este método es utilizar una medida del ajuste del modelo que mida la discrepancia entre los datos ajustados por el modelo y los datos. Para datos normales esta medida es la *Suma de cuadrados residual* , para datos no-Normales es el **Deviance**. En general, el deviance se basa en valor (maximizado) de la verosimilitud. El test se basa en la reducción de esta medida de ajuste del modelo cuando incluimos una nueva variable. Una buena aproximación para la distribución que sigue esta reducción de la verosimilitud es la distribución Chi-cuadrado. Por lo tanto, otra forma de hacer contrastes sobre los parámetros es mediante un test Chi-cuadrado.
   En general, supongamos que queremos contrastar $\boldsymbol{\beta} \subset \omega_r$ (el modelo *reducido*, es decir, $\boldsymbol{\beta}_1 = 0$) frente a $\boldsymbol{\beta} \subset \omega_f$ (el modelo *completo*) donde $\omega_r \subset \omega_f$. Consideremos la diferencia entre los máximos del logaritmo de la función de verosimilitud para el modelo completo y el reducido. Sea L_f el valor maximizado del logaritmo de la verosimlitud bajo el modelo completo (todos los β_1 presentes) y sea L_r el valor maximizado del logaritmo de la verosimiitud bajo el modelo reducido (con $\beta_1 = 0$). Es claro que L_f es al menos tan grance como L_r (¿por qué?). El test estadístico es:

$$S(\omega_r, \omega_f) = -2(L_r - L_f) = 2\sum_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right] / \phi$$

   donde $\tilde{\theta}_i$=eml (estimador máximo verosímil) bajo ω_r y $\hat{\theta}_i$=eml bajo ω_f.

   La distribución de $S(\omega_r, \omega_f)$ se aproxima por una χ^2 con grados de libertad igual a la diferencia entre el número de parámetros en el modelo completo y el reducido (si estamos contrastando $\beta_1 = 0$, entonces, χ^2_1). Los intervalos de confianza para los parámetros se calculan también usando la distribución χ^2.

Deviance escalado: Un modelo saturado, es un modelo que ajusta perfectamente los datos (es decir, que tiene tantos parámetros como observaciones y los valores ajustados son iguales a los observados). Sea l_s el logaritmo de la verosimilitud de modelo saturado, y sea l_m el valor maximizado del logarirmo de la verosimilitud del modelo de interés. El *Deviance escalado* del modelo de interés es

$$\text{Deviance escalado} = S(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2(L(\hat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - L(\boldsymbol{y}, \phi, \boldsymbol{y}) = -2(l_m - l_s)$$

y el deviance es

$$Deviance = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = S(\boldsymbol{y}, \hat{\boldsymbol{\mu}})\phi$$

Por lo tanto, el deviance escalado se puede ver como un test estadístico para contrastar la hipótesis de que los parámetros del modelo saturado (y que no está en el modelo de interés) son iguales a 0. En el caso de datos de Poisson y Binomial, $\phi = 1$, y por lo tanto ambos deviance coinciden. En el caso de la Normal, $\phi = \sigma^2$, entonces, el deviance escalado sigue una distribución F.

5. Diagnósticos para GLMs

Los métodos para comprobar si se cumple las hipótesis de un GLM son fundamentalmente los mismos que para el caso regresión lineal, pero basados en la última iteración del algoritmo IRLS.

5.1. Residuos

En el modelo de rgersión lineal ordinario, los residuos $y_i - \hat{\mu}_i$ se usan para detecta la violación de las hipótesis del modelo, como son la varianza no homgenea, independencia, etc. En GLMs hay tres tipos de residuos:

1. **Residuos respuesta:** $y_i - \hat{\mu}_i$, que no son apropiados ya que $Var(y_i)$ no es constante.

2. **Residuos de Pearso:**
$$r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var(y_i)}}}$$
Tienen varianza constante y media cero. Son útiles para comprobar si la varianza no se ha especificado correctamente.

3. **Residuos deviance:**
$$sign(y_i - \hat{\mu}_i)\sqrt{d_i^2}$$
donde d_i es la contribución de la i-ésima observación al deviance del modelo. estos residuos tienen la propiedad de tener el mismo signo $y_i - \hat{\mu}_i$, y su suma de cuadrados es el deviance. En muchos modelos, los residuos deviance están más próximos a una Normal que los residuos de Pearson, por eso se usan más frecuentemente para construir gráficos de diagnóstico.

4. **residuos estandarizados:** Los residuos deviance y perason pueden ser corregidos para tener varianza estandarizada y evitar los efectos del leverage si se dividen por $\sqrt{\phi(1-h_{ii})}$, en muchos casos $\phi = 1$, y cuando no lo es, se reemplaza por un estimador. Estos residuos deberían ser aproximadamente $N(0,1)$ en el caso de datos de Poisson y Biomiales con valores altos de conteo, y por lo tanto los residuos devería estar entre -2 y +2. McCullagh and Nelder (1989, chap. 12) recomienda hacer el siguiente chequeo:

   a) **Chequeo informal usando los residuos:** Dibujar los residuos deviance estandarizados frente al predictor lineal $\hat{\boldsymbol{\eta}}$ o frente a los valores ajustados $\hat{\boldsymbol{\mu}}$, en este gráfico no debería aparecer ningún patrón. las deviaciones más comunes son:

   - Aparece curvatura en la media.
   - Cambio sistemático de rango en los valores ajustados.

   La curvatura puede aparecer por: elección equivocada de la función link, elección equivocada de la escala de la covariables.

b) **Comprobar la función varianza:** Dibujar los valores absolutos de los residuos frente a los valores ajustados, si la función varianza no se ha elegido correctamente esto dará lugar a una tendencia en la media. Por ejemplo, podemos elegir que la varianza sea una función lineal de la media, pero podría ser en realidad una función cuadrática.

c) **Comprobar la función link:** Gráfico del working vector $\boldsymbol{z}$ que se describio en la página 8, frente al predictor lineal $\boldsymbol{\eta}$. Si la función link es correcta, el patrón del gráfico debería ser una línea recta, (¿por qué?).

d) **Media del leverage:** En regresión lineal, usamos los elementos diagonales de la matriz de proyección como una medida del leverage. Ahora, la matriz de proyección se obtiene en la última iteración del algoritmo IRLS. ¿Qué forma tiene esta matriz?.

e) **Medidas de infuencia:** Lo equivalente a la distnacia de Cook en GLMs es

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\phi}}$$

Bibliografía

Fisher, R. (1934). Thow new properties of mathematical likelihood. *Proceedings of the Royal Statistical Society of London, A*, 144:285–307.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the gauss newton method. *Biometrika*, 61:439–447.