Modelos Lineales Generalizados, Modelos Aditivos generalizados (GAM) y visión del análisis de deviance.

Documento para el curso MGL 2017. ALT

Los GLM (McCullagh y Nelder, 1989) constituyen un procedimiento para describir el efecto de uno o más factores (cuantitativos o cualitativos, cruzados o anidados, fijos o aleatorios) en una o más variables dependientes que se deterioran cuando aparecen contravenciones a sus supuestos. Asumen, dada su semejanza con la regresión, errores con distribución normal o no, cuando la varianza no es constante. Se utilizan en muchos tipos de estudios experimentales especialmente los relacionados con Regresión Múltiple, ANAVA (con sus Sumas de Cuadrados Corregidas (SS..) Tipo III, que verifican la significancia marginal de cada factor, asumiendo que este fue entrado de último al modelo, o la SS.. Tipo I que asumen los factores en el orden en que fueron entrados al modelo), modelos mixtos, etc.

Cuando la varianza no es constante, o cuando los errores no son normalmente distribuidos, específicamente se consideran variables respuestas como: datos de conteos (expresados como proporciones: regresiones logísticas); datos de conteos crudos (modelos log-lineales); variables respuestas binarias; datos de tiempos de muerte, en que la varianza se incrementa más rápida que linealmente con respecto a la media (datos de tiempo con errores Gamma). Figura 1

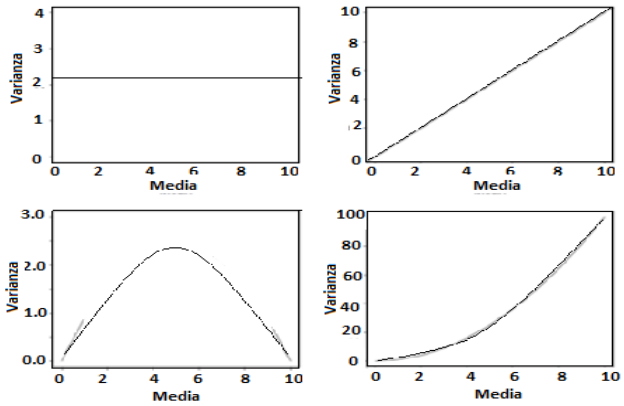


Figura 1. Comportamiento de la varianza con respecto a la media

El anava convencional asume homocedasticidad (gráfica superior izquierda), en conteos (variable respuesta de números enteros y a menudo con muchos ceros) la varianza puede incrementarse linealmente con la media (grafica superior derecha), con datos en proporciones (conteos de éxitos y fracasos), la varianza a veces aparece como una función con forma de U invertida con respecto a las medias (inferior izquierda) y, cuando *Y* sigue una distribución Gamma (datos al tiempo de muerte) la varianza se incrementa más rápido que linealmente con la media (abajo derecha).

Así asumidos los MLG tienen 3 propiedades importantes: 1- La estructura del error; 2- El predictor lineal y 3- Una función de encadenamiento.

Estructura de los errores.

Cuando los errores no se comportan normalmente pueden resultar: fuertemente sesgados o con curtosis, acotados, como en las proporciones o, imposibilitados de alcanzar valores

negativos como en los conteos, lo cual se manejaba con transformaciones de la *Y* o con métodos no paramétricos. A cambio, un MLG permite definir la estructura del error por medio de "*familias directrices*" como parte del modelo (fórmula), por ejemplo, como lo hace el R, lo que propicia mirar una gran variedad de especificaciones para el error, así:

glm(y ~ z, family = poisson); o sea que *Y* tiene errores tipo Poisson.
glm(y ~ z, family = binomial); datos de respuesta binaria y errores ídem.

en que la variable respuesta y se modela contra Z (llamada así por su composición con unos y ceros) que, como hemos visto, puede ser continua (Análisis de regresión) o categórica (como en el ANOVA, o su similar análisis de desvianzas, antes descrito).

El predictor lineal

La estructura del modelo relaciona cada $y_{observado}$ con su $y_{predicho}$, por medio del predictor lineal así, por ejemplo: $\ell_i = \sum_{j=0}^{p} x_{ij}\beta_j$ o $\ell_i = \sum_{j=0}^{p} \beta_j x_{ij}$ suma lineal de efectos, ya vista para p variables explicatorias $\boldsymbol{X}$ y $\boldsymbol{\beta}_j$ parámetros por estimar, es decir por una regresión. La parte derecha corresponde a la estructura lineal.

Se dan tantos términos en el predictor como parámetros (p) por estimar. Así para la regresión simple, el predictor lineal es una suma de dos términos cuyos parámetros son el intercepto y la pendiente.

Un ANAVA con 4 tratamientos, tiene un predictor lineal como la suma de 4 términos para lograr la estimación de la media para cada nivel del factor. Sí, aparecieran covariables estas adicionan un término a cada predictor lineal, (la pendiente de cada relacionamiento), los términos para la interacción, en un ANAVA, en arreglo factorial adicionan uno o más parámetros al predictor lineal, dependiendo de los grados de libertad de cada factor (ejemplo, habría 3 parámetros extras para la interacción entre un factor con dos niveles y uno con cuatro porque (2−1)*(4−1)=3.

Grado de ajuste.

Para medir el grado de ajuste logrado por el modelo, el GLM evalúa el predictor lineal para cada valor de la respuesta, luego compara el valor predicho con un valor de *y* transformado (cuya transformación se especifica en una función de encadenamiento). El valor ajustado se calcula por medio de la recíproca de tal función (las llamadas inversas por ejemplo en el EXCELL), para regresar a la escala original de los datos y medidos.

Función encadenante (*link*).

Una dificultad para los MLG son las relaciones entre los valores de la variable respuesta (tanto la observada como la predicha por el modelo ajustado) y el predictor lineal, para lo cual conviene recordar que la función encadenante relaciona los valores medios de *y* a su predictor lineal algo como $\ell_i = g(\mu)$, o sea como una función de medias.

Esto, aunque simple, necesita reflexiones al respecto por lo sabido, el modelo predictor lineal se compone de una suma de los términos para cada uno de los *p* parámetros, lo cual ya no es un valor de *y* (excepto en el caso especial del encadenante *identidad* que se ha usado implícitamente hasta ahora). El valor de ℓ se obtiene por transformaciones del valor de y por la función encadenante, y los valores predichos de *y* $\left(\hat{y}\right)$ por medio de la inversa de la función encadenante aplicada a ℓ.

Un buen criterio, en la selección de esta función, aseguraría que los valores ajustados queden entre límites razonables, por ejemplo, los conteos para valores > 0, pues los negativos carecen de sentido, las proporciones entre 0 y 1. En el primer caso, un encadenante *logarítmico*, es el apropiado puesto que los valores ajustados son los *antilogs* del predictor lineal, y todo *antilog* es mayor o igual a 0. En el segundo caso, el encadenamiento *logit* es más apropiado pues los valores ajustados se calculan como los *antilogs* de las razones de Odss (p/q), o sea log(p/q), como se vio para la logística.

Al usar diferentes encadenantes se pueden comparar las ejecutorias de varios modelos, ya que al mantenerse la *deviance total* constante, se puede investigar las consecuencias en las alteraciones de los modelos. Como regla de a puño se tiene que la encadenante más apropiada es la que produzca la mínima *deviance* residual.

Funciones encadenantes canónicas.

Las funciones canónicas al respecto son las opciones empleadas por defecto, cuando una estructura particular del error se especifica en la familia directriz dentro de la función *model.* Cuando se omite la directriz encadenante significa que los siguientes ajustes serán los usados (canónicos o por defecto):

Error	encadenante	nombre funcional: R
normal	identidad	*identity*
poisson	log	*log*
binomial	logit	*logit*
Gamma	recíproca	*reciprocal*

Se debe notar que cada encadenante está asociada a la distribución del error y que sólo el error gamma se escribe con mayúscula en su primera letra en R.

Escoger una u otra (ejemplo la encadenante log) y la transformación de la *y* (ej, log(*y*)) en vez de la *y* original, requiere alguna experiencia, pues la decisión se toma con base en, sí la varianza es constante en la escala original de las mediciones. Sí la varianza fuera constante, usaríamos una función encadenante pero, si se incrementara con respecto a la media, probablemente lo mejor es la transformación logarítmica de la Y.

Datos en proporciones y error binomial.

Se basan en elementos ya conocidos y estudiados. Tienen 3 características importantes que afectan sus análisis:
1. Son estrictamente acotados;
2. Son heterocedásticos;

3. Los errores no siguen la normal.

Por ejemplo solo se podría tener una $0 \leq p \leq 1$. Entonces, no tiene sentido un modelo lineal con una pendiente negativa o positiva para datos proporcionales porque se podría llegar a altos niveles de la variable X con proporciones predichas negativas o mayores que 1 y otros sucesos semejantes cuando la probabilidad de éxito es 0 o 1.

Todo lo anterior (acotamiento, heterocedasticidad y errores no normales) se supera con un GLM con estructura binomial para los errores como esta:
glm(y~x, family=binomial) o simplificada como glm(y~x,binomial).

Conteos y errores Poisson.

Se debe considerar que estos:
1. Son acotados por debajo, no hay valores <0;
2. Son heterocedásticos (la varianza se incrementa con la media);
3. Los errores no siguen la normal;
4. Por ser los conteos número enteros afectan la distribución de los errores

Entonces en R basta con:
glm(y~x,poisson)
y así, el modelo es ajustado con encadenante logarítmico lo cual asegura que los valores ajustados serán acotados por lo bajo y los errores tipo Poisson para la no normalidad.

Ejemplo: se desea modelar el número de escarabajos muertos contra dosis de algún producto, lts/m^{3}agua, para controlar un escarabajo xilófago en arrumes de trozas de pino pátula atacadas, controladas por un xilocida, para lo cual se recolectaron los siguientes datos de escarabajos muertos a la hora de aplicarse con las dosis recomendadas, por el fabricante:

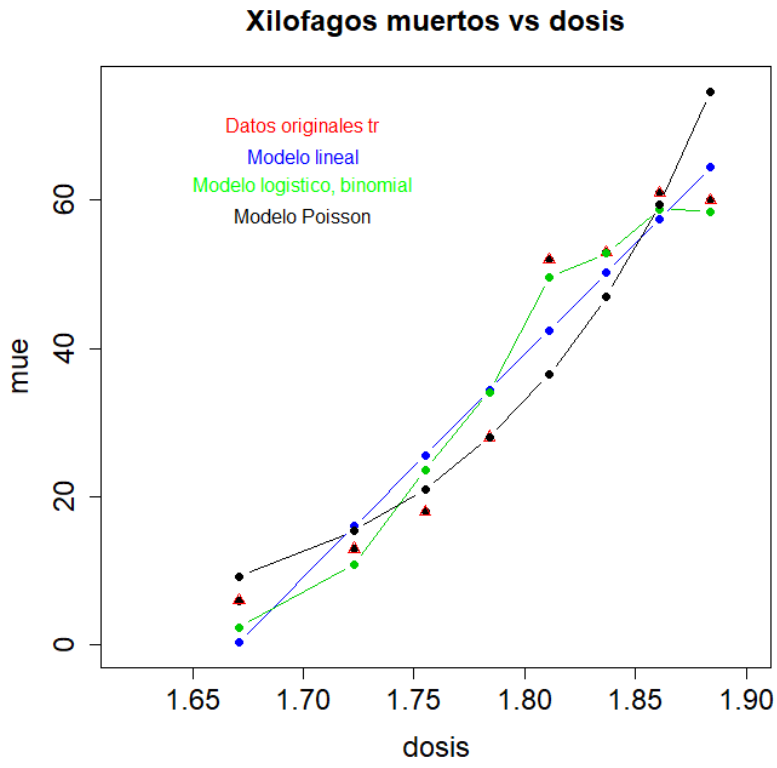
	dosis	mue	n	frac
1	1.671	6	59	53
2	1.723	13	60	47
3	1.755	18	62	44
4	1.784	28	56	28
5	1.811	52	63	11
6	1.837	53	59	6
7	1.861	61	62	1
8	1.884	60	60	0

```
xilof<-read.table("clipboard")
attach(xilof)
names(xilof)
[1] "dosis" "mue"   "n"     "frac"

mode1<-lm(mue~dosis)#Modelo lineal con conteos, no correcto (azules).
summary(mode1)

Call:
lm(formula = mue ~ dosis)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9884 -5.4625  0.7971  2.7739  9.8327
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -544.77      60.50  -9.005 0.000105 ***
dosis         324.04      33.71   9.612 7.25e-05 ***
Residual standard error: 6.017 on 6 degrees of freedom
Multiple R-squared:  0.939,     Adjusted R-squared:  0.9289
F-statistic: 92.39 on 1 and 6 DF,  p-value: 7.255e-05
```

A pesar de estos ajustes el modelo no sería adecuado. La siguiente gráfica resulta de correr diferentes modelos, mostrados luego de ella.

```
plot(dosis,mue,pch=16,cex.lab=1.5,cex.axis=1.5,cex.main=1.5,main="Xilofagos muertos vs
dosis",xlim=c(1.65,1.9),ylim=c(0,70))
lines(dosis,mode1$fitted.values,col="blue",pch=16,type="b")
lines(dosis,mode2$fitted.values*xilof$n,col=3,pch=16,type="b")
lines(dosis,prem3,col="black",pch=16,type="b")#luego del modelo 3
lines(dosis,mue,col="red",pch=2,type="p")

text(1.70,70, "Datos originales tr",col="red",pch=2)
text(1.70,66, "Modelo lineal",col="blue")
text(1.70,62, "Modelo logistico, binomial",col="green")
text(1.70,58, "Modelo Poisson",col="black")
```




```
mode2<-glm(xilof$mue/xilof$n ~ xilof$dosis, family=binomial(link=logit))
mode2<-glm(mue/n ~ dosis, family=binomial(link=logit))
Warning message:
In eval(expr, envir, enclos) : non-integer #successes in a binomial glm!
summary(mode2)

Call:
glm(formula = mue/n ~ dosis, family = binomial(link = logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.22095  -0.03902   0.09298   0.19683   0.26461
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -56.68      38.42  -1.475    0.140
dosis          32.02      21.58   1.484    0.138

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4.71186  on 7  degrees of freedom
Residual deviance: 0.25897  on 6  degrees of freedom
AIC: 8.1782

Number of Fisher Scoring iterations: 5

mode3<-glm(mue ~ dosis, family=poisson)
> summary(mode3)

Call:
glm(formula = mue ~ dosis, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8016  -0.8106  -0.2087   0.3635   2.4852

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -14.693      2.000  -7.346 2.04e-13 ***
dosis         10.091      1.093   9.232  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 111.836  on 7  degrees of freedom
Residual deviance:  13.309  on 6  degrees of freedom
AIC: 58.785

Number of Fisher Scoring iterations: 4

> prem3<-mode3$fitted.values
```

El ajuste glm, mejora ostensiblemente el modelo (rojo y verde se parecen más que rojo azul). Además el AIC creció en el modelo 3, pasó de 8.18 a 58.79.

Deviance: **Medida de bondad de ajuste, en modelos, para un MLG.**

Los valores ajustados, producidos por el modelo, difícilmente igualarán perfectamente a los valores de los datos. El tamaño de la discrepancia entre datos modelados y observados, es una medida de lo inadecuado del modelo, por lo cual un pequeño valor podrá ser tolerable, pero no uno muy grande. Al final veremos un concepto ampliado del Análisis de deviance.

La medida de la discrepancia en un MLG para evaluar la falta de ajuste entre modelo y datos es llamada la *deviance*, ya conocida y definida como −2 veces la diferencia en la función logarítmica de verosimilitud entre el modelo actual y el saturado (o sea un modelo que ajusta perfectamente los datos):

$$\boldsymbol{Deviance} = -2LN\left[\frac{Ver.Modelo\ propuesto}{Ver.Modelo\ saturado}\right]$$

Esto lo hemos reiterado varias veces por su importancia. Como el saturado no depende de los parámetros del modelo, minimizar *deviance* es lo mismo que maximizar la verosimilitud.

Además, debe notarse que la *deviance* debe estimarse de diferentes maneras para las diferentes familias de la estructura con MLG, Tabla siguiente:

Tabla 01. Fórmulas para la *Deviance* en diferentes familias MLG.

(y dato observado, $\bar{y}$, valor medio de y, μ valores ajustados de *y* con el modelo de máxima verosimilitud, y n el denominador binomial en un MLG binomial).

Familia Err Est	Deviance	Funcion de varianza
normal	$\sum(y-\bar{y})^2$	1
poisson	$2\sum y\ln(y/u)+(y-u)$	u
binomial	$2\sum y\ln(y/u)+(n-y)\ln(n-y)/(n-u)$	$\frac{u(n-u)}{n}$
Gamma	$2\sum(y-u)/y-\ln(y/u)$	u^2
gaus inversa	$\sum(y-u)^2/(u^2y)$	u^3

Casi verosimilitud (*Quasi-likelihood*).

En algunos casos es difícil especificar la forma precisa de la distribución del error. Podríamos saber que no es normal, pero no asegurar con certidumbre de que distribución se trata, por ejemplo ¿una Binomial negativa??.

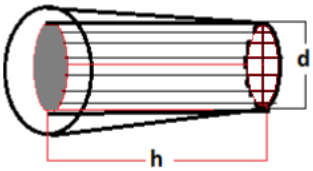
Aparece una alternativa simple y robusta (Wedderburn, 1974), que usa la más elemental información acerca de las Y, denominada "*relación media-varianza"*, la cual sorprendentemente es suficiente para retener en forma cercana la eficiencia de los estimadores máximo verosímiles.

Supongamos que sabemos que siempre la *Y>0*, por lo cual los datos serán invariablemente sesgados a derecha y la varianza se incrementa con la media, pero no posibilita especificar una distribución particular (que podría ser o Poisson o binomial negativa para los errores). En este caso la *Quasi-likelihood* nos libra de especificar la distribución de los errores y solo nos pide dar la *relación media-varianza*, estimada de los datos: $\sigma^2(y_i)\propto\sigma^2(\mu_i)$. Por ejemplo, para datos normalmente distribuidos, la SSE se distribuye como una chi-cuadrada.

Ajustes o compensaciones (*Offsets*)

Una compensación es un componente del predictor lineal que es conocido de antemano (por teoría o como un modelo mecánico del proceso) y, por ser conocido, no requiere estimar parámetros de los datos. Para los modelos lineales con errores normales una compensación es redundante, pues se puede simplemente substraer la compensación de los valores de la variable Y, y trabajar con los residuales en vez de los valores y. En cambio, para los GLM, si es necesario especificar la compensación; que se mantiene constante mientras otras variables explicatorias son evaluadas. Por ejemplo, un símil con datos de trozas de árboles. La teoría antecedente es simple. Asumimos que las trozas son

aproximadamente cilíndricas (entonces la conicidad es despreciable entre la base y la cima de ellas).



Así, el volumen, v, en relación al diámetro de la troza *d*, y la altura, *h*, será dada por $v = \frac{\pi d^2}{40000} h$, y al tomar logaritmos:

$$\log(v) = \log\left(\frac{\pi}{40000}\right) + 2\log(d) + \log(h)$$

Podríamos esperar entonces que si hiciéramos una regresión múltiple de *log(v)* sobre *log(h)* y *log(d)* tuviéramos pendientes estimadas de 1.0 para log(h) y 2.0 para (d). Para verlo sean los siguientes datos de volumen diámetro y altura de unas trozas casi cilíndricas:

	v	d	h	esmo1	esmo2	esmo3		v	d	h	esmo1	esmo2	esmo3
1	0.05	12	1.8	0.048	0.049	0.041	15	1.37	43.8	4.8	1.366	1.370	1.447
2	0.91	36	4.7	0.941	0.937	0.957	16	0.34	25	3.7	0.380	0.378	0.361
3	2.54	52	6.3	2.487	2.468	2.687	17	2.17	50.7	6	2.243	2.231	2.419
4	0.14	17	2.5	0.128	0.128	0.115	18	0.38	24.3	3.9	0.380	0.377	0.359
5	0.05	12.6	1.9	0.056	0.057	0.048	19	0.35	23.8	3.7	0.348	0.346	0.328
6	0.14	17.4	2.6	0.138	0.139	0.124	20	1.02	39.4	4.4	1.031	1.035	1.074
7	0.12	16.1	2.6	0.117	0.118	0.104	21	0.71	33.4	4.2	0.741	0.740	0.745
8	1.44	43.2	4.9	1.370	1.370	1.445	22	1.74	49.2	4.8	1.697	1.709	1.843
9	1	37.4	4.6	0.983	0.982	1.008	23	0.25	18.6	3.9	0.239	0.234	0.213
10	0.7	32.3	4.3	0.707	0.705	0.706	24	1.93	51.1	5	1.870	1.883	2.044
11	1.68	46.2	5	1.580	1.582	1.687	25	1.48	45.4	4.8	1.469	1.474	1.567
12	1.25	41.4	4.7	1.226	1.227	1.284	26	0.99	38.8	4.4	1.010	1.013	1.047
13	0.7	33.6	4	0.695	0.698	0.704	27	1.97	49.6	5.7	2.039	2.033	2.196
14	1.04	39.3	4.5	1.062	1.063	1.102	28	1.55	48.2	4.7	1.597	1.609	1.729

```
fore<-read.table("clipboard")Las columnas esmoi, fueron añnadidas despues de las modelaciones
attach(fore)
names(fore)
[1] "v" "d" "h"
str(fore)
'data.frame':   28 obs. of  3 variables:
 $ v: num  0.049 0.905 2.542 0.137 0.054 ...
 $ d: num  12 36 52 17 12.6 ...
 $ h: num  1.8 4.68 6.3 2.52 1.94 2.62 2.56 4.9 4.58 4.28 ...
```

d<-d/100#modificamos las unidades del d de cm a m

```
model1<-glm(log(v)~log(d)+log(h))
summary(model1)

Call:
glm(formula = log(v) ~ log(d) + log(h))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.103628  -0.022537   0.003476   0.021344   0.068089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11660    0.16767   0.695    0.493
```

```
log(d)       1.78133    0.05363  33.212  < 2e-16 ***
log(h)       1.06461    0.07752  13.733 3.79e-13 ***
(Dispersion parameter for gaussian family taken to be 0.001573751)

    Null deviance: 34.156871  on 27  degrees of freedom
Residual deviance:  0.039344  on 25  degrees of freedom
AIC: -96.433
Number of Fisher Scoring iterations: 2
```

Los estimados son razonablemente cercanos a los esperados (1.064 en vez de 1.0 para log(h) y 1.7813 en vez de 2.0 para log(d)). Vamos a usar la compensación para especificar la segunda respuesta teórica de log(v), para log(h); o sea una pendiente de 1.0 diferente a la estimada 1.06461.

```
model2<-glm(log(v)~log(d)+offset(log(h)))
> summary(model2)

Call:
glm(formula = log(v) ~ log(d) + offset(log(h)))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.099232  -0.023987   0.001109   0.020643   0.080886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25528    0.02054   12.43 1.91e-12 ***
log(d)       1.82375    0.01682  108.43  < 2e-16 ***

(Dispersion parameter for gaussian family taken to be 0.001555268)

    Null deviance: 18.326336  on 27  degrees of freedom
Residual deviance:  0.040437  on 26  degrees of freedom
AIC: -97.665

Number of Fisher Scoring iterations: 2
```

Naturalmente la *deviance* residual se hizo solo un poco mayor. El AIC disminuyó de-96.433 a -97.665 por lo que la simplificación se justifica. El modelo queda entonces como:
=EXP(0.25528+1.82375*LN($D2/100)+LN(E2))

Ahora tratemos de incluir la pendiente teórica (2.0) para log(d) como compensación:
model3<-glm(log(v)~1+offset(log(h)+2*log(d)))

```
model3<-glm(log(v)~1+offset(log(h)+2*log(d)))
summary(model3)

Call:
glm(formula = log(v) ~ 1 + offset(log(h) + 2 * log(d)))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.10945  -0.05555  -0.04217   0.05026   0.18100

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45581    0.01671   27.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.007822778)

    Null deviance: 0.21121  on 27  degrees of freedom
Residual deviance: 0.21121  on 27  degrees of freedom
```

```
AIC: -53.378
```

De Nuevo la deviance residual se hizo solo un poco mayor. El AIC creció por lo que la simplificación ya no se justifica. El modelo:=EXP(0.45581+2*LN($L2/100)+LN(M2))

model4<-glm(log(v) ~ offset(log(pi/4)+log(h)+2*log(d))-1)# Ya no Habrá modelo.

El efecto de las compensaciones se muestra con los estimados en la tabla inicial que muestra grandes semejanzas entre los estimados.

Regresión y modelos de diseños experimentales (DE).

Se encuentra una gran relación entre estos dos conceptos, por cuanto cualquiera de los DE puede escribirse como un modelo lineal con variables explicatorias o independientes ficticias, para lo cual ya se conocen los problemas que conlleva, sobre todo la singularidad. Por ejemplo, el modelo básico de un ANAVA podría presentarse como

$$y_{ij} = \mu + \mu_{ij}; i = 1 \cdots \ell \quad (1),$$

en el cual las perturbaciones $\mu_{ij} = y_{ij} - \mu_i$ cumplen los postulados de la regresión. De acuerdo con ello, este último podría presentarse como:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{ZU} \quad (2),$$

Z con ℓ columnas, escrito matricialmente como:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{\ell 1} \\ y_{\ell 2} \\ \vdots \\ y_{\ell n_\ell} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1n_1} \\ \mu_{21} \\ \vdots \\ \mu_{2n_2} \\ \vdots \\ \mu_{\ell 1} \\ \mu_{\ell 2} \\ \vdots \\ \mu_{\ell n_\ell} \end{bmatrix} + \begin{bmatrix} 1 & 0 & \vdots & 0 \\ 1 & 0 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \vdots & 0 \\ 0 & 1 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & 1 \\ 0 & 0 & \vdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_\ell \end{bmatrix}$$

las columnas de **Z** ortogonales y $\boldsymbol{Z'Z}$ diagonal.

El diseño experimental con un factor lo habíamos expresado como: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ o en nuestra notación como $y_{ij} = \mu + \alpha_i + \mu_{ij}$, μ la media general, α_i el efecto de tratamientos o efecto incremental del grupo i, y μ_{ij} el error.

$\boldsymbol{Z'Z}$ no podría utilizar una variable *dummy* para cada grupo pues no sería invertible. Por ejemplo, con dos grupos, si se definieran dos variables ficticias $\boldsymbol{Z_1}$ y $\boldsymbol{Z_2}$, entonces $\boldsymbol{1} = \boldsymbol{Z_1} + \boldsymbol{Z_2}$ y el modelo o la regresión:

$$\hat{\boldsymbol{Y}} = \boldsymbol{1}\hat{\mu} + \boldsymbol{Z_1}\hat{\alpha}_1 + \boldsymbol{Z_2}\hat{\alpha}_2 \quad (3)$$

se queda sin solución única por la posibilidad de múltiples formas para descomponer $\hat{\boldsymbol{Y}}$, que si resulta única, lleva a una multicolinealidad exacta, Figura 2, pues $\hat{\boldsymbol{Y}}$ queda en el mismo plano de z_1 y z_2 .

Para resolver el problema anterior hay dos soluciones:
1- Hacer restricciones sobre los parámetros o,
2- eliminar variables.

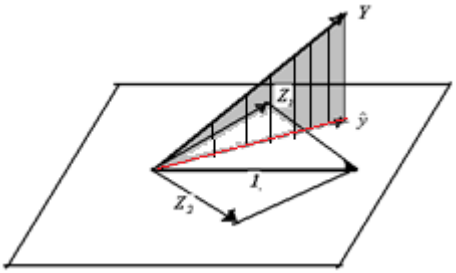


Figura 2 Descomposición de Y con dos variables ficticias (multicolinealidad exacta)

En el diseño experimental se opta por la primera y se supone que. $\sum \tau_i = \sum \alpha_i = 0$, en cuyo caso: $\alpha_1 + \alpha_2 = 0$ con lo cual

$$\hat{\boldsymbol{Y}} = \boldsymbol{1}\hat{\mu} + \boldsymbol{Z_1}\hat{\alpha}_1 + \boldsymbol{Z_2}\left(-\hat{\alpha}_1\right) = \boldsymbol{1}\hat{\mu} + \left(\boldsymbol{Z_1} - \boldsymbol{Z_2}\right)\hat{\alpha}_1 = \boldsymbol{1}\hat{\mu} + \boldsymbol{Z_3}\hat{\alpha}_1 \quad (4)$$

con $\boldsymbol{Z_3}$ variable con valor 1 para el primer grupo y -1 para el segundo por lo que quedan $\boldsymbol{1}\ y\ \boldsymbol{Z_3}$ ortogonales, con lo cual:

$$\hat{\alpha}_1 = \frac{\sum yz_3}{\left(\sum z_3\right)^2} = \frac{1}{n}\left(n_1\hat{y}_{1\bullet} - n_2\hat{y}_{2\bullet}\right),$$

mide la diferencia entre los grupos.

La segunda solución plantea la eliminación de una de las variables en la ecuación (3), por ejemplo:

$$\hat{\boldsymbol{Y}} = \boldsymbol{1}\mu^* + \boldsymbol{Z_2}\alpha^* \quad (5)$$

en que las variables no son ortogonales, pero permiten relacionar los parámetros de este modelo con los del original (3) por medio de $\boldsymbol{E}\left[\boldsymbol{Y}\,|\,\boldsymbol{grupo\ 1}\right] = \mu^* = \mu + \alpha_1$; y $\boldsymbol{E}\left[\boldsymbol{Y}\,|\,\boldsymbol{grupo\ 2}\right] = \mu^* + \alpha^* = \mu + \alpha_2$, en las cuales si los estimadores mínimo cuadráticos de (5) son $\hat{\mu}^*\ y\ \hat{\alpha}^*$,se puede obtener: $\hat{\mu} = \hat{\mu}^* = \mu + \frac{\hat{\alpha}^*}{2}$; $\quad y \quad \hat{\alpha}_1 = -\frac{\hat{\alpha}^*}{2} = -\hat{\alpha}_2$

que reproduce el análisis de varianza con un factor al estudiarlo como una regresión.

Modelo lineal generalizado

Se presenta como:

$$\boldsymbol{Y} = \boldsymbol{Z\theta} + \boldsymbol{U} \quad (6)$$

$$\boldsymbol{H\theta} = \boldsymbol{C} \quad (7)$$

Y, vector conocido de dimensión nx1 de valores independientes y con σ^2 igual a los residuales, $\boldsymbol{Z}_{nxp}$, de rango $h \leq p$,p número de variables explicatorias, $\boldsymbol{\theta}$ vector de p parámetros desconocidos, $\boldsymbol{U}_{nx1}$ variables independientes no observadas con distribución $N\left(o, \sigma^2\right)$, $\boldsymbol{H}_{r\,xp}$, matriz de restricciones lineales con *r* < *p* y su rango = r y, ***C*** un vector conocido de constantes. El número de parámetros independientes es por tanto k = p – r.

Se resaltan dos posibles modelos más importantes o conocidos:
1- El modelo clásico de regresión, ***Z*** incluye valores de variables continuas tales que $\boldsymbol{Z'Z}$ resulten no singulares, $p \leq n$
2- ***Z*** incluye valores de variables discretas (0, 1 y -1) con $\boldsymbol{Z'Z}$ singular, $p \geq n$ por lo cual necesita restricciones en que $n \geq p - r$. Ejemplo, el modelo de dos factores con interacción, i muestras en el primero y j en el segundo, por lo que $\boldsymbol{\theta}$ queda con dimensión *p = i + j + ij.*

Estimación.

El modelo será estimable cuando existan, al menos, tantos datos como parámetros, en cuyo caso la estimación máximo verosímil se reduce a la de mínimos cuadrados, por la normalidad.
Lo anterior implica encontrar un vector $\hat{\boldsymbol{\theta}}$ tal que:

$$\boldsymbol{Z'}\left(\boldsymbol{Y} - \boldsymbol{Z}\hat{\theta}\right) = \boldsymbol{Z'e} = \boldsymbol{0} \quad (8)$$

o sea con residuales independientes. Con ello, como en cualquier modelo de regresión

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'Y} \quad (9)$$

Sí ***Z*** es singular pero existen restricciones, $\hat{\boldsymbol{\theta}}_{\boldsymbol{t}}^{*} = \hat{\boldsymbol{\theta}} - \boldsymbol{A}\left(\boldsymbol{H}\hat{\boldsymbol{\theta}} - \mathbf{C}\right)$ estimador con restricciones, $\boldsymbol{\theta}$ será el hallado en (4.9) y **A** para correcciones del estimador. Se debe entonces encontrar un estimador que satisfaga (4.7) y (4.8) al mismo tiempo. Para ello

$$\boldsymbol{H}\hat{\boldsymbol{\theta}}_{\mathrm{r}}^{*} = \boldsymbol{C} \quad (4.10)$$

Multiplicando por ***H'***

$$\boldsymbol{H'H}\hat{\boldsymbol{\theta}}_{\mathrm{r}}^{*} = \boldsymbol{H'C} \quad (4.11)$$

y si el estimador satisface (4.8),

$$\boldsymbol{Z'Z}\hat{\boldsymbol{\theta}}_{\mathrm{r}}^{*} = \boldsymbol{Z'Y} \quad (4.12)$$

(4.11) y (4.12) no tienen solución única, ya que ***H'H*** y ***Z'Z*** son singulares pero, si se suman como:

$$\left(\boldsymbol{Z'Z} + \boldsymbol{H'H}\right)\hat{\boldsymbol{\theta}}_{\mathrm{r}}^{*} = \boldsymbol{Z'Y} + \boldsymbol{C} \quad (4.13)$$

y la matriz $\left(\boldsymbol{Z'Z + H'H}\right)$ tiene rango p, (cuando las filas de la matriz $\boldsymbol{Z}$ sean linealmente independientes de las de $\boldsymbol{H}$, o sea que la matriz $\begin{bmatrix}\boldsymbol{Z}\\ \boldsymbol{H}\end{bmatrix}$ sea de rango p), entonces se alcanza la solución

$$\hat{\boldsymbol{\theta}}^{*}_{\mathrm{r}} = \left(\boldsymbol{Z'Z + H'H}\right)^{-1}\left(\boldsymbol{Z'Y + C}\right) \quad (4.14)$$

que cubre todos los diseños experimentales conocidos.

Contrastes. El contraste principal en los modelos lineales es que un coeficiente o vector de coeficientes sea 0, realizables por dos métodos distintos:

1- Utilizando la distancia del estimador; dado un vector

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} \rightarrow \boldsymbol{N}_{\boldsymbol{h}}\left(\theta, \boldsymbol{M}\sigma^{2}\right)^{-1} \quad (4.15)$$

su contraste $\boldsymbol{\theta}_{\boldsymbol{1}} = 0$ se analiza con una F:

$$\mathrm{F} = \frac{\hat{\boldsymbol{\theta}}'_{\boldsymbol{1}}\boldsymbol{M}^{-1}\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}}{\mathrm{h}\hat{\mathrm{s}}^{2}} \quad (4.16)$$

2- Se descompone la variabilidad total en fuentes de variación y se comparan las variabilidades explicadas por cada factor. Ejemplo, sean variables de media 0, $\hat{\boldsymbol{\theta}}$ vector de parámetros ortogonales entre sí:

$$\mathrm{Y}=\mathrm{Z}\hat{\boldsymbol{\theta}}+\mathrm{e}=\mathrm{Z}_{1}\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} + \mathrm{Z}_{2}\hat{\boldsymbol{\theta}}_{\boldsymbol{2}} + \ldots + \mathrm{Z}_{\mathrm{m}}\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} + e \quad (4.17)$$

en que $\boldsymbol{Z'Z = 0}$ es el procedimiento clásico de ANAVA y la variabilidad total VT:

$$\boldsymbol{VT = Y'Y} = \hat{\boldsymbol{\theta}}'_{\boldsymbol{1}}\left(\boldsymbol{Z}'_{\boldsymbol{1}}\boldsymbol{Z}_{\boldsymbol{1}}\right)\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} + \hat{\boldsymbol{\theta}}'_{\boldsymbol{2}}\left(\boldsymbol{Z}'_{\boldsymbol{2}}\boldsymbol{Z}_{\boldsymbol{2}}\right)\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} + \cdots \hat{\boldsymbol{\theta}}'_{\mathrm{m}}\left(\boldsymbol{Z}'_{\boldsymbol{m}}\boldsymbol{Z}_{\boldsymbol{m}}\right)\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} + \boldsymbol{e'e} \quad (4.19)$$

y, la variablidad explicada (MSTr) por el vector de parámetros $\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ de dimensión h será:

$$\mathrm{VE}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}\right) = \hat{\boldsymbol{\theta}}'_{\boldsymbol{1}}\left(\boldsymbol{Z}'_{\boldsymbol{1}}\boldsymbol{Z}_{\boldsymbol{1}}\right)\hat{\boldsymbol{\theta}}_{\boldsymbol{1}} \quad (4.20)$$

cuyo contraste =0 se verifica por F:

$$\mathrm{F} = \frac{\hat{\boldsymbol{\theta}}'_{\boldsymbol{1}}\left(\boldsymbol{Z}'_{\boldsymbol{1}}\boldsymbol{Z}_{\boldsymbol{1}}\right)\hat{\boldsymbol{\theta}}_{\boldsymbol{1}}}{\mathrm{h}\hat{\mathrm{s}}^{2}_{\mathrm{R}}} \quad (4.21)$$

MANOVA

Cuando se usan más de dos variables dependientes, que ya escapa al alcance de este curso, se usa el MANOVA. Cuando los factores tengan apenas dos niveles se pueden entrar tanto datos categóricos como cuantitativos. Sin embargo, se adelantan algunos conceptos.

Para cada efecto se han diseñado estadísticos para verificar si o no hay efectos entre todas las VD atribuibles a tal factor. Las matrices se basan en las ***SS*.**. sumas de cuadrados corregidas de productos cruzados, atribuibles a las hipótesis consideradas (H) y con respecto a las **SS** de residuales (E) como los siguientes:

Lambda de Wilks, basado en la relación de dos determinantes:

$\frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$ =?* al cual le encontró $\frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \Lambda^{*}$ que se comporta como una χ^{2}

La **traza de Pillai** calculada como $\mathrm{tr}|\mathbf{H(H+E)}^{-1}|$,

la **traza de Hotelling-Lawley** calculada como $\mathrm{tr}|\mathbf{HE}^{-1}|$,

o la **mayor raíz de Roy** = $MMR = \frac{\lambda_1}{1+\lambda_1}$, λ_1 el mayor valor propio de la matriz de Hotelling-Lawley. Este estadístico se acompaña de los valores de *s, m, y n*, (desviaciones, número de factores y número de datos) usados para calcular valores de F y aun para otros estadísticos. Se debe notar que las pruebas son exactas sí *s = 1 o 2* y *aproximadas cuando se acercan estos valores.* Los 3 primeros estadísticos se muestran junto con las pruebas de F. Pequeños valores del *P-Value* < 0.05 indican efectos significantes al 95% de confianza. De todos modos este tema debe abordarse dentro de la estadística multivariada.

Ejemplo, se muestra un experimento para ver cómo influyen dos tipos de luminosidad (alta y baja), dos humedades (<30% y >al 70%) y temperatura (<15º y >25 º) en la biomasa de la raíz de algunas plántulas a los dos meses de sembradas, pero se perdieron algunos datos y un factorial 2^3 no se pudo efectuar, resultando la siguiente información, Tabla 1:

Tabla 1, Biomasa de raíz a dos meses de sembradas en distintas condiciones de X1=iluminación, X2=humedad y X3=temperatura.

	Y	x1	x2	x3
1	81.90	-1	-1	-1
2	85.68	-1	-1	-1
3	71.82	1	-1	-1
4	126.00	-1	1	-1
5	75.60	1	1	-1
6	69.30	-1	-1	1
7	63.00	1	-1	1
8	103.32	1	-1	1
9	161.28	-1	1	1
10	90.72	1	1	1
11	137.34	1	1	1

Al correr el modelo convencional se encuentra que no se dan interacciones significativas como se muestra enseguida:

```
X1<-as.factor(X1)
X2<-as.factor(X2)
X3<-as.factor(X3)

model1<-glm(Y~X1*X2*X3)
summary(model1)

Call:
glm(formula = Y ~ X1 * X2 * X3)

Deviance Residuals:
     1       2       3       4       5       6       7       8       9      10
 -1.89    1.89    0.00    0.00    0.00    0.00  -20.16   20.16    0.00  -10.71
    11
 10.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    83.79      13.22   6.336  0.00795 **
```

```
X11           -11.97      22.91  -0.523  0.63741
X21            42.21      22.91   1.843  0.16260
X31           -14.49      22.91  -0.633  0.57193
X11:X21       -38.43      34.99  -1.098  0.35231
X11:X31        25.83      32.39   0.797  0.48352
X21:X31        49.77      34.99   1.422  0.25006
X11:X21:X31   -35.28      47.68  -0.740  0.51301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 349.8012)

    Null deviance: 8590.4  on 10  degrees of freedom
Residual deviance: 1049.4  on  3  degrees of freedom
AIC: 99.356

Number of Fisher Scoring iterations: 2
```

Muestra que parece imponerse el modelo nulo, pero a pesar de ello se corre el modelo sin interacciones

```
model2<-glm(Y~X1+X2+X3)
> summary(model2)

Call:
glm(formula = Y ~ X1 + X2 + X3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-30.6600  -14.2553    0.9882    9.8329   29.6718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    84.69      12.53   6.757 0.000263 ***
X11           -26.31      14.10  -1.866 0.104298
X21            35.08      13.66   2.568 0.037091 *
X31            15.27      14.10   1.083 0.314792
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 500.7833)

    Null deviance: 8590.4  on 10  degrees of freedom
Residual deviance: 3505.5  on  7  degrees of freedom
AIC: 104.62

Number of Fisher Scoring iterations: 2
```

Se actualiza el model 2, puesto que X2 resultó significativa

```
model3<-update(model2,~.-X1 -X3)
> summary(model3)

Call:
glm(formula = Y ~ X2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-37.548  -13.020   -1.008    9.681   48.132

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    79.17      10.04   7.887 2.48e-05 ***
X21            33.98      14.89   2.282   0.0484 *
 (Dispersion parameter for gaussian family taken to be 604.6345)

    Null deviance: 8590.4  on 10  degrees of freedom
```

```
Residual deviance: 5441.7  on  9  degrees of freedom
AIC: 105.46

Number of Fisher Scoring iterations: 2
```

Coeficientes del Modelo. Como se mencionó al principio en este diseño subyace un modelo lineal de la forma:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i \quad (12)$$

en el cual Y es la variable dependiente, las variables Xs, portadoras de la información de cada uno de los efectos en el modelo y, los ε, asumidos como variables aleatorias normalmente e independientemente distribuidas con media cero, que al acudir al Modelo Lineal Generalizado para una variable dependiente y 3 factores categóricos, **A = X_1, B = X_2 y C= X_3**, mostró que el modelo no fue significativo.

ANALISIS DE DEVIANCE. 3a interpretación (Como similar a un anava)

Es una generalización del ANAVA para los GLM, obtenido para una secuencia de modelos anidados (cada uno incluye más términos que el anterior). Para analizar tal secuencia, usamos la *deviance* como una medida de discrepancia entre ellos, así podemos hablar de diferencias de *deviance* (generalmente van a una tabla), Se explicará más de una vez en diferentes contextos.

Sea Mp_1, Mp_2, . . ., Mp_r, una sucesión de modelos anidados de dimensión $p_1 < p_2 < \ldots < p_r$ (suponemos en todos ellos la misma distribución y la misma función encadenante), unas matrices de diseño: Xp_1, Xp_2, . . . , Xp_r y, las deviances $Dp_1 > Dp_2 > \ldots > Dp_r$.

Matrices de diseño para el modelo lineal general (GLM)

El estimador lineal del GLM (una regresión) se basa en una matriz de diseño a partir de los factores, las covariables y el modelo especificado, en que las bloque-columnas son las variables predictores de la regresión en sus diversos niveles).

La matriz, tiene n filas (número de observaciones), y un bloque de columnas (uno, con cada variable indicadora, para cada término del modelo). Tantas columnas como grados de libertad tengan el término.

La primera columna bloque corresponde a la constante y está formada por unos (1). La columna bloque para una covariable también solo tiene a ella.

Ejemplo: Sea **A**, factor con 4 niveles y una codificación con (#niveles -1) ej: -1, 0, +1. Entonces tiene 3 grados de libertad y su bloque contiene 3 columnas, llamémoslas a1, a2, a3. Cada columna se codifica de la siguiente manera:

Nivel de A	a1	a2	a3
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

Sea el factor **B** con 3 niveles anidados dentro de cada nivel de **A**. Entonces su bloque contiene (3 - 1) x 4 = 8 columnas, llamémoslas b_{11}, b_{12}, b_{21}, b_{22}, b_{31}, b_{32}, b_{41}, b_{42}, con la siguiente codificación:

Nivel de A	Nivel de B	b11	b12	b21	b22	b31	b32	b41	b42
1	1	1	0	0	0	0	0	0	0
1	2	0	1	0	0	0	0	0	0
1	3	-1	-1	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0	0
2	2	0	0	0	1	0	0	0	0
2	3	0	0	-1	-1	0	0	0	0
3	1	0	0	0	0	1	0	0	0
3	2	0	0	0	0	0	1	0	0
3	3	0	0	0	0	-1	-1	0	0
4	1	0	0	0	0	0	0	1	0
4	2	0	0	0	0	0	0	0	1
4	3	0	0	0	0	0	0	-1	-1

Para calcular las variables indicadoras para un término de la interacción se multiplican todas las variables simuladas correspondientes por los factores y/o las covariables en ella. Por ejemplo, sea el factor A con 6 niveles, C con 3, D con 4 y, Z y W covariables. Entonces el término A * C * D * Z * W * W tiene 5 x 2 x 3 x 1 x 1 x 1 = 30 variables indicadoras resultado de multiplicar cada variable indicadora de A por cada una de C, por cada una de D, por las covariables Z una vez y W otra.

Deviance. cuarta interpretación.

La diferencia **Dpi − Dpj** , pj > pi, es interpretable como una medida de la variación de los datos explicada por los términos que están en Mpj y no estan en Mpi, incluidos los efectos de los términos que están en Mpi e ignorando los efectos de cualquier término que no está en Mpj . De esta manera, ya conocido el contraste de verosimilitud:

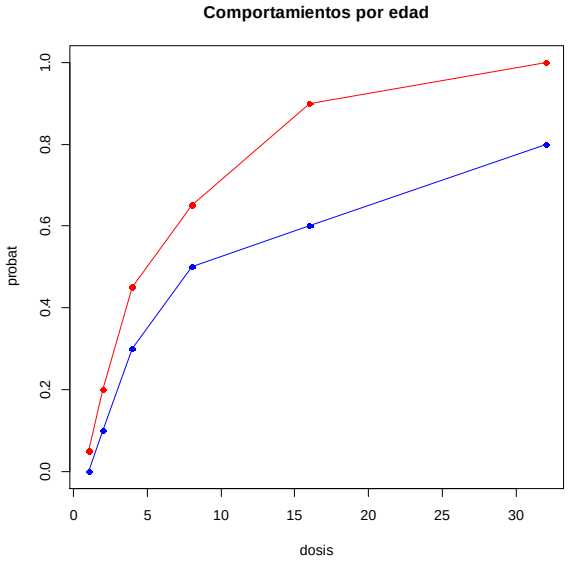
$$G = D_{MV}(M_1) - D_{MV}(M_2) = -2 \ln\left[\frac{\text{Verosimilitud 1}}{\text{Verosimilitud 2}}\right],$$

el cual se distribuye como una χ^2 cuyos grados de libertad equivalen a la diferencia de parámetros entre modelos, si $D_{pi} - D_{pj} > \chi^2_{pj-pi,\alpha}$ los efectos de los términos que están en Mpj y no están en Mpi son significativos. Cada secuencia de modelos corresponde a una tabla de análisis de la varianza diferente. La secuencia de los modelos estará determinada por el interés del investigador.

Ejemplo: se hizo un experimento para mirar el resultado de un fungicida para el control de la mancha azul, en trozas recientemente cortadas de P. patula de hasta 10 años de edad, dejadas en el bosque, regadas con agua en diluciones de, 1, 2, 4, 8, 16 y 32 (µg/lit) del fungicida, por edad (antejuveniles=A < 5 años y juveniles=J) para lo cual se dispuso en el bosque de 40 árboles en cada uno de 20 sitios (numerados de 0 a 20) de los cuales se seleccionaron 12 para instalarlos por edad así: juveniles en los sitios 1, 4, 9, 13, 18, 20; antejuveniles en los 0, 2, 6, 10, 12 y 16 con los siguientes resultados de del ataque del hongo, con ataque y sin ataque:

	dosis	atac	edad	noatac
1	1	2	J	38
2	2	8	J	32
3	4	18	J	22
4	8	26	J	14
5	16	36	J	4
6	32	40	J	0
7	1	0	A	40
8	2	4	A	36
9	4	12	A	28
10	8	20	A	20
11	16	24	A	16
12	32	32	A	8

```
fungi<-read.table("clipboard")
attach(fungi)
names(fungi)
[1] "dosis"  "atac"   "edad"   "noatac"
ldosis<-rep(0:5,2)#creamos una variable entera entre 0 y 5 para llamar las dosis
probat<- atac/40#probabilidad de elementos atacados.
plot(dosis, probat,type="n",xlab="dosis",ylab="probat",main="Comportamientos por edad")
lines(dosis[edad=="J"],type="p", probat[edad=="J"],col="red",pch=16)
lines(dosis[edad=="J"],type="l", probat[edad=="J"],col="red",pch=16)
lines(dosis[edad=="A"],type="p", probat[edad=="A"],col="blue",pch=16)
lines(dosis[edad=="A"],type="l", probat[edad=="A"],col="blue",pch=16)
```



Según esta gráfica la probabilidad de ataque se ve diferente, más severo para los juveniles que para los árboles menores de 5 años.

Queremos investigar la posibilidad de que haya diferentes pendientes para las dos edades. Para ello plantearemos y ajustaremos el modelo logístico, ya conocido:

$\text{logit}(\pi i) = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{ldosis} + \beta_3 \text{ edad:ldosis}$

atyna<-cbind(atac,noatac)#generamos ambas variables conjuntamente con atacados y no atacados.

modlo<- glm(atyna~edad*ldosis, family=binomial)

```
summary(modlo)

Call:
glm(formula = atyna ~ edad * ldosis, family = binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9778  -0.4539  -0.1074   0.5405   1.5609

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.9935     0.3908  -7.660 1.86e-14 ***
edadJ          0.1750     0.5503   0.318   0.7505
ldosis         0.9060     0.1182   7.668 1.75e-14 ***
edadJ:ldosis   0.3529     0.1909   1.849   0.0645 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 249.7512  on 11  degrees of freedom
Residual deviance:   9.9875  on  8  degrees of freedom
AIC: 54.755

Number of Fisher Scoring iterations: 4
```

Los estimados (coeficientes de los predictores dosis y edad están ahora en las unidades logit vistas en la regresión logística.

Deviance: Segunda interpretación.

Ya vimos que es una medida de bondad de ajuste de un GLM, o por el contrario de maldad de ello: mientras mayor indica peor ajuste. El modelo muestra: *null deviance* y *la residual deviance*. La nula muestra que tan bien es predicha la respuesta(y) por un modelo que apenas incluye el intercepto (o sea la media total).

Para nuestro ejemplo obtuvimos: *Null deviance* 249.8 con 11 grados de libertad, pero al incluir las variables predictoras (edad y ldosis) decreció la *Residual deviance* a 9.9 con 8 grados de libertad, o sea una gran reducción, cerca de 240 puntos con una pérdida de 3 grados de libertad, mostrando la mejoría del modelo.

.

Indice de Fisher (*Fisher Scoring*).

Este indicador es una derivada del método de Newton para la solución numérica de problemas de máxima verosimilitud. Acá nos muestra que se necesitaron 4 iteraciones para alcanzar el ajuste, lo que realmente no dice mucho más que, el modelo no tiene problemas para converger.

Criterios de información.

El AIC, ya visto propicia un método para evaluar la calidad de nuestro modelo mediante la comparación de dos modelos relacionados, basados en la *Deviance*, pero penalizados por hacer los modelos más complicados o sea por alterar la parsimonia, similar a nuestro R2 ajustado para prevenir de variables predictoras irrelevantes. Así elegimos el modelo que entre modelos similares más complejos presente el menor AIC.

Para el ejemplo dado, aparentemente de la lectura de la tabla, el efecto de la edad parece no significativo, sin embargo, debemos ser cuidadosos al interpretar esto. Dado que estamos ajustando distintas pendientes para cada edad, el test individual para este

parámetro verifica la hipótesis de que las curvas no difieren cuando ldosis es 0. Vamos a reparametrizar de manera de incluir el intercepto en una dosis central (3).

Queremos investigar la posibilidad de que haya diferentes pendientes para las dos edades. Para ello plantearemos y ajustaremos el modelo:

$\text{logit}(\pi_i) = \beta\ 0 + \beta\ 1\text{edad} + \beta\ 2\text{ldosis} + \beta\ 3\ \text{edad:ldosis}$

```
pat2<- glm(atyna~edad*I(ldosis-3), family=binomial)
summary(pat2)

Call:
glm(formula = atyna ~ edad * I(ldosis - 3), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9778  -0.4539  -0.1074   0.5405   1.5609

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.2754     0.1630  -1.690   0.0911 .
edadJ                1.2337     0.2666   4.628 3.69e-06 ***
I(ldosis - 3)        0.9060     0.1182   7.668 1.75e-14 ***
edadJ:I(ldosis - 3)  0.3529     0.1909   1.849   0.0645 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 249.7512  on 11  degrees of freedom
Residual deviance:   9.9875  on  8  degrees of freedom
AIC: 54.755

Number of Fisher Scoring iterations: 4
```

que muestra los mismos parámetros del anterior, o sea no se deterioró, además de, una diferencia significativa entre los dos sexos en la dosis 3 (central). El modelo ajusta muy bien ya que la DR se comporta como una chi2, entonces al buscar en el R:

(1-pchisq(**9.9875** ,8)= 0.7582464

nos lleva a la hipótesis nula de modelos iguales). Comparamos distintos modelos mediante la instrucción ANOVA y para este solo, nos muestra que ambas variables fueron importantes

Anova(modlo,test="Chisq")

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: atyna

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          11    249.751
edad        1   12.154        10    237.597 0.0004898 ***
ldosis      1  224.083         9     13.514 < 2.2e-16 ***
edad:ldosis 1    3.527         8      9.987 0.0603888 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
nalysis of Deviance Table

Model: binomial, link: logit
```

Ahora ajustamos una pendiente para cada ~~sexo~~:
pat3<- glm(atyna~edad+ldosis-1, family=binomial)

```
> summary(pat3)

Call:
glm(formula = atyna ~ edad + ldosis - 1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.56327  -0.92409  -0.03146   0.68548   2.02154

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
edadA  -3.47316    0.33129 -10.484   <2e-16 ***
edadJ  -2.37241    0.27260  -8.703   <2e-16 ***
ldosis  1.06421    0.09269  11.482   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 252.454  on 12  degrees of freedom
Residual deviance:  13.514  on  9  degrees of freedom
AIC: 56.281

Number of Fisher Scoring iterations: 4
```

Muestra entonces que difieren al calcular el modelo sin intercepto.

Modelos Aditivos generalizados (GAM).

Los *GAM* en R son extensiones no paramétricas de los GLM, utilizables cuando no se tiene una razón a priori para escoger una función de respuesta particular (tales como la lineal, cuadrática, polinomial, etc.) y, uno espera que los datos hablen por sí mismos. Los GAM hacen sus procesos vía funciones de suavización, similar a lo ya conocido acerca de las regresiones ponderadas

Se comportan algo similar a los GLM con diferentes estructuras para el error y diferentes funciones de encadenamiento para conteos y proporciones. La diferencia radica en que la forma de las relaciones entre Y y una variable continua X no es especificada por alguna forma funcional explícita. En lugar de ello se acude a suavizadores no paramétricos para describir las relaciones. Esto es particularmente útil para relaciones que exhiben formas complicadas tales como curvas arqueadas. El modelo es muy similar a un GLM excepto que las relaciones que deseamos son suavizadas en R con el prefijo s (*spline o smooth*): entonces si deseamos por ejemplo una regresión múltiple con 3 variables explicatorias continuas: *w*, *x* y *z* sobre una *Y* (de conteo de datos) acudimos a la función:
model<-gam(y~s(w)+s(x)+s(z),poisson)

Son modelos jerárquicos, de tal manera que la inclusión de una interacción de alto orden como A:B:C, necesariamente implica la inclusión de todas las demás de más bajo orden, marginales a ella, o sea. A:B, A:C y B:C, mediante los efectos principales para A, B y C.

Ya que los modelos son anidados, el modelo más complicado necesariamente explica por lo menos tanto de la variabilidad como lo hace el más simple de los modelos y usualmente, aún más. Lo que resulta importante conocer es, si los parámetros extras en el modelo más complejo se justifican en el sentido de adicionar más poder explicatorio a los modelos. Si ello no ocurre, por parsimonia debemos aceptar el más simple de los modelos. En general así trabaja el alm:

.

Model<-gam(v~s(d)+s(h), family=gaussian(link=identity),data=fore))

Ejemplo, se mostrará con unos datos de ciprés, supongamos una parcela de 1/10 de ha surgida de un inventario, en la cual solo medimos los datos de inventario en una parcela central de 1/40. Los datos son:

	dap	h	v	vest		dap	h	v	vest
1	17.93	19.44	0.198	0.207	17	18.58	18.06	0.219	0.206
2	19.01	17.5	0.189	0.209	18	22.68	20	0.345	0.33
3	23.11	22.5	0.387	0.385	19	23.33	23.06	0.401	0.403
4	23.76	18.33	0.294	0.335	20	23.76	20.83	0.371	0.377
5	23.98	22.22	0.423	0.41	21	24.19	20.83	0.368	0.391
6	24.41	21.94	0.452	0.42	22	24.62	21.11	0.453	0.411
7	24.62	21.11	0.41	0.411	23	25.27	19.17	0.416	0.396
8	25.92	20.83	0.438	0.452	24	27.86	20.56	0.498	0.518
9	27.86	23.61	0.621	0.598	25	28.73	23.89	0.621	0.646
10	29.59	19.72	0.542	0.563	26	29.81	17.78	0.562	0.521
11	30.24	21.67	0.629	0.644	27	30.67	22.22	0.701	0.68
12	31.32	20.56	0.681	0.655	28	34.56	20	0.801	0.764
13	35.21	21.39	0.881	0.842	29	37.37	22.5	0.864	0.983
14	37.8	22.78	0.895	1.016	30	38.66	22.22	1.148	1.029
15	38.88	22.22	1.082	1.039	31	38.88	22.22	1.003	1.039

acip<-read.table("clipboard")
attach(cip)
names(cip)
[1] "dap" "h" "v"
library(mgcv)#bajarla previamente

mod1 <- gam(v~ s(dap) + s(h), family=Gamma(link=log))

```
 mod1
 Family: Gamma
Link function: log
Formula:
v ~ s(dap) + s(h)
Estimated degrees of freedom:
2.75 1.00  total = 4.75
 GCV score: 0.005559895
summary(mod1)
Family: Gamma
Link function: log

Formula:
v ~ s(dap) + s(h)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.64507    0.01222   -52.8   <2e-16 ***
Approximate significance of smooth terms:
         edf Ref.df      F  p-value
```

```
s(dap) 2.752  3.437 282.97  < 2e-16 ***
s(h)   1.000  1.000  30.06 5.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.974   Deviance explained = 98.4%
GCV = 0.0055599  Scale est. = 0.0046265  n = 31
```

Los θ_i son escogidos para minimizar bien sea el indice GCV (GCV score) que se calcula:

$V_g = n \| W(y - Ay) \|^2 / [tr(I - gA)]^2$

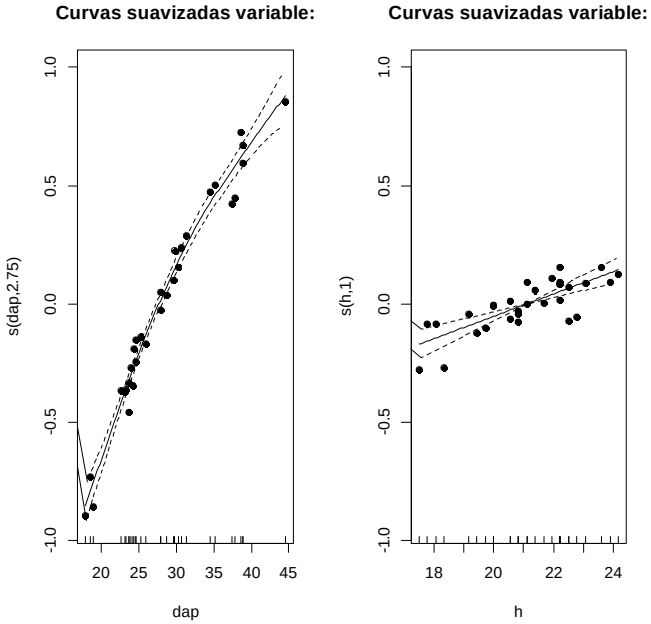
o el UBRE score calculado como:

$V_u = \| W(y - Ay \|^2 / n - 2str(I - gA) / n + s$

En que:
g es el factor de inflación gamma para los gl (usualmente ajustado a 1)
s el parámetro de escala,.
A la matriz sombrero (matriz de influencia o hat matrix) para ajustar el problema (o sea la matriz que mapea los datos con los valores ajustados), y ya sabemos que la dependencia de las coordenadas con los parámetros de suavización es através de esta matriz A.
Veamos los gráficos de los datos, en un ventanazo de 2 columnas:

```
par(mfrow=c(1,2))
plot(mod1, residuals=TRUE,pch=19,main="Curvas suavizadas variable:")
```

Con el gráfico queremos conocer el comportamiento de cada variable explicatoria original y suavizada por modelo:



muestran los efectos estimados (líneas continuas) de cada una de las variables (h en la gráfica 2 y dap en la gráfica 1) y las bandas de confianza al 95 %.

Los puntos de cada grafica son los residuos parciales de cada observación (que en el caso de la Altura (h) y la observación (i), sería el residuo de Pearson de la observación i sumado a $\hat{f}_1(h_i)$

Comentario de graficas (plots) y consecuencias restricción identificabilidad en variables.

En un modelo GLM el residuo de Pearson correspondiente a la observación y_i se define como:

$$\hat{e}_{pi} = \frac{yi - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

siendo el estimador de $\mu_i = E(y_i \mid X)$, y el estimador de V (μi) que se define como V(μi) en que $V(\mu_i) = V(y_i \mid X)/\phi$. Si el modelo ajustado es adecuado estos errores son aproximadamente normales con media 0 y varianza φ.

La opción por defecto de *gam* trabaja siempre con bases de splines de regresión *thin plate* (Banda delgada). Vamos a considerar un nuevo modelo con una base de *splines cúbicos* para la variable h y la variable dap.

```
mod2 <- gam(v~s(dap, bs="cr") + s(h, bs="cr"),
family=Gamma(link=log))
mod2
 Family: Gamma
Link function: log

Formula:
v ~ s(dap, bs = "cr") + s(h, bs = "cr")

Estimated degrees of freedom:
2.73 1.00  total = 4.73

GCV score: 0.005564391
```

Podemos observar que solo se produjeron unos leves cambios en el modelo ajustado. Un buen resultado, pero puede ser por los datos utilizados. No siempre será así. *gam* por defecto trabaja siempre con una base de dimensión k=10 (el numero de nodos). Vamos a considerar un nuevo modelo con una base de *splines* cúbicos de dimensión k=20 para la variable dap y continuamos con la base de splines con la que trabaja gam por defecto para la variable h.

```
mod3 <- gam(v ~ s(h)+s(dap, bs="cr", k=20),family=Gamma(link=log))
mod3

Family: Gamma
Link function: log
Formula:
v ~ s(h) + s(dap, bs = "cr", k = 20)

Estimated degrees of freedom:
1.00 2.76  total = 4.76

GCV score: 0.005560806
```

Trabajamos ahora con una base de mayor dimensión por lo que estamos considerando inicialmente muchos más grados de libertad que si la base fuera de menor dimensión. Nuevamente los resultados apenas cambiaron con respecto a los de los modelos anteriores. Tampoco ocurre siempre. Luego de las primeras valoraciones del modelo, cuando decidimos que es aceptable, podemos profundizar en su análisis:

summary(mod1)# ya lo habíamos visto

```
Family: Gamma
Link function: log

Formula:
v ~ s(dap) + s(h)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.64507    0.01222   -52.8   <2e-16 ***

Approximate significance of smooth terms:
         edf Ref.df      F  p-value
s(dap) 2.752  3.437 282.97  < 2e-16 ***
s(h)   1.000  1.000  30.06 5.79e-06 ***

R-sq.(adj) =  0.974   Deviance explained = 98.4%
GCV = 0.0055599  Scale est. = 0.0046265  n = 31
```

Si queremos los *p-values* para valorar las variables predictoras podemos usar el comando anova:

```
anova(mod1)
Family: Gamma
Link function: log

Formula:
v ~ s(dap) + s(h)
Approximate significance of smooth terms:
         edf Ref.df      F  p-value
s(dap) 2.752  3.437 282.97  < 2e-16
s(h)   1.000  1.000  30.06 5.79e-06
```

Coeficientes estimados($\hat{\beta}_j$)

coef(mod1)#coeficientes estimados

```
  (Intercept)      s(dap).1      s(dap).2      s(dap).3      s(dap).4
-6.450668e-01 -1.685495e-02  3.299203e-02 -1.376863e-02 -3.722404e-02
     s(dap).5      s(dap).6      s(dap).7      s(dap).8      s(dap).9
 1.776909e-02  3.550784e-02  6.125088e-03  2.353423e-01  4.552935e-01
       s(h).1        s(h).2        s(h).3        s(h).4        s(h).5
-1.682880e-08 -1.417975e-07 -1.146927e-07  2.004284e-07  1.632025e-08
       s(h).6        s(h).7        s(h).8        s(h).9
 2.094991e-07 -6.861583e-08  1.087633e-06  8.232264e-02
```

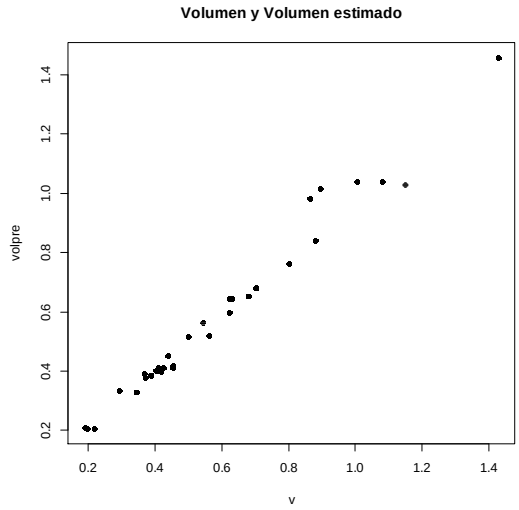
Matriz de varianzas-covarianzas de ($\hat{\beta}_j$)

Vcov <- vcov(mod1)

No se muestra porque la entrega punto por punto.

Predicciones. Si queremos las predicciones en la escala de la variable respuesta:

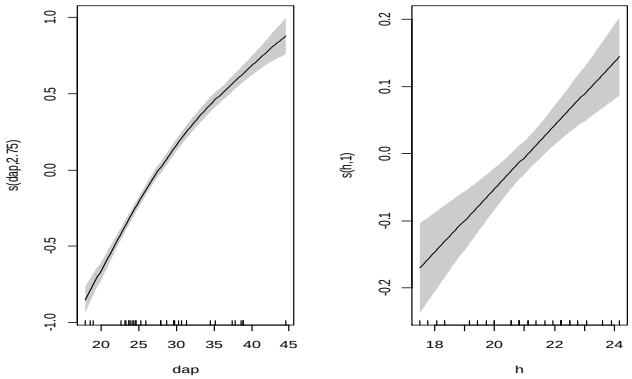
```
volpre<- predict(mod1,type=”response”) # los resultados en la tabla inicial presentada
plot(v,volpre, pch=16, lty=1,lwd=1,main="Volumen y Volumen estimado")
```



El uso más importante de *predict.gam* es para predecir el predictor lineal del modelo para nuevos valores de las variables predictoras:

#creamos un banco de datos con los nuevos valores de las variables predictoras

```
pd <- data.frame(Altura=c(9.34,23.8),dap=c(12.23,31.4))
predict(mod1,newdata=pd)
plot.gam tiene m_as opciones, por ejemplo:
par(mfrow=c(1,2))
plot(mod1,shade=TRUE,seWithMean=TRUE, scale=0)
```



A veces es conveniente visualizar la respuesta esperada en términos de las dos variables predictoras, solo por mostrarlo:

```
vis.gam(mod1,theta=30, type=”response”, ticktype=”detailed”)
vis.gam(mod1,theta=10, type="response", ticktype="detailed")
```