Paquete VGAM. Aplicaciones biomédicas con R

Jenifer Espasandín-Domínguez[1],Carmen Cadarso-Suárez[1], Thomas Kneib[2], Francisco Gude[3].

[1] *Unit of Biostatistics, (CiMUS). University of Santiago de Compostela.*
[2] *Chair of Statistics, Georg-August-Universität Göttingen, Germany.*
[3] *Clinical University Hospital of Santiago de Compostela, Santiago de Compostela.*

Introducción

¿Qué necesitamos para modelizar de forma flexible un conjunto de datos?

- **Distribuciones flexibles** para modelar la variable respuesta,
- Conocer la **distribución completa** de la variable respuesta $\Rightarrow$ Modelar **todos sus parámetros** (no únicamente la media),
- Detectar la **heterogeneidad** existente en los datos,
- **Funciones flexibles** para estudiar la relación existente entre los parámetros de la respuesta y las **covariables**.

Regresión multivariante

- Además, en muchos estudios biomédicos, el interés radica en la modelización de respuestas que conforman un **vector multivariante**:
  - Misma patología en ambos ojos (Retinopatía Diabética).
  - Diferentes patologías en un mismo paciente (Diabetes, HTA).
  - Tensión sistólica y diastólica en el mismo individuo...
- También es de interés la estructura de dependencia entre las respuestas, así como el efecto de covariables en dicha estructura.

VGAM, Yee and Wild (1996)

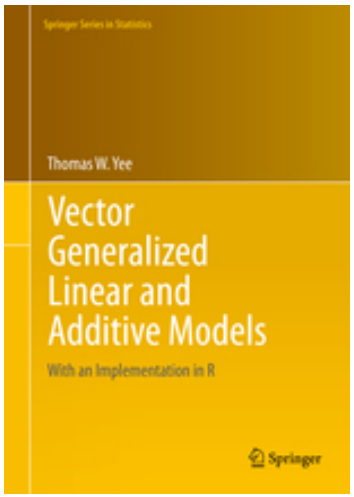
En la literatura estadística se han propuesto varias **alternativas flexibles** que permiten incorporar:

- *respuestas multivariantes*,
- *modelar la distribución completa de la variable respuesta*,
- *preditores flexibles*.

En este trabajo, se presentarán modelos **GAMs vectoriales**

⬇

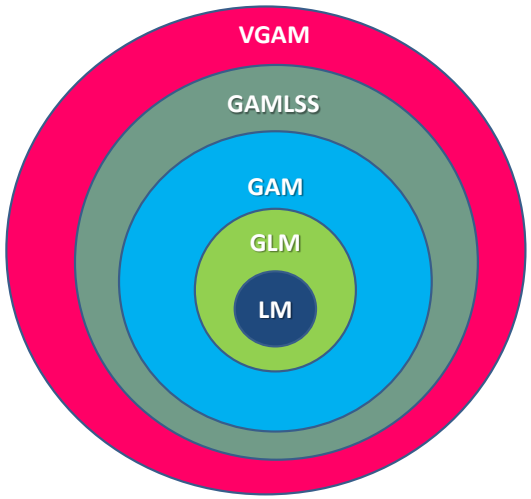
VGAMs, **Vector Generalized Linear and Additive Models**.



VGAM

Yee y Wild (1996) extienden la formulación GAM de Hastie y Tibshirani (1990), en dos direcciones:

- **Respuestas Univariantes Generalizadas:** distribuciones más allá de la familia exponencial.
  - *Supone una extensión de los GAMLSS (GAM for Location, Scale and Shape, Rigby and Stasinopoulos, 2005).*
  - *Respuestas*: tasas, cuantiles, proporciones, supervivencia...

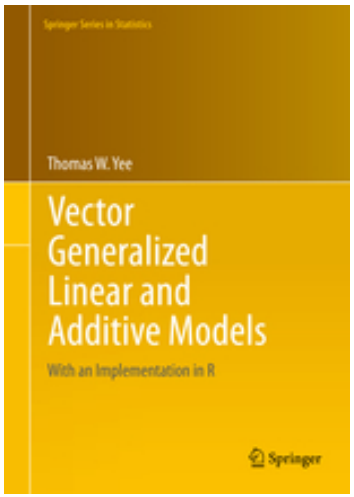


- **Respuestas Multivariantes:** La respuesta es un vector.

Monografía VGAM

Yee TW. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer.

- **Algoritmo**: Backfitting vectorial (Yee and Wild, 1996).
- **Suavizadores**: Vector Splines (Fessler, 1991; Yee, 1993).
- **Selección automática del grado de suavización:** No.



Aplicación en R de los modelos VGAM

- **Diabetes:** Estudio de la correlación entre proteínas glicadas $\Rightarrow$ (*Modelo VGAM con respuesta Gaussiana*).
- **Software:** R $\Rightarrow$ Package: VGAM.

Diabetes: Correlación entre proteinas glicadas

- Los endocrinólogos están interesados en la diagnosis y control de la Diabetes.
- Existen dos métodos para la diagnosis de pacientes diabéticos:
  - Glucosa en sangre.
  - Hemoglobina glicada (A1c).
- Otras proteínas pueden glicar, entre ellas la Fructosamina (Fru). Sin embargo, la correlación entre estas proteínas, y la glucosa, no es perfecta.
- Existen factores que pueden modular $(A1c, Fru)$ y su correlación...

Presentación de los datos

Base de datos:

1. n=612 individuos de la población general (A Estrada, Galicia).

Variables Respuesta:

1. **Vector Respuesta:** $\Rightarrow (A1c, Fru)$.

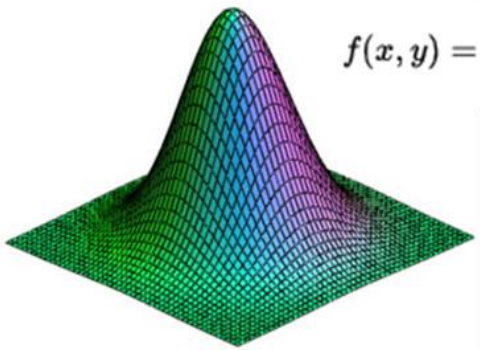
Covariables:

1. Edad (*edad*) y género de los pacientes (*sex*).
2. Volumen Corpuscular medio (mcv).
3. Glucosa basal (glub).

Modelo bivariante Normal

Vector de Respuestas

Supondremos que el vector de respuestas $(A1c, Fru)$ sigue una distribución **normal bivariante**, con 5 parámetros $(\mu_1, \sigma_1, \mu_2, \sigma_2, \rho)$, siendo ρ el coeficiente de correlación de Pearson.

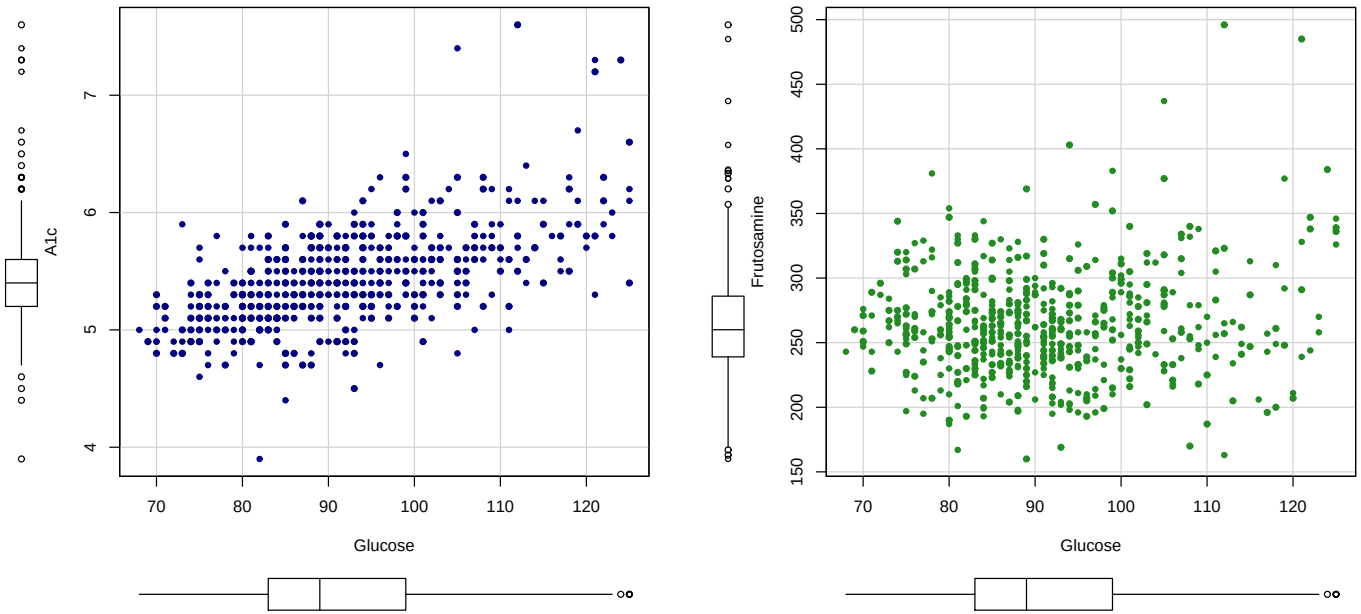


$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

Respuesta

Vector Respuesta: $(Y_{i1}, Y_{i2})' = (A1c_i, fru_i)$ medidos en % y $\frac{mol}{l}$.



Especificación del modelo

Modelo

$$\eta_i^{\mu_1} = \beta_0^{\mu_1} + f_1^{\mu_1}(edad)_i + f_2^{\mu_1}(glu)_i + f_3^{\mu_1}(mcv)_i,$$
$$\eta_i^{\mu_2} = \beta_0^{\mu_2} + f_1^{\mu_2}(edad)_i + f_2^{\mu_2}(glub)_i + f_3^{\mu_2}(mcv)_i,$$

$$\eta_i^{\sigma_1^2} = \beta_0^{\sigma_1^2} + f_1^{\sigma_1^2}(edad)_i + f_2^{\sigma_1^2}(glub)_i + f_3^{\sigma_1^2}(mcv)_i,$$
$$\eta_i^{\sigma_2^2} = \beta_0^{\sigma_2^2} + f_1^{\sigma_2^2}(edad)_i + f_2^{\sigma_2^2}(glub)_i + f_3^{\sigma_2^2}(mcv)_i,$$

$$\eta_i^{\rho} = \beta_0^{\rho} + f_1^{\rho}(edad)_i + f_2^{\rho}(glub)_i + f_3^{\rho}(mcv)_i.$$

Notas

- **Links**: identity (para μ), log (para σ), rhobit (para ρ).
- **Software**: VGAM (Yee and Wild, 1996) con Splines Vectoriales.

Modelización

R Código

```
> y =  cbind(data$a1c,data$fru)
> fit =  vgam(y~sex+bs(edad)+bs(mcvc)+bs(gluc),
na.action=na.omit,family=binormal(zero=NULL),
 control=vgam.control(maxit=100),trace=T,data)
```

Binormal

```
> args(binormal)
binormal(lmean="identitylink", lmean2="identitylink",
lsd1="loge", lsd2="loge", lrho= "rhobit",
imean1=NULL, imean2=NULL, isd1=NULL, isd2=NULL, irho=NULL,
eq.mean = FALSE, eq.sd=FALSE, zero=c("sd", "rho"))
```

Summary

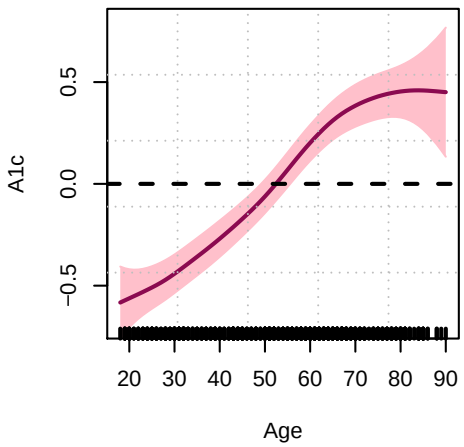
```
> summary(fit)

Call:
vgam(formula = y ~ sex + s(edad) + s(mcvc) + s(glub),
family = binormal(zero = NULL),
data = data, na.action = na.omit, control = vgam.control(maxit = 100),
model = T, trace = T)

Number of linear predictors: 5
Names of linear predictors: mean1, mean2, loge(sd1), loge(sd2), rhobit(rho)
Dispersion Parameter for binormal family: 1

Log-likelihood: -956.8785 on 2684.513 degrees of freedom

Number of iterations: 79
```

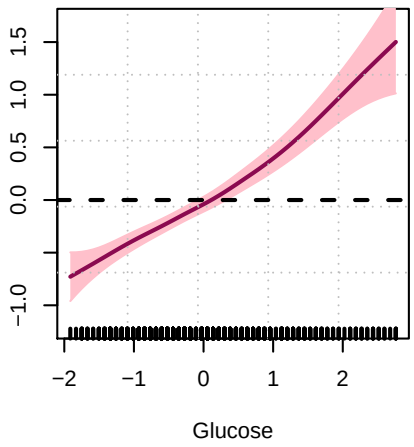
Summary

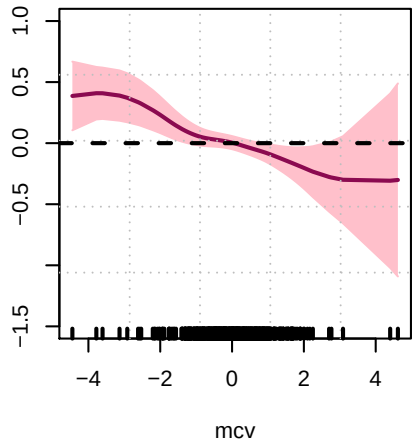
DF for Terms and Approximate Chi-squares for Nonparametric Effects

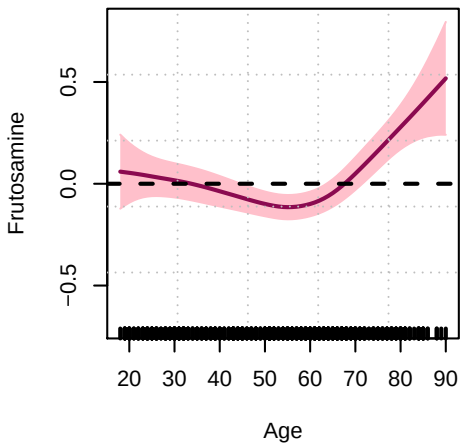
	Df	Npar Df	Npar Chisq	P(Chi)
(Intercept):1	1			
(Intercept):2	1			
(Intercept):3	1			
(Intercept):4	1			
(Intercept):5	1			
sex:1	1			
sex:2	1			
sex:3	1			
sex:4	1			
sex:5	1			
s(edad):1	1	3.0	6.422	0.09476
s(edad):2	1	3.0	32.910	0.00000
s(edad):3	1	3.0	10.562	0.01445
s(edad):4	1	3.0	8.614	0.03511
s(edad):5	1	3.0	1.687	0.63952
s(mcvc):1	1	3.4	4.205	0.28872
s(mcvc):2	1	3.3	22.680	0.00007
s(mcvc):3	1	3.0	12.033	0.00736
s(mcvc):4	1	3.0	11.737	0.00844
s(mcvc):5	1	3.0	5.967	0.11398
s(glub):1	1	2.8	5.359	0.12728
s(glub):2	1	2.9	15.669	0.00124
s(glub):3	1	3.0	0.779	0.85660
s(glub):4	1	3.0	14.775	0.00206
s(glub):5	1	3.0	3.404	0.33550

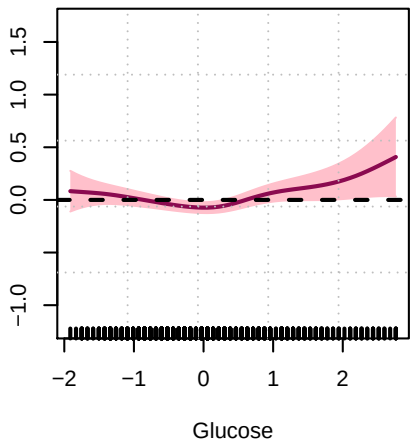
Medias

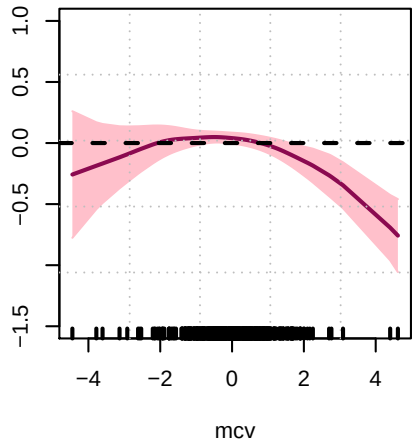

A1c
0.5
0.0
−0.5
20 30 40 50 60 70 80 90
Age


1.5
1.0
0.5
0.0
−1.0
−2 −1 0 1 2
Glucose

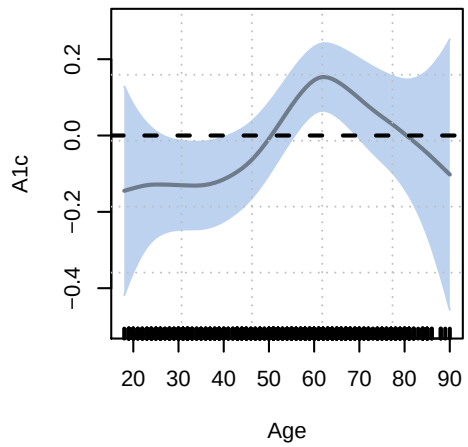

1.0
0.5
0.0
−0.5
−1.5
−4 −2 0 2 4
mcv

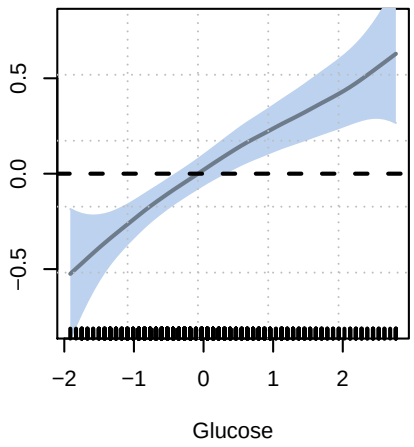

Frutosamine
0.5
0.0
−0.5
20 30 40 50 60 70 80 90
Age

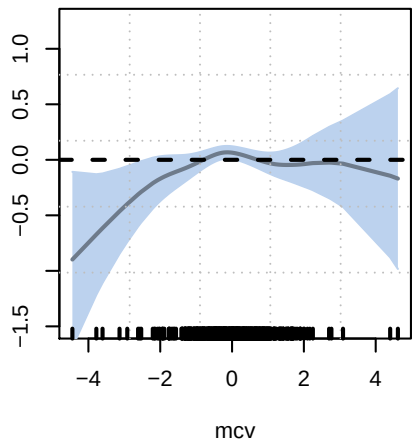

1.5
1.0
0.5
0.0
−1.0
−2 −1 0 1 2
Glucose

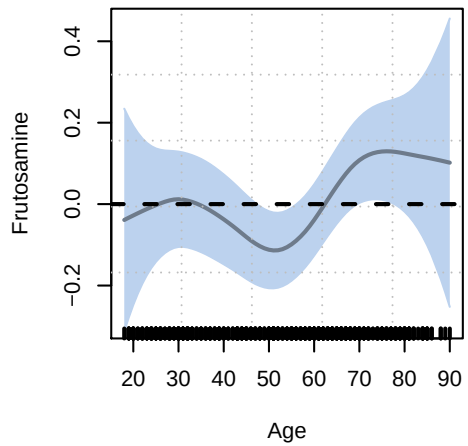

1.0
0.5
0.0
−0.5
−1.5
−4 −2 0 2 4
mcv

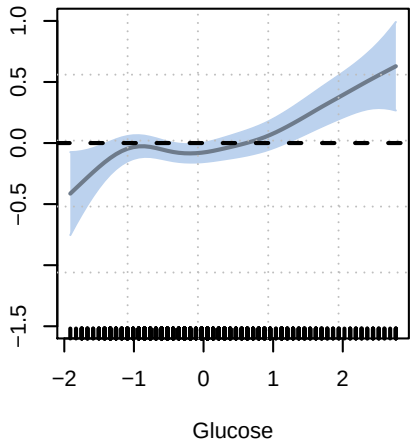
Desviaciones típicas

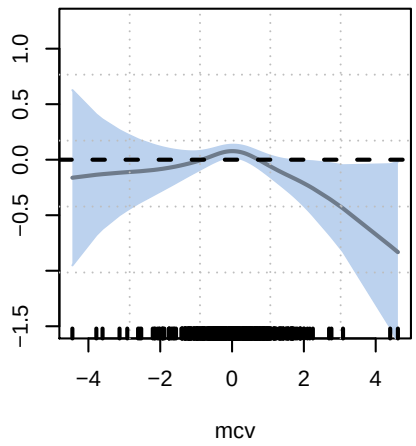




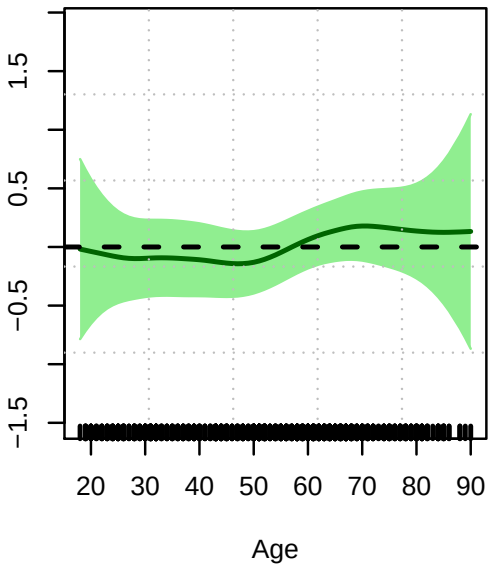


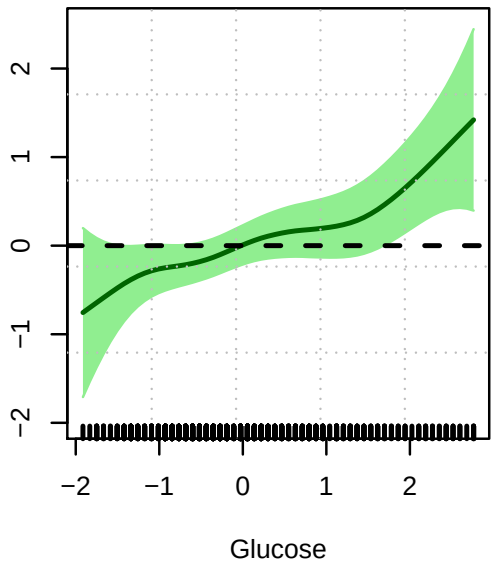


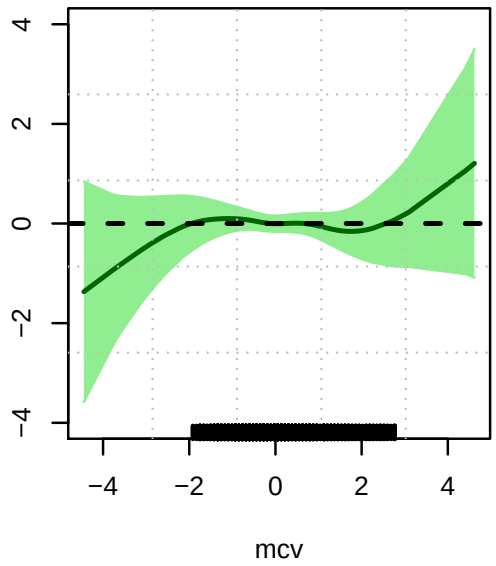




Correlaciones



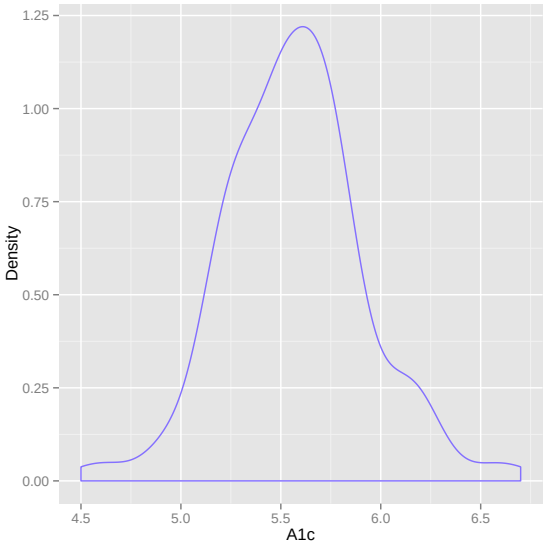


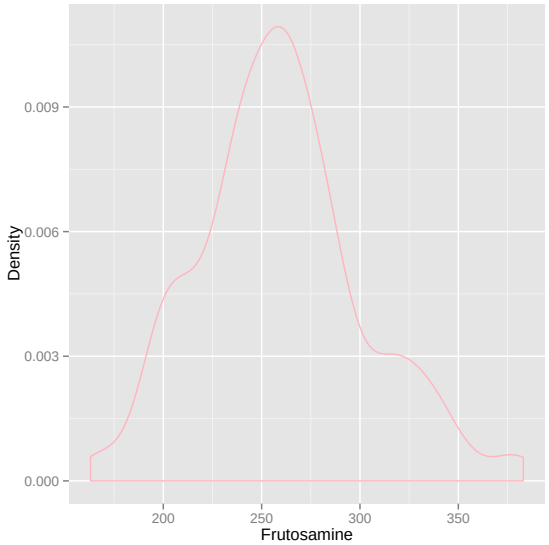


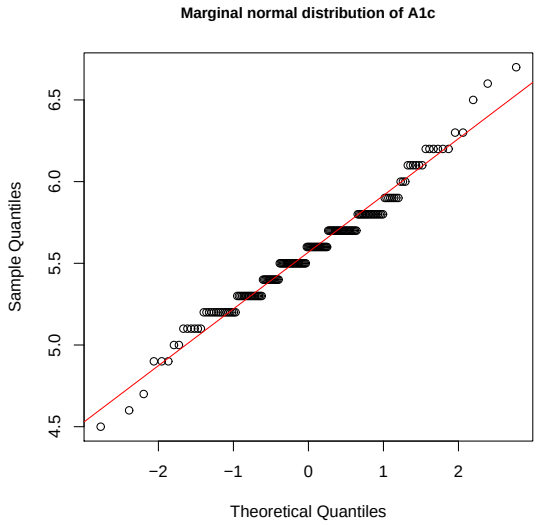
Ejemplo de Código para gráficas

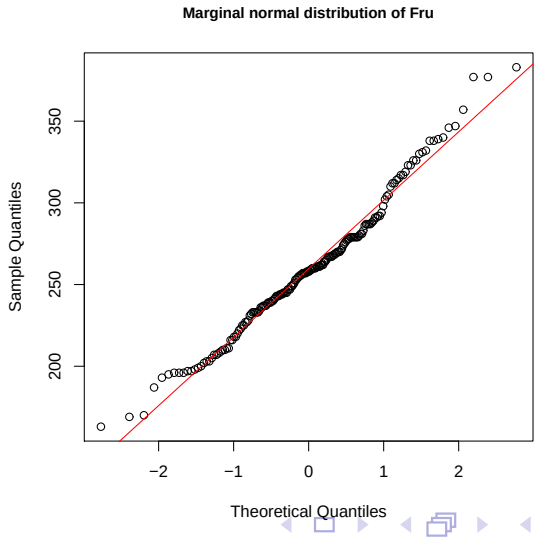
```
#-------------------------------------
>columna=6
>coef=ter$fitted.values[,columna]
>ooo <- order(with(data,edad))
>variable=data[ooo, ]$edad
#-------------------------------------

>plot(coef[ooo] ~ variable, data =data[ooo, ], col = "darkgreen",
 type = "l", xlab=.edad",ylab = .A1c",ylim=c(-0.7,0.8))
>sd=ter$se[,columna]
>b1=coef-sd*1.96
>b2=coef+sd*1.96
>lines(b1[ooo]~variable,type="l",col="pink",data=data[ooo, ])
>lines(b2[ooo]~variable,type="l",col="pink",data=data[ooo, ])
>polygon(c(variable,rev(variable)),c(b2[ooo],rev(b1[ooo])),col="pink",
border=NA)
>lines(coef[ooo] ~ variable, data =data[ooo, ], col = "deeppink4",lwd=2,
 type = "l", xlab=,ylab = )
>grid(NA, 5,col="grey") # grid only in y-direction
>grid(5, NA, lwd = 1,col="grey") # grid only in y-direction
>abline(h=0,lty=2,lwd=2)
>rug(variable,lwd=2)
```

¿Marginales normales?





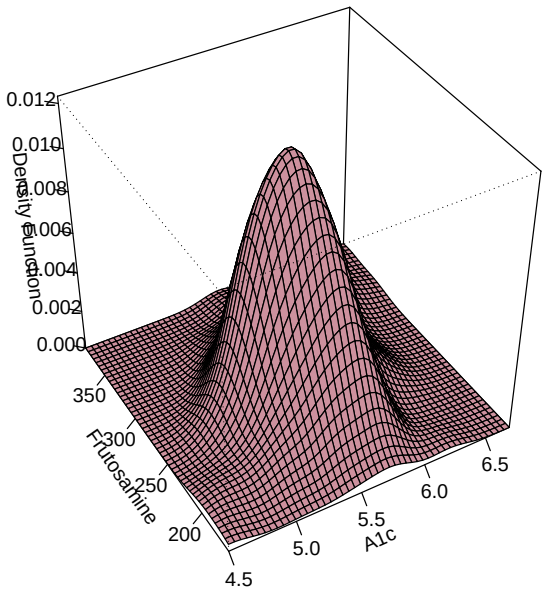




¿Normal Bivariante?

Nuestro estudio en glicación de proteinas (y otras aplicaciones biomédicas) sugieren que:

- La suposición de normalidad multivariante y marginal, puede ser muy restrictiva en la práctica.



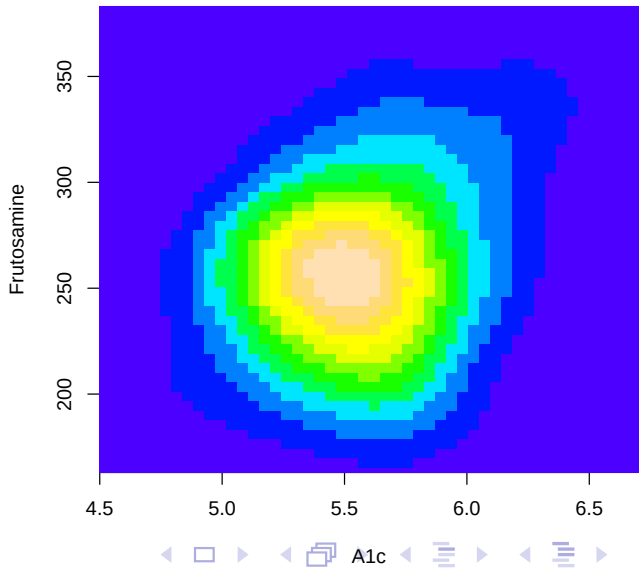


Table 13.1 Some bivariate distributions currently implemented in VGAM. Here, $\boldsymbol{y} = (y_1, y_2)^T$, and they have been separated according to their support.

Distribution	CDF $F(y_1, y_2; \theta)$ (or PDF $f(y_1, y_2; \theta)$)	Support	Range of θ	VGAM family
Freund (1961)'s exponential	$f(\boldsymbol{y}) = \alpha\beta' \exp\left\{-\beta' y_2 - (\alpha + \beta - \beta')y_1\right\}$	$0 < y_1 < y_2 < \infty$	$(\alpha, \alpha', \beta, \beta') \in (0, \infty)^4$	`freund61()`
	$f(\boldsymbol{y}) = \beta\alpha' \exp\left\{-\alpha' y_1 - (\alpha + \beta - \alpha')y_2\right\}$	$0 < y_2 < y_1 < \infty$	$(\alpha, \alpha', \beta, \beta') \in (0, \infty)^4$	`freund61()`
McKay's bivariate gamma	$f(\boldsymbol{y}) = \frac{y_1^{s_1-1}\,(y_2 - y_1)^{s_2-1}}{b^{s_1+s_2}\,\Gamma(s_1)\,\Gamma(s_2)} \exp(-y_2/b)$	$0 < y_1 < y_2 < \infty$	$(b, s_1, s_2) \in (0, \infty)^3$	`bigamma.mckay()`
Gamma hyperbola	$f(\boldsymbol{y}) = \exp\left\{-\frac{y_1}{\theta}e^{-\theta} - \theta y_2\right\}$	$(0, \infty) \times (1, \infty)$	$0 < \theta$	`gammahyperbola()`
Gumbel's Type I exponential	$\exp\{-y_1 - y_2 + \alpha y_1 y_2\} + 1 - e^{-y_1} - e^{-y_2}$	$(0, \infty)^2$	$\alpha \in \mathbb{R}$	`bigumbelIexp()`
FGM exponential	$e^{-y_1-y_2}\left[1 + \alpha\left(1 - e^{-y_1}\right)\left(1 - e^{-y_2}\right)\right] + 1 - e^{-y_1} - e^{-y_2}$	$(0, \infty)^2$	$-1 < \alpha < 1$	`bifgmexp()`
Bivariate logistic	$\left[1 + \exp\left\{-\left(\frac{y_1 - a_1}{b_1}\right)\right\} + \exp\left\{-\left(\frac{y_2 - a_2}{b_2}\right)\right\}\right]^{-1}$	$\mathbb{R}^2$	$0 < b_1,\ 0 < b_2$	`bilogistic(dpr)`
Bivariate normal, N_2	$\Phi_2(y_1, y_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$; see Sect. 13.2.1	$\mathbb{R}^2$	$-1 < \rho < 1,\ 0 < \sigma_j$	`binormal(dpr)`
Bivariate Student-t	$f(\boldsymbol{y}) = \frac{1}{2\pi\sqrt{1-\rho^2}} \frac{(1 + y_1^2 + y_2^2 - 2\rho y_1 y_2)}{(\nu(1-\rho^2))^{(\nu+2)/2}}$	$\mathbb{R}^2$	$-1 < \rho < 1$	`bistudentt(d)`

Table 13.2 Bivariate copulas currently implemented by VGAM. Notes: (i) The support is $(u_1, u_2) \in [0,1]^2$, and α is the association parameter (apar). (ii) Non-Archimedian copulas have no generator functions φ (—). (iii) See (A.61) for the Debye function $D_n(x)$ definition. (iv) Much of this table was adapted from Trivedi and Zimmer (2005) and Nelsen (2006). (v) The EIMs of some of these families appear in Schepsmeier and Stöber (2014).

Copula	$C(u_1, u_2; \alpha)$	α-domain	ρ_S	ρ_τ	Generator $\varphi(t)$	VGAM family
AMH	$\frac{u_1 u_2}{1-\alpha(1-u_1)(1-u_2)}$	$(-1,1)$	Complicated	Complicated	$\log \frac{1-\alpha(1-t)}{t}$	`bianhcop(dpr)`
Clayton	$\left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-1/\alpha}$	$(0,\infty)$	No closed form	$\frac{\alpha}{\alpha+2}$	$(1+t)^{-1/\alpha}$	`biclaytoncop(dr)`
Frank	$-\frac{1}{\alpha}\log\left(1 - \frac{(1-e^{-\alpha u_1})(1-e^{-\alpha u_2})}{1-e^{-\alpha}}\right)$	$\mathbb{R}\backslash\{0\}$	$1 - \frac{D_1(\alpha) - D_2(\alpha)}{\alpha/12}$	$1 - \frac{1-D_1(\alpha)}{\alpha/4}$	$-\log\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$	`bifrankcop(dpr)`
Gaussian	$\Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho = \alpha)$	$(-1,1)$	$\frac{6}{\pi}\sin^{-1}\left(\frac{\alpha}{2}\right)$	$\frac{2}{\pi}\sin^{-1}(\alpha)$	—	`binormalcop(dpr)`
FGM	$u_1 u_2 [1 + \alpha(1-u_1)(1-u_2)]$	$(-1,1)$	$\frac{\alpha}{3}$	$\frac{2\alpha}{9}$	—	`bifgmcop(dpr)`
Plackett	Eq. (13.13) with $F_j = u_j$	$(0,\infty)$	$\frac{\alpha+1}{\alpha-1} - \frac{2\alpha\log\alpha}{(\alpha-1)^2}$	No closed form	—	`biplackettcop(dpr)`

Referencias Principales

[1] Yee, T.W. (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

[2] Yee, T.W. and Wild, C. J. (1996). Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B, 58(3), 481-493.

[3] Yee, T.W. (2016). VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-2. URL http://CRAN.R-project.org/package=VGAM

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CiMUS Center for Research in Molecular Medicine and Chronic Diseases

XUNTA DE GALICIA

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN