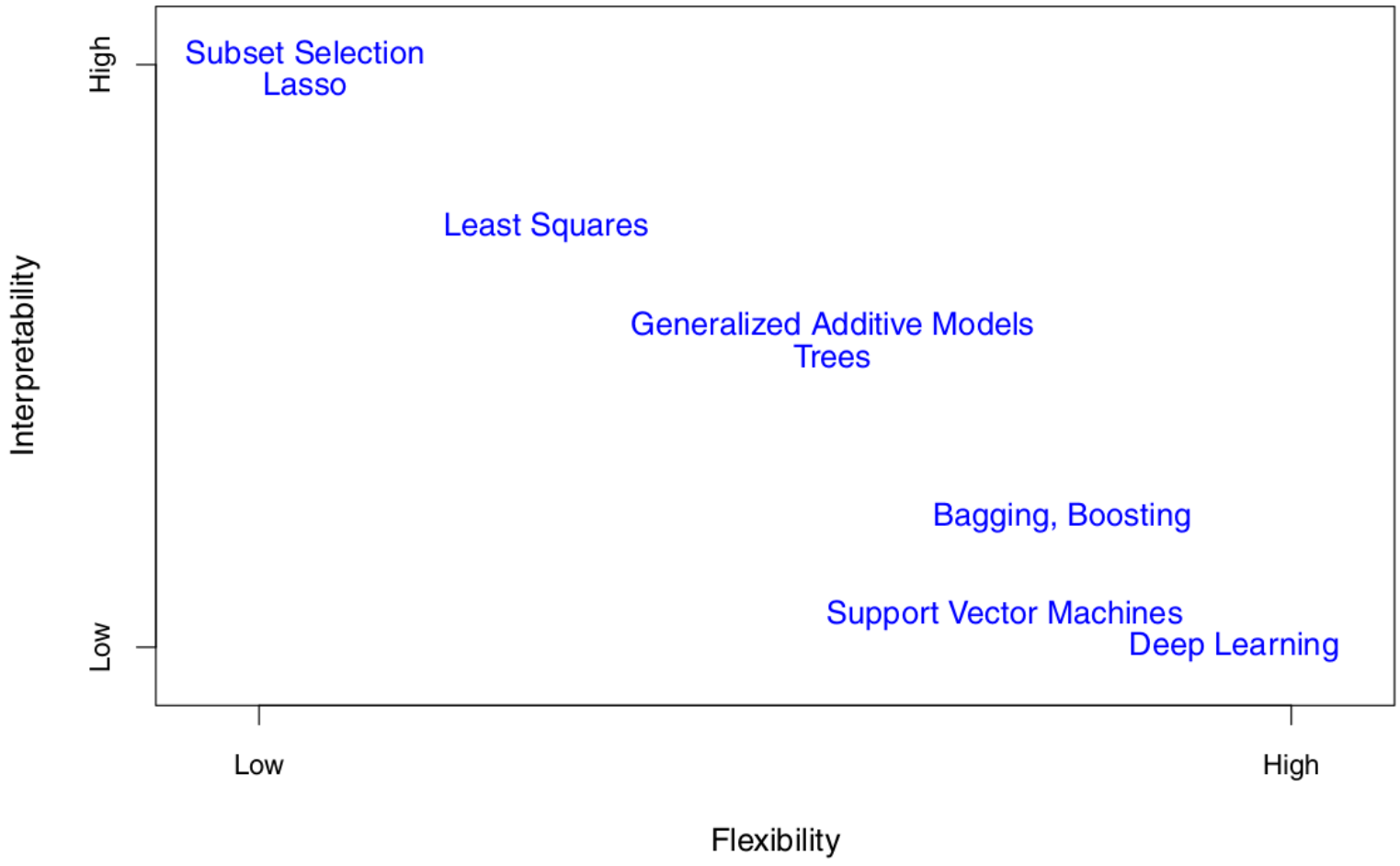
Modelos Aditivos Generalizados

Modelos Estadísticos Avanzados

Santiago Benitez-Vieyra

Where the wild things are

Where the wild things are



Quitando supuestos

Supuesto	solución
Normalidad	GLM
Independencia	LMM
Normalidad e independencia	GLMM
Normalidad y linealidad	GAM
Normalidad, independencia y linealidad	GAMM

Cómo quitamos la linealidad?

$$g(\mu_i) = \eta(X_{i1}, \ldots, X_{ik})$$

GLM

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$$

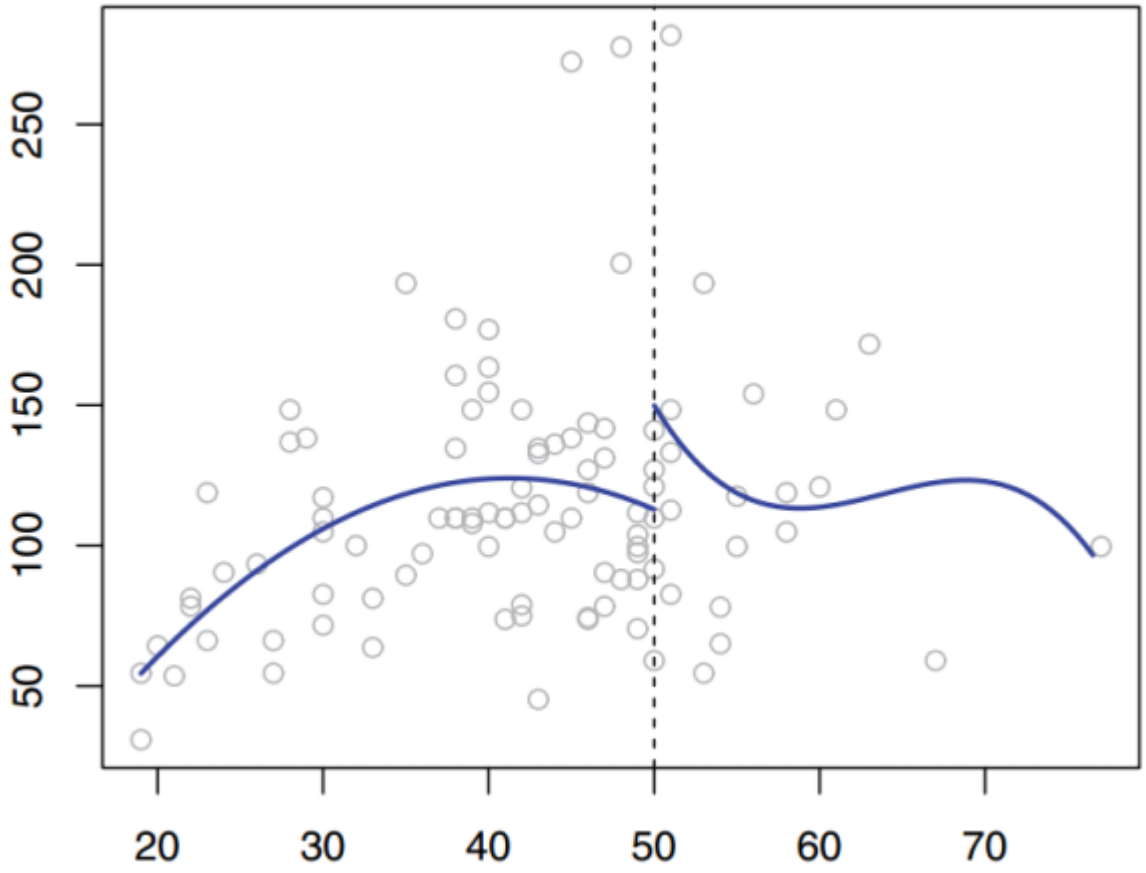
GAM

$$g(\mu_i) = \beta_0 + f_1 X_{1i} + \ldots + f_k X_{ki}$$

Primero: polinomios por piezas

Basis functions

¿Qué colocamos en *f*? la idea es tener una familia de funciones o trasformaciones que puedan ser aplicadas a las variables X.
En vez de ajustar un modelo lineal en X, ajustamos el modelo incluyendo estas funciones.



pero hay ciertos resquisitos…

Existen requisitos a cumplir por las funciones básicas

- Debe ser ***continuo***.
- Debe ser ***suave***.

En la práctica esto se logra imponiendo contricciones sobre la continuidad de la primera y la segunda derivada de la curva. El resultado de estas contricciones es un ***spline***. Existen muchas funciones básicas como *thin plate splines* y *cubic splines*.

Complejidad

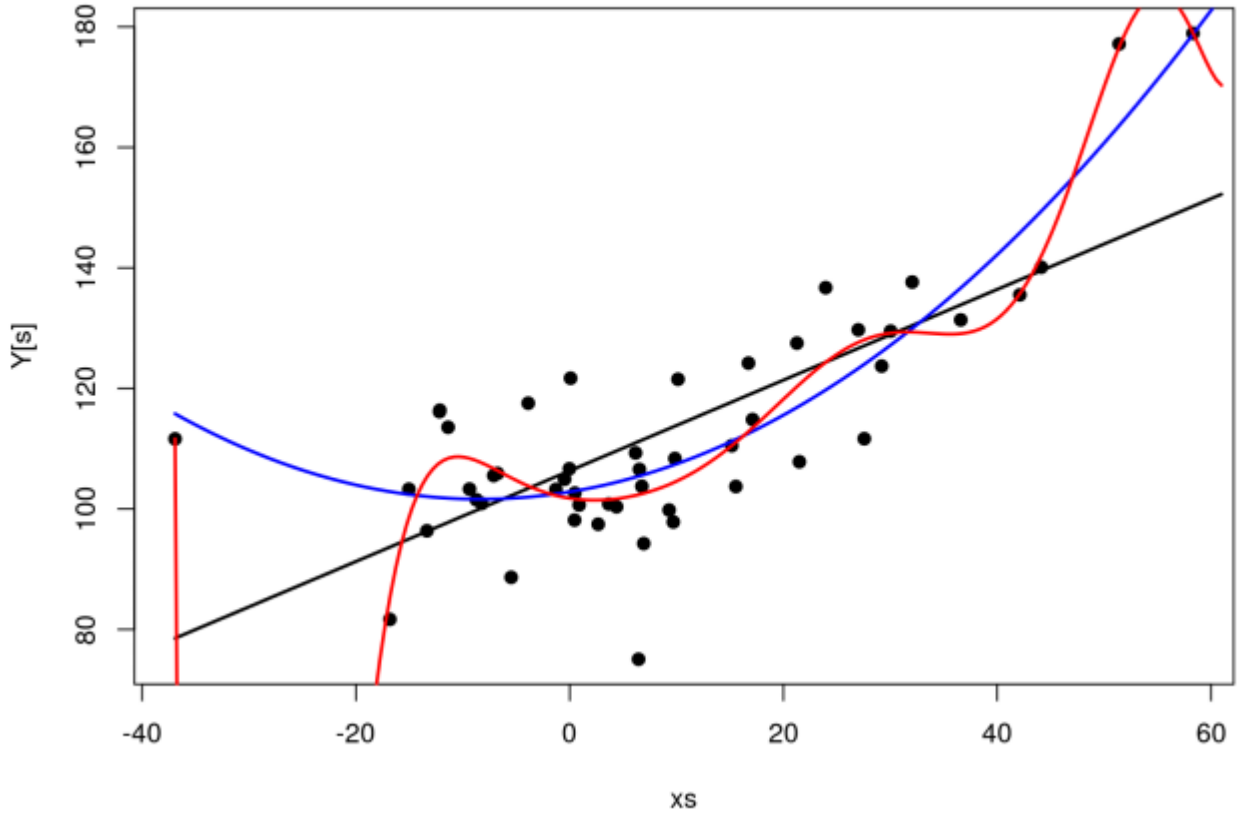
¿Cómo determinar la complejidad de la curva que ajustamos? Si usamos más nudos (y por lo tanto, más grados de libertad), la curva se va a ajustar cada vez mejor a los datos, pero va a resultar un modelo demasiado complejo.

Bias-variance trade-off

En general, los modelos se ajustan minimizando algún tipo de medida del error (por ejemplo, los mínimos cuadrados.

$$1/n \sum (y_i - f(x_i))^2$$

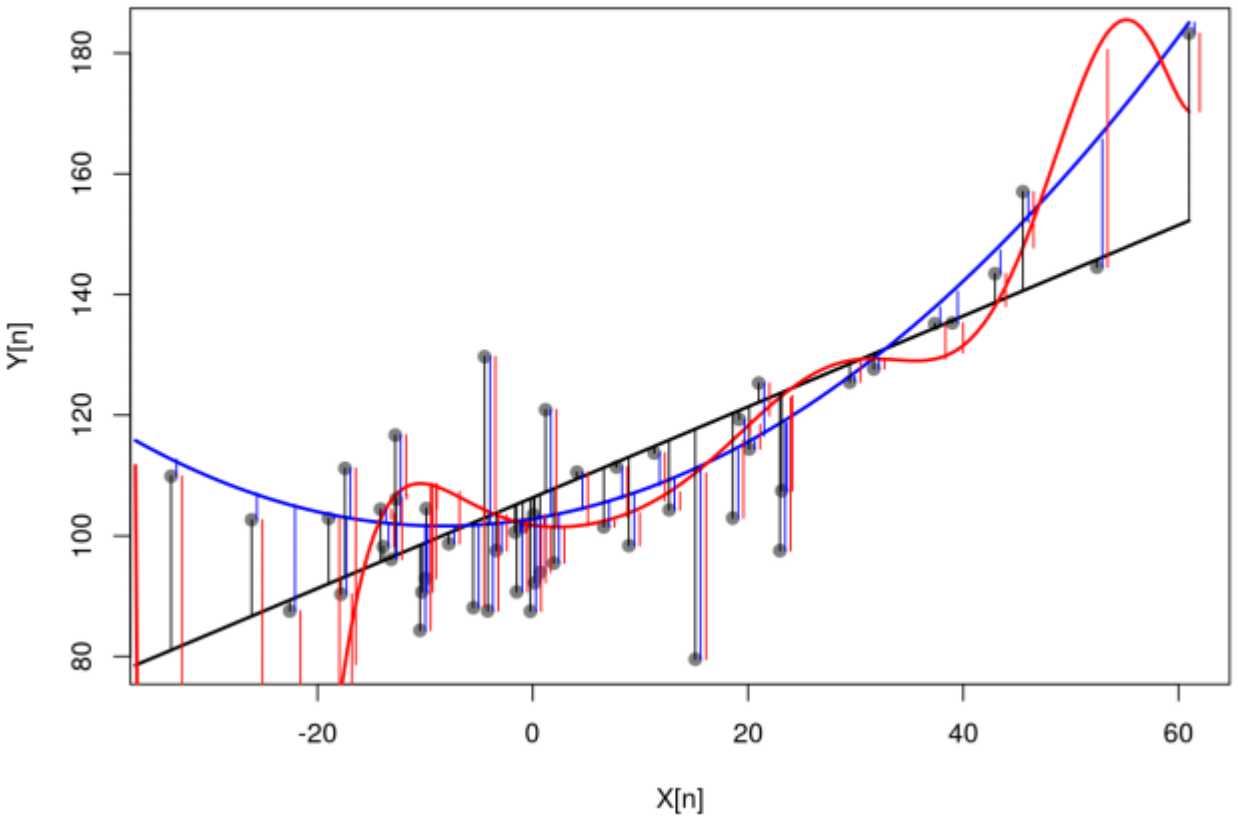
Pero esta medida decrece a medida que se incrementa la ***flexibilidad*** del modelo (medida de su complejidad, como los grados de libertad).
¿Qué pasa con un modelo complejo si luego intentamos utilizarlo para predecir?

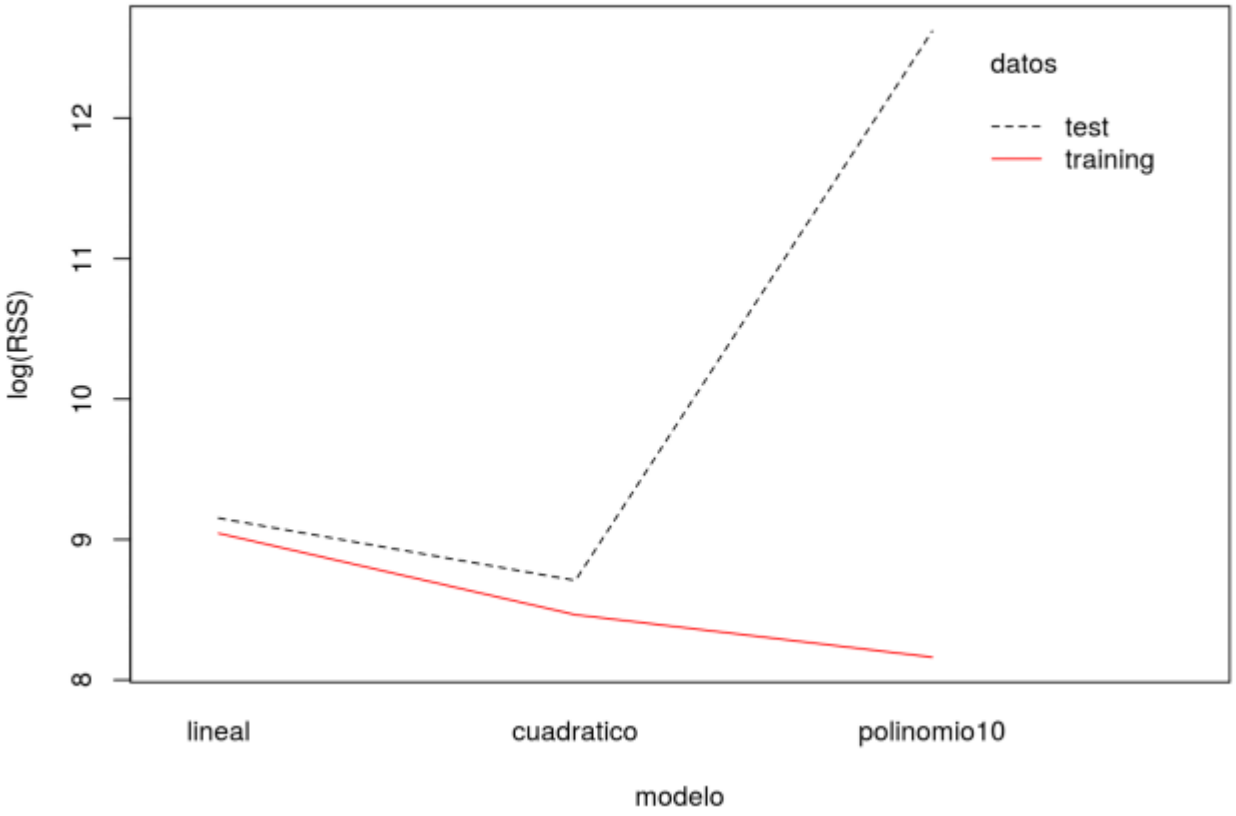


negro, *d.f.* = 2; rojo *d.f.* = 3; azul *d.f.* = 10.

Esto sucede porque estamos utilizando los residuos del set de ***DATOS DE ENTRENAMIENTO*** (training data).

Si usamos los residuso sobre ***DATOS DE PRUEBA*** (test data)

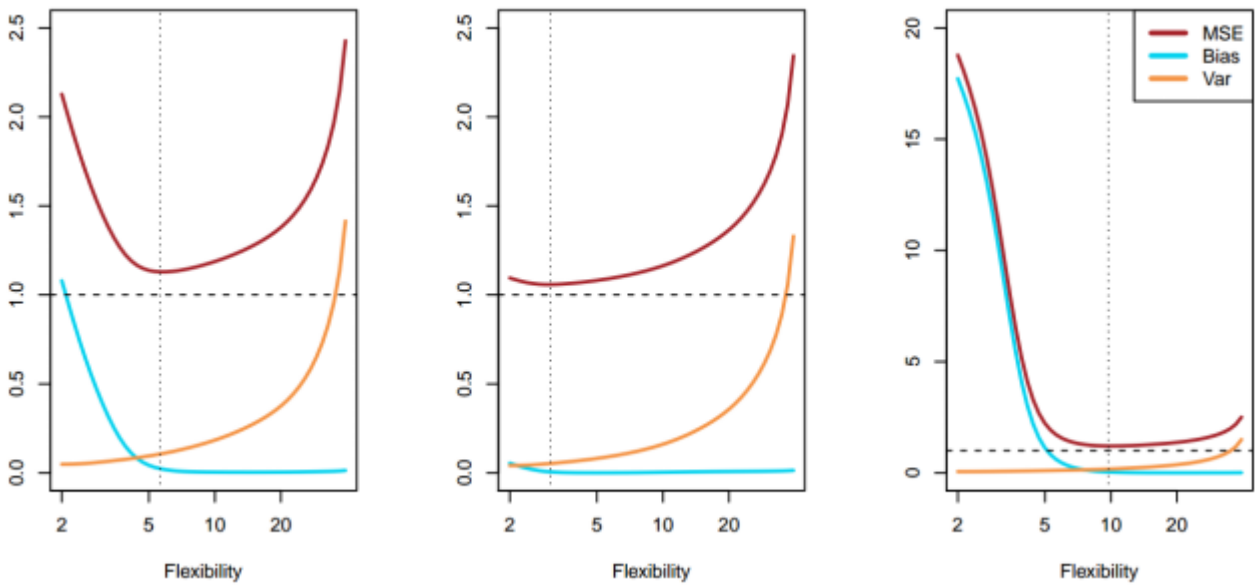



log(RSS)
8
9
10
11
12
lineal
cuadratico
polinomio10
modelo
datos
test
training

El valor esperado de *test MSE* (residuos de prueba)

$$E(MSE_{test}) = Var(\hat{f}\,(x_0)) + [Bias(\hat{f}\,(x_0))]^2 + Var(\epsilon)$$

La ***varianza*** surge de los cambios en $\hat{f}$ con diferentes datos de entrenamiento.
Métodos más flexibles tienen más varianza.
El ***sesgo*** es el error introducido por tratar de aproximar el comportamiento de la vida real con un modelo. Modelos más sencillos tienen más sesgo.
Este trade-off es la base de la ***PENALIZACIÓN***.



* de James et al. 2014

Volviendo a los smoothing splines…

minimizar

$$\sum (y_i - f(x_i))^2 + \lambda \int f''(t)^2 d(t)$$

donde el segúndo término es una penalización contra la variabilidad en f (f'' es la segunda derivada, el cambio en la pendiente). λ es un ***parámetro de suavizado***.

- Si $\lambda = 0$ la penalidad se anula y el modelo puede ser tan complejo como sea necesario para que los residuos tengan el mínimo valor posible (en la práctica, puede volver los residuos iguales a cero).
- Si $\lambda \to \infty$ la penalidad se vuelve muy grande y el modelo debe ser tan simple como sea posible.
- Si un spline es minimizado de esta manera se vuelve un ***smoothing spline***.
- λ controla el trade-off entre sesgo y varianza.
- λ controla los ***efective degreees of freedom***.

Selección de λ. Validación cruzada.

Para cada valor de λ en una serie se calculan los errores de la **predicción** partiendo el set de datos en datos de entrenamiento y prueba.

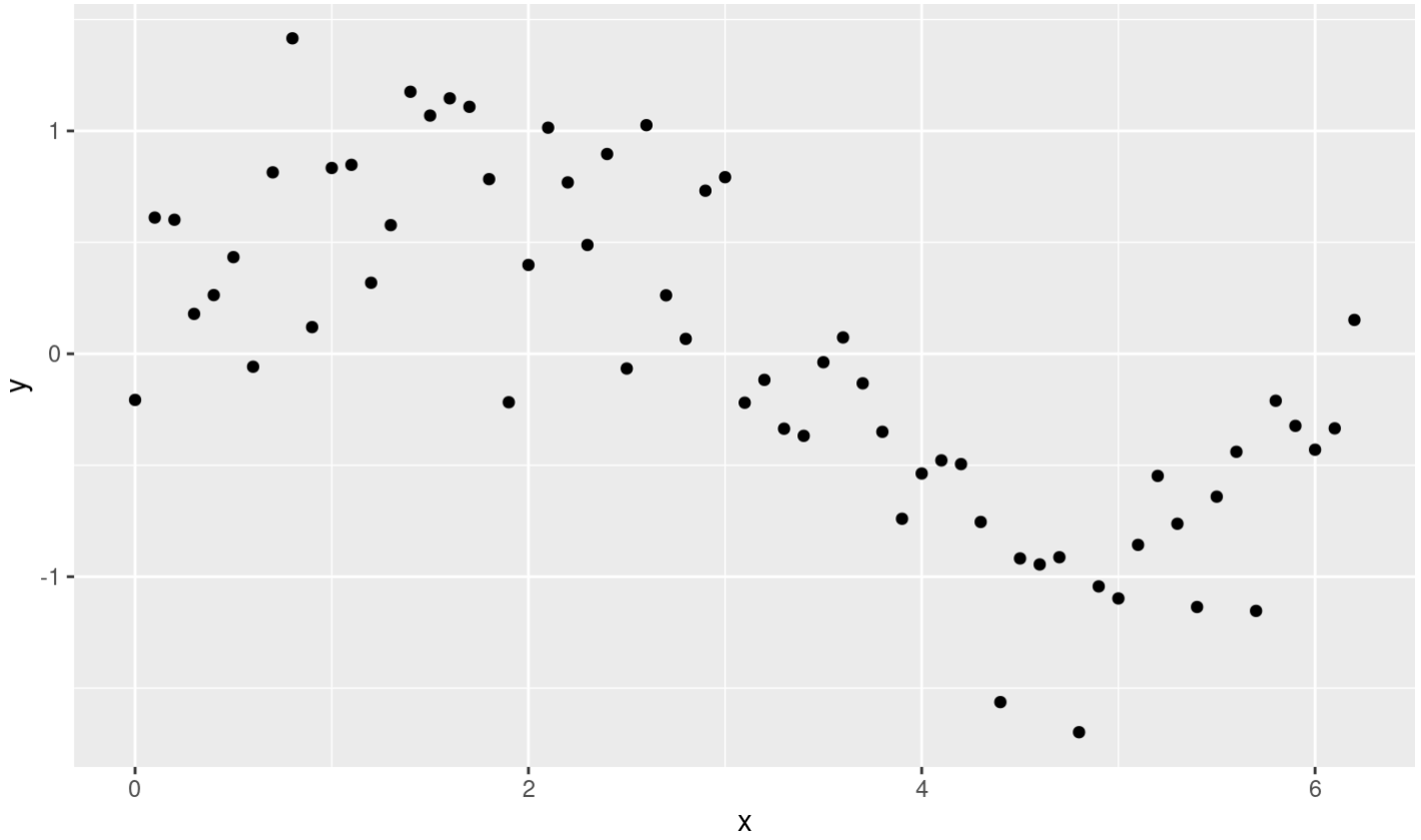
- Ordinary Cross Validation (OCV). Parte en set de datos *k* veces (habitualmente 5 o 10). *BIC* es una aproximación a OCV.
- Leave One Out Cross Validation (LOOCV). Parte el set de datos *n* veces (en cada ocasión sólo queda un dato de pueba). *AIC* es una aproximación a LOOCV.
- Hay muchas versiones de CV, como GCV.

Selección de λ. Verosimilitud: ML y REML.

- Los métodos basados en los errores de la predicción tienen cierta tendencia a encontrar más de un mínimo local.
- Los métodos basados en ML o REML tratan las funciones de suavizado como efectos aleatorios, de tal forma que λ puede ser tomado como un parámatro de varianza.
- Esto surge de una interpretación Bayesiana de los GAMs…

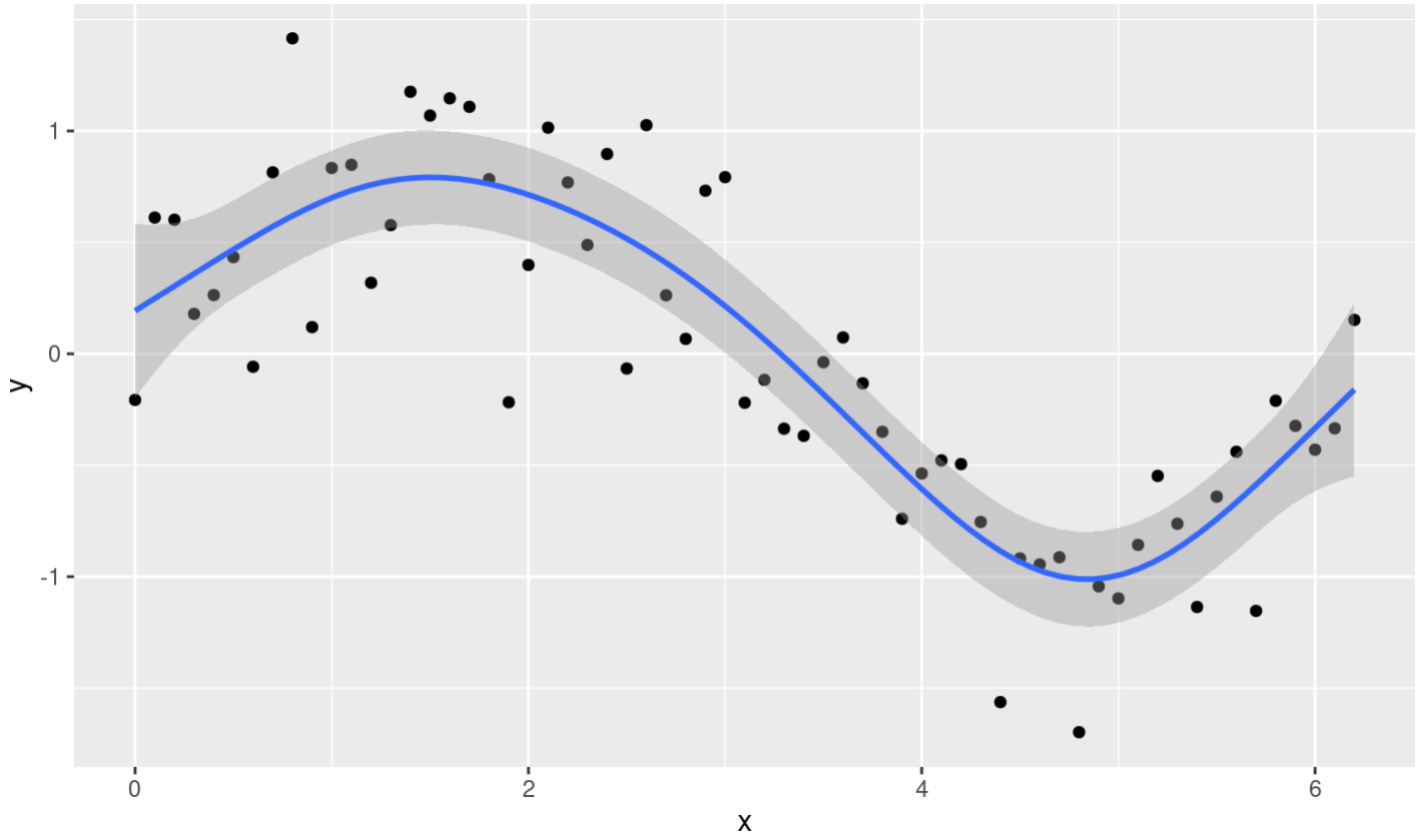
Paquete *mgcv* de Simon Wood

- Amplia variedad de splines.
- Selección de variables por *shrinkage*.
- Soporta métodos de verosimilitud y de validación cruzada para obtener λ.
- Tiene una función especial (*bam*) para grandes sets de datos.
- Omite las observaciones sin datos.
- Soporta suavizados multidimensionales utilizando *tensores* y *thin plate splines*.
- Posee herramientas de diagnóstico generales y de diagnóstico de la *concurvidad*.

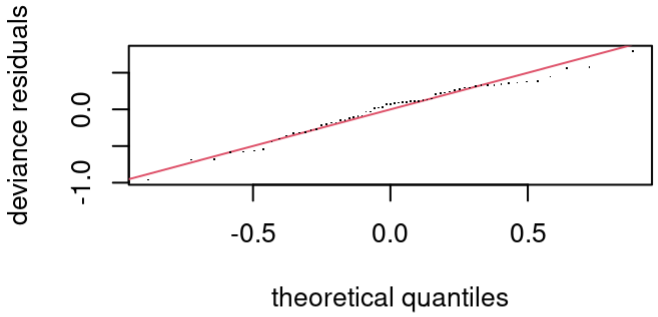

1
0
-1
y
0
2
4
6
x

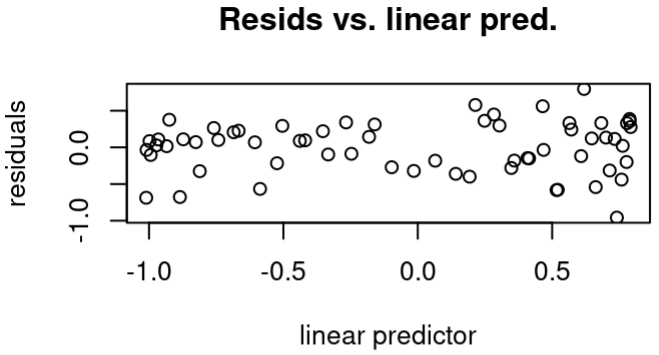
```r
library(mgcv)
fit <- gam(y ~ s(x), method = "REML")
summary(fit)
```

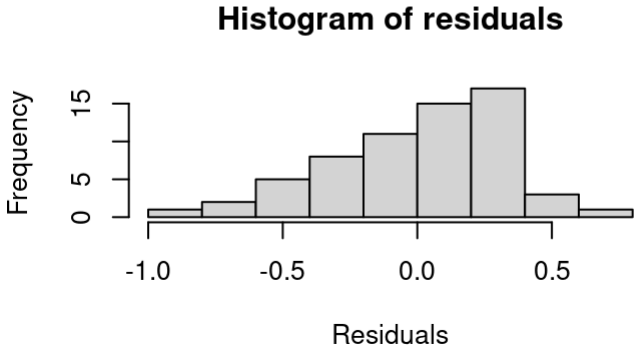
```
Family: gaussian
Link function: identity

Formula:
y ~ s(x)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01827    0.04604  -0.397    0.693

Approximate significance of smooth terms:
       edf Ref.df     F p-value
s(x) 5.384  6.517 29.98  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.758   Deviance explained = 77.9%
-REML =  35.16  Scale est. = 0.13356   n = 63
```

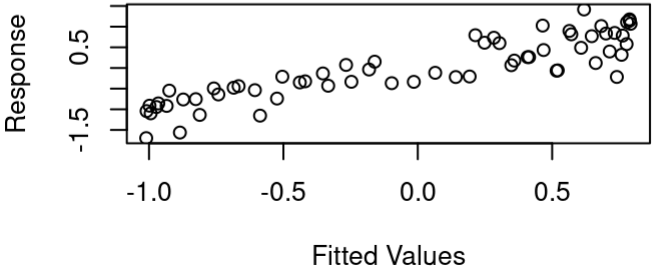

1
0
-1
y
0
2
4
6
x

```
gam.check(fit)


Method: REML   Optimizer: outer newton
full convergence after 5 iterations.
Gradient range [-5.408989e-06,3.208159e-06]
(score 35.15953 & scale 0.1335646).
Hessian positive definite, eigenvalue range [2.066374,30.66895].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

       k'  edf k-index p-value
s(x) 9.00 5.38    1.09    0.69
```


deviance residuals
-1.0
0.0
-0.5
0.0
0.5
theoretical quantiles


Resids vs. linear pred.
residuals
-1.0
0.0
-1.0
-0.5
0.0
0.5
linear predictor


Histogram of residuals
Frequency
0
5
15
-1.0
-0.5
0.0
0.5
Residuals


Response vs. Fitted Values
Response
-1.5
0.5
-1.0
-0.5
0.0
0.5
Fitted Values

bases	**utilidad**
bs = ‘tp’	*thin plate splines* por defecto
bs = ‘cr’	*cubic splines*
bs = ‘cc’ o ‘cp’	splines cíclicos
bs = ‘re’	efectos random
bs = ‘fs’	factores (igual λ por nivel)
bs = ‘sos’	splines esféricos
y hay mas	…

Familias

- `binomial()`
- `poisson()`
- `Gamma()`
- `inverse.gaussian()`
- `nb()`
- `tw()`
- `mvn()`
- `multinom()`
- `betar()`
- `scat()`
- `gaulss()`
- `ziplss()`
- `twlss()`
- `cox.ph()`
- `gamals()`
- `ocat()`

END