Modelización de regresión de series temporales con R.

XIV Summer School MESIO UPC-UB

modelo dlnm

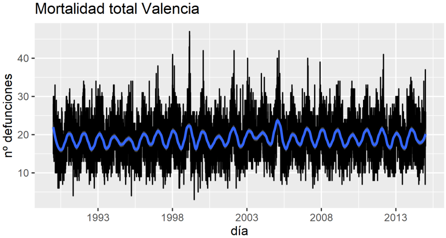
Carmen Iñíguez
Depto. Estadística i I.O., UV

ÍNDICE

DLNM

DLNM: **D**istributed **L**ag **N**on **L**inear **M**odel.

- Diseño: Series temporales



- Permite describir relaciones no lineales y retardadas.
  - Relación dosis-respuesta no lineal
  - El efecto tiene lugar o se prolonga varios días (lags) después de la exposición.
- Paradigma de aplicación: **temperatura y mortalidad**.

Relación temperatura-mortalidad

- Relación no lineal

American Journal of Epidemiology
Copyright © 2002 by the Johns Hopkins Bloomberg School of Public Health

Vol. 155, No. 1
Printed in U.S.A.

Temperature and Mortality in 11 Cities of the Eastern United States

Frank C. Curriero,[1] Karlyn S. Heiner,[1] Jonathan M. Samet,[2] Scott L. Zeger,[1] Lisa Strug,[1] and Jonathan A. Patz[3]

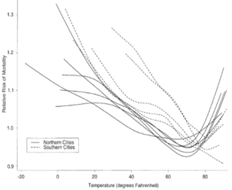


FIGURE 1. Temperature-mortality relative risk functions for 11 US cities, 1973–1994. Northern cities: Boston, Massachusetts; Chicago, Illinois; New York, New York; Philadelphia, Pennsylvania; Baltimore, Maryland; and Washington, DC. Southern cities: Charlotte, North Carolina; Atlanta, Georgia; Jacksonville, Florida; Tampa, Florida; and Miami, Florida. °C = 5/9 × (°F − 32).

- Relación retardada

American Journal of Epidemiology
Copyright © 1993 by The Johns Hopkins University School of Hygiene and Public Health

Vol. 137, No. 3
Printed in U.S.A.

Outdoor Air Temperature and Mortality in the Netherlands: A Time-Series Analysis

Anton E. Kunst, Caspar W. N. Looman, and Johan P. Mackenbach

TABLE 2. Association between temperature and mortality, controlling for a number of variables, the Netherlands, 1979–1987

Temperature and control variable	% effect, by lag period† 0	1–2	3–6	7–14	15–30	Aggregate effect‡
Cold						
None	−0.25***	0.30***	0.44***	0.50***	0.18***	1.17
Influenza incidence	−0.27***	0.29***	0.42***	0.40***	−0.07	0.77
Sulphur dioxide density	−0.25***	0.29***	0.50***	0.58***	0.22***	1.34
Season	−0.26***	0.30***	0.43***	0.49***	0.22***	1.18
All 3 variables together	−0.27***	0.26***	0.45***	0.41***	0.06	0.91
Heat						
None	1.74***	1.24***	−0.22	−0.53***	−0.81***	1.42
Influenza incidence	1.77***	1.26***	−0.23	−0.56***	−1.14***	1.10
Sulphur dioxide density	1.77***	1.24***	−0.12	−0.45**	−0.69***	1.75
Season	1.72***	1.24***	−0.20	−0.47***	−0.43*	1.86
All 3 variables together	1.76***	1.23***	−0.14	−0.49***	−0.51*	1.85

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

† Percent mortality increase per change in degrees Celsius. (The effect of 1 degree of change during 1 individual day within a lag period is about equal to the effect for the entire period divided by the number of days.) Estimated from regression analysis of mortality on the average values of "cold," "heat," and the respective control variables in the five lag periods. Regression coefficients are transformed according to the formula in Materials and Methods.

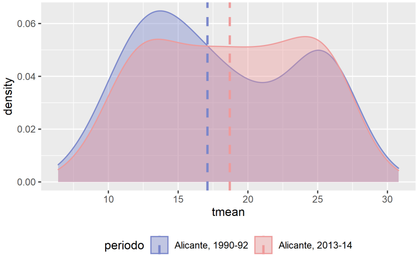
‡ The sum of the percent effects associated with the five lag periods.

Relación temperatura-mortalidad

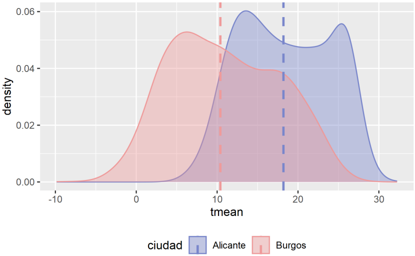
Otras **complejidades**...

- exposición cambiante:

\- Variación geográfica



\- Variación temporal



- adaptación, aclimatación: **la propia relación es cambiante!!**

Aparición del modelo DLNM

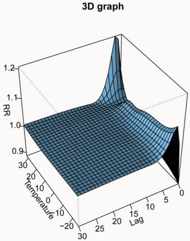
- Abordaje anterior: gam + agrupación arbitaria de lags
- En 2010: se desarrolla el marco teórico del dlnm y en 2011 se implementa en **R**
  - Gasparrini A & Armstrong B. (2010). Distributed lag non-linear models. *Stat Med.* **29** (2): 2224–34.
  - Gasparrini A (2011). Distributed Lag Linear and Non-Linear Models in R:The Package dlnm. *J Stat Softw.* **43** (8): 1–20.
- Entorno de trabajo:
  - Gasparrini A (2013)& Armstrong B. Reducing and meta-analysing estimates from distributed lag non-linear models. *BMC Med Res Methodol.* **13**: 1.
  - Gasparrini A (2014)& Leone M. Attributable risk from distributed lag models. *BMC Med Res Methodol.* **14**: 55.
  - Armstrong BG, Gasparrini A, Tobias A (2014). Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Med Res Methodol.* **14**: 122.
  - Tobías A, Armstrong B, Gasparrini A. Investigating Uncertainty in the Minimum Mortality Temperature: Methods and Application to 52 Spanish Cities. Epidemiology. 2017 Jan;28(1):72-76.
  - Gasparrini A, Scheipl F, Armstrong B, Kenward MG (2017). A penalized framework for distributed lag non-linear models. *Biometrics.* **73**(3): 938–948.
  - Sera F, Armstrong B, Blangiardo M, Gasparrini A (2019) M. An extended mixed-effects framework for meta-analysis. *Stat Med.* **38**(29): 5429–5444.

Aparición del modelo DLNM

- Abordaje anterior: gam + agrupación arbitaria de lags
- En 2010: se desarrolla el marco teórico del dlnm y en 2011 se implementa en **R**
  - Gasparrini A & Armstrong B. (2010). Distributed lag non-linear models. *Stat Med.* **29** (2): 2224–34.
  - Gasparrini A (2011). Distributed Lag Linear and Non-Linear Models in R:The Package dlnm. *J Stat Softw.* **43** (8): 1–20.
- Entorno de trabajo:
  - Gasparrini A (2013)& Armstrong B. Reducing and meta-analysing estimates from distributed lag non-linear models. *BMC Med Res Methodol.* **13**: 1.
  - Gasparrini A (2014)& Leone M. Attributable risk from distributed lag models. *BMC Med Res Methodol.* **14**: 55.
  - Armstrong BG, Gasparrini A, Tobias A (2014). Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Med Res Methodol.* **14**: 122.
  - Tobías A, Armstrong B, Gasparrini A. Investigating Uncertainty in the Minimum Mortality Temperature: Methods and Application to 52 Spanish Cities. Epidemiology. 2017 Jan;28(1):72-76.
  - Gasparrini A, Scheipl F, Armstrong B, Kenward MG (2017). A penalized framework for distributed lag non-linear models. *Biometrics.* **73**(3): 938–948.
  - Sera F, Armstrong B, Blangiardo M, Gasparrini A (2019) M. An extended mixed-effects framework for meta-analysis. *Stat Med.* **38**(29): 5429–5444.

DLNM. Formulación.

- El **dlnm** representa un espacio de trabajo para una modelización flexible relaciones no lineales, potencialmente retardadas, en series temporales.
- La metodología se basa en la definición de una **crossbasis**, espacio funcional bidimensional que se expresa mediante la combinación de dos sets de funciones bases.
- El primer set describe la relación en la dimensión del predictor (relación **dosis-respuesta**)y el segundo la relación en la dimensión del tiempo (relación **lag-respuesta**).
- Ancestro: **pdlm**: (Almond(1965), Zanobetti y Schwartz (2000), Armstrong (2006))



Formulación en un caso sencillo: pDLM.

- Supongamos una relación **dosis-respuesta** lineal que se prologa hasta L días:

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_L x_{t-L} + \epsilon_t, \quad \forall t = 1, 2, \ldots, n \tag{1}$$

$$\mathbf{D} := \left[X, X_{(-1)}, \ldots, X_{(-L)}\right]_{n\times(L+1)}, \quad \boldsymbol{\beta} = [\beta_0, \beta_1 \ldots, \beta_L]^t,$$

$$Y = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

- $\boxed{\text{Un pdlm es la reparametrización que resulta al asumir una relación funcional entre } \beta_l \text{ y } l}$

- Es la llamada relación **lag-respuesta**. Por ejemplo si la asumimos **parabólica**:

$$\beta_l = \eta_0 + \eta_1 \cdot l + \eta_2 \cdot l^2, \quad \forall l = 1, 2, \ldots, L \tag{3}$$

$$R_{l.} := \left[1, l, l^2\right], \mathbf{R} = [R_{l.}]_{(L+1)\times 3}, \quad \boldsymbol{\eta} = [\eta_0, \eta_1, \eta_2]^t, \quad \nu_l = 3$$

$$\boldsymbol{\beta} = \mathbf{R}\boldsymbol{\eta} \tag{4}$$

- Sustituyendo (4) en (2):

$$Y = \mathbf{D}\mathbf{R}\boldsymbol{\eta} + \boldsymbol{\epsilon} = \mathbf{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{5}$$

$$\mathbf{W}_{n\times 3} : \textbf{crossbasis}$$

- Tras estimar $\boldsymbol{\eta}$ en (5) y se recupera $\boldsymbol{\beta}$ usando (4):

$$\begin{cases} \hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}} \end{cases} \rightarrow \begin{cases} \hat{\boldsymbol{\beta}} = \mathbf{R}\hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \mathbf{R}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}\mathbf{R}' \end{cases} \tag{6}$$

Estimación en pDLM.

- ¿Cual es el impacto (**retardado**) de una exposición concreta, $x_t = x^*$, en el día t, l días después de haber ocurrido?

$$\hat{y}_t = \hat{\beta}_0 x_t + \hat{\beta}_1 x_{t-1} + \ldots + \hat{\beta}_l x_{t-l} + \ldots + \hat{\beta}_L x_{t-L}$$

$$\boxed{\hat{\beta}_l \cdot x^*}$$

- ¿Cuál es impacto total estimado (**overall**) de una exposición $x_t = x^*$?

$$\begin{aligned}
\hat{y}_t &= \hat{\beta}_0 x_t + \hat{\beta}_1 x_{t-1} + \ldots + \hat{\beta}_L x_{t-L} \\
\hat{y}_{t+1} &= \hat{\beta}_0 x_{t+1} + \hat{\beta}_1 x_t + \ldots + \hat{\beta}_L x_{t-L+1} \\
&\vdots \qquad \ddots \\
\hat{y}_{t+L} &= \hat{\beta}_0 x_t + \hat{\beta}_1 x_{t-1} + \ldots + \hat{\beta}_L x_t
\end{aligned}$$

$$\boxed{\left(\sum_{l=0}^{L} \hat{\beta}_l\right) \cdot x^*}$$

Ventajas pDLM

¿Qué hemos ganado asumiendo una distribución en los retardos?

Sin

- nº coefs $(L+1)\uparrow$: sobreparametrización
- Colinealidad $(D.j$ muy correladas$)\uparrow$: inestabilidad numérica

$\leftrightsquigarrow$

Con

- nº coefs: $(\nu_l)\downarrow$: reducción dimensionalidad
- No colinealidad $(R.j$ independientes$)$: estabilidad numérica

DLNM

Generalización de pdlm: relación dosis y lag respuesta no paramétricas (splines)

- Formulación algo complicada: intervienen dos splines: $\mathrm{s}_{\mathbf{1}}$, spline de dimensión ν_x modeliza la relación dosis-respuesta y $\mathrm{s}_{\mathbf{2}}$, spline de dimensión ν_l modeliza la relación lag-respuesta.

$$Y = \mathrm{s}_{\mathbf{1}}(X_0) + \mathrm{s}_{\mathbf{1}}(X_{(-1)}) + \ldots + \mathrm{s}_{\mathbf{1}}(X_{(-L)}) + \boldsymbol{\varepsilon} \quad (7)$$

$$\mathbf{R} = \mathrm{s}_{\mathbf{2}}((0, \ldots, l, \ldots L)) \quad (8)$$

- que pasa por definir un array $\dot{\boldsymbol{D}}_{n\times\nu_x\times(L+1)}$ tridimensional:

$$Y = \sum_{j=1}^{L+1} \dot{\boldsymbol{D}}_{\cdot\cdot j}\boldsymbol{\beta}^j + \boldsymbol{\varepsilon} = \sum_{i=1}^{\nu_x} \dot{\boldsymbol{D}}_{\cdot i \cdot}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}; \quad \boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_{v_x}^t)^t \quad (9)$$

$$Y = \sum_{i=1}^{\nu_x} \dot{\boldsymbol{D}}_{\cdot i \cdot}\boldsymbol{R}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon} = \sum_{i=1}^{\nu_x} \boldsymbol{W_i}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon} \quad (10)$$

$$\boldsymbol{W} = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{v_x}]_{n\times(v_l\cdot v_x)} := \text{Crossbasis}$$

$$\begin{cases} \hat{\boldsymbol{\eta}}_{v_l\cdot v_x\times 1} \\ \hat{\boldsymbol{\Sigma}}_\eta \end{cases} \rightarrow \begin{cases} \hat{\boldsymbol{\beta}}_{(L+1)\cdot v_x\times 1} = (\boldsymbol{I}\otimes\boldsymbol{R})\hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\Sigma}}_\beta = (\boldsymbol{I}\otimes\boldsymbol{R})\hat{\boldsymbol{\Sigma}}_\eta(\boldsymbol{I}\otimes\boldsymbol{R})^t \end{cases} \quad (11)$$

DLNM EN R.

- El entorno dlnm está implementado en la librería **dlnm** de **R** y es compatible con un amplio rango de familias de regresión (`glm`, `gam`, `lme`, `clogit`,...)
  - entre ellas la regresión de **Poisson**, familia estándar para una respuesta de conteo.
- El ajuste se realiza en dos pasos:
  1. Se crea la crossbasis (`cb`), matriz de diseño, con el comando homónimo:

```
nlag<-l
mi.argvar=list(fun="lin" int=F)
mi.arglag=list(fun="poly" degree=2,int=T)
cb<-crossbasis(datos$predictor, lag=nlag, argvar=mi.argvar,
arglag=mi.arglag)
```

  2. Se ajusta el modelo simplemente incluyendo `cb` como predictor:

```
model.glm<-glm(nonext∼cb+pred1+...+predm,...)
```

- Las estructuras disponibles (polinomios, tipos de splines, etc.), así como su sintaxis pueden encontrarse en:

```
help(onebasis)
```

DLNM EN R.

- Implementación del dlnm dirigida a la **interpretación de resultados**, por ello la salida numérica es muy detallada. Incluye:
  1. coeficientes η y su matriz de covarianza
  2. matriz de efectos para cada valor de l y x (tipo: link y respuesta)
  3. vector de efectos totales, overall (tipos link y respuesta)

  - En reg. de Poisson el tipo "respuesta" son riesgos relativos (RR).

- La predicción se obtiene mediante los comandos **`crosspred`** y **`crossreduce`** (mañana!!))
  1. Completa:
     ```
     pred<-crosspred(cb,model.glm, at=valorpred1, cen=valorpred0)
     ```
  2. Reducida en el rango del predictor:
     ```
     tpred<-quantile(predictor, probs)
     bvar <- do.call("onebasis",c(list(x=tpred),attr(cb,"argvar")))
     pred<-crosspred(basis=bvar,coef=coef(cb),vcov=vcov(cb),at=...,cen=...)
     ```
  3. Reducida en el rango de los retardos:
     ```
     xlag <- 1:L
     blag <- do.call("onebasis",c(list(x=xlag),attr(cb,"arglag")))
     pred<-crosspred(basis=blag,coef=coef(cb),vcov=vcov(cb),at=...,cen=...)
     ```

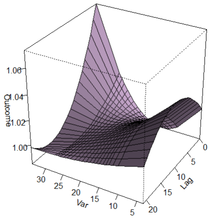
  - En reg. de Poisson, la opción `model.link='log'` proporciona RR.

DLNM en R.

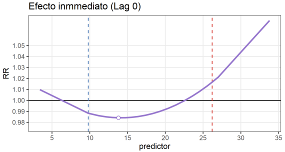
- Implementación dlnm dirigida a la interpretación de resultados con un énfasis importante en la **representación gráfica** .
- Por ello proporciona, mediante el comando `plot`, varios gráficos de interés, eso sí, en formato base (de momento):
  1. Grafico 3D de la superficie de ajuste:
     `plot(pred)`
  2. Para un retardo concreto en el rg del predictor:
     `plot(pred,lag=1)`
  3. El efecto total (overall):
     `plot(pred,ptype="overall")`
  4. Por retardo para un valor concreto del predictor:
     `frio<-pred$predvar[5]`
     `plot(pred,ptype="slices", var=frio,...)`
- Ayuda en `plot.crosspred`.

DLNM EN R.

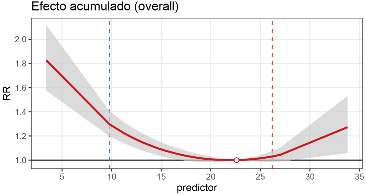
- Plano de prediccion:



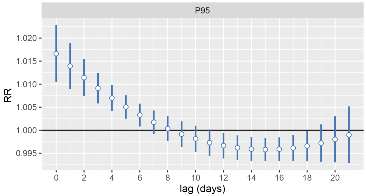
- Efecto en un lag concreto



- **Efecto total**:



- **Efecto a lo largo de los lags** :



VENTAJAS DLNM

¿Qué hemos ganado respecto al pdlm?

pdlm

- dosis- respuesta lineal: no siempre realista
- lag-respuesta polinómica: pre-supuesta

$\leftrightsquigarrow$

dlnm

sobretodo FLEXIBILIDAD

ALGUNAS NOTAS

- Comparación de modelos y diagnóstico son los estándar (AIC, BIC, LRT, pacf)
- Para la interpretación, conviene utilizar como punto de referencia, el valor de exposición asociado a la mínima predicción.
  - En el caso de temperatura como exposición y mortalidad como respuesta, sería el valor de temperatura asociado a la mínima mortalidad: **TMM**
- Cuidado con los **perdidos**: cada perdido aislado supone L perdidos en el análisis.
- Estudios **multi-ciudad**: interesa curva resumen (mañana!!)
- Estudios para **subperiodos** (por ejemplo: verano)
  - Ajuste:

    ```
    dats<-subset(dat2,mm %in% 5:9)
    cb <- crossbasis(pred, lag=L, argvar, arglag, group=dats$yy)
    ```
  - Se perderán los L primeros datos (retardos) en cada grupo