Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 3

Robbie M Parks, PhD

22nd July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Outline from previous lecture

- Case crossover design
- Time series design

Outline

- Distributed lag non-linear models (DLNM)
- Case crossover with hybrid DLNM
- Treed distributed non-linear models (TDLNMs)

Critical windows of exposure

- When is exposure most important?
- Think of the examples we've seen so far in class:
- County-level annual $PM_{2.5}$ and BMI
- Daily $PM_{2.5}$ and CVD admissions
- 3-day average $PM_{2.5}$ and CVD admissions

- What assumption were we making?

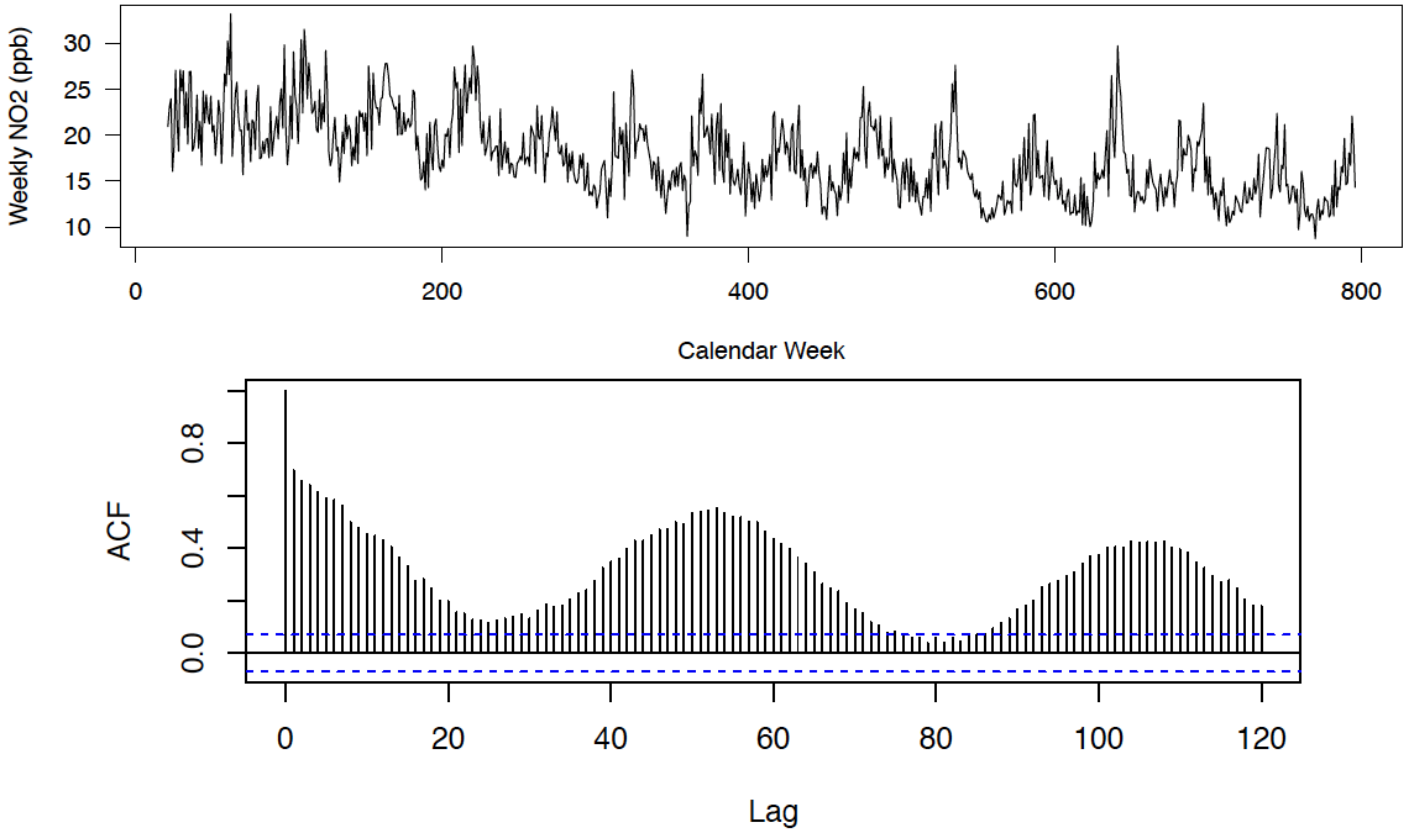
- What if we do not know?

Critical windows of exposure

- Examine multiple different windows of exposure
  - Given some expert knowledge (hopefully!) and prior hypothesis
  - E.g. look at short-term exposure to $PM_{2.5}$ and CVD
- How?
- Before we go into that, what is a lag?

Critical windows of exposure

- Run a different model for each lag of interest
  - Any issues with that?



- Other options?

Unconstrained distributed lag models

- Include multiple lags in the same model

$$\log(\mathrm{E}[\mathrm{CVD}_t] = \beta_0 + \beta_1 PM2.5_t + \beta_2 PM2.5_{t-1} + \beta_3 PM2.5_{t-2} + \cdots$$

  - β_1 effect estimate for lag 0 (same day)
  - β_2 effect estimate for lag 1 (day before the event)
  - β_3 effect estimate for lag 2 (2 days before the event)
  - …

- Independent effect for each lag, adjusting for other lags

Unconstrained distributed lag models

Include multiple lags in the same model

$$\log(\mathrm{E}[\mathrm{CVD}_t] = \beta_0 + \beta_1 PM2.5_t + \beta_2 PM2.5_{t-1} + \beta_3 PM2.5_{t-2} + \cdots$$

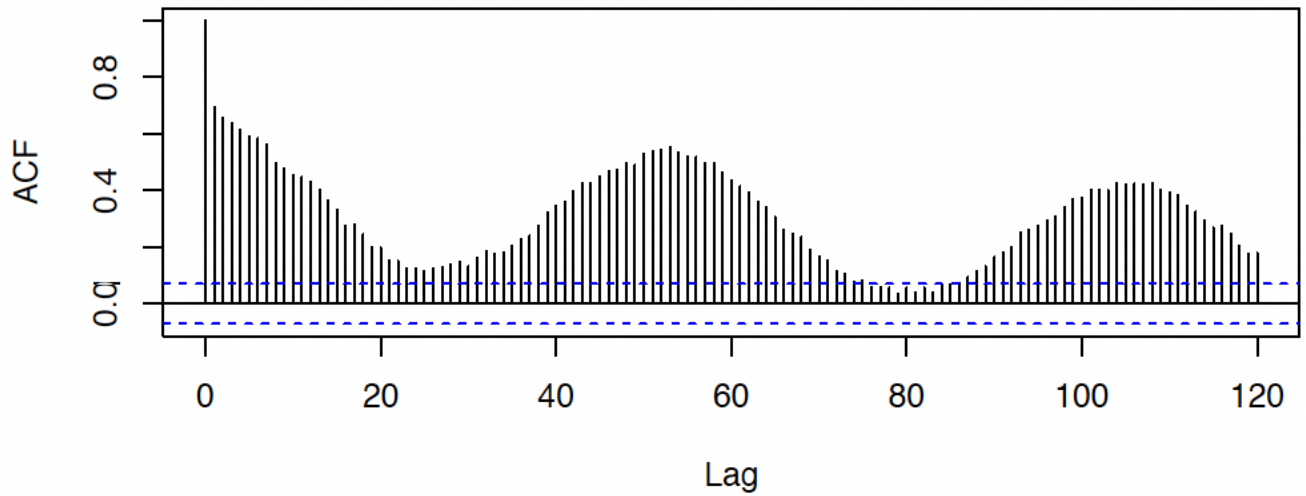
Cumulative effect of exposure over K lags $\sum_{k=1}^{K} \beta_k$

- Can I do this with lag-specific models?

- Variance of the cumulative effect?
  - $Var(\beta_1 + \beta_2) = Var(\beta_1) + Var(\beta_2) + 2Covar(\beta_1, \beta_2)$
  - (Can be extended for more than two variables
- For small in magnitude effect estimates, the estimated cumulative effect will be very similar to the effect of the average exposure of the same period

Unconstrained distributed lag models

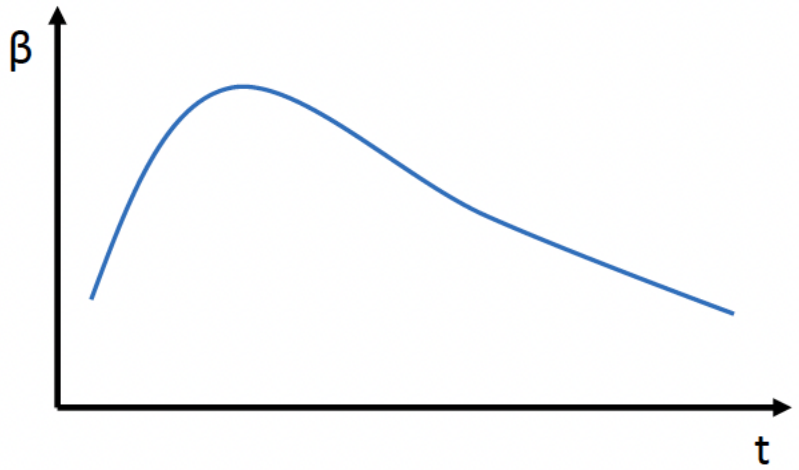
- Any issues now with this approach?



- Other options?

Constrain the distributed lag models

- Add an additional constrain on how the effect estimates can vary over time (i.e., lags)
- k = f (t)
- Underlying biology suggesting that day-to-day (or
- week-to-week etc) effects k likely have some structure
- How do we select f (t)? Any ideas?



Constrain the distributed lag models

Polynomials

- Up until recently that people had to hard code these, a polynomial of 4th degree was most commonly used
- Up to how many changes in direction for a 4th degree polynomial?

- Recent coding advances allow for more parameterizations
  - dlnm package in R, provided by Antonio Gasparrini
- E.g. natural splines

Constrain the distributed lag models

- Now add non-linear element: dlnm:
- We can also now fit non-linear functions for the exposure-response curve
  - What does this mean?
  - E.g. polynomials and natural splines
- Also, allowing a different shape of the exposure-response curve at each lag

Few more notes on DLNMs

- The exposure windows need to have equal duration
  - E.g. one day, one week, 5 weeks, etc
- In lab we'll learn about DLNMs in time series
- DLNMs can also be used in other study designs as well
  - Cohort, Case crossover, Survival, etc
- For time series, the dlnm package will create the exposure matrix for us
- For other designs, we have to create it ourselves
- In Bayesian world, can spatially smooth for small areas (cutting edge)
- As always - emphasis on interpretation and biological plausibility!

Case crossover with hybrid DLNM

- Combining case crossover study design with distributed lag non-linear terms:
- In practise, this becomes:
  - Long table with case and crossover controls
  - Wide table with lagged exposures

Case crossover with hybrid DLNM

- Combining case crossover study design with distributed lag non-linear terms:
- In practise, this becomes:
  - Long table with case and crossover controls
  - Wide table with lagged exposures

Case crossover with hybrid DLNM

nr	zcta	DayName	date_control	ck	lag0	lag1	lag2	lag3	lag4	lag5	lag6
1	14727	CaseDay_0	1995-08-03	1	23.70	23.29	22.65	20.98	20.10	23.58	23.32
1	14727	After_1	1995-08-10	0	19.95	17.85	17.48	19.84	20.31	20.25	23.03
1	14727	After_2	1995-08-17	0	22.37	24.54	25.13	24.57	20.65	22.10	21.48
1	14727	After_3	1995-08-24	0	17.03	17.15	16.16	21.36	21.13	20.78	22.40
1	14727	After_4	1995-08-31	0	21.52	16.93	19.41	17.91	18.33	17.32	12.99
2	14739	CaseDay_0	1995-10-12	1	14.46	12.96	12.41	9.12	9.19	14.69	17.48
2	14739	After_1	1995-10-19	0	13.51	11.10	5.79	4.41	5.37	12.94	15.75
2	14739	After_2	1995-10-26	0	5.65	5.95	10.69	10.49	5.43	8.09	12.56
2	14739	Before_1	1995-10-05	0	13.20	15.50	11.93	14.66	14.41	12.84	12.67
3	14895	CaseDay_0	1995-12-27	1	-9.89	-8.96	-6.41	-5.98	-5.41	-6.96	-9.85
3	14895	Before_1	1995-12-20	0	-12.05	-9.78	-6.82	-6.70	-4.06	-1.96	-3.60
3	14895	Before_2	1995-12-13	0	-11.10	-12.85	-12.95	-13.27	-5.60	-6.46	-6.47
3	14895	Before_3	1995-12-06	0	-4.05	-1.34	-1.98	-0.33	-2.92	0.36	-4.41
4	14715	CaseDay_0	1995-05-01	1	6.46	6.09	6.86	6.12	11.09	6.61	5.25
4	14715	After_1	1995-05-08	0	4.09	6.20	7.58	8.79	9.51	8.84	7.84
4	14715	After_2	1995-05-15	0	11.28	13.06	12.85	10.27	11.54	13.27	7.20
4	14715	After_3	1995-05-22	0	9.70	12.81	11.49	9.31	10.33	12.23	11.72
4	14715	After_4	1995-05-29	0	15.72	11.95	12.13	11.76	12.25	15.92	13.71
6	14895	CaseDay_0	1995-07-07	1	20.05	22.30	21.94	18.85	13.68	14.19	19.22

Treed distributed non-linear models (TDLNMs)

- **What Are Treed DLNMs?**
  - Extension of DLNMs using recursive partitioning (tree-based methods)
  - Allows for *data-driven identification* of effect heterogeneity
  - Combines exposure–lag–response modeling with subgroup detection
- **Why Use Treed DLNMs?**
  - Capture **non-linear**, **lagged**, and **heterogeneous** exposure–response relationships
  - Identify **subgroups** with differing temporal risk profiles
  - Avoids pre-specification of strata or interaction terms
- **How Do They Work?**
  - Fit a DLNM at each node of a regression tree
  - Recursive splits based on covariates (e.g., age, SES, geography)
  - Each terminal node represents a subgroup with its own DLNM

Treed distributed non-linear models (TDLNMs)

- **Strengths**
  - Flexible modeling of **complex temporal effects**
  - Uncovers **effect modification** without a priori assumptions
  - Useful for high-dimensional data or unknown structure
- **Limitations**
  - Computationally intensive
  - Risk of **overfitting** without cross-validation
  - Interpretation can be complex for large trees
- **Applications**
  - Environmental epidemiology: temperature, air pollution, etc.
  - Useful when **time-varying effects differ across populations**

Outline

- Distributed lag non-linear models (DLNM)
- Case crossover with hybrid DLNM
- Treed distributed non-linear models (TDLNMs)

Getting ready for the lab

- This lab will involve taking some models and concepts from the **Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 3** lecture and introduce you to the way non-linear regression works:

Application

- How can you imagine applying this learning to your data and your research questions?

Questions

- Questions?

Modelling of complex, non-linear relationships in time series data while accounting for delayed effects 3

Robbie M Parks, PhD

22nd July 2025

Email: robbie.parks@columbia.edu

BlueSky: @robbiemparks

Website: sparklabnyc.github.io

Art by Amy Wolfe

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH