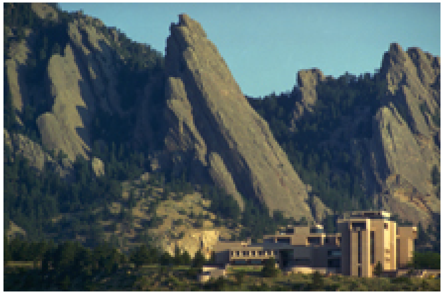
Modern regression and Mortality

Doug Nychka
National Center for Atmospheric Research
www.cgd.ucar.edu/~nychka

- **Statistical models**
- **GLM models**
- **Flexible regression**
- **Combining GLM and splines.**



Statistical tools for the analysis of Mortality related to air pollution and temperature.

The goal of this talk is to give a gentle introduction to the models used by Zeger, Dominici et al. SPH Johns Hopkins for explaning mortality based on environmental factors.

The Johns Hopkins group is focused on the effect of high levels of particulates (PM10) on short term, non-accidental mortality rates.

This leverages the National Morbidity, Mortality and Air Pollution Study Database (NMMAPS). (Welty, Peng and McDermott)

Schematic of PM10 model:

Mortality in an urban center =

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

+ seasonal effects

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

+ seasonal effects

+ particulates + temperature/humidity stress +

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

+ seasonal effects

+ particulates + temperature/humidity stress +

+ interaction of temperture with age

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

+ seasonal effects

+ particulates + temperature/humidity stress +

+ interaction of temperture with age

+ random component.

Schematic of PM10 model:

Mortality in an urban center =

Calendar effects + Age categories

+ seasonal effects

+ particulates + temperature/humidity stress +

+ interaction of temperture with age

+ random component.

This work is a useful motivation for understanding

Poisson Regression

Flexible regression models

Some data atributes:

- A response depends on other variables: (Mortality depends ontemperature and PM.)

$$Y_k = f(X_k) + noise$$

  Here we interpret the noise as having a mean of zero.

- The explanatory variables can be either catgorical (age category, day-of-week) or continuous (daily average temperature, PM)
- The response may be discrete or NonGaussian

Issues for data analysis

How does one specify all of these components in a simple and unambigous fashion?

How does one estimate a multivariate functional relationship?

Formula's

Some variables:

`mort`: **Daily, nonaccidental mortality for three age categories.**
`time`: **Day**

`tmp`: **Daily average temperature**
`tmp3`: **Daily average temperature for past three days**

`PM`: **Daily Particulate ($< 10\mu$g)**

`AgeCat`: **Three age categories <65, 65-74, >75**

Formula for a linear regression of mortality on 3-day temperature

```
mort ~ tmp3
```

3-day temperature and PM

```
mort ~ tmp + PM
```

Dependence on the age categories

```
mort ~ AgeCat + tmp + PM
```

These are additive models because the variables appear by themselves and the contribution of each can be inferred from their individual values.

Interactive dependence on the age categories

```
mort ~ AgeCat+ tmp+ AgeCat:tmp
```

or just

```
mort ~ AgeCat*tmp
```

Three different slopes and intercepts, the : is an interaction. * includes all possible terms. There are several conventions that make this not as transparent in the fit.

This is more interpretable.

```
mort ~  AgeCat + AgeCat:tmp -1
```

```
Coefficients:
    AgeCat1      AgeCat2      AgeCat3  tmp:AgeCat1  tmp:AgeCat2  tmp:AgeCat3
    66.804        44.522       121.158   -0.156        -0.139       -0.473
```

All models are wrong, but some are useful ...

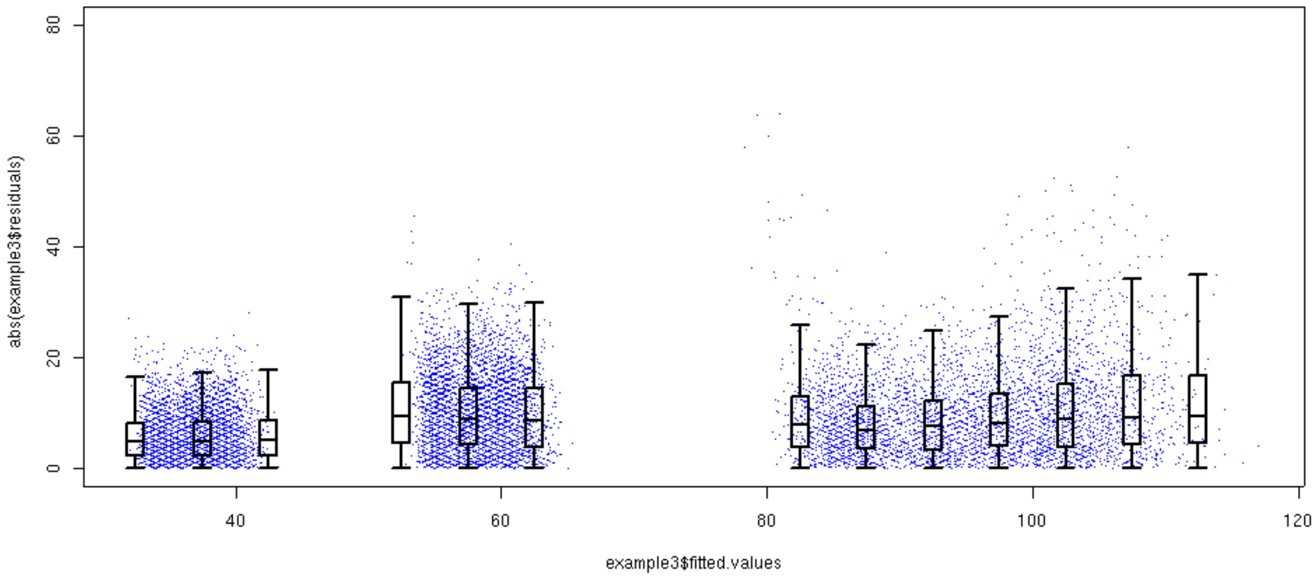
All models are wrong, but some are useful ...

but some models are more wrong than others!

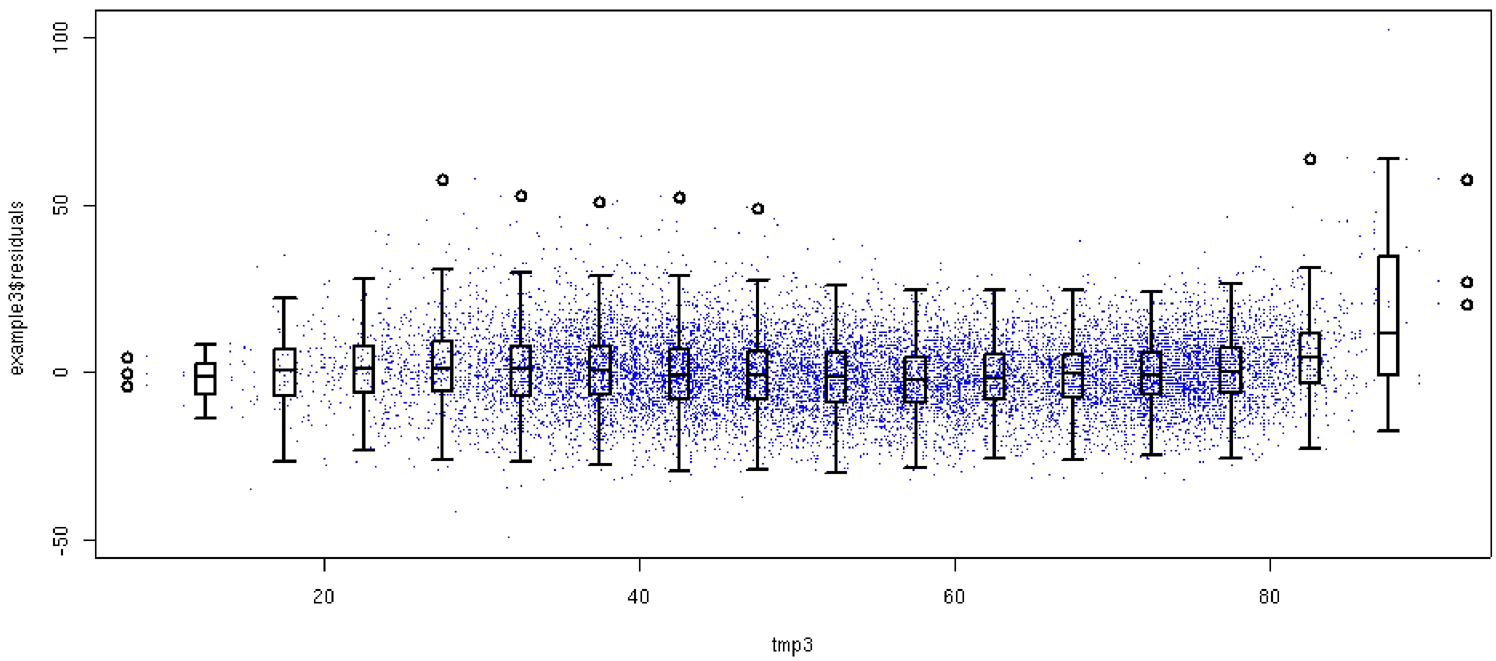
All models are wrong, but some are useful ...

but some models are more wrong than others!

Changing variance is missed.
Absolute residuals from 3 slopes model.



Subtleties of the temperature response are missed!



A Generalized Linear Model (GLM)

Assume that mortality (M_t) is approximately Poisson distributed

Mean model with log link function

$$M_t \sim Poisson(\mu_t)$$

$$E(M_t) = \mu_t$$

model the *log* of the mean.

$$log(\mu_t) = f(\mathbf{covariates})$$

Variance model with over-dispersion

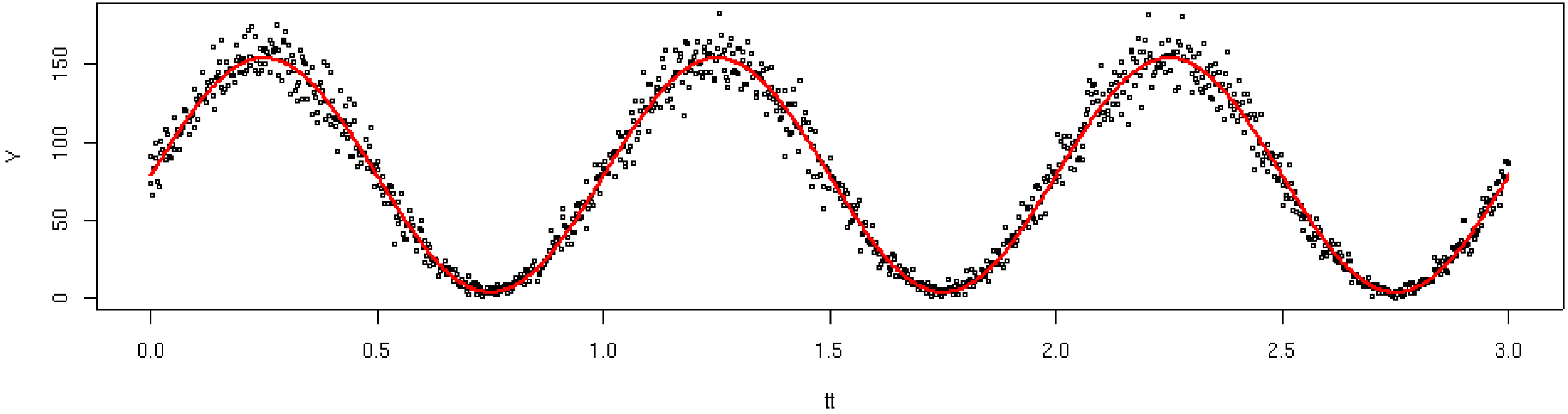
The Poisson distribution has a variance equal to the mean:
$Var(M_t) = \mu_t$. **This allows modeling of both the mean and variance simultaneously.**

Often the Poisson does not include enough variability so an additional parameter is included:

$$Var(M_t) = \phi\mu_t$$

ϕ **the over-dispersion parameter.**

Example of a Poisson Response



GLM fit to mortality

First consider the linear model but using the Poission regression. $log(\mu_t)$ **has three different slopes as a function of temperature based on the age categories.**

In R-code:

```
glm( mort~ AgeCat*tmp3, family=quasipoisson)
```

Results in residuals that better fit model assumptions.

Flexible Regression

Response of mortality is not linear, not clear what functional form to choose.

One strategy is to represent the unknown function as a linear combination of basis functions

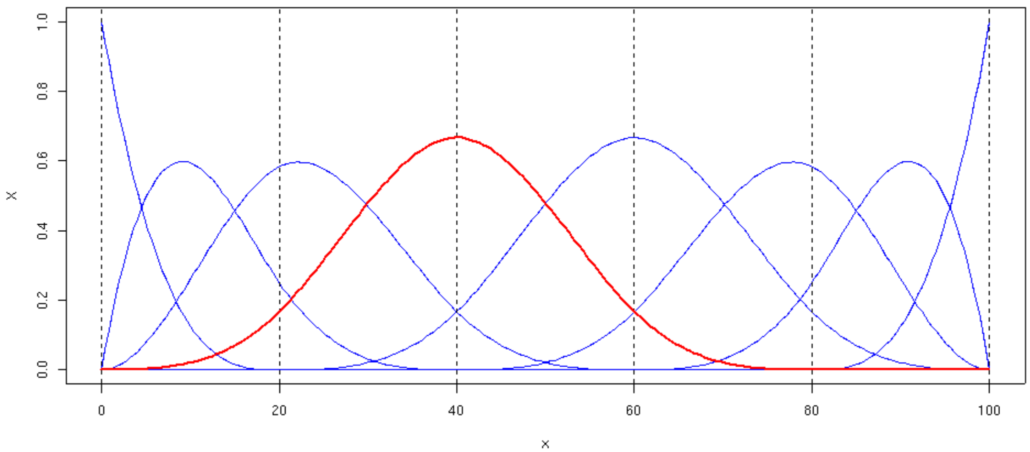
$$f(temp) = \sum_{j=1} M\psi_j(temp)\beta_j$$

Splines: **Local basis functions that allow one to control the flexibility of the shape of f.**

Three factors that control splines:

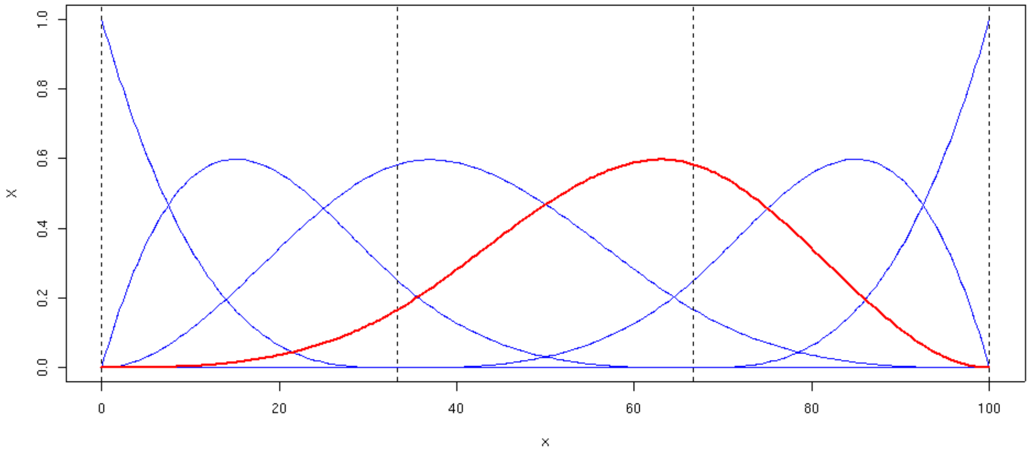
- **number of knots**
- **location of knots (usually equally spaced)**
- **order of fit (usually cubic)**

A cubic B-spline basis for temperature (6 knots)

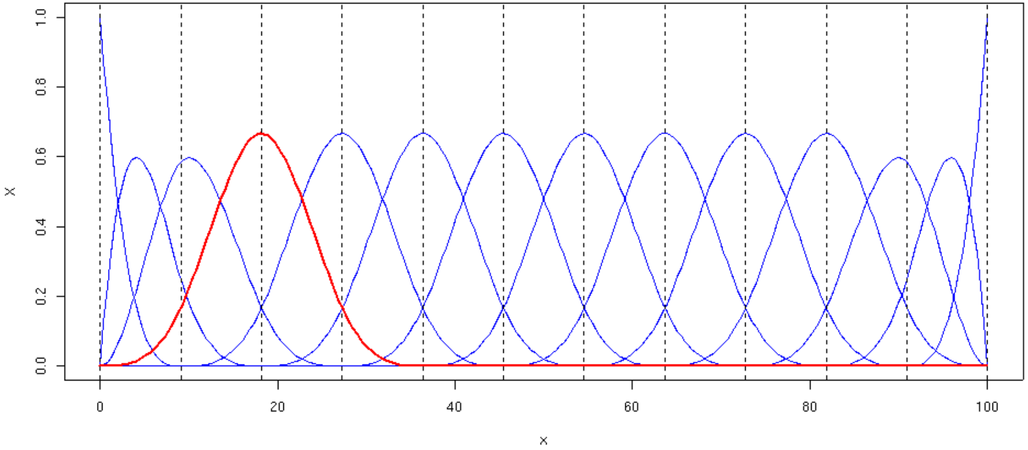


Functions are piecewise cubic in between knots.

Controlling the resolution



4 knots



12 knots

Applying the spline model to temperature

The simplest model is now

$$log(\mu_t) = AgeCat + f(tmp3) = AgeCat + \sum_{j=1}^{M} \psi_j(tmp3_t)\beta_j$$

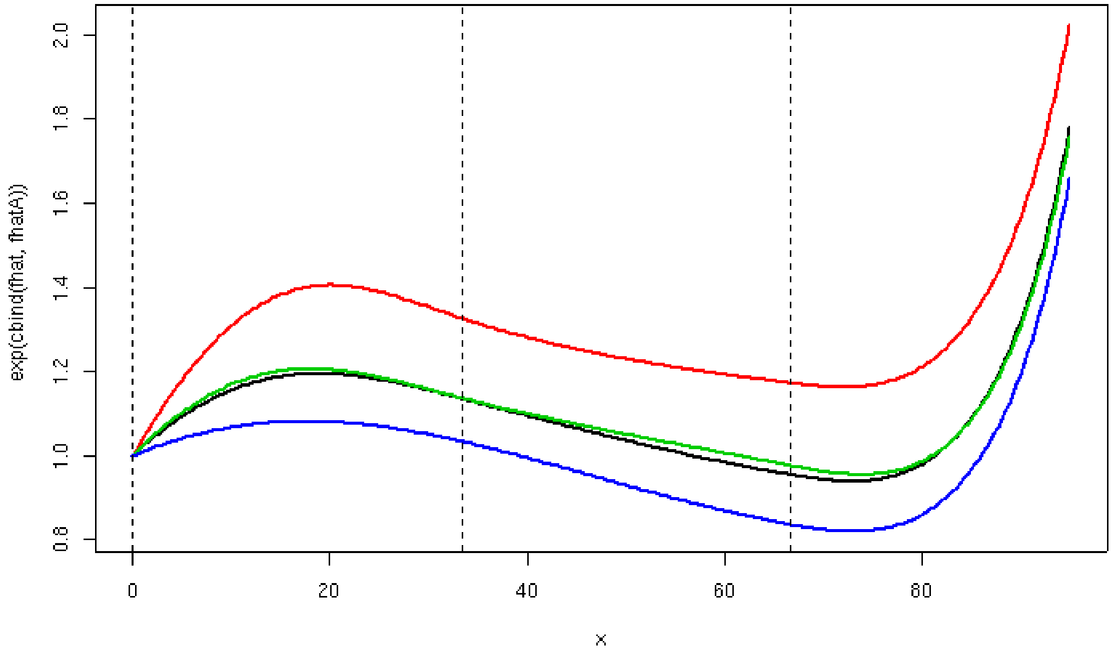
where ψ_j are the B-splines.

No interaction here between age category and the temperature response.

An extension is to have a distinct response to temperature for each age category. This gives three seperate B-spline curves with different coefficients.

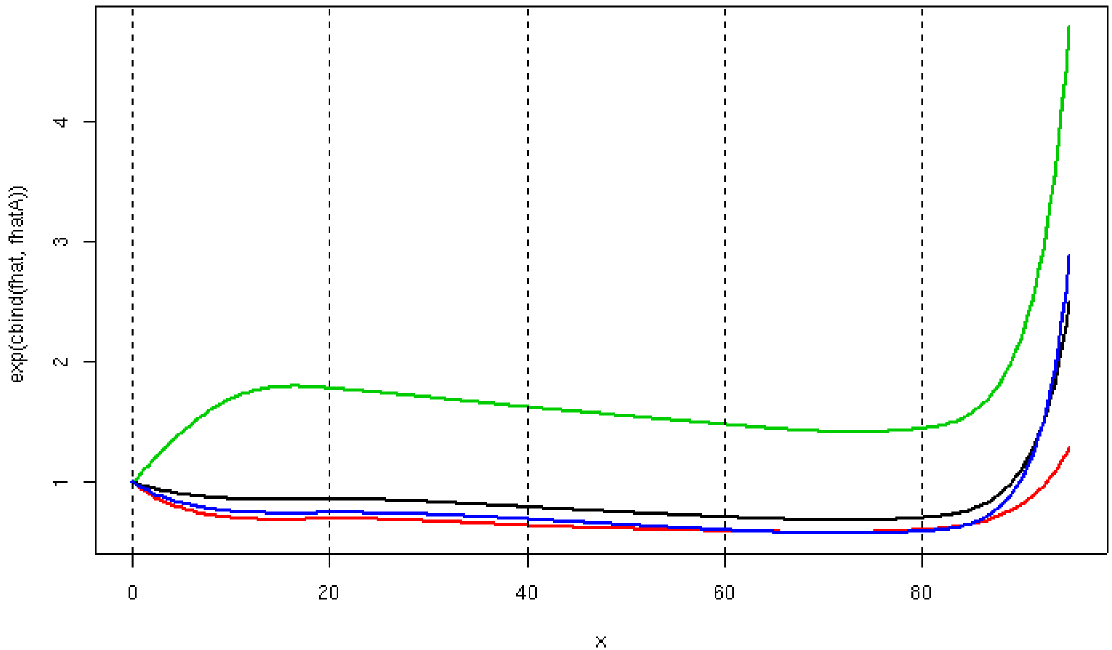
Results with 4 knots

Relative Risk estimates:



Results with 6 knots

Relative Risk estimates:



Issues of inference

Selecting the amount of curve flexibility

There are information criteria and cross-validation techniques to do this.

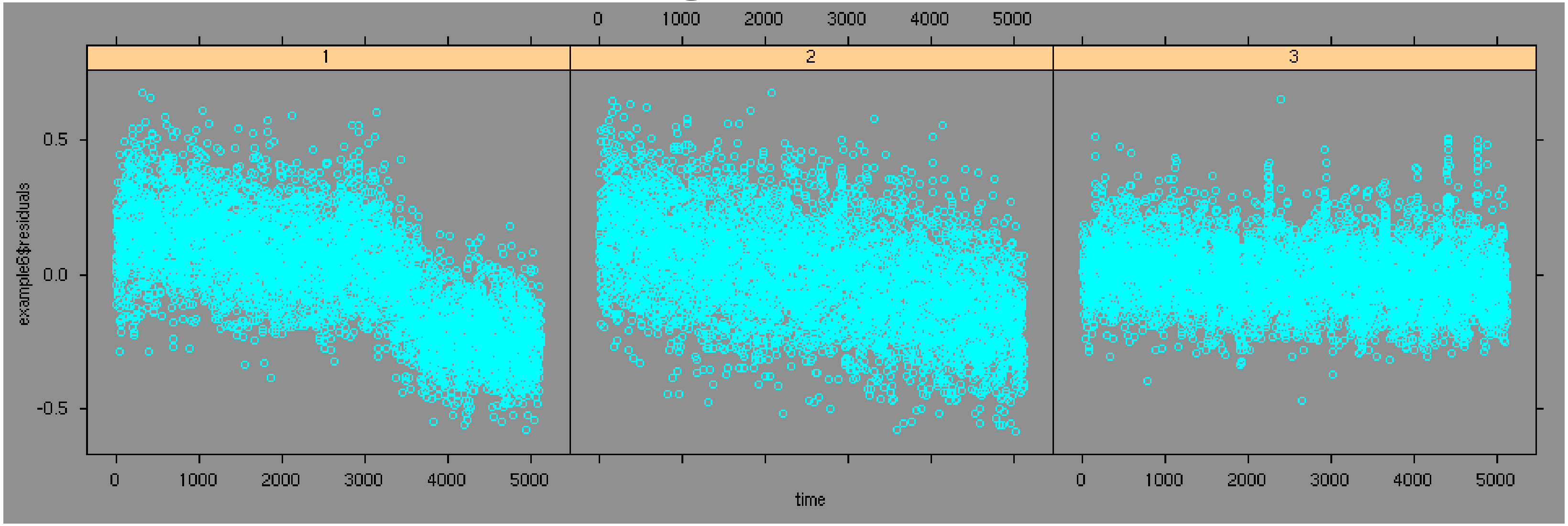
Uncertainty in estimates

Parametric bootstrapping, fitting synthetic data sets generated from the fitted model gives a useful measure of error.

Bayesian methods can also give an idea of the uncertainty of the estimated relationships.

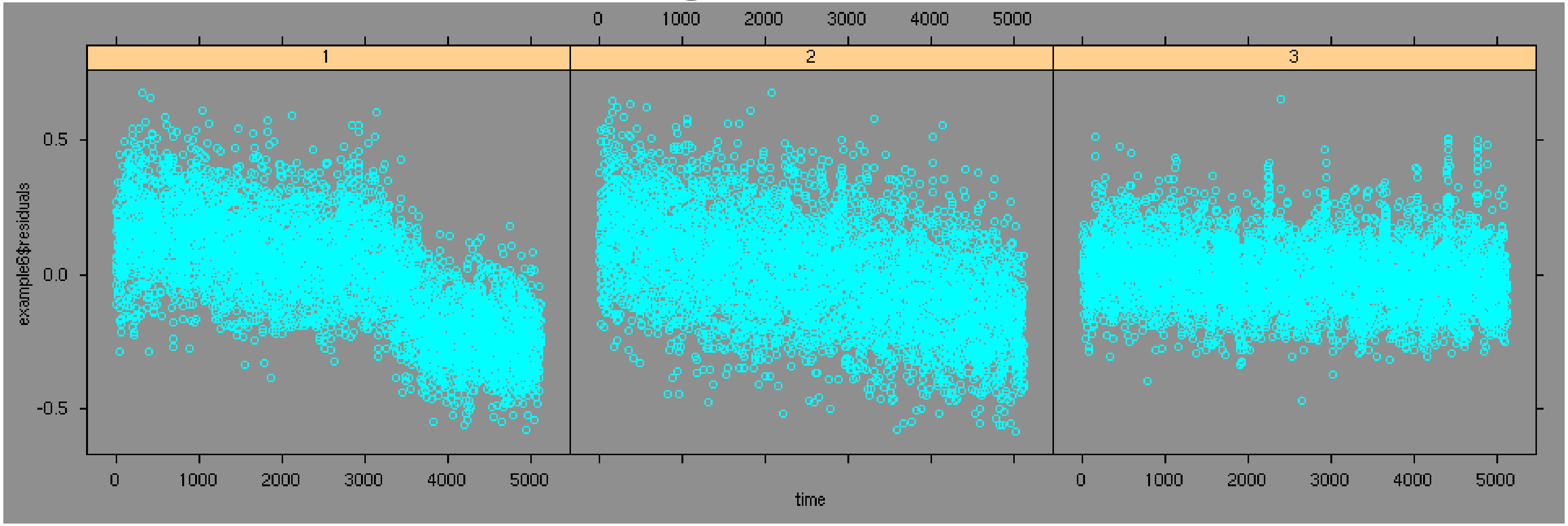
This talk is ending too soon!

Standardized residuals against time:



This talk is ending too soon!

Standardized residuals against time:

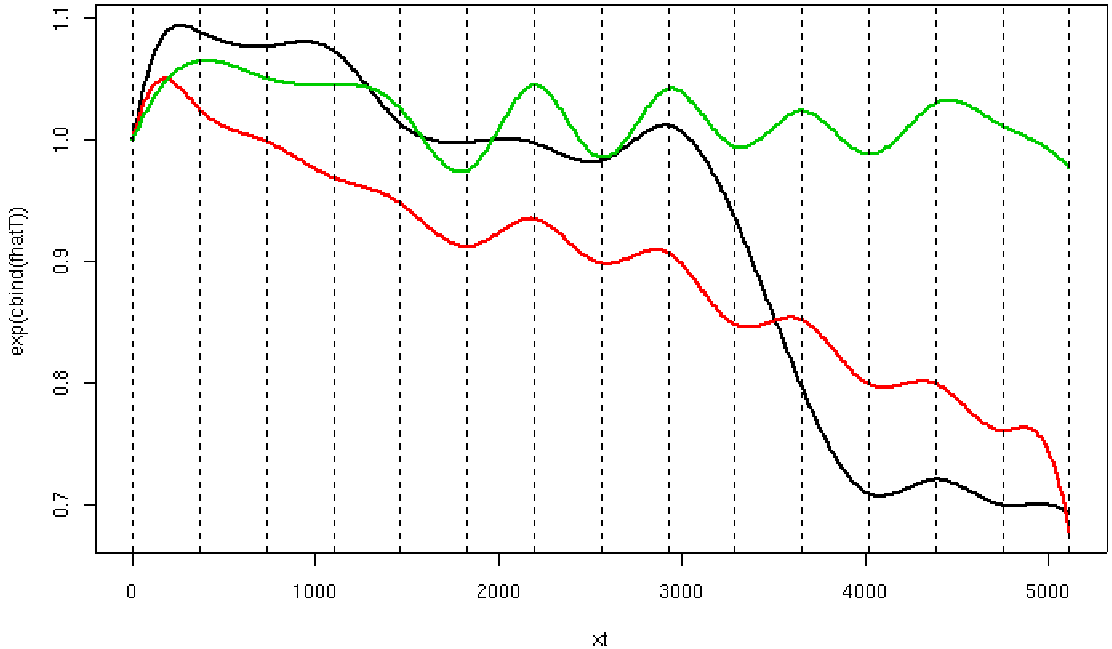


Use an additive model:

$$log(\mu_t) = f_1(\mathbf{tmp3}_t) + f_2(\mathbf{t})$$

f_1 **and** f_2 **are both modeled using B-splines**

Additive curves for time trends



Remarks

- There are many new statistical tools to discern structure in complex data sets.
- Inference can be formalized by Bayesian methods
- One challenge is to combine models across cities.