Introduction to Modelling II Extensions to GLMs for spatially structured data

ISAIR

Yang Liu

January 23, 2025

Linear models

- Recall that we can fit a *linear model* to show how $\boldsymbol{y}$ varies with some variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

where $X = [\boldsymbol{x}_1 | \boldsymbol{x}_2 | \ldots | \boldsymbol{x}_m]$. We'll see how to extend this model to be more useful.

"Essentially, all models are wrong but some models are useful" - George E. P. Box

Linear models - assumptions

- Linearity: every $\boldsymbol{x}_k$ has a linear effect on $\boldsymbol{y}$
- Multicollinearity: the $\boldsymbol{x}_k$ have little covariance between them
- Independence: all y_i and ε_i do not give information about y_j and ε_j
- Homogeneity of errors - all errors described by the same normal distribution

Linear models - reality

- Real epi/trial/health data is collected with some structure
- Effects not necessarily linear
- Measured variables might not explain *all* variation
- We might need to account for some group-specific effect
  - group might be an individual in a repeated measures design
  - group might be spatial location in a surveillance design

GAMs

- We can replace the linear terms in a GLM with some sort of scatterplot smoother (Hastie and Tibshirani 1986)
- No need to supply, *a priori* what we expect the functional form to be
- e.g. Eilers and Marx (1996) use penalised B-splines (De Boor 1978)

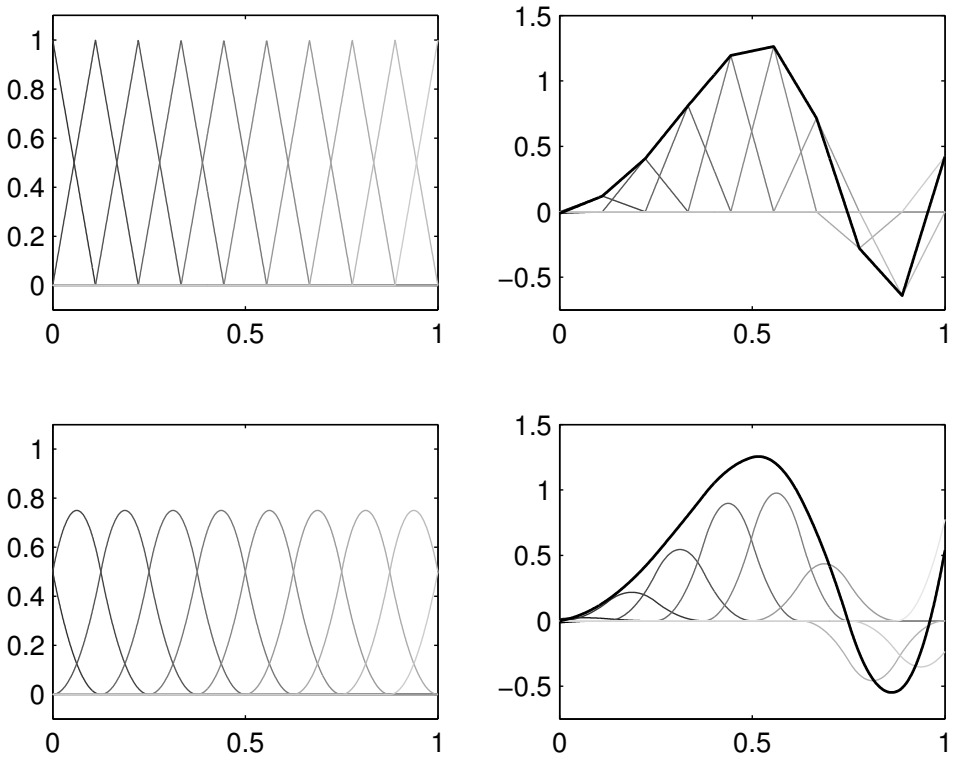


Figure 1: B-spline basis functions (left) and a linear combination thereof to reconstruct a curve

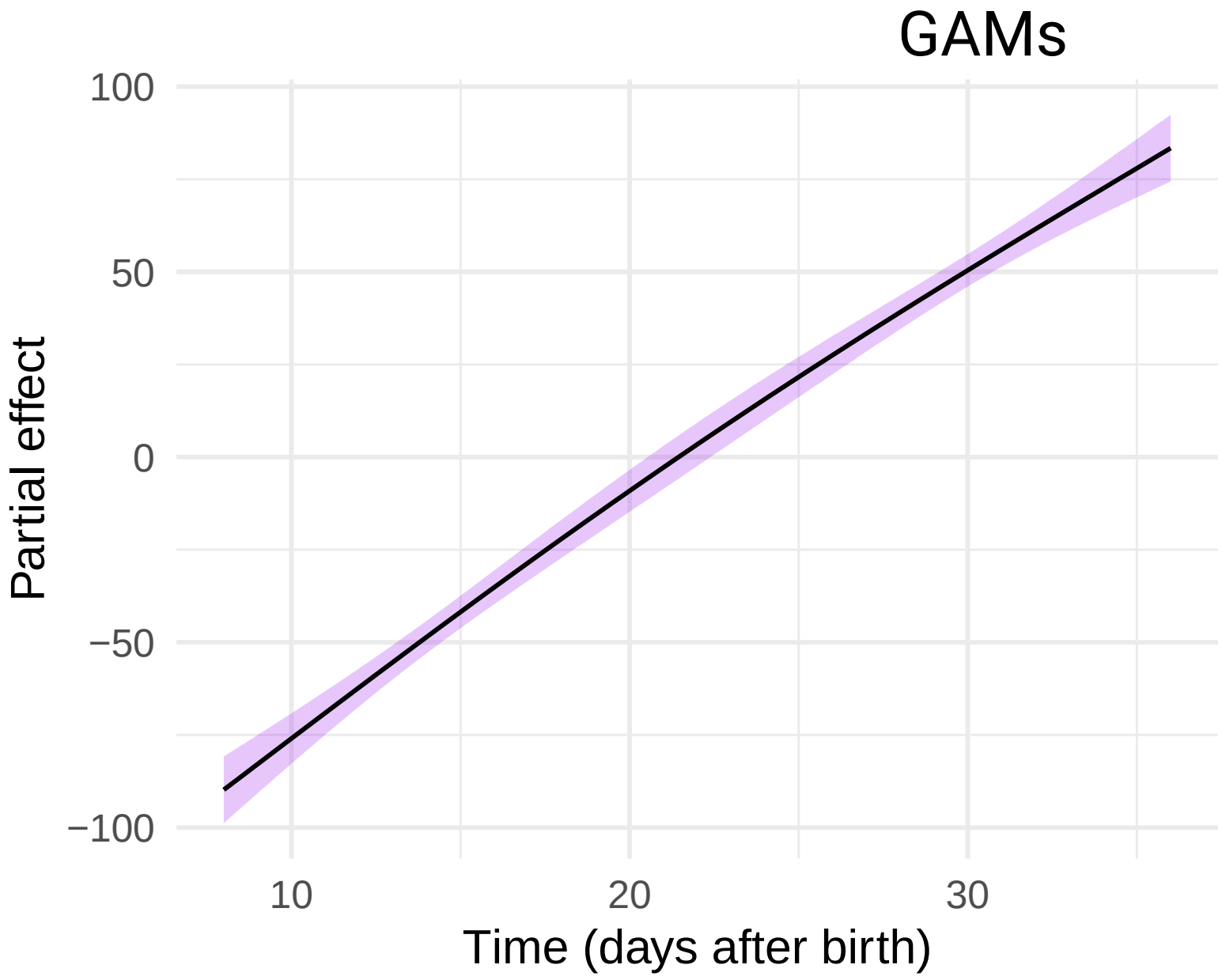
GAMs

- Replace linear terms in `glm()` with **s**mooth functions

```
library(mgcv)
rats_gam <- gam(data = rats_df,
                weight ~ s(time, k = 3))

summary(rats_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## weight ~ s(time, k = 3)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  242.653      1.295   187.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df    F p-value
## s(time) 1.839  1.974 1148  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Our smooth term is equivalent to a 1.84 degree polynomial

Model comparison

- AIC (Akaike 1974) a criterion for model comparison
- Trades off goodness of fit (log-likelihood, $\log \mathcal{L}$) and number of parameters in the model, k

$$\mathrm{AIC} = 2k - 2\log \mathcal{L}\left(\boldsymbol{y}\middle|\boldsymbol{x}, \widehat{\boldsymbol{\theta}}\right)$$

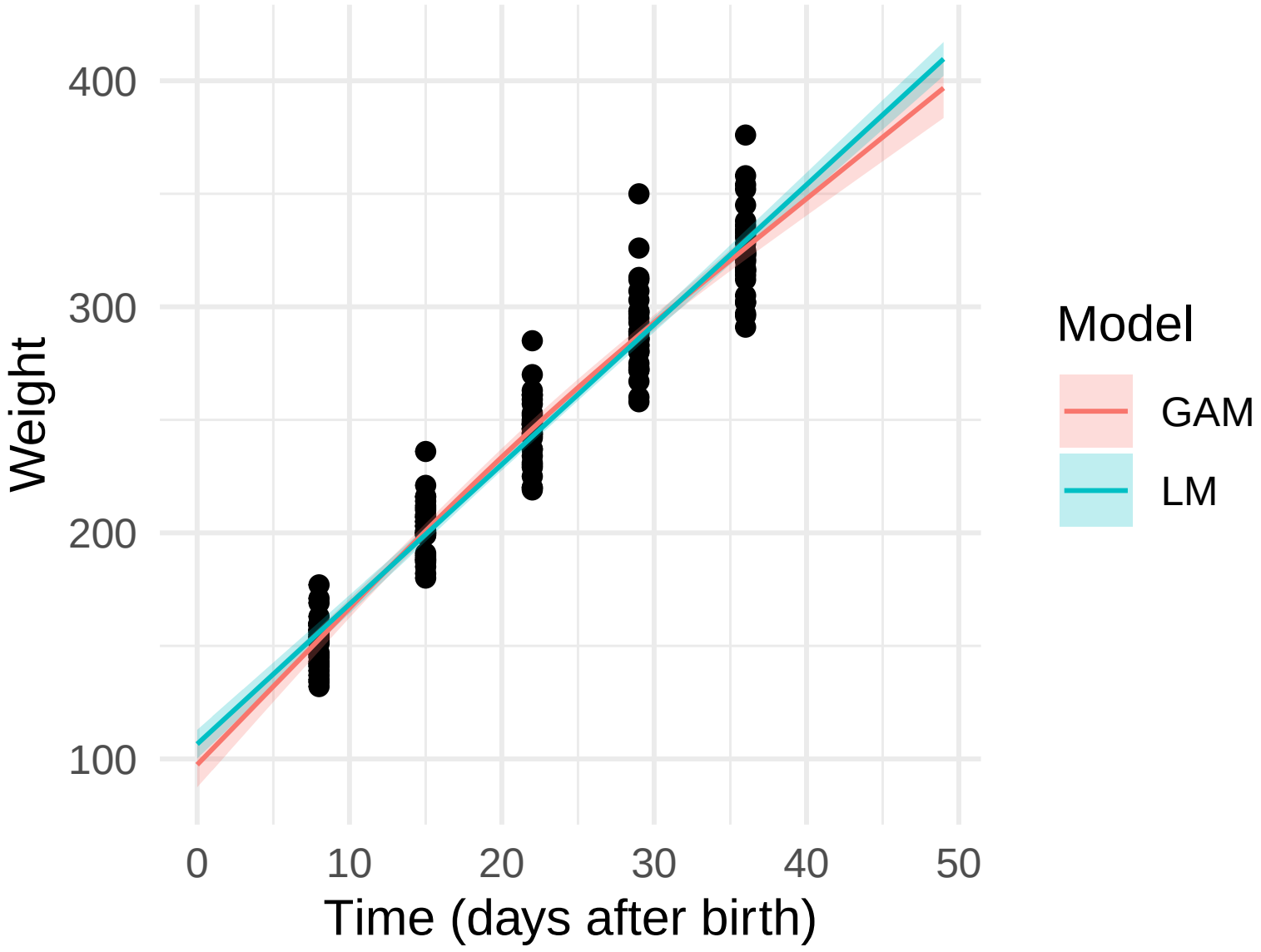
- $\widehat{\boldsymbol{\theta}}$ are parameter estimates for given model
- Relative measure of goodness of fit for $\boldsymbol{y}$ among a set of models
- Choose a model with lower AIC value
- Other ways to choose among a set of models, typically via cross validation

GAMs

Model	logLik	df	AIC	BIC
GAM	-625.93	2.84	1259.54	1271.10
LM	-628.96	2.00	1263.91	1272.95

Under AIC, GAM is preferred as AIC is smaller, though not by much

GAMs

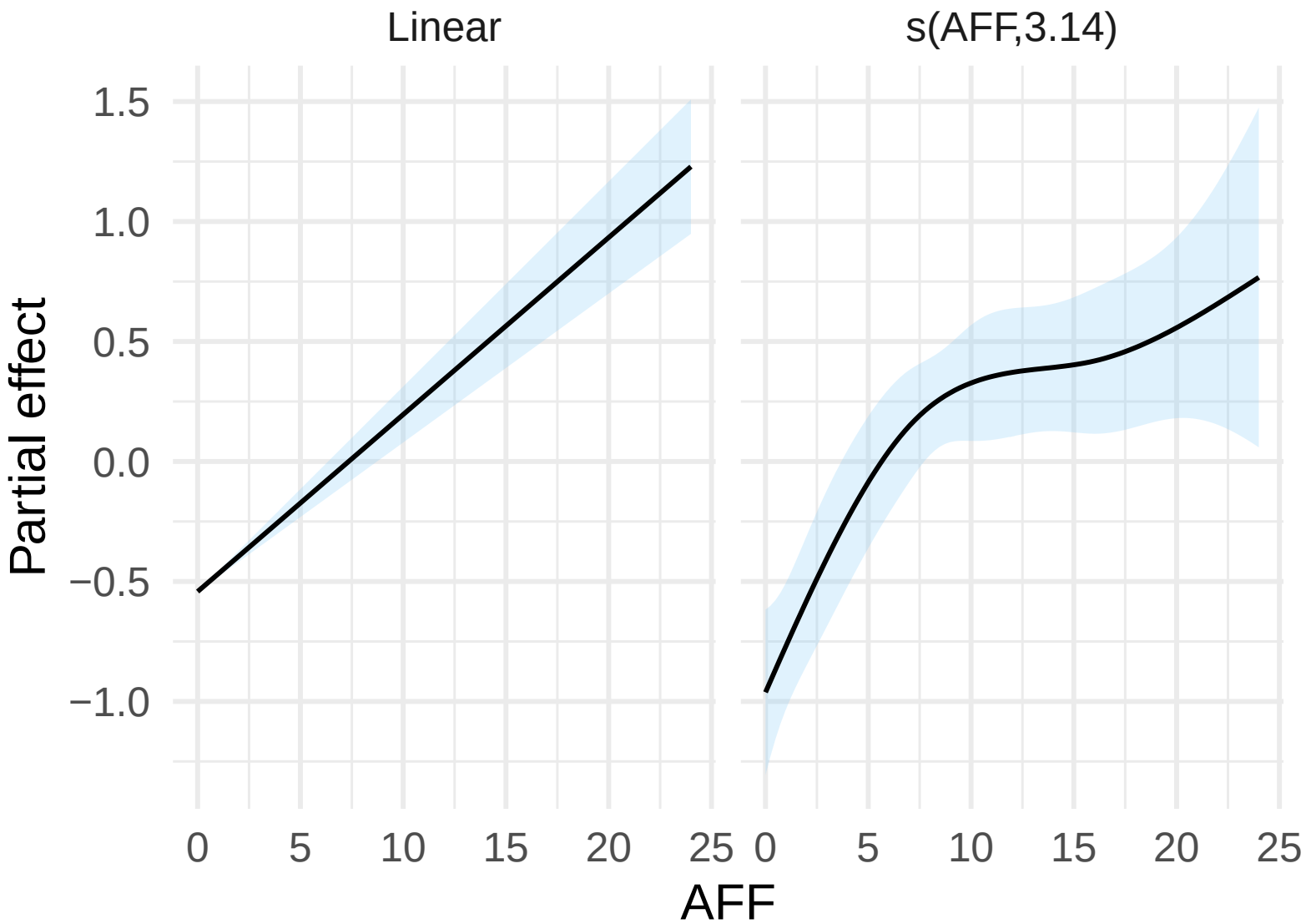

Weight
400
300
200
100
0
10
20
30
40
50
Time (days after birth)
Model
GAM
LM

GAMs - Poisson

- Perhaps the effect of AFF on lip cancer is non-linear, let's use a **s**mooth function instead
- `k=6` controls the size of the basis we wish to use, here a maximum of six degrees of freedom
- default `k` gives error due to few unique values of `AFF`

```
scotland_gam <- gam(data = scotland_data,
                    CANCER ~ offset(log(CEXP)) +
                      s(AFF, k = 6),
                    family = poisson())
```

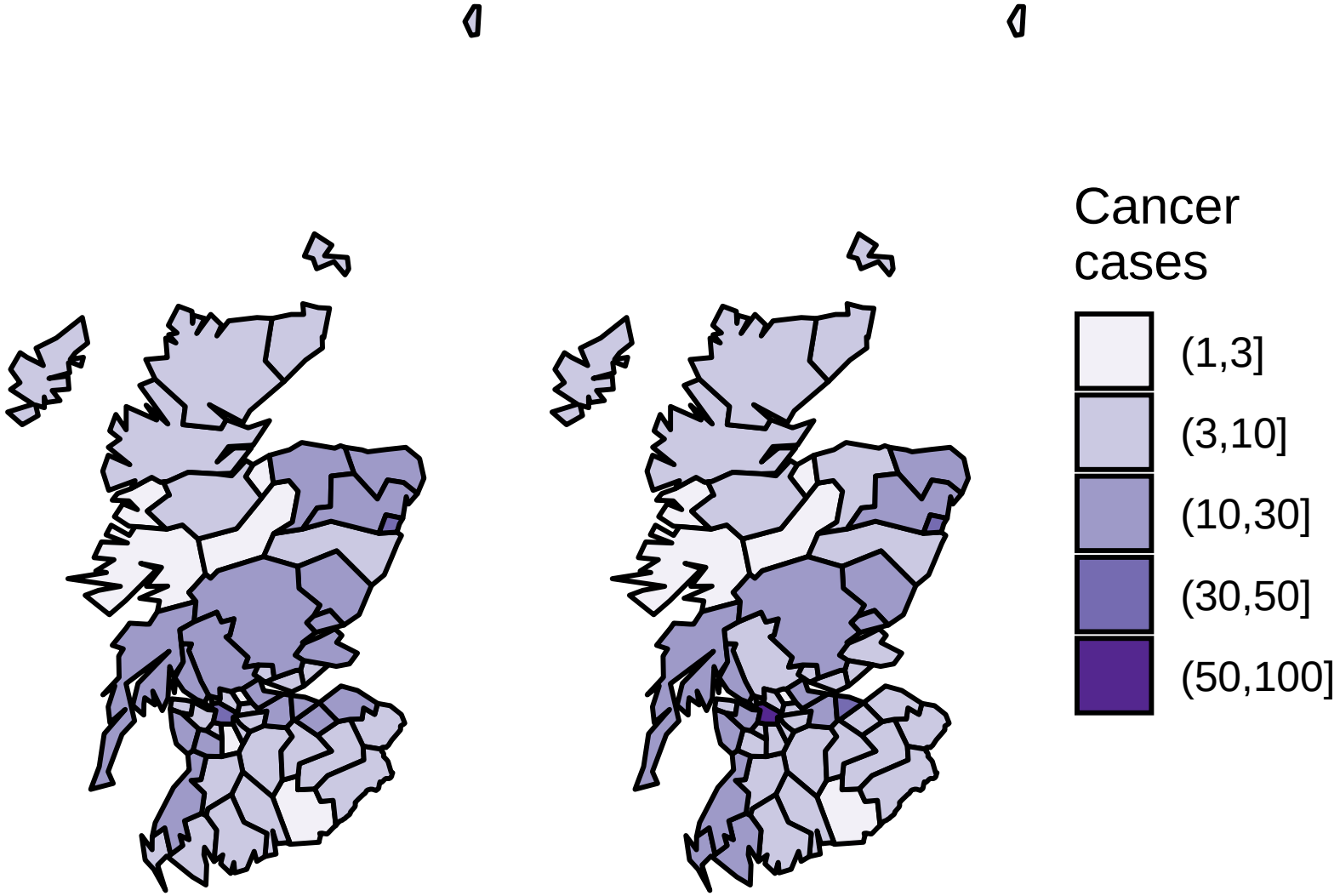
GAMs


Linear
s(AFF,3.14)
Partial effect
1.5
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
AFF

GAMs

GAM

GLM



GAMs

- GAMs are incredibly flexible
- Replace linear effects with additive effects
- Interested in overall effect rather than individual basis coefficients
- Penalisation of changes in second derivative can stop spline being too wiggly

GAMs

Complete the exercise fitting a GAM to the Haiti vaccine coverage data

Linear models

Recall the linear model in matrix form

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$$

We can expand this as

$$y_i = X_{i,*}\boldsymbol{\beta} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$$
$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

where we have n observations of m predictor variables

Linear models - random effects

- A random effect on the intercept if we believe that within each group there is a little *je nes sais quoi*,

$$y_{ij} = \beta_{0j} + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

- We assume that the β_{0j} are iid with a mean of zero and common variance,

$$\beta_{0j} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_0^2\right)$$

- For a model with a single explanatory variable, this is parallel lines

Linear models - random effects

- If there's a group-level adjustment required to account for between-group differences in the effect of an x, a *random effect* can be used:

$$y_{ij} = \beta_0 + \beta_{1j} x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_m x_{mij} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

where the data are now measurement i in group j and each group, $j \in \{1, 2, \ldots, J\}$ has its own slope,

$$\beta_{1j} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_1^2\right)$$

- For a single explanatory variable, this model is a group of lines that all have the same y intercept

Linear models - random effects

- **Fixed** effects: effects are fixed across groups
- **Random** effects: effects are random, but related, across groups, indicating that the influence of a predictor (or intercept) on outcome depends on which group the observation belongs to
- **Mixed** effects: combination of the two, allowing effects that are common to all groups and specific to each group

Linear models - random effects

- In general, we write linear mixed effects models as

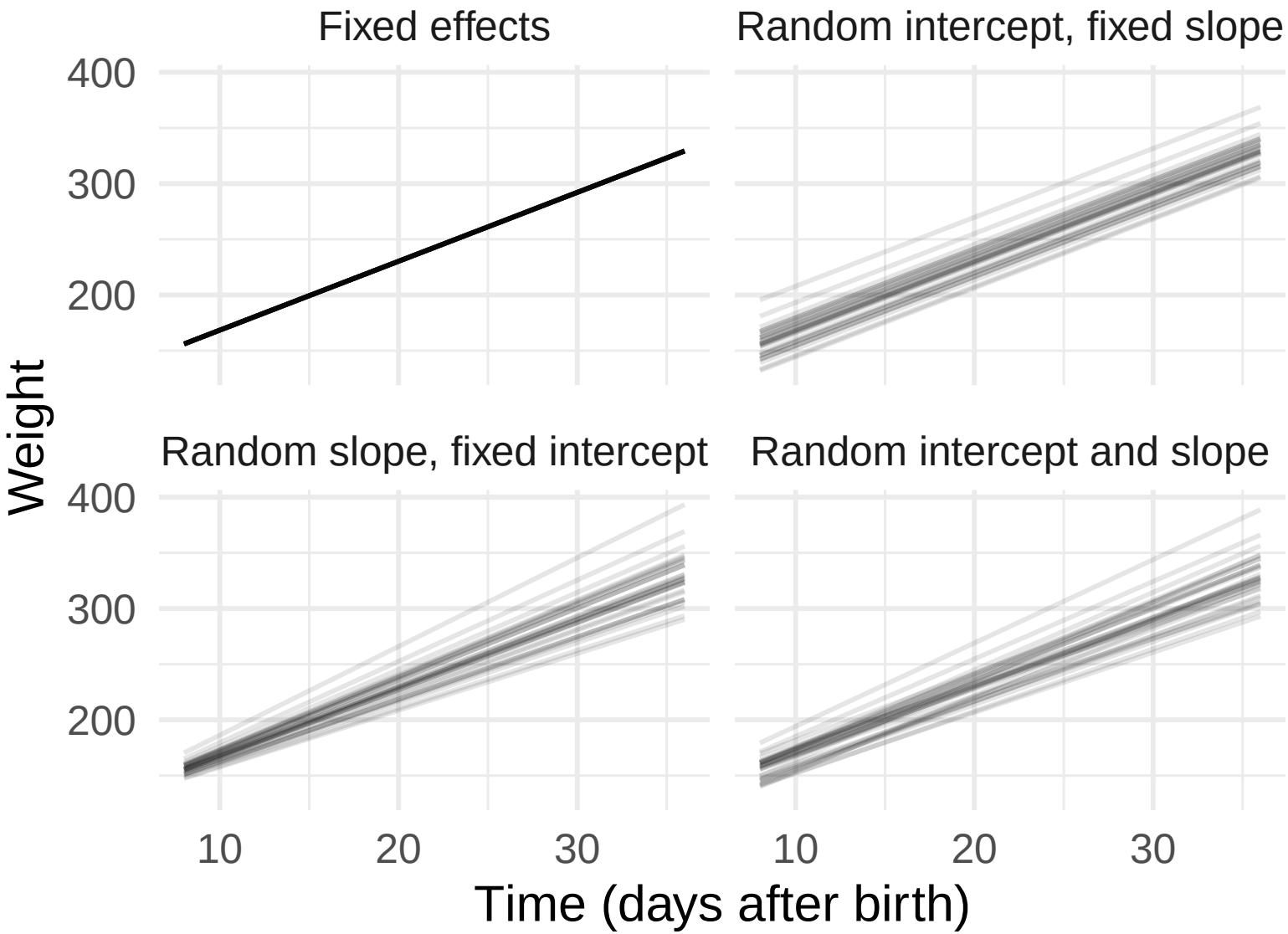
$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_m x_{mij} + u_{0j} + u_{1j} z_{1ij} + u_{2j} z_{2ij} + \ldots + u_{lj} z_{lij} + \varepsilon_{ij}$$

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\varepsilon}$$

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

where

- $X \in \mathbb{R}^{n \times m}$ is fixed effect matrix of predictors with coefficients $\boldsymbol{\beta}$
- $Z \in \mathbb{R}^{n \times lJ}$ is same for l random effects variables for J groups with coefs $\boldsymbol{u}$


Fixed effects
Random intercept, fixed slope
Random slope, fixed intercept
Random intercept and slope
Weight
400
300
200
10
20
30
Time (days after birth)

- Residuals are smaller and look more iid
- Be careful not to overparameterise model

Linear models - random effects

Complete the exercises fitting linear mixed models to the rats data

GLMMs

- GLM with mixed effects

$$g\left(\mathbb{E}(\boldsymbol{y})\right) = X\boldsymbol{\beta} + Z\boldsymbol{u}.$$

- z_{ij} is only non-zero if observation i is a member of the level that column j corresponds to
- R package mgcv (Wood 2017) fits mixed effects models with `gamm()`
- Specify `formula` as normal, and also `random` for structure of random effects
  - e.g. `random = list(ID = ~1)` for random intercept

GLMMs

Complete the exercises on vaccine coverage in Haiti

Summary

- GAMs provide more flexible estimation of mean
- Random effects provide way of handling structure in data
- We can have random effects structure on splines (GAMM)
- All model choices must be defensible
- Model comparison and diagnostics important aspect of fitting
- Most of the spatial models for the rest of the week can be fit as GAMs

References I

Akaike, H. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/tac.1974.1100705.

De Boor, Carl. 1978. *A Practical Guide to Splines*. Vol. 27. springer-verlag New York.

Eilers, Paul H. C., and Brian D. Marx. 1996. “Flexible Smoothing with b -Splines and Penalties.” *Statistical Science* 11 (2): 89–121. https://doi.org/10.1214/ss/1038425655.

Hastie, Trevor, and Robert Tibshirani. 1986. “Generalized Additive Models.” *Statist. Sci.* 1 (3): 297–310. https://doi.org/10.1214/ss/1177013604.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC.