Generalized additive models II

Patrick Breheny

October 13

Coronary heart disease study

- Today's lecture will feature several case studies involving the application of smoothing splines and GAMs to real data sets
- Our first study looks at many potential risk factors for coronary heart disease (CHD):
  - `sbp`: Systolic blood pressure
  - `tobacco`: Cumulative lifetime tobacco consumption (kg)
  - `ldl`: Low-density lipoprotein
  - `adiposity`: % of weight from fat
  - `obesity`: Body mass index
  - `famhist`: Family history of CHD
  - `typea`: A measure of stress
  - `alcohol`: Current alcohol consumption
  - `age`

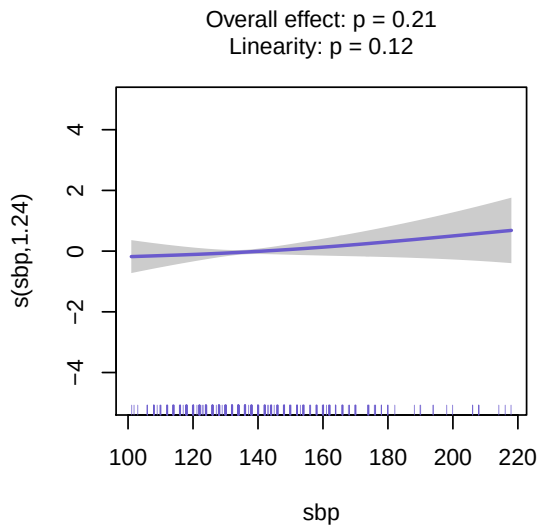
CHD study: Plan of analysis

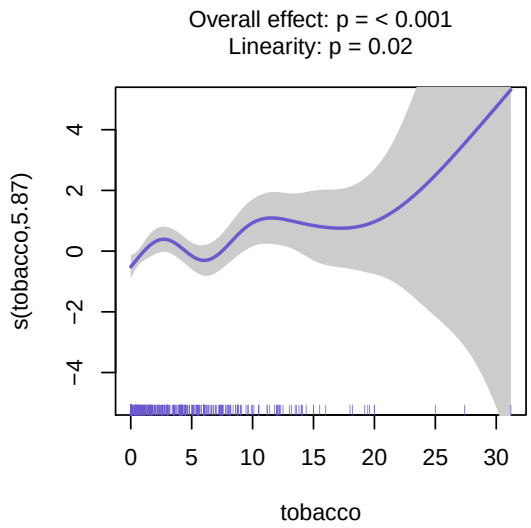
- There are many ways one could analyze this data, but since we are interested in all the potential risk factors, one reasonable way to proceed is to model each of the continuous factors using smoothing splines, with the degrees of freedom for each smooth estimated by GCV
- Here there is one categorical factor, family history, which is highly significant ($p < 0.0001$) with an odds ratio of 2.6 (1.6, 4.1)

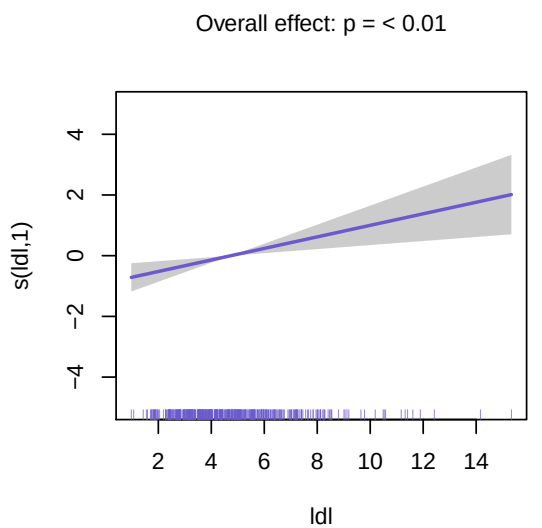
CHD study: Smoothing term tests

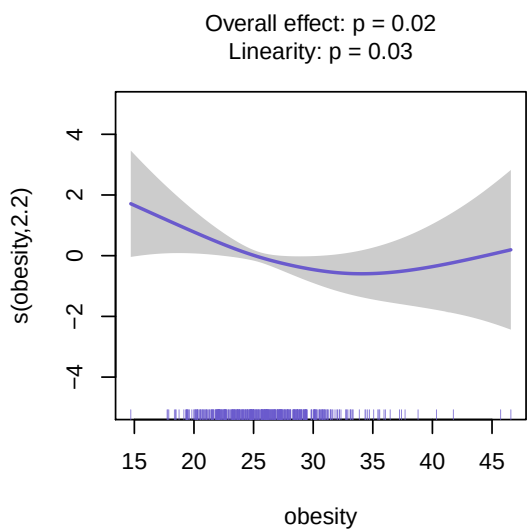
- The remaining 8 predictors are continuous; cumulatively they add 19.02 degrees of freedom to the model
- We are interested in carrying out likelihood ratio tests for these terms
- Because there are so many smooth terms, it is desirable to fix the smoothing parameters $\{\lambda_j\}$ at their full-model values; otherwise the hypothesis tests will become muddled by the fact that the reduced model differs not only in terms of its components, but also how flexible each term is allowed to be

CHD GAM

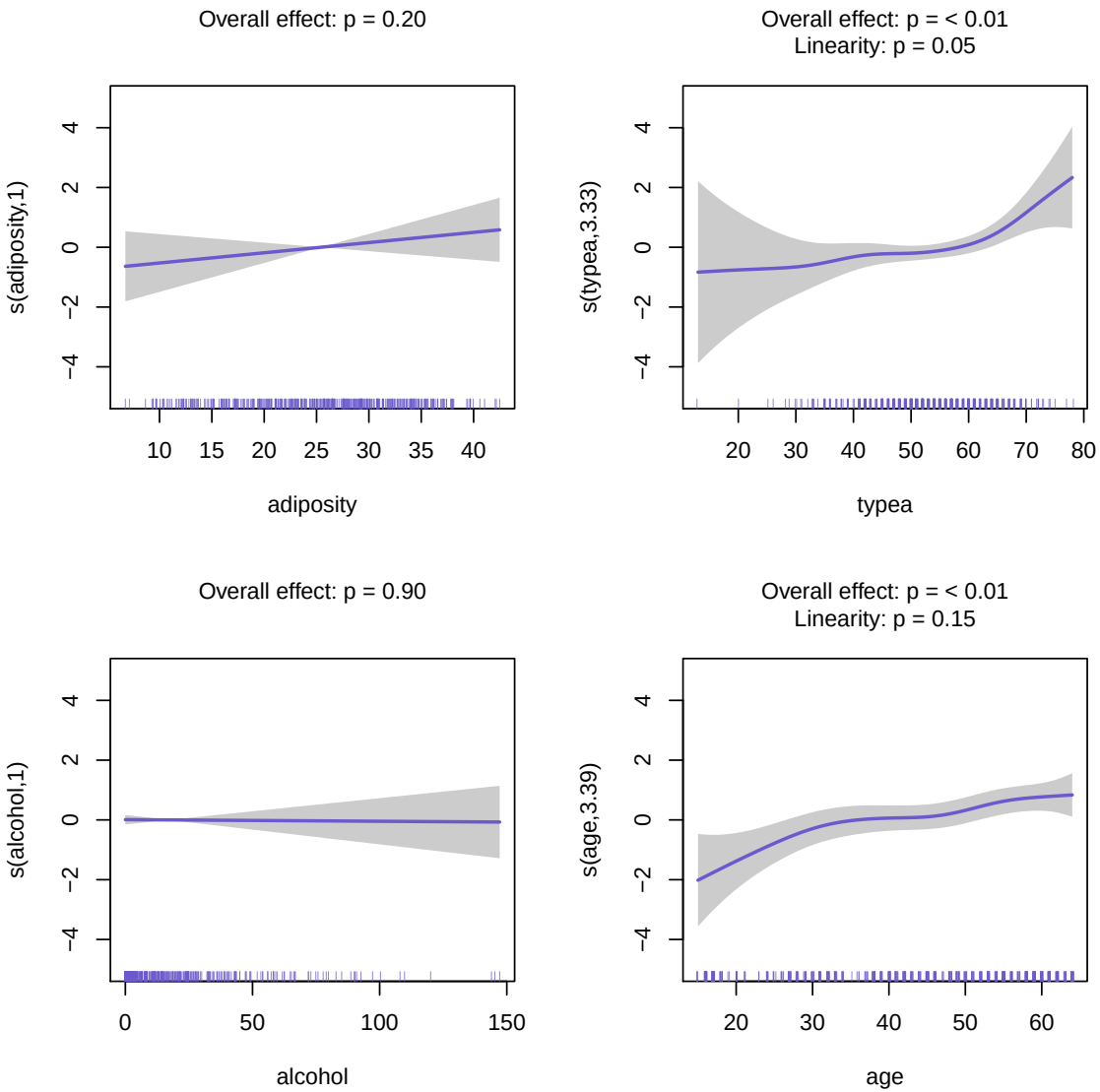

Overall effect: p = 0.21
Linearity: p = 0.12
s(sbp,1.24)
sbp


Overall effect: p = < 0.001
Linearity: p = 0.02
s(tobacco,5.87)
tobacco


Overall effect: p = < 0.01
s(ldl,1)
ldl


Overall effect: p = 0.02
Linearity: p = 0.03
s(obesity,2.2)
obesity

CHD GAM (cont'd)



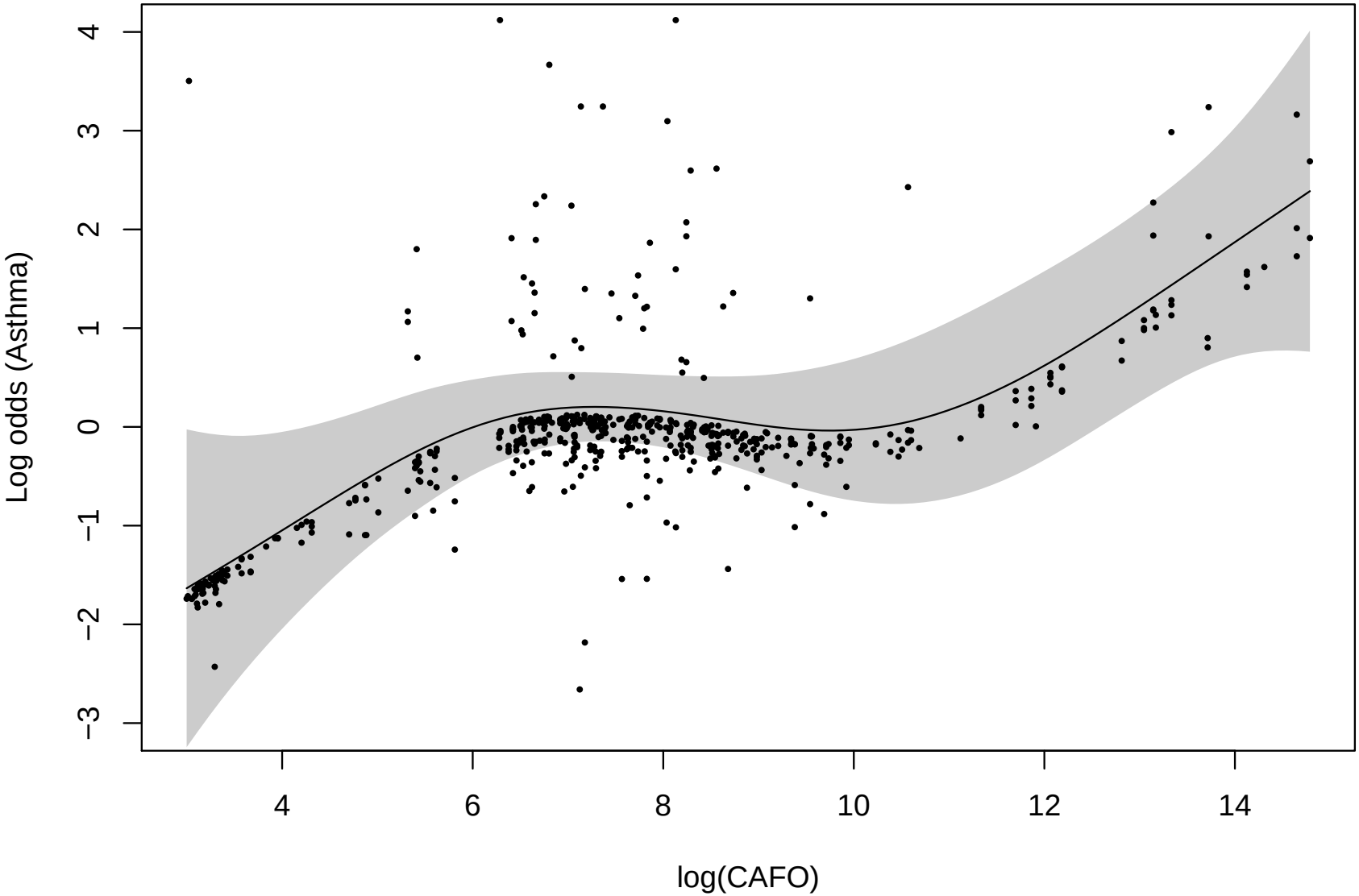
Asthma CAFO study

- As another example, I collaborated with Dr. Sanderson on a project involving exposure to air pollution from concentrated animal feeding operations (CAFOs) and its effect on developing asthma in children
- The exposure metric was calculated based on the proximity of the home to a CAFO (or CAFOs), the size of those CAFOs, and adjusted for wind patterns to develop an overall CAFO exposure score
- Based on previous studies of asthma, we also adjusted for a number of risk factors: age, gender, socioeconomic status, allergies, premature birth, early respiratory infection, and whether the individual worked with pigs

CAFO study: Plan of analysis

- We could once again model each continuous term using smoothing splines, although since the focus of the study was on a single risk factor, I modeled CAFO exposure nonparametrically and included the others as simple linear terms
- The estimation of how CAFO exposure affects asthma risk was the purpose of the study; an interesting additional question is whether or not any interactions are present involving this exposure
- In the course of the study, we found some evidence for an interaction with early respiratory illness (the test for interaction had a p-value of 0.06)

CAFO study: Results

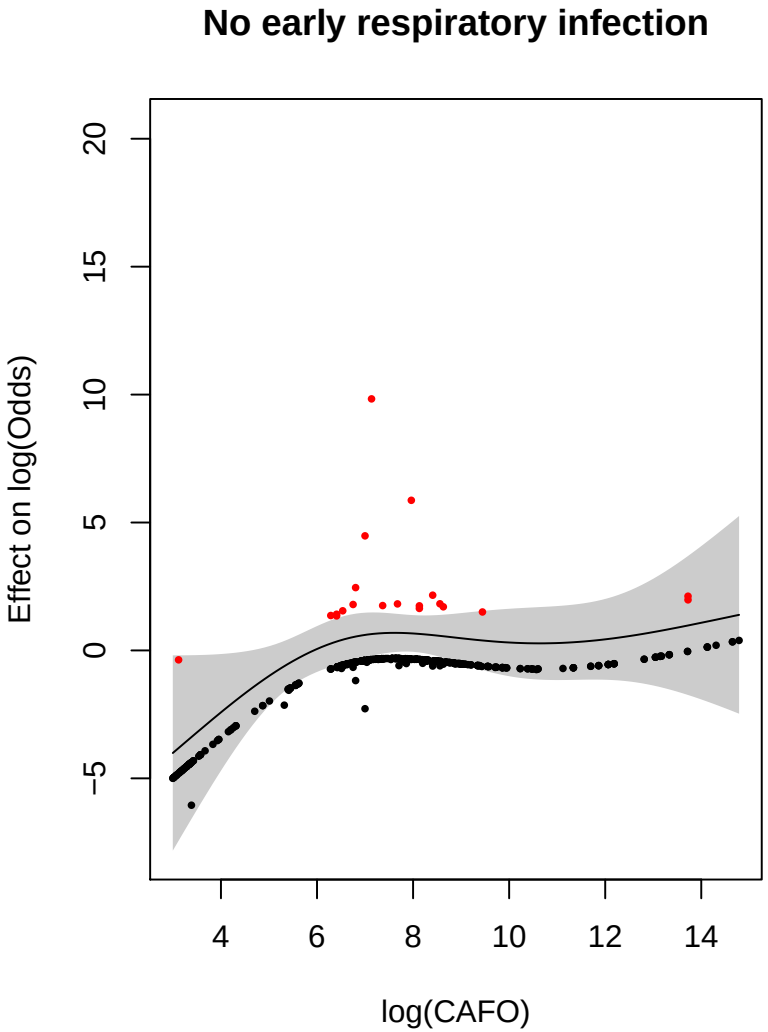


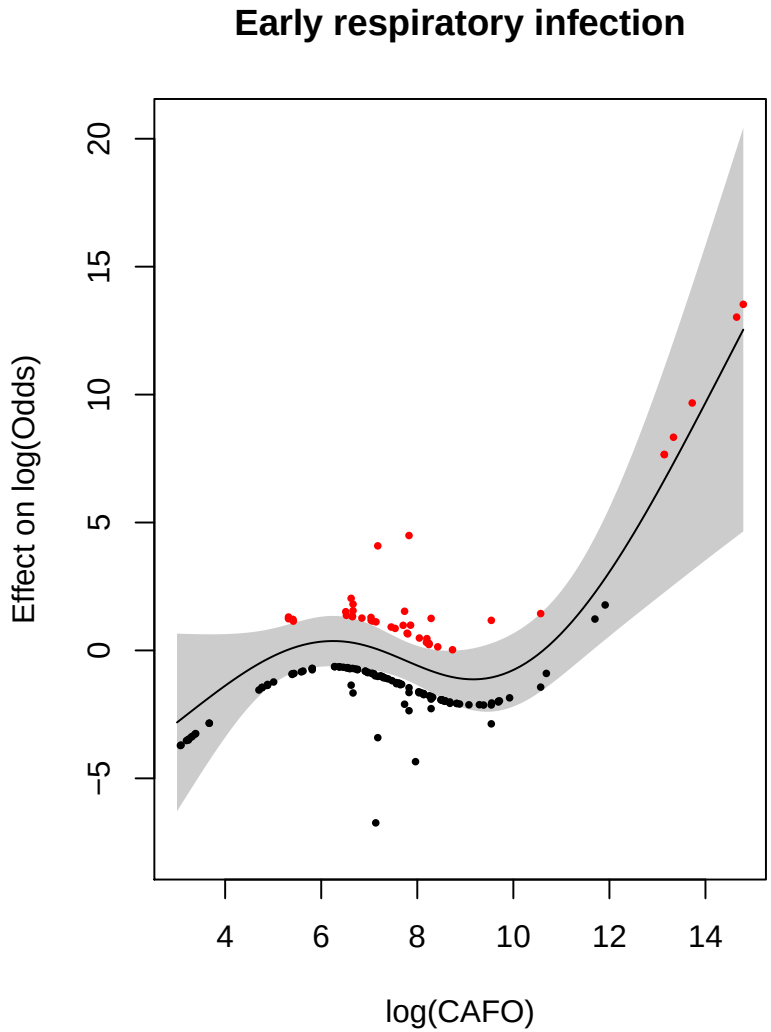
CAFO study: Results (cont'd)

Model	Deviance explained
Base	28.8%
Linear	31.1%
Spline	33.0%

p-value for overall effect of CAFO: .001
p-value for nonlinear component of CAFO: .03

CAFO study: Possible interaction





GAMMs

- I am glossing over a complication present in the CAFO study: the observations were not all independent
- Specifically, sampling was done at the household/family level and one would expect some correlation among subjects within a household
- Possibilities for dealing with this are to include a random intercept (a Generalized Additive Mixed Model, or GAMM), or to use generalized estimating equations

Multidimensional splines

- GAMs are wonderful tools, but they do rely on that assumption of additivity; what if effects are not additive?
- To put it another way, suppose we have x_1 and x_2, both continuous variables; can we estimate

$$\mathrm{E}(y|\mathbf{x}) = f(x_1, x_2)$$

in a completely nonparametric way, without making any assumptions on f?

Tensor product splines

- One approach to constructing multidimensional splines is using the *tensor product basis*
- Suppose we specify a set of basis functions $\{h_{1k}\}$ for x_1 and $\{h_{2k}\}$ for x_2, with M_1 and M_2 elements, respectively
- The tensor product basis for the two-dimensional smooth function of x_1 and x_2 is given by

$$g_{jk}(x_1, x_2) = h_{1j}(x_1)h_{2k}(x_2)$$

and has $M_1 \times M_2$ elements

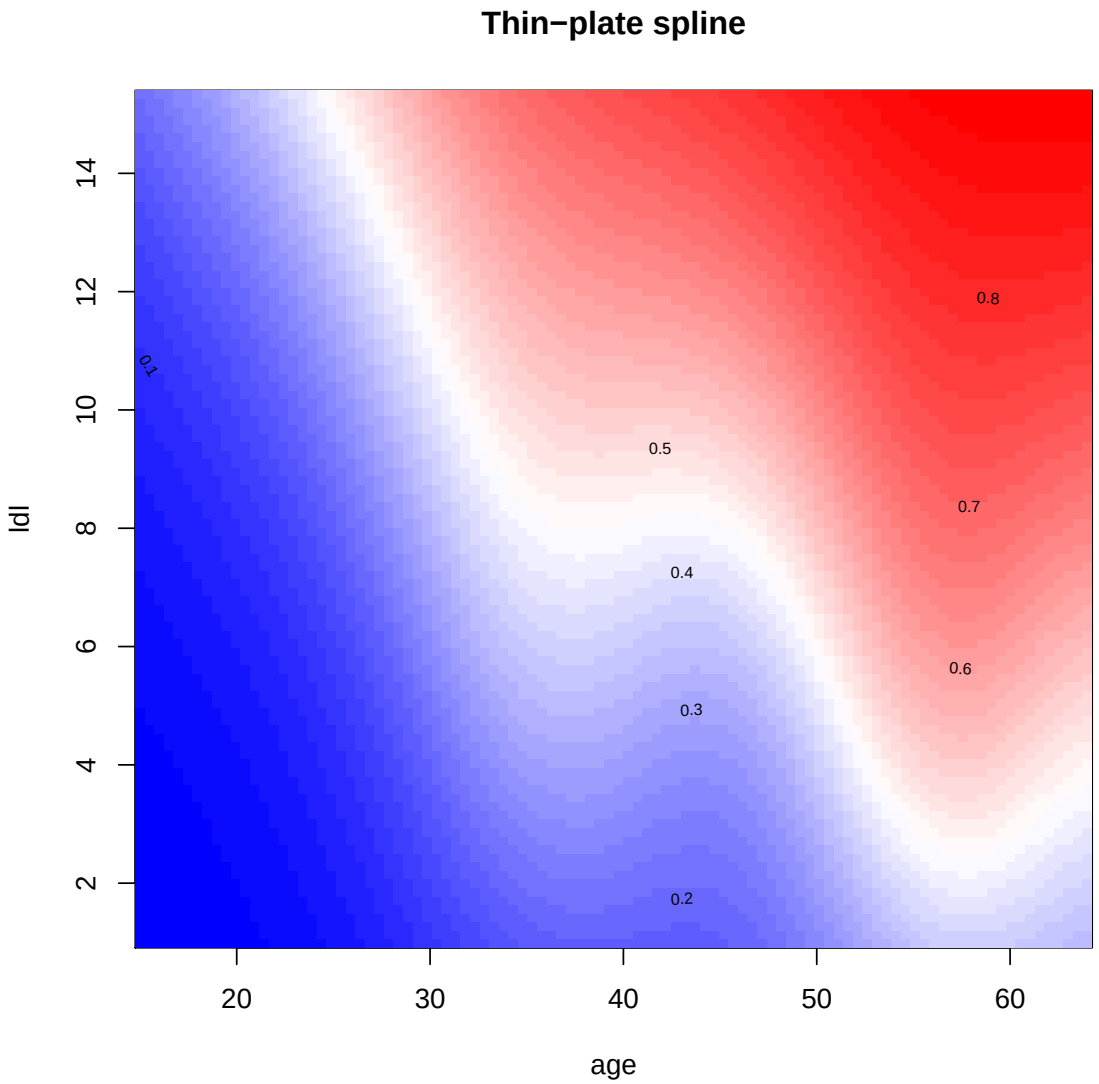
Thin plate splines

- The multidimensional analog of smoothing splines are called *thin plate regression splines*
- For two dimensions, we find $f(x_1, x_2)$ that minimizes

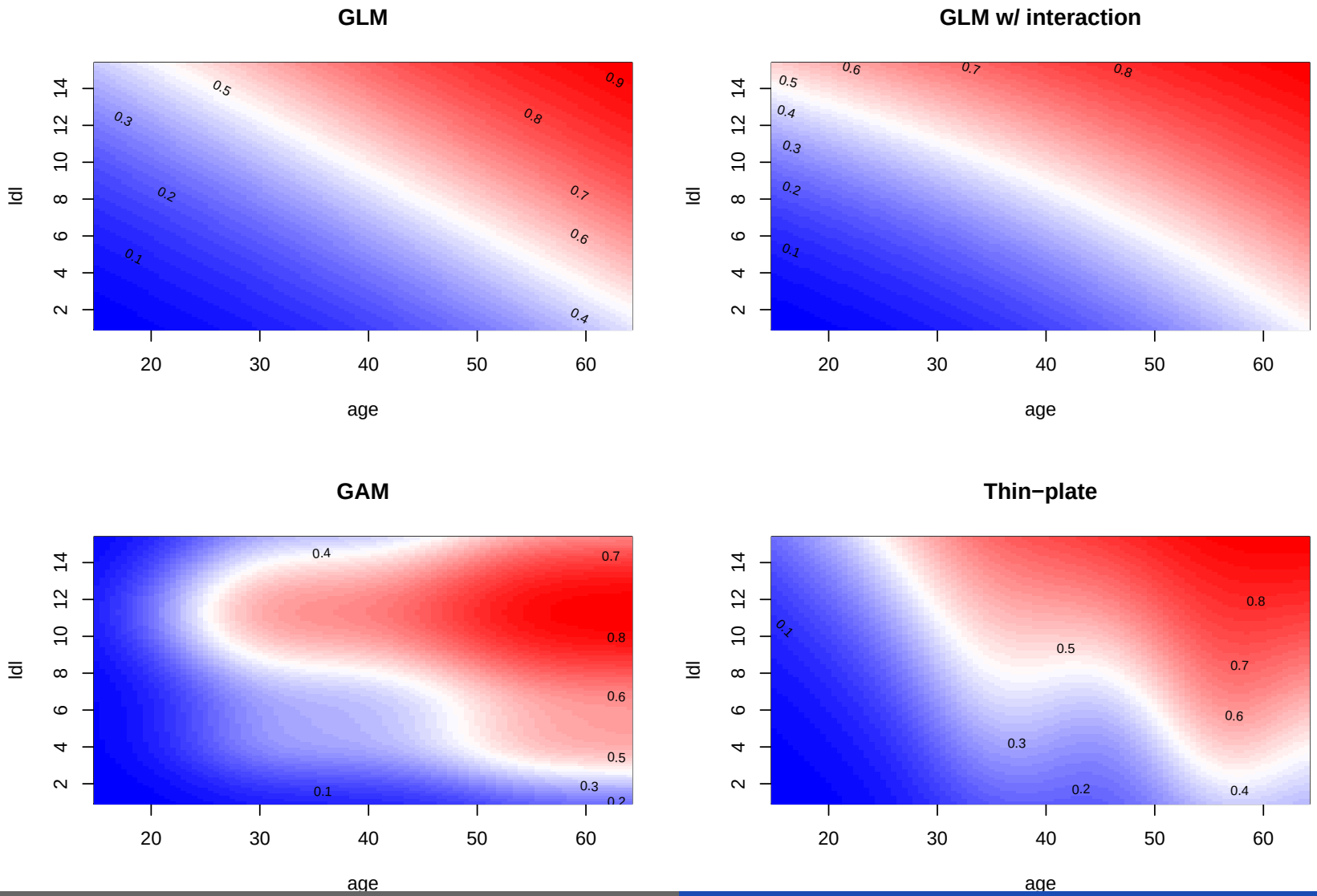
$$-\sum_{i=1}^{n} \ell\{y_i, f(x_{1i}, x_{2i})\} + \lambda \int\int \left[\frac{\partial^2 f}{\partial u^2}\right]^2 + 2\left[\frac{\partial^2 f}{\partial u \partial v}\right]^2 + \left[\frac{\partial^2 f}{\partial v^2}\right]^2 du dv$$

- Thin plate regression splines have fairly complicated basis functions

2D smoothing



Restrictions imposed by various models



The curse of dimensionality

- Thin plate regression splines can be extended further into higher dimensions, but they become rather computationally intensive as the dimension exceeds 2
- Also, the curse of dimensionality implies that we need an exponentially increasing amount of data to maintain accuracy as p increases
- Furthermore, multidimensional smooth functions are harder to visualize and interpret
- In other words, it is nice to have the option of including interactions, but one generally cannot stray too far from the structure provided by the additive model

WISEWOMAN

- Our final case study involves WISEWOMAN, a project initiated by the CDC offering early screenings of risk factors for heart disease, diabetes, and other chronic diseases in financially disadvantaged women
- In Iowa, the project contained an intervention component designed to effect lifestyle changes in at-risk individuals

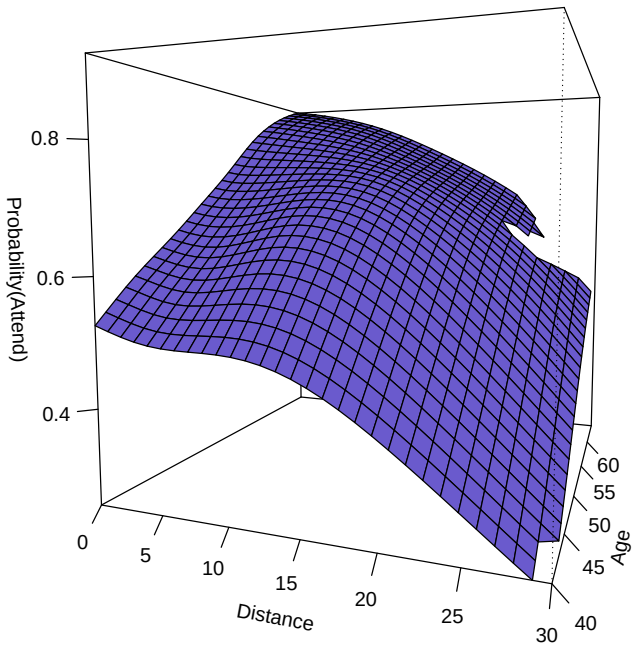
Attendance

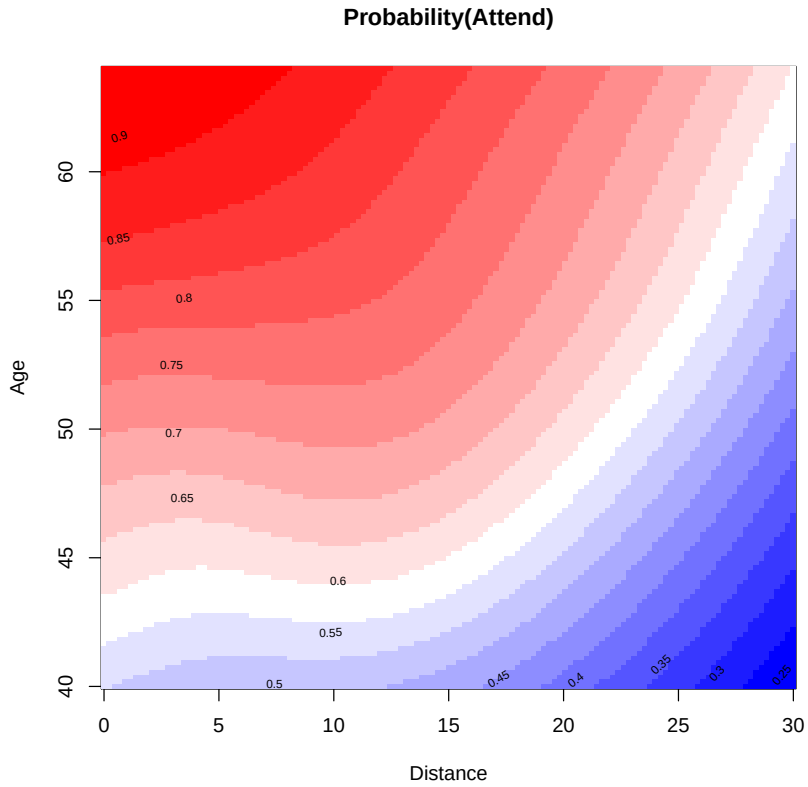
- However, no intervention can be successful if the individuals targeted do not attend
- In order to design better interventions, it is of interest to study factors associated with attendance vs. nonattendance
- In particular, investigators are interested in the role that travel distance plays in attendance

WISEWOMAN study: Analysis

- As with the CAFO study, the emphasis was on a particular term, so distance was modeled using a smoothing spline while other factors affecting attendance were controlled for using linear terms
- In this study, we found significant evidence for an interaction between travel distance and age
- Furthermore, we found evidence of a three-way interaction, in that the above two-dimensional smooth function differed in urban and rural settings

WISEWOMAN: Rural results





WISEWOMAN: Urban vs. rural (cross-sections)