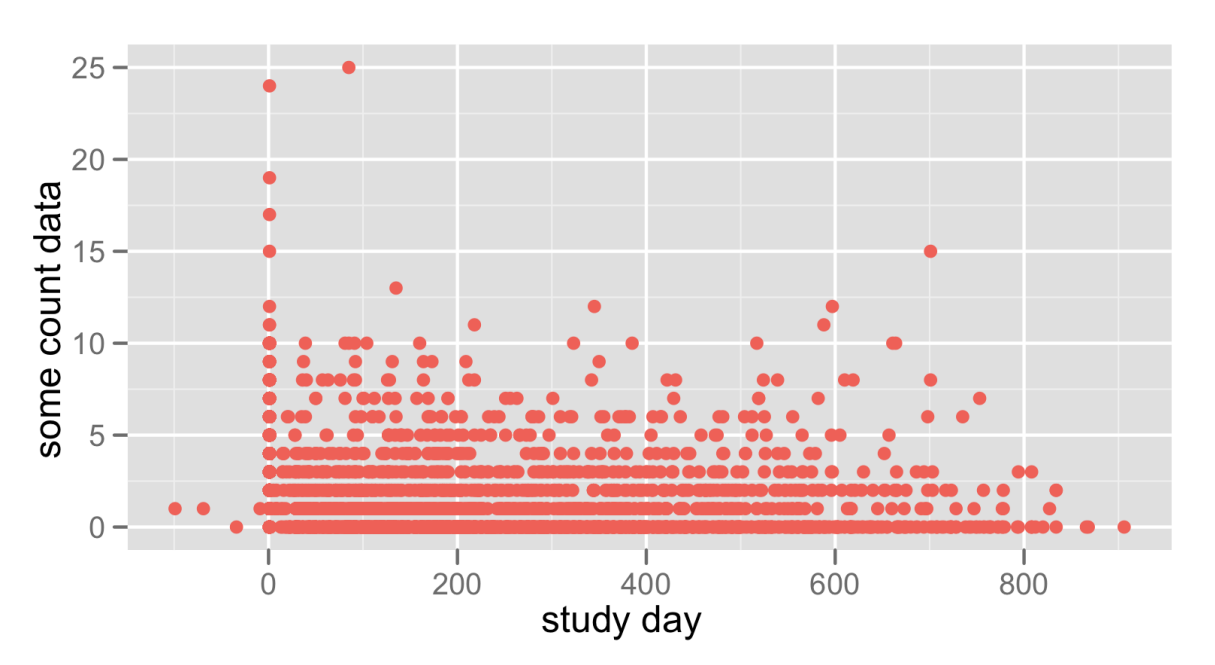
A Tale of Two GAMs

Generalized additive models as a tool for data exploration

Mariah Silkey, Actelion Pharmacueticals Ltd.

Some observational data

- Progressive disease
- Lots of missing data



- Time between visits, or whether visits take place at all is uncontrolled
- Patients may be on the study drug, or not, or on a competing treatment
- Patients can switch treatments at will

Linear Regression

$$Y_i = \alpha + \beta_1 \times X_i + \beta_2 \times X_i^2 + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

Ordinary least squares used to fit parameters $\alpha + \beta_i$

General additive models

Additive models fit smooth functions through some terms, so rather than optimizing β_i we optimize smoothing functions f_1

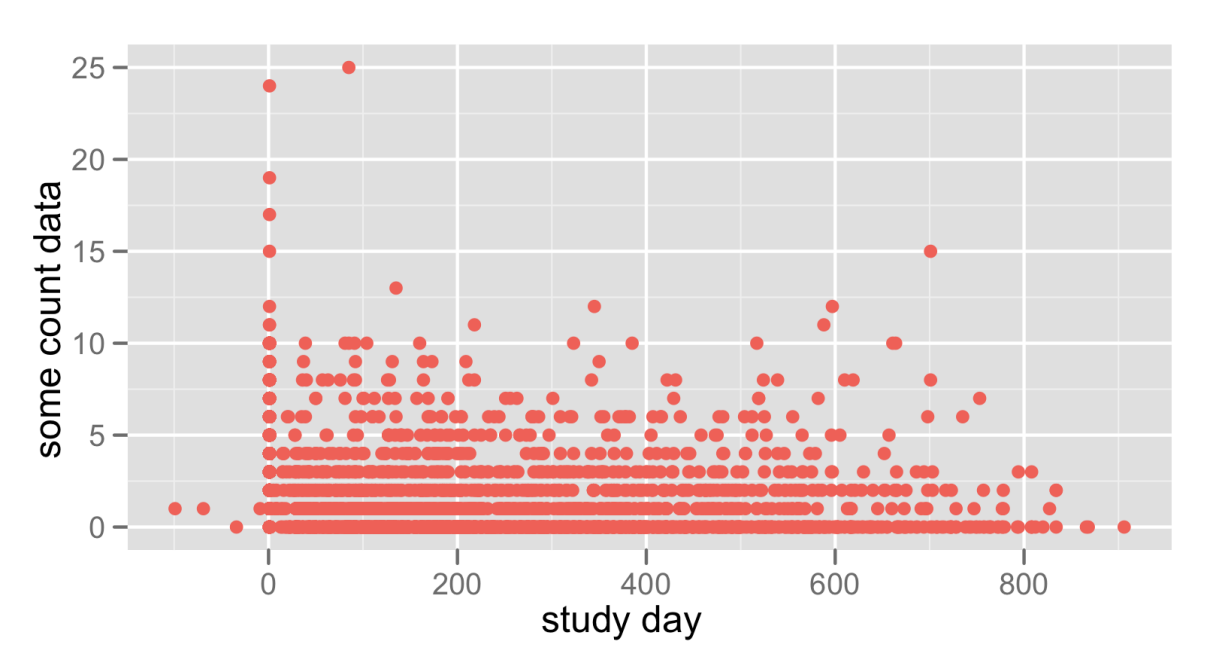
$$Y_i = \alpha + f_1(X_i) + f(X_i^2) + \varepsilon_i \qquad \text{where } \varepsilon_i \sim N(0, \sigma^2)$$

optimize

- span for loess
- # and location of knots for a regression spline
- # and location of knots, and degree of polynomial for quadratic and cubic regression splines
- Λ for penalized splines

Some observational data

- Progressive disease
- Lots of missing data



- Time between visits, or whether visits take place at all is uncontrolled
- Patients may be on the study drug, or not, or on a competing treatment
- Patients can switch treatments at will

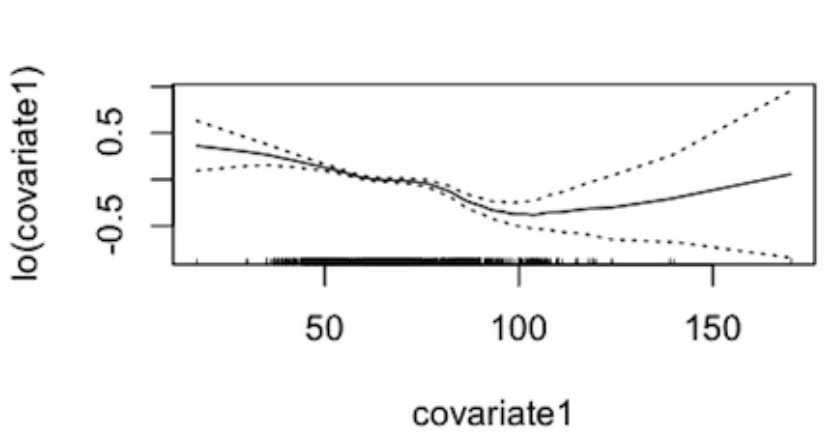
R package gam

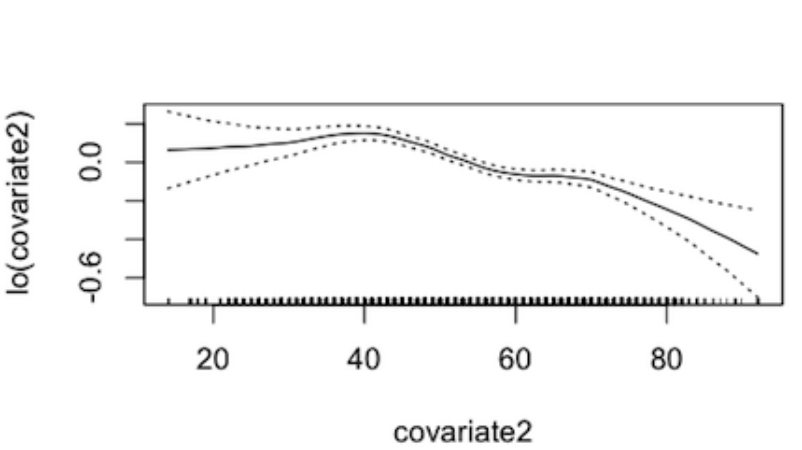
gam written by Trevor Hastie and Robert Tibshirani
uses loess smoothers or smoothing splines

Optimization by AIC, or visual inspection of residuals

gam alá gam

Call: gam(formula = counts ~ studyday + lo(covariate2) + lo(covariate1), family = poisson, data = x5)
Deviance Residuals:
    Min   1Q Median   3Q   Max
-2.1366 -1.5843 -0.6925 0.4581 9.7668
AIC: 17494.9
DF for Terms and Chi-squares for Nonparametric Effects
                 Df Npar Df Npar Chisq   P(Chi)
lo(covariate2) 1   2.2   27.117 1.819e-06
lo(covariate1) 1   3.0   18.284 0.000387





R for GAM - mgcv

mgcv written by Simon Wood

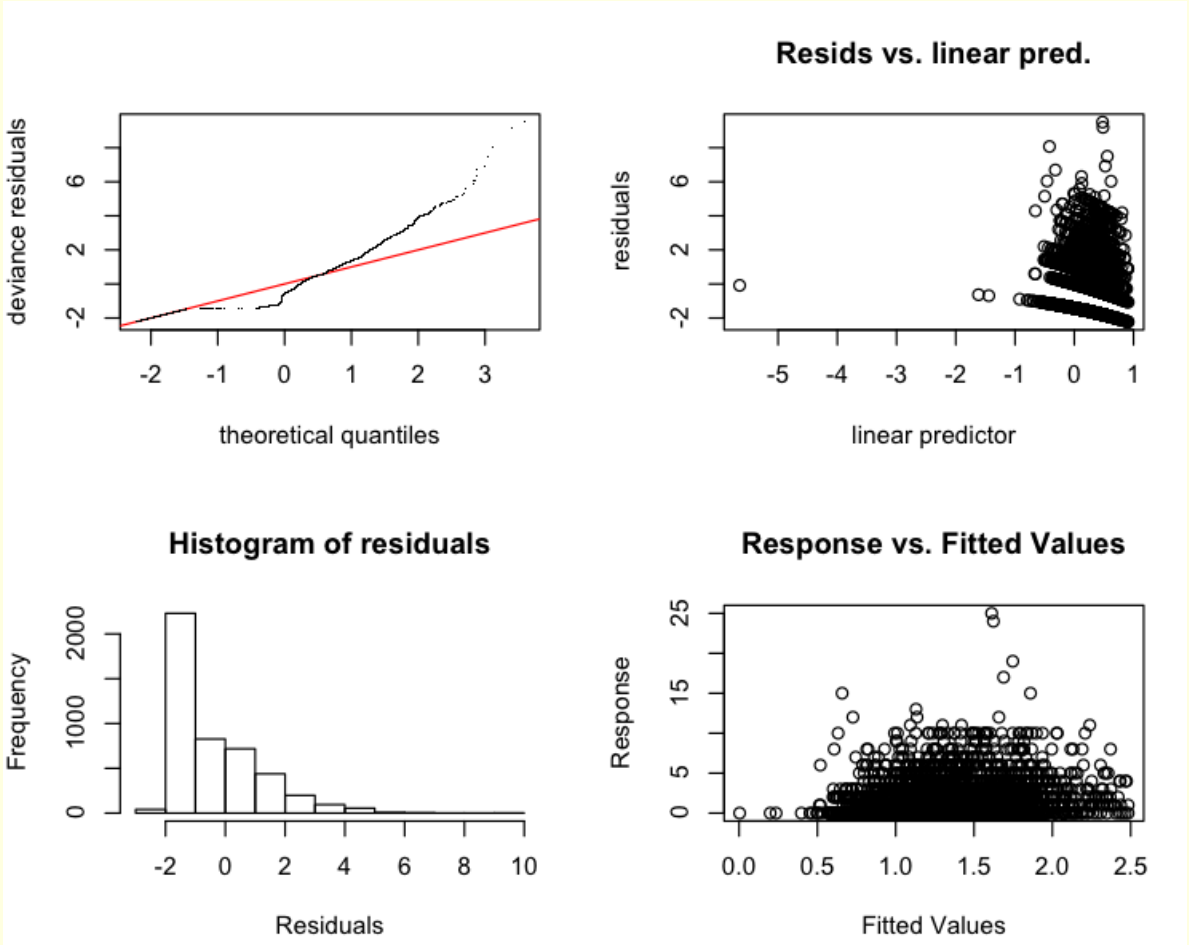
Uses penalized regression spline methods, using lots of knots and minimizing the penalized sum of squares equation below

$$\| Y - X\beta \|^2 + \lambda \int f''(x)^2 \, dx$$

Either for a given λ , or uses cross validation to choose the optimal penalty.

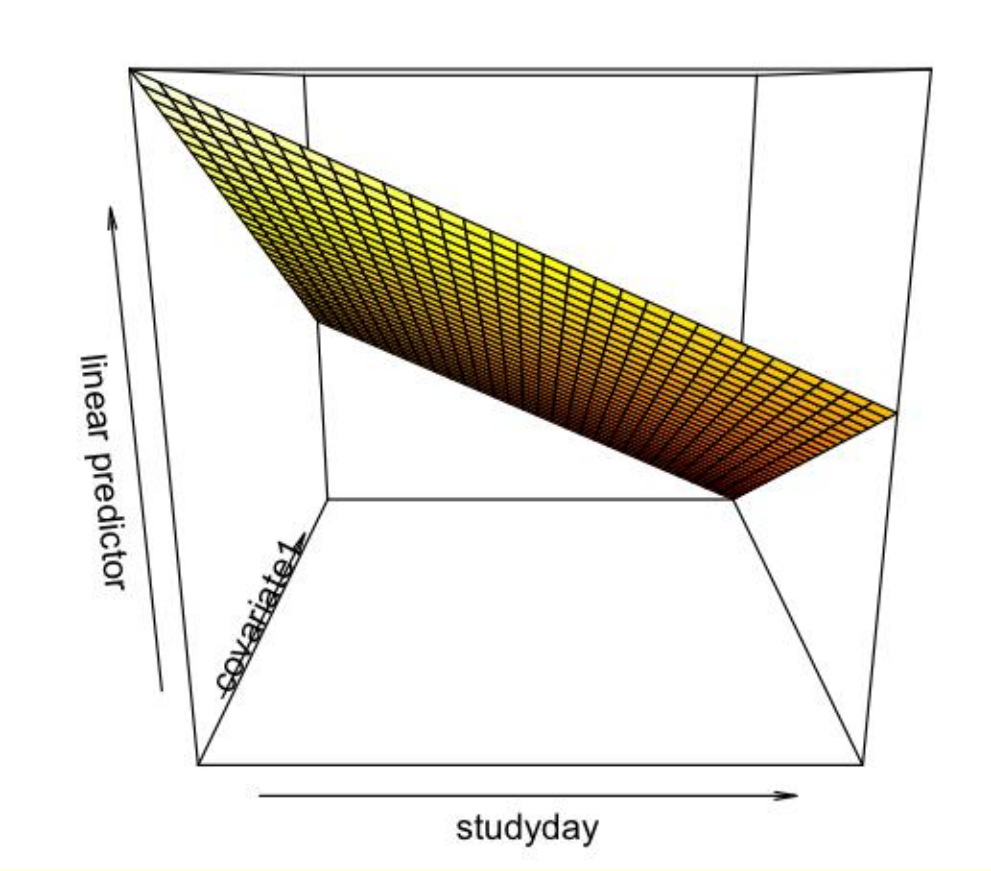
gam alá mgcv

b <- gam(counts ~ study day + s(covariate1)+s(covariate2),
data=dat, family = poisson)



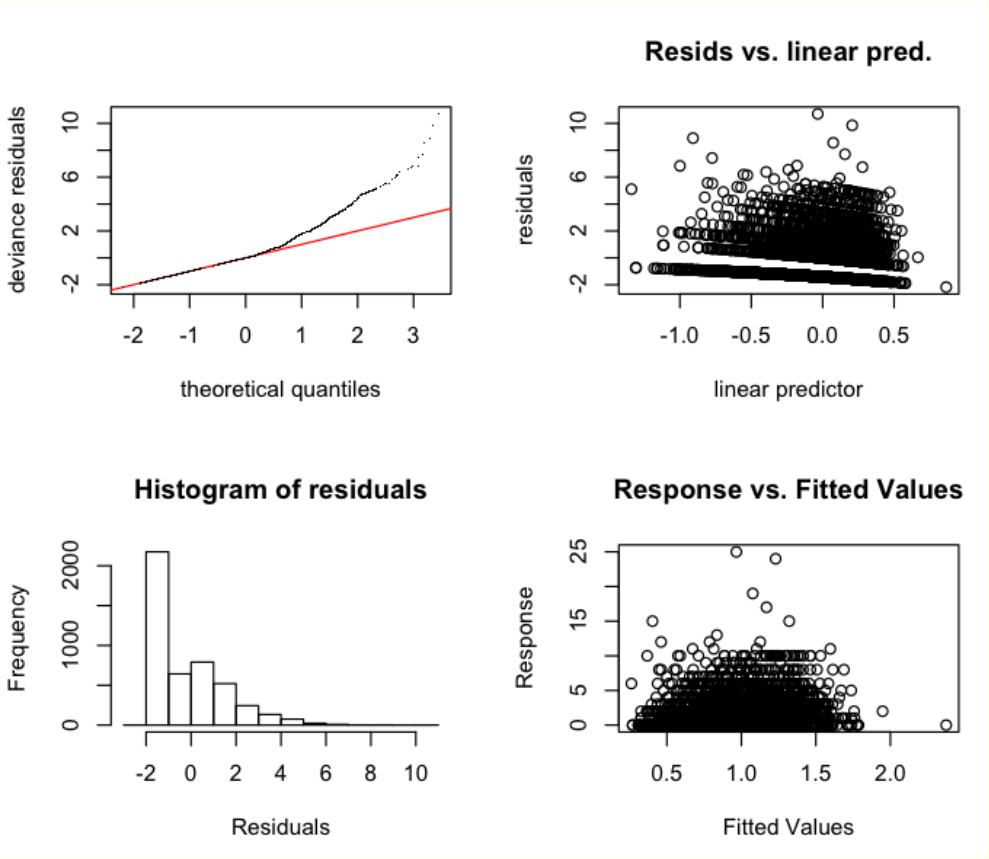
gam alá mgcv

b <- gam(counts ~ study day + s(covariate1)+s(covariate2), data=dat, family = poisso



gam alá mgcv

```
b2 <- gamm(counts ~ studyday + s(covariate1)+s(covariate2),
        random =list (subject =~1), data=dat, family = poisson)
```



gam and mgcv

Readers familiar with the classical textbook from Hastie and Tibshirani may prefer the **gam** package as it follows the theory described in the book.

The gam in **mgcv**: allows for a wider variety of covariance structures, including repeated measures, and spatial and temporal correlations. It's the bee's knees, in short.

References

- *Mixed Effects Models and Extensions in Ecology with R*. 2007. Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, Graham M. Smith

- *Generalized Additive Models: An Introduction with R*, 2006 Simon N. Wood

- *The Elements of Statistical Learning: Data Mining, Inference and Prediction*,2009 Trevor Hastie, Robert Tibshirani, and Jerome Friedman

- *R: A language and environment for statistical computing*. R Development Core Team (2009). R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

- Wood, S.N. (2008) Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society (B) 70(3):495-518 ;more information at http://people.bath.ac.uk/sw283/mgcv.

- Packages **mgcv** and gam can be downloaded http://stat.ethz.ch/CRAN.