Statistics for the Sciences

Generalized Additive Models

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Generalized Additive Models (GAM)
- Generalized Additive Mixed Models (GAMM)
- Example

Generalized Additive Models

- Recall that we discussed `Local linear regression` for flexibly predicting a response Y on the basis of a single predictor X.
  - The method is nonparametric.
- Recall linear models

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

- Generalized additive models (GAMs) provide a general framework for extending a linear model by **allowing non-linear functions** of each of the predictors, while maintaining **additivity**.

Generalized Additive Models

- MLR model: $E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, i = 1, \ldots, n$
- A GAM fits a more flexible model:

$$E(y_i) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}), i = 1, \ldots, n$$

where the f_js are (nonparametric) smooth nonlinear functions, such as Loess smoothers and smoothing splines.

Generalized Additive Models

- Generalized Linear Models (GLMs) fit models constrained to a parametric (linear) form.
- Generalized Additive Models (GAMs) fit a broader range of non-parametric models determined from the observed data.
- Additivity in GAMs: The response variable is modeled as the sum of functions of each predictor.
- GAMs can be analyzed within the same framework as linear and generalized linear models using goodness-of-fit measurements deviance and AIC.

Generalized Additive Models

- Advantages:
  - GAMs allow us to fit a non-linear f_j to each X_j , so that we can automatically model non-linear relationships that standard linear regression will miss.
  - Because the model is additive, we can examine the effect of each X_j on Y individually while holding all of the other variables fixed.
- The main limitation of GAMs is that the model is restricted to be additive. With many variables, important interactions can be missed.
  - However, we can manually add interaction terms to the GAM model.

Generalized Additive Mixed Models

- Recall Linear Mixed Models

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}b + \varepsilon$$

  where $\boldsymbol{b}$ and ε are independent normal vectors.

- A Generalized Additive Mixed Model (GAMM) just generalizes the Linear Mixed Models above by replacing the linear predictor $\boldsymbol{X}\beta$ with smooth nonlinear functions of the predictors $\boldsymbol{X}$.

$$Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \boldsymbol{Z}_i b$$

  In matrix notation,

$$\boldsymbol{Y} = \boldsymbol{f}(X\beta) + \boldsymbol{Z}b + \varepsilon$$

  , where $\boldsymbol{f}$ is the vector of smooth nonlinear functions.

Example

- Both GAM and GAMMS can be applied to Generalized Linear Models, that is when the responses are not normally distributed. For example, the GAMs for binary responses with logit link

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p)$$

- Unfortunately, SPSS cannot do GAM or GAMMs. We just check the results produced by R.
- Go to https://mjkeough.github.io/examples/cabanellas.nb.html check the complete analysis.

Example

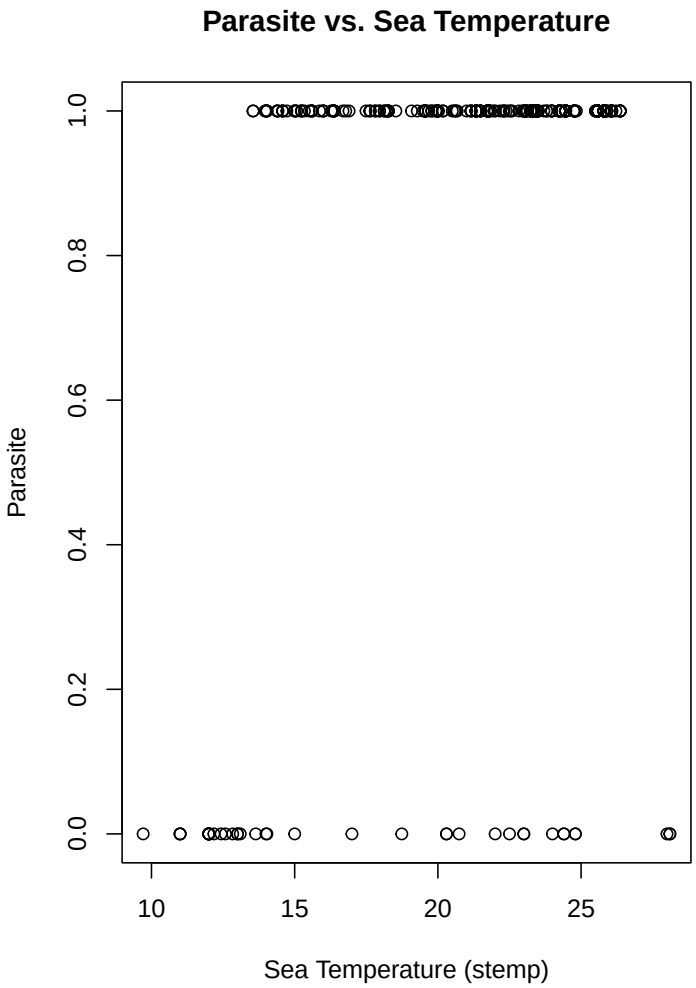
- Example (`cabanellas.csv`): Cabanellas-Reboredo et al. (2019) studied the spread of a disease in a large bivalve (Pinna nobilis) in the Mediterranean caused by a protozoan endoparasite. They collated observations of dead or unwell bivalves from many sites using information from scientific surveys and citizen science contributions. They only used observations from sites that their dispersal models indicated the disease could have spread to. They focused on relating the presence of the disease at a site to salinity and temperature.
  - Binary response `parasite`
  - Predictors: `stemp` and `ssal`

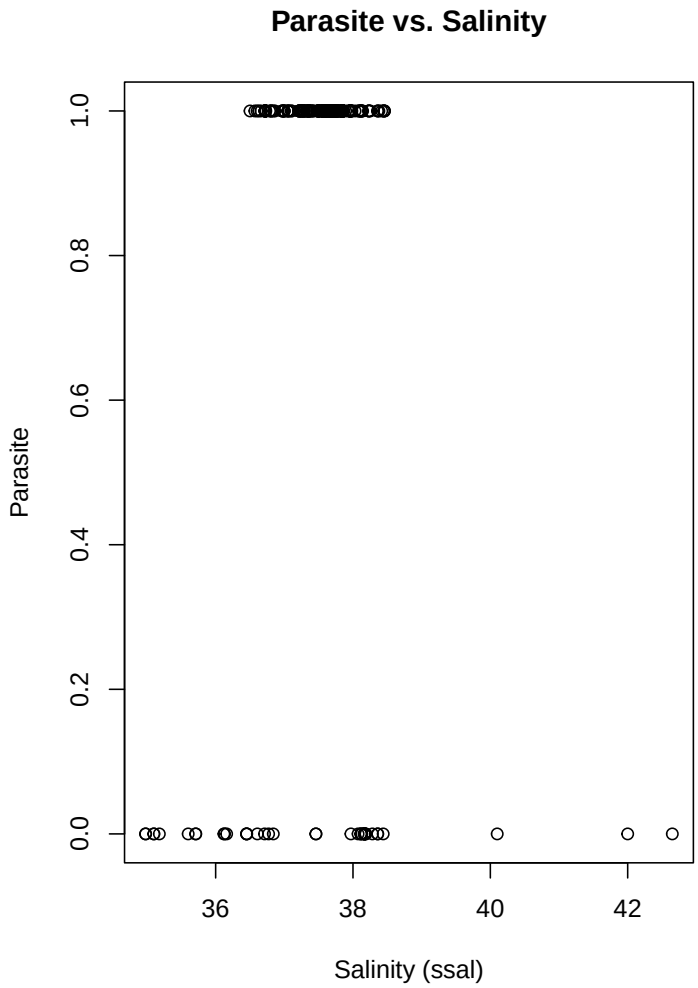
Example

- First 5 rows of data

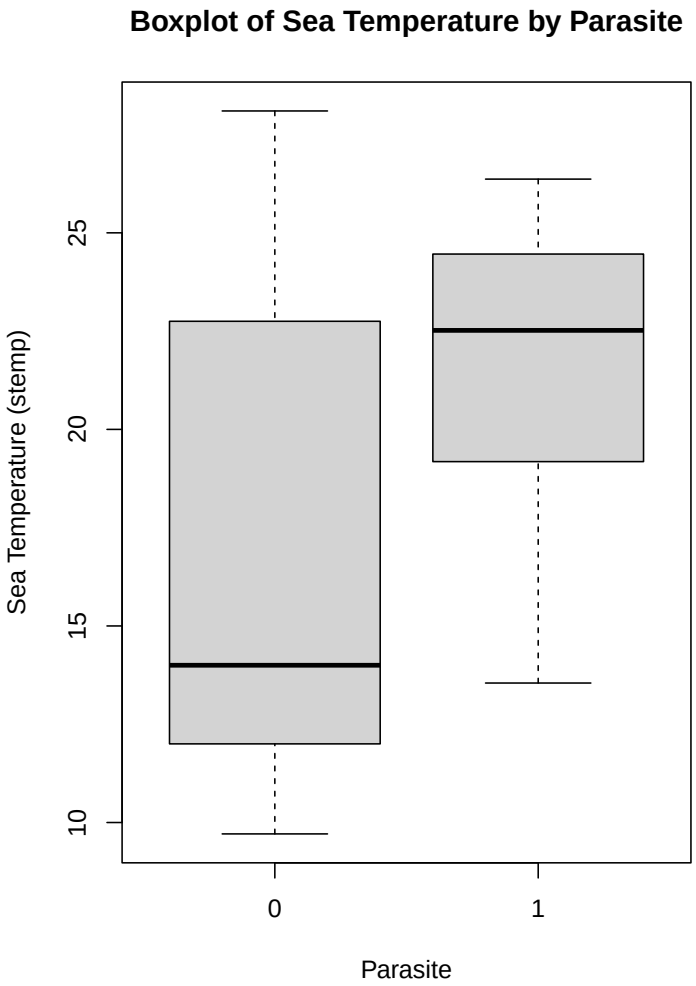
```
##      id day month year     lat   long depth     habitat parasite syear sm
## 1 7258  30    11 2016 39.1266 2.9403     8 Phanerogams         1  2016
## 2 7259   1    12 2016 39.1504 2.9438    10 Phanerogams         1  2016
## 3 7283  31     3 2017 39.1538 2.9441    20 Phanerogams         1  2017
## 4 7414   5     4 2017 39.1500 2.9441     9       Rocky         1  2017
## 5 7643  28     5 2017 39.8220 4.2230    12       Mixed         1  2017
##   sday slong slat   sdepth    stemp     ssal
## 1   15     3   39  5.02159 20.61909 37.52667
## 2   15     3   39  5.02159 17.96152 37.39895
## 3   15     3   39 15.07854 15.23615 37.51138
## 4   15     3   39  5.02159 16.78289 37.35170
## 5   15     4   40 15.07854 17.64399 37.78589
```

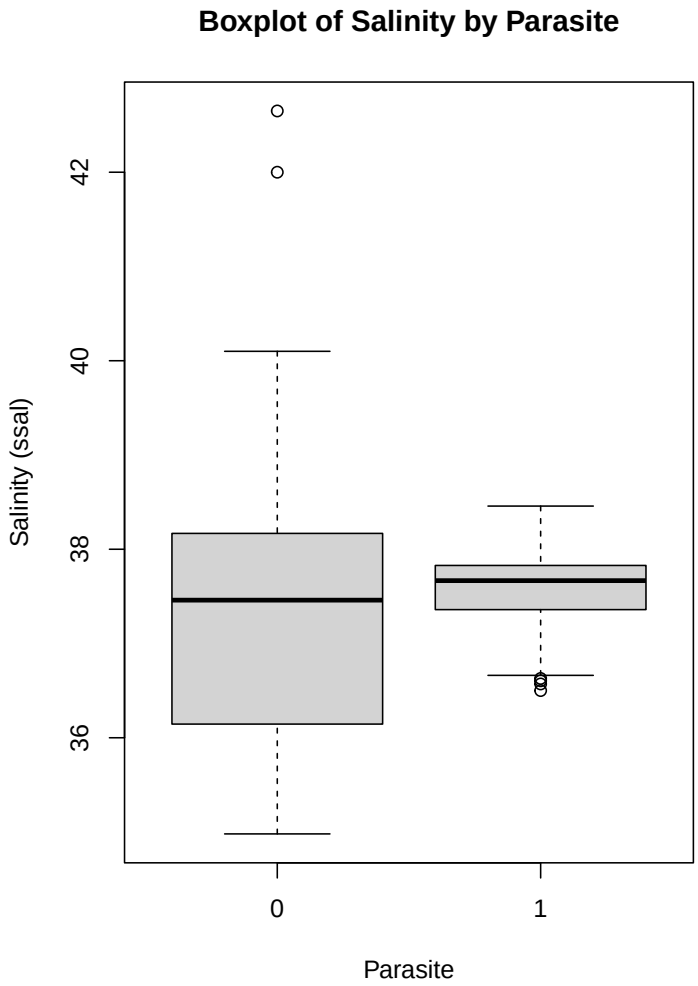
Example


Parasite vs. Sea Temperature
Parasite
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
Sea Temperature (stemp)


Parasite vs. Salinity
Parasite
0.0
0.2
0.4
0.6
0.8
1.0
36
38
40
42
Salinity (ssal)

Example


Boxplot of Sea Temperature by Parasite
Sea Temperature (stemp)
10
15
20
25
0
1
Parasite


Boxplot of Salinity by Parasite
Salinity (ssal)
36
38
40
42
0
1
Parasite

Example

- Fit full glm logistic regression `parasite ~ stemp+ssal`

```
##
## Call:
## glm(formula = parasite ~ stemp + ssal, family = binomial, data = cabanellas)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -33.64169   11.08048  -3.036  0.00240 **
## stemp         0.26706    0.04878   5.475 4.38e-08 ***
## ssal          0.80687    0.28499   2.831  0.00464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 223.30  on 270  degrees of freedom
## Residual deviance: 182.28  on 268  degrees of freedom
## AIC: 188.28
##
## Number of Fisher Scoring iterations: 5
```

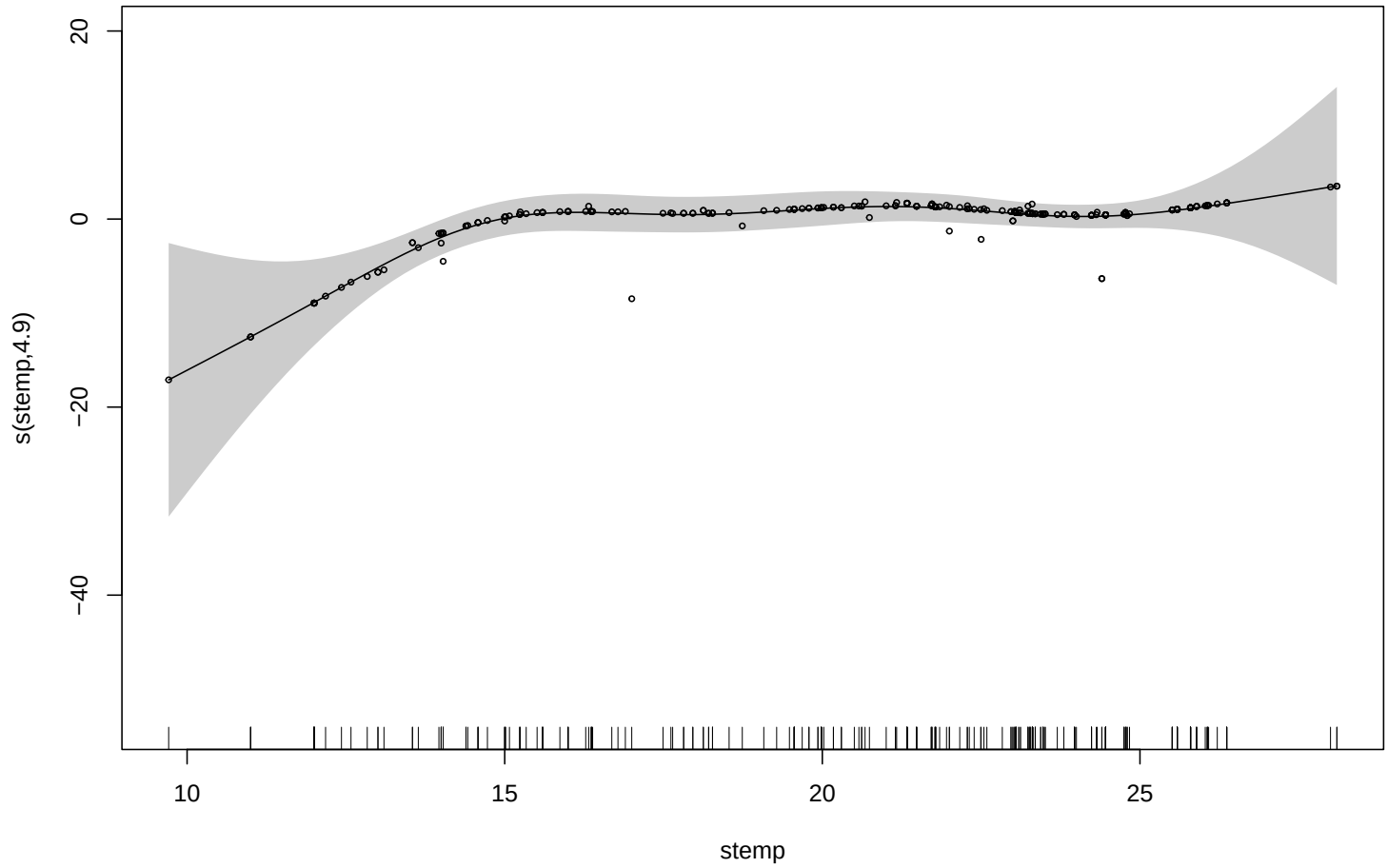
Example

- The R package `mgcv` offers generalized additive modelling functions
- Fit GAM with thin plate regression spline

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## parasite ~ s(stemp) + s(ssal)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.6728     0.5004   5.341 9.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df Chi.sq  p-value
## s(stemp) 4.895  5.877  18.13 0.004770 **
## s(ssal)  2.910  3.588  22.57 0.000757 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.755   Deviance explained = 71.1%
## UBRE = -0.69724  Scale est. = 1         n = 271
```

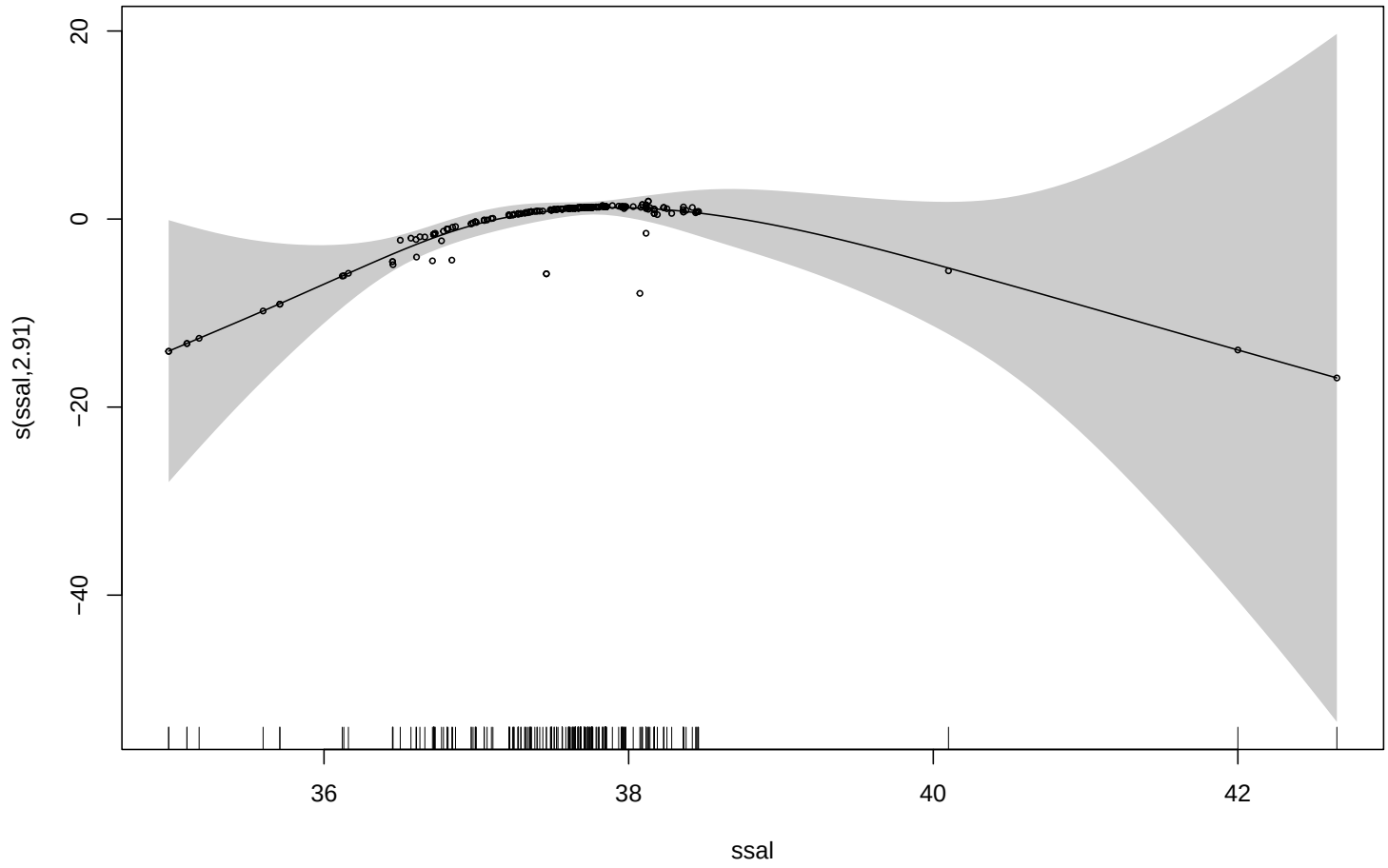
Example

- Diagnostic plots for the fitted GAM model



Example

- Diagnostic plots for the fitted GAM model



License