Generalized Additive Models (GAMs)

Israel Borokini

Advanced Analysis Methods in Natural Resources and Environmental Science (NRES 746)

October 3, 2016

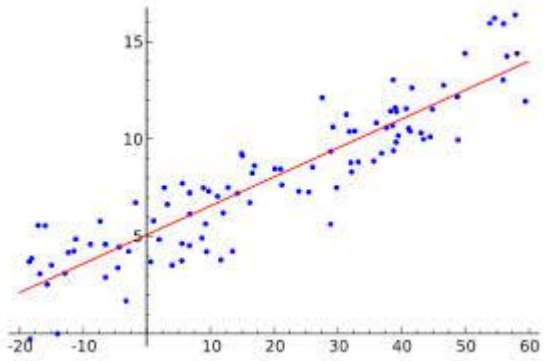
Outline

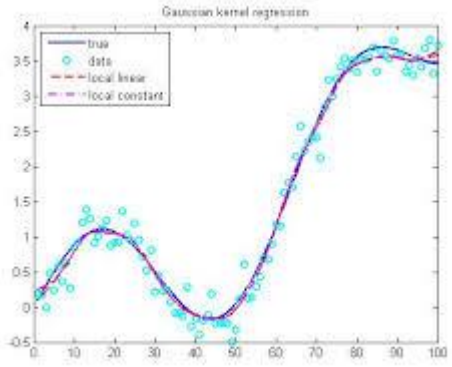
- **Quick refresher on linear regression**
- **Generalized Additive Models**
  - Statistical expression
  - Operations
  - Research Applications
  - R packages for GAMs
  - Examples
  - K selection

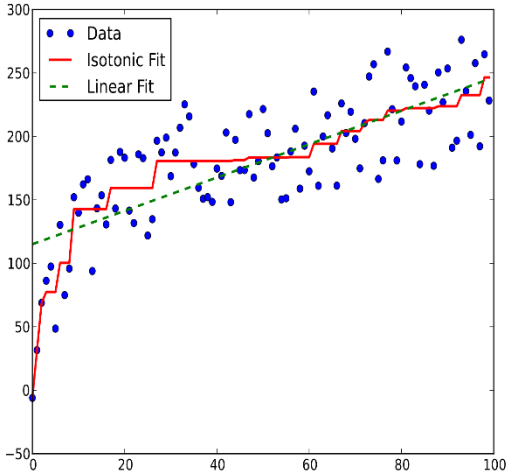
Regression

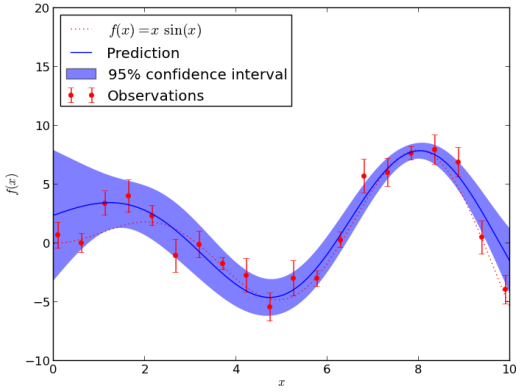
- Regression methods are used to investigate relationships between predictors and response variables

- A good model should perform three functions: description, inference and predictions









Linear Regression Model

- Bivariate regression: $\mathbf{Y = \alpha + \beta X + \varepsilon}$ **where $\varepsilon \sim N(0,\delta^2)$**
- Multivariate regression: $\mathbf{Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon}$
- Quadratic regression: $\mathbf{Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \varepsilon}$
- Polynomial regression: $\mathbf{Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3 + \beta_n X_n^n + \varepsilon}$

- **Y** - response variable **X** - explanatory variable
- **ε** - residual error, to cover unexplained information, assumed to be normally distributed with mean of 0 and δ^2
- **α** and **β** are intercept and slope respectively, to be determined at CI = 95%
- N – sample size

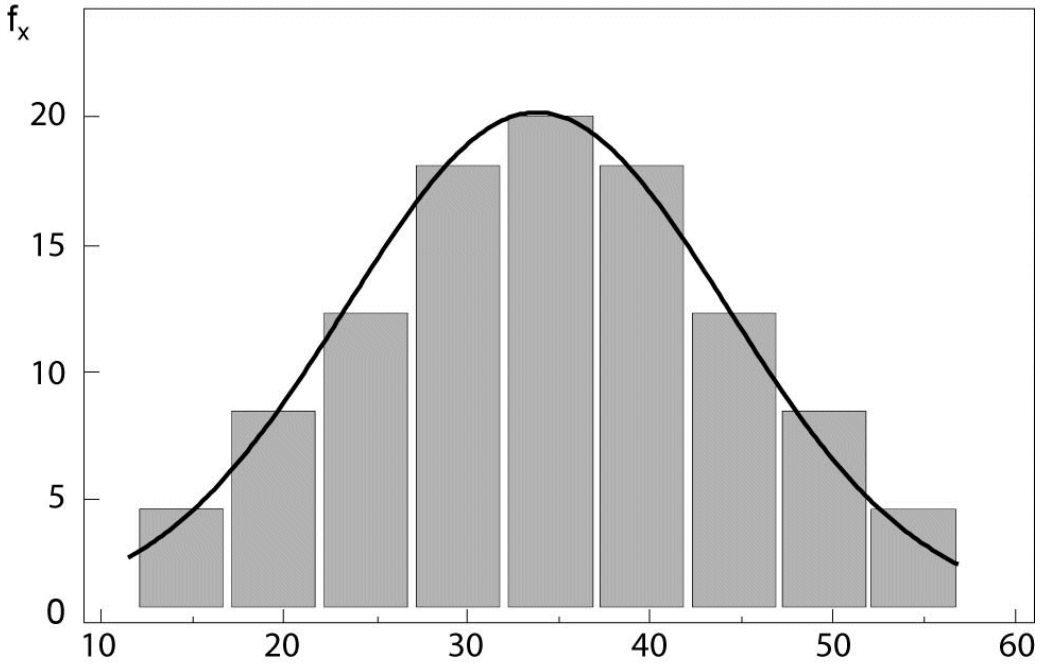
- OLS regression computes values of **α** and **β** that best fit the response by minimizing sum of squared errors (assuming linearity and homoscedasticity)

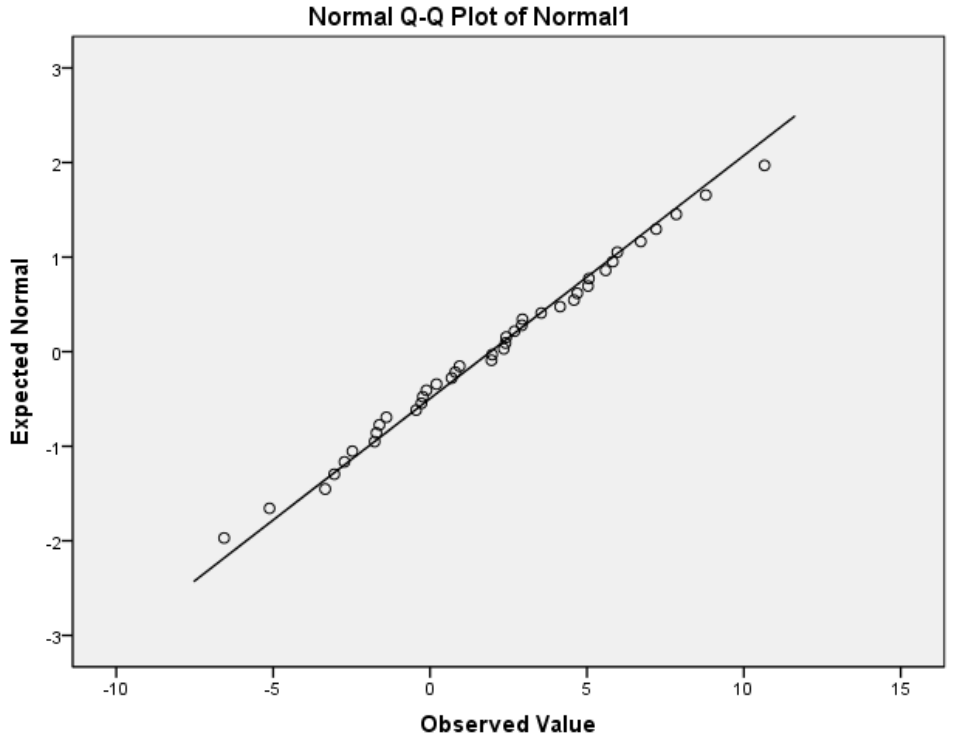
Assumptions of Linear regression models

- Linearity (sensitive to outliers & data inaccuracy)
- Multivariate normality
- Little or no multicollinearity & singularity
- No auto-correlation
- Homoscedasticity
- Prefers large response variable (20:1)

Normality

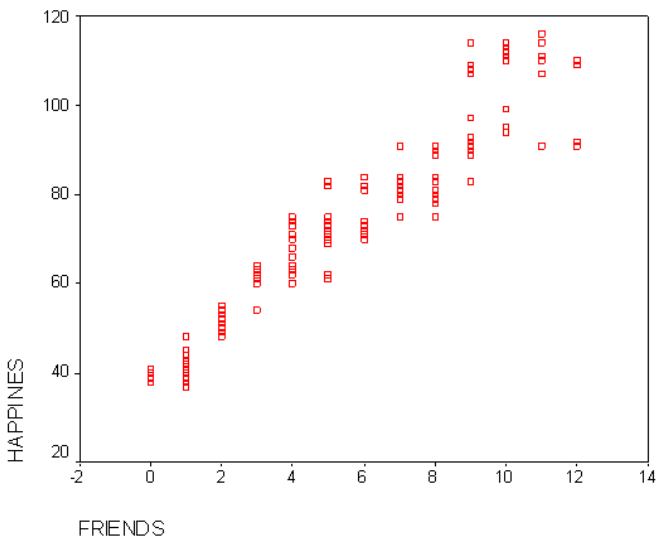
- histogram and fitted normal curve
- QQ plot
- Partial residual plots
- Kolmogorov-Smirnov test (less powerful)
- Shapiro-Wilk test
- Anderson-Darling test

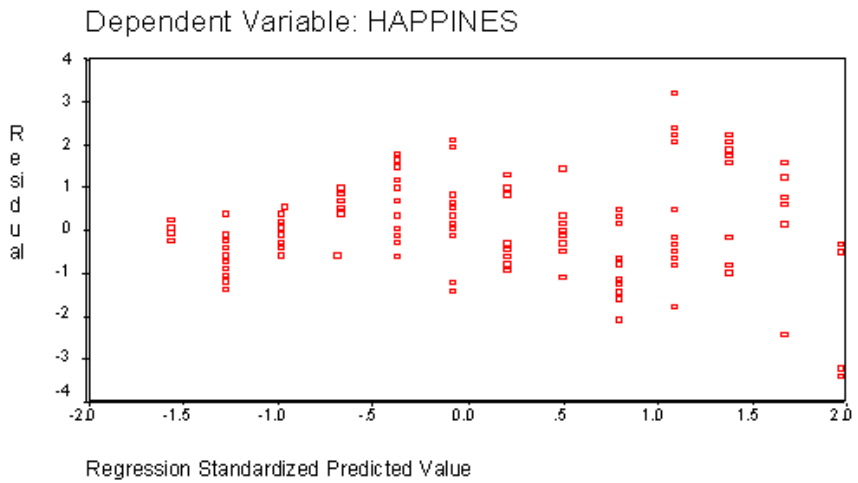


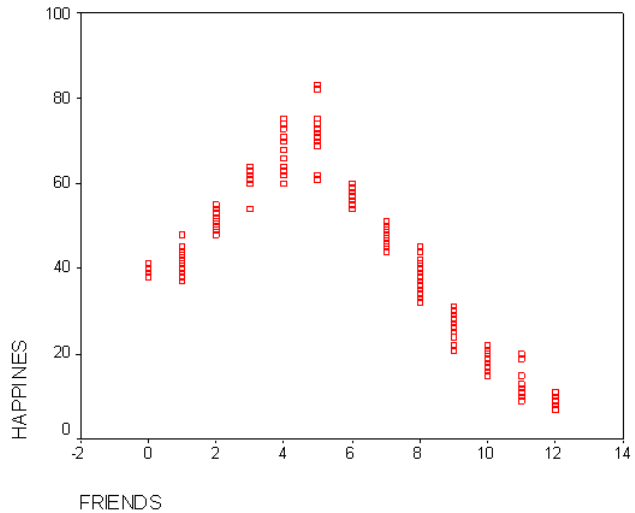


Linearity

- Linear relationship between response and predictors
  - bivariate scatterplots







Multicollinearity and singularity

- Multicollinearity – strong correlations between (or among) predictors
- Singularity – when predictors are perfectly correlated, that is r = 1.0
- Effects: bias predictions
- Solutions: remove some variables or factor analysis
- Detected with the following tests
  - Correlation matrix (correlation values >1 indicates multicollinearity)
  - Tolerance measures: $T = 1 - R^2$ ($T < 0.1$ indicates multicollinearity)
  - Variance inflation factor: $VIF = 1/T$ ($VIF > 100$ indicates multicollinearity)
  - Condition index (values ≥10 indicates multicollinearity)

Autocorrelation

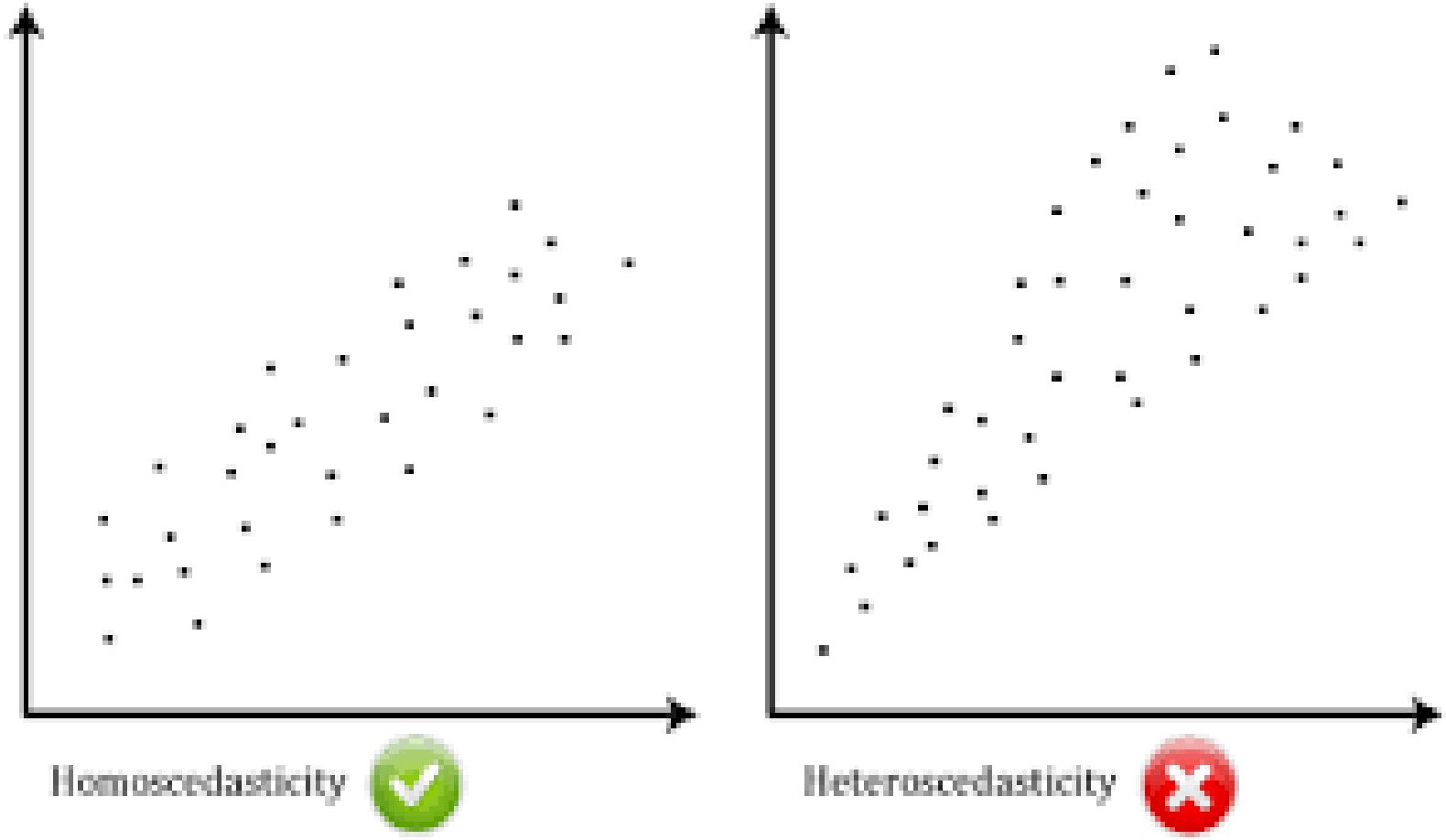
- There is no statistical independence among residuals:

$$y(x + 1) = y(x)$$

- Detected by
  - Scatter plots
  - Durbin-Watson’s d test: d values > 2.5 indicates autocorrelation

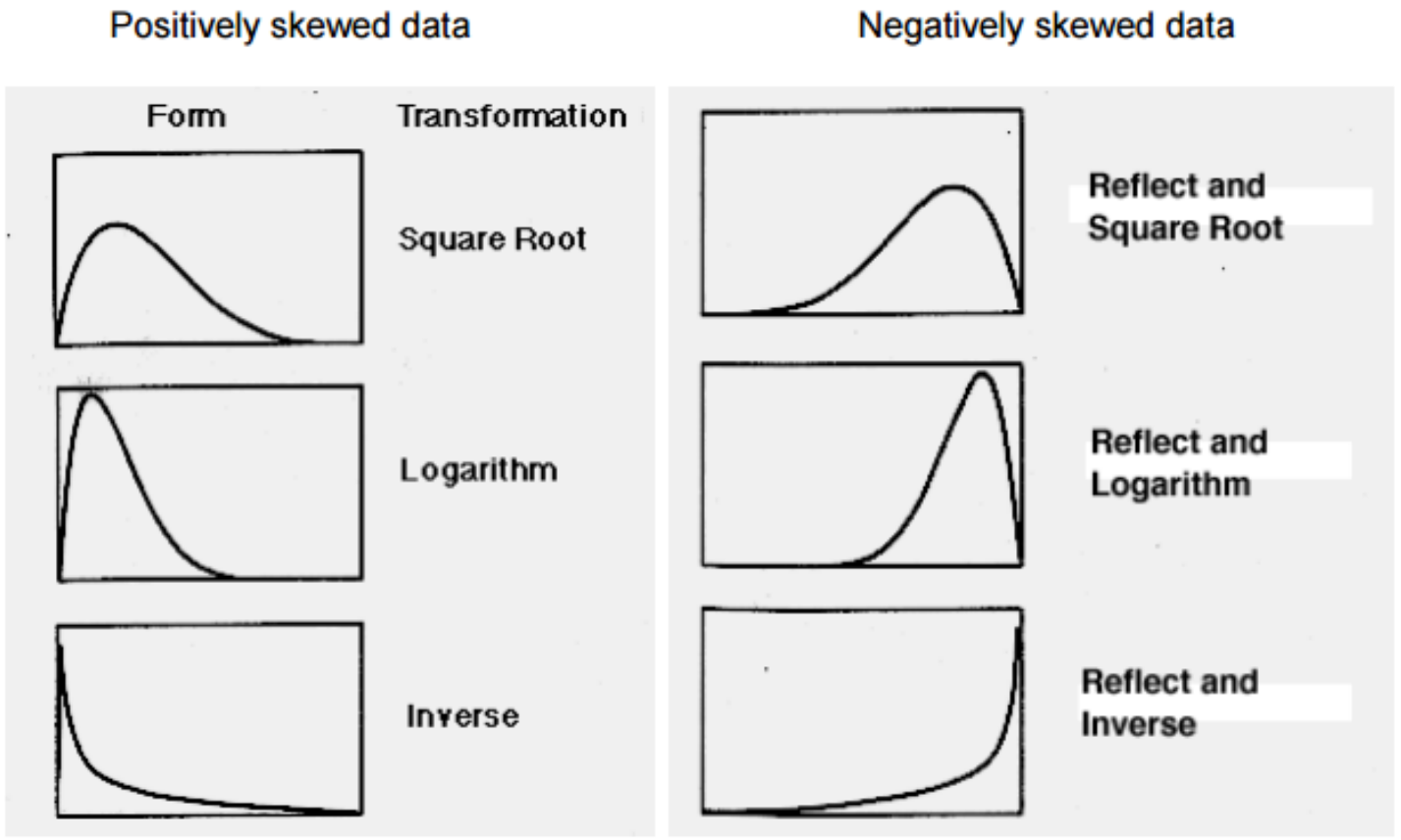
Assumptions: Homoscedasticity

- Data are homoscedastic if the residuals plot is the same width for all values of the response variable
- Detected by:
  - Scatterplot
  - Goldfeld-Quandt test



Transformations

- Moderate deviation: square root transformation
- Substantial non-normal: log transformation
- Severe non-normal: inverse transformation
- Negative skew: data reflection before transformation
- Heteroscedasticity: Use general least squares
- Non-linear: non-linear least squares or MLE



Transformation should be considered during model interpretations

http://rogeriofvieira.com/wp-content/uploads/2016/05/Data-Transformations-1.pdf

Model types

- Parametric: strong parametric assumptions. Average change in response variable is proportional to change in predictor variable- LMs, GLMs

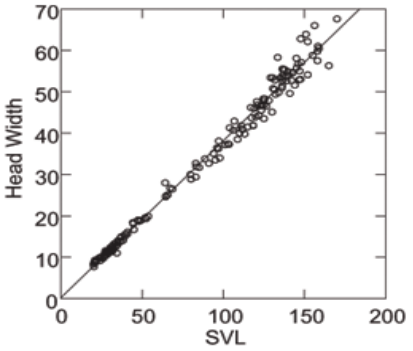
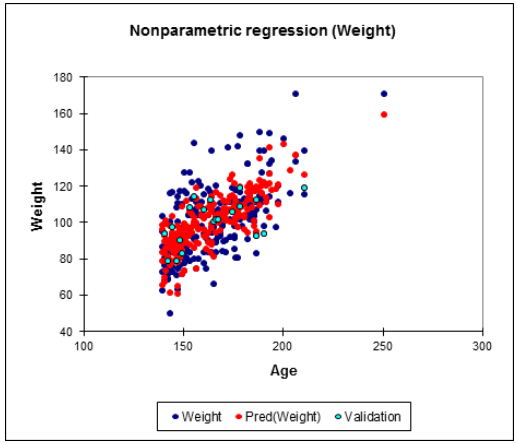
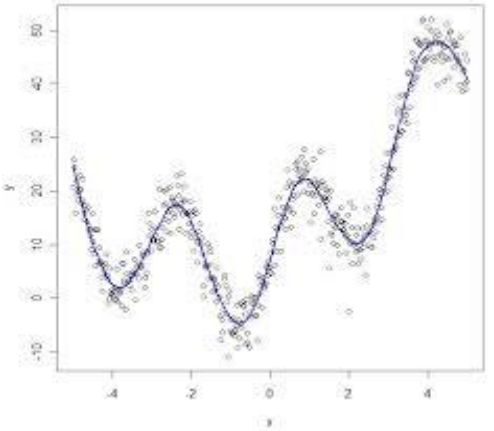


Fig. 5 – Regression of head width on SVL for juvenile and adult male specimens of *Leptodactylus knudseni*. Adjusted multiple $r^2 = 0.989$, $p = 0.000$. Values in mm.

- Non-parametric: no assumptions on relationships among variables- kernel smoothing



- Semi-parametric: general assumptions, such that relationships among variables are not restricted to any shape – additive models, GAMs.



Additive Models

- Developed by Stone (1985)
- Estimates additive approximation to multivariate regression function
- Advantages:
  - Avoids “curse of dimensionality” by using univariate smoother
  - Individual terms estimates explain relationship among variables

https://support.sas.com/rnd/app/stat/topics/gam/gam.pdf

Generalized Additive Models (GAMs)

- GAMs (Hastie & Tibshirani 1986, 1990) are semi-parametric extensions of GLMs, **only making assumption that the functions are additive and the components are smooth**
- GAMs have the ability to deal with highly non-linear and non-monotonic relationships between the response and explanatory variables

Their mentors, at Stanford, Drs. Nelder and Wedderburn developed GLMs

Etymology – what's in a name?

- From Italian word "gamba"
- In those days, it is a slang for a person's leg, especially an attractive woman's leg

Linear Regression Models

Recall…

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

$$\begin{aligned}
y_1 &= b_0 + b_1 X_{11} + \cdots + b_p X_{1p} + e_1 \\
y_2 &= b_0 + b_1 X_{21} + \cdots + b_p X_{2p} + e_2 \\
y_3 &= b_0 + b_1 X_{31} + \cdots + b_p X_{3p} + e_3 \\
&\vdots \qquad\qquad\qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
y_n &= b_0 + b_1 X_{n1} + \cdots + b_p X_{np} + e_n
\end{aligned}$$

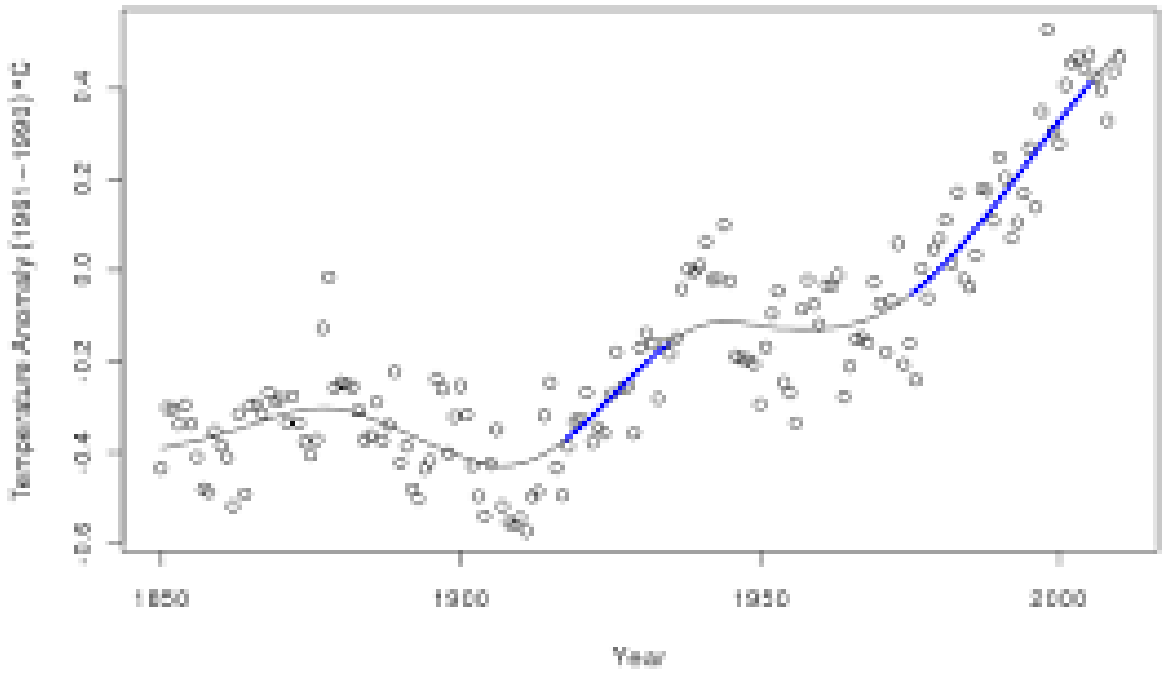
- **Y** - response variable **X** - explanatory variable
- **ε** - residual error, to cover unexplained information, assumed to be normally distributed with mean of 0 and δ^2
- **α** and **β** are intercept and slope respectively, to be determined at CI = 95%
- N – sample size

When to use GAMs

- When assumptions cannot be made on specific link function for error distribution
- Non-linearity in partial residual plots may suggest semi-parametric modeling
- Priori hypothesis or theory suggest non-linear or skewed relationship among variables
- ***Shape of predictor functions is determined by the data (Data speak for themselves!!)***

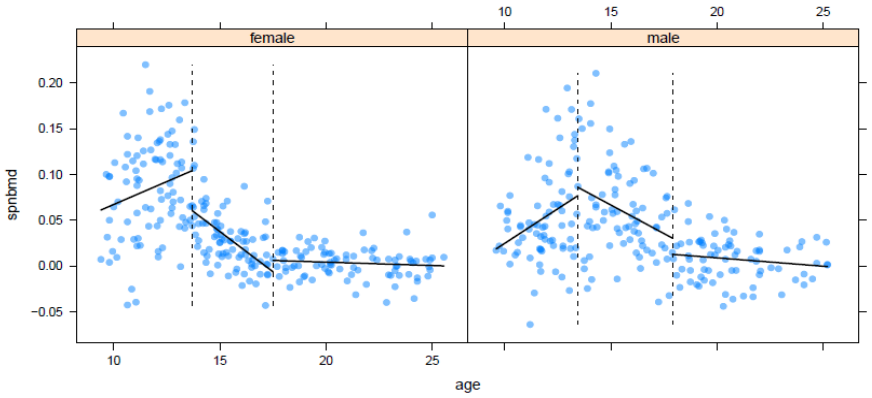
Generalized Additive Models

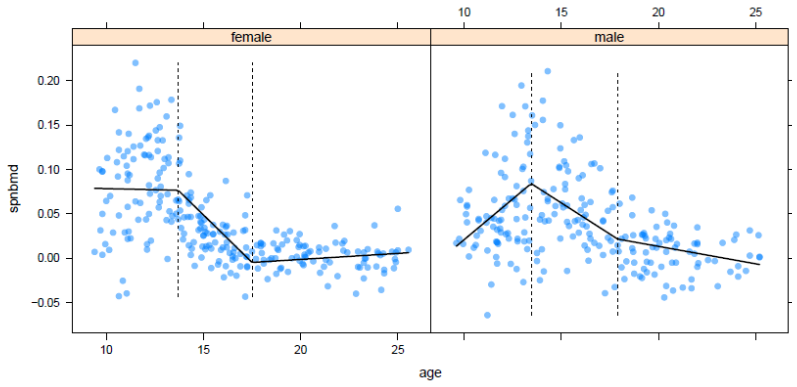
- **Expressed as:**
  - $\mathbf{Y = \alpha + \mathit{f}(X) + \varepsilon}$ where $\varepsilon \sim N(0, \delta^2)$
- Where $\boldsymbol{\beta X}$ are replaced with the smoothing curve $\boldsymbol{\mathit{f}(X)}$ which is not defined by an equation, but can be predicted from the model

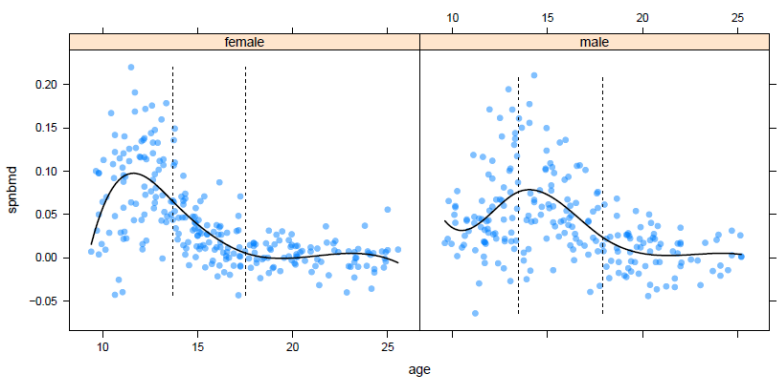


What GAMs do to your data?

- Separate each predictor into knots, k (sections)
- Fitting of data in each section independently using low order polynomial or spline functions
- Adds functions of all knots to predict the link function (smoothing): that's why it is called "additive" model
- Smoothing of knots is done by functions in "loess" and "splines" depending on R package used
- Model fitting is based on likelihood (e.g. AIC scores)







http://plantecology.syr.edu/fridley/bio793/gam.html

Uniqueness of GAMs

- A unique aspect of generalized additive models is the non-parametric (unspecified) function ***f*** of the predictor variables ***x***

- Generalized additive models are very flexible, and provide excellent fit for both linear and nonlinear relationships (multiple link functions)

- GAMs can be applied normal distribution as well as Poisson, binomial, gamma and other distributions…

- Regularization of predictor functions helps to avoid over-fitting

Advantages and application of GAMs

- Very powerful for prediction and interpolation
- Highly used in SDMs and ENMs (Elith et al. 2006)
- Analogous to hinge feature of maxent algorithm (Phillips et al. 2006)
- Building optimization models
- Comparatively GAMs shows lower AIC scores and explained higher deviance than GLMs
- Applied in Genetics, epidemiology, molecular biology, air quality and medicine (Dominici et al. 2002)

http://plantecology.syr.edu/fridley/bio793/gam.html

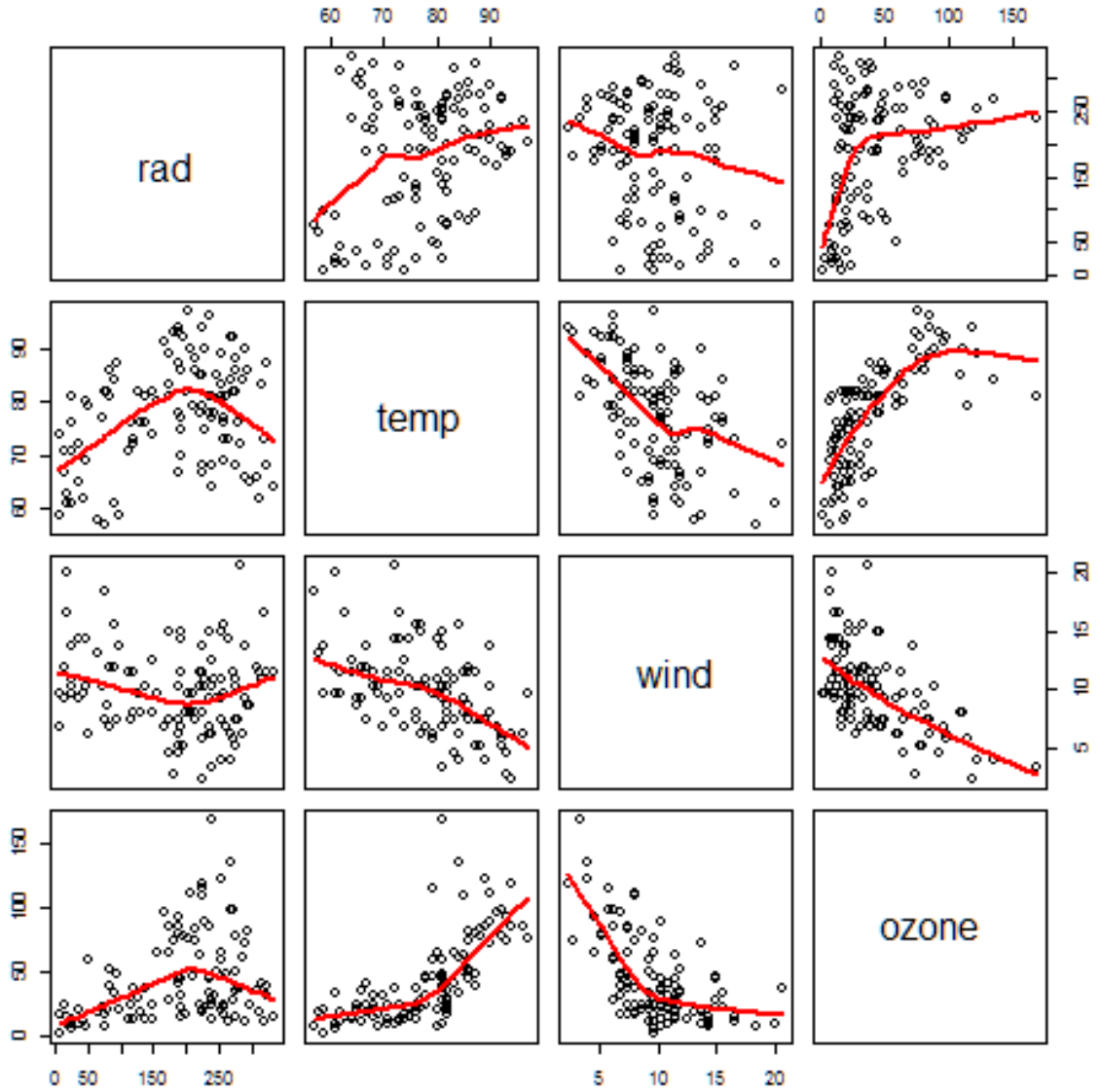
Packages that implement GAMs in R

- gdxrrw (can read or write GDX files)
- mgcv
- gam (old version of mgcv) – requires “splines” package
- mda – “bruto” function
- gamstools

Basic example

```
attach(ozone.data)
pairs(ozone.data, panel = function(x, y) {
    points(x, y)
    lines(lowess(x, y), lwd = 2, col = "red")
    }
```

http://geog.uoregon.edu/GeogR/topics/gamex1.html



Data and code from Crawley, M.J. (2005) *Statistics, An Introduction using R*, Wiley.

```
ozone.gam1 <- gam(ozone ~ s(rad) + s(temp) + s(wind))
summary(ozone.gam1)
```

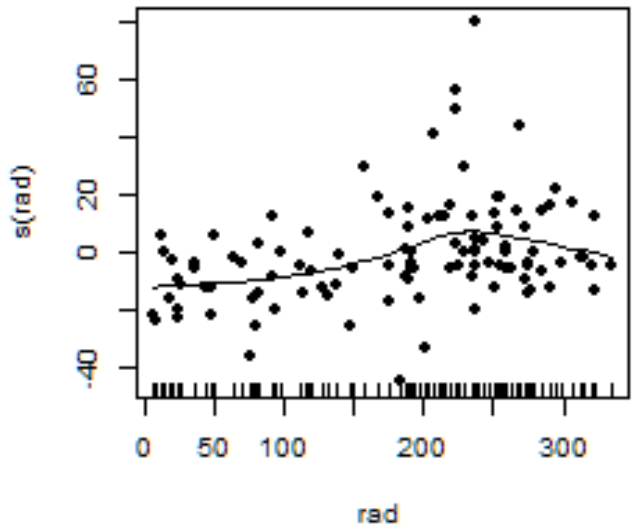
```
##
## Call: gam(formula = ozone ~ s(rad) + s(temp) + s(wind))
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -45.69 -10.05  -3.35    7.99   73.12
##
## (Dispersion Parameter for gaussian family taken to be 303)
##
##     Null Deviance: 121802 on 110 degrees of freedom
## Residual Deviance: 29698 on 98 degrees of freedom
## AIC: 963.4
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##           Df Sum Sq Mean Sq F value  Pr(>F)
## s(rad)     1  14374   14374    47.4 5.5e-10 ***
## s(temp)    1  37222   37222   122.8 < 2e-16 ***
## s(wind)    1  11464   11464    37.8 1.7e-08 ***
## Residuals 98  29698     303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F   Pr(F)
## (Intercept)
## s(rad)            3   2.05 0.11118
## s(temp)           3   6.67 0.00038 ***
## s(wind)           3   9.26 1.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

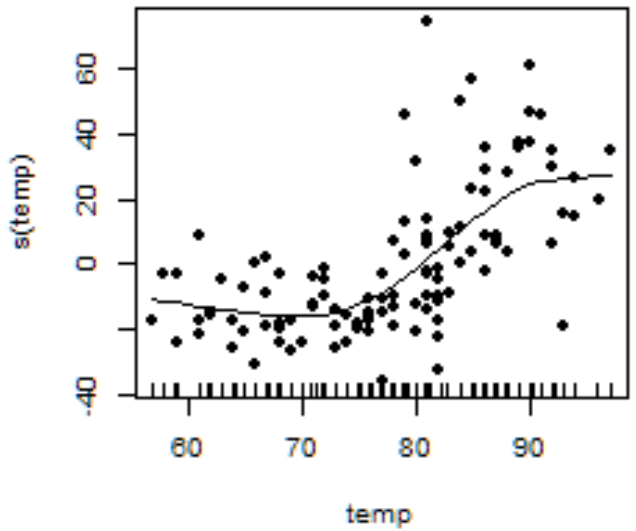
"s" is the smoother function added to the covariates

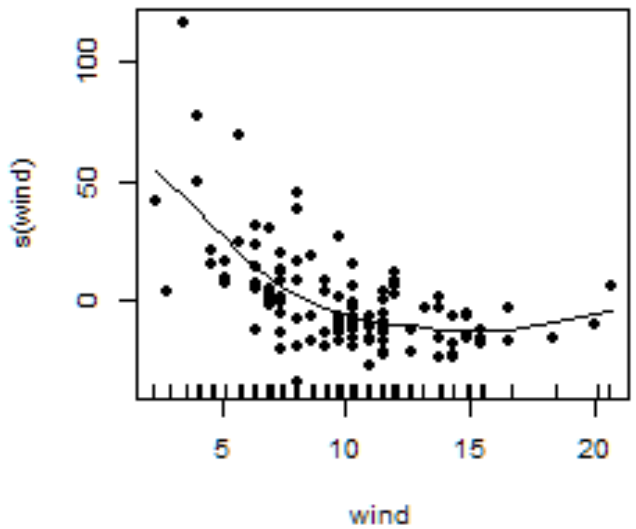
Significant effect shows evidence of non-linear relationship

http://geog.uoregon.edu/GeogR/topics/gamex1.html

```
plot(ozone.gam1,
resid = T, pch = 16)
```







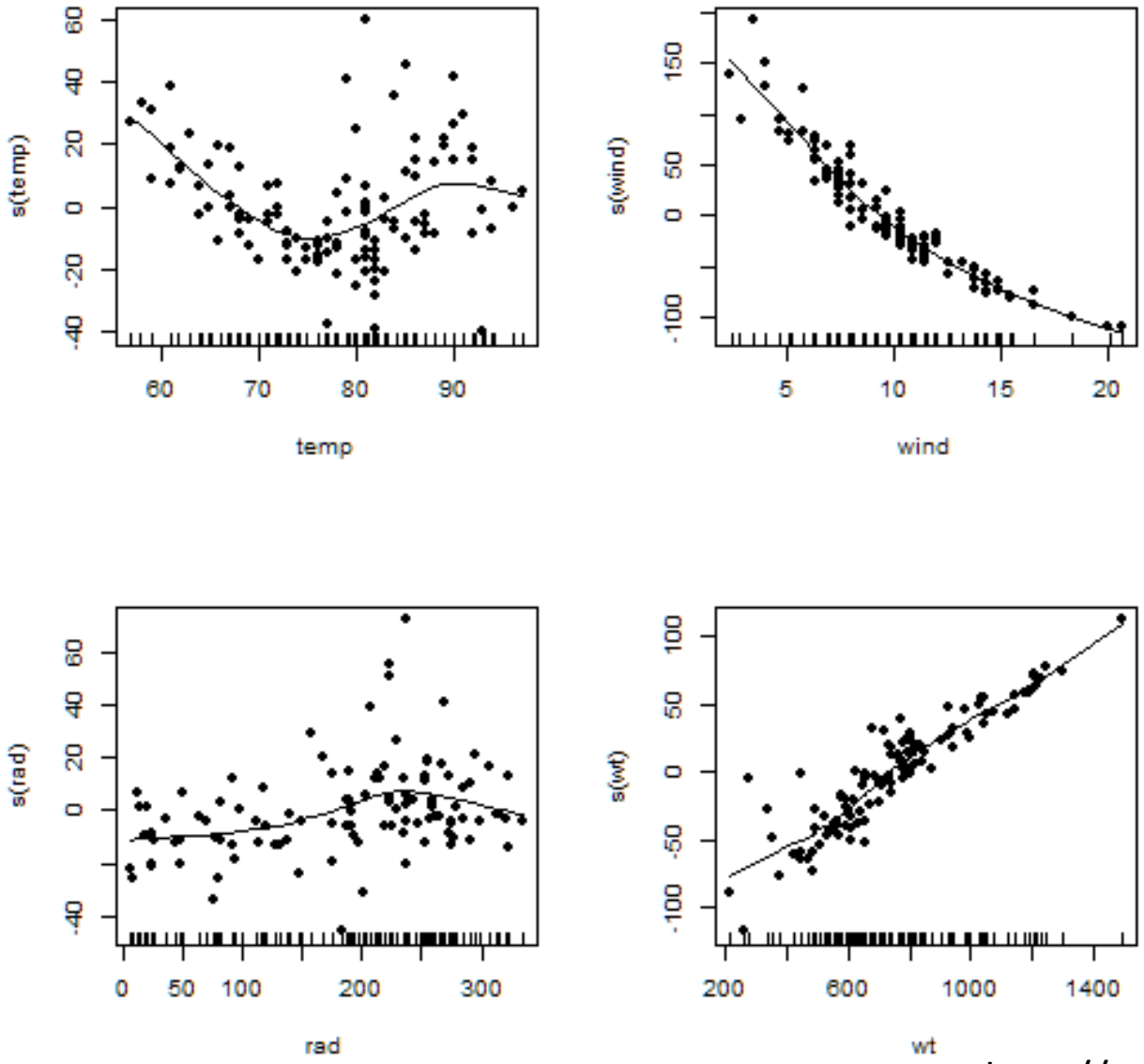
http://geog.uoregon.edu/GeogR/topics/gamex1.html

```
wt <- wind * temp
ozone.gam2 <- gam(ozone ~ s(temp) + s(wind) + s(rad) + s(wt))
summary(ozone.gam2)
```

```
##
## Call: gam(formula = ozone ~ s(temp) + s(wind) + s(rad) + s(wt))
## Deviance Residuals:
##    Min     1Q Median     3Q    Max
## -46.54  -9.18  -2.51   7.34  65.09
##
## (Dispersion Parameter for gaussian family taken to be 288.6)
##
##     Null Deviance: 121802 on 110 degrees of freedom
## Residual Deviance: 27132 on 94 degrees of freedom
## AIC: 961.4
##
## Number of Local Scoring Iterations: 8
##
## Anova for Parametric Effects
##           Df Sum Sq Mean Sq F value  Pr(>F)
## s(temp)    1  46069   46069   159.6 < 2e-16 ***
## s(wind)    1  11625   11625    40.3 7.7e-09 ***
## s(rad)     1   3694    3694    12.8 0.00055 ***
## s(wt)      1   3270    3270    11.3 0.00111 **
## Residuals 94  27132     289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F   Pr(F)
## (Intercept)
## s(temp)           3   8.42 5.1e-05 ***
## s(wind)           3  11.31 2.1e-06 ***
## s(rad)            3   2.04    0.11
## s(wt)             3   3.12    0.03 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

http://geog.uoregon.edu/GeogR/topics/gamex1.html

plot(ozone.gam2, resid = T, pch = 16)



GAMs has been reported not to handle interactions very well

http://geog.uoregon.edu/GeogR/topics/gamex1.html

Another example with more functions…

Ecology, 83(10), 2002, p. 2942

COASTAL ECOLOGICAL DATA FROM THE VIRGINIAN BIOGEOGRAPHIC PROVINCE, 1990–1993

***Ecological Archives* E083-057**

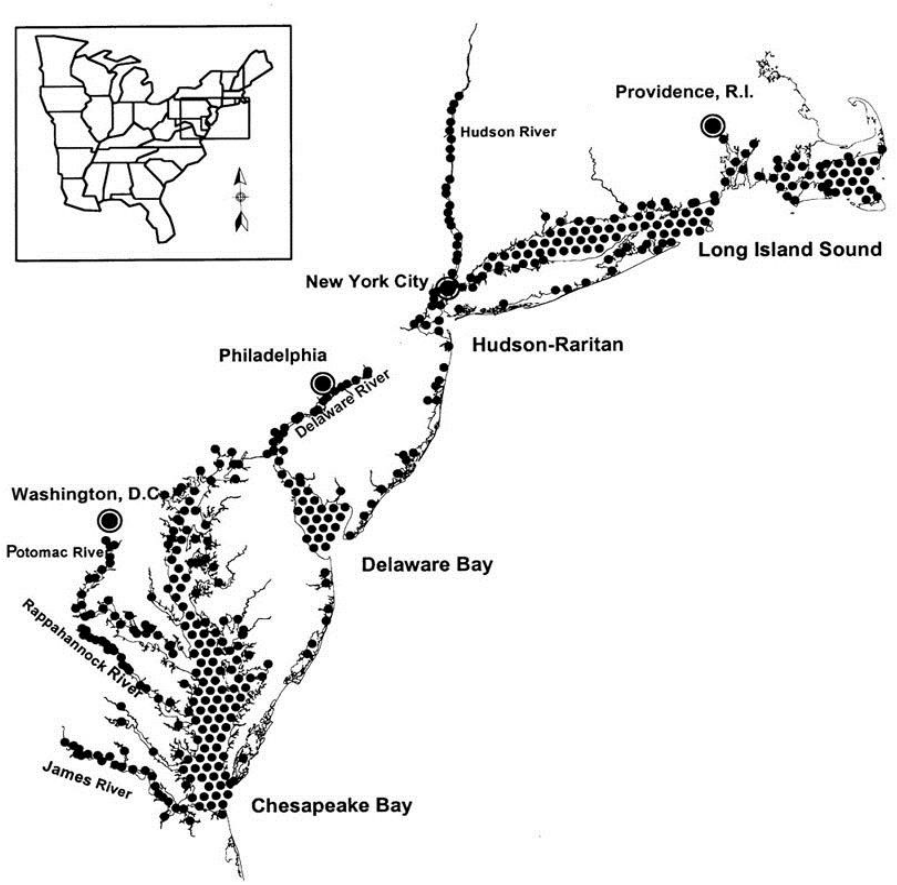
STEPHEN S. HALE,[2] MELISSA M. HUGHES, CHARLES J. STROBEL, HENRY W. BUFFUM, JANE L. COPELAND, and JOHN F. PAUL

[1]*Atlantic Ecology Division, U.S. Environmental Protection Agency, 27 Tarzwell Drive, Narragansett, Rhode Island 02882 USA*

To investigate how community diversity (measured by Shannon's Index) is influenced by environmental variables like water quality and sediment

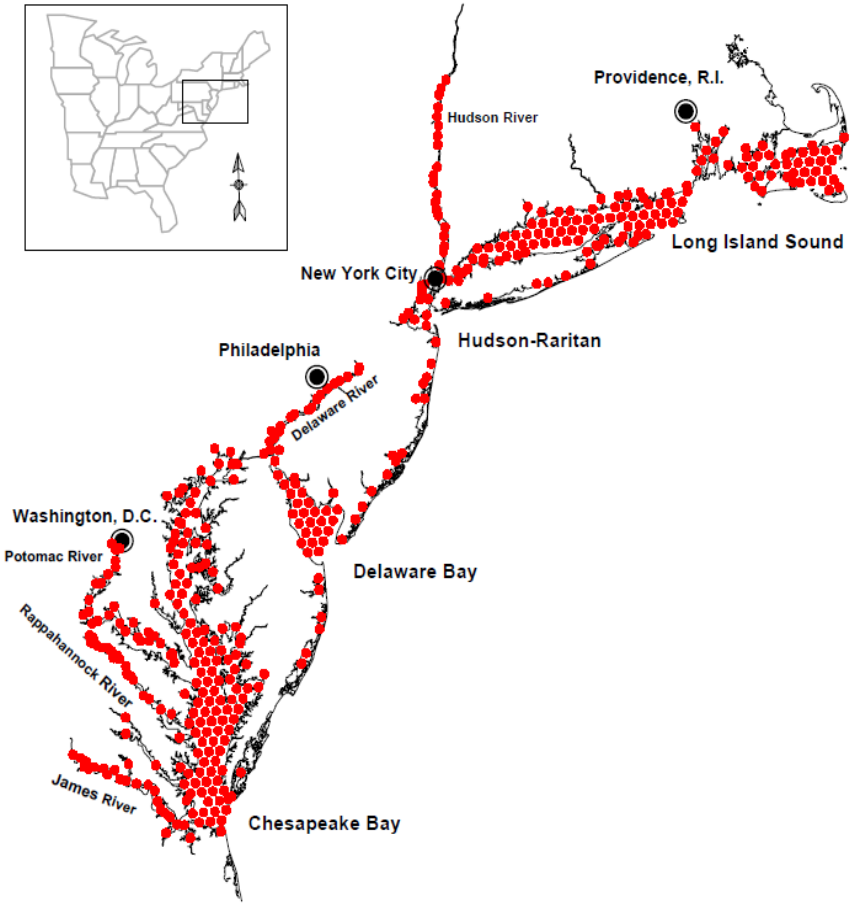
Getting started

Data set collected from 303 stations in estuaries, bays, and tidal rivers located in the Virginian Biogeographic Province (Cape Cod MA to Cape Henry VA) by the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program



Variables

Parameters collected include: dissolved oxygen (DO) estuary strata, pH, salinity, temperature, fluorescence, depth, photosynthetically active radiation [PAR] (mE/m^2/s), density and frequency of fish diversity, total organic carbon (TOC) and transmissivity.



GAM fitting

Here, k is specified

```
library(mgcv, quietly = TRUE)
#?mgcv: You will use this...
Form1 <- formula(H.c ~ s(DO, k = 10) + #Dissolved oxygen
                       s(SAL, k = 10) + #Salinity
                       s(PH, k = 10) + #pH
                       s(FLUOR, k = 10) + #Fluorescence
                       s(TRANS, k = 10) + #Transmissivity (%)
                       s(log(PAR), k = 10) + #Photosynthetically active radiation
                       #s(DENSITY) + #Density... omitted b/c correlated w/ SAL
                       s(DEPTH, k = 10) + #DEPTH
                       s(TOC, k = 10) + #Total Organic Carbon (%)
                       s(LON, LAT, k = 25)) #Longitude and Latitude

# Note that LON and LAT are included within the same s() term, this means that they
# are not additive, rather interacting.
```

Commands

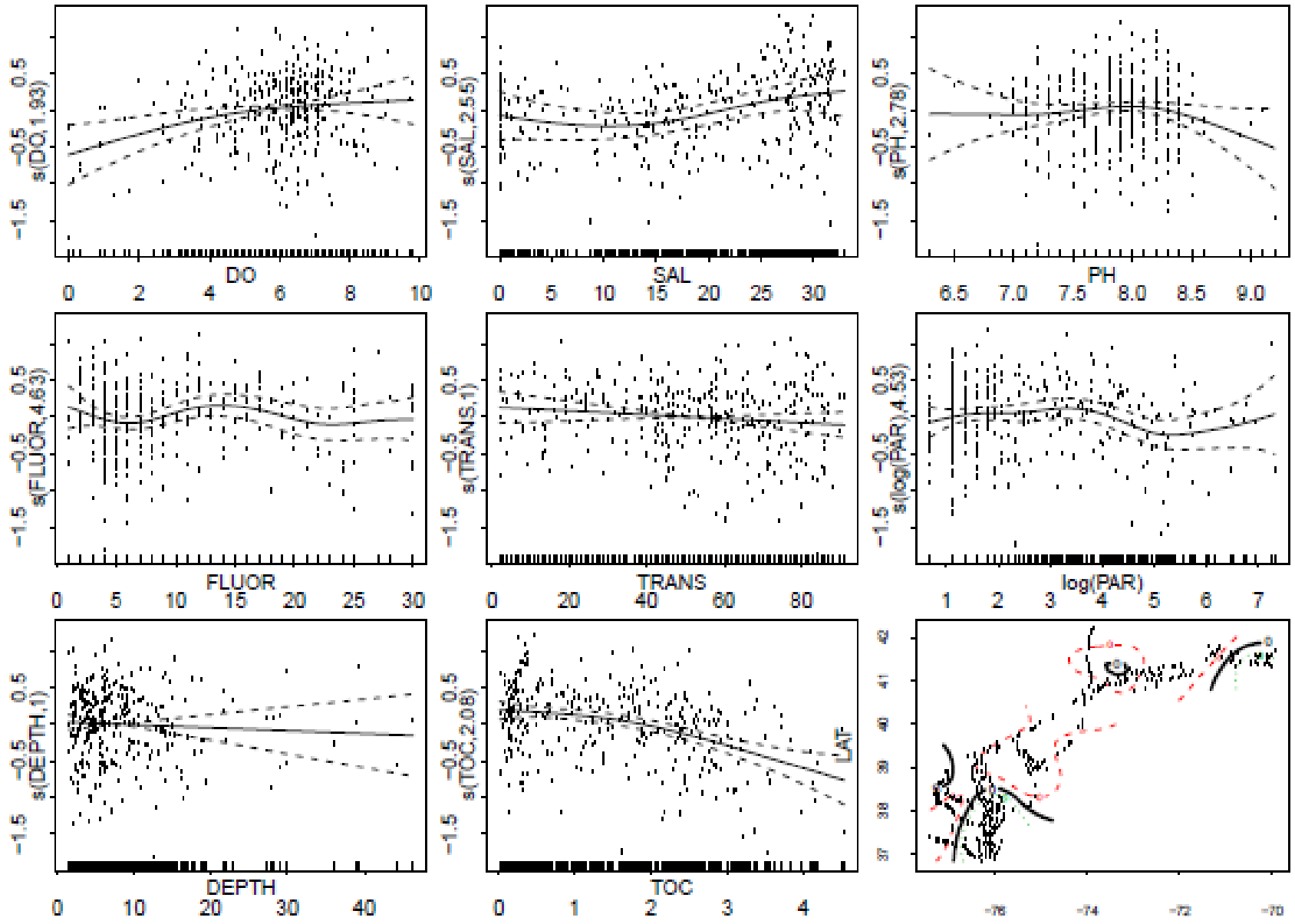
- Independent (i.e., additive): **$s(x_1) + s(x_2)$, ...** Where x_1 and x_2 are covariates that the smooth is a function of.
- Interaction: If covariates are on same scale: **s(x1, x2)...**, for example, longitude and latitude (use isotropic smoothing): s(LON, LAT, k = 25). If covariates aren't on the same scale: **te(x1, x2, ...)** formulation of tensor product smoothers
- Removing the s() from a term: x1 + x2,... removes the smoother, and it effectively becomes a linear component.
- Knots, k: specifies the dimension of the basis function used to represent the smooth term (also called smoothing parameter, **λ or α**)

```
G1 <- gam(Form1, data = dd)
summary(G1)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## H.c ~ s(DO, k = 10) + s(SAL, k = 10) + s(PH, k = 10) + s(FLUOR,
##     k = 10) + s(TRANS, k = 10) + s(log(PAR), k = 10) + s(DEPTH,
##     k = 10) + s(TOC, k = 10) + s(LON, LAT, k = 25)
## <environment: 0x028f2ee8>
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0883     0.0283    73.8   <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(DO)         1.93   2.44  3.56   0.023 *
## s(SAL)        2.55   3.17  2.18   0.087 .
## s(PH)         2.78   3.57  1.51   0.202
## s(FLUOR)      4.63   5.63  1.64   0.141
## s(TRANS)      1.00   1.00  1.52   0.218
## s(log(PAR))   4.53   5.51  1.68   0.133
## s(DEPTH)      1.00   1.00  0.29   0.588
## s(TOC)        2.08   2.60 14.54 9.5e-08 ***
## s(LON,LAT)   17.86  21.33  4.68 2.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.465   Deviance explained = 52.8%
## GCV score = 0.29571  Scale est. = 0.25991   n = 325
```

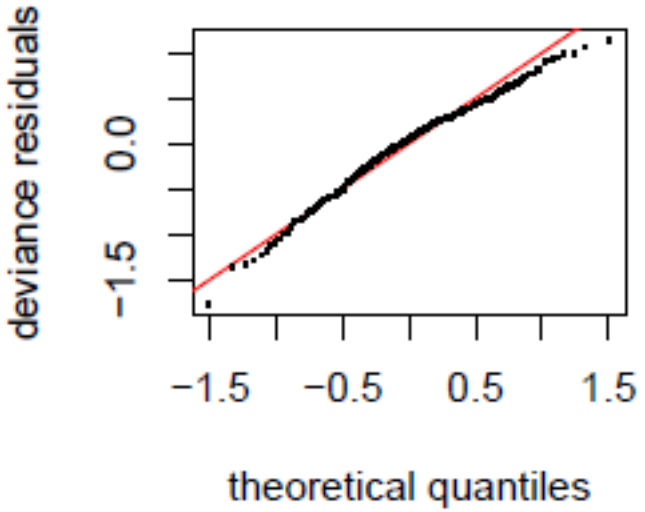
AIC = 523.2

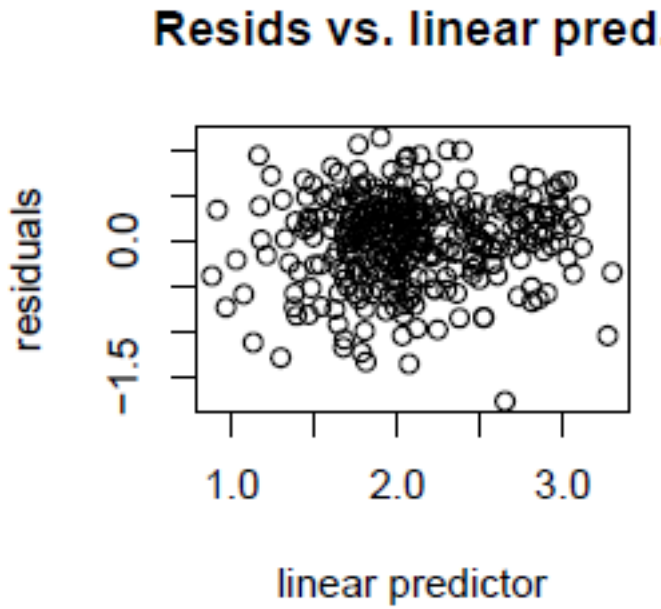
plot(G1)

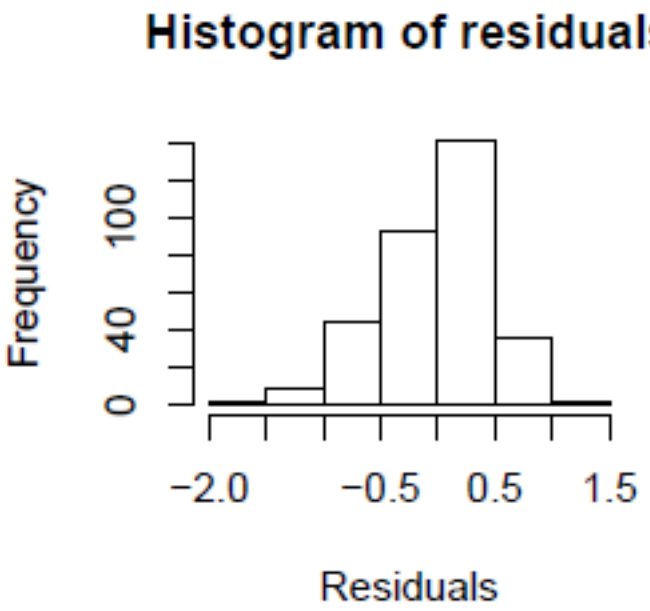

s(DO,1.93)
DO
s(SAL,2.55)
SAL
s(PH,2.78)
PH
s(FLUOR,4.63)
FLUOR
s(TRANS,1)
TRANS
s(log(PAR),4.53)
log(PAR)
s(DEPTH,1)
DEPTH
s(TOC,2.08)
TOC
LAT

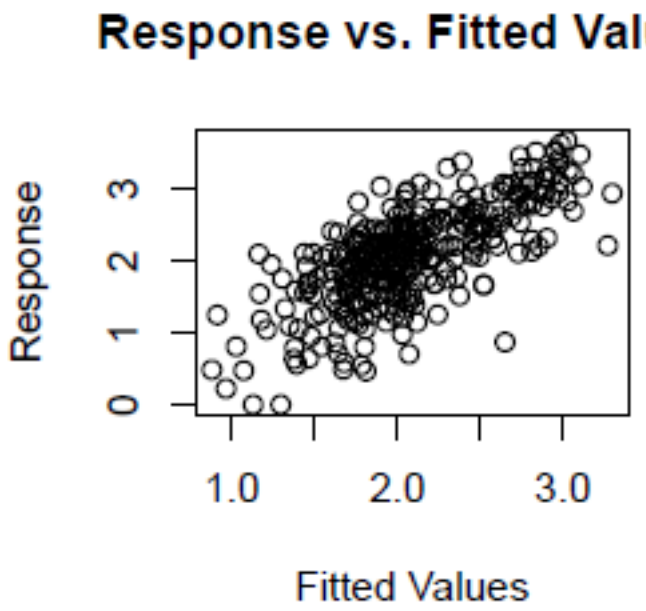
Using *gam.check* function, we check how k selection fits the predictors: is it too low or too high?

gam.check or *qq.gam* produces residual plots



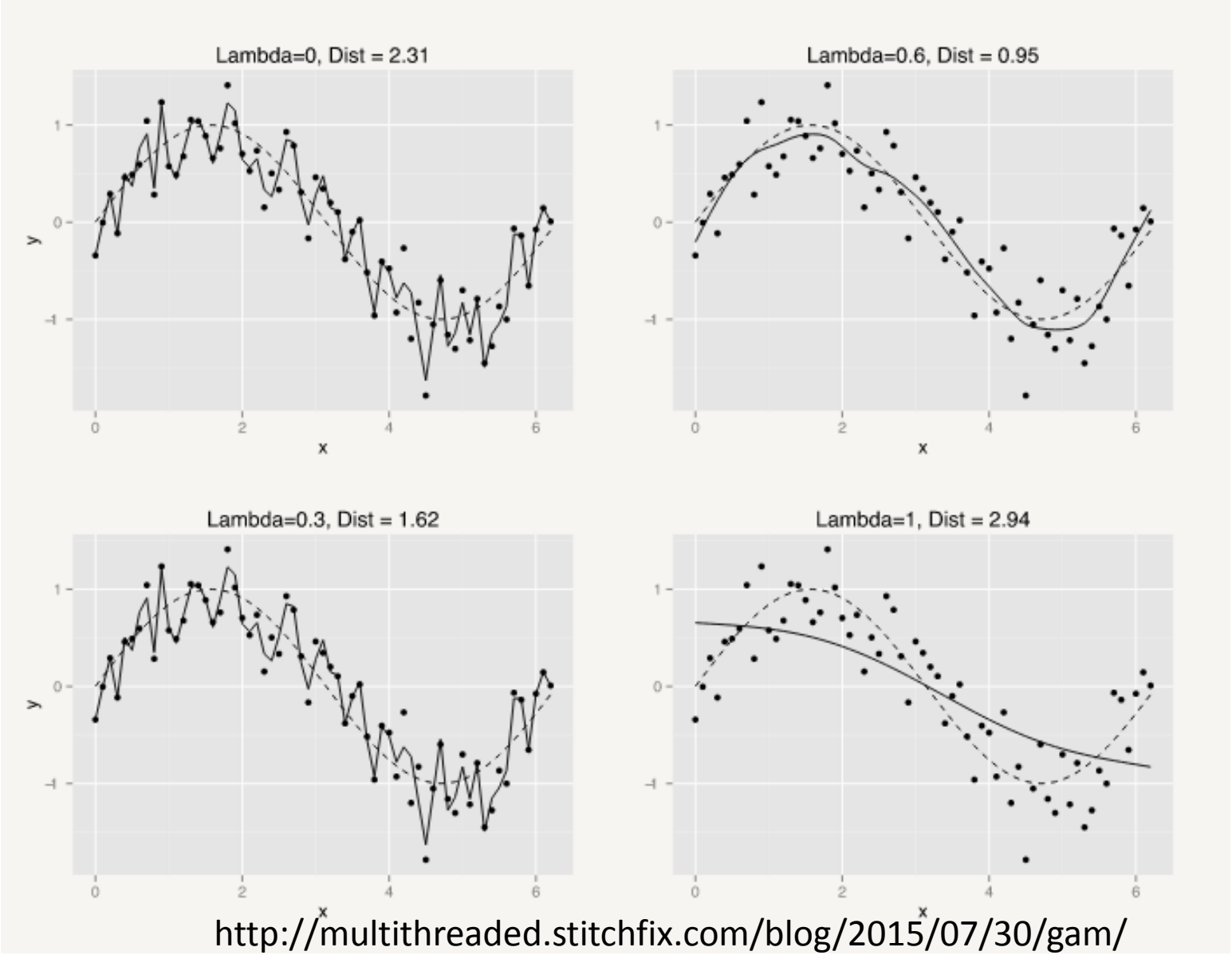






K selection and overfitting

- If $\boldsymbol{\alpha}$ is too large, we run risk of underfitting, and if $\boldsymbol{\alpha}$ is too small, overfitting can occur.
- Trade-off in bias (in-sample error) and variance
- Curves with less variance are good for prediction



http://multithreaded.stitchfix.com/blog/2015/07/30/gam/

Smoothing Parameter (λ, k or α)

- There are different methods used to select k:
  - **Cross-validation methods** *(found in R package mgcv)*
    - Cross-validation (CV)
    - Generalized Cross-validation (GCV)
    - Unbiased Risk Estimator (UBRE)
  - **Likelihood Methods**
    - Restricted Maximum Likelihood (REML)
    - Maximum Likelihood (ML)

Explore smooth.terms in "mgcv" package for thorough explanations

How to deal with over-fitting in GAMs

- Model selection with AIC or BIC
- Simple models vs. complex models: curse of dimensionality
- Predictor selection: backward or forward
- Cross validation: 4 or 5-folds (training data)
- Regularization: penalize sources of over-fitting
- Reduce feature space using tools like PCA
- Use bagging (bootstrap aggregation)
- Iterative modelling and play around with k

https://www.quora.com/How-can-I-avoid-overfitting

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 12 iterations.
## The RMS GCV score gradiant at convergence was 3.573e-08 .
## The Hessian was positive definite.
## The estimated model rank was 97 (maximum possible: 97)
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                 k'    edf k-index p-value
## s(DO)        9.000  1.927   1.022    0.66
## s(SAL)       9.000  2.545   1.067    0.89
## s(PH)        9.000  2.778   1.065    0.90
## s(FLUOR)     9.000  4.630   0.958    0.22
## s(TRANS)     9.000  1.000   1.036    0.74
## s(log(PAR))  9.000  4.526   1.016    0.52
## s(DEPTH)     9.000  1.000   1.096    0.96
## s(TOC)       9.000  2.082   1.003    0.50
## s(LON,LAT)  24.000 17.863   1.103    0.97
```

Iterative modelling until you produce the best fit and optimal k

Degrees of Freedom (df or K')

- Df is equal to the number of parameters needed to produce the curve, and is calculated by:
  - Df = number of knots – 1
- The – 1 part is caused by identification constraint which ensures that all possible predictions from every smoother included in GAM equal to zero
- We use effective degrees of freedom (edf), which is inversely linked with λ, to compare smoothers

- High edf (≥8) means that the curve is non-linear (low λ), edf = 1 is a straight line (high λ)

Very useful resources

- **https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/summary.gam.html**
- **http://multithreaded.stitchfix.com/blog/2015/07/30/gam/**
- **https://support.sas.com/rnd/app/stat/topics/gam/gam.pdf**
- **http://plantecology.syr.edu/fridley/bio793/gam.html**
- **http://geog.uoregon.edu/GeogR/topics/gamex1.html**
- **https://rpubs.com/ryankelly/GAMs**