Generalized Additive Model when Variables are not Normal and Associations are not Linear

Diana Šimić

CECIIS — Workshop on Data Analysis
28 September 2017

Workshop on Data Analysis

- What do you expect from a workshop?
  Information or Education?
- What is data analysis to you?
  Depends on who you are:
  - statistician
  - informatics professional
  - computer scientist
  - data scientist
  - economist
  - health professional ...

Statistician vs. Data Scientist

	Statistician	Data Scientist
Goal	Explain or model variation	Predict values
Evaluation	Parsimony, fit, interpretation	Prediction errors
Generalization	Randomization (Experiment or Sample)	Cross-validation (Big Data)
Models	Base models on theory	Infer models from data

inspired by Bojana Dalbelo Bašić @ BIOSTAT 2017

Statistical Thinking 1

Model variation of “dependent variable” y using distribution function

$$y \sim F(y, \theta)$$

where $\theta = [\theta_1, \theta_2, ..., \theta_k]$

Explain variation in “dependent variable” y given predictors $x = [x_1, x_2, ..., x_p]$ as

$$y|x \sim F(y, \theta(x))$$

Statistical Thinking 2

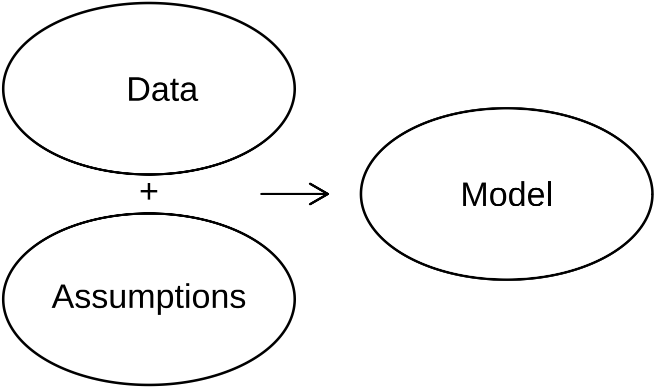
- World is inherently stochastic (random).
- Given a good model and predictors we can significantly reduce "unexplained variation".
- Given data on a representative (random) sample we can reliably estimate parameters of the model so that model describes the target population well.
- Good models reflects "real" relationships in target population.

Data Science Thinking

- World is inherently deterministic.
- Given a good algorithm and predictors we can significantly reduce "prediction error".
- Given enough (big) data predictions will be accurate.
- We do not aim to draw inference on the nature of relationships in target population from DS algorithms. After all, if prediction errors become too large in the future, we will find a better algorithm.

From the point of view of positivistic epistemology . . . this is not science at all.

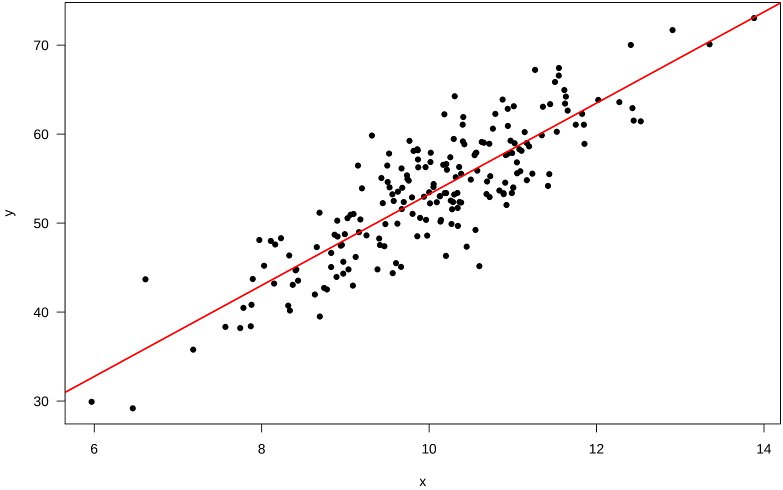
Statistical Modeling



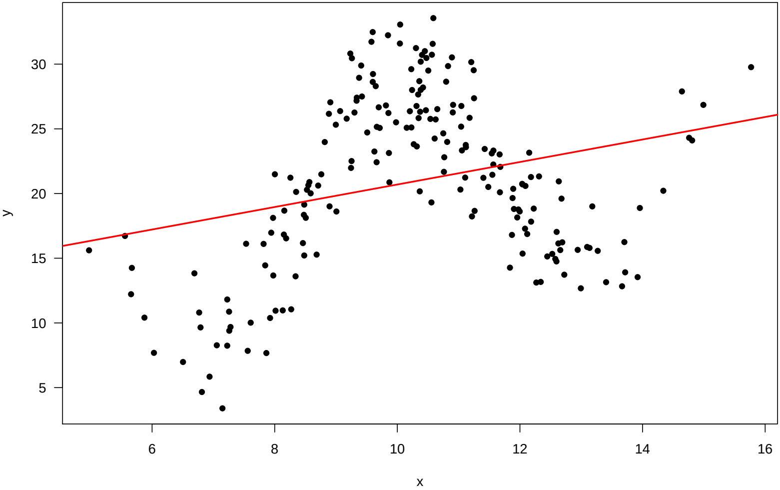
Linear Regression

- One of the most popular statistical models
- Model variation in a numerical (dependent) variable given values of one (or more) “independent” variables (predictors)
- Usually introduced as the best line in the sense of minimum sum of squared errors

Is this good?



How about this?



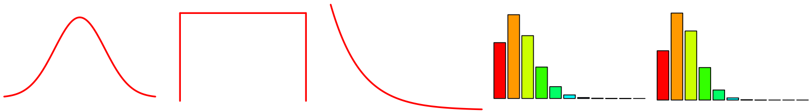
A different angle

- Instead of asking: How to minimize sum of squared errors?
- Think: How to model variation in y given x?

Use an appropriate distribution function.

Some distribution functions

Distribution	Domain	Function	Expectation	Variance
Normal	$y \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform	$y \in [a, b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$y \in [\alpha, \infty)$	$\frac{1}{\beta} e^{-\frac{y-\alpha}{\beta}}$	$\alpha + \beta$	β^2
Poisson	$y \in \{0, 1, \dots\}$	$\frac{e^{-\lambda}\lambda^y}{y!}$	λ	λ
Binomial	$y \in \{0, 1, \dots, n\}$	$\binom{n}{y} p^y (1-p)^{(n-y)}$	np	$np(1-p)$



Normal Uniform Exponential Poisson Binomial

Back to linear regression 1

Data:

- Dependent variable (quantitative)
- One or more independent variables (quantitative or indicator)

Assumptions:

- Relationship between the dependent and independent variables is linear
- Residuals follow normal distribution
- Residuals are independent from prediction and any of the independent variables
- Residuals are homoscedastic (i.e. have constant variance)

Model:

- Conditional distribution of the dependent variable, given values of the independent variables is normal, with constant variance and mean that is a linear combination of independent variables.

Back to linear regression 2

Let $y = [y_1, y_2, \ldots, y_n]^T$ be a column vector representing the dependent variable.

Let

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

be a matrix with columns representing the independent variables, where the first column contains number 1 in all rows, and x_i^T represents the i-th row.

Let $b = [\beta_0, \beta_1, \ldots, \beta_p]$ be a column vector of regression coefficients.

Linear regression model can be stated as:

$$y_i | x_i \sim N(x_i^T b, \sigma^2) \quad \text{or} \quad y_i | x_i = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \epsilon_i ; \epsilon_i \sim N(0, \sigma)$$

From linear regression to general linear model

- Matrix X is usually called the design matrix.
- Independent variables can be qualitative. Such variables are represented by a set of indicator columns in the design matrix.
- Such a model includes:
  - Simple linear regression
  - Multiple linear regression
  - t-test
  - analisis of variance
  - analysis of covariance
  - ...

What can go wrong?

Starting from the model:

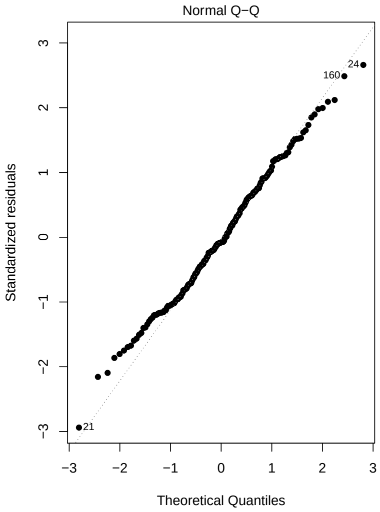
$$y|x \sim N\left(x^T b, \sigma\right)$$

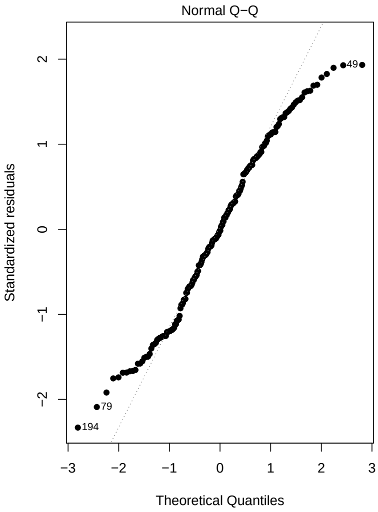
1. Conditional distribution of y given x might not be normal
2. Expectation of y given x might not be a linear combination of the elements of x
3. Variance might not be constant
4. Outliers may influence parameter estimates.

We can check these assumptions using diagnostic graphs.

Checking Assumptions: Normality

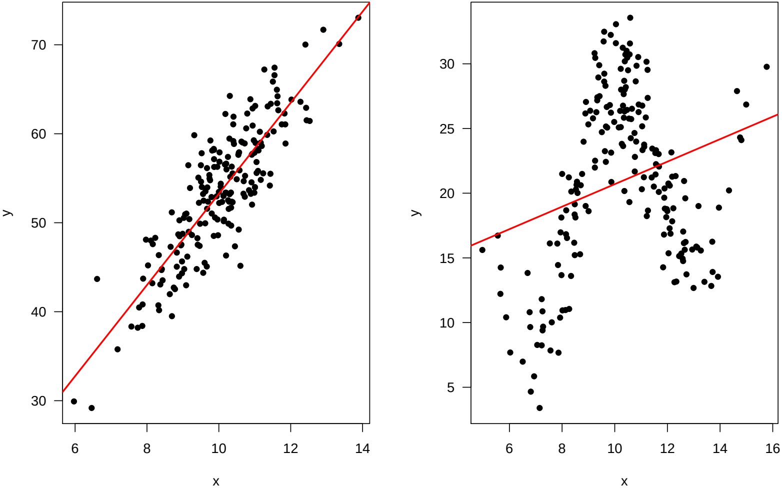
Residual qq-plot





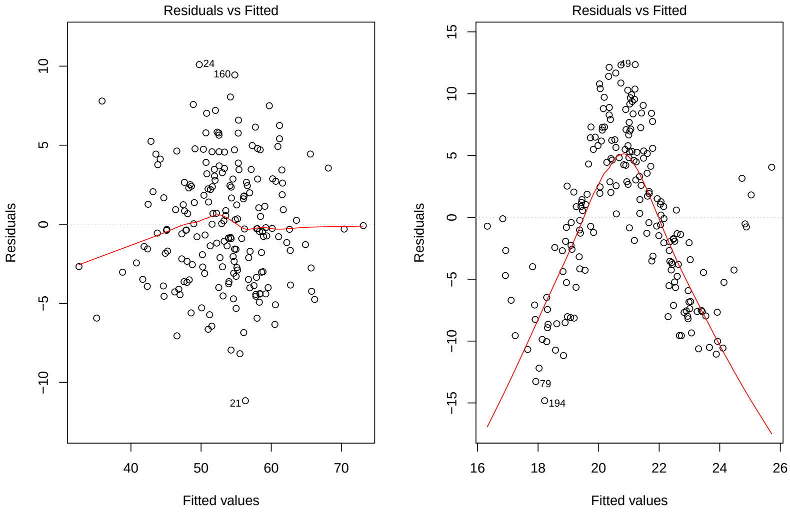
Checking Assumptions: Linearity

Scatterplot with line of linear regression.



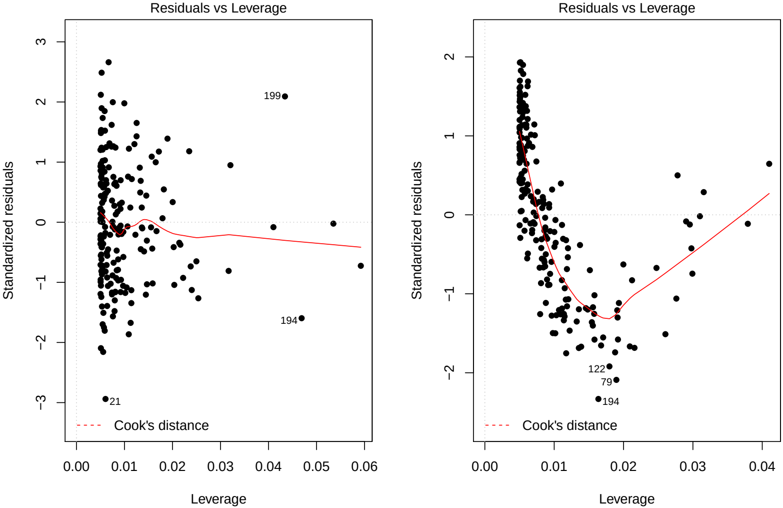
Checking Assumptions: Independence, homoscedasticity

Residual scatterplot



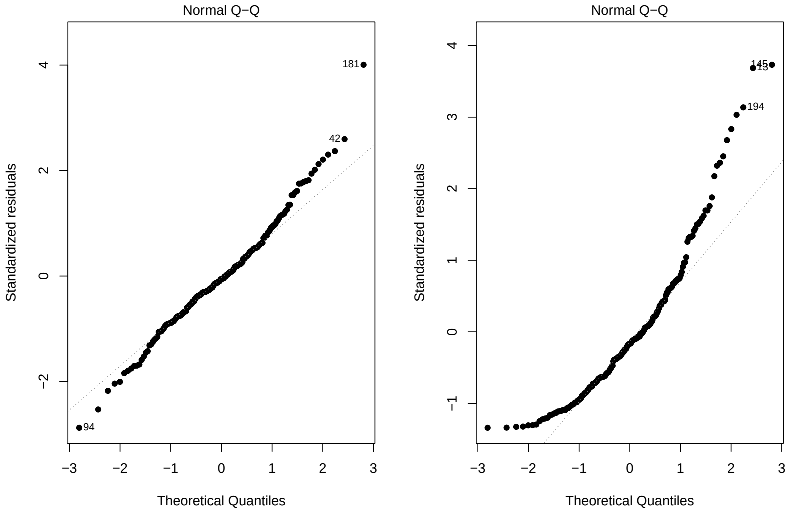
Checking Assumptions: Outliers

Leverage plot

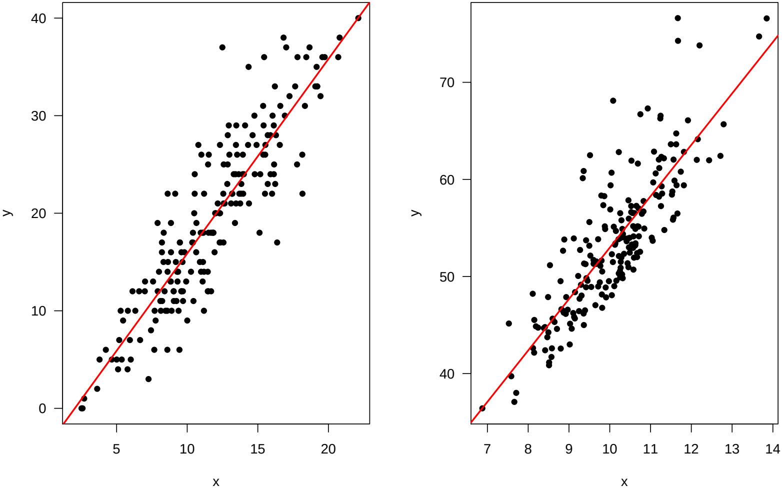


More examples: Normality

Residual qq-plot

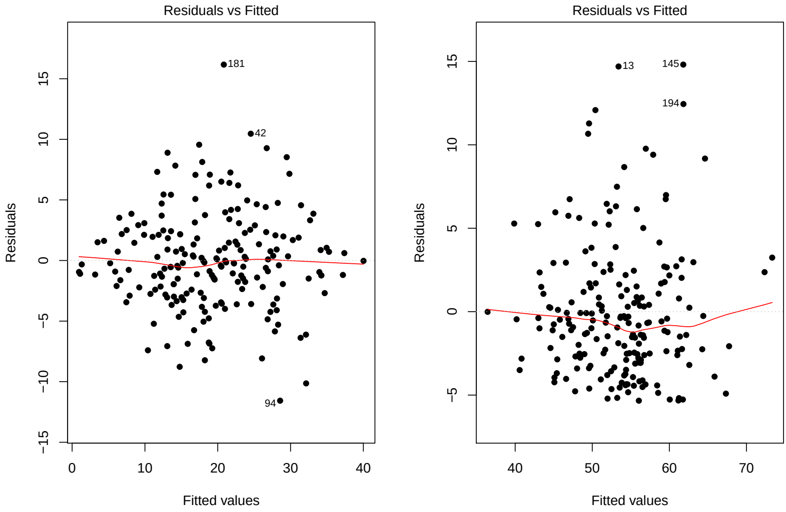


More examples: Linearity



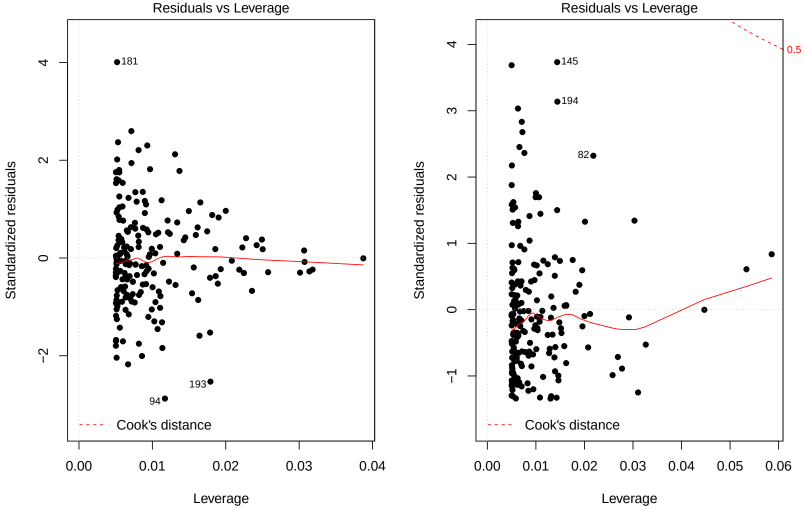
More examples: Independence, homoscedasticity

Residual scatterplot



More examples: Outliers

Leverage plot



When distribution is not normal . . .

Use model with a different distribution: generalized linear model

Instead of model

$$y|x \sim N\left(x^T b, \sigma\right)$$

use model

$$y|x \sim F(\theta)$$

and

$$E(y|x) = g^{-1}\left(x^T b\right)$$

where F is a distribution function from the exponential family, and g is a link function.

Exponential family

Distributions that can be written in the form

$$f(x|\theta) = A(x)B(\theta)e^{\eta(\theta)T(x)}$$

Notice that components that depend on the value of the variable and on the value of the parameter can be separated.

This family contains many distribution functions: normal, Poisson, beta, gamma, exponential, binomial and multinomial with fixed number of trials, negative binomial with fixed number of failures ...

$\eta(\theta)$ is called natural parameter.

Function η is a natural link function.

Logistic regression

- outcome is proportion of succeses for a known number of trials
- prediction is expected probability of successes given predictors ($\hat{p}$)
- natural link function is logit:

$$\eta(p) = ln\left(\frac{p}{1-p}\right)$$

- quantity under the logarithm is called odds ratio
- log odds ratio is modeled as a linear combination of predictors

Poisson regression

- outcome is a number of events in a given time interval
- prediction is expected rate of succes given predictors ($\hat{\lambda}$)
- natural link function is natural logarithm

$$\eta(\lambda) = ln\lambda$$

- log Poisson rate is modeled as a linear combination of predictors

When association is not linear ...

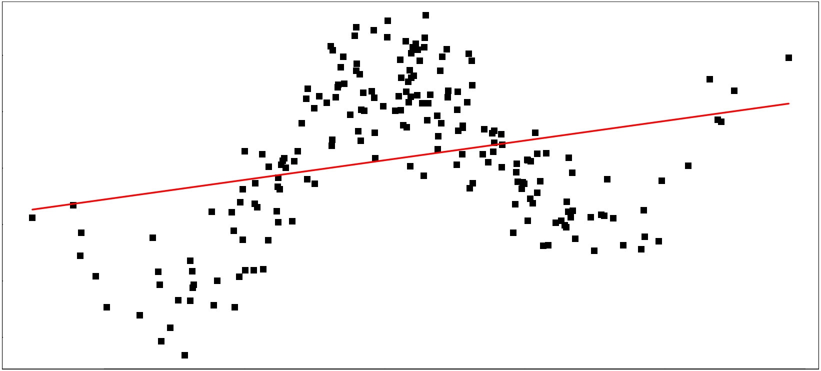
- transform independent variables (e.g. polinomial terms, logarithms, exponential function etc.)
- transform dependent variable (e.g. logarithm etc.)

These approaches are constrained to known functional forms ...

- use nonparametric smoother

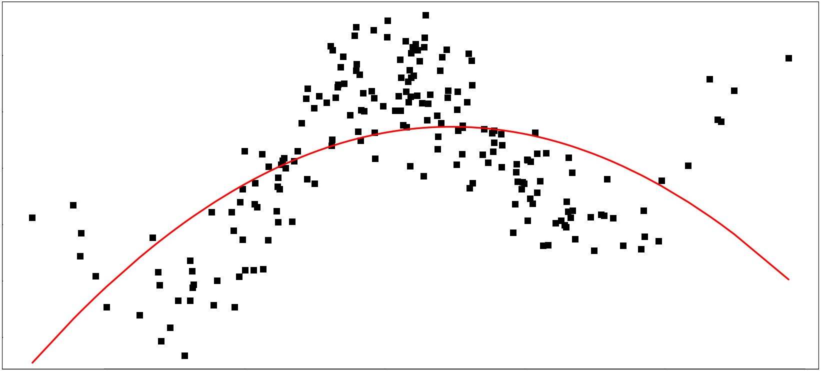
Linear regression

```
pom <- lm(y ~ x)
plot(x,y, pch=15)
lines(x[order(x)], pom$fitted.values[order(x)],
      lwd=2, col="red")
```



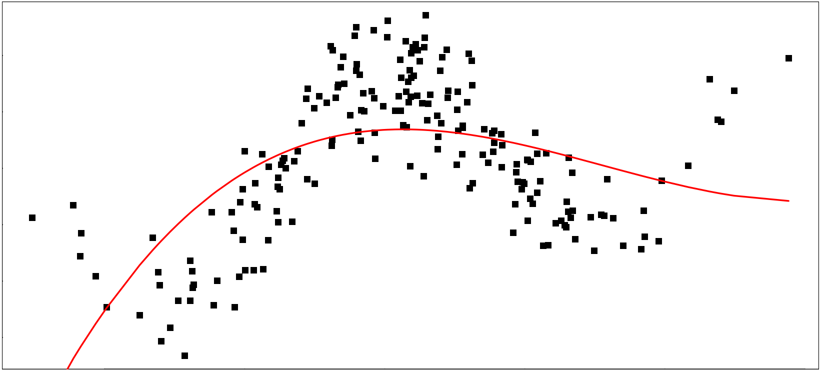
Quadratic regression

```
x2 <- (x - mean(x))^2
pom <- lm(y ~ x + x2)
plot(x,y, pch=15)
lines(x[order(x)], pom$fitted.values[order(x)],
      lwd=2, col="red")
```



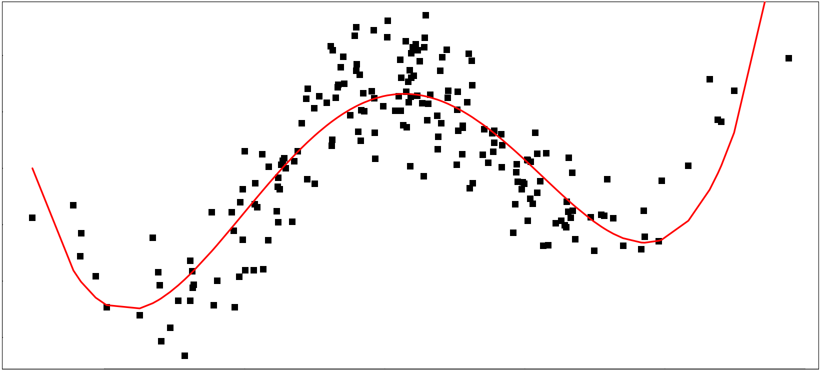
Cubic regression

```
x3 <- (x - mean(x))^3
pom <- lm(y ~ x + x2 + x3)
plot(x,y, pch=15)
lines(x[order(x)], pom$fitted.values[order(x)],
      lwd=2, col="red")
```



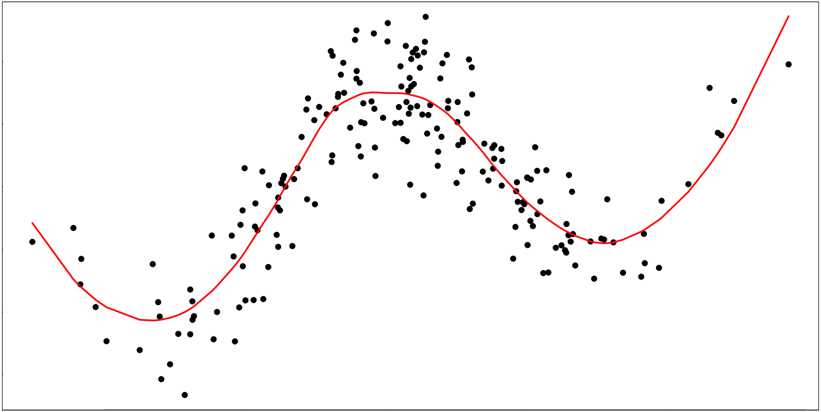
4th degree regression

```
x4 <- (x - mean(x))^4
pom <- lm(y ~ x + x2 + x3 + x4)
plot(x,y, pch=15)
lines(x[order(x)], pom$fitted.values[order(x)],
      lwd=2, col="red")
```



What is a smoother

- function that sumarizes relationship between dependent and independent variable
- values of the function follow the trend, but exibit less variation than the dependent variable

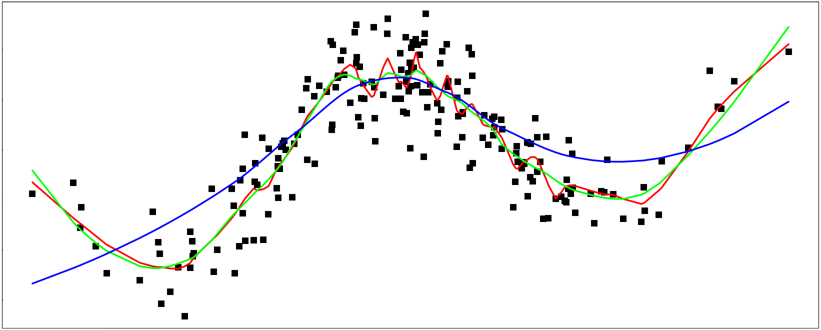


How smoothers smooth?

- Take a “window” from the range of values of the independent variable
- Choose a summary function (e.g. mean, linear regression, weighted mean . . .)
- Move the window across the range of independent variable
- Prediction for the center of the window is the chosen summary function
- Result is a smooth curve
- Wider window -> smoother curve
- Narrower window -> more wrigly curve

Degree of smoothness

```
pom1 <- loess(y~x, span=0.1)
pom2 <- loess(y~x, span=0.2)
pom3 <- loess(y~x, span=0.9)
plot(x,y, pch=15)
lines(x[order(x)], pom1$fitted[order(x)],
      lwd=2, col="red")
lines(x[order(x)], pom2$fitted[order(x)],
      lwd=2, col="green")
lines(x[order(x)], pom3$fitted[order(x)],
      lwd=2, col="blue")
```



From linear to additive models

Instead of using a linear combination of independent variables:

$$b_0 + \sum_i x_i b_i$$

use sum of smooth functions:

$$\sum_i s_i(x_i)$$

Thus:

$$y|x \sim F(\theta)$$

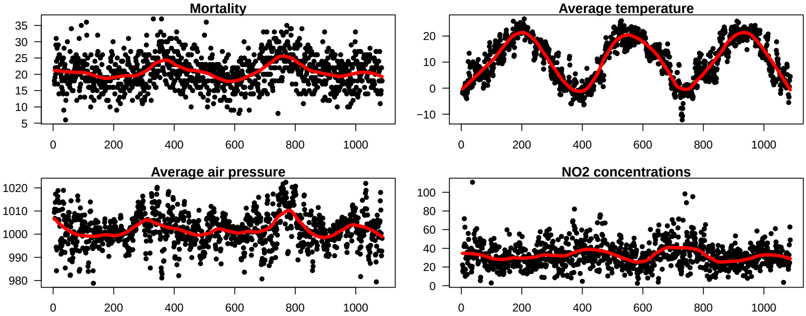
and

$$E(y|x) = g^{-1}\left(\sum_i s_i(x_i)\right)$$

Association between mortality and air pollution

Daily data on:

- number of deaths in the city of Zagreb in 1995 to 1997
- meteorological conditions (minimum, average, maximum of daily temperature, relative humidity, air pressure)
- common epidemics (cases of influenza)
- air pollution (concentrations of NO_x, SO_2, black smoke)

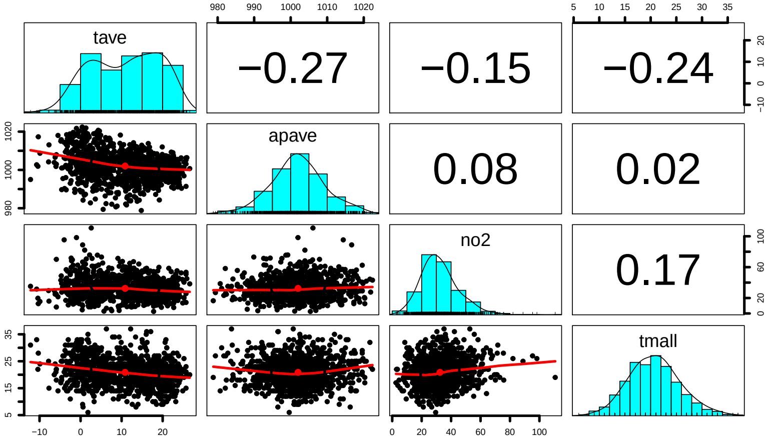


Some associations . . .

```r
require(psych)
```

```
## Loading required package: psych
```

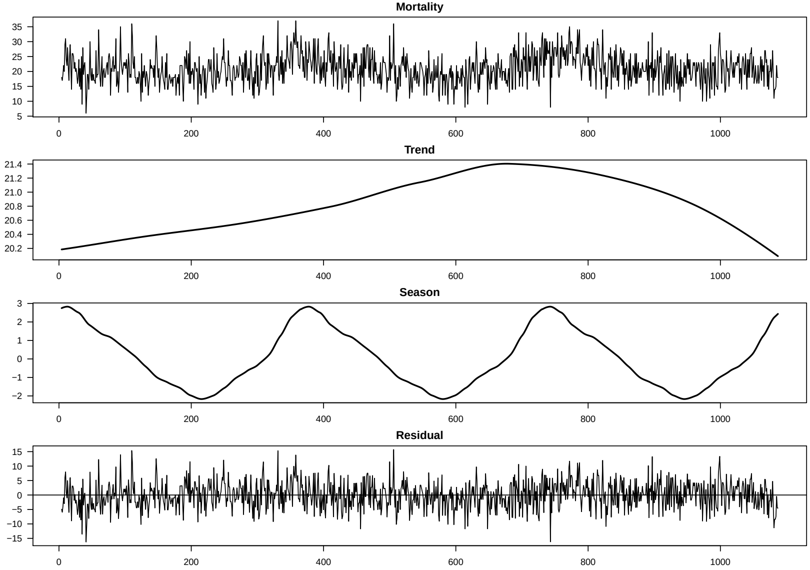
```r
pairs.panels(ts.podaci[,c("tave", "apave", "no2", "tmall")], lwd=3)
```



Approach to the analysis

- Poisson regression (outcomes are counts)
- Decomposition of the time series into trend, seasonal, and residual components (additive)
- Model association with air pollution, meteorological and epidemiological data for current and previous days with trend and seasonal components as offset

Trend and seasonality



Building the model

```
require(gam)
tmall.model<-gam(tmall ~ wday + s(dang) + s(rhmin) + s(tmax.l1) +
                   s(tmin.l2) + s(rhmin.l2) + s(apmax.l2) + s(no2),
                 offset=log(trend+season),
                 data=ts.podaci,
                 family=poisson(log))
```

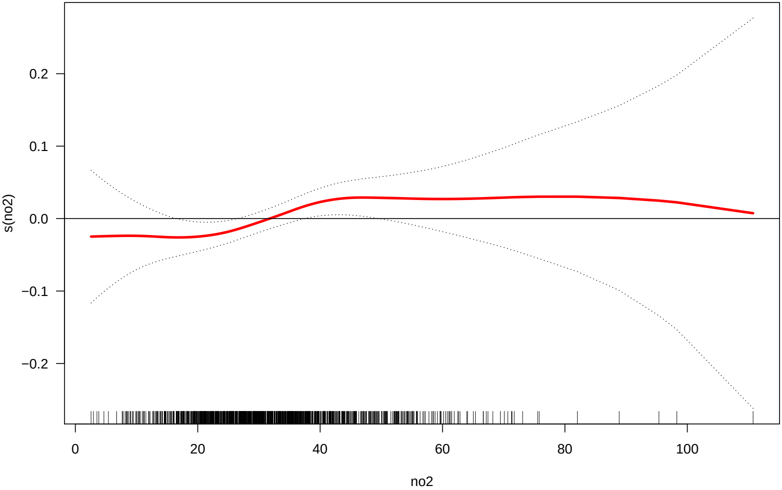
Model summary

```
## Anova for Parametric Effects
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## wday            6   16.66   2.777  2.7687  0.011248 *
## s(dang)         1    8.17   8.168  8.1426  0.004409 **
## s(rhmin)        1   10.09  10.093 10.0611  0.001558 **
## s(tmax.l1)      1   29.03  29.034 28.9438 9.187e-08 ***
## s(tmin.l2)      1   42.04  42.044 41.9130 1.463e-10 ***
## s(rhmin.l2)     1    7.60   7.596  7.5722  0.006030 **
## s(apmax.l2)     1    0.14   0.145  0.1445  0.703897
## s(no2)          1    5.36   5.365  5.3481  0.020938 *
## Residuals    1049 1052.28   1.003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

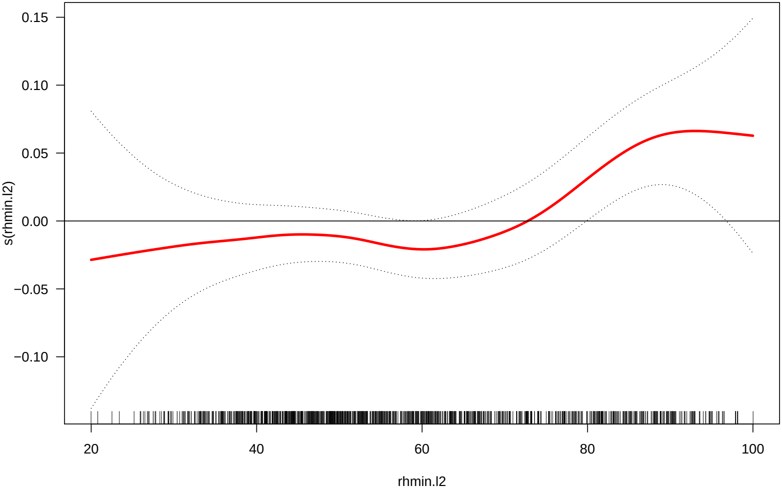
Model summary - continued

```
## Anova for Nonparametric Effects
##             Npar Df Npar Chisq    P(Chi)
## (Intercept)
## wday
## s(dang)           3    30.2234 1.239e-06 ***
## s(rhmin)          3     1.2769  0.734658
## s(tmax.l1)        3     2.7809  0.426628
## s(tmin.l2)        3     6.7324  0.080944 .
## s(rhmin.l2)       3     9.8330  0.020045 *
## s(apmax.l2)       3    11.4256  0.009635 **
## s(no2)            3     4.1430  0.246444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

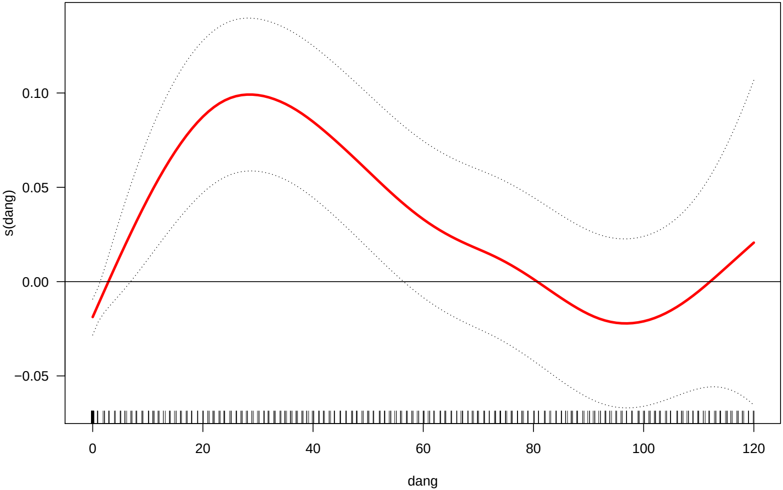
Mortality vs. NO2 - partial effect



Mortality vs. relative humidity



Mortality vs. day of influenza epidemic



Interpretation of results

- Regression coefficient in Poisson regression is a logarithm of the relative risk
- Exponential function will transform a coefficient into relative risk
- Range of effects in our model is ca. 0.4
- That transforms into relative risk of 1.4918247.

Conclusions

- Statistical models enable capturing shape of data distributions.
- Contemporary statistics provides a wide range of statistical models that can deal with:
  - Non-normality (generalized models)
  - Non-linearity (additive models)
- It is also possible to take into account dependence among observations (with mixed models) etc.
- Generalized additive (mixed) models provide a versatile tool for modeling wide range of outcomes that do not meet requirements of a linear model.

Questions?

Thank You!