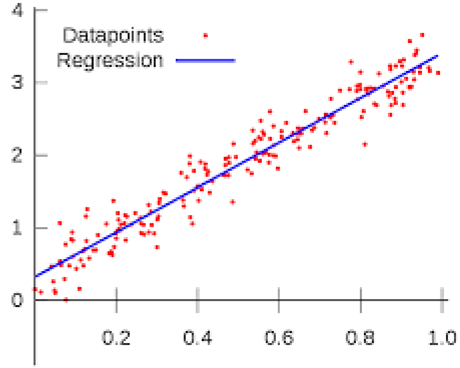
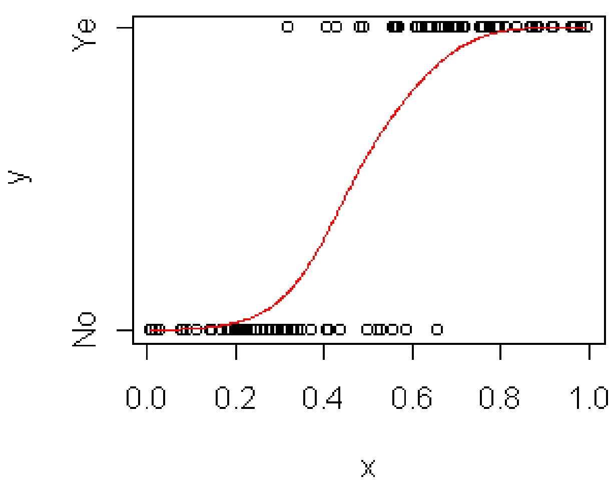
Nelineární vztah mezi x a y

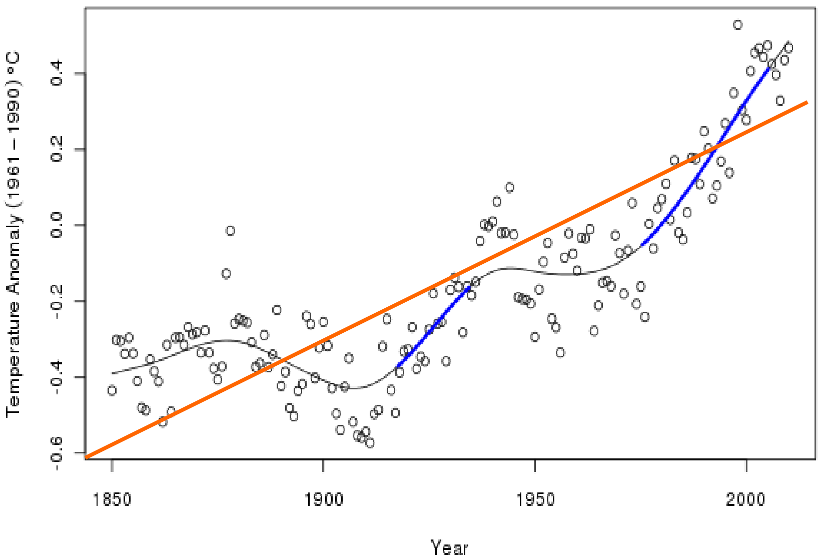
- Doposud jsem předpokládaly lineární vztah mezi vysvětlující a vysvětlovanou proměnnou:



- A nebo nelineární vztah, který je možno linearizovat transformací y:



V přírodě jsou komplexní nelineární vztahy mezi x a y naprosto běžné. Použití lineárního vztahu by mohlo vést k nesmyslným predikcím modelu a hodnotám testovací statistiky



Nelineární vztah mezi x a y

Analýza v R (nejčastější možnosti):

Aproximace pomocí polynomů

Piecewise Regression

Nelineární regrese

GAM – zobecněné aditivní modely

(1) Aproximace pomocí polynomů:

Lineární regrese:

$y = a + bx$

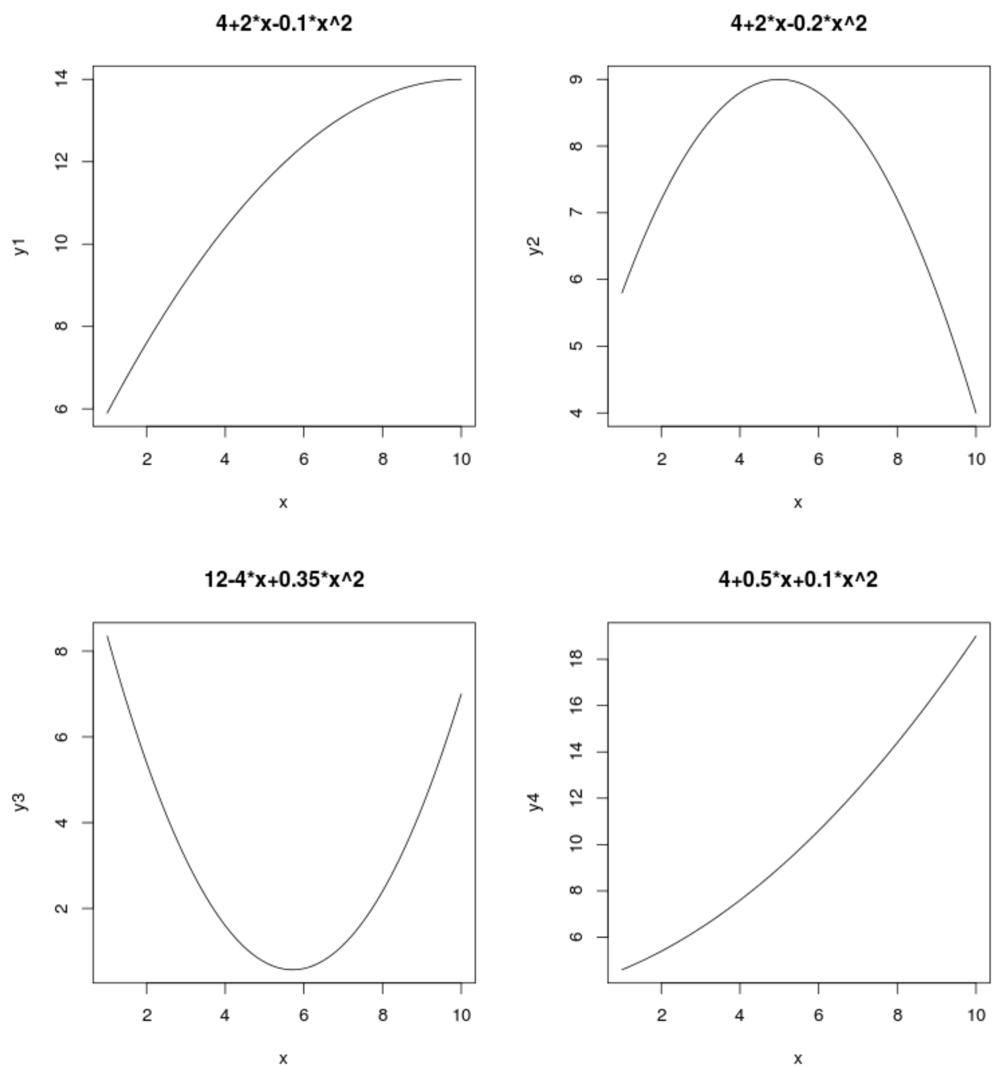
Polynomická regrese

$y = a + b_1x^1 + b_2x^2 \dots\dots\dots b_zx^z$

V praxi se používá většinou do polynomu třetího stupně (**kubický polynom:** $y = a + b_1x^1 + b_2x^2 + b_3x^3$)

Vhodný např. aproximaci růstových křivek

Příklady vztahů, které lze fitovat pomocí kvadratické polynomiální regrese:



(1) Aproximace pomocí polynomů:

Lineární regrese:

$y = a + bx$

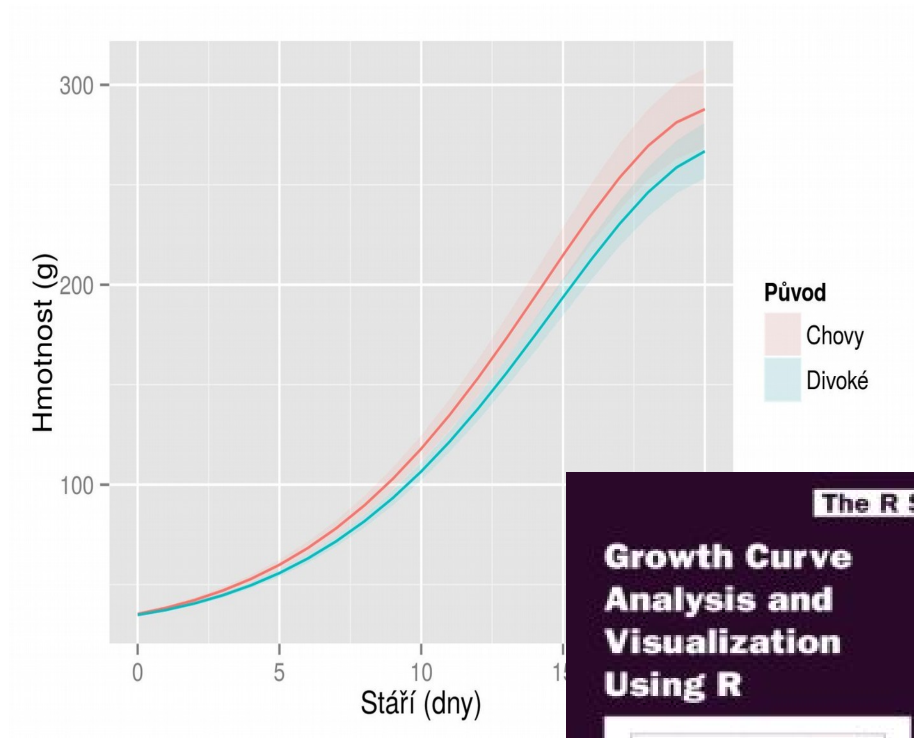
Polynomická regrese

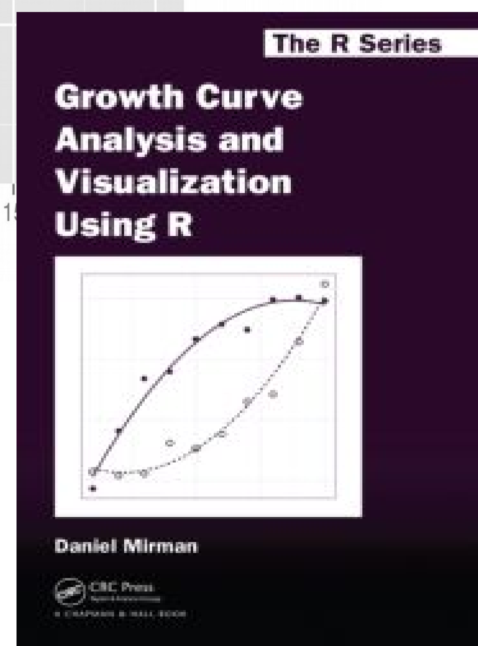
$y = a + b_1x^1 + b_2x^2 \ldots\ldots b_zx^z$

V praxi se používá většinou do polynomu třetího stupně (**kubický polynom:** $y = a + b_1x^1 + b_2x^2 + b_3x^3$)

Vhodný např. aproximaci růstových křivek

Příklad růstové křivky pomocí kubického polynomu:





(1) Aproximace pomocí polynomů:

Výpočet v R:

•Chceme zjistit, jestli naše data fitují líp polynomický než lineární model

- V R předem definujeme stupeň polynomu”

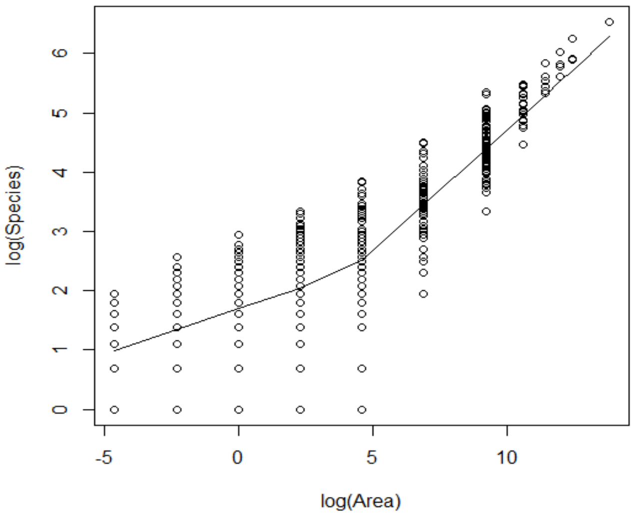
```
(a): a~poly(b,2)
(b): a~ b + I(b^2))
```

- druhá možnost pro nás obvykle vhodnější (typ paramerizace na kterou jsme zvyklí), naproti tomu první možnost je vhodnější z výpočetního hlediska
- Optimalizace odhadu parametrů na základě minimalizace sumy čtverců
- Porovnání jednodušších a složitějších modelů pomocí delečních testů (postupně odstraňujeme polynomy vyšších rádů)

```
model1<-lm(x~poly(y,2))
model2<-lm(x~y)
anova(model1,model2)
```

(2) Piecewise regrese

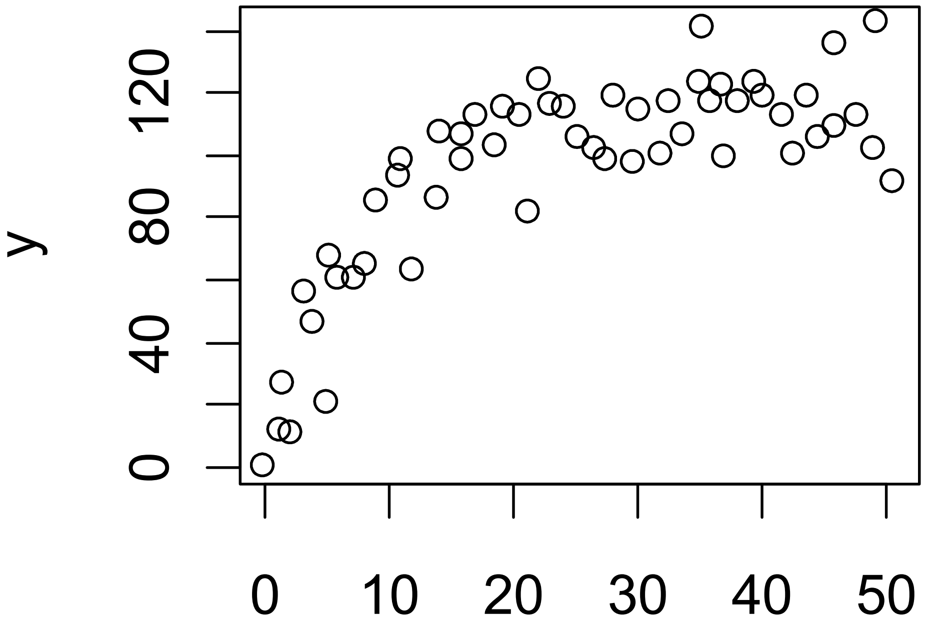
- Fitujeme rozdílné regresní rovnice pro různé oblasti dat



- Jak najít optimální break point?
  1) zkoušíme ho umístit v různých hodnotách pro x
  2) počítáme devianci pro jednotlivé modely
  3) vybereme model s optimální hodnotou BP

(3) Explicitně definované nelineární fce. pomocí nls

- Někdy ale máme explicitní funkci (a nebo sérii funkcí) které chceme fitovat na data a porovnávat, které líp fituje data



V R příkaz nls, musí se zadat hrubé odhady parametrů, které slouží jako počáteční hodnoty v následných iteracích

```
nls(y~a-b*exp(-c*x),start=list(a=100,b=90,c=0.15))

Nebo specifikovat pomocí tzv. Self starting funkcí
```

Nelineární regrese

- Pokud použijeme tzv. Self starting funkce tak nemusíme zadávat počáteční odhady parametrů

```
model<-nls(rate~SSmicmen(conc,a,b))
```

Table 20.1. Useful non-linear functions.

Name	**Equation**
Asymptotic functions	
Michaelis–Menten	$y = \dfrac{ax}{1+bx}$
2-parameter asymptotic exponential	$y = a(1-\mathrm{e}^{-bx})$
3-parameter asymptotic exponential	$y = a - b\mathrm{e}^{-cx}$
S-shaped functions	
2-parameter logistic	$y = \dfrac{\mathrm{e}^{a+bx}}{1+\mathrm{e}^{a+bx}}$
3-paramerter logistic	$y = \dfrac{a}{1+b\mathrm{e}^{-cx}}$
4-parameter logistic	$y = a + \dfrac{b-a}{1+\mathrm{e}^{(c-x)/d}}$
Weibull	$y = a - b\mathrm{e}^{-(cx^d)}$
Gompertz	$y = a\mathrm{e}^{-b\mathrm{e}^{-cx}}$
Humped curves	
Ricker curve	$y = ax\mathrm{e}^{-bx}$
First-order compartment	$y = k\exp(-\exp(a)x) - \exp(-\exp(b)x)$
Bell-shaped	$y = a\exp(-\lvert bx\rvert^2)$
Biexponential	$y = a\mathrm{e}^{bx} - c\mathrm{e}^{-dx}$

Generalized Additive Models

Co to je?

Jak to funguje?

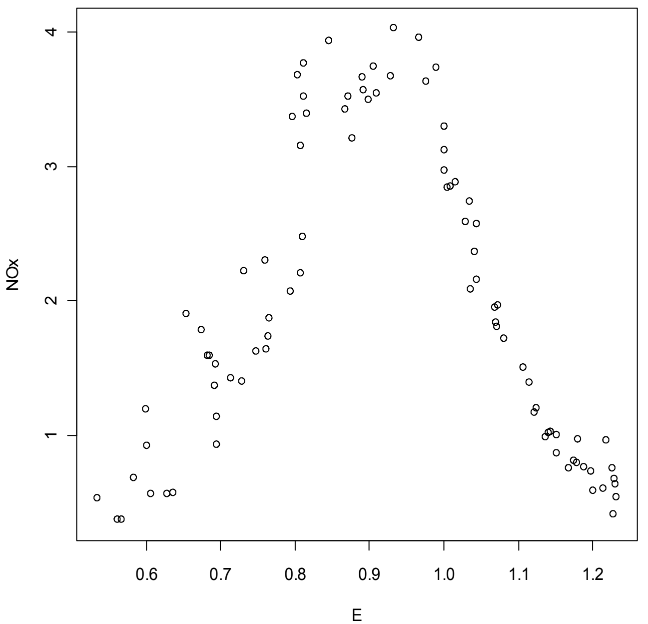
Jak to používat (v R)?

GAM – Generalized Additive Models

Často netušíme podle jaké matematické funce by se měla naše data chovat

Nebo je vztah mezi vysvětlující a vysvětlovanou proměnnou tak komplexní, že nejsme schopni odhadnout odpovídající rovnici

V takových případech je dobré použít GAM

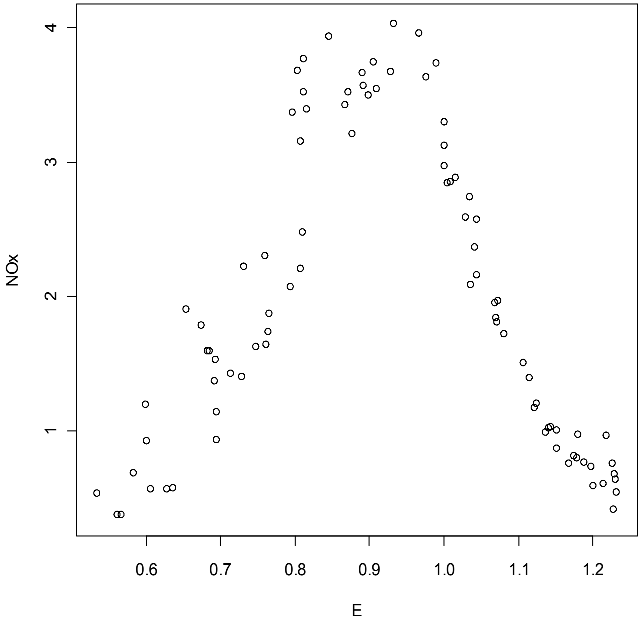


GAM – Generalized Additive Models

Neparametrická metoda – cílem není přesně odhadnout funkci a její parametry

Sofistikovaná obdoba piecewise regression

- Data se rozdělí na optimální počet segmentů
- Pro jednotlivé segmenty samostatný fit
- Jednotlivé fity se spojí dohromady v jednolitou funkci



Generalized Additive Models: Na co to je?

V lineárních modelech je vztah mezi vysvětlovanou a vysvětlující proměnnou popsán např. touto rovnicí

$$Y = \alpha + \beta X + \varepsilon$$

V GAM se vztah fituje pomocí neparametrické vyhlazovací funkce a žádnou rovnici tudíž nedostanem (to neznamená ale, že není možné dělat predikce...)

Zobecněné aditivní modely - GAM

Vyhlazovacích funkcí existuje celá řada nejjednodušší je LOESS vyhlazování

Princip LOESS

- lokální regrese prováděná na rámečku
- Dostaneme predikovanou hodnotu pro hloubku 1500
- Potom posuneme okénko o kus dál a celý proces opakujeme
- Pomocí span se dá specifikovat jaký podíl dat bude v okénku

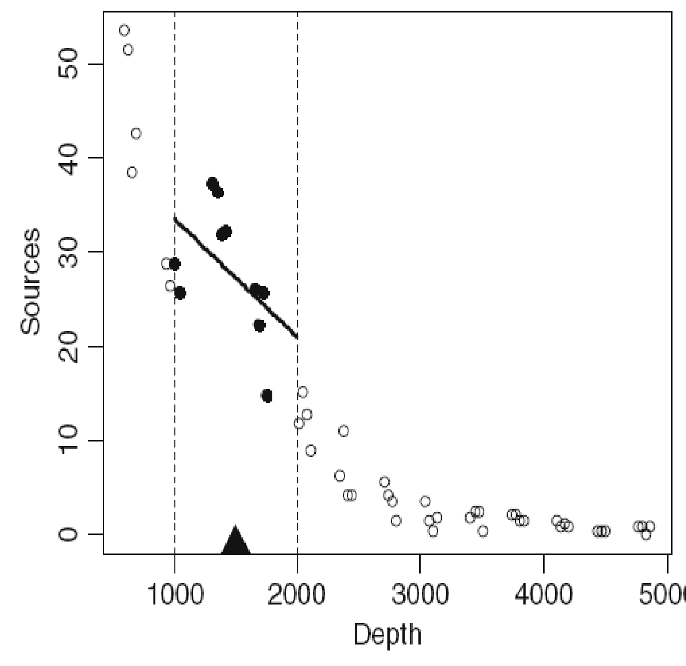


Fig. 3.2 Illustration of LOESS smoothing. We want to predict a value of sources at the target value of 1500 m depth (denoted by the *triangle*). A window around this target value is chosen (denoted by the *dotted lines*) and all points in this window (the *black dots*) are used in the local linear regression analysis

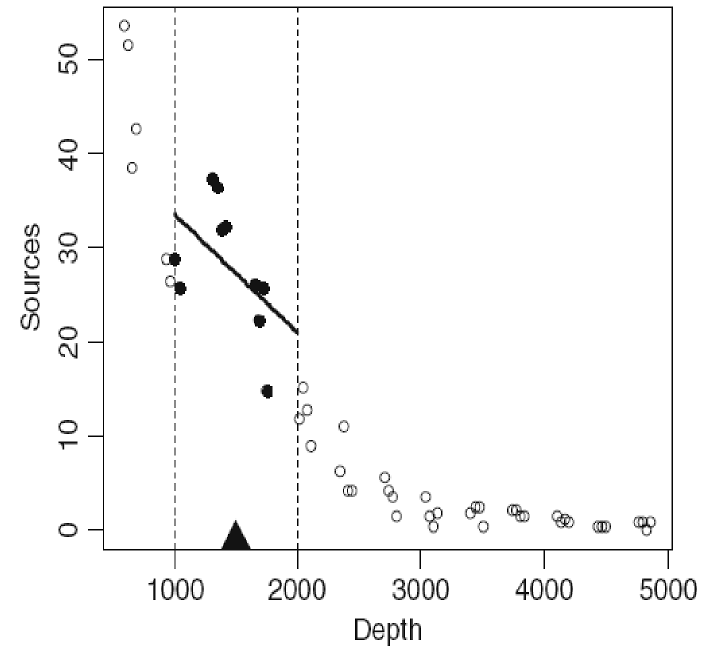
Zobecněné aditivní modely - GAM

Vyhlazovacích funkcí existuje celá řada nejjednodušší je LOESS vyhlazování

Princip LOESS

– Pokud bude mít span malou hodnotu tak tak náš výsledný model bude příliš složitý

Fig. 3.2 Illustration of LOESS smoothing. We want to predict a value of sources at the target value of 1500 m depth (denoted by the *triangle*). A window around this target value is chosen (denoted by the *dotted lines*) and all points in this window (the *black dots*) are used in the local linear regression analysis

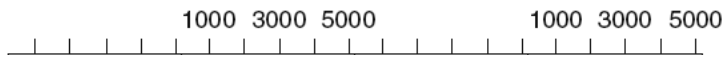


Zobecněné aditivní modely - GAM

Sofistikovanější typy smootherů – **balíček mcgv**

Pracují většinou na principu kubické regrese uvnitř jednotlivých segmentů dat.

Algoritmy k optimalizaci počtu segmentů a složitosti daných fitů



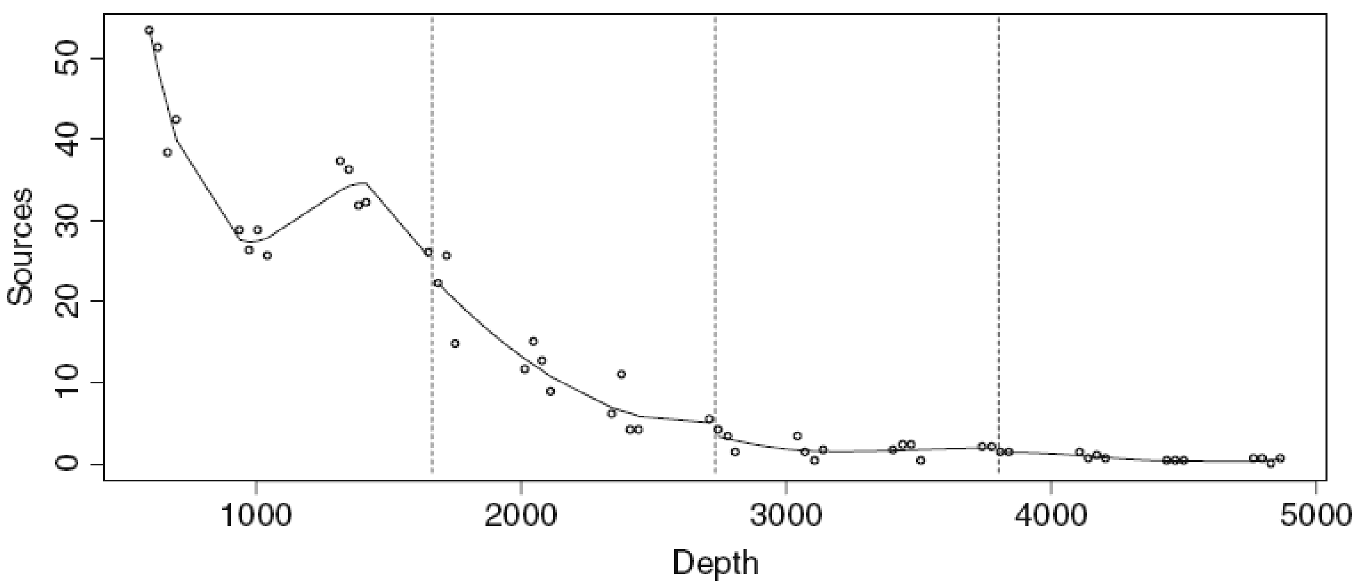


Fig. 3.8 Illustration of fitting a cubic polynomial on four segments of data using the ISIT data from station 19. We arbitrarily choose four segments along the depth gradient. The dotted lines mark these segments, and the line in each segment is the fit from the cubic polynomial model. R code to create this graph can be found on the book website

Zobecněné aditivní modely - GAM

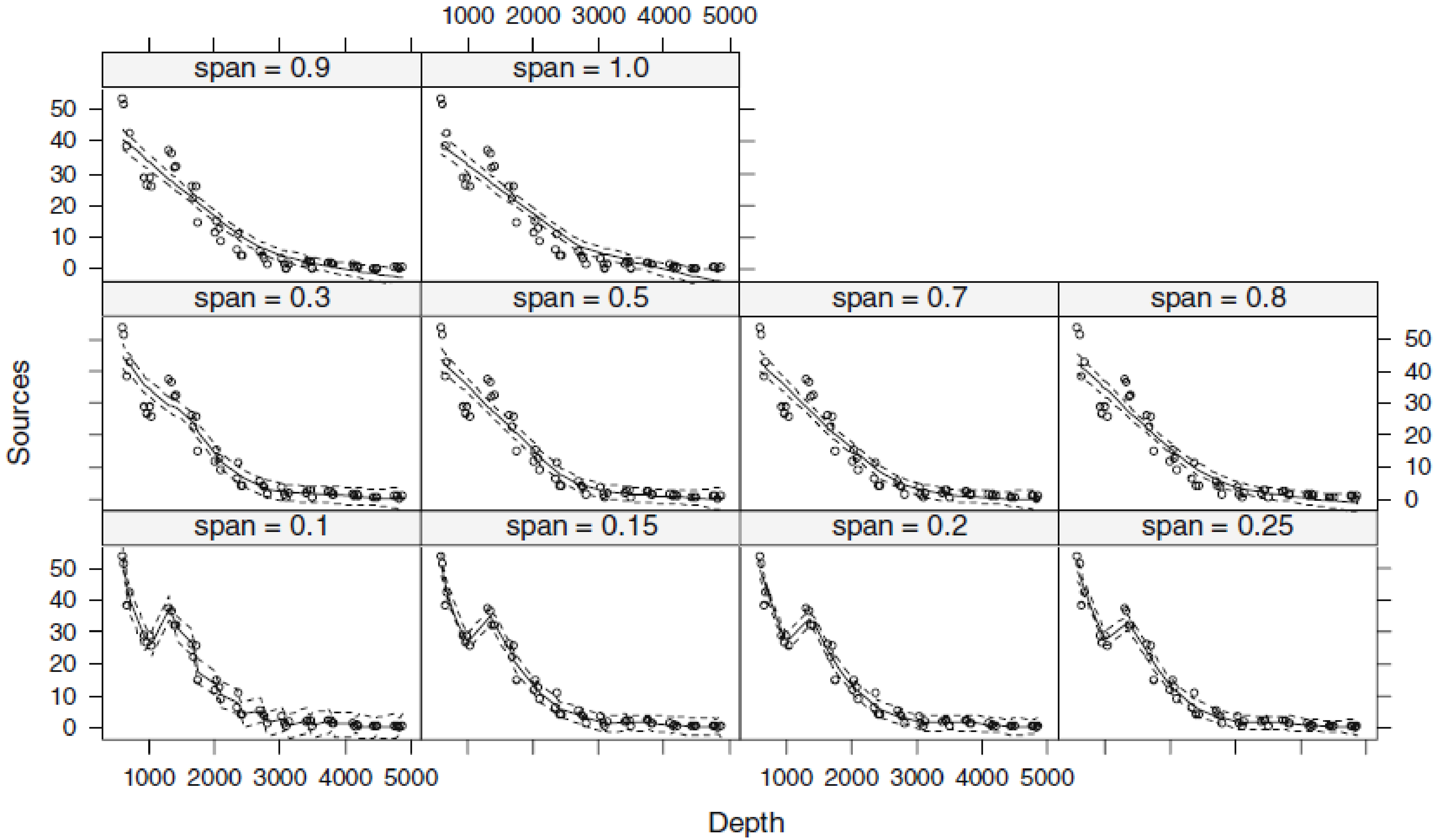


Fig. 3.3 LOESS smoothers for different span values. The solid line is the LOESS smoother and the dotted lines are 95% confidence bands. A visual inspection of the fitted lines and observed values indicates that a span of 0.2 seems to be optimal. The R code to make this graph is about two pages and is therefore presented on the book website

Zobecněné aditivní modely - GAM

V R většinou automatický odhad komplexity křivky

můžeme fitovat pomocí balíčků

gam – jednoduchý intuitivně pochopitelný typ vyhlazování křivky – LOESS smoothing - **lokální regrese**

Odhad SPAN manuálně (okometricky, AIC)

mgcv – Používá většinou složitější typ vyhlazování (Penalised smoothing spines)

Umožňuje počítat složitější typy příkladů, náhodné efekty, časová a prostorová autokorelace apod.

Automatický odhad SPAN (kompexita křivky)

Zobecněné aditivní modely - GAM

Zápis gam v R (balíček mgcv):

Vztah mezi vysvětlovanou a všemi vysvětlujícími proměnnými je fitován pomocí neparametrické vyhlazovací funkce:

gam(y~s(x) + s(y) + s(z))

Pouze pro x je fitován nelineární vztah u ostatních proměnných je lineární:

gam(y~s(x) + y + z)

Takto specifikujeme, že nás zajímá dvourozměrný vztah mezi dvěma proměnnýma („interakce“)

gam(y~s(x) + s(y,z))

Typ rozdělení dat specifikujeme tak jak jsme na to zvyklí z GLM:

gam(y~s(x) + s(y,z),falmily=Poisson)

Zobecněné aditivní modely - GAM

V R (balíček mgcv):

```
M3 <- gam(Sources16 ~ s(Depth16, fx = FALSE, k=-1,
          bs = "cr"))
```

Specifikuje typ smootheru

LOESS vs. spine regression

s – specifikuje které proměnné

budou „vyhlazovány“

Míra smoothingu není fixně nastavená

ale odhaduje se na základě cross-validace

GAM se dají aplikovat na různé typy rozdělení dat, gaussovské binomické, poisson apod.

Zobecněné aditivní modely - GAM

V R (balíček mgcv):

Všechny proměnné mohou být v GAM fitovány nelineárně:

```
y~s(x)+s(z)
```

Nebo jen některé z nich

```
y~s(x)+z
```

GAM se dají aplikovat na různé typy rozdělení dat, gaussovské binomické, poisson apod.

Zobecněné aditivní modely - GAM

V R (balíček mgcv):

```
Family: gaussian
Link function: identity

Formula:
ozone ~ s(rad) + s(temp) + s(wind)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    42.10       1.66   25.36   <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(rad)  2.763  3.451  3.964  0.00736 **
s(temp) 3.841  4.762 11.612 1.28e-08 ***
s(wind) 2.918  3.666 13.770 1.53e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.724   Deviance explained = 74.8%
GCV score =    338  Scale est. = 305.96    n = 111
```