Generalized Additive Models

Dealing with non-linearity

Not everything is linear

Linearity in the parameters : $Y = \alpha + \beta_1 Z$

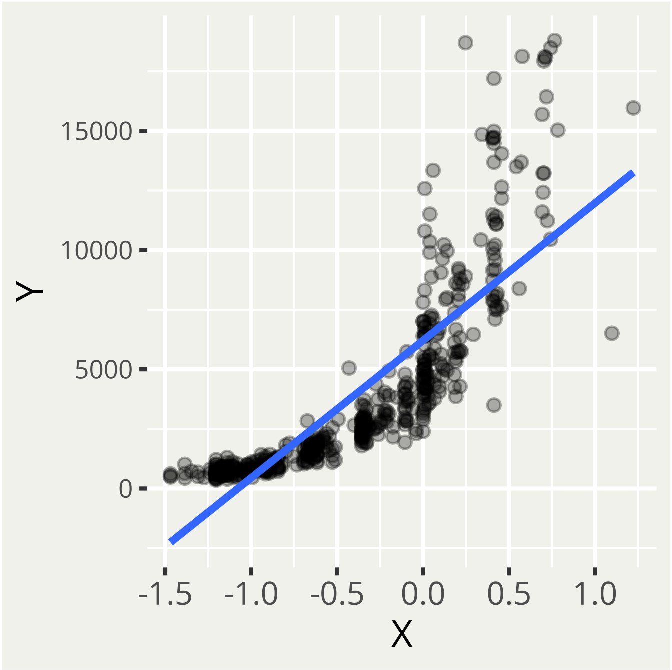
$$Y = \alpha + \beta_1 X + \beta_2 X^2 \qquad\qquad Y = \alpha + \beta_1 (XW)$$

$$Y = \alpha + \beta_1 \log(X) \qquad\qquad Y = \alpha + \beta_1 \exp(X)$$

However : $Y = \alpha + \beta_1 X_1 e^{\beta_2 X_2 + \beta_3 X_{3i}}$ is not linear

Not everything is linear

Inappropriate (Generalized) Linear Regression



What to do with non-linearity?

Include interactions

Include quadratic effect : $\alpha + \beta_1 X + \beta_2 X^2$

More explanatory variables

Transform to linearise (avoid)

Use a smoother

Smoothing methods

Beyond linearity

Polynomial regression

Step function, segmented / piecewise regression

Splines / smoothing

Generalized Additive Model

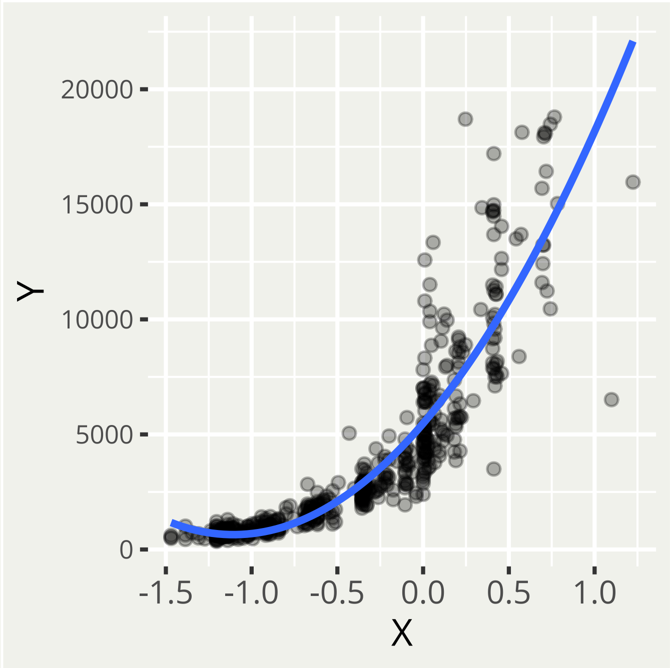
Polynomial regression

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

$$\eta = \mu = \beta_0 + \beta_1 x + \beta_2 x^2$$

Still linear model

Imposes a global structure



In R :

```
m1 <- glm(y ~ poly(x, 2), data = dd)
```

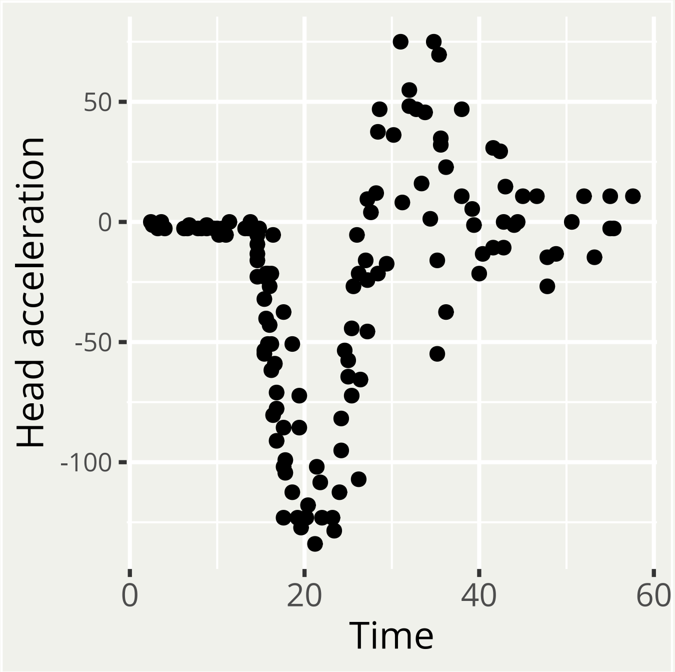
Non-linear regression

Difficult to specify

Reason for cubic polynomial?

In R :

```
stats::nls()
nlme::nlme()
```



1970's US census with polynomial : predicted a population crash in 2015!

Segmented / piecewise regression

Break range in bins

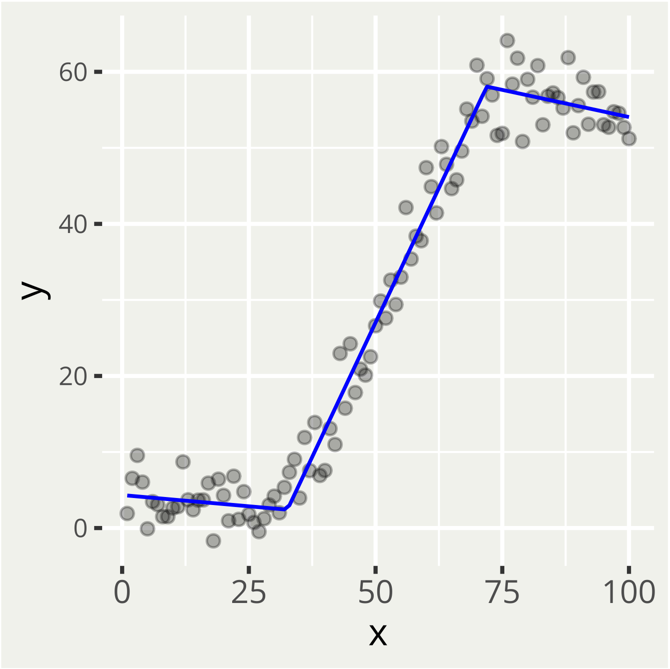
Fit LM in each bin

```r
library(segmented)

# Usual (G)LM
out_lm <- lm(y ~ x, data = dati)

# Get the segments
o <- segmented(out_lm,
               seg.Z = ~ x,
               psi = list(x = c(30,60)))

# Pass to plotting function...
```



Splines / Smoothing

Split X in regions (bins)

Fit low-degree polynomial on each region of X

Compute fit at target x_0 with nearby observations

Possibly : smoothness penalty

Stable estimates, more flexible

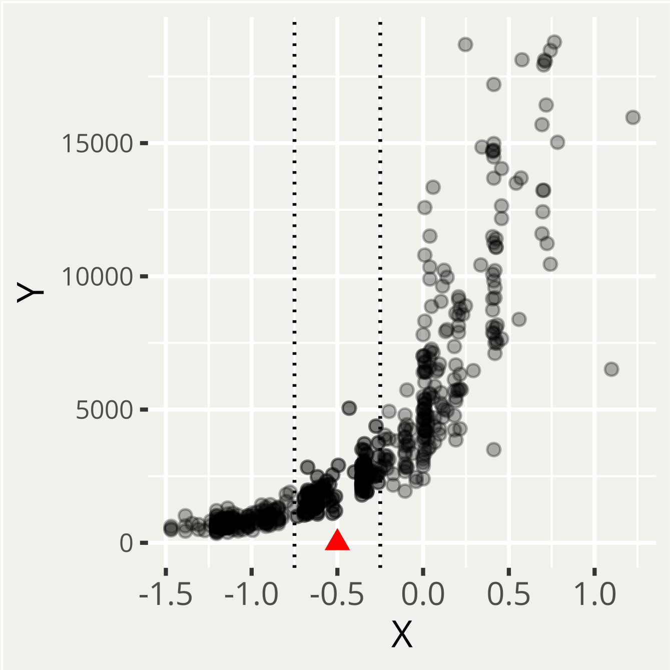
Splines : LOESS smoothing

Target value x

Window around target

Value of y at target x?

Mean / median



Splines : LOESS smoothing

Value of y at target x?

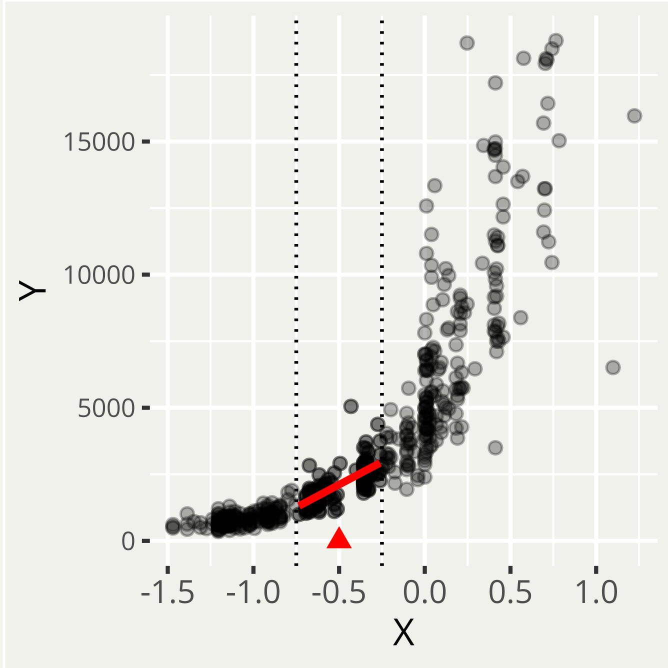
(Weighted) linear regression

Predict y from regression

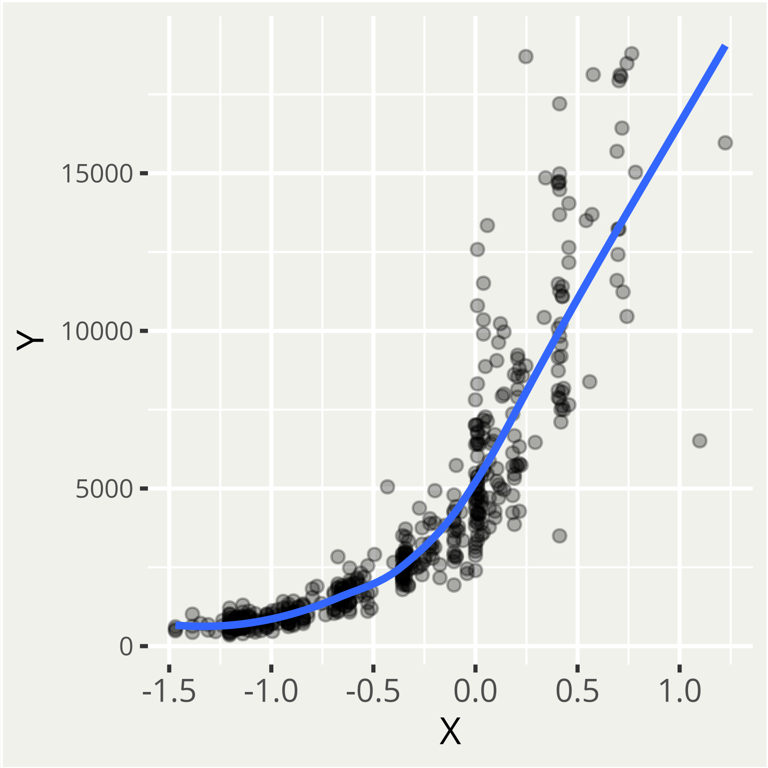
Repeat for sequence of x

In R :

```
loess(dist ~ speed,
  data = cars,
  span = 0.75, # degree of smoothing
  degree = 1   # linear / polynomial
  )
```



Splines : LOESS smoothing



Moving a window along X

- What happens to the edges?
  - Not much to do...
  - Be careful with interpretation
- What window size?
  - Trial and error : variance/bias trade-off
  - Residuals graph : pattern $\rightarrow$ increase width
  - Compare models (*e.g.* with AIC)

Generalized Additive Models

Generalized Additive Models

Extension of GLMs

Allow non-linear function for each predictor

Maintain additivity

For quantitative and qualitative variables

Incorporating non-linear predictors

GLM : $g(E(Y)) = \beta_0 + \beta_i x_i + \ldots$

Replace $\beta_i x_i$ with non-parametric function :

$f_i(x_i)$: non-linear *smooth function* of covariates

LOESS, LOWESS, cubic splines *etc.*

Estimate each f_i and *add* them up

GAM : $g(E(Y)) = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip})$

GAMs in R

Function `gam::gam()`

- Local regression and smoothing splines

Function `mgcv::gam()`

- Penalized regression splines / smoothers
- More technical (derivatives) but more flexible
- Cross-validation (GCV): automatic selection of smoothing parameters

Large collection of smoothers

- Thin plate smoother
  - Generally quite good (not for large data sets)
  - Default in R
- Cyclic cubic regression spline
  - Far left = far right *e.g.* X is *month*
- Shrinkage smoothers
  - Smoothing "shrinks" towards 0 $\rightarrow$ variable selection
- In practice, small differences...

GAM Smoothers : Cubic Spline

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

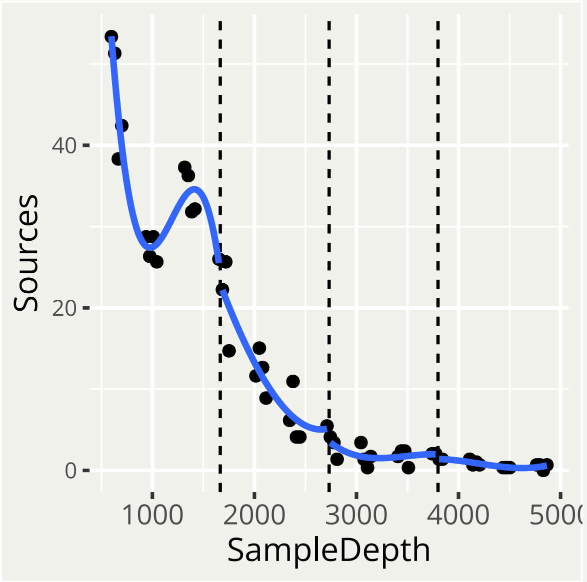
$$\mu = \alpha + f(X_i)$$

Divide into intervals

Cubic polynomial on each interval

Fitted values $\rightarrow$ smoothing curve

Conditions for smooth connections



More on splines

More knots $\rightarrow$ less smooth

Find optimal # of knots visually or AIC

General recommendation :

- < 50 observations : 3 knots
- > 100 observations : 5 knots

GAM smoothers : P-Spline

Approximate $f(z)$ with polynomial spline and many knots

Introduce *penalty* on wiggliness

Penalized likelihood

Avoids overfitting

Don't accept cross-validation blindly...

GAMs with multiple predictors

Different smoothers for different variables

$$g(\mu) = \alpha + f_1(X_i) + f_2(Z_i)$$

Hybrid models

$$g(\mu) = \alpha + f_1(X_i) + \beta Z_i + factor(W_i)$$

Random effects *etc.*

Some problems (may) remain

- Violation of independence
- Heterogeneity of variance
- Collinearity (concurvity)
- Nested data

Advantages of GAMs

Non-linear...

Very flexible

Additive :

- Effect of X_i while holding other predictors constant
- Interpretable

Extension with random effects *etc.*

Disadvantages of GAMs

Beware of over-fitting

Is added complexity (generality) necessary?

May be a bit "tricky"...

Going (a bit) further

Generalized Additive Models - Michael Clark