

Why mgcv is awesome

✉ chris.maine@uhb.nhs.uk

🌐 mainard.co.uk

🐙 github.com/chrismainey

🐦 twitter.com/chrismainey

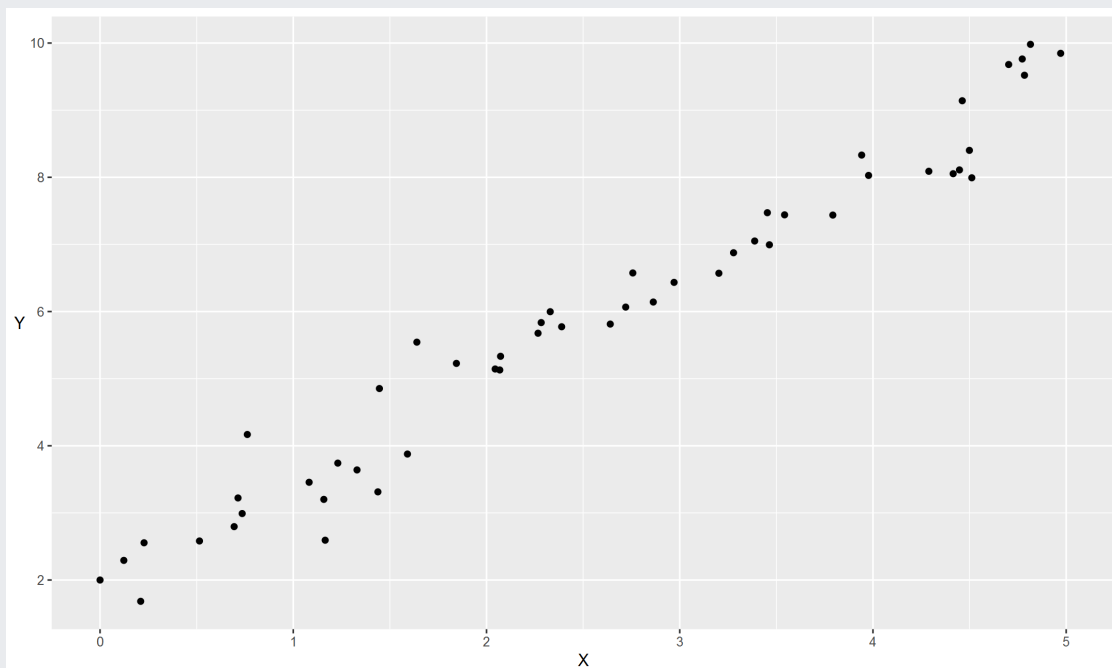


Don't think about it too hard... 😊

1 / 19

Regression models on non-linear data

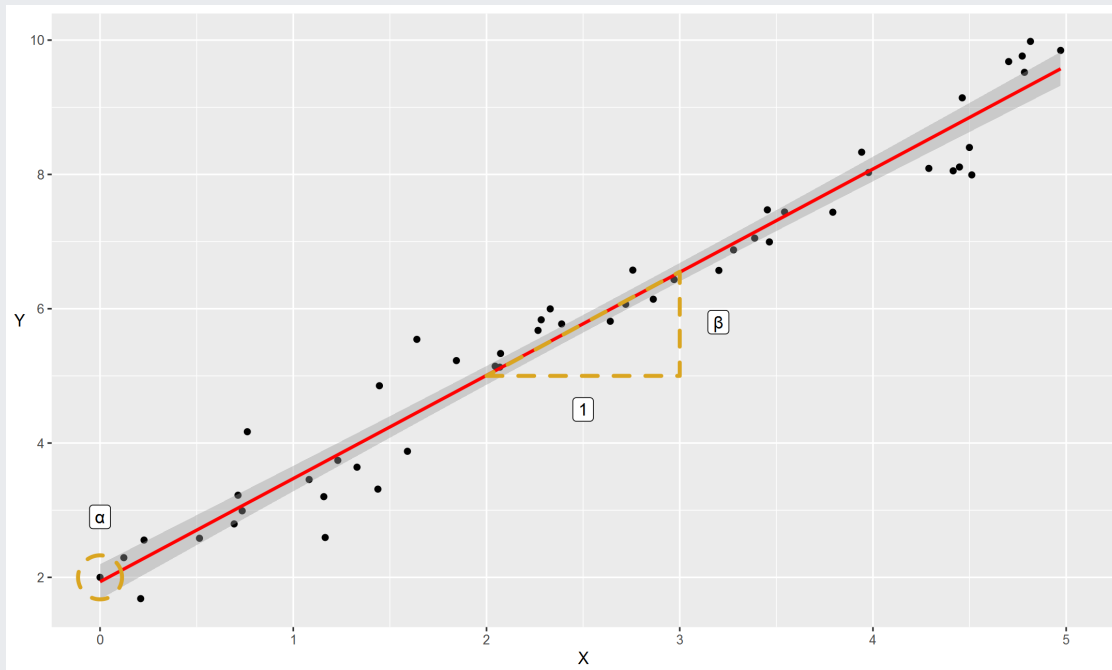
- Regression is a method for predicting a variable, Y , using another, X



2 / 19

Equation of a straight line (1)

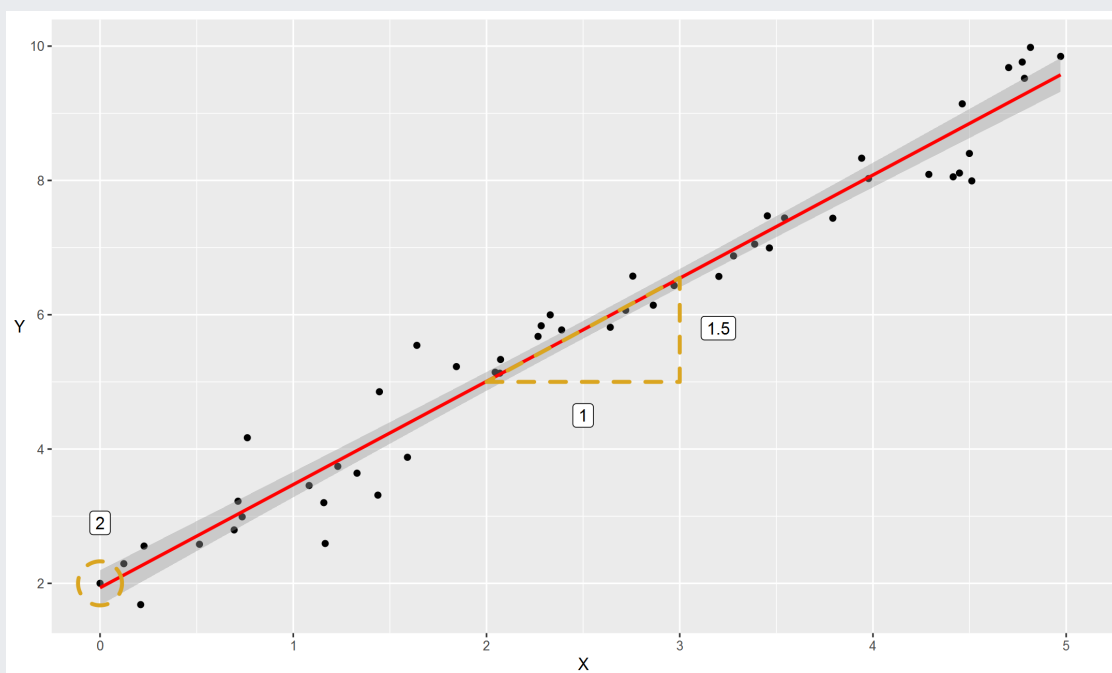
$$y = \alpha + \beta x + \epsilon$$



3 / 19

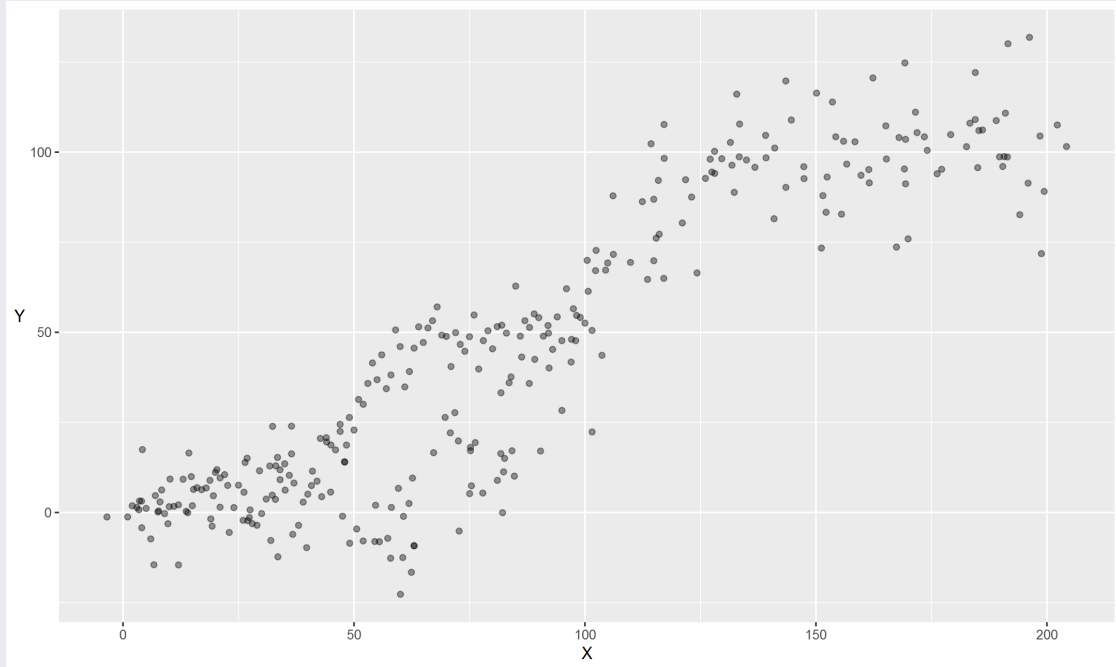
Equation of a straight line (2)

$$y = 2 + 1.5x + \epsilon$$



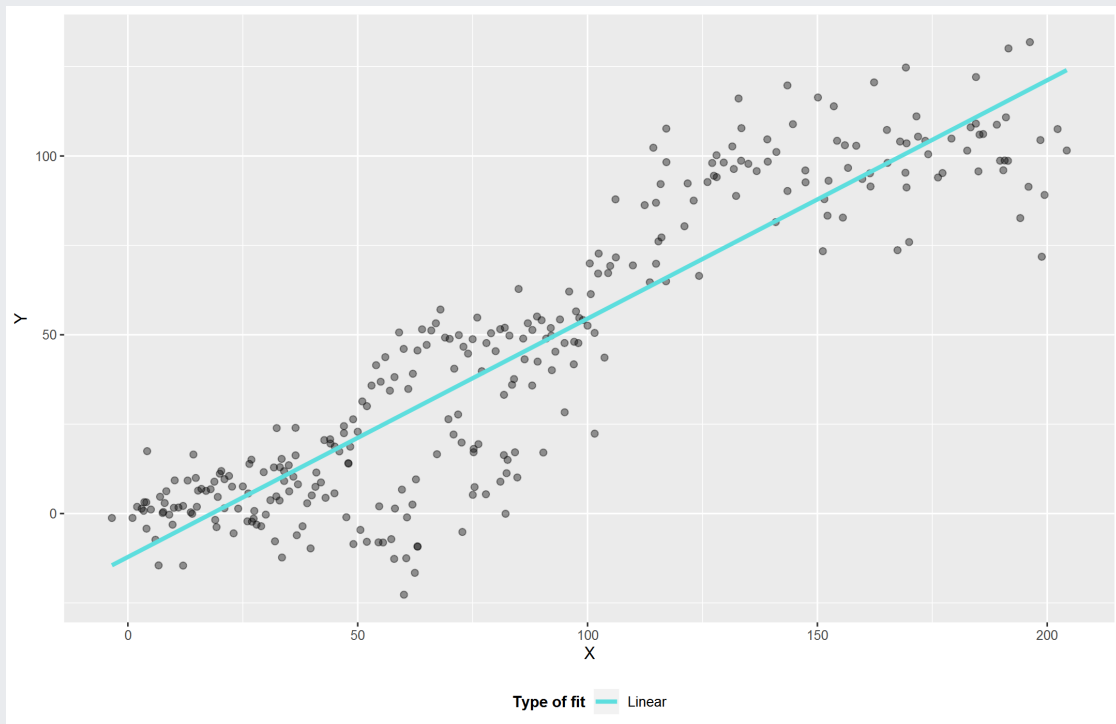
4 / 19

What about nonlinear data? (1)



5 / 19

What about nonlinear data? (2)



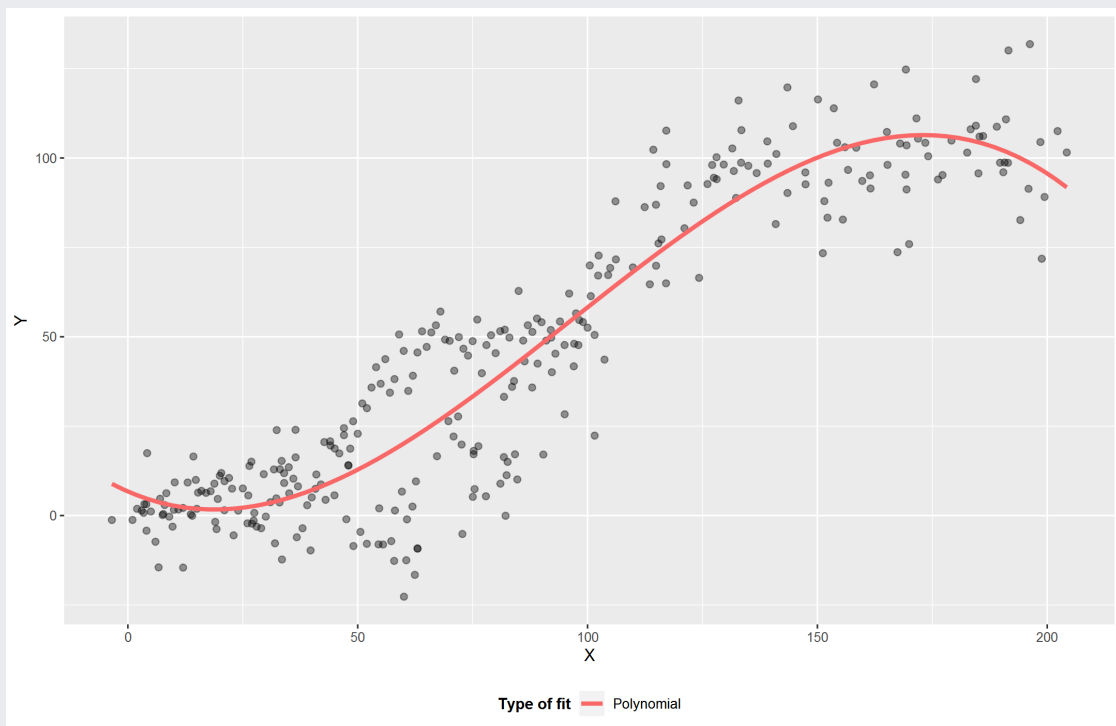
6 / 19

What about nonlinear data? (3)



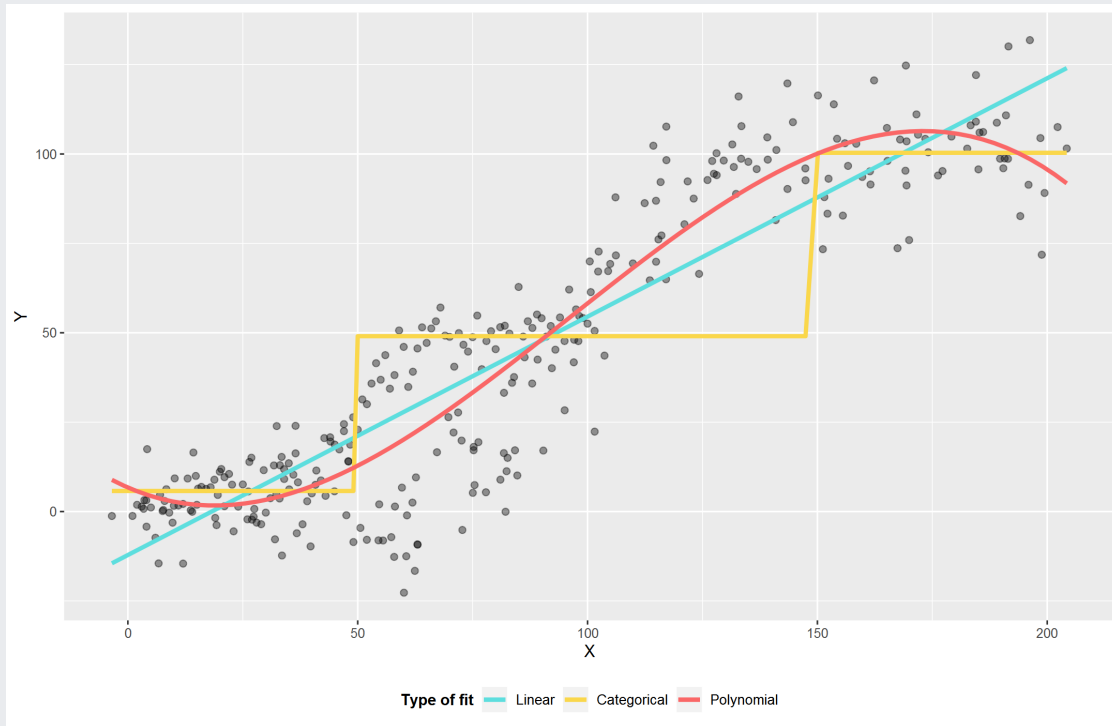
7 / 19

What about nonlinear data? (4)



8 / 19

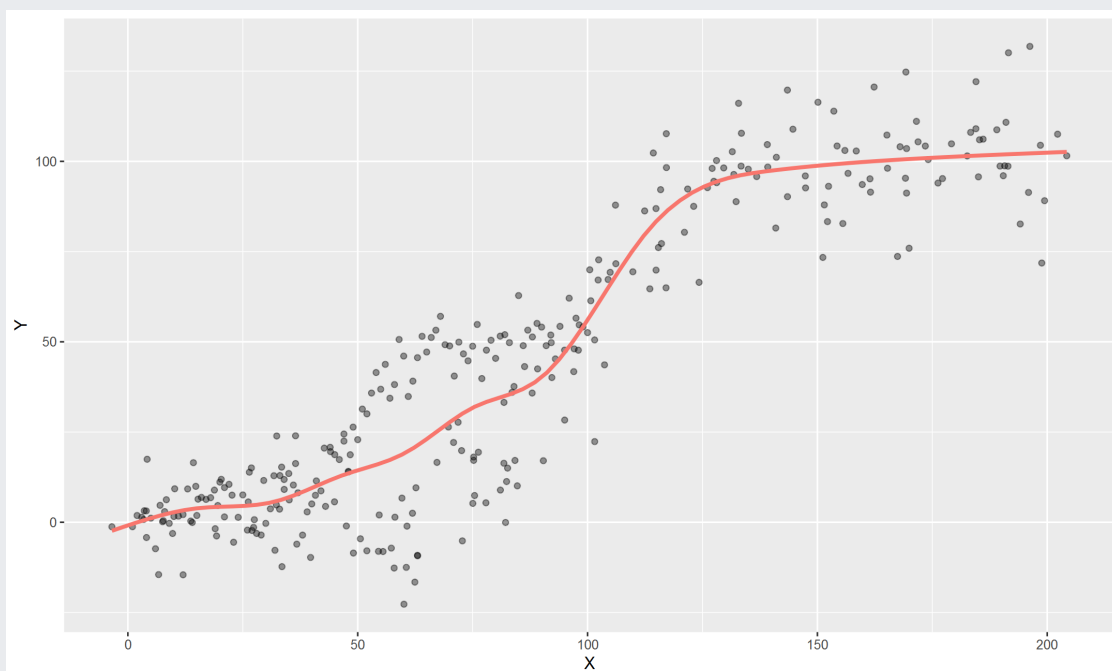
What about nonlinear data? (5)



9 / 19

Splines

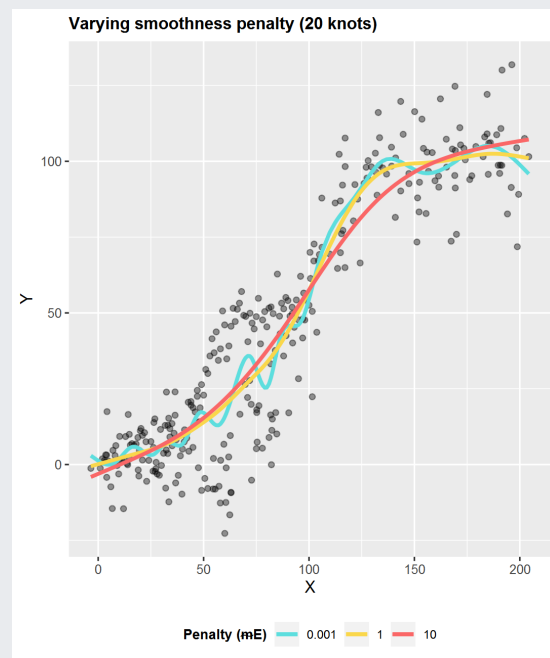
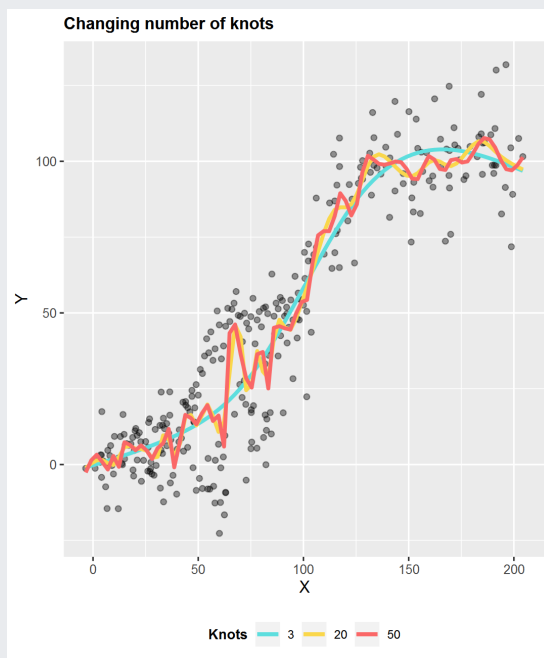
- Smooth, piece-wise polynomials, like a flexible strip for drawing curves.
- 'Knot points' between each section



10 / 19

How smooth?

Can be controlled by number of knots (k), or by a penalty γ .



11 / 19

Generalized Additive Model

- Regression models where we fit smoothers (like splines) from our data.
- Strictly additive, but smoothers can describe complex relationships.
- In our case:

$$y = \alpha + f(x) + \epsilon$$

Or more formally, an example GAM might be (Wood, 2017):

$$g(\mu_i) = A_i\theta + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4) + \dots$$

Where:

- $\mu_i \equiv E(Y_i)$, the expectation of Y
- $Y_i \sim EF(\mu_i, \phi_i)$, Y_i is a response variable, distributed according to exponential family distribution with mean μ_i and shape parameter ϕ .
- A_i is a row of the model matrix for any strictly parametric model components with θ the corresponding parameter vector.
- f_i are smooth functions of the covariates, x_k , where k is each function basis.

12 / 19

What does that mean for me?

- Can build regression models with smoothers
- Suited to non-linear, or noisy data
- *Hastie (1985)* used knot every point, *Wood (2017)* uses reduced-rank version

mgcv: mixed gam computation vehicle

- Prof. Simon Wood's package, pretty much the standard
- Included in standard R distribution, used in `ggplot2` `geom_smooth` etc.

```
library(mgcv)
my_gam <- gam(Y ~ s(X, bs="cr"), data=dt)
```

- `s()` control smoothers
- `bs="cr"` telling it to use cubic regression spline ('basis')
- Default is 10 knots (`k=10` argument), but you can alter this

13 / 19

Model Output:

```
summary(my_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Y ~ s(X, bs = "cr")
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.9659    0.8305   52.94   <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(X)  6.087   7.143 296.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.876   Deviance explained = 87.9%
## GCV = 211.94   Scale est. = 206.93      n = 300
```

14 / 19

Check your model:

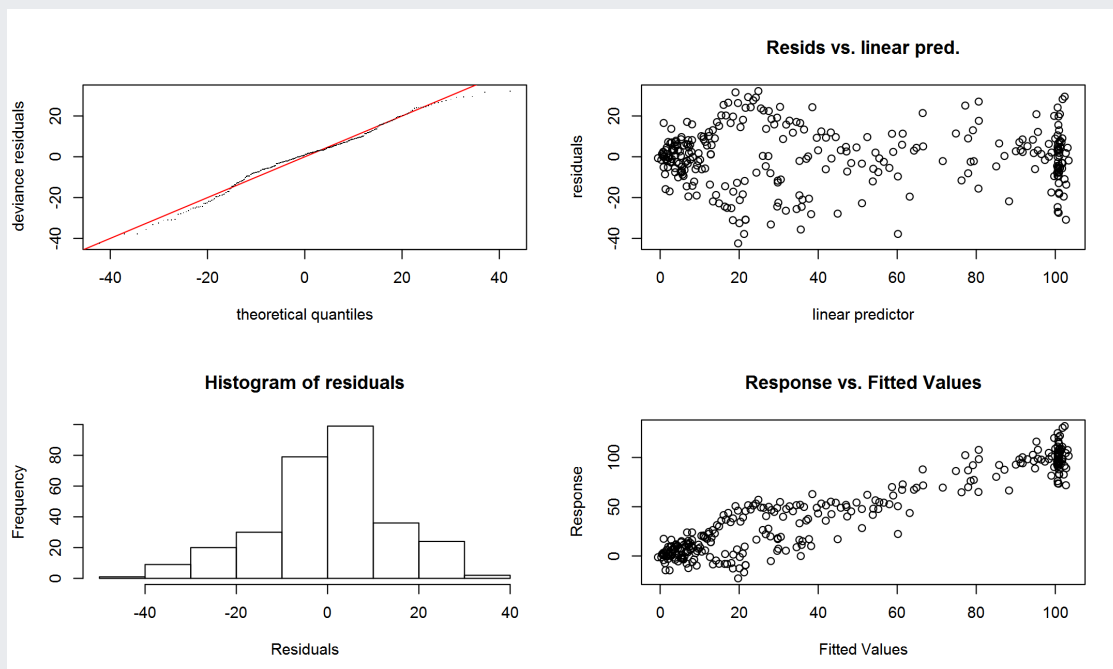
```
gam.check(my_gam)
```

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 4 iterations.
## The RMS GCV score gradient at convergence was 1.107369e-05 .
## The Hessian was positive definite.
## Model rank = 10 / 10
##
## 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(X) 9.00 6.09    1.1   0.97
```

15 / 19

Check your model:

```
gam.check(my_gam)
```



```
##
```


Is it any better than linear model?

```
my_lm <- lm(Y ~ X, data=dt)
anova(my_lm, my_gam)

## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ s(X, bs = "cr")
##   Res.Df  RSS      Df Sum of Sq      F      Pr(>F)
## 1 298.00 88154
## 2 292.91 60613  5.0873    27540 26.161 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, yes it is!

17 / 19

Summary

- Regression models are concerned with explaining one variable: y , with another: x
- This relationship is assumed to be linear
- If your data are not linear, or noisy, a smoother might be appropriate
- Splines are ideal smoothers, and are polynomials joined at 'knot' points
- GAMs are a framework for regressions using smoothers
- `mgcv` is a great package for GAMs with various smoothers available
- `mgcv` estimates the required smoothing penalty for you
- `gratia` or `mgcViz` packages are good visualization tool for GAMs

18 / 19

References and Further reading:

GitHub code:

https://github.com/chrismainey/Why_mgcv_is_awesome

Simon Wood's comprehensive book:

- WOOD, S. N. 2017. Generalized Additive Models: An Introduction with R, Second Edition, Florida, USA, CRC Press.

Noam Ross free online GAM course:

<https://noamross.github.io/gams-in-r-course/>

- HARRELL, F. E., JR. 2001. Regression Modeling Strategies, New York, Springer-Verlag New York.
- HASTIE, T. & TIBSHIRANI, R. 1986. Generalized Additive Models. Statistical Science, 1, 297-310. 291
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. The Elements of Statistical Learning : Data Mining, Inference, and Prediction, New York, NETHERLANDS, Springer.