Model checking

“perhaps the most important part of applied statistical modelling”

Simon Wood

Model checking

- Checking $\neq$ validation!
- As with detection function, checking is important
- Want to know the model conforms to assumptions
- What assumptions should we check?

What to check

- Convergence
- Basis size
- Residuals

Convergence

Convergence

- Fitting the GAM involves an optimization
- By default this is REstricted Maximum Likelihood (REML) score
- Sometimes this can go wrong
- R will warn you!

A model that converges

```
gam.check(dsm_tw_xy_depth)
```

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-3.468176e-05,1.090937e-05]
(score 374.7249 & scale 4.172176).
Hessian positive definite, eigenvalue range [1.179219,301.267].
Model rank =  39 / 39

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,y)   29.00 11.11    0.65  <2e-16 ***
s(Depth)  9.00  3.84    0.81    0.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A bad model

```
Error in while (mean(ldxx/(ldxx + ldss)) > 0.4) { :
  missing value where TRUE/FALSE needed
In addition: Warning message:
In sqrt(w) : NaNs produced
Error in while (mean(ldxx/(ldxx + ldss)) > 0.4) { :
  missing value where TRUE/FALSE needed
```

This is **rare**

The Folk Theorem of Statistical Computing

"most statistical computational problems are due not to the algorithm being used but rather the model itself"

Andrew Gelman

Basis size

Basis size (k)

- Set k per term
- e.g. s(x, k=10) or s(x, y, k=100)
- Penalty removes “extra” wigglyness
  - *up to a point!*
- (But computation is slower with bigger k)

Checking basis size

```
gam.check(dsm_x_tw)
```

```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-3.08755e-06,4.928064e-07]
(score 409.936 & scale 6.041307).
Hessian positive definite, eigenvalue range [0.7645492,302.127].
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 4.96    0.76    0.44
```

Increasing basis size

```
dsm_x_tw_k <- dsm(count~s(x, k=20), ddf.obj=df,
                  segment.data=segs, observation.data=obs,
                  family=tw())
gam.check(dsm_x_tw_k)
```

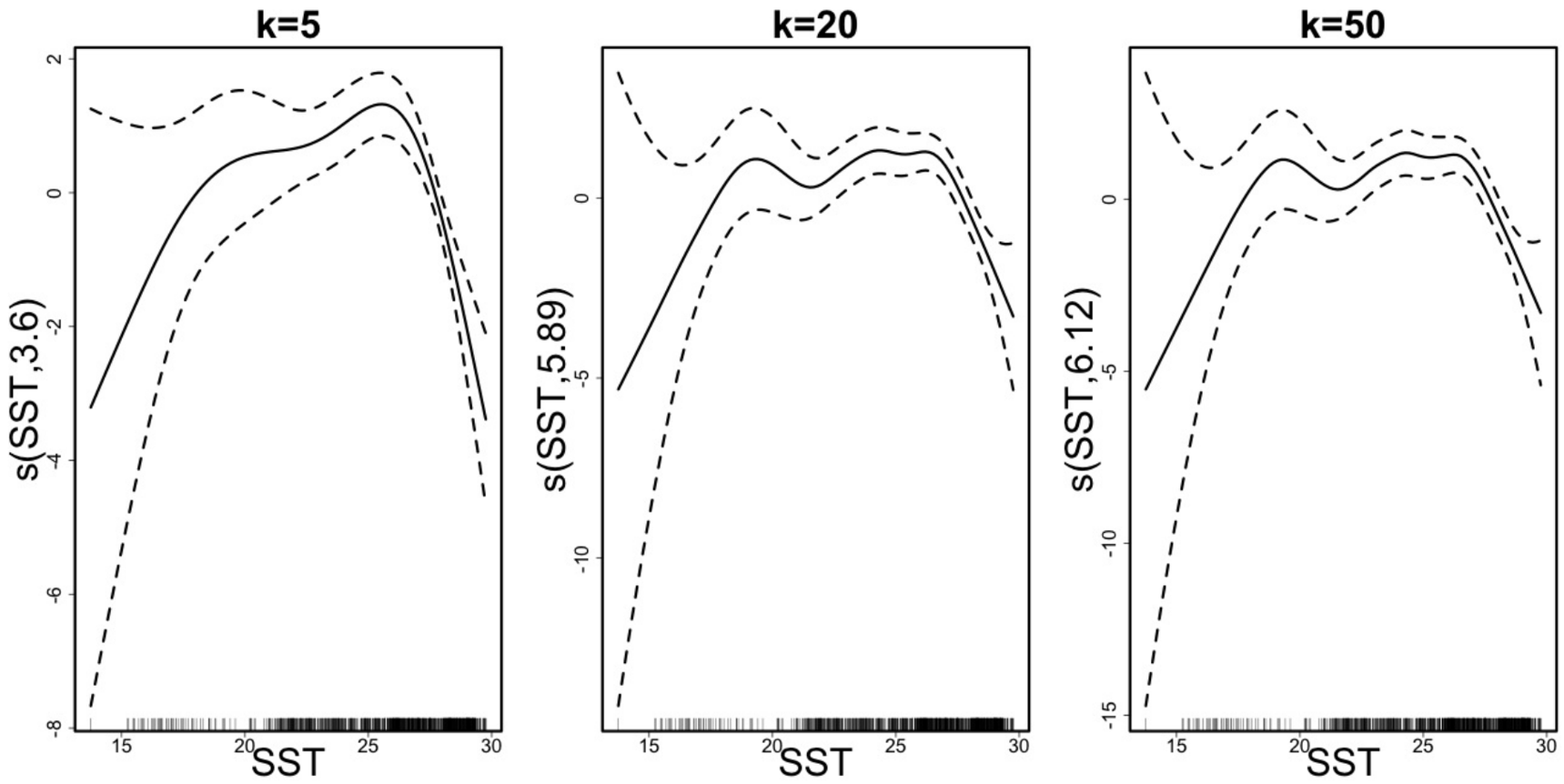
```
Method: REML   Optimizer: outer newton
full convergence after 7 iterations.
Gradient range [-2.301238e-08,3.930667e-09]
(score 409.9245 & scale 6.033913).
Hessian positive definite, eigenvalue range [0.7678456,302.0336].
Model rank =  20 / 20

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) 19.00  5.25    0.76    0.39
```

Sometimes basis size isn’t the issue…

- Generally, double k and see what happens
- Didn't increase the EDF much here
- Other things can cause low “p-value” and “k-index”
- Increasing k can cause problems (nullspace)

k is a maximum

- (Usually) Don't need to worry about things being too wiggly
- k gives the maximum complexity
- Penalty deals with the rest

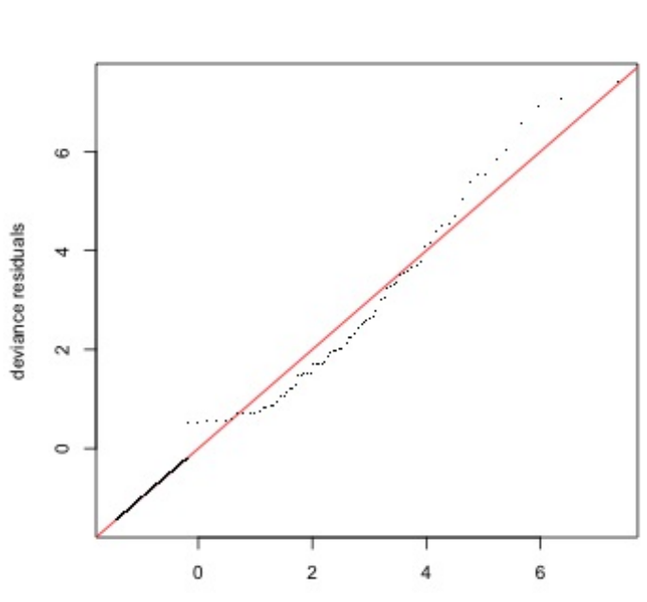


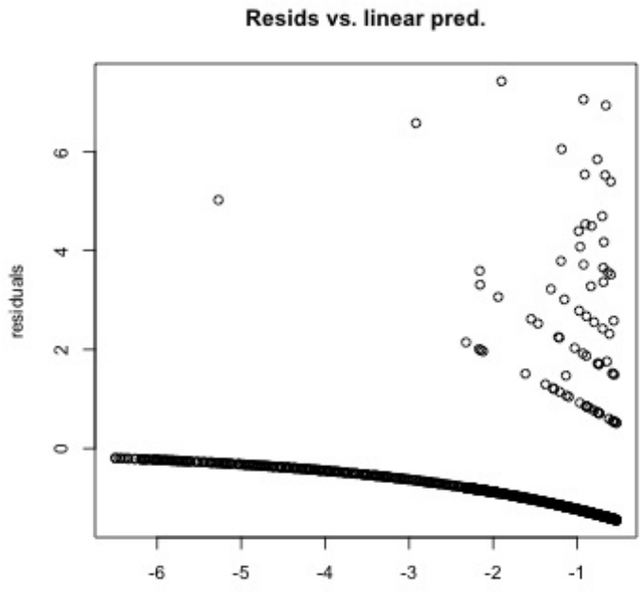
Residuals

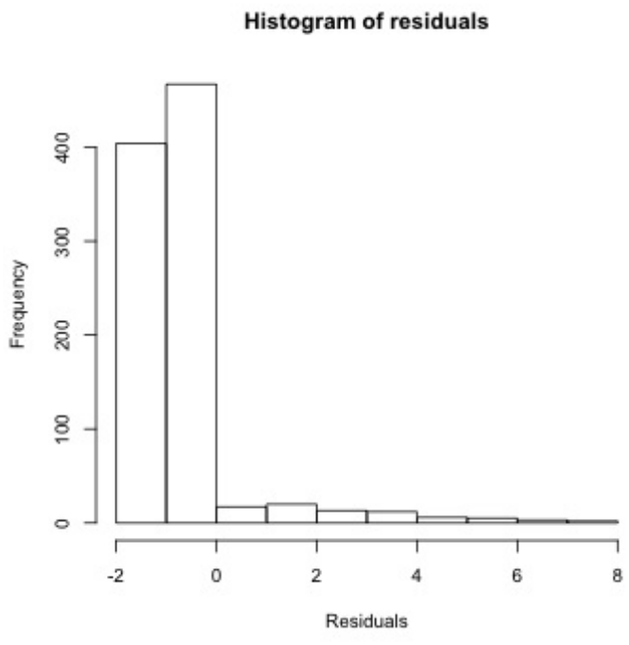
What are residuals?

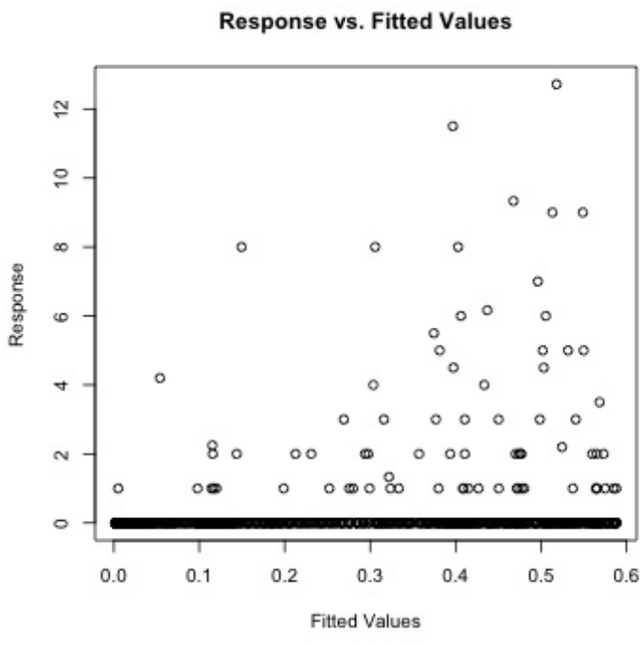
- Generally residuals = observed value - fitted value
- BUT hard to see patterns in these "raw" residuals
- Need to standardise $\Rightarrow$ **deviance residuals**
- Residual sum of squares $\Rightarrow$ linear model
  - deviance $\Rightarrow$ GAM
- Expect these residuals $\sim N(0, 1)$

Residual checking





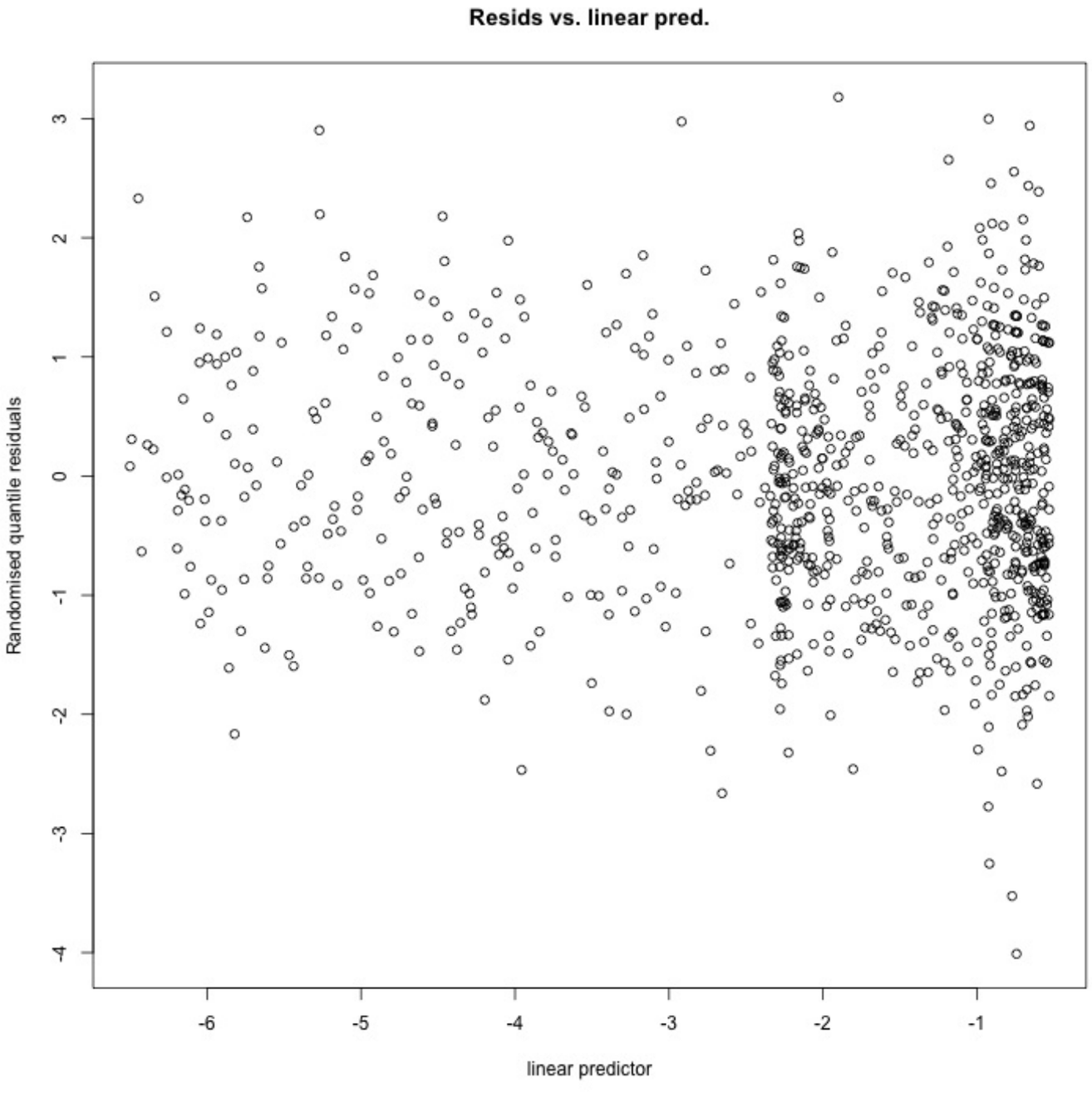




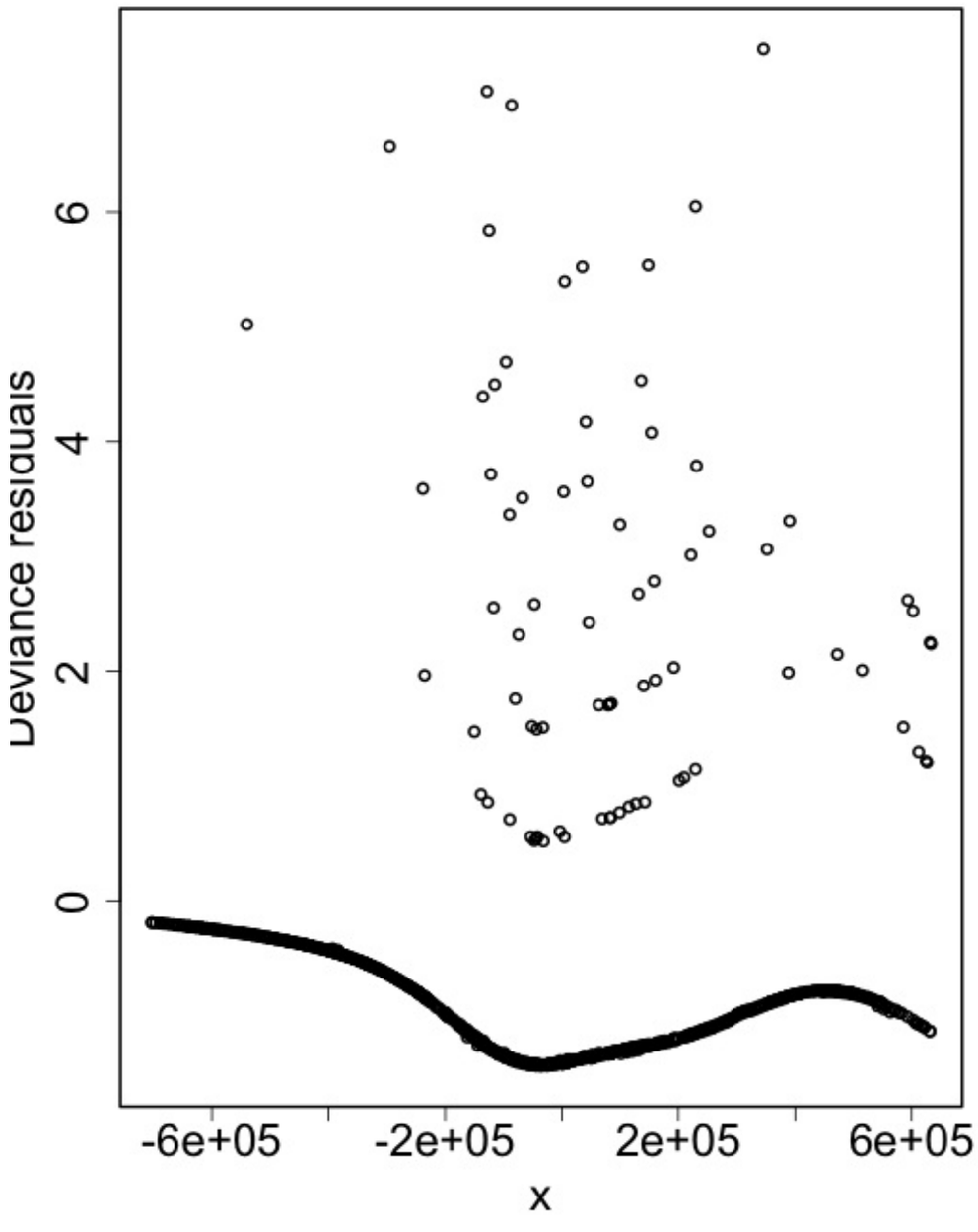
Shortcomings

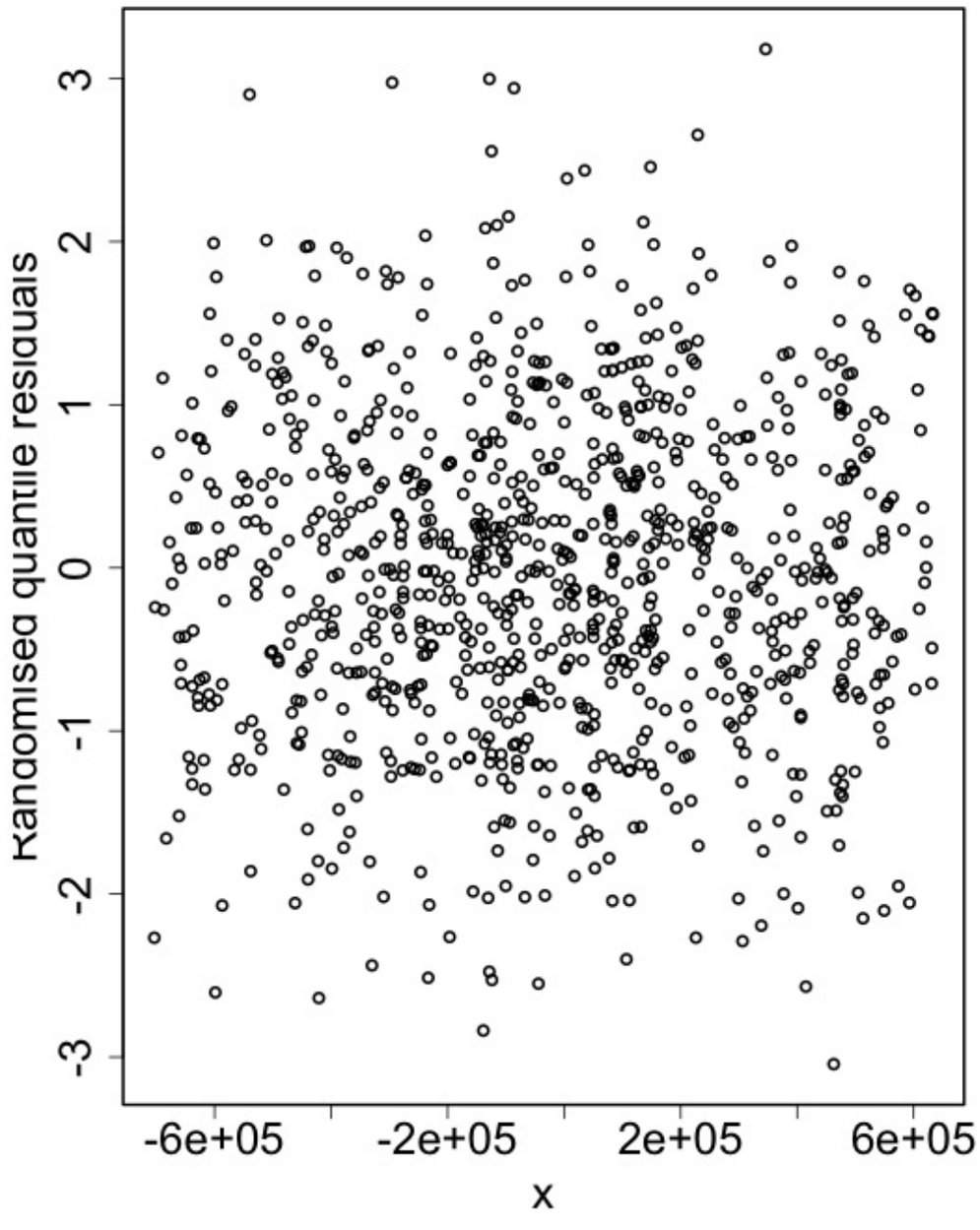
- gam.check can be helpful
- “Resids vs. linear pred” is victim of artifacts
- Need an alternative
- “Randomised quanitle residuals” (*experimental*)
  - rqgam.check
  - Exactly normal residuals

Randomised quantile residuals

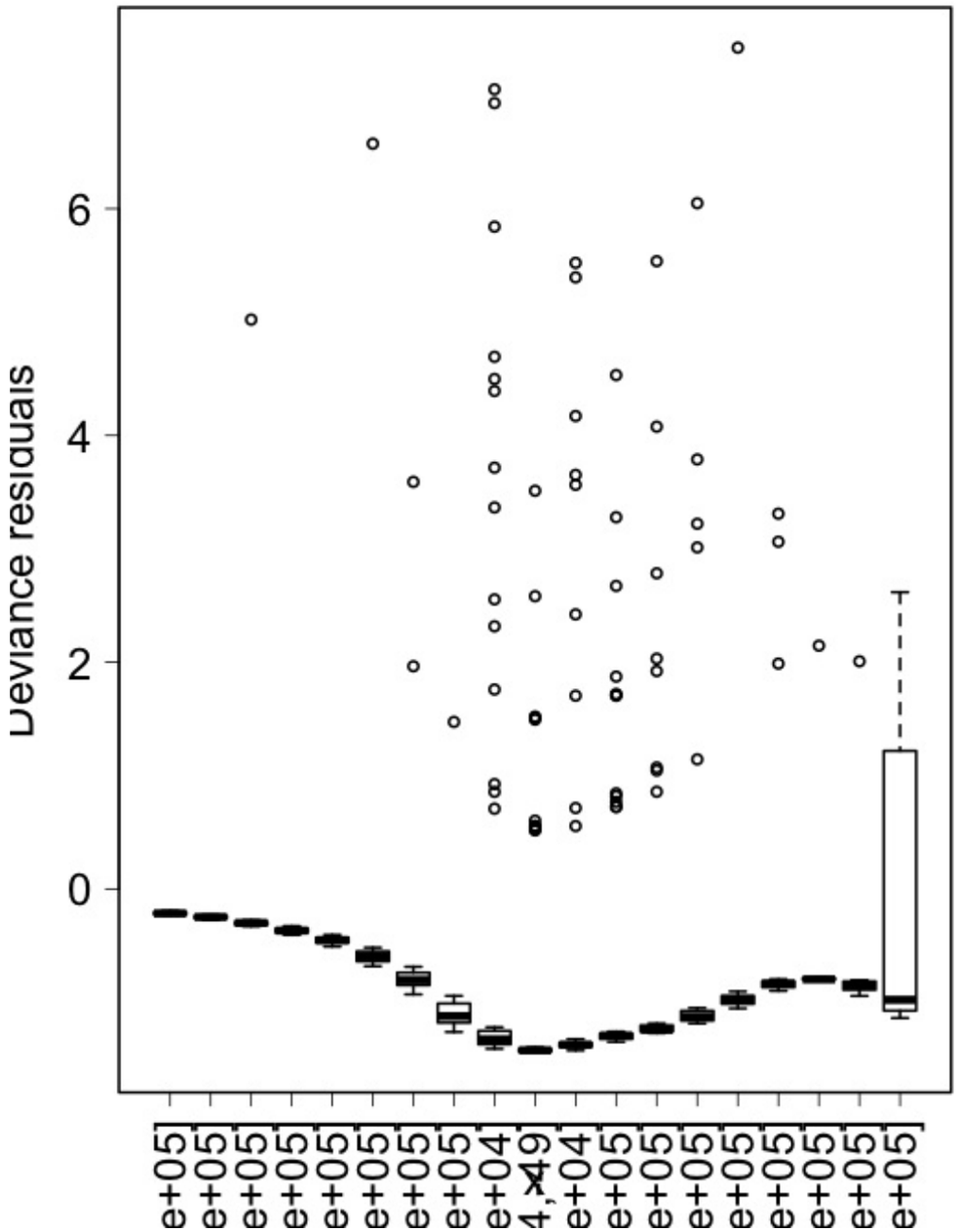


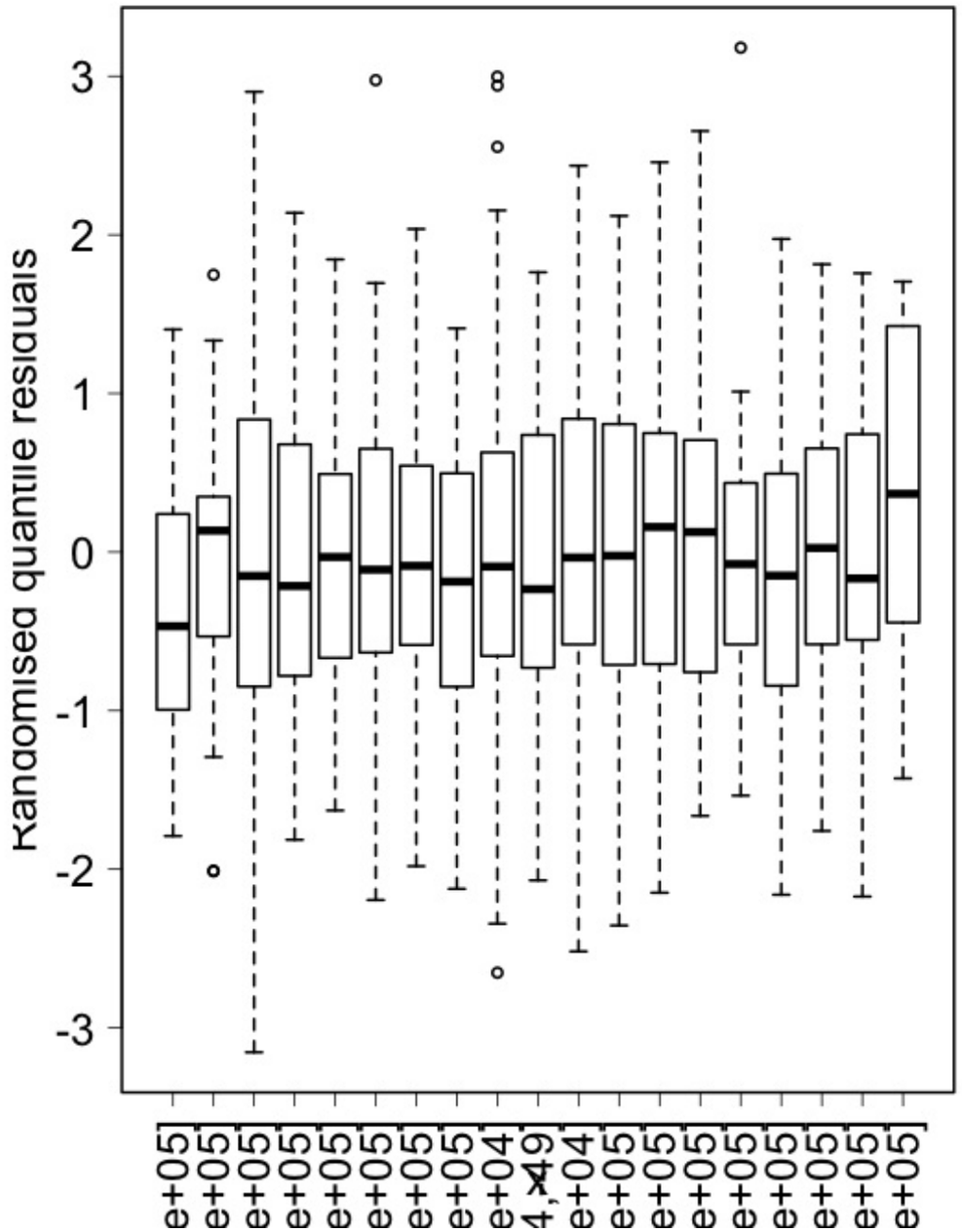
Residuals vs. covariates



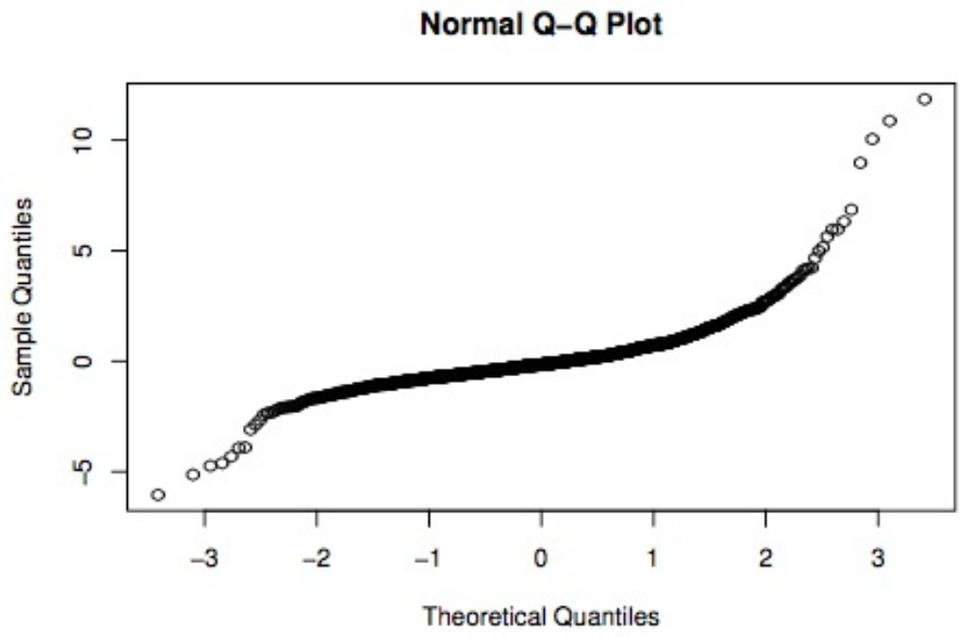


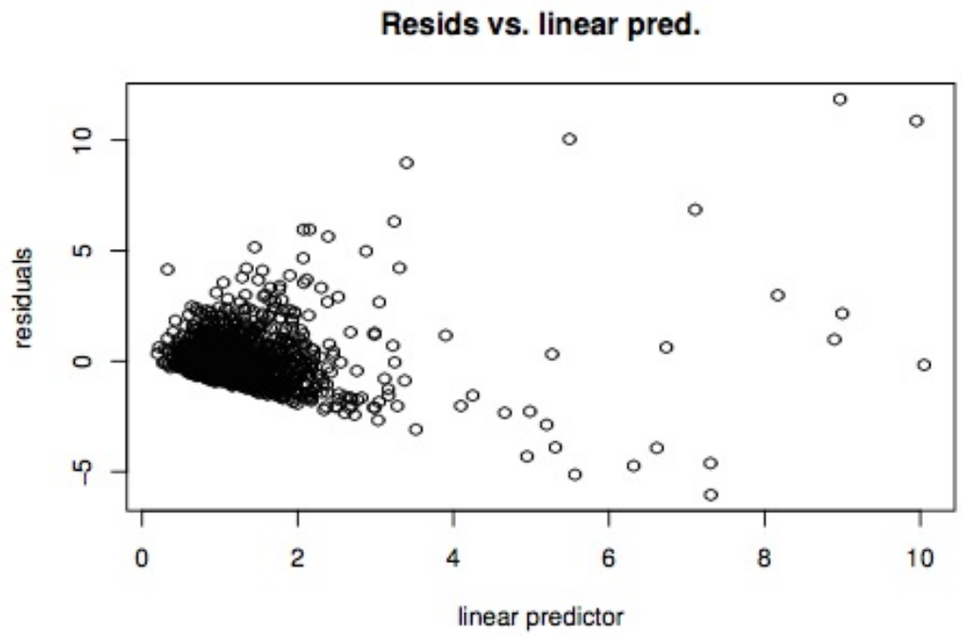
Residuals vs. covariates (boxplots)

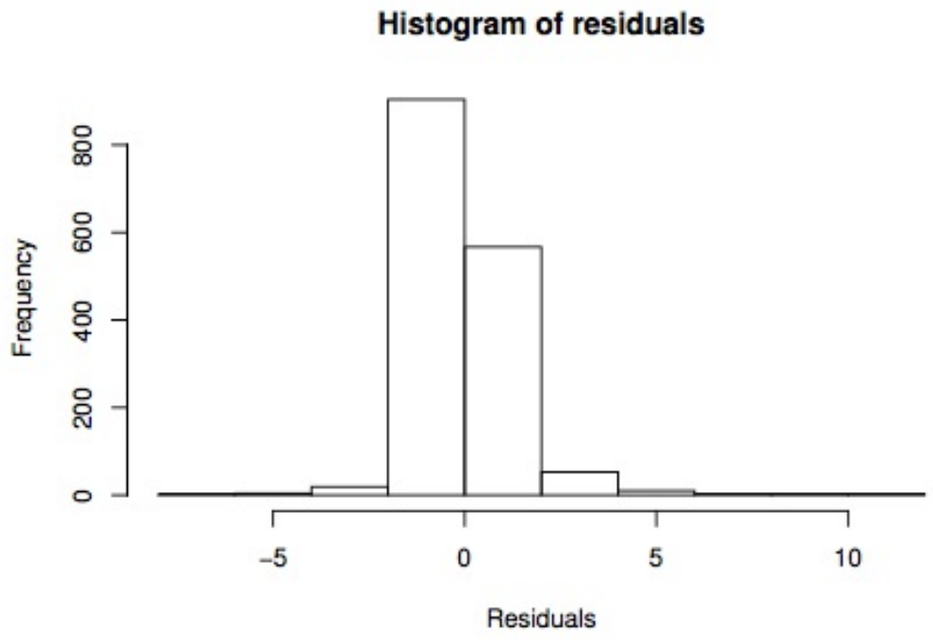


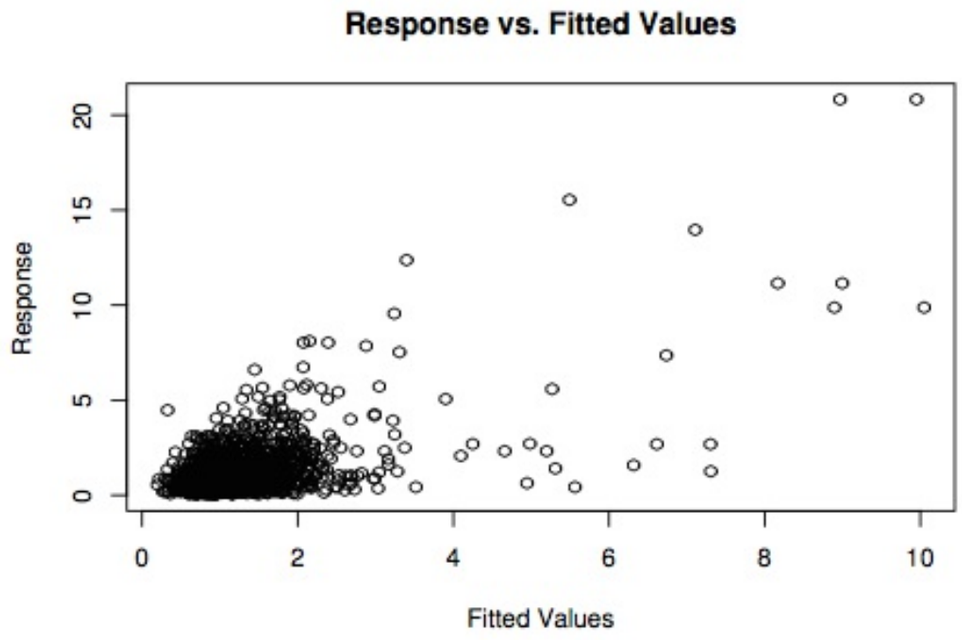


Example of ”bad” plots

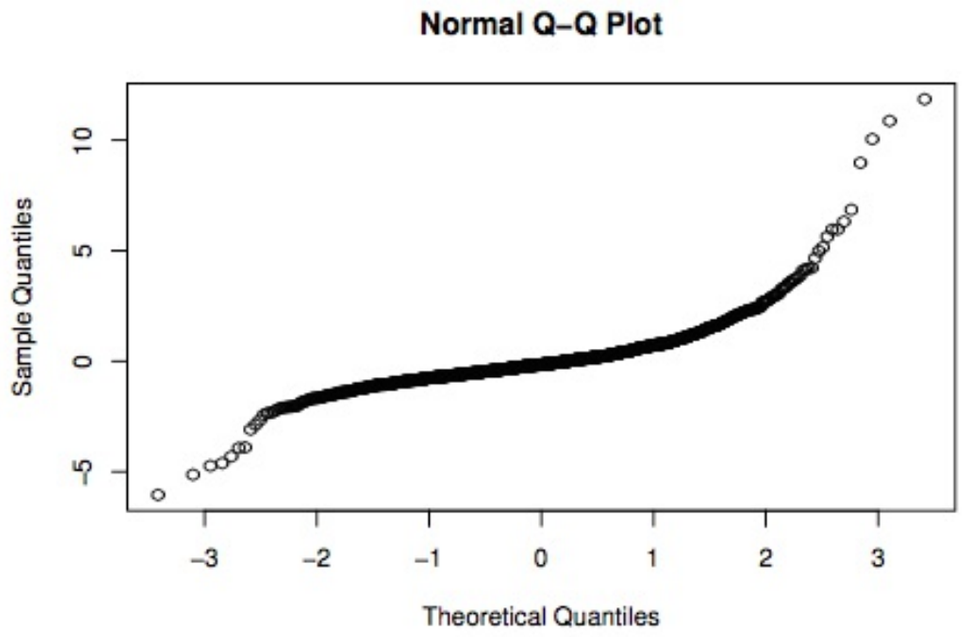


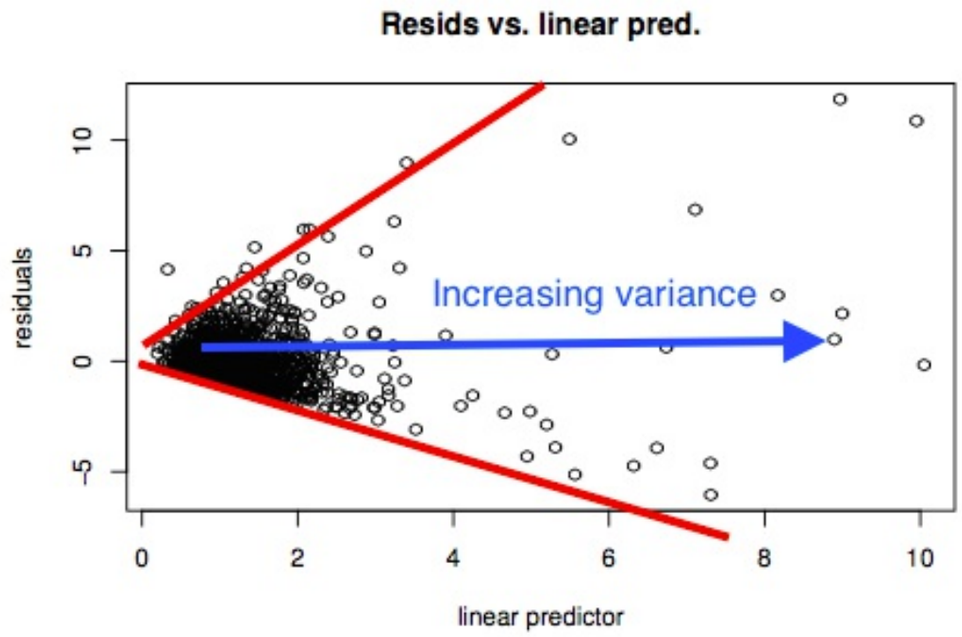


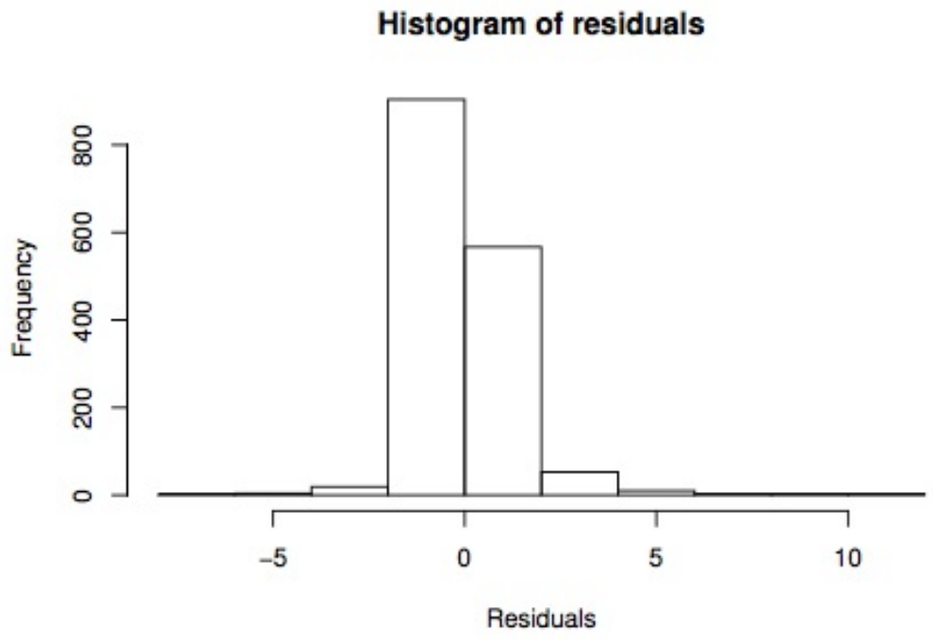


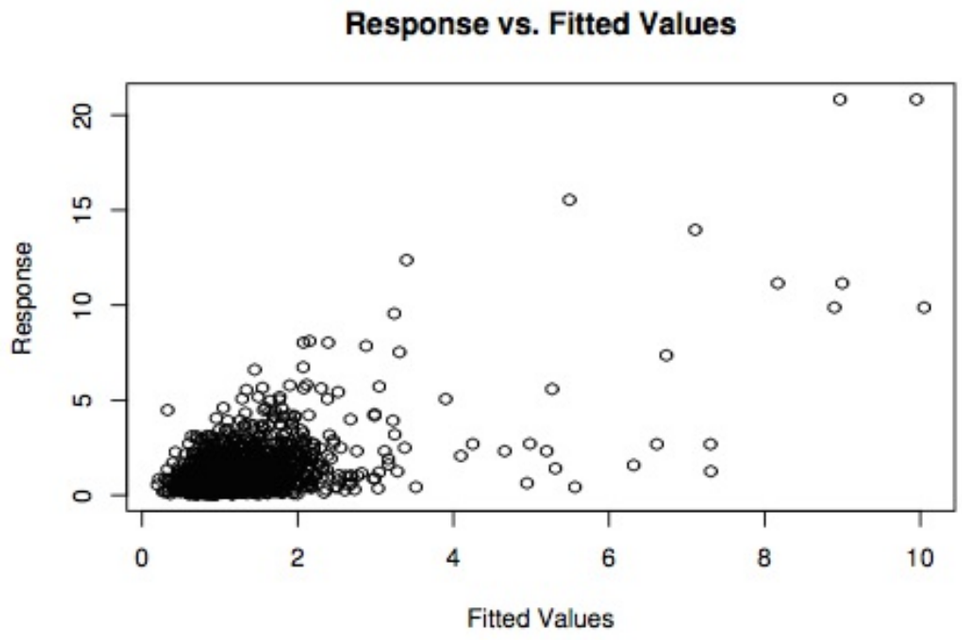


Example of ”bad” plots









Residual checks

- Looking for patterns (not artifacts)
- This can be tricky
- Need to use a mixture of techniques
- Cycle through checks, make changes recheck
- Each dataset is different

Summary

- Convergence
  - Rarely an issue
  - Check your thinking about the model
- Basis size
  - k is a maximum
  - Double and see what happens
- Residuals
  - Deviance and randomised quantile
  - check for artifacts
- gam.check is your friend