Model checking

University of
St Andrews

"perhaps the most important part of applied statistical modelling"

Simon Wood

Model checking

- Checking ≠ 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

A bad model

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 per term
- e.g. or
- Penalty removes "extra" wigglyness
  -
- (But computation is slower with bigger)

Checking basis size

Increasing basis size

Sometimes basis size isn't the issue...

- Generally, double and see what happens
- Didn't increase the EDF much here
- Other things can cause low " " and " "
- Increasing can cause problems (nullspace)

k is a maximum

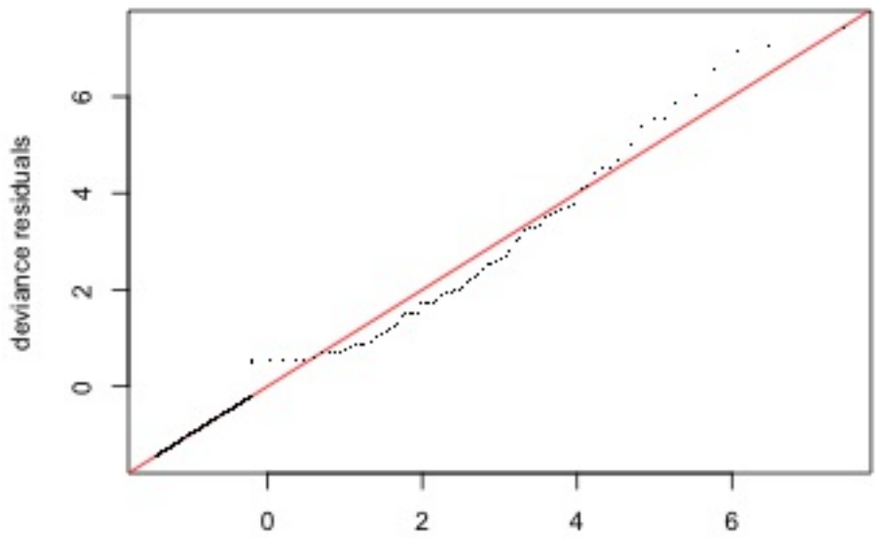
- Don't worry about things being too wiggly
- k gives the maximum complexity
- Penalty deals with the rest

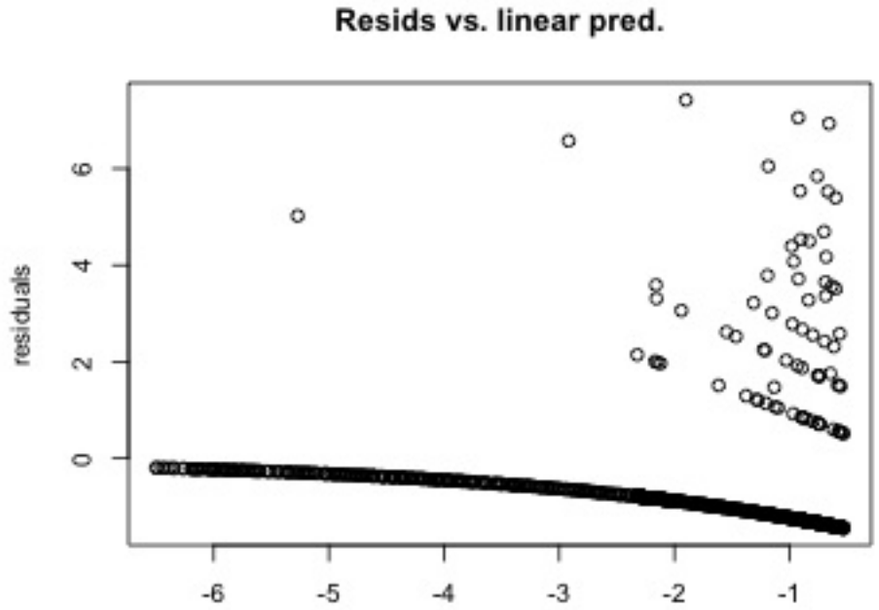
Residuals

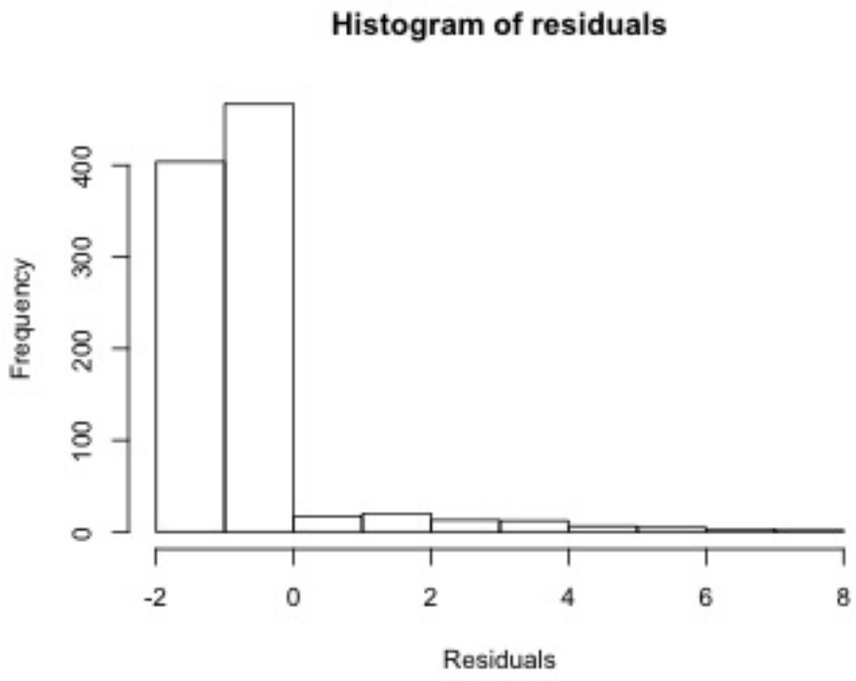
What are residuals?

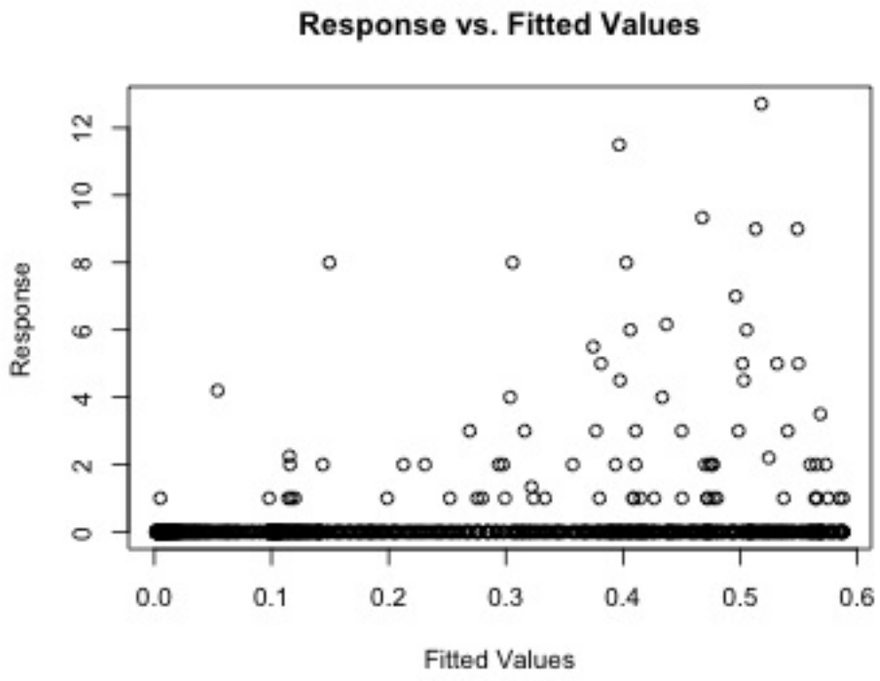
- Generally residuals = observed value - fitted value
- BUT hard to see patterns in these "raw" residuals
- Need to standardise $\Rightarrow$ **deviance residuals**
- Residual sum of squares $\Rightarrow$ linear model
  - deviance $\Rightarrow$ GAM
- Expect these residuals $\sim \mathrm{N}(0, 1)$

Residual checking





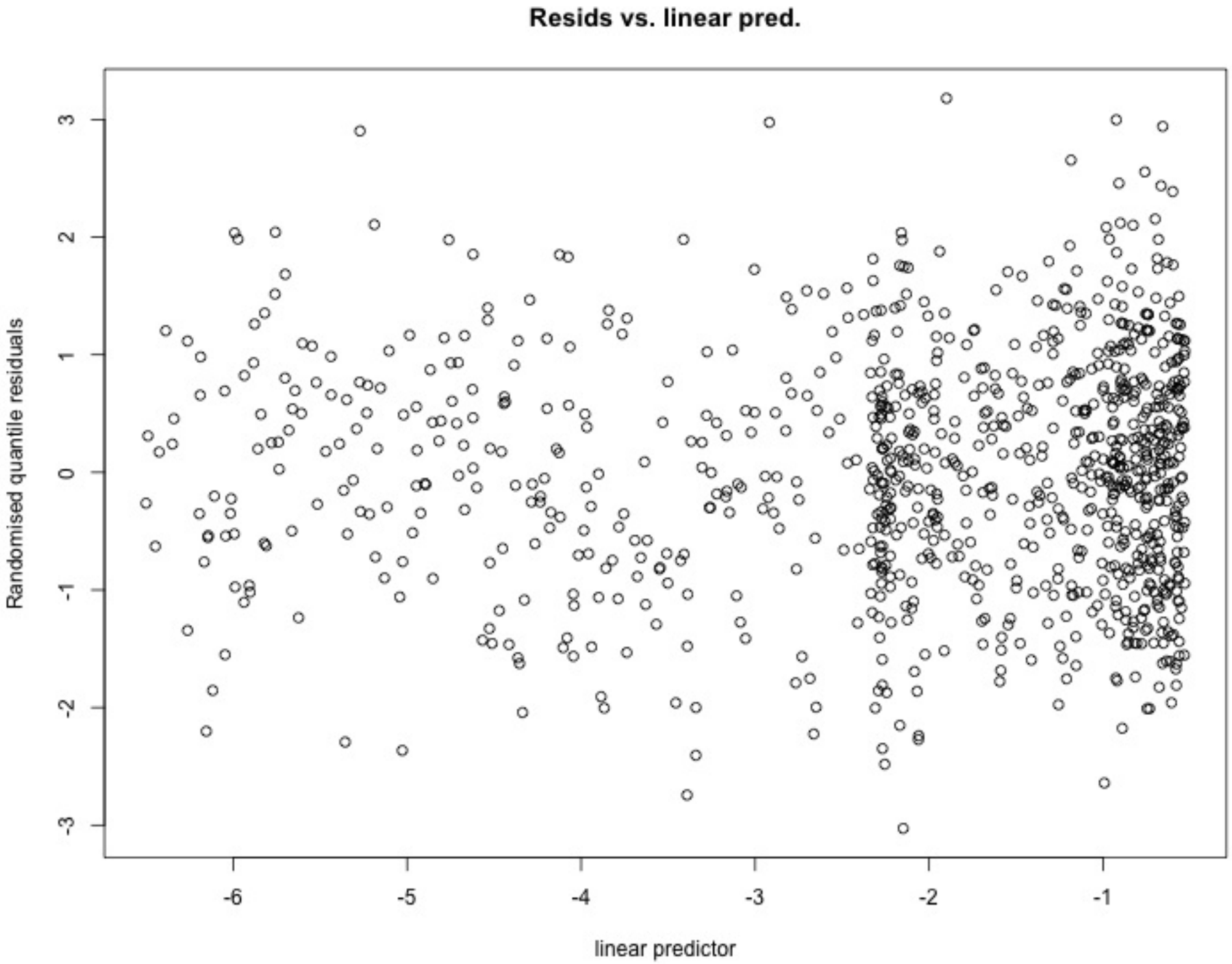




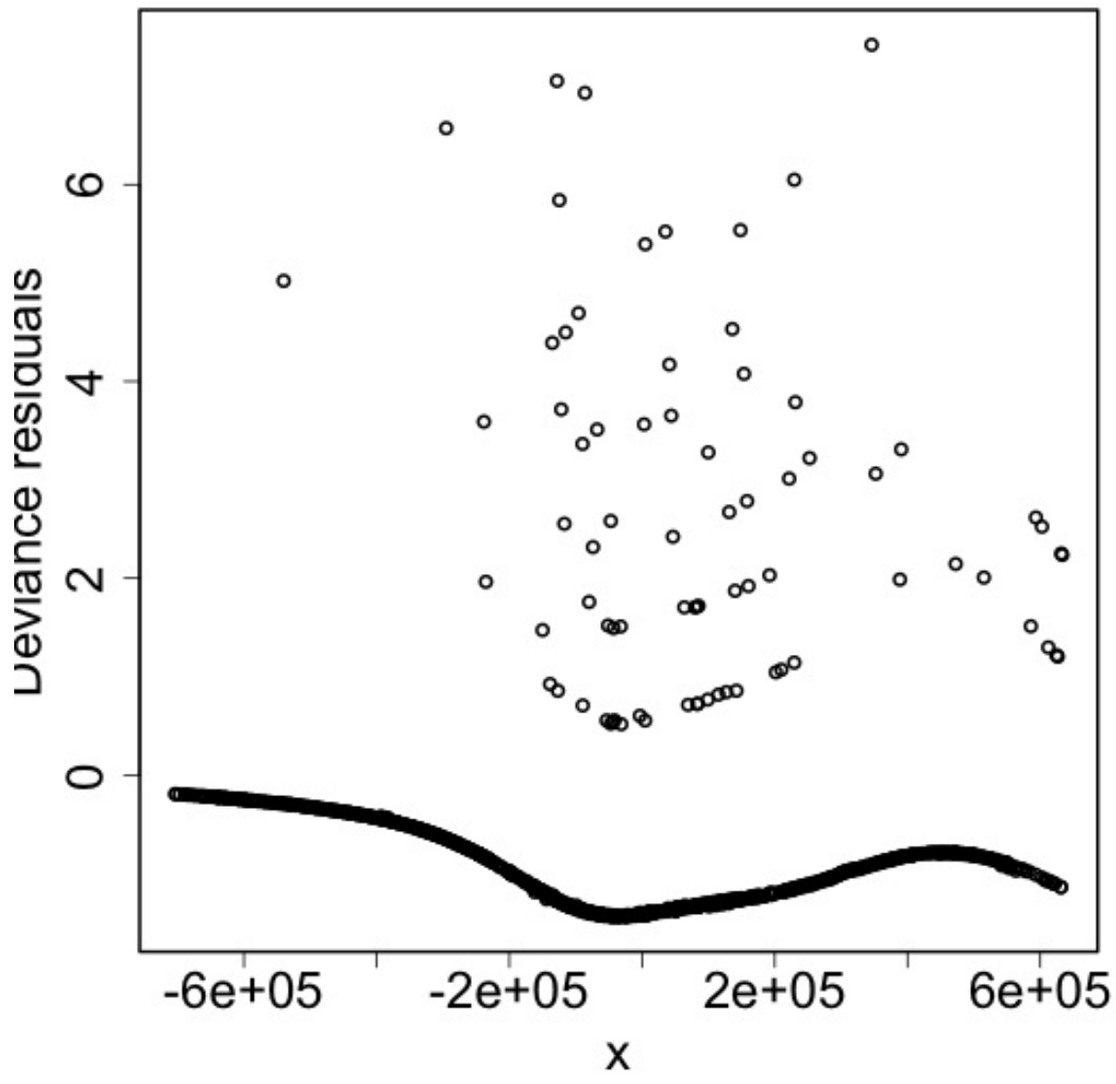
Shortcomings

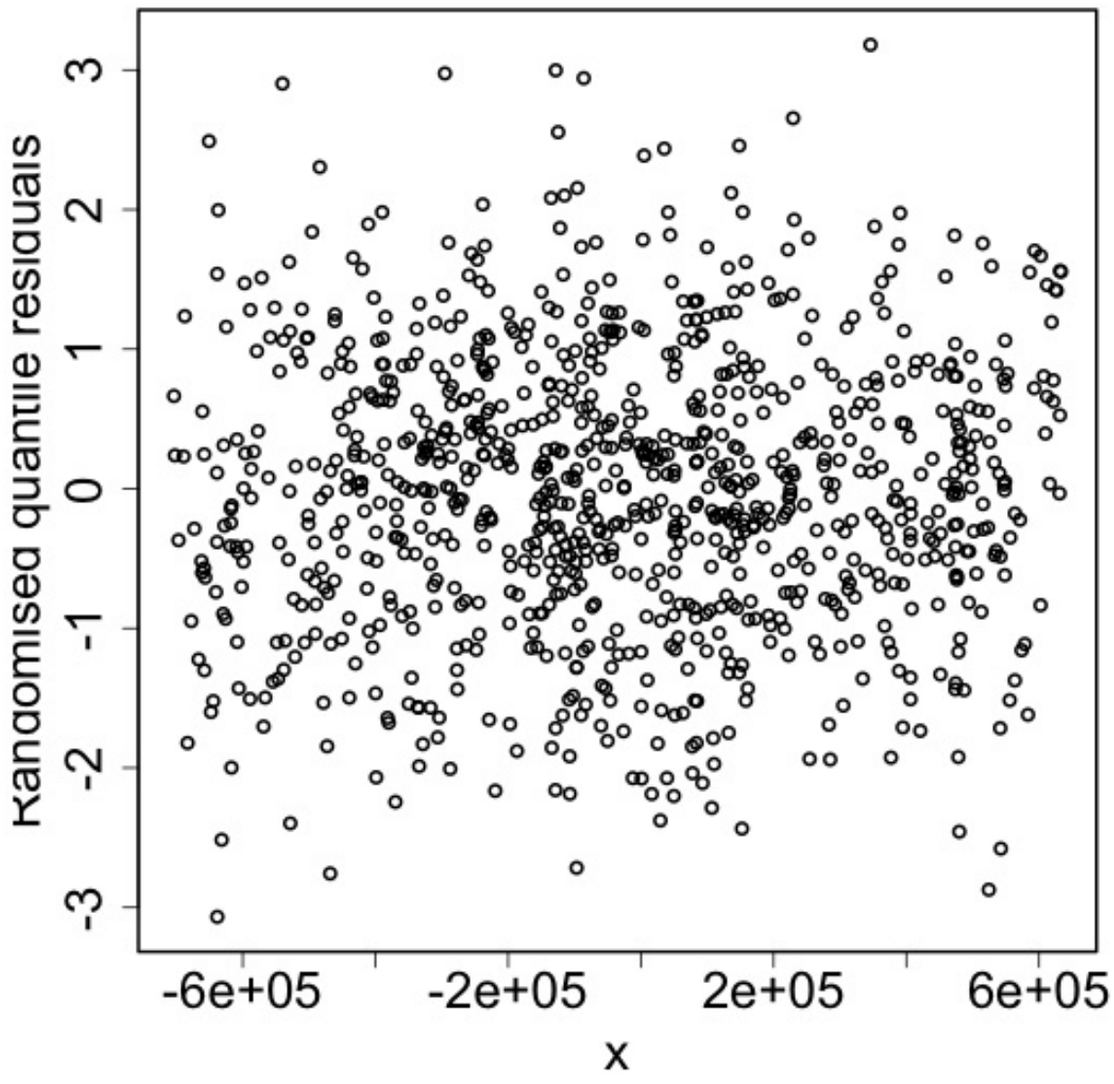
- can be helpful
- "Resids vs. linear pred" is victim of artifacts
- Need an alternative
- "Randomised quanitle residuals" ( )
  -
  - Exactly normal residuals

Randomised quantile residuals

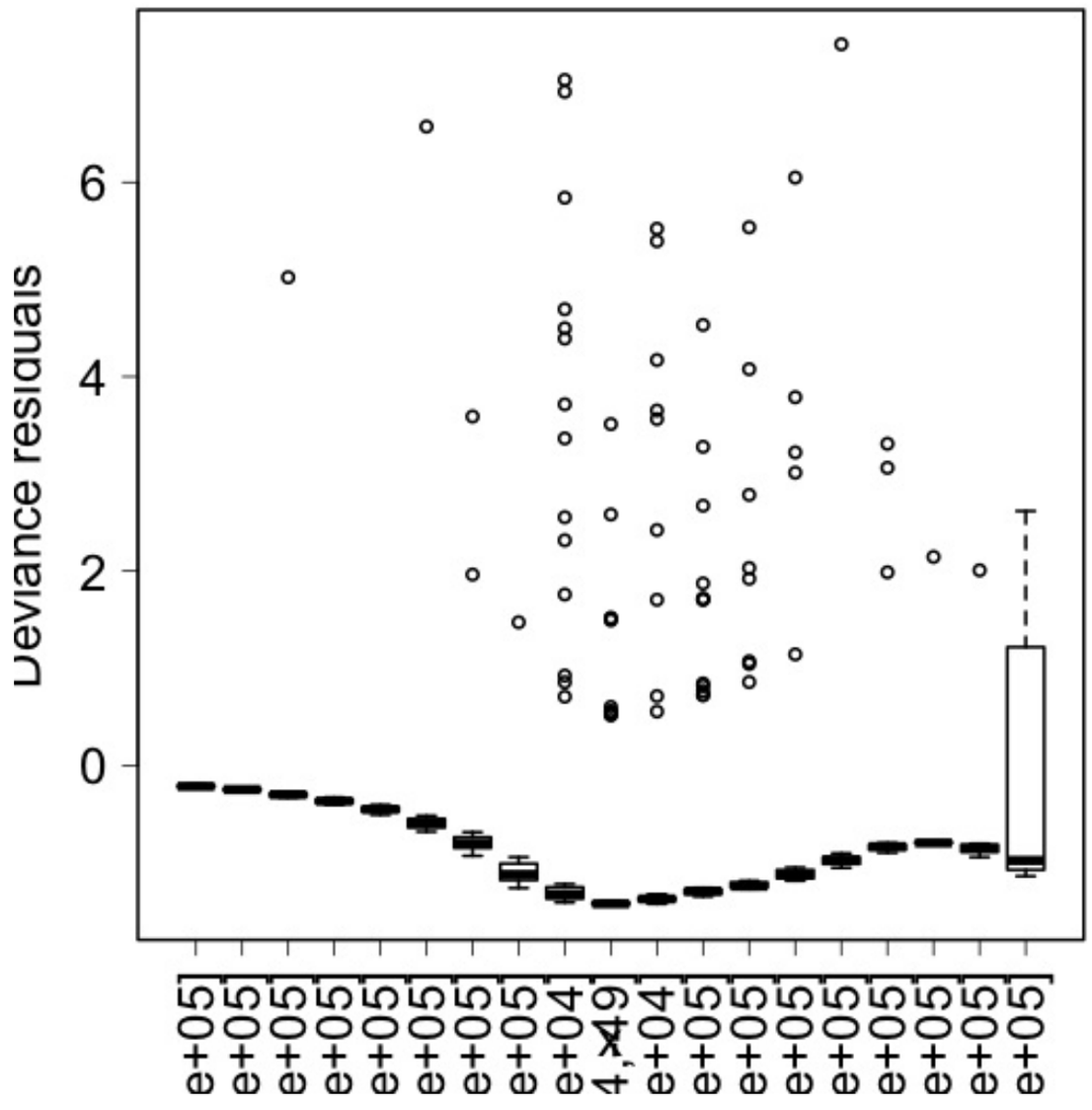


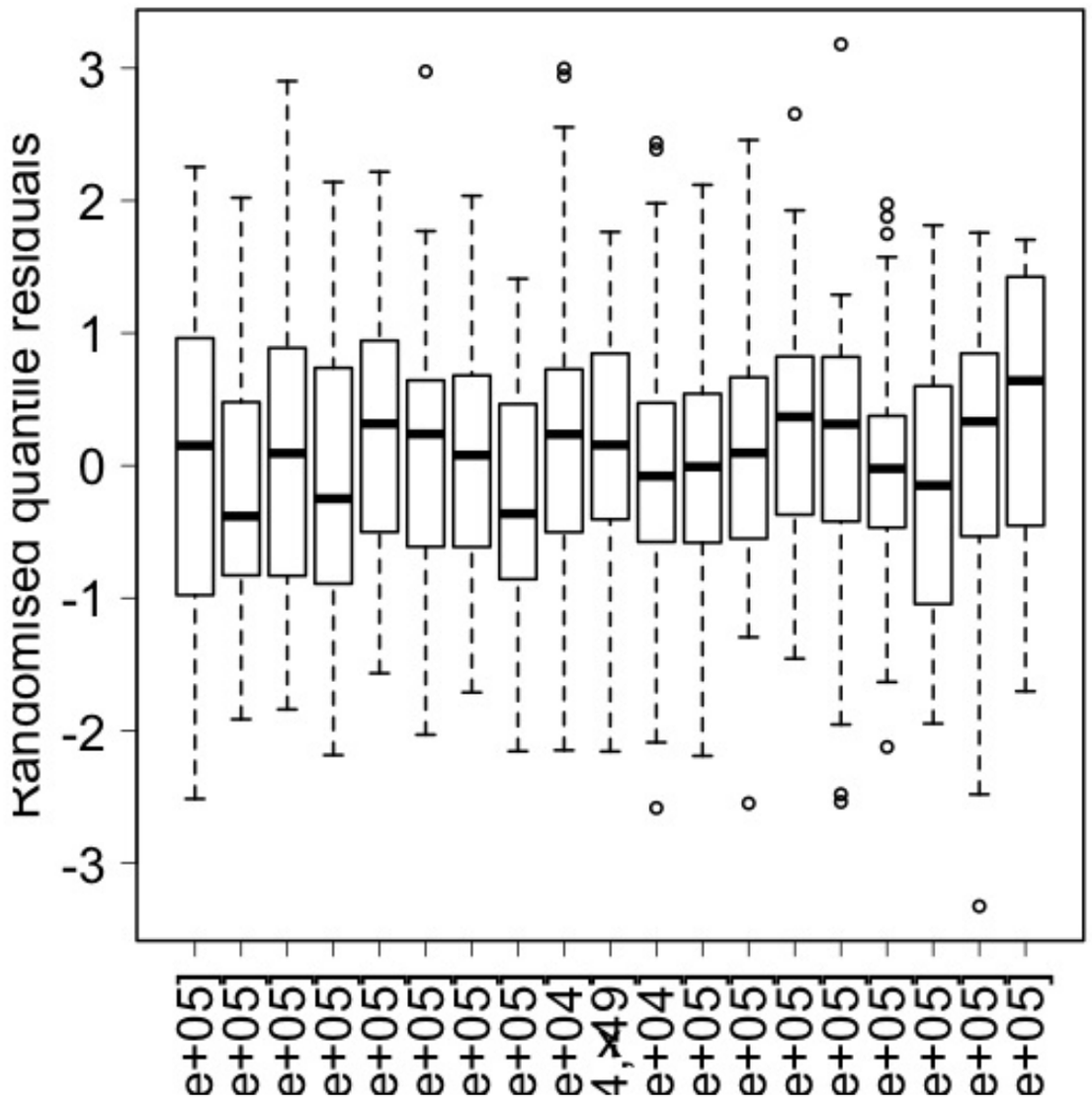
Residuals vs. covariates



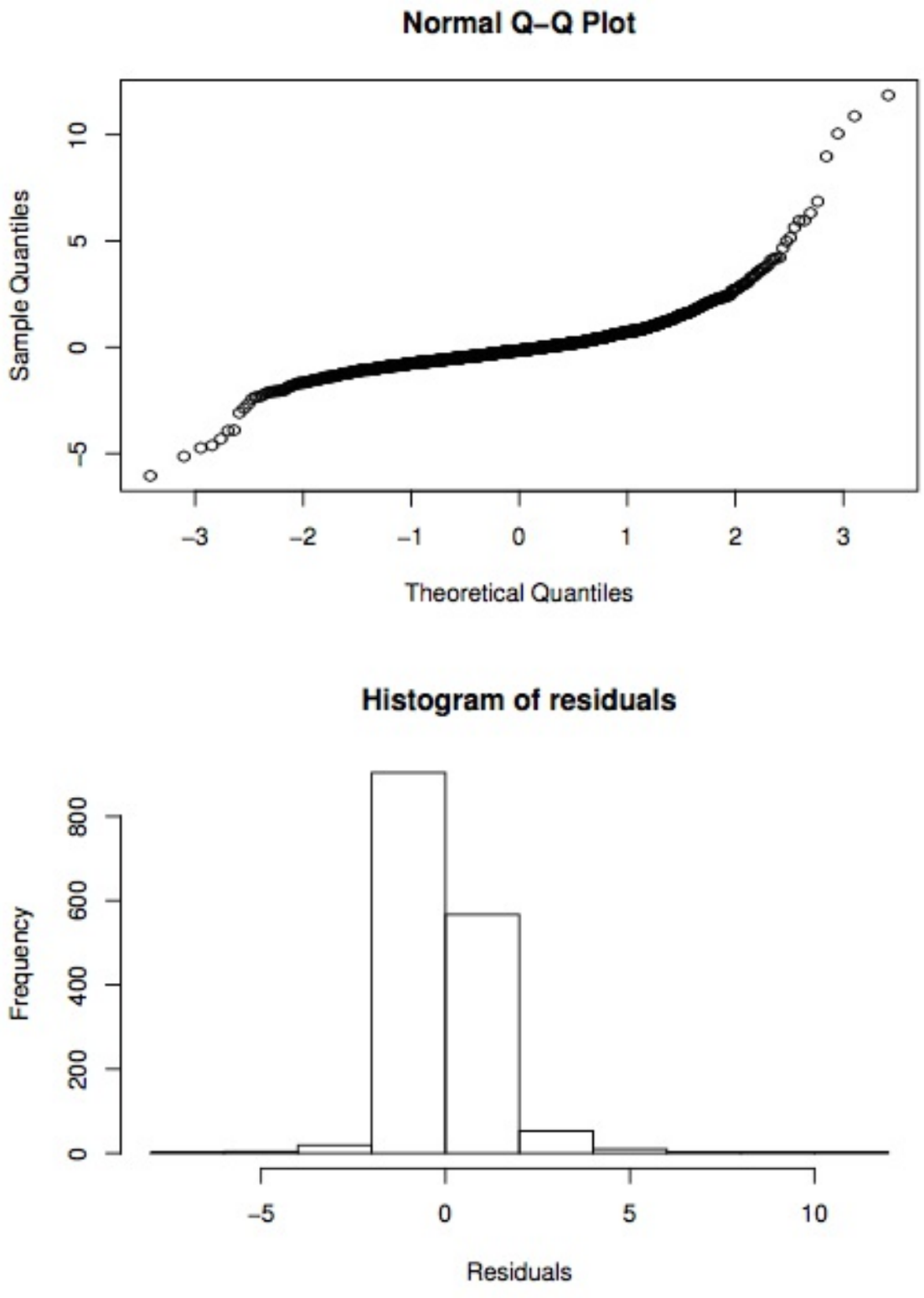


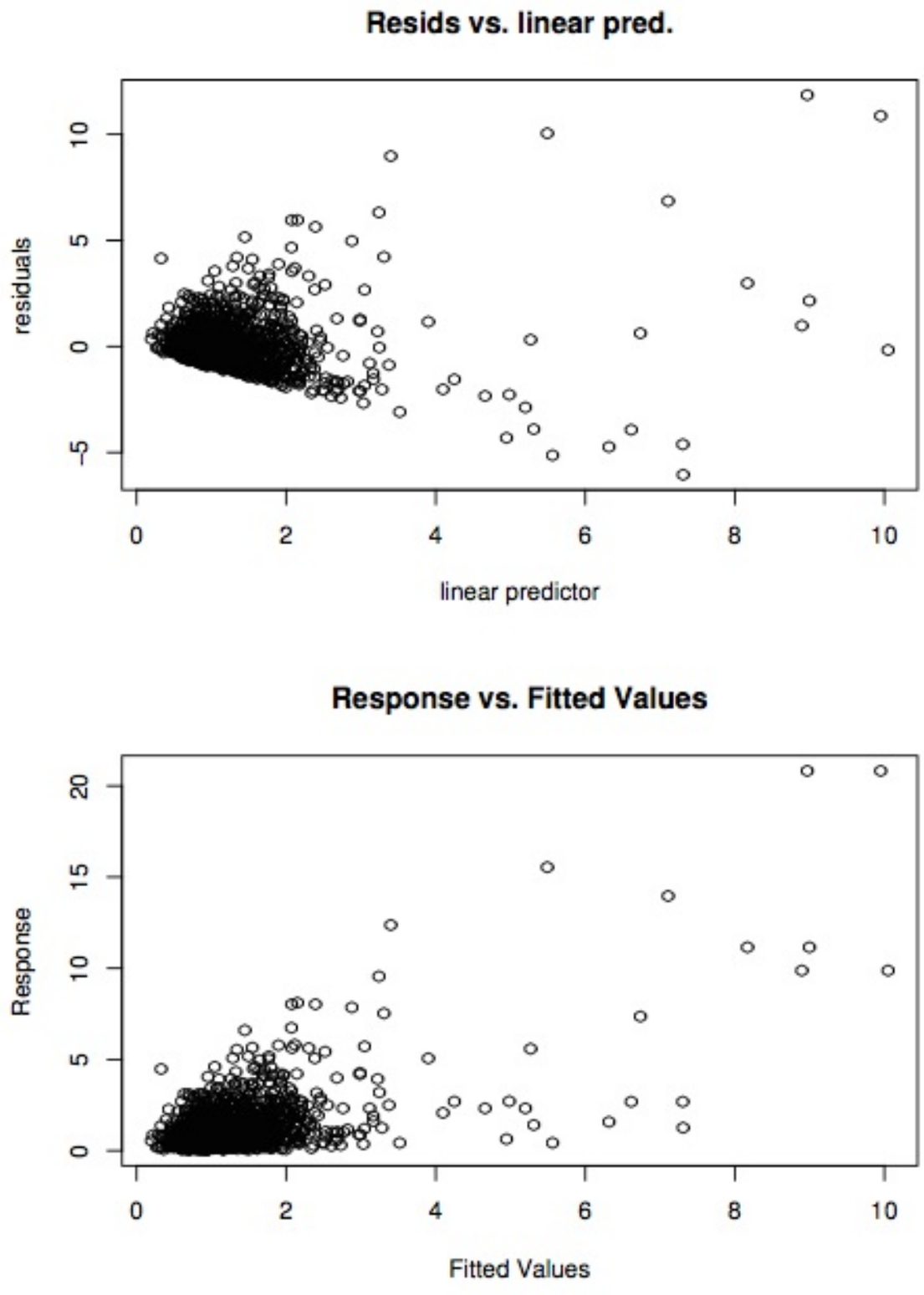
Residuals vs. covariates (boxplots)



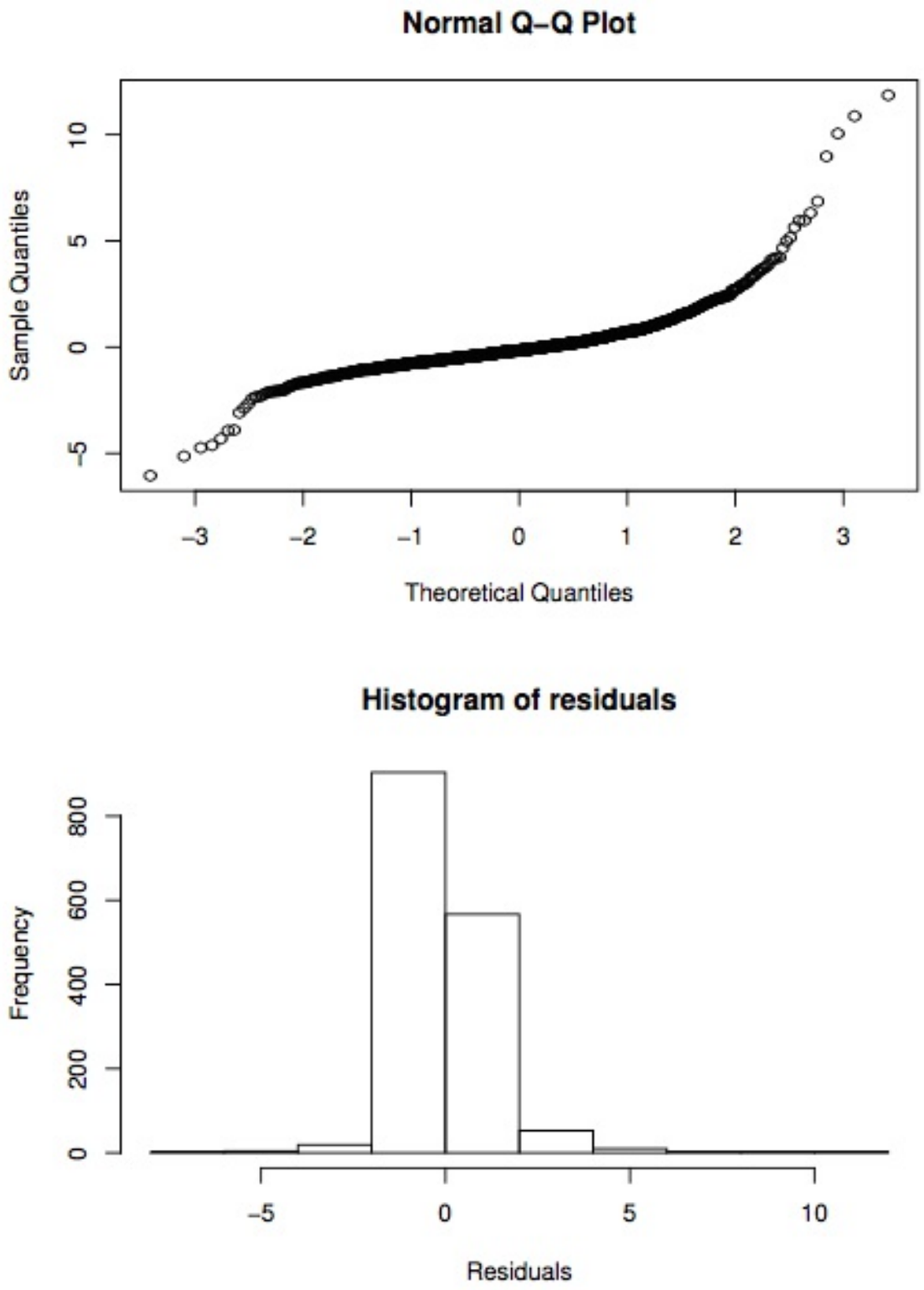


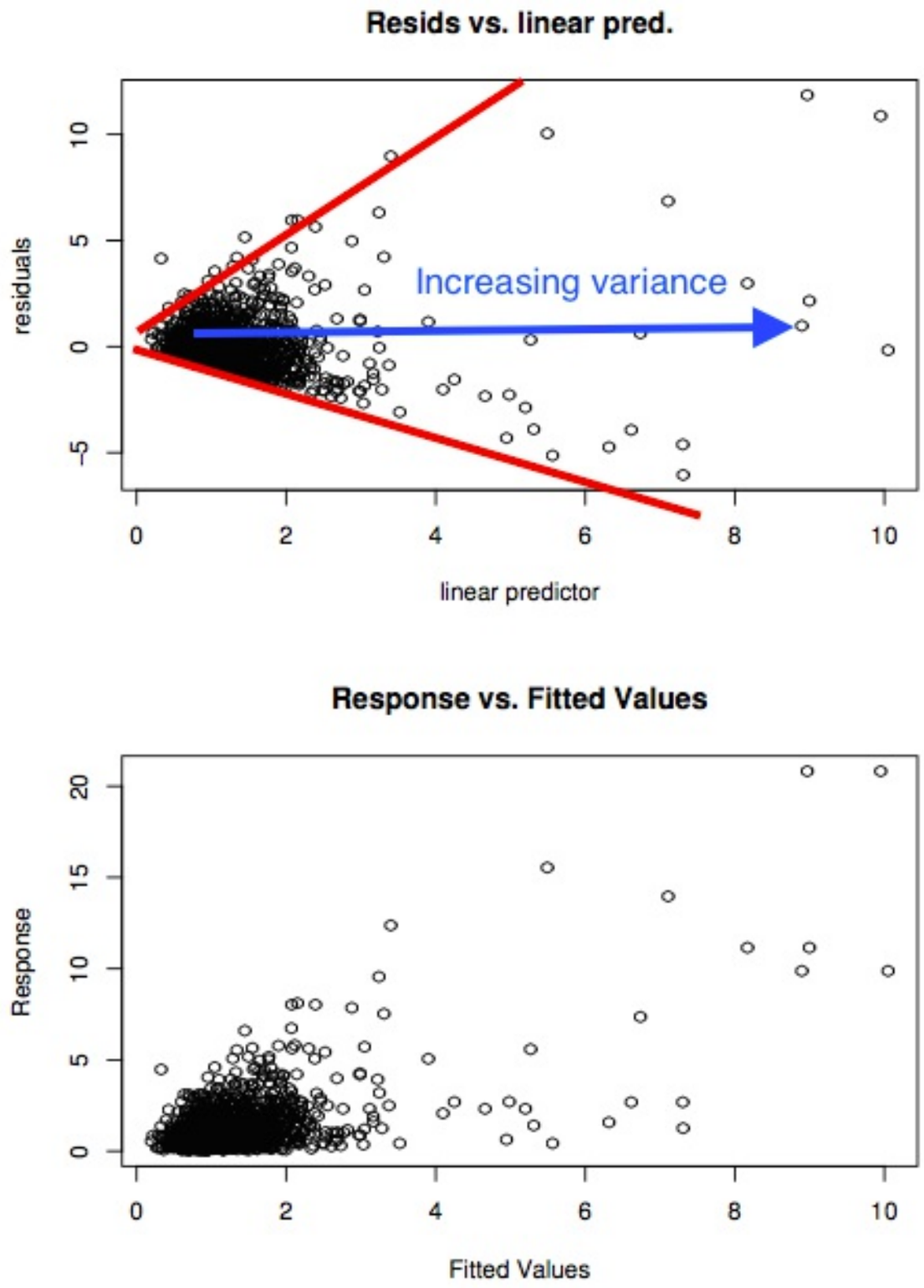
Example of "bad" plots





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
- is your friend