

# Model checking

David L Miller

# Outline

You fitted a GAM, everything is fine, right? *Right?*

But what about?

- Smooth terms flexibility?
- Non-constant variance?
- Response distribution selection?
- Correlated covariates?

*“perhaps the most important  
part of applied statistical  
modelling”*

Simon Wood, Generalized Additive Models: An Introduction in R

# Basis size (k)

- $k \approx$  number of basis functions
- 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)

# Default basis size

```
b <- gam(Nhat ~ s(day), data=pop_unhappy, method="REML")
gam.check(b)
```

```
Method: REML    Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-0.004427235,0.003510967]
(score 6424.346 & scale 844.0966).
Hessian positive definite, eigenvalue range [3.647717,668.0272].
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(day) | 9.000 | 8.778 | 0.638   | 0       |

# Increasing basis size

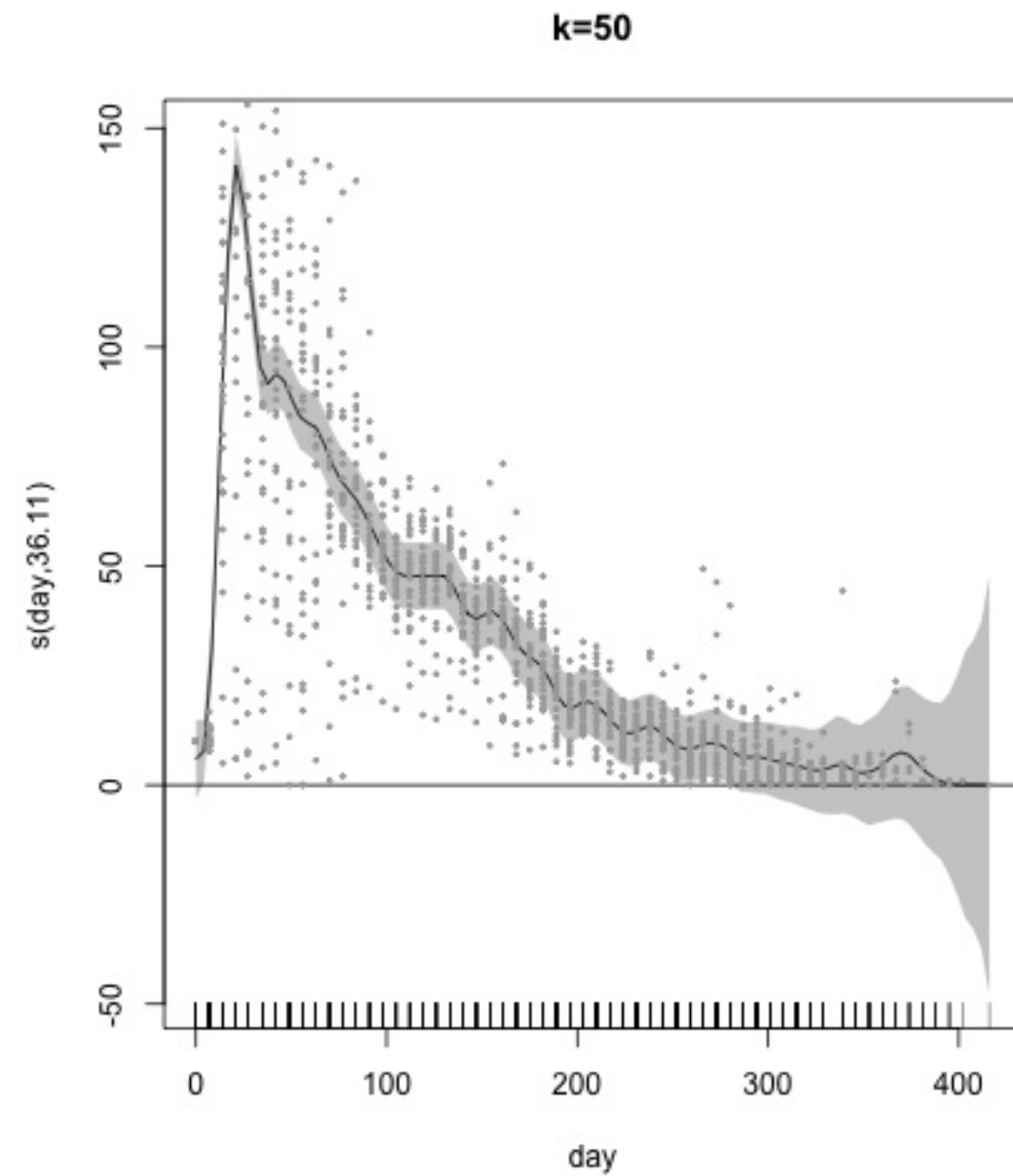
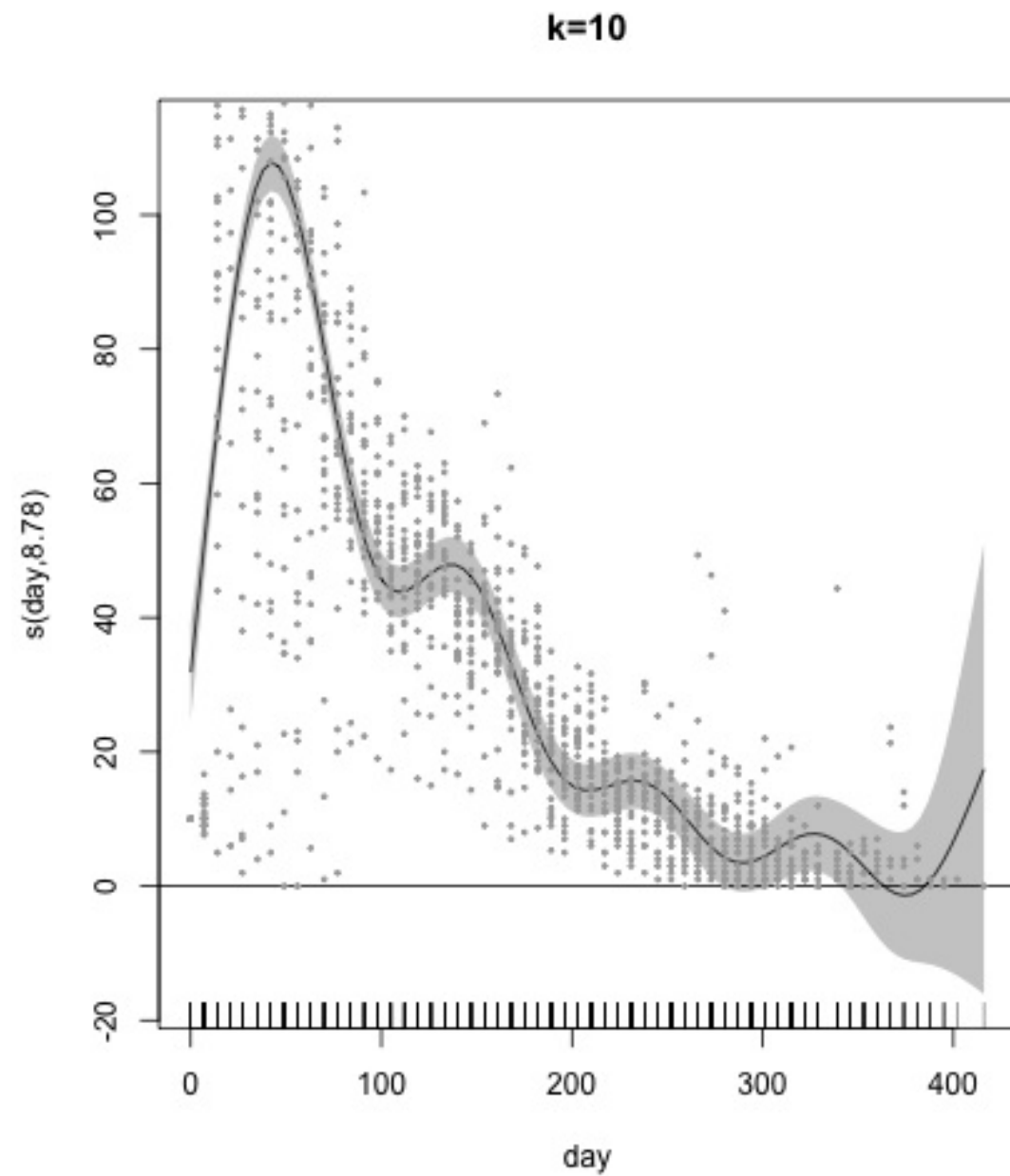
```
b_50 <- gam(Nhat ~ s(day, k=50), data=pop_unhappy, method="REML")
gam.check(b_50)
```

```
Method: REML    Optimizer: outer newton
full convergence after 6 iterations.
Gradient range [-1.841561e-07,6.686406e-09]
(score 6296.597 & scale 647.8836).
Hessian positive definite, eigenvalue range [10.49952,668.4687].
Model rank =   50 / 50
```

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(day) | 49.000 | 36.110 | 0.847   | 0       |

# Does it make a difference?!



# Historical/philosophical note

- “Keep relationships simple and interpretable”
  - What does this mean?
  - Bias confirmation?
  - Limit model to get “clean” relationships?
- Some literature suggests “limit  $k=5$ ” or somesuch
  - Original gam package for S+ had a default  $k=5$
  - Coincidence?
  - (Simon Wood, pers. comm.)



# Residual checks

# Residuals

- Deal with 2 types of residuals
  - Deviance
  - Randomized quantile
- Raw residuals are just (observed - fitted)
  - Analog to  $R^2$
  - Difficult to assess mean-variance relationship graphically
  - Need to rescale so mean-variance is constant

# Deviance residuals

- Deviance  $\approx$  “ $R^2$  for GAMs”
- Per-observation deviance  $\approx$  raw residuals?
- Multiply by sign of (observed-fitted)
- Should be Normal(0, 1) distributed

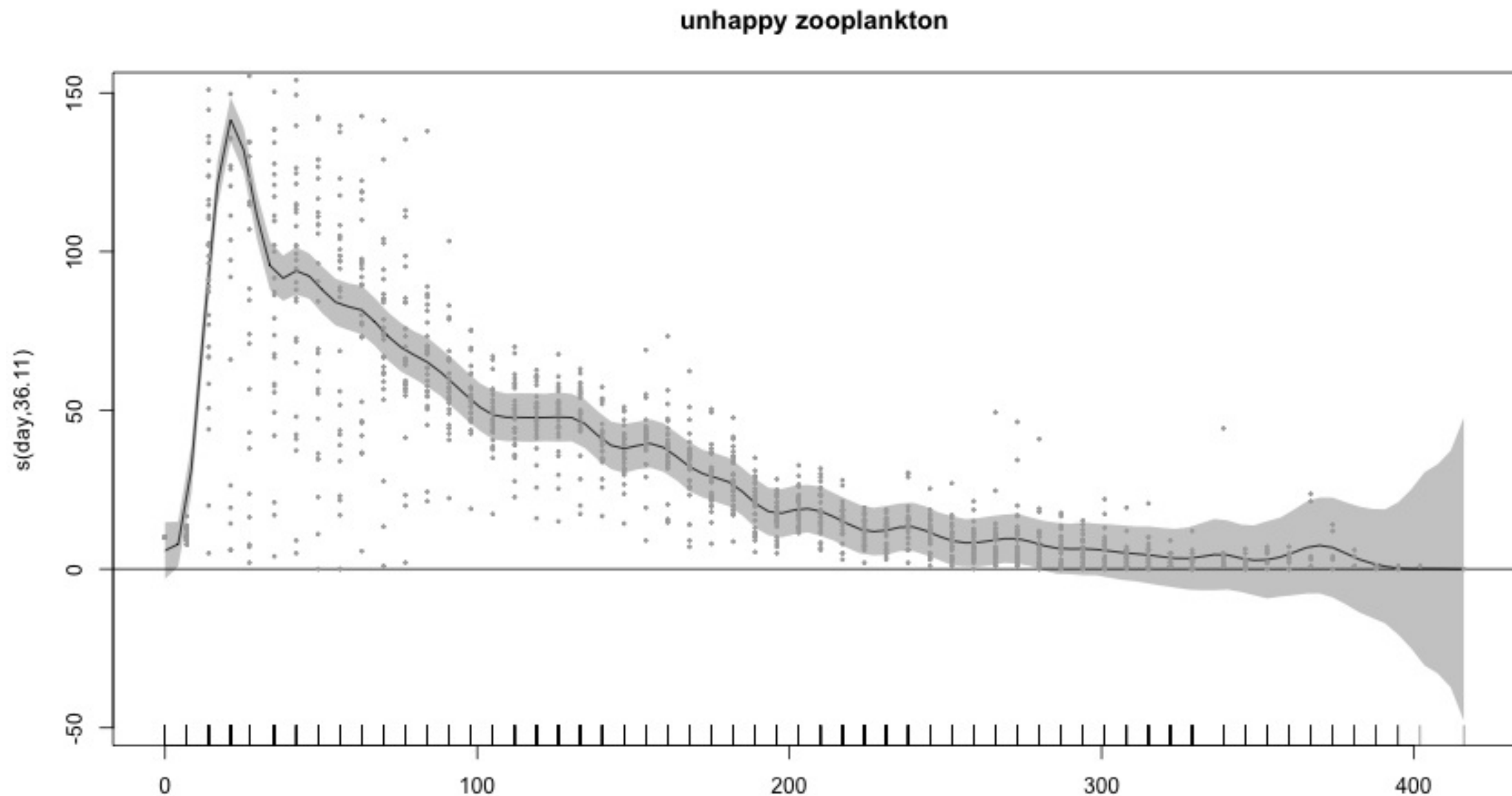
# gam.check() plots

`gam.check()` creates 4 plots:

1. Quantile-quantile plots of residuals
2. Histogram of residuals
3. Residuals vs. linear predictor
4. Observed vs. fitted values

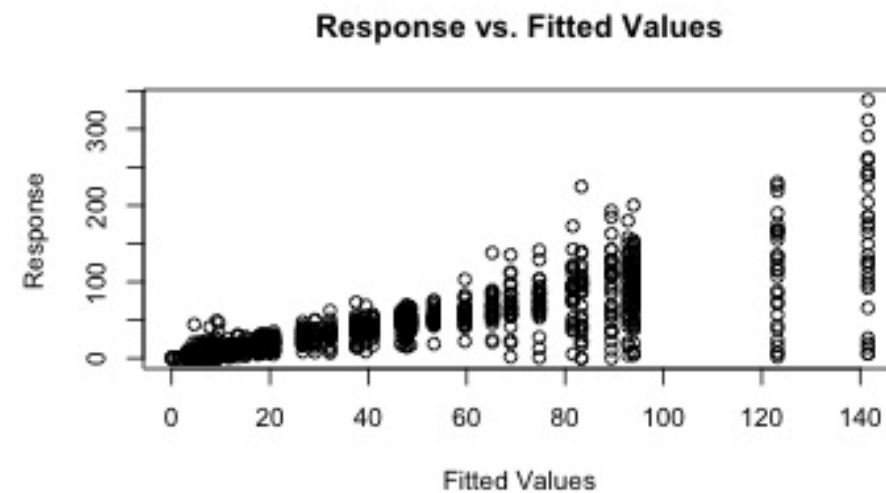
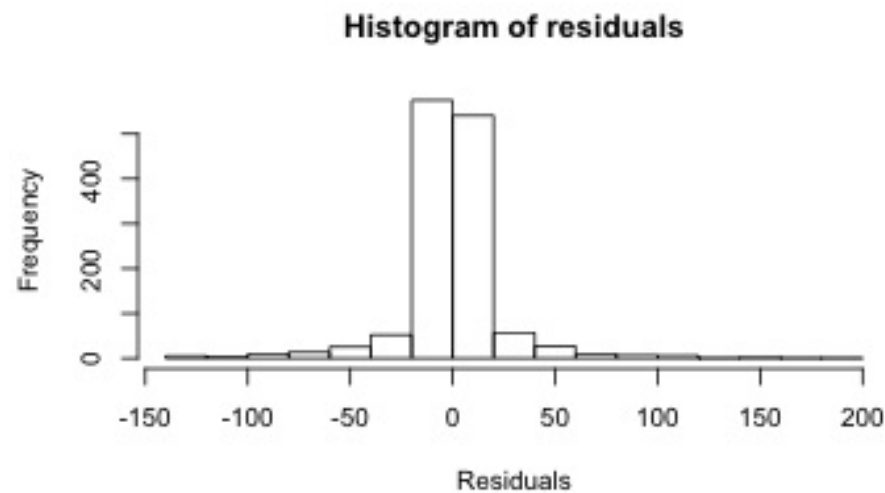
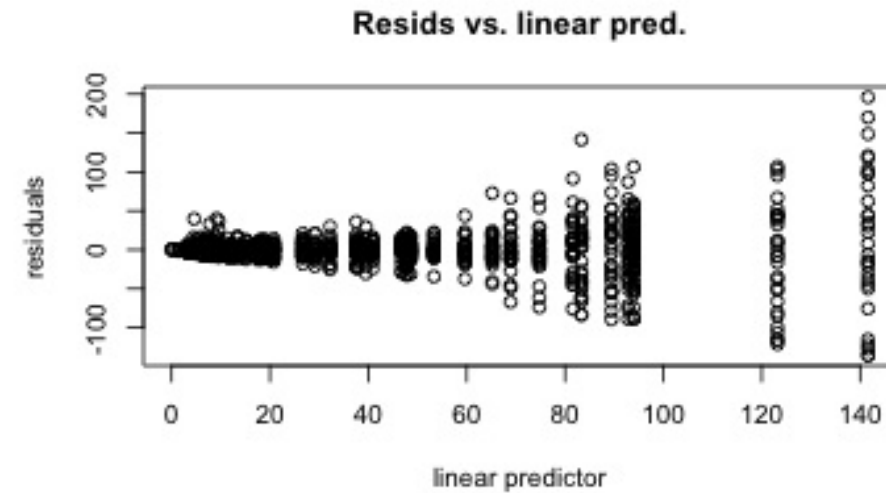
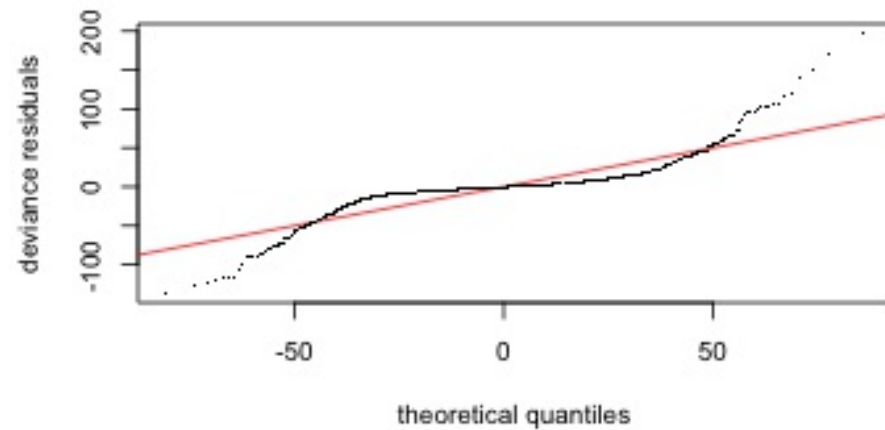
# Checking response distribution

- Left side of `gam.check` plots
- Examples from the Drake & Griffen data
- Looking at the “deteriorating” populations



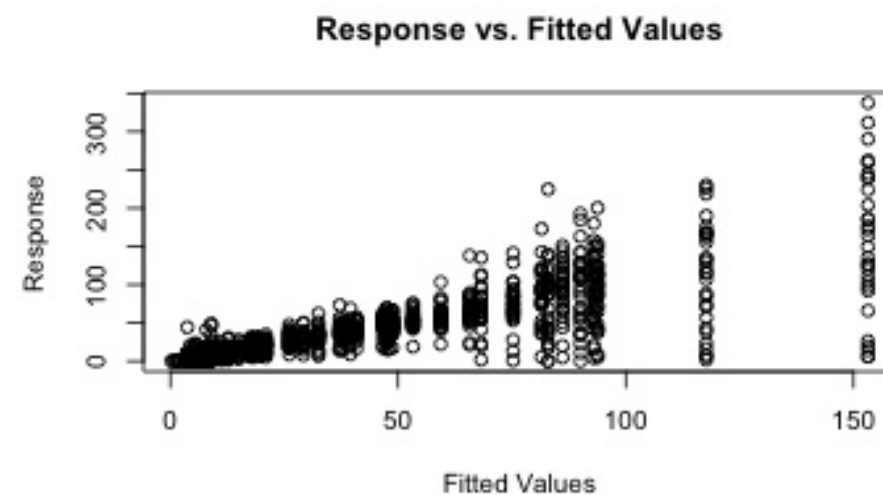
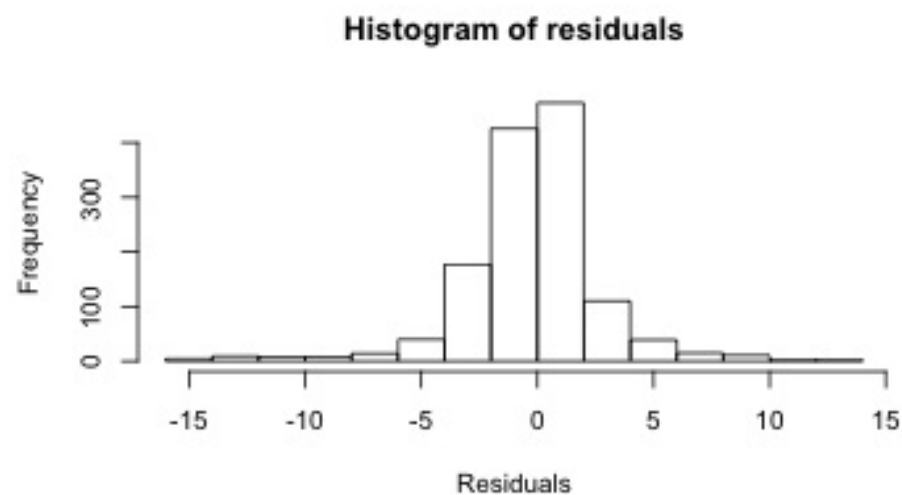
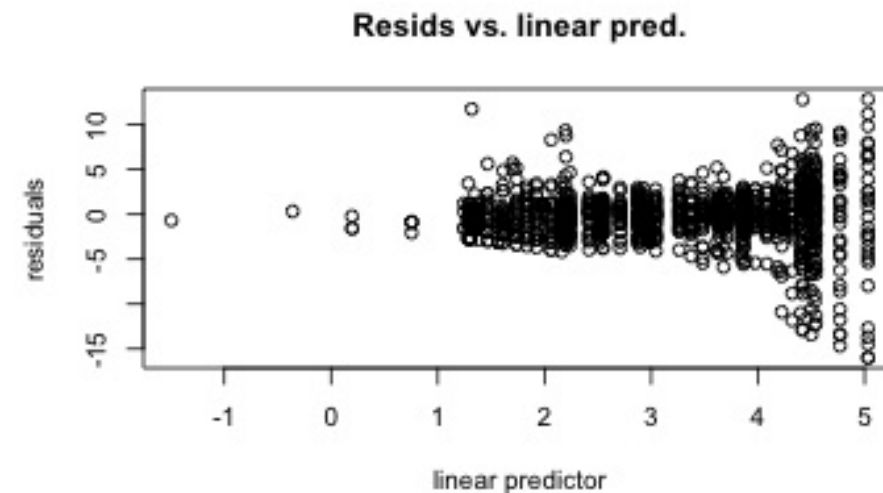
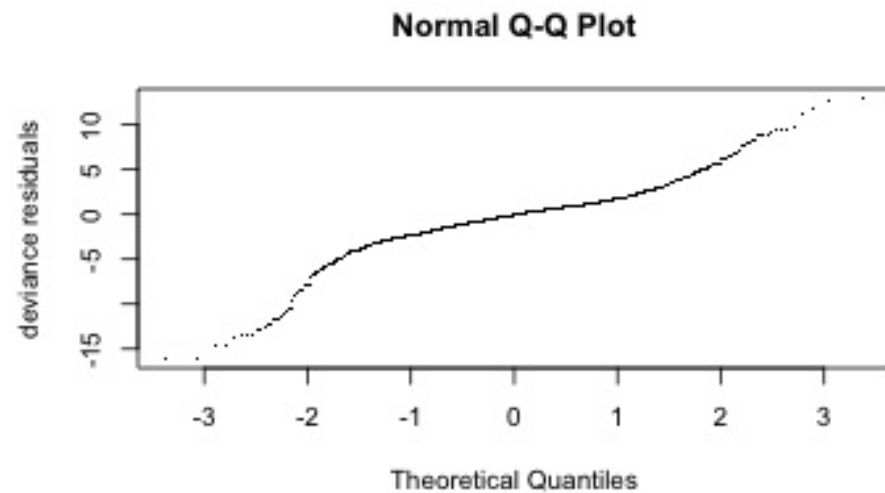
# Normal response with count data

```
b <- gam(Nhat ~ s(day, k=50), data=pop_unhappy, method="REML")
```



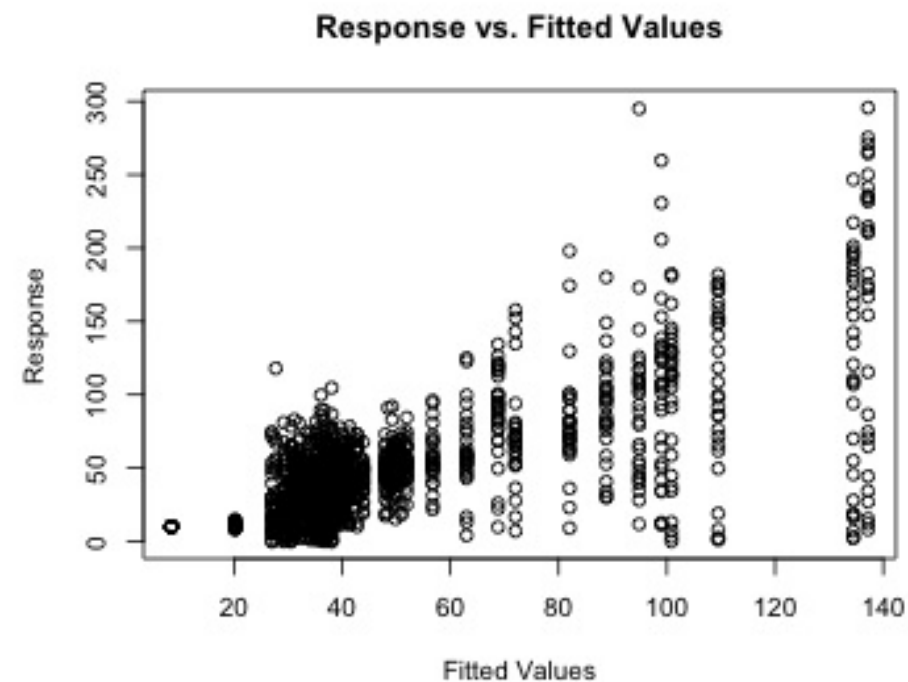
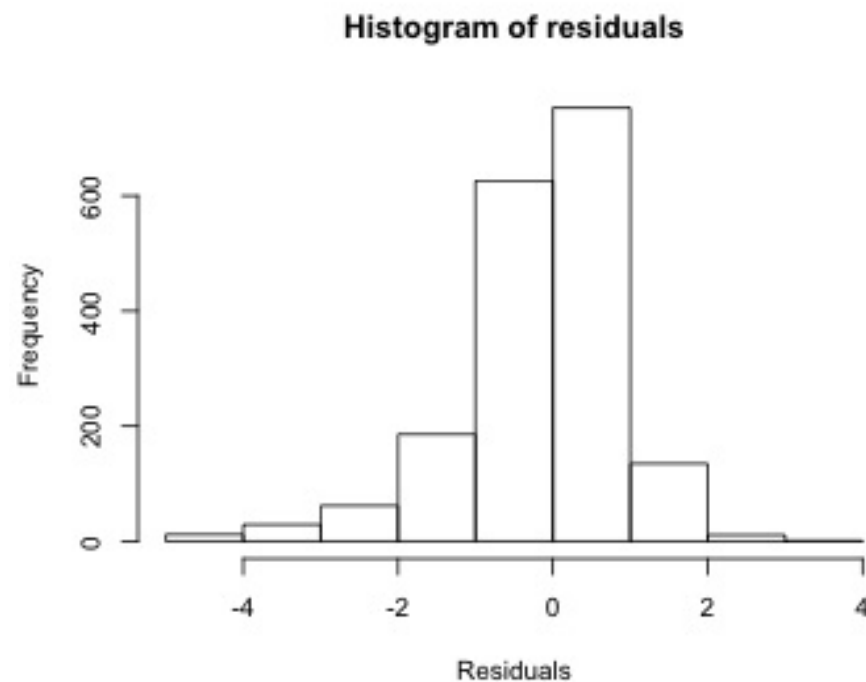
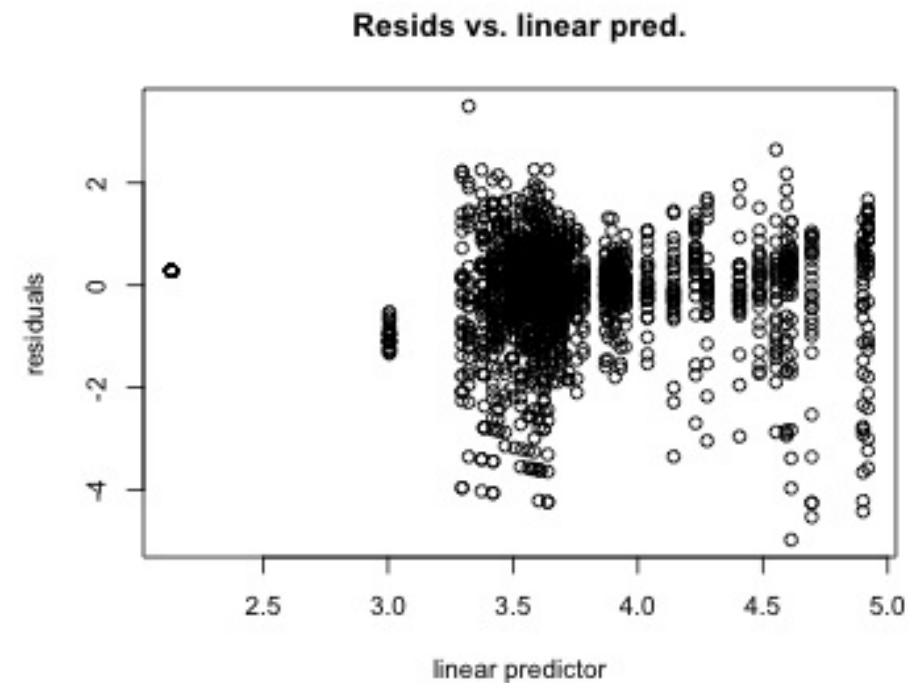
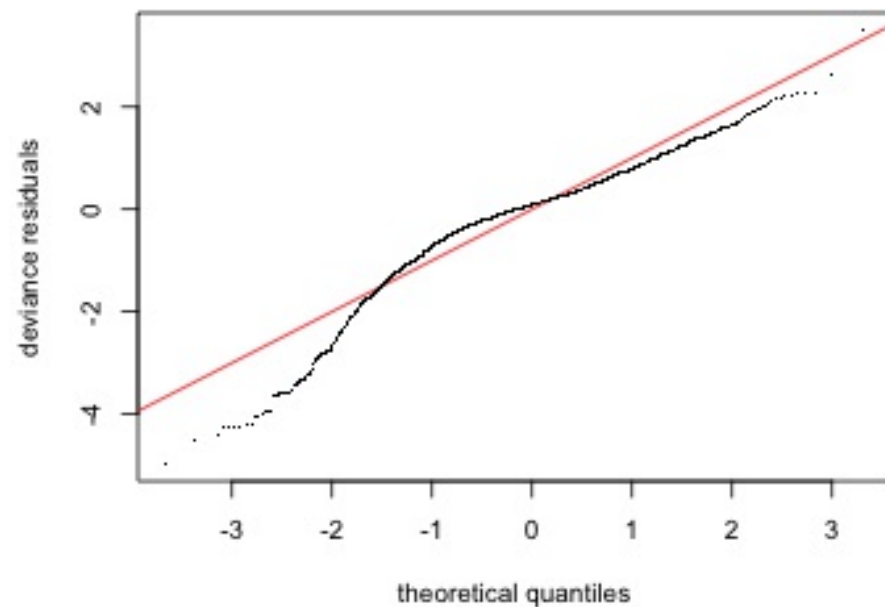
# What about a count distribution?

```
b_quasi <- gam(Nhat ~ s(day, k=50), data=pop_unhappy,  
method="REML", family=quasipoisson())
```



# What about a fancier count distribution?

```
b_nb <- gam(Nhat ~ s(day, k=50), data=pop_happy, method="REML",  
family=nb())
```

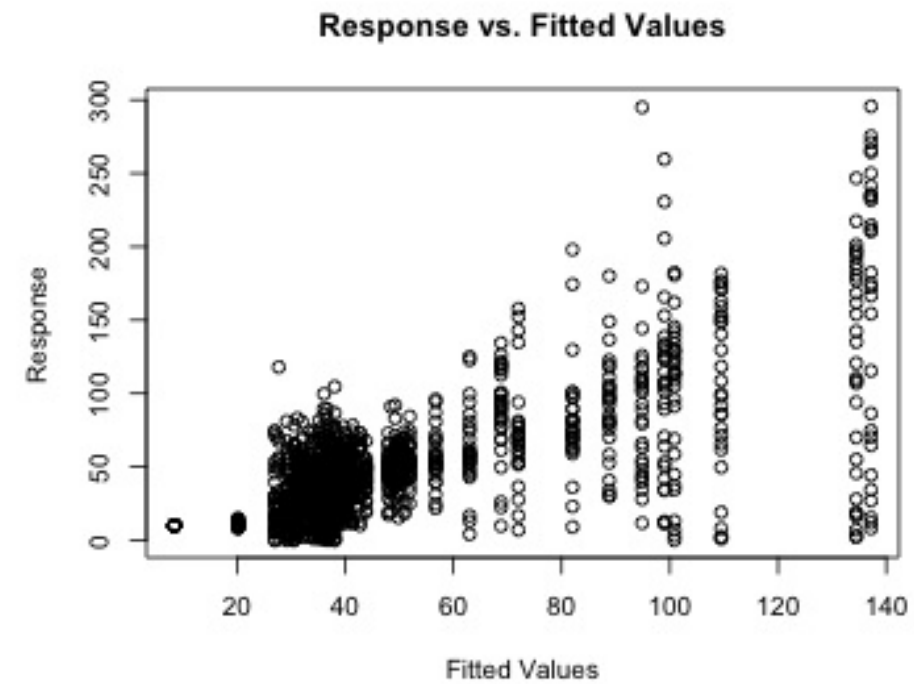
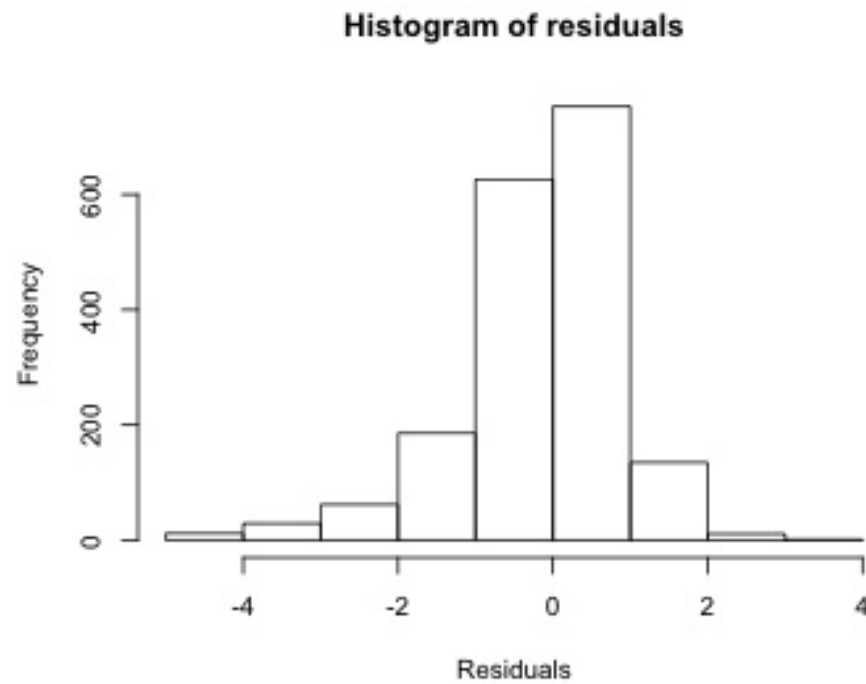
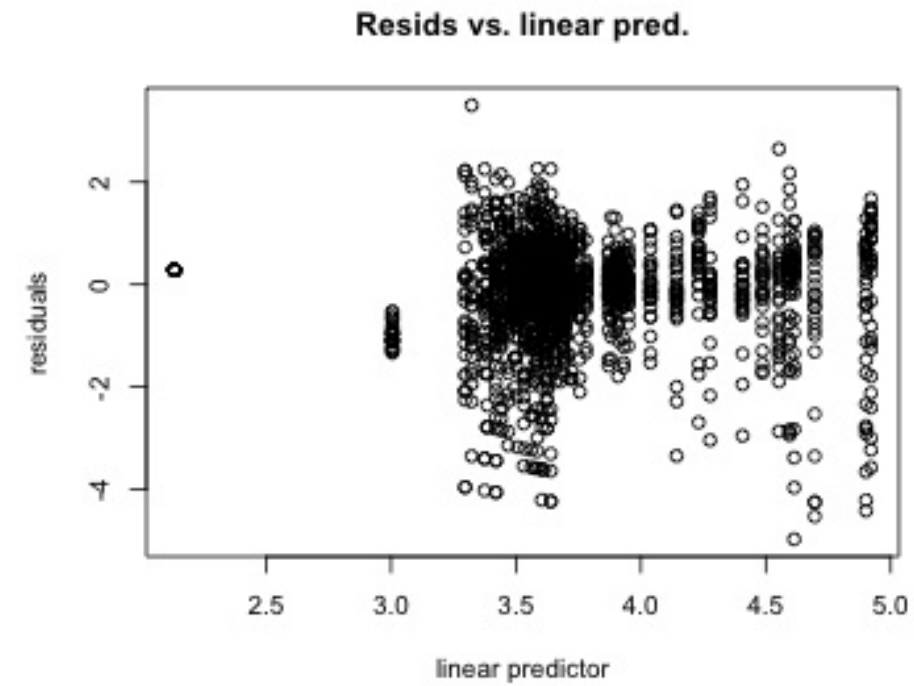
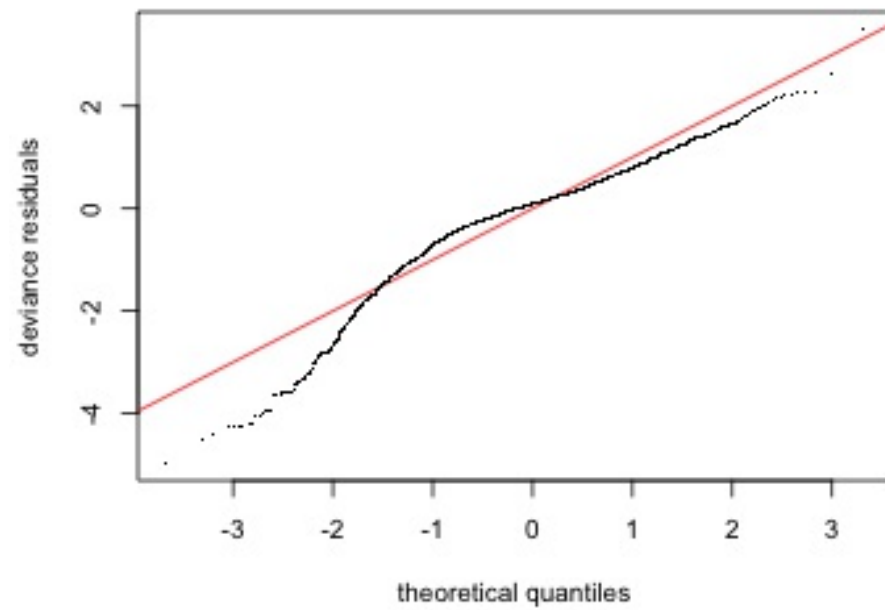




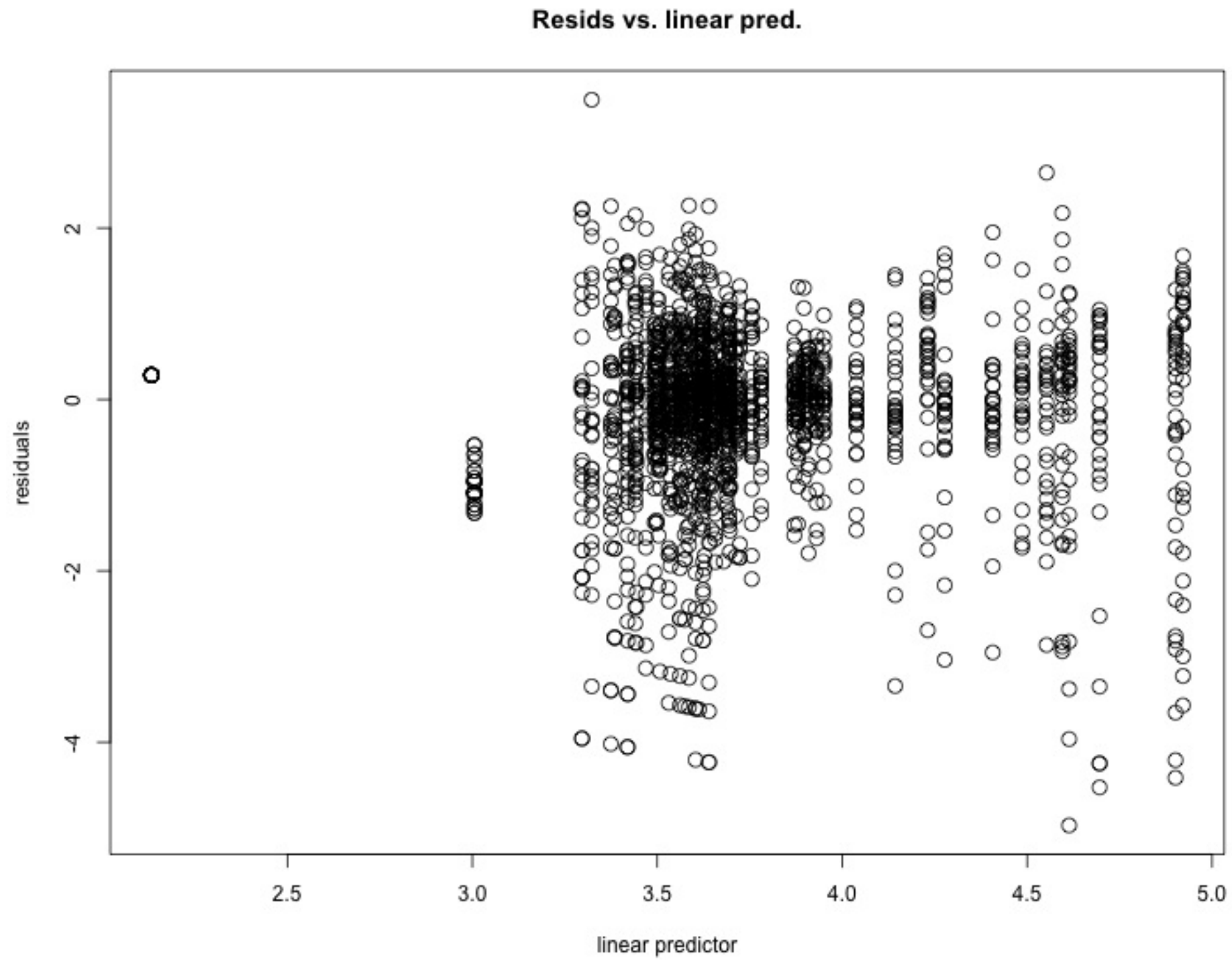
# Variance relationships

- Heteroskedasticity
- Do we know that the mean-variance relationship is right?
- Deviance resids should give us constant variance if model correct?
- Right column of `gam.check`:
  - residuals vs. linear prediction == cloud
  - Response vs. fitted == line-ish

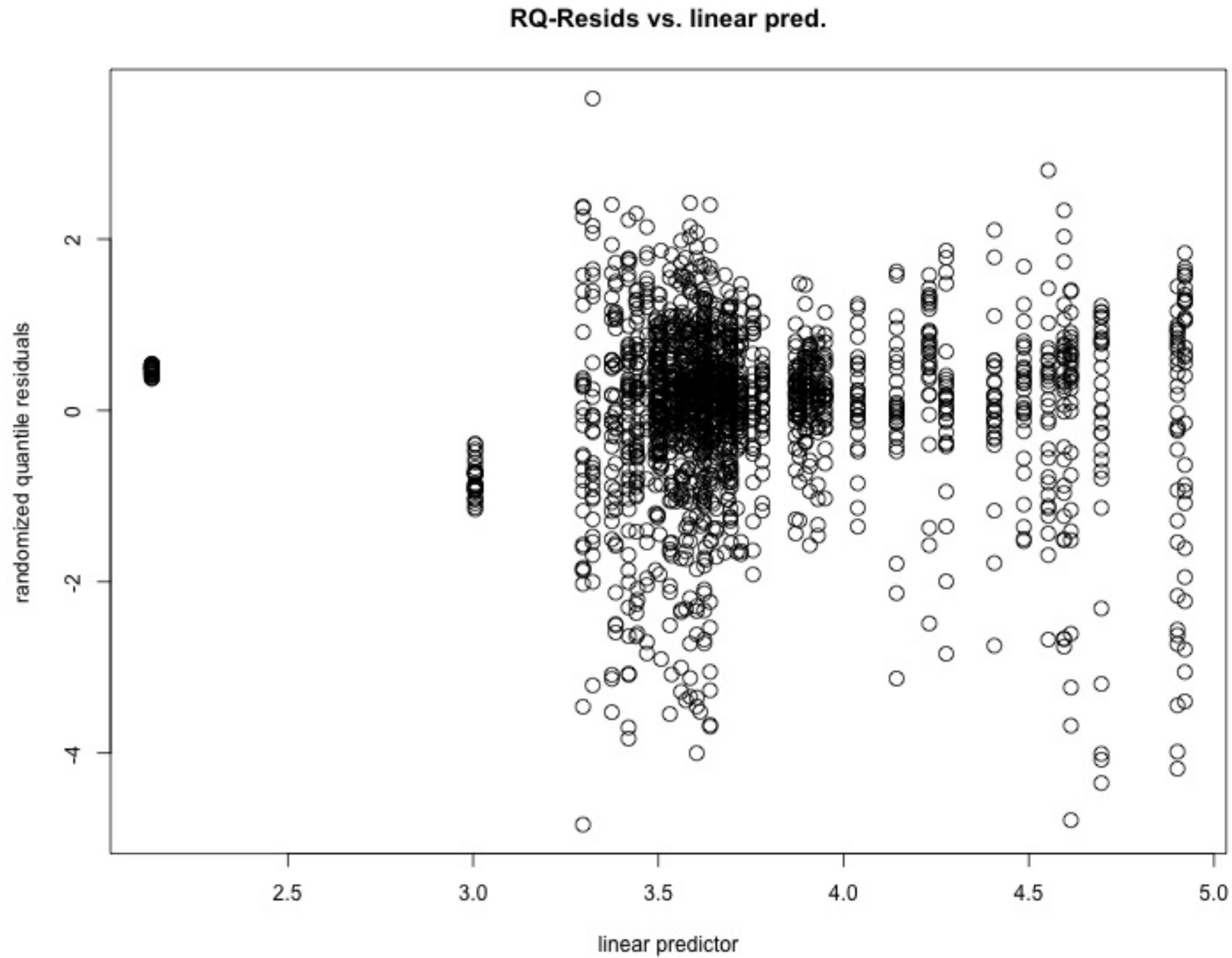
# Mean-variance incorrect



# Close up



# Randomized quantile residuals

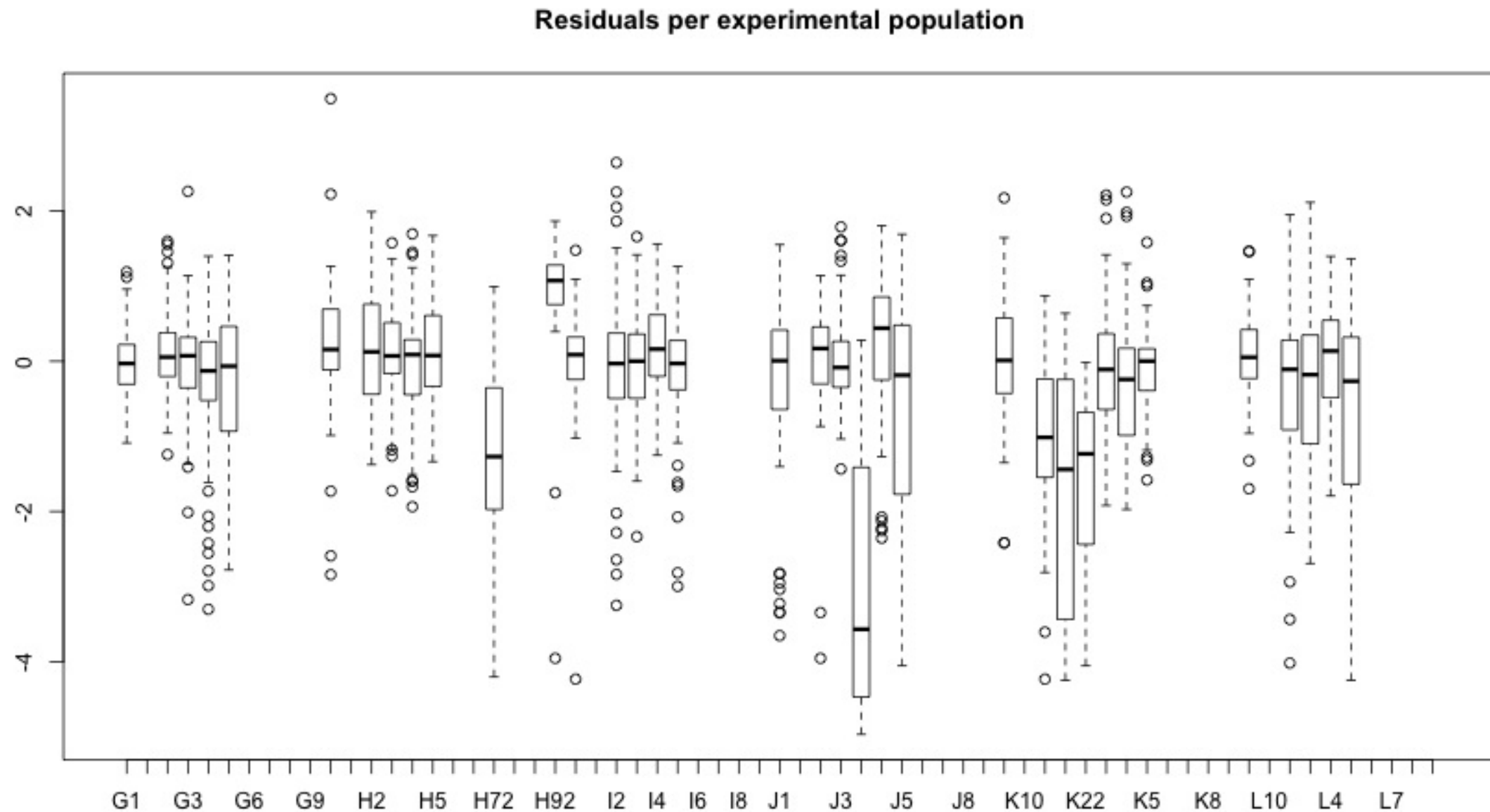


# Shortcomings

- Deviance residuals vs. linear predictor is victim of artifacts
- Need an alternative
- “Randomised quantile residuals” (*experimental*)
  - `rqresiduals` in `statmod`
  - Exactly normal residuals ... if the model is right!
  - `rqgam.check` in `dsm` (ignore left side plots!)

# These plots are just the start

- Need to go further
- Look at aggregations of residuals by other variables



# Residual checking as art form

- Residuals can tell you a **lot** about your model
- No general method
  - Depends on data
  - Depends on inferential goals
- Highlight model deficiencies
- Inform what to do next; which other questions are interesting

# Tobler's first law of geography

“Everything is related to everything else, but near things are more related than distant things”

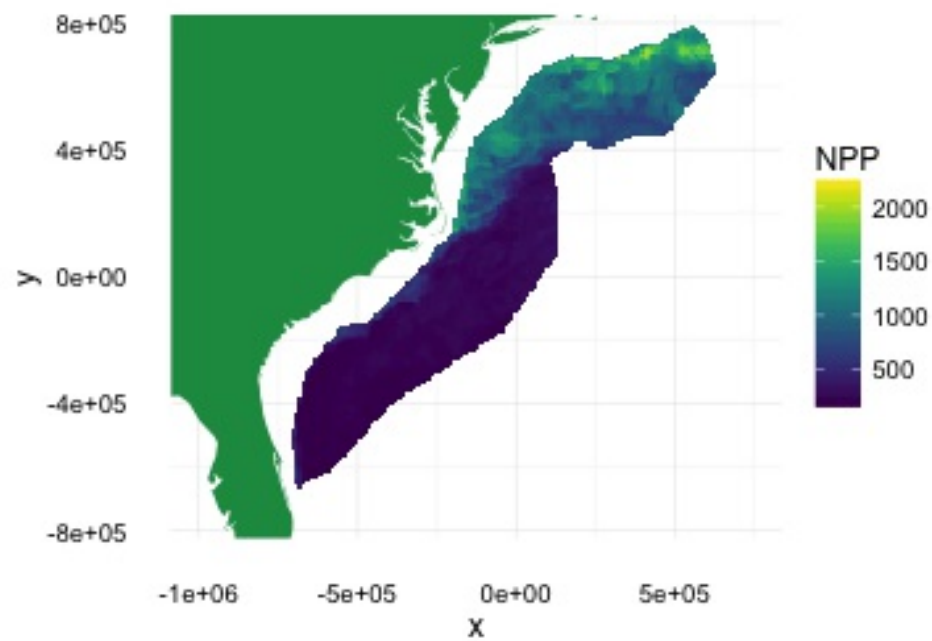
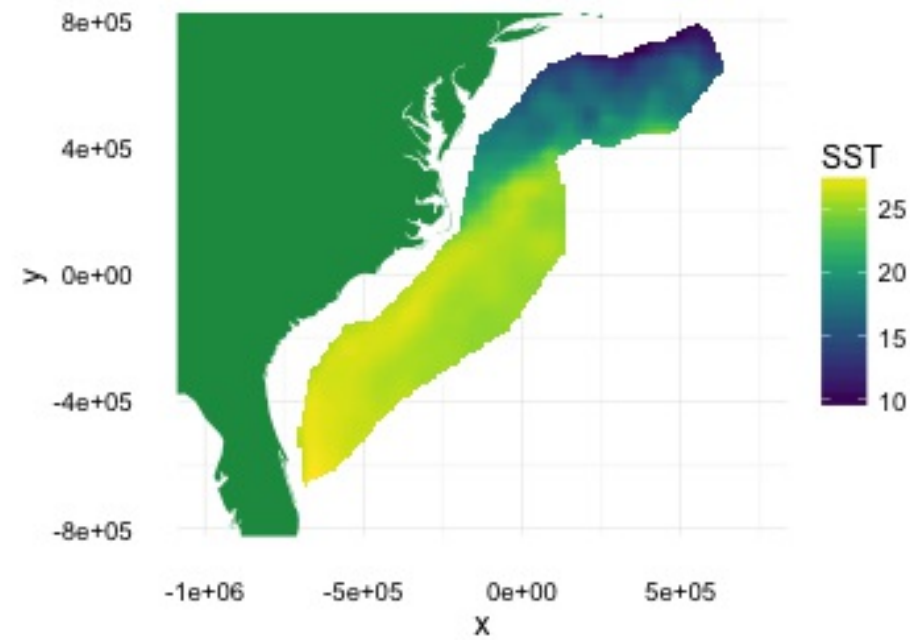
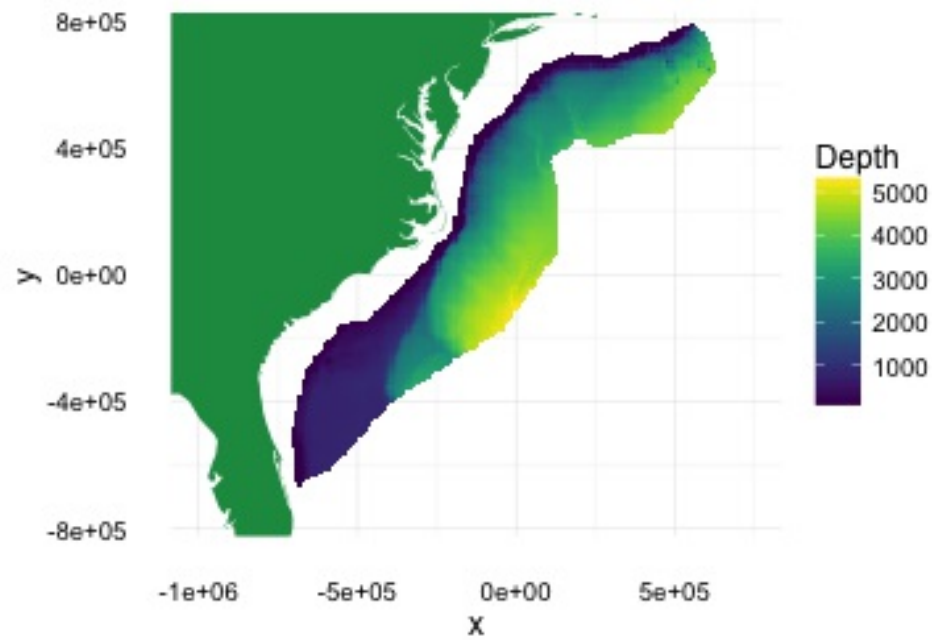
Tobler (1970)



# Concurvity

- We all know correlated covariates are bad
- What about non-linear correlations?
- Can we describe covariates as functions of each other?
- Important for model selection — sensitivity analysis

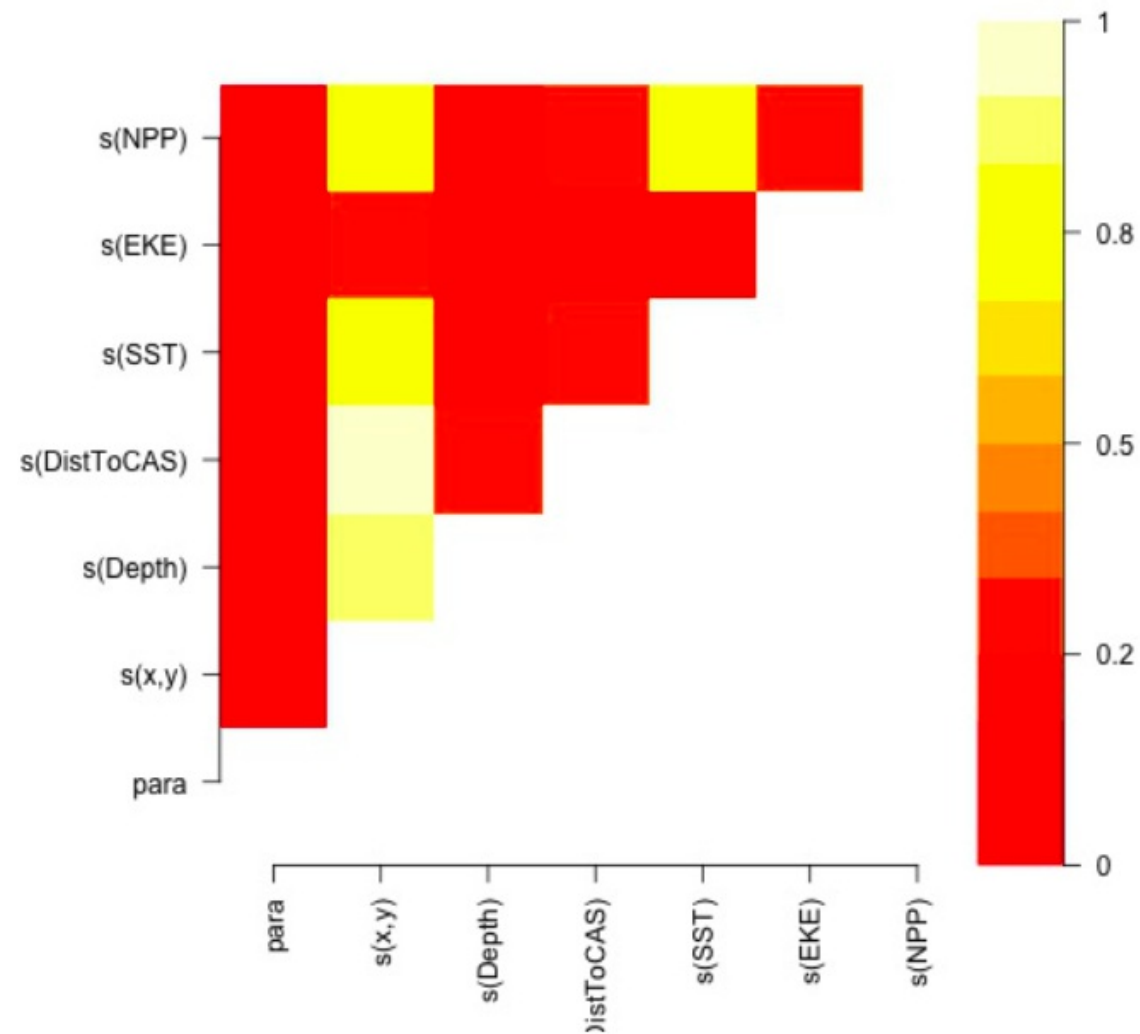
# Example - US East coast



# Checking for concurvity

- Measures, for each smooth term, how well this term could be approximated by
  - `concurvity(model, full=TRUE)`: some combination of all other smooth terms
  - `concurvity(model, full=FALSE)`: each of the other smooth terms in the model (useful for identifying which terms are causing issues)

# Plotting concavity



- We can visualise
- `vis.concavity` on course site

# Concurvity: things to remember

- Can make your model unstable to small changes
- `cor(data)` not sufficient: use the `concurvity(model)` function
- Not always obvious from plots of covariates or smooths