overview of process so far

List of predictor variables

- Forest height: FHT (m)

  range: 0 - 20

  year: 2019

- Length of growing season: LGS (days)

  calculated by subtracting date of onset of greenness increase (15% increase in greeness) from onset greenness decrease (10% decrease in greenness)

  range: 28 - 244 days

  year: 2020

  *there are quite a few spots without data

- Canopy cover: cc (%)

  range: 0 - 100

  year:

  Terra Nova: 2010 with gaps filled in with 2005 (2015 had most of the park missing)

  Gros Morne: 2015 with gaps filled in with 2010

- Enhanced vegetation index: EVI (NA)
  - amplitude: EVIamp, maximum minus minimum EVI during that year “interpreted as indicating the degree to which the leaf life cycles of individual plants and species are synchronized” (Gerard et al. 2020)
  - median: EVImed, median EVI during that year

  range: 16 - 9600

  year: 2019

- Aspect: ASP (degrees)

  range: 0-360

  year: 2011

*when slope is 0 there can be no aspect, in these instances I put 0.01˚as aspect

- Slope: SLO (percent)

  range: 0 - 260

- Elevation: ELE (m)

- Dominant species: SPC (categorical) * decided to switch from dominant species to forest type because I didn't sample all dominant species

  range: 8 categories (CS = coniferous shrub, DS = deciduous shrub, NS = not sufficiently restocked, bf = balsam fir, bs = black spruce, tA = trembling aspen, tL = tamarack, wB = white birch)

  year: 2010

- Land cover type: CEC_LCT (categorical)

  range: 5 categories (1 = temperate or sub-polar needle leaf forest, 6 = mixed forest, 8 = temperate or sub-polar shrubland, 14 = wetland, 16 = barrenland)

  year: 2020

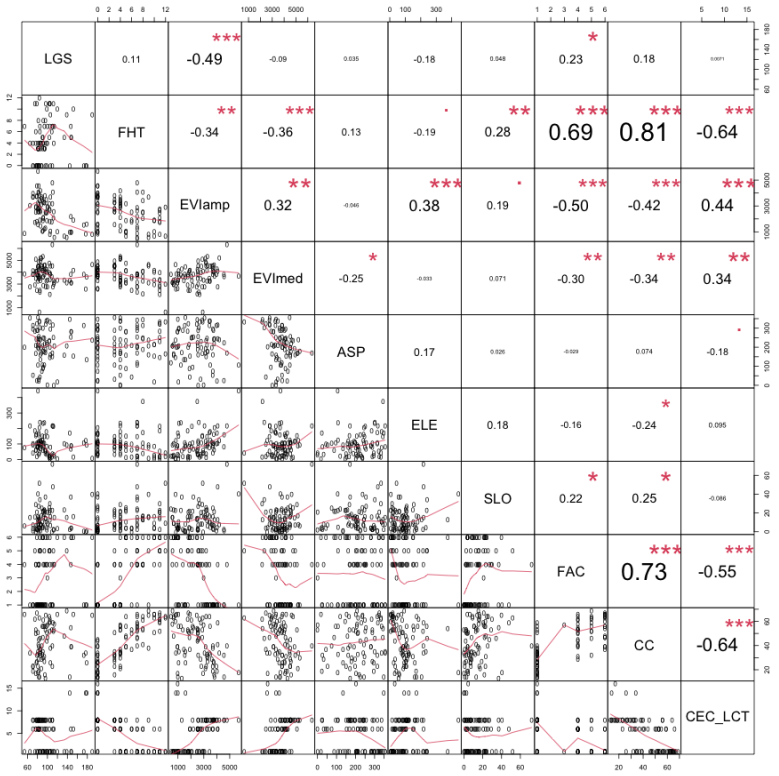
- Age class: FAC (categorical)

  range: 5 categories (1 = 0 - 20, 2 = 21 - 40, 3 = 41 - 60, 4 = 61 - 80, 5 = 81+)

Correlation charts

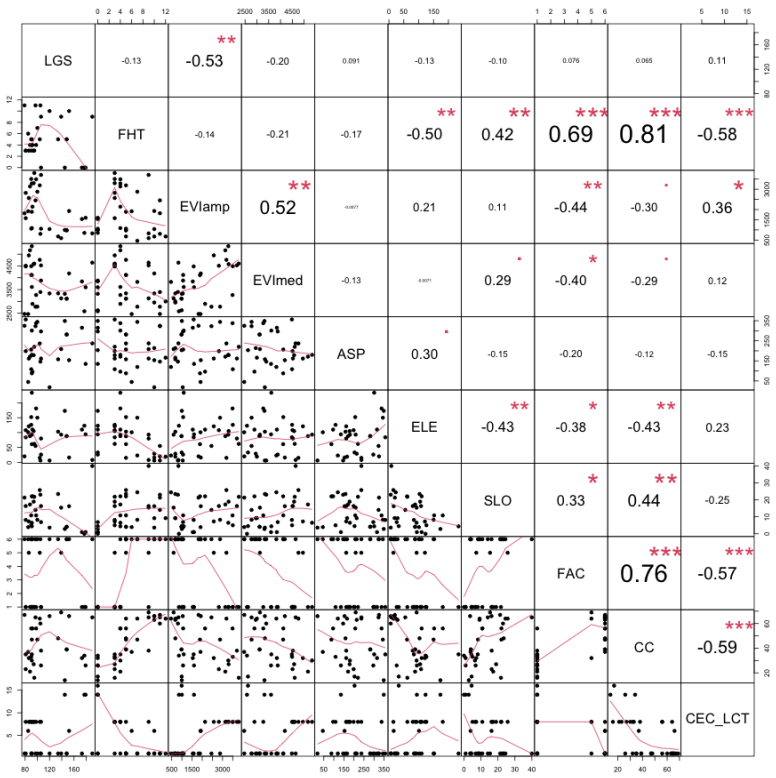
at sites

across both parks



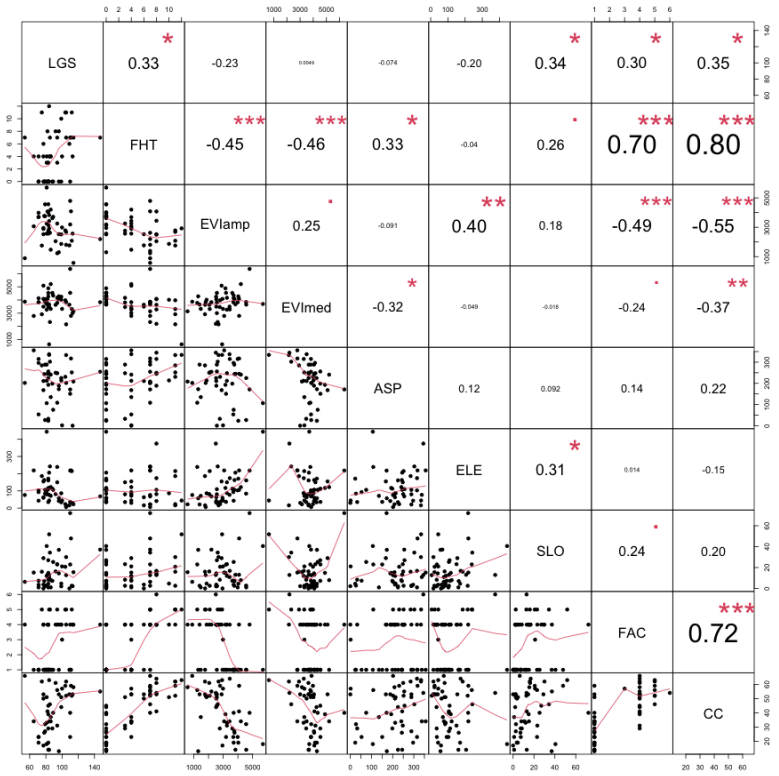
Variables <chr>	VIF <dbl>
LGS	1.4141…
FHT	3.1142…
EVIamp	3.0911…
EVImed	1.4256…
ASP	1.1722…
ELE	1.3655…
SLO	1.3745…
FAC	2.5918…
CC	3.3890…
CEC_LCT	2.9248…

TN



Variables <chr>	VIF <dbl>
LGS	1.4395…
FHT	3.6099…
EVIamp	1.9945…
EVImed	1.3777…
ASP	1.1939…
ELE	1.4747…
SLO	1.6086…
FAC	2.7403…
CC	3.6019…

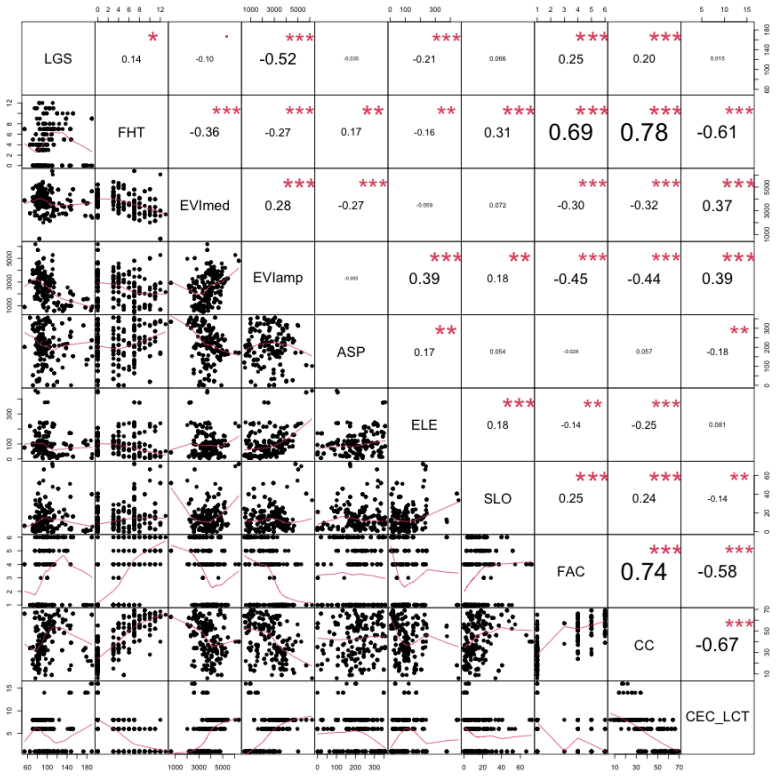
GM



Variables <chr>	VIF <dbl>
LGS	1.7306...
FHT	3.5352...
EVIamp	4.0959...
EVImed	2.1202...
ASP	2.0052...
ELE	1.6722...
SLO	1.7053...
FAC	3.3277...
CC	3.7500...
CEC_LCT	4.1000...

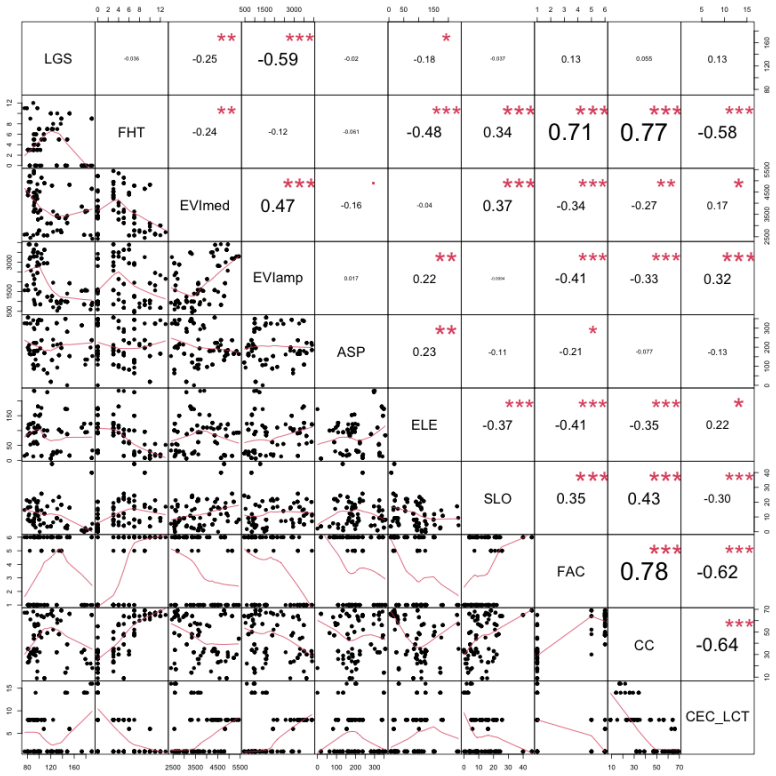
for subplots

across both parks



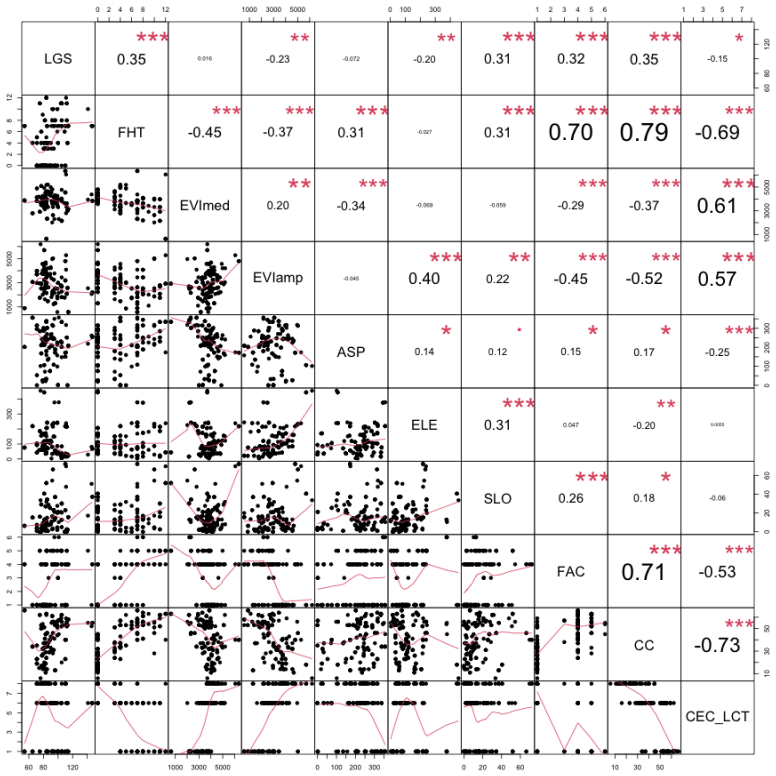
Variables <chr>	VIF <dbl>
LGS	1.5367...
FHT	2.9675...
EVImed	1.4739...
EVIamp	2.6100...
ASP	1.1919...
ELE	1.3763...
SLO	1.3949...
FAC	2.7363...
CC	3.5036...
CEC_LCT	2.6100...

TN



Variables <chr>	VIF <dbl>
LGS	1.4528…
FHT	3.9317…
EVImed	1.8377…
EVIamp	2.5037…
ASP	1.2309…
ELE	1.7139…
SLO	1.6009…
FAC	2.9003…
CC	4.1071…
CEC_LCT	4.1376…

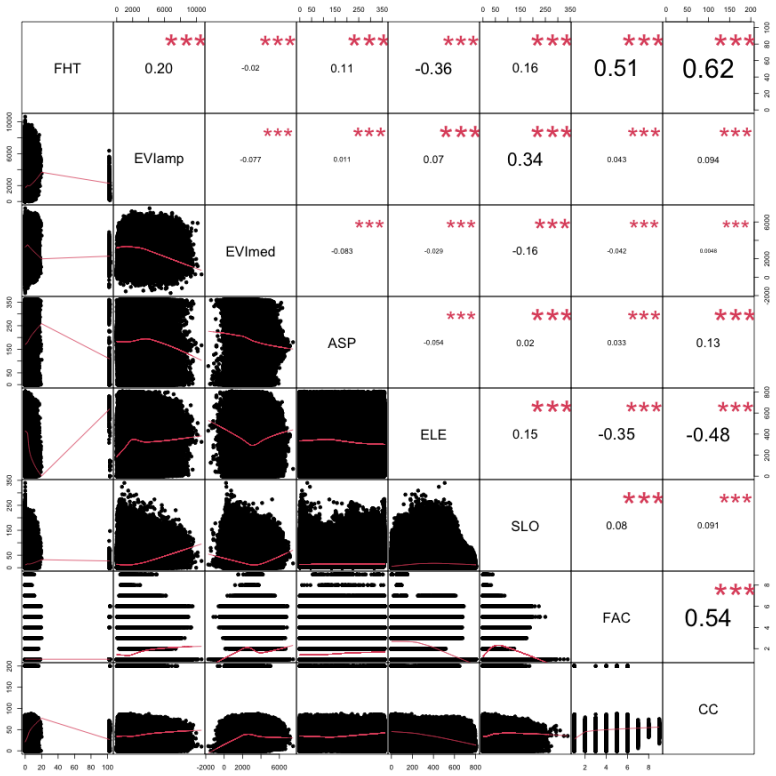
GM



Variables <chr>	VIF <dbl>
LGS	1.9334…
FHT	3.1910…
EVImed	2.2809…
EVIamp	3.2996…
ASP	1.4758…
ELE	1.6684…
SLO	1.8180…
FAC	3.6897…
CC	3.5895…
CEC_LCT	2.9254…

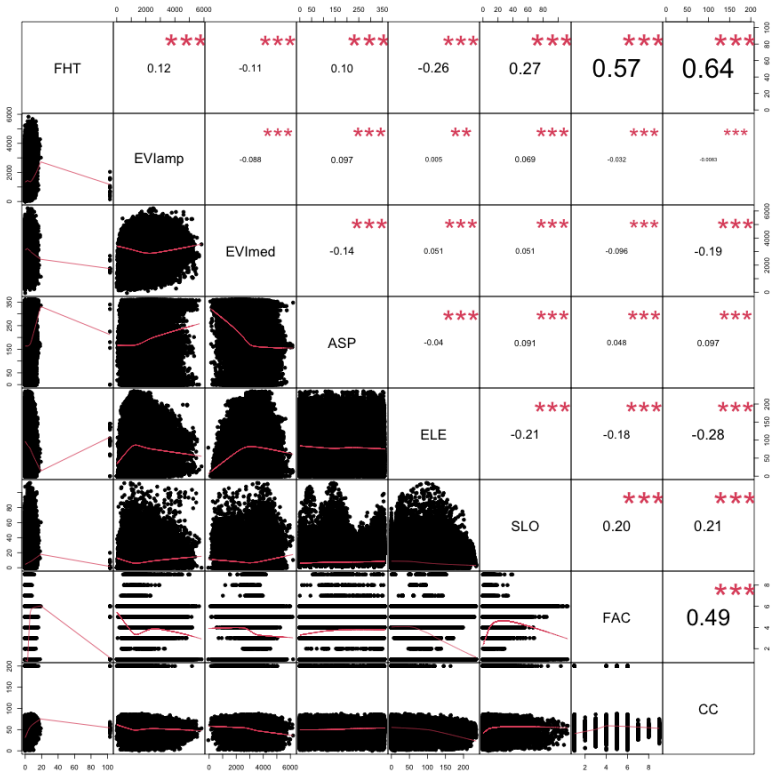
for whole parks

across both parks



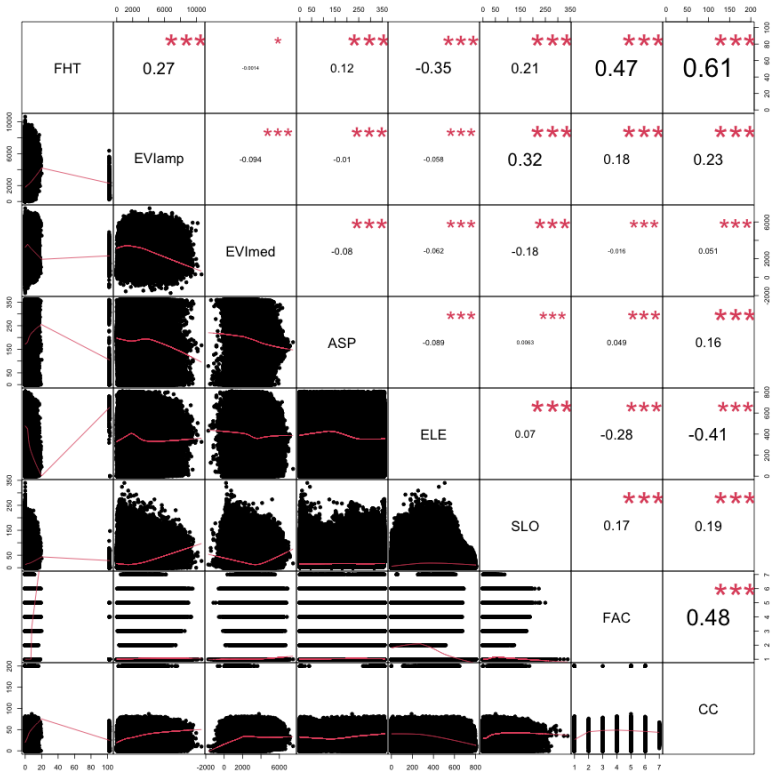
Variables <chr>	VIF <dbl>
FHT	2.3911…
EVIamp	1.2241…
EVImed	1.0364…
ASP	1.0305…
ELE	1.4404…
SLO	1.2438…
FAC	1.6075…
CC	2.2329…

TN



Variables <chr>	VIF <dbl>
FHT	1.8760…
EVIamp	1.1895…
EVImed	1.0542…
ASP	1.0400…
ELE	1.3228…
SLO	1.2188…
FAC	1.4365…
CC	1.9178…

GM



Variables <chr>	VIF <dbl>
FHT	2.2183…
EVIamp	1.0677…
EVImed	1.0796…
ASP	1.0346…
ELE	1.1259…
SLO	1.1311…
FAC	1.6113…
CC	2.0727…

Conclusions

The patterns are quite consistent across each subset of the data.

My conclusions are predominantly based on the correlation charts, but I also am influenced a bit by the fact that none of the variance inflation factors are above 4.2, with most being below 2.

Remove CC

Canopy cover is very correlated with forest height. And although I have no biological reasoning for keeping forest height over canopy cover, the forest height data set is more complete and recent.

Remove FAC

The forest age is technically categorical, but the categories are ranges of a continuous variable. Is it ok to use it in a correlation chart?

If it is ok to use I propose removing FAC. It is very correlated with height. And I would prefer to keep height because it is actually a continuous variable.

Remove LCT

Land cover type is also actually a categorical variable so it's interesting that it is so correlated. It is probably because the higher numbers describe less vegetated land cover types (eg; 1 is temperate needle forest, 8 is shrubland and 16 is barrenland).

The only data we are really losing by removing it is whether the forested areas are needle leaf or mixed forest.

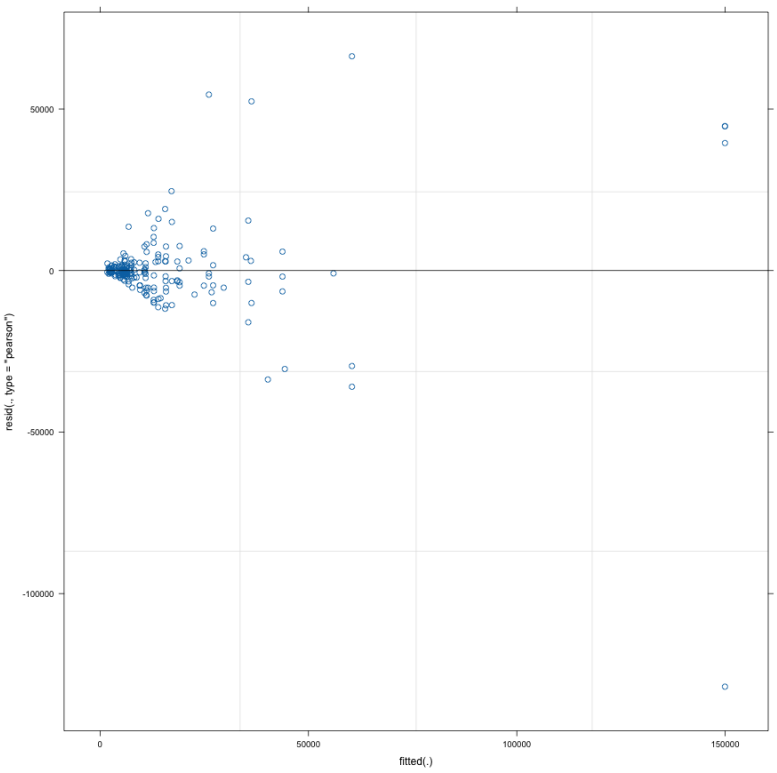
Residual distributions from GzLMs

for full model in Gros Morne with C per m2 of each subplot used as response variable and a random effect of site. Predictor variables have been scaled.

C ~ FHT + EVIamp + EVImed + ELE + SLO + ASP + (1|site_id)

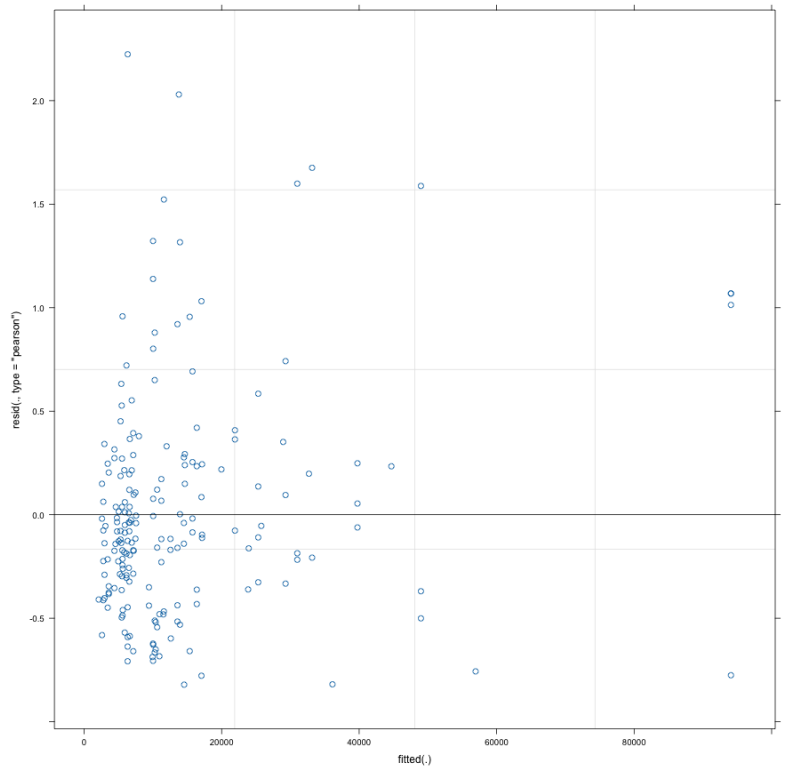
Gaussian

Log link



Gamma

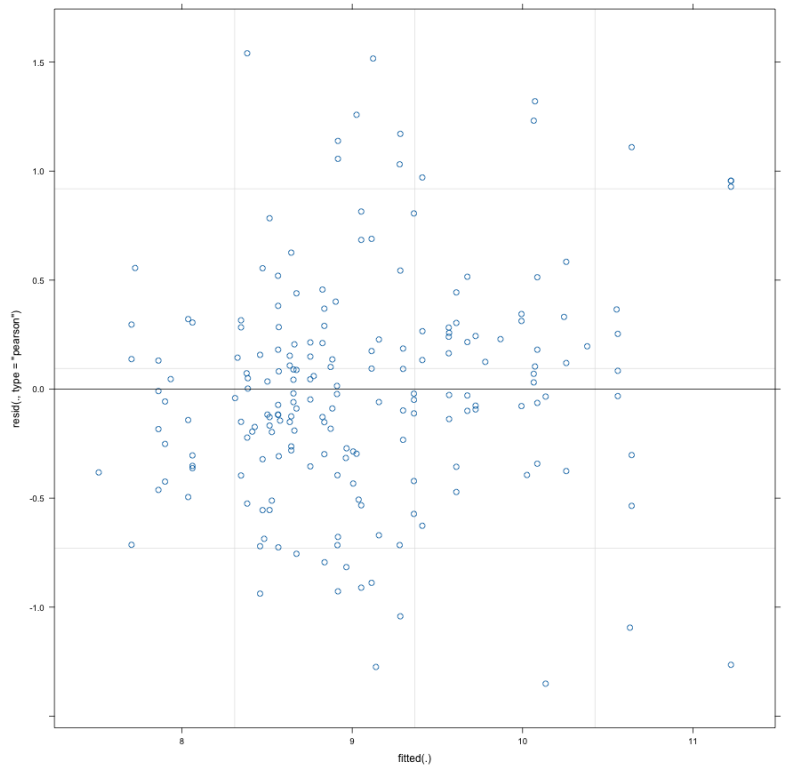
log link



Inverse and identity links don't converge

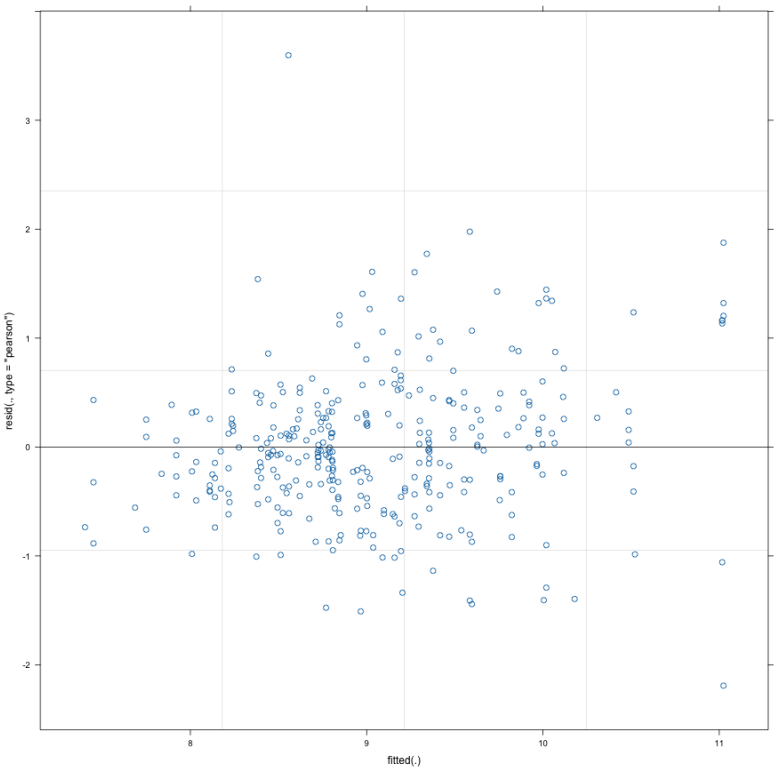
log tranformed linear model

scaled predictors



not scaled predictors: give warning

Warning: Some predictor variables are on very different scales: consider rescaling



Linear Models

subsetting data

park level

by park -> GM and TN get own models

subplot or site

sites -> average C per m2 across the 4 subplots at each site

subplots -> use C per m2 of each subplot and have random effect of site to account for pseudoreplication

structure

i. log(C) ~ ε

ii. log(C) ~ (1|site) + ε

iii. log(C) ~ all env. + ε

iv. log(C) ~ all env. + (1|site) + ε

v. log(C) ~ reduced env. + ε

vi. log(C) ~ reduced env. + (1|site) + ε

vii. log(C) ~ univariate + ε

viii. log(C) ~ univariate + (1|site) + ε
