Linear Mixed Models - Questions

Lance Stasinski

9/27/2021

Linear Mixed Models - Questions

This document walks through the full process of creating linear mixed models with the lichen spectroscopy data, and the 9 questions I currently have can be found in bold along the way.

Research Questions

1. How do reflectance measurements taken from a lichen thallus change as lichen specimens age?

- Do lichen show a general pattern? i.e. Do the underlying traits change in a consistent manner between lichen species, such that only the starting values differentiate species rather than the rates of change for those corresponding traits.
- Are there differences in the rate of reflectance change between species? I'm not concerned with exactly how each species responds to age.

2. Where does much of the variation in either the rate of reflectance change or the starting reflectance values arise? i.e. Does much of the variation occur between species or at higher taxonomic ranks such as order or class?

Linear mixed models appear to be a good choice for answering these questions.

Data structure

Full-range spectra (400 to 2400nm) are available at 1nm resolution for 30 species that represent 19 families, 16 orders, and 6 classes. The spectra per each species were taken from a type II chronosequence (trading space for time) of aging lichen thalli. Each thallus is represented by 4 reflectance measurements from various parts of the thallus to capture spectral variation (lichen surfaces can be quite heterogenous!). The amount of data is not equal among species or higher taxonomic ranks or equal across the time scale (imbalanced).

Number of measurements per species

```
setwd("~/GitHub/Lichen-Herbarium-Spectra")

library(spectrolab)
```

```
## Warning: package 'spectrolab' was built under R version 4.1.1
```

```
## spectrolab version: 0.0.14
##
## Please cite:
## Meireles J, Schweiger A, Cavender-Bares J (2017). spectrolab: Class
## and Methods for Spectral Data in R. doi: 10.5281/zenodo.3934575 (URL:
## https://doi.org/10.5281/zenodo.3934575), R package version 0.0.14,
## <URL: https://CRAN.R-project.org/package=spectrolab>.DOI: https://doi.org/10.5281/zenodo.3934575

##
## Attaching package: 'spectrolab'

## The following objects are masked from 'package:stats':
##
##     sd, smooth, var
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:spectrolab':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
spectra = readRDS('spectra/lichen_spectra.rds')
spec_df = as.data.frame(spectra)
spec_df %>% count(scientificName)
```

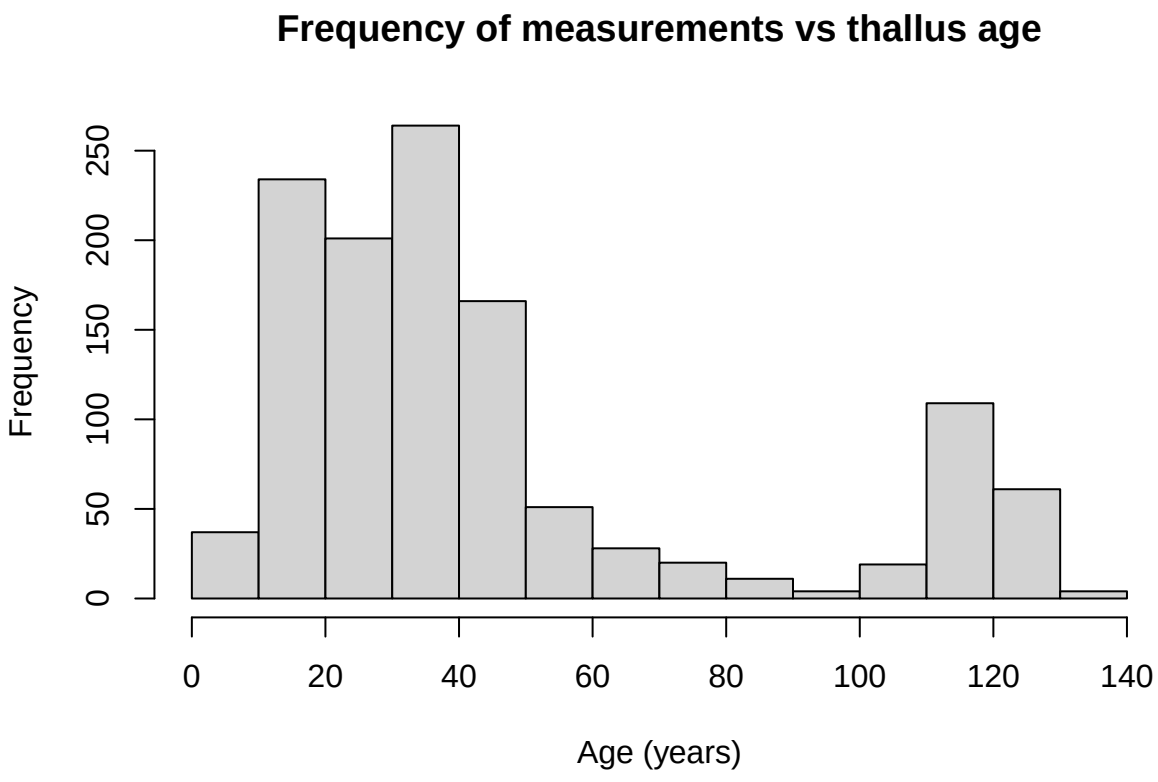
```
##                 scientificName   n
## 1         Acarospora_americana  32
## 2              Baeomyces_rufus  61
## 3     Caloplaca_flavovirescens  60
## 4          Candelaria_concolor  62
## 5       Chrysothrix_candelaris  52
## 6             Dimelaena_oreina  60
## 7              Ephebe_ocellata  24
## 8  Flavoparmelia_baltimorensis  17
## 9       Flavoparmelia_caperata 119
## 10      Flavoparmelia_euplecta   8
## 11      Flavoparmelia_haysomii  20
## 12      Flavoparmelia_rutidota   4
## 13     Flavoparmelia_soredians   8
## 14     Flavopunctelia_darrowii   4
```

```
## 15   Flavopunctelia_flaventior  12
## 16   Flavopunctelia_praesignis  21
## 17     Flavopunctelia_soredica  16
## 18             Graphis_scripta  44
## 19          Ionaspis_lacustris  52
## 20          Lecidea_tessellata  44
## 21           Loxospora_elatina  56
## 22   Neofuscelia_verruculifera   4
## 23       Peltigera_elisabethae  68
## 24      Pertusaria_ophthalmiza  64
## 25          Rhizocarpon_grande  55
## 26      Strigula_submuriformis  48
## 27         Trypethelium_virens  79
## 28   Umbilicaria_muehlenbergii  48
## 29         Verrucaria_fuscella  51
## 30     Xanthoparmelia_darrowii  16
```

NOTE: This table shows the number of measurements per species, not number of individual thalli.

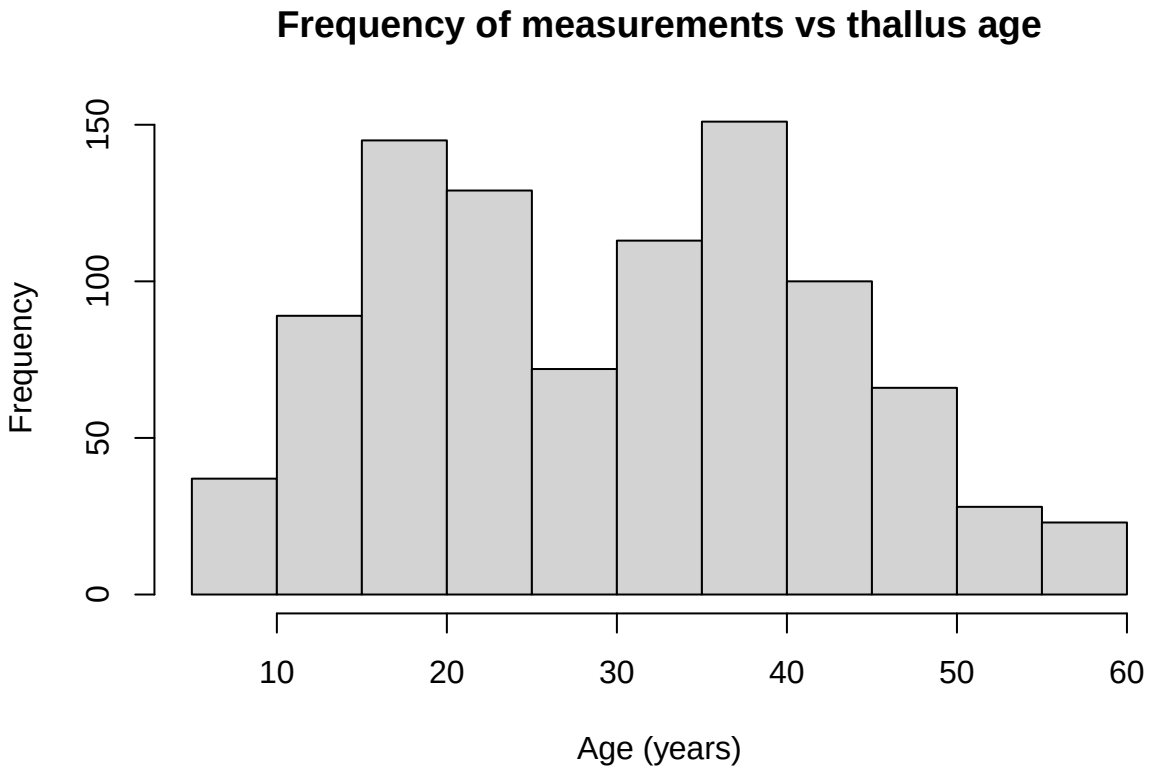
Histogram of measurements across speciemn age

```
hist(spec_df$age, main = 'Frequency of measurements vs thallus age', xlab = 'Age (years)')
```



Clearly there is an issue with the bimodal distribution of age, so I will elect to limit the dataset to only measurements taken on specimens 60 years old and younger.

```r
spec_df = spec_df[spec_df$age <= 60, ]
hist(spec_df$age, main = 'Frequency of measurements vs thallus age', xlab = 'Age (years)')
```



This seems a little bit better for understanding how reflectance changes as thalli age.

Question 1: Does this age range seem appropriate? Should the age range be trimmed further? Perhaps a range of 10-50 years old would allow for more balance in the data (there would be loss of information on the early part of the aging process, yet there's not much to begin with).

Model Assumptions

Do the data fit the assumptions required by linear mixed models/the underlying linear model? I will use 3 wavelengths, 1 from each of the three spectral regions (VIS, NIR, SWIR), to get a rough estimate as to how the data fit the model assumptions. I'll use wavelengths 550, 850, and 1550 because they are near the middle of their respective spectral regions and are not part of major or minor waterbands which could potentially throw off estimates.

The underlying OLS model assumptions

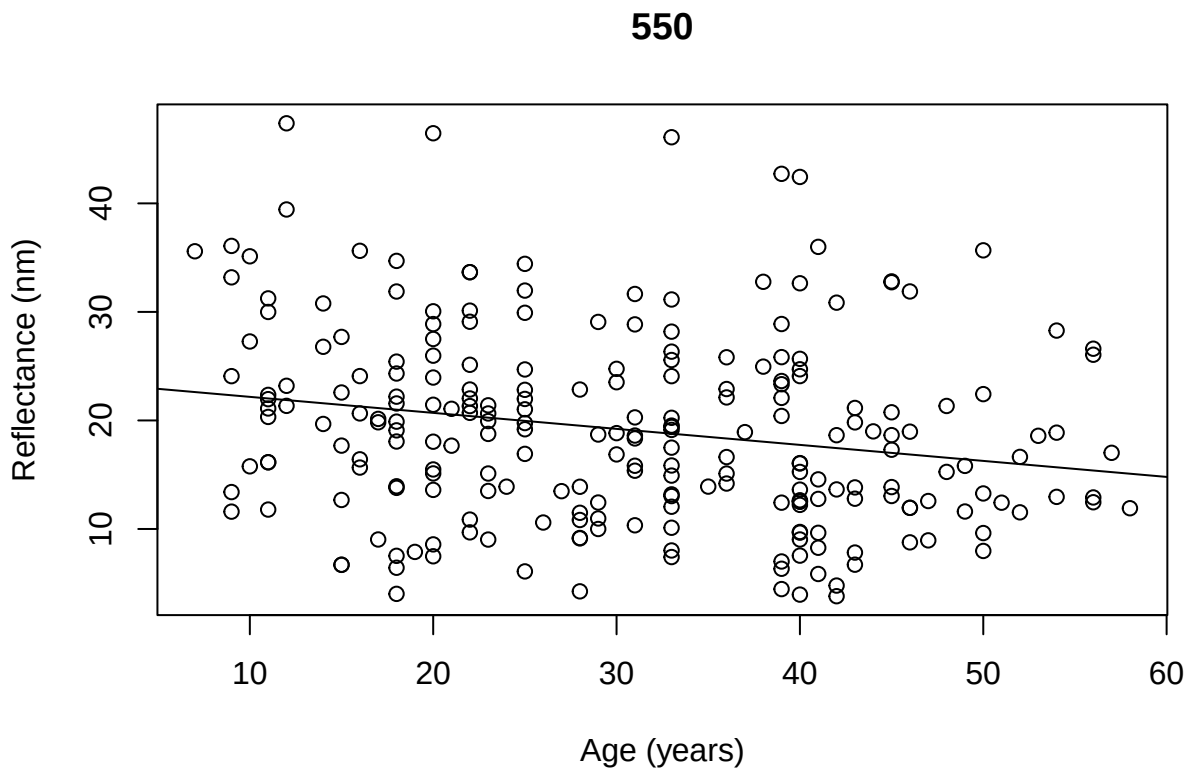
1. Independence - The individual measurements from single thallus are NOT independent. Thus, it may make more sense to reduce the data to the mean reflectance per thallus. Further, the reflectance between individual thalli are not independent within a species - they share evolutionary history. However, this should be fixed by treating species as a random effect in the linear mixed model.

Question 2: Is using the mean spectra acceptable? Should I account for the variation in the spectra per individual in some way?

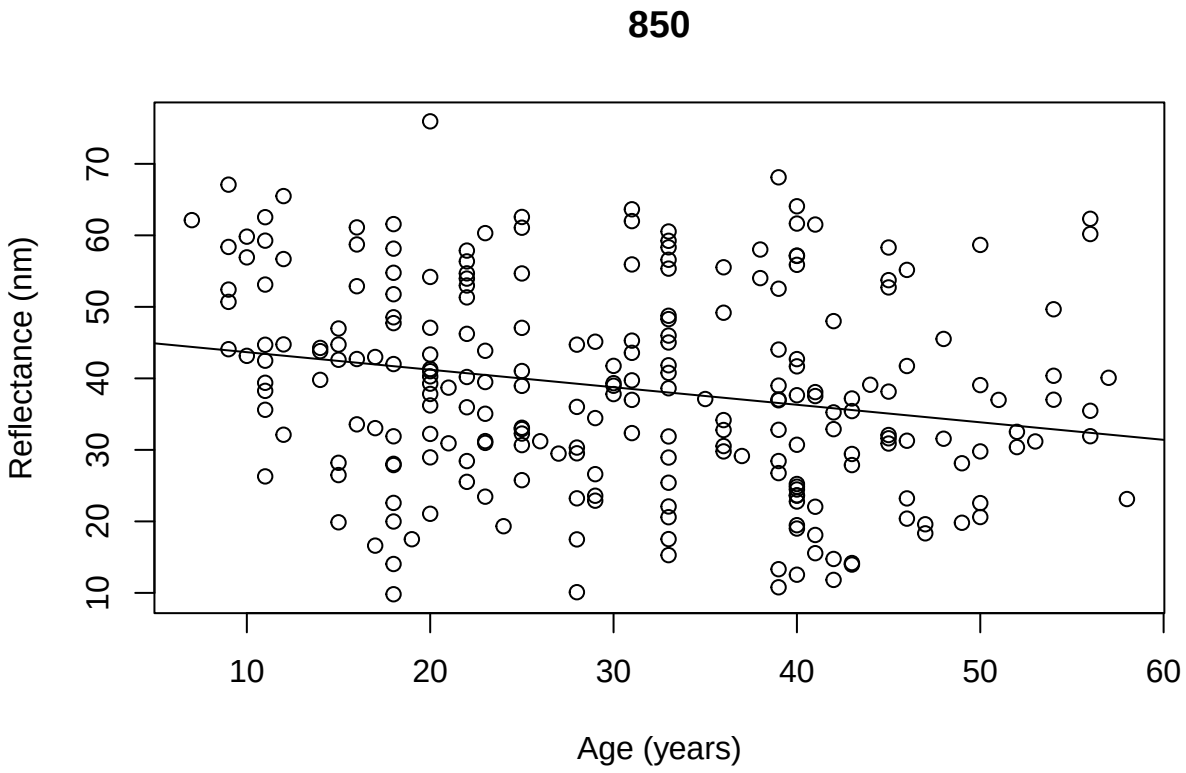
```
spectra = spectra[meta(spectra)$age < 60,]
spectra = aggregate(spectra, meta(spectra)$X, mean, try_keep_txt(mean)) #X indicates individual thallus
data = meta(spectra)
spec.m = as.matrix(spectra) * 100 #convert reflectance to a percentage to help with interpretation
spectra_percent = as_spectra(spec.m)
meta(spectra_percent) = data
spec_df = as.data.frame(spectra_percent)
```

2. Linearity

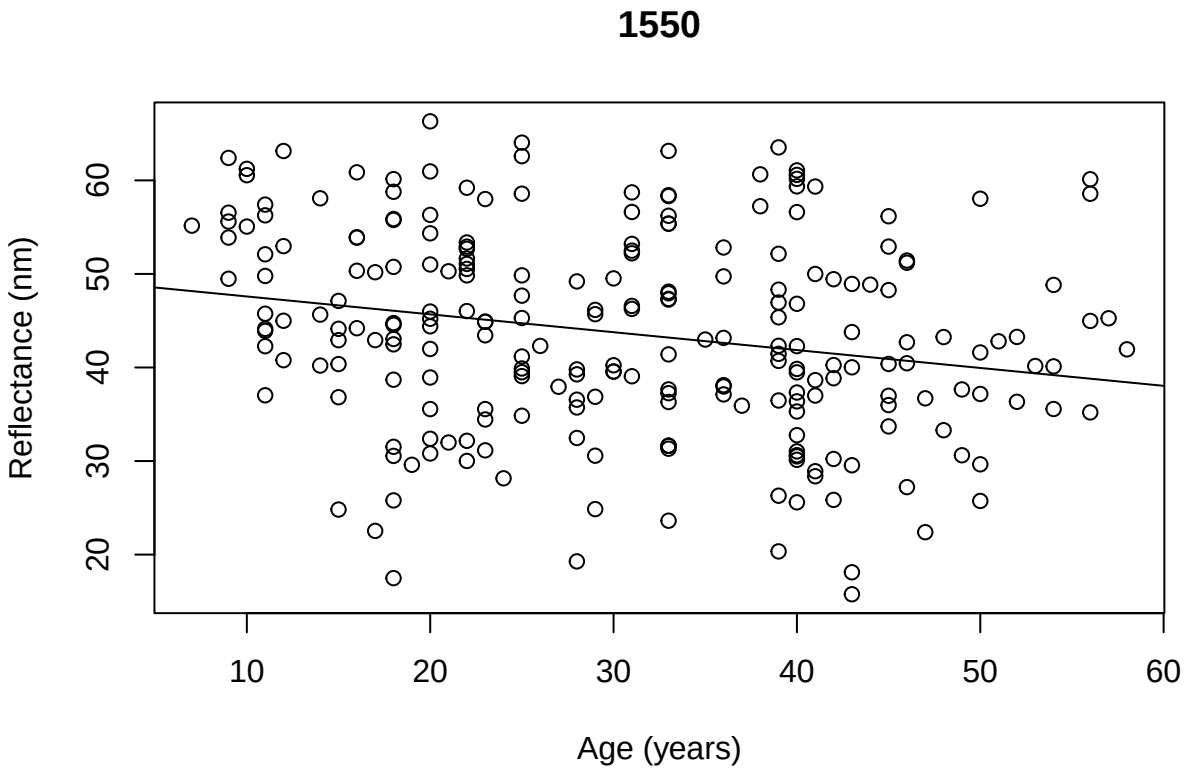
```
lm550 = lm(spec_df[, '550'] ~ spec_df$age)
lm850 = lm(spec_df[, '850'] ~ spec_df$age)
lm1550 = lm(spec_df[, '1550'] ~ spec_df$age)

plot(spec_df$age, spec_df[, '550'], main = '550', ylab = 'Reflectance (nm)', xlab = 'Age (years)')
abline(lm550)
```




```
plot(spec_df$age, spec_df[, '850'], main = '850', ylab = 'Reflectance (nm)', xlab = 'Age (years)')
abline(lm850)
```

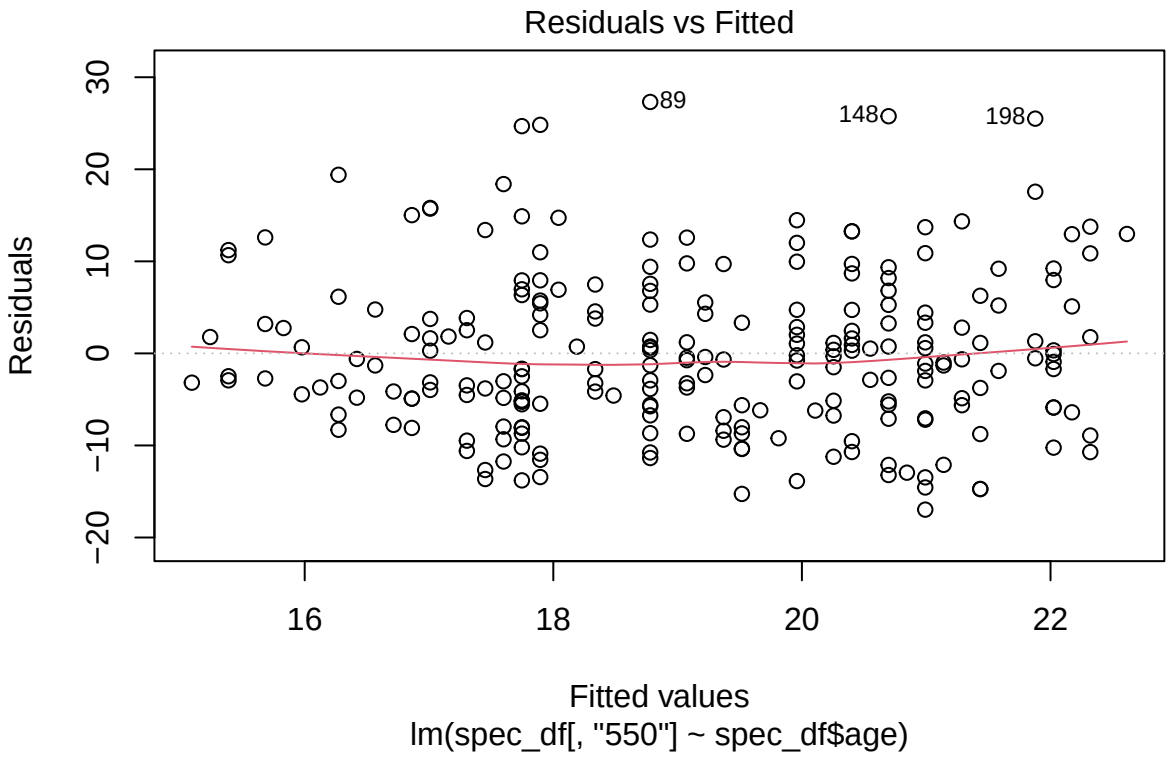



```
plot(spec_df$age, spec_df[, '1550'], main = '1550', ylab = 'Reflectance (nm)', xlab = 'Age (years)')
abline(lm1550)
```

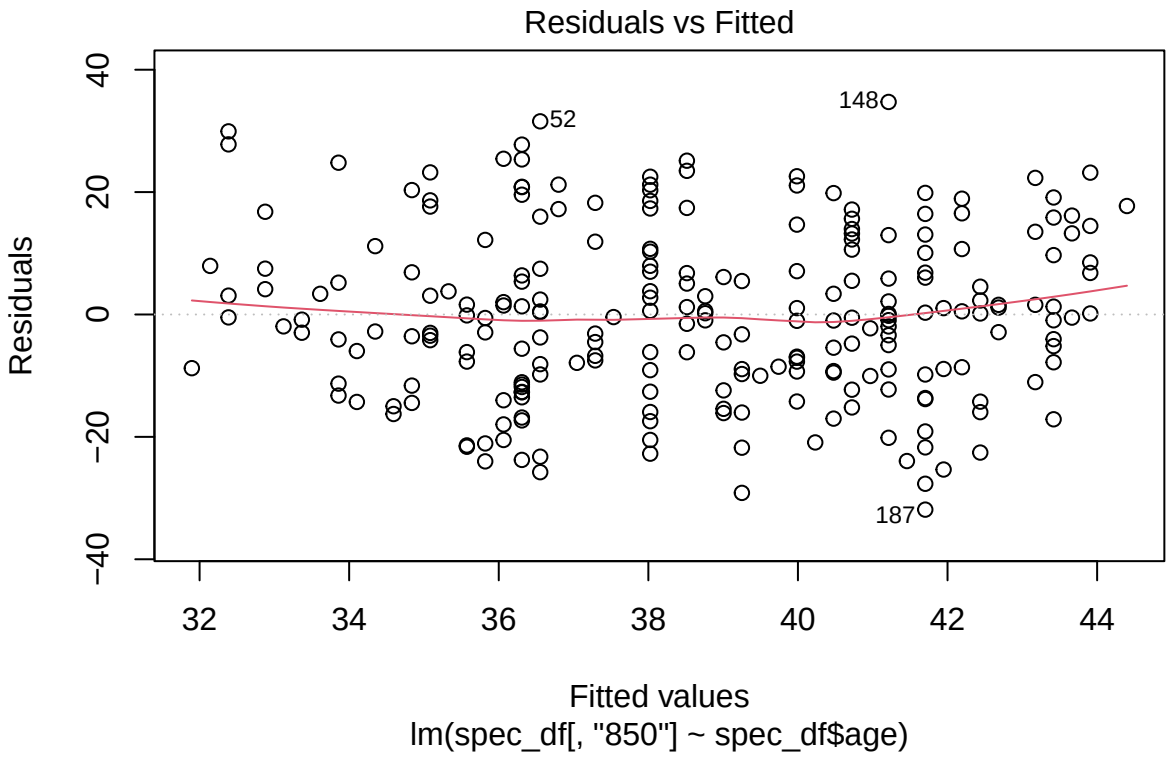


It seems like linearity is met. At least it does not look like any other function type would fit the data better.

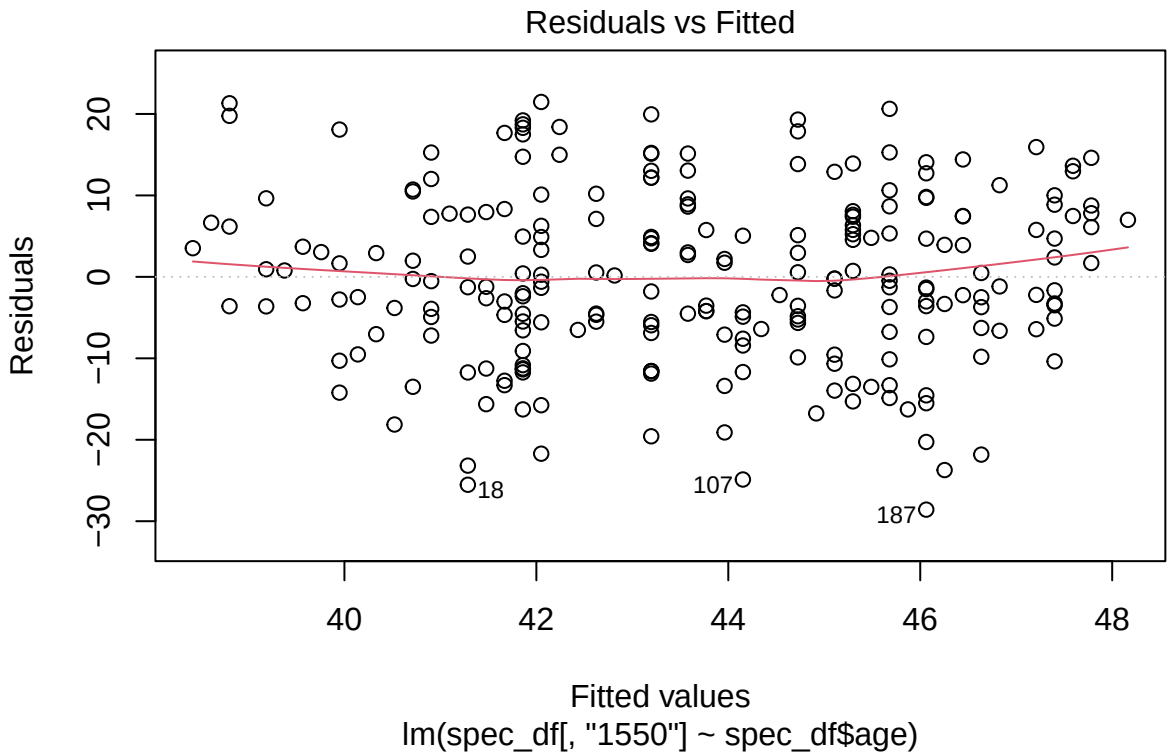
```
plot(lm550, 1)
```


Residuals vs Fitted
89
148
198
Residuals
30
20
10
0
−10
−20
16
18
20
22
Fitted values
lm(spec_df[, "550"] ~ spec_df$age)

```
plot(lm850, 1)
```



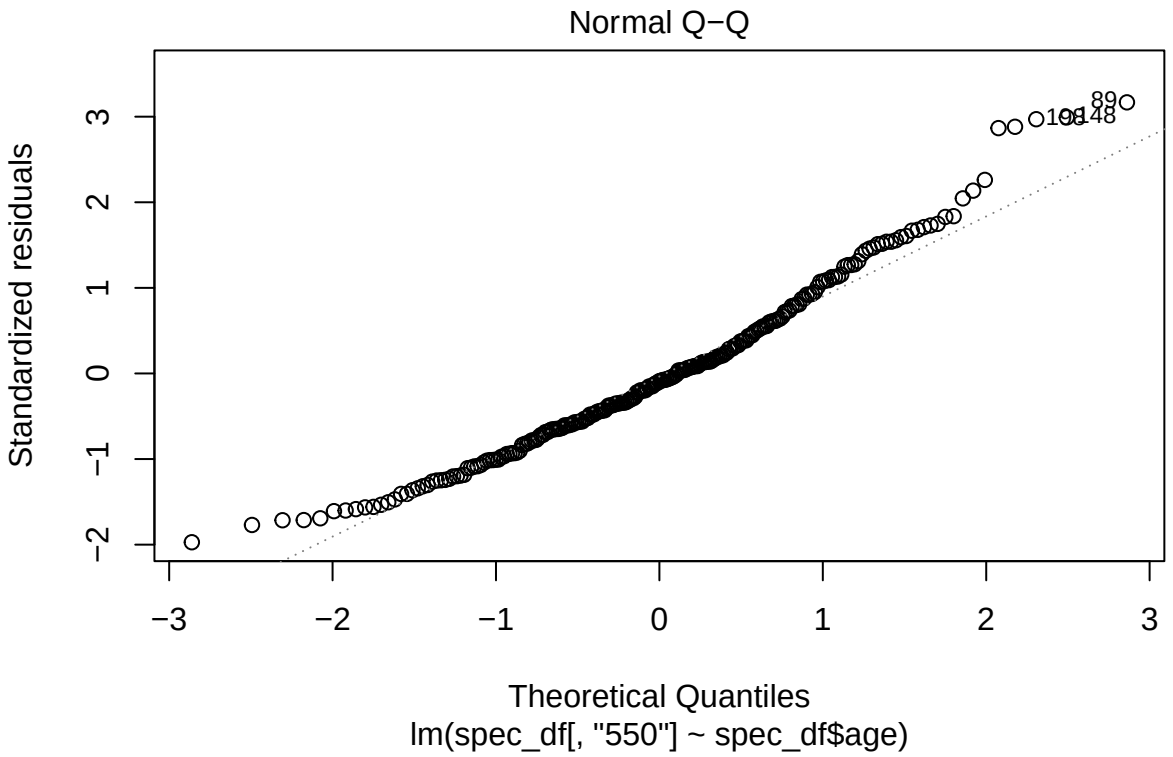

```
plot(lm1550, 1)
```



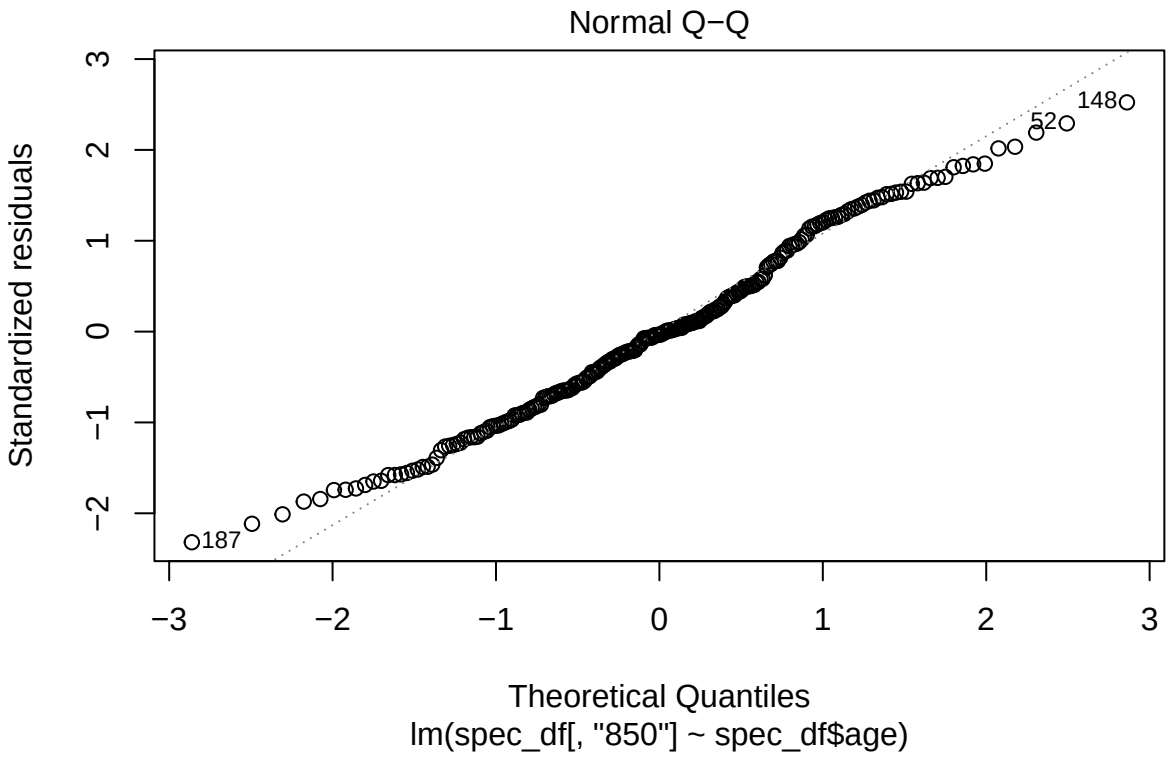
The residuals versus fitted plots also indicate that a linear function is a decent fit for the data.

3. Normality

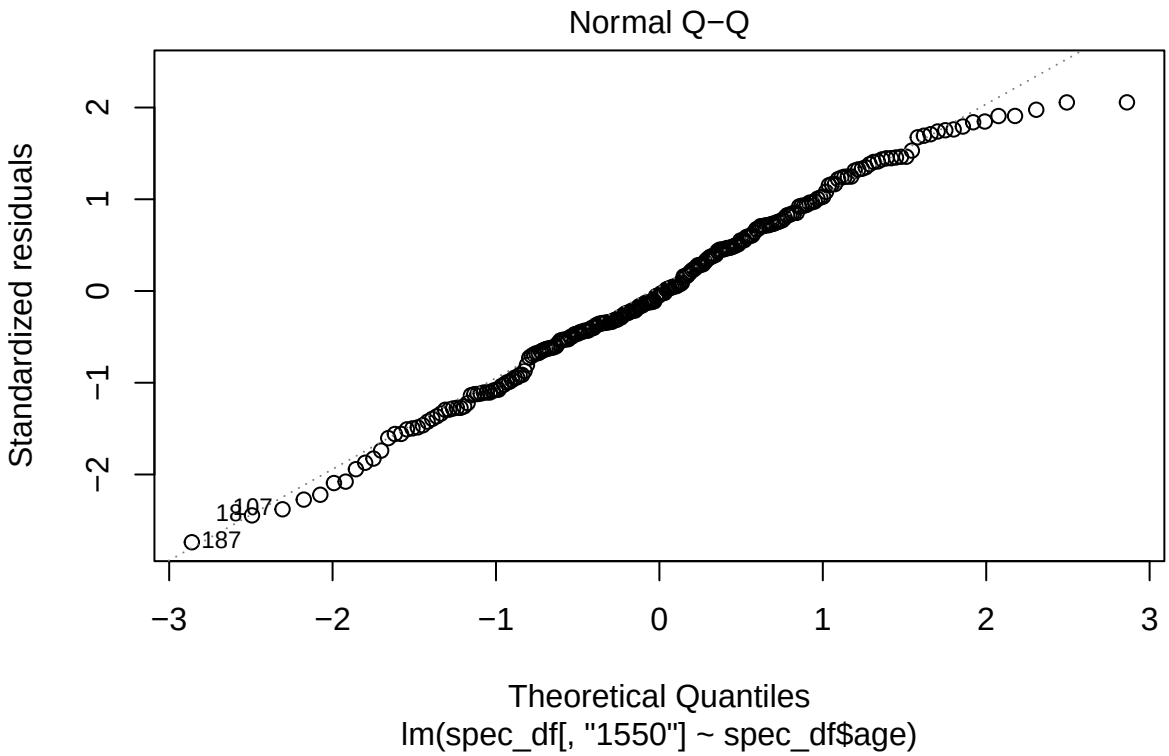
```
plot(lm550, 2)
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(spec_df[, "550"] ~ spec_df$age)
−3
−2
−1
0
1
2
3
89
148

```
plot(lm850, 2)
```



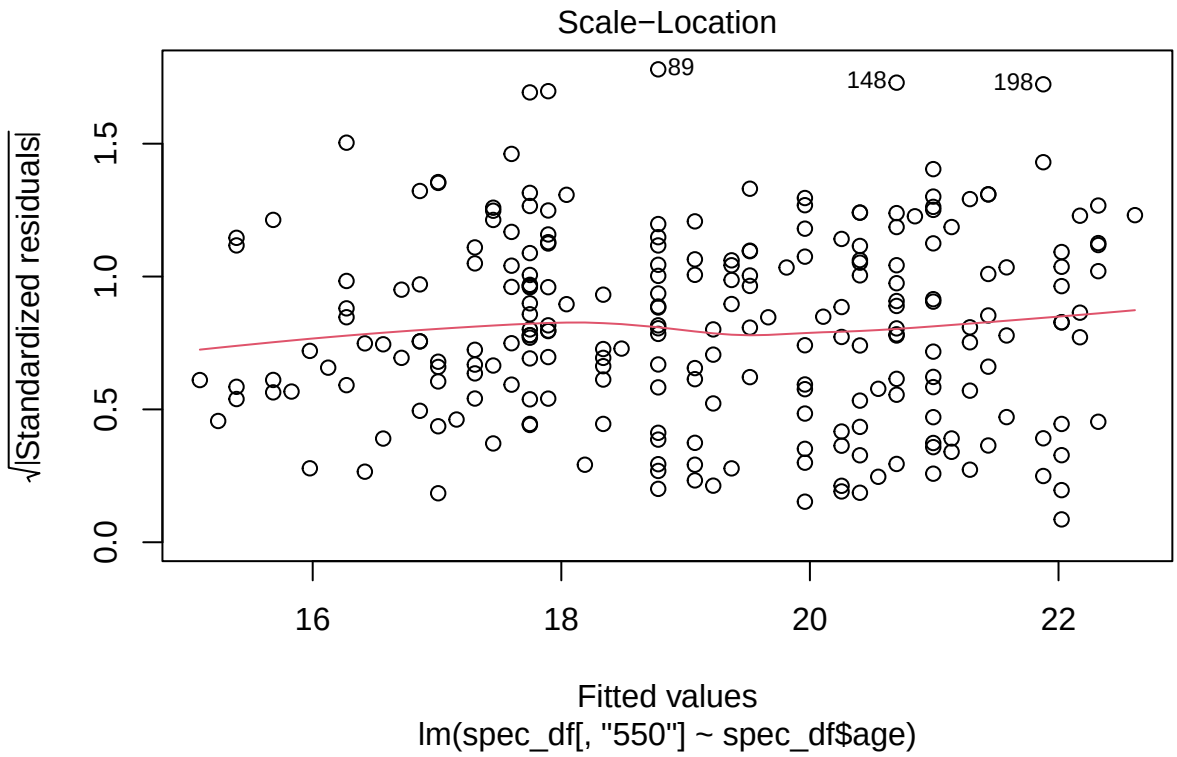

```
plot(lm1550, 2)
```



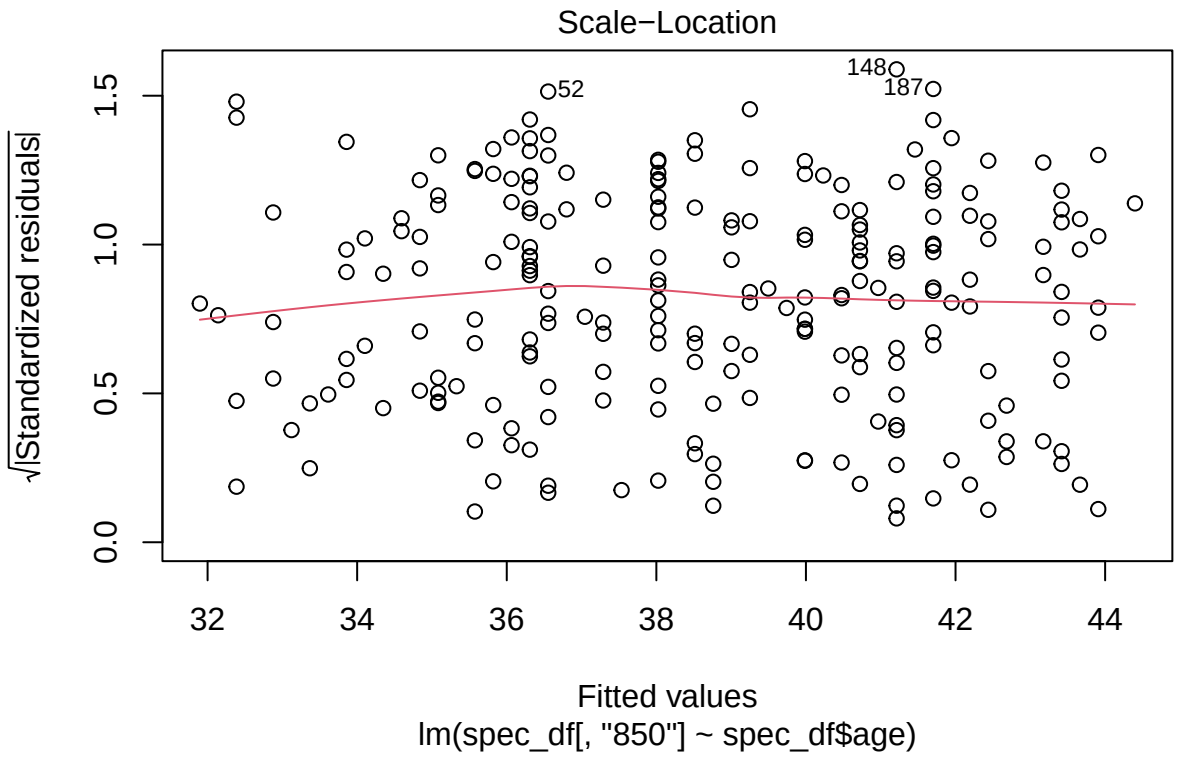
The tails are a bit skewed from the reference line, but it seems like the residuals are close to normally distributed.

4. Homoscedasticity

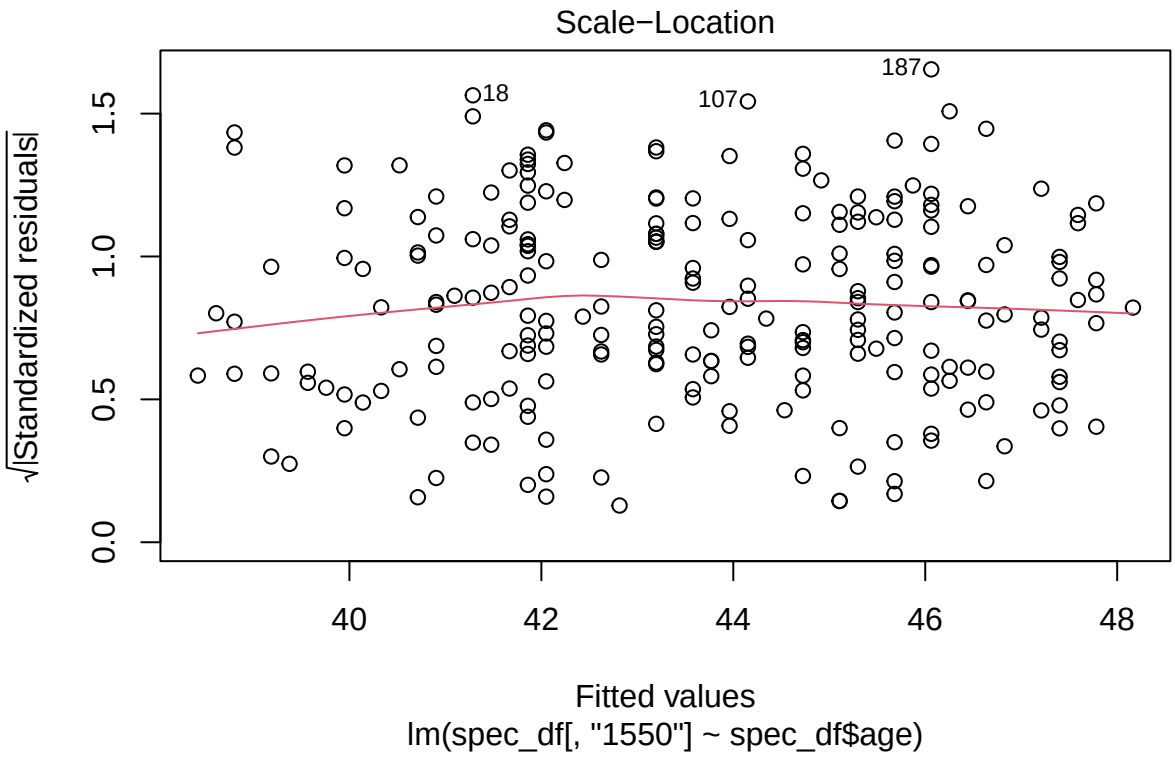
```
plot(lm550, 3)
```


Scale−Location
89
148
198
√|Standardized residuals|
0.0
0.5
1.0
1.5
16
18
20
22
Fitted values
lm(spec_df[, "550"] ~ spec_df$age)

```
plot(lm850, 3)
```




```
plot(lm1550, 3)
```



Looks pretty homoscedastic to me.

Overall, it looks like the assumptions of the linear model are generally met.

Linear Mixed Model Assumptions

Before checking model assumptions, let's create a linear mixed model by treating species (denoted as scientificName in the dataframe) as a random effect. Let's also compare a variable intercept - fixed slope model to a variable intercept - variable slope model for each of the selected 3 wavelengths.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
varInt550 = lmer(spec_df[, '550'] ~ age + (1|scientificName),
                         data = spec_df, REML = T,
                   lmerControl(optimizer ='bobyqa', #prevents convergence error
                               boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
varSlope550 = lmer(spec_df[, '550'] ~ age + (1+age|scientificName),
                         data = spec_df, REML = T,
                    lmerControl(optimizer ='bobyqa',
                               boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
anova(varInt550, varSlope550)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: spec_df
## Models:
## varInt550: spec_df[, "550"] ~ age + (1 | scientificName)
## varSlope550: spec_df[, "550"] ~ age + (1 + age | scientificName)
##             npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## varInt550      4 1510.5 1524.4 -751.27   1502.5
## varSlope550    6 1514.3 1535.1 -751.14   1502.3 0.262  2     0.8772
```

```
library(lme4)
varInt850 = lmer(spec_df[, '850'] ~ age + (1|scientificName),
                        data = spec_df, REML = T,
                  lmerControl(optimizer ='bobyqa', #prevents convergence error
                              boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
varSlope850 = lmer(spec_df[, '850'] ~ age + (1+age|scientificName),
                        data = spec_df, REML = T,
                   lmerControl(optimizer ='bobyqa',
                              boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(varInt850, varSlope850)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: spec_df
## Models:
## varInt850: spec_df[, "850"] ~ age + (1 | scientificName)
## varSlope850: spec_df[, "850"] ~ age + (1 + age | scientificName)
##             npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## varInt850      4 1679.2 1693.1 -835.6   1671.2
## varSlope850    6 1683.2 1704.0 -835.6   1671.2 0.0018  2     0.9991
```

```
library(lme4)
varInt1550 = lmer(spec_df[, '1550'] ~ age + (1|scientificName),
                        data = spec_df, REML = T,
                  lmerControl(optimizer ='bobyqa', #prevents convergence error
                              boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
varSlope1550 = lmer(spec_df[, '1550'] ~ age + (1+age|scientificName),
                        data = spec_df, REML = T,
                   lmerControl(optimizer ='bobyqa',
                              boundary.tol = 1e-5, optCtrl = list(maxfun = 1e5)))
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(varInt1550, varSlope1550)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: spec_df
## Models:
```

```
## varInt1550: spec_df[, "1550"] ~ age + (1 | scientificName)
## varSlope1550: spec_df[, "1550"] ~ age + (1 + age | scientificName)
##              npar    AIC  BIC  logLik deviance  Chisq Df Pr(>Chisq)
## varInt1550      4 1621.2 1635 -806.58   1613.2
## varSlope1550    6 1625.2 1646 -806.58   1613.2 0.0035  2     0.9983
```

NOTE: The `boundary (singular) fit: see ?isSingular` warning arises from the variable intercept - variable slope models. - Indicates "some dimensions of the variance-covariance matrix have been estimated as exactly zero."

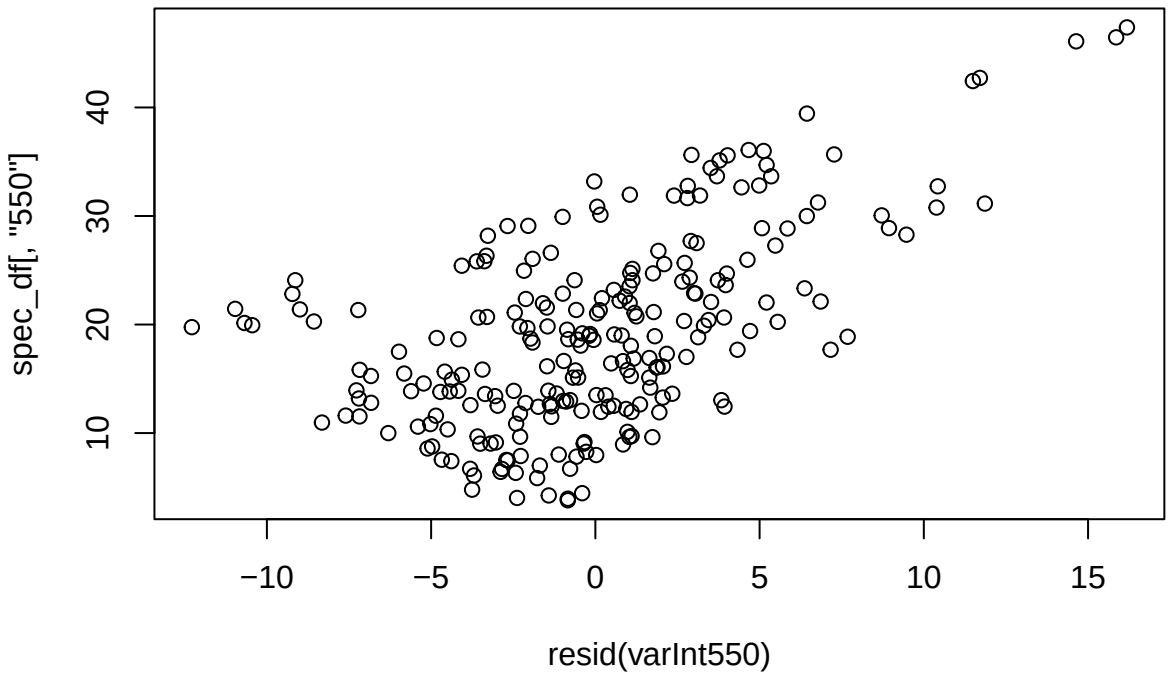
Question 3: Is a singular fit something I need to fix? If so, how? I'm not finding any clear answers on any blogs/documentation.

It looks like the variable intercept - fixed slope model is a better fit for each of these three wavelengths. This also turns out to be true for most wavelengths (not presented here).

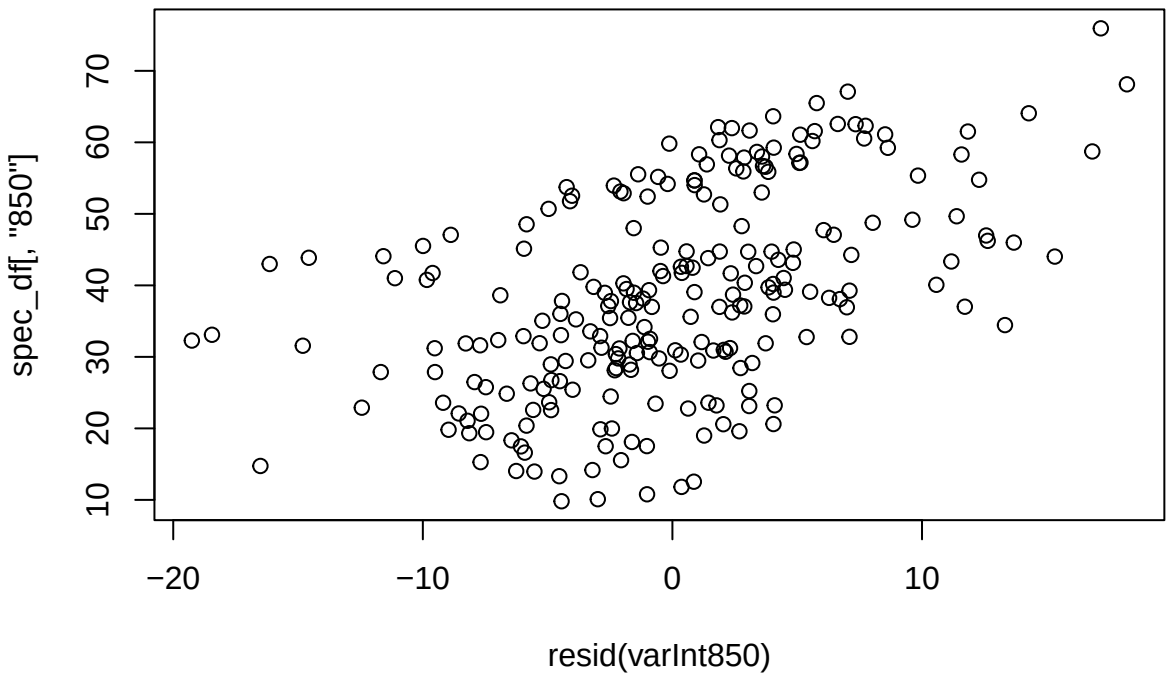
Variable intercept - fixed slope models

1. Linearity

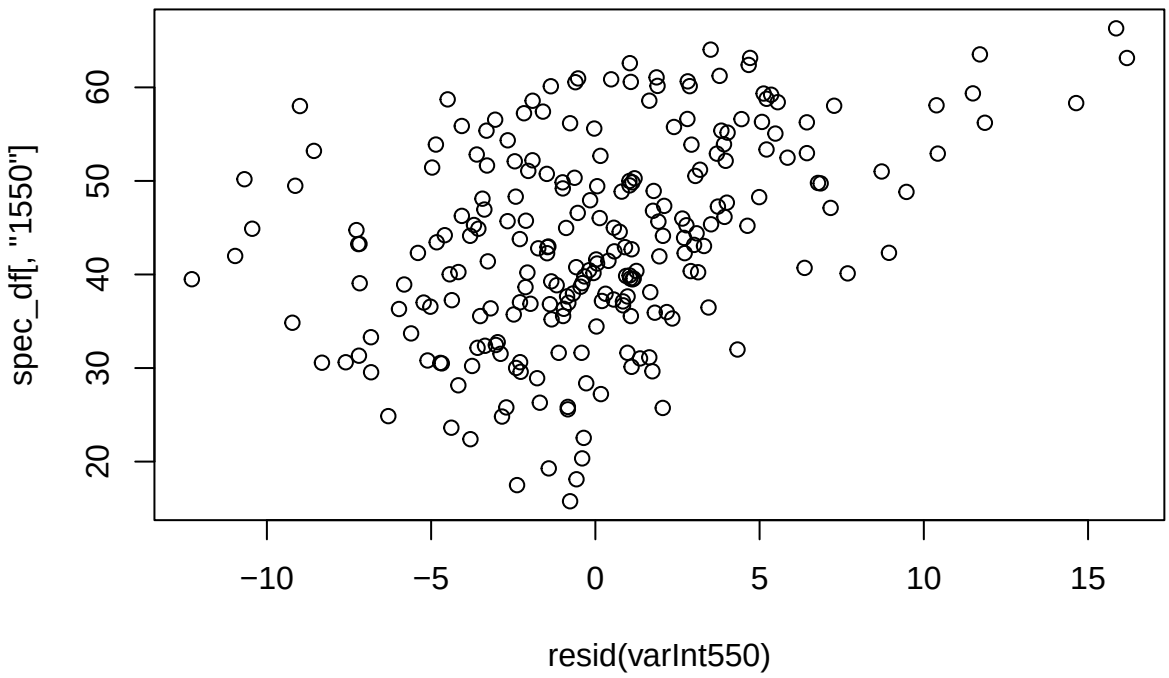
```
plot(resid(varInt550), spec_df[, '550'])
```




```
plot(resid(varInt850), spec_df[, '850'])
```



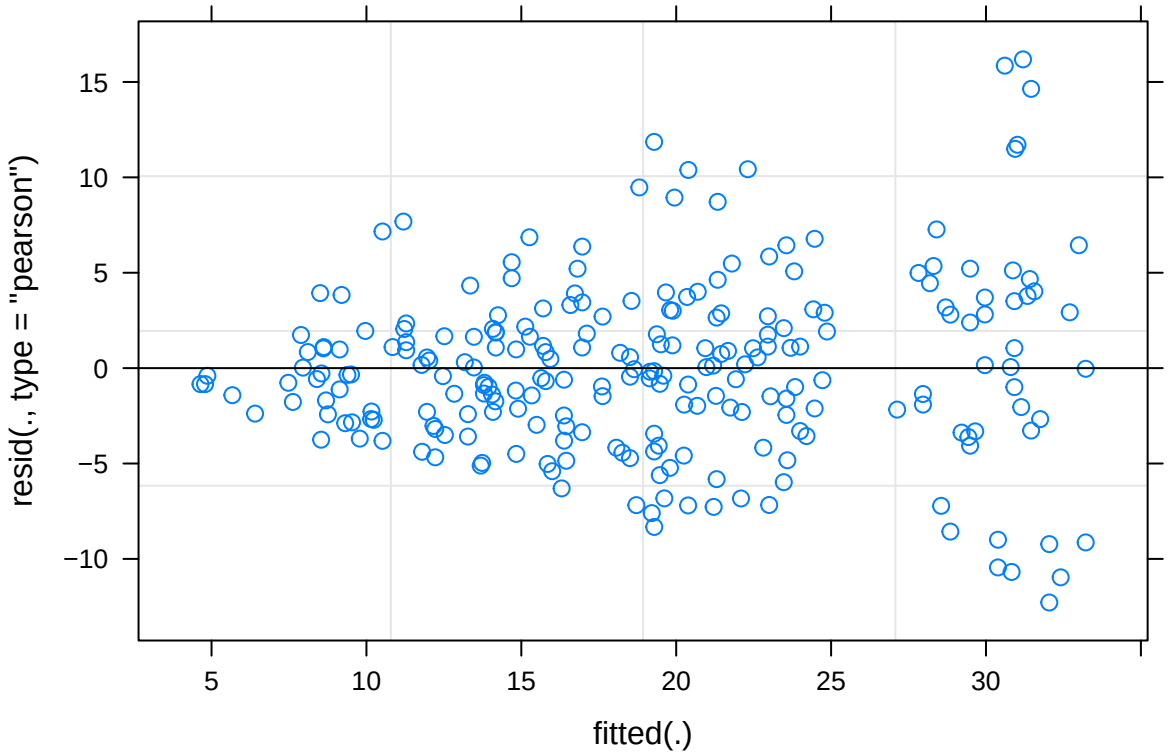

```
plot(resid(varInt550), spec_df[, '1550'])
```



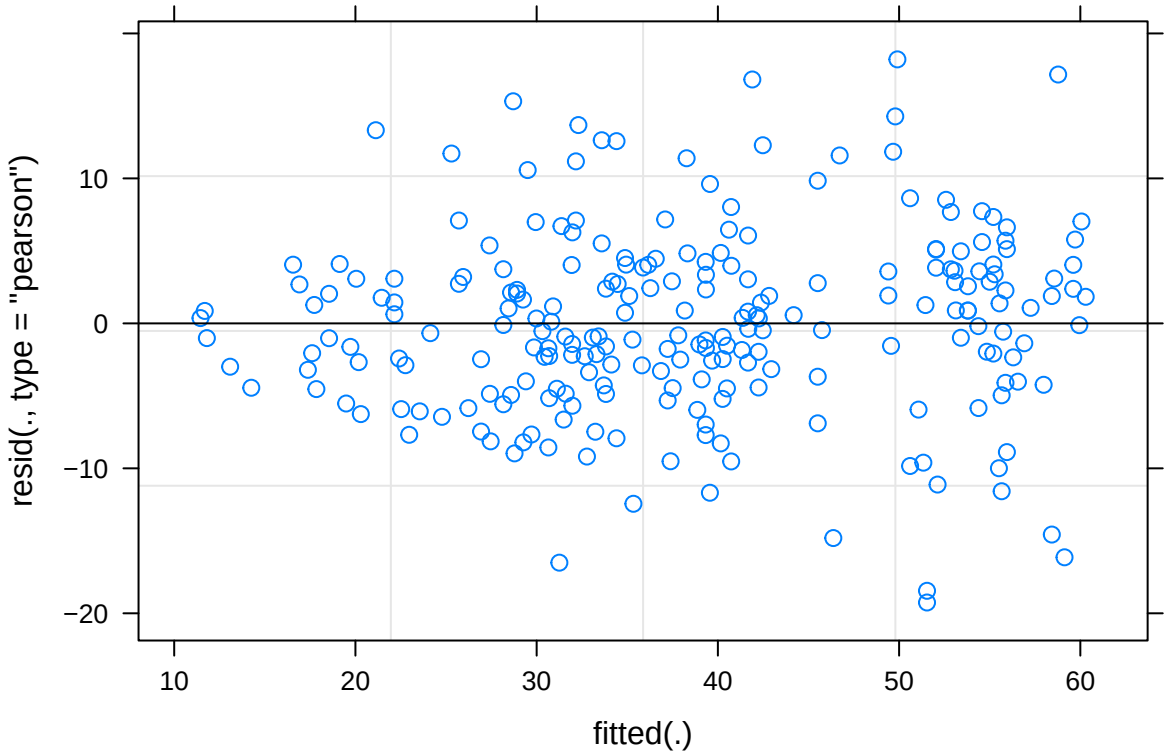
According to Michael Palmeri, a plot of random points indicates linearity. These plots look pretty random to me, but perhaps some sort of pattern can be seen on the right side of the plot.

2. Homoscedasticity

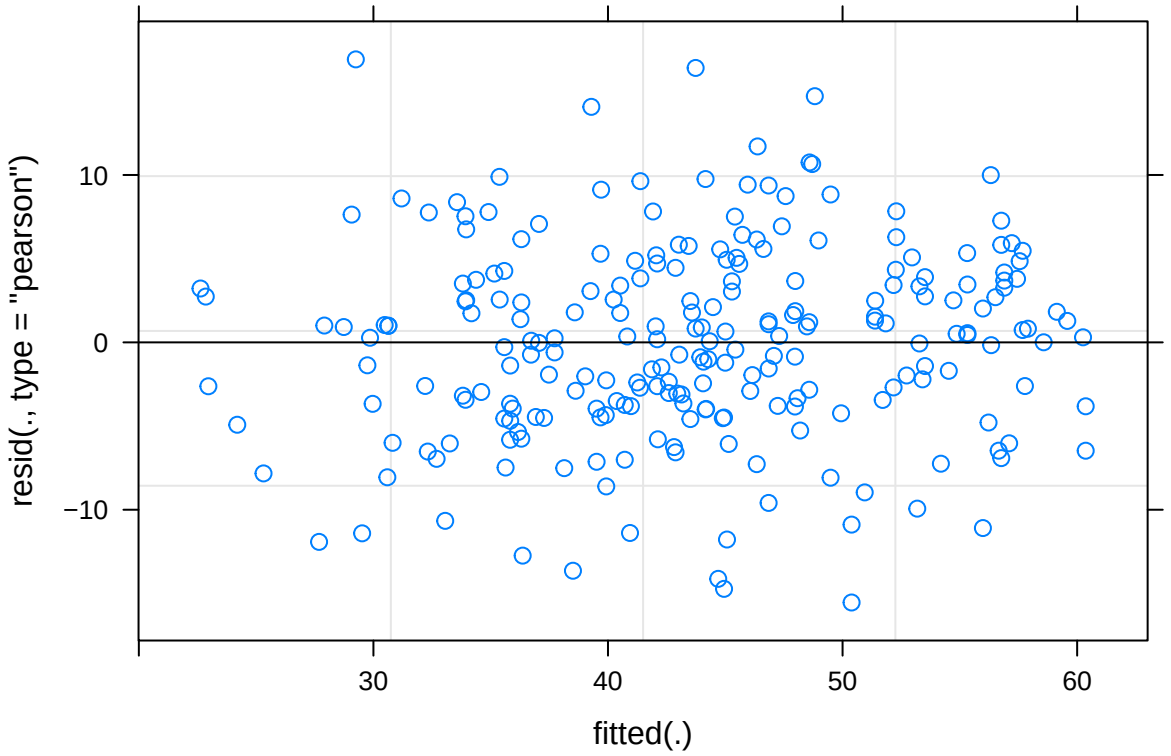
```
plot(varInt550)
```




```
plot(varInt850)
```




```
plot(varInt1550)
```

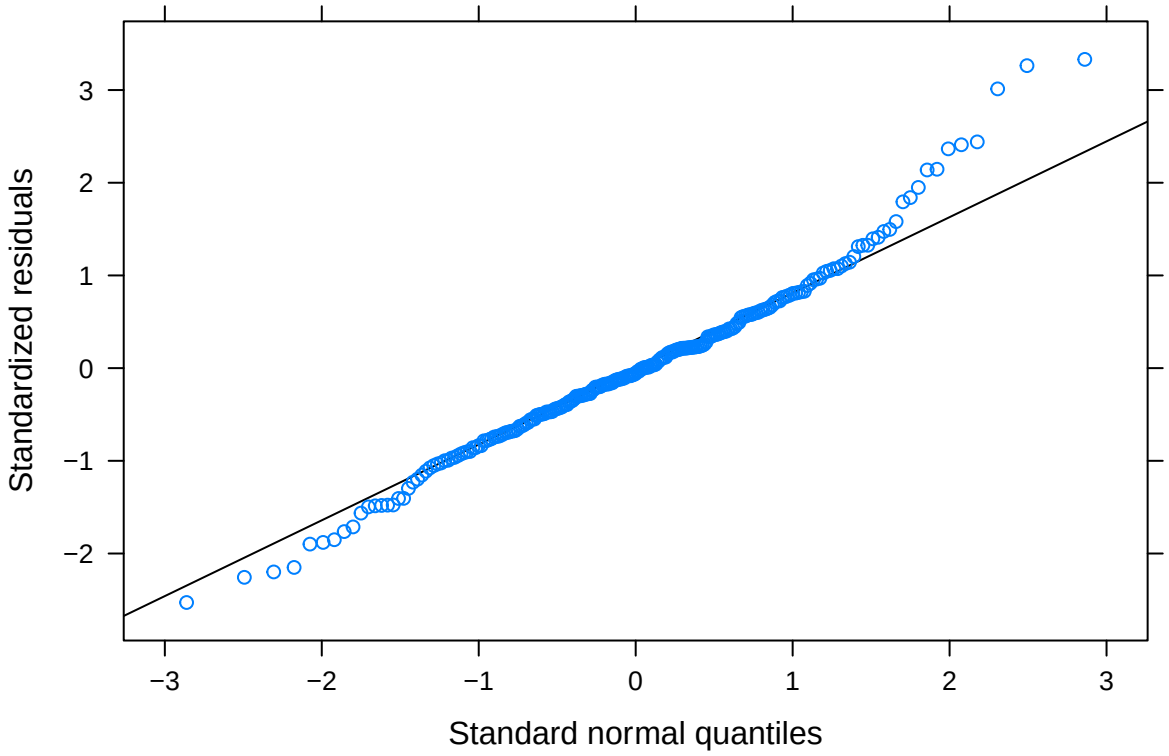


The 550 and 850 models appear to show heteroscedasticity, but the 1500 model looks pretty homoscedastic (sausage shaped).

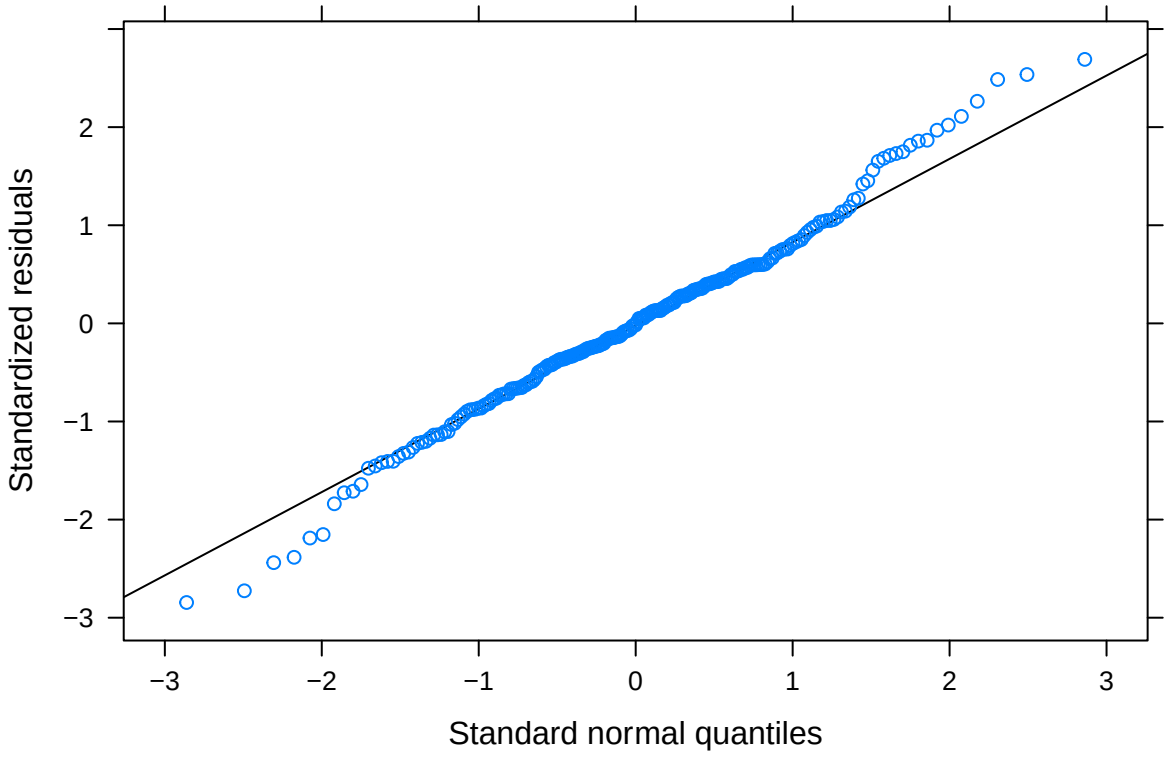
Question 4: Should I transform reflectance values? Some wavelengths show heteroscedasticity while others don't. Transforming could also make interpretation more difficult. Further, Schielzeth et al. (2020) demonstrate linear mixed models to be robust to heteroscedasticity.

3. Normally distributed residuals

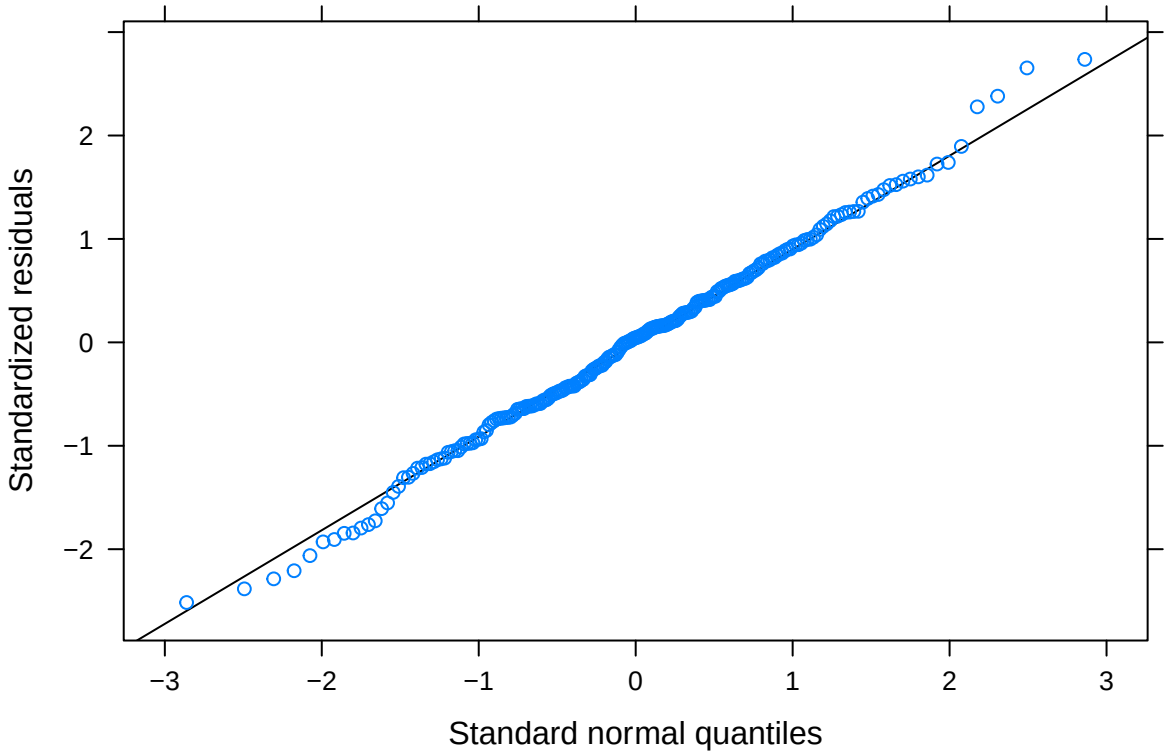
```
library(lattice)
qqmath(varInt550)
```




```
qqmath(varInt850)
```




```
qqmath(varInt1550)
```



Not the best, not the worst. Model 550 is the furthest from a true normal distribution.

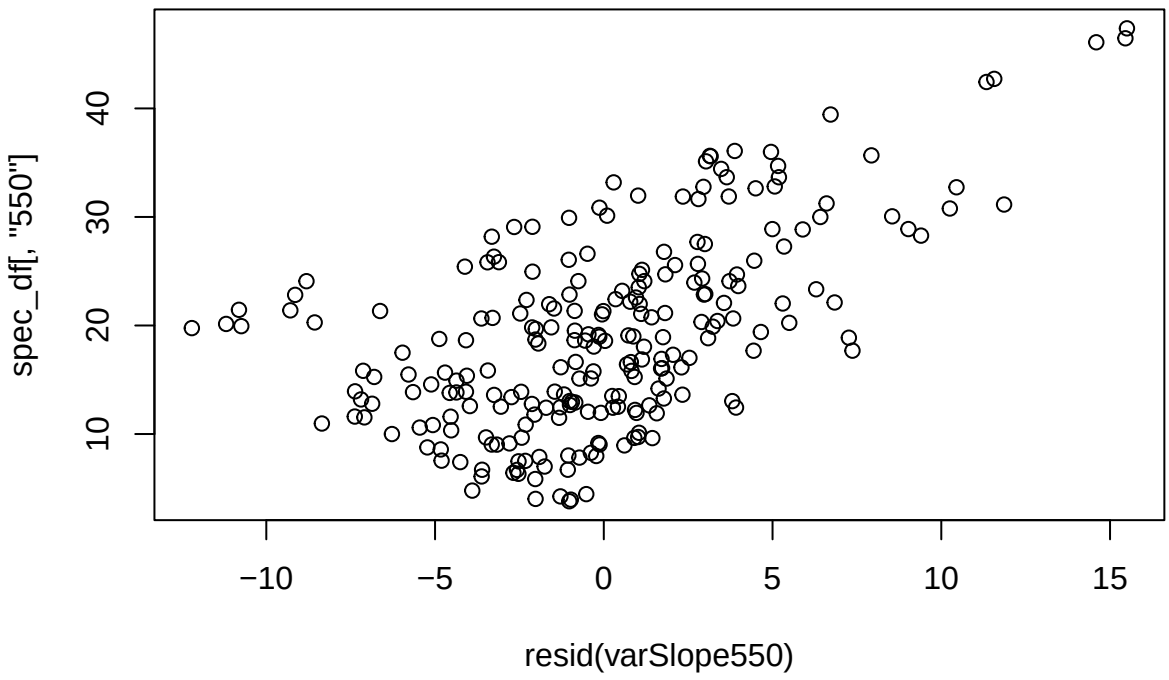
Variable intercept - variable slope models

Although the previous ANOVAs show the variable intercept - variable slope model to be worse than the variable intercept - variable slope model, I'm still interested if slopes vary between species. I'll add the code and results for assessing assumptions, but feel free to skim this since the results are nearly identical to those presented for the variable intercept - fixed slope model.

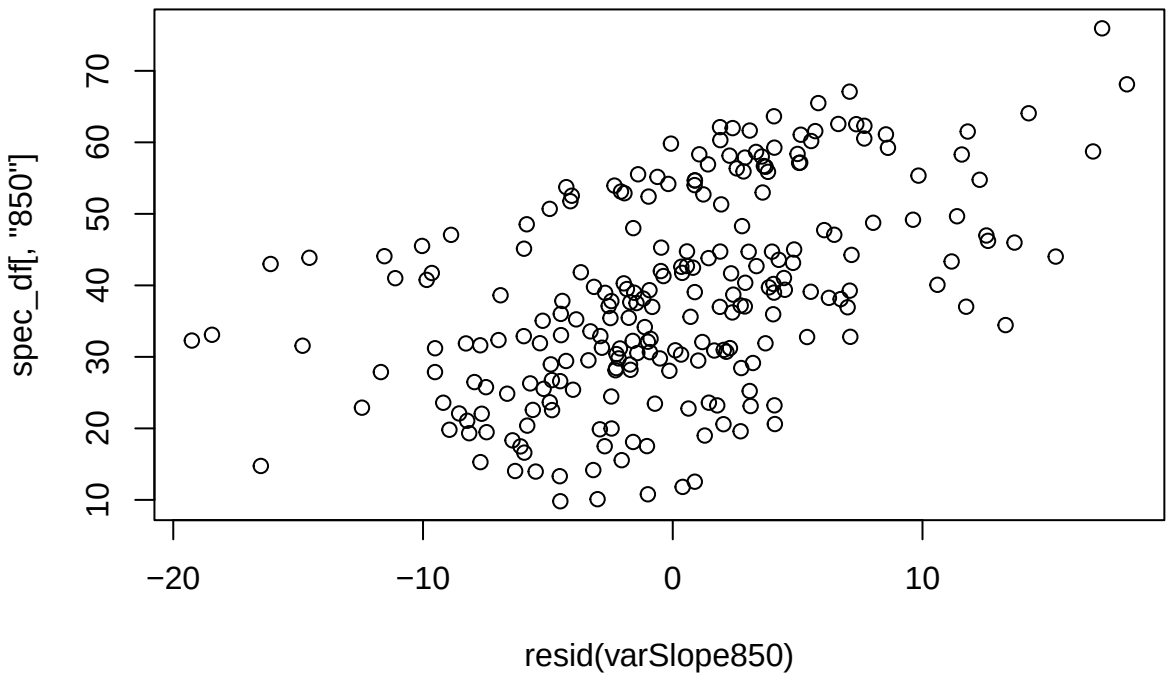
Assess model assumptions

1. Linearity

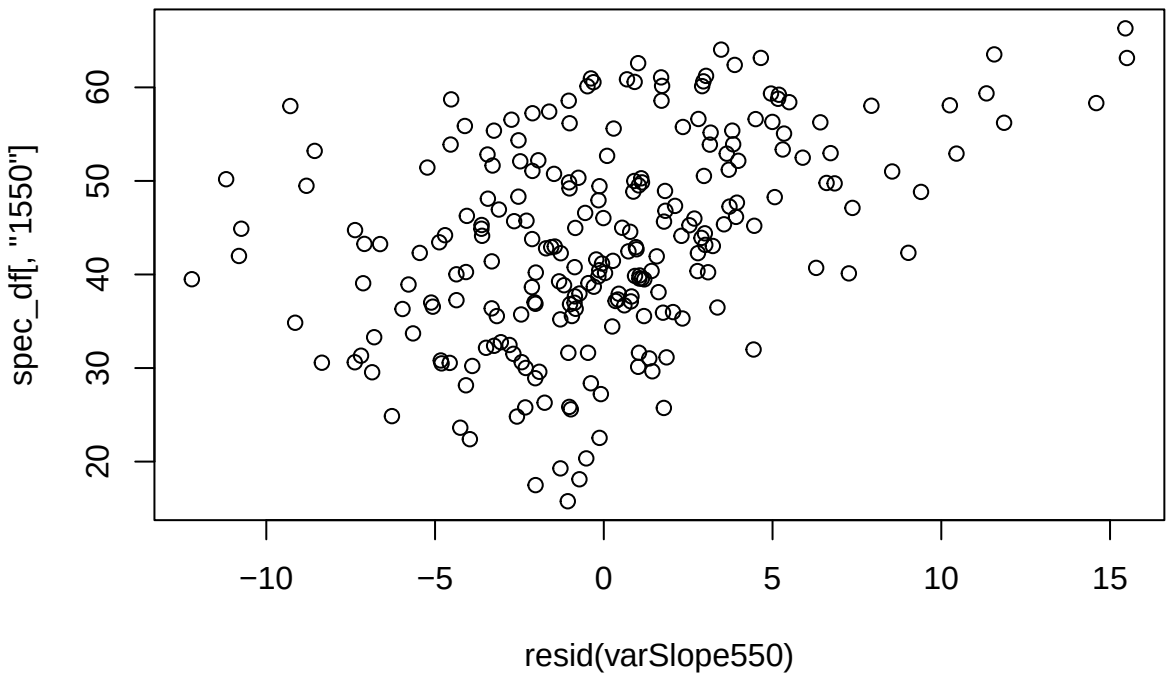
```
plot(resid(varSlope550), spec_df[, '550'])
```




```
plot(resid(varSlope850), spec_df[, '850'])
```



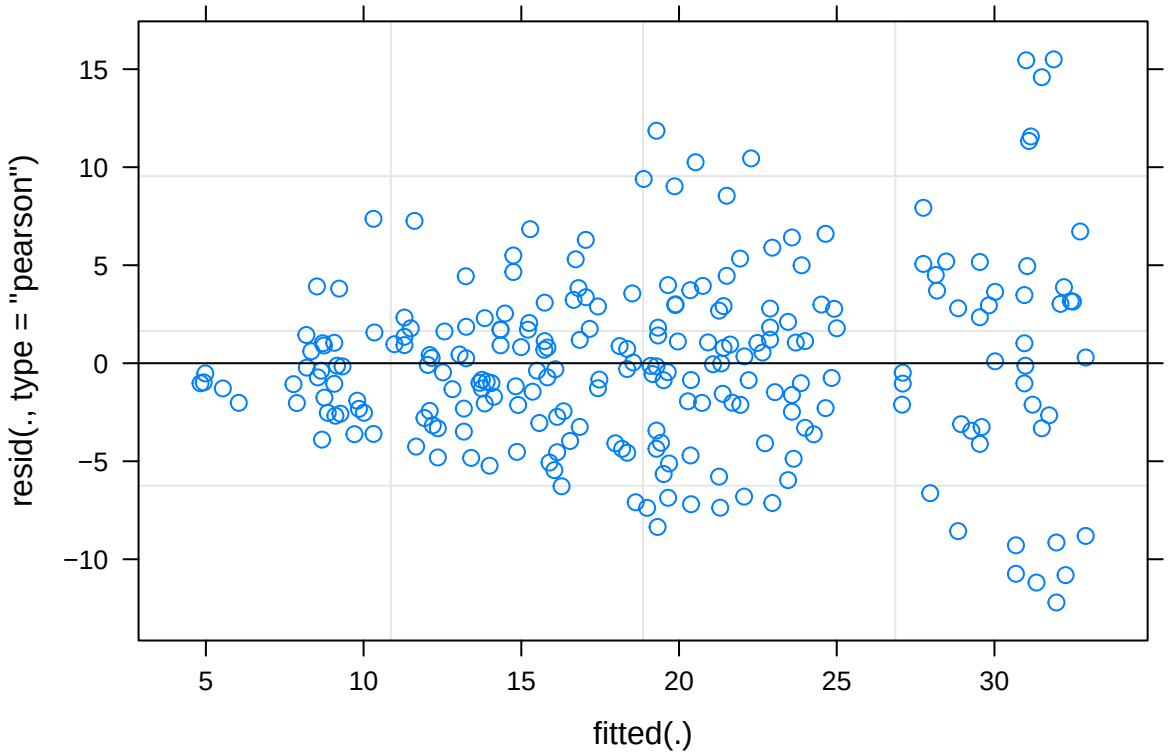

```
plot(resid(varSlope550), spec_df[, '1550'])
```



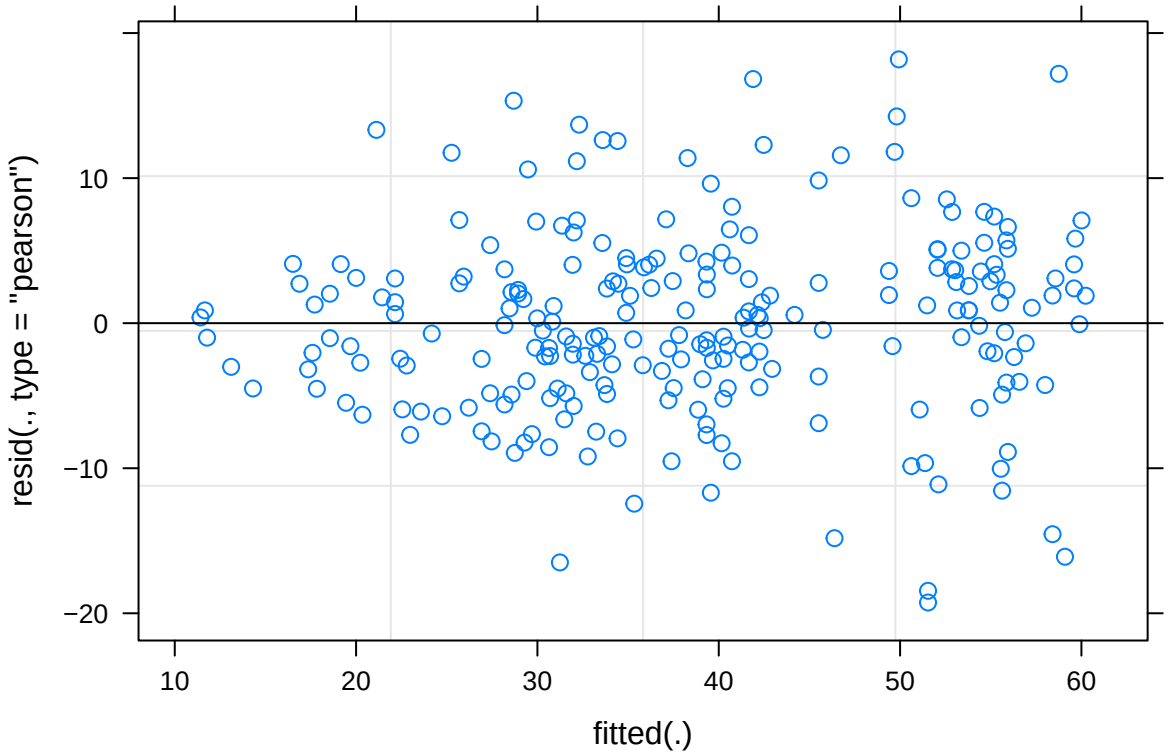
Seems pretty random.

2. Homoscedasticity

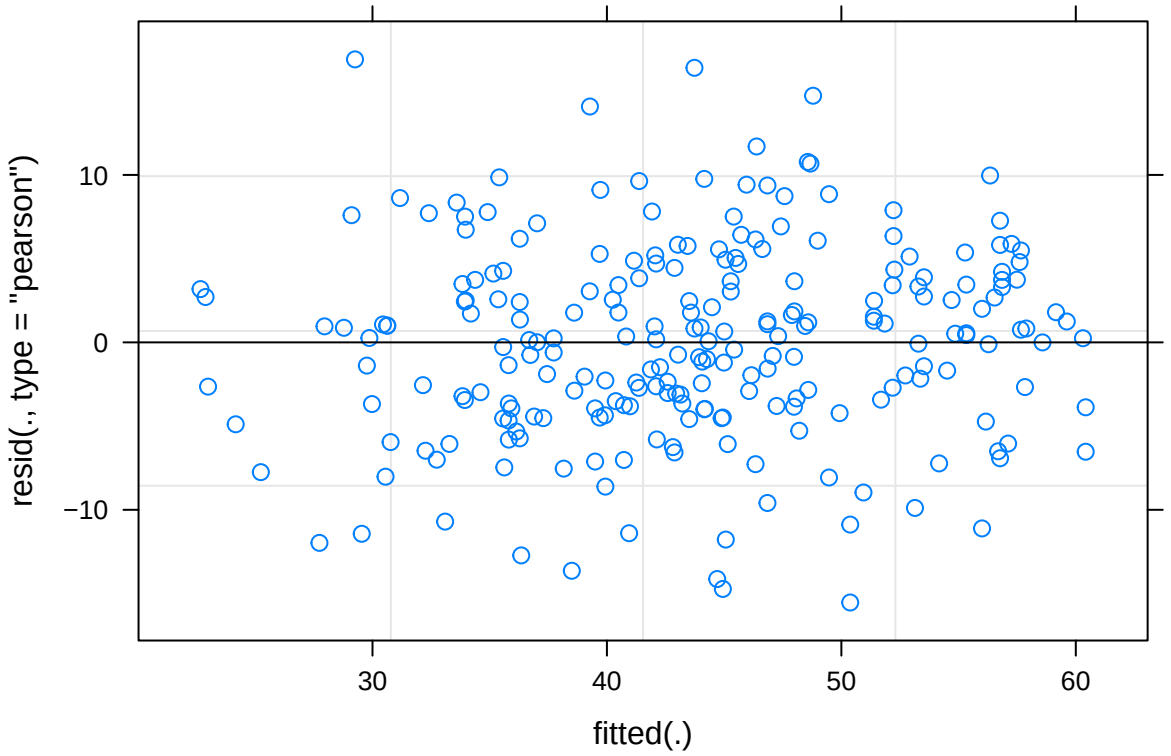
```
plot(varSlope550)
```




```
plot(varSlope850)
```

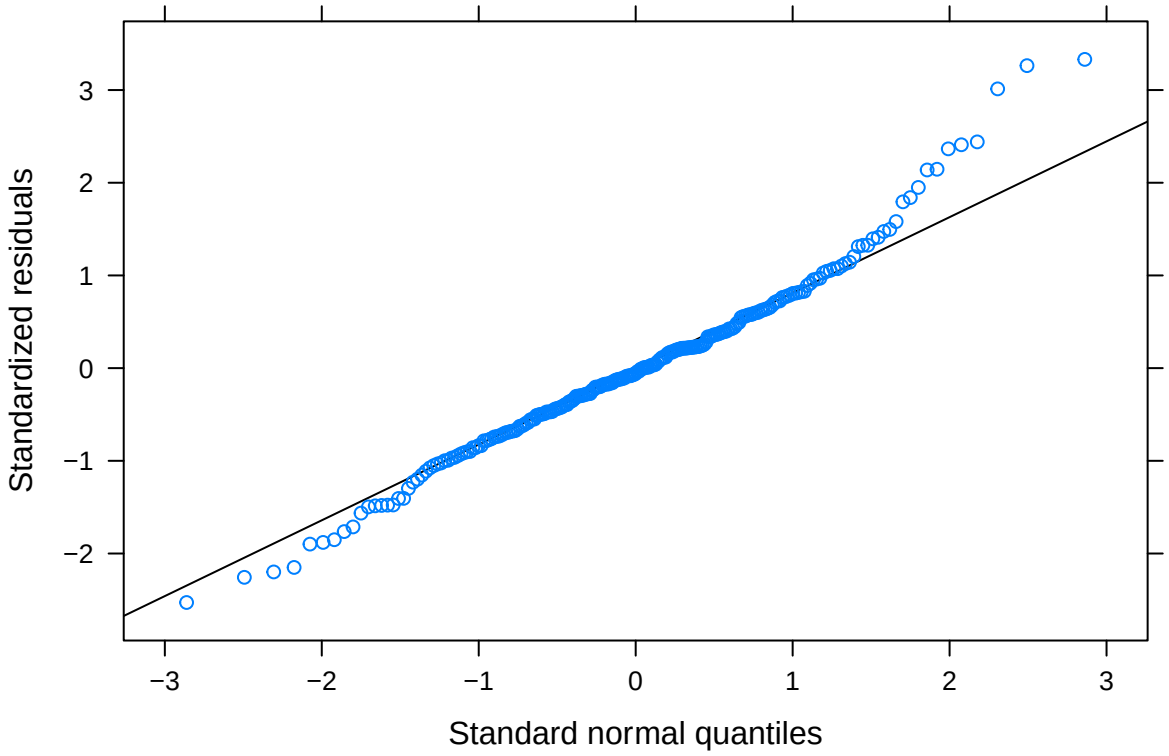



```
plot(varSlope1550)
```

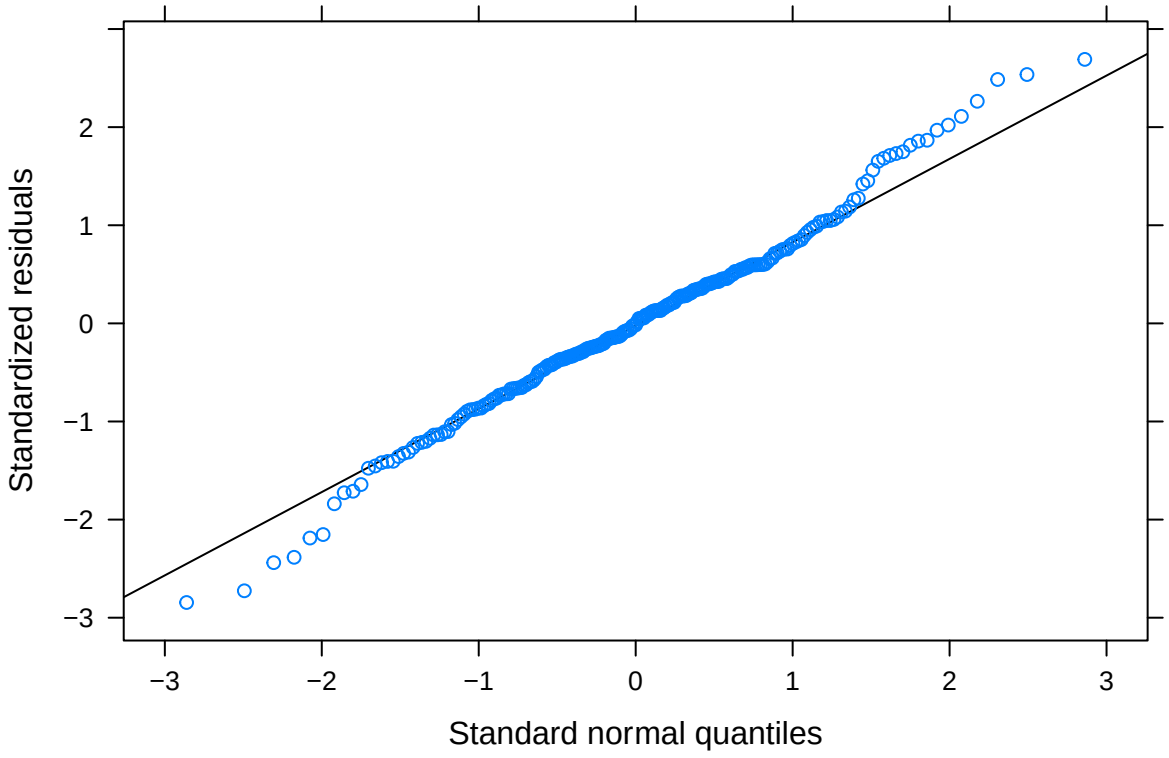


Heteroscedasticity in 550 and 850.

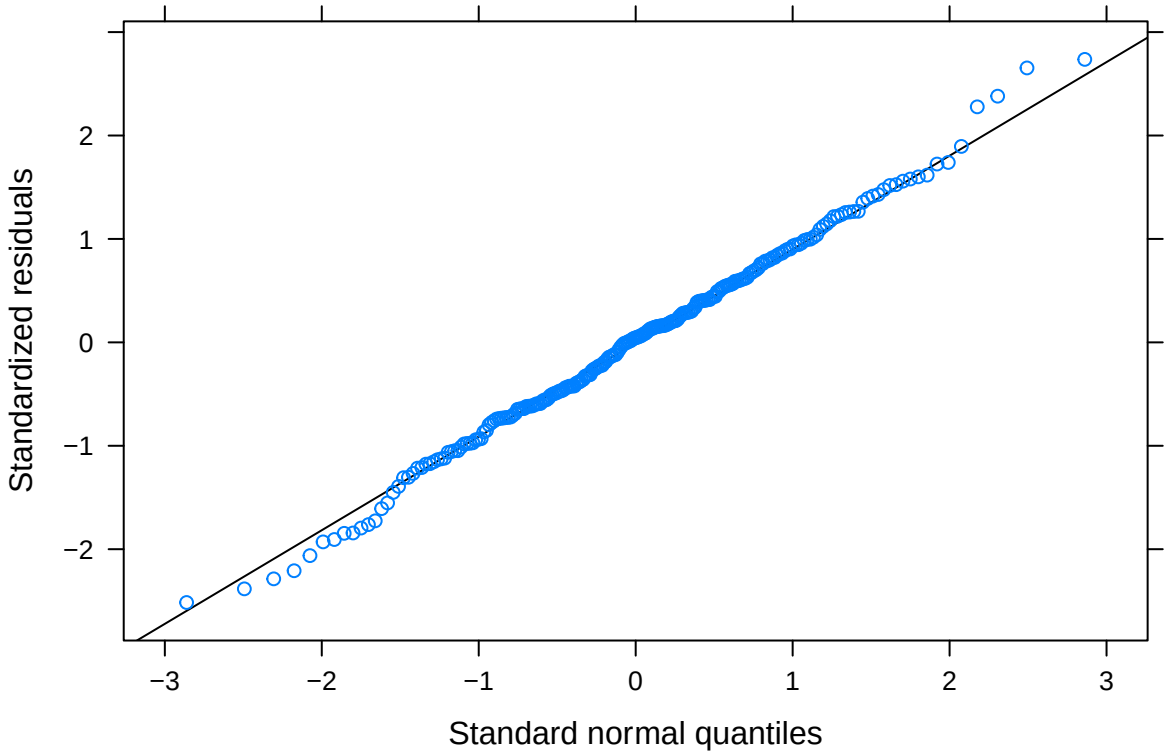
```
library(lattice)
qqmath(varInt550)
```




```
qqmath(varInt850)
```




```
qqmath(varInt1550)
```



Tails skewed for 550 and 850, ok for 1550.

Overall, these are the same results as for the variable intercept - fixed slope models.

Interpreting the model summary

Let's just look at the model summary for one model:

```
summary(varSlope550)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: spec_df[, "550"] ~ age + (1 + age | scientificName)
##    Data: spec_df
## Control:
## lmerControl(optimizer = "bobyqa", boundary.tol = 1e-05, optCtrl = list(maxfun = 1e+05))
##
## REML criterion at convergence: 1505
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5258 -0.5325 -0.0332  0.5251  3.2077
##
## Random effects:
##  Groups         Name        Variance Std.Dev. Corr
##  scientificName (Intercept) 64.59452 8.03707
##                 age          0.00152 0.03898  -0.42
```

```
##  Residual                   23.35346 4.83254
## Number of obs: 237, groups:  scientificName, 29
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 22.07581    1.74015  12.686
## age         -0.07220    0.02805  -2.574
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.537
```

The summary states that the fixed effects consist of an intercept of 22.08% reflectance and a slope of -0.072 % reflectance/year.

Question 5: I'm pretty sure slope would be expressed as change in percent reflectance per year. Do you agree with this?

Gabriela Hajduk explains that variance listed under in the Random effects section is the left over variance that is not explained by the fixed effects. She then states that you can get estimates of the variance explained by random effect by dividing the variance of that random effect by the sum of the random effect variance. In this case, the intercept that varies between species (scientificName) accounts for 64.6/(64.6 + 0.0015) + 23.4) = 73.4% of the variance not explained by the fixed effects.

Question: 6 Is this how the summary should be interpreted? If so, we see that the random effects slope can vary by a decent amount when compared to the fixed slope, but the variable slope will never account for much of the variance because it should either be expressed in different units or intercept variance and slope variance occur at two different scales. How do I reconcile this?

Coefficients per species

I can obtain the regression coefficients for each species, but should I trust the estimates considering some species are only represented by a few individuals (low sample size)?

```
coef(varSlope550)
```

```
## $scientificName
##                              (Intercept)         age
## Acarospora_americana           10.063136 -0.05316257
## Baeomyces_rufus                22.675291 -0.07026369
## Caloplaca_flavovirescens       14.502466 -0.06016599
## Candelaria_concolor            18.133937 -0.06408150
## Chrysothrix_candelaris         22.724069 -0.07880225
## Dimelaena_oreina               23.251313 -0.08699309
## Ephebe_ocellata                 6.952838 -0.05053982
## Flavoparmelia_baltimorensis    25.632005 -0.07435187
## Flavoparmelia_caperata         33.181624 -0.10872963
## Flavoparmelia_euplecta         30.393576 -0.08743442
## Flavoparmelia_haysomii         33.116308 -0.08668396
## Flavoparmelia_rutidota         23.544996 -0.07486043
## Flavoparmelia_soredians        31.957672 -0.08805189
## Flavopunctelia_flaventior      31.111218 -0.08745335
## Flavopunctelia_praesignis      21.632224 -0.07112771
## Flavopunctelia_soredica        24.427215 -0.07707090
```

```
## Graphis_scripta               25.574425 -0.08441539
## Ionaspis_lacustris            14.467153 -0.05284364
## Lecidea_tessellata            19.479761 -0.06245287
## Loxospora_elatina             33.413197 -0.05792092
## Neofuscelia_verruculifera     24.765806 -0.07605011
## Peltigera_elisabethae         16.657246 -0.05788947
## Pertusaria_ophthalmiza        26.139970 -0.08129887
## Rhizocarpon_grande            10.789637 -0.05209909
## Strigula_submuriformis        21.648791 -0.08028076
## Trypethelium_virens           18.038052 -0.07657339
## Umbilicaria_muehlenbergii     11.219725 -0.06039737
## Verrucaria_fuscella           13.412735 -0.05293292
## Xanthoparmelia_darrowii       31.292184 -0.07883280
##
## attr(,"class")
## [1] "coef.mer"
```

Hierarchical Models

Now I want to know from which taxonomic level much of the variation in slope and intercepts comes from. I have several questions about how to implement this.

Question 7: Which of the following models would be correct, if any?

1. spec_df[, '550'] ~ age + (1 + age|Class) + (1 + age|Class:Order) + (1 + age|Class:Order:Family) + (1 + age|Class:Order:Family:scientificName)
2. spec_df[, '550'] ~ age + (1 + age|Class) + (1 + age|Class:Order) + (1 + age|Class:Order:Family)
3. spec_df[, '550'] ~ age + (1 + age|Class) + (1 + age|Order) + (1 + age|Family) + (1 +age|scientificName)

7a: Model #1 differs from Model #2 by including species (scientificName) in the model; however, in Biometrics, we were told to leave off the lowest level of replication. Does that make sense here?

7b: User Macro on stackExchange states that you really should have thorough replication for each combination of levels indicated by the interaction terms. I don't think these data have that. For example, most of the specimens fall into a single class, and many families and orders only contain one species.

7c: If Model #1 or Model #2 is the best option, how do I interpret the interaction terms in the summary?

```
summary(lmer(spec_df[, '550'] ~ age + (1 + age|Class) + (1 + age|Class:Order) + (1 +        age|Class:(
           data = spec_df, REML = T,
           lmerControl(optimizer ='bobyqa', boundary.tol = 1e-5,
                       optCtrl = list(maxfun = 1e5))))
```

```
## boundary (singular) fit: see ?isSingular
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## spec_df[, "550"] ~ age + (1 + age | Class) + (1 + age | Class:Order) +
##     (1 + age | Class:Order:Family) + (1 + age | Class:Order:Family:scientificName)
##    Data: spec_df
## Control:
```

```
## lmerControl(optimizer = "bobyqa", boundary.tol = 1e-05, optCtrl = list(maxfun = 1e+05))
##
## REML criterion at convergence: 1491.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4973 -0.5429 -0.0580  0.5417  3.2442
##
## Random effects:
##  Groups                              Name        Variance  Std.Dev.  Corr
##  Class:Order:Family:scientificName (Intercept) 1.866e+01 4.320e+00
##                                     age         1.494e-03 3.866e-02 -0.82
##  Class:Order:Family                 (Intercept) 3.357e+01 5.794e+00
##                                     age         3.997e-05 6.322e-03 -1.00
##  Class:Order                        (Intercept) 5.947e+00 2.439e+00
##                                     age         1.148e-04 1.071e-02 1.00
##  Class                              (Intercept) 4.719e-07 6.870e-04
##                                     age         1.292e-11 3.594e-06 -1.00
##  Residual                                       2.328e+01 4.825e+00
## Number of obs: 237, groups:
## Class:Order:Family:scientificName, 29; Class:Order:Family, 19; Class:Order, 16; Class, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 18.58063    1.96197   9.470
## age         -0.06675    0.02789  -2.394
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.503
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

Final Questions

Question 8: What are the most important statistics to report for linear mixed models?

Question 9: Are there alternative models I should consider? Bayesian Hierarchical models?

References

Hajduk, G.K. (2019). Introduction to linear mixed models. https://ourcodingclub.github.io/tutorials/mixed-models/

Macro (https://stats.stackexchange.com/users/4856/macro), Questions about how random effects are specified in lmer, URL (version: 2013-08-11): https://stats.stackexchange.com/q/31634

Palmeri, M.(n.d.) Chapter 18: Testing the assumptions of multilevel models. https://ademos.people.uic.edu/Chapter18.html#1_preface

Schielzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Allegue, H., Teplitsky, C., Reale, D., Dochtermann, N.A., Garamszegi, L.Z., Araya-Ajoy, Y. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11:*1141-1152. https://besjournals.onlinelibrary.wiley.com/doi/epdf/10.1111/2041-210X.13434