Mixed effects Modelling 3 – checking

Starting point:

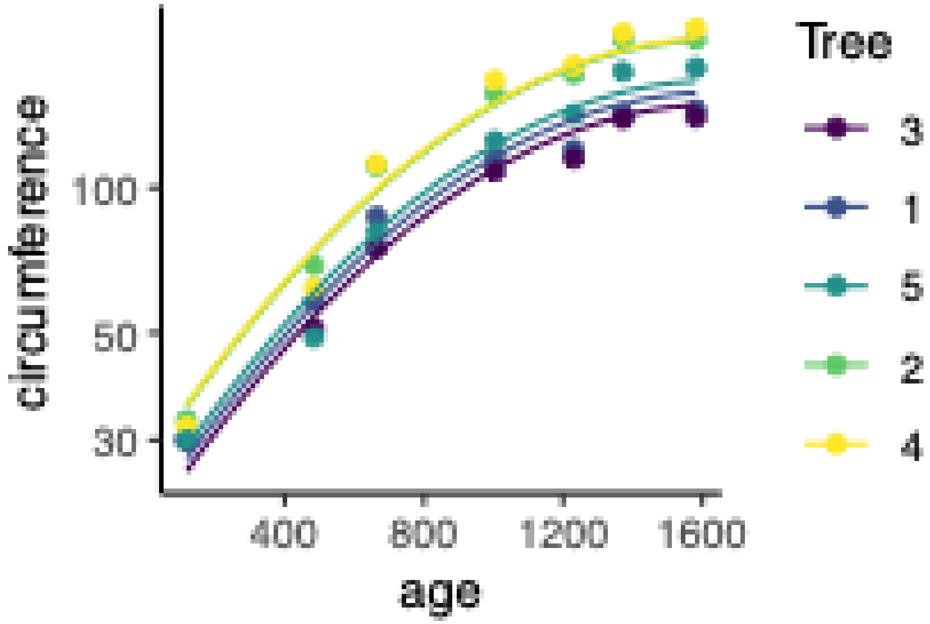
- Fitted mixed-effects model
- Scientific questions about the relationship between response and explanatory variables

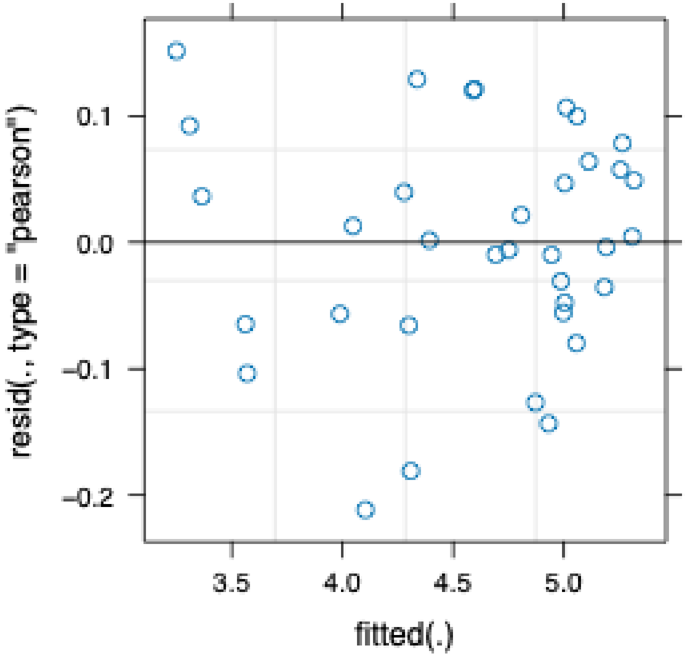
Example Experiment – variables

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)
- Data looked a bit problematic
  - Didn't seem very linear
  - Really bunched at low levels and spread out at high ones
- Addressed with log transform
- *Is the model ok?*
- If so, is that curviness real?

```
> plot(model)
```

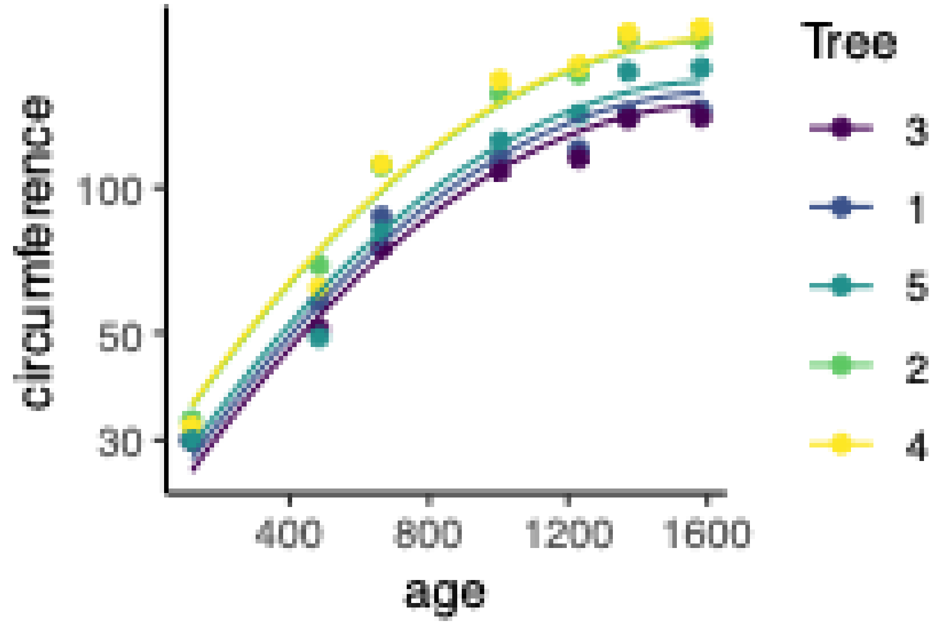
- Not *too* horrible





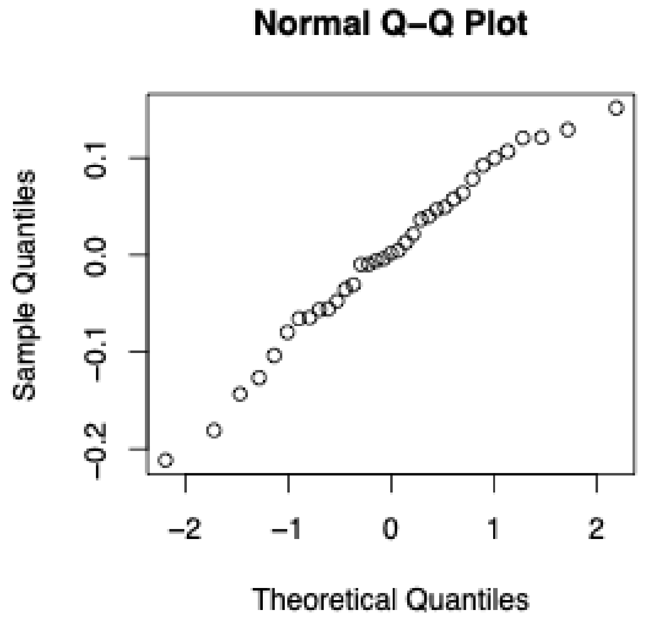
Example Experiment – variables

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)
- Data looked a bit problematic
  - Didn't seem very linear
  - Really bunched at low levels and spread out at high ones
- Addressed with log transform
- *Is the model ok?*
- If so, is that curviness real?



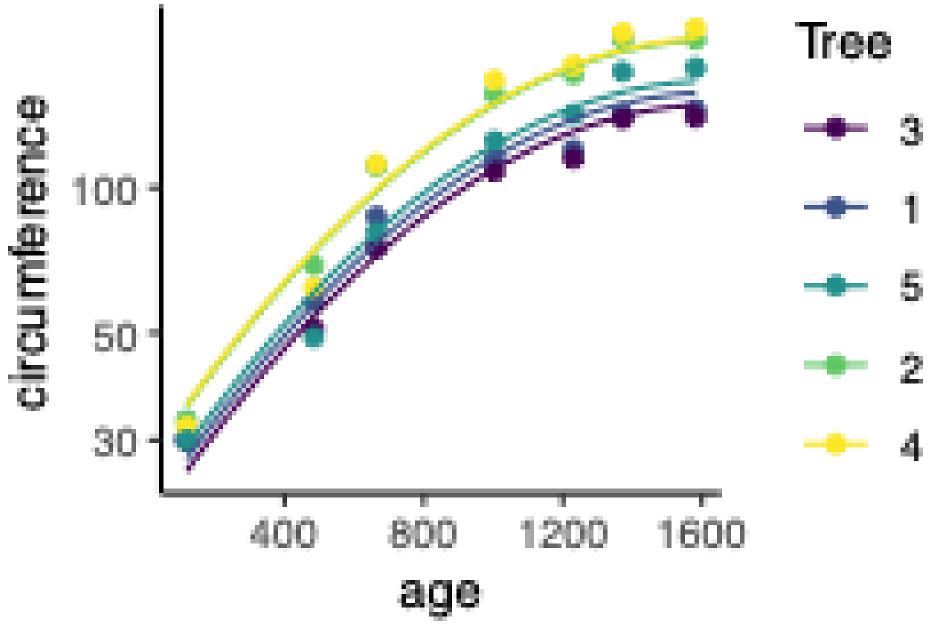

```
> qqnorm(resid(model))
```

- Not *too* horrible



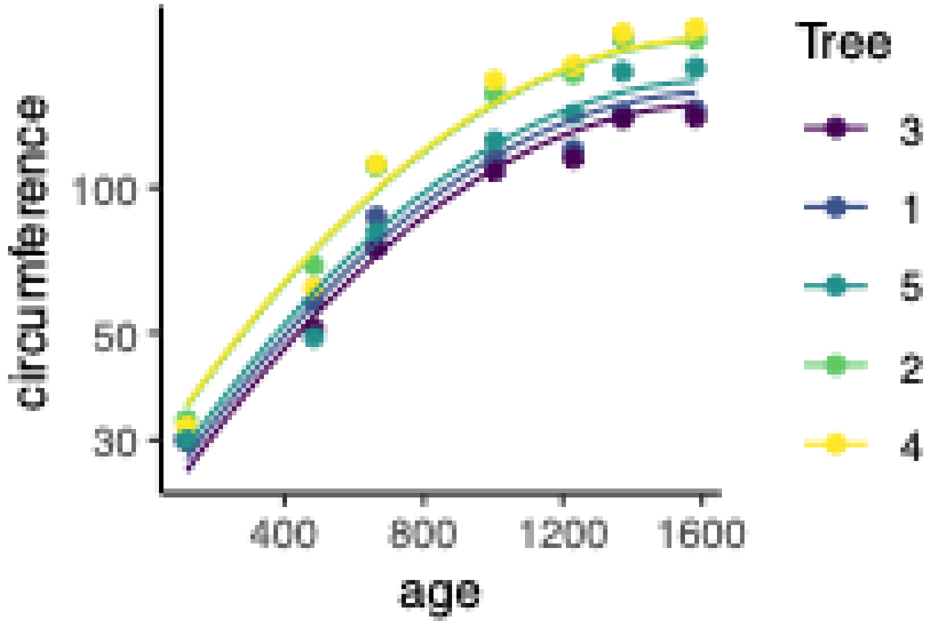
Example Experiment – variables

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)
- Data looked a bit problematic
  - Didn't seem very linear
  - Really bunched at low levels and spread out at high ones
- Addressed with log transform
- *Is the model ok?*
- Not *too* horrible
- If so, is that curviness real?



Example Experiment – variables

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)

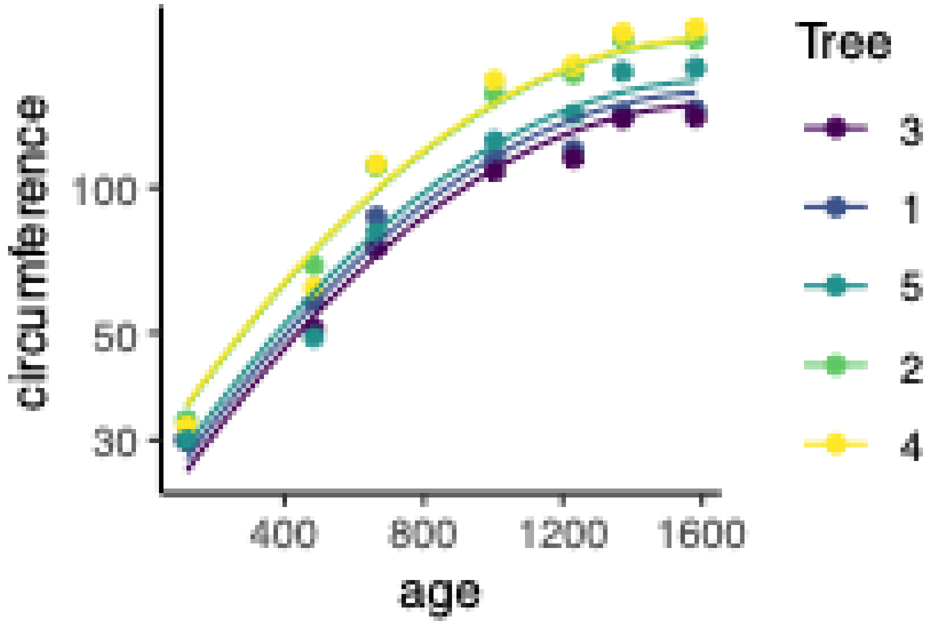



```
> summary(model)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(circumference) ~ age + I(age^2) + (1 | Tree)
   Data: Orange

REML criterion at convergence: 1.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.15991 -0.57319  0.01656  0.61931  1.54516

Random effects:
 Groups   Name        Variance Std.Dev.
 Tree     (Intercept) 0.022835 0.15111
 Residual             0.009594 0.09795
Number of obs: 35, groups:  Tree, 5

Fixed effects:
              Estimate Std. Error t value
(Intercept)  3.130e+00  8.654e-02  36.166
age          2.470e-03  1.448e-04  17.056
I(age^2)    -7.507e-07  8.206e-08  -9.148

Correlation of Fixed Effects:
         (Intr) age
age      -0.543
I(age^2)  0.470 -0.972
fit warnings:
Some predictor variables are on very different scales: consider rescaling
```

Example Experiment – questions

- Growth of orange trees
- Three variables:
    - Circumference (numeric, continuous)
    - Age (numeric, continuous, fixed)
    - Tree (numeric, categorical, random)




```
> summary(model)
```

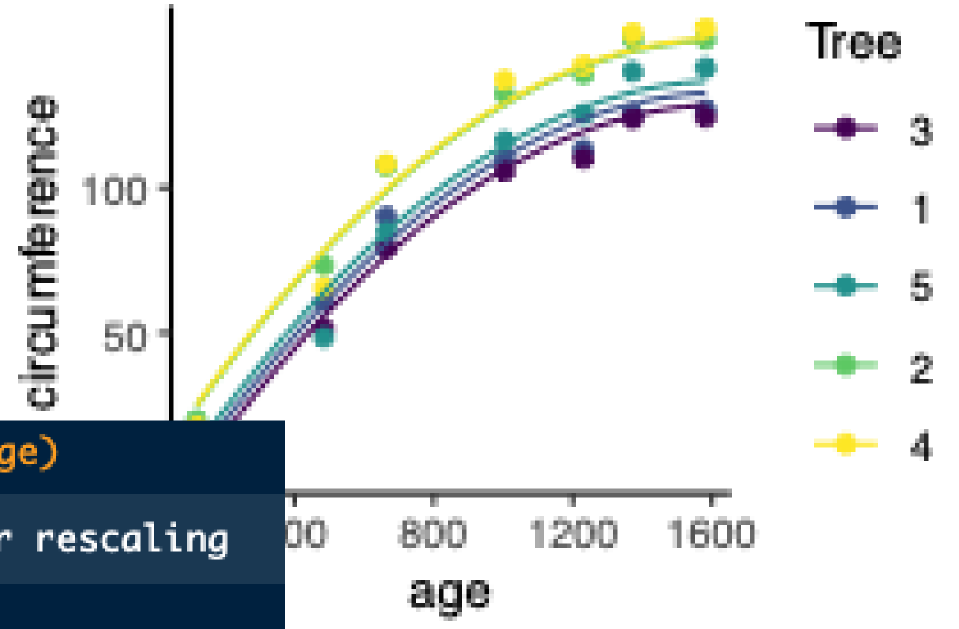
```
Fixed effects:
              Estimate Std. Error t value
(Intercept)  3.130e+00  8.654e-02  36.166
age          2.470e-03  1.448e-04  17.056
I(age^2)    -7.507e-07  8.206e-08  -9.148
```

- No Wald tests of fixed effects!
    - Deliberate choice (see https://is.gd/glmmFAQ)
    - Are approximations (e.g. via emmeans, lmerTest)
- Take alternative approach
    - Compare to a reduced model (without the curvature)
    - Test difference via likelihood ratio test, or AIC

```
m1 <- lmer(log(circumference) ~ age + I(age^2) + (1|Tree), data = Orange)
m2 <- lmer(log(circumference) ~ age + (1|Tree), data = Orange)
anova(m1,m2)
```

Example Experiment – questions

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)

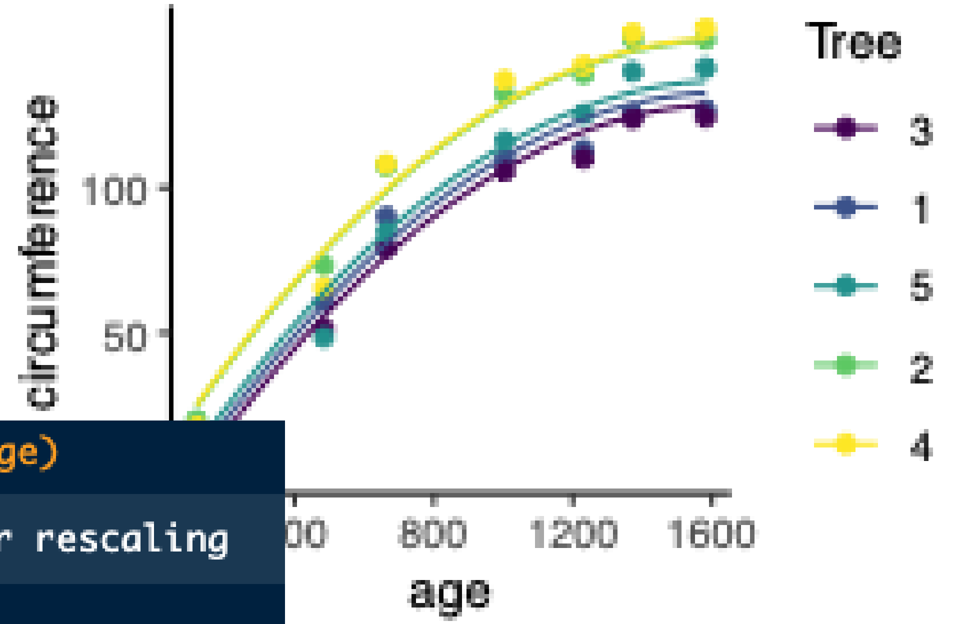



```
> m1 <- lmer(log(circumference) ~ age + I(age^2) + (1|Tree), data = Orange)
Warning: Some predictor variables are on very different scales: consider rescaling
> m2 <- lmer(log(circumference) ~ age + (1|Tree), data = Orange)
> anova(m1,m2)
refitting model(s) with ML (instead of REML)
Data: Orange
Models:
m2: log(circumference) ~ age + (1 | Tree)
m1: log(circumference) ~ age + I(age^2) + (1 | Tree)
   npar     AIC     BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m2    4  -2.631   3.590  5.3155   -10.631
m1    5 -42.135 -34.358 26.0674   -52.135 41.504  1  1.176e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Expect log likelihood to go up with more complex model
- But does it go up more than we expect? Yes $P = 1.2 \times 10^{-10}$
- This is a 'Likelihood Ratio Test'

Example Experiment – questions

- Growth of orange trees
- Three variables:
    - Circumference (numeric, continuous)
    - Age (numeric, continuous, fixed)
    - Tree (numeric, categorical, random)

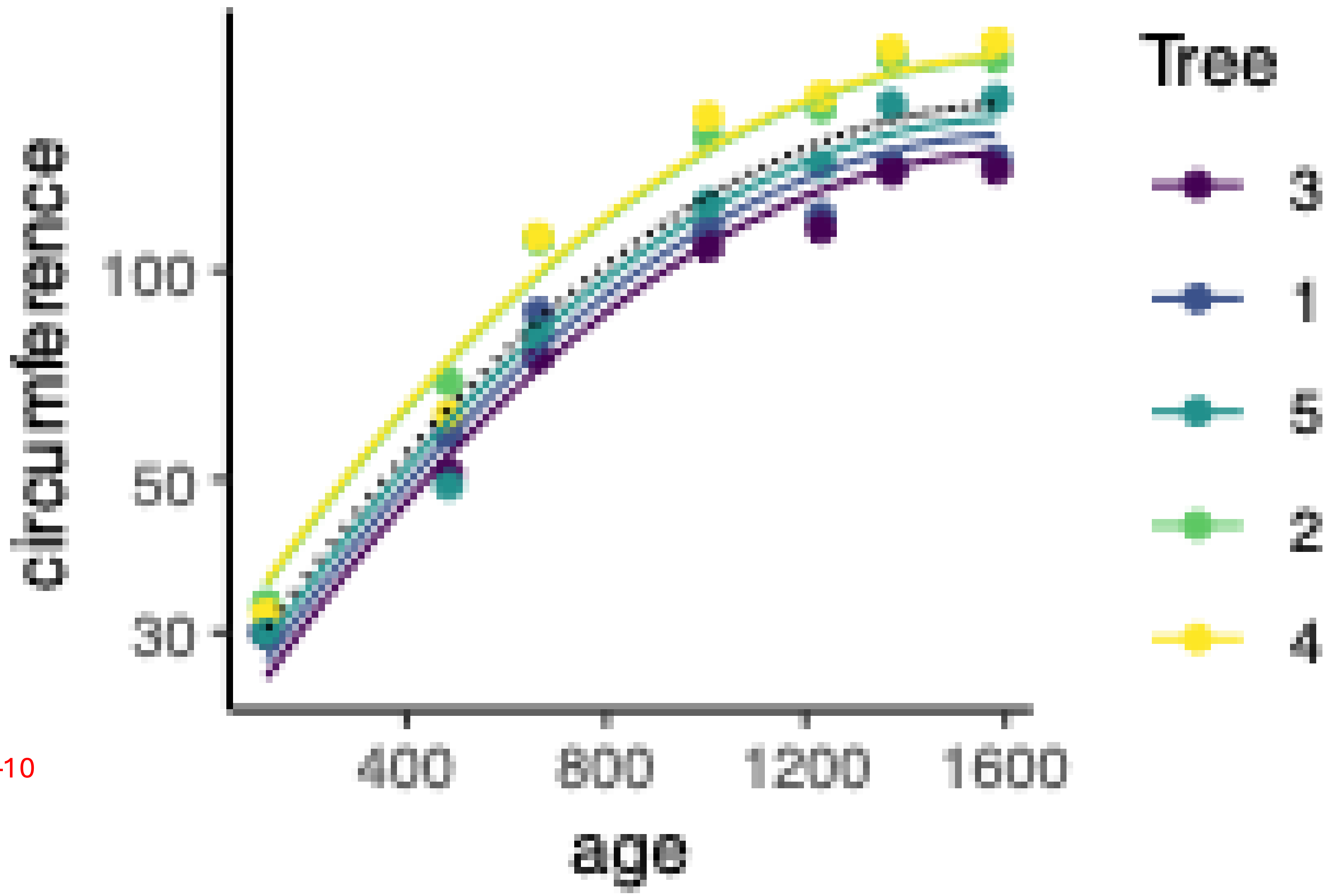



```
> m1 <- lmer(log(circumference) ~ age + I(age^2) + (1|Tree), data = Orange)

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

> m2 <- lmer(log(circumference) ~ age + (1|Tree), data = Orange)
> anova(m1,m2)

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

Data: Orange
Models:
m2: log(circumference) ~ age + (1 | Tree)
m1: log(circumference) ~ age + I(age^2) + (1 | Tree)
   npar     AIC     BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
m2    4  -2.631   3.590  5.3155   -10.631
m1    5 -42.135 -34.358 26.0674   -52.135 41.504  1  1.176e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- AIC (Akaike Information Criterion) gives balance of numbers of parameters and goodness of fit – lower is better (can be positive or negative)
- Usually agrees with LRT (tends to prefer slightly more complex models)
- Good for comparing a set of models

Example Experiment – questions

- Growth of orange trees
- Three variables:
  - Circumference (numeric, continuous)
  - Age (numeric, continuous, fixed)
  - Tree (numeric, categorical, random)



- *Is the model ok?*
  - Not *too* horrible
- If so, is that curviness real?
  - Yes, $P = 1.2 \times 10^{-10}$
- Go on to use the model, e.g. for understanding what an average tree will do (black line)