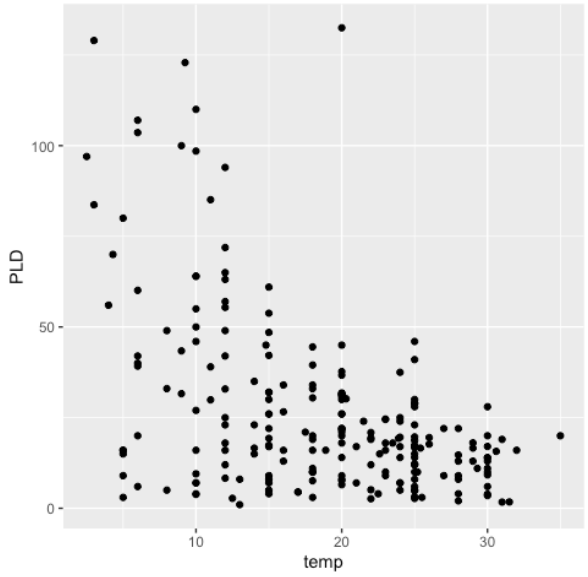
Lecture 22. Mixed models IV.

In the last lecture we looked at an example of an interaction between a fixed effect and a random effect (Treatment effect varied by Year). Now I'm going to go through a nice example showing how we can let regression slopes vary randomly as well.

This data is from a paper by O'Connor et al. (2007, PNAS, "Temperature control of larval dispersal…"). The authors compiled many lab studies where the pelagic larval duration of a benthic marine fish or invertebrate was measured at multiple temperatures (74 species total). Pelagic larval duration (PLD) is often measured in the laboratory by rearing larvae and measuring the time until the larvae settle. PLD is ecologically significant, because a longer time spent in the water column may result in greater dispersal distances, a greater likelihood of mortality during the sensitive larval stage, and a greater likelihood of being advected away from suitable habitat. It is already known that PLD tends to decline as temperature increases, because an increased metabolic rate speeds development to 'competency', or the stage at which larvae are ready to settle out of the water column. It is also known that species at higher latitudes are more likely to have either direct development (no pelagic larval stage) or lecithotrophic larvae (larvae are provided with a yolk sac, and tend to be larger), as opposed to planktotrophic larvae that get their nutrition solely by planktonic feeding (Thorsen's rule).

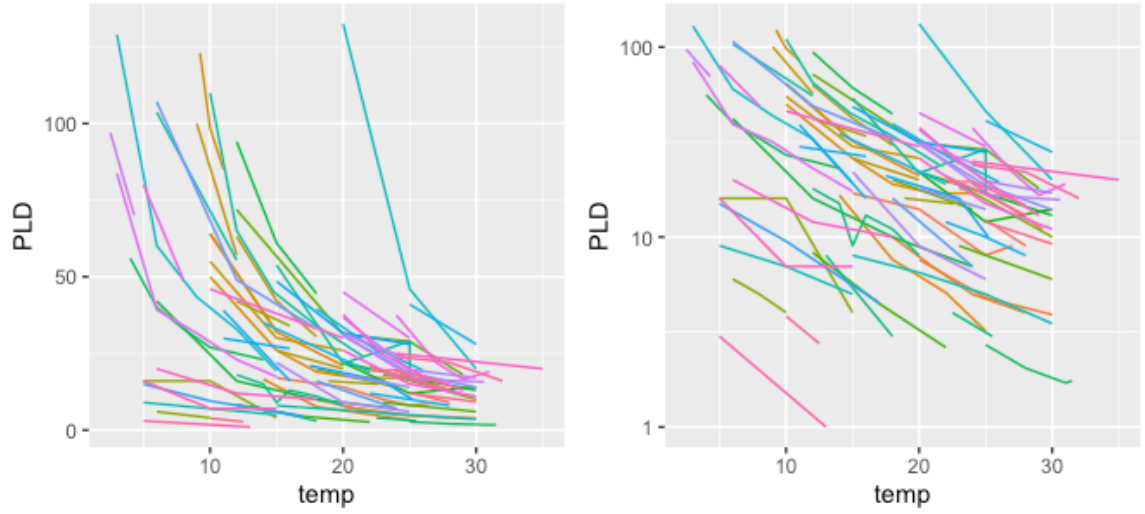
The authors were interested in testing whether the relationship between PLD and temperature was 'universal', i.e. do all species have the same temperature dependency? If true this would provide great predictive power for understanding these systems. They were also interested in quantifying how much cross-species variation there is in PLD, relative to the temperature effect, and whether some of this variation could be predicted by ecological factors (polar vs temperate vs tropical range; lecithotrophic vs planktotrophic larvae).

The raw data look like this:



PLD tends to decline nonlinearly with temperature, and prior work suggests that a power law is a good fit for the relationship. The compilation has a lot of data, but each species only has 2-6 measurements:

```
p1 = ggplot(pldata, aes(temp, PLD, col = species)) + geom_line(show.leg
end = FALSE)
p2 = ggplot(pldata, aes(temp, PLD, col = species)) + geom_line(show.leg
end = FALSE) + scale_y_log10()
grid.arrange(p1, p2, nrow = 1)
```



The plot on the left shows lines that connect the measurements for each species; the plot on the right is the same data on a log-log scale. As we discussed back at the beginning of the course, a power law looks like this: $Y = a*X^b$. If we take the log of both sides, we get:

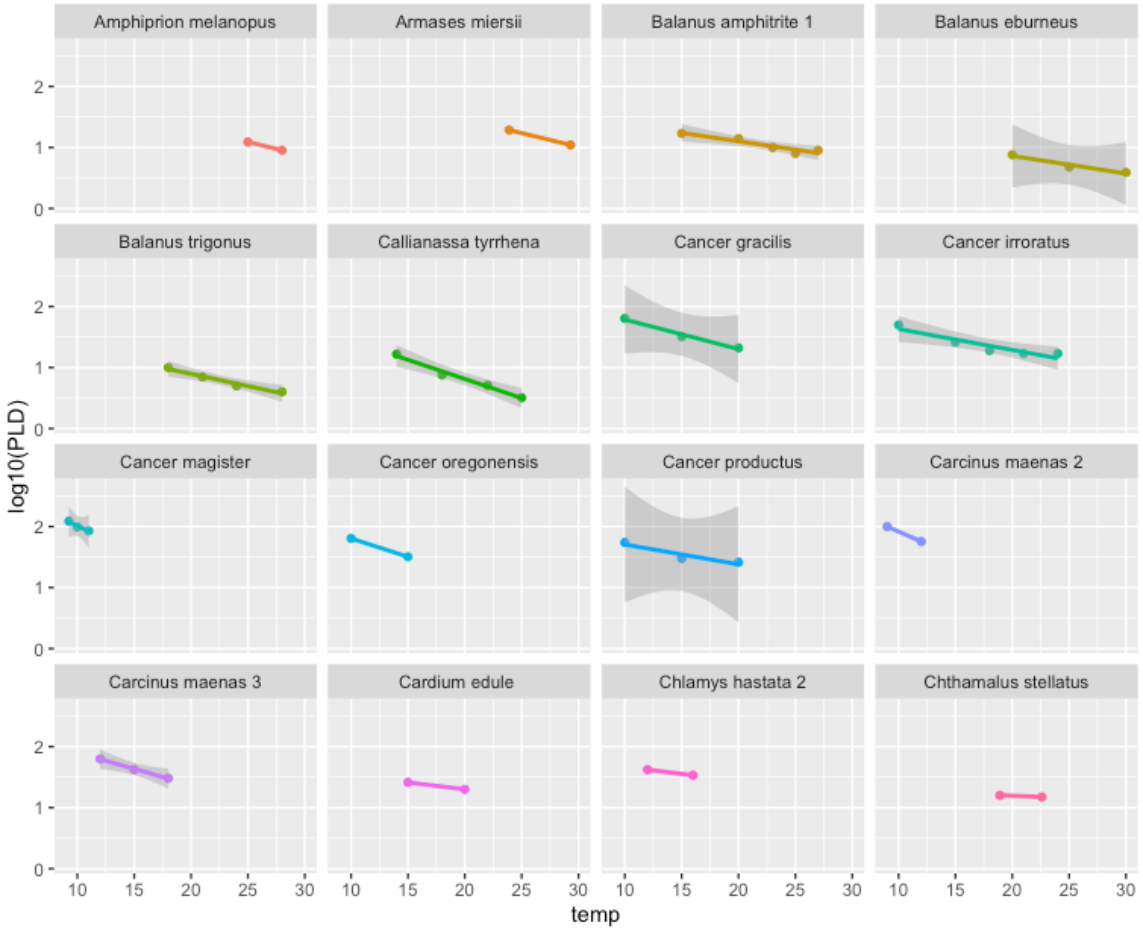
$$\log(Y) = \log(a) + b * \log(X)$$

This means we can esimate such a relationship with a linear model, if we assume that log(*Y*) is normally distributed (equivalent to *Y* being lognormally distributed).

These data, and the questions at hand, are perfect for mixed models. We don't want to fit 74 regressions, one for each species, especially because some species have little data. Alternatively, we can use random effects for the intercept and slope parameters, to quantify the amount of variation across species, while also allowing the species-specific estimates to be shrunk towards the overall mean, depending on how much data there is for each species. And we can include predictors to see if the among-species variation in intercept and slope can be explained.

The plot above makes it look like the power law (log-log relationship) is pretty good. We can get a better sense by fitting a linear regression for a subset of the species:

```
ggplot(subset(pldata, species %in% levels(factor(pldata$species))[1:16]), aes(temp, log10(PLD), col = species)) + geom_point(show.legend = FALSE) + geom_smooth(method = 'lm', show.legend = FALSE) + facet_wrap(~species)
```



This isn't informative for the species with 2 observations, but the others seems to fit a line well. Now we can specify a model with random effects for the intercept and slope:

```
mod = lmer(log10pld ~ log10temp + (1+log10temp|species), data = pldata)
```

Note how the random effect is specified: (1+log10temp|species). This is saying 'let the intercept and slope of the regression vary randomly by species'. I don't actually need to include the '1', as the random intercept will automatically be calculated when I include the slope, but I've specified it here to be clear. This is the same syntax we used in the butterfly example, where we wanted to see if Treatment varies by Year. Now instead of letting the treatment effect vary randomly by group, we are letting the slope in a regression vary randomly by group. For this reason, this kind of model is often called a 'varying slopes model'. Technically, I should call it a 'varying intercepts and slopes model', because both are varying. It is possible to let either or both vary by group, as illustrated in this plot from Gelman & Hill:

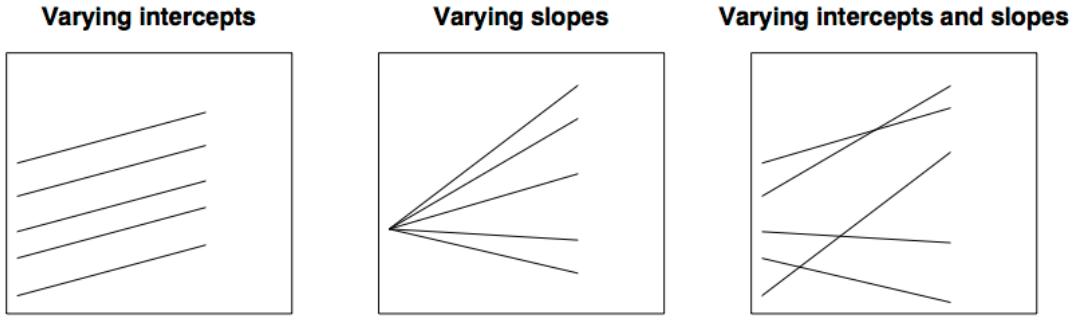


Figure 11.1 *Linear regression models with (a) varying intercepts ($y = \alpha_j + \beta x$), (b) varying slopes ($y = \alpha + \beta_j x$), and (c) both ($y = \alpha_j + \beta_j x$). The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between x and the group indicators.*

So we can compare how much species vary in PLD, regardless of temperature (intercept), and also how much they vary in their response to temperature (slope). Also note that I've transformed PLD and temperature using the base-10 logarithm. This is just to help the interpretation of the parameters. For a base 10 logarithm, if $\log_{10}(X)$ increases by 1, that means X increases by a factor of 10 (and vice versa).

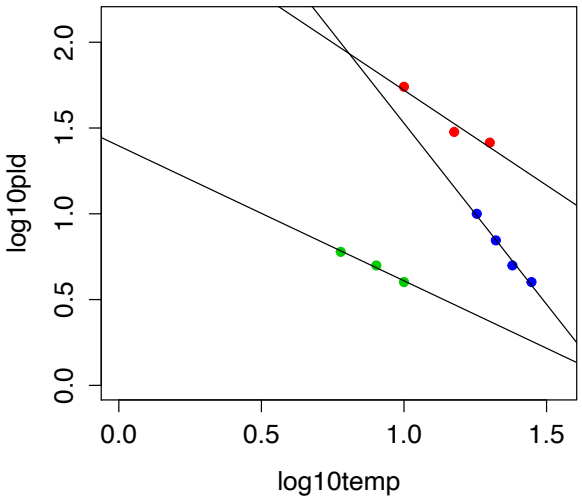
Let's look at the output:

```
summary(mod)
```

```
## Linear mixed model fit by REML ['merModLmerTest']
## Formula: log10pld ~ log10temp + (1 + log10temp | species)
##    Data: pldata
```

```
## ## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  species  (Intercept) 0.81483  0.9027
##           log10temp   0.28947  0.5380   -0.92
##  Residual             0.00435  0.0659
## Number of obs: 218, groups:  species, 74
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   3.0816     0.1289 60.4000    23.9   <2e-16 ***
## log10temp    -1.4566     0.0851 55.8000   -17.1   <2e-16 ***
```

As discussed in last lecture, the model fits three parameters for the random effects: the variance for intercept, the variance for log10temp, and the correlation between these two random effects. In this case the correlation is quite high, -0.92. This is a common occurrence in this kind of model, and it happens because the temperatures measured are typically quite far from zero, which induces a correlation between the intercept and the slope. Let's visualize this using three species:



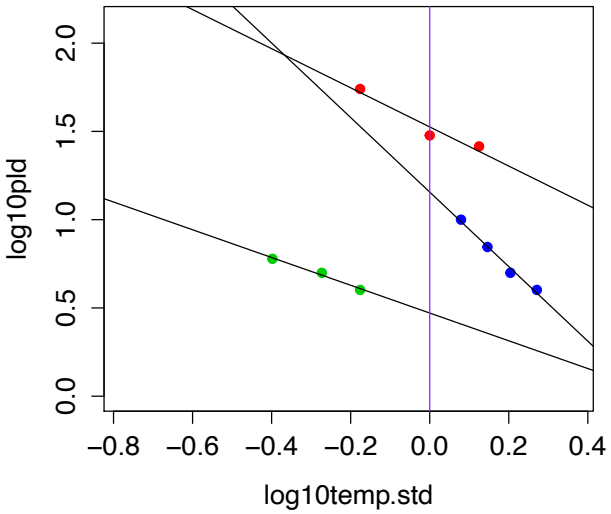
When a line is fit to each of these three species (the three colors), the intercept is the point at which the line crosses log10temp = 0. But the data actually ranges from 0.7-1.5 on the x-axis. That means that a species with a steep slope (blue species) is going to have a really high intercept, and a species with a shallow slope (green) is going to have a lower intercept. Effectively, the intercept and the slope aren't really telling us different things.

Centering the predictor

To fix this, we can *center* the predictor variable, so that it contains both positive and negative values:

```
pldata$log10temp.std = pldata$log10temp - log10(15)
```

Here I've centered log10temp around log(15), because 15ºC is roughly in the middle of temperatures used in these experiments. Now we can fit lines using the centered predictor:



The same relationships are fit (centering the x-axis doesn't change where the points are relative to each other), but now the intercept (where the x-axis is zero) lies in the middle of the data. This means that the intercept is now doing a better job of telling use whether as species has a high or low PLD, after controlling for the effect of temperature. So in this case the red species has a higher intercept than the blue species, whereas in the previous plot the blue species had a higher intercept.

Now let's fit a model using the centered predictor.

```
mod = lmer(log10pld ~ log10temp.std + (1+log10temp.std|species), data = pldat
a)
summary(mod)
```

```
## Linear mixed model fit by REML ['merModLmerTest']
## Formula: log10pld ~ log10temp.std + (1 + log10temp.std | species)
##    Data: pldata
##
## Random effects:
##  Groups   Name          Variance Std.Dev. Corr
##  species  (Intercept)   0.16409  0.4051
##           log10temp.std 0.28947  0.5380   -0.49
##  Residual               0.00435  0.0659
## Number of obs: 218, groups:  species, 74
```

```
##
## Fixed effects:
##                Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)     1.3685     0.0482 72.4000    28.4   <2e-16 ***
## log10temp.std  -1.4566     0.0851 55.8000   -17.1   <2e-16 ***
```

Now the correlation between the intercept and the slope is -0.5. This is less likely to be an artifact, and more likely to be actual information: Species with a higher overall PLD also tend to respond more steeply to temperature (have a more negative slope).

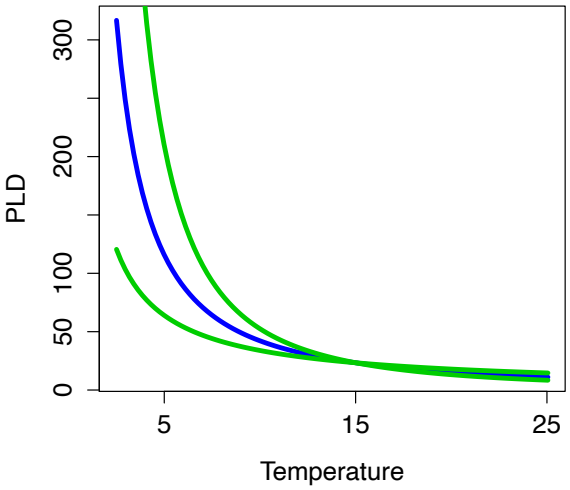
Let's think about what the variance components mean: the fixed effect for intercept is 1.37 (on the original scale, $10^{1.37} = 23$ days). The random effect has a standard deviation of 0.4, which means species typically range from $10^{0.97}$ to $10^{1.77}$, or 9 to 59 days. So there is clearly considerable variation among species in pelagic larval duration, which is already well known. The random effect for the slope is a little trickier to parse. The fixed effect for log10temp.std is -1.46. Our regression is intended to fit a power law, PLD = a*Temperature^b. But now that we've centered the predictor, what we've fit is PLD = a*(Temperature/15)^b, because taking the log of both sides gives

$$\log(PLD) = \log(a) + b * (\log(Temp) - \log(15))$$

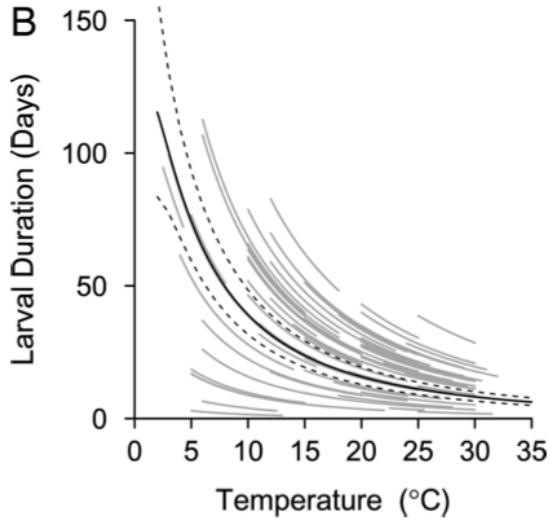
And the part multiplied by b is our centered predictor. So now let's plot the fixed effects fit with curve:

```
a = 10^fixef(mod)[1]
b = fixef(mod)[2]
curve(a*x^b, from = 0.167, to = 1.67, xaxt = 'n', col = 'blue', lwd = 4, xlab
= 'Temperature', ylab = 'PLD')
axis(1, at = c(5/15, 15/15, 25/15), labels = c(5, 15, 25))
b.plusSD = fixef(mod)[2] + 0.54
curve(a*x^b.plusSD, from = 0.167, to = 1.67, col = 'green3', lwd = 4, add = T)
b.minusSD = fixef(mod)[2] - 0.54
curve(a*x^b.minusSD, from = 0.167, to = 1.67, col = 'green3', lwd = 4, add =
```

```
T)
```

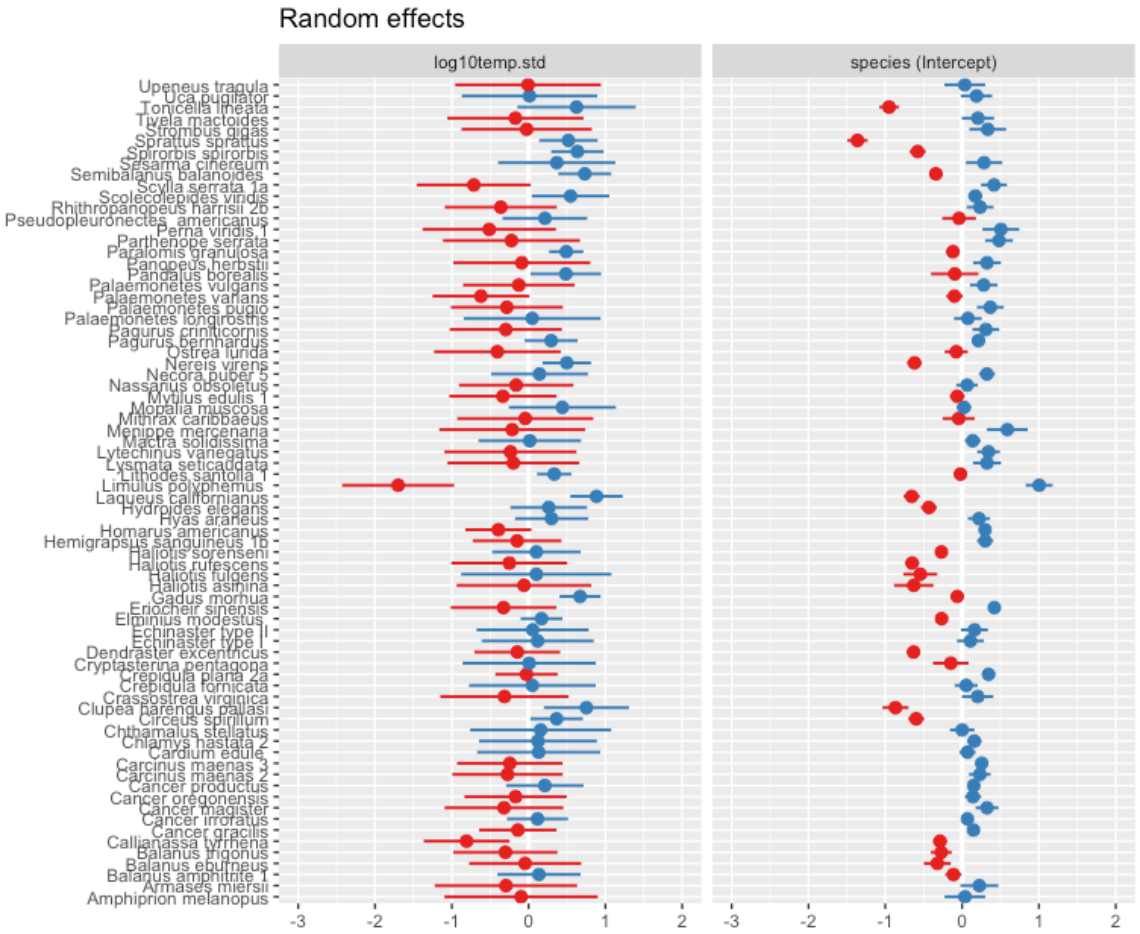


The blue line is from the fixed effects intercept and slope, and the green lines come from adding or subtracting 1 random effect Std. Dev. from the slope. So we can see that the variation in slope across species is not insubstantial; some species are definitely more temperature-sensitive than others, but at the same time all species still have a pretty important sensitivity of PLD to temperature. We could also use curve() to plot the species-specific model predictions, i.e. combine the intercept and slope estimate for each species, as I did in a previous example. But for now I'll just include the plot from the paper:



We also can visualize the random effects variation a caterpillar plot:

```
plot_model(mod, type = 're')
```



The species-specific intercepts have pretty tight confidence intervals, indicating we have enough data to get a good sense for variation in PLD across species. The species-specific slopes have a bit more uncertainty; they do vary, as indicated by the random effects variance, but each individual slope is fairly uncertain, which is not surprising due to the limited number of temperatures used in each experiment. The authors of the paper used this plot to say that the temperature-dependence of PLD is pretty 'universal', with minor variation among species. Personally I might emphasize the variation a bit more, but I can see their point. If one wanted to test whether the variation in slopes is significant, a likelihood ratio test would be appropriate:

```
mod = lmer(log10pld ~ log10temp.std + (1+log10temp.std|species), data = pldat
a, REML = FALSE)
mod.noslope = lmer(log10pld ~ log10temp.std + (1|species), data = pldata, REML
 = FALSE)
anova(mod, mod.noslope)

## Data: pldata
## Models:
## ..1: log10pld ~ log10temp.std + (1 | species)
## object: log10pld ~ log10temp.std + (1 + log10temp.std | species)
##        Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## ..1     4 -124 -110   66.0     -132
```

```
## object  6 -153 -133   82.6     -165  33.1      2    6.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly the variation in slopes across species is supported by the data.

Adding in predictors for the random intercepts and slopes

The model already has a fixed effect predictor that varies within species, as well as across species (temperature). We can also add in predictors to test ideas about variation across species in their adaptation to temperature. The dataset includes both lecithotrophic and planktotrophic species, 18 and 56 respectively. These life history / dispersal strategies may explain some variation in pelagic larval duration (i.e. the intercepts). It's possible that these strategies affect the sensitivity of PLD to temperature as well, so out of curiosity we can test that interaction. Another interesting predictor is the climate 'biome' of the species, i.e. they have been coded into polar / temperate / tropical species (there are 8, 53, and 13 of each). This variable should reflect general patterns of adaptation to different temperature regimes, and so we can see if it predicts intercepts and/or slopes as well.

```
mod.level2 = lmer(log10pld ~ log10temp.std + log10temp.std*climate +
log10temp.std*feeding.type + (1+log10temp.std|species), data = pldata)
```

As we've seen in prior examples, adding in these 'level 2' predictors is specified in the same way as 'level 1' predictors. Because we're testing two interactions here, I first want to see if these interactions are significant; if not, I will drop them from the model, because it is much easier to interpret/test main effects without interactions in the model.

```
anova(mod.level2, ddf = "Kenward-Roger")

## Analysis of Variance Table of type 3  with  Kenward-Roger
## approximation for degrees of freedom
##                             Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)
## log10temp.std                1.420   1.420     1  71.0   164.1 <2e-16 ***
## climate                      0.002   0.001     2  78.5     1.2 0.2957
## feeding.type                 0.039   0.039     1  70.6     6.5 0.0127 *
## log10temp.std:climate        0.053   0.026     2  60.6     5.9 0.0044 **
## log10temp.std:feeding.type   0.002   0.002     1  60.7     0.3 0.5569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks like climate affects the slope of response to temperature, but feeding type does not. So I'll remove the latter interaction, leaving the main effect of feeding type:

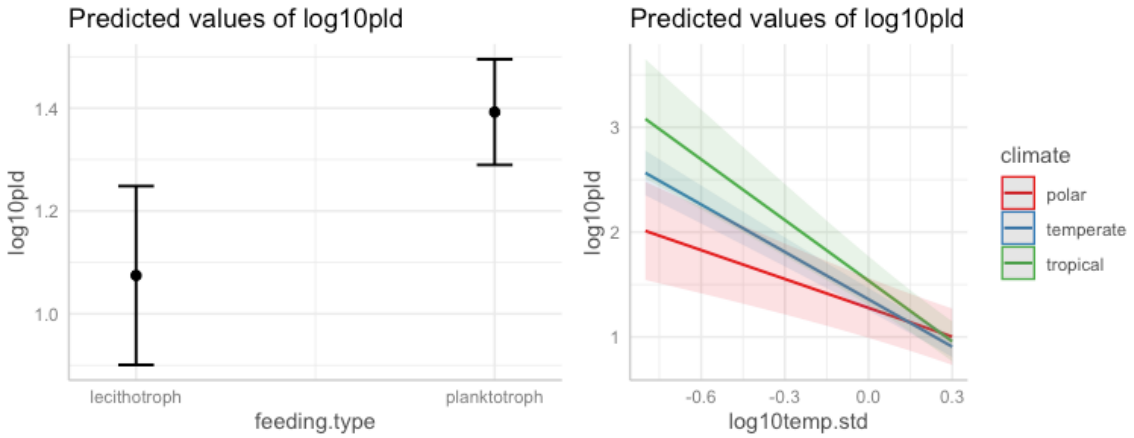
```
mod.level2 = lmer(log10pld ~ log10temp.std + log10temp.std*climate +
feeding.type + (1+log10temp.std|species), data = pldata)
```

```
anova(mod.level2, ddf = "Kenward-Roger")
```

```
## Analysis of Variance Table of type 3  with  Kenward-Roger
## approximation for degrees of freedom
##                        Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)
## log10temp.std          1.458   1.458     1  72.8   176.2 <2e-16 ***
## climate                0.002   0.001     2  79.8     1.2 0.3082
## feeding.type           0.040   0.040     1  70.6     9.4 0.0031 **
## log10temp.std:climate  0.054   0.027     2  53.4     6.2 0.0038 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So it looks like mean PLD varies by feeding type, while there is an interaction between climate and temperature. Note that the tests for the main effects of log10temp.std and climate are kind of sketchy, because we know there is an interaction, and that's what I will focus on. Let's visualize these results in a few different ways.

```
grid.arrange(grobs = list(plot(ggeffect(mod.level2, terms = "feeding.type")), plot(ggeffect(mod.level2, terms = c("log10temp.std", "climate")))), nrow = 1)
```



The model says that planktotrophs tend to have a higher PLD than lecithotrophs. The difference is about 0.3 on the log10 scale, which is a factor of two. So on average planktotrophic larvae stay in the water column twice as long as lecithotrophic larvae. The cause of this is not clear from the data, but lecithotrophic larve may develop faster due to the nutritional supplement, and they also tend to be larger, and therefore may require less time to mature. The interaction between temperature and climate is intriguing. It looks like polar species have the shallowest slope, while tropical species have the steepest slope. We don't have a ton of representatives from polar and tropical environments, but the implication is that species in colder environments have evolved to be less sensitive to temperature. Based on the effects plot, this specifically means that polar species do not suffer as much from low temperature, in terms of having a longer PLD, compared to tropical species, and temperature species are intermediate. This is consistent with the idea that having a very long PLD will increase a larva's probability of dying or being advected away from suitable habitat, and therefore adaptations to develop faster at low temperature are advantageous.

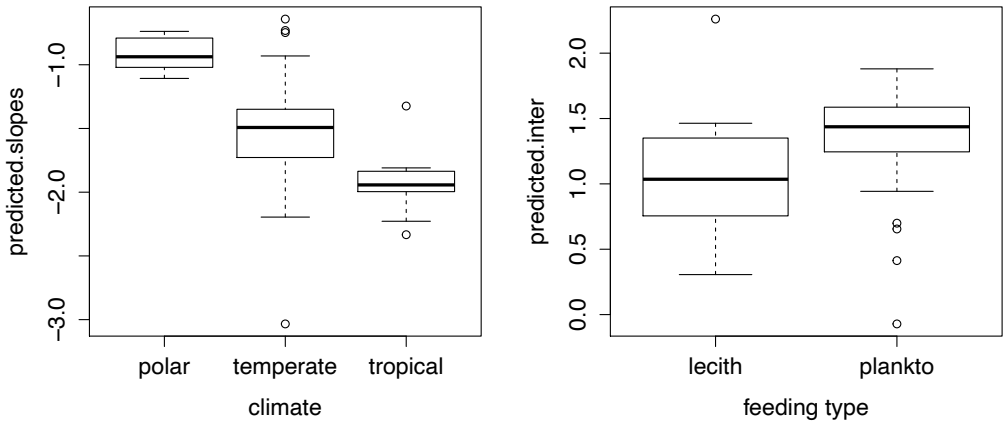
To get a sense for how much variation these predictors explain, we could use the variance components to get multi-level R^2, as I described in a previous lecture. For now I'll just report the whole-model estimate from r.squaredGLMM:

```
> r.squaredGLMM(mod.level2)
      R2m       R2c
0.3621605 0.9823193
```

The two predictors in total explain 36% of the variation, which is decent but clearly there is plenty of unexplained variation in PLD. Including the random effects bumps this up to 98%. This is because we are effectively fitting a regression for each species, and the residual noise around these regressions is low (the variance component for the residuals is only 0.06). This means the unexplained variation is due to across-species variation, not measurement noise. We can also get a sense for how well across-species variation is predicted by plotting the fixed+random species-specific predictions vs. the predictors, as we've seen previously:

```
which.climate = tapply(pldata$climate, pldata$species, function(x) x[1])
predicted.slopes = ranef(mod.level2)$species[["log10temp.std"]] +
fixef(mod.level2)["log10temp.std"] +
fixef(mod.level2)["log10temp.std:climatetemperate"]*(which.climate == 2) +
fixef(mod.level2)["log10temp.std:climatetropical"]*(which.climate == 3)
plot(predicted.slopes ~ factor(which.climate))

which.feeding = tapply(pldata$feeding.type, pldata$species, function(x) x[1])
predicted.inter = ranef(mod.level2)$species[["(Intercept)"]] +
fixef(mod.level2)["(Intercept)"] +
fixef(mod.level2)["feeding.typeplanktotroph"]*(which.feeding == 2)
plot(predicted.inter ~ factor(which.feeding))
```



The species-specific slopes are definitely pretty strongly differentiated by the climate variable; in contrast, feeding type has an effect but there is much more residual overlap in the species-specific intercepts.