Week 6 Computer Lab Introduction: Quantitative Genetics

VIDEO 1

Scene 1: Quantitative Genetics

Hello again! This week, we'll do quantitative genetics to see whether Western white pine shows adaptation in a trait called d13C. This is the ratio of 13C to 12C isotopes and correlates with water use efficiency. We have data from a common garden experiment, where seeds were grown under the same condition and measured for d13C. From these data, we'll estimate the heritable trait of the mother tree. Then we test whether the trait shows a pattern of variation that is different from gene flow in neutral markers, and we'll correlate the trait with environmental data.

This video addresses some of the methods challenges related to this lab. We'll need to work with linear mixed models, which is a step up from previous week's regression models, and we'll want to cover a number of concepts that will help you better understand the worked example.

After an overview of the design of the Western white pine study, I'll give a quick introduction to SNP data. Then, we'll cover the types of models available and how to choose between them, and how to tell R what model we want to fit. In fact, R uses model formulas that are kind of cool once you got the hang of it.

When we go from basic regression or ANOVA to linear mixed models, things can get confusing. We won't go into details here, but in a second video, we'll have a quick look under the hood to get a better idea why we can't use standard tests for these models, and how to interpret them.

Scene 2: Experimental Design

Here's the study area around Lake Tahoe in the Western United States. The study used a hierarchical sampling design with 10 populations. In each population, 3 plots were selected, and we have bioclimatic data at the plot level. In each plot, between 2 and 8 trees were sampled, these are called 'Families'. For each tree, we have genotypic data from 164 SNP markers. From each tree, between 5 and 35 seeds were collected. An average of five seeds per family were randomly assigned to five blocks in a common garden experiment. The seeds were germinated and seedlings were measured for the d13C trait.

The lab starts with the seedling data, for which we have trait measurements but no molecular data. From the offspring trait data, we want to estimate the trait. Remember that trait expression is a combination of the genotype, environment, and any interaction between genotype and environment. So, if we had measured the trait of the adult trees in the field, we would not know what part is due to the genotype. With the seedlings grown in a common garden, we assume that the environment is constant, hence all variation is due to the genotype.

That simplifies the analysis. However, we do need to account for the hierarchical sampling design, and for the blocks used in the common garden to account for any variation in micro-environmental conditions.

Once we have estimated the heritable trait of the mother trees, we can test whether differentiation at the quantitative trait, measured by Qst, is different from differentiation at neutral markers, measured by Fst. That's where we will use the genetic differentiation observed for a set of SNP markers as a null model to represent patterns of gene flow.

At the plot level, we also have 7 bioclimatic variables, so we can find out whether the heritable trait is indeed related to environmental selection, or if it can better be explained by geographical patterns that we'll approximate by including the spatial coordinates in our model.

Scene 3: Introducing SNP Data

Here's a preview of the SNP data set. Each row represents one adult tree sampled in the field. Then we have the columns with the SNP data. SNPs are codominant, so that we have two alleles for each diploid individual. In this case, we only know the bi-allelic genotype, but we don't know the haplotype. That means, we don't know which alleles of different markers were inherited together on the same chromosome copy. Each SNP is a combination of A and C, or G and T values. One of the two represents the wildtype, the other the mutant. Some SNPs may be monomorphic, that means, there are no mutants in this sample. Each SNP is less variable than a typical microsatellite marker, but we often have hundreds or thousands of SNPs, which makes them very informative. We don't usually know a priori whether each SNP behaves as a neutral marker or shows a signal of selection, but a vast majority of markers will likely reflect neutral patterns of gene flow. We'll look into this later in the course in Week 11.

In summary, we have 164 SNP markers, four of which are monomorphic. The markers have up to 20% missing values, which is common for this kind of data. In this case, we already know that the SNPs are unrelated to the quantitative trait of interest.

Scene 4: What Type of Linear Model to Fit?

We will fit a number of different models in today's lab. The first question is about the variable types of the response and predictor variables. As an example, we want to model the measured trait d13C of the seedlings as a function of 'Family', 'Population' and 'Block'. Later in the example, we'll add a covariate 'AWS', which is a quantitative variable that represents soil available water supply.

Our response variable 'd13C' is quantitative. The factors 'Family' and 'Population' and 'Block' are categorical. According to this table, we should consider an ANOVA. If we add AWS as a covariate, it will become an ANCOVA, which means an analysis of variance with a covariate. Similarly, if we add a categorical factor to a regression analysis, we end up with a regression with dummy variables.

It's interesting to see that with this approach, most of your standard stats course ends up in this table. You may have learned each method separately, but they are just variations on a theme, the theme being a linear model. For the rest of this lab, we'll discuss things in the context of an ANOVA.

You'll probably remember from your intro stats course that each method has a set of conditions and assumptions. Among other things, basic ANOVA requires that the residuals approximate a normal distribution, and that the design is balanced, which means that each group has the same sample size. That's shown here in the line for 'LM', or linear model. If the design is unbalanced, we can use a linear mixed model, or 'LMM'. In other words, 'LMM' does not have this condition, hence you can use it for balanced or unbalanced designs. Similarly, if the data do not follow a normal distribution, even after transformation, we can use a Generalized linear mixed model, or 'GLMM'. GLMM can handle balanced or unbalanced designs.

The next question to address is whether our factors are nested or crossed. Nested factors are common in hierarchical sampling designs. Here, trees are nested within plots, and the plots are nested within populations. For now, we'll drop the plots and consider 'Family' to be nested within 'Population'. It is nested because each tree can occur in only one population, not in all of them.

The use of blocks in the common garden experiment is a typical case of a factorial design with crossed factors. The seeds from each family were randomly assigned to blocks. Ideally, all combinations of family and block are equally frequent, as shown here with one seed per family per block. That would be a balanced design. However, the design here comes close to it but there are deviations, as shown here, which makes the design unbalanced.

Now we get to a difficult topic: fixed or random factors. I've read that "one person's fixed effect is another person's random effect", and if you read more about this, you will find that there are different philosophies that can be very confusing.

I follow this argumentation: think about what would happen if you replicated the study in a different study area. Would you observe the same factor levels, or different ones? Let's consider 'Family'. Each tree represents a genotype, and clearly, we would be sampling a different set of genotypes in a different study area. That would make 'Family' a random factor, and the same goes for 'Population'. We can make a similar case for 'Block'. Each block represents a specific micro-environment, and clearly, we would be sampling different micro-environments with blocks in a different common garden. That would make 'Block' a random factor.

However, now a second criterion kicks in. We need to have many levels to fit a random effect. Why is that? For a fixed factor, we would estimate a parameter for each factor level. That means if we treat 'Family' as a fixed factor, we need to estimate 157 group means. That's a lot of parameters! We can reduce that to all but two parameters if we treat 'Family' as a random

factor. Then we'll only estimate a mean and standard deviation of the group means. That's two instead of 157 parameters! No worries, we can still estimate an average trait for each family from the results.

Estimating the standard deviation among group means requires a reasonably large number of groups, or factor levels. I've seen six listed as a minimum, but I'm not sure there is a consensus about that. In general, standard deviations are much harder to estimate with any precision than means, so even six sounds like a very low number. Here we have only five blocks, hence it may be safer to treat 'Block' as a fixed factor.

The last point here is about interactions. In a controlled experiment, we are usually interested in interactions between crossed factors. In this interaction plot, we have two Blocks A and B on the x-axis, and two populations shown by different colors. For each group, that is, for each combination of Block and Population, we have error bars that show the group mean plus/minus one standard error. The group means for the same population are connected with a line. Here, the effect of 'Block' depends on the population, and vice versa. For population Pop 1, there was no difference between blocks, but for population Pop 2, trait values were higher on average in Block B than in Block A.

What would this plot look like if there was no interaction? If neither Block nor Population had an effect, all means would be the same. If Block had an effect but not Population, the two lines would fall on top of each other. If Population had an effect but not Block, we would get two parallel horizontal lines. And if Population and Block both had effects, but their effects were independent of each other, we would get parallel lines with a slope. This indicates that the effects are additive.

If we combine one categorical factor and one quantitative covariate, we are looking at a scatterplot with separate regression lines for each group. Here, the lines have the same slope but different intercepts, which means that population Pop 1 has higher d13C levels than Pop 2 but the association with available water supply is the same. Parallel lines mean that the effects are additive, which is to say that the effect of one predictor does not depend on the value of the other.

Here we have an example of non-additive effects, which means a true interaction. The two groups have different slopes, and in this case, Pop 1 shows no association between available water supply and the trait, whereas there is a strong correlation between the two in Pop 2. When we specify the model, we will have to tell R whether or not we want to consider such interactions.

You can now use the five questions on this slide to decide for your own data set what kind of model is most appropriate, whether your factors are nested or crossed, fixed or random, and what kind of interactions you want to include, if any.

Scene 5: Specify Model Formula in R

So how do we tell R about all these decisions? R uses a special type of language for this, called model formulas. Many functions can interpret model formulas and thus it is really worth learning about them.

Let's consider our categorical factors first. The simplest model is simply the global mean of the trait d13C among all seedlings from the common garden experiment. Here's the formula for fitting a global mean. On the left side, we have the response, and on the right side of the tilde symbol, we have a number one, which represents the intercept. If we fit one intercept only, this will be equal to the mean of all values.

If we want to fit a different mean for the seedlings from each population, we simply add 'Pop' on the right side. To add 'Block' as a second factor with additive effects, no interaction, we use the plus symbol. If we want to include the interaction between 'Pop' and 'Block', we use an asterisk as a multiplication symbol. This is a short form, and R will interpret this as "fit individual effects for 'Pop' and for 'Block', plus their interaction term." The interaction alone, without the individual or marginal effects, would be represented by a colon instead of an asterisk.

Finally, nested factors are indicated with a slash, where the highest level of the hierarchy comes first. Here, 'Family' is nested within 'Pop'. This is short form for "fit an individual effect for 'Pop' and an interaction term for 'Pop' and 'Family'. No individual effect is fitted for the nested factor 'Family'!

So far, we have treated all factors as fixed. How would we modify the formulas if they all were random factors? Most R packages that can fit linear mixed models or GLMMs use the following notation: We use round brackets to indicate each random factor. Inside the bracket, we use a vertical bar to separate two parts. On the right of the vertical bar, we list the random factor, or nested factors. On the left of the bar, we list what should be fitted for each level. The number '1' indicates that we want to fit a different mean for each group.

What happens if we include AWS as a covariate? It is a good idea to standardize each covariate so that it has a mean of zero and a variance of one, especially if you want to include interactions.

The simplest model here is a simple regression. I'm not listing a number one for the intercept, but it is implied. Indeed, R will interpret this short form as if I had added the number one. Here's an illustration of what an intercept means. If we start the plot at the origin, where both variables are zero, then the intercept is the value where the regression line crosses the y-axis.

If I want to specify that no intercept should be fitted, I need to specify this by adding '-1'. This will force the regression line to go through the origin.

Now I can add my factors just like before. Here, the '+' means that I'm fitting parallel lines with different intercepts, and the asterisk means that I'm fitting a different regression slope for each Block.

If we treat 'Block' as a random factor, it gets a little tricky. To fit separate slopes, I need to repeat my covariate 'AWS' on the left side of the bar in the random term for 'Block'.

With the help of this slide, you can now tell R what model to fit to your own data. We'll practice this a bit in the tutorial. For completeness, I want to add just one more thing. You have probably noted that the arithmetic operators, like the plus symbol and the multiplication symbol, have a different meaning inside of R formulas. So, what if you want to actually take the sum, or the product, of the values in two vectors? To do so, you'll need to tell R to interpret the symbols as arithmetic operators. Simply write the expression inside a function 'I()', which stands for 'AsIs'. The same applies if you want to add a quadratic term, like here.

Scene 1: Quantitative Genetics

This brings us to the end of the lab introduction video. A second video takes a look under the hood of linear models, which may help you fully understand the worked example.

VIDEO 2

Scene 1: Quantitative Genetics

This video is the second part of the lab introduction. Here we'll look at some of the statistical issues of fitting and interpreting linear mixed models.

Scene 6: Under the Hood

With the correct formula, we can ask R to fit the model and perform significance tests for model parameters. For your standard ANOVA, we would use an F-test for the overall model and for each factor or interaction. What does this do?

The F-statistic in an ANOVA compares the variability between groups to the variability within groups. In this first example, it does not matter whether I sample two seedlings from the same group or from different groups, on average, they would show equal trait variation. If like here the variability between groups is the same as the variability within groups, the F-statistic is 1.

In the second example, seedlings in the same group have much more similar trait values than seedlings from different groups. Because the variability between groups is larger than the variability within groups, the F-statistic will be quite a bit larger than 1. If all assumptions are met, we can use an F-test to test for significant differences among groups. For this, we compare the observed F statistic to the known F distribution with two parameters, the model degrees of freedom and the residual degrees of freedom.

What are the assumptions? We want to see similar variability in all groups, and a balanced design with equal group size. The residuals should follow a normal distribution, and the observations should be independent. What happens if that last point is not the case, and how would we know?
First, the total degrees of freedom are calculated as sample size n minus one. The minus one reflects the fact that we used up one degree of freedom for estimating a global mean. The model degrees of freedom are the number of parameters fitted, minus one. This is because we already subtracted one for the global mean when calculating the total degrees of freedom, and we only need k - 1 additional parameter estimates.

The residual degrees of freedom then are simply the difference, n minus k. We know the number of parameters in the model, and we know sample size, so that should be easy. But wait: n here is not just sample size, it is the number of independent observations.

Think of it this way: if I sample two trees at different sites, each tree will contribute a fully new set of information. On the other hand, if I sample the same tree twice, I get a second measurement, but I get very little new information. That second measurement may be a little different due to measurement error, but it contributes only a fraction of the amount of new information from a completely independent new sample, and thus only a fraction of an additional degree of freedom. If we knew what fraction, we could actually calculate our degrees of freedom based on these fractions.

What are other ways that can make observations more similar and thus reduce their information content and the degrees of freedom?

- This can be due to the study design, if we used a hierarchical sample, blocking, paired samples or repeated measures.
- We can have autocorrelation, where nearby observations in space or in time are more similar than distant ones. Similarly, if each data point represents a species, species that are closely related tend to be more similar than phylogenetically distant ones.
- Another source is co-ancestry, in the form of kinship among individuals, or population history or phylo-geographic patterns.

It is clear from this list that non-independence is common in landscape genetic data, we can't ignore it but we need solutions how to deal with it. If the assumptions of F-tests are not met, we can go down this list of alternatives that make less and less stringent assumptions. Unfortunately, this comes at a cost of increasing complexity of the methods, and I won't explain them here. For instance, if we know the fractional degrees of freedom we can account for them in conditional F-tests. The most complex ones are Markov chain Monte Carlo and bootstrap methods.

Scene 7: Interpreting LMM results

So, linear mixed models allow us to account for non-independence in the data, but this comes at the cost of more complexity. That can make model interpretation quite confusing. Here, we'll look at the same four steps that we covered for regression models in Week 4: is the model valid, are the effects statistically significant, and what is the size and direction of the effects? As a simple rule, where we had one thing to consider in regression, we now have two.

With regression, we used a normal probability plot, or qqplot, to check that the residuals more or less follow a normal distribution. Here, the normal assumption also applies to the random effects. That can be random intercepts or random slopes, and we may have multiple random effects. We can create a normal probability plot for each of them.

Now back to the residuals. In addition to the normal probability plot, we'll want a plot of residuals against fitted values to check that the variance is constant, and a plot of residuals against any covariate to check for non-linear relationships, and Cook's distance to check for influential observations. For regression, we got the plots automatically, but here we'll have to create some of them ourselves. One tricky bit is that there are two types of residuals with linear mixed models. The conditional residuals account for both the fixed and the random effects, and the marginal residuals account only for the fixed effects.

If there are no major problems, we can move on to statistical significance testing. Here it matters what algorithm we use to fit the model, and that depends on the purpose, whether we want to compare different models, or interpret the fitted model. When we fit a model with 'lm', this does not matter because we use the least squares algorithms anyways. For linear mixed models, there is no algorithm that is valid for both goals. We use maximum likelihood, 'ML', to compare models. We use this to get a valid AIC and to test hypotheses about fixed effects. We use restricted maximum likelihood, 'REML', to calculate an R squared, assess variance components or to test random effects.

So, to test fixed effects, we use the model fitted with maximum likelihood, which means that we set the option 'REML' to 'FALSE'. The simplest test to use is Wald's chi-square test, which we can get with the function 'Anova', with capital A, from the package 'car'. Here we have only one factor, block, and we see that the p-value is smaller than 5%, so there is a significant block effect. More specifically, the null hypothesis is that block does not explain any variance in the trait, and the alternative is that block does explain some variance in the trait, more than one might expect just by chance.

For testing a random effect, we would use the model fitted with 'REML', and perform a likelihood ratio test. This test compares two models, one is the full model and the other is the same model but without the term that we want to test. For example, we might drop the random effect for population to test the population effect. The likelihood ratio test compares

the two models and tests whether adding population, given all other terms in the model, significantly increases the variance explained.

However, for the random effects, we are usually more interested in how much variance they explain, rather than just testing whether they have an effect. We start with getting a measure for R-squared, though what we get here is not exactly the same as the R-squared from regression analysis. Again, this comes in two parts. The marginal R-squared is the variance explained by the fixed effects, and the conditional R-squared is the total variance explained. Hence the difference between the two is what is explained by the random effects. We can go further and assess the variance component explained by each random factor. Indeed, that's what our quantitative genetics example is all about: how much of the variation in the seedling trait is explained by population, and how much by family? For this, we use the function VarCorr.

If we had several fixed effects, we would probably want to know which one is the most important. Similar to multiple regression, we can look at standardize beta coefficients, and there is a function to extract them.

Finally, we will want to know which populations, families or blocks had high seedling trait values and which ones had low values. In a linear model, all effects are treated as additive, which means that we predict a seedling trait value as the sum of a global intercept, its block effect, population effect, and family effect. Each effect can be positive or negative, here for example, the block effect is negative. The function 'fixef' extracts the fixed effects and returns a list with the additive effect of each factor level. The first level is treated as a reference level, and its value becomes the global mean. This means that the seedlings in block 1 get this value, plus their population and family effects, and the seedlings for example in block 5 get the intercept plus the value for block 5.

The function 'ranef' extracts the random effects. These are all relative to the global mean. Here, seedlings from population 'black cyn' showed an increased trait value by 0.56 on average.

Scene 8: Linear Models in R

Now that we know what to look for, let's check the model summary produced by R. It does not show all the things we discussed, that's why I listed additional functions. The first line tells us how the model was fitted, with maximum likelihood or REML, and now we know why we should care. Then R lists the model call, so we can check how the model was specified.

Then we get the variance components, both as variances and as standard deviations. These are absolute values, so they don't add up to 1, but we can already compare the population and family effects to the unexplained residual variance. The next line summarizes the sample size at each level, and we can use this to double check that the model has been specified correctly.

Then we get a table with the estimates of the fixed effects. The first value is the global intercept, which is also the mean for block 1. We should ignore the t-statistics reported here for

two reasons: the model has been fitted with REML, not maximum likelihood, and we should use a different test anyways.

Scene 1: Quantitative Genetics

Well, we have covered quite a bit of stuff in this two-part video. The interactive tutorial will help you get practice with R model formulas, and the worked example shows how to apply the methods and interpret the results. Don't panic if you don't understand every detail, just try to link back to the concepts we've covered here. A few additional topics relevant to fitting linear models are covered in other weeks, such as residual analysis, spatial linear models, and model selection.