Notes for: Walsh and Lynch. Genetics and Analysis of Quantitative Traits

Thomas Battram

2020-04-25

Contents

Preface

This is a good book, but if I make it through the whole thing I deserve several medals and some cake.

Chapter 1

An overview of quantitative genetics

BORWANG!!!

This chapter just introduces the book and some simple concepts.

Chapter 2

Properties of distributions

ALSO BORWANG!

You can guess what this chapter was on and also how much of a hoot it was...

Chapter 3

Covariance, regression, and correlation

3.1 Covariance

Covariance is a measure of association and the covariance between x and y would be denoted by $\sigma(x, y)$. If x and y are independent then $\sigma(x, y) = 0$, BUT if $\sigma(x, y) = 0$, x and y aren't necessarily independent.

3.1.1 Useful identities for covariance

Covariance of x with itself = variance of x:

$$\sigma(x, x) = \sigma^2(x) \tag{3.1}$$

For constants (here represented by a) see (3.2) below

$$\begin{aligned} \sigma(a, x) &= 0 \\ \sigma(ax, y) &= a\sigma(x, y) \\ \sigma^2(a, x) &= a^2\sigma^2(x) \\ \sigma[(a + x), y] &= \sigma(x, y) \end{aligned} \tag{3.2}$$

The covariance of 2 sums can be written as the sum of covariances, i.e. just multiply out the brackets (I've left this blank, do it yourself or check book):

$$\sigma[(x + y), (w + z)] = ...$$

Variance of a sum is sum of variances and covariances (figure this out):

$$\sigma^2(x + y) = ...$$

3.2 Least squares linear regression

Linear model:

$$y = \alpha + \beta x + e$$

Continuing on, α and β will be the true population values and a and b will be the intercept and slope for the line of best fit derived from observed data. The derivation of a and b using the least-squares model can be found on pages 39-41. Buuut, who cares about that, here are the results:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{Cov(x, y)}{Var(x)}$$

3.2.1 Properties of least squares

6 in the book, just writing down important/not obvious ones.

- The mean residual ($\bar{e}$) is 0
- Residual errors are uncorrelated with predictor variable x (see book for why)
  - BUT e and x may not be independent if the relationship between x and y is non-linear. If it is truly non-linear $E(e|x)! = 0$
- Variance of e can vary with x, in this situation the the regression is said to display heteroscedasticity (see Figure 3.4 for great illustration)
- The regression of y on x is different to the regression of x on y!

3.3 Correlation

Correlation coefficient between x and y:

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$

The correlation coefficient is a dimensionless measure of association and it is symmetrical (i.e. $r(x, y) = r(y, x)$).

Scaling x or y by constants does not change the correlation coefficient, but it does affect variances and covariances.

The correlation coefficient is a standardised regression coefficient -> the regression coefficient resulting from rescaling x and y such that each has unit variance).

r^2 assumes $E(y|x)$ is linear!

3.4 Differential selection (brief intro)

The directional selection differential, S, is the difference between the mean phenotype within that generation before selection (μ_s) and the mean phenotype within that generation after (μ) selection.

$$S = \mu_s - \mu$$

If all individuals have equal fertility and viability then selecting individuals won't change anything so $\mu_s = \mu$ and $S = 0$.

If $W(z)$ is the probability that individuals with phenotype z survive to reproduce and $p(z)$ is the density of z (pretty much means distribution) before selection, then the density after selection is:

$$p_s(z) = \frac{W(z)p(z)}{\int W(z)p(z)dz}$$

The denominator here is the mean individual fitness ($\bar{W}$). The relative fitness of z is $w(z) = \frac{W(z)}{\bar{W}}$.

After some sweet derivation (see page 46), you finish with:

$$S = \sigma[z, w(z)]$$

Therefore the directional selection is equivalent to the covariance of the phenotype and the relative fitness.

If you regress offspring phenotype on the midparent phenotype and that relationship is linear with slope β, a change in mean midparent phenotype induces an expected change in mean phenotype across generations equal to:

$$\Delta\mu = \mu_0 - \mu = \beta(\mu_s - \mu) = \beta S$$

This is the breeders' equation!

3.5 Correlation between genotype and phenotype (brief intro)

Only when there is no gene-environment interaction is the variance explained by genetics (broad-sense heritability) the equation below:

$$H^2 = \frac{\sigma_G^2}{\sigma_z^2}$$

,

where z is the phenotype and G is the sum of the total effects (not just additive) at all loci on the trait.

The slope of a midparent-offspring regression provides an estimate of the proportion of the phenotypic variance that is attributable to additive genetic factors (the narrow-sense heritability).

$$h^2 = \frac{\sigma_A^2}{\sigma_z^2}$$

So as h^2 is just the regression of offspring phenotype on midparent phenotype it can actually be used in the breeders' equation!

$$\Delta\mu = h^2 S$$

So the narrow-sense heritability can be thought of as the efficiency of the response to selection. If $h^2 = 0$ there can be no evolutionary change regardless of strength of selection. Although this should be obvious because if h^2 is 0 then there is clearly no passing of genetic material onto the next generation that is influencing that trait.

Chapter 4

Properties of single loci

4.1 Introduction

too easy

4.2 Allele and genotype frequencies

too easy

4.3 The transmission of genetic information

4.3.1 The Hardy-Weinberg principle

$$p^2 + 2pq + q^2 = 1$$

where p = allele frequency of first allele at a locus and q = allele frequency of the second allele at that same locus.

Assumptions of H-W:

- No selection
- No mutation
- Random mating
- No differential migration
- No random drift

Even though these assumptions will never be met completely in the real world, for the majority of the time the H-W prinicple holds regardless.

Assuming assumptions are met, 2 important points from H-W:

1. It takes no more than a single generation equilibriate and stabilize gene frequencies in the two sexes.
2. Only one additional generation is required for the stabilisation of the genotype frequencies into the predictible Hardy-Weinberg proportions.

4.3.2 Sex-linked loci

Alleles on sex chromosomes in diploid organisms are obviously different. Sons can only receive an X chromosome from their mother so the frequency of X linked loci in the sons is equal to that of their mothers. Daughters receive both an X chromosome from Mum + from Dad.

Overall this means allele frequencies oscilate around an equilibrium state, but continually get closer to that state over the generations (see Figure 4.2 and page 56 for equation).

4.3.3 Polyploidy

Skipped over this section because it's not relevant to human quant gen. Buuut, essentially it just details how to derive allele frequencies under a certain case of polyploidy. Also, it should be noted that of course H-W does not hold under polyploidy!

4.3.4 Age structure

Age structure also complicates our idealised model of H-W. In populations composed of several age classes, the generations overlap, and this causes the approach of genotype frequencies towards the H-W expectations to be gradual (rather than just by 1/2 generations), even in the case of an autosomal locus. Doesn't explain this very much, but it's covered elsewhere. Importantly, when newly founded populations have significant age structure, fluctuations in both gene and genotoype frequencies may occur for a substantial period of time even in the abscence of selection!

4.3.5 Testing for Hardy-Weinberg proportions

Says in the book that LRT can be used to test for departures from HWE, buuuut I'm pretty sure that even now the most common method is the chi-squared test (or Fisher's exact test if the sample size is tiny and the allele is rare.). Essentially, in a population, at a specific locus, you can calculate the allele frequencies (and from that expected genotype frequencies) from the observed genotype frequencies then test if there is a difference between the observed and expected values. LRT equation for it given on page 60. **RECREATE THE CHI-SQUARED TEST IN CODE!!!!**

Should remember (as pointed out above), that just because some assumptions are violated, doesn't mean you'd get a departure from HWE!

4.4 Characterising the influence of a locus on the phenotype

If a trait is entirely influenced by a single locus then the genetic effect on that trait can be characterised pretty easily and the dominance and additive effects of the alleles can be calculated. So if a locus has genotypes B_1B_1, B_1B_2, B_2B_2, then the values given to these genotypes can be said to be: $-a$, $(1+k)a$ and $+a$. Now if you have genotype data at that locus and data on the trait you can work out the effect of the B_2 allele by taking the mean phenotypic value of individuals with B_2B_2 and subtracting the mean phenotypic value of individuals with B_1B_1 and dividing by 2 i.e.

$$B_{2eff} = \frac{p_{B2} - p_{B1}}{2}$$

where B_{2eff} is the effect of allele B_2, p_{B2} is the mean phenotypic value of individuals with B_2B_2 and p_{B1} is the mean phneotypic value of individuals with B_1B_1.

As $B_{2eff} = a$ you can then substitute this in to $(1+k)a$ to get the dominance coefficient k. Of course if $k = 0$ then there is no dominance (in reality you would calculate probability of dominance).

4.5 The basis of dominance

Confusing part… Don't really get the enzyme activity bit…

Main point (I think) is that new deleterious mutations are very likely to be recessive and new mutations with a slight deleterious effect interact in an almost entirely additive fashion (no dominance!).

4.6 Fisher's decomposition of the genotypic value

Recalling that the phenotypic value can be partitioned like so:

$$z = G + E$$

where z is the phenotype, G is the genotypic value and E is the environmental value.

The genotypic value of a specific locus can be partitioned into it's "expected" values based on there being only additive effects ($\hat{G}$) and the deviations from the expected values or dominance effects (δ). So for genotype B_iB_J:

$$G_ij = \hat{G_ij} + \delta_ij$$

This can be formalised (whatever the fuck that means) by regressing the genotypic values on the number of B_1 and B_2 alleles in the genotype (N_1 and N_2):

$$G_ij = \hat{G_ij} + \delta_ij = \mu_G + \alpha_1 N_1 + \alpha_2 N_2$$

$\mu_G =$ the mean genotypic value in the population, α_1 and α_2 are the slopes of the regression, N_1 and N_2 are the number of B_1 and B_2 alleles. So the regression is:

$$G_ij \; N_2 + N_1$$

By noting that for any individual, $N_1 = 2 - N_2$ you can reduce the multiple regression model into an easier to work with univariate model. Give it a go:

$$G_ij = ...$$

If you plotted the genotypic value (G) against gene content (N_2 or number of B_2 alleles) and calculated residuals these residuals would be δ (see Figure 4.6).

The rest of the chapter uses this regression and what we know about genotype frequencies to derive a formula for the average effect of allelic substitution:

$$\alpha = a[1 + k(p_1 - p_2)]$$

where $a =$ genotypic value of B_2 (see above), k is the dominance coefficient and p_1 and p_2 are the frequencies of B_1 and B_2. This value α represents the average change in genotypic value that results when a B_2 allele is randomly substituted for a B_1 allele. If no dominance ($k = 0$) then $\alpha = a$. Except in the case of additivity, the average effect of allelic substitution is not simply a function of the inherent physiological properties of the allele. It can only be defined in the context of the population!

4.7 Partioning the genetic variance.

Deriving variance of G:

$$G = \hat{G} + \delta$$
$$\sigma_G^2 = \sigma^2(\hat{G} + \delta)$$
$$\sigma_G^2 = \sigma^2(\hat{G}) + \sigma(\hat{G} + \delta) + \sigma^2(\delta)$$

The top equation is just like a regression, with δ being the residual error and we know that for least-squares there is no correlation between the residual error and the predictor. So there is no correlation between $\hat{G}$ and δ. Therefore:

$$\sigma_G^2 = \sigma^2(\hat{G}) + \sigma^2(\delta)$$

OR more commonly

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

σ_A^2 is the variance of G explained by regression on N_2 (or N_1), and σ_D^2 is the residual variance of that regression. The variance of the additive and dominance effects!

For a diallelic locus we can do some rearranging of equations in Table 4.1 of the book and get these equations:

$$\sigma_A^2 = 2p_1p_2\alpha^2$$

$$\sigma_D^2 = (2p_1p_2ak)^2$$

From these we can clearly see that both components depend on allele frequencies, the dominance coefficient and the homozygous effect (remember α is just the slope of the G N_2 regression!).

By plotting how genetic variance changes with gene frequency under different scenarios (**DEFINITELY DO THIS**). You see some interesting patterns. Firstly, at a single diallelic locus, you see that σ_A^2 reaches it's peak when $p_1 = p_2 = 0.5$. Secondly, it's clear that, even in the case of overdominance (which is rare!), additive genetic variance will almost always be much higher than genetic variance from dominance effects, even when the frequency of the dominant allele is high. (btw I think the scale on the top left graph is wrong. . .).

4.8 Additive effects, average excesses and breeding values

The dominance deviation of a parent, which is a function of the interaction between the two parental alleles, is eliminated when gametes are produced. Thus, one can think of $\hat{G}$ and δ as the heritable and nonheritable components of an individual's genotypic value.

Fisher proposed two different measures of the effect of an allele: one being the additive effect (α_i) and then the average excess α_i^x. The average excess α_2^x of allele B_2 is the difference between the mean genotypic value of individuals carrying at least one copy of B_2 and the mean genotypic value of a random individual from the entire population:

$$\alpha_2^x = (G_{12}P_{12|2} + G_{22}P_{22|2}) - \mu_G$$

where P_{ij} is the conditional probability of a B_iB_j genotype given that one allele is B_i. Under random mating $P_{ij|i} = p_j$ (p_j = frequency of allele B_j). THINK ABOUT HARDY-WEINBERG AND IT MAKES SENSE!

So under random mating,

$$\alpha_2^x = G_{12}p_1 + G_{22}p_2 - \mu_G$$

$G_{12} = a(1+k)$ and $G_{22} = 2a$. By substituting these into the equation above for α_1^x and α_2^x and then calculating α_1 and α_2 (additive effects) by the method previously mentioned (regressing genotypic value G on the number of B_2 alleles N_2), we will see they're equivalent (shown on page 72):

$$\alpha_2 = p_1\alpha$$

$$\alpha_1 = -p_2\alpha$$

The breeding value of an individual (A) is the sum of the additive effects of its genes. So the breeding value of a B_1B_1 homozygote is just $2\alpha_1$. In randomly mating populations the breeding value of a genotype is equivalent to twice the expected deviation of its offspring mean phenotype from the population mean. Soooo, no genotype information is needed to calculate the breeding value. All we have to do is mate an individual to many randomly chosen individuals from the population and taking twice the deviation of its offspring mean from the population mean. EASY IN HUMANS!!!

In Chapter 13 this will be discussed wrt candidate gene studies.

4.9 Extensions for multiple alleles and non random mating

So this section seems mostly unrelevant as we're unlikely to deal with situations with more than 2 alleles. Non-random mating could be encountered if we're interested in some phenotypes (e.g. alcohol intake). Buuuut, it's still good to note some of the generalised equations for what we've been discussing so far in the chapter.

4.9.1 Average excess

When n alleles are present, the average excess, α_i^x, for any allele B_i is given by

$$\alpha_i^x = \sum_{j=1}^{n} P_{ij|i} G_{ij} - \mu_G$$

Remember, under random mating $P_{ij|i} == p_j$

4.9.2 Additive effects

The genotypic value can also be obtained using regression as before, but in it's generalised form is a multivariate regression. For n alleles

$$G = \mu_G + \sum_{i=1}^{n} \alpha_i N_i + \delta$$

After some re-arranging can derive the regression coefficients and finally end with

$$\alpha_i = \sum_{j=1}^{n} p_j G_{ij} - \mu_G$$

i.e. under random mating, the average effects (α_i) are equal to the conditional mean deviations from the mean genotypic value of the population (μ_G).

For non-random mating we need the inbreeding coefficient, f to define our genotype frequencies:

$$P_{ii} = (1-f)p_i^2 + f p_i$$

$$P_{ij} = 2(1-f)p_i p_j$$

Unsure of why, but this means

$$\alpha_i = \frac{\alpha_i^x}{1+f}$$

so f is the fractional reduction of heterozygote frequencies relative to those expected under random mating. This means you can kind of do a test for random mating by checking heterozygote and homozygote frequencies in a population!

4.9.3 Additive genetic variance

The additive genetic variance across n alleles is

$$\sigma_A^2 = 2\sum_{i=1}^{n} p_i \alpha_i \alpha_i^x$$

In general inbreeding inflates the additive genetic variance by causing correlations among the effects of alleles within the same individuals.

The broad sense heritability, even under scenarios of non-random mating can be given by

$$\sigma_G^2 = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij})$$

although it should be noted that the definitions of α_i and $\delta_i j$ change with the degree of inbreeding! Random mating means α_i and α_j are uncorrelated so we get back to the good old equation

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

Importantly, under random mating, σ_A^2 is equivalent to the variance of breeding values of individuals in the population.

Summarising some key terms

The homozygous effect, a, and the dominance coefficient, k, are intrinsic properties of allelic products. They are not functions of allele frequencies, but may vary with genetic background

The additive effect, α_i, and the average excess, α_i^x, are properties of alleles in a particular population. They are functions of a, k and genotype frequencies (p_i).

The breeding value, A, is a property of a particular individual in reference to a particular population. It's equivalent to the sum of the additive effects of an individual's alleles.

The additive genetic variance, σ_A^2 is a property of a particular population. It is equivalent to the variance of the breeding values of individuals within the population.

Chapter 5

Sources of genetic variation for multilocus traits

5.1 Epistasis

Epistasis describes the nonadditivity of effects between loci, i.e. the alleles of one loci influence the effects of another loci.

The genotypic value, G_{ijkl}, needs to take into account all the interaction terms that can arrive between loci, for two loci it's additive x additive effects ($\alpha\alpha$), additive x dominance effects ($\alpha\delta$), and dominance x dominance effects ($\delta\delta$). As the number of loci increases the number of interaction terms increase steadily e.g. $\alpha\alpha\alpha$ will be there for three loci.

5.2 A general least-squares model for genetic effects

This is just an extension of the one-locus linear model introduced in Chapter 4.

For this section, imagine we are interested in measuring the genetic effects of two loci, G_{ijkl}, which can easily be extended to more. The additive effect of an allele on a phenotype is just the phenotypic value in people with that allele minus the mean phenotypic value of the population. When considering epistatic effects we can define it in the same way.

$$\alpha_i = G_{i...} - \mu_G \tag{5.1}$$

$G_{i...}$ is just the conditional mean phenotype of individuals with allele i at the first locus without regard to the other allele at that locus or to the genotype at the second locus. The other additive terms (for α_j, α_k, α_l) are defined in the same way. Within each locus, the mean value of average effects (weighted by allele frequency) = 0.

Dominance effects can be defined in a similar way, complete these equations by recalling the equation for G_{ij} at the end of Chapter 4:

$$\delta_{ij} = G_{ij} - ... \tag{5.2}$$

$$\delta_{lk} = G_{lk} - ... \tag{5.3}$$

Like with the additive effects, the mean dominance deviation at each locus is equal to zero.

Epistatic effect terms proceed in a similar fashion. Letting $G_{i.k.}$ be the mean phenotype of individuals with gene i at locus 1 and k at locus 2, without regard to the other two genes, the ikth additive x additive effect is:

$$(\alpha\alpha)_{ik} = G_{i.k.} - \mu_G - \alpha i - \alpha k \tag{5.4}$$

So $(\alpha\alpha)_{ik}$ is the deviation of the conditional mean $G_{i.k.}$ from the expectation based on the population mean μ_G and the additive effects α_i and α_k. An additive x dominance effect measures the interaction between an allele at one locus with a genotype of another locus (see equation 5.5 in book) and the dominance x dominance effect involves an interaction between the genotypes at each locus (see equation 5.6 in book).

The complete genotypic value, $G_{ijkl...}$ can be found in equation 5.7 in the book. These parameters depend on genotype frequencies in the population, but the mean value of each type of effect is always equal to zero.

The genotypic value of an individual is often impossible to quantify because of variation in the phenotype due to the environment, but the genotypic value for an individual equation can be extended to populations. Providing mating is random and segregation of loci is independent, there is no statistical relationship between the genes found within or among loci. So the total genetic variance is just the sum of the variance of the individual effects, simplified this is:

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD} + ... \tag{5.5}$$

... here and in other cases just symbolises more terms can be added if more than two loci are used.

Epistatic effects are expected to be common throughout the genome and Wright thought they were the rule, rather than exception. See example two in the book for calculations of epistatic effects and how much variance they contribute to the overall genetic variance component. Overall, it is clear that even with large epistatic effects, additive genetic variance, σ_A^2 will pretty much always (if not always) contribute to the vast majority of overall genetic variance σ_G^2. This is important for two reasons:

1. Variance components provide limited insight into the physiological mode of gene action, i.e. just because genetic variance is explained by additive effects (which means you essentially count each gene separately), it does not mean the interaction between genes is not important in terms of their function!
2. When interested in the variance of a trait that is explained by genetics, you can expect the vast majority of that variance to be explained by additive genetic effects, which makes things like estimating heritability far easier.

5.2.1 Extension to haploids and polyploids

Skipped this section as not relevant to humans.