Chapter 3: Covariance, regression, and correlation

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.

Useful identities for cov

Covariance of x with itself = variance of x:

$$\sigma(x, x) = \sigma^2(x)$$

For constants (here represented by a):

$$\sigma(a, x) = 0$$
$$\sigma(ax, y) = a\sigma(x, y)$$
$$\sigma^2(a, x) = a^2\sigma^2(x)$$
$$\sigma[(a + x), y] = \sigma(x, y)$$

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):

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

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:

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

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!

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!

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!

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.