Genet. Res., Camb. (1976), **28**, *pp.* 199–203
Printed in Great Britain

SHORT PAPER

Regressions between relatives

By M. G. BULMER
Department of Biomathematics, Pusey Street, Oxford

(*Received* 26 *January* 1976)

SUMMARY

A metric character determined by a large number of loci without epistasis is normally distributed. In the absence of linkage the joint distribution in two or more relatives is multivariate normal, so that all regressions are linear and have constant residual variance. In the presence of linkage this is no longer true except in the case of parent and child; for all other types of relatives the regression line is unaffected by linkage but the residual variance about this line is no longer constant but increases away from the mean.

The theory of correlations between relatives is well established, but it is sometimes necessary to possess more information than is given by the correlation coefficients; in particular one might want to know whether the regression of an individual on one or more relatives is linear and whether the variance about the regression is constant.

Consider a character determined by n loci without epistasis and assume random mating without selection or mutation in an effectively infinite population. To find the joint distribution of the character in a number of related individuals, write g_{ij} for the genetic contribution from the jth locus in the ith individual, $\mathbf{g}_j$ for the vector of these contributions at the jth locus in the different individuals, $G_i = \sum_j g_{ij}$ for the genotypic value of the ith individual and $\mathbf{G}$ for the vector of genotypic values.

In the absence of linkage loci assort independently of one another so that the vectors $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ are statistically independent. It follows from the multivariate form of the central limit theorem that, under rather general conditions, their sum $\mathbf{G}$ will have a multivariate normal distribution as $n \to \infty$. In particular all regressions will in the limit be linear with constant variance about them, whether or not there is dominance.

In the presence of linkage the vectors $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ will not in general be independent of each other. The reason is that knowledge of $\mathbf{g}_1$, for example, provides some information about the number of genes which are identical by descent in the different individuals at the first locus, which provides information about $\mathbf{g}_2$ if the first and second loci are linked. However, if the individuals are related as parent and child the number of identical loci is known to be exactly 1 at each locus, so that no further information about it can be obtained. Thus the joint distribution of parent and child, or of mother, father and child, will become multivariate normal in the limit even in the presence of linkage. This will not be true if the individuals are related in any other way (Bulmer, 1971).

To investigate the effect of linkage in more detail consider a pair of related individuals. The marginal distributions of the two genotypic values, G_1 and G_2, will in the limit be

normal, with mean μ and variance σ_G^2, say. The conditional mean and variance of G_2 given G_1 can be expressed as

$$\left.\begin{aligned} E(G_2|G_1) &= \sum_{j=1}^{n} E(g_{2j}|G_1), \\ \operatorname{var}(G_2|G_1) &= \sum_{j=1}^{n} \operatorname{var}(g_{2j}|G_1) + 2\sum_{j<k} \operatorname{cov}(g_{2j}, g_{2k}|G_1). \end{aligned}\right\} \quad (1)$$

The distribution of g_{2j} only depends on G_1 through the information contained in G_1 about g_{1j}; furthermore the distribution of g_{1j} given G_1 is unaffected by linkage provided that the population is in linkage equilibrium. Hence the distribution of g_{2j} given G_1, and in particular its mean and variance, are unaffected by linkage. On the other hand the joint distribution of g_{2j} and g_{2k}, and in particular their covariance, conditional on G_1, will be affected by linkage for the reasons discussed above. It can be concluded that the regression of G_2 on G_1 is unaffected by linkage and will thus become linear in the limit, but that the variance about this regression line will be affected by linkage and will only become constant in the limit in the absence of linkage or in the case of parent and child.

To evaluate the effect of linkage on the variance we first consider an extreme situation with complete linkage between all pairs of loci but with linkage equilibrium; no assumptions are made about the genetic model except that there is a large number of loci without epistasis. Denote by P_s ($s = 0, 1, 2$) the probability that the two related individuals have s identical genes at a particular locus. With complete linkage the number of identical genes must be the same at all loci, so that the conditional distribution of G_2 given G_1 is a mixture of three distributions: (i) with probability P_0 there are no genes identical by descent at any locus and G_2 will be normal with mean μ and variance σ_G^2; (ii) with probability P_1 there is one identical gene at every locus, so that the distribution of G_2 given G_1 will be the same as that of child given parent and will be normal with mean $\mu + \frac{1}{2}h^2(G_1 - \mu)$ and with variance $(1 - \frac{1}{4}h^4)\sigma_G^2$, where h^2 is the ratio of the additive to the total genetic variance; (iii) with probability P_2 there are two identical genes at every locus so that $G_2 = G_1$. The variance about the regression line is easily found to be

$$\operatorname{var}(G_2|G_1) = [1 - (\tfrac{1}{2}h^2P_1 + P_2)^2]\sigma_G^2 + [\tfrac{1}{4}h^4P_1 + P_2 - (\tfrac{1}{2}h^2P_1 + P_2)^2][(G_1 - \mu)^2 - \sigma_G^2]. \quad (2)$$

If there is environmental as well as genetic variance, then

$$\operatorname{var}(Y_2|Y_1) = [1 - (\tfrac{1}{2}h^2P_1 + h_w^2P_2)^2]\sigma_p^2 + [\tfrac{1}{4}h^4P_1 + h_w^4P_2 - (\tfrac{1}{2}h^2P_1 + h_w^2P_2)^2][(Y_1 - \mu)^2 - \sigma_p^2]. \quad (3)$$

In this equation Y_1 and Y_2 are the phenotypic values in the two relatives, σ_p^2 is the phenotypic variance, h^2 is the (narrow) heritability (ratio of additive genetic to phenotypic variance) and h_w^2 is the wide heritability (ratio of total genetic to phenotypic variance).

In the absence of linkage the variance about the regression line is given by the first term on the right hand side of (2) or (3). The effect of the second term is to increase the variance about the regression line as the first individual's genotypic (or phenotypic) value departs from the mean. For parent and child, $P_0 = P_2 = 0$, $P_1 = 1$; thus the second term vanishes and the variance is constant, as already shown. By contrast, in sibs $P_0 = P_2 = \frac{1}{4}$, $P_1 = \frac{1}{2}$; putting $h^2 = 1$ in (2) and $h^2 = h_w^2 = \frac{1}{2}$ in (3) to represent a situation with equal amounts of additive genetic and environmental variance, we find that

$$\left.\begin{aligned} \operatorname{var}(G_2|G_1) &= \tfrac{5}{8}\sigma_G^2 + \tfrac{1}{8}(G_1 - \mu)^2, \\ \operatorname{var}(Y_2|Y_1) &= \tfrac{29}{32}\sigma_p^2 + \tfrac{1}{32}(Y_1 - \mu)^2. \end{aligned}\right\} \quad (4)$$

There is thus a substantial amount of heteroscedasticity, particularly in the genotypic regression.

In the more realistic case with partial linkage, a solution has only been found under a simplified genetic model in which there are two alleles, + and −, at each locus which

contribute 1 and 0 respectively to the character and in which the + allele has the same frequency, p, at each of the n loci. Under this model it is shown in the Appendix that

$$\text{var}\,(G_2|G_1) = (1-\overline{\pi}_2)\,\sigma_G^2 + (\overline{\pi}_2 - R^2)\,(G_1-\mu)^2 + \text{terms of smaller order of magnitude in } n. \tag{5}$$

In this equation $R \equiv \frac{1}{2}P_1 + P_2$ is the coefficient of relationship between the individuals. π_2 is the probability that if one gene is chosen at random at each of two loci in one individual, then the other individual will possess genes identical to both of them; π_2 will depend on the recombination fraction between the loci, and $\overline{\pi}_2$ is its average value over all pairs of loci. In the absence of linkage $\overline{\pi}_2 = R^2$ so that the variance is constant, but in the presence of linkage $\overline{\pi}_2 > R^2$ for any relatives except parent and child so that the variance about the regression line increases as G_1 moves away from its mean value. Equation (5) has been obtained by assuming genes with equal effects and frequencies, and with only two alleles per locus, but it seems likely that it would remain valid under a more general model without dominance provided that an appropriate averaging process for calculating $\overline{\pi}_2$ was used. It should be noted that if there is complete linkage at all loci, then $\overline{\pi}_2 = \frac{1}{2}P_1 + P_2$ so that (5) is equivalent to (2) in the absence of dominance. If there is environmental as well as additive genetic variance, equation (5) is replaced by

$$\text{var}\,(Y_2|\,Y_1) = (1-\overline{\pi}_2 h^4)\,\sigma_p^2 + (\overline{\pi}_2 - R^2)\,h^4(G_1-\mu)^2. \tag{6}$$

To evaluate the magnitude of the effect shown in (5) and (6) it is necessary to estimate a typical value for $(\overline{\pi}_2 - R^2)$. For sibs it can be shown that $(\pi_2 - R^2) = \frac{1}{8} - \frac{1}{2}r(1-r)$ for a pair of loci with recombination fraction r. Consider a chromosome with a length of 1 morgan. If loci are distributed at random along the chromosome, then the map distance, x, between two randomly chosen loci will have the density function $f(x) = 2(1-x)\,(0 \leqslant x \leqslant 1)$. If the recombination fraction is related to map distance by the standard mapping function $r = \frac{1}{2}\,[1-\exp(-2x)]$, the Expected value of $\frac{1}{8} - \frac{1}{2}r(1-r)$ is found by a simple integration to be 0·047. In a species with k chromosomes each of unit length, $(\overline{\pi}_2 - R^2)$ for sibs is thus $0{\cdot}047/k$, and equation (5) becomes

$$\text{var}\,(G_2|G_1) = (0{\cdot}75 - 0{\cdot}047/k)\,\sigma_G^2 + \frac{0{\cdot}047}{k}\,(G_1-\mu)^2 \tag{7}$$

for sibs. The effect is likely to be negligible unless the number of chromosomes is small. The absence of crossing-over in the male in *Drosophila* can be taken into account by using the mapping function $r = \frac{1}{4}[1-\exp(-2x)]$. In this case the Expected value of $\frac{1}{8} - \frac{1}{2}r(1-r)$ is 0·079, so that $(\overline{\pi}_2 - R^2)$ for sibs is $0{\cdot}079/3 = 0{\cdot}026$ if *Drosophila* is assumed to have three chromosomes of unit length. Thus for sibs in this species

$$\text{var}\,(G_2|G_1) = 0{\cdot}724\sigma_G^2 + 0{\cdot}026(G_1-\mu)^2. \tag{8}$$

It will be seen from equation (6) that for the phenotypic regression the factor $(\overline{\pi}_2 - R^2)$ must be multiplied by the square of the heritability.

It is concluded that the increase in the residual variance away from the mean is likely to be too small to be experimentally detectable in a single generation. This factor may nevertheless have an appreciable cumulative effect, particularly in an organism such as *Drosophila*, if stabilizing or disruptive selection is continued over many generations. It should therefore be taken into account in any theoretical analysis of the effect of selection on genetic variability.

REFERENCE

BULMER, M. G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.

APPENDIX. DERIVATION OF EQUATION (5)

Let $x_{ijl}(j = 1, 2, \ldots, n; l = 1, 2)$ be the contribution from the lth gene at the jth locus in the ith individual. Each x takes the value 1 or 0 with probability p or $q\,(= 1-p)$. The joint distribution of the x's in a particular individual given their sum, G_i, is

$$P(x_{i\,11}, x_{i\,12}, x_{i\,21}, x_{i\,22}, \ldots, x_{in\,1}, x_{in\,2} | \sum_{j,l} x_{ij\,l} = G_i) = 1 \Big/ \binom{2n}{G_i}.$$

This is the distribution of G_i balls in $2n$ boxes; the balls are placed at random in boxes, subject to not more than one ball per box.

The genetic effect at the jth locus in the ith individual is $g_{ij} = x_{ij\,1} + x_{ij\,2}$. Consider the distribution of g_{2j} given the genotypic value in a related individual, G_1. This distribution only depends on G_1 through the information contained in G_1 about g_{1j}. It can be seen from the 'balls-in-boxes' model that this distribution is as follows:

Value of g_{2j}	Probability of this value given G_1
0	$q^2P_0 + q\dfrac{(2n-G_1)}{2n}P_1 + \dfrac{(2n-G_1)}{2n}\dfrac{(2n-G_1-1)}{(2n-1)}P_2$
1	$2pqP_0 + \left[q\dfrac{G_1}{2n} + p\dfrac{(2n-G_1)}{2n}\right]P_1 + \dfrac{G_1}{n}\dfrac{(2n-G_1)}{(2n-1)}P_2$
2	$p^2P_0 + p\dfrac{G_1}{2n}P_1 + \dfrac{G_1}{2n}\dfrac{(G_1-1)}{(2n-1)}P_2$

Hence

$$\mathrm{var}\,(g_{2j}|G_1) = 2pq - \frac{(G_1-\mu)}{n}(p-q)R + \frac{(G_1-\mu)^2}{n^2}(\tfrac{1}{2}P_2 - R^2) - \frac{G_1(2n-G_1)}{(2n-1)\,2n^2}P_2,$$

where $\mu = 2np$ is the genotypic mean and $R = \frac{1}{2}P_1 + P_2$ is the coefficient of relationship between the individuals.

To evaluate the covariance of g_{2j} and g_{2k} given G_1 we first define the quantity $\pi_s\,(s = 0, 1, 2)$ as the probability that if one gene is chosen at random at each of the two loci in the first individual, then the other individual will possess s genes identical to one or other of them. π_s will depend on the recombination fraction (r) between the two loci, but it should be noted that $\pi_1 + 2\pi_2 = P_1 + 2P_2\,(= 2R)$ from the additive property of Expected values. For example, for two full sibs $\pi_0 = \pi_2 = \frac{3}{8} - \frac{1}{2}r(1-r)$, $\pi_1 = \frac{1}{4} + r(1-r)$; and for grandparent–grandchild $\pi_0 = \frac{1}{2} + \frac{1}{8}(1-r)$, $\pi_1 = \frac{1}{4}(1+r), \pi_2 = \frac{1}{8}(1-r)$.

The conditional covariance between g_{2j} and g_{2k} can be found from the following facts:

$$\begin{aligned}
\mathrm{cov}\,(g_{2j}, g_{2k}|G_1) &= 4\,\mathrm{cov}\,(x_{2j1}, x_{2k1}|G_1),\\
\mathrm{cov}\,(x_{2j1}, x_{2k1}|G_1) &= E(x_{2j1}.x_{2k1}|G_1) - E^2(x_{2j1}|G_1),\\
E(x_{2j1}.x_{2k1}|G_1) &= \mathrm{prob}\,[x_{2j1} = x_{2k1} = 1|G_1],\\
&= p^2\pi_0 + p\frac{G_1}{2n}\pi_1 + \frac{G_1(G_1-1)}{2n(2n-1)}\pi_2,\\
E(x_{2j1}|G_1) &= p + \frac{R(G_1-\mu)}{2n}.
\end{aligned}$$

Hence

$$\mathrm{cov}\,(g_{2j}, g_{2k}|G_1) = \frac{(G_1-\mu)^2}{n^2}(\pi_2 - R^2) - \frac{G_1(2n-G_1)}{(2n-1)n^2}\pi_2.$$

https://doi.org/10.1017/S0016672300016876 Published online by Cambridge University Press

Finally the conditional variance of G_2 given G_1 can be found from the summation shown in equation (1). Thus

$$\text{var}\,(G_2|G_1) = 2npq-(G_1-\mu)\,(p-q)R+\frac{(G_1-\mu)^2}{n}\,[n(\overline{\pi}_2-R^2)+\tfrac{1}{2}P_2-\overline{\pi}_2]$$
$$-\frac{G_1(2n-G_1)}{n(2n-1)}\,[n\overline{\pi}_2+\tfrac{1}{2}P_2-\overline{\pi}_2],$$

where $\overline{\pi}_2$ is the average value of π_2 over all pairs of loci. Treating $G_1-\mu$ as a quantity of order $n^{\frac{1}{2}}$ and writing $\sigma_G^2 = 2npq$, we find that

$$\text{var}\,(G_2|G_1) = (1-\overline{\pi}_2)\,\sigma_G^2+(\overline{\pi}_2-R^2)\,(G_1-\mu)^2+o(n).$$

https://doi.org/10.1017/S0016672300016876 Published online by Cambridge University Press