African Journal of Biotechnology Vol. 10(41), pp. 7932-7938, 3 August, 2010
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB09.1907
ISSN 1684–5315 © 2010 Academic Journals

Review

Estimating and testing the effect of allelic recombination on the correlation between genotypic relatives

Oyeka I. C. A[1]., Oyeka C. A.[2] and Uzuke C. A.[1*]

[1]Department of Statistics Nnamdi Azikiwe University, Awka.
[2]Department of Applied Microbilogy and Brewing Nnamdi Azikiwe University Awka.

Accepted 21 January, 2011

This paper provides estimates of the correlation between genotypic relatives and the effect of allelic recombination on the correlation assuming random mating. It is shown that the correlation is a non negative quantity and that allelic recombination has the effect of reducing total variation and doubling the correlation between genotypic relatives with respect to measurements on the character of interest. The significance of the correlation coefficient as well as the fitted regression model was obtained using Analysis of Variance method.

Key words: Allele, genotype, regression, correlation, F-ratio, analysis of variance.

INTRODUCTION

Genetic recombination is an effective means of combining one individual trait of two parents, permitting the comparison of one expression of a character with another expression of the same traits (Burns, 1976, Alberts, 1994, Maloy S, 1994).

Although much work has been done on the correlation between relatives for various physical characteristics starting with the pioneering work by Fisher (1918), very little has been written on the effect of genetic recombination on these correlations (Ewens, 1979, Oyeka and Oyeka, 1988). These writers also failed to provide a statistic for testing the significance of the estimated correlation coefficient. In this paper, the work by (Oyeka and Oyeka 1988) is modified to include a test statistic for the estimated correlation coefficient and to know if the hypothesised model fits.

CORRELATION BETWEEN RELATIVES

We will assume that a certain population has a gene locus with possible alleles A and a. We also assume that the probability of occurrence of allele A in the population of interest is p and that the corresponding probability of the allele a is q = 1-p. We further assume that a certain characteristic or factor V of the population of interest is completely determined by the genotype at the locus in such a way that the individuals of genotype AA have a value or measurement of (V = x) for the character or factor of interest; all individuals of genotype Aa have a measurement of (V = y) and all individuals of genotype aa have measurement (V = z). We finally assume that our population obeys the Hardy-Weinberg law of random mating (Stein, 1943; Clavel et al., 1989).

Assuming n-pairs of relatives are studied, let R_1 and R_2 be a pair of relatives whom we know for sure have at least one allele in common. Then under the law of random mating, the occurrence of the allele A and a in the genotypes are independent. Hence the probabilities of occurrence of AA, Aa, or aA and aa are, respectively,

$pp = p^2$

$p(1-p) + (1-p)p = 2p(1-p)$

and

$(1-p)\,(1-p) = (1-p)^2$

We first derive an estimate of the correlation between the

*Corresponding author. E-mail: chinwe_uzuke@yahoo.com.

measurements on the character of genotypic relatives. To do this we first find the conditional probabilities that R_2 say, is of a certain genotype given the genotype of R_1 and then proceed to derive the joint probability distribution for the genotypes of R_1 and R_2.

Now if R_1 is of genotype AA, then R_2 must have allele A in common with R_1. Also since the occurrence of the second allele in R_2 is independent of the occurrence of the known allele A, the second allele in R_2 is either A with probability p or a with probability q = 1-p.

Now, since it is assumed that the relatives R_1 and R_2 must have at least one allele in common, the relative R_2 cannot be of genotype aa if R_1 is of genotype AA.

Hence, the required conditional probabilities are:

$$P(R_2 = AA / R_1 = AA) = p$$

$$P(R_2 = Aa / R_1 = AA) = 1 - p$$

$$P(R_2 = aa / R_1 = AA) = 0$$

If now R_1 is of genotype Aa, then R_2 must have either the allele A or the allele a in common with R_1. Since the second allele occurs independently of the first allele in R_2, the second allele is either A with probability p or a with probability (1-p). Hence if R_1 is of genotype Aa, then R_2 is of genotype AA with probability p; of genotype Aa (or aA) with probability 1-p +p = 1 and of genotype aa with probability 1-p.

Hence,

$$P(R_2 = AA / R_1 = Aa) = p$$

$$P(R_2 = Aa / R_1 = Aa) = 1$$

$$P(R_2 = aa / R_1 = Aa) = 1 - p$$

Similarly,

$$P(R_2 = AA / R_1 = aa) = 0$$

$$P(R_2 = Aa / R_1 = aa) = p$$

$$P(R_2 = aa / R_1 = aa) = 1 - p$$

Now to find the joint probability distribution of R_1 and R_2, we apply the multiplication law of probability (Miller J 1996), which states that for any two events x and y,

$$P(X = x, Y = y) = P(Y = y / X = x)P(X = x)$$

Hence,

$$P(R_1 = AA, R_2 = AA) = P(R_2 = AA / R_1 = AA)P(R_1 = AA) = p(p.p) = p^3$$

since the alleles in the genotype occur independently.

Similarly,

$$P(R_1 = AA, R_2 = Aa) = P(R_2 = Aa / R_1 = AA)P(R_1 = AA) = (1\text{-}p)(p.p) = p^2(1\text{-}p).$$

Other probabilities are:

$$P(R_1 = AA, R_2 = aa) = P(R_2 = aa / R_1 = AA)P(R_1 = AA) = (0)\ (p^2) = 0$$

$$P(R_1 = Aa, R_2 = AA) = P(R_2 = AA / R_1 = Aa)P(R_1 = Aa) = p \times p\ (1\text{-}p) = p^2 (1\text{-}p)$$

$$P(R_1 = Aa, R_2 = Aa) = P(R_2 = Aa / R_1 = Aa)P(R_1 = Aa) = 1 \times p\ (1\text{-}p) = p\ (1\text{-}p)$$

$$P(R_1 = Aa, R_2 = aa) = P(R_2 = aa / R_1 = Aa)P(R_1 = Aa) = (1\text{-}p)\ (p\ (1\text{-}p)) = p(1-p)^2$$

$$P(R_1 = aa, R_2 = AA) = P(R_2 = AA / R_1 = aa)P(R_1 = aa) = 0 \times (1-p)^2 = 0$$

$$P(R_1 = aa, R_2 = Aa) = P(R_2 = Aa / R_1 = aa) \times P(R_1 = aa) = p \times (1-p)^2 = p\ (1-p)^2$$

$$P(R_1 = aa, R_2 = aa) = P(R_2 = aa / R_1 = aa) \times P(R_1 = aa) = (1\text{-}p).\ (1-p)^2 = (1-p)^3$$

These calculations yield the results of Table 1 which shows the joint probability distribution of R_1 and R_2, the marginal probability distribution and the corresponding measurements on the character of interest in the population.

Hence from the Table 1,

$$E(R_1) = p^2x + 2p(1-p)y + (1-p)^2z = m$$

And

$$E(R_2) = p^2x + 2p(1-p)y + (1-p)^2z = m \qquad (1)$$

Table 1. Joint probability distribution of R_1 and R_2.

Genotype for relative R_1	Measurement V	Genotype for relative R_2			Marginal probability $P(R_1)$
		AA	Aa	aa	
		X	Y	Z	
AA	X	P^3	$P^2(1-p)$	0	P^2
Aa	Y	$P^2(1-p)$	$P(1-p)$	$P(1-p)^2$	$2P(1-p)$
Aa	Z	0	$P(1-p)^2$	$(1-p)^3$	$(1-p)^2$
Marginal probability $P(R_2)$		P^2	$2P(1-p)$	$(1-p)^2$	1

Therefore, the expectation of R_1 is equal to that of R_2. The corresponding variance S_t^2 on the measurement R_1 is given as:

$V(R_1)= E(R_1^2) – (E(R_1))^2$

$$= p^2X^2 - 2p(1-p)Y^2 + (1-p)^2Z^2 - (p^2X + 2P(1-P)Y + (1-P)^2Z)^2$$

$$= P^2(1-P^2)X^2 + 2P(1-P)(1-2P-2P^2)Y^2 + (1-P)^2(2P(1-P))Z^2 - 2P^2(1-P^2)XZ - 4P(1-P)^3ZY - 4P^3(1-P)XY$$

And that of R_2 is $V(R_2)= E(R_2^2) – (E(R_2))^2$

$$= P^2X^2 - 2p(1-p)Y^2 + (1-p)^2Z^2 - (P^2X + 2P(1-P)Y + (1-P)^2Z)^2$$

$$= P^2(1-P^2)X^2 + 2P(1-P)(1-2P-2P^2)Y^2 + (1-P)^2(2P(1-P)Z^2 - 2P^2(1-P^2)XZ - 4P(1-P)^3ZY - 4P^3(1-P)XY$$

$= V(R_1)$

Hence, the variance of the measurement on R_1 which is the same as the measurement on R_2 is given as

$$S_t^2 = p^2x^2 + 2p(1-p)y^2 + (1-p)^2z^2 - m^2 \quad (2)$$

Where $m = E(R_1) = p^2x + 2p(1-p)y + (1-p)^2z$ from Equation 1.

The covariance between R_1 and R_2 is also calculated in the usual way from the table as:

$S_{12} = Cov(R_1,R_2) = E(R_1,R_2) – E(R_1)E(R_2)$ (Uche, 2004).

$$S_{12} = p^3x^2 + p^2(1-p)xy + p^2(1-p)xy + p(1-p)y^2 + p(1-p)^2yz + p(1-p)^2yz + (1-p)^3z^2 - m^2$$

Where $m = E(R_1) = E(R_2)$

$$= p^3x^2 + 2p^2(1-p)xy + p(1-p)y^2 + 2p(1-p)^2yz + (1-p)^3z^2 - m^2$$

$$= p^3x^2 + 2p^2(1-p)xy + p(1-p)y^2 + 2p(1-p)^2yz + (1-p)^3z^2 - (p^2x + 2p(1-p)y + (1-p)^2z)^2$$

$$S_{12} = p(1-p)((p^2x^2 + 2p(1-2p)xy + (1-2p)^2y^2 - 2p(1-p)xz - 2(1-p)(1-2p)yz + (1-p)^2z^2)$$

$$S_{12} = p(1-p)[px + (1-2p)y - (1-p)z]^2 \quad (3)$$

The correlation, r, between R_1 and R_2 is found by dividing Equation (3) by Equation (2) since the variance of R_1 is the same as that of R_2.

Thus,

$r = \frac{cov(R_1R_2)}{var(R_1)}$, since $var(R_1) = var(R_2)$

Then,

$$r = \frac{S_{12}}{S_t^2} \quad (4)$$

Note that since p≥0.

The covariance, S_{12} and hence the correlation, r, is a non negative quantity, and for 0<p<1, has a value zero only when

$$p = \frac{x-y}{x+z-2y} \quad (5)$$

Provided x and z are both greater than y or x and z are both less than y.

EFFECT OF ALLELIC RECOMBINATION

Let us now examine what would happen to the measurements of the character of interest and hence the correlation if we recombine the alleles by replacing one allele of a genotype by another allele. Specifically, and without loss of generality, suppose we replace an allele A by an allele a in a genotype determining a certain character in an individual, by this allelic replacement model, the original individual must possess an A allele and hence must be of genotype AA or of genotype Aa.

Hence, if the replacement is being made in an AA individual, the resulting effect on the measurement on the character of interest is to reduce y by x; that is

$y - x$,

while if the allelic replacement is being made on an Aa individual the effect would be

$z - y$.

Interest is now on finding the differential effect of this allelic recombination on the measurement of the character concerned and its significance. We propose to do this using the method of least squares. Let us, therefore, find the best estimates, in the least square sense, of the parameters $\mu, \alpha \text{ and } \beta$ that would minimize the expected sum of squared deviations of x, y, and z from $\mu + 2\alpha, \mu + \alpha + \beta, \text{and } \mu + 2\beta$, respectively, assuming random mating and subject to the constraint

$$P\alpha + (1-p)\beta = 0 \quad (6)$$

Where α= the effect of allele A on the character of interest and β =the effect of allele a on the character of interest.

The expected sum of squared deviations of observed from their true values of the measurements using the marginal probability distribution of Table 1.

$$S^2e = p^2(x-\mu-2\alpha)^2 + 2p(1-p)(y-\mu-\alpha-\beta)^2 + (1-p)^2(z-\mu-2\beta)^2 \quad (7)$$

Differentiating S^2e with respect to μ and minimizing we have

$$\frac{dS^2e}{d\mu} = 2p^2(x-\mu-2\alpha) + 4p(1-p)(y-\mu-\alpha-\beta) + 2(1-p)^2(z-\mu-2\beta) = 0.$$

When solved replacing μ, α, and β by their estimates $\hat{\mu}, \hat{\alpha} \text{ and } \hat{\beta}$, respectively, and simplifying yields

$$p^2x + 2p(1-p)y + (1-p)^2z = p^2(\hat{\mu}+2\hat{\alpha}) + 2p(1-p)(\hat{\mu}+\hat{\alpha}+\hat{\beta}) + (1-p)^2(\hat{\mu}+2\hat{\beta}).$$

The right hand side of the above equation reduces to

$$p^2\mu + 2p^2\hat{\alpha} + 2p\hat{\mu} + 2p\alpha + 2p\hat{\beta} - 2p^2\hat{\mu} - 2p^2\hat{\alpha} - 2p^2\hat{\beta} + \hat{\mu} + 2\hat{\alpha} - 2p\hat{\mu} - 4p\hat{\alpha} + p^2\hat{\mu} + 2p^2\hat{\beta}$$

$$= 2[p\hat{\alpha} + (1-p)\beta] + \hat{\mu}$$

$$= \hat{\mu}.$$

Since from equation (6)

$$P\alpha + (1-p)\beta = 0.$$

Hence we have that

$$\hat{\mu} = m = p^2x + 2p(1-p)y + (1-p)^2z \quad (8)$$

Also, differentiating S_e^2 with respect to α, yields

$$\frac{dS_e^2}{d\alpha} = -4p^2(x-\mu-2\alpha\alpha) - 4p(1-p)(y-\mu-\alpha-\beta) = 0$$

$$= -4[p^2(x-\mu-2\alpha) + p(1-p)(y-\mu-\alpha-\beta)] = 0$$

Therefore,

$$p^2(x-\mu-2\alpha)+p(1-p)(y-\mu-\alpha-\beta)=0$$

Replacing the values of α and β by their estimates $\hat{\alpha}$ and $\hat{\beta}$ and solving, we have:

$$p^2(x-m)+p(1-p)(y-m)$$
$$=p[p\hat{\alpha}+(1-p)\hat{\beta}+\hat{\alpha}]$$
$$=p(0+\hat{\alpha})=p\hat{\alpha}.$$

Hence, we then have that

$$\hat{\alpha}=p(x-m)+(1-p)(y-m) \quad (9)$$

Differentiating S_e^2 with respect to β we have:

$$\frac{dS^2e}{d\beta}=-2p(1-p)(y-\mu-\alpha-\beta)-4p(1-p)^2(2-\mu-2\beta)$$
$$=-2p[(1-p)(y-\mu-\alpha-\beta)+2(1-p)^2(z-\mu-2\beta)]=0$$

$$p(1-p)(y-m)+2(1-p)^2(z-m)=(1-P)\beta$$

Replacing β by its estimate $\hat{\beta}$ and solving we obtain

Hence, the differential effect of replacing the allele A by the allele a on the measurement of the character concerned is estimated using Equations (9) and (10) as:

$$\hat{\beta}=p(y-m)+2(1-p)(z-m) \quad (10)$$
$$\hat{\beta}-\hat{\alpha}=[p(y-m)+(1-p)(z-m)]-[p(x-m)+(1-p)(y-m)]$$
$$=[py-pm+z-pz-m+pm]-[px-pm+y-py-m+pm]$$
$$=py-pm+z-pz-m+pm-px+pm-y+py+m-pm$$
$$=py-px+py-pz+z-y$$

$$=p(y-x)+(1-p)(z-y) \quad (11)$$

Substituting equations (8) – (10) into equation (7), we obtain, after simplification, the minimum value of the sum of squared deviation, S^2e, or the so-called "error sum of squares" in regression anal-ysis parlance, as

$$S_e^2=p^2(1-p)^2[x+z-2y]^2 \quad (12)$$

For 0<p<1, S_e^2 assumes the value of zero only if

$$y=\tfrac{1}{2}(x+z) \quad (13)$$

That is if the measurement on individuals who are heterozygous on the character of interest is equal to half the sum of the measurements on homozygous individuals. The difference between the total sum of square S_t^2 of equation (2) and S^2e of equation 12 namely, the sum of squares due to regression model that is that part of the total variation S_t^2 accounted for or explained by the regression model.

$$S_o^2=S_t^2-S_e^2=2p(1-p)(px+(1-2p)y+(1-p)z)^2 \quad (14)$$

S_o^2 is similar to a regression sum of squares and may be interpreted as that part of the total variation S_t^2 in the measurement of the characters of interest attributable to the average effects of α and β of the alleles A and a. We note from equation (4) that recombination has the effect of reducing the value of the total variation S_t^2 by S_e^2, the residual variance. Also, it can be seen from Equation (3) that:

$$S_o^2=2S_{12} \quad (15)$$

Thus the allelic recombination doubles the covariance between genotypic relatives.

Now the proportion of total variance, S_t^2 explained by α and β is

$$Q=\frac{S_o^2}{S_t^2}=R^2 \quad (16)$$

Table 2. Analysis of variance table for the hypothesised regression model.

Source of variation	Sum of square	Degrees of freedom	Mean square	F-ratio
Regression	SSR= $S_R^2 = 2p(1-p)(px+(1-2p)y+(1-p)z)^2$	2	MSR = $\frac{SSR}{2} = \frac{S_R^2}{2}$	F = $\frac{MSR}{MSE} = \frac{(n-3)S_R^2}{2S_E^2} = \frac{(n-3)(2p(1-p)(px+(1-2p)y+(1-p)z)^2)}{2p^2(1-p)^2[x+z-2y]^2}$
Error	SSE = $S_E^2 = p^2(1-p)^2[x+z-2y]^2$	n-3	MSE = $\frac{SSE}{n-3} = \frac{S_E^2}{n-3}$	
Total	SST = $S_T^2 = p^2x^2 + 2p(1-p)y^2 + (1-p)^2z^2 - m^2$	n-1		

Where R^2 is the usual coefficient of determination in regression parlance.
Or equivalently from equation (4) we have that

$$Q = \frac{2S_{AB}}{S_E^2} = 2r \qquad (17)$$

In other words, the proportion of total variance in the measurements on the character of interest that is accounted for by the effect of manipulation of the alleles is equal to twice the correlation between genotypic relatives in the absence of allelic recombination.

TESTING THE SIGNIFICANCE OF THE FIT OF THE REGRESSION MODEL AND THE CORRELATION BETWEEN GENOTYPIC RELATIVES

One may be interested in testing the hypothesis that the regression model fits. That is, testing whether the differential effects of the allelic recom-bination on the character of interest are statistically different from zero. To test the null hypothesis Ho, we may the Analysis of Variance method (Montgomery and Peck, 1992). The three sums of squares, their associated degrees of freedom, their mean squares, and the resulting F-ratio are summarised in Table 2.

The F-ratio =

$$\frac{(n-3)(2p(1-p)(px+(1-2p)y+(1-p)z)^2)}{2p^2(1-p)^2[x+z-2y]^2} \qquad (18)$$

which has an F-distribution with (2, n-3) degrees of freedom and may be compared at an α significance level with tabulated critical F-value to test that the regression model fits. If:

F-ratio > $F_{(1-\alpha),\,(2,\,n-3)}$

We reject the null hypothesis of no differential effects of allelic recombination on the genotypic relatives.
We may also wish to test the null hypothesis that allelic recombination has no significant effect on the correlation between genotypic relatives that is that the population correlation coefficient ρ due to allelic recombination of genotypic relatives is zero versus the alternative hypothesis that ρ is different from zero.

Now, note that testing that the regression model fits is equivalent to testing the hypothesis that, in the sampled population, the values of Q which is equal to R^2 is equal to 2r not equal to zero.
The significance of the sample estimates of these population parameters is tested using the usual F-test of equation (18). The rejection of the null hypothesis implies that the population values of Q equals R^2 is not equal to or equivalently, Q equals 2r is not equal to zero in the population implying that r is not equal to zero in the population sampled. Hence the usual F-test provides a test statistic for testing the null hypothesis that the correlation between genotypic relatives is zero versus the alternative hypothesis that the correlation between genotypic relatives is different from zero. The null hypothesis is rejected at an appropriately chosen significance level α.

CONCLUSION

This paper provided estimates of the correlation between genotypic relatives both in the presence and in the absence of allelic ecombination. It is shown that the correlation between genotypic relatives in the absence of allelic recombination is double the correlation between genotypic relatives in the presence of allelic recom-bination. The correlation obtained is a non negative quantity and except for trivial cases (p = 0 or 1), assuming the value zero only for some critical value p. The significance of the correlation obtained and the regression model fitted are tested using the analysis of variance technique.

REFERENCES

Alberts B (1994). Molecular Biology of Cells (3rd edition)

Garland Publishing Company.
Burns G (1976). The Science of Genetics: An Introduction to Heredity (3rd edition) Macmillan Publishing Company, New York.
Clavel F, Hoggan MD, Willey RL, Strebel K, Martin M A, Rapaske R (1989). Genetic Recombination of Human Immunodeficiency Virus. J. Virol., 63(3): 1455-1459.
Ewens WJ (1979). Mathematical Population Genetics, Springer – Varlang, New York.
Fisher RA (1918). The Correlation Between Relatives on the Supposition of Mendalan Inheritance. Transition Royal Society of Edinburgh. 52: 399–433.
Maloy S, Cronan J, Freifeldev (1994). Microbial Genetics (2nd Edition), Jones and Bartlet MA.
Miller J (1996). Statistics for Advanced Level (2nd Edition) Cambridge University Press.
Montgomery DC, Peck EA (1992). Introduction to Linear Regression Analysis. John Willey and Sons, 2nd Edition.
Oyeka ICA (2009). Applied Statistical Methods in Sciences Norben Avocation. pp: 135–137.
Oyeka ICA, Oyeka CA (1988). Estimating the Effect of Allelic Recombination on the Correlation Between Genotypic Relatives. Nigerian J. Biotechnol., Enugu, Nigeria.
Uche PI (2004). Probability: Theory and Practice Longman Nig PLC.