Theor Appl Genet (1988) 75:720–722

Offspring-parent regression for a binary trait

S. Im* and D. Gianola**

Department of Animal Sciences, University of Illinois, Urbana, IL 61801, USA

Received August 31, 1987; Accepted November 11, 1987
Communicated by D. Van Vleck

Summary. Offspring-parent regression is often used to estimate the heritability of a quantitative trait. It is shown that for a purely binary trait, the regression of offspring on one parent is always linear, while that on both parents or mid-parent is generally nonlinear. However, the regressions are linear on a logistic scale.

Key words: Offspring-parent regression – Heritability – Binary traits

Introduction

Offspring-parent regression is often used to estimate the heritability of a quantitative trait. The assumption of linearity of the regression function is usually accepted by plant and animal breeders. However, the problems caused by nonlinearity have been recognized and discussed by some workers (e.g., Robertson 1977; Bulmer 1980; Gimelfarb 1986). As an example of experimental evidence, Robertson (1977) and Maki-Tanila (1987) discovered curvature in offspring-parent regression for bristle number in *Drosophila*, and proposed several models of gene action to explain this phenomenon.

In this article, we are concerned with the offspring-parent regression for a binary trait. Some authors have studied this problem assuming an underlying normally distributed "liability" variable (Falconer 1965; Thompson et al. 1985). However, we assume in this paper that the trait is purely binary, without invoking the existence of "liability". It is shown that the regression of offspring on one parent is always linear, while that on both parents or mid-parent is generally nonlinear.

Regression on one parent

Consider a binary trait such as "alive" versus "dead", or "diseased" versus "healthy". Suppose that individuals are given a score 1 if they have the attribute and 0 if they do not. Let the random variable (X, Y) be the scores of a parent-offspring pair. The mean value of an offspring given the phenotypic value of a parent can be written as:

$$E(Y|X=x) = E(Y|X=0) + [E(Y|X=1) - E(Y|X=0)]\,x. \quad (1)$$

When $x=0$, the above becomes $E(Y|X=0)$; when $x=1$, one obtains $E(Y|X=1)$. Thus, the regression of offspring on one parent is always linear.

If p_{xy} denote the joint probabilities $Pr(X=x, Y=y)$ for $x, y = 0, 1$, then the slope of the regression is:

$$\beta = E(Y|X=1) - E(Y|X=0)$$
$$= [(0)\,Pr(Y=0|X=1) + (1)\,Pr(Y=1|X=1)]$$
$$- [(0)\,Pr(Y=0|X=0) + (1)\,Pr(Y=1|X=0)]$$
$$= Pr(Y=1|X=1) - Pr(Y=1|X=0) = \frac{p_{11}}{p_{1+}} - \frac{p_{01}}{p_{0+}} \quad (2)$$

where $p_{x+} = p_{x0} + p_{x1}$; $(x = 0, 1)$.

Maximum likelihood estimation of β

Suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ are the realized values of the binary trait for n unrelated parent-offspring pairs in a randomly mating population. Let n_{xy} be the frequencies of occurrence of (x, y) for $x, y = 0, 1$. Then the distribution of (n_{xy}) is multinomial with parameters n and (p_{xy}). The maximum likelihood estimator of p_{xy} is n_{xy}/n_{++},

* Permanent address: Laboratoire de Biometrie, INRA-Toulouse, B.P. 27, F-31326 Castanet-Tolosan, France
** To whom correspondence should be addressed

where $n_{++} = n_{00} + n_{01} + n_{10} + n_{11}$. Using the invariance property of maximum likelihood estimates in (2), it follows that the maximum likelihood estimates in (2), it follows that the maximum likelihood estimator of the regression is

$$\hat{\beta} = \frac{n_{11}}{n_{10}+n_{11}} - \frac{n_{01}}{n_{00}+n_{01}}. \tag{3}$$

The standard error of $\hat{\beta}$ can be readily obtained using the δ-method (Aickin 1983). The statistic $2\hat{\beta}$ gives an estimate of the heritability of the binary trait.

Relationship with the coefficient of correlation

Under normality of (X,Y), the slope of the regression and the correlation coefficient are equal when the variances are the same. Such a relationship also holds true for a bivariate (0,1) distribution. Suppose that X and Y have as joint distribution

$$\begin{aligned}&\Pr(X = x, Y = y) \qquad (4)\\ &= p^x (1-p)^{1-x} q^y (1-q)^{1-y} \left[1 + \varrho \frac{(x-p)(y-q)}{\sqrt{p(1-p)\,q(1-q)}}\right],\end{aligned}$$

where ϱ is the correlation coefficient between X and Y, and p and q are marginal probabilities of 1 for X and Y, respectively (Hamdan and Martinson 1971). Note in (4) that when $\varrho = 0$ and $x = y = 1$, $\Pr(X = 1, Y = 1) = pq$. This is the result that one would obtain under independence assumptions.

If $E(X) = E(Y) = p = q$, then $\mathrm{Var}(X) = \mathrm{Var}(Y) = p(1-p)$. Using this in (4):

$$\begin{aligned}&\Pr(X = x, Y = y)\\ &= p^{x+y}(1-p)^{2-x-y}\left[1 + \varrho \frac{(x-p)(y-p)}{p(1-p)}\right]. \qquad (5)\end{aligned}$$

Application of (5) gives

$$p_{11} = p^2 [1 + \varrho(1-p)/p]$$

$$p_{10} = p_{01} = p(1-p)(1-\varrho)$$

$$p_{00} = (1-p)^2 [1 + \varrho\, p/(1-p)].$$

If (2) is written in terms of the above probabilities, one obtains as slope of the regression line:

$$\begin{aligned}\beta = &\frac{p^2[1+\varrho(1-p)/p]}{p(1-p)(1-\varrho) + p^2[1+\varrho(1-p)/p]}\\ &- \frac{p(1-p)(1-\varrho)}{(1-p)^2[1+\varrho\, p/(1-p)] + p(1-p)(1-\varrho)} = \varrho, \qquad (6)\end{aligned}$$

with $\beta = \varrho$ following after algebra. Therefore, if X and Y have the same mean, and thus the same variance, the correlation coefficient and the slope of the regression are equal.

If (x_i, y_i) are records of n indepent individuals, and $E(X) = E(Y)$, then $\beta = \varrho$ is repeatability. Rutledge (1977) compared several estimators of repeatability of a threshold trait by Monte Carlo methods. Not surprisingly, he concluded that $\hat{\beta}$, referred to as Lush's estimator, was a poor estimator of the correlation on the liability scale and that estimates of correlation on the two scales should not be used interchangeably. However, $\hat{\beta}$ is a sensible estimator if variation is purely binary.

Regression on both parents

Let X_1, X_2 be scores on the two parents, Y be the offspring score and let

$$p_{x_1x_2y} = \Pr(X_1 = x_1, X_2 = x_2, Y = y).$$

Define

$$\pi_{x_1x_2} = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2),$$

so that

$$\pi_{x_1x_2} = \frac{p_{x_1x_21}}{p_{x_1x_20} + p_{x_1x_21}}; \quad (x_1, x_2) = 0, 1. \tag{7}$$

The mean value of an offspring given the score of the parents is

$$\begin{aligned}&E(Y | X_1 = x_1, X_2 = x_2) \qquad (8)\\ &= (0)\left[\frac{p_{x_1x_20}}{p_{x_1x_20} + p_{x_1x_21}}\right] + (1)\left[\frac{p_{x_1x_21}}{p_{x_1x_20} + p_{x_1x_21}}\right] = \pi_{x_1x_2}\end{aligned}$$

with values:

π_{00} when $x_1 = x_2 = 0$

π_{01} when $x_1 = 0, x_2 = 1$

π_{10} when $x_1 = 1, x_2 = 0$

π_{11} when $x_1 = 1, x_2 = 1$.

Thus, the regression of offspring on both parents can be written as

$$\begin{aligned}&E(Y | X_1 = x_1, X_2 = x_2)\\ &= \pi_{00} + (\pi_{10} - \pi_{00})x_1 + (\pi_{01} - \pi_{00})x_2\\ &\quad + (\pi_{11} - \pi_{10} - \pi_{01} + \pi_{00})x_1 x_2. \qquad (9)\end{aligned}$$

For example, when $x_1 = x_2 = 1$, (9) gives π_{11}, as it should be. Unlike the usual normal model, the regression of offspring on both parents is generally nonlinear. There is an interaction between the scores of the parents on the probability that their offspring has the trait.

If parameters are the same ($\pi_{01} = \pi_{10}$) in males and females, (9) reduces to

$$\begin{aligned}&E(Y | X_1 = x_1, X_2 = x_2) \qquad (10)\\ &= \pi_{00} + (\pi_{10} - \pi_{00})(x_1 + x_2) + (\pi_{11} - 2\pi_{10} + \pi_{00})x_1 x_2.\end{aligned}$$

The regression on the mid-parent value, $\bar{x} = (x_1 + x_2)/2$, is then

$$\begin{aligned}&E(Y | \bar{X} = \bar{x})\\ &= \pi_{00} + 2(\pi_{10} - \pi_{00})\bar{x} + (\pi_{11} - 2\pi_{10} + \pi_{00})\delta_1(\bar{x}), \qquad (11)\end{aligned}$$

where

$$\delta_1(\bar{x}) = \begin{cases} 1 & \text{if } \bar{x} = 1 \\ 0 & \text{if } \bar{x} \neq 1. \end{cases}$$

Thus, the regression on both parents and the regression on mid-parent are linear only if the difference in conditional probabilities is additive, that is:

$$\pi_{10} = \frac{\pi_{00} + \pi_{11}}{2}$$

or, equivalently

$$\pi_{11} - \pi_{10} = \pi_{10} - \pi_{00}.$$

In general, these regressions are not linear and the offspring mid-parent regression cannot be used to estimate the heritability of a binary trait.

Regression on a logistic scale

Applying a log-linear model to the joint probabilities $p_{x_1x_2y}$, we show that the regression on both parents or on mid-parent values is linear on a logistic scale but not on the observed scale. Consider the model

$$\log p_{x_1x_2y} = \mu + \alpha(x_1 + x_2 + y) + \theta(x_1y + x_2y), \tag{12}$$

which is a log-linear model with no interaction between the parents (x_1x_2). With this model, and employing (7) the conditional probabilities are given by:

$$\pi_{00} = \frac{e^{(\mu+\alpha)}}{e^{\mu} + e^{(\mu+\alpha)}} = \frac{e^{\alpha}}{1+e^{\alpha}}, \tag{13a}$$

$$\pi_{10} = \pi_{01} = \frac{e^{\mu+2\alpha+\theta}}{e^{\mu+\alpha} + e^{\mu+2\alpha+\theta}} = \frac{e^{\alpha+\theta}}{1+e^{\alpha+\theta}}, \tag{13b}$$

$$\pi_{11} = \frac{e^{\mu+3\alpha+\theta}}{e^{\mu+2\alpha} + e^{\mu+3\alpha+\theta}} = \frac{e^{\alpha+2\theta}}{1+e^{\alpha+2\theta}}. \tag{13c}$$

Using (8)

$$E(Y|X_1 = x_1, X_2 = x_2) = \frac{p_{x_1x_21}}{p_{x_1x_20} + p_{x_1x_21}} = \frac{e^{\alpha+\theta(x_1+x_2)}}{1+e^{\alpha+\theta(x_1+x_2)}}. \tag{14}$$

Also

$$E(Y|\bar{X} = \bar{x}) = \frac{e^{\alpha+2\theta\bar{x}}}{1+e^{\alpha+2\theta\bar{x}}}. \tag{15}$$

Thus, the regressions on both parents or on mid-parental values are linear on a logistic scale, the functions being $\alpha + \theta(x_1 + x_2)$ or $\alpha + 2\theta\bar{x}$, but not so on the observed scale, as can be readily ascertained by differentiating (14) of (15) with respect to the parental scores.

The parameters α and θ can be estimated by maximum likelihood. The distribution of the frequencies $n_{x_1x_2y}$ is multinomial with parameter n and probabilities $p_{x_1x_2y}$. The likelihood function is

$$l(\alpha, \theta) = \frac{n!}{\prod_{x_1,x_2,y} n_{x_1x_2y}!} \prod_{x_1,x_2,y} e^{\mu+\alpha(x_1+x_2+y)+\theta(x_1y+x_2y)} \tag{16}$$

and estimates of parameters and of their standard errors can be obtained under the constraint

$$\sum_{x_1,x_2,y} e^{\mu+\alpha(x_1+x_2+y)+\theta(x_1y+x_2y)} = 1.$$

The statistic $2\hat{\theta}$ gives an estimator of heritability on the logistic scale.

Note in (13a)–(13c) that a heritability of 0 implies $\pi_{00} = \pi_{01} = \pi_{11}$. From (8), this also implies that the mean value of the offspring is expected to be the same irrespective of the scores of the parents. Also, using (8) and (12) it can be seen that if $\theta = 0$

$$E(Y|X_1 = x_1, X_2 = x_2) = \pi_{x_1x_2} = \frac{e^{\mu+\alpha(x_1+x_2+1)}}{e^{\mu+\alpha(x_1+x_2)} + e^{\mu+\alpha(x_1+x_2+1)}} = \frac{e^{\alpha}}{1+e^{\alpha}}. \tag{17}$$

Acknowledgements. This research was supported by the Illinois Agriculture Experiment Station and by Grant US-805-84 from BARD, the United States-Israel Binational Agricultural Research and Development Fund.

References

Aickin M (1983) Linear statistical analysis of discrete data. Wiley, New York, 585 pp

Bulmer MG (1980) The mathematical theory of quantitative genetics. Clarendon Press, Oxford, 255 pp

Falconer DS (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann Hum Genet 29:51–76

Gimelfarb A (1986) Offspring-parent genotypic regression: How linear is it? Biometrics 42:67–71

Hamdam MA, Martinson EO (1971) Maximum likelihood estimation in the bivariate binomial (0,1) distribution: application to 2 × 2 tables. Aust J Stat 13:154–158

Maki-Tanila A (1987) Non-linearity of the regression of offspring on parent's phenotype. Abstr 2nd Int Conf Quant Genet. Raleigh, North Carolina, p 75

Robertson A (1977) The non-linearity of offspring-parent regression. In: Pollak E, Kempthorne O, Bailey TB Jr (eds) Proc Int Conf Quant Genet. Iowa State University Press, Ames, pp 297–304

Rutledge JJ (1977) Repeatability of threshold traits. Biometrics 33:395–399

Thompson R, McGuirk BJ, Gilmour AR (1985) Estimating the heritability of all-or-none and categorical traits by offspring-parent regression. J Anim Breed Genet 102:342–354