COVARIANCE ANALYSIS WITH A COVARIATE INTERACTION: AN EXAMPLE OF A SIMPLE LINEAR REGRESSION COMPARISON TECHNIQUE

D. E. Palmquist

C. A. Stockwell

COVARIANCE ANALYSIS WITH A COVARIATE INTERACTION: AN EXAMPLE OF A SIMPLE LINEAR REGRESSION COMPARISON TECHNIQUE

D.E. Palmquist and C.A. Stockwell
USDA/ARS, 920 Valley Road, Reno, NV 89512 and
Dept. of Biology, University of Nevada Reno

ABSTRACT

Many real data sets that would normally lend themselves to being analyzed by an analysis of covariance, have a covariate interaction present with one or more of the factors in the experiment. Because this violates the assumption of same-slope covariate effect across all treatments, an analysis of covariance should not be performed. The course normally taken when there is such an interaction is to derive regression equations for the dependent variable as a function of the covariate, at each level of the factor(s) being tested. A general linear model F-test can then be used to test whether there are any overall differences between the regression lines. A technique that uses two mathematical distance measures to detect regression line differences once a significant general linear model F-test is obtained is illustrated. Applying these distance measures enables us to perform modified multiple comparisons of the regressions without resorting to the use of multiple pairwise general linear model F-tests, which inflate the Type I error rate. With this method, we are able to incorporate both factor and covariate information into the analysis to overcome the covariate-factor interaction problem.

1. INTRODUCTION

Mosquitofish (Gambusia affinis) are small, guppy-like fish native to the southeastern United States that have been widely introduced worldwide for mosquito abatement control (Courtenay and Meffe 1989). They exhibit striking differences in life history characteristics depending on where the populations are located (Stockwell and Mulvey in review). This study focuses on four isolated populations in warm springs of the western Great Basin of Nevada who share a common ancestor from approximately 60 years ago. Around 1940, mosquitofish were introduced to the Wabuska Hot Springs, a geothermal-fed spring, from Fallon, Nevada (Stockwell and Mulvey in review), and to the Garret Ranch in the Black Rock Desert. They were then introduced into both artificially drilled and natural artesian springs to the Parker and Bonham Ranches in the Smoke Creek Desert in Northwestern Nevada (Stockwell and Vinyard in review). These four sites are located about 50 linear km apart and lack any hydrologic connections between the springs, so no natural gene flow occurs (Figure 1).

Each of the waters at the four sites differ in thermal and chemical/physical properties. Water properties such as temperature, conductivity, salinity, and pH were measured at each of the four sites (Table 1), along with mosquitofish characteristics such as brood size of pregnant females, mother length, mother somatic weight, fat content, and embryo weight, for two different months (April and May) in 1993. Thermally stable sites, Bonham and Garrett, have females with similar life history traits such as: maturity at larger sizes, relatively high fat reserves, low to moderate reproductive allotment, and small embryos (Stockwell and Mulvey in review). In the Wabuska site, where thermal fluctuations are the norm due to the activities of a geothermal plant, the females reach maturity at smaller sizes, have lower fat reserves, moderate to high reproductive allotment, and large embryos. The Parker Ranch site mosquitofish live in a pool with minimal thermal input and are characterized by moderate size at maturity, low fat reserves, moderate to high reproductive allotment, and have large embryos (Stockwell and Vinyard in review).

Female body size measurements, such as length and weight, are good covariates in explaining some of the variation in these site-specific population differences. Before any such analyses were run, we checked for homogeneity of slopes. For all reproductive attributes, except brood size, the homogeneity of slopes assumption was not violated by the mother size covariate. For these attributes, an analysis of covariance was used to analyze site differences. This was not possible for the dependant variable of brood size. Fecundity, expressed as brood size, was a confounding of the female mother weight covariate and the population site. The researchers were unwilling to discard the important relationship between brood size and the female mother weight covariate, and could not look for differences due to site alone. Of course, we could have made each site-covariate combination into a factor and partitioned the degrees of freedom in the ANOVA table to examine each one individually, as is sometimes done, but we wanted to try this method for its ease of use and because we believe we get more information from the regressions themselves. The information provided by the regression lines adds even further to the study of the factor-covariate interaction than a partitioning of the ANOVA table would provide.

Simple linear regressions of brood size as a function of the female weight covariate were calculated so that more information on the covariate effect on brood size could be incorporated into the analysis. A general linear model F-test was performed to determine if any of the regression lines were different from the rest (Neter, Wasserman, and Kutner 1990). If a significant F-test result was obtained, indicating that at least one of the regression lines was different from the others, we used two

mathematical distance measures as a modified regression multiple comparison technique to determine which equations were different from and which were similar to each other (Palmquist et al. 1993, and Palmquist 1993).

Using this technique, we were able to incorporate covariate information into the analysis. We were able to confirm our visual interpretation of graphs of the simple linear regressions without resorting to the use of multiple general linear model pairwise tests, which inflate the Type I error rate. We were able to statistically determine which line or lines should be removed so that a covariance analysis could be conducted on only those treatments having equal covariate slope effects. The use of this technique allows us to check for site differences while incorporating the effects of the covariate on fecundity. As the number of treatment levels and comparisons increase, this method becomes even more useful as a post hoc grouping technique for regression analysis.

2. METHODS

Separate covariance analyses, one each for April and May of 1993, were run with site as the independent variable, brood size as the dependent variable, and using female somatic weight as the covariate, using Proc GLM in SAS (SAS Institute Inc. 1989). A site-covariate interaction term was included in these preliminary analyses to check for homogeneity of slopes across sites before proceeding with a regular analysis of covariance. The month of April had a significant site by female weight covariate interaction ($p \leq 0.0008$), as did the month of May ($p \leq 0.0001$). Simple linear regressions were computed for each site within each month:

April 1993

(1)	Bonham	$Y = 9.8 + 92.3*X$	$R^2 = 0.17$
(2)	Garrett	$Y = 1.9 + 111.4*X$	$R^2 = 0.43$
(3)	Parker	$Y = 1.3 + 110.9*X$	$R^2 = 0.63$
(4)	Wabuska	$Y = 0.9 + 273.5*X$	$R^2 = 0.75$

May 1993

(5)	Bonham	$Y = -0.9 + 94.8*X$	$R^2 = 0.54$
(6)	Garrett	$Y = 13 + 23.7*X$	$R^2 = 0.04$
(7)	Parker	$Y = 1.7 + 253.5*X$	$R^2 = 0.82$
(8)	Wabuska	$Y = 4.1 + 171.7*X$	$R^2 = 0.54$

Where the dependent variable Y is brood size, and the independent variable X is the female weight covariate (Figure 2). In Equation (6), not only is the R^2 extremely low, but the slope coefficient is not significant. This site was removed from further May analysis. All of the other equations have significant slope coefficients. Since the Garrett Ranch site was removed from May 1993, the analysis of covariance, including the factor-covariate interaction term, was rerun and again found to be significant ($p \leq 0.0001$).

A general linear model test was conducted to test the hypothesis of equality of the four simple linear regressions for April and the remaining three for May. Results show that for April, at least one of the four regression lines is different from the rest ($p \leq 0.01$), and for May, at least one of the three regression lines tested is different from the others ($p \leq 0.01$).

At this point, the distance measures were calculated and used to associate the simple linear regressions into groups having similar and those having unlike slopes. The distance measures, originally used for quadratic multiple linear regression comparisons, have an intuitive appeal that is much easier to see when the form of the equations is simple linear. The first, measure A, is based on the standard mathematical property of orthogonality. In the simple linear case, it relates directly to the angle between two lines, or, the slope (Figure 3). To satisfy covariance analysis assumptions of equal slope effect across all factors without resorting to multiple pairwise tests or removing one variable at a time and re-testing, this measure provides a simple way of determining which slopes are similar and which are disparate. The second measure, D_p, is a measure of the area between two lines. It assumes nothing about equality of slopes, but gives a general picture of how far apart two lines or surfaces are (Figure 3). In combination with the measure A, it provides an overall grouping for a set of linear equations, without having to divide the experimentwise alpha level by the number of comparisons being made. This is a modified multiple comparison, or grouping technique because we do not know how far apart or how orthogonal two lines have to be from one another in order to be declared significantly different. The experimentwise

alpha level used is that of the general linear model F-test that precedes these calculations. Since we know from a significant F-test that at least one of the lines is different from the rest, we make only one significance statement with the distance measures. We say that the only significant difference is that between the pair of regression lines having the largest value of the distance measures (Palmquist et al. 1993).

Information from measure A will be used and weighted more than for measure D_p, since it has the bigger role to play in terms of direct slope comparisons (Palmquist 1993). But, information from D_p will be utilized as a cross check to measure A. The calculation of both of the distance measures does not involve an iterative process. This is an advantage over multiple t or F tests and even over multiple comparison tests where one tries to control the experimentwise error rate. The measures provide a grouping tool whereby slopes, in this example, can be classified as either the same or not. This allows us to eliminate one or more locations at a time from the covariance analysis plus allows us to see covariate-location relationships via the individual regressions themselves.

3. RESULTS AND DISCUSSION

Values of the measure A for all pairwise site comparisons within each month were computed (Table 2). The largest value for a pair of sites occurred between Bonham and Wabuska for April 1993 data ($p \leq 0.01$). This is the only significantly declared slope difference among all six of the pairwise simple linear regression site comparisons for this month. Even though we cannot make any statements of significance concerning the other five comparisons, it is easy to see how they group together. The next-to-largest values of measure A occur between Parker and Wabuska, and between Garrett and Wabuska, respectively. This indicates that Wabuska is the site showing the most difference in the covariate slope effect of the four sites. The two sites showing the smallest value of measure A, hence exhibiting the most similar slopes, are Garrett and Parker. Bonham doesn't appear to be too different from either Garrett or Parker, based on measure A values. This interpretation is presented with the standard use of multiple comparison letters, whereby significant differences exist only for those comparisons with no letters in common (Table 4). This agrees with the visual interpretation one would expect from examining the graphs of the regression lines for April (Figure 2).

The largest value of the measure A for May 1993 data is between the Bonham and Parker sites (Table 2), implying that the slopes of the Bonham and Parker lines are significantly different ($p \leq 0.01$). The next-to-largest measure A value occurs between Bonham and Wabuska, while the most similar slopes are between the

Parker and Wabuska site regression lines. As in April, the multiple comparison letter interpretation of these differences for May (Table 4) agrees with the visual interpretation of the graph of the lines (Figure 2).

The information from measure D_p, while not as pertinent to the covariate-slope issue as is the measure A, is used as a check in this situation to see if similar results can be obtained. For April 1993 data, the largest value of D_p occurs between the Parker and Wabuska regressions (Table 3). Hence, these two site regression lines are significantly different ($p \leq 0.01$), based on a metric of the area between them. The D_p values of Garrett and Wabuska, and Bonham and Wabuska do not appear to be much different from that of Parker and Wabuska. As was the case for measure A, the Garrett and Parker lines exhibit the least difference while the Bonham regression line appears to be equally similar to both Garrett and Parker (Table 3). Even though a different pair of regression lines are declared significantly different by the measure D_p than by the measure A, the interpretation remains the same (Table 4). In both cases, Wabuska's regression line is clearly the least similar of the four lines tested.

For May 1993 data, the results from measure D_p mirror those from measure A. The Bonham and Parker regression lines are significantly different ($p \leq 0.01$), while the most similarity occurs between the Parker and Wabuska lines (Table 3). The multiple comparison letter interpretation (Table 4) agrees with that of measure A, and also is visually verified by examining the graphs (Figure 2). For both measures A and D_p, the regression line for the Bonham site is the least similar of the three lines tested.

We applied the information from our distance measures to covariance analysis by deciding to rerun the covariance analysis test for April 1993 data without the Wabuska site, and for May 1993 data without the Bonham site (the Garrett Ranch site was already deleted due to nonsignificance of it's regression line). We found that there is no longer a weight by site covariate-factor interaction for April 1993 when the Wabuska site is deleted from the analysis. We also found no covariate-factor interaction for the May 1993 analysis when both Garrett and Bonham were deleted. The researchers now have the option of running a covariance analysis they can use without assumption violations, plus information from the individual regressions themselves.

4. SUMMARY

The presence of a covariate-factor interaction violates the same-slope covariate effect across all treatments assumption needed for an analysis of covariance. After deriving regression lines of the covariate as a function of the dependent variable for all treatments, a standard general linear model F-test approach is usually performed to determine whether or not there are differences between the lines. To detect regression line differences once a significant general linear model F-test is obtained, we use two mathematical distance measures. These measures, A and D_p, allow us to perform modified multiple pairwise regression line comparisons. Measure A in the simple linear regression case is a direct measurement of slope similarity, whereas measure D_p is a general metric of the area between two lines.

Using this technique as a modified multiple comparison method for regression equations enables us to group the simple linear regressions in this experiment to statistically determine which one(s) should be excluded from an analysis of covariance so that the same-slope covariate assumption is not violated.

We are able to incorporate covariate information into the analysis, as well as have the more specific information that the simple linear regressions provide for each treatment. We could confirm our visual interpretation of the graphs of the simple linear regressions without resorting to the use of multiple general linear model pairwise tests which inflate the Type I error rate.

ACKNOWLEDGEMENTS

Special thanks to Dr. Sitadri Bagchi of the Mathematics Dept. at the University of Nevada Reno, Dr. James Young, Research Leader of the USDA/ARS Conservation Biology unit, Reno, NV, and to Dr. Gary Vinyard of the Biology Dept. at the University of Nevada Reno for all their help from concepts to photographs.

REFERENCES

Courtenay, W.R. Jr. and G.K. Meffe. 1989. Small fishes in strange places: A review of introduced Poeciliids. PP. 319-331 in G.K. Meffe and F.F. Snelson Jr, eds., Ecology and Evolution of livebearing fishes (Poeciliidae). Prentice Hall, Englewood Cliffs, New Jersey.

Neter, J., W. Wasserman, and M.H. Kutner. 1990. Applied Linear Statistical Models. Third Edition, Richard D. Irwin, Inc.

Palmquist, D.E. 1993. The use of two distance measures as a modified multiple comparison technique for obtaining response surface differences. MS thesis, University of Nevada Mathematics Dept., Reno, NV.

Palmquist, D.E., S.N. Bagchi, J.A. Young, and R.D. Davis. 1993. Distance measures in post hoc comparisons of temperature-germination quadratic response surfaces. Proceedings of the 1993 Kansas State University Conference on Applied Statistics in Agriculture, pp. 31-39, Manhattan, KS.

SAS Institute Inc., SAS/STAT Users Guide, Version 6, Fourth Edition, Volume 2, Cary, NC:SAS Institute Inc., 1989. 846 pp.

Stockwell, C.A. and M. Mulvey. In review. Preserving allelic diversity: are translocations successful? Submitted December 1994 to Conservation Biology.

Stockwell, C.A. and G.L. Vinyard. In review. Life history variation in recently isolated populations of mosquitofish (Gambusia affinis): A case of rapid evolution? Submitted to Copeia.

Table 1. Physical and chemical characteristics of the four Nevada study site springs.

SITE	TEMPERATURE (° C)	CONDUCTIVITY (μmohs)	SALINITY (%)	pH
BONHAM		4139	2.14	8.00
Pool	29.14 (25-33)			
Side	21.36 (13-33)			
GARRETT	33.90 (29-41)	1638	0.73	8.78
PARKER	27.50 (26-29)	323	0.00	8.95
WABUSKA	28.70 (12-40)	1693	0.90	8.47

Table 2. Values of Measure A for all pairwise site comparisons for April and May 1993.

APRIL 1993

Pairwise Site Comparisons	Measure A	
Bonham-Garrett	1.72×10^{-6}	
Bonham-Parker	1.65×10^{-6}	
Bonham-Wabuska	2.58×10^{-5}	*
Garrett-Parker	8.19×10^{-10}	
Garrett-Wabuska	1.42×10^{-5}	
Parker-Wabuska	1.44×10^{-5}	

May 1993

Pairwise Site Comparisons	Measure A	
Bonham-Parker	2.18×10^{-5}	*
Bonham-Wabuska	1.12×10^{-5}	
Parker-Wabuska	1.77×10^{-6}	

Table 3. Values of Measure D_p for all pairwise site comparisons for April and May 1993.

APRIL 1993

Pairwise Site Comparisons	Measure D_p	
Bonham-Garrett	4.77	
Bonham-Parker	5.40	
Bonham-Wabuska	29.71	
Garrett-Parker	0.69	
Garrett-Wabuska	32.50	
Parker-Wabuska	33.14	*

May 1993

Pairwise Site Comparisons	Measure D_p	
Bonham-Parker	30.02	*
Bonham-Wabuska	18.00	
Parker-Wabuska	12.20	

Table 4. Interpretation from Measures A and D_p. Sites followed by the same letter(s) are not significantly different at the alpha = 0.01 level.

APRIL 1993

SITE	MEASURE A	MEASURE D_p
Bonham	a	ab
Garrett	ab	ab
Parker	ab	a
Wabuska	b	b

MAY 1993

SITE	MEASURE A	MEASURE D_p
Bonham	a	a
Parker	b	b
Wabuska	ab	ab

Figure 1. Introduction history for mosquitofish in Nevada

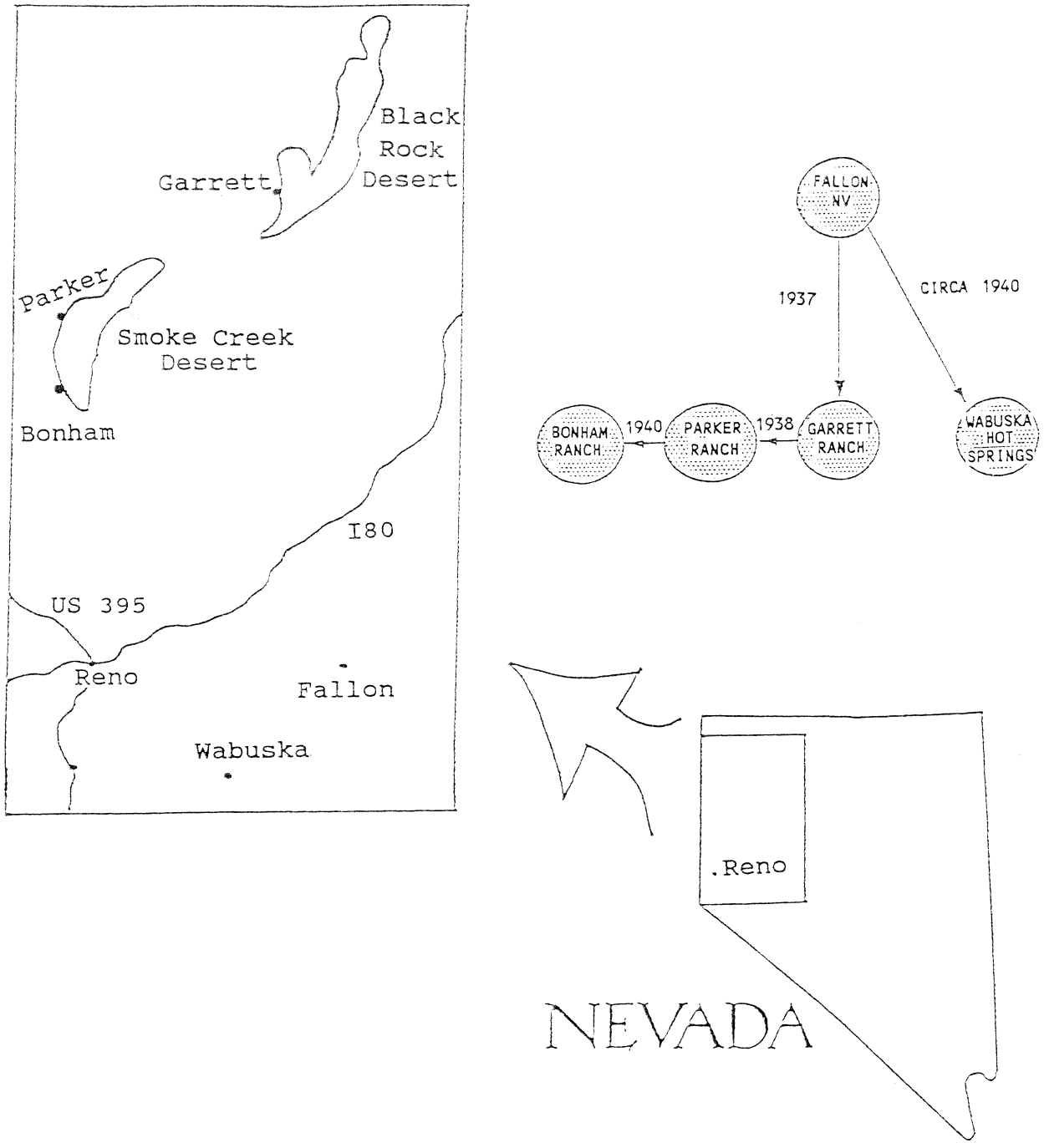
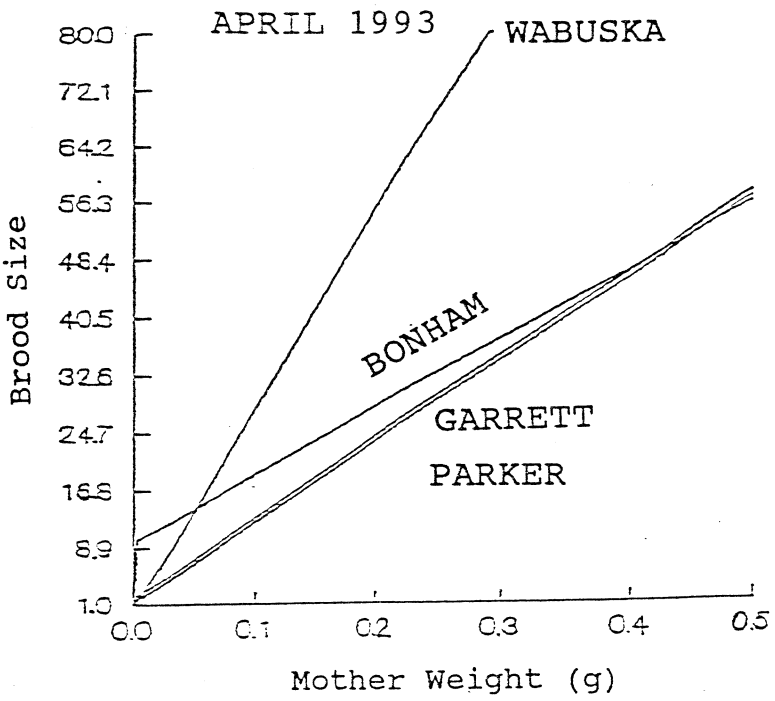


Figure 2. Simple linear regression graphs of the four Nevada study sites for each month.



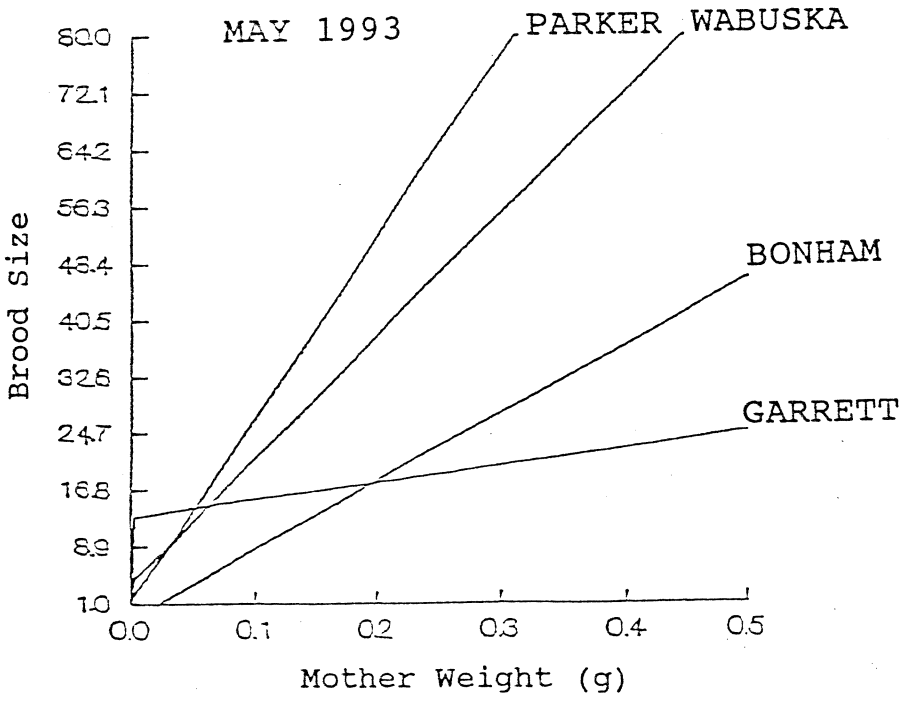


Figure 3. DEFINITION OF DISTANCE MEASURES A AND D_P.

MEASURE A

$$A = 1 - |\cos \theta|, \qquad 0 \le A \le 1,$$

where

$$\cos \theta = \frac{(\beta^{(i)}, \beta^{(j)})}{\|\beta^{(i)}\| \|\beta^{(j)}\|}, \qquad 0 \le \theta \le \pi,$$

$\beta^{(n)}$ = Coefficient vector of the slope in the n^{th} simple linear regression equation.

θ = Angle between lines i and j.

MEASURE D_P

$$D_P = \left[\int_x \frac{(f^i - f^j)^2}{C} \, dx \right]^{1/2}$$

f^n = The n^{th} simple linear regression equation.

C = Standardization constant for comparison of all n regression lines. For SLR, C = X_{MAX} (of all the lines) - X_{MIN} (of all the lines).