Psychological Bulletin
1978, Vol. 85, No. 6, 1317-1322

To Understand Regression From Parent to Offspring, Think Statistically

Lloyd G. Humphreys
University of Illinois at Urbana-Champaign

Incorrect inferences are drawn by many psychologists about regression from one generation to another. These appear to be caused by confusion between observations and theory. Regression is basically statistical. Biological regression represents an interpretation of the data. If regression is always conceptualized in the first instance as a statistical phenomenon, fewer mistakes will occur. It is appropriate to make theoretical interpretations only after one is clear about the empirical observations.

After the report "Educational Uses of Tests with Disadvantaged Children" (Cleary, Humphreys, Kendrick, & Wesman, 1975) appeared, I received several letters that criticized the section of the report that discussed regression from parent to child. We were told, for example, that we had assumed genetic determination of intelligence in that discussion. We were also told that there was no regression from midpoint IQ to child. These and related errors have also been made frequently by my students and colleagues. More dramatically, errors of interpretation have been made in recent publications by two research workers in the field of behavior genetics. There is apparently a need for an exposition concerning the nature of regression from parent to child.

Only one principle is required to clear up misunderstandings: Regression is a statistical phenomenon. Biological regression is only one possible interpretation. The whole gamut of causes that produce correlations less than unity between parents and offspring may be and typically are involved. The basic statistical observation must not be confused with a possible theoretical interpretation.

Regression in Intelligence

Regression in intelligence was the topic discussed in the report by Cleary et al. (1975) and represents a useful area in which to apply the general principle just described. If one starts with a sample of individual parents and individual children and computes the correlation between paired scores (IQs for convenience) on a standard test of intelligence, a value of .50 is a representative finding. The sex of either the parent or the child appears to be immaterial. Control of the birth order effect would be desirable and can readily be done, but birth order attenuates the size of this correlation only slightly. The estimation of the IQ of either the parent or the child is made by means of the usual linear regression equation:

$$\hat{Y} = r_{xy}\frac{S_y}{S_x}(X - \bar{X}) + \bar{Y}. \quad (1)$$

No genetic assumptions or estimates are needed or desirable. One does make one important statistical assumption: Regression is approximately linear. The expected variance of obtained scores about expected scores of all persons in the sample is as follows:

$$S_{y \cdot x}{}^2 = S_y{}^2(1 - r_{xy}{}^2). \quad (2)$$

If the bivariate distribution is homoscedastic, Equation 2 can also be used to estimate the

This research was supported by National Institute of Mental Health Grant MH 23612-04, Studies of Intellectual Development and Organization. The author is indebted to N. W. Norton of the Department of Animal Science, University of Illinois at Urbana-Champaign for many helpful suggestions.

Requests for reprints should be sent to Lloyd G. Humphreys, 425 Psychology Building, University of Illinois, Champaign, Illinois 61820.

distribution of discrepancies about each and every predicted score. It is not necessary to assume that the distributions are bivariately normal unless one wishes to make interpretations beyond those of Equations 1 and 2.

The estimation of child's IQ from parent's IQ, or the reverse, is made from the obtained correlation, which reflects determinants from three distinct areas of causation, namely, genetics, environment, and measurement error. The contribution of the last of these can be estimated quite accurately, given an adequate research design, but the relative contributions of the first two are uncertain. This uncertainty does not, however, affect in the slightest the accuracy of the prediction.

The regression formula shows very clearly the requirements for the statement that children tend to regress halfway back to the mean of the population of the parents. In addition to knowledge of the correlation, one must have confidence in the identity of the two means and the two standard deviations. If $\bar{X} = \bar{Y} = 100$, $S_x^2 = S_y^2 = 256$, $r = .50$, the regression is linear, and the arrays are homoscedastic, the mean IQ of the children of a large group of parents all having the same extreme IQ of 150 will be 125 with variance of 192.

When one wishes to predict a child's IQ and knows the IQs of both parents, the principles do not differ. Only the constants used in the regression equation need to be changed. One must have the standard deviation of midparent IQs and the correlation between midparent IQ and child IQ. These constants are rarely reported, but one finds more frequently means, standard deviations, and intercorrelations of husband, wife, and child. From these statistics the standard deviation and the correlation required can readily be computed by means of formulas for the sum of two components (see Appendix).

For purposes of this exposition it is assumed that standard deviations for husband and wife are equal and that the correlation between husband and wife is also .50. The last figure is a reasonably representative finding, but note that it reflects the present degree of assortative mating in a particular country. It is complexly determined by a variety of social customs and institutions. Given the parent–child correlation of .50 discussed above, the combination of assumptions leads to a variance of the distribution of midpoint IQs of 192 and a correlation with child's IQ of $\sqrt{\frac{1}{3}}$. Making the appropriate substitutions in the regression equation, the prediction from a midparent IQ of 150 is 133 with variance of 171. Thus, knowledge of the IQs of both parents does lead to less regression by the children toward the mean of the midparent distribution, but regression is still far from zero in amount.

Table 1
Parent–Child Correlations (N = *105 Familes*)
From Jones (*1928*)

Child	Father	Mother	Raw score weights	Regression weights
Son	.524	.544	.596	.598
Daughter	.505	.557	.592	.597

Note. Raw score weights are the observed standard deviations of the IQs; regression weights are those obtained in multiple regression. Jones reported .586 for both midparent correlations, which involve raw score weights. The discrepancy is most probably due to grouping errors in frequency distributions. In 1928, correlations were typically computed from scatter plots.

It is of interest both genetically and environmentally that the raw score linear composite of father and mother (midparent IQ) predicts the child's IQ as well as does the linear composite that uses optimum weights. Data to illustrate this are available in the article by Jones (1928) and are presented in Table 1.[1] These data also illustrate the lack of importance of knowing the sex of the child in estimating the child's IQ. Clearly, there is in these data no sex linkage of a genetic sort in the development of individual differences in intelligence. Although the *N* is too small for us to be certain, there

[1] I am greately indebted to Arthur R. Jensen, who informed me about these data, which are more complete than most similar data sets in the literature. The correlations in Table 1 are congruent with the assumed population correlations used in this discussion.

is in the sample a difference that could represent a small environmental linkage.

One can also estimate the correlation between one parent and two or more offspring in a parallel fashion. The correlation between parent and midchild in a two-child family will be higher than the correlation between parent and one child. As the number of children increases, the correlation of parent with the composite will increase, but there will be decreasing returns in predictability as the number of children increases as a function of the intercorrelations of the IQs of the offspring. A sibling–sibling correlation of .50 is a representative figure. Note also that the midparent–midchild correlation will be higher still.

When one keeps in mind the statistical nature of regression, it is not at all confusing that regression takes place, seemingly backward in time, from children to parents. Given linearity, regression is symmetrical. Everything depends on whether one selects extreme parents and wishes to predict a child's IQ or one selects extreme children and wishes to predict a parent's IQ. If a group of children are selected, each of whom has an IQ of 150, the predicted IQ of one parent or of the midparent is 125. This seeming paradox becomes less paradoxical when one recalls that 125 is more extreme in the midparent than in the individual-parent distribution. Thus, the same numerical score represents different amounts of regression in the two cases. Variability about the predicted score also differs. It is even possible that on occasion, extreme families could be selected, in which case the usual regression from generation to generation would not occur.

It is useful to compare these computations with those that would be expected given perfect (narrow) heritability, random mating, perfect reliability of measurement, and no sex linkage. The parent–child correlation would be .50, but the midparent–child correlation would be .707. This is a statistical result of the assumptions made concerning the size of correlations, and it also fits genetic theory perfectly. The variance of midparent IQs would be 128, smaller than before because husband and wife are paired at random. The prediction of a child's IQ from a midparent IQ of 150 would be 150 with variance of 128. Thus, in the hypothetical situation described, the expected regression would seemingly shrink to zero, but midparent IQs of 150 would be much more extreme in random mating than in assortative mating. There would also be a good deal of error in making the prediction. Husband and wife determine one half of the variance in children's IQs under these hypothetical conditions but less than one half under actual present-day conditions. The lack of independence of husband and wife in intelligence, however, does not carry over to certain quantitative physical traits of very high heritability that are also measured with very high reliability.

It is noteworthy that data on intelligence do not quite fit the genetic model. One can allow for the contribution of measurement error by correcting all correlations for attenuation, but the corrected correlation of the composite of husband and wife with child is less than .707. Perhaps the present degree of assortative mating has not been stable at its present level for a sufficient number of generations to increase parent–child correlations to appropriate levels. On the other hand, there may be environmental pressures that reduce the size of parent–child correlations below the level expected on the basis of the genetic model. For example, the childhood environments of the two parents are not identical with the environment they have provided for their children.

Applications to Recent Publications

These principles are quite straightforward, but as noted above, mistakes have been made by persons working in this area. One example occurs in Jensen's (1973) book. On page 170, he undertook to explain the mean IQs of the children of Terman's gifted group. The basic data he used were as follows: The mean IQ for gifted parents was 152, and that for their spouses was 125.

From these data Jensen obtained a midparent mean IQ (138.5), applied an assumed narrow heritability coefficient of .71, and computed the predicted mean (126). Since this estimate was lower than the obtained mean (132.7), he explained the dif-

Table 2
Mean IQs of Parents and Their Biological Children

	Parents				Children		
Measure	N	Father	Mother	Midparent	N	Obtained	Expected[a]
WAIS	99	120.8	118.2	119.5	14	118.9	113
WISC					82	117.9	113
Stanford-Binet					48	113.8	113
Total sample	99	120.8	118.2	119.5	144	116.6[b]	113[b]

Note. WAIS = Wechsler Adult Intelligence Scale full-scale IQ; WISC = Wechsler Intelligence Scale for Children full-scale IQ; Stanford-Binet = Stanford-Binet Intelligence Scale IQ.
[a] Based on the assumption that the three scales of measurement are equivalent.
[b] Mean value for total sample.

ference as being due to the expected environmental contribution to individual differences in intelligence that would be associated with any set of high-IQ parents.

The computations and the chain of reasoning contain two errors. In the first place, the mean IQ of the spouses furnishes no information that will increase the accuracy of the prediction. It merely confirms within an error of one IQ unit that there has been the expected amount of assortative mating in the gifted group. It does not matter whether one predicts a child's IQ from a single parent's IQ of 152 or from a midparent's IQ of 138.5. It is 126 in the former case, 125.7 in the latter. The second error derives from the fact that the prediction is made from empirical data that incorporate all of the causal elements involved: genetics, environment, and random error. If the gifted group represents the extreme tail of the distribution of IQs in the unselected population, one cannot invoke the usual (in that tail) environmental component in children's IQs as an added component. The environmental component is included in the predicted value. Jensen's use of a biological construct, narrow heritability, in place of a statistical prediction produced the error.

There is, however, a significant discrepancy between the predicted and obtained mean that requires an explanation. Since there are several different tests involved, an obvious source of the discrepancy lies in the test norms. The discrepancy, in other words, might be an artifact of the scales of measurement used. If this explanation can be ruled out, then one may turn to more complex possibilities. The gifted group and their children may not be an unselected sample of the tail of the population distribution of IQs, in spite of the evidence to the contrary furnished by the IQs of the spouses. Was there experimenter bias in the testing of the children of the gifted? Was the original selection of the gifted group based on grounds much broader than test score, so the typical parent–child correlation was exceeded by a substantial amount? Did the gifted parents coach their children in answering questions on intelligence tests? Did the children know that their parents were gifted, and was their own behavior influenced by that knowledge? Did the parental knowledge that they themselves were gifted tend to produce an unusually stimulating intellectual environment in the home? Note that this last hypothesis requires the home environment to have stimulating properties beyond those expected in families that have one parent testing at the 152 level.

The question concerning the original selection of the gifted group raised in the preceding paragraph is particularly pertinent to a discussion of regression. Students nominated by teachers were tested, and among these, only those children who exceeded the IQ cutoff for the definition of *gifted* entered the study. Thus, they were a selected subset of all students who would have met the test criterion had all students been tested. Was the selection of students to be nominated based in large part on information available to the teacher about individual parents? Was this information highly correlated with parental IQs? If the gifted group had parents who were substantially more intelligent than the expected regressed level, this would also reduce somewhat the amount of regression expected in their children. Did the teacher nominations add sufficiently to the reliability of the test scores, even though they were based solely on the behavior of the children, that the amount of regression would be reduced? Though both possibilities are realistic to some degree, it seems doubtful that either alone or in combination they could produce a discrepancy of the size involved in these data.

A second example of deficient understanding of the mechanism of regression occurred in Scarr and Weinberg's (1976) article. Their discussion was based on the intelligence test scores of white parents and their biological children in families who had adopted black or racially mixed children. Their data appear in Table 2.

The authors made the following comment concerning the data in Table 2: "As expected from polygenic theory, when both parents have high IQ scores, there is less regression toward the population mean than under conditions of random mating" (Scarr & Weinberg, 1976, p. 731). This statement is confusing at best and is clearly wrong in the implication it leaves with the reader—that there is nothing remarkable about the findings. By their inadequate analysis the authors have overlooked something potentially important.

The parents in these families were not directly selected on the basis of their intelligence, although they were clearly well above average in this regard. The primary selection was based on a combination of attitudes and domestic economic circumstances that led to the adoption of a child of another race and the decision to participate in the research project. The amount of indirect selection for intelligence is the resultant of two factors: the correlation between intelligence and the complex continuum underlying the decisions to adopt and to participate, and the distance above the population mean of the cutoff score on the adoption-participation continuum. It is interesting to note, therefore, that indirect selection operated about equally on husband and wife, since their means are approximately equal to each other. The decisions to adopt and to participate were apparently shared equally.

If the biological children did not participate in the decisions in any way, their mean IQ in our present society should be about two thirds of the distance from the population mean to the mean of the two parents. In these data, however, their mean is significantly higher than expected. Again, if one is confident that the scales of measurement are not responsible for the discrepancy, one can speculate psychologically concerning causes. Was there a sufficiently large number of biological children of an age that would have allowed meaningful participation in the family's decisions to adopt and to participate? Could the decision to participate have been based in part on the degree of satisfaction felt by the parents in the progress of their own children? (The authors raised a related possibility with respect to satisfaction felt in the progress of their adopted children.) Might the decision to *adopt* children of another race have been based in part on the degree of satisfaction felt by the parents in the progress of their own children? This last speculation is a particularly interesting one from the point of view of possible environmental consequences for the adopted children.

Conclusion

The concluding statement for this note is very brief: Always conceptualize regression problems in terms of statistical regression. Once clarity has been achieved concerning the empirical observations, theoretical explanations are in order.

References

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist,* 1975, *30,* 15–41.

Jensen, A. R. *Educability and group differences.* New York: Harper & Row, 1973.

Jones, H. E. A first study of parent–child resemblance in intelligence. *Yearbook of the National Society for the Study of Education* (Part 1), 1928, *27,* 61–72.

Scarr, S., & Weinberg, R. A. IQ test performance of black children adopted by white families. *American Psychologist,* 1976, *31,* 726–739.

Appendix

Let x and y represent any two measures, for example, IQs of father and mother, and let z represent a third measure, such as the IQ of a child. The variance of the mean of the first two measures is derived as follows:

$$S_{(x+y)/2}{}^2 = \frac{\Sigma x^2 + \Sigma y^2 + 2\Sigma xy}{4N}$$

$$= \frac{S_x{}^2 + S_y{}^2 + 2r_{xy}S_xS_y}{4}.$$

Making the appropriate substitutions, the variance of midparent IQs is readily computed:

$$S^2 = \frac{256 + 256 + 2(.50)256}{4} = \frac{768}{4} = 192.$$

If, however, mating was random, the variance of midparent IQs would be smaller:

$$S^2 = \frac{256 + 256}{4} = \frac{512}{4} = 128.$$

The correlation between the mean of the two measures and a third measure is derived as follows:

$$r_{[(x+y)z]/2} = \frac{\Sigma xz + \Sigma yz}{2NS_{(x+y)/2}S_z}$$

$$= \frac{r_{xz}S_x + r_{yz}S_y}{\sqrt{S_x{}^2 + S_y{}^2 + 2r_{xy}S_xS_y}}.$$

Making the appropriate substitutions, the correlation between midparent IQ and child IQ is easily obtained:

$$r = \frac{(.5)16 + (.5)16}{\sqrt{768}} = \sqrt{\frac{256}{768}} = \sqrt{\tfrac{1}{3}}.$$

If, however, mating was random and the identical parent-child correlations were observed, the correlation between midparent IQ and child's IQ would be larger:

$$r = \frac{(.5)16 + (.5)16}{\sqrt{512}} = \sqrt{\frac{256}{512}} = \sqrt{\tfrac{1}{2}}.$$

Received August 5, 1977 ■