DEMOGRAPHY© Volume 23, Number 2 May 1986

IS THE DENSITY-FERTILITY RELATION A STATISTICAL ARTIFACT?: A REPLY TO ERIC JENSEN

Glenn Firebaugh
Department of Sociology and Anthropology, Vanderbilt University, Nashville, Tennessee 37235

In a regression analysis of the 1961–1972 crude birthrates for 22 farm villages in the Punjab, India, I consistently obtained a negative coefficient for population density. Therefore I wrote that "Population density apparently has an inhibiting effect on fertility in these villages" (Firebaugh 1982, abstract; hereafter "Punjab study"). Professor Jensen suggests a different explanation for the negative coefficient: that it is artifactual. He gives two types of arguments—one logical, one statistical—for his artifact interpretation. *The logical argument*: An inverse relation between density and fertility is "puzzling" and "counterintuitive." If high density dampens fertility, (a) how did lower-fertility villages become more densely populated by 1961, and (b) why is there no evidence of convergence on the crude birth rate (CBR) during 1961–1972? *The statistical argument*: The two key variables, the CBR and population density, have a common term (population). As a result, the observed relation between the CBR and density "is illusory, a statistical artifact" (Jensen 1986:284).

Both arguments are wrong. Because Jensen devotes much more attention to the statistical argument, I begin with it.

RATIO VARIABLES AND THE ARTIFACT ISSUE

At least since Pearson (1897; but see Yule 1910) many readers have been wary of results based on the correlation of ratio variables with common terms (e.g., percentages, rates); the use of such variables in correlation or regression analysis presumably introduces special problems. Like many others in that tradition, Jensen relies heavily on intuition. There is something unsettling about correlating ratio variables that share a component; many social scientists intuit a "built-in" relation between such variables, and they refer to such relations as "artifactual," "spurious," "definitionally dependent," and "tautological" (e.g., Logan 1982; Pendleton 1984; but see Belsley 1972).

That common intuition notwithstanding, the use of ratio variables with common terms introduces no special "artifact" problems in regression analysis (at most, the use of ratios might obscure problems that already exist). This point can be demonstrated in a straightforward manner. Consider the following population model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \quad (1)$$

where ε has a zero mean and is independent of the Xs. The parameters of (1) can be estimated by regressing Y on the Xs:

$$Y = b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + e \quad (2)$$

Since ε has a zero mean and is independent of the Xs, (2) provides unbiased estimates of (1). But (2) is not the only unbiased estimator of (1). The estimator created by dividing equation (1) by one of the regressors is also unbiased. If X_1 is selected as the divisor, the estimator is:

$$Y^* = c_1 + c_2 X_2^* + c_3 X_3^* + c_4 X_4^* + c_5 X_5^* + e^* \quad (2^*)$$

where $Y^* = Y/X_1$, $X_2^* = X_2/X_1$, and so on ($X_1 \neq 0$). Observe that equation (2*) uses ratio variables with a common term (X_1), yet the regression coefficients provide unbiased estimates of the parameters of (1). *Division by a regressor does not introduce bias.* Indeed, division by a regressor is a well-known method of dealing with heteroskedastic error terms, and the method does *not* create bias (e.g., Johnston 1972: 215–216). In short, the "ratio estimator" is unbiased when the corresponding "component estimator" is unbiased (e.g., [2*] is unbiased when [2] is unbiased).

Despite these well-established principles, the artifact myth is firmly entrenched. Jensen, for example, claims that "the use of ratio variables and pooled data create substantial statistical problems," the result being "inflated" estimates (1986: 284). That claim is false. Regardless of the form of the data—pooled or not—the ratio estimator is biased *only when* the corresponding component estimator is biased.[1] Put differently, the use of ratio variables does *not* "create" bias (I made that point in note 10 [Firebaugh 1982: 493], but, as his note 2 makes clear, Jensen missed the point entirely).

Since the ratio estimator is biased only when the corresponding component estimator is biased, allegations about bias in the former also apply to the latter. Hence, in the Punjab study, if the density coefficient is biased in the ratio form of the model, then it is also biased in the component form of the model. To make this concrete, consider equation (2*) and the corresponding component equation, (2). Assume that X_1 and X_2 are village population and area, respectively. Then b_2 and c_2 are the "density coefficients,"[2] and, since c_2 (ratio model) is biased only when b_2 (component model) is biased, we can focus on b_2 in evaluating Jensen's allegation (1986: 284) that I obtained an "inflated estimate of the effect of population density."

Does, then, b_2 in fact give an inflated estimate of the density effect? Jensen thinks so, based on three claims. The first claim is of "pooling bias": "Pooling the data will create spurious correlation between [village density and the CBR]" (1986: 284). Jensen bases this claim on a rather involved argument about weighted coefficients, ratio variables, and relative variances. But there is a more direct way to determine whether pooling creates spurious correlation: compare correlation coefficients before and after pooling. Those coefficients are reported in tables 2 and 4 of Firebaugh (1982: 489, 490): $r = -.22$ before pooling and $-.23$ after pooling. Thus Jensen's fear that pooling "will create spurious correlation" between density and the CBR is unfounded.[3]

The second claim is of omitted-variable bias: "The simultaneous responses of the CBR and village population size to outside-of-model variation yield an inflated estimate of the effect of population density on fertility" (1986: 284). Empirical analyses are almost always vulnerable to the allegation that they suffer from omitted-variable bias. What is new here is the further claim that one can determine a priori the *direction* of the bias. That claim is false. Since the correlations between measured and unmeasured variables are unknown, it is not possible to determined whether omitted-variable bias, if present, has inflated or *attenuated* b_2.[4]

The third claim is of simultaneity bias: "The problem is this: as population grows over time, density increases. . . . Thus . . . the effect of outside-of-model changes in the CBR will be reflected in the coefficient of density" (1986: 283). I certainly do not deny that the CBR affects population density. Nor am I unaware of the estimation problems posed by that fact. That is why I lagged density in the pooled model, and used *household* density (households per km^2 of land) in the cross-section model. Apparently those steps to avoid bias do not satisfy Jensen.

If Jensen is right—if "the effect of outside-of-model changes in the CBR" are

"reflected" in the density coefficient—then presumably that coefficient would be deflated if a simultaneous equation model were used. In this instance, then, Jensen's claim can be tested empirically. Accordingly, I reanalyzed the cross-section data, using two-stage least squares to estimate a two-equation model in which the CBR affects density (population/land) and vice versa.[5] Contrary to Jensen's argument about simultaneity bias, the density coefficient is *larger* in this analysis than it was in the single-equation ordinary least squares (OLS) analysis: in the OLS analysis, b^* (standardized regression coefficient) ranged from $-.23$ to $-.35$ (Firebaugh 1982: table 5); in the two-stage least squares (2SLS) analysis, $b^* = -.36$ ($b = -.043$, $p < .01$). The observed effect of density cannot be dismissed as a "statistical artifact" stemming from simultaneity bias.

IS THE DENSITY EFFECT COUNTERINTUITIVE?

Jensen also objects to the Punjab study on the grounds that an effect of density on fertility is "counterintuitive." To understand that objection, let us begin where Jensen begins, with figure 1 of Firebaugh (1982: 483). Based on annual data for 1961–1972, figure 1 depicts the trends in the average CBR for higher versus lower density villages. There is a general downward trend in the CBR for both groups, with the higher density villages having the lower birthrates every year. But if density has a negative effect on the CBR then why, Jensen asks, is there no apparent convergence on the CBR? Moreover, why do denser villages have lower CBRs the initial year (1961)?

The answer to the first question is that density is not the only determinant of fertility. Observe that figure 1 depicts the relation between density (higher/lower) and CBR—a *bivariate* relation. After *eyeballing* the graph of the *bivariate* density-CBR relation, Jensen is convinced that density does not affect fertility. But when the *same data* are statistically analyzed using a regression model with control variables, and a density effect is observed, Jensen objects that the result cannot be trusted because it is subject to omitted-variable bias!

Nor should it be puzzling that denser villages had lower birthrates in 1961. That fact is quite consistent with my conclusion: "The results indicate that population density has a moderate inhibiting effect on fertility in these villages" (Firebaugh 1982: 491). Observe that I do not claim that the density effect overwhelms all other causes of fertility. Nor do I claim that density has an inverse effect on fertility at all times in all places. High density has become a significant problem in the Indian villages only in recent times.

Finally, it should be noted that, far from being "puzzling" and "counterintuitive," a negative density effect is precisely what I had hypothesized. First, high density makes inheritance problematic. Second, high density reduces child labor value. Third, high density encourages migration, which tends to erode the social supports for high fertility. Those arguments are detailed in Firebaugh (1982: 484–485), yet Jensen does not mention them. In fact, he does not provide *any* evidence, theoretical or empirical, to rebut my theoretical rationale for a negative density effect.[6]

IS THE DENSITY EFFECT COUNTERINTUITIVE?

Jensen also objects to the Punjab study on the grounds that an effect of density on fertility is "counterintuitive." To understand that objection, let us begin where Jensen begins, with figure 1 of Firebaugh (1982: 483). Based on annual data for 1961–1972, figure 1 depicts the trends in the average CBR for higher versus lower density villages. There is a general downward trend in the CBR for both groups, with the higher density villages having the lower birthrates every year. But if density has a

negative effect on the CBR then why, Jensen asks, is there no apparent convergence on the CBR? Moreover, why do denser villages have lower CBRs the initial year (1961)?

The answer to the first question is that density is not the only determinant of fertility. Observe that figure 1 depicts the relation between density (higher/lower) and CBR—a *bivariate* relation. After *eyeballing* the graph of the *bivariate* density-CBR relation, Jensen is convinced that density does not affect fertility. But when the *same data* are statistically analyzed using a regression model with control variables, and a density effect is observed, Jensen objects that the result cannot be trusted because it is subject to omitted-variable bias!

Nor should it be puzzling that denser villages had lower birthrates in 1961. That fact is quite consistent with my conclusion: "The results indicate that population density has a moderate inhibiting effect on fertility in these villages" (Firebaugh 1982: 491). Observe that I do not claim that the density effect overwhelms all other causes of fertility. Nor do I claim that density has an inverse effect on fertility at all times in all places. High density has become a significant problem in the Indian villages only in recent times.

Finally, it should be noted that, far from being "puzzling" and "counterintuitive," a negative density effect is precisely what I had hypothesized. First, high density makes inheritance problematic. Second, high density reduces child labor value. Third, high density encourages migration, which tends to erode the social supports for high fertility. Those arguments are detailed in Firebaugh (1982: 484–485), yet Jensen does not mention them. In fact, he does not provide *any* evidence, theoretical or empirical, to rebut my theoretical rationale for a negative density effect.[6]

CONCLUSION

Is the density-fertility relation a statistical artifact? The answer, I conclude, is no. Like all empirical findings, the observed coefficients might misstate the true effects, so the density coefficient could be "wrong." Moreover, the density coefficient does not identify the mechanism(s)—inheritance, child labor value, or migration—through which density affects fertility. But the density coefficient is not *artifactual*. Jensen's claims are specious. His principal claim, that the use of ratio variables with common components creates bias, is based on a statistical myth. Two other claims, that pooling the village data creates a spurious correlation between density and the CBR and that the density coefficient has been inflated by simultaneity bias are shown to be false. His final claim, that omitted-variable bias must inflate the density coefficient, is also wrong. I see no reason, then, to modify the conclusions of the Punjab study.

NOTES

[1] In the case of aggregate data the ratio estimator often outperforms the component estimator since the disturbance term in (2*) is more likely to be homoskedastic (or nearly so). This seemed most likely for the village data, so I used ratios rather than components in the Punjab analysis.

[2] Actually, X_2/X_1 is the inverse of density, so c_2 represents the effect of "hectares per person" rather than "persons per hectare" (density). Thus a positive sign for c_2 (or b_2) implies a negative effect for density. But that does not affect the bias issue raised by Jensen; if the estimated effect of density is inflated, then the estimated effect of density inverse is also inflated, and vice versa.

[3] Jensen's erroneous conclusion about "pooling bias" is not surprising, since he makes a serious mistake in his argument. Specifically, Jensen argues that the smaller t-values for the density coefficients in the non-pooled data are due to the fact that "variability of land area has not been artifically reduced (by pooling)" (p. 284). That argument makes no sense. Other things constant, standard errors shrink (thus t-values *increase*) as the variance of a regressor increases. For example, in equation (2), the t-value for b_2 increases as the variance of X_2 increases. Recall that X_2 represents area. Thus, if the variance of area is larger in the non-pooled data, the t-value for b_2 should be *larger* (not smaller) for that data.

Moreover, Jensen's statement that the within-village relation between density and the CBR is "based on the . . . correlation of population size with its lagged inverse" (p. 284) is incorrect. That would be true only if density were the dependent variable and the CBR were the lagged independent variable. Thus, in the middle of his argument, Jensen transposes the independent and dependent variables.

[4] If an omitted variable is constant over time for a village, its effect in the error-components analysis is absorbed by the dummy variable for the village, so the method I used *does* "relieve" omitted-variable bias of that sort. In addition to the village dummies, the pooled model includes controls for village caste composition, female literacy, agricultural production, and year. As noted in Firebaugh (1982), the lack of reliable data on other variables precluded a more elaborate fertility model for the Punjab villages. The Punjab results are suggestive—sufficiently so, I hope, that others will want to refine the Punjab model and test the density hypothesis with other data sets.

[5] Prior (1961) population density is added as an exogenous variable to the basic cross-section model (Firebaugh 1982: table 5) and a panel model is formed in which the endogenous variables are 1970 population density and 1970 CBR (1970 CBR is average CBR over 1969–1971). To obtain estimates, it is assumed that there are no "cross-lag" effects, i.e., it is assumed (a) that 1961 density has no direct effect on 1970 CBR (its effect is assumed to be indirect, e.g., 1961 density affects 1970 density, and 1970 density affects 1970 CBR), and (b) that the effect of 1961 CBR on 1970 density is also indirect (through, e.g., 1961 density and 1970 CBR). The coefficients for caste composition and farm output are trivial in the CBR equation, so that structural equation was estimated without those variables.

[6] That high density inhibits fertility is also suggested by evidence from other Indian villages. In Rampur, a village near Delhi, population almost quadrupled between 1911 and 1975; as a result, "incentives for high fertility . . . [have been] counteracted by the rapid subdivision of landholdings" (Das Gupta 1978: 181). Caldwell et al. (1982) make a similar argument based on their study of nine villages in South India.

ACKNOWLEDGMENTS

I thank Barrett Lee and Timothy Bartik for their comments on an earlier draft of this reply.

REFERENCES

Belsley, D. A. 1972. Specification with deflated variables and specious spurious correlation. *Econometrica* 40:923–927.

Caldwell, J. C., P. H. Reddy, and P. Caldwell. 1982. The causes of demographic change in rural South India: a micro approach. *Population and Development Review* 8:689–727.

Das Gupta, M. 1978. Production relations and population: Rampur. *Journal of Development Studies* 14:177–185.

Firebaugh, G. 1982. Population density and fertility in 22 Indian villages. *Demography* 19:481–494.

Jensen, Eric. 1986. A Comment on Glenn Firebaugh's "Population Density and Fertility." *Demography* 23:283–284.

Johnston, J. 1972. *Econometric Methods*. Second Edition. New York: McGraw-Hill.

Logan, C. H. 1982. Problems in ratio correlation: the case of deterrence research. *Social Forces* 60:791–810.

Pearson, K. 1897. Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 60:489–498.

Pendleton, B. F. 1984. Correcting for ratio variable correlation: examples using models of mortality. *Social Science Research* 13:268–286.

Yule, G. U. 1910. On the interpretation of correlations between indices or ratios. *Journal of the Royal Statistical Society* 73:644–647.