Comment

Kirk M. Wolter

The main point of the paper by Freedman and Navidi seems to be that statistical models and inferences derived from them cannot be trusted, and indeed might be misleading, unless the underlying assumptions are made explicit and are shown to be appropriate after careful testing and verification. The paper illustrates the general point with an analysis of some data concerning coverage errors in the 1980 Decennial Census.

Who among us would disagree with the general point? I certainly would not. The stating and checking of assumptions should be an integral part of any scientific investigation conducted by competent professionals. This is so particularly for the modeling of statistical data.

But assumptions always fail to some degree, and when they do, I wish to reserve the right to consider a statistical model useful if it can be demonstrated that decisions made on the basis of the model are better in some sense than the decisions that would be made in the absence of the model. In other words, I believe the notion of model robustness is of central importance. Would Freedman and Navidi go so far as to disagree with this general philosophy? I doubt it. Indeed, I suppose most statisticians would tend to agree with the general philosophy. At the margin, however, there will always be disagreements between statisticians about the acceptability or usefulness of any given model in any specific application.

Now I turn to the analysis of the 1980 data by Freedman and Navidi. These data are concerned generally with the completeness of the 1980 Decennial Census in respect to the population count, and specifically with issues that arose in U. S. district court in *Cuomo v. Baldrige*. Much has already appeared in the statistical literature about this celebrated case, and as background information for new readers I summarize some of the salient features.

In this lawsuit the State and City of New York complained they had been undercounted in the 1980 Census disproportionately to the balance of the nation, and therefore their voting strength was diluted and they were denied their fair share of federal grants to local areas. As a remedy, New York asked the court to compel the Census Bureau to adjust the census population counts for the estimated undercount. The lawsuit first went to trial in 1980. It was decided in favor of New York but was later remanded for a new trial by the Second Circuit because of an improper order entered in the original trial precluding the Census Bureau from introducing evidence in its defense and also because the lower court failed to recognize and consider the important competing interests of other jurisdictions in the census process. The new trial commenced in January 1984 and proceeded in three parts. In part one, Barbara Bailar, Vincent Barabba, Ansley Coale, Charles Cowan, Leon Gilford, Nathan Keyfitz, Richard Nathan, Jeffrey Passel, Jacob Siegel, Michael Stoto, James Trussell, Kenneth Wachter, and Kirk Wolter presented expert testimony on behalf of the Census Bureau (defendants) and Eugene Ericksen, Philip Hauser, Charles Keeley, Samuel Preston, Karl Taeuber, and John Tukey presented expert testimony on behalf of New York (plaintiffs). This initial phase of the trial discussed the 1980 undercount estimates themselves, of which there were 12 sets derived from the Census Bureau's Post Enumeration Program (PEP) for the nation and for each of 66 geographic areas, and a 13th set derived by demographic analysis and available only at the national level. Also discussed was the applicability of statistical loss functions to the census adjustment problem and the precise meaning of the term "better than the census."

Plaintiffs' rebuttal, phase 2 of the trial, commenced in February 1984 and expert testimony was offered by Eugene Ericksen, Franklin Fisher, and Joseph Kadane. The regression analysis discussed by Freedman and Navidi was first presented during this phase of the trial, with the plaintiffs' experts asserting that the models were successful in removing problems from the PEP data and that the resulting regression or Bayes estimates were more accurate than the census, and thus should replace the census. The proposal, in brief, was to replace the census population count of each small area, j, by

$$\text{Adjusted Census}_j = \frac{\text{Census}_j}{1 - \hat{y}_j/100},$$

where $\hat{y}_j = \hat{a} + \hat{b}\ \text{min}_j + \hat{c}\ \text{crime}_j + \hat{d}\ \text{conv}_j$ is a prediction of the percentage of undercount in area j; the generalized least squares estimator (GLS) $\hat{\beta}$ of $\beta = (a, b, c, d)$ is given in equation (12); min_j, crime_j, and conv_j are values of the predictor variables specific to the small area j; and the data used in obtaining $\hat{\beta}$ are at the 66-point level of aggregation.

Kirk M. Wolter is Chief of the Statistical Research Division, U. S. Bureau of the Census, Room 3536, FB #3, Washington, D. C. 20233.

No proposal was put forth to adjust the census tabulations regarding characteristics of the population or of the nation's housing stock.

Defendants' surrebuttal, phase 3 of the trial, involved expert testimony by David Freedman, Gary Koch, and Kirk Wolter. This occurred in May 1984 and involved a critique of the modeling results presented by plaintiffs on rebuttal. The paper by Freedman and Navidi, with few exceptions, is a summary of Freedman's testimony during this phase of the trial.

During the trial, I basically agreed with Freedman's conclusions about the 1980 data and about the application of regression methods to that body of data, and I still agree. I am, perhaps, more optimistic than the authors about the prospect for successful applicationof such methods to 1990 Census data. In what follows I explain both the 1980 and 1990 situations as I understand them.

1. The 1980 Census has not been adjusted for the measured undercount because the undercount estimates themselves are subject to error. I believe this error is larger than the degree of error in the census itself, and, therefore, conclude that an adjusted census would be less accurate than the original unadjusted census. For the new reader, it is worth noting that the Census Bureau decided not to adjust, and I participated in that decision, in advance of the regression modeling work presented at trial. The conclusions of Freedman and Navidi about the regression models corroborate the Census Bureau's prior findings. Put simply, the decision *not* to adjust was *not* made in response to the regression results, yet the results confirm the decision.

2. The levels of missing data encountered in the Post Enumeration Program (PEP) are much higher than the supposed level of undercount. In the April Current Population Survey (CPS) sample, 8.6% of the cases could not be assigned an enumeration status either because of nonresponse or because after considerable checking and follow-up it could not be determined whether or not they were counted in the census. The comparable level of missingness in the August CPS sample was 9.7%. At the local level these problems were more acute than at the national level. For example, in New York City and Houston there was considerable nonresponse in the April CPS, and even for those who did respond, 41.5 and 64.6%, respectively, of the total nonmatches were imputed as such rather than determined clerically. There was also a significant degree of missingness in the E sample, which was designed to measure overenumerations. Given that the true net coverage error may be in a range between a slight overcount to a 1 or 2% undercount, I conclude that the resolution of the measuring instrument, namely the PEP, is not adequate to permit a precise measurement of the net coverage error. Because of the considerable uncertainties associated with the missing data, the Census Bureau produced 12 sets of estimates of net coverage error, each relying implicitly on a different model of the missingness mechanism. The large observed variation among the 12 sets confirms the imprecision of the measuring instrument, and leads me to further conclude that none of the sets is likely to be more accurate than the original census.

3. In a related problem, the Census Bureau also encountered enormous difficulties in the clerical matching operation of the CPS to the census. False nonmatches tended to be the primary problem, thus tending, in the absence of other errors, to bias upward the estimators of net coverage error.

4. Do plaintiffs' regressions help? Do they diminish some of the problems with the PEP estimates and offer a good means of adjusting the census? Freedman and Navidi conclude not. And so do I. There are two statistical questions here:

a. Do the seven assumptions (1)–(7) obtain or not in this application?

b. Even if the seven assumptions fail to some degree, are the models useful in the sense that the resulting adjusted values are closer to the truth than the original census values?

The analysis by Freedman and Navidi will convince most statisticians that the seven regression assumptions fail in important ways. My own analyses of these and other data, which in many respects parallel the analyses of Freedman and Navidi, likewise suggest important failures of key assumptions. I also believe the failures occur to an *excessive degree* and, thus, I believe any adjusted values prepared as a result of these analyses will be further from the truth than the census values. The analyses by Freedman and Navidi will convince many statisticians of the wisdom of this nonadjustment position, but obviously not all will be convinced. All should remember, however, that issue b is fundamentally intractable because no one knows the true population counts. Thus, to some extent at least, issue b must be decided by professional judgment after careful analysis of the available evidence.

5. While I agree generally with the analysis of the seven assumptions by Freedman and Navidi, I have the following additional comments, numbered to correspond to the assumptions.

Assumption 2. At the time of the proposed adjustment, the γ_i are fixed. That is, there is a fixed true undercount for each area i. On the other hand, the use of the linear model for γ_i implies that the user is willing to entertain the notion of a conceptual population of all possible ways the 1980 Census could have come out but didn't. This concept will need to be explained to and accepted by the Congress, governors, mayors, and other nontechnical people. This will be challenging, but should not in itself defeat the question of adjustment. Freedman and Navidi may go too far when they assert that "... randomness in the census is too complex to model." They are right that the randomness is complex, but we do model it, rightly or wrongly. The model is a type of capture–recapture model, and given the model, the census count $\hat{N}$ is stochastic with mean N_p and variance $N_p(1 - p)$, where p is the probability of capture by the census.

Assumption 3. There are very important biases in the PEP data y_i. Thus, $E\delta_i \neq 0$. Furthermore, there is no reason to suppose that min_i, crime_i, and conv_i are the only independent variables affecting the true undercount γ_i, nor even to suppose that they are the most important variables. Thus, $E\epsilon_i \neq 0$. The comments on this assumption by Freedman and Navidi are particularly poignant.

Assumption 5. Granting (2) and (3), it seems to me that the variance of ϵ_i must generally increase with increasing values of the independent variables. This is so in most regression models involving survey data.

Assumption 6. The error terms are not independent, as is well explained by Freedman and Navidi. But failure of this assumption, by itself, is unlikely to bias the regression predictions. Estimated variances will be too small, however.

Assumption 7. I believe that y_i is asymptotically normal under a wide range of realistic conditions. I am uncertain, however, about whether the approach to normality is sufficiently rapid to validate this assumption for PEP data.

6. More on the troublesome assumption 3. The bias in PEP can be expressed in two parts: that which is in the column space of $X = (\text{min, crime, conv})$ and that which is orthogonal to this space. The first part biases both the regression and the Bayes predictions, while the second part biases only the Bayes predictions. In addition, the bias in ϵ_i due to important omitted variables is in two parts: that which is in the column space of X and that which is not. The former biases the estimated coefficients while the latter biases the predictions of the percentage of undercount. Freedman and Navidi explain this clearly.

7. Bias in ϵ_i is difficult to prove or disprove on the basis of the 1980 data. Freedman and Navidi replaced "crime" by "urbanization" and saw little difference in the quality of fit. Similarly, I replace "crime" by "mobility" (i.e., percentage of persons living in the same location as 5 years earlier) and saw little difference in the quality of fit. The GLS equation is

$$\begin{aligned} \text{PEP } 5/9 = {} & 8.101 + 0.081 \text{ min} \\ & + .015 \text{ conv} - .141 \text{ mob} + \text{error}. \end{aligned}$$

So is "mobility" or "urbanization" an important omitted variable? Are there others? And, it potentially makes an important difference in the predicted undercount percentages. For example, the predicted undercount percentages (and standard errors) for New York are

Area	Equation containing crime	Equation containing mobility
New York State	2.54 (.239)	1.30 (.235)
New York City	5.19 (.425)	3.23 (.368)
Balance of New York	0.67 (.288)	−0.005 (.240)

Results for smaller areas of the nation often show even larger differences between the two equations.

8. It is relatively easy to establish the existence of the bias component in δ_i which is in the column space of X. Freedman and Navidi present the results of regressing the difference (3/8 minus 5/8) on X. I produced many similar regressions, and they tend generally to support the conclusions of Freedman and Navidi. In addition, I produced a number of surrogates for matching bias and regressed each on X. Freedman and Navidi give the results for the case where the surrogate is "percentage unresolved." An additional example is where the surrogate is "percentage of total matches achieved after follow-up (%FUM)." For the April data, by ordinary least squares (OLS),

$$\begin{aligned} \%\text{FUM} = {} & \underset{(1.19)}{13.4} + \underset{(.031)}{0.67} \text{ min} \\ & - \underset{(.021)}{.084} \text{ crime} + \underset{(.017)}{.020} \text{ conv} + \text{error}. \end{aligned}$$

These results provide confirmation that bias in PEP is related to X and thus that PEP biases in turn bias the regression and Bayes predictions.

9. For those of us who have actual experience with the implementation of undercount studies, it is almost axiomatic that the greatest difficulties in measuring the undercount are encountered in the same areas where the census itself encounters great difficulties. The results referred to in 8 confirm this axiom.

10. Now for some additional points. Although the regression analysis does not filter out the PEP biases, it may tend to reduce the effects of sampling error (and random, nonsampling errors) on the PEP estimates. At least this is so in an average sense. Thus, regression models can be thought of as a kind of smoothing device. In the case of 1980, the models calibrate the PEP, not the truth, and they smooth the random effects.

11. There are measurement errors in the independent variables, particularly in crime. Such errors bias the estimated coefficients. Further, important biases will occur in local area adjustments where the distri-

bution of measurement error will be different than at the aggregate level at which the model is fit.

12. A potentially important problem that Freedman and Navidi mention but do not stress adequately is the so-called Simpson's paradox. The model fit to the 66-point data set may or may not bear any relationship to the model that applies to local areas. I would be substantially more satisfied with a model where the unit of analysis is more nearly the same as the local area units receiving the adjustment. Achieving this was difficult at best for the 1980 data given the design of the PEP. Also, in the context of a proposed 1980 adjustment, local area adjustments would force one to extrapolate the regression model far outside of the data set used in fitting the model.

13. Another important point concerns average error or average expected loss. Let I denote a set of areas to be adjusted, indexed by i. Then, one possible measure of average loss, although by no means the only measure, is

$$\sum_{i \in I} | E_i - P_i | / \#\{I\},$$

where P_i denotes true population for area i and E_i denotes an estimate of P_i. One estimate E_i is the original census for area i and another would be the adjusted census. Granting assumptions (1)–(7), it will be necessary to adopt such ideas of average loss in order to conclude that adjusted values are better than census values. Looking at each area individually almost always recommends against adjustment, because no model can guarantee improvement for each and every local area in the country.

14. So where does this leave us in terms of the future? It seems to me that models of the kind discussed here may indeed be useful for adjusting the 1990 Census. It also seems to me that the critical work involves the direct undercount estimators y_i themselves. If the undercount programs for 1990 can be conducted with little nonsampling error (particularly bias) relative to the presumed size of the undercount, then there is a good likelihood that models of the kind discussed here will be acceptable and adjusted population counts derived from such models will be closer to the truth than the unadjusted census counts. In addition to this critical work, we must carefully check the remaining assumptions and modify the methodology where necessary and use a unit of analysis that is comparable to the areas receiving adjustments.

15. A possible 1990 scenario would be to base the undercount program for 1990 on the concept of a Post Enumeration Survey (PES), designed expressly for coverage measurement purposes. According to standard survey principles, stratify the PES according to known characteristics of the population so as to achieve homogeneity (with respect to coverage error) within strata. Produce a capture–recapture estimate of the undercount for each stratum, say y_h for $h = 1, \cdots, L$. Regress the y_h on variables thought to be well correlated with undercount, probably on the same variables used in stratification (although perhaps updated by some 1990 short form information). Use the fitted equation to predict the undercount for each stratum, say $\hat{y}_h$ $(h = 1, \cdots, L)$. The $\hat{y}_h$ are smoothed versions of the y_h. Adjust the population counts at the block level, using $\hat{y}_h$ and applying it to the census counts of each block within stratum h $(h = 1, \cdots, L)$. If appropriate, do the last several points individually for several age–race–sex groups. Also adjust personal and housing characteristics in an acceptable manner, but give primary emphasis in the adjustment to the population counts.

By careful design and conduct of the PES, I believe the Census Bureau can achieve a situation where many of the assumptions (1)–(7) hold approximately. I also believe Simpson's paradox may be avoided because the fitted model is not necessarily used to predict values for smaller areas. A remaining problem is bias in ϵ_i that is orthogonal to the chosen independent variables. I am not sure we can solve this problem. But we may get closer to the truth even though we have not achieved an exact solution.

16. The Census Bureau's plans for 1990 include development of a set of *standards* for census adjustment. The standards will describe in a rather precise way the conditions under which an adjustment of the census will be undertaken. The standards will necessarily involve 1) consideration of what it means to be "more accurate" than the census, 2) consideration of the quality of the basic census enumeration, 3) consideration of the quality of the basic PES (or whatever other undercount programs are undertaken for 1990), and 4) consideration of the status of the assumptions underlying any models (regression or other) that might be used as part of the adjustment process.

The standards must explicitly recognize degrees of failure of the assumptions. I generally adhere to the philosophy that all models are wrong to one degree or another. The critical question, therefore, is whether the degree of failure is sufficiently minor that the model moves decisions closer to truth than in the absence of the model. In other words, the standards need to address the range of failures of assumptions under which the adjusted census would be considered more accurate than the unadjusted census.

In summary, it seems likely to me that regression models may be an integral part of the 1990 Census process, possibly according to the scenario in 15. If so, it seems certain that assumptions (1)–(7) and others like them will form the core of the standards for adjustment. The paper by Freedman and Navidi provides valuable early discussion on this important topic and contributes importantly to the continuing debate about census coverage error and the wisdom of census adjustment. Most statisticians should find their discussions informative, amusing, and provocative. I certainly did.

Comment

Albert Madansky

My comments on the Freedman–Navidi paper are of two sorts, one directed specifically to the content of the paper and the other a set of general remarks directed at the common theme of Freedman's recent papers (Freedman, 1981, 1985, Freedman et al., 1983, as well as this one), critiqueing the use of statistics in modeling.

1. COMMENTS ON THE FREEDMAN–NAVIDI PAPER

They describe the Post Enumeration Program (PEP) studies (Section 3) and point out that "about two dozen different series of PEP estimates were developed," each based on a different set of imputation rules for treating the missing data. They claim that the Bureau of the Census "was unwilling to use PEP to adjust the population counts" because 1) there was considerable variation across the series, 2) the probabilistic basis for the estimate was open to serious question, and 3) the standard errors of the estimate turned out to be quite large.

To impute these as the reasons for the "unwillingness" of the Bureau of the Census to use PEP to adjust the population counts lends a greater aura of ratiocinativity to that decision than actually was the case. In truth, the Bureau of the Census was unwilling *by any means* to adjust the population count, and never considered in a constructive way how one might use PEP to adjust the population counts. The Bureau of Census stance was more in the nature of "we don't want to adjust the raw census counts" and "even if we wanted to adjust, we don't know *how* to adjust using PEP data" than in the nature of the authors' imputed scenario, namely an implied willingness to adjust, recognition that methodology was available for effecting that adjustment, but, taking the view that "the PEP data are so problematical that we don't want to use them to adjust the raw census," and rationally deciding not to embark on an adjustment program. Indeed, Mitroff et al. (1982, 1983; see also Kadane, 1984) indicate that the principal motivating factor for the Bureau of the Census decision not to adjust was that the Bureau has historically been "nonpolitical and objective" and that use of any adjustment procedure would be in violation of that standard of Bureau behavior.

The positive contribution made by Ericksen and Kadane was to set forth an approach by which the PEP data could be used to adjust the census. Their paper merely suggested an approach toward adjustment; the work they did to implement their approach was in the nature of a constructive proof of an "existence theorem," used in an advocacy proceeding partially for the numbers it produced but primarily to make the point that indeed adjustment was feasible with the data at hand.

But let us get to the substance of the Freedman–Navidi paper. What should one make of the three "warts" in the PEP data? That the standard error of the PEP estimates turned out to be high is no reason not to use them if, in combination with the raw census data, one can produce demonstrably better estimates of the population counts than those achievable by using merely the raw census data. Let us see by a quick calculation whether this is in fact potentially the case.

The essence of the procedure for estimating the population count using the results of a postcensus sample (e.g., PEP) can be seen from a consideration of the following:

	Census	Sample
Respondents	n	n'
Nonrespondents	m	m'
Total	N	N'

Here N is the true census count, n is the observed census count, N' is the postcensus sample size, n' is the number in the postcensus sample who were also in the census, and $m' = N' - n'$ is the number in the postcensus sample who were not counted in the census. Now let $\theta = m/N$, the fraction undercount in the

Albert Madansky is Professor of Business Administration and Associate Dean for Ph.D. Studies, Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, IL 60637.