Statistical Science
1986, Vol. 1, No. 1, 3–39

Regression Models for Adjusting the 1980 Census

D. A. Freedman and W. C. Navidi

Abstract. After the 1980 Census, New York State sued to compel the Bureau of the Census to adjust the population counts, using a regression model. The appropriateness of such models is considered in this paper.

Keywords and phrases: Modeling, regression, hierarchical Bayesian regression.

1. INTRODUCTION

Models are often used to decide issues in situations marked by uncertainty. However, statistical inferences from data depend on assumptions about the processes which generated those data. If the assumptions do not hold, the inferences may not be reliable either. This limitation is often ignored by applied workers who fail to identify crucial assumptions or subject them to any kind of empirical testing. In such circumstances, using statistical procedures may only compound the uncertainty. To paraphrase Freedman, Rothenberg, and Sutch (1983):

> It ain't what you don't know that gets you into trouble, it's what you think you know that ain't so.

Statistical modeling seems likely to increase the stock of things you think you know that ain't so. For this reason among others, we do not accept the proposition that statistical models are useful, even compared to nothing—unless the assumptions are made explicit and shown to be appropriate. Our object is to illustrate the general point by discussing an example.

2. THE CENSUS

Every 10th year since 1790, a census has been taken to count the people of the United States. The total population is determined, and even more important nowadays, so are subtotals for each of the 50 states, the 3,000-odd counties, and the 39,000 minor civil divisions. These subtotals are used to apportion Congress and to allocate entitlement funds, amounting in the early 1980s to about $100 billion a year. Thus, great interest has been attached to these counts.

Demographic analysis suggested that in the 1970 Census, about 2% of the total population was missed: the *undercount.* Also, there was some evidence to show that the undercount varied across areas, with rural areas and central cities having greater undercounts than suburbs. Similarly, the undercount was thought to vary across population groups: poor people, minorities, and illegal immigrants were considered the groups hardest to count.

As a result, the Bureau made intensive efforts to eliminate the undercount for the 1980 Census, especially in problem areas. There were extensive community outreach programs as well as flying squads to count people in pool halls, bus terminals, and flop houses. These efforts seem to have met with some success. Indeed, demographic analysis indicated that at the national level, there was an overcount of about ¼ of 1% of the legal population, although some of the illegal population was probably missed.

Even so, many local governments were dissatisfied with the results and sued to compel the Bureau to revise its counts. One such suit was filed by the State of New York (*Cuomo v. Baldrige, SDNY*) and turned out to involve a number of important statistical issues. (One author of the present paper appeared for the government in surrebuttal, as did Professor G. Koch. Expert statisticians for New York in rebuttal included Professors E. Ericksen, F. Fisher, and J. Kadane.)

3. THE PEP ESTIMATES

Much of the detailed information about the undercount in the 1980 Census came from the Post Enumeration Program (PEP). This involved two kinds of studies. The first attempted to estimate the overcov-

D. A. Freedman is Professor of Statistics and W. C. Navidi is a graduate student in the Department of Statistics, University of California, Berkeley, CA 94720. *This paper describes statistical work done in connection with a law suit (Cuomo v. Baldrige, SDNY). Freedman gave expert testimony on behalf of the defense.*

erage due to double counting, coding households to the wrong areas, and inclusion of fictitious persons (called "curbstone cases" in the jargon of the Bureau). It was based on the *E sample*, a probability sample of about 100,000 records drawn from the 1980 Census. To verify census information for persons in the E sample, interviewers were sent into the field to locate and interview each of the 100,000 people in the sample. (The check for double counting was done only within neighboring enumeration districts.)

For about 3% of the records in the E sample, it turned out to be impossible on the basis of census records and field interviews to determine whether or not the person had been correctly counted on Census Day. Additional information (not necessarily accurate) was available in many such cases from local post offices, and decisions had to be made whether or not to use this information.

The second kind of study attempted to estimate undercoverage using a capture-recapture model. In this model, people who are counted in the census are deemed captured. Then a probability sample of the population, called a *P sample*, is taken. The people in the P sample who were counted in the census are deemed recaptured. The percentage of people in the P sample who were not counted in the census is used (along with other information) to estimate the census undercount. This procedure assumes that being counted in the census and being counted in the P sample are in probabilistic terms independent events after stratification on suitable covariates.

Two P samples were used in this study. They were the April and August samples from the 1980 Current Population Survey (CPS). The CPS is a monthly survey done by the Bureau of the Census for the Bureau of Labor Statistics; in 1980, the sample size was about 150,000 persons. For more information on the CPS, see Freedman, Pisani, and Purves (1978, Chapter 22) or Bureau of the Census (1978).

An attempt was made to match each person in the P sample against the census to see if he or she had been counted. Those cases for which a match could not be made were followed up be sending an interviewer into the field to obtain additional information, for instance, an exact address or the correct spelling of a name. After follow-up most of the cases were declared to be matched or nonmatched to the census. However, for about 4% of the cases, match status could not be determined on the basis of records or field interview. Several sets of imputation rules were developed for handling the unresolved cases: for example, use the match status of the last resolved case with the same sex, race, age, and area of residence. Each set of rules led, of course, to a different undercount estimate. For an additional 4% of the cases, the CPS interview was not completed. Since the CPS is a panel study, information (not necessarily accurate) was available from previous interviews, and a decision had to be made whether or not to use that information.

About two dozen different series of PEP estimates were developed. Each series contained undercount estimates for each of 66 areas (states and large central cities). There was considerable variation across the series. About half the series were discarded for one reason or another, but even the remaining dozen were quite inconsistent with one another: for instance, the estimated undercount for New York City ranged from 1–8%. This large variation may seem surprising, but in a major area like a central city or a state, the undercount will only be a few percentage points. With around 10% of the data missing, the choice of imputation rules has a serious impact on the estimates, as does the decision on use of previous CPS interviews, or post office responses, or the choice between the two P samples.

There were other problems too. The probabilistic basis for the estimate, independence of being counted in the census and being counted in the CPS, was open to serious question: someone hard to find for the census may also be hard to find for the CPS; on the other hand, persons interviewed in the April CPS may have been less willing to participate in the census.

Since the P and E samples were drawn by probability methods, standard errors for the estimates could be computed by variants of the split-sample technique: these turned out to be quite large. For all these reasons, the Bureau of the Census was unwilling to use PEP to adjust the population counts.

4. THE REGRESSION ESTIMATES

The PEP estimates, as discussed in the previous section, were judged unreliable by the Bureau. In an attempt to improve the reliability, New York turned to the regression model described in Ericksen and Kadane (1985). In the large, the idea is to borrow strength by a kind of generalized averaging. The analysis involved 66 areas to be indexed by i. The study areas were of three types: states, like Alaska or Wyoming; central cities, like New York, Los Angeles, or Chicago; and states apart from their central cities, like New York State apart from New York City or California apart from Los Angeles, San Diego, and San Francisco. The regression model was applied to three of the PEP series, the principal one favored by New York's experts being PEP 2/9 (the 2 refers to the treatment of the P sample, the 9 to the E sample).

Let y_i denote the PEP estimate for undercount in area i, expressed as a percentage. Thus, $y_i = 3.12$ indicates an estimated undercount of 3.12% for area i, while $y_i = -1.79$ indicates an estimated overcount

of 1.79%. Let γ_i denote the true undercount in area i, expressed as a percentage; γ_i is not observable.

New York's experts began the analysis with eight explanatory variables and built a regression model using the subset of three that best fit the PEP estimates by ordinary least squares. These three variables were: min_i, the percentage of the population in area i who were black or Hispanic; crime_i, the crime rate in area i; and conv_i, the percentage of the population in area i who were *conventionally enumerated.* In certain counties (mostly rural), the Census Bureau enumerated the population conventionally, that is, in a face-to-face interview with an enumerator. In other counties, the population was enumerated unconventionally, by mail. Thus, for each county, the percentage conventionally enumerated was either 0% or 100%. Different states have different mixes of counties, so the percentage conventionally enumerated in a state ranged from 0–100%.

The model had two equations: the first says that PEP gave an unbiased estimate of the true undercount; the second, that the true undercounts are linearly related to the explanatory variables. To begin with, we will discuss these equations somewhat informally. The first says

$$\begin{pmatrix}\text{PEP estimate}\\ \text{for area } i\end{pmatrix} = \begin{pmatrix}\text{true undercount}\\ \text{in area } i\end{pmatrix} + \delta_i.$$

The second says

$$\text{true undercount in area } i = a + b\begin{pmatrix}\text{percent}\\ \text{minority}\\ \text{in area } i\end{pmatrix} + c\begin{pmatrix}\text{crime}\\ \text{rate}\\ \text{in area } i\end{pmatrix} + d\begin{pmatrix}\text{percent}\\ \text{conventionally}\\ \text{enumerated}\\ \text{in area } i\end{pmatrix} + \epsilon_i.$$

Informally, the assumptions on the disturbance terms δ and ϵ can be stated as follows: there are two boxes of tickets for each area, one representing the possible δs and the other the possible ϵs. These tickets follow the normal curve, with mean 0. For area i, the δ box has variance K_i, the split-sample variance estimate produced by the Bureau. All the ϵ boxes have the same variance σ^2. The δ_i is drawn at random from the δ box for area i; the ϵ_i, from the ϵ box.

The first equation says that the PEP estimated undercount for area i equals the true undercount for that area plus a draw made at random from the δ box for that area: PEP is unbiased and the errors in PEP are unrelated from area to area. The second equation says that the true undercount for an area equals the displayed linear function of the explanatory variables, but is driven off this expected value by a random error drawn from the ϵ box. These errors are unrelated from area to area and unrelated to the δs.

More formally, the model can be stated as follows:

$$y_i = \gamma_i + \delta_i, \tag{1}$$

$$\gamma_i = a + b\,\text{min}_i + c\,\text{crime}_i + d\,\text{conv}_i + \epsilon_i. \tag{2}$$

The assumptions on the disturbance terms are as follows:

$$E(\delta_i) = E(\epsilon_i) = 0, \tag{3}$$

$$\text{var } \delta_i = K_i, \tag{4}$$

$$\text{var } \epsilon_i = \sigma^2, \tag{5}$$

$$\delta_1, \delta_2, \cdots, \delta_{66}, \quad \epsilon_1, \epsilon_2, \cdots, \epsilon_{66} \text{ are independent,} \tag{6}$$

$$\delta_i \text{ and } \epsilon_i \text{ are normally distributed.} \tag{7}$$

New York's experts did not make these assumptions explicit, nor did they give any empirical foundation for them. We defer our critique of the assumptions to the next section, and explain here what was done with the model. The first objective was to find an estimate more accurate than PEP for $\gamma = (\gamma_1, \cdots, \gamma_{66})$, the 66×1 column vector of true undercounts. The suggested procedure was as follows: Let X be the 66×4 design matrix, whose ith row lists the explanatory variables $(1, \text{min}_i, \text{crime}_i, \text{conv}_i)$ for the ith area. Let $H = X(X^TX)^{-1}X^T$, the projection matrix onto the column space of X. Let $K = \text{cov } \delta$ be the covariance matrix of the column vector $\delta = (\delta_1, \cdots, \delta_{66})$, so $K_{ii} = K_i$ and $K_{ij} = 0$ for $i \neq j$. Let $\hat{\sigma}^2$ be an estimate of σ^2. (Some methods for computing $\hat{\sigma}^2$ are discussed in Section 7.)

Define the 66×66 matrix Γ as follows:

$$\Gamma^{-1} = K^{-1} + \hat{\sigma}^{-2}(I - H), \tag{8}$$

where I is the 66×66 identity matrix. New York proposed the estimator

$$\hat{\gamma} = \Gamma K^{-1} y, \tag{9}$$

where y is the 66×1 column vector of PEP estimates $(y_1, \cdots, y_{66})$.

This was justified on Bayesian grounds, as in Lindley and Smith (1972). If a, b, c, and d are given a diffuse prior, and if K and σ^2 are treated as known constants with $\hat{\sigma}^2 \equiv \sigma^2$, then the posterior distribution of γ is normal with mean $\hat{\gamma}$ and covariance matrix Γ. This covariance matrix has a frequentist interpretation: if the specification is correct, and if $\hat{\sigma}^2 \equiv \sigma^2$, it is easily seen that

$$\text{cov } \hat{\gamma} = \Gamma. \tag{10}$$

For a frequentist justification of $\hat{\gamma}$, see Lemma 7.3 below. The focus of the present paper is not the Bayesian-frequentist controversy, but the use of inappropriate statistical models no matter how they may be estimated.

The second objective was to estimate undercounts for subareas (e.g., counties) of the 66 areas in the study. To accomplish this, New York's experts proposed the following regression procedure: Combine equations (1) and (2) to get

$$(11)\quad y_i = a + b\,\mathrm{min}_i + c\,\mathrm{crime}_i + d\,\mathrm{conv}_i + \eta_i,$$

where $\eta_i = \delta_i + \epsilon_i$ has mean 0 and variance $K_i + \sigma^2$; the η_i are independent in i by assumption. Let $\beta = (a, b, c, d)$ be the column vector of parameters in (11). Then a generalized least squares (GLS) estimator of β is

$$(12)\qquad \hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y,$$

where $\Sigma = K + \hat{\sigma}^2 I$; implicitly, $\hat{\sigma}^2 \equiv \sigma^2$ is assumed. (The displayed $\hat{\beta}$ is the Bayes estimate too, with a diffuse prior on the parameters.) If $\hat{\sigma}^2 \equiv \sigma^2$, then

$$(13)\qquad \mathrm{cov}\,\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}.$$

The proposal was to estimate undercounts for subareas by substitution into (2): if j is a subarea with percentage minority min_j, crime rate crime_j, and percent conventionally enumerated conv_j, the proposed estimate for the undercount there is

$$(14)\qquad \hat{a} + \hat{b}\,\mathrm{min}_j + \hat{c}\,\mathrm{crime}_j + \hat{d}\,\mathrm{conv}_j.$$

The implications of this idea will be discussed below. (There were other methods suggested for dealing with subareas not so relevant to our theme.)

In effect, New York was offering an existence proof: there are ways to adjust the census to make it more accurate, starting from PEP and using regression. Equation (10) was crucial to the argument, since the standard errors were used by New York to gauge the accuracy of the regression estimates. In the law case, New York sued to compel the Bureau to adjust the counts (using regression or some better procedure if one could be found). We turn now to the validity of the regression procedure.

5. A CRITIQUE

Granting assumptions (1–7), New York did have a good way of adjusting the census. However, no evidence was presented to show the assumptions were true, and all seemed suspect. This section will discuss the assumptions from a theoretical point of view; empirical evidence will be presented in the next section.

Bias. Most persons in the CPS were counted in the census; only a small percentage were not. On the other hand, matching huge data files is a complex and error-prone process, and imputations for missing data leave plenty of room for error as well. Since the great majority of persons were counted in the census, there is a strong tendency for mistakes to inflate rather than deflate the estimated undercount. Thus, some bias in the PEP estimates is almost inevitable, compounded by dependence between capture and recapture in the census and the CPS. Indeed, it is the differences in bias which seem to cause such a spread in the estimates across the different PEP series. Bias in the PEP estimates that is well related to the explanatory variables is not removed by New York's adjustment method. Nor is bias in the ϵs of equation (2) orthogonal to those variables. Furthermore, the impact of bias is not measured in New York's standard error formula.

More technically, assume (1) and (2) but not (3). Instead, let $\mu = E(\delta)$ and $\nu = E(\epsilon)$ be the 66×1 vectors of means. As is easily seen (Lemma 7.1)

$$(15)\qquad \Gamma K^{-1}X = X.$$

So

$$(16)\quad \hat{\gamma} - \gamma = \Gamma K^{-1}y - \gamma = \Gamma K^{-1}\delta + (\Gamma K^{-1} - I)\epsilon,$$

$$(17)\qquad E\{\hat{\gamma} - \gamma\} = \Gamma K^{-1}\mu + (\Gamma K^{-1} - I)\nu.$$

If μ is in the column space of X, or has a relatively large projection into that space, $\hat{\gamma}$ will be badly biased; the parallel discussion of ν is omitted. In effect, New York was taking the position that bias in the census—the undercount—was well related to the three explanatory variables, but bias in PEP was not.

Subareas. New York's experts wanted to apply the model to subareas not in the study. However, this is almost a logical contradiction. For example, subdivide each of the 66 study areas into blocks of 100,000 people. Suppose (2) and the assumptions on ϵ hold for the blocks. We would get the analog of (2) for a state by averaging the equations for the blocks making up that state. The more populous states would have more blocks, and their ϵs will have smaller variances. In short, if the model holds for the blocks, it cannot hold for the states. This only sharpens the basic question: why do assumptions (1–7) hold for the 66 study areas? Compare Ericksen and Kadane (1985, pp. 104–105).

Omitted variables and measurement error. Suppose the variable x_i belongs in equation (2) and is omitted. One of New York's experts conceded that this could bias the coefficient estimates (and hence the adjustment process for subareas although this was not made explicit). On the other hand, this expert argued that $\hat{\gamma}$ would not be biased. To reach this conclusion, he made the assumption that the omitted variable was linearly related to the variables in the equation, plus a random error:

$$(18)\quad x_i = \alpha_0 + \alpha_1\mathrm{min}_i + \alpha_2\mathrm{crime}_i + \alpha_3\mathrm{conv}_i + \xi_i,$$

the ξ_i having mean 0, constant variance, and being independent of other errors.

His argument seems weak. Equation (18) is just as suspect as (2): in fact, it is exactly the same equation

with x_i in place of y_i on the left hand side. In short, (18) is just another assumption, with no foundation in theory or in fact. If that assumption is wrong, omitted variables can cause serious bias in $\hat{\gamma}$. And there is a long list of powerful-looking omitted variables, including for example: the percentages of elderly people, of women, of single-parent families; the rate of population change; response rates on other surveys or on census questionnaire items; geographical location; interactions. Similar considerations apply to the possibility of measurement error in the right hand side variables of (2): compare Ericksen and Kadane (1985, p. 104).

The standard errors. New York argued that $\hat{\gamma}$ was a good estimator for γ because the standard errors were small. Indeed, treating K and $\hat{\sigma}^2$ as constant, ϵ and δ as independent,

$$\text{(19)} \quad \begin{aligned} \text{cov } \hat{\gamma} &= \Gamma K^{-1}(\text{cov } \delta)K^{-1}\Gamma \\ &\quad + (I - \Gamma K^{-1})(\text{cov } \epsilon)(I - K^{-1}\Gamma). \end{aligned}$$

If further cov $\delta = K$ and cov $\epsilon = \sigma^2 I$ and $\hat{\sigma}^2 \equiv \sigma^2$, then

$$\text{(20)} \quad \text{cov } \hat{\gamma} = \Gamma K^{-1}\Gamma + \sigma^2(I - \Gamma K^{-1})(I - K^{-1}\Gamma) = \Gamma.$$

(The second equality can be verified on multiplying by Γ^{-1} and simplifying: Lemma 7.2.)

On the other hand, the standard errors do not measure the impact of bias in PEP. Furthermore, the computations really ride on the assumptions. If ϵ and δ are dependent, or cov $\delta \neq K$, or cov $\epsilon \neq \sigma^2 I$, then cov $\hat{\gamma} \neq \Gamma$, and printing out the diagonal elements of Γ tells us very little about the size of the random errors in $\hat{\gamma}$, even leaving bias aside. Just to indicate possibilities: the Census Bureau has three data processing centers and twelve regional offices servicing different parts of the country. Mistakes in administrative procedures or in data processing (including, for instance, spilling coffee on a box of forms) will inevitably affect more than one area. Likewise, random events that affect the census or the CPS in one area seem very likely to affect adjacent areas: for example, a snowstorm in the northern Rocky Mountains in April could easily affect the CPS in Idaho, Montana, and Wyoming. Finally, specification errors in (2), e.g., omitted variables or region-specific coefficients, will cause correlation across areas. For such reasons, the independence assumption (6) seems quite suspect.

New York's procedure for computing standard errors treats var $\delta_i = K_i$ as known and fixed. This is contrary to fact. The Bureau estimates K_i by the split-sample technique, so that estimate is itself affected by sampling error. (In some of the 66 study areas, there were only a few hundred people in the CPS, of whom only a dozen or so were not matched to the census, so sampling error in K_i is nontrivial.) Too, the Bureau's technique for estimating K_i is known to miss some components of nonsampling error, and even some components of sampling error in smaller areas. Furthermore, New York's procedure for computing standard errors treats σ^2 as known. However, even granting assumptions (1–7), the parameter σ^2 must be estimated from the data, and that estimate is subject to appreciable random error, a component of variance missing from New York's formula.

In short, the stochastic assumptions in the model were far from logically inevitable. New York did not make these assumptions explicit, let alone justifying them or quantifying the impact of departures.

6. EMPIRICAL RESULTS

Recall that y is the 66×1 vector of PEP estimated undercounts, and $\Gamma = [K^{-1} + \hat{\sigma}^{-2}(I - H)]^{-1}$ is a 66×66 matrix. The estimator proposed by New York was $\hat{\gamma} = \Gamma K^{-1} y$: with this estimator, $\hat{\gamma}$ for each area is a linear combination of the PEP-estimated undercounts for all 66 areas. To make this a bit more vivid, Table 1 shows (for the PEP 2/9 series) a 10×10

TABLE 1

New York's estimator: $\hat{\gamma}$ for each area is a signed linear combination of the PEP estimates for all 66 areas, with the coefficients specified by ΓK^{-1}; coefficients for 10 areas are shown below

	Alabama	Alaska	SF	Fla	Idaho	NYC	NYS	ND	SD	Wyo
Alabama	.366	−.016	−.003	.000	−.004	.008	.034	.017	.004	−.006
Alaska	−.027	.333	.006	.019	.104	.016	−.046	.221	.187	.086
San Francisco	−.028	.036	.071	.105	.011	.066	−.001	−.110	−.062	−.010
Florida	.000	.008	.008	.429	.004	.027	.014	−.038	−.023	.000
Idaho	−.003	.040	.001	.003	.467	.000	.003	−.000	.084	.033
New York City	.016	.019	.013	.068	.001	.263	−.012	−.060	−.035	.007
New York State	.005	−.004	−.000	.003	.001	−.001	.782	.001	−.001	−.003
North Dakota	.002	.015	−.001	−.006	.019	−.004	.001	.859	.042	.014
South Dakota	.002	.039	−.002	−.011	.046	−.006	−.003	.127	.672	.035
Wyoming	−.013	.107	.002	.001	.106	.007	−.039	.255	.207	.273

Note: Alabama and Alaska are the first two areas in the study, Wyoming, the last. San Francisco was chosen as being of interest to us, while New York was the centerpiece of the case. Florida, Idaho, North Dakota, and South Dakota were chosen because they made heavy contribution to the other areas.

matrix of ΓK^{-1}. Take San Francisco, for example. To compute $\hat{\gamma}$ for that area, take −.028 times the PEP estimate for Alabama, +.036 times the PEP estimate for Alaska, and so forth, all the way through to −0.010 times the estimate for Wyoming. The algebraic sum of these 66 products is $\hat{\gamma}$ for San Francisco, and the corresponding row of Table 1 shows 10 of the 66 coefficients.

As Table 1 shows, San Francisco contributes relatively little to its own estimate (its coefficient is .071); Florida contributes rather more (.105); and North Dakota takes away (−.110). The numbers for Wyoming caught the Court's attention: North Dakota and South Dakota contribute fairly heavily to Wyoming, but get very little in return. To sum up, New York's procedure does not have much intuitive appeal. A strong theory is needed to establish relationships among the PEP estimates for the 66 areas, in order to derive the coefficients reported in Table 1 and the standard errors of the resulting estimates.

In the previous section, we argued on *a priori* grounds that New York's theory was quite weak; this section will present some empirical evidence. The first point is that ΓK^{-1} preserves the column space of X, so any bias in y that is linearly related to X will not be corrected by New York's procedure. On this score, there was general agreement among the experts. Does such bias exist, and if so, how much impact does it have on the estimates?

New York preferred the PEP 2/9 series; we wanted a comparison series that had not been so intensively studied, and chose the PEP 10/8 series because it was listed next to 2/9 on the computer printout we were given. (This may seem a bit cavalier, but the idea was exactly to prevent any accusation that we had picked an extreme series.)

Fitting the two series by GLS gives:

$$\begin{aligned}\text{PEP } 2/9 = {} & -3.09 + \underset{(.52)}{.058}\,\text{min} \\ & + \underset{(.010)}{.056}\,\text{crime} + \underset{(.006)}{.026}\,\text{conv} + \text{error},\end{aligned} \tag{21}$$

$$\begin{aligned}\text{PEP } 10/8 = {} & -2.23 + \underset{(.45)}{.022}\,\text{min} \\ & + \underset{(.009)}{.032}\,\text{crime} + \underset{(.005)}{.031}\,\text{conv} + \text{error}.\end{aligned} \tag{22}$$

The numbers appearing in parentheses below the estimated coefficients are the standard errors of those estimates. A preliminary ordinary least squares (OLS) fit is used to estimate σ^2; details are explained in the next section. There was little to distinguish the quality of fit between (21) and (22).

We then examined the difference between the two series. An OLS fit gives

$$\begin{aligned}&(\text{PEP } 10/8) - (\text{PEP } 2/9) \\ &\quad = 1.79 - \underset{(.56)}{.010}\,\text{min} - \underset{(.010)}{.044}\,\text{crime} \\ &\qquad + \underset{(.01)}{.14}\,\text{conv} + \text{error},\end{aligned} \tag{23}$$

$$R^2 = .4, \qquad F = 14, \qquad P < 5 * 10^{-7}.$$

The difference between the two series is therefore well related to New York's explanatory variables. We conclude that at least one of the two series is biased, and the bias is related to X.

We then ran New York's estimation process, starting from 10/8 rather than 2/9. The estimates turned out to be quite different and on the whole the results from 10/8 had smaller standard errors (see Table 2).

Of course, if 2/9 were unbiased and all the bias were in 10/8, Table 2 would be irrelevant to New York's case. However, it is hard to see why that should be so. From this perspective, our choice of 10/8 was lucky, because the two series may be compared as follows: 2/9 was based on the April CPS, 10/8 on August. Using the August survey may reduce the dependence of CPS and census, but increase the errors caused by people moving between Census Day and the CPS interview. There were minor differences in handling certain kinds of nonresponse (use of post office data or prior CPS interviews), but the same imputation rules after follow-up were used in the two series. All in all, it is by no means obvious which series is better, if either. Even more to the point, it is hard to see which series is more likely to obey the assumptions of the model.

TABLE 2
New York's estimator starting from PEP 2/9 or PEP 10/8

Area	Estimate		Standard error	
	2/9	10/8	2/9	10/8
Alabama	.78	−.10	.55	.24
Alaska	3.30	2.96	.69	.45
Los Angeles	5.56	1.90	.65	.33
San Diego	2.61	.86	.64	.20
San Francisco	4.35	1.70	.75	.38
Rest of California	3.10	.76	.44	.16
Chicago	3.96	1.04	.72	.43
Rest of Illinois	1.01	−.53	.42	.14
New York City	5.57	1.94	.65	.33
Rest of New York	−1.12	−.53	.31	.14
Wyoming	2.88	2.58	.69	.42

Note: Los Angeles, Chicago, and New York are the largest cities in the U.S. We elected to fill out the table with the other cities in California and the rest of that state and of Illinois.

Of course, direct evidence on bias in 2/9 is not available, because the true undercounts are unknown. However, the Bureau did have good evidence to show that the strengths of various likely sources of bias were well related to X. For example, let imp_i be the percentage of cases in area i with imputed CPS-census match status. For April series (including PEP 2/9), by OLS,

$$\begin{aligned}\text{imp} = {} & 1.40 + .047\ \text{min} \\ & (.61) \quad (.016) \\ & + 0.24\ \text{crime} + .014\ \text{conv} + \text{error}, \\ & \quad (.011) \qquad (.0085)\end{aligned} \tag{24}$$

$$R^2 = .4, \qquad F = 12, \qquad P < 2 * 10^{-6}.$$

Imputation is clearly a potential source of bias, and the amount of imputation by area is well related to the explanatory variables chosen by New York. This completes our discussion of bias.

New York's choice of variables and of functional form seemed quite arbitrary. Too, New York seemed to ignore the problems created by measurement error in the explanatory variables. Since crime rate statistics are notoriously unreliable, it occurred to us to run New York's adjustment process with crime rate replaced by *urbanization*, that is, the percentage of the population in an area living in urban communities. (There is strong opinion in the Bureau that urban and rural areas present very different kinds of enumeration problems.) The equation fitted by GLS is

$$\begin{aligned}\text{PEP } 10/8 = {} & -2.27 + .031\ \text{min} + .025\ \text{urb} \\ & (.57) \quad (.011) \qquad (.009) \\ & + .031\ \text{conv} + \text{error}. \\ & \quad (.005)\end{aligned} \tag{25}$$

We saw little difference in quality of fit between this and (22). The adjustments and standard errors for certain areas are shown in Table 3. On the whole, replacing crime by urbanization seems to represent an improvement, in terms of reducing bias due to measurement error, and even making a marginal reduction in standard errors.

For the 66 areas in the study, the choice of variables has some impact on the adjustments, but not a major one since both sets of variables span essentially the same column space. On the other hand, when extrapolating to subareas, the choice of variables matters a lot. The point is illustrated in Table 4 for 14 counties. The first two are hypothetical: A is a suburb with a low crime rate; B is a rural high-crime area. The next 12 counties are real, and were chosen to match county A or B, with some variation in the conventionality variable. As will be seen, when it comes to subareas, the explanatory variables make quite a difference. There seems to us no rational ground for choosing crime rate over urbanization, and the choice has major political implications: the former variable helps high-crime areas, like central cities; the latter, low-crime areas like suburbs. This completes our discussion of the impact of New York's rather arbitrary choice of specification.

Finally, we discuss a simulation experiment which makes three points: i) The variables which belong in the equation cannot be identified from the data; ii) σ^2 cannot be reliably estimated from the data; iii) New York's standard errors are much too optimistic. The experiment takes the vector γ of true undercounts as fixed not random (more about this later); indeed, γ was taken equal to the PEP 10/8 series. The PEP estimates were simulated as

$$y^* = \gamma + \delta^*, \tag{26}$$

where the δ_i^* are independent normal variables with mean 0, and var $\delta_i^* = K_i$, the Bureau's split-sample variance for 10/8. In effect, this grants (1) and the

TABLE 3

New York's estimator based on PEP 10/8: *crime rate vs. urbanization*

Area	Estimate		Standard error	
	Crime	Urb	Crime	Urb
Alabama	−.10	.02	.24	.24
Alaska	2.96	2.59	.45	.42
Los Angeles	1.90	1.56	.33	.30
San Diego	.86	.93	.20	.24
San Francisco	1.70	.97	.38	.24
Rest of California	.76	.80	.16	.18
Chicago	1.04	1.85	.43	.37
Rest of Illinois	−.53	−.11	.14	.15
New York City	1.94	1.55	.33	.30
Rest of New York	−.53	−.19	.14	.14
Wyoming	2.58	2.55	.42	.42

TABLE 4

New York's estimation process applied to subareas, starting from PEP 10/8, *using crime rate or urbanization*

County	Data				Estimate	
	Min	Conv	Crime	Urb	Crime	Urb
Hypothetical	0	0	0.0	100	−2.2	.2
Hypothetical	0	0	100.0	0	1.0	−2.3
Treutlen, GA	32	0	1.5	49	−1.5	−.1
St. Bernard, LA	12	0	3.6	96	−1.9	.5
Chickasaw, MS	36	0	4.2	40	−1.3	−.2
Nuckolls, NE	0	100	3.3	36	1.0	1.7
Pierce, ND	0	100	4.9	54	1.0	2.2
Tripp, SD	0	100	4.6	48	1.0	2.0
Alpine, CA	4	100	252.0	0	9.0	1.0
Giltin, CO	5	100	102.0	0	4.2	1.0
Summit, CO	2	100	171.0	0	6.4	.9
Lake, MI	18	0	105.0	0	1.5	−1.7
Menominee, WI	0	100	153.0	0	5.8	.8

assumptions on δ, but takes an agnostic position on (2) and ϵ: the latter set of assumptions holds in the simulation world to the extent that it does for 10/8 in the real world.

We then generated 100 artificial data sets from (26). For each data set, we followed New York's procedure of choosing the best subset of three variables out of the given eight by OLS, and then estimating $\sigma^2 = \text{var } \epsilon_i$ in (2). The results for the first ten data sets are shown in Table 5. It will be seen that there is no consistency in the choice of variables or the estimate of σ^2.

For each of the 100 data sets, we have so far chosen the three variables that define the design matrix X and have estimated σ^2. We then use (8), (9), and (10) to get an adjustment $\hat{\gamma}_i^*$ for each area, with nominal variance Γ_{ii}^*. The covariance matrix K for δ was fixed throughout the simulation; for each artificial data set, however, there was a separate estimate $\hat{\sigma}^*$ for σ, and a separate projection matrix H^* because the best three variables change from data set to data set: then $\Gamma^* = [K^{-1} + \hat{\sigma}^{*-2}(I - H^*)]^{-1}$, according to (8).

The real rms error for area i is $[E_*\{(\hat{\gamma}_i^* - \gamma_i)^2\}]^{1/2}$, estimated by taking the root mean square over the 100 replications; likewise, the nominal rms error is $[E_*\{\Gamma_{ii}^*\}]^{1/2}$. These quantities are shown in Table 6, along with $K_i^{1/2}$, the PEP SE. As will be seen, the nominal standard errors are sometimes quite misleading. Summary statistics for all 66 areas may therefore be of interest. Taking the root mean square over all 66 areas:

$$\text{Real rms error} = \sqrt{1/66 \sum_{i=1}^{66} E_*\{(\hat{\gamma}_i^* - \gamma_i)^2\}} = 1.17,$$

$$\text{Nominal rms error} = \sqrt{1/66 \sum_{i=1}^{66} E_*\{\Gamma_{ii}^*\}} = .82,$$

$$\text{rms PEP SE} = \sqrt{1/66 \sum_{i=1}^{66} K_i} = 1.59.$$

Thus, the nominal rms error is too small by about 40%. The reduction in error is exaggerated by a factor of $(1.59 - .82)/(1.59 - 1.17)$, which is nearly 2. (This bias seems to be due in part to the usual problems caused by variable selection; and in part to components of error not recognized by New York, such as bias or the variability in $\hat{\sigma}$.) Whatever improvement $\hat{\gamma}$ does make on y is contingent on the assumed independence of the δs and knowledge of their variances, both assumptions being open to serious question.

TABLE 5
A simulation experiment on variable selection and estimation of σ^2

Run	Min	Crime	Conv	cc	mu	pov	ed	eng-d	$\hat{\sigma}^2$
1	X		X			X			.61
2	X		X	X					.00
3			X	X	X				.23
4			X	X	X				1.20
5		X	X	X					1.64
6			X	X		X			1.46
7					X	X	X		3.44
8	X		X		X				3.65
9		X	X		X				.40
10	X		X		X				4.53

Note: cc is an indicator for central cities; mu is the percentage of multiple-unit housing; pov is the percentage below the poverty line; ed is the percentage with a high school degree; eng-d is the percentage who have difficulty speaking English.

TABLE 6
A simulation experiment on standard errors

	Real rms error	Nominal rms error	PEP SE
Alabama	.90	.69	.90
Alaska	1.77	1.26	2.80
Los Angeles	.65	.84	1.20
San Diego	1.92	1.13	2.62
San Francisco	1.61	.88	1.30
Rest of California	.49	.41	.45
Chicago	1.35	1.05	1.61
Rest of Illinois	.49	.55	.67
New York City	.71	.77	1.03
Rest of New York	.59	.67	.88
Wyoming	.73	.77	.96

To sum up, the simulation results indicate that even granting large parts of the regression model, the variables which drive the undercount cannot be reliably identified from the data; the impact of this on subareas has already been discussed. Furthermore, the nominal standard errors are on the whole much too small, in part because σ^2 is hard to estimate, and in part because the nominal standard errors ignore bias.

A comment may be in order on the randomness in γ. For area i, let N_i be the population on Census Day, and $\hat{N}_i$ the census estimate. Then $\gamma_i = (N_i - \hat{N}_i)/N_i \times 100\%$. From a Bayesian point of view, it may be proper to treat γ_i as unknown and therefore random. Even from the frequentist viewpoint, it may conceivably be proper to view the 1980 Census as a measuring device, subject to bias $X\beta$ and random error ϵ. New York's calculations then relate not only to the actual 1980 Census, but to all the other ways it could have turned out but did not. From our point of view, this seems a bit fanciful. We prefer to consider N_i as fixed and unknown, while $\hat{N}_i$ is fixed and known, so γ_i is not stochastic. This preference is partly because the census was in fact taken and did come out the way it came out. But the main reason is that the indeterminacy of the census numbers seems to us far too complicated to model in terms of random draws from a box. Our simulation was done conditionally on $\hat{N}_i$ because the randomness in the census is too complex to model.

7. TECHNICAL DETAILS

We begin by describing more carefully the procedure for estimating $\hat{\sigma}^2$. Consider the OLS regression of y

on X; let H be the projection matrix; the residual vector is $e = (I - H)(\epsilon + \delta)$. Using (3–6), it is easy to compute

$$E\{\| e \|^2\} = E\{\text{trace } ee^T\}$$
$$= \text{trace } (I - H)(K + \sigma^2)(I - H)$$
$$= [\text{trace } (I - H)K] + (n - 4)\sigma^2.$$

Now

$$\hat{\sigma}_0^2 = \frac{1}{n-4}\{\| e \|^2 - \text{trace } (I - H)K\}.$$

This estimator is the one used in the simulation experiments.

We also considered an iterative procedure like one suggested by New York. Starting from say σ_0^2 one forms $\hat{\Sigma}_0 = K + \hat{\sigma}_0^2 I$, and regresses $\hat{\Sigma}_0^{-1/2} y$ on $\hat{\Sigma}_0^{-1/2} X$; let λ_0 be the mean square for error; if λ_0 is close to 1, stop; if not, revise $\hat{\sigma}_0^2$ and iterate. Convergence seemed fairly rapid, but the procedure was hard to automate for the simulation. Tables 1–4 were computed this way with 51 study areas (the states and D.C.), since that was our understanding of New York procedure.

LEMMA 1. $\Gamma K^{-1} X = X$.

PROOF. Let $M = I + \hat{\sigma}^{-2} K(I - H)$. Thus $\Gamma K^{-1} = M^{-1}$. If v is in the column space of X, clearly $Mv = v$, so $\Gamma K^{-1} v = v$. □

LEMMA 2. *If* (1–6) *hold, and* $\hat{\sigma}^2 \equiv \sigma^2$, *then* $cov(\hat{\gamma} - \gamma) = \Gamma$.

PROOF. By Lemma 1, $\Gamma K^{-1} H = H$, so

$$\Gamma K^{-1} + (I - \Gamma K^{-1})(I - H) = I.$$

Multiply on the right by Γ and note

$$\Gamma - H\Gamma = \hat{\sigma}^2 (I - K^{-1}\Gamma)$$

because

$$\hat{\sigma}^{-2}(I - H) = \Gamma^{-1} - K^{-1}. \qquad \square$$

LEMMA 3. *If* (1–6) *holds, and* $\hat{\sigma}^2 \equiv \sigma^2$, *then* $\hat{\gamma}$ *is the minimum-variance unbiased linear estimate of* γ.

PROOF. This follows from Goldberger (1962). There is an unfortunate conflict of notation: the ϵ in his (2.1) is our $\epsilon + \delta$, so Ω in his (2.3) is $K + \sigma^2 I$. The y_* in his (2.4) is γ_i, and $\epsilon_* = \epsilon_i$: thus, his w is all 0s except for σ^2 in position i. His optimal predictor is $\hat{c}'y$, where $\hat{c}'$ is defined in his (3.11); to verify that $\hat{c}'$ is the ith row of ΓK^{-1} amounts to proving that

$$\Gamma K^{-1} = G + \sigma^2 \Sigma^{-1}(I - G), \tag{27}$$

where G is the GLS projection matrix

$$G = X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}. \tag{28}$$

This reduces to verifying that

$$K\Gamma^{-1} G + \sigma^2 K \Gamma^{-1} \Sigma^{-1}(I - G) = I. \tag{29}$$

Now $K\Gamma^{-1} X = X$ as in Lemma 1, so $K\Gamma^{-1} G = G$. One also verifies that

$$\sigma^2 K \Gamma^{-1} \Sigma^{-1} = I - KH\Sigma^{-1}. \tag{30}$$

Substitution into (29) reduces the problem to showing

$$KH\Sigma^{-1} G = KH\Sigma^{-1}. \tag{31}$$

This can be verified by substituting the definitions of G and H.

ACKNOWLEDGMENTS

We would like to thank M. L. Eaton and D. Lane for their help. Variations on the quote in Section 1 have been attributed to Mark Twain, Artemus Ward, and even Will Rogers (during the 1984 presidential debate).

REFERENCES

BUREAU OF THE CENSUS (1978). *The Current Population Survey: Design and Methodology.* Technical Paper No. 40, U. S. Dept. of Commerce, Washington, D.C.

ERICKSEN, E. P. and KADANE, J. B. (1985). Estimating the population in a census year: 1980 and beyond. *J. Amer. Statist. Assoc.* **80** 98–131.

FREEDMAN, D. and PETERS, S. (1984). Bootstrapping a regression equation: Some empirical results. *J. Business Economic Statist.* **2** 150–158.

FREEDMAN, D., PISANI, R. and PURVES, R. (1978). *Statistics.* Norton, New York.

FREEDMAN, D., ROTHENBERG, T. and SUTCH, R. (1983). On energy policy models. *J. Business Economic Statist.* **1** 24–36.

GOLDBERGER, A. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Amer. Statist. Assoc.* **57** 369–375.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc.* **67** 1–19.