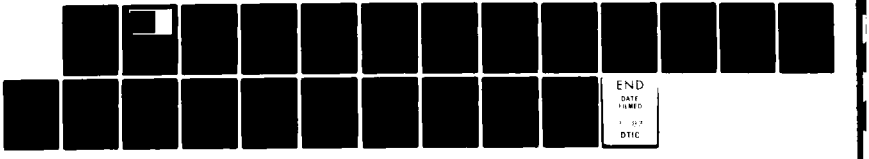
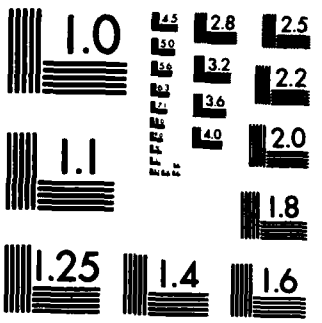
AD-A124 381 DIFFICULTIES WITH REGRESSION ANALYSIS OF AGE-ADJUSTED 1/1
RATES(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH
CENTER P R ROSENBAUM ET AL. SEP 82 MRC-TSR-2428
UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL





MRC Technical Summary Report #2428

DIFFICULTIES WITH REGRESSION ANALYSES OF AGE-ADJUSTED RATES

Paul R. Rosenbaum and
Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

September 1982

(Received April 27, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

DIFFICULTIES WITH REGRESSION ANALYSES OF AGE-ADJUSTED RATES

Paul R. Rosenbaum* and Donald B. Rubin**

Technical Summary Report #2428
September 1982

ABSTRACT

A common type of observational study compares population rates in several regions having differing policies in an effort to assess the effects of those policies. In many studies, particularly in public health and epidemiology, age-adjusted rates are regressed on predictor variables to obtain a covariance adjusted estimate of effect; we show that this estimate is generally biased for the appropriate regression coefficient. The analysis of crude rates with age as a covariate can, under familiar models, lead to unbiased estimates, and therefore can be preferrable. Several other regression methods are also considered.

AMS(MOS) Subject Classification: 62P99, 62J05, 62F12, 62F11

Key Words: Observational studies, covariance adjustments, regression analysis, direct adjustment, adjusted rates.

Work Unit Number 4 - Probability and Statistics

* Departments of Statistics and Human Oncology, University of Wisconsin-Madison.
**Mathematics Research Center, University of Wisconsin-Madison.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, and in part by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center.

DIFFICULTIES WITH REGRESSION ANALYSES OF AGE-ADJUSTED RATES

Paul R. Rosenbaum* and Donald B. Rubin**

1. Introduction: A Common Type of Observational Study

A common and inexpensive type of observational study uses previously collected population data, such as census data, to assess the effects of policies which are specific to certain counties, states or nations. An example is the comparison of motor vehicle mortality rates in states with and without required automobile inspection, (Fuchs and Leveson 1967; Colton and Buxbaum 1968). Note that in this example, all people living in the same state are subject to the same law.

A related though distinct type of observational study involves an exposure or treatment that is more prevalent in some states than in others: the relationship between the extent of exposure and the outcome is studied in an effort to assess the effects of exposure. Examples include (a) studies which examine site-specific cancer mortality rates in various counties and their relationship to environmental factors in these counties (e.g., Blair, Fraumeni, and Mason 1980) and (b) studies of the socioeconomic correlates of mortality (e.g., Kitagawa and Hauser 1973). Our discussion here is relevant to both types of studies, and demonstrates that standard analyses, such as those in the above references, are generally inappropriate. The problem arises because the

* Departments of Statistics and Human Oncology, University of Wisconsin-Madison.
**Mathematics Research Center, University of Wisconsin-Madison.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, and in part by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center.

outcome variables used in those analyses, such as death rates in various states, have been age adjusted, whereas the predictor variables have not been age adjusted. The use of crude state death rates as the outcome variable with crude covariates and age as predictors can avoid the problem, at least under some simple linear models. The use of age-specific rates as the outcome variable is generally inappropriate unless age-specific predictors are used.

2. A Motivating Simple Case: Age Adjustment By Regression

Suppose we wish to estimate the regression coefficient $\beta_{YX_1 \cdot X_2}$ of Y on X_1 in the multiple regression with two predictors, X_1 and X_2. It is well known that the least squares estimate of this coefficient may be found by, first, regressing Y on X_2 and calculating the residuals $Y \cdot X_2$, then regressing X_1 on X_2 and calculating the residuals $X_1 \cdot X_2$, and finally calculating the estimate of $\beta_{YX_1 \cdot X_2}$ as the estimated slope in the regression of the first set of residuals $Y \cdot X_2$ on the second $X_1 \cdot X_2$. An example is given by Mosteller and Tukey (1977, p.271); the formal argument is given by Seber (1977, p.65). This process of "sweeping out" one variable at a time forms the basis for several of the algorithms used for multiple regression, particularly the Gaussian pivoting in Beaton's sweep operator (Dempster 1969, p.62).

We can now give a rough description of the difficulty with the regression analysis of age-adjusted rates; the argument is formalized in the next section. Suppose that Y is an age and state specific mortality rate, that X_2 is the corresponding age, and that X_1 is any variable that varies with both age and state, say X_1 = per capita personal income. Roughly speaking, $Y \cdot X_2$ is the age-adjusted

mortality. To find the least squares estimate of $\beta_{YX_1 \cdot X_2}$ we should regress age-adjusted mortality $Y \cdot X_2$ on age-adjusted income $X_1 \cdot X_2$. However, that is not what is often mistakenly done; rather age-adjusted mortality $Y \cdot X_2$ is regressed on income X_1, giving a biased estimate unless income X_1 and age X_2 are orthogonal. The point is: if we adjust mortality for age, we must adjust the covariates for age as well.

Although age-adjusted mortality rates are commonly available, it is uncommon to find covariates such as income that have been age adjusted before tabulation. If the available data consist of adjusted mortality rates and unadjusted per capita income for each state, we cannot generally adjust income for age, and therefore cannot determine the partial regression coefficient of mortality on income adjusting for age.

An alternative solution would be to regress adjusted mortality $Y \cdot X_2$ on crude per capita income X_1 and crude age X_2, when the age information, X_2, is available. It is easily shown that the coefficient of income in this regression is the usual unbiased least squares estimate of $\beta_{YX_1 \cdot X_2}$. Unfortunately this procedure is not generally applicable to age-adjusted rates, for reasons described in §5 below.

3. Regression Analysis of Adjusted Rates

Let Y_{asi} be the response of the i^{th} person with age a in state s, for $i = 1,2,\ldots,n_{as}$. For purposes of this discussion, we assume the following linear model for Y_{asi} which includes polynomial terms in age:

$$E(Y_{asi}|D) = \alpha + \sum_{j=1}^{J} \beta_j a^j + \Delta z_{asi} + \underset{\sim}{\gamma}^T \underset{\sim}{x}_s + \underset{\sim}{\xi}^T \underset{\sim}{w}_{asi} \qquad (1)$$

for $i = 1,2,\ldots,n_{as}$, $a = 1,2,\ldots,A$ $s = 1,2,\ldots,S$,

where

Z_{asi} = 1 if individual i was exposed to the treatment and

0 otherwise,

$\underset{\sim}{X}_s$ is a vector of characteristics of state s (e.g., minimum driving age in the state)

$\underset{\sim}{W}_{asi}$ is a vector of characteristics of the individual (eg, income, marital status), excluding features of the state as a whole since these are included in $\underset{\sim}{X}_s$, but possibly including characteristics which are constant for all members of certain counties (e.g.; source of drinking water: city supplied vs. private well),

$\alpha, \beta_1, \beta_2, \ldots, \beta_J$, Δ, $\underset{\sim}{\gamma}$, $\underset{\sim}{\xi}$, are parameters, and D is short-hand for the age information and all the Z's, $\underset{\sim}{X}$'s and $\underset{\sim}{W}$'s. The polynomial in age can be replaced by other linear structures such as an indicator variable for each age or age category, a polynomial in the logarithm or exponential of age, or a combination of a polynomial in age and indicator variables for extreme age categories.

If Y_{asi} is binary, the linear logistic model (Cox 1970) is more attractive than the linear model for most purposes; however, the logit model does not lead to straightforward conclusions about the common practice of linearly regressing adjusted rates on predictors, nor would use of the logit model eliminate the problems that we describe which result from the use of age-adjusted rates.

The age and state specific mean response (or rate if Y_{asi} is binary) is $\bar{Y}_{as+} = \frac{1}{n_{as}} \sum_{i=1}^{n_{as}} Y_{asi}$. By (1), the expectation of $\bar{Y}_{as+}$ is

$$E(\bar{Y}_{as+}|D) = \alpha + \sum_{j=1}^{J} \beta_j a^j + \Delta\bar{Z}_{as+} + \underset{\sim}{\chi}^T \underset{\sim}{X}_s + \underset{\sim}{\xi}^T \bar{\underset{\sim}{W}}_{as+} \qquad (2)$$

where $\bar{Z}_{as+}$ and $\bar{\underset{\sim}{W}}_{as+}$ are averages of the Z_{asi} and the $\underset{\sim}{W}_{asi}$, respectively. Clearly, the parameters of model (1) may be estimated from a suitable weighted regression of the age and state specific rates $\bar{Y}_{as+}$ on the age and state specific averages in (2). For example, if the conditional variances given D of the Y_{asi}'s are all equal to a common value σ^2, and if the Y_{asi}'s are conditionaly uncorrelated, the appropriate weight for $\bar{Y}_{as+}$ in regression model (2) is n_{as}. Other choices for weights are described by Pocock, Cook and Beresford (1981).

Now consider the crude unadjusted rates for state s, namely $\bar{Y}_{+s+} = (\sum_a n_{as}\bar{Y}_{as+})/(\sum_a n_{as})$, with expectations

$$E(\bar{Y}_{+s+}|D) = \alpha + \sum_{j=1}^{J} \beta_j m_{sj} + \Delta\bar{Z}_{+s+} + \underset{\sim}{\chi}^T \underset{\sim}{X}_s + \underset{\sim}{\xi}^T \bar{\underset{\sim}{W}}_{+s+} \qquad (3)$$

where $\bar{Z}_{+s+}$ and $\bar{\underset{\sim}{W}}_{+s+}$ are averages of Z_{asi}, $\underset{\sim}{W}_{asi}$ over all individuals in state s, and $m_{sj} = (\sum_a n_{as}a^j)/(\sum_a n_{as})$ is the j^{th} moment of age in state s. If the first J moments of the age distribution are available from each state, then the parameters of model (1) may be estimated by a suitable weighted regression of the crude rates $\bar{Y}_{+s+}$ on the crude predictors $(m_{sj}, j=1,\ldots, J; \bar{Z}_{+s+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{+s+})$ for the states. For example, under the simple assumption of the previous paragraph, the weight for $\bar{Y}_{+s+}$ would be the population of state s, namely $\sum_a n_{as}$.

In practice, the moments m_{sj} of age distributions may be approximated from frequency tabulations of age distributions for each state, using, for example, the EM algorithm of Dempster, Laird and Rubin (1977) to correct for grouping. If a linear structure other than a polynomial is used for age in (1), then the corresponding averages would appear in (3). For example, if indicators are used for each age category, then the proportion of individuals in each age category in each state, $\bar{p}_{as} = n_{as}/\sum_a n_{as}$, would appear in (3).

Now consider the age-adjusted rates

$$\tilde{Y}_{+s+} = \sum_a f_a \bar{Y}_{as+}$$

where f_a is the fraction of the reference population with age a. Note that the same weights f_a are applied in all states. For example, the total population age distribution might be used as weights, so that $f_a = n_{a+}/n_{++}$. Now,

$$\begin{aligned} E(\tilde{Y}_{+s+}|D) &= \alpha + \sum_{j=1}^{J} \beta_j \sum_a f_a a^j + \Delta \sum_a f_a \bar{Z}_{as+} + \underset{\sim}{\gamma}^T \underset{\sim}{X}_s + \underset{\sim}{\xi}^T \sum_a f_a \underset{\sim}{\bar{W}}_{as+} \\ &= \alpha + \sum_{j=1}^{J} \beta_j \tilde{m}_j + \Delta \tilde{Z}_{+s+} + \underset{\sim}{\gamma}^T \underset{\sim}{X}_s + \underset{\sim}{\xi}^T \underset{\sim}{\tilde{W}}_{+s+} \\ &= \tilde{\alpha} + \Delta \tilde{Z}_{+s+} + \underset{\sim}{\gamma}^T \underset{\sim}{X}_s + \underset{\sim}{\xi}^T \underset{\sim}{\tilde{W}}_{+s+} \end{aligned} \qquad (4)$$

say, where $\tilde{m}_j$ is the j^{th} moment of age in the reference population, and $\tilde{Z}_{+s+}$ and $\underset{\sim}{\tilde{W}}_{+s+}$ are the <u>age-adjusted averages of Z and W for state s</u>. Note that the constant $\tilde{\alpha}$ includes the age component, $\Sigma\beta_j\tilde{m}_j$, which is the same for all states; this would be true no matter what linear structure is assumed in (1) for the regression on age.

Equation (4) formally describes the difficulty, mentioned in the last section, that is encountered when age-adjusted rates are regressed on predictors. To estimate the parameters of the model (1), we must regress the adjusted rates $\tilde{Y}_s$ on the age-adjusted treatment indicator $\tilde{Z}_{+s+}$, the age-adjusted covariates $\tilde{\underset{\sim}{W}}_{+s+}$, and $\underset{\sim}{X}_s$. Note that there is no difficulty when both (a) treatment, Z_{asi}, is constant within a state, as is the case when Z represents a state law, and (b) the only covariates involved are the descriptors $\underset{\sim}{X}_s$ of the state as a whole, such as other state laws or policies. However, there is a difficulty if there are covariates $\underset{\sim}{W}_{asi}$ such as personal income that describe individuals within a state, or when there are covariates such as pollution levels that describe areas within a state, because in such cases age-adjusted income or pollution levels are required to fit equation (4), and these quantities are rarely tabulated in official publications. Moreover, the difficulty also occurs if treatment, Z_{asi}, varies within a state, for in such cases, the age-adjusted rate $\tilde{Y}_{+s+}$ should be regressed on age-adjusted exposure $\tilde{Z}_{+s+}$.

Although age-specific death rates, $\bar{Y}_{as+}$, may be available, it is often difficult to obtain age-specific predictors $(\bar{Z}_{as+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{as+})$. As a result, another common practice is to regress age-specific rates $\bar{Y}_{as+}$ on crude predictors $(\bar{Z}_{+s+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{+s+})$. An example is a study of the association in 18 countries between wine consumption and cardiovascular mortality among men and women aged 55 to 64 (St. Leger, Cochrane, and Moore 1979). However, inspection of equation (2) shows that this procedure is generally inappropriate, unless the age-specific predictors $(\bar{Z}_{as+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{as+})$ equal the crude predictors $(\bar{Z}_{+s+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{+s+})$.

4. An Example

This section presents an example to illustrate the problem described in §3. The data used are a mixture of real and artificial data, because the true values for the age-adjusted covariates were not available, and we wished to dramatize possible effects. As a result, although the studies from which the data were drawn may have been affected by the problems we describe, our numerical results do not necessarily contradict the qualitative conclusions of those studies.

Table 1 contains (a) age-adjusted motor vehicle accident mortality rates ($\tilde{Y}_{+s+}$) for white males in 1960 for the 48 contiguous states of the United States, (b) a variable $\bar{Z}_{+s+}$ indicating whether the state requires motor vehicle inspections, (c) the percent of the state living in urban areas $\bar{W}_{+s+}$, and (d) the (artificial) age-adjusted percent of urbanization, $\tilde{W}_{+s+}$. Since the state law affects everyone in a state, the inspection indicator is not altered by age-adjustment; i.e., $\bar{Z}_{+s+} = \tilde{Z}_{+s+}$.

Presumably, an individual's risk of accident mortality (e.g., prob($Y_{asi} = 1$), say), depends less on the statewide degree of urbanization $\bar{W}_{+s+}$ than on whether the individual himself lives in an urbanized area (i.e., whether $W_{asi} = 1$, say). For example, an individual living outside Massena, New York, far from Manhattan, may be no more affected by the high percent of urbanization in New York State than are residents of, say, Vermont. If the age distributions in urban and rural areas differ, then $\bar{W}_{+s+}$ and $\tilde{W}_{+s+}$ will generally differ, generally leading to a biased estimate of the coefficient of automobile inspection $\bar{Z}_{+s+}$ when adjusted mortality $\tilde{Y}_{+s+}$ is regressed on $\bar{Z}_{+s+}$ and crude urbanization $\bar{W}_{+s+}$.

TABLE 1. Data For The Example: Mortality and Motor Vehicle Inspections

State	Age-adjusted Motor Vehicle Mortality* $\tilde{Y}_{+s+}$	Inspection State* (1 = yes) (0 = no) $\bar{z}_{+s+} = \tilde{z}_{+s+}$	Percent Urban** $\bar{w}_{+s+}$	Age-adjusted Percent Urban*** $\tilde{w}_{+s+}$
1	57.5	0.	26.7	26.7
2	57.7	0.	50.1	50.1
3	56.2	0.	13.4	13.4
4	47.7	0.	35.2	35.2
5	21.0	0.	34.4	34.4
6	40.9	0.	25.6	25.6
7	51.1	0.	24.1	24.1
8	52.6	0.	.0	.0
9	31.3	0.	42.1	42.1
10	47.7	0.	30.0	30.0
11	43.0	0.	22.0	22.0
12	44.6	0.	17.2	17.2
13	53.8	0.	16.0	16.0
14	49.0	0.	32.5	32.5
15	30.3	0.	30.3	30.3
16	40.7	0.	32.9	32.9
17	41.2	0.	27.1	27.1
18	66.5	0.	6.6	6.6
19	47.9	0.	32.4	32.4
20	62.4	0.	16.0	16.0
21	39.8	0.	30.5	30.5
22	95.3	0.	40.6	40.6
23	53.0	0.	16.0	16.0
24	55.5	0.	7.4	7.4
25	38.0	0.	35.2	35.2
26	49.9	0.	27.8	27.8
27	50.3	0.	24.0	24.0
28	55.4	0.	9.6	9.6

State	Age-adjusted Motor Vehicle Mortality* $\tilde{Y}_{+s+}$	Inspection State* (1 = yes) (0 = no) $\bar{Z}_{+s+} = \tilde{Z}_{+s+}$	Percent Urban** $\bar{W}_{+s+}$	Age-adjusted Percent Urban*** $\tilde{W}_{+s+}$
29	62.4	0.	9.6	9.6
30	45.0	0.	25.5	25.5
31	35.5	0.	31.1	31.1
32	47.6	0.	28.4	28.4
33	96.0	0.	.0	.0
34	49.9	1.	37.4	67.4
35	37.5	1.	21.5	51.5
36	29.6	1.	14.2	44.2
37	21.0	1.	34.7	64.7
38	37.4	1.	14.5	44.5
39	20.9	1.	18.7	48.7
40	79.1	1.	21.2	51.2
41	23.2	1.	55.8	85.8
42	28.1	1.	31.1	61.1
43	13.4	1.	33.6	63.6
44	47.7	1.	46.3	76.3
45	42.4	1.	35.3	65.3
46	51.4	1.	.0	30.0
47	35.7	1.	25.1	55.1
48	42.9	1.	13.5	42.5

* From Colton and Buxbaum (1968). Rate is for white males in 1960, adjusted to the total population age distribution in 1960.

** From Kitagawa and Hauser (1973)

*** Artificial. $\tilde{W}_{+s+} = \bar{W}_{+s+} + 30\bar{Z}_{+s+}$.

Table 2 summarizes the results of (a) regressing $\tilde{Y}_{+s+}$ on $\tilde{Z}_{+s+}$ and $\tilde{W}_{+s+}$ and (b) regressing $\tilde{Y}_{+s+}$ on $\bar{Z}_{+s+}$ and $\bar{W}_{+s+}$. For purposes of illustration only, no attention has been paid to the important questions of weighting the rates (Pocock, Cook and Beresford 1981) or to regression diagnostics (Draper and Smith 1966, chapter 3; Seber 1977, section 6.6). By construction of the age-adjusted urbanization variable, the two estimates of the coefficient of inspection differ markedly: with age-adjusted covariates, the coefficient is positive; with unadjusted covariates, the coefficient is significantly negative.

TABLE 2. Results Of Two Regressions: least squares estimate and 95% confidence intervals

Parameter	(a) Regression With Age-adjusted Covariates	(b) Regression With Unadjusted Covariates
α	60.54	60.54
	(50.77, 70.31)	(50.77, 70.31)
Δ	.27	-12.14
	(-14.20, 14.74)	(-21.38, -2.89)
ξ	-.41	-.41
	(-.76, -.07)	(-.76, -.07)

5. TECHNICAL ISSUES

5.1 A Formal Expression for the Bias of the Estimator of Δ.

We now obtain an expression for the bias that results from regressing adjusted mortality $\tilde{Y}_{+s+}$ on crude predictors $\bar{Z}_{+s+}$, $\underset{\sim}{X}_s$, and $\bar{\underset{\sim}{W}}_{+s+}$. Let

$$\bar{\underset{\sim}{V}} = \begin{bmatrix} 1 & \bar{Z}_{+1+} & \underset{\sim}{X}_1^T & \bar{\underset{\sim}{W}}_{+1+}^T \\ 1 & \bar{Z}_{+2+} & \underset{\sim}{X}_2^T & \bar{\underset{\sim}{W}}_{+2+}^T \\ & & \vdots & \\ 1 & \bar{Z}_{+S+} & \underset{\sim}{X}_S^T & \bar{\underset{\sim}{W}}_{+S+}^T \end{bmatrix} \quad \text{and} \quad \tilde{\underset{\sim}{V}} = \begin{bmatrix} 1 & \tilde{Z}_{+1+} & \underset{\sim}{X}_{+1+}^T & \tilde{\underset{\sim}{W}}_{+1+}^T \\ 1 & \tilde{Z}_{+2+} & \underset{\sim}{X}_{+2+}^T & \tilde{\underset{\sim}{W}}_{+2+}^T \\ & & \vdots & \\ 1 & \tilde{Z}_{+S+} & \underset{\sim}{X}_{+S+}^T & \tilde{\underset{\sim}{W}}_{+S+}^T \end{bmatrix}$$

and let $\underset{\sim}{\theta}^T = (\alpha, \Delta, \underset{\sim}{\gamma}^T, \underset{\sim}{\xi}^T)$. Moreover, let $\tilde{\underset{\sim}{Y}} = (\tilde{Y}_{+1+}, \tilde{Y}_{+2+}, \ldots, \tilde{Y}_{+S+})^T$. For any full rank matrix $\underset{\sim}{Q}$ that will be used to weight the adjusted mortality rates, the estimator $\tilde{\underset{\sim}{\theta}} = (\tilde{\underset{\sim}{V}}^T \underset{\sim}{Q} \tilde{\underset{\sim}{V}})^{-1} \tilde{\underset{\sim}{V}}^T \underset{\sim}{Q} \tilde{\underset{\sim}{Y}}$ that results from regressing adjusted $\tilde{Y}_{+s+}$ on adjusted covariates $\tilde{Z}_{+s+}$, $\underset{\sim}{X}_s$, $\tilde{\underset{\sim}{W}}_{+s+}$, is unbiased for $\underset{\sim}{\theta}$ since $E(\tilde{\underset{\sim}{Y}}|D) = \tilde{\underset{\sim}{V}}\underset{\sim}{\theta}$. However, the estimator $\bar{\underset{\sim}{\theta}} = (\bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \bar{\underset{\sim}{V}})^{-1} \bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \tilde{\underset{\sim}{Y}}$ that results from regressing adjusted $\tilde{Y}_{+s+}$ on crude covariates $\bar{Z}_{+s+}$, $\underset{\sim}{X}_s$, $\bar{\underset{\sim}{W}}_{+s+}$, has bias

$$E(\bar{\underset{\sim}{\theta}} - \underset{\sim}{\theta}|D) = [(\bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \bar{\underset{\sim}{V}})^{-1} \bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \tilde{\underset{\sim}{V}} - \underset{\sim}{I}] \underset{\sim}{\theta}$$

where $\underset{\sim}{I}$ is the identity matrix. Let $\underset{\sim}{t}^T = (t_1, \ldots t_s)$ be the second row of $(\bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \bar{\underset{\sim}{V}})^{-1} \bar{\underset{\sim}{V}}^T \underset{\sim}{Q} \tilde{\underset{\sim}{V}}$; the bias in the estimator of Δ from $\bar{\underset{\sim}{\theta}}$ is $\underset{\sim}{t}^T \underset{\sim}{\theta} - \Delta$, and so, as we would expect, the bias in the estimator of Δ is affected by all the variables.

If Z_{asi} is constant within each state, as in the case of a state law, then t_i is the mean difference, in the i^{th} column of $\tilde{V}$,

between states with the law ($\tilde{z}_{+s+} = 1$) and states without the law ($\tilde{z}_{+s+} = 0$) after covariance adjustment for $\underset{\sim}{x}_s$ and $\bar{\underset{\sim}{w}}_{+s+}$. For instance, in the example in §4, t_3 is the mean difference between inspection and noninspection states in age-adjusted urbanization after covariance adjustment for crude urbanization.

5.2 Properties of an Alternative Estimator

An alternative estimator, mentioned at the end of §2, involves regressing adjusted mortality $\tilde{Y}_{+s+}$ on crude predictors $\bar{z}_{+s+}$, $\underset{\sim}{x}_s$, $\bar{\underset{\sim}{w}}_{+s+}$ and age. Age may be represented either by moments of the age distributions within the states, m_{sj}, or by the proportions $\bar{p}_{as}$ of people in state s with age a. From (4) we have

$$E(\tilde{Y}_{+s+}|D) = \tilde{\alpha} + \Delta\bar{z}_{+s+} + \underset{\sim}{\gamma}^T\underset{\sim}{x}_s + \underset{\sim}{\xi}^T\bar{\underset{\sim}{w}}_{+s+} + \Delta(\tilde{z}_{+s+} - \bar{z}_{+s+}) + \underset{\sim}{\xi}^T(\tilde{\underset{\sim}{w}}_{+s+} - \bar{\underset{\sim}{w}}_{+s+}) + 0 \cdot \sum_{j=1}^{J}\bar{p}_{as+} \tag{6}$$

where the $\bar{p}_{as}$'s have zero coefficients since the expectation in (4), which is conditional on all the age information in D, does not depend on age. If the differences $(\tilde{z}_{+s+} - \bar{z}_{+s+})$ and $(\tilde{\underset{\sim}{w}}_{+s+} - \bar{\underset{\sim}{w}}_{+s+})$ can be written as linear functions of the $\bar{p}_{as+}$'s, then (6) can be rewritten

$$E(\tilde{Y}_{+s+}|D) = \tilde{\alpha}^* + \Delta\bar{z}_{+s+} + \underset{\sim}{\gamma}^T\underset{\sim}{x}_s + \underset{\sim}{\xi}^T\bar{\underset{\sim}{w}}_{+s+} + \sum_a \phi_a\bar{p}_{as} \tag{7}$$

for some parameters $\tilde{\alpha}^*$ and ϕ_a, $a = 1,2,\ldots,A$; in this case, the alternative estimator leads to unbiased estimates of Δ.

The differences $(\tilde{z}_{+s+} - \bar{z}_{+s+})$ and $(\tilde{\underset{\sim}{w}}_{+s+} - \bar{\underset{\sim}{w}}_{+s+})$ will indeed be linear functions of the proportions $\bar{p}_{as+}$ if the age-specific regressors

$\bar{z}_{as+}$ and $\bar{\underset{\sim}{w}}_{as+}$ can be written as the sum of an age and a state component, i.e. if

$$\bar{z}_{as+} = m_a + r_s$$

and (8)

$$\bar{\underset{\sim}{w}}_{as+} = \underset{\sim}{u}_a + \underset{\sim}{v}_s$$

for some scalars m_a and r_s, and some vectors $\underset{\sim}{u}_a$ and $\underset{\sim}{v}_s$, for all a and s. If $\bar{w}_{as+}$ is average income in state s at age a, and (8) is true, then the difference in average income between New York and Virginia, say, is the same at all ages. To see that (8) implies the required linear dependence, note that

$$\begin{aligned} \tilde{z}_{+s+} - \bar{z}_{+s+} &= \sum_a \bar{z}_{as+}(f_a - \bar{p}_{as+}) \\ &= \sum_a (m_a + r_s)(f_a - \bar{p}_{as+}) \\ &= (\sum_a m_a f_a) - (\sum_a m_a \bar{p}_{as+}) \end{aligned} \qquad (9)$$

since $\sum_a f_a = \sum_a \bar{p}_{as+} = 1$. As required, (9) is a linear function of the $\bar{p}_{as+}$'s. Analagous arguments apply to the $\bar{\underset{\sim}{w}}_{as}$'s.

The condition that $(\tilde{z}_{+s+} - \bar{z}_{+s+})$ and $(\tilde{\underset{\sim}{w}}_{+s+} - \bar{\underset{\sim}{w}}_{+s+})$ must be linear functions of the proportions $\bar{p}_{as+}$ is quite restrictive. Even random deviations from linear dependence would constitute errors in the predictor variables, leading to biased estimates by analogy with standard arguments (e.g. Seber 1977, p.155; Johnston 1972, p.281).

6. Summary

We have considered the following seven procedures:

(a) Regression of the responses of individuals, Y_{asi}, on the age of individuals and the predictors $(Z_{asi}, \underset{\sim}{X}_s, \underset{\sim}{W}_{asi})$ describing individuals.

(b) Weighted regression of the age-specific response rates $\bar{Y}_{as+}$ on the age-specific predictor averages $(\bar{Z}_{as+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{as+})$.

(c) Weighted regression of the crude response rates $\bar{Y}_{+s+}$ on the crude predictor averages $(\bar{Z}_{+s+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{+s+})$.

(d) Weighted regression of the age-adjusted rates $\tilde{Y}_{+s+}$ on the age-adjusted predictors $(\tilde{Z}_{+s+}, \underset{\sim}{X}_s, \tilde{\underset{\sim}{W}}_{+s+})$.

(e) Weighted regression of age-adjusted rates $\tilde{Y}_{+s+}$ on age and crude predictors $(\bar{Z}_{+s+}, \underset{\sim}{X}_{+s+}, \bar{\underset{\sim}{W}}_{+s+})$.

(f) Weighted regression of age-adjusted rates $\tilde{Y}_{+s+}$ on crude predictors $(\bar{Z}_{+s+}, \underset{\sim}{X}_s, \bar{\underset{\sim}{W}}_{+s+})$.

(g) Weighted regression of age-specific rates $\bar{Y}_{as+}$ on crude predictors $(\bar{Z}_{+s+}, X_s, \bar{\underset{\sim}{W}}_{+s+})$.

Under the simple linear model for (a), that is equation (1), methods (a) through (d) yield unbiased estimates of the parameters of the model; however, the data required for methods (a), (b), and (d) are often unavailable in official tabulations. The crude rates required for (c) are available in some but not all official tabulations; for example, homicide death rates are rarely age-adjusted, whereas, coronary disease mortality rates are usually age-adjusted. Method (e) can yield unbiased estimates under restrictive assumptions defined in §5.2. Methods (f) and

(g), although popular techniques in practice, do not generally lead to unbiased estimates under the linear model for (a).

REFERENCES

Blair, A., Fraumeni, J.F., and Mason, T.J. (1980) Geographic patterns of leukemia in the United States. Journal of Chronic Diseases, 33, 251-260.

Colton, T. and Buxbaum, R.C. (1968) Motor vehicle inspection and motor vehicle accident mortality. American Journal of Public Health, 58: 1090-1099.

Cox, D.R. (1970) The Analysis of Binary Data. London: Metheun.

Dempster, A.P. (1969) Elements of Continuous Multivariate Analysis. Reading, Massachusetts: Addison-Wesley.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39: 1-38.

Draper, N.R. and Smith, H. (1966) Applied Regression Analysis. New York: John Wiley.

Fuchs, V.R., and Leveson, I. (1967) Motor accident mortality and compulsory inspection of vehicles. Journal of the American Medical Association, 201: 657-661.

Johnston, J. (1977) Econometric Methods. New York: McGraw-Hill.

Kitagawa, E.M. and Hauser, P.M. (1973) Differential Mortality in the United States. Cambridge, Massachusetts: Harvard University Press.

Mosteller, F. and Tukey, J.W. (1977) Data Analysis and Regression. Reading, Massachusetts: Addison Wesley.

Pocock, S.J., Cook, D.G., and Beresford, S.A.A. (1981) Regression of area mortality rates on explanatory variables: what weighting is appropriate? Applied Statistics 30: 286-295.

Seber, G.A.F. (1977) Linear Regression Analysis. New York: John Wiley.

St. Leger, A.S., Cochrane, A.L., and Moore, F. (1979) Factors associated with cardiac mortality in developed countries with particular reference to the consumption of wine. Lancet 1: 1017-1020.

PRR/DBR/jik

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2428	2. GOVT ACCESSION NO. AD-A124 381	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DIFFICULTIES WITH REGRESSION ANALYSES OF AGE ADJUSTED RATES		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Paul R. Rosenbaum and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041 and P30-CA-14520
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Probability and Statistics
11. CONTROLLING OFFICE NAME AND ADDRESS * see below		12. REPORT DATE September 1982 13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

* U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Observational studies, covariance adjustments, regression analysis, direct adjustment, adjusted rates.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A common type of observational study compares population rates in several regions having differing policies in an effort to assess the effects of those policies. In many studies, particularly in public health and epidemiology, age-adjusted rates are regressed on predictor variables to obtain a covariance

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (cont.)

adjusted estimate of effect; we show that this estimate is generally biased for the appropriate regression coefficient. The analysis of crude rates with age as a covariate can, under familiar models, lead to unbiased estimates, and therefore can be preferrable. Several other regression methods are also considered.

END

DATE
FILMED

3-83

DTIC