IIA.E.25

THE JOHNS HOPKINS UNIVERSITY

SCHOOL OF HYGIENE AND PUBLIC HEALTH

DEPARTMENT OF BIOSTATISTICS

615 North Wolfe Street • Baltimore, Maryland 21205

August 21, 1985

David Weil, Ph.D.
ECAO (MD 52)
U. S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

Dear Dr. Weil:

Here is the report which I have prepared for your office under purchase order number 5D4076NASA. Please let me know if you need any further clarification.

Yours truly,

Richard Royall

Richard M. Royall
Professor

RM/pah
Enclosure

ON THE RELATIONSHIP BETWEEN BLOOD LEAD AND BLOOD PRESSURE

IN THE NHANES II DATA

August 20, 1985

Richard M. Royall

This is in response to a request for my evaluation of the comments of DuPont and the reply by J. Schwartz concerning the use of site variables in regression analyses of the relationship of blood pressure to blood lead in the NHANES II data.

My questions and comments can be put in terms of a series of simplified models. The first model states that for fixed values of ℓn blood lead (x) and other regressors (z) (and within a race × sex subgroup) a blood pressure for the j^{th} individual at site i has expected value of the form:

$$E(Y_{ij}) = \alpha + \beta x_{ij} + \gamma z_{ij} \quad . \qquad (1)$$

As I understand it, model (1) is essentially that used by Pirkle et al. (1985). We are particularly interested in β, the blood pressure/ℓn blood lead slope.

Pirkle et al. used various combinations of regressors (z) in analyses under model (1) and estimated β (diastolic pressure, white males aged 40-59 years) at nearly 4, with a standard error of 1.4 (Table 2, p. 250). Because these analyses (and the related ones of Harlan et al. (1985)) included a vast array of regressors (z), including most of those physiologic and dietary variables known or reasonably suspected to have a non-negligible relationship to blood pressure, their result looks solid.

Since the data were collected from 64 different geographic sites (and observations at different sites were made at different times) it seems natural to check the above findings using another model:

$$E(Y_i) = \alpha_i + \beta x_{ij} + \gamma z_{ij} \quad . \tag{2}$$

If indeed model (1) is adequate and there are no important regressors missing, model (2) should give about the same results as (1) concerning β, so long as there is adequate x and z variation within sites. If the x-variation within sites is small, however, the estimate of β, while still unbiased, will have a larger standard error than that obtained under model (1). On the other hand, if there is substantial x-variation within sites, and if the new (model (2)) estimate of β is much smaller than the old (model (1)) estimate, then we must worry about possible inadequacy of (1) even if we are unable to think of a better model.

As another check on the model (1) results we might analyze the data under the model

$$E(Y_{ij}) = \alpha + \beta_i x_{ij} + \gamma z_{ij} \tag{3}$$

which allows the x-slope to change from site to site. If we see little variation among the estimated β_i's then we have more confidence in (1). But too much variation (relative to that expected if the β_i are all in fact equal) would raise new questions. In that case, if $(\bar{x}_i, b_i)$ are the mean and the slope estimate for site i, then a plot of these 64 points suggesting a smooth curve would indicate that the regression of y on x is

curvilinear. But variation in the b_i which is not related to the x_i would be very hard to interpret. Either way, large variation among the b_i's would make us reluctant to accept the results from (1). Again the within-site x-variation is important, because smaller x-variation not only gives a larger standard error to the estimate (b_i) of slope for that site, but also gives a greater bias towards zero if the x's are measured with error.

Yet another check would be given by the model

$$E(Y_{ij}) = \alpha_i + \beta_i x_{ij} + \gamma z_{ij} \qquad (4)$$

which combines the generality of (2) and (3). But now the number of parameters is 128+k, where k is the number of z-regressors, and we begin to be limited by the size of the NHANES II data set. For example, in the subgroup analyzed by Pirkle et al. (white males aged 40-59) there are fewer than 600 observations, or fewer than 5 per parameter in model (4).

Finally we might allow the z-coefficients to vary with site

$$E(Y_{ij}) = \alpha_i + \beta_i x_{ij} + \gamma_i z_{ij} \quad . \qquad (5)$$

But this model, with 64(2+k) parameters, looks quite beyond the reach of our data set.

I am not sure which of these models (2) - (5) corresponds to that used in the DuPont analysis (May 9, 1985) in which the estimated β_1 in blood lead coefficient turned out to be less than half as big as the Pirkle et al. estimate (and not statistically significantly different from zero). For that reason I do not know how to interpret the new result. The key paragraph seems to be the following (DuPont Report p.7):

> "The blood pressure-log blood lead coefficient in models with 64 site terms represents the average of the within-site coefficients, weighted by the number of observations from each site. The within-in site coefficients alternatively could have been obtained by running separate analyses for each of the 64 sites, but, by introducing the X_i site indicator variables, only one model was necessary."

The first sentence suggests model (4). But the second suggests that it might have model (5) instead. Has anyone used models (2) and (3)? How much within-site variation among x's (ℓn blood lead) is there? It seems to me that the only valid reasons for not trying to confirm the conclusions from model (1) by analyses under more general models like (2) and (3) would be that after adjusting for the z-regressors there remains little within-site x-variation.

My concern is simply with the internal consistency of the evidence. If blood lead is associated with increased blood pressure, then we should see this relationship within sites unless there is little within-site blood lead variation or we introduce so many parameters (as in model (5)) that the whole picture goes out of focus.

I remain confused. While DuPont raised the question of consistency of results across sites, I am unsure about which model they used and about how much within site variation there is for x and the other regressors. Thus I don't know how to interpret the DuPont findings. The reponses of EPA (Schwartz memos to Weil "Response to DuPont Critique..." and to Ware "Time trends and site effects...") have not focussed on the points about which I am concerned.

REFERENCES

DuPont. Comments on Corrigenda to the Second External Review Draft of the Revised Criteria Document, Air Quality Criteria for Lead (EPA-600/8-83-028B), Docket Number ECAO-CD-81-2), May 9, 1985.

Harlan, W. R., Landis, J. R., Schmouder, R. L., Goldstein, N. G., and Harlan, L. C. (1985). Blood lead and blood pressure. JAMA, 253, 530-534.

Pirkle, J. L., Schwartz, J., Landis, J. R., and Harlan, W. R. (1985). The relationship between blood lead levels and blood pressure and its cardiovascular risk implications. American Journal of Epidemiology, 121, 246-258.

Schwartz, J. (1985). Blood lead, blood pressure, and site effects. Memo to David Weil, ECAO, dated June 26, 1985.

Schwartz, J. (1985). Time trends and site effects in the NHANES blood lead-blood pressure relationship. Memo to James Ware, CASAC.