EN-1027 REV. 7-70

E. I. du Pont de Nemours & Company
Incorporated
Wilmington, Delaware 19898

Engineering Department

CC: A. J. Pahnke-PD-Pet Lab
E. N. Cantwell-PD-Pet Lab
J. M. Pierrard-PD-Pet Lab
M. C. Olson-ESD Field
W. F. Wagner-ESD
D. W. Marquardt-ESD/ASG File
~~R.~~ D. Snee-ESD
IC 31 - Applied Statistics

October 7, 1982

E. S. JACOBS
PETROCHEMICALS DEPARTMENT
CHAMBERS WORKS, PETROLEUM LAB

<u>Draft Section V on Examination of CDC Data</u>

Attached for your records is my redraft of E. N. Cantwell's Section V for the October 8, 1982 submission on EPA's proposed gasoline lead rule that we discussed earlier this week. I reviewed the results and draft with R. D. Snee on Monday and we both feel comfortable with the statistical approach and presentation of findings.

ENGINEERING SERVICE DIVISION

C. G. Pfeifer

C. G. Pfeifer
Applied Statistics Group

CGP/ssd
Atch

N33791

TEH 0531516

Section V

EXAMINATION OF THE INFLUENCE OF THE SUCCESS OF THE CDC CHILDHOOD LEAD POISONING PREVENTION PROGRAMS AND GASOLINE LEAD USE ON CHILDREN IDENTIFIED TO HAVE ELEVATED BLOOD LEAD LEVELS

Summary

Explanatory variables reflecting lead paint exposure and gasoline lead use were related to percent of children with blood lead levels above 30 μg/dl identified in the CDC Childhood Lead Poisoning Prevention Programs. It was found that lead paint exposure variables were significantly correlated with the incidence of elevated blood lead levels, and by themselves, could account for all of the decreasing chronological trend in blood lead. Gasoline lead use by itself could not account for all of the decreasing trend in the blood lead data. When both lead paint exposure variables and gasoline lead use were considered simultaneously, gasoline lead use was not significantly correlated with the incidence of blood lead levels above 30 ug/dl (or above 49 ug/dl) but lead paint exposure variables were. These results demonstrate the importance of considering all sources of lead exposure in any correlation analysis.

The quarterly data reported by the CDC on children with elevated blood lead levels and on the actions taken to reduce lead exposure in dwelling were combined with quarterly data on the amount of lead used in gasoline. Data were available for the time period from the first quarter 1976 through the third quarter 1981 for 31 states in the U.S. Multiple linear regression analyses were employed to investigate the relationships among percent of children with elevated blood lead levels, lead paint exposure variables, and gasoline lead use.

Data Bases

The data bases used to examine possible relationships between the percent of children with blood lead levels above 30 μg/dl and lead-based paint and/or gasoline lead were:

1. The quarterly reports of the CDC Lead Poisoning Prevention Program published in the CDC Morbidity and Mortality Weekly Reports.

2. The reports by the Ethyl Corporation of the amount of leaded gasoline sold each month by state.

3. The reports by the Department of Energy of their semiannual Motor Gasoline Surveys.

The above information was used to derive explanatory variables employed in multiple regression analyses.

CDC Lead Poisoning Prevention Program - The CDC Program reports the following information for each calendar quarter. (Shown in parenthesis for each data entry are variable names for subsequent reference.)

1. The number of children screened for possible high lead levels. (Screened)

2. The number of children with confirmed blood lead levels between 30 and 49 μg/dl. (Class 2)

3. The number of children with confirmed blood lead levels at or above 50 μg/dl. (Classes 3,4)

4. The number of dwellings inspected for possible high lead exposure. (Inspected)

5. The number of dwellings found to have high levels of lead exposure due to the presence or residues from lead-based paint. (Pb Found)

6. The number of dwellings in which the high level of lead exposure was reduced by some action. (Pb Reduced)

The above information is given for various cities or for an entire state. Thirty-one states have data recorded for at least some time periods since the inception of this program. Because of the many cases where data is reported only for an entire state, the data from various cities in any given state were combined to give state totals for each quarter.

From the primary data discussed above, the following variables were calculated.

7. The children with blood lead levels between 30 and 49 μg/dl, CDC Class 2, expressed as a percent of those screened. (% Class 2)

8. The children with blood lead levels at or above 50 μg/dl, CDC Classes 3 and 4, expressed as a percent of those screened. (% Classes 3,4)

9. The total of children with blood lead levels above 30 μg/dl, expressed as a percent of those screened. (% Classes 2,3,4)

10. The dwellings found to have high levels of lead exposure expressed as a percent of those inspected. (% Pb Found)

11. The dwellings in which the high level of lead exposure was reduced, expressed as a percent of those inspected. (% Pb Reduced)

12. The dwellings in which the high level of lead exposure was reduced, expressed as a percent of those in which high level of lead exposure was found. (% Reduced/Found)

To account for the possible long-term effect of the continuing programs to reduce the high levels of lead exposure, the following variables were also calculated for each calendar quarter.

13. The cumulative number of dwellings in which a high level of lead exposure was reduced. (Cum. Pb Reduced)

14. The cumulative percent of dwellings in which a high level of lead exposure was reduced. (Cum. % Pb Reduced).

Thus, the CDC Screening Program yields a measure of the children with blood lead levels above 30 $\mu g/dl$, a measure of their exposure to high lead levels due to high levels of lead in their dwellings (primarily from lead-based paint) and a measure of the effectiveness of the CDC programs to reduce exposure to high lead levels in dwelling.

Gasoline Lead Use - Gasoline lead use for each of the states in the U.S. was derived from combining the Ethyl monthly gasoline data with the DOE semiannual survey data. The Ethyl data were first combined into calendar quarters to give the amount of leaded regular and leaded premium grade gasoline sold each quarter in each state. The DOE semiannual data were used to derive the lead concentration in regular and premium gasolines for the winter (1) and summer (3) quarters for each state. The lead concentrations for the spring (2) and fall (4) quarters were estimated by linear interpolation between the appropriate winter and summer quarters. The total lead used by quarter by state was calculated by multiplying the gasoline volumes by the corresonding gasoline lead concentration values for the regular and premium grades and summing the two values (Gas Pb Used).

Time Frame - The time period chosen for this study was from first quarter of 1976 through third quarter 1981 or 23 consecutive quarters. The CDC changed the blood lead levels defining the different classes of children in March 1975 and was in the process of changing the reporting format in 1975. The first year in which all of the data was reported fully was 1976. The latest published report is for the third quarter of 1981.

Data List - A listing of the data base, showing CDS data items 1-6 and gasoline lead use by state and calendar quarter is given in Appendix A.

General Approach

Inspection of the data revealed a general downward trend in percent of children with blood lead levels above 30 μg/dl (% Classes 2,3,4) when the data were combined over all 31 states and plotted against calendar quarter (Figure 1). This downward trend could be interpreted as due to one or more of the following:

- Success of the continuing CDC lead paint exposure reduction campaigns
- Downward trend in gasoline lead use
- Trends in other sources of lead exposure for which data are not presently available
- Nonhomogeneous sampling
- Improved or better controlled analytical procedures for measuring blood lead.

Since Du Pont is unaware of any quantitative information that would enable the latter three reasons to be addressed, attention was focused on the effects of the lead paint reduction effort and gasoline lead use on explaining variation in percent of children in elevated blood lead Class 2, Classes 3,4, and Classes 2,3,4.

Predictor variables considered for multiple regression analysis were

Lead Paint	Gasoline Pb	Classification
Inspected	Gas. Pb Used	State
% Pb Found		Season
% Pb Reduced		
% Reduced/Found		
Cum. Pb Reduced		
Cum. % Pb Reduced		

The number of lead paint variables chosen for consideration is fewer than the total number available because a number of the variables were calculated from two of the others. The classification variables, state and season, were included because of expected state-to-state and season-to-season differences that would contribute to variation in percents of elevated blood levels.

A chronological trend term was not included in the basic regression analyses because, presumably, whatever time trend exists is due to one or more of the five reasons above. However, analyses were performed on percent elevated blood lead data corrected for lead exposure variables to determine if any chronological trend remained.

All statistical analyses reported are regressions run on the Statistical Analysis System (SAS), a commercial statistical software package. Preliminary analyses of regression residuals strongly suggested that it was not appropriate to use % Class 2, % Classes 3,4, or % Classes 2,3,4 as such as dependent variables because of lack of normality and heteroscedasticity of variances. Consequently, unless otherwise noted, the dependent variables for all analyses were $\log_{10}$(% Class 2), $\log_{10}$(% Classes 3,4), $\log_{10}$(% Classes 2,3,4).

Analyses and Results

Regression analyses on each of the three dependent variables for three models, denoted by A, B and C, were initially performed. Model A included terms for all lead exposure variables; Model B included terms accounting for lead paint exposure sources but not gasoline lead use; and, Model C included terms accounting for gasoline lead use but not for lead paint exposure sources. All models include terms for state and season. These models are simple additive models (ie, no interactions among terms are accounted for) and are of the form previously used in blood lead correlation studies (eg, Billick et al _____).

The results of these analyses are summarized in Table 1. The R^2-value denotes the percent of the total variation in the dependent variable that is explained by all the terms in the model. The decremental F-ratio is a statistic measuring the correlation of a model term with the dependent variable after accounting for the effects of all other model terms. The significance level reflects the likelihood of no correlation. For significance levels $\leq .05$ a conclusion of significant correlation is generally made.

From the results shown in Table 1 the following observations may be made:

- For Model A and all categories of elevated blood leads, several lead paint exposure sources and gasoline lead use were statistically significant.
- % Pb Reduced and Cum. Pb Reduced are the most influential lead paint exposure variables.
- Analyses ignoring gasoline lead use (Model B) lead to the conclusion that several lead paint exposure variables are significantly associated with each category of elevated blood lead.
- Analyses ignoring lead paint (Model C) exposure variables lead to the conclusion that gasoline lead use is significantly associated with each category of elevated blood lead.

Because of the nature of historical data bases such as the one being used, there was concern that correlations among the predictor variables might disguise some key information. Clearly, throughout the 1970's most sources of lead exposure have decreased, and therefore, it was expected that gasoline lead would be correlated with lead paint variables and, further, that lead paint dust variables would be correlated with each other.

One way to identify the seriousness of such suspected multicollinearities among exposure variables, is to regress each, in turn, on all the others. The results of these regression analyses are given in Table 2. These results clearly show that for each of seven exposure variables considered, at least two, and as many as four of the remaining six are significantly correlated with the dependent variable, as evidenced by estimated coefficients that are significantly different from zero.

Therefore, it was decided to determine a subset of the seven exposure variables that gave the best explanatory power (highest R^2-value). For this analysis SAS's MAX R option of PROC STEPWISE was used. The model chosen was the one that included the largest subset, subject to the condition that each exposure variable was statistically significant when evaluated in the presence of the others. The analysis was performed for each of the three elevated blood lead dependent variables corrected for state-to-state and season-to-season effects.

For all these elevated blood lead categories, the following four exposure variables satisfied the above conditions.

% Pb Reduced
Cum. % Pb Reduced
% Pb Found
Gas Pb Use

These variables provided additional explanatory power (incremental R^2) to that of state and season factors as summarized below

Category	Incremental R^2
% Class 2	19.7%
% Classes 3,4	13.5%
% Classes 2,3,4	19.5%

Following this selection procedure another set of regression analyses were performed using only the above four exposure variables, in addition to state and season terms. However, unlike the previous analyses summarized in Table 1, all possible two-way interactions were included in the model. This was done because frequently an overly simplified model can give misleading results, especially when partial confounding (multicollinearities) among predictor variables is present. It was also suspected that significant interactions (such as the effect of gasoline lead use on blood lead categories might be different for different states) might exist. If they did they would help improve the explanatory power of the model. Although 227 degrees of freedom were associated with estimating all two-way interactions, approximately 350 degrees of freedom (depending on the model) remained for estimation of error.

Three models, denoted by A', B' and C', were investigated. They correspond to Models A, B, C discussed above except, as noted, only four lead exposure variables are considered along with all two-way interactions.

The results of these analyses are presented in Table 3 and are displayed analogous to those of Table 1. The significance levels of the decremental F-ratios of the interaction terms are not given for space considerations. Confirming suspicions, in each case several statistically significant interactions were noted. For example, for Model A' and Category % Classes 2,3,4, interactions, Gas. Pb Use vs State, % Pb Reduced vs State, Cum. % Pb Reduced vs State, and State vs Season were all statistically significant.

From the results shown in Table 3, the following observations may be made:

- The predictive power (R^2) of these models, which included interactions and a parsimonious set of exposure variables, was greatly improved over those results in Table 1.

- Results from Model A', which includes terms reflecting both lead paint exposure and gasoline lead use, suggest that lead paint exposure variables are more influential in explaining variation in elevated blood lead levels than gasoline lead use. Gasoline lead use is not significant for, % Classes 3,4 and % Classes 2,3,4.

- Analyses ignoring gasoline lead use (Model B') lead to the conclusion lead paint exposure variables are significantly associated with each category of elevated blood lead.

- Analyses ignoring lead paint (Model C') lead to the conclusion gasoline lead use is significantly associated with each category of elevated blood lead.

To evaluate the extent to which the lead exposure variables considered could account for the chronological time trend, the residuals from the analyses on Models A', B' and C' were used. There corresponds a residual for each data record used in a regression analysis; a residual reflects the difference between the observed dependent variable value and the value predicted from the model. The average value of all the residuals is zero. By adding the average value of the dependent variable analyzed to each residual, a data set with the same average value as the original data is obtained. These data are sometimes referred to as corrected data -- corrected for effects associated with terms in the regression model. Thus, corrected data from Models A'-C' are data where the effects of the lead exposure sources state, season, and all two-way interactions have been removed. By regressing these corrected data (actually taking the antilogarithm of the corrected data) on chronological quarter (1-23) one can assess whether any time trend remains.

The results of these analyses are given in Table 4. The entries under "Uncorrected Data" reflect regression slopes associated with chronological time; these slopes summarize the overall linear time trend (Δ % class category/quarter) in the actual data (not subjected to logarithmic transformation).

From the results shown in Table 4, the following observations may be made:

- Analyses on the uncorrected or actual data (ie, % Class 2, % Classes 3,4, and % Classes 2,3,4) show a statistically significant decreasing trend.
- Based on corrected data from Model A', no significant time trend remained after the effect of lead paint exposure variables and gasoline lead use were accounted for.
- Based on corrected data from Model B', no significant time trend remained after the effects of lead paint exposure variables were accounted for.
- Based on corrected data from Model C', a significant time trend remained (% Classes 3,4 slope was significant at level .055) after the effect of gasoline lead was accounted for.

On the basis of the foregoing points, these results suggest that lead paint exposure variables are more associated with variation in elevated blood lead categories than gasoline lead use.

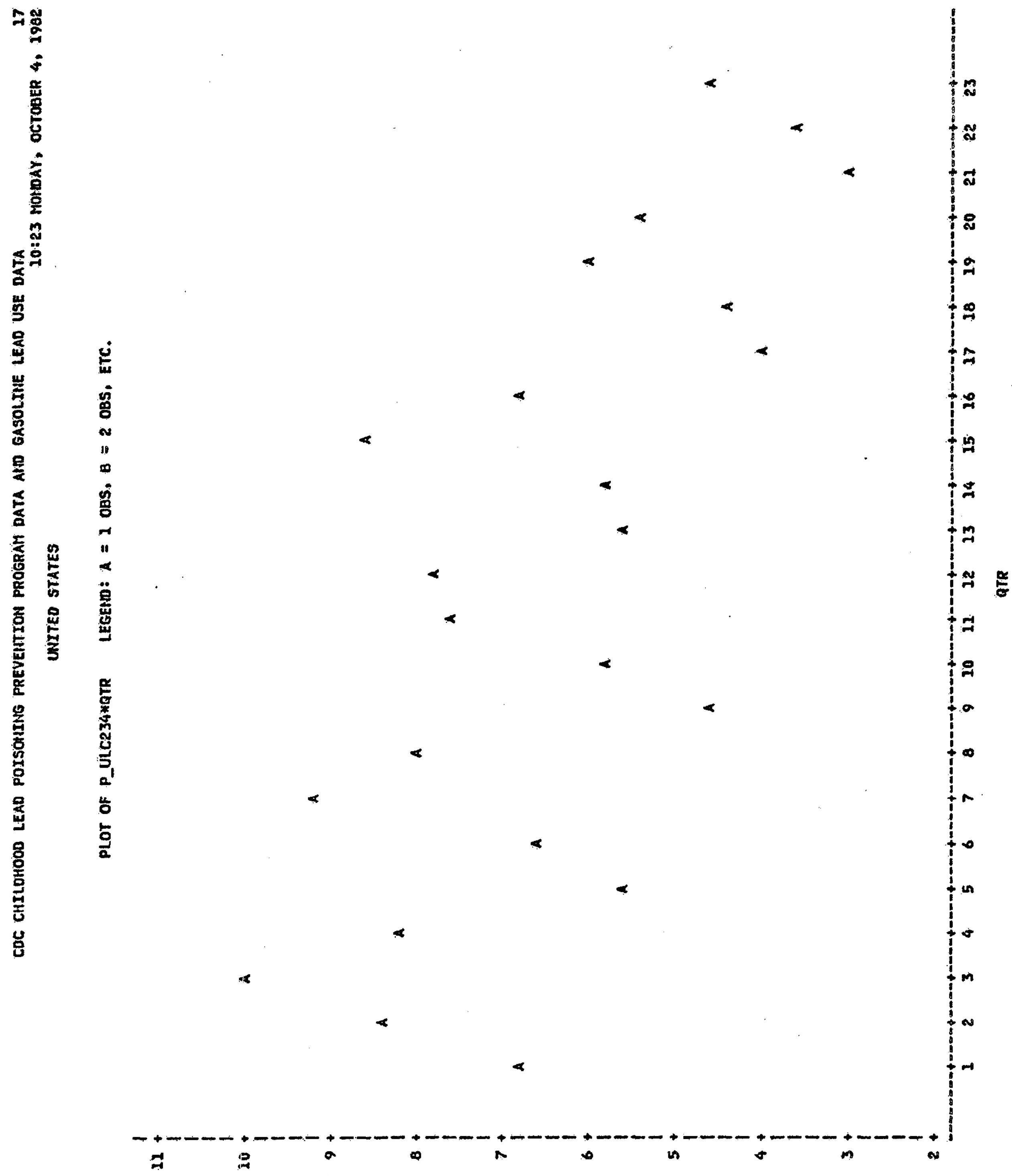
Discussion

The results of the foregoing analyses put in better perspective the importance that sources of lead exposure (other than gasoline lead) can have on statistically explaining blood lead variation. The findings presented suggest that the CDC cleanup programs are not only successful in reducing elevated blood lead levels but are more influential than reductions in gasoline lead use. It is exasperating to Du Pont, that analyses correlating elevated blood lead levels to lead paint exposure variables have not been reported and submitted to the docket on the present issue. It is especially disconcerting because neither CDC nor HUD have investigated these data -- data which are directly related to their chartered responsibilities. Rather, instead, both CDC and HUD have chosen to spend their energies correlating gasoline lead use (to the exclusion of other sources of lead exposure) to blood lead levels.

It is recognized that inferring casual relationships and relative contributions of each exposure source to variation in blood lead responses from analyses such as these can be very misleading. When only lead paint exposure variables were considered, one might be tempted to conclude that lead paint exposure is the primary influence on children with elevated blood leads; conversely, when only gasoline lead is considered one might be tempted to conclude that it has the primary influence on children with elevated blood leads. The problem evidenced here and also inherent in all studies relating blood lead to a single lead exposure source such as gasoline lead, is that all lead exposure variables are presumably highly correlated with one another over time (not to mention the inherent problems in the quality of the blood lead data itself), and consequently, the effects and contributions of each can never be unambiguously sorted out. Focusing in on one or a few exposure sources, to the exclusion of others, only exaggerates the confusion. Granted regulatory decisions still have to be made, but it indeed would be unfortunate if more decisions depend heavily on such correlation analyses without being aware of their limitations and the high likelihood of incorrect conclusions being drawn from them.

The results presented serve to effectively uncover, by way of a concrete example, the kinds of problems and limitations associated with correlating blood lead with only gasoline lead use that Du Pont has repeatedly admonished the EPA on in earlier submissions to this docket. Based on these preliminary results and the fact that this data base is unique in that it includes information on more than one lead exposure source, Du Pont will continue to investigate these data and strongly encourages EPA to do the same.

FIGURE 1



TEH 0531527

DUP050032738

Table 1

Regression Results (1) Comparing Explanatory Power of Lead Paint Exposure and Gasoline Lead Use

	% Class 2			% Class 3,4			% Classes 2,3,4		
	Model A	Model B	Model C	Model A	Model B	Model C	Model A	Model B	Model C
% Variation Explained, R^2	64.4	55.9	60.3	63.5	58.3	60.5	66.8	58.6	62.8
Significance Levels of Decremental F-ratios									
State	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001
Season	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001	≤.0001
Inspected	.1131	.1732		.3731	.4071		.1810	.2505	
% Pb Found	.9743	.6101		.9512	.6556		.6324	.8852	
% Pb Reduced	≤.0001	≤.0001		≤.0001	≤.0001		≤.0001	≤.0001	
% Reduced/Found	.1123	.1091		.0560	.0546		.0501	.0589	
Cum. Pb Reduced	≤.0001	.5542		.0004	.5240		≤.0001	.5215	
Cum % Pb Reduced	.0014	.0008		.0280	.0147		.0019	.0011	
Gas Pb Used	≤.0001		≤.0001	≤.0001		≤.0001	≤.0001		≤.0001

(1)Dependent variable was $\log_{10}$ of percent of elevated blood leads found in each class.

TEH 0531528

DUP050032739

Table 2

Results of Each Lead Explanatory Variable Regressed on Others

	Coefficients of Lead Exposure Predictor Variables							
Dependent Variable	Inspected	% Pb Found	% PB Reduced	% Reduced/ Found	Cum. Pb Red.	Cum. % Pb Red.	Gas Pb	% Variable Explained R2
Inspected	--	103.3*	-6.58*	.390	-36.5	.201*	176.8*	20.4%
% Pb Found	.0009*	--	.006	-.037*	-.0002*	1.024*	-.069	59.5%
% Pb Reduced	-.004	.403	--	.017	.011*	1.949*	-10.7*	40.8%
% Pb Reduced/Found	.0005	-4.92*	.031	--	.0001	8.63*	3.34	29.9%
Cum. Pb Reduced	.360*	-37.2*	30.65	.200	--	-19.1	322.4*	37.7%
Cum. % Pb Reduced	-.000	.500*	.013*	.032*	.000	--	.208	62.0%
Gas Pb	.00005*	-.002	-.004*	.001	.00005*	.012	--	6.6%

*denotes coefficient significantly different from zero at level ≤.05

TEH 0531529

DUP050032740

Table 3

Regression Results (1) Comparing Explanatory Power of Lead Paint Exposure and Gasoline Lead Use Using a Model Accounting for Two-way Interactions

	% Class 2			% Classes 3,4			% Classes 2,3,4		
	Model A'	Model B'	Model C'	Model A'	Model B'	Model C'	Model A'	Model B'	Model C'
% Variation Explained, R^2	86.7	82.4	77.1	85.8	82.9	73.6	88.2	84.4	77.9
Significance Levels of Decremental F-ratios									
State	.0022	≤.0001	≤.0001	.0011	≤.0001	≤.0001	.0007	≤.0001	≤.0001
Season	.4160	.7857	.0387	.0569	.0002	≤.0001	.8667	.2555	.0002
% Pb Found	.6686	.0980		.1111	.1326		.3796	.0398	
% Pb Reduced	.0154	≤.0001		.3813	.0018		.0622	≤.0001	
Cum % Pb Reduced	.0030	.1687		.0317	.0334		.0065	.0802	
Gas Pb Used	.0493		≤.0001	.8938		.0003	.1429		≤.0001

(1) Only results of linear terms are shown; each model accounted for effects of all possible two-way interactions among linear terms. The dependent variable was $\log_{10}$ of percent of elevated blood leads found in each class.

Table 4

Intercepts and Slopes of Residuals(1) and Uncorrected Data(2) Regressed on Chronological Quarter (1-23)

		Uncorrected Data(2)	Based on Residuals of Model A'	B'	C'
% Class 2	Intercept	2.883*	3.005*	3.215*	
	Slope	.00159	-.00469	-.0173*	-.180*
% Classes 3,4	Intercept	1.431*	1.479*	1.685*	
	Slope	.00169	-.000586	-.0119**	-.0862*
% Classes 2,3,4	Intercept	4.333*	4.498*	4.870*	
	Slope	.00354	-.00532	-.0266*	-.266*

(1)Overall mean response was added to each residual from Models A', B', and C' and then these values were converted to original metric by taking antilogarithms.

(2)Model using observed percents of elevated blood leads in each class category include terms to account for state-to-state and season-to-season differences.

* Significantly different from zero at level $\leq .05$
**Significantly different from zero at level $\leq .06$

TEH 0531531

DUP050032742