1-6

Date: 10/19/82

Subject: Comments on the Department of Energy Analysis of the NHANES II Lead Data

From: James L. Pirkle M.D. Ph.D.
JLP Clinical Chemistry Division
Center for Environmental Health
Centers for Disease Control

The Department of Energy (DOE) model for predicting the number of children with lead toxicity (defined as whole blood lead of 30 ug/dl or more) from gasoline lead levels is associated with significant amounts of random and systematic error which seriously affect is predictive capabilities. After correcting these sources of error as much as possible, the random error in predicting the number of children with lead toxicity at a 95% confidence interval is still at least as large as ± several hundred thousand children even at low levels of gasoline lead consumption.

The following is first a summary then a discussion of these errors:

(1) Centers for Disease Control (CDC) lead screening data indicates that the DOE model underestimates the number of children with toxic lead levels by approximately a factor of 10.

(2) As the mean blood lead level decreases, the DOE assumption of a lognormal fit to the tail of the lead distribution causes the number of children with lead toxicity to be underestimated by approximately 144,000 children. In addition, the random error just from using the lognormal model dictates that estimates can be no more precise than ± 190,000.

(3) The first step of the DOE model is to calculate a mean blood lead level for children (ages .5-5 years) from the amount of gasoline lead being used. Standard regression analysis shows that the mean blood lead level can be estimated no better than about ±2 ug/dl at a 95% prediction confidence interval. This random error then results in 10 to 30 fold differences in calculations of the number of children with lead toxicity.

(4) The second step of the DOE model estimates the number of children with lead toxicity from the mean blood lead level. Even at low mean blood lead levels, the amount of random error associated with this step alone causes the estimate of the number of children with lead toxicity to be no more precise than approximately 48,000 to 560,000 children (95% prediction confidence interval). This random error would be in addition to those noted in items (2) and (3).

(5) A more direct analysis which makes no distribution assumptions, utilizes the entire NHANES II lead database, and accounts for the complex survey design, indicates that estimating the number of children with lead toxicity

from gasoline lead values results in a consistently large random error. This random error decreases somewhat as the amount of lead in gasoline decreases. But even for low gasoline lead levels, the 95% prediction confidence interval for the number of children with lead toxicity is still at least as wide as from 45,000 to 500,000 children!

(6) A cutoff value of 30 micrograms per deciliter (ug/dl) to define "lead toxicity" almost certainly underestimates the number of children having harmful health effects from lead.

Further discussion of these points:

(1) Prediction of the DOE model disagrees with the CDC lead screening data.

CDC lead screening data from 55 cities and/or counties in the U.S.A. provides useful external criteria to evaluate the predictive accuracy of the DOE model. NHANES II data indicated that during 1978, approximately 675,000 children .5-5 years old had lead toxicity. That same year, the CDC screening program screened 399,000 children uncovering 26,000 with lead toxicity. During 1981, the screening program screened 525,000 children and found 22,000 with lead toxicity. After adjusting for an increased screening population, the 1981 screening data strongly suggest a population of at least 400,000 children with lead toxicity in 1981. This is in sharp contrast to only 40,000 children predicted from the DOE model. The DOE model certainly is markedly underestimating the number of children with lead toxicity and probably by a factor of 10 according to the CDC screening data.

(2) Error in the DOE model from assuming a lognormal distribution.

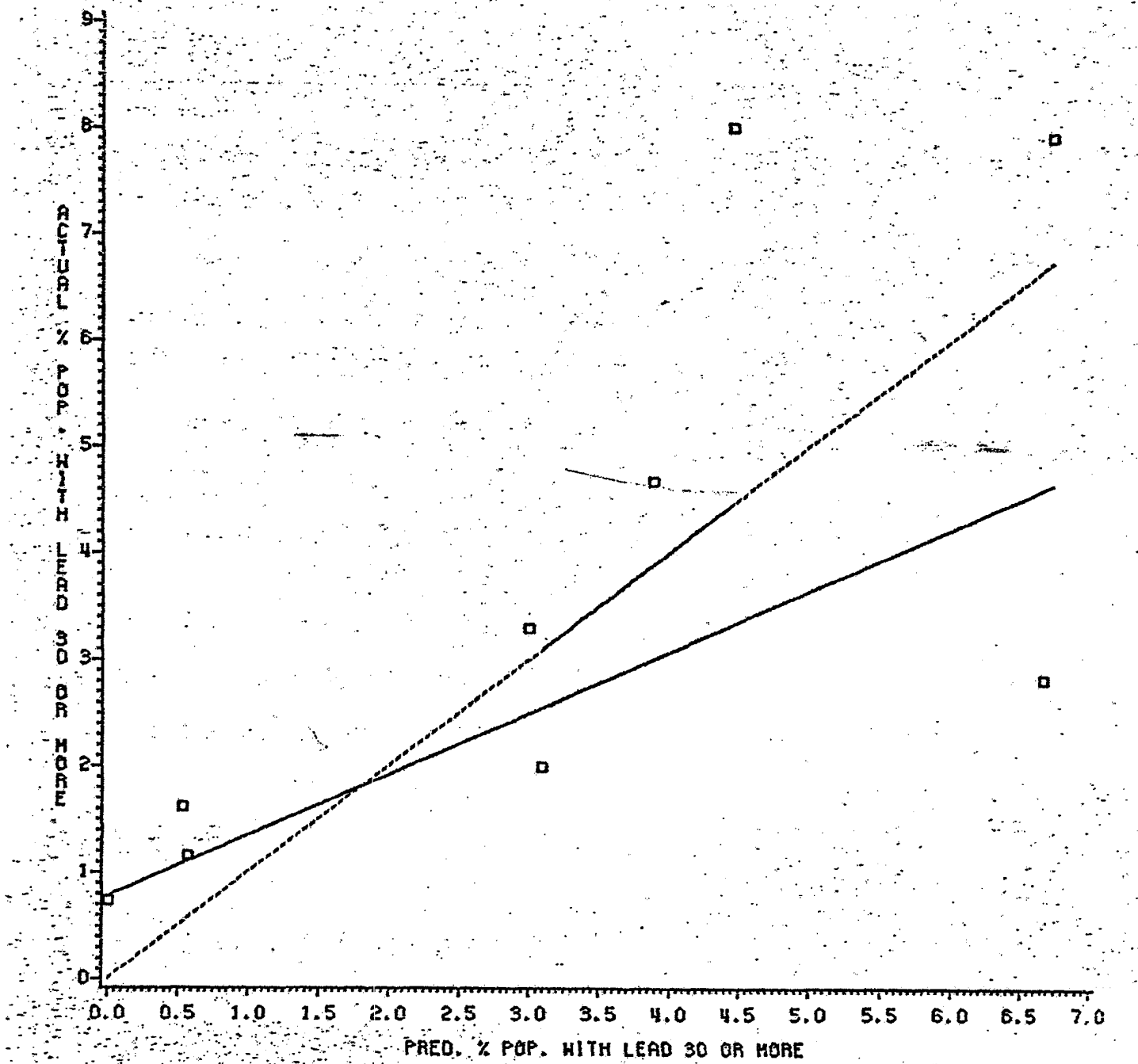
When compared to a normal or square-root-normal distribution, the lognormal distribution provides a superior fit to the NHANES II lead data. As with every model, there is some error intrinsically associated with its representation of actual data. An absolutely crucial assumption of the DOE model is that the lognormal distribution gives an accurate estimate of the tail area where lead is 30 ug/dl or more. For a given population, the NHANES II data allow the calculation of how much error is associated with making this assumption that a lognormal distribution fits the tail. Since children (defined as .5-5 years old) are the most sensitive population in relation to lead's harmful effect, the analysis presented here shows how well a lognormal fit predicts the number of children with lead levels of 30 ug/dl or more (i.e. have lead toxicity).

Figure 1 shows the actual versus the lognormal-predicted percent of children with lead toxicity for January-June and July-December 6-month periods over the duration of the NHANES II survey. So, for example, the y-value of one point is the actual percent of children found to have lead toxicity from January 1976 through June 1976 and its corresponding x-value is the percent of children predicted by the lognormal distribution assumption. The other points are determined similarly but during other 6-month periods.

FIGURE 1

ACTUAL VS PREDICTED PERCENT POPULATION WITH LEAD 30 OR MORE

AGES .5-5 YEARS
JANUARY-JUNE AND JULY-DECEMBER INTERVALS



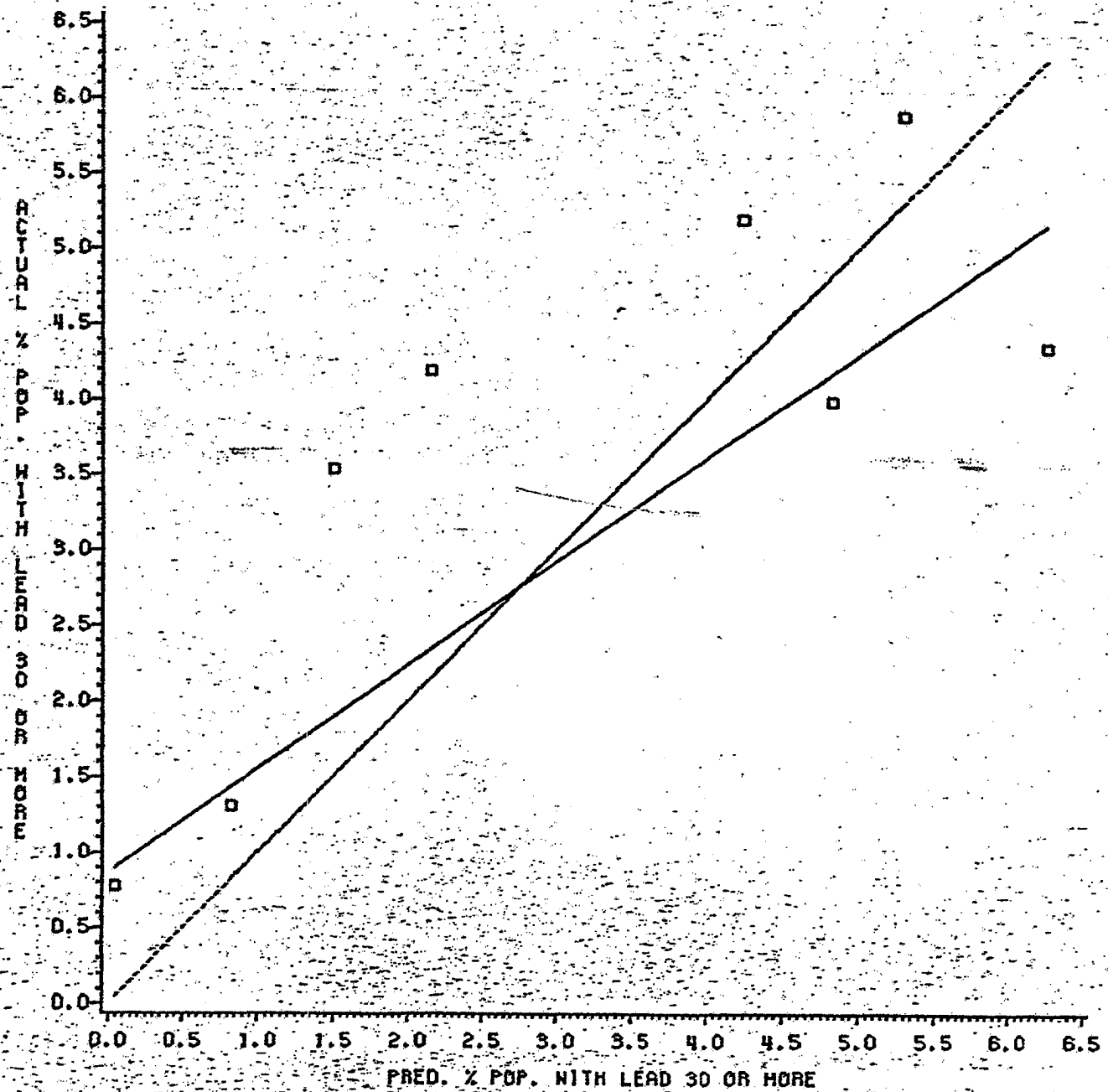
TEH 0533315

DUP050034580

FIGURE 2

ACTUAL VS PREDICTED PERCENT POPULATION WITH LEAD 30 OR MORE

AGES .5-5 YEARS
APRIL-SEPTEMBER AND OCTOBER-MARCH INTERVALS



TEH 0533316

DUP050034581

The following method was used to calculate the lognormal-predicted percent of the children with lead toxicity. The lead levels of the children were transformed by taking the natural logarithm. The mean and standard deviation of these logarithms was then calculated in log space. The percent of the population with lead of 30 or more was calculated from the following formula:

$$\text{Pct. with 30 or more} = \{1 - \text{Probnorm}[(\ln(29.5) - \text{geometric mean})/\text{geom. std. dev.}]\} \times 100$$

where probnorm is the inverse of the probit function (1).

The dashed line in Figure 1 is the ideal line if the actual and lognormal-predicted matched perfectly. The solid line is the actual weighted linear regression fit of the actual percent population measured in the NHANES II vs the lognormal prediction (see Appendix A - Weighted Linear Regression). Note that as the lognormal-predicted percent approaches zero, the actual percent population is still at 0.8%. The 1980 census data indicate that the U.S. population has slightly more than 18,000,000 children ages .5-5 years. So as the lognormal model predicts nearly zero children at toxic levels, the best fit from the NHANES II data shows that the lognormal prediction will be underestimating the actual number by about 144,000.

In addition to the above mentioned systematic error, there is random error associated with predicting with the lognormal distribution. The weighted regression fit of the actual on the predicted (as shown in Figure 1) calculated 95%-prediction confidence intervals for the points plotted. The confidence intervals ranged from ± 1.1% to ± 8.0%. This means that at a 95% confidence interval, ANY PREDICTION OF NUMBERS OF CHILDREN WITH LEAD TOXICITY THAT RELIES ON THE ASSUMPTION OF A LOGNORMAL DISTRIBUTION WILL HAVE AN ERROR OF AT LEAST ± 190,000 CHILDREN.

Figure 2 is identical to Figure 1 except the 6-month intervals are staggered to April-September and October-March periods. The interpretation is the same as Figure 1, indicating the conclusions drawn above were not dependent on the choice of the 6-month sampling period.

Thus rather large unavoidable systematic and random errors are associated with the use of a lognormal distribution to approximate the blood lead level distribution of children.

(3) Error in the DOE model due to estimating children's mean blood lead levels from the amount of lead used in gasoline production.

The path of the DOE calculation is shown below:

lead used in gasoline $\xrightarrow{A}$ mean blood lead level of children $\xrightarrow{B}$ number of children with lead toxicity

Error in estimating the amount of lead in gasoline adds to the error in step A to produce error in the mean blood lead level estimate. The error in this

estimate in turn adds to the error in step B to produce error in the final estimate of the number of children with lead toxicity. Each source of error must be estimated along with how it propagates through the calculation.

This section deals with the error in step A. EPA takes quarterly accountings of the amount of lead used in gasoline production across the U.S. To be conservative in estimating error, it is assumed that these lead in gasoline estimates are free from error. (Any actual error in EPA estimates of gasoline lead will increase the error in estimating mean blood lead levels).

Mean blood lead levels can be estimated from lead used in gasoline by using weighted linear regression. Figure 3 illustrates the results of the weighted linear regression of mean blood lead levels on lead in gasoline (see Appendix A - Weighted Linear Regression). As in Figures 1 and 2, the total child population has been subdivided into groups which were sampled during 6-month periods over the course of NHANES II (Feb. 1976 - Feb. 1980).

The amount of lead used in gasoline during each of these 6-month periods is known from EPA reports. The error bars shown are 95% prediction confidence intervals (2). The center points of the bars are connected and represent the actually observed mean blood lead levels. Figure 4 shows an analogous graph where the 6-month intervals are staggered to April-September and October-March. An important observation from these figures is that the 95% prediction confidence interval is always at least as large as 4 ug/dl. Therefore, ANY PREDICTION OF THE MEAN BLOOD LEAD LEVEL OF CHILDREN FROM GASOLINE LEAD WILL BE NO BETTER THAN ± 2 ug/dl AT A 95% CONFIDENCE INTERVAL. Especially at the lower mean lead levels, this error results in a 10 to 30 fold difference in the estimated number of children with lead toxicity.

(4) Error in the DOE model due to estimating the number of children with lead toxicity from the mean blood lead level.

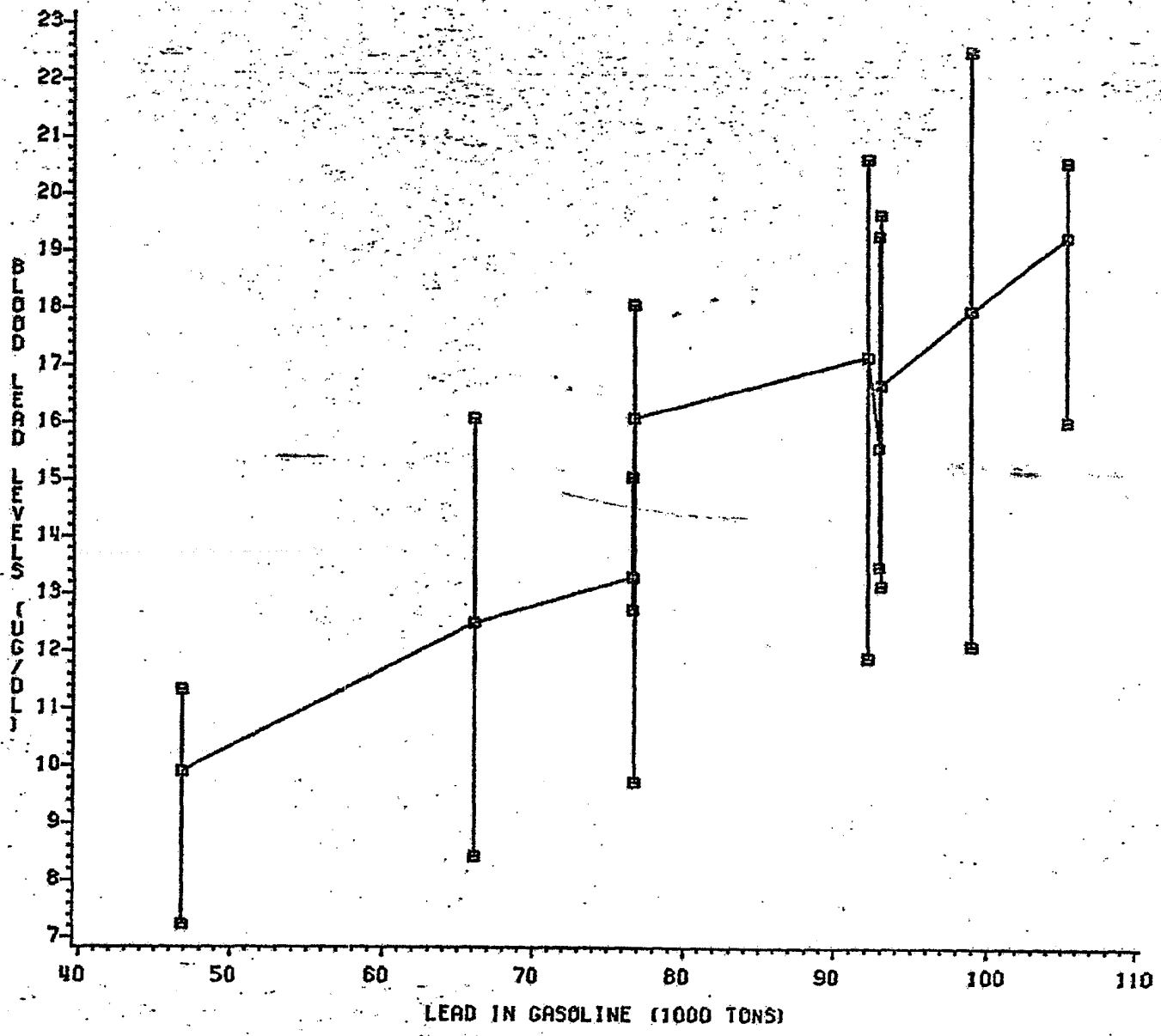
The second part of the DOE model estimates the number of children with lead toxicity from the children's mean blood lead level (step B in item (3)). Since the DOE analysis was limited to 6 percentile points of the lead distribution, the actual number of children with lead of 30 ug/dl or more could not be determined. Instead, the six points were fit to a lognormal distribution by nonlinear regression and the best fit lognormal distribution was then used to calculate how many children had elevated lead levels. (The error associated with the use of this lognormal distribution model is discussed in item (2)).

Employing <u>all</u> of the NHANES II lead data, the error due to estimating the number of children with lead toxicity from the mean blood lead level can be determined. A plot of the actual percent of children with lead toxicity versus blood lead levels revealed a nonlinear relationship. The logit transformation (3) is a common linearizing transformation used with dependent variables which are proportions. Figure 5 shows the logit of the percent of children with lead toxicity versus mean blood lead levels for populations of children defined by 6-month periods of the NHANES II survey exactly as

FIGURE 3

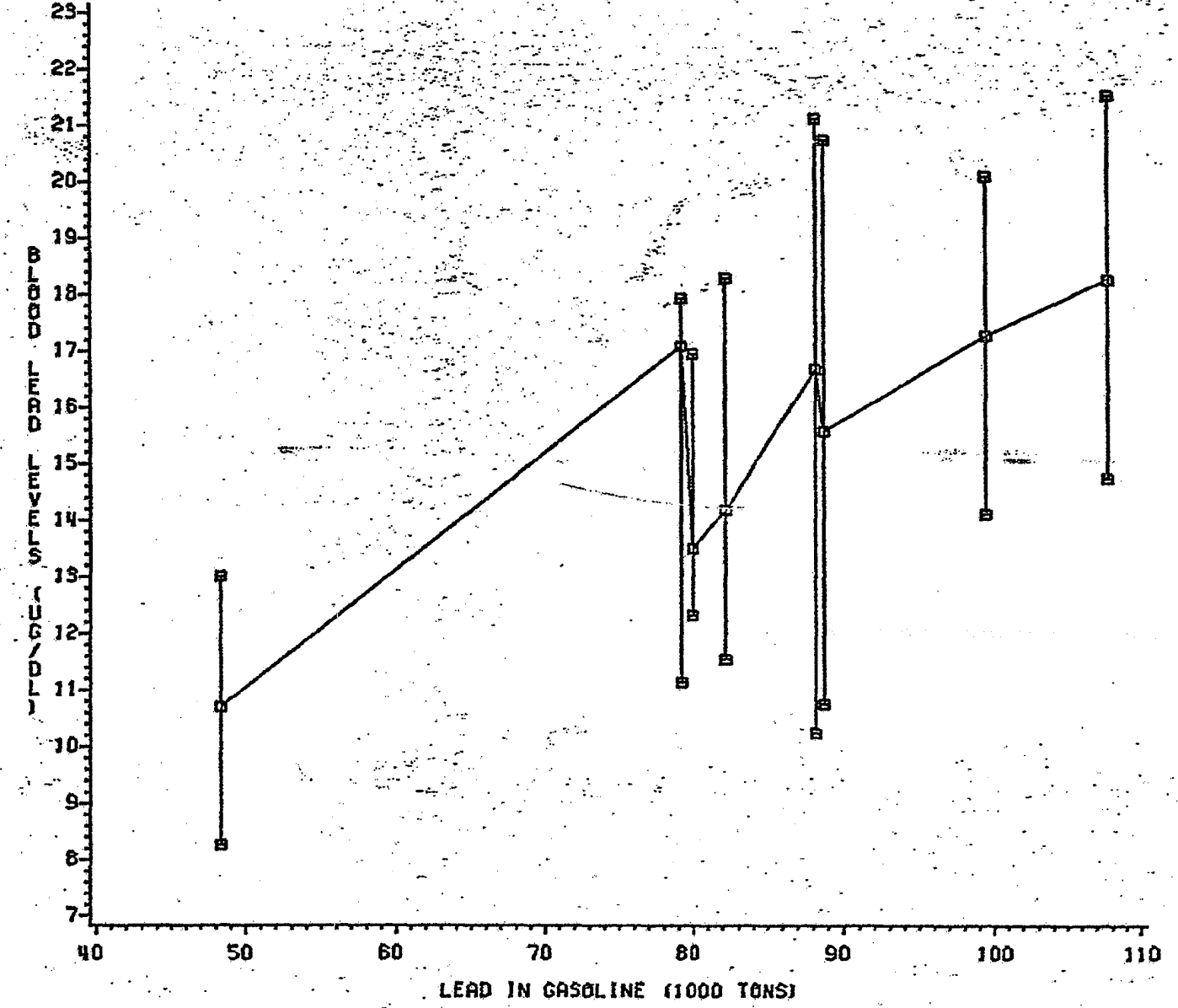
MEAN BLOOD LEAD LEVELS VS. LEAD USED IN GASOLINE

AGES .5-5 YEARS
JANUARY-JUNE AND JULY-DECEMBER INTERVALS



TEH 0533319

DUP050034584



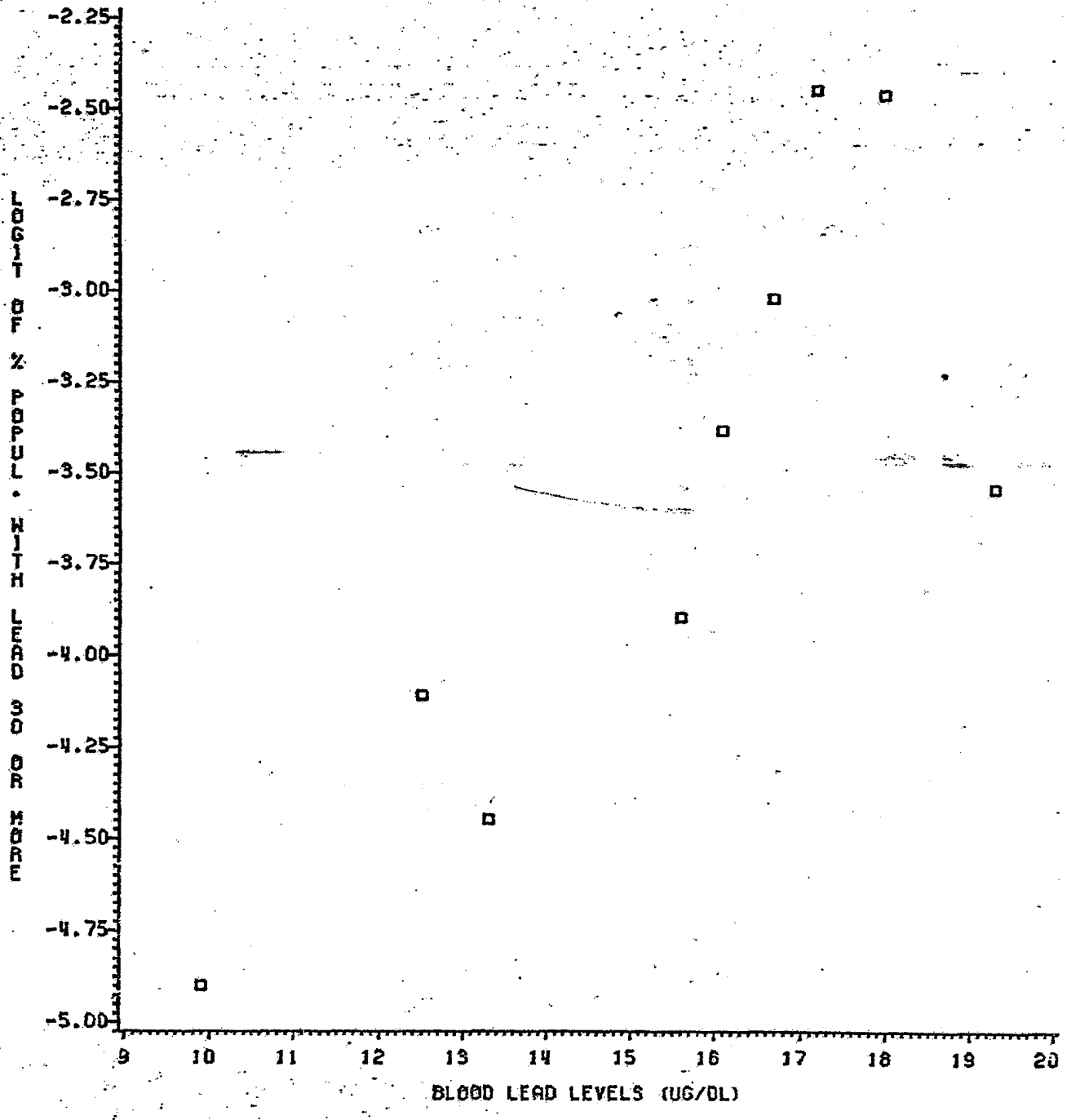
TEH 0533320

DUP050034585

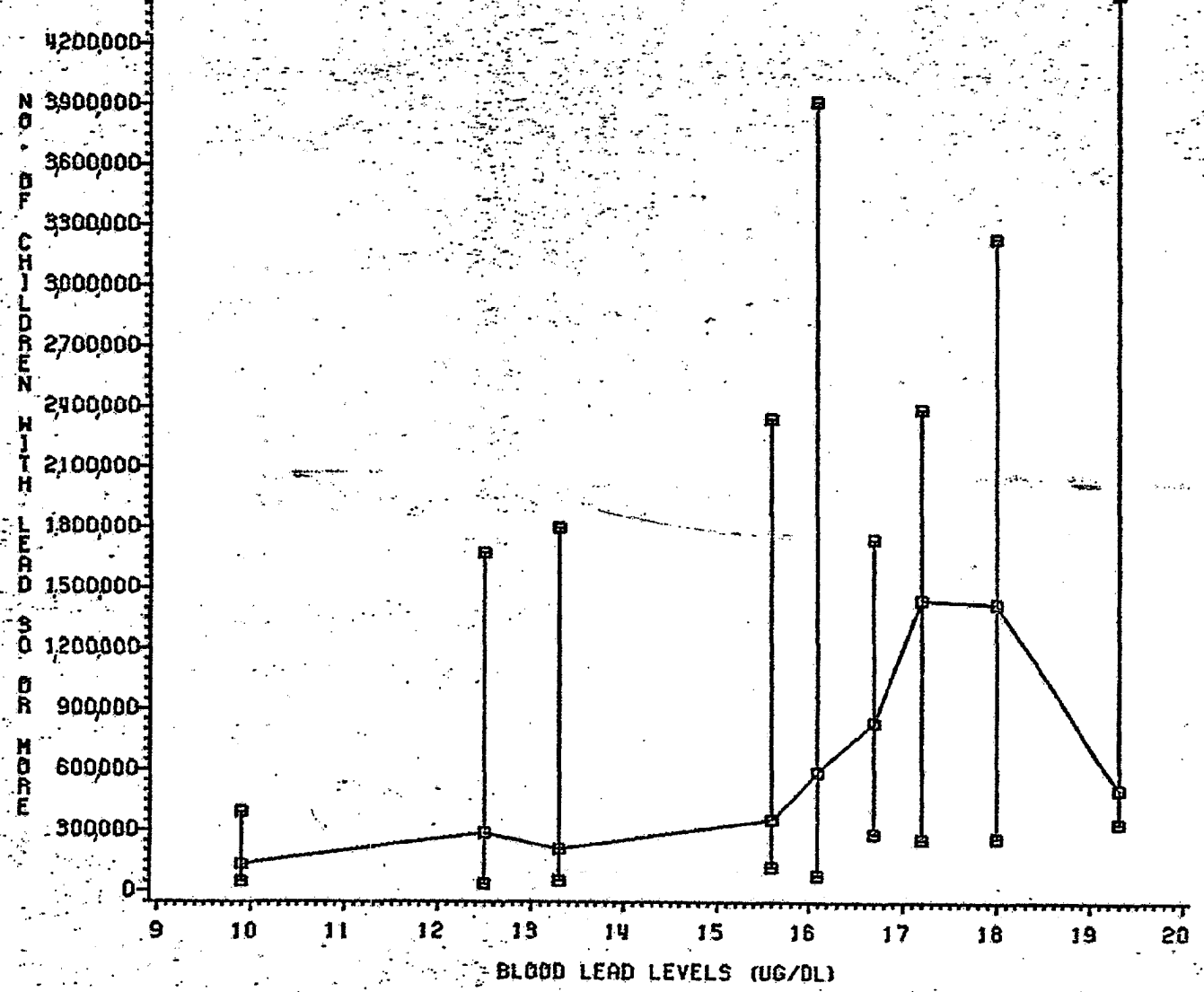
FIGURE 5

LOGIT OF % POP. WITH LEAD OF 30 OR MORE VS. MEAN BLOOD LEAD LEVELS

AGES .5-5 YEARS
JANUARY-JUNE AND JULY-DECEMBER INTERVALS



TEH 0533321



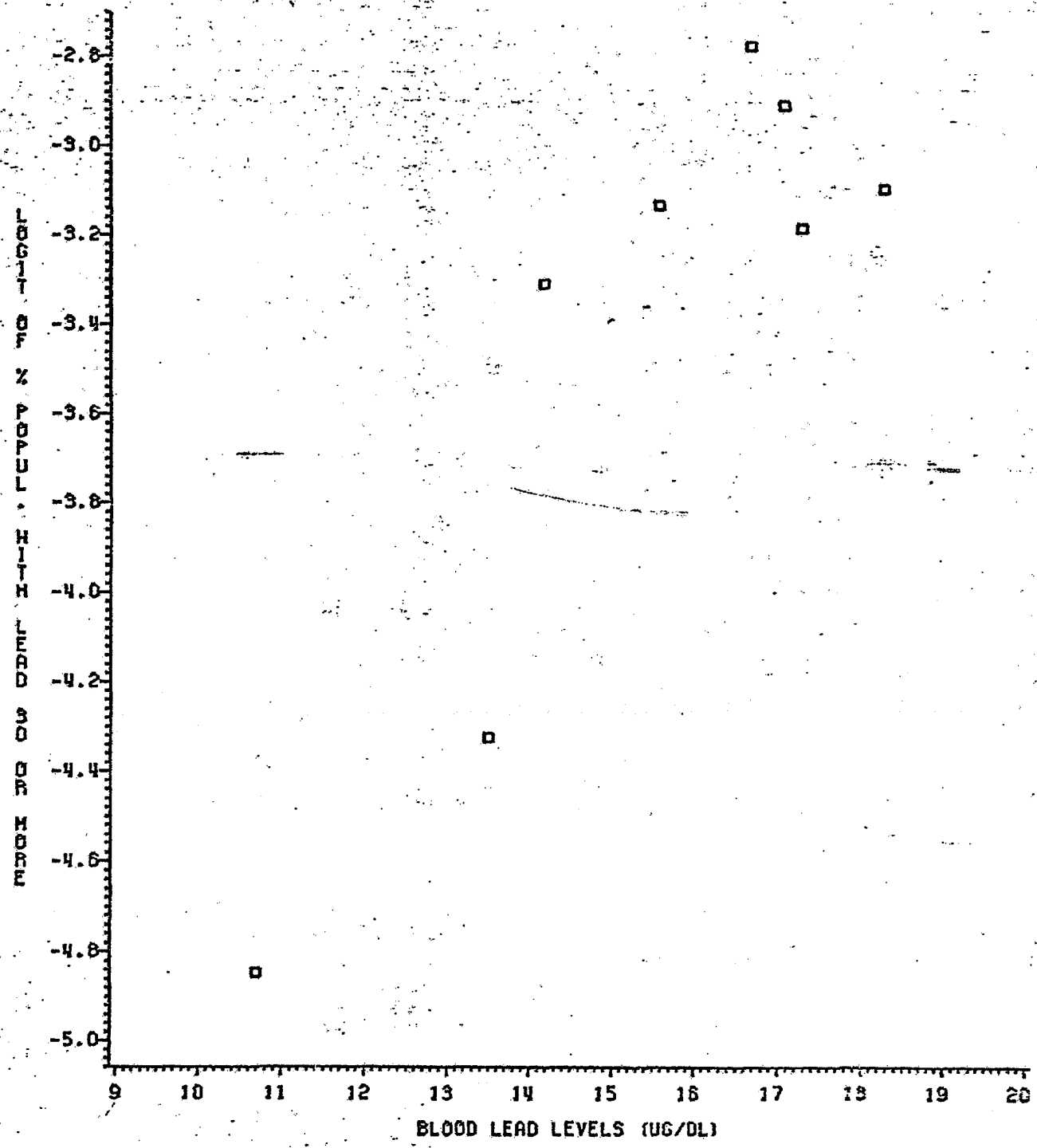
TEH 0533322

DUP050034587

FIGURE 7

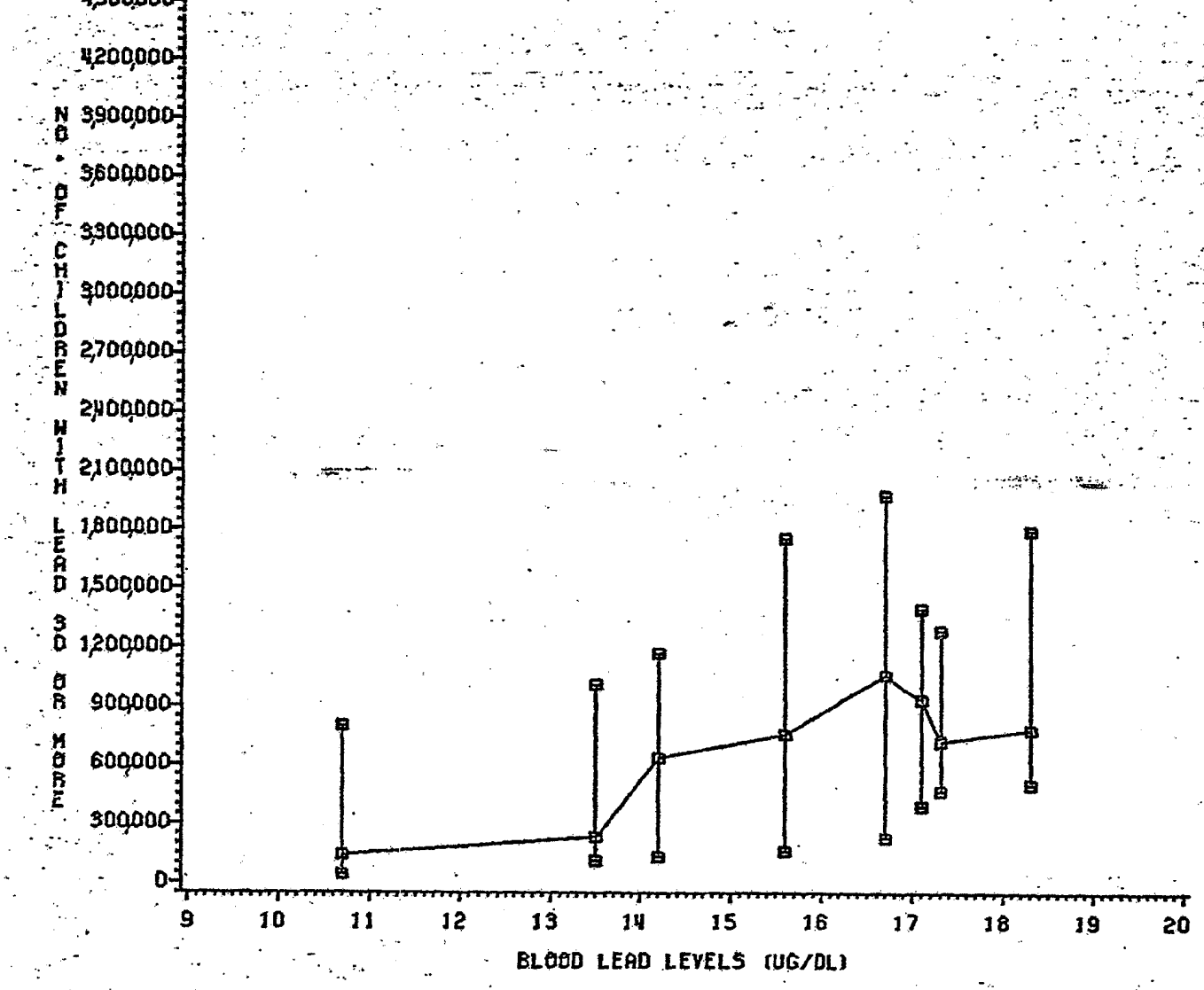
LOGIT OF % POP. WITH LEAD OF 30 OR MORE VS. MEAN BLOOD LEAD LEVELS

AGES .5-5 YEARS
APRIL-SEPTEMBER AND OCTOBER-MARCH INTERVALS



TEH 0533323

DUP050034588



TEH 0533324

DUP050034589

described in items (2) and (3). A weighted linear regression was done of the logit on the mean blood lead levels. The weighting included the influence of the complex sample design of NHANES II and the effect of the logit transformation on the variance of the observations (see Appendix B -- Weighted Linear Regression With the Logit Transformation).

From the weighted regression, the 95% prediction confidence intervals were calculated. The confidence intervals and the logits of observed percentages were back transformed to percentages. Then the percentages were multiplied by the 1980 census estimate of children .5-5 years old (i.e. 18,000,000) to arrive at a final estimate of the number of children at toxic levels along with the 95% prediction confidence interval about that estimate. These results are plotted in figure 6. The interpretation of the graph is analogous to figure 3 and 4. The bars are 95% prediction confidence intervals and the center connected points are the actual observed percentages from the NHANES II data. The 95% prediction confidence intervals are distinctly asymmetric about the observed values. Figures 7 and 8 are analogous to 5 and 6 except the 6-month period has been staggered to April-September and October-March to be sure the interval choice does not affect the results.

From figure 6 and 8, even at low mean lead levels, the 95% prediction confidence interval is at least as broad as 50,000 to 550,000 children. This is a conservative estimate of the amount of random error associated with estimating the number of children with lead toxicity from the mean blood lead level even if no lognormal distribution assumption is made.

Propagating the error from the two steps in the DOE model:

The random error in estimating the number of children with lead toxicity from mean blood lead levels assumes there is no error in the mean blood lead level estimate. From item (3), the random error in the mean blood leal level estimates is at least ± 2 ug/dl (95% prediction confidence interval). Examining figure 6 and 8 illustrates a straightforward way to estimate how much this error influences the final predicted number of children with lead toxicity. One can select a blood lead level and note the range of the confidence intervals at ± 2 ug/dl away from the level of interest. It is seen that estimating the mean blood lead level to only within + 2 ug/dl can markedly multiply the uncertainty of the final prediction of the number of children with lead toxicity.

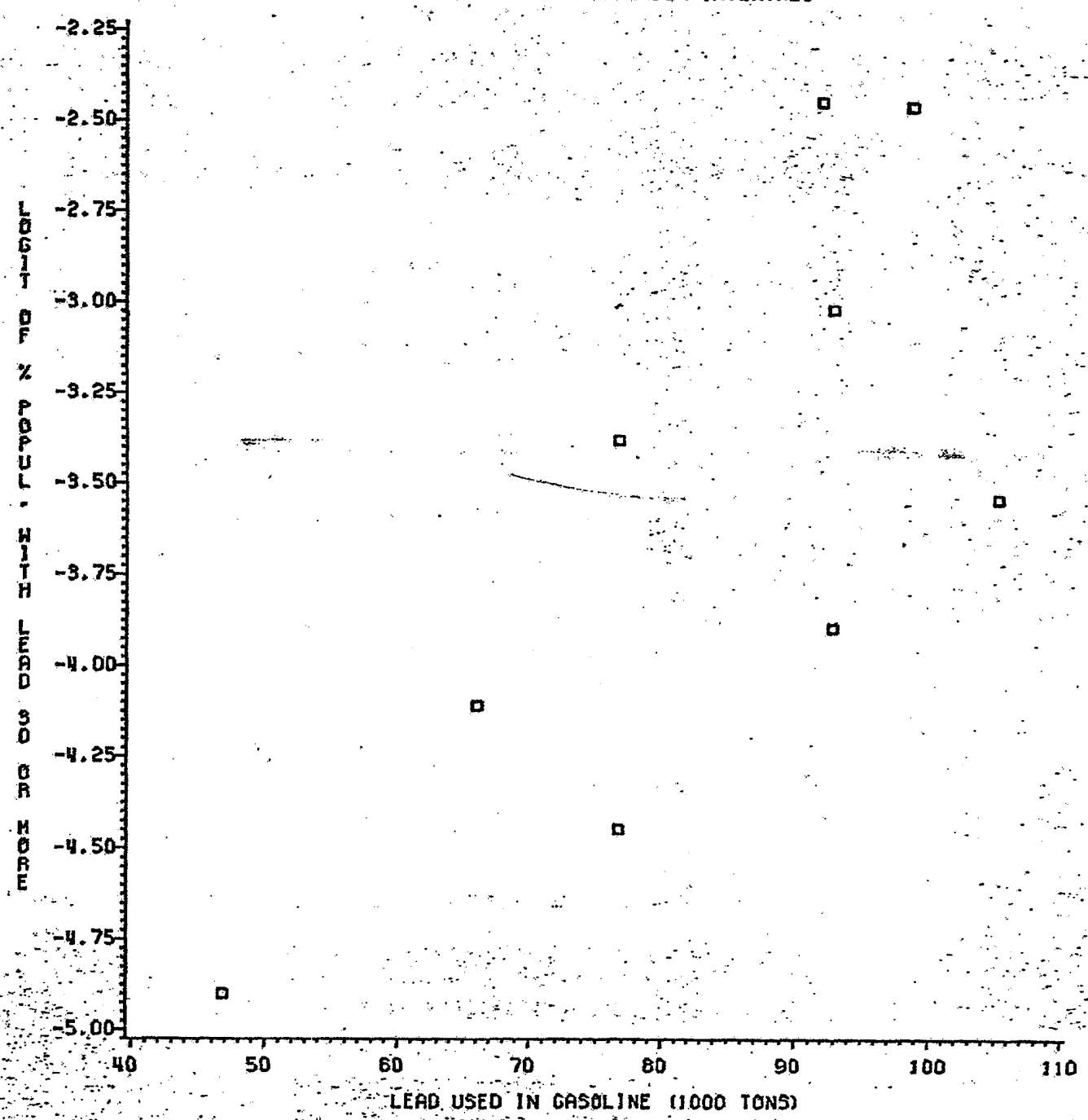
(5) A more direct analytical approach which makes no distribution assumption, utilizes the entire NHANES II lead data base and accounts for the complex sample design indicates that using gasoline lead values to predict the number of children with lead toxicity results in a consistently large random error.

A direct analysis was performed which was not encumbered with any distribution assumption or with the need for multiple-step propagation of error analysis. The basis of the approach is to start with the amount of lead used in gasoline and determine in one step how many children had toxic levels from the NHANES II lead data. First, the NHANES II data was divided into 6-month periods as described in items (2) and (3). The amount of lead in gasoline during each period was known from EPA reports. The percent of children with lead toxicity

FIGURE 9

LOGIT OF % OF POPUL. WITH LEAD OF 30 OR MORE VS. GASOLINE LEAD

AGES .5-5 YEARS
JANUARY-JUNE AND JULY-DECEMBER INTERVALS



TEH 0533326

DUP050034591

for the period was calculated directly from the NHANES II data. A plot of percent of children with lead toxicity versus lead used in gasoline showed a nonlinear relationship. As in item (2), the logit transformation was chosen to linearize the data because of its established usefulness in linearizing proportions. Figure 9 shows a plot of the logit of the percent of children with lead toxicity versus the amount of lead used in gasoline. A weighted linear regression was done of the logit on lead used in gasoline. The weighting included the influence of the complex sample design of NHANES II and the effect of the logit transformation on the variances of the observations (see Appendix B -- Weighted Linear Regression with the Logit Transformation).

From the weighted regression, the 95% prediction confidence intervals were calculated. The confidence intervals and the logits of observed percentages were then back transformed to percentages. These percentages were multiplied by the 1980 census estimate of children .5-5 years old (i.e. 18,000,000) to arrive at a final estimate of how many children have toxic lead levels and the 95% prediction confidence interval about this estimate. The result is shown in Figure 10. The bars are 95% prediction confidence intervals and the center connected points are the actual observed percentages from the NHANES II data. Figures 11 and 12 are analogous to 9 and 10 except the 6-month period has been staggered to April-September and October-March to be sure the interval choice doe not affect the results.

Figure 10 shows that even at low gasoline lead levels the 95% prediction confidence interval is at least as large as 45,000 to 500,000 children. Figure 12 shows a very large confidence interval at low gasoline lead levels extending up over 2,500,000 children. It is interesting to note the variation in observed values and confidence intervals across the domain of gasoline lead values. This variability is most likely due to the small sample sizes of children with lead toxicity which determine the percentage estimates. These data do not provide a clear extrapolation to lower gasoline lead values.

In summary, after eliminating distribution assumptions and the need for muliple-step propagation of error analysis, the 95% prediction confidence interval for the number of children with lead toxicity is still at least as wide as 45,000 to 500,000 children, accounting only for random error.

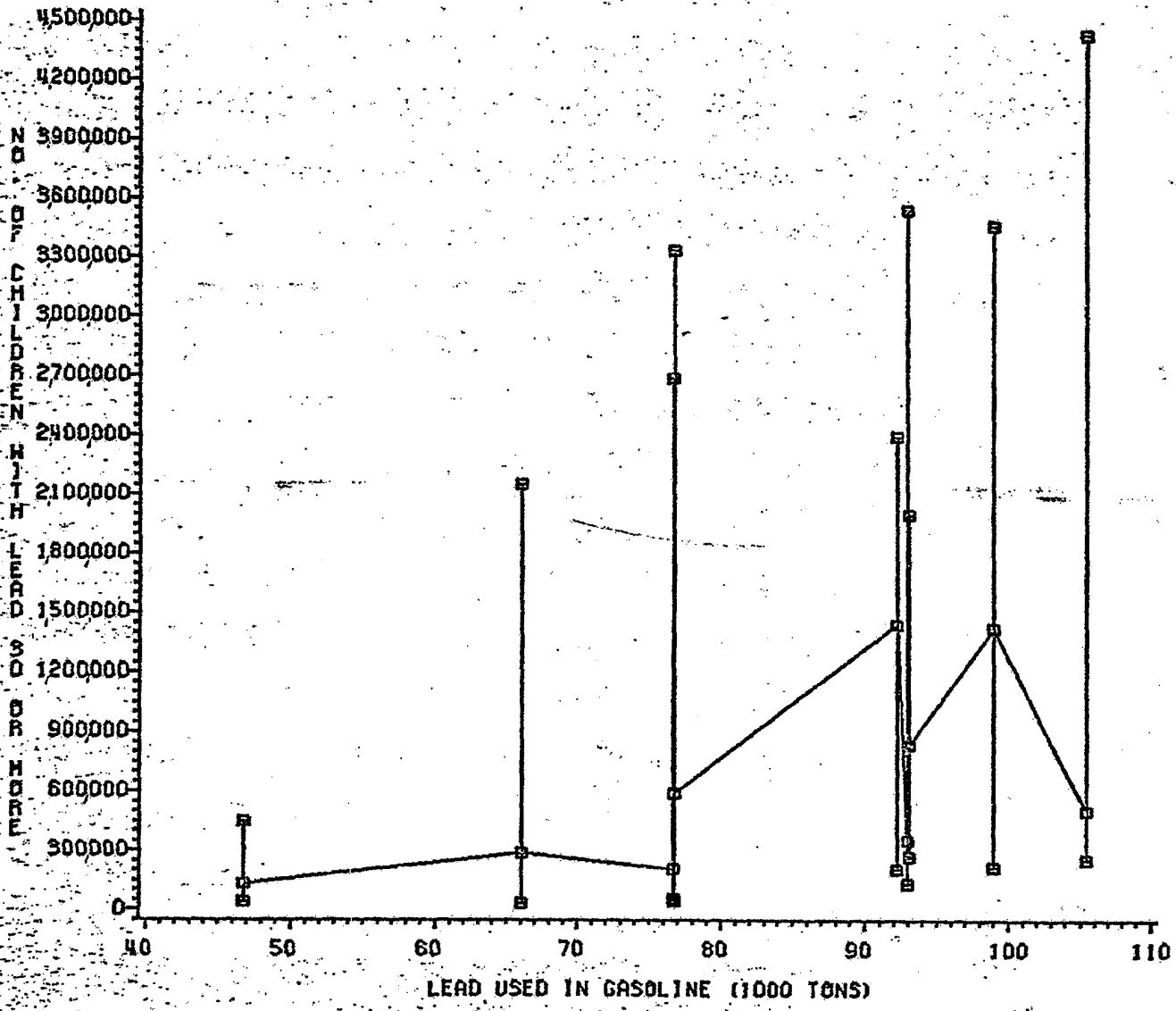
(6) Underestimating the number of children with "lead toxicity" from selecting too high a cutoff blood lead level in defining "toxic".

Mounting scientific evidence indicates that levels of lead previously considered safe actually may be harmful to human health. This evidence has been extensively presented to EPA. The previous toxic threshold value of 30 ug/dl is almost certainly too high and children are probably affected by lead at lower levels. If the calculation presented in this paper are redone with lower definitions of toxicity (e.g. 20 or 25 ug/dl), the number of children with "lead toxicity" will increase by roughly 5 to 10 fold.

FIGURE 10

NO. OF PERSONS WITH LEAD OF 30 OR MORE VS. GASOLINE LEAD

AGES .5-5 YEARS
JANUARY-JUNE AND JULY-DECEMBER INTERVALS



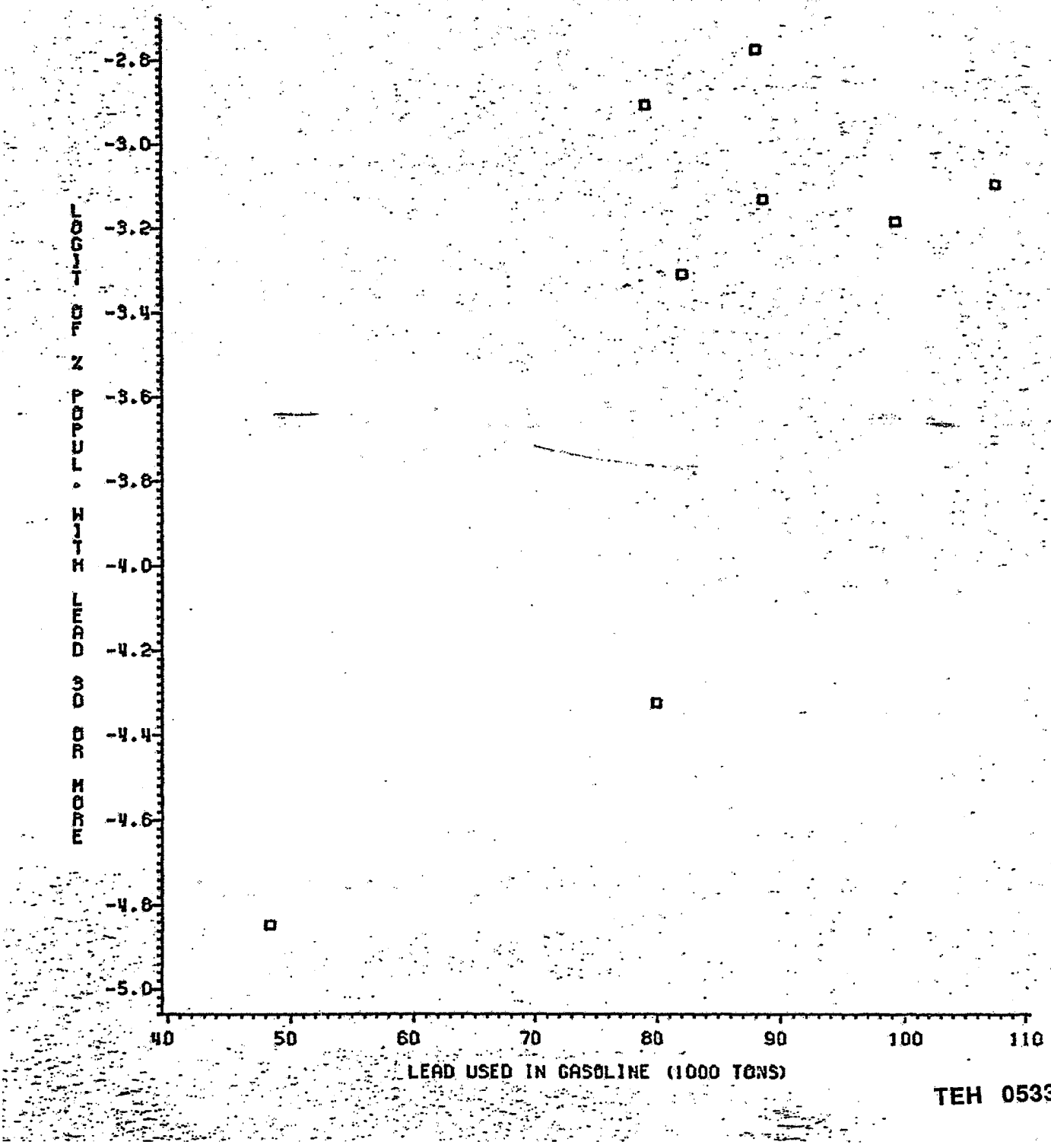
TEH 0533328

DUP050034593

FIGURE 11

LOGIT OF % POPUL. WITH LEAD OF 30 OR MORE VS. GASOLINE LEAD

AGES .5-5 YEARS
APRIL-SEPTEMBER AND OCTOBER-MARCH INTERVALS



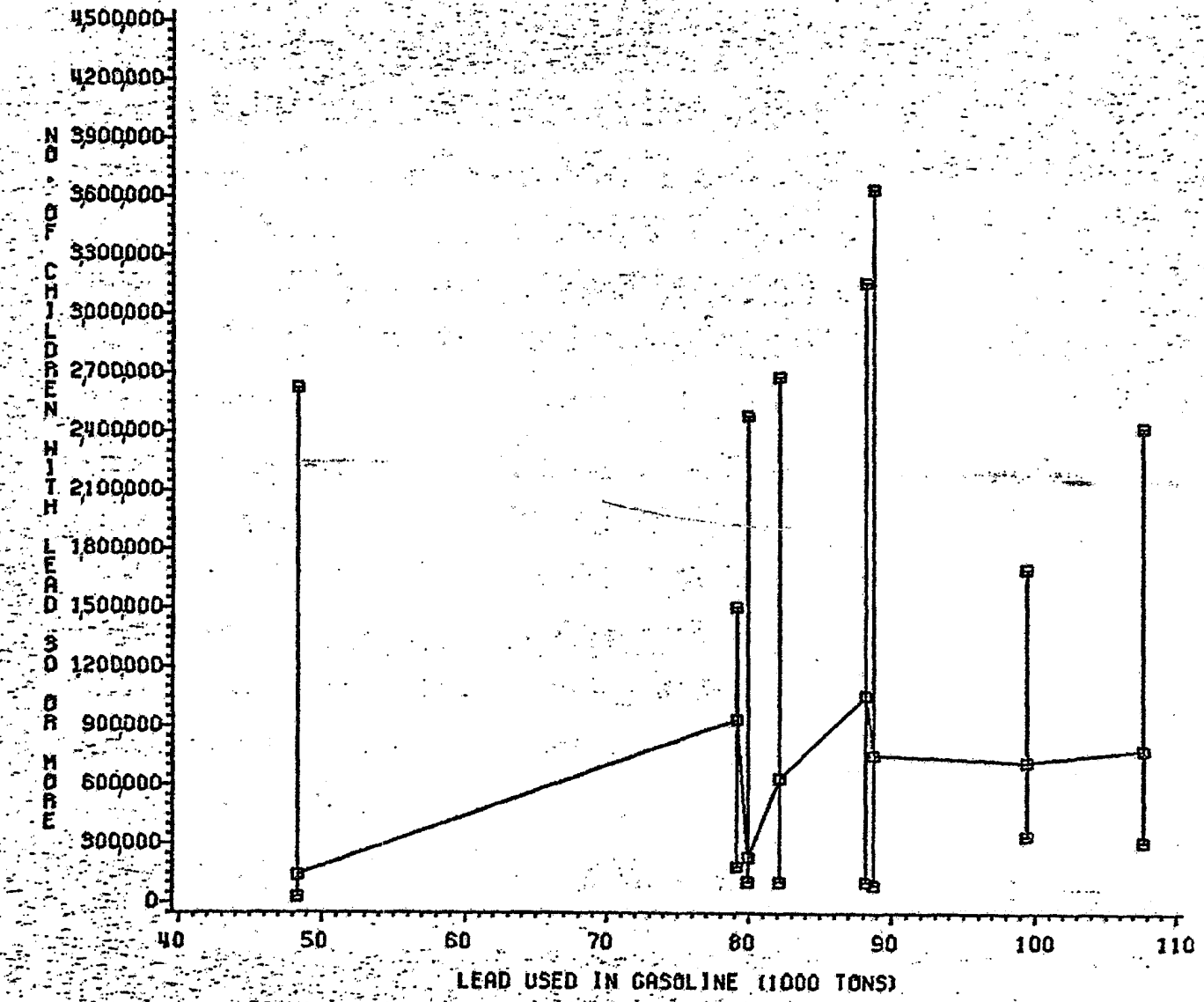
TEH 0533329

DUP050034594

FIGURE 12

NO. OF PERSONS WITH LEAD OF 30 OR MORE VS. GASOLINE LEAD

AGES .5-5 YEARS
APRIL-SEPTEMBER AND OCTOBER-MARCH INTERVALS



TEH 0533330

DUP050034595

Appendix A - Weighted Linear Regression

In figures 1 through 4, results from weighted linear regression needed to be done in order to achieve homoscedasticity. In order to obtain the best linear unbiased estimates, the weighting used was the reciprocal of the variance. The variance of the dependent variables in these graphs (mean lead levels and percentage of children with lead of 30 ug/dl or more) were calculated using the computer programs SESUDAAN and SUKREGR and take into account the complex sample design.

Appendix B – Weighted Linear Regression with the Logit Transformation

The logit transformation of percentages was used several times in this analysis. It is important to account for the effect of the transformation on the variance of the observations so proper weights can be calculated for weighted linear regression. In all of the regressions done in this analysis, the weighting took into account the complex sample design and the effect of the logit transformation on the variance of the percentages. A derivation of the variance of the logit transformation is given below:

Recall Taylor's formula $f(x) = f(a) + (x-a) f'(a) + \ldots$

Applying Taylor's formula to any random variable z and taking "a" to be $E(z)$, gives for any transformation:

$$f(z) \doteq E + (z-E) f'(E) \quad \text{where} \quad E = E(z) \qquad (1)$$

Recalling the formula for variance:

$$\text{Var}(f(z)) = E((f(z) - E(f(z)))^2) \qquad (2)$$

Substituting (1) into the right side of equation (2) and reducing yields:

$$\text{Var}(f(z)) \doteq f'(E)^2 \, \text{Var}(z) \qquad (3)$$

Applying (3) to the $\text{logit}(\hat{p})$ where $\hat{p}$ is a percentage:

$$\text{Var}(\text{logit}(\hat{p})) \doteq \left[\left(\frac{d \log\left(\frac{p}{1-p}\right)}{dp} \right|_{p = E(\hat{p})} \right]^2 \text{Var}(\hat{p}) \qquad (4)$$

After working through the derivative, (4) reduces to

$$\text{Var}(\text{logit}(\hat{p})) \doteq \text{Var}(\hat{p}) / (\hat{p}^2 (1-\hat{p})^2) \qquad (5)$$

The weighting was then done by reciprocal variance.
$\text{Var}(\hat{p})$ was calculated as described in Appendix A.

REFERENCES

1) Statistical Analysis System Manual, 1979 Edition, p.443

2) Applied Linear Statistical Models, John Neter and William Wasserman, (1974) p. 69

3) Ibid. p. 330