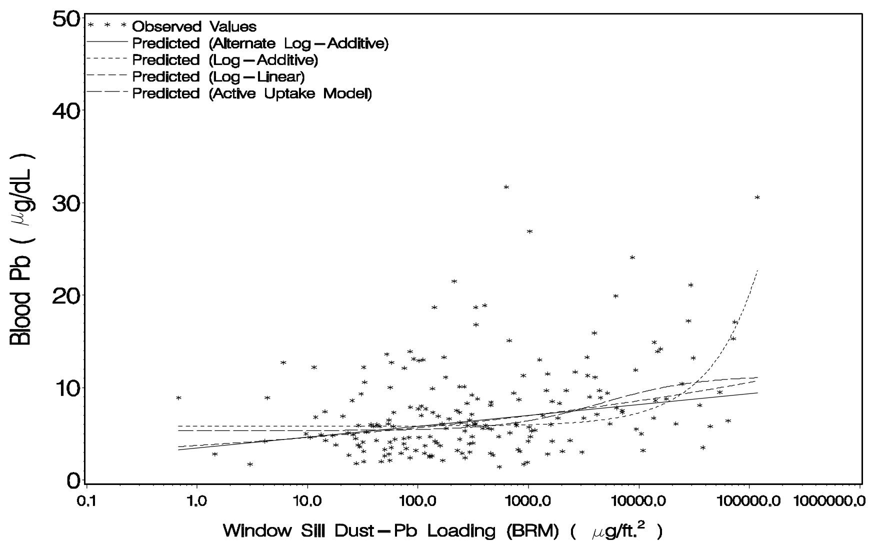
| Statistical Model | Parameter Estimates and (Associated Standard Errors) <br> $^0$ s.e. ( $_0$) | <br> $^1$ s.e. ( $_1$) | <br> $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | $^2$ | $R^2$ | Estimated Log-Likelihood | Number of Observations |
|---|---|---|---|---|---|---|---|
| Log-Linear | 1.32 (0.1114) | 0.09 (0.0177) | -- | 0.3382 | 0.1176 | -171.75 | 197 |
| Log-Additive | 5.84 (0.2692) | 0.00 (0.0000) | -- | 0.3460 | 0.0974 | -173.98 | 197 |
| Alternate Log-Additive | 3.46 (0.6138) | 0.51 (0.1123) | -- | 0.3428 | 0.1055 | -173.09 | 197 |
| Active Uptake | 10.15 (1.7822) | 0.00 (0.0041) | 11.29 (2.0248) | 0.3389 | 0.1249 | -170.94 | 197 |



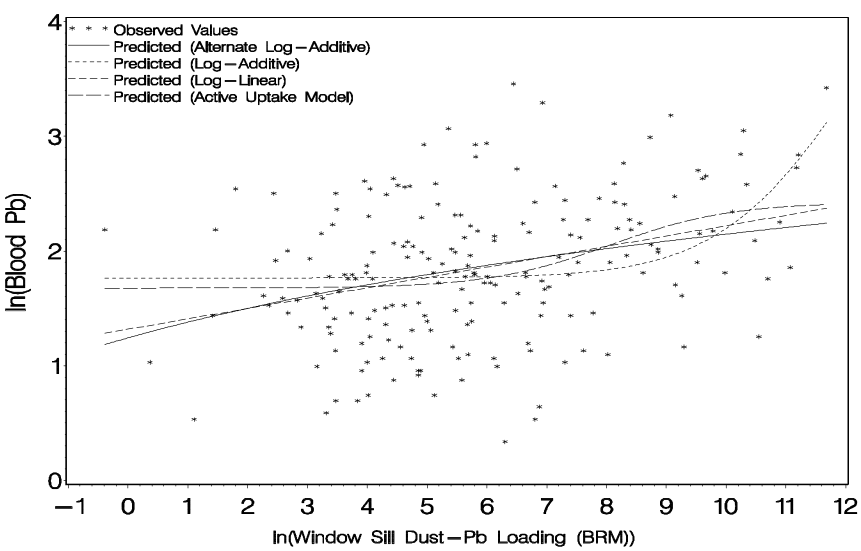




**Figure G11-13. Bivariate Relationship Between Blood-Lead Concentration and Window Sill Dust-Lead Loading (BRM Samples).**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0<br>s.e. ( $_0$) | 1<br>s.e. ( $_1$) | $\theta_{HSP}$<br>s.e. ($\theta_{HSP}$) | 2 | $R^2$ | Estimated Log-Likelihood | Number of Observations |
| Log-Linear | 1.08<br>(0.1444) | 0.08<br>(0.0138) | -- | 0.3366 | 0.1402 | -164.29 | 189 |
| Log-Additive | 5.54<br>(0.2938) | 0.00<br>(0.0000) | -- | 0.3530 | 0.0983 | -168.78 | 189 |
| Alternate Log-Additive | 2.41<br>(0.6968) | 0.40<br>(0.0758) | -- | 0.3414 | 0.1279 | -165.63 | 189 |
| Active Uptake | 9.35<br>(1.6300) | 0.00<br>(0.0001) | 9.83<br>(1.4255) | 0.3331 | 0.1581 | -162.30 | 189 |

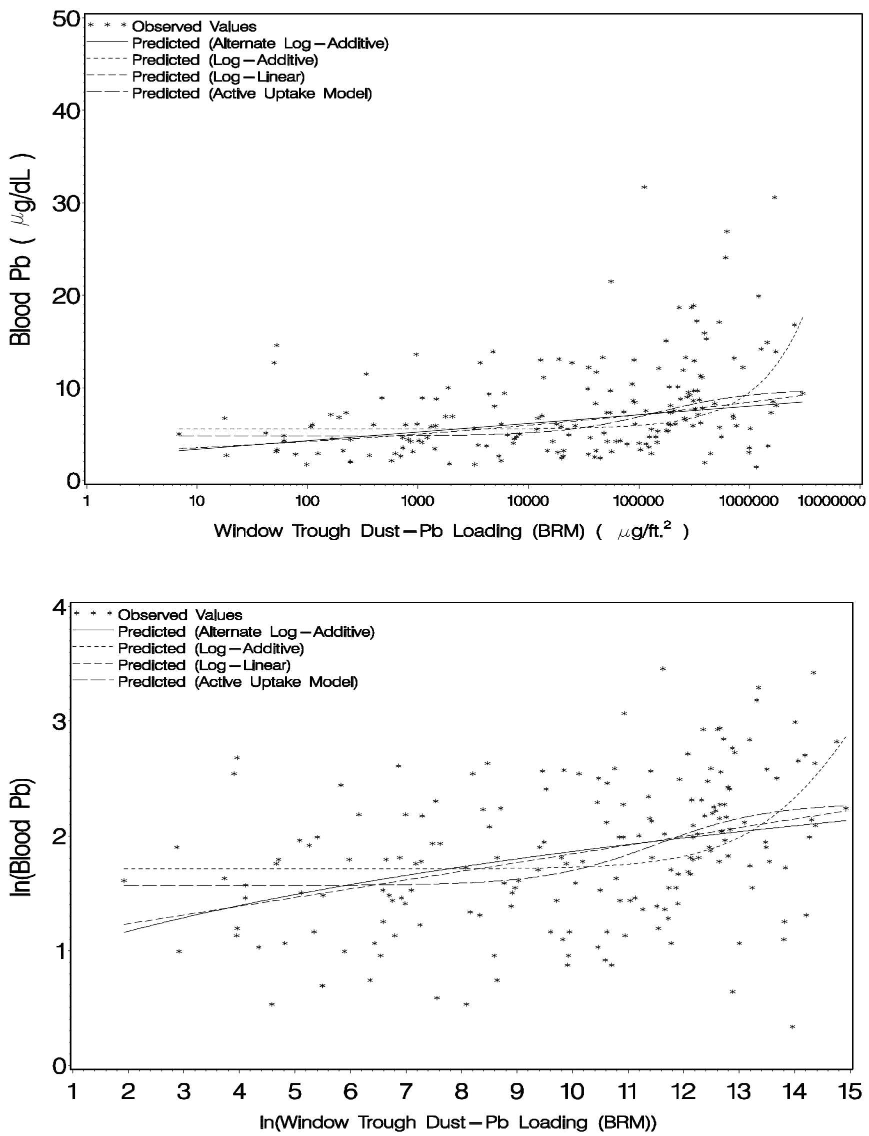
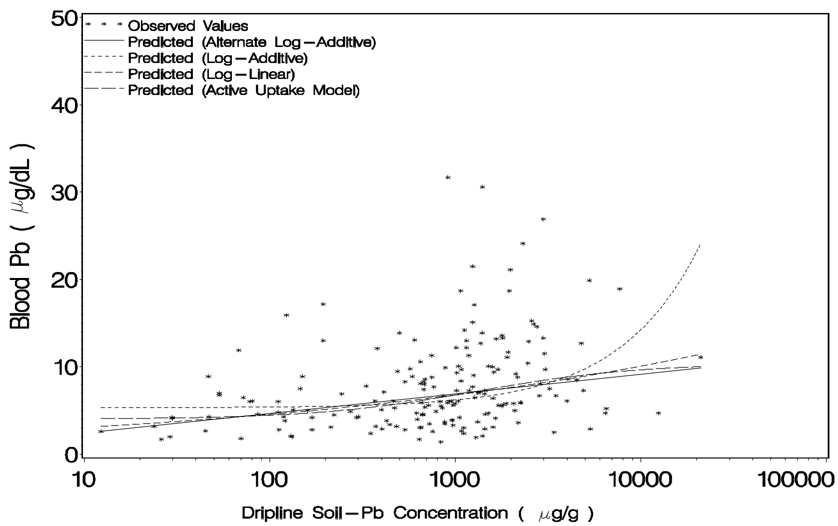



**Figure G11-14. Bivariate Relationship Between Blood-Lead Concentration and Window Well Dust-Lead Loading (BRM Samples).**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) | | | 2 | $R^2$ | Estimated Log-Likelihood | Number of Observations |
|---|---|---|---|---|---|---|---|
| | 0 s.e. ( 0) | 1 s.e. ( 1) | $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | | | | |
| Log-Linear | 0.73 (0.2216) | 0.17 (0.0330) | -- | 0.3410 | 0.1288 | -162.86 | 186 |
| Log-Additive | 5.31 (0.3700) | 0.00 (0.0003) | -- | 0.3647 | 0.0683 | -169.11 | 186 |
| Alternate Log-Additive | 0.16 (1.0046) | 0.97 (0.1655) | -- | 0.3414 | 0.1278 | -162.97 | 186 |
| Active Uptake | 6.53 (1.4373) | 0.01 (0.0092) | 10.40 (2.1910) | 0.3443 | 0.1299 | -162.75 | 186 |



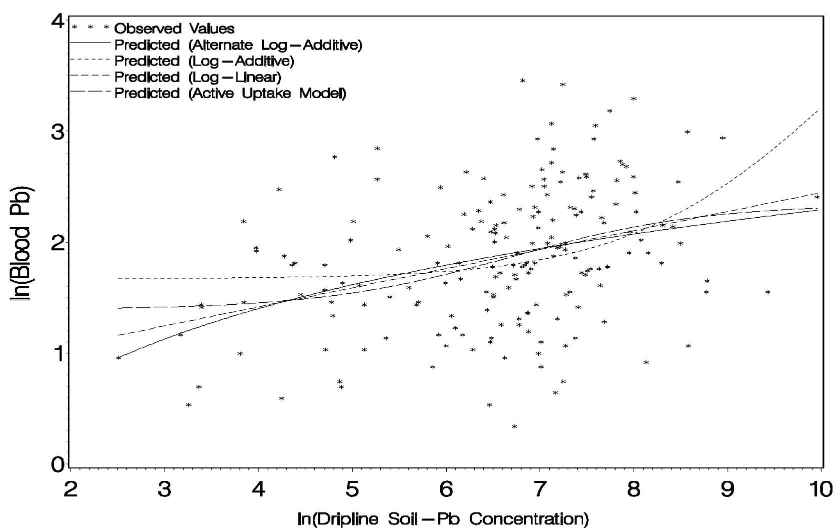
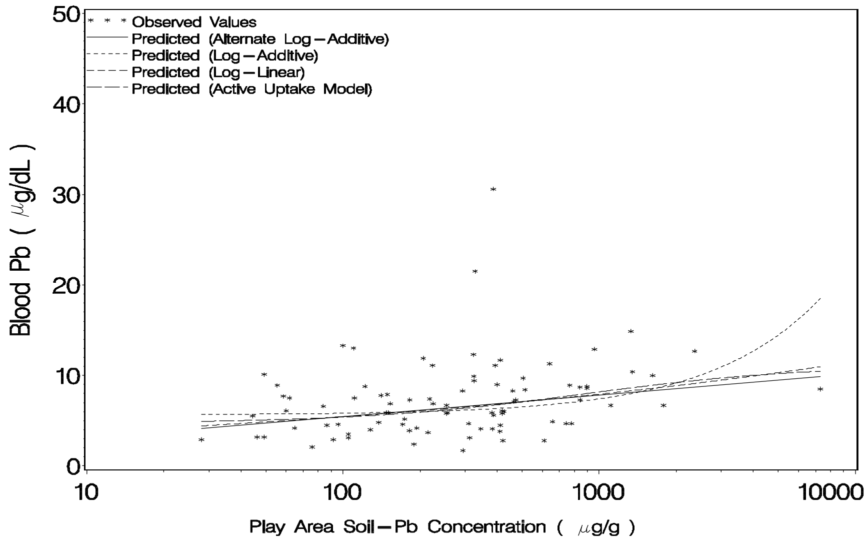




**Figure G11-15. Bivariate Relationship Between Blood-Lead Concentration and Dripline Soil-Lead Concentration.**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) | | | | | Estimated Log-Likelihood | Number of Observations |
|---|---|---|---|---|---|---|---|
| | $^0$ s.e. ( $_0$) | $^1$ s.e. ( $_1$) | $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | $^2$ | $R^2$ | | |
| Log-Linear | 0.94 (0.2941) | 0.16 (0.0518) | -- | 0.2419 | 0.1044 | -60.71 | 87 |
| Log-Additive | 5.65 (0.4364) | 0.00 (0.0008) | -- | 0.2528 | 0.0642 | -62.62 | 87 |
| Alternate Log-Additive | 0.74 (1.7617) | 1.03 (0.3294) | -- | 0.2422 | 0.1032 | -60.77 | 87 |
| Active Uptake | 8.23 (2.5304) | 0.02 (0.0293) | 11.15 (4.5624) | 0.2475 | 0.1050 | -60.69 | 87 |



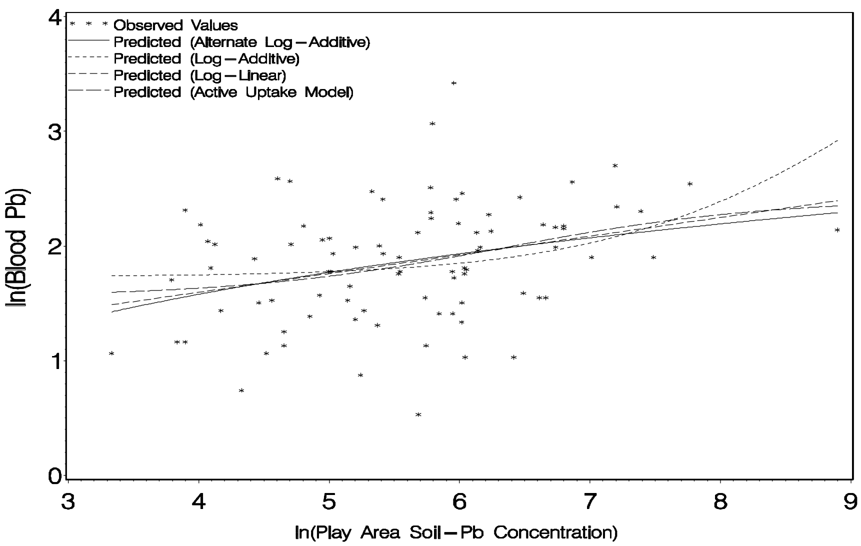




**Figure G11-16. Bivariate Relationship Between Blood-Lead Concentration and Play Area Soil-Lead Concentration.**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) $_0$ s.e. ( $_0$) | $_1$ s.e. ( $_1$) | $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | $^2$ | $R^2$ | Estimated Log-Likelihood | Number of Observations |
|---|---|---|---|---|---|---|---|
| Log-Linear | 3.15 (0.7032) | 0.66 (0.1487) | -- | 0.3483 | 0.0903 | -180.00 | 205 |
| Log-Additive | 6.14 (0.2867) | 0.00 (0.0002) | -- | 0.3752 | 0.0203 | -187.53 | 205 |
| Alternate Log-Additive | 1.31 (0.1269) | 0.11 (0.0238) | -- | 0.3476 | 0.0922 | -179.79 | 205 |
| Active Uptake | 9.70 (2.1860) | 0.06 (0.0491) | 9.29 (1.2981) | 0.3488 | 0.0981 | -179.14 | 205 |

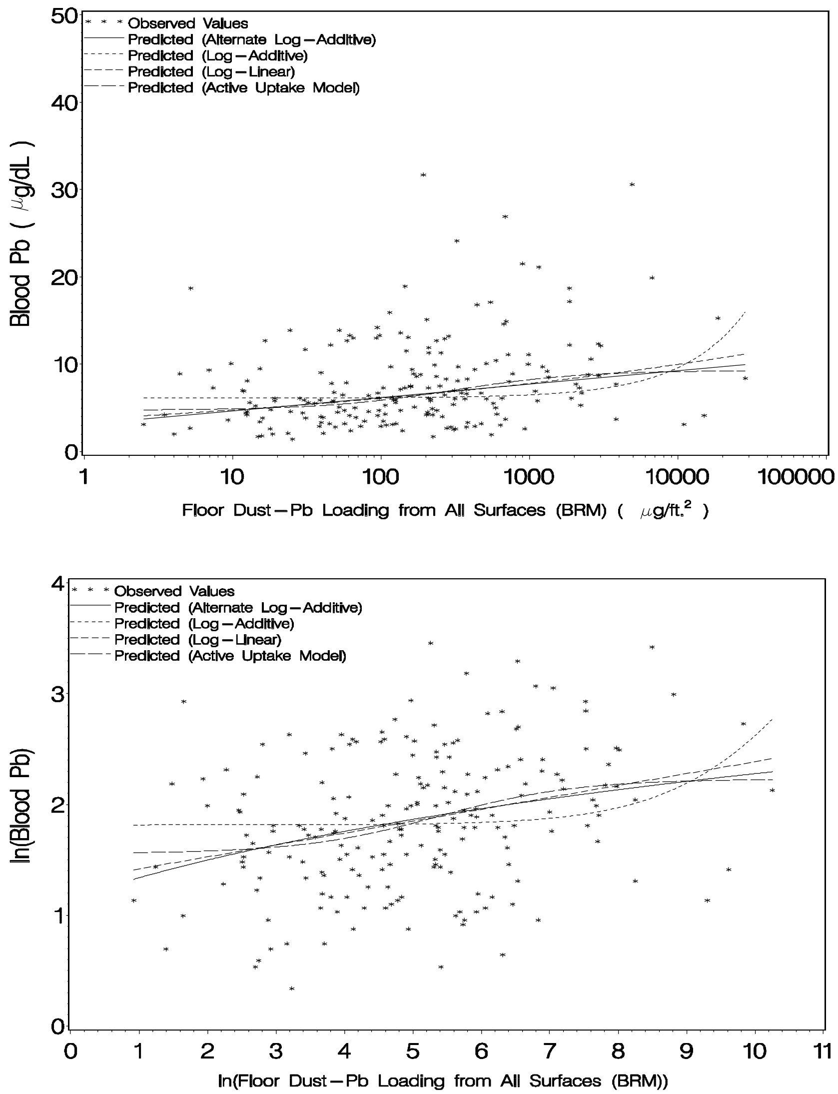
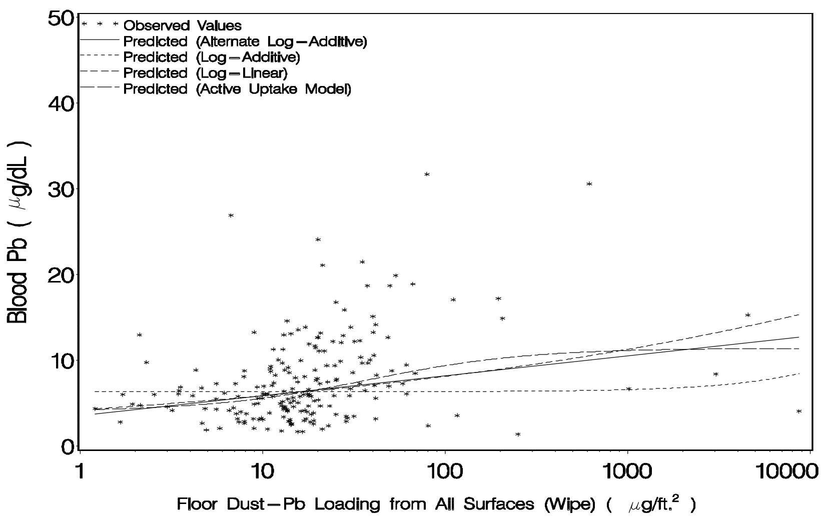



**Figure G11-17. Bivariate Relationship Between Blood-Lead Concentration and the Total Effect of Floor Dust-Lead Loading from All Surfaces (Carpeted or Uncarpeted) (BRM Samples).**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) | | | | | | |
|---|---|---|---|---|---|---|---|
| | $_0$ s.e. ( $_0$) | $_1$ s.e. ( $_1$) | $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | $^2$ | $R^2$ | Estimated Log-Likelihood | Number of Observations |
| Log-Linear | 3.57 (0.6925) | 1.01 (0.2547) | -- | 0.3511 | 0.0771 | -182.59 | 203 |
| Log-Additive | 6.36 (0.2776) | 0.00 (0.0005) | -- | 0.3800 | 0.0011 | -190.70 | 203 |
| Alternate Log-Additive | 1.45 (0.1109) | 0.14 (0.0358) | -- | 0.3531 | 0.0717 | -183.18 | 203 |
| Active Uptake | 6.33 (1.6813) | 0.47 (0.2797) | 11.44 (2.3671) | 0.3438 | 0.1051 | -179.44 | 203 |



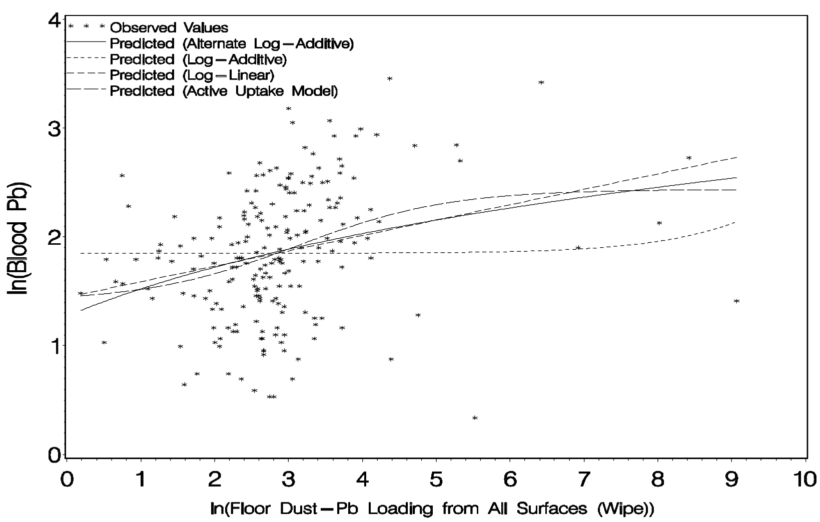
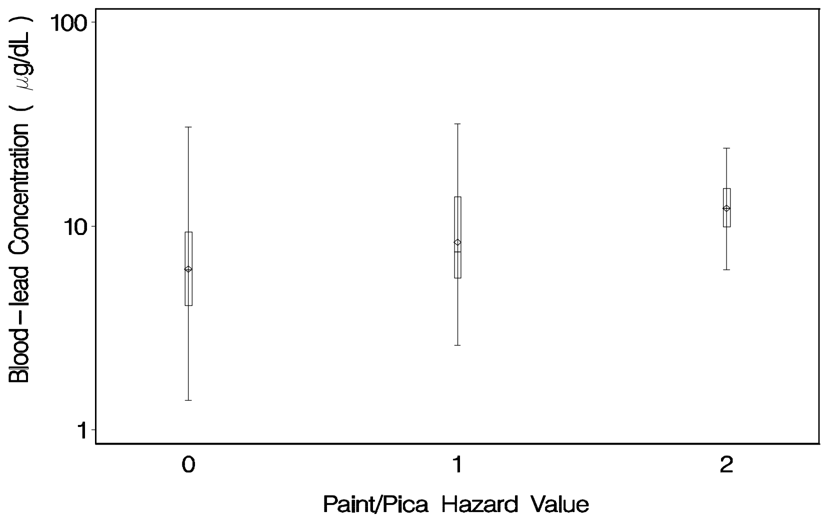




**Figure G11-18. Bivariate Relationship Between Blood-Lead Concentration and the Total Effect of Floor Dust-Lead Loading from All Surfaces (Carpeted or Uncarpeted) (Wipe Samples).**

| Statistical Model | Parameter Estimates and (Associated Standard Errors) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 s.e. ( 0) | 1 s.e. ( 1) | $\theta_{HSP}$ s.e. ($\theta_{HSP}$) | 2 | $R^2$ | Estimated Log-Likelihood | Number of Observations |
| Log-Linear | 1.82 (0.0437) | 0.33 (0.1058) | | 0.3628 | 0.0463 | -185.95 | 205 |
| Log-Additive | 6.15 (0.2696) | 2.67 (1.0765) | | 0.3630 | 0.0457 | -186.02 | 205 |
| Alternate Log-Additive | 6.15 (0.2696) | 2.67 (1.0765) | | 0.3630 | 0.0457 | -186.02 | 205 |
| Active Uptake | 6.15 (11.3986) | 2.66 (13.2990) | 2.49E8 (1.8727E16) | 0.3666 | 0.0457 | -186.02 | 205 |



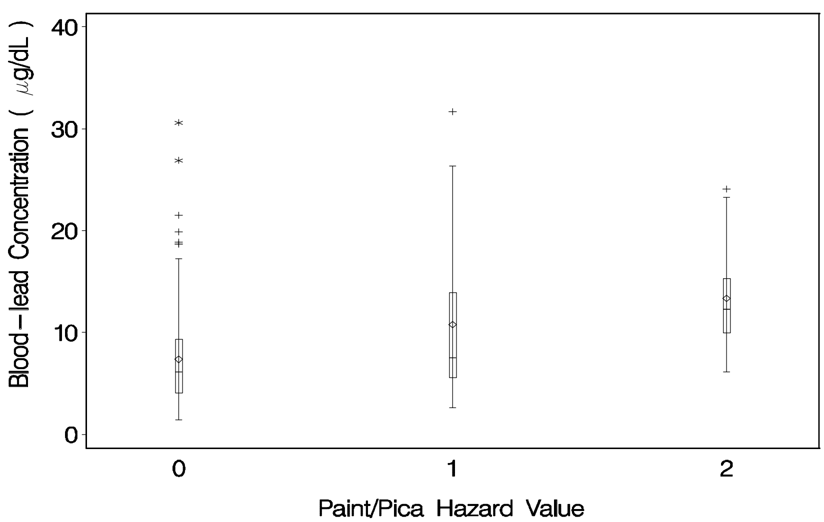




**Figure G11-19. Bivariate Relationship Between Blood-Lead Concentration and Paint/Pica Hazard Variable (Interior).**

# G12: Appendix on Regression Diagnostics

## Regression Diagnostics

This section of the appendix describes the diagnostic analyses performed as part of development of a multimedia exposure model using data from the Rochester Lead-In-Dust Study. Through the use of regression diagnostics, the adequacy of fit of the various candidate models developed (including the multi-media predictive model) to the data observed can be determined, and model assumptions can be verified. Results are presented for the final chosen model in particular, for which the following regression diagnostic "stages" were performed:

1. A normal quantile plot of the residuals was created. A normal quantile plot which can best be described by a straight line indicates that residuals (errors) are approximately normally distributed, as assumed. The quantile plot given in Figure G12-1 can best be described by a straight line, and therefore the assumption of normal errors is satisfied.

2. Residual values were plotted versus predicted values. This scatterplot could indicate signs of nonconstant variance (if points spread out or tighten up as you move from left to right) or nonlinearity (if points look quadratic or bow-shaped). A scatterplot exhibiting no pattern indicates no such problems. Similarly, plots of residuals versus predictors should indicate no discernible pattern. A plot of residuals versus predicted values is given in Figure G12-2. A plot of residuals versus predictor variables are given in Figure G12-3. Note that none of these plots indicate any relationship and each resembles a somewhat random scattering of points.

3. A plot of Cook's distance and DFFITS (both measures of influence) versus studentized residuals (a measure of how far an observation deviates from the modeled relationship) can indicate potential outliers - points with undue influence and points lying far outside the model's prediction. A plot of these two influence statistics are given in Figure G12-4. Each of these plots point to two possible outliers: observations with Child Identification Number (CID) 00166 and 04072. The observation with CID 00166 is also the observation with the lowest PbB level, while the observation with CID 04072 has the largest PbF level and the fifth smallest PbS level, and thus may require further examination. Note that DFFITS and Cook's distance are related to the studentized residuals and by definition are themselves similar, so observable patterns in these plots indicate nothing. However, typically those points with large studentized residuals (larger than 3 in absolute value) or DFFITS (larger than 1 in absolute value), or Cook's distance (larger than 1) possibly require further examination.

4. For a closer examination of how points influence model parameter estimates, the models were fit while excluding a single point at a time. Analysis of the coefficients adjusted for their standard error (intercept, and coefficients of PbS, PbF, PbW and PbP), including plots, can provide information about the influence of specific observations. Plots of the scaled measure of change in each parameter estimate are provided in the scatterplot matrix of Figure G12-5. Typically, values exceeding 1 in absolute value are suspect points. Note that none of the points in the multi-media

predictive model analysis is suspect by this criteria.  Table G12-1 below provides the parameter estimates while excluding the potential outliers flagged in stage (3).

**Table G12-1.  Influence of Possible Outlying Observations**

| Param. | Description | Estimate (deleting CIDs 00166, 04072) | Estimate (deleting CID 00166) | Estimate (Model with no deletions) |
|---|---|---|---|---|
| $_0$ | Intercept | 0.427484 (0.234447) | 0.403628 (0.234713) | 0.417648 (0.240347) |
| $_1$ | log (PbS):  Drip-line Soil-Lead Concentration (fine soil fraction) | 0.101146 (0.035592) | 0.115042 (0.034462) | 0.114038 (0.035294) |
| $_2$ | PbP:  Indicator of Interior Paint/Pica Hazard | 0.229457 (0.097897) | 0.236655 (0.098118) | 0.248043 (0.100421) |
| $_3$ | log (PbF):  Area-Weighted Arithmetic Mean (Wipe) Dust-Lead Loading from Any Floor (Carpeted or Uncarpeted) | 0.119694 (0.044423) | 0.090483 (0.039976) | 0.066338 (0.040151) |
| $_4$ | log (PbW): Area-weighted Arithmetic Mean (Wipe) Dust-Lead Loading from Window Sills | 0.075433 (0.035178) | 0.077318 (0.035277) | 0.087010 (0.035987) |
| $R^2$ | Coefficient of Determination | 23.98% | 23.23% | 21.67% |
| | Root Mean-Square Error (Residual Error) | 0.54670 | 0.54861 | 0.56188 |

This table indicates that excluding these points changes the parameter estimates only slightly.

5. Partial regression leverage plots were created for the environmental measures of lead exposure: dripline soil, floor dust from carpeted and uncarpeted floors, paint/pica hazard, and window sill dust.  A partial regression leverage plot that exhibits a linear relationship between blood-lead and the variable under consideration is indicative of  a linear relationship between blood lead and the environmental measure of lead exposure while controlling for all the other variables in the model.  The partial regression leverage plots given in Figure G12-6 indicate adjusted linear relationships for the lead-exposure variables included in the log-linear multimedia exposure model fitted to the data from the Rochester Lead-in-Dust Study.  Note that a partial regression leverage plot is produced by plotting the residuals from a regression of the response variable $(LPbB_{ijk})$ on all predictor variables excluding the lead exposure variable under consideration, versus the residuals from a regression of the lead exposure variable under consideration on the remaining predictor variables.

6. Partial $R^2$ comparisons can be made between predictor variables included in the model. A high partial $R^2$ indicates greater importance in predicting blood-lead concentration. Table G12-2 below provides the coefficient of determination ($R^2$) for a series of models in which one of the four predictor variables is excluded from the log-linear model.  The additional amount of variability in blood-lead concentrations explained by the excluded predictor variable once added to the model is also provided.

**Table G12-2.  Partial R-squared Comparisons.**

| Variable Excluded from the Model | Coefficient of Determination($R^2$) | Partial Coefficient of Determination (Partial $R^2$)[a] | Additional Variability Explained = (21.67% - $R^2$)[b] |
|---|---|---|---|
| Paint/Pica Hazard | 18.93% | 3.38% | 2.74% |
| Floor Dust-Lead | 20.44% | 1.54% | 1.23% |
| Dripline Soil-Lead | 16.97% | 5.67% | 2.63% |
| Window Sill Dust-Lead | 19.04% | 3.25% | 4.70% |

[a] Partial $R^2$ gives the contribution to the percent variation explained by adding in the variable of interest. It is calculated as: $\frac{R^2\ (FULL) - R^2\ (REDUCED)}{1 - R^2\ (REDUCED)}$ .

[b] 21.67% denotes the coefficient of determination ($R^2$) for the full multi-media predictive model.

The multi-media predictive model explains 21.67% of the variability in childhood blood-lead concentrations. Exposure from soil is the best predictor of blood-lead concentration, with the highest partial $R^2$ of around five percent.

7. An analysis into the effects of collinearity using several methods was conducted during the development of the multi-media predictive model. Issues pertaining to collinearity and strong correlation among potential lead-exposure predictor variables had a prominent role in the variable selection for the multi-media predictive model. Estimates of the tolerance statistic and variance inflation factor associated with each predictor variable in the model are provided in Table G12-3, together with a single value decomposition for the design matrix of observed predictor variables in the Rochester Study.

   To aid in the interpretation of these collinearity diagnostics, note that a large condition index indicates the data are ill-conditioned, or when extremely large, that parameter estimates are subject to substantial numerical error. A collinearity problem occurs whenever a variable with a high condition index is also a chief contributor to the variability between two or more variables.

   Variance inflation factors measure how much the variability associated with a particular parameter estimate is inflated due to collinearity between the predictors in a regression model. Although no formal criteria exists for establishing a critical variance inflation factor, it is common practice to associate a condition index of 10 with the notion that weak dependencies may be starting to affect the regression estimates. Condition indices of 30 to 100 indicate moderate to strong dependencies, and indices of greater than 100 indicate serious collinearity problems. The number of condition indices in the critical range indicates the number of near dependencies contributing to the collinearity problem.

Finally, another collinearity diagnostic is the condition number , defined by = (largest eigenvalue / smallest eigenvalue)$^{½}$, where large values suggest collinearity.

**Table G12-3. Collinearity Diagnostics**

| Index | Eigenvalue | Condition Index | PbF | PbW | PbS | PbP |
|---|---|---|---|---|---|---|
| | | | Proportion of Variability Explained | | | |
| 1 | 1.70803 | 1.00000 | 0.1395 | 0.1659 | 0.1295 | 0.0380 |
| 2 | 0.95248 | 1.33912 | 0.0264 | 0.0360 | 0.0013 | 0.9450 |
| 3 | 0.81482 | 1.44783 | 0.4116 | 0.0009 | 0.6107 | 0.0105 |
| 4 | 0.52466 | 1.80430 | 0.4225 | 0.7972 | 0.2585 | 0.0065 |
| | | | Tolerance | | | |
| | | | 0.820436 | 0.736608 | 0.855715 | 0.976984 |
| | | | Variance Inflation | | | |
| | | | 1.218864 | 1.357574 | 1.168613 | 1.0235585 |

Note that the largest condition index in Table G12-3 is 1.8, and the largest inflation factor is 1.36 (PbW). Therefore, the multi-media predictive model (in its current form) does not appear to suffer from a severe collinearity problem, nor does it appear to be ill-conditioned (numerically unstable or fragile). The following matrix contains the correlation coefficients among the four predictor variables used in the multi-media predictive model. The coefficients are based on a sample size of 179 children/ households included in the current model.

| | PbF | PbW | PbS | PbP |
|---|---|---|---|---|
| **PbF** | 1.000 | 0.417 | 0.186 | 0.110 |
| **PbW** | 0.417 | 1.000 | 0.370 | 0.101 |
| **PbS** | 0.186 | 0.370 | 1.000 | 0.119 |
| **PbP** | 0.110 | 0.101 | 0.119 | 1.000 |

Plots are provided in Figure G12-7 of each continuous predictor variable versus another continuous predictor variable, where each observation is coded for values of the paint/pica hazard variable (0, 1 or 2). These plots provide insight into the range of possible values over which the multi-media predictive model was constructed, and over which inferences can be drawn.

Based on the regression diagnostics on the multi-media predictive model it was concluded that: