MODELING PERFORMANCE OF ENGINEERING CONTROLS WHEN REDUCTIONS ARE LARGEST AT THE HIGHEST ENVIRONMENTAL CONCENTRATIONS OF THE HAZARDOUS CONTAMINANT

Stanley A. Shulman, Kenneth R. Mead, and R Leroy Mickelsen
National Inst. for Occupational Safety and Health, 4676 Columbia Parkway, MS-R3, Cincinnati, OH 45226

Key Words: Multiplicative Interaction, Regression on Column Sums, Lognormal Environmental Data

Abstract

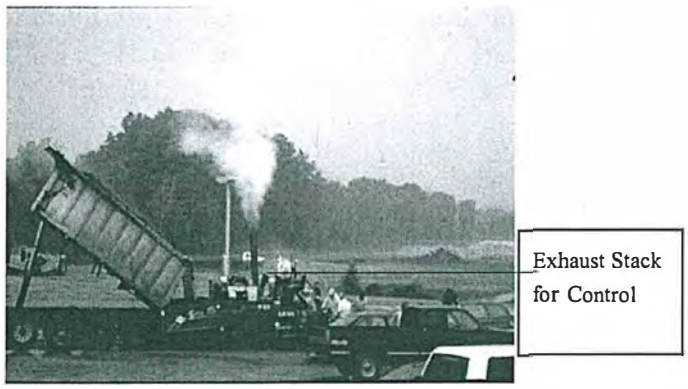
Engineering controls are often compared outdoors using randomized pairs to determine if a new control design reduces worker exposure to airborne hazardous contaminants when compared to an uncontrolled work environment. A common occurrence is that the ratio of means of controlled to uncontrolled work environments depends upon concentration. When the uncontrolled environment is at its highest concentrations, the engineering control has greatest impact and the ratio is smallest (reduction is largest). Such interaction may arise outdoors because of wind and other natural conditions that vary contaminant concentrations. The following approaches to model this interaction are compared (all on natural log scale): 1) model estimates reduction separately for upper 25% of uncontrolled samples and for lower 75%; 2) model is based on Tukey's one degree of freedom model for interaction, equivalent to regression of control differences on pair means; 3)model regresses control differences on control-off values. Results based on lognormality indicate that for many situations model 3) may better describe the interaction than model 2). Since model 3) has greater power than model 1), model 3) may be preferable.

1) Introduction

The data considered are outdoor data, for which the comparison of interest is an uncontrolled to a controlled environment. A common result is that the reduction in contaminant is greatest when the uncontrolled environment (referred to as "control-off") is highest. This may make sense, in that the highest control-off measurements may occur when environmental factors such as wind have the least impact. The example data here are air measurements of organic compounds on a highway paving machine. The control ventilation system draws contaminant from the source at the auger, and discharges it out of the workers' breathing zones through a vertical stack above the paver and the paver driver.

2) Example Data

Since the control could be switched on or off, changing between the two control settings was very simple. In

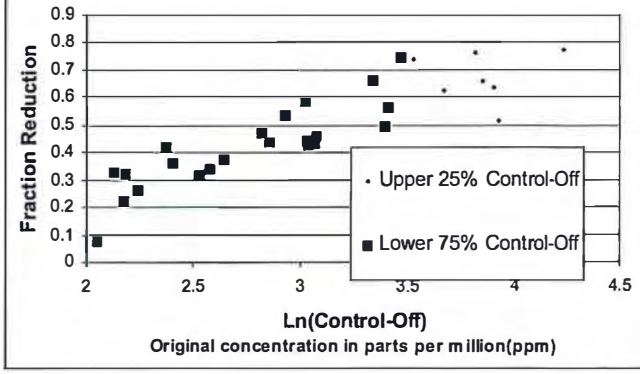


these studies data were collected in randomized pairs of (control-off, control-on), each trial in each pair lasting at least one and a half minutes. The pairs were collected over five days of sampling. The data consisted of organic compound concentrations averaged over four second intervals. Medians were computed for each trial because medians are much less correlated than individual readings. Deletion of data collected near transitions in paving status also reduced correlation. (For example, half a minute of data was deleted in trials which were either preceded or succeeded by a period of no paving that lasted for at least 25 seconds.)

The measure of control effectiveness used is the fraction reduction of airborne organic compound concentrations, defined as:

Fraction Reduction (1)
= 1- (control-on median)/(control-off median).

Figure 1: Fraction Reduction in Airborne Organic Concentration



Thus, the smaller the ratio of control-on median to control-off median, the closer the fraction reduction is to 1, the maximum value.

For the example data the estimated fraction reductions are plotted above. It is clear from the figure that there is increasing reduction with increasing control-off values.

For all models presented, the data are transformed to the natural log scale, since that scale is convenient for ratios, as are needed for estimating the fraction reduction.

3) Model 1: Upper 25% Model

The data are divided into two groups, based on the values of the control-off median in the pair. Those pairs with the control-off value in the upper 25% of all control-off values are in one group, and the remaining pairs in a second group. The two groups are labeled in Figure 1, where it can be seen that for the 25% pairs, the fraction reduction was about 0.68, compared to about 0.44 for the lower 75% group. Even though the overall average of about 0.50 is statistically significant at the 5% level, this difference in fraction reduction between the two groups is large. In other situations, when the overall average is not statistically significant, the reduction for the upper 25% pairs may be. A weakness of this approach is that the division into two groups is based on a somewhat arbitrary dividing point: 25%. The conclusions can differ if a different point is used.

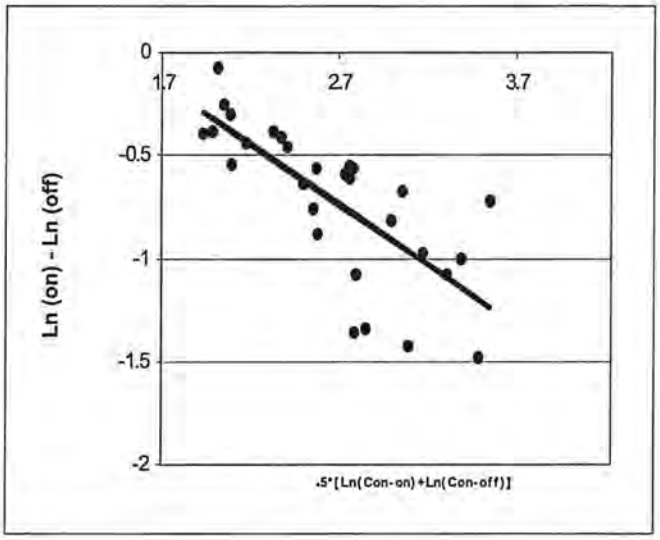
4) Model 2- Regression on Pair Means

The trend seen in Figure 1 of increasing fraction reduction with increasing control-off values is a form of interaction. The model proposed by Tukey for multiplicative interaction in a two-way design allows for the dependence of method differences on the pair means. This can be shown as follows. In the model of Tukey [1], the control-pair interaction is a multiplier of a factor for the pair mean and a factor for the control type mean:

$$Ln(y_{p,con})=\mu+ \beta_p+\alpha_{con}+G\beta_p\alpha_{con} + e_{p,con},\ p=1,2,..P, \quad (2)$$

where p=pair, con=control setting ("c" =control-on, "nc"= control-off). β_p =[(mean for pair p)-μ]; $\Sigma_p\beta_p=0$; α_{con}=[(mean for control c)- μ]; $\alpha_c=-\alpha_{nc}$; $e_{p,con} \sim N(0,\sigma^2)$. Estimates for the parameters are obtained by substituting the corresponding means from the data. The estimate for G is $\Sigma_{con,p}\ \alpha_{con}\ \beta_p\ y_{p,con}/[\Sigma_{con}\ \alpha^2_{con}\ \Sigma_p\beta^2_p]$, where the sample estimates for α_{con} and β_p replace the parameter values.

Figure 2
Ln(Con-on)-Ln(Con-off) vs
0.5[Ln(Con-on+Ln(Con-off)];
Actual Data and Fitted Line



The model given in eq.(2) is appropriate for any two-way design. For a design for which the control factor has just two levels, the estimate for G may be written in a simpler form:

$$G= \Sigma_p\beta_p\ (y_{p,c}- y_{p,nc})/[2\ \alpha_c\ \Sigma_p\beta^2_p], \quad (3)$$

with sample estimates in place of α_c and β_p. Taking differences under eq. (2), as in references [2,3], we obtain:

$$Ln(y_{p,c})-Ln(y_{p,nc})=2\ \alpha_c + 2\ \alpha_c\ G\beta_p+(e_{p,c}- e_{p,nc}) \quad (4)$$

Thus, differences are linear functions of pair means. Since the estimate of G is produced by substitution of the sample values for the parameters in (3), and since $(2\alpha_cG)= \Sigma_p\beta_p\ (y_{p,c}- y_{p,nc})/\ \Sigma_p\beta^2_p$, the estimate of the slope from the linear regression (using the sample values in place of the β_ps) can be used to obtain the value of G. The regression slope can be tested to determine whether the value of G is zero. This test can be used, even though the estimate of G may be biased, since the estimates of the β_ps differ from the true values because of measurement error[4].

For the example data, the differences of the log-

transformed data (control-on -control-off) are shown in Figure 2, in addition to the fitted line. The multiplicative interaction is statistically significant at the 5% level.

5) Model 3: Regression on Control-off

Figure 1 suggests the following modification of eq. (4). Since the fraction reduction appears to be dependent on the control-off values, it seems sensible to regress the log scale differences on the log of the true control-off values:

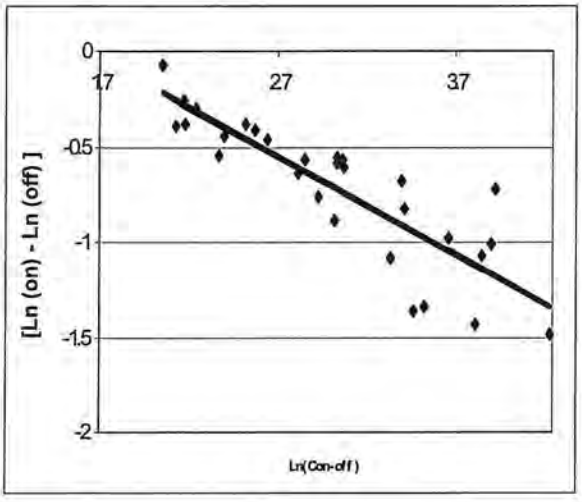
$$\text{Ln}[y_{p,c}]-\text{Ln}[y_{p,nc}]= \gamma+\delta\ \mu_{p,nc}+f_p, \qquad (5)$$

where $\mu_{p,nc}$=mean of control-off for pair p, on log scale, γ= line's intercept, δ= slope, and $f_p \sim N(0, \sigma_f^2)$.

In eq. (5) measurement error is assumed small compared to environmental variability, in which case the model may be fitted by linear regression by replacing $\mu_{p,nc}$ by $\text{Ln}[y_{p,nc}]$[4]. If $\delta<0$, $[\text{Ln}(y_{p,c})- \text{Ln}(y_{p,nc})]$ decreases with increasing control-off. ($y_{p,c}/y_{p,nc}$ decreases.)

For the example data, shown with the fitted line in Figure 3, the estimated slope is significant at the 5% level.

Figure 3
Ln(Con-on)-Ln(Con-off) vs Ln(Con-off)
Actual Data and Fitted Line



6) Why Increasing Control Effectiveness with Increasing Control-off Values?

The previous sections have focused on recognition and statistical modeling of the phenomenon. Here we consider possible reasons for its occurrence. Two possibilities to consider are:

a) Environmental data are often lognormal[5]. The consequences should be examined. Greater reductions at higher control-off values may relate to lognormality. Since lognormality is associated with environmental variability, the higher control-off values could correspond to less environmental control, and statistical results which show greater reduction at higher control-off values can be interpreted to mean that reduction due to control is greatest when environmental control is least.

b) What characteristics of the measurement process itself could lead to this phenomenon? One possibility is the presence of background levels, which are difficult to estimate. The idea is that when the control-off values are high, they are far above background, and the background would, therefore, have little effect on the denominator for the ratio of control-on to control-off.

Only a) will be examined here; b) will be investigated in a future study.

7) Lognormality as an Explanation of the Phenomenon

For the example data, since the pairs were collected over five days, and since there was substantial day to day variation, lognormality of the data was assessed by examination of the residuals from random effect models fitted separately to the two control types. Both control-on and control-off distributions appear to be lognormal.

Under bivariate lognormality, $[\text{Ln}(y_{con\text{-}on})-\text{Ln}(y_{con\text{-}off})]$ can be written as a linear regression on $[a\text{Ln}(y_{con\text{-}on}) + (1-a)\text{Ln}(y_{con\text{-}off})]$ for the variable "a" in the interval [0, 1]. Interest here is when a=0.5 (regression on pair means) or a=0 (regression on control-off).

For the regression on pair means model the slope is proportional to $(\sigma^2_{con\text{-}on}-\sigma^2_{con\text{-}off})$. If $\sigma^2_{con\text{-}on} \cong \sigma^2_{con\text{-}off}$, then slope$\cong$0, where the variances are on the log scale.

(The exact form of the regression, obtained by transformation of variables, is:
$\text{Ln}(y_{con\text{-}on})- \text{Ln}(y_{con\text{-}off})=$ (6)

$(\mu_{con\text{-}on} - \mu_{con\text{-}off})$
$+ \xi (.5)[\ Ln\ (y_{con\text{-}on}) + Ln(y_{con\text{-}off}) - (\mu_{con\text{-}on} - \mu_{con\text{-}off})] + \pi\ h,$
where correlation $[Ln(y_{con\text{-}on}), Ln(y_{con\text{-}off})] = \rho,$
h is distributed as normal(mean=0,variance=1),
where $\xi = 2\ (\sigma^2_{con\text{-}on} - \sigma^2_{con\text{-}off})/\ \sigma^2_t\ ,$
$\pi = \{[1 - (\sigma^2_{con\text{-}on} - \sigma^2_{con\text{-}off})^2 / (\sigma^2_d\ \sigma^2_t)]\sigma^2_d\}^{0.5},$
$\sigma^2_d = \sigma^2_{con\text{-}on} + \sigma^2_{con\text{-}off} - 2\ \rho\ \sigma_{con\text{-}on}\ \sigma_{con\text{-}off}$
$\sigma^2_t = \sigma^2_{con\text{-}on} + \sigma^2_{con\text{-}off} + 2\ \rho\ \sigma_{con\text{-}on}\ \sigma_{con\text{-}off})$

For the regression on control-off the slope is
$\nu = [\ \rho\sigma_{con\text{-}on}/\sigma_{con\text{-}off} - 1]$. If $\rho\sigma_{con\text{-}on} < \sigma_{con\text{-}off}$, then the slope<0.

(The exact form of the regression, obtained by transformation of variables, is:
$Ln\ (y_{con\text{-}on}) - Ln(y_{con\text{-}off}) =$ (7)
$(\mu_{con\text{-}on,L} - \mu_{con\text{-}off,L}) + \nu[Ln(y_{con\text{-}off}) - \mu_{con\text{-}off,L})$
$+ (1 - \rho^2)^{0.5}\ \sigma_{con\text{-}on,L}\ d$, where d is distributed as normal(mean=0,variance=1))

For the example data, the slope estimated from the data under the regression on pair means model was (-0.6), and under the regression on control-off model was (-0.5). The corresponding estimates based on the lognormal distribution were within 0.05 of the above estimates, when sample values were substituted for parameter values..

8) Which Model to Prefer?

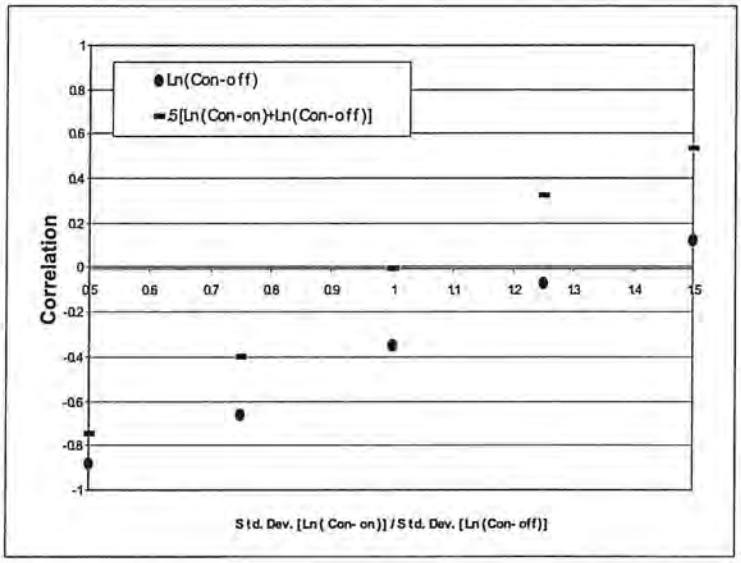
One way to compare the two models is by a comparison of the correlations between the dependent and independent variables. These correlations are plotted as in Figure 4 as functions of the correlations ρ of the data on the natural log scale, and as functions of the ratio of log scale standard deviation of control-on to control-off. These calculated correlations do not require lognormality. However, the importance of the log-normality in the discussion is the linearity associated with it [6], and the higher the correlation the greater the tendency of the data to lie on a straight line.

The data plotted in Figure 4 are for the case that ρ is 0.75. Similar figures result for correlations=0.5 or 0.9. For the standard deviation ratio (on/off) $\leq 1/\rho$, the regression on control-off model has more negative correlation than the regression on pair means model. Since for the upper 25% phenomenon, we expect to see negative correlation, this indicates that the regression on control-off model has greater tendency to lie on a negatively sloped line than the regression on pair means model.

Other comments on Figure 4 are that for control ratio 0.5, there is little difference in correlations. As the ratio increases, the difference between correlations increases.

To compare the power of the three models a limited simulation study was carried out, as shown in Table 1. The example 1 (Ex. 1) in the table refers to the example studied here, for which the ratio of standard deviations was about 0.6, and ρ=0.85. The example 2 (Ex. 2) refers to a second example, not shown here, in which the ratio of standard deviations is approximately 1, and ρ=0.8.

Figure 4
Correlation between:
[Ln(on)-Ln(off)] &{Ln(off) or 0.5[Ln(on)+Ln(off)]}
vs Ratio[Std. Dev. Ln(on)/ Std. Dev. Ln(off)] when
Correlation [Ln (On), Ln (Off)]=0.75



The simulation results shown in Table 2 confirm that the two regression models can yield quite different results for statistical significance of their slopes. Setting a) yields significant results for all three models, as the example data do. Setting b) indicates the effect of standard deviation ratio ~ 1, and also the relatively low power for the regression on control-off model, since ρ is large, and , therefore, $(\rho\ \sigma_{con\text{-}on} - \sigma_{con\text{-}off})$ is not so different from 0. Setting c) is a redo of setting a) , which shows that when the standard deviation ratio is close to 1, and the standard deviations are much smaller than in

b), the power continues to be low for regression on control-off. Setting d) is a redo of setting b). Since the correlation is high, the control-off standard deviation must be reduced sufficiently so that the regression on pair means model has considerable power.

Also, the fraction yielding a statistically significant result for regression on control-off always exceeds that for the upper 25% model, suggesting that the regression model has greater power than the upper 25% model. This result holds for the different variance relationships used. This is a result that would be expected, since the regression on control-off model uses all the data to fit a single line.

Table 1: Design of Simulation Study
Lognormal Simulations: 160 Samples of Size 25

Parameter Values Used	Response Variables for a-d:
a) sample values, Ex. 1 $\sigma_{con\text{-}on}$~0.36, $\sigma_{con\text{-}off}$~0.63 ρ~0.85	Fraction of samples for which: i) upper 25% model had significantly different result from lower 75% control-off pairs ii)regression on pair means model yielded statistically significant slope (eq. (4)) iii)regression on control-off model yielded statistically significant slope. (eq. (5))
b) sample values, Ex. 2 $\sigma_{con\text{-}on}$ ~ $\sigma_{con\text{-}off}$~1.5, ρ~0.8	
c) $\sigma_{con\text{-}on}$ ~ $\sigma_{con\text{-}off}$, using $\sigma_{con\text{-}on}$, from Ex. 1, $\sigma_{con\text{-}on}$~0.36, ρ~0.85	
d)In Ex. 2, $\sigma_{con\text{-}off}$ modified so that (ρ $\sigma_{con\text{-}on}$ - $\sigma_{con\text{-}off}$)~0, ρ~0.8	

Table 2: Power Calculations from Simulations
Fractions of Samples that Yielded Statistically Significant Results at 5% Level

Parameter Value Settings, from Table 1	i) Upper 25% Model	ii) Regression on Pair Means Model	iii) Regression on Control-Off Model
a	0.99	1	1
b	0.20	0.01	0.4
c	0.2	0.06	0.3
d	0.04	0.6	0.07

9) Conclusions and Recommendations

It is common in outdoor studies of engineering controls that the reduction in concentration due to the control is highest when uncontrolled measurements are highest.

Results presented suggest this may be due to the lognormal distribution of airborne contaminant data. Higher control-off values can correspond to less environmental control; results showing greater reduction at these values can mean reduction due to control is greatest when environmental control is least. Another possible explanation is the effect of background concentrations, which will be investigated in future research.

Three models were compared: the upper 25% model, the regression on pair means model, and the regression on control-off model. For standard deviation ratio(control-on / control- off) $\leq 1/\{\text{correlation}[\text{Ln}(y_{con\text{-}on}),\text{Ln}(y_{con\text{-}off})]\}$, the regression on control-off model yields greater negative correlation than the regression on pair means model. Also, the upper 25% model does not have as much power as the regression on control-off model. Thus, the regression on control-off model can give a better estimate of how well the control system is functioning.

10) Acknowledgments

The authors wish to thank James Deddens, Michael Gressel, Edward Krieg, Jr., and Paul Schlecht of NIOSH for their helpful reviews of this document. The recognition of the upper 25% phenomenon is based on work of former NIOSH employees Dennis O'Brien and Thomas Fischbach, and we appreciate Dennis

O'Brien's suggestion that we consider its applicability to the asphalt data. The authors appreciate helpful comments by Steve Simon of Children's Mercy Hospital, Kansas City, and Michael Butterworth of CBS, both of whom viewed the poster session at the Joint Statistical Meetings.

11) References

1) Scheffe, H. The Analysis of Variance, Wiley, 1959, p. 129.

2) Mandel, J. A Method for Fitting Empirical surfaces to Physical and Chemical Data," Technometrics, 1969, V. 11, pp 411- 429.

3) Mandel, J. "The Partitioning of Interaction in Analysis of Variance," Journal of Research of the National Bureau of Standards- B Mathematical Sciences, V. 738, 1969, pp. 309-328.

4) Draper, N. and Smith, H. Applied Regression Analysis, 2nd Ed., Wiley, 1981, p. 122.

5) Rappaport, S.M. Assessment of Long-Term Exposures to Toxic Substances in Air. Annals of Occupational Hygiene, Vol 15, 1991, pp.61-121.

6) Anderson, T.W. An Introduction to Multivariate Statistical Analysis. Wiley, 1958, p.30.