ORAU TEAM
Dose Reconstruction Project for NIOSH

Oak Ridge Associated Universities | NV5|Dade Moeller | MJW Technical Services

DOE Review Release 08/25/2023

Applications of Regression Models and Ratios	ORAUT-RPRT-0106	Rev. 00
	Effective Date:	08/03/2023
	Supersedes:	None

Subject Expert(s): Nancy Chalmers, Daniel Sheanshang, and Thomas R. LaBone

Document Owner Approval:	Signature on File Nancy Chalmers, Document Owner	Approval Date:	08/01/2023
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	08/02/2023
Concurrence:	Vickie S. Short Signature on File for Kate Kimpan, Project Director	Concurrence Date:	08/01/2023
Approval:	Signature on File Timothy D. Taulbee, Associate Director for Science	Approval Date:	08/03/2023

**FOR DOCUMENTS MARKED AS A TOTAL REWRITE OR REVISION,
REPLACE THE PRIOR REVISION AND DISCARD / DESTROY ALL COPIES OF THE PRIOR REVISION.**

☒ New ☐ Total Rewrite ☐ Revision

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
08/03/2023	00	New report generated to illustrate the benefits and detriments of using regression models versus ratios. Incorporates formal internal and NIOSH review comments. Training is not required. Initiated by John M. Byrne and authored by Nancy Chalmers.

TRADEMARK INFORMATION

Minitab® is a registered trademark of Minitab in the United States and/or other countries.

All other trademarks are the property of their respective owners.

TABLE OF CONTENTS

SECTION **TITLE** **PAGE**

LIST OF TABLES

TABLE **TITLE** **PAGE**

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ABRWH	Advisory Board on Radiation and Worker Health
Bq	becquerel
GM	geometric mean
GSD	geometric standard deviation
mrem	millirem
NIOSH	National Institute for Occupational Safety and Health
NS	Neutron Source
ORAU	Oak Ridge Associated Universities
ORAUT	ORAU Team
ROS	regression on order statistics
SRDB Ref ID	Site Research Database Reference Identification (number)
yr	year

1.0 INTRODUCTION

Datasets often have a set of paired quantities *Y* and *X*, and it is sometimes helpful to determine a predictive relationship between them. For example, suppose a dataset contains the neutron dose *Y* and the paired photon dose *X* for a number of dosimeters. A relationship can be established between the measured neutron doses and photon doses so that, when given a new measured photon dose, the model can be used to predict the associated unmeasured neutron dose.

If a proportional linear relationship exists between *X* and *Y*, a linear regression of *Y* on *X* through the origin is used to calculate the ratio of *Y* to *X*:

$$Y = \beta X + \varepsilon \qquad (1\text{-}1)$$

where

Y = the response variable
β = the slope parameter
X = the predictor variable
ε = a random error term, assume ε is normally distributed with mean 0 and standard deviation σ, which can be denoted as $\varepsilon \sim N(0,\sigma)$

To estimate *Y* given *X*, simply multiply *X* by the slope β, which is the mean ratio *Y*/*X*.

If a linear relationship exists between *X* and *Y* but is not proportional, a linear regression of *Y* on *X* (not through the origin) is used to calculate the relationship of *Y* to *X* [Weisberg 2005, pp. 48–49]:

$$Y = \mu + \beta X + \varepsilon \qquad (1\text{-}2)$$

where

μ = the intercept parameter

To estimate *Y* given *X*, simply multiply *X* by the slope β and add the intercept μ.

When a linear relationship does not exist between *X* and *Y*, one option is to transform *X* and/or *Y* so that linear modeling techniques are appropriate. For data typically encountered in health physics, the log transform of *X* and *Y* is often a good choice:

$$\log(Y) = \mu + \beta \log(X) + \varepsilon \qquad (1\text{-}3)$$

where

log = natural log (base e)

It is possible that the relationship between *X* and *Y* is nonlinear and cannot be transformed into a linear form. Those relationships would be handled with other statistical techniques that are outside of the scope of this report. The remainder of this report focuses on linear relationships like the one in Equation 1-3, which are referred to in this report as the “regression method.” For simplicity, the regression analyses for this report use linear regression but the task can be accomplished using quantile regression or other techniques [Oak Ridge Associated Universities (ORAU) Team (ORAUT) 2018].

In the dose reconstruction program, the standard method for estimating *Y* from *X* has historically been fitting a lognormal model to *Y*/*X* using the regression on order statistics (ROS) [Helsel 2011, pp. 52–56], which is referred to in this report as the "ratio method." This analysis gives the median of *Y*/*X*, which is the geometric mean (GM) of the fitted lognormal distribution. To estimate *Y* given *X*, simply multiply *X* by this GM. Lognormal ROS with ratios is widely used in the dose reconstruction program but does not appear to be used outside the program.

2.0 PURPOSE

The remainder of this report demonstrates that using ROS with ratios can be problematic[1] and should only be used with caution. The main problem with this method is that it collapses a bivariate relationship (*Y* versus *X*) to a univariate relationship (*Y*/*X*), which results in a loss of information. In practice, ROS with ratios causes the following problems:

1. Gives the correct answer only for one specific linear relationship (i.e., Equation 1-3 with $\beta = 1$).

2. For $\beta \neq 1$, will usually give a seemingly useful answer that is incorrect. This includes the case where there is no relationship between *X* and *Y* (i.e., when *X* and *Y* are independent).

   This problem can only be diagnosed by looking at the bivariate relationship between *X* and *Y* (i.e., it cannot be diagnosed by examining the result of the ROS).

Simulated data are used to illustrate the issues with ROS with ratios for various values of β (using Equation 1-3):

- X and *Y* have a relationship and $\beta = 1$ (Section 3.0),
- *X* and *Y* have a relationship and $0 < \beta < 1$ (Section 4.0),
- *X* and *Y* have a relationship and $\beta > 1$ (Section 5.0), and
- *X* and *Y* have no relationship, which implies $\beta = 0$ (Section 6.0).

Values of $\beta < 0$ are not addressed in this report. Mathematically, if $\beta < 0$, switching the roles of *X* and *Y* will result in positive β. In practice, if $\beta < 0$, there is no need to switch the roles of *X* and *Y*; model the bivariate relationship with negative β.

Simulated data are used in Sections 3.0 to 6.0 so that the true values of the parameters are known and can be compared with the estimated values from the models to show that the regression method is superior to the ratio method. Each section refers to an attachment that illustrates the same concepts with real data, where the true values of the parameters are not known but the regression estimates are taken to be the best estimates. No evaluation of the quality or source of the data is included in the attachments. The examples in the attachment are for illustrative purposes only. The supporting code and data for each example is in ORAUT [2023]. Section 7.0 describes how the value of β affects the performance of the ratio method as illustrated by the simulated examples in Sections 3.0 to 6.0. Section 8.0 summarizes the concepts in the report.

Note that the issues discussed in the following sections do not apply to modeling the distribution of *Y* or the distribution of *X* separately – they apply only to modeling the two together. For example, there

---

[1] Note that the issue is with the use of ratios, not ROS. Ratios are problematic even if the average (or some other summary statistic) of a group of observed ratios is used. The ROS technique is statistically sound and can be legitimately used in other applications. This report focuses on the use of ROS with ratios because that technique is so widely used in the dose reconstruction program.

are no problems associated with modeling the photon doses in a population with a lognormal distribution as long as the lognormal distribution provides an adequate fit to those data.

3.0 *X* AND *Y* HAVE A RELATIONSHIP AND β = 1

Assuming that there is a well-defined linear relationship between the logarithm of *Y* and the logarithm of *X*, the linear model in Equation 1-3 can be rearranged to give:

$$\log\left(\frac{Y}{X^{\beta}}\right) = \mu + \varepsilon \quad (3\text{-}1)$$

which means that:

$$\log\left(\frac{Y}{X^{\beta}}\right) \sim N(\mu, \sigma) \quad (3\text{-}2)$$

In the special case where β = 1, the logarithm of the ratio *Y*/*X* is normally distributed, which means that the ratio *Y*/*X* is lognormally distributed. This means that for β = 1, the regression in Equation 1-3 is equivalent to lognormal ROS of the ratio *Y*/*X* regardless of the value of μ or σ.

To illustrate this point, the 1,000 pairs of data shown in Figure 3-1 are simulated with β = 1, μ = 3, and σ = 0.2.

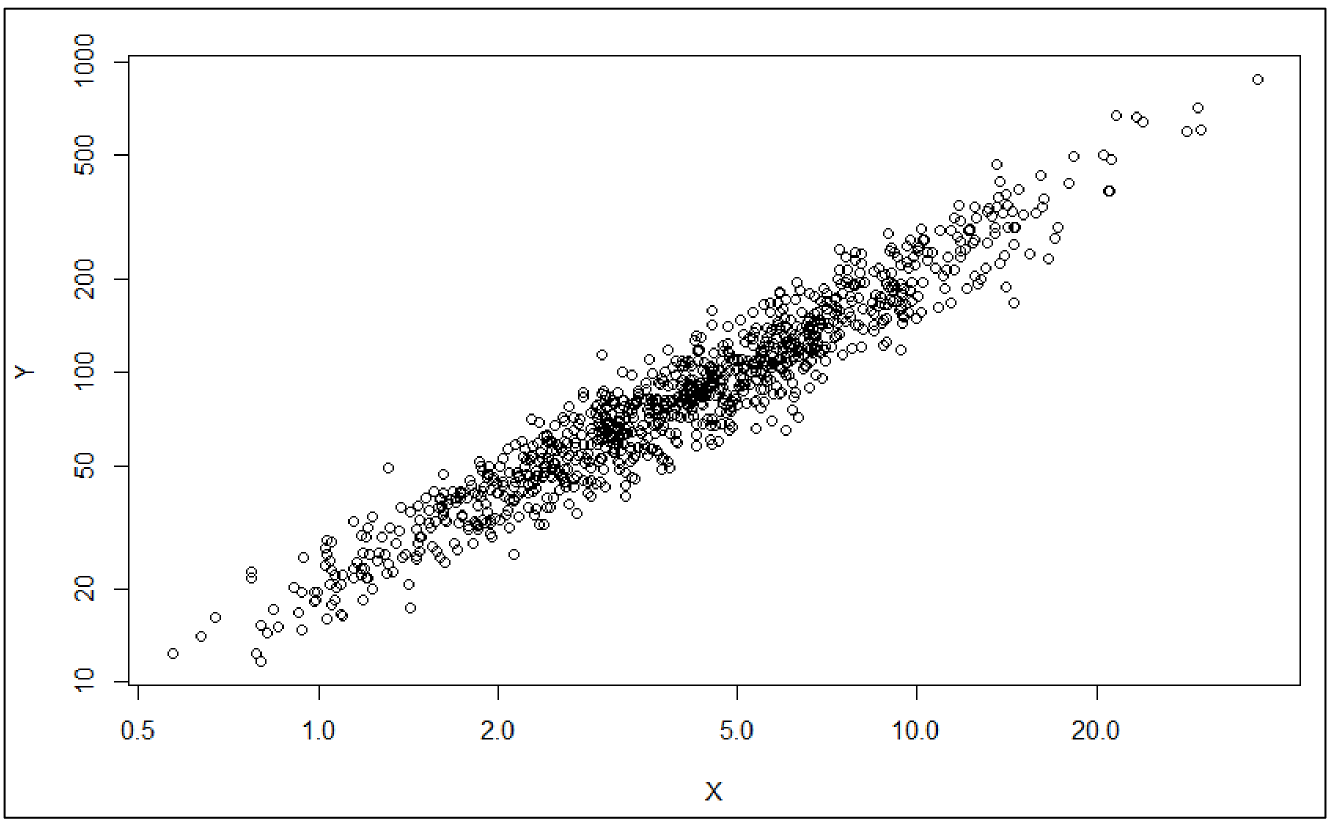

Figure 3-1. Data simulated with β = 1, μ = 3, and σ = 0.2.

3.1 LINEAR REGRESSION

Fitting the simulated data from Figure 3-1 using ordinary least squares and Equation 1-3 gives the parameter estimates in Table 3-1.

Table 3-1. Parameter estimates for regression using Equation 1-3.

Parameter	True Value	Estimate
β	1	1.001
μ	3	3.014
σ	0.2	0.1980

As expected, the estimated parameters match the true parameters very well. Figure 3-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

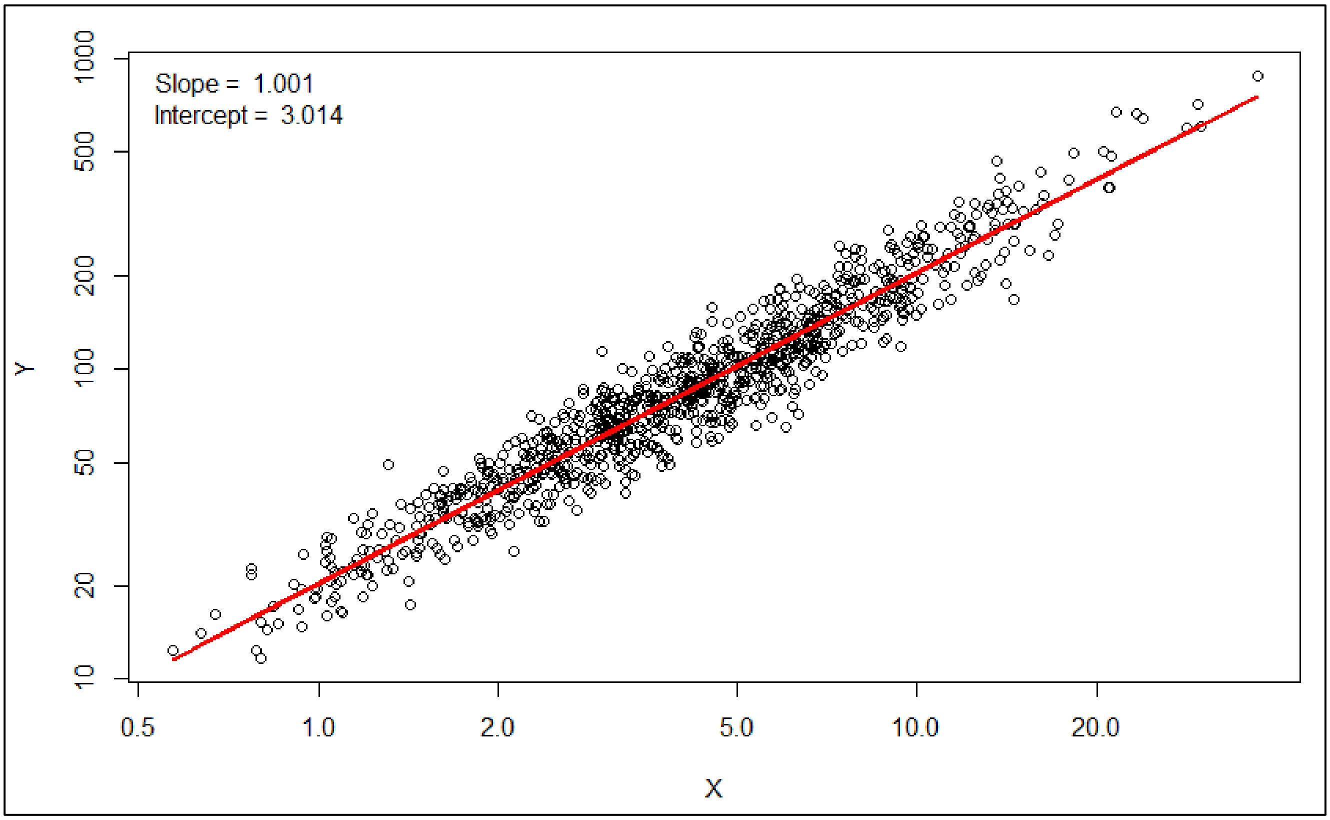

Figure 3-2. Data simulated with β = 1, μ = 3, and σ = 0.2 and the associated linear regression fit.

3.2 ROS WITH RATIOS

Taking the simulated data from Figure 3-1, calculating the ratio of *Y*/*X* for each of the 1,000 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table 3-2, including the GM, the geometric standard deviation (GSD), and the lognormal probability plot in Figure 3-3.

Table 3-2. Parameter estimates for lognormal ROS with ratios.

Parameter	True Value	Estimate
β	1	(a)
μ	3	3.015
σ	0.2	0.198
GM	e^{μ} = 20.1	20.39
GSD	e^{σ} = 1.2	1.219

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

As expected, the estimated parameters in Table 3-2 match the true parameters very well.

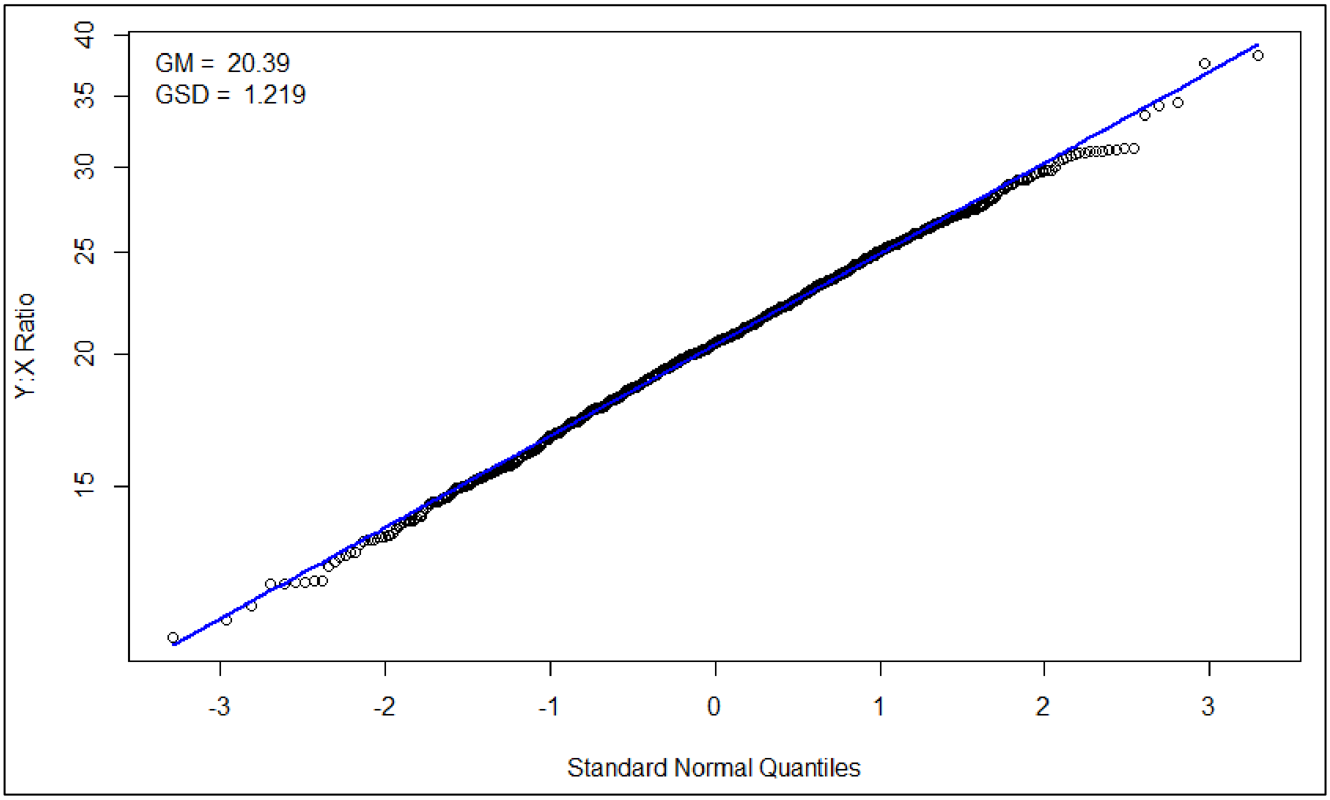

Figure 3-3. Data simulated with β = 1, μ = 3, and σ = 0.2 and the associated lognormal ROS with ratios fit.

3.3 COMPARISON OF REGRESSION AND RATIOS

As mentioned in Section 3.0, for β = 1 the linear regression in Section 3.1 and the ROS with ratios in Section 3.2 are equivalent. The parameter estimates in Tables 3-1 and 3-2 match the true values very well. To further illustrate the equivalence of the two methods when β = 1, Figure 3-4 adds the line resulting from the ROS with the ratios to Figure 3-2. The two lines are essentially identical, which means the two methods are equivalent.

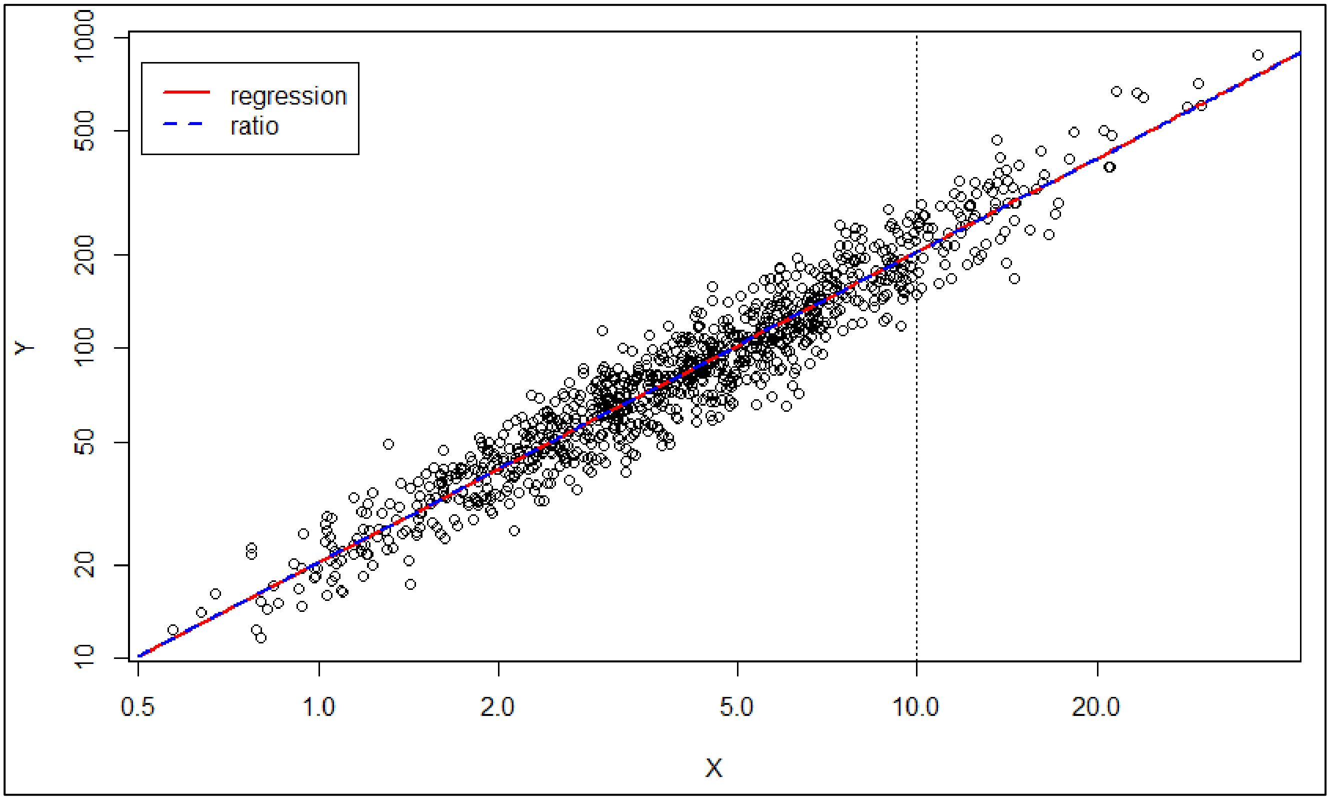

Figure 3-4. Data simulated with β = 1, μ = 3, and σ = 0.2 and the associated linear regression and lognormal ROS with ratios fits.

To illustrate how each of the models could be used, consider predicting *Y* for an *X* value of 10 (i.e., X_0 = 10; dotted vertical line in Figure 3-4). The predicted value of $\hat{Y}_{reg}$ from the regression model in Equation 1-3 is:

$$\hat{Y}_{reg} = \exp\left[\hat{\mu}_{reg} + \hat{\beta}_{reg}\log\left(X_0\right)\right] \quad (3\text{-}3)$$

where

$\hat{\mu}_{reg}$ = the estimate of μ from the regression in Table 3-1

$\hat{\beta}_{reg}$ = the estimate of β from the regression in Table 3-1

Therefore

$$\hat{Y}_{reg} = \exp\left[3.014 + 1.001 \times \log\left(10\right)\right] \quad (3\text{-}4)$$

$$\hat{Y}_{reg} = 204.2 \quad (3\text{-}5)$$

The predicted value of $\hat{Y}_{rat}$ from the ROS with ratios in Equation 3-1 is:

$$\hat{Y}_{rat} = \exp\left(\hat{\mu}_{rat}\right) \times X_0 \quad (3\text{-}6)$$

where

$\hat{\mu}_{rat}$ = the estimate of μ from the ROS with ratios from Table 3-2

Therefore

$$\hat{Y}_{rat} = \exp(3.015) \times 10 \quad (3\text{-}7)$$

$$\hat{Y}_{rat} = 203.9 \quad (3\text{-}8)$$

Because the two models agree very well when β = 1, there is little difference in the predicted values for X_0 = 10. See Attachment A for an example that illustrates these concepts with Fernald thorium data.

When β ≠ 1, the linear regression and lognormal ROS with ratios models will not agree and will produce different predicted values. The degree and direction of the differences depends on the value of β. Sections 4.0, 5.0, and 6.0 present simulations to illustrate how changing β affects predictions.

4.0 *X* AND *Y* HAVE A RELATIONSHIP AND 0 < β < 1

To illustrate the scenario when *X* and *Y* have a relationship and 0 < β < 1, the 1,000 pairs of data shown in Figure 4-1 are simulated with β = 0.5, μ = 3, and σ = 0.2.

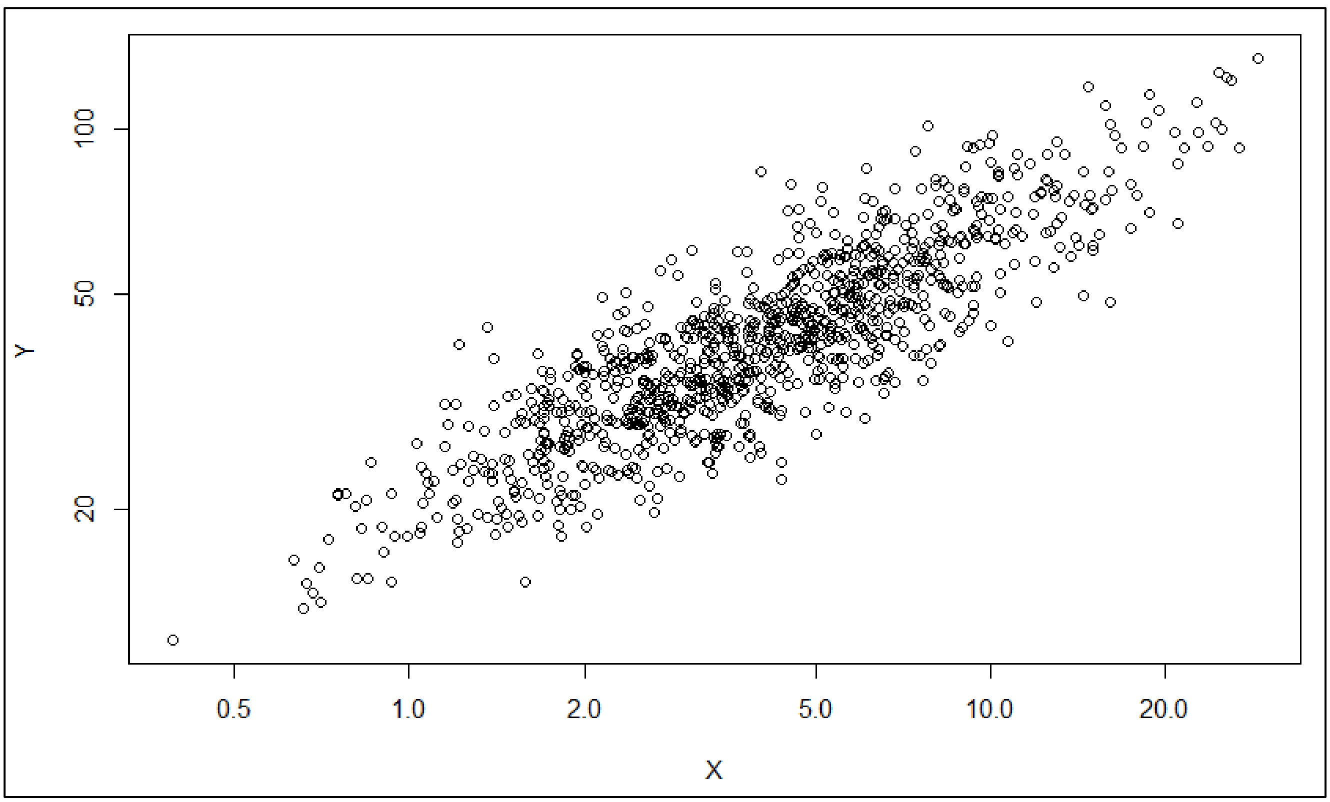

Figure 4-1. Data simulated with β = 0.5, μ = 3, and σ = 0.2.

4.1 LINEAR REGRESSION

Fitting the simulated data from Figure 4-1 using ordinary least squares and Equation 1-3 gives the parameter estimates in Table 4-1.

Table 4-1. Parameter estimates for regression using Equation 1-3.

Parameter	True Value	Estimate
β	0.5	0.4991
μ	3	3.007
σ	0.2	0.2048

As expected, the estimated parameters match the true parameters very well. Figure 4-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

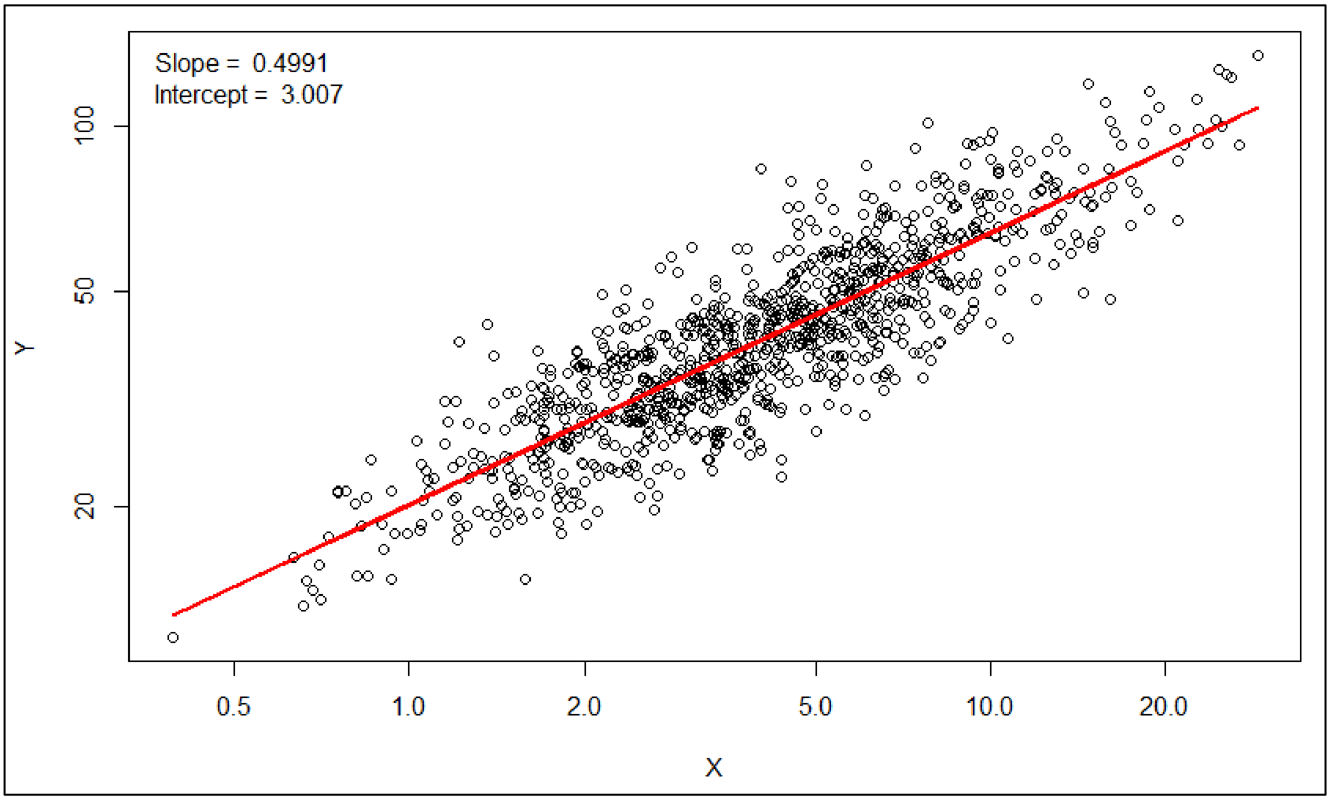

Figure 4-2. Data simulated with β = 0.5, μ = 3, and σ = 0.2 and the associated linear regression fit.

4.2 ROS WITH RATIOS

Taking the simulated data from Figure 4-1, calculating the ratio of *Y*/*X* for each of the 1,000 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table 4-2 and the lognormal probability plot in Figure 4-3.

Table 4-2. Parameter estimates for lognormal ROS with ratios.

Parameter	True Value	Estimate
β	0.5	(a)
μ	3	2.314
σ	0.2	0.405
GM	$e^{\mu} = 20.1$	10.12
GSD	$e^{\sigma} = 1.2$	1.499

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

Because β ≠ 1, the estimated parameters in Table 4-2 do not match the true parameters very well. The fit in Figure 4-3 looks good, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers.

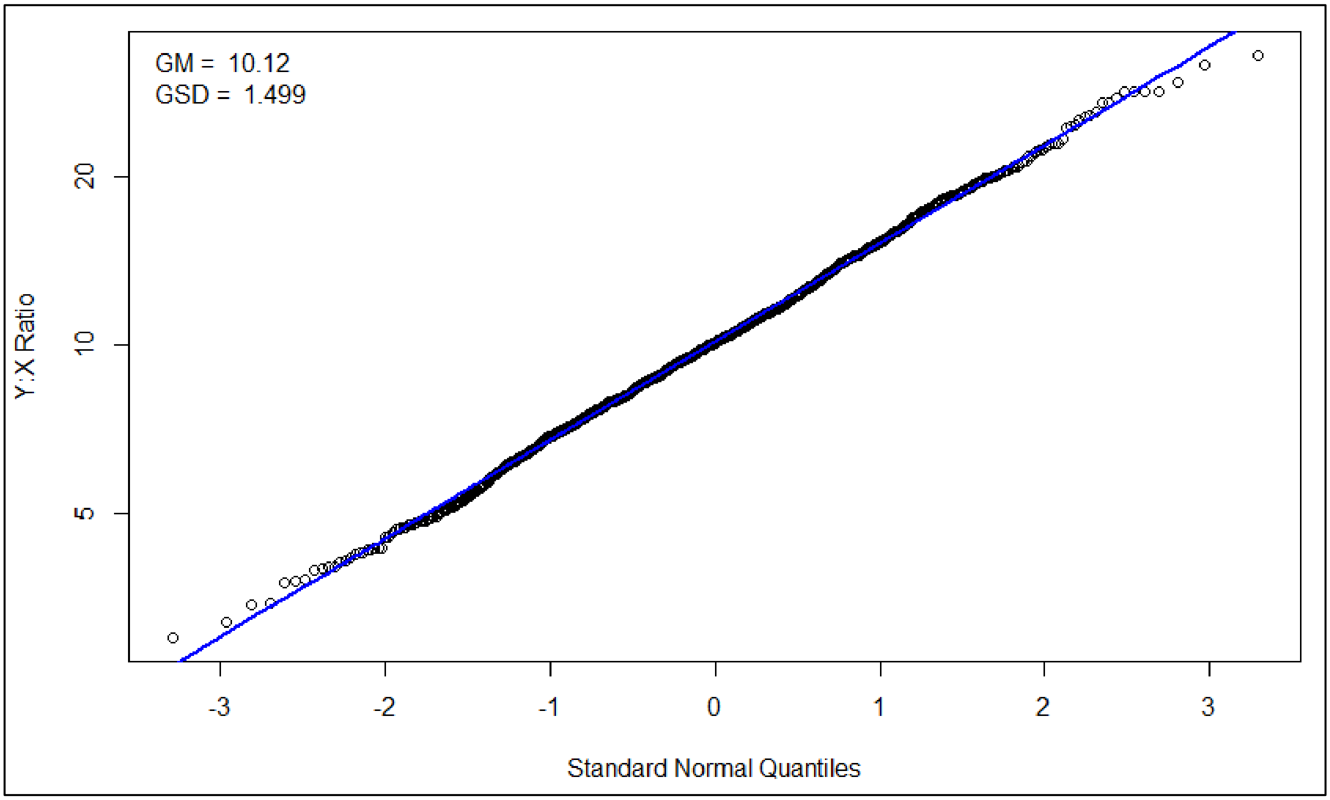

Figure 4-3. Data simulated with β = 0.5, μ = 3, and σ = 0.2 and the associated lognormal ROS with ratios fit.

4.3 COMPARISON OF REGRESSION AND RATIOS

The parameter estimates in Table 4-1 match the true values very well, but the estimates in Table 4-2 do not. To illustrate the performance of the two methods when β = 0.5, Figure 4-4 adds the line resulting from ROS with the ratios to Figure 4-2. The two lines intersect roughly in the middle of the horizontal axis. When β ≠ 1, the lines from the linear regression fit and the lognormal ROS with ratios will intersect at X_{int} [ORAUT 2023] where:

$$X_{int} = \exp\left(\frac{\hat{\mu}_{reg} - \hat{\mu}_{rat}}{1 - \hat{\beta}_{reg}}\right) \tag{4-1}$$

which for this simulated example is:

$$X_{\text{int}} = \exp\left(\frac{3.007 - 2.314}{1 - 0.4991}\right) \tag{4-2}$$

$$X_{\text{int}} = 3.985 \tag{4-3}$$

This means that for any *X* value less than 3.985, the ratio method will underestimate the predicted value of *Y* by some amount. For any *X* value greater than 3.985, the ratio method will overestimate the predicted value of *Y* by some amount. The amount of under- or overestimation depends on how far the value of *X* is from 3.985.

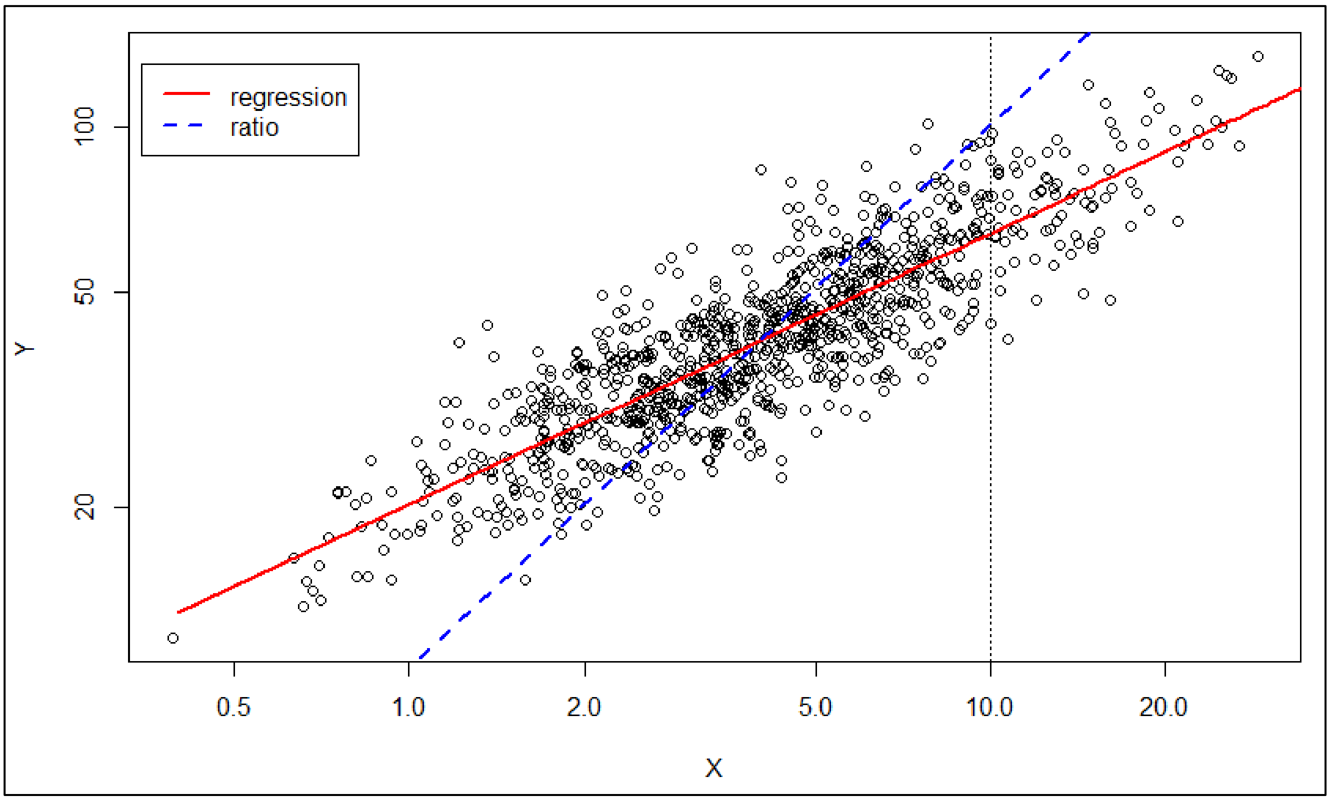

Figure 4-4. Data simulated with β = 0.5, μ = 3, and σ = 0.2 and the associated linear regression and lognormal ROS with ratios fits.

For example, suppose X_0 = 10 (dotted vertical line in Figure 4-4). The predicted value of *Y* from the regression model in Equation 3-3 is:

$$\hat{Y}_{\text{reg}} = \exp\left[3.007 + 0.4991 \times \log(10)\right] \tag{4-4}$$

$$\hat{Y}_{\text{reg}} = 63.83 \tag{4-5}$$

The predicted value of *Y* from the ROS with ratios in Equation 3-6 is:

$$\hat{Y}_{\text{rat}} = \exp(2.314) \times 10 \tag{4-6}$$

$$\hat{Y}_{rat} = 101.2 \tag{4-7}$$

See Attachment B for an example that illustrates these concepts with Paducah ^{99}Tc data.

5.0 *X* AND *Y* HAVE A RELATIONSHIP AND β > 1

To illustrate the scenario when *X* and *Y* have a relationship and β > 1, the 1,000 pairs of data shown in Figure 5-1 are simulated with β = 2, μ = 3, and σ = 0.2.

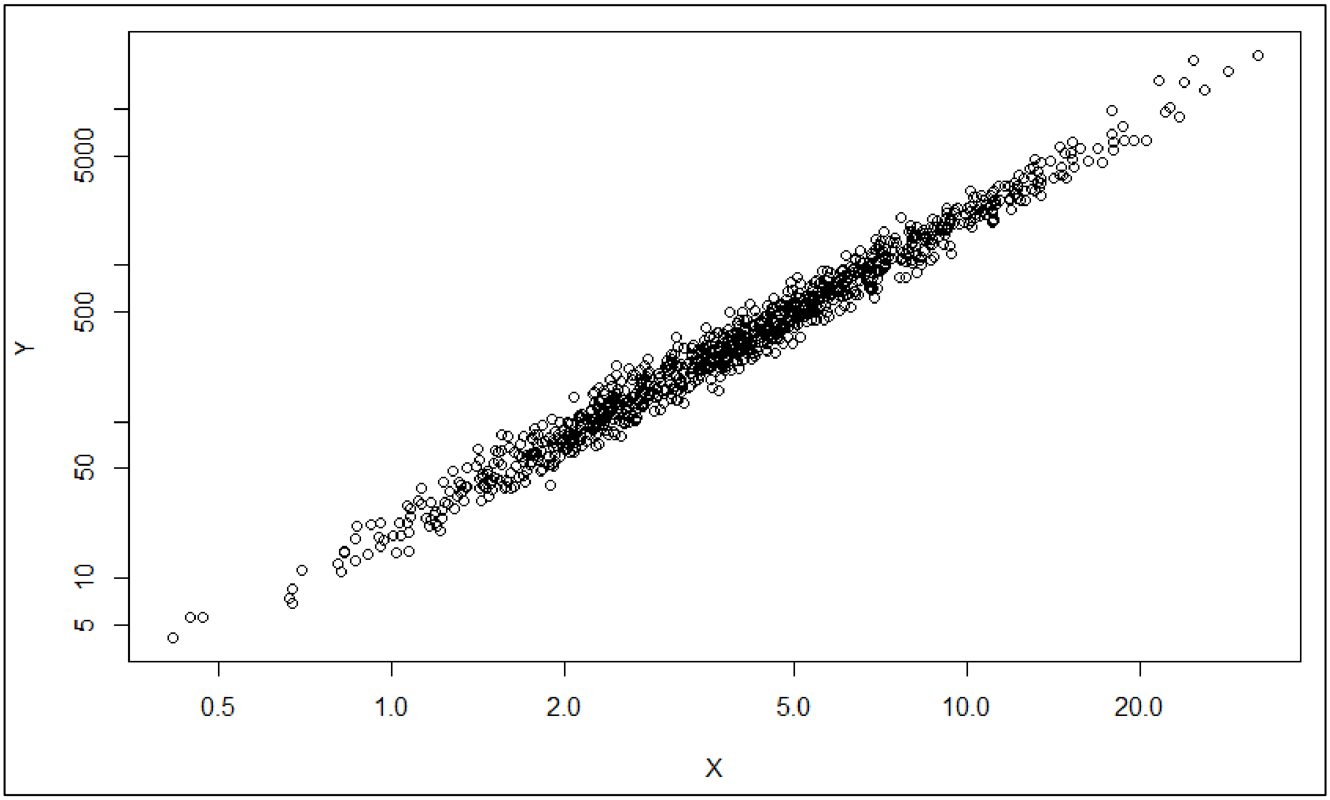

Figure 5-1. Data simulated with β = 2, μ = 3, and σ = 0.2.

5.1 LINEAR REGRESSION

Fitting the simulated data from Figure 5-1 using ordinary least squares and Equation 1-3 gives the parameter estimates in Table 5-1.

Table 5-1. Parameter estimates for regression using Equation 1-3.

Parameter	True Value	Estimate
β	2	2.007
μ	3	2.995
σ	0.2	0.1976

As expected, the estimated parameters match the true parameters very well. Figure 5-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

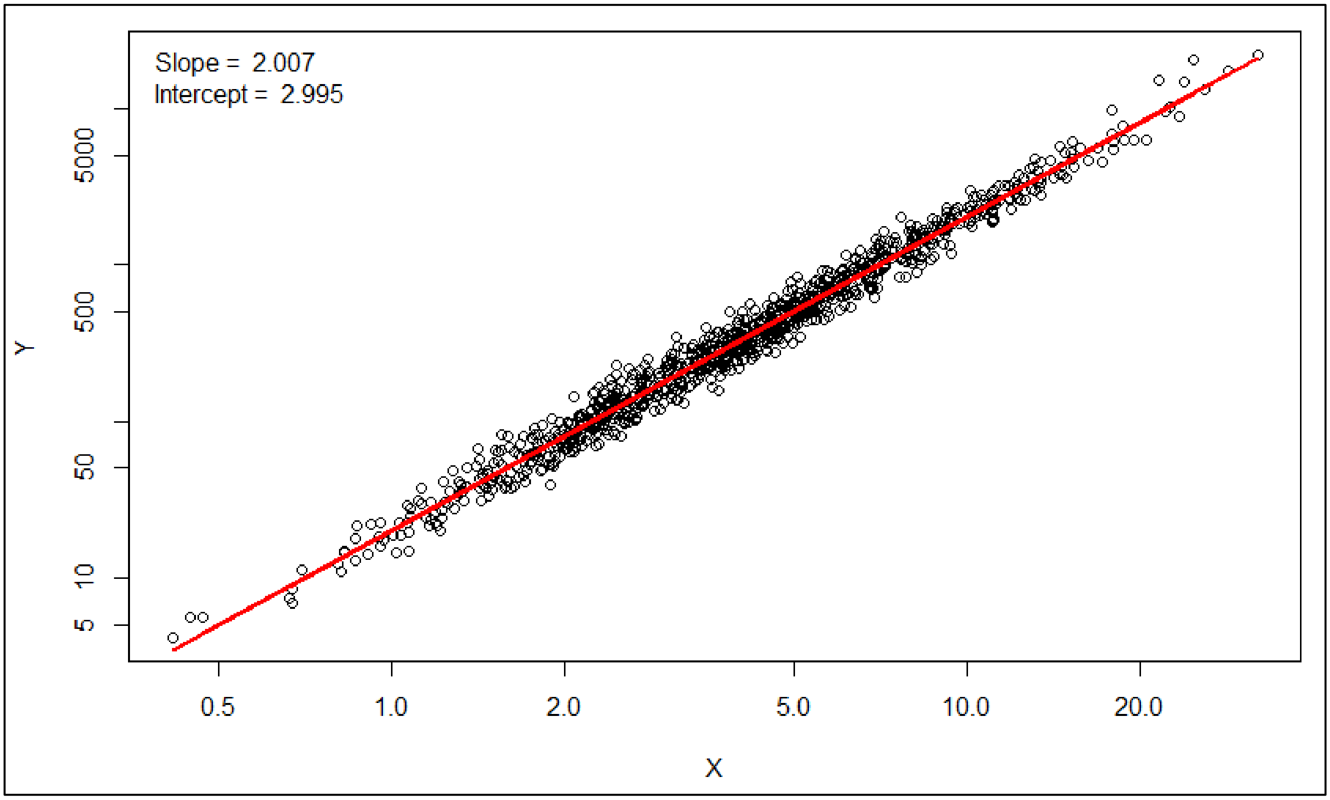

Figure 5-2. Data simulated with β = 2, μ = 3, and σ = 0.2 and the associated linear regression fit.

5.2 ROS WITH RATIOS

Taking the simulated data from Figure 5-1, calculating the ratio of *Y*/*X* for each of the 1,000 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table 5-2 and the lognormal probability plot in Figure 5-3.

Table 5-2. Parameter estimates for lognormal ROS with ratios.

Parameter	True Value	Estimate
β	2	(a)
μ	3	4.387
σ	0.2	0.7234
GM	e^{μ} = 20.1	80.43
GSD	e^{σ} = 1.2	2.061

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

Because β ≠ 1, the estimated parameters in Table 5-2 do not match the true parameters very well. The fit in Figure 5-3 looks good, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers.

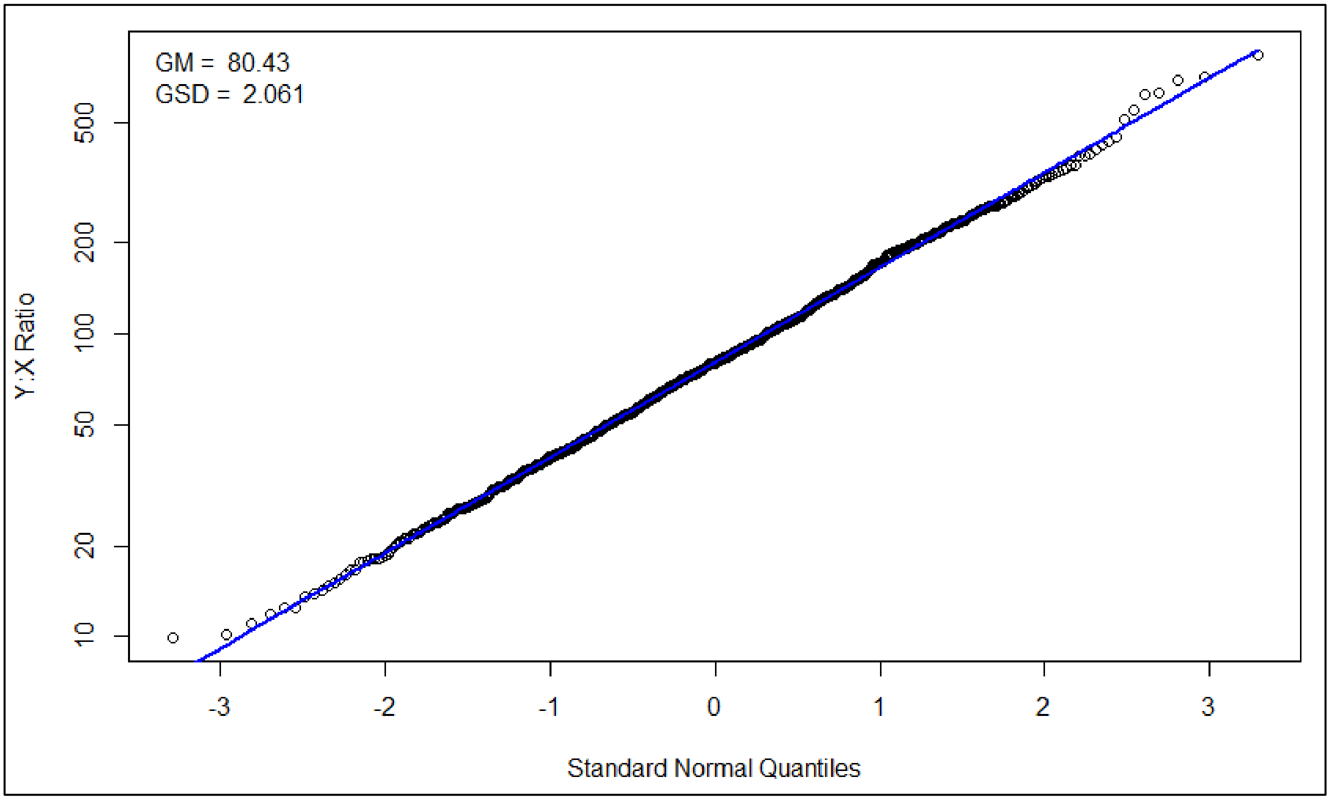

Figure 5-3. Data simulated with β = 2, μ = 3, and σ = 0.2 and the associated lognormal ROS with ratios fit.

5.3 COMPARISON OF REGRESSION AND RATIOS

The parameter estimates in Table 5-1 match the true values very well, but the estimates in Table 5-2 do not. To illustrate the performance of the two methods when β = 2, Figure 5-4 adds the line resulting from ROS with the ratios to Figure 5-2. The two lines intersect at:

$$X_{\text{int}} = \exp\left(\frac{2.995 - 4.387}{1 - 2.007}\right) \tag{5-1}$$

$$X_{\text{int}} = 3.986 \tag{5-2}$$

which is essentially the same value in Equation 4-3 because the true value of β was changed from 0.5 to 1/0.5 = 2 and the true values of μ and σ were not changed.

This means that for any *X* value less than 3.986, the ratio method will overestimate the correct predicted value of *Y* by some amount. For any *X* value greater than 3.986, the ratio method will underestimate the correct predicted value of *Y* by some amount.

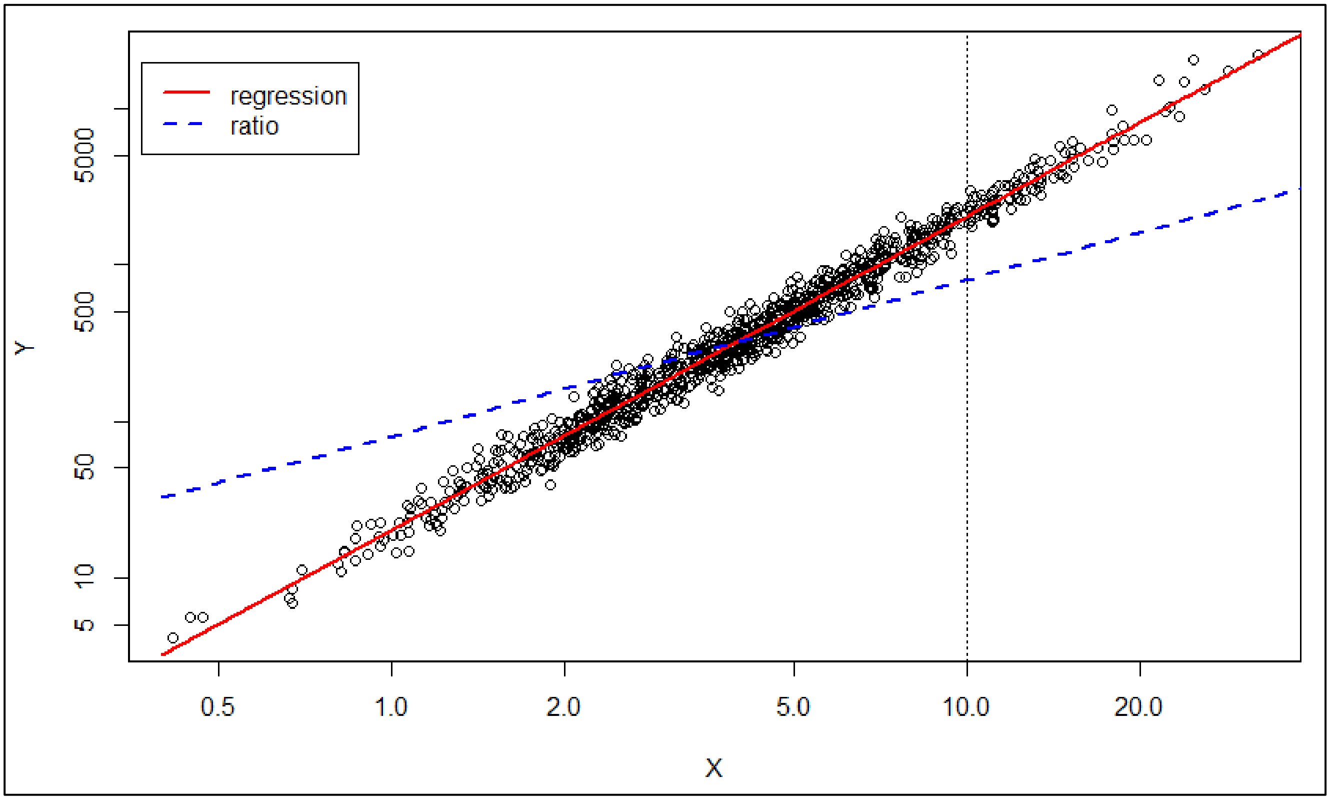

Figure 5-4. Data simulated with β = 2, μ = 3, and σ = 0.2 and the associated linear regression and lognormal ROS with ratios fits.

For example, suppose X_0 = 10 (dotted vertical line in Figure 5-4). The predicted value of *Y* from the regression model in Equation 3-3 is:

$$\hat{Y}_{reg} = \exp\left[2.995 + 2.007 \times \log(10)\right] \quad (5\text{-}3)$$

$$\hat{Y}_{reg} = 2{,}031 \quad (5\text{-}4)$$

The predicted value of *Y* from the ROS with ratios in Equation 3-6 is:

$$\hat{Y}_{rat} = \exp(4.387) \times 10 \quad (5\text{-}5)$$

$$\hat{Y}_{rat} = 804.3 \quad (5\text{-}6)$$

See Attachment C for an example that illustrates these concepts with Mound Neutron Source (NS) data.

6.0 *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0

To illustrate the scenario when *X* and *Y* have no relationship, which implies β = 0, 1,000 *X* values are randomly drawn from a lognormal distribution with GM = 4 and GSD = 2. One thousand *Y* values are randomly drawn from a lognormal distribution with GM = 3 and GSD = 3. Drawing randomly from these two distributions makes the pairs of data independent, meaning they have no relationship, and implies β is 0. The pairs of data are shown in Figure 6-1.

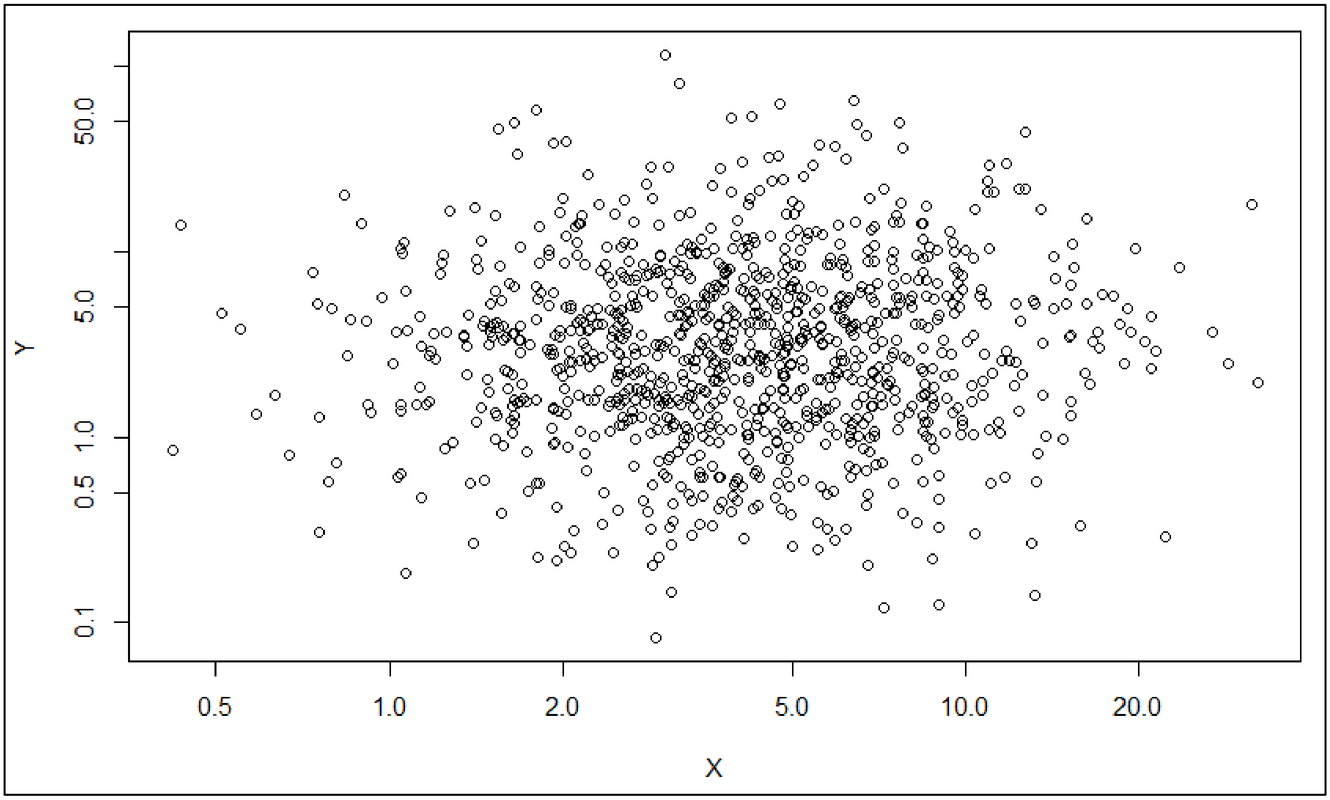

Figure 6-1. Data simulated from two independent lognormal distributions.

6.1 LINEAR REGRESSION

The data in Figure 6-1 clearly have no relationship. Essentially, the predicted value of *Y* will be the mean of *Y* despite the value of *X*, but for the sake of illustration the parameter estimates are included in Table 6-1. Note that the true values of μ and σ are simply the logarithm of the GM and logarithm of the GSD of *Y*, respectively, since β = 0.

Table 6-1. Parameter estimates for regression using Equation 1-3.

Parameter	True Value	Estimate
β	0	0.069
μ	log(3) = 1.099	0.9806
σ	log(3) = 1.099	1.127

As expected, the estimated β parameter is fairly close to zero. Figure 6-2 shows the data and regression line on logarithmic scales, so the relationship appears linear and essentially flat.

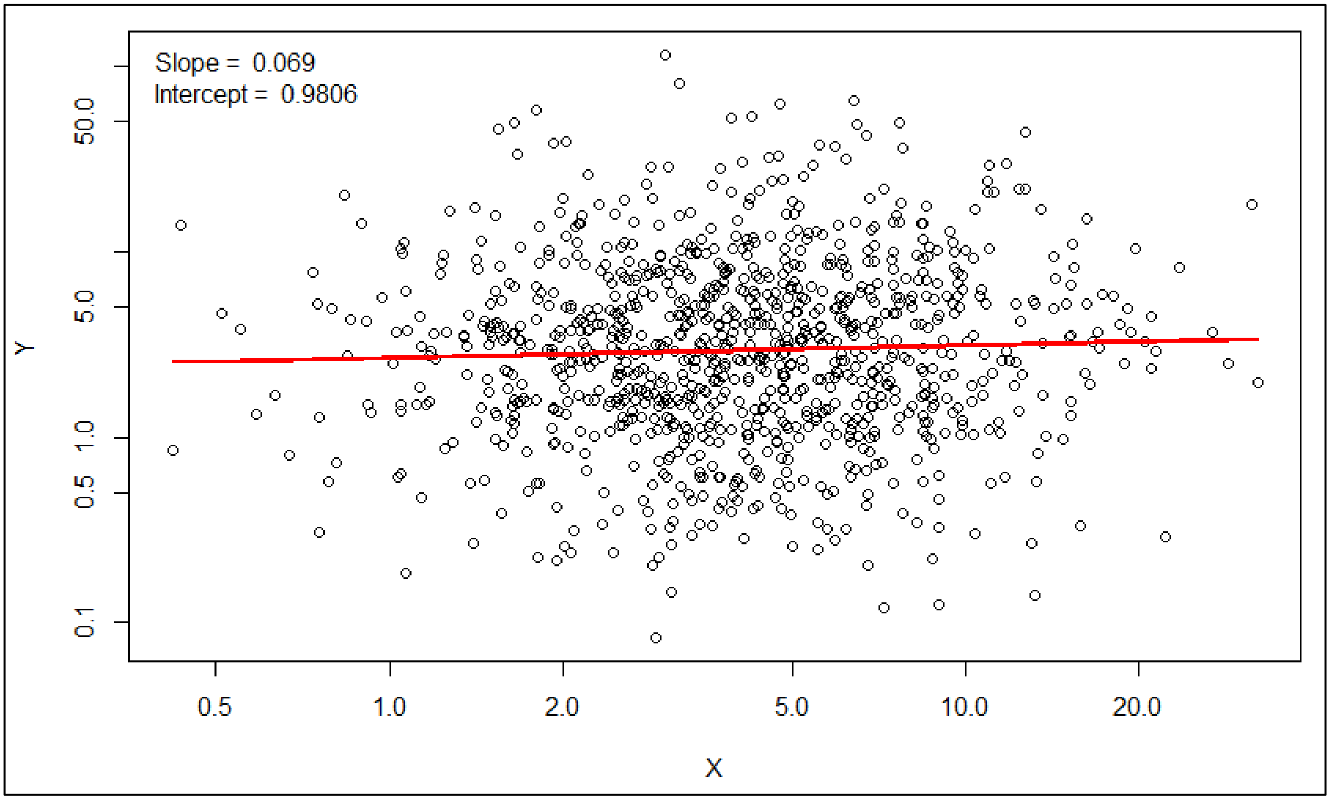

Figure 6-2. Data simulated from two independent lognormal distributions and the associated linear regression fit.

6.2 ROS WITH RATIOS

Taking the simulated data from Figure 6-1, calculating the ratio of *Y/X* for each of the 1,000 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table 6-2 and the lognormal probability plot in Figure 6-3.

Table 6-2. Parameter estimates for lognormal ROS with ratios.

Parameter	True Value	Estimate
β	0	(a)
μ	log(3) = 1.099	−0.3142
σ	log(3) = 1.099	1.296
GM	3	0.7303
GSD	3	3.656

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

The fit in Figure 6-3 looks good, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers. Nothing about the parameters in Table 6-2 or the fit in Figure 6-3 indicates that *X* and *Y* have no relationship. By collapsing independent *X* and *Y* into a one-dimensional ratio, two variables that have no relationship now appear to have a meaningful relationship.

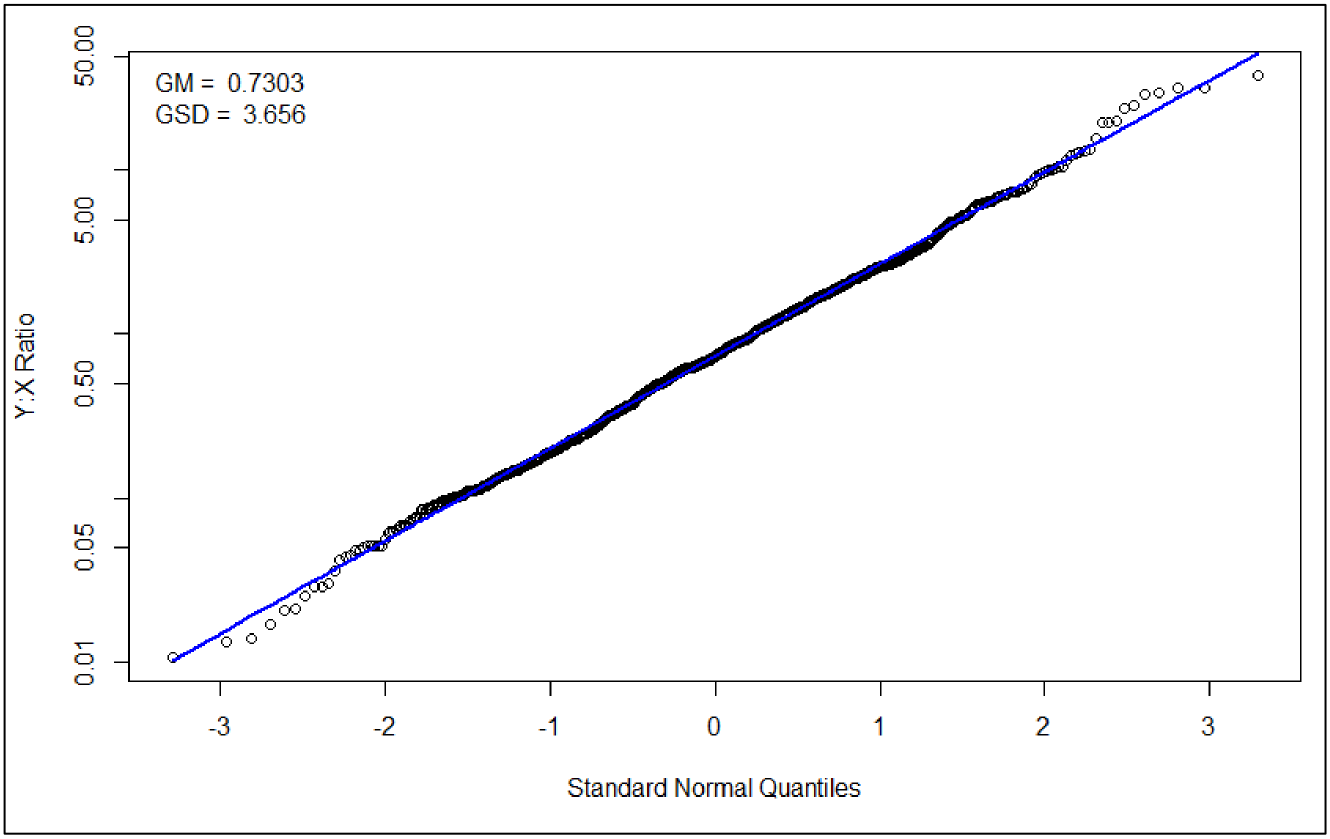

Figure 6-3. Data simulated from two independent lognormal distributions and the associated lognormal ROS with ratios fit.

In fact, if *X* is lognormally distributed with parameters GM_X and GSD_X and *Y* is lognormally distributed with parameters GM_Y and GSD_Y, then:

$$\log(X) \sim N(\mu_X, \sigma_X) \tag{6-1}$$

$$\log(Y) \sim N(\mu_Y, \sigma_Y) \tag{6-2}$$

If *X* and *Y* are independent, which is assumed throughout this section, it follows that [Casella and Berger 2002, p. 211]:

$$\log\left(\frac{Y}{X}\right) \sim N\left(\mu_Y - \mu_X, \sqrt{\sigma_Y^2 + \sigma_X^2}\right) \tag{6-3}$$

This means that, mathematically, two independent lognormal distributions will always produce a ratio distribution that is also lognormal.

6.3 COMPARISON OF REGRESSION AND RATIOS

To illustrate the performance of the two methods when X and Y have no relationship and β = 0, Figure 6-4 adds the line resulting from the ROS with the ratios to Figure 6-2. The two lines intersect at:

$$X_{\text{int}} = \exp\left(\frac{0.9806-(-0.3142)}{1-0.069}\right) \tag{6-4}$$

$$X_{\text{int}} = 4.018 \tag{6-5}$$

This means that for any X value less than 4.018, the ratio method will underestimate the correct predicted value of Y by some amount. For any X value greater than 4.018, the ratio method will overestimate the correct predicted value of Y by some amount.

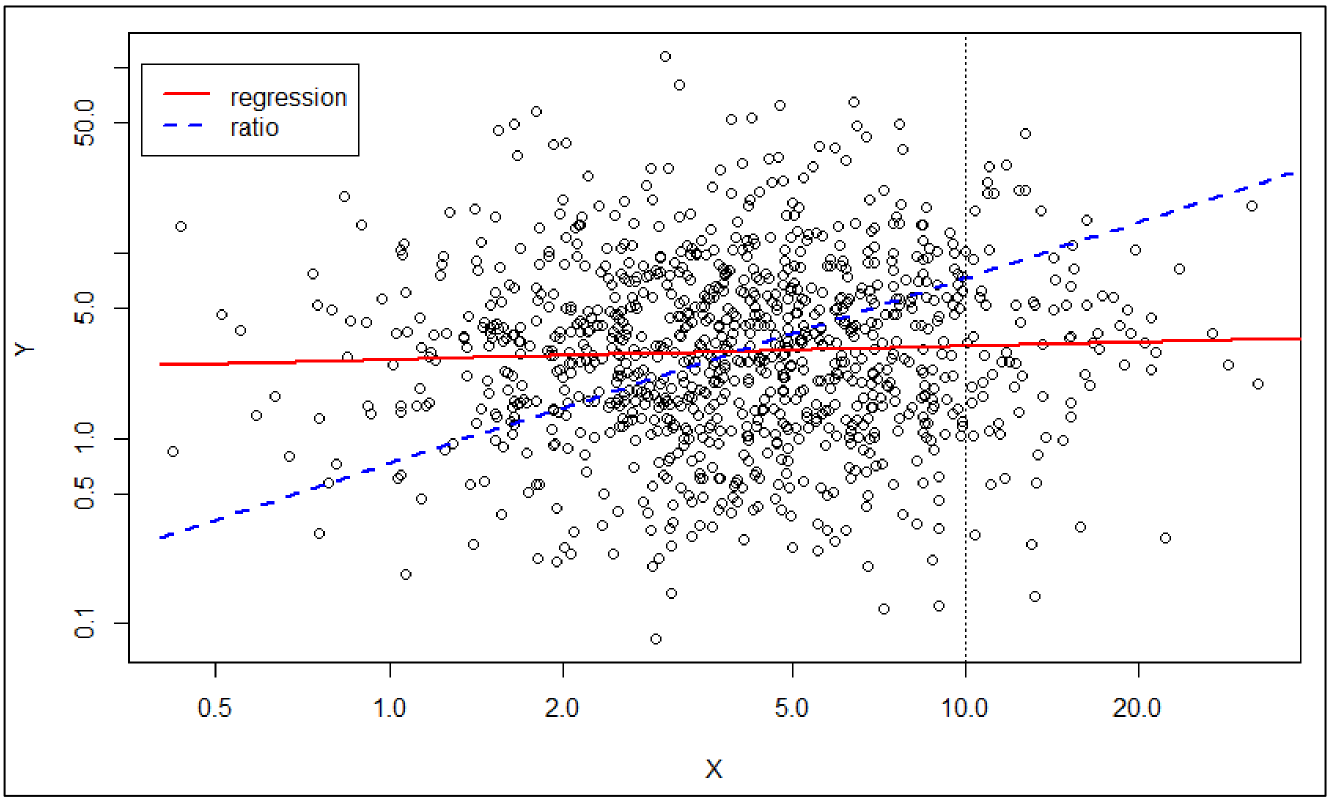

Figure 6-4. Data simulated from two independent lognormal distributions and the associated linear regression and lognormal ROS with ratios fits.

For example, suppose X_0 = 10 (dotted vertical line in Figure 6-4). The predicted value of Y from the regression model in Equation 3-3 is:

$$\hat{Y}_{\text{reg}} = \exp\left[0.9806+0.069\times\log(10)\right] \tag{6-6}$$

$$\hat{Y}_{\text{reg}} = 3.125 \tag{6-7}$$

The predicted value of *Y* from the ROS with ratios in Equation 3-6 is:

$$\hat{Y}_{\text{rat}} = \exp\left(-0.3142\right) \times 10 \quad (6\text{-}8)$$

$$\hat{Y}_{\text{rat}} = 7.303 \quad (6\text{-}9)$$

The essentially flat linear regression fit indicates the predicted value of *Y* does not depend on the observed value of *X*. However, the lognormal fit to the ratio indicates that the predicted value of *Y* should change as a function of *X*, which is obviously inappropriate based on Figure 6-2 and the fact that *X* and *Y* have no relationship. In this case, the ratio creates the illusion of a meaningful relationship where none exists. See Attachment D for an example that illustrates these concepts with Rocky Flats neutron and gamma dose data.

In general, before statistical analysis of any kind is done, subject matter experts should consider whether there should be a relationship between two variables. Based on subject knowledge, if there is no reason to suspect that a relationship exists, there should be no attempt to do any modeling. If there is reason to suspect that a relationship should exist, and the data look similar to Figure 6-4, more exploratory analyses should be done to determine whether relationships exist amongst subgroups of the dataset.

7.0 EFFECT OF β ON THE RATIO METHOD

The value of β affects the degree and direction of the under- or overestimation of *Y* using lognormal ROS with ratios. For the remainder of this section, assume μ and σ are held fixed. The worst-case scenario, in terms of the amount of underestimation of *Y* from the ratio method for *X* values less than X_{int} and overestimation of *Y* for values larger than X_{int}, is when β = 0 (Section 6.0). For 0 < β < 1, the amount of under- and overestimation of *Y* will decrease as β increases (Section 4.0), with the ratio method being correct only for β = 1 (Section 3.0, no under- or overestimation). For β > 1, as β increases, the ratio will overestimate *Y* for *X* values less than X_{int} and underestimate *Y* for values greater than X_{int} (Section 5.0). These effects of changing β are also summarized using a simulation [ORAUT 2023].

8.0 SUMMARY

Overall, the key points of this discussion are:

- The standard method of fitting a lognormal distribution to the ratio of *Y*/*X* (i.e., ROS) is valid only for the model in Equation 1-3 when β = 1.

- For data generated by any statistical model other than that in Equation 1-3, ROS and other methods that might reasonably be used with ratios (like empirical quantiles) will give biased estimates of the GM, GSD, and other quantiles of interest (such as the 95th percentile). The lognormal probability plot of the ratio cannot indicate if the parameter estimates are biased.

- ROS with ratios will generate estimates of the GM and GSD even when there is no meaningful relationship between *X* and *Y*. The lognormal probability plot of the ratio cannot be used to diagnose this problem.

The main recommendations to draw from these conclusions are:

- Always examine the scatterplot of *Y* versus *X* to determine if a relationship exists between *Y* and *X*.

- If there is no meaningful relationship (linear or nonlinear) between *Y* and *X*, avoid creating the appearances of one by modeling the *Y*/*X* ratio.

- If a meaningful relationship exists, it is preferable to model the bivariate relationship between *Y* and *X* using standard regression or quantile regression [ORAUT 2018] rather than model the relationship of *Y*/*X* as a univariate (e.g., lognormal) distribution.

- Reserve modeling the ratio *Y*/*X* for situations in which all other bivariate methods have been exhausted, and use it only with great care.

REFERENCES

Casella G, Berger RL [2002]. Statistical inference. 2nd ed. Pacific Grove, CA: Duxbury. [SRDB Ref ID: 195943]

Helsel DR [2012]. Statistics for censored environmental data using Minitab and R. 2nd ed. Hoboken, NJ: John Wiley & Sons. [SRDB Ref ID: 150353]

ORAUT [2018]. Applications of regression in external dose reconstruction. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-RPRT-0087 Rev. 00, March 13. [SRDB Ref ID: 170100]

ORAUT [2023]. ORAUT-RPRT-0106 Rev. 00 support files. Zip file. Oak Ridge, TN: Oak Ridge Associated Universities Team. February 28. [SRDB Ref ID: 195944]

Weisberg S [2005]. Applied linear regression. 3rd ed. Hoboken, NJ: John Wiley & Sons. [SRDB Ref ID: 169440]

ATTACHMENT A
FERNALD THORIUM DATA:
X AND *Y* HAVE A RELATIONSHIP AND β = 1

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ATTACHMENT A
FERNALD THORIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β = 1 (continued)**

A.1 DATA

Figure A-1 shows a scatterplot of Fernald thorium data from 1994 to 1996.

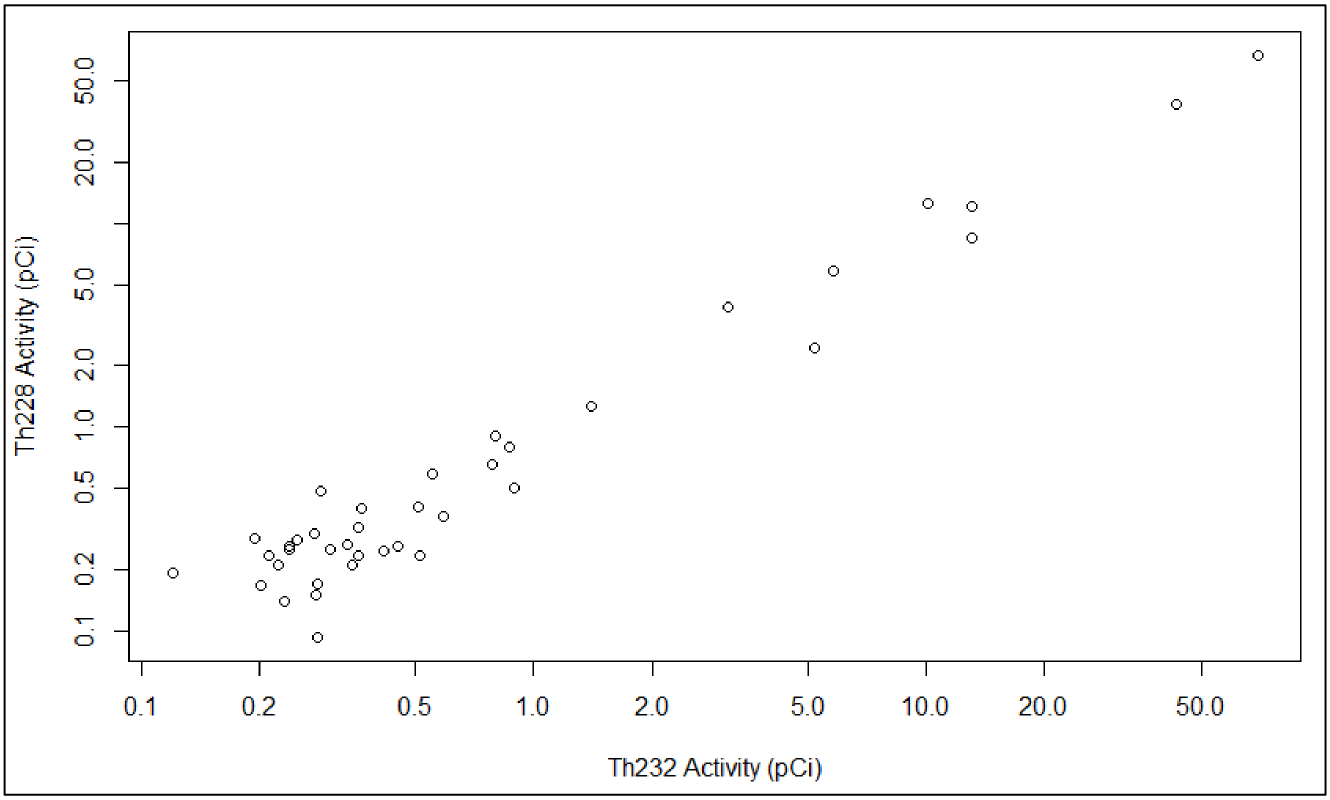

Figure A-1. Fernald thorium data from 1994 to 1996.

A.2 LINEAR REGRESSION

Fitting the data from Figure A-1 using ordinary least squares gives the parameter estimates in Table A-1.

Table A-1. Parameter estimates for regression.

Parameter	Estimate
β	0.9997
μ	−0.1775
σ	0.3635

Figure A-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

ATTACHMENT A
FERNALD THORIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β = 1 (continued)**

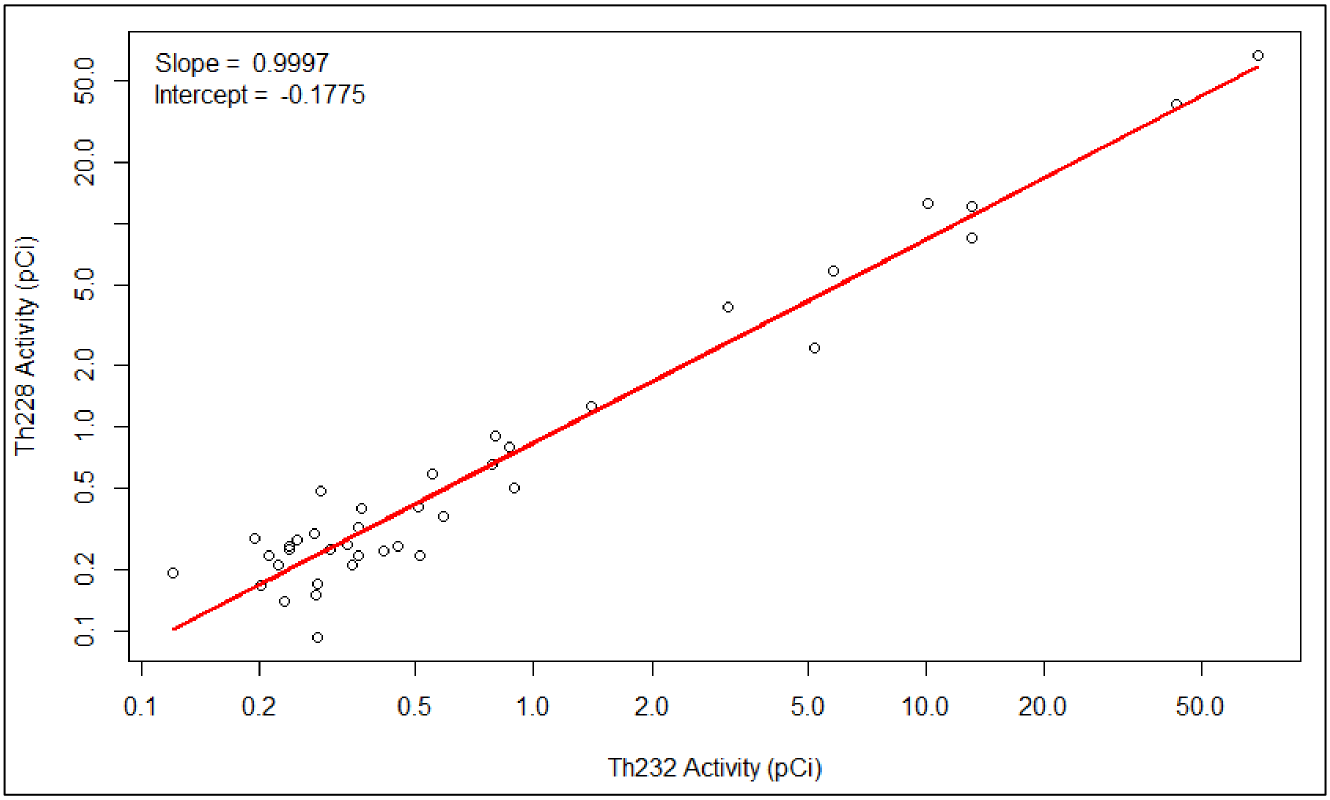

Figure A-2. Fernald thorium data from 1994 to 1996 and the associated linear regression fit.

A.3 ROS WITH RATIOS

Taking the data from Figure A-1, calculating the ratio of *Y*/*X* for each of the 39 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table A-2 and the lognormal probability plot in Figure A-3.

Table A-2. Parameter estimates for lognormal ROS with ratios.

Parameter	Estimate
β	(a)
μ	−0.1774
σ	0.3551
GM	0.8374
GSD	1.426

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

ATTACHMENT A
FERNALD THORIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β = 1 (continued)**

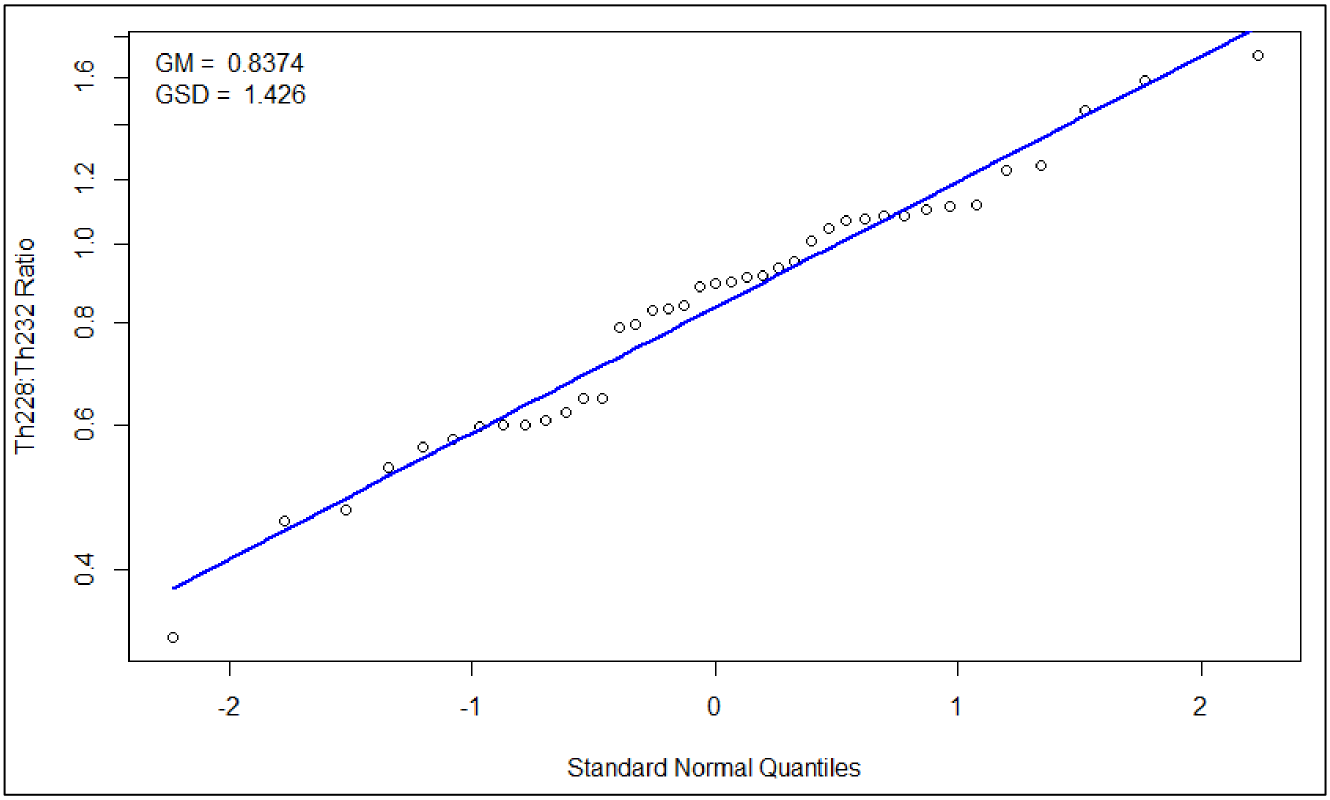

Figure A-3. Fernald thorium data from 1994 to 1996 and the associated lognormal ROS with ratios fit.

A.4 COMPARISON OF REGRESSION AND RATIOS

For β = 1, the linear regression in Section A.2 and the ROS with ratios in Section A.3 are equivalent. The parameter estimates of μ and σ in Tables A-1 and A-2 match each other very well. To further illustrate the equivalence of the two methods when β = 1, Figure A-4 adds the line resulting from ROS with the ratios to Figure A-2. The two lines are essentially identical, which means the two methods are equivalent.

ATTACHMENT A
FERNALD THORIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β = 1 (continued)**

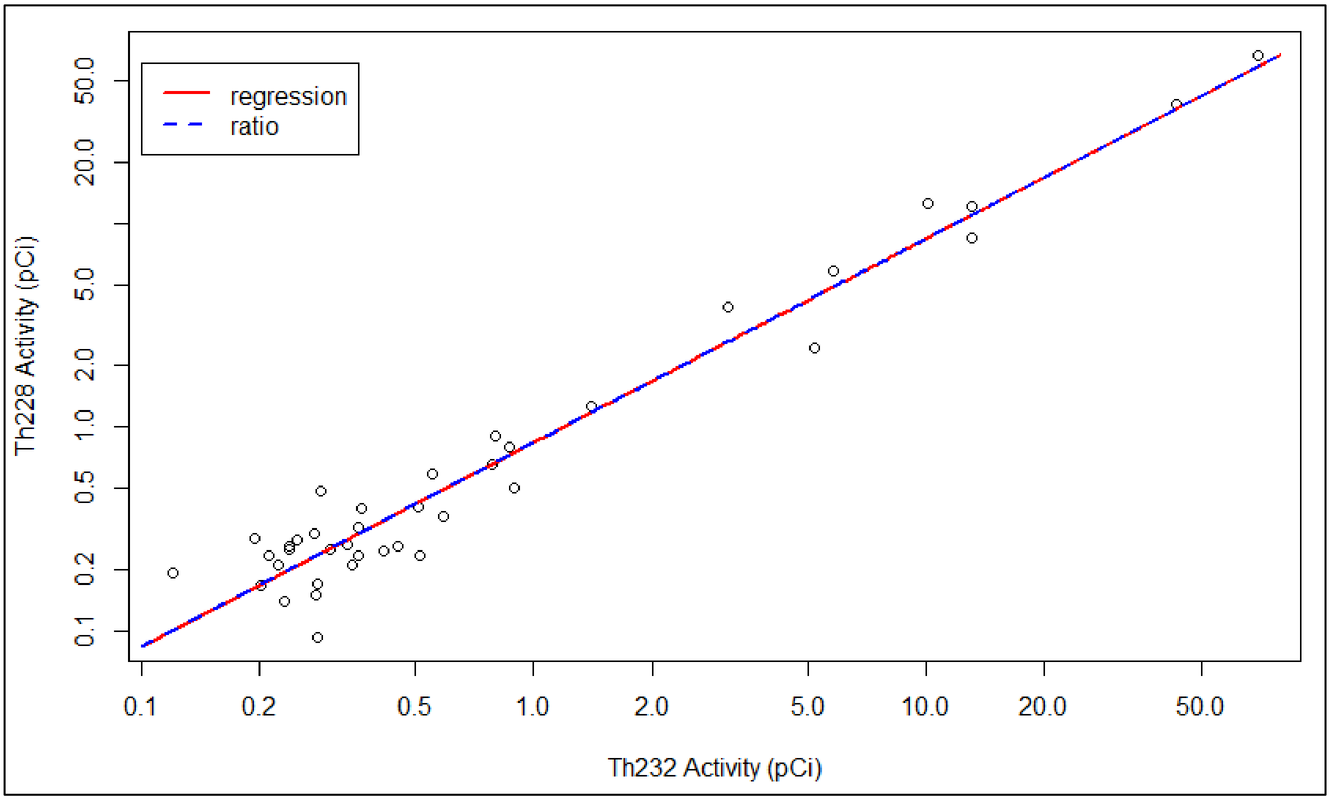

Figure A-4. Fernald thorium data from 1994 to 1996 and the associated linear regression and lognormal ROS with ratios fits.

Because the two models agree very well when β = 1, the models would produce little difference in the predicted values for any value of *X*, as seen in Section 3.3, so no example calculations are done here.

ATTACHMENT B
PADUCAH TECHNETIUM DATA:
X AND *Y* HAVE A RELATIONSHIP AND $0 < \beta < 1$

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ATTACHMENT B
PADUCAH TECHNETIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND 0 < β < 1 (continued)**

B.1 DATA

Figure B-1 shows a scatterplot of Paducah ^{99}Tc annual releases and intakes from 1980 to 1999, excluding 1996 and 1997.

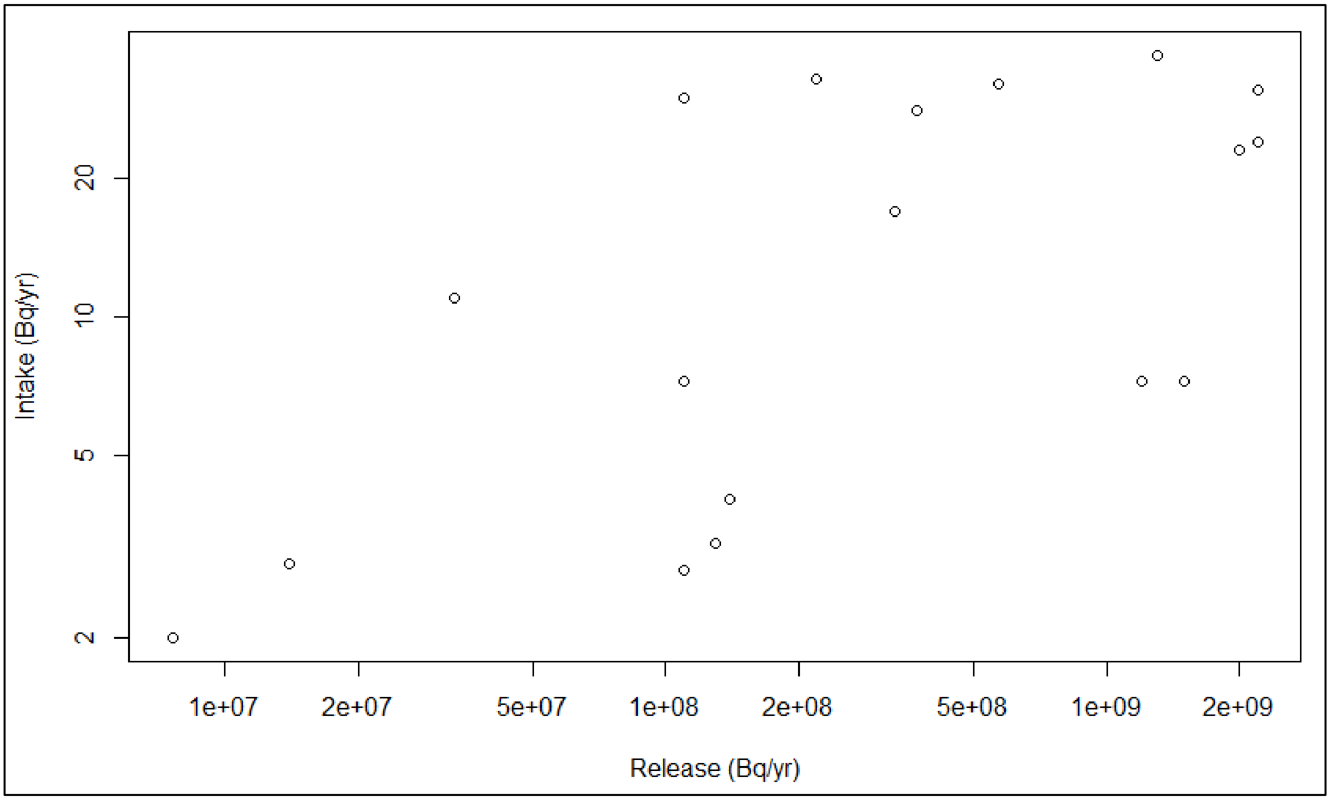

Figure B-1. Paducah annual air data from 1980 to 1999 (excluding 1996 and 1997).

B.2 LINEAR REGRESSION

Fitting the data from Figure B-1 using ordinary least squares gives the parameter estimates in Table B-1.

Table B-1. Parameter estimates for regression.

Parameter	Estimate
β	0.3787
μ	−4.918
σ	0.8105

Figure B-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

ATTACHMENT B
PADUCAH TECHNETIUM DATA:
X AND *Y* HAVE A RELATIONSHIP AND 0 < β < 1 (continued)

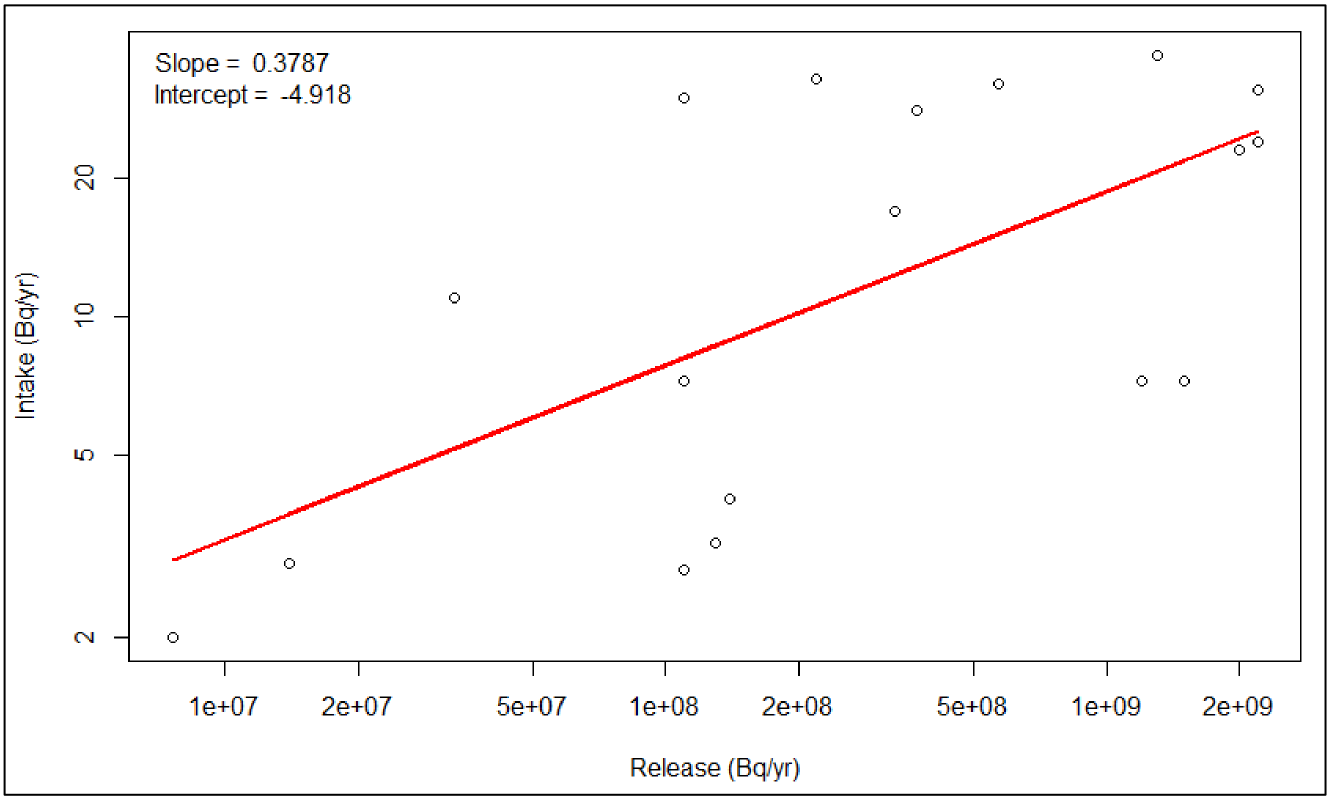

Figure B-2. Paducah annual air data from 1980 to 1999 (excluding 1996 and 1997) and the associated linear regression fit.

B.3 ROS WITH RATIOS

Taking the data from Figure B-1, calculating the ratio of *Y*/*X* for each of the 18 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table B-2 and the lognormal probability plot in Figure B-3.

Table B-2. Parameter estimates for lognormal ROS with ratios.

Parameter	Estimate
β	(a)
μ	−16.96
σ	1.313
GM	4.321E−8
GSD	3.718

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

The fit in Figure B-3 looks good, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers.

ATTACHMENT B
PADUCAH TECHNETIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND 0 < β < 1 (continued)**

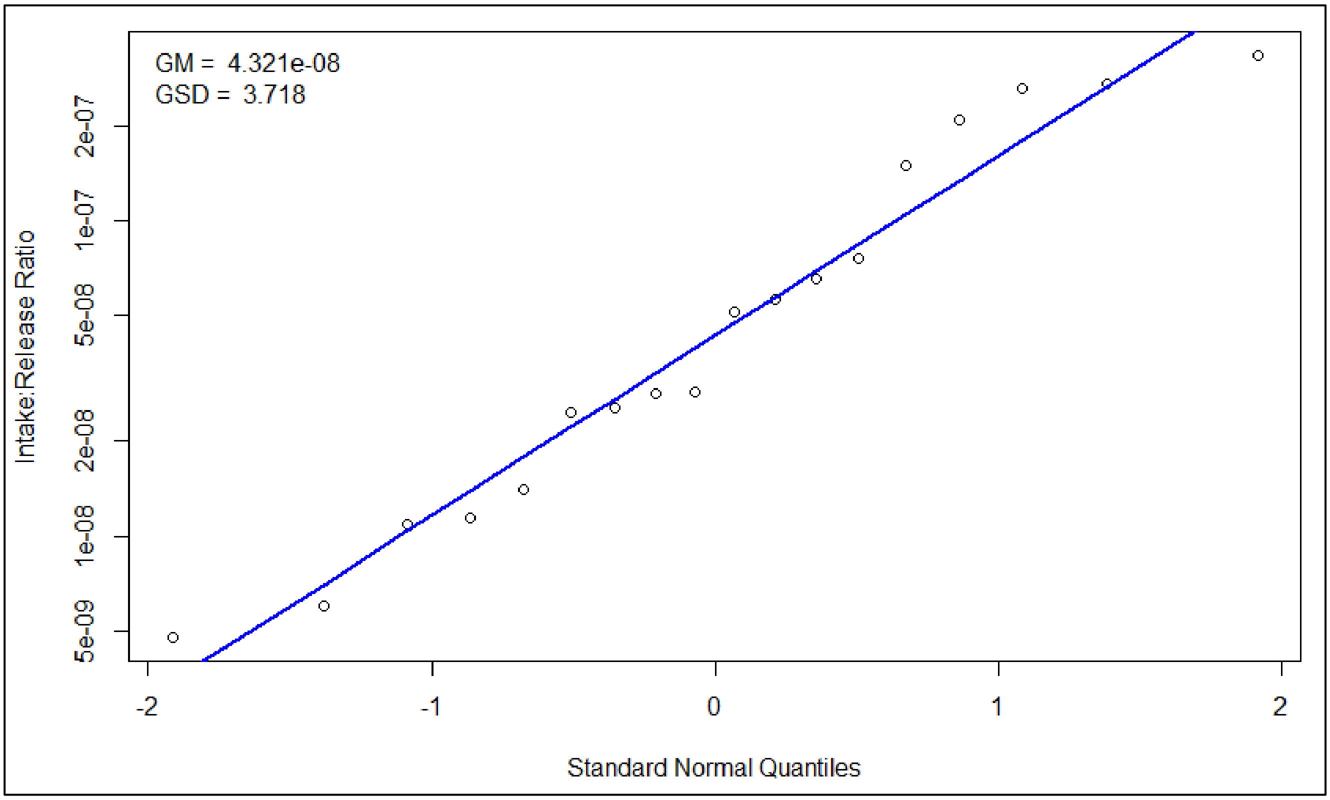

Figure B-3. Paducah annual air data from 1980 to 1999 (excluding 1996 and 1997) and the associated lognormal ROS with ratios fit.

B.4 COMPARISON OF REGRESSION AND RATIOS

To illustrate the performance of the two methods for this dataset, Figure B-4 adds the line resulting from ROS with the ratios to Figure B-2. The two lines intersect at:

$$X_{\text{int}} = \exp\left(\frac{-4.918 - (-16.96)}{1 - 0.3787}\right) \tag{B-1}$$

$$X_{\text{int}} = 2.603 \times 10^{8} \text{ Bq/yr} \tag{B-2}$$

This means the ratio method will underestimate the best estimate of predicted intake for releases less than 2.603×10^{8} Bq/yr and overestimate the best estimate of predicted intake for releases greater than 2.603×10^{8} Bq/yr.

ATTACHMENT B
PADUCAH TECHNETIUM DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND 0 < β < 1 (continued)**

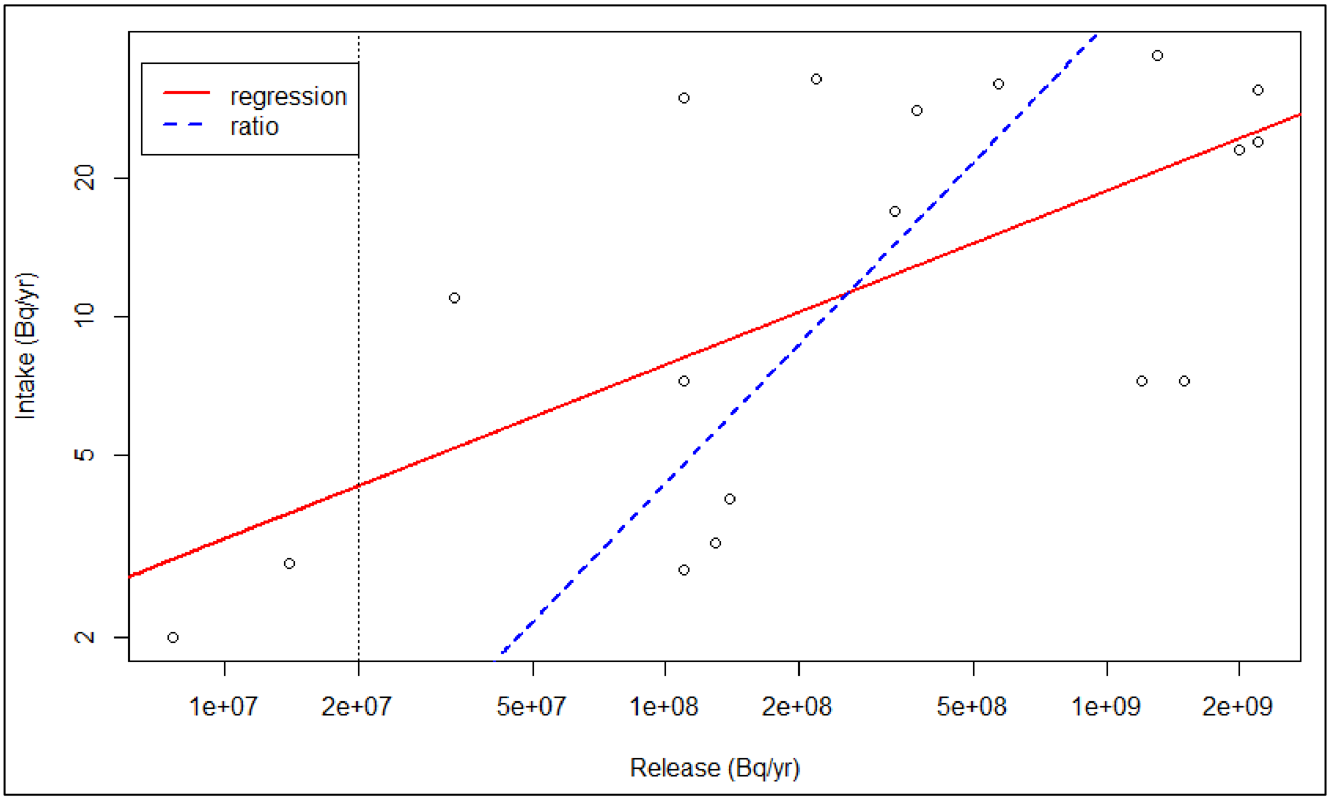

Figure B-4. Paducah annual air data from 1980 to 1999 (excluding 1996 and 1997) and the associated linear regression and lognormal ROS with ratios fits.

For example, for a release of 2 × 10^7 Bq/yr (dotted vertical line in Figure B-4), the predicted intake from the regression model is:

$$\hat{Y}_{reg} = \exp\left[-4.918 + 0.3787 \times \log\left(2 \times 10^7\right)\right] \quad \text{(B-3)}$$

$$\hat{Y}_{reg} = 4.256 \text{ Bq/yr} \quad \text{(B-4)}$$

The predicted intake from the ROS with ratios is:

$$\hat{Y}_{rat} = \exp(-16.96)\left(2 \times 10^7\right) \quad \text{(B-5)}$$

$$\hat{Y}_{rat} = 0.8642 \text{ Bq/yr} \quad \text{(B-6)}$$

For a release of 2 × 10^7 Bq/yr, the ROS with ratio model underestimates by a factor of approximately 5.

ATTACHMENT C
MOUND NEUTRON SOURCE DATA:
X AND *Y* HAVE A RELATIONSHIP AND β > 1

TABLE OF CONTENTS

SECTION **TITLE** **PAGE**

LIST OF TABLES

TABLE **TITLE** **PAGE**

LIST OF FIGURES

FIGURE **TITLE** **PAGE**

ATTACHMENT C
MOUND NEUTRON SOURCE DATA:
X AND *Y* HAVE A RELATIONSHIP AND β > 1 (continued)

C.1 DATA

Figure C-1 shows a scatterplot of neutron and photon doses from the NS group at Mound in 1954 and 1957.

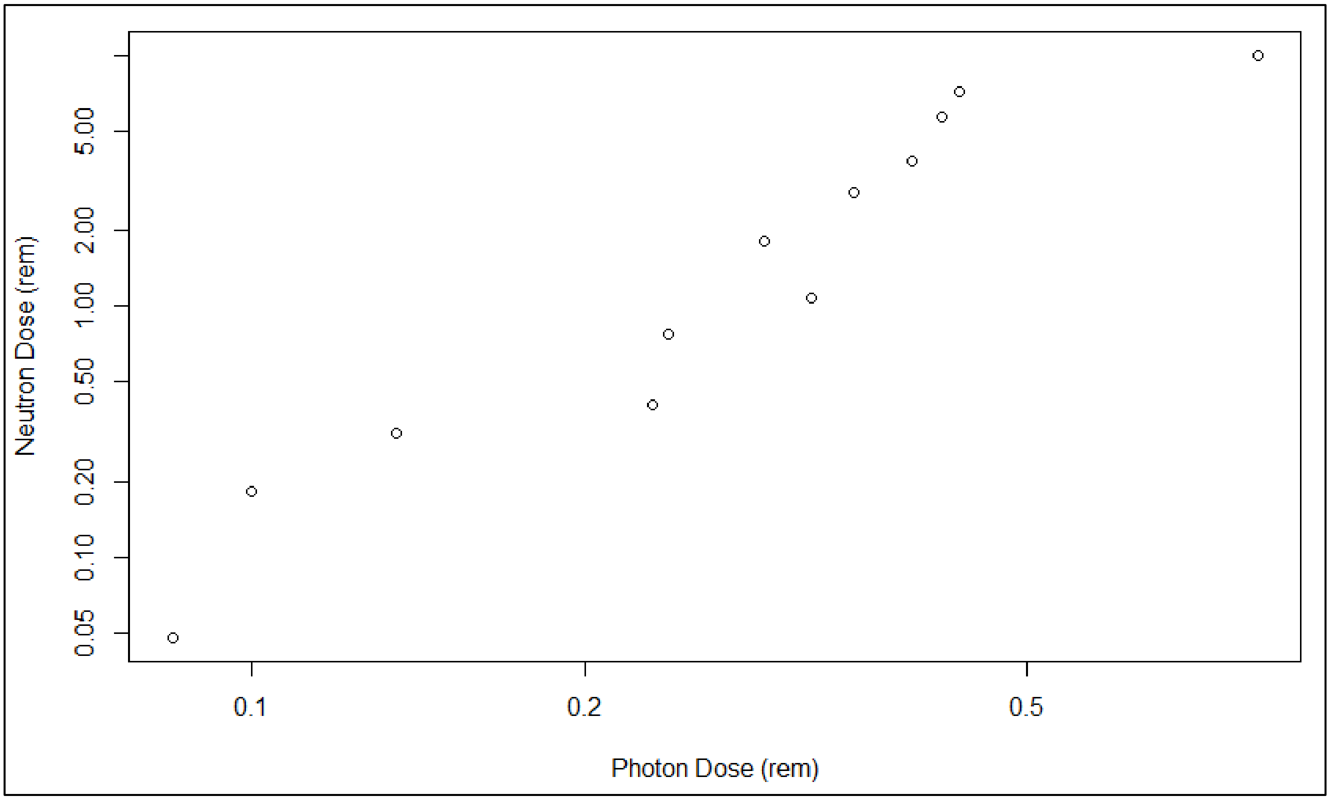

Figure C-1. Mound neutron and photon data from NS in 1954 and 1957.

C.2 LINEAR REGRESSION

Fitting the data from Figure C-1 using ordinary least squares gives the parameter estimates in Table C-1.

Table C-1. Parameter estimates for regression.

Parameter	Estimate
β	2.396
μ	3.353
σ	0.490

Figure C-2 shows the data and regression line on logarithmic scales, so the relationship appears linear.

ATTACHMENT C
MOUND NEUTRON SOURCE DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β > 1 (continued)**

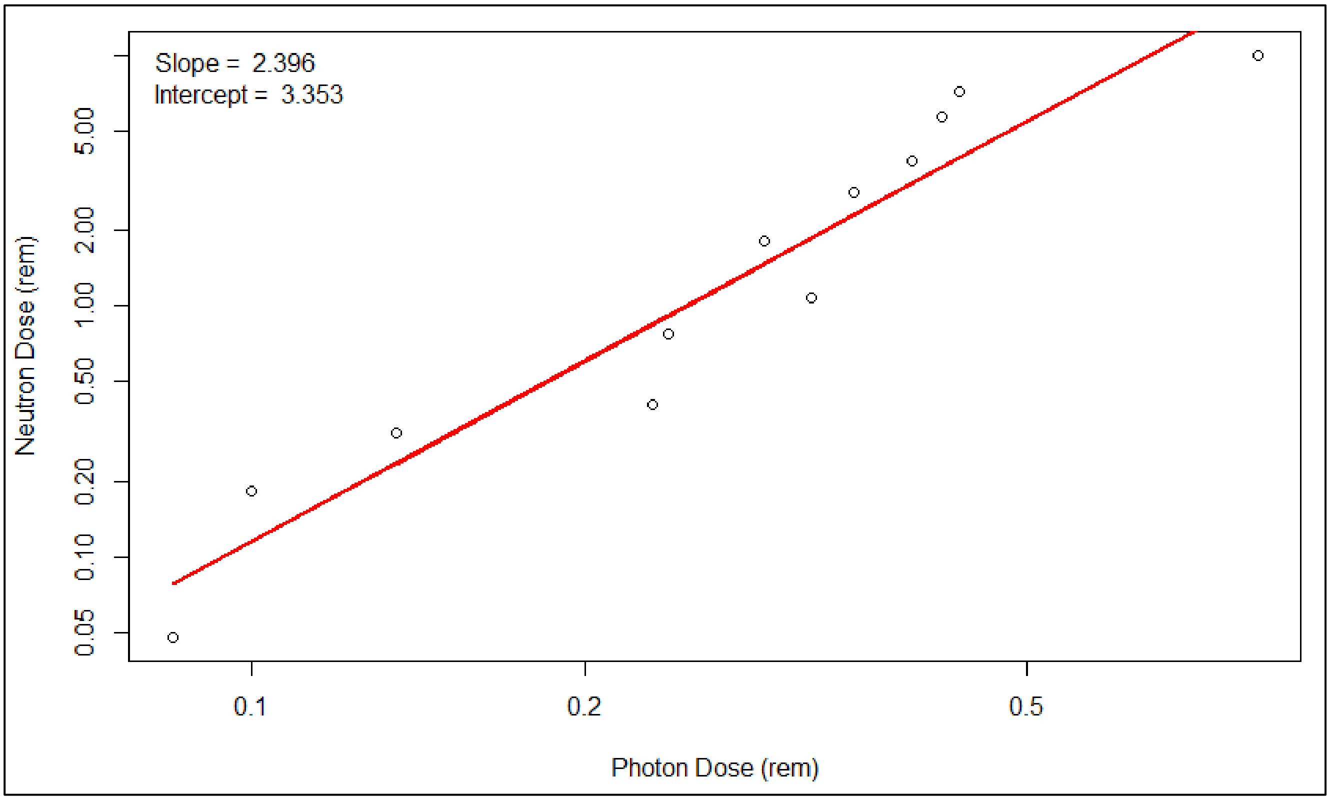

Figure C-2. Mound neutron and photon data from NS in 1954 and 1957 and the associated linear regression fit.

C.3 ROS WITH RATIOS

Taking the data from Figure C-1, calculating the ratio of *Y*/*X* for each of the 12 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table C-2 and the lognormal probability plot in Figure C-3.

Table C-2. Parameter estimates for lognormal ROS with ratios.

Parameter	Estimate
β	(a)
μ	1.498
σ	1.008
GM	4.471
GSD	2.74

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

The fit in Figure C-3 looks okay, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers.

ATTACHMENT C
MOUND NEUTRON SOURCE DATA:
***X* AND *Y* HAVE A RELATIONSHIP AND β > 1 (continued)**

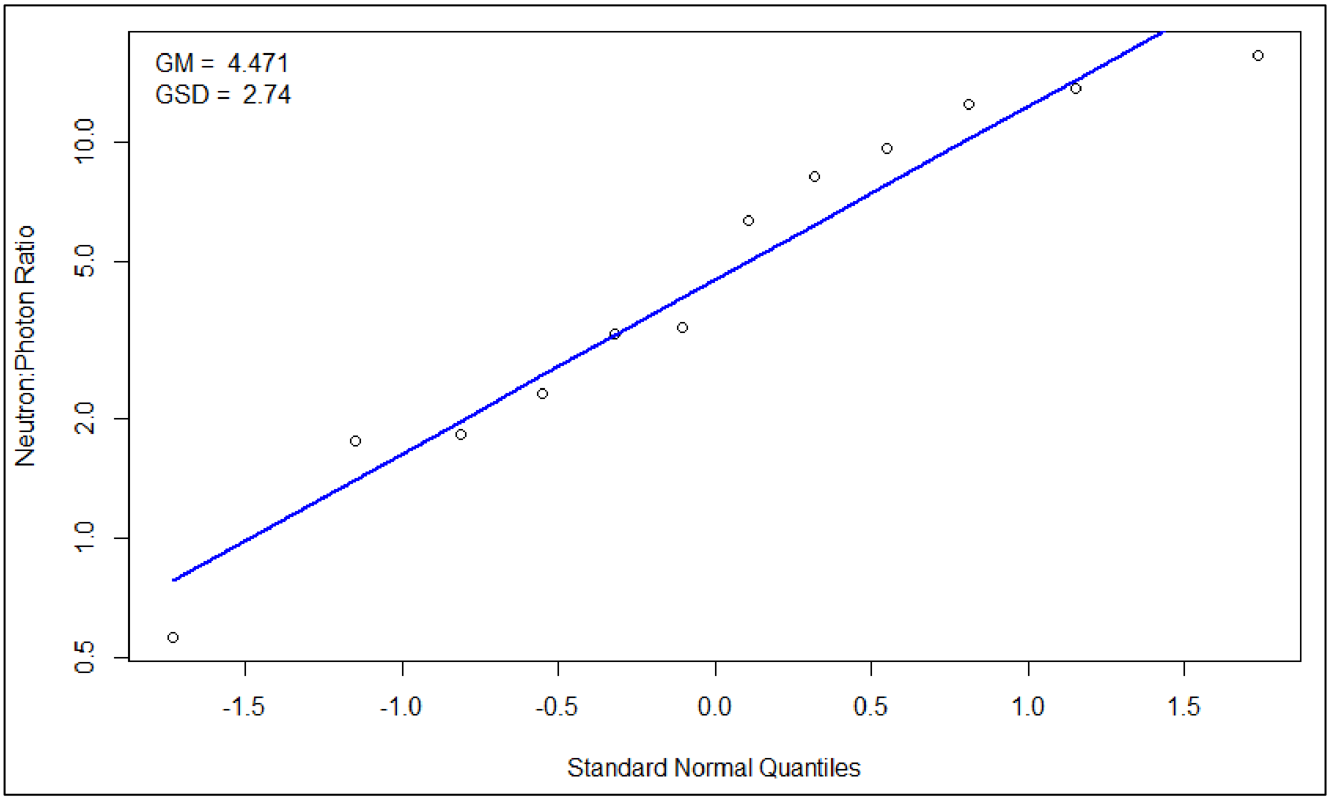

Figure C-3. Mound neutron and photon data from NS in 1954 and 1957 and the associated lognormal ROS with ratios fit.

C.4 COMPARISON OF REGRESSION AND RATIOS

To illustrate the performance of the two methods for this dataset, Figure C-4 adds the line resulting from ROS with the ratios to Figure C-2. The two lines intersect at:

$$X_{\text{int}} = \exp\left(\frac{3.353 - 1.498}{1 - 2.396}\right) \tag{C-1}$$

$$X_{\text{int}} = 0.2647\ \text{rem} \tag{C-2}$$

This means the ratio method will underestimate the best estimate of predicted neutron dose for photon doses less than 0.2647 rem and overestimate the best estimate of predicted neutron dose for photon doses greater than 0.2647 rem.

ATTACHMENT C
MOUND NEUTRON SOURCE DATA:
X AND *Y* HAVE A RELATIONSHIP AND β > 1 (continued)

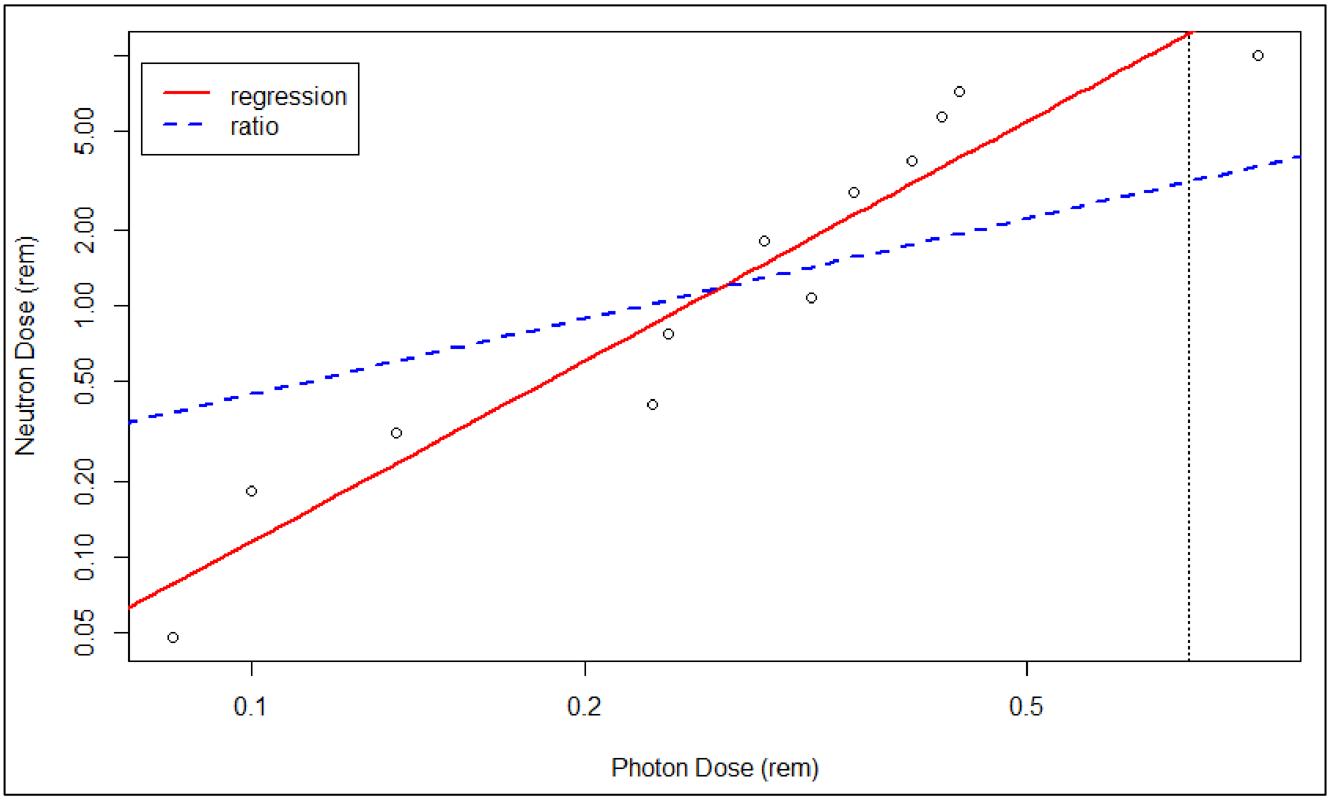

Figure C-4. Mound neutron and photon data from NS in 1954 and 1957 and the associated linear regression and lognormal ROS with ratios fits.

For example, for a photon dose of 0.7 rem (dotted vertical line in Figure C-4), the predicted neutron dose from the regression model is:

$$\hat{Y}_{reg} = \exp\left[3.353 + 2.396 \times \log(0.7)\right] \quad \text{(C-3)}$$

$$\hat{Y}_{reg} = 12.16 \text{ rem} \quad \text{(C-4)}$$

The predicted neutron dose from the ROS with ratios is:

$$\hat{Y}_{rat} = \exp(4.471) \times 0.7 \quad \text{(C-5)}$$

$$\hat{Y}_{rat} = 3.13 \text{ rem} \quad \text{(C-6)}$$

For a photon dose of 0.7 rem, the ROS with ratio model underestimates by a factor of approximately 4.

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0 (continued)

D.1 DATA

Figure D-1 shows a scatterplot of neutron and gamma doses from Rocky Flats in 1969.

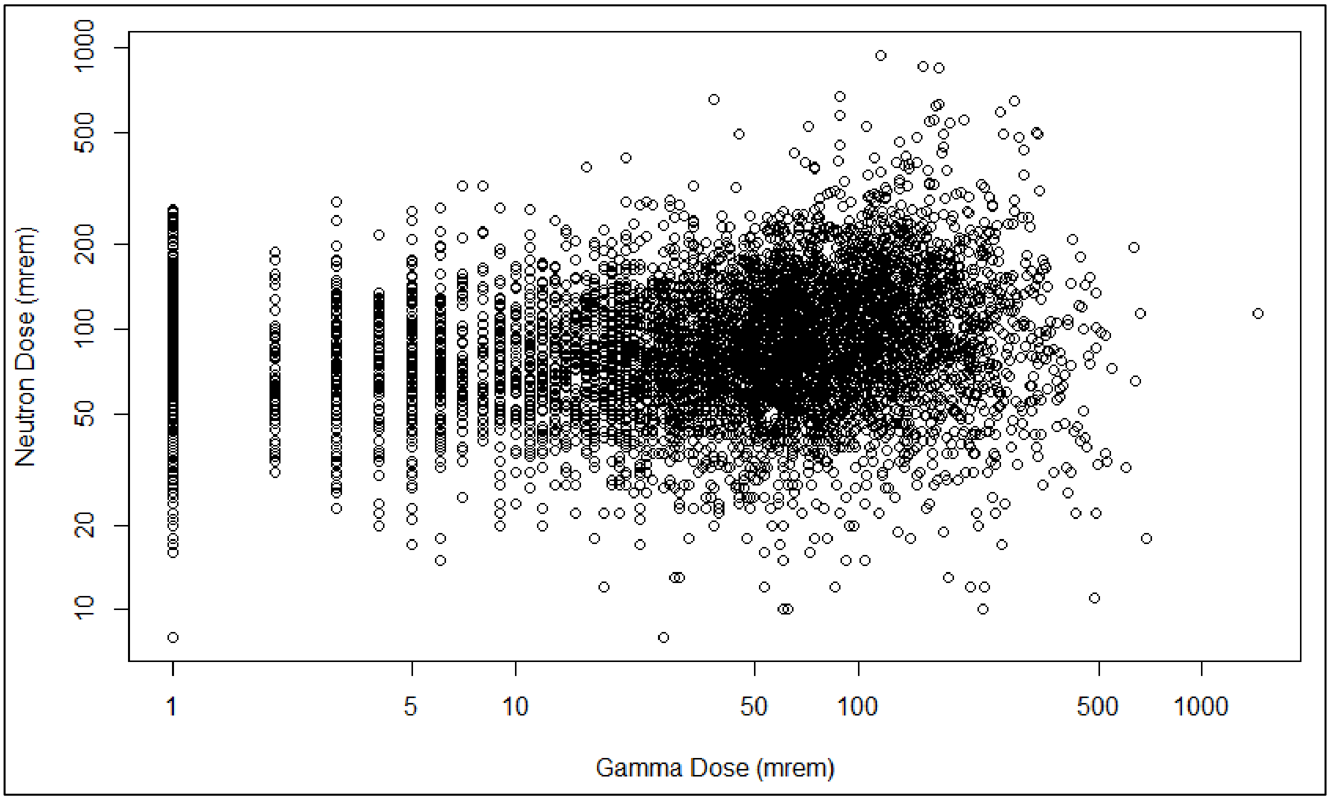

Figure D-1. Rocky Flats neutron and gamma data from 1969.

D.2 LINEAR REGRESSION

Clearly the data in Figure D-1 have no relationship. Essentially, the predicted neutron dose will be the mean of the neutron doses despite the value of *X*, but for the sake of illustration, the parameter estimates are included in Table D-1.

Table D-1. Parameter estimates for regression.

Parameter	Estimate
β	0.06333
μ	4.21
σ	0.5244

As expected, the estimated β parameter is fairly close to zero. Figure D-2 shows the data and regression line on logarithmic scales, so the regression line appears fairly flat.

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0 (continued)

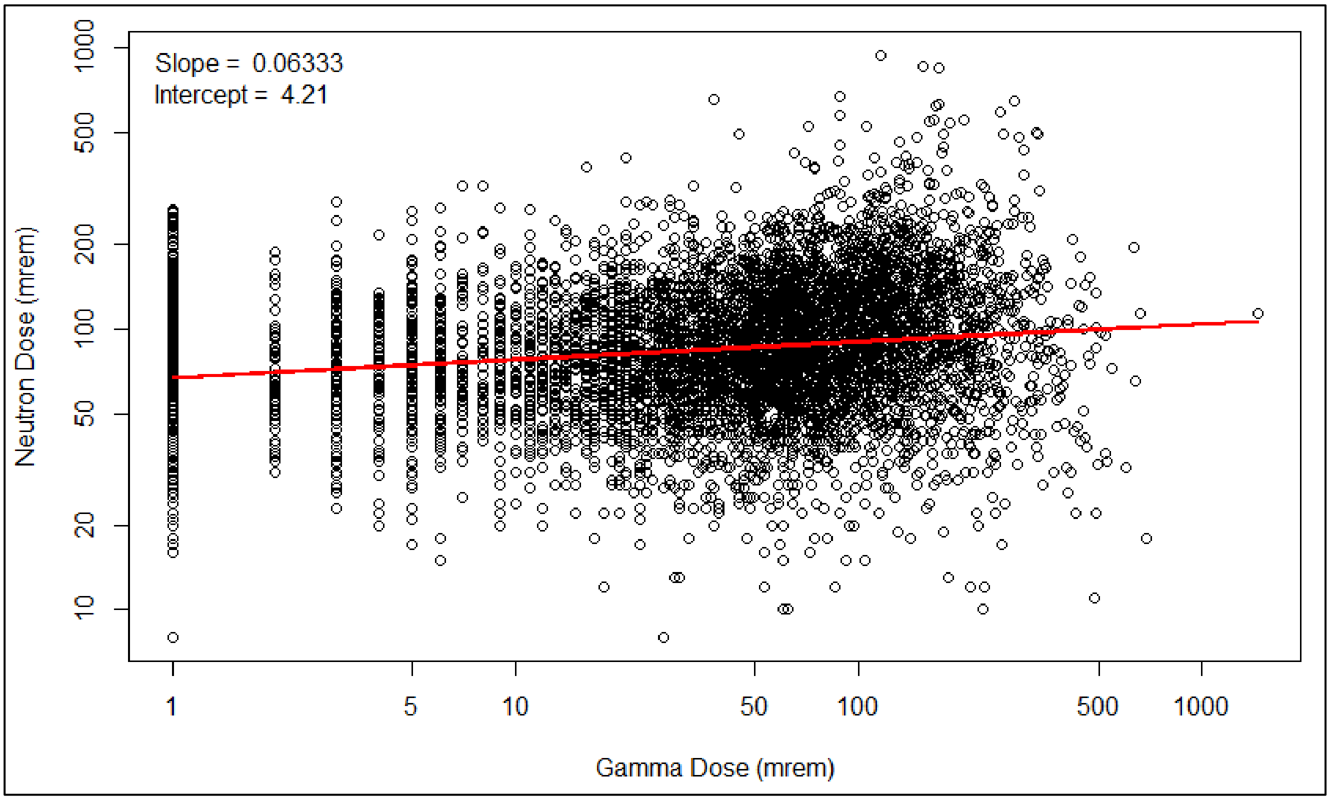

Figure D-2. Rocky Flats neutron and gamma data from 1969 and the associated linear regression fit.

D.3 ROS WITH RATIOS

Taking the data from Figure D-1, calculating the ratio of *Y*/*X* for each of the 6,948 pairs, and modeling those ratios with lognormal ROS results in the estimates in Table D-2 and the lognormal probability plot in Figure D-3.

Table D-2. Parameter estimates for lognormal ROS with ratios.

Parameter	Estimate
β	(a)
μ	1.106
σ	1.556
GM	3.023
GSD	4.737

a. β is assumed to be 1 when using ratios. The loss of information when collapsing bivariate data to univariate ratios will not allow estimation of β.

The fit in Figure D-3 looks okay, and there is no visible indication of the issue with the ROS with ratios giving the incorrect answers. Nothing about the parameters in Table D-2 or the fit in Figure D-3 indicates that *X* and *Y* have no relationship. By collapsing independent gamma and neutron doses

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0 (continued)

into a one-dimensional ratio, two variables that have no relationship now appear to have a meaningful relationship.

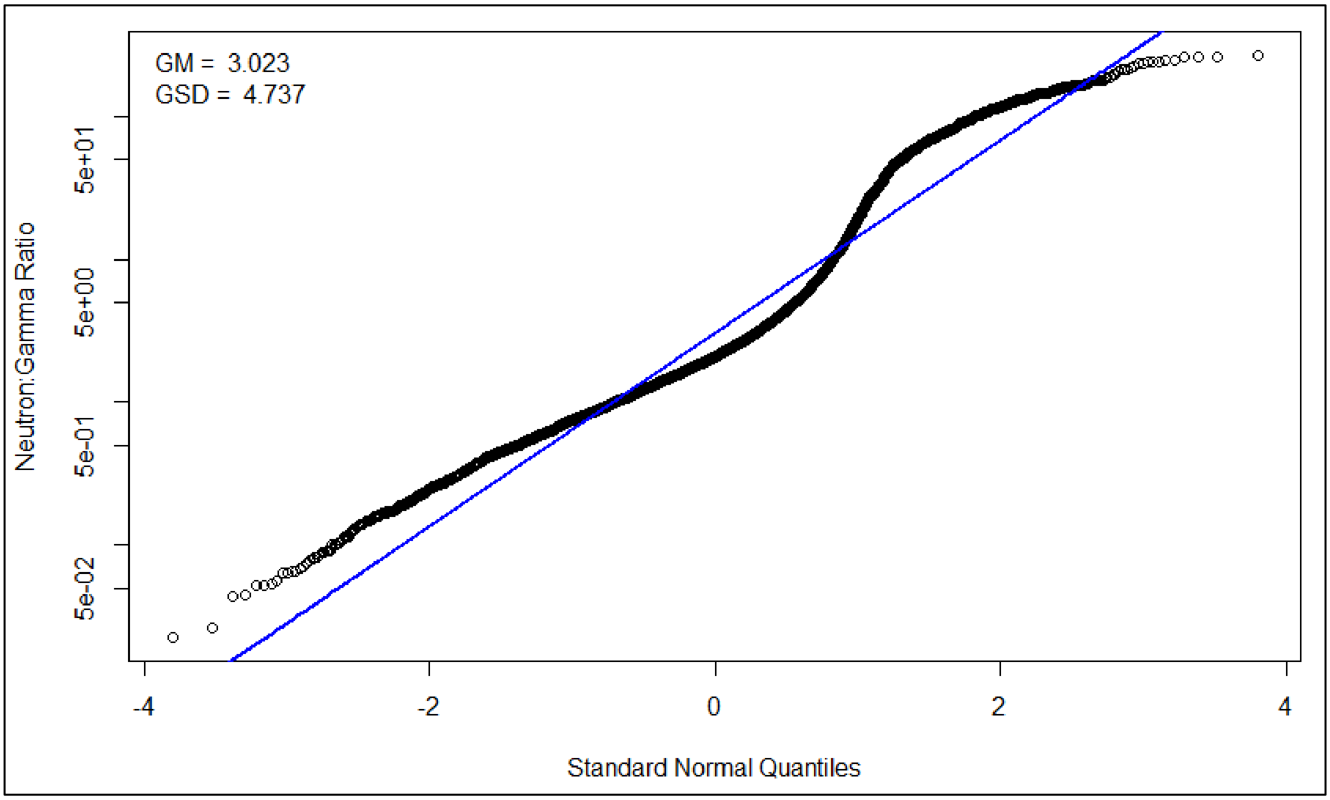

Figure D-3. Rocky Flats neutron and gamma data from 1969 and the associated lognormal ROS with ratios fit.

D.4 COMPARISON OF REGRESSION AND RATIOS

To illustrate the performance of the two methods when gamma and neutron doses have no relationship, Figure D-4 adds the line resulting from ROS with the ratios to Figure D-2. The two lines intersect at:

$$X_{int} = \exp\left(\frac{4.21 - 1.106}{1 - 0.06333}\right) \quad \text{(D-1)}$$

$$X_{int} = 27.49 \text{ mrem} \quad \text{(D-2)}$$

This means the ratio method will underestimate the best estimate of predicted neutron dose for gamma doses less than 27.488 mrem and overestimate the best estimate of predicted neutron dose for gamma doses greater than 27.488 mrem.

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES β = 0 (continued)

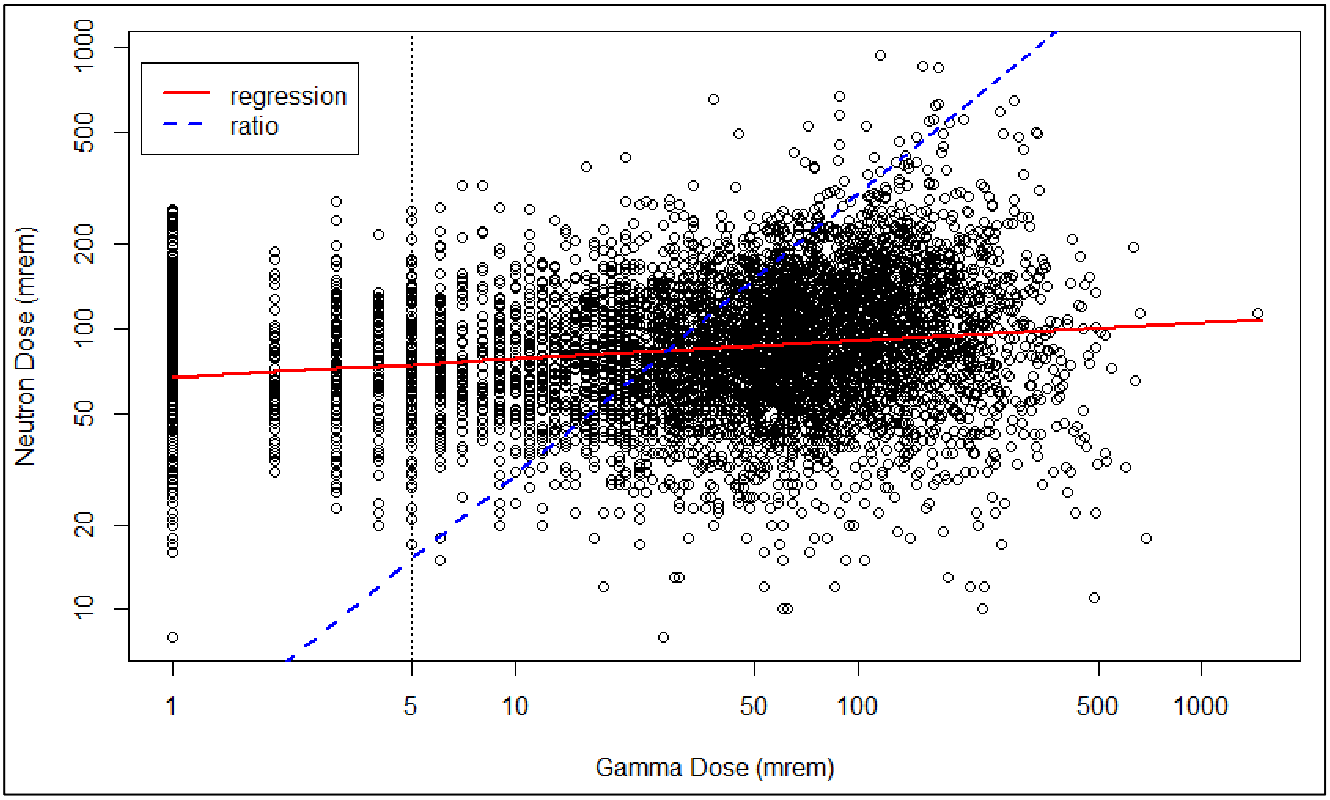

Figure D-4. Rocky Flats neutron and gamma data from 1969 and the associated linear regression and lognormal ROS with ratios fits.

For example, for a gamma dose of 5 mrem (dotted vertical line in Figure D-4), the predicted neutron dose from the regression model is:

$$\hat{Y}_{reg} = \exp\left[4.21 + 0.06333 \times \log(5)\right] \tag{D-3}$$

$$\hat{Y}_{reg} = 74.58 \text{ mrem} \tag{D-4}$$

The predicted neutron dose from the ROS with ratios is:

$$\hat{Y}_{rat} = \exp(1.106) \times 5 \tag{D-5}$$

$$\hat{Y}_{rat} = 15.12 \text{ mrem} \tag{D-6}$$

For a gamma dose of 5 mrem, the ROS with ratio model underestimates by a factor of approximately 5 and is smaller than all of the observed neutron doses with associated gamma doses of 5 mrem. The essentially flat linear regression fit would indicate that the predicted neutron dose does not depend on the observed gamma dose. However, the lognormal fit to the ratio indicates that the predicted neutron dose should change as a function of gamma dose, which is obviously inappropriate based on the points in Figure D-4 and the fact that these Rocky Flats neutron and gamma doses have no relationship.

ATTACHMENT D
ROCKY FLATS NEUTRON AND GAMMA DOSE DATA: *X* AND *Y* HAVE NO RELATIONSHIP, WHICH IMPLIES $\beta = 0$ (continued)

As mentioned in Section 6.3, if there is reason to believe that neutron and gamma doses should have a relationship, exploratory analyses should be done to determine whether relationships exist amongst subgroups of the dataset.