WILEY

Spurious Correlation and the Fallacy of the Ratio Standard Revisited
Author(s): Richard A. Kronmal
Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 156, No. 3 (1993), pp. 379-392
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983064
Accessed: 23-06-2017 21:12 UTC

J. R. Statist. Soc. A (1993)
156, *Part* 3, *pp.* 379–392

Spurious Correlation and the Fallacy of the Ratio Standard Revisited

By RICHARD A. KRONMAL[†]

University of Washington, Seattle, USA, and University of Southampton, UK

[Received February 1992. Final revision December 1992]

SUMMARY

Spurious correlation refers to the correlation between indices that have a common component. A 'per ratio' standard is based on a biological measurement adjusted for some physical measurement by division. Renowned statisticians and biologists (Pearson, Neyman and Tanner) have warned about the problems in interpretation that ratios cause. This warning has been largely ignored. The consequences of using a single ratio as either the dependent or one of the independent variables in a multiple-regression analysis are described. It is shown that the use of ratios in regression analyses can lead to incorrect or misleading inferences. A recommendation is made that the use of ratios in regression analyses be avoided.

Keywords: INTERACTIONS IN REGRESSION; MULTIPLE LINEAR REGRESSION; RATIOS IN REGRESSION; RATIO STANDARDS; SPURIOUS CORRELATION

1. INTRODUCTION

Pearson (1897) unleashed an imp on a group of complete skeletons. The imp randomly mixed the bones into new 'complete' skeletons making sure that each skeleton was complete. An anthropologist, to verify that the bones making up each skeleton were from the same individual, computed the correlation coefficient between the length of various bones from each skeleton divided by the length of the skeleton. This correlation was large and highly statistically significant; on this basis the anthropologist concluded that the bones were from the same individuals.

Much later, a fictitious friend of Neyman (1952), in an empirical attempt to verify the theory that storks bring babies, computed the correlation of the number of storks per 10000 women to the number of babies per 10000 women in a sample of counties. He found a highly statistically significant correlation and cautiously concluded that '. . . although there is no evidence of storks actually bringing babies, there is overwhelming evidence that, by some mysterious process, they influence the birth rate'!

Using the ratio of stroke volume of the heart divided by body weight, Tanner (1949) described a disease of thin people caused by 'statistical artifact'. He warned of the danger of attempting to adjust for a body size variable by dividing the variable of interest by it. It is still common practice to define a diseased state based on similarly constructed ratios.

Widely disseminated by the world press in 1985 were the results reported in an article in the *Lancet* by investigators from Harvard Medical School and the Boston

[†] *Address for correspondence*: Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, USA.

0035–9238/93/156379

University School of Public Health (Woolhandler and Himmelstein, 1985). They computed the regression equation relating the percentage of the gross national product spent on the military to infant mortality rates in the countries of the world and concluded that their findings '... support the hypothesis that arms spending is causally related to infant mortality'.

Ostlund *et al.* (1990) and Freedman *et al.* (1990) describe how the waist-to-hip circumference ratio provides an explanation of why women have levels of high density lipoprotein that are higher than in men. Wingard (1990) speculates that waist-to-hip ratio may '... help explain gender differences in lipoprotein levels as well as coronary heart disease (CHD) risk'.

All these examples are illustrations of the use of ratios, which can lead to what Pearson (1897) called 'spurious correlation'. He used this term to describe the correlation between ratios that exists even if all the component variables of the ratios are uncorrelated. Neyman (1952) gave a formula for evaluating the degree of correlation when the numerator variables are linearly related to the common denominator variable.

These examples are a small sample of the many ratios that are used in various fields. In sociology, where the ratios are usually rates with population size as the denominator, there has been considerable debate about the merits of using ratios, with some papers defending their use (Kasarda and Nolan, 1979; MacMillan and Daft, 1980), some warning of the problem of spurious correlation (Chilton, 1982; Logan, 1982) and still others taking a middle ground (Kuh and Meyer, 1955; Long, 1980; Pendelton, 1984; Pendelton *et al.*, 1979; Schuessler, 1974). These papers have provided confusing and contradictory results. This issue has been clarified in an excellent paper (Firebaugh and Gibbs, 1985) for the case in which all variables appearing in the regression equation are divided by a common variable or one of the independent variables also appears as the denominator of the dependent variable (a ratio).

This paper explores the problems associated with using a ratio as a dependent variable and as independent variables in multiple regression. The results are given for multiple linear regression but the principles are applicable to generalized linear models. The focus throughout is on the partial regression coefficients rather than the correlation coefficient, as that is the most common usage today.

In Section 2 the results for the case in which both the independent and the dependent variables are divided by a common variable are reviewed. Neyman's 'storks and babies' example is used to illustrate the effects of the use of rates on the inferences from these data.

Division of only the dependent variable by an independent variable is shown in Section 3 to result in estimates of the partial regression coefficients of other independent variables that often cause investigators to reach incorrect interpretations about the effects of these variables on the dependent variable. This problem is illustrated with the widely used ratio of forced expiratory volume FEV1 divided by the square of the height (Cole, 1975; Dockery *et al.*, 1985).

Section 4 demonstrates that the use of a ratio as an independent variable provides an inadequate adjustment for each of the variables appearing in the ratio and thus can lead to incorrect inference about the effects of the independent variables on the dependent variable. To illustrate this phenomenon, the ratio of the weight to the height squared, often referred to as body mass index (BMI), is used.

Finally, Section 5 shows that models which include ratios are special cases of a properly specified linear model and that there is little to be gained by using ratios in this setting.

2. DIVISION OF BOTH INDEPENDENT AND DEPENDENT VARIABLES BY COMMON VARIABLE

2.1. *Analytical Results*

Firebaugh and Gibbs (1985) discuss the pros and cons of controlling for a 'dominant confounding variable', often called a deflator, by dividing both the dependent and the independent variable by it. Friedlander (1980) studied this problem in the multiple-regression context in detail. She showed that dividing both the dependent and the independent variables by a common factor results in estimates of the partial regression parameters that take a simple form. These results are reviewed below.

Let Y be an $n \times 1$ vector, Z be an $n \times n$ diagonal matrix and X be an $n \times p$ centred matrix, e.g. $X_{i,j} = x_{i,j} - \bar{x}_j$, where $\bar{x}_j = \Sigma_{i=1}^n x_{i,j}/n$. Assume that the correct model relating Y to X and Z is

$$Y = \mathbf{1}_n\beta_0 + X\beta_X + Z\mathbf{1}_n\beta_Z + \epsilon, \tag{1}$$

where X and Z are fixed quantities with the diagonal elements of $Z \neq 0$, β_0 and β_Z are scalars and β_X is a $p \times 1$ vector, ϵ is independent of X and Z and $E(\epsilon) = 0$. Define $\mathbf{1}_{t\times t}$ to be the $t \times t$ identity matrix and $\mathbf{1}_t$ the $t \times 1$ vector of 1s. The notation $\hat{\beta}_{S.T}$ will be used to represent the least squares estimate of the regression coefficient vector of S on T, where either S or T may be a vector or matrix.

However, if instead of using model (1) both the dependent variable X and the independent variable Y in the model are divided by Z, the resulting model is

$$Z^{-1}Y = \mathbf{1}_n\alpha_0 + Z^{-1}X\alpha_X + \epsilon. \tag{2}$$

Note that neither the constant term nor the error term is divided by Z. We shall return to this omission later.

If we find estimates $\hat{\alpha}_X$ for the coefficients in model (2) by least squares we obtain

$$\hat{\alpha}_X = (X'Z^{-1}MMZ^{-1}X)^{-1}X'Z^{-1}MZ^{-1}Y$$

where $M = I_{n\times n} - \mathbf{1}_n\mathbf{1}_n'/n$. It can be shown that $E(\hat{\alpha}_X) = \beta_X + \hat{\beta}_{Z^{-1}.Z^{-1}X}\beta_0$, where $\hat{\beta}_{Z^{-1}.Z^{-1}X}$ is a $p \times 1$ vector of least squares regression coefficient estimates,

$$\hat{\beta}_{Z^{-1}.Z^{-1}X} = (X'Z^{-1}MMZ^{-1}X)^{-1}X'Z^{-1}MZ^{-1}.$$

$\hat{\alpha}_X$ is a biased estimate of β_X with the bias a function of β_0.

The $E(\hat{\alpha}_X)$ can be expressed in a simple form by making the assumption that each column of X is a random variable that is linearly related to Z and the variance of X is independent of Z, e.g.

$$X_j = \mathbf{1}_n\delta_{0,j} + Z\mathbf{1}_n\delta_{Z,j} + \epsilon_j, \qquad j = 1, 2, \ldots, p.$$

Here, $\delta_{0,j}$ and $\delta_{Z,j}$ are scalars and ϵ_j is an $n \times 1$ random vector with $E(\epsilon_j) = \mathbf{0}_n$. Then, $E(\hat{\alpha}_X) = \beta_X + \Delta\beta_0$, where Δ is a $p \times 1$ vector with the jth element equal to $\delta_{0,j}/\{\delta_{0,j}^2 + \mathrm{var}(\epsilon_j)\}$.

$\hat{\alpha}_X$ is an unbiased estimate of β_X when $\beta_0 = 0$. The jth element of $\hat{\alpha}_X$ will also be an unbiased estimate of the jth element of β_X whenever $\delta_{0,j} = 0$. This is equivalent to the result that Neyman derived for the correlation between Y/Z and X/Z when given that Z, Y and X are independent. When β_0 and $\delta_{0,j}$ are not equal to 0 then the expected value of the jth element of $\hat{\alpha}_X$ will be non-zero even when the jth element of X is unrelated to Y conditional on Z and thus the jth element of β_X is equal to 0. This is an example of spurious correlation.

If we divide equation (1) by Z (multiply both sides of the equation by Z^{-1}), then it becomes obvious why $\hat{\alpha}_X$ is usually a biased estimate of β_X:

$$Z^{-1}Y = Z^{-1}\mathbf{1}_n\beta_0 + Z^{-1}X\beta_X + \mathbf{1}_n\beta_Z + Z^{-1}\epsilon. \tag{3}$$

The least squares estimates of the parameters in equation (3) will be unbiased estimates of the parameters in equation (1) and will be the estimates that would be obtained if equation (1) were assumed and weighted least squares carried out with weights equal to Z_j^{-2}.

Equation (3) is the same as equation (2) when $\beta_0 = 0$. Thus, equation (2) gives estimates of the regression coefficients that are equivalent to those that would be obtained from equation (1) if the intercept were forced to be 0. It is well known that, if β_0 is non-zero, then regression through the origin will result in very biased estimates of the regression coefficients. This has led to the recommendation that the Z^{-1}-term be added to equation (2), thus allowing the ratios to be used (Firebaugh and Gibbs, 1985; Prather, 1976; Friedlander, 1980). Although this solves the problem of possible biases due to the omission of the intercept term in equation (1), the division by Z affects the variance of the error term which can be either beneficial or detrimental depending on the relationship between the error variance and Z. Although this is important both here and in Section 3, it will not be discussed further because it can be dealt with by weighted least squares without the need to use ratios.

2.2. *Neyman's Example*

These results will be illustrated by Neyman's example of storks and babies. This is not done because of any shortage of real life examples, but to make Neyman's data more readily available for teaching and because the example nicely illustrates the potential magnitude of the spurious association.

In Neyman's example both the number of storks and the number of babies are divided by the number of women to form the birth-rate and the stork population rate. This is both the most natural and universally used way to adjust for population size.

Table 1 shows the data collected by Neyman's friend. A plot of the birth-rate *versus* the stork population rate and the least squares regression line is given in Fig. 1. Even without any formal regression analysis, the linearity of the relationship is apparent.

If a least squares regression analysis is done by using equation (2), the result is birth-rate = 3.3 stork-rate + 2.4. The correlation between the birth-rate and stork-rate is 0.63 ($p < 0.00001$). On the basis of this result, Neyman's friend reached the conclusion given earlier that somehow the presence of storks was influencing the birth-rate. He went on to suggest that a randomized clinical trial be instituted to

TABLE 1
Babies and storks by county

County	*Women (× 10000)*	*No. of storks*	*No. of babies*	*Stork-rate (per 10000)*	*Birth-rate (per 10000)*
1	1	2	10	2.00	10.00
2	1	2	15	2.00	15.00
3	1	2	20	2.00	20.00
4	1	3	10	3.00	10.00
5	1	3	15	3.00	15.00
6	1	3	20	3.00	20.00
7	1	4	10	4.00	10.00
8	1	4	15	4.00	15.00
9	1	4	20	4.00	20.00
10	2	4	15	2.00	7.50
11	2	4	20	2.00	10.00
12	2	4	25	2.00	12.50
13	2	5	15	2.50	7.50
14	2	5	20	2.50	10.00
15	2	5	25	2.50	12.50
16	2	6	15	3.00	7.50
17	2	6	20	3.00	10.00
18	2	6	25	3.00	12.50
19	3	5	20	1.67	6.67
20	3	5	25	1.67	8.33
21	3	5	30	1.67	10.00
22	3	6	20	2.00	6.67
23	3	6	25	2.00	8.33
24	3	6	30	2.00	10.00
25	3	7	20	2.33	6.67
26	3	7	25	2.33	8.33
27	3	7	30	2.33	10.00
28	4	6	25	1.50	6.25
29	4	6	30	1.50	7.50
30	4	6	35	1.50	8.75
31	4	7	25	1.75	6.25
32	4	7	30	1.75	7.50
33	4	7	35	1.75	8.75
34	4	8	25	2.00	6.25
35	4	8	30	2.00	7.50
36	4	8	35	2.00	8.75
37	5	7	30	1.40	6.00
38	5	7	35	1.40	7.00
39	5	7	40	1.40	8.00
40	5	8	30	1.60	6.00
41	5	8	35	1.60	7.00
42	5	8	40	1.60	8.00
43	5	9	30	1.80	6.00
44	5	9	35	1.80	7.00
45	5	9	40	1.80	8.00
46	6	8	35	1.33	5.83
47	6	8	40	1.33	6.67
48	6	8	45	1.33	7.50
49	6	9	35	1.50	5.83
50	6	9	40	1.50	6.67
51	6	9	45	1.50	7.50
52	6	10	35	1.67	5.83
53	6	10	40	1.67	6.67
54	6	10	45	1.67	7.50

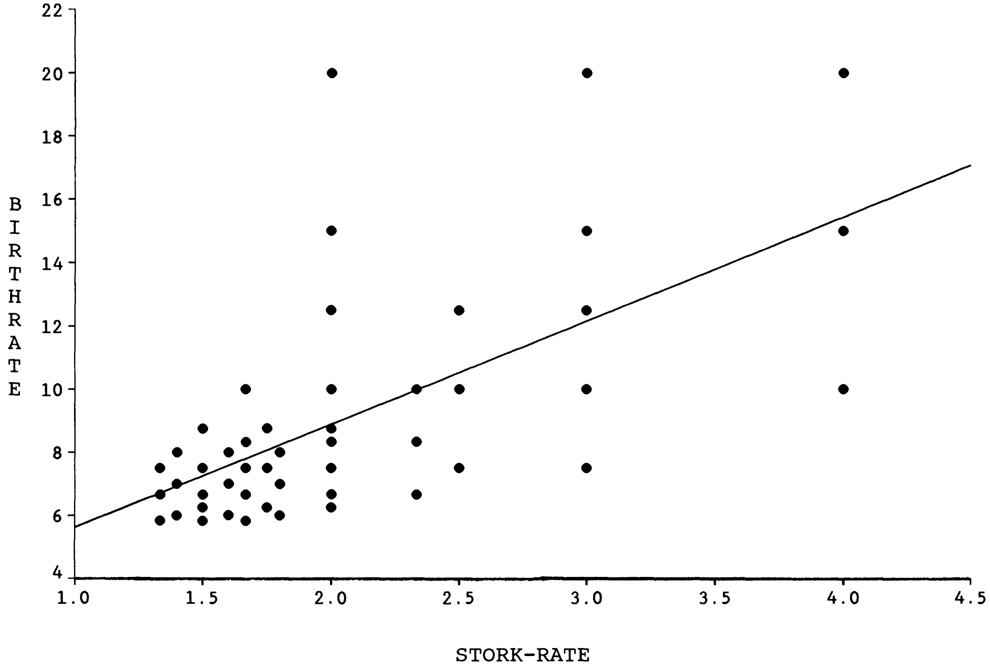

Fig. 1. Birth-rate (per 10000 women) *versus* stork-rate (per 10000 women): data taken from Neyman (1952)

find out whether the removal of storks would be an effective method of birth control.

However, if equation (1) is used, the regression and correlation coefficients are both exactly equal to 0. When equation (3) is used the same result is obtained. Neyman did not compute these estimates of the regression parameter but simply pointed out the obvious independence of the number of storks to the number of babies in counties with the same number of women.

The example given previously of infant mortality rate and military spending (Woolhandler and Himmelstein, 1985) is similar to the storks and babies example. Although infant mortality may be related to military spending, this association is more probably a consequence of the improper adjustment for population size. Even if it were not, it would not have justified the causal interpretation given by the authors of that paper.

3. ONLY DEPENDENT VARIABLE A RATIO

3.1. *Analytical Results*

Regression analysis is often done where only the dependent variable is a ratio. This is usually justified on the grounds of simplicity or because it is felt that the quantity of interest is the ratio. In most cases the purpose of the division is to adjust the numerator variable for the effect of the denominator variable.

Tanner (1949) expressed concern that the construction of 'normal' standards based on such ratios could lead to the classification of some individuals as abnormal simply because of the inadequate adjustment for the denominator variable. In spite of Tanner's warnings, standards based on the division of a measured biological quantity by a variable related to human size, such as height, weight or BMI, are common today. Since these quantities have become established as the 'quantity of interest', investigators often attempt to determine the relationship of other variables to them or to use them as explanatory variables in regression equations.

Again assume that the model given by equation (1) describes the true relationship between the Y and X and Z. Now fit by least squares regression

$$Z^{-1}Y = \mathbf{1}_n\alpha_0 + X\alpha_X + \epsilon. \tag{4}$$

If $\hat{\alpha}_X$ is the least squares estimate of α_X, then

$$E(\hat{\alpha}_X) = \hat{\beta}_{Z^{-1}X.X}\beta_X + \hat{\beta}_{Z^{-1}.X}\beta_0.$$

$\hat{\beta}_{Z^{-1}X.X}$ is a $p \times p$ matrix and $\hat{\beta}_{Z^{-1}.X}$ is a $p \times 1$ vector of least squares partial regression coefficient estimates.

To illustrate the potential difficulties associated with the use of the ratio in equation (4), the estimated regression coefficient is examined when $p = 1$ and the vectors X and Z are linearly related, e.g. $X = \delta_0 + Z\delta_Z + \epsilon_X$, where ϵ_X and Z are independent random vectors and $E(\epsilon_X) = \mathbf{0}_n$ and $\operatorname{var}(X) = \sigma_X^2$. Then

$$E(\hat{\alpha}_X) = E(Z^{-1})(1 - R_{X.Z}^2)\beta_X + \frac{\delta_Z\{1 - E(Z)\,E(Z^{-1})\}}{\sigma_X^2}(\delta_0\beta_X + \beta_0), \tag{5}$$

where $R_{X.Z}^2$ is the square of the correlation coefficient of X with Z.

This equation shows some of the obvious problems associated with the use of the ratio. As was the case in Section 2, the expected value of $\hat{\alpha}_X$ will usually be non-zero even when X is independent of Y given Z. Further, $E(\hat{\alpha}_X)$ is a function of the mean value of Z^{-1}. Thus, if we compare coefficients based on equation (4) from samples from populations with different means for Z^{-1}, these will differ even if the coefficient β_X is the same for these populations. This will often lead to the incorrect conclusion that the relationships between the dependent and independent variables are different in the populations when in fact they may be the same.

Two examples will be given. The first is a fictitious embellishment of Neyman's example.

3.2. *Relationship between Number of Storks and Birth-rate*

An investigator employed by the Zero Population Group (ZPG), who also was a charter member of the Trust for the Preservation of Wildlife Species, was appalled when he saw the results of the analysis of Neyman's friend. Some of his colleagues in the ZPG were considering a campaign to cut the birth-rate by eradicating the world's stork population. However, he felt that the analysis of the stork data was flawed since the division of the number of storks by the number of women in the county had no biological rationale. He felt that the number of storks in a community has little to do with the number of women who live there. Thus he reanalysed the data by using the number of storks as the independent variable. The resulting regression equation is

$$\text{birth-rate} = -1.15(\text{number of storks}) + 16.4.$$

The correlation between the number of storks and the birth-rate is -0.70 ($p < 0.00001$).

He was both relieved and delighted to find that the number of storks was inversely related to the birth-rate. Thus he immediately pointed out to his colleagues that the best way to cut the birth-rate was to *increase* the number of storks. The two fictitious investigators in this example illustrate the kinds of diametrically opposite and incorrect inferences that can be made from the same data set when ratios are used.

This example exemplifies the problem encountered when the dependent variable is a ratio. Even though Y, the numerator of the ratio, is uncorrelated with X, the independent variable, conditional on Z, the ratio is significantly correlated to X through its relationship to Z, the denominator of the ratio.

3.3. *FEV1/Height^2 as Dependent Variable*

Cole (1975), by analytical arguments and through data analysis, and Dockery *et al.* (1985), by using data analysis, recommend the use of FEV1/height^2 as the variable of choice in determining age-specific standards for pulmonary function for males and females for many populations.

Although it is beyond the scope of this paper to examine their arguments in detail, using data from a large study of the elderly, we shall see that the use of this ratio can lead to possibly faulty conclusions that are exactly those expected from the results given in Section 3.1.

The Cardiovascular Health Study (CHS) is a large on-going study of a sample of elderly people from four US communities (Fried *et al.*, 1991). Pulmonary function data were collected from over 5000 of the participants by trained technicians in a standardized fashion by using a computerized spirometry device. The participants ranged from 65 to 99 years of age.

Both Cole (1975) and Dockery *et al.* (1985) fit the equation

$$\text{FEV1} = \text{height}^2(\alpha_0 + \alpha_1\,\text{age}) \qquad (6)$$

relating FEV1 to age, separately for males and females. Both indicate, however, that this is the same as the ratio model, and they are only fitting it in this fashion to allow for comparisons with the linear model in FEV1 and, once fitted, the coefficients will be used to estimate FEV1/height^2. Using equation (6) for the data from the CHS we obtain

$$\text{FEV1/height}^2 = 1.42 - 0.008\,\text{age}$$

for males and

$$\text{FEV1/height}^2 = 1.47 - 0.010\,\text{age}$$

for females. Testing the hypothesis that the age coefficients are equal for males and females gives a statistically significant p-value of 0.02. The coefficients are negative, with that for females larger in absolute value than for males. Cole also finds statistically significant male–female differences in his data sets for the regression coefficients for age, with the females having negative coefficients that are larger in absolute value than those for males. He concluded from this that, as they age, females lose lung capacity at a faster proportional rate than do males.

If, instead of using the ratio, we fit a linear regression model with FEV1 as the dependent variable and age and height squared as the independent variables, we obtain

$$\mathrm{FEV1} = 1.61 - 0.023\,\mathrm{age} + 0.869\,\mathrm{height}^2$$

for males and

$$\mathrm{FEV1} = 1.84 - 0.026\,\mathrm{age} + 0.727\,\mathrm{height}^2$$

for females. The test of the hypothesis of equal partial regression coefficients for age for males and females in these equations is not rejected ($p > 0.65$). Thus the model using the ratio gives rise to a different conclusion about the relationship of lung capacity and age for males and females from that given by the model employing height squared as an independent variable. The reason for this difference is that the coefficient in equation (6) is measuring the joint effect of varying age and height squared (an interaction) whereas the coefficient obtained from the linear regression model given by equation (1) measures the effect of age after adjusting for height squared. Equation (5) predicts that the population with the smaller average height would tend to have the larger (in absolute value) coefficient for age when equation (4) is used as the model, even if the coefficient for age in equation (1) is the same for males and females. Although this does not prove that the model given in equation (4) is incorrect or that there is no gender–age interaction in this or other populations, it is consistent with what we would expect from the results given in Section 3.1 when model (1) is the correct model.

4. ONE INDEPENDENT VARIABLE A RATIO

4.1. *Analytical Results*

Suppose that the true model is

$$Y = \beta_0 + X\beta_X + Z^{-1}\mathbf{1}_n\beta_{Z^{-1}} + W\beta_W + \epsilon. \tag{7}$$

X and Z are defined as before, W is a $p \times 1$ vector and β_W is a scalar parameter. To simplify the notation, X, Z^{-1} and W are centred about their means.

However, instead of fitting model (7) we compute estimates of the regression parameters in the model

$$Y = \mathbf{1}_n\alpha_0 + X\alpha_X + Z^{-1}W\alpha_{Z^{-1}W} + \epsilon. \tag{8}$$

The expected values for the least squares estimates of $\hat{\alpha}_{Z^{-1}W}$ and $\hat{\alpha}_X$ are

$$E(\hat{\alpha}_{Z^{-1}W}) = \hat{\beta}_{X.(Z^{-1}W.X)}\beta_X + \hat{\beta}_{Z^{-1}.(Z^{-1}W.X)}\beta_{Z^{-1}} + \hat{\beta}_{W.(Z^{-1}W.X)}\beta_W$$

and

$$E(\hat{\alpha}_X) = \hat{\beta}_X + (\hat{\beta}_{Z^{-1}.X} - \hat{\beta}_{Z^{-1}W.X}\hat{\beta}_{Z^{-1}.(Z^{-1}W.X)})\beta_{Z^{-1}} + (\hat{\beta}_{W.X} - \hat{\beta}_{Z^{-1}W.X}\hat{\beta}_{W.(Z^{-1}W.X)})\beta_W,$$

where $(Z^{-1}W.X)$ is the residual resulting from fitting by least squares $Z^{-1}W$ with X. This simplifies when only Z^{-1} and W are used to fit Y, to

$$E(\hat{\alpha}_{Z^{-1}W}) = \hat{\beta}_{Z^{-1}.Z^{-1}W}\beta_{Z^{-1}} + \hat{\beta}_{W.Z^{-1}W}\beta_W. \tag{9}$$

Thus the expected value of $\hat{\alpha}_{Z^{-1}W}$ is a linear combination of the coefficients $\beta_{Z^{-1}}$ and β_W. This equation illustrates why interaction terms are usually statistically significant predictors of Y when they are included in a prediction equation that does not include the variables making up the interaction. This is often the case when ratio variables are used and sometimes happens when stepwise regression techniques are used with interaction terms included among the possible predictor variables. Since the regression coefficient for the ratio includes factors measuring the relationship of both components of the ratio on the dependent variable, the ratio alone is often a better predictor than either of the individual variables that make it up. But using the ratio is usually not better than the model that includes both components. It is also possible for both W and Z^{-1} to be statistically significant predictors of Y and yet the product by itself to be not significant.

4.2. *Body Mass Index as Independent Variable*

Probably the most widely used ratio in medical research is the BMI. In most cases it is used in regression equations without including either height or weight in the equation. It will be shown using data from the CHS that this practice can lead to incorrect inferences and a loss of important information about the relative contribution of height and weight to the prediction of the dependent variable.

Suppose that we were interested in exploring the relationship between obesity and waist circumference. The BMI will be used to measure obesity. In this example we shall standardize (subtract the mean and divide by the standard deviation) waist size, 1/height^2, weight and BMI. This will allow us to see better the relative contributions of the variables in the regression analysis. We obtain, for the regression coefficient relating the BMI to waist size, $\hat{\alpha}_{\mathrm{BMI}} = 0.788$ ($p < 0.00001$).

If we use equation (8) to estimate $\hat{\alpha}_{\mathrm{BMI}}$, we obtain

$$\hat{\alpha}_{\mathrm{BMI}} \approx -0.0103\beta_{\mathrm{height}^{-2}} + 0.7482\beta_{\mathrm{weight}}.$$

Since $\beta_{\mathrm{height}^{-2}} = 0.2623$ and $\beta_{\mathrm{weight}} = 0.9515$, we see that height only weakly contributes to the magnitude of the regression coefficient for the BMI, which is almost completely determined by weight.

The regression equation relating weight and height^{-2} to waist circumference is

$$\mathrm{waist} = 0.2623\,\mathrm{height}^{-2} + 0.9515\,\mathrm{weight}.$$

If we add the BMI to this equation, R^2, the coefficient of multiple determination, increases by only 0.002. Finally, the R^2-value for the equation including the BMI alone is 0.621, whereas for the equation using both height^{-2} and weight it is 0.706.

Clearly, we have a much better understanding of the relationship between body size and waist size from the equation using the two variables than when we use the BMI. When the BMI alone is used, we have no idea about how much of the waist size is due to height and how much to weight. As might be expected, it is principally the weight that determines waist size, with the height making a lesser contribution.

The next example shows that a ratio can be judged to be unrelated to the dependent variable even when one of the variables making up the ratio is strongly related to the dependent variable. For this example we use FEV1 and the BMI. If we compute the linear regression equation relating BMI to FEV1, we find that the regression coefficient is not statistically significant ($p > 0.20$) and $R^2 = 0.0001$. Yet in the multiple regression using weight and height as dependent variables height is a highly statistically significant predictor of FEV1 ($p < 0.00001$) and weight is not statistically significant ($p > 0.15$) and $R^2 = 0.30$. Thus had we followed the widely accepted practice of only using the BMI in our equation for FEV1 we would have missed the well-established strong relationship of height to FEV1.

Many researchers either use and/or recommend the use of BMI as a measure of obesity. Two recent papers provide some rationale for this recommendation (Cole, 1991; Gray and Fujioka, 1991). Cole (1991) argued that among the indices proposed it is the best on several grounds. Gray and Fujioka (1991) related the BMI to body fat measured by underwater weighing and total body water. Although these researchers may be right that it is the 'best' single number to measure obesity, it suffers when compared with using height and weight as independent variables, as already shown.

Criqui *et al.* (1982) support their argument that the BMI is the best measure of obesity by comparing it with other proposed measures on the degree of correlation with risk factors for CHD. Using data from the CHS, Table 2 shows the results of comparing the prediction of several risk factors for CHD with BMI as opposed to using height and weight.

Regression parameter estimates were calculated for models with the risk factor as the dependent variable and with height and weight or height^{-2} and weight as the independent variables. In each case, the statistical significance of the BMI was tested after the height and weight variables were entered into the equation. In no instance did the BMI add significantly. More importantly, the relative size and statistical significance of the coefficients vary considerably for the various risk factors. No single function of the height and weight, such as BMI, is likely to capture fully the ways in which height and weight are related to these risk factors. The BMI is an interaction between weight and height^{-2} and can only be adequately interpreted in an equation that includes both of these variables (the main effects). These results show that the use of height and weight is preferable to using the BMI alone.

TABLE 2
Relationship of BMI to risk factors for CHD

Risk factor	*Coefficients for the following variables in the model*†:			R^2	*p-value for BMI*
	Height (m)	*1/height^2*	*Weight (kg)*		
Systolic blood pressure	−43.5		0.11	0.015	0.43
(mm Hg)		90.8	0.10	0.014	0.96
Diastolic blood pressure	2.8‡		0.09	0.022	0.38
(mm Hg)		−6.4‡	0.09	0.022	0.32
Cholesterol (mg/dl^3)	−36.9		0.10	0.086	0.59
		73.2	0.09	0.086	0.28
Log(triglyceride)	−0.73		0.009	0.063	0.85
(mg/dl^3)		1.53	0.009	0.062	0.44
Glucose (mg/dl^3)	−43.7		0.63	0.056	0.21
		93.0	0.63	0.056	0.10

†All regression equations included gender coded as a 0–1 variable.

‡Not statistically significantly different from 0 at the $\alpha = 0.05$ level. All other coefficients were significant at the $\alpha < 0.05$ level.

5. FULL LINEAR MODEL INCLUDING INTERACTIONS AS ALTERNATIVE TO USE OF RATIOS

In each of the cases where ratios are used it is clear that the models could have been written in such a way that the ratios were part of the full linear regression model with interactions. In the case of division of both the dependent and the independent variables by a common variable, this required only the addition of Z^{-1} to make the model a full linear model including an intercept term; for example, when equation (1) is the correct model,

$$Y = \mathbf{1}_n\beta_0 X\beta_X + Z\mathbf{1}_n\beta_Z + \epsilon,$$

and multiplying both sides of the equation by Z^{-1}

$$Z^{-1}Y = Z^{-1}\mathbf{1}_n\beta_0 + Z^{-1}X\beta_X + \beta_Z + Z^{-1}\epsilon.$$

Now β_0 is the partial regression coefficient for Z^{-1} and β_Z is the intercept. Least squares regression using this equation will give unbiased estimates of the coefficients in model (1). As shown earlier, when the dependent variable Y is divided by Z, this affects the error distribution and variance, which can have unpredictable effects on the estimated regression coefficients. Whether this is desirable or not depends on the distribution of the error term and the Z-variable. The decision to use these ratios, even with the Z^{-1}-variable included, requires a careful consideration of the distribution of the dependent variable and Z.

When only the dependent variable is a ratio, model (4) is equivalent to a linear model with the intercept term and the first-order terms involving the other independent variables left out of the equation and includes regression terms for Z and for the interaction of Z with the other dependent variables.

This can be seen by rewriting model (4) as

$$Y = Z\mathbf{1}_n\alpha_0 + ZX\alpha_X + Z^{-1}\epsilon.$$

This model is a subset of the linear model that includes the missing first-order terms β_0 and β_X:

$$Y = \beta_0 + \beta_X X + Z\mathbf{1}_n\alpha_0 + ZX\alpha_X + Z^{-1}\epsilon.$$

If β_0 and β_X are near 0 it may be acceptable to eliminate them from the equation. This should only be done after determining their value and not by assuming in advance that they are 0.

Finally, for the model where only an independent variable is a ratio, it is clear that the ratio is an interaction and could be included along with its constituent variables as shown in the regression equation

$$Y = \beta_0 + X\beta_X + Z^{-1}\mathbf{1}_n\beta_{Z-1} + W\beta_W + Z^{-1}W\beta_{Z^{-1}W} + \epsilon.$$

If the ratio adds to the prediction (e.g. $\beta_{Z^{-1}W}$ is non-zero) then there is nothing wrong with including it. It is not good statistical practice to include interactions in an equation without first including the variables that comprise it as first-order terms in the model.

6. CONCLUSIONS AND RECOMMENDATIONS

The message of this paper is that ratio variables should only be used in the context of a full linear model in which the variables that make up the ratio are included and the intercept term is also present. The common practice of using ratios for either the dependent or the independent variable in regression analyses can lead to misleading inferences and rarely results in any gain. This practice is widespread and entrenched, however, and it may be difficult to convince some researchers that they should give up their most prized ratio or index.

The statistician who agrees with the recommendations of this paper will often be forced to answer arguments put forward by these researchers. Given below are some of these arguments, and suggested responses to them.

(a) 'Everyone else uses this ratio. If we don't, we won't be able to compare our results with those of others.' Although this argument sounds like what one hears from a teenager to justify particularly noxious behaviour, it is equally difficult to counter in this context. The reviewers of the investigator's paper may have already published articles using the ratio and will sometimes insist that this is the only correct way to present the data. Thus the burden falls on the investigator and the statistician to explain why the presentation of the results without the ratio is preferable. The point can also be made that comparisons with the results reported in other papers would be difficult, precisely because of the use of the ratio. As shown here, the results of regression analysis using ratios are not readily comparable across studies, because of possible distribution differences of the Z-variable. It is hoped that reference to this paper will provide support for these arguments.

(b) 'The ratio may provide a better model.' This may be true in some instances. It requires justification, which is best accomplished by first fitting the usual linear model without the ratio, followed by a careful comparison of the results when the ratio is used instead. The ratio model will rarely prove to be superior.

(c) 'Ratio models are simpler.' This is the argument made by Cole (1975) for the ratio of FEV1 and forced vital capacity with height. Although it may be true that the reduction by one of the number of parameters results in a simplification, it is certainly not a major or important argument. Considering the possibility of misleading inferences caused by using the ratio, this slight simplification is not worth the potential cost.

(d) 'The ratio is the "natural" quantity of interest.' In the case of population rates, this may be true. Even here, however, the division by population size is to remove its effect from the numerator variable. Whether this is the optimal way to accomplish this is unclear. Even when such rates are used, there is no reason not to include the reciprocal of the population size as a covariate. For other ratios, the purpose of the denominator is usually to adjust for it. In these instances, there is little to commend the use of this method of adjustment.

Neyman (1979) said, 'Spurious correlations have been ruining empirical statistical research from times immemorial'. In my opinion, the continued widespread use of ratios makes Neyman's sentiments still true today.

ACKNOWLEDGEMENT

This paper is dedicated to the memory of Donovan Thompson, PhD (1919–91), esteemed mentor and colleague.

REFERENCES

Chilton, R. (1982) Analyzing urban crime data: deterrence and the limitations of arrests per offense ratios. *Criminology*, **19**, 590-607.

Cole, T. J. (1975) Linear and proportional regression models in the prediction of ventilatory function. *J. R. Statist. Soc.* A, **138**, 297-324.

——— (1991) Weight-stature indices to measure underweight, overweight, and obesity. In *Anthropometric Assessment of Nutritional Status*, pp. 83-111. New York: Wiley-Liss.

Criqui, M. H., Klauber, M. R. *et al.* (1982) Adjustment for obesity in studies of cardiovascular disease. *Am. J. Epidem.*, **116**, 685-691.

Dockery, D. W., Ware, J. H., Ferris, B. G., Glicksberg, D. S., Fay, M. E., Spiro, A. and Speizer, F. E. (1985) Distribution of forced expiratory volume in one second and forced vital capacity in healthy, white, adult never-smokers in six U.S. cities. *Am. Rev. Resp. Dis.*, **131**, 511-520.

Firebaugh, G. and Gibbs, J. (1985) User's guide to ratio variables. *Am. Sociol. Rev.*, **50**, 713-722.

Freedman, D. S., Jacobsen, S. J. *et al.* (1990) Body fat distribution and male/female differences in lipids and lipoproteins. *Circulation*, **6**, 1498-1506.

Fried, L. P., Borhani, N. D. *et al.* (1991) The Cardiovascular Health Study: design and rationale. *Ann. Epidem.*, **1**, 263-276.

Friedlander, L. J. (1980) A study of the correlation between ratio variables. *PhD Dissertation*. University of Washington, Seattle.

Gray, D. S. and Fujioka, K. (1991) Use of relative weight and body mass index for the determination of adiposity. *J. Clin. Epidem.*, **44**, 545-550.

Kasandra, J. D. and Nolan, P. D. (1979) Ratio measurement and theoretical inference in social research. *Socl Forces*, **52**, 108-121.

Kuh, E. and Mayer, J. R. (1955) Correlation and regression estimates when the data are ratios. *Econometrica*, **23**, 400-416.

Logan, C. H. (1982) Problems in ratio correlation: the case of deterrence research. *Socl Forces*, **60**, 791-810.

Long, S. B. (1980) The continuing debate over the use of ratio variables: facts and fiction. In *Sociological Methodology* (ed. K. F. Schuessler). San Francisco: Jossey-Bass.

MacMillan, A. and Daft, R. L. (1980) Relationships among ratio variables with common components: fact or artifact. *Socl Forces*, **58**, 1109-1128.

Neyman, J. (1952) *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd edn, pp. 143-154. Washington DC: US Department of Agriculture.

——— (1979) Human cancer: radiation and chemicals, complete. *Science*, **205**, 259-260.

Ostlund, R. E., Staten, M., Kohrt, W. M., Schulz, J. and Malley, M. (1990) The ratio of waist-to-hip circumference, plasma insulin level, and glucose intolerance as independent predictors of the HDL2 cholesterol level in older adults. *New Engl. J. Med.*, **322**, 229-234.

Pearson, K. (1897) Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc. R. Soc. Lond.*, **60**, 489-497.

Pendelton, B. F. (1984) Correction for ratio variable correlation: examples using models of mortality. *Socl Sci. Res.*, **13**, 268-286.

Pendelton, B. F., Warren, R. D. and Chang, H. C. (1979) Correlated denominators in multiple regression and change analyses. *Sociol. Meth. Res.*, **7**, 451-474.

Prather, J. E. (1976) Spurious correlation and social science research: the effect of using ratio variables. *Proc. Am. Statist. Ass. Socl Statist. Sect.*, 684-688.

Schuessler, K. F. (1974) Analysis of ratio variables: opportunities and pitfalls. *Am. J. Sociol.*, **80**, 379-396.

Tanner, J. M. (1949) Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J. Appl. Physiol.*, **2**, 1-15.

Wingard, D. L. (1990) Sex differences and coronary heart disease: a case of comparing apples to pears? *Circulation*, **6**, 1710-1712.

Woolhandler, S. and Himmelstein, D. U. (1985) Militarism and mortality—an international analysis of arms spending and infant death rates. *Lancet*, 1375-1378.