AMERICAN
Journal of Epidemiology
Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 104 NOVEMBER, 1976 NO. 5

Reviews and Commentary

THE EFFECT OF REGRESSION TO THE MEAN IN EPIDEMIOLOGIC AND CLINICAL STUDIES

C. E. DAVIS

INTRODUCTION

Regression to the mean is the phrase used to identify the phenomenon that a variable that is extreme on its first measurement will tend to be closer to the center of the distribution for a later measurement. In studies based on biological measurements, this variability can be attributed to both the inherent variation in the phenomenon being measured and the variability of the measurement itself. The concept of regression to the mean is an important consideration in studies where subjects are chosen because of a biological variable above or below a specified level. For example, in a trial of an anti-hypertensive drug, subjects with elevated blood pressure would be chosen, and, if regression to the mean is not properly accounted for, the results of the study may be biased.

At least three types of studies are affected by this potential problem. One type is a survey in which subjects are selected for subsequent screening or follow-up based on an initial extreme value. An example of such a survey is a portion of the Lipid Research Clinics (LRC) Prevalence Study (1). In this study, well-defined populations are screened for cholesterol and triglycerides. A random sample of 15 per cent of those screened are asked to return for more extensive clinical evaluations. In addition, subjects with elevated cholesterol and/or triglycerides are also asked to return for the additional evaluation. In this survey, it can be expected that subjects recalled for additional lipid determinations because of elevated lipids will, on the average, have lower lipid values at the second screen. This is not true for the randomly selected subjects. Since the random sample covers the full range of lipid determinations, the mean lipid levels at the two screens for this group will be identical; i.e., since these subjects are not chosen based on extreme values, regression to the mean will not be a factor. It will thus be imperative that in analyses of the results of this survey which compare lipid distributions of the randomly chosen group with those for the group chosen for elevated lipids, the effect of regression on the lipid values of the latter be taken into account.

A second type of study which should take account of effects of regression to the mean is a study in which no control group is formed as a basis for comparison. As an

Abbreviations: LDL, low-density lipoprotein; LRC, Lipid Research Clinics.

From the Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514.

This work was supported by US National Heart and Lung Institute Contract NIH-NHLI-71-2243 from the National Institutes of Health.

example, suppose a group of subjects is chosen because the subjects have elevated blood pressure, and treatment with a drug is initiated. If a control group which is given placebo is included in the study, the effect of regression can be estimated from this group. However, even when no such group is available, procedures for estimating the effect of regression exist (2), as outlined below.

Regression can affect controlled clinical studies in two ways. One problem is that in many studies subjects are identified in two stages. At the first screen, subjects with extreme values are identified and asked to return for a second evaluation. At this second evaluation they are also required to have an extreme observation before they are included in the study. An example of such a study is the Hypertension Detection and Follow-Up Study (3). Regression affects such studies since the second measure will tend to be lower; hence, the subject may be excluded if required to retain the same elevated level. The second possible problem associated with controlled studies is the choice of a baseline from which to measure the effect of treatment.

As noted above, the control group can be used to adjust for regression effects. In addition, the study can be designed so that the effect of regression can be reduced. Ederer (4) suggests taking two pretreatment assignment measures, using the first to classify a subject and the second to measure baseline. Gardner and Heady (5) suggest using the average of two or more measures for classification purposes.

In the remainder of this paper, various methods of accounting for regression to the mean will be described and illustrated. After the introduction of some statistical notation, two methods of adjusting for the effect of regression will be discussed. The final section of the paper will compare two methods of designing a study in such a manner as to reduce the impact of regression. In what follows we will limit the discussion to high extreme values. The arguments are, of course, equally applicable to low extreme values.

The statistical model

Let y_i, $i = 1, 2, 3$ be the ith measure of the variable of interest and k_i be the cutpoint. We assume that y_i is normally (Gaussian) distributed with mean μ and standard deviation σ. Let ρ_{ij} be the correlation between the ith and jth measure. Finally, let $z_i = (y_i - \mu)/\sigma$ and $a_i = (k_i - \mu)/\sigma$ be the standardized variables and cutpoints, respectively. It then follows that (see, for example, Tallis (6)) the mean of the distribution of observations chosen because they exceed k_1, is

$$E(y_1 | y_1 > k_1) = \mu + c_1 \sigma, \qquad (1)$$

where

$$c_1 = \phi(a_1)/[1 - \Phi(a_1)],$$
$$\phi(a_1) = \exp(-\tfrac{1}{2} a_1^2)/\sqrt{2\pi}$$

and

$$\Phi(a_1) = \int_{-\infty}^{a_1} \phi(x)\, dx.$$

Similarly, the expected value of the distribution of y_2, the second measure, given that y_1 exceeds k_1 is

$$E(y_2 | y_1 > k_1) = \mu + \rho_{12} c_1 \sigma. \qquad (2)$$

It follows, then, that the effect of regression will be the difference

$$E(y_1 | y_1 > k_1) - E(y_2 | y_1 > k_1) = c_1 \sigma (1 - \rho_{12}). \qquad (3)$$

Examination of this equation indicates that if $\rho_{12} = 1$, there is no regression to the mean effect. As ρ_{12} becomes smaller in absolute value the effect of regression becomes larger in absolute value. Finally, we note that if subjects are chosen without regard to the extremity of the variable of interest, this is identical to choosing $k_1 = -\infty$, in which case $c_1 = 0$, and there is no regression to the mean effect.

Estimating the effect of regression

In this section, two alternatives for estimating the effect of regression are considered. Both of these methods are applicable in uncontrolled studies, but the example used will demonstrate the utility in the conduct of trials which employ more than one measurement to classify a subject.

The first method which can be used to estimate the effect of regression to the mean is to obtain external estimates of the parameters μ, σ and ρ_{12}. Estimates of these parameters in combination with the cutpoint k_1 can be used through formula 3 to estimate the expected effect of regression. Suppose we have estimates from another related study which we denote by $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\rho}_{12}$. From formula 3 we estimate the effect of regression as $c_1\hat{\sigma}\ (1 - \hat{\rho}_{12})$.

As an example of the use of this procedure we will consider an example using cholesterol. In addition to the prevalence surveys mentioned previously, the Lipid Research Clinics (7) are conducting a trial to determine if lowering cholesterol will lower the incidence of coronary disease. We shall refer to this study as the LRC Prevention Trial. In this study, cholesterol eligible men aged 35–59 years are identified in two stages. At a preliminary screening the potential subject must have a cholesterol greater than or equal to 265 mg%. If the subject qualifies by this criterion, he is seen in the clinic, at which time his LDL cholesterol must be greater than or equal to 190 mg% if he is to be considered further for the study. Shortly after the study began, it was observed that a sizeable proportion of the subjects who met the first criterion did not meet the second, and indeed there was a considerable drop in cholesterol between the two measurements. Two explanations were advanced for this change: regression to the mean and a change in dietary pattern. It is readily apparent that regression to the mean contributes to loss of subjects at the second screen, but the effect of a change in dietary pattern and its relation to regression is not clear. Data from the LRC Prevalence survey was used to obtain estimates of the relevant parameters. In obtaining the estimates, log cholesterol was used; i.e., cholesterol was assumed to be log normally distributed. From the data on the first 1346 men 35–59 years of age screened in the LRC Prevalence Study, the following parameter estimates were obtained: $\hat{\mu} = 5.331$; $\hat{\sigma} = 0.170$; and $\hat{\rho}_{12} = 0.82$. In formula 1 we have then $k_1 = \log 265 = 5.580$ and hence $a_1 = 1.46$, $\Phi(a_1) = 0.137$ and $\Phi(a_1) = 0.928$. Since $\Phi(a_1) = 0.928$, one would expect approximately 7 percent of the population of men aged 35–59 to exceed 265 mg%. Applying these values gives the results of table 1. It is apparent that regression to the mean accounts for the change in cholesterol between the two screens. Indeed, there was less regression observed than this procedure predicted.

Table 1

Expected and observed effect of regression to the mean in the LRC Prevention Trial using LRC prevalence estimates

Screening value	Expected mean	Observed mean
First screen (y_1)	5.656	5.676
Second screen (y_2)	5.597	5.638
Difference (y_1-y_2)	0.059	0.038

Often in studies such as the example noted above, external estimates of the parameters will not be available. In such cases, James (2) has described a method of obtaining estimates based on the data at hand. In order to apply this procedure one calculates from the data $\bar{y}_1$, s_1^2, s_2^2 and $b_{2.1}$, where s_1^2 and s_2^2 are the sample variances of y_1, and y_2, respectively, and $b_{2.1}$ is the regression coefficient for the regression of y_2 on y_1. From these sample values compute:

$$\hat{\mu} = \bar{y}_1 - c_1\hat{\sigma} \qquad (4)$$

$$\hat{\sigma}^2 = s_1^2/[c_1(k_1 - c_1) + 1], \qquad (5)$$

and

$$\hat{\rho}_{12} = [b_{2.1}^2[c_1(k_1 - c_1) + 1] - s_2^2/\hat{\sigma}^2 + 1]^{1/2}. \quad (6)$$

Since this formulation requires knowledge of c_1, it must in turn require assumptions concerning the proportion of the distribution selected. For example, in the cholesterol example given above, the cutpoint 265 mg% was chosen as the 95th percentile; from tables of the normal distribution one can then obtain $c_1 = 2.06$. Using the data from the first 3484 men screened in the LRC Prevention Trial the following estimates were obtained: $\bar{y}_1 = 5.676$; $s_1^2 = 0.00728$; $s_2^2 = 0.01348$; and $b_{1.2} = 0.792$. Plugging these into formulas 4, 5 and 6 we obtain: $\hat{\mu} = 5.207$; $\hat{\sigma} = 0.228$; and $\hat{\rho}_{12} = 0.91$. Table 2 gives the results of applying these estimates in formulas 1, 2 and 3. In view of the observed and expected change in cholesterol, it is apparent that this method also indicates that the change in cholesterol between the two visits is attributable to regression to the mean.

The choice of a baseline value

In many clinical studies the object of treatment is to effect a change in an important variable. As noted above, if a single observation is used to classify a subject as having a disorder and is also used as a baseline from which to measure the effect of treatment, the treatment effect will be over-estimated since regression will occur. Two proposals have been made for reducing the effect of regression.

Table 2

Expected and observed effect of regression to the mean in the LRC Prevention Trial using estimates of James (1973) (2)

Screening value	Expected mean	Observed mean
First screen (y_1)	5.676*	5.676*
Second screen (y_2)	5.634	5.638
Difference (y_1–y_2)	0.042	0.038

* Since $\bar{y}_1$ is used in the computations of $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\rho}_{12}$, the observed and expected means are identical.

Unfortunately, neither can eliminate the effect except in special cases.

The first method we shall investigate is to take the average of several values as a classification and baseline measure. This method is intuitively appealing since it reduces the variability in the criterion measurement and thus should reduce the effect of regression in subsequent measures. In order to investigate the properties of this method we introduce new notations. Let the ith observation on an individual be denoted by

$$y_i = Y + e_i;\ i = 1, 2, \cdots,$$

where Y is the individual's "true" value and e_i is the error in measuring Y. We shall assume that Y is normally distributed with mean μ and variance δ^2, while e_i is normally distributed with mean 0 and variance γ^2. Moreover, assume that e_i is independent of Y and of e_j, where e_j is the error associated with another observation. This model has been described by Gardner and Heady (5). However, they did not consider methods of reducing the effect of regression on subsequent measures. From the above assumptions, it follows that the parameters used in our previous model can be written as

$$\sigma^2 = \delta^2 + \gamma^2$$

and

$$\rho_{12} = \frac{\delta^2}{\delta^2 + \gamma^2}.$$

It also follows that, if $\bar{y}$ is the mean of several observations, $\bar{y}$ is normally distributed with mean μ and variance $\delta^2 + \gamma^2/n$. Thus, from formula 1 we can write

$$E(\bar{y} \mid \bar{y} > k_1) = \mu + c_3\sqrt{\delta^2 + \gamma^2/n},$$

where

$$c_3 = \phi\left(\frac{k_1 - \mu}{\sqrt{\delta^2 + \gamma^2/n}}\right) \Big/ \left[1 - \Phi\left(\frac{k_1 - \mu}{\sqrt{\delta^2 + \gamma^2/n}}\right)\right].$$

It can be shown that the correlation between $\bar{y}$ and a subsequent measure y^* is

$$\rho^* = \frac{\delta^2}{\sqrt{\delta^2 + \gamma^2/n)\ (\delta^2 + \gamma^2)}}.$$

Hence

$$E(y^* | \bar{y} > k) = \mu + c_3\rho^*\sqrt{\delta^2 + \gamma^2}$$
$$= \mu + c_3\delta^2/\sqrt{\delta^2 + \gamma^2/n}$$

and

$$E(\bar{y} - y^* | \bar{y} > k)$$
$$= c_3\left\{\sqrt{\delta^2 + \frac{\gamma^2}{n}} - \delta^2/\sqrt{\delta^2 + \frac{\gamma^2}{n}}\right\}$$
$$= c_3\gamma^2/n\sqrt{\delta^2 + \frac{\gamma^2}{n}}.$$

Note that as n becomes large, the effect of regression becomes small; i.e.,

$$\lim_{n \to \infty} E(\bar{y} - y^* | \bar{y} > k) = 0.$$

Thus, a method of reducing the effect of regression on the baseline measure is to take the average of a number of measures. Figure 1 illustrates the effect using the cholesterol example from the LRC Prevention Trial. It is apparent that in this case the reduction in regression effect is quite steep until four observations are used, after which each additional observation results in little gain. Of course, the choice of the number of observations to use will also depend on the cost of obtaining them.

Ederer (4) proposed reducing the effect of regression by using the first observation to classify a subject and a second observation as the baseline from which to measure treatment effects. From formula 2 we find that

$$E(y_2 | y_1 > k_1) = \mu + \rho_{12}c_1\sigma$$

and

$$E(y_3 | y_1 > k_1) = \mu + \rho_{13}c_1\sigma,$$

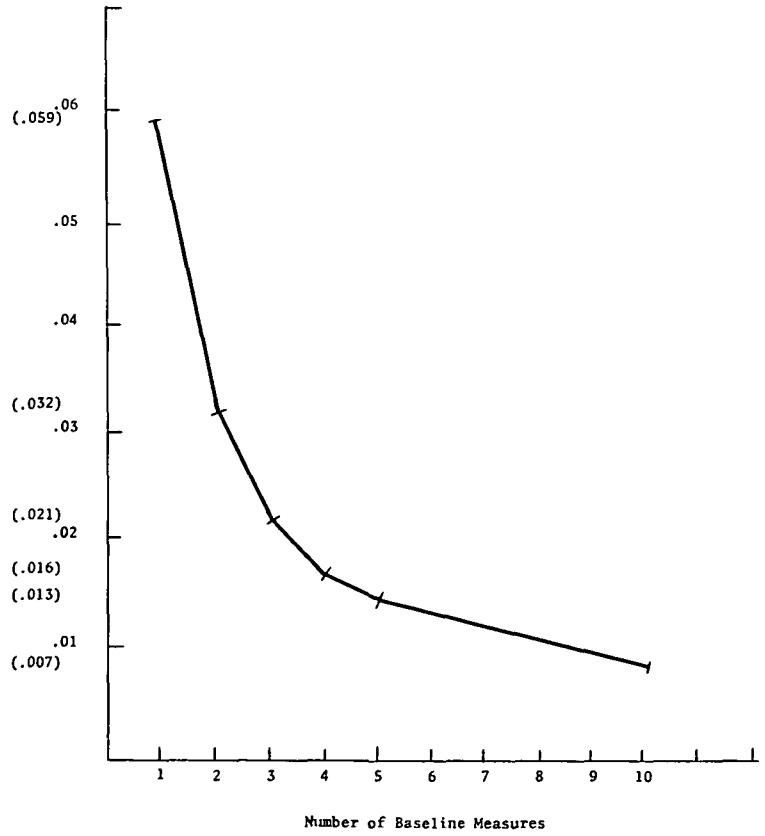

FIGURE 1. Expected regression to the mean of log cholesterol by the number of baseline measures, LRC prevalence estimates.

where ρ_{13} is the correlation between the first and third observations. Under this situation, the effect of regression is

$$E(y_2 - y_3 | y_1 > k_1) = c_1\sigma(\rho_{12} - \rho_{13}).$$

Hence, the effect of regression will be zero if $\rho_{12} = \rho_{13}$, i.e., if the correlation between the first and second observations is identical with that between the first and third observations. Based on this, we can see that it is not necessary to take the average of several observations if the difference $\rho_{12} - \rho_{13}$ is small. Thus, in the planning stages, it would be worthwhile to examine this difference to determine the method of classification and baseline. However, in terms of the probability of misclassification, the use of the average is clearly superior to the use of a single observation. For the details in terms of the model used here see Gardner and Heady (5).

If one has $k(\geq 2)$ measures available, the average of k_1 can be used for classification purposes and the average of k_2 for the baseline measure ($k = k_1 + k_2$). This will lead to a small probability of misclassification as well as reduce the regression to the mean.

Summary

In this paper, we have noted the ways in which regression to the mean can affect the measurement of treatment effects in clinical and epidemiologic studies. It is apparent that regression can have a sizeable effect and may lead to erroneous conclusions concerning treatment effects. Thus, the procedures outlined here should be useful in taking regression into account at the planning stage as well as at the time of analysis.

References

1. Lipid Research Clinics: Reference Manual for Lipid Research Clinics Program Prevalence Study. Unpublished protocol. Chapel Hill, NC, Central Patient Registry and Coordinating Center, University of North Carolina, 1972
2. James KE: Regression toward the mean in uncontrolled clinical studies. Biometrics 29:121-130, 1973
3. Remington RD: The Hypertension Detection and Follow-Up Program (USA). Bull Inst Natl Sante Rech Med 21:185-194, 1973
4. Ederer F: Serum cholesterol: effects of diet and regression toward the mean. J Chronic Dis 25:277-289, 1972
5. Gardner MJ, Heady JA: Some effects of within-person variability in epidemiological studies. J Chronic Dis 26:781-795, 1973
6. Tallis GM: The Moment Generating Function of the Truncated Multinormal Distribution. Journal of the Royal Statistical Society, B. 23:223-229, 1961
7. Lipid Research Clinics: Protocol for the Lipid Research Clinics Type II Coronary Primary Prevention Trial. Unpublished protocol. Chapel Hill, NC, Central Patient Registry and Coordinating Center, University of North Carolina, 1973