IJE vol.34 no.1 © International Epidemiological Association 2004; all rights reserved.
Advance Access publication 27 August 2004

International Journal of Epidemiology 2005;**34**:215–220
doi:10.1093/ije/dyh299

EDUCATION

Regression to the mean: what it is and how to deal with it

Adrian G Barnett,[1] Jolieke C van der Pols[1] and Annette J Dobson[1]

Accepted 1 July 2004

Background Regression to the mean (RTM) is a statistical phenomenon that can make natural variation in repeated data look like real change. It happens when unusually large or small measurements tend to be followed by measurements that are closer to the mean.

Methods We give some examples of the phenomenon, and discuss methods to overcome it at the design and analysis stages of a study.

Results The effect of RTM in a sample becomes more noticeable with increasing measurement error and when follow-up measurements are only examined on a sub-sample selected using a baseline value.

Conclusions RTM is a ubiquitous phenomenon in repeated data and should always be considered as a possible cause of an observed change. Its effect can be alleviated through better study design and use of suitable statistical methods.

Keywords Regression to the mean, repeated measures, intervention, clinical trials, observational studies, longitudinal studies, statistics, epidemiological research design

In this tutorial style paper we give an introduction to the problem of regression to the mean (RTM) and then use examples to highlight practical methods for dealing with the problem at the design and analysis stages of a study.

RTM at the subject level

RTM is a statistical phenomenon that occurs when repeated measurements are made on the same subject or unit of observation. It happens because values are observed with random error. By random error we mean a non-systematic variation in the observed values around a true mean (e.g. random measurement error, or random fluctuations in a subject). Systematic error, where the observed values are consistently biased, is not the cause of RTM. It is rare to observe data without random error, which makes RTM a common phenomenon.

Figure 1 illustrates a simple example of RTM using an artificial but realistic[1] distribution of high density cholesterol (HDL) cholesterol in a single subject. The first panel shows a Normal distribution of observations for the same subject. The true mean for this subject (shown here as 50 mg/dl) is unknown in practice and we assume it remains constant over time. We assume that the variation is only due to random error (e.g. fluctuations in the HDL cholesterol measurements, or the subject's diet).

In the second panel we show an observed HDL cholesterol value (from this Normal distribution) of 30 mg/dl, a relatively low reading for this subject. If we were to observe another value in the same subject it would more likely be >30 mg/dl than <30 mg/dl (third panel). That is, the next observed value would probably be closer to the mean of 50 mg/dl (third panel).

In general, when observing repeated measurements in the same subject, relatively high (or relatively low) observations are likely to be followed by less extreme ones nearer the subject's true mean. This phenomenon was first discussed by Sir Francis Galton in 1877 (see Stigler[2] for an historical account of RTM), and it was Galton who coined the phrase 'regression to the mean'. The practical problem caused by RTM is the need to distinguish a real change from this expected change due to the natural variation. For example, in the third panel of Figure 1 we might think that the subject's HDL cholesterol has increased when in fact the first measurement was just unusually low and the subject's true mean HDL cholesterol has remained constant.

RTM at the group level

The problem of RTM is not restricted to individual measurements. We now give an example where the effect of RTM is compounded by categorizing subjects into groups based on their baseline measurement(s).

[1] School of Population Health, The University of Queensland, Herston, Brisbane QLD 4006, Australia.

Correspondence: Adrian Barnett, School of Population Health, The University of Queensland, Brisbane QLD 4006, Australia. E-mail: a.barnett@uq.edu.au

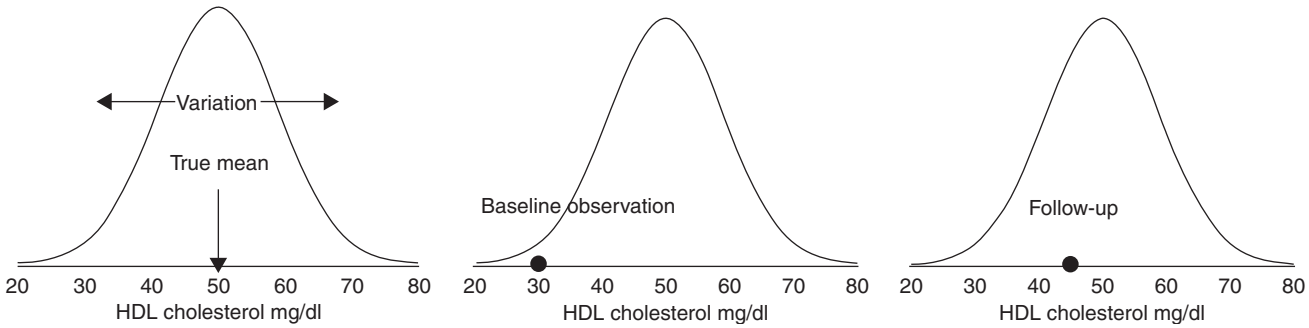

Figure 1 Graphical example of true mean and variation, and of regression to the mean using a Normal distribution. The distribution represents high density lipoprotein (HDL) cholesterol in a single subject with a true mean of 50 mg/dl and standard deviation of 9 mg/dl

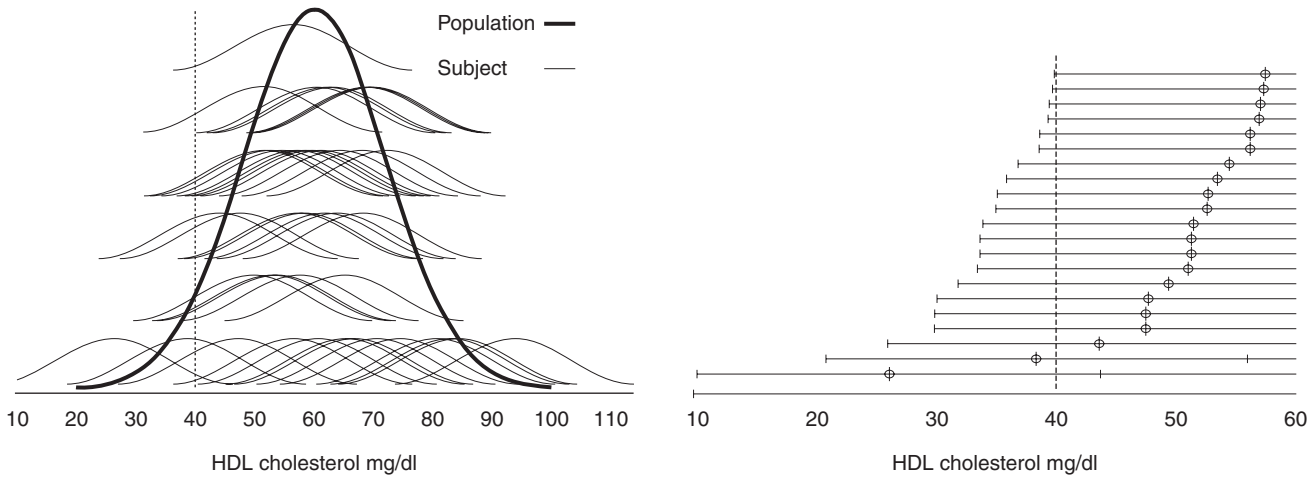

Figure 2 Hypothetical population distribution of high density lipoprotein HDL cholesterol and distributions for 50 simulated subjects. The cut-point at 40 mg/dl identifies those with a low HDL cholesterol. The second panel focuses on the left tail of the population and hence on those 21 subjects who may be included in the 'low' group. The second panel shows the mean and variation in HDL cholesterol for these 21 subjects, and is a bird's-eye view of the first panel. Adapted from Yudkin & Stratton[15]

Consider measuring HDL cholesterol in a random sample of subjects from a defined population. The distribution of HDL cholesterol in this population is Normal with some mean (say 60 mg/dl) and standard deviation (say 12 mg/dl). Subjects in the population each have their own mean (which is within a population range of about 20–100 mg/dl). Due to measurement error and random fluctuation we cannot expect that subjects will have the same HDL cholesterol reading if measured at two different times.

The first panel of Figure 2 shows a hypothetical population distribution of HDL cholesterol and the simulated distributions of 50 subjects generated from this population. We assumed that at the population level HDL cholesterol was Normally distributed with mean 60 mg/dl and standard deviation 12 mg/dl. At the subject level HDL cholesterol is also Normally distributed but with a smaller standard deviation (9 mg/dl). We further assumed that the variation was the same for all subjects. Chesher discusses the RTM effect in non-Normal data.[3]

There are more subjects with a true mean close to 60 mg/dl (the population mean) and fewer near the population tails (e.g. with means of <30 or >90 mg/dl).

After an initial measurement, suppose a subgroup of subjects is identified with undesirably low HDL cholesterol. In this example we use a value of <40 mg/dl to define a group of people targeted for treatment.[4] The second panel of Figure 2 shows the subjects who may be included in this low HDL cholesterol group. There are only 2 subjects whose true mean is <40 mg/dl, and 19 whose true mean is >40 mg/dl. On re-measurement these 19 subjects with a true mean >40 mg/dl will most likely return results closer to their true mean. This will have the effect of increasing the overall mean of the low group. Such a change in the group mean may wrongly be attributed to a true change in HDL cholesterol when the real cause would be RTM. The issue becomes particularly important when an intervention is applied between the two measurements.

The left panel of Figure 2 highlights another important consequence of random variation: the variability in individual measurements is greater than the variability in the true means. Suppose we wanted to use individuals with levels of HDL cholesterol similar to those in Figure 2 to estimate the increasing risk of coronary heart disease with decreasing HDL cholesterol. A simple method would be to fit a regression line to data comprising coronary event rates (y-axis) plotted against HDL (x-axis). If we used the individual measurements of HDL the x-axis data would have a greater range than if we used the true means, whereas the data on the y-axis remains unchanged.

Therefore the slope of the line would be shallower than if the true means had been known and used. This attenuation of association is known as the regression dilution bias.[5]

Quantifying the effect of RTM

The above example highlighted how both the within-subject variance σ_w^2 and the population or between-subject variance σ_b^2 contribute to RTM. The formula to calculate the expected RTM effect, for Normally distributed data, is defined as:[6,7]

$$\text{RTM effect} = \frac{\sigma_w^2}{\sqrt{\sigma_w^2 + \sigma_b^2}} C(z), \\ = \sigma_t(1 - \rho)C(z), \quad -1 \leqslant \rho \leqslant 1, \qquad (1)$$

where $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$ is the total variance, $\sigma_w^2 = (1 - \rho)\sigma_t^2$ is the within-subject variance, $\sigma_b^2 = \rho\sigma_t^2$ is the between-subject variance, ρ is the correlation and,

$$C(z) = \phi(z)/\Phi(z),$$

where $z = (c - \mu)/\sigma_t$ if the subjects are selected using a baseline measurement greater than c, and $z = (\mu - c)/\sigma_t$ if the subjects are selected using a baseline measurement less than c; μ is the population mean. The terms $\phi(z)$ and $\Phi(z)$ are respectively the probability density and the cumulative distribution functions of the standard Normal distribution.

In the HDL cholesterol example the cut-off was a baseline measurement less than $c = 40$ mg/dl, so we use $z = (60 - 40)/15 = 1.33$. From tables of the standard Normal distribution $\phi(z) = 0.16$ and $\Phi(z) = 0.09$. From the within ($\sigma_w^2 = 9^2$) and between-subject variances ($\sigma_b^2 = 12^2$) we can calculate $\rho = 0.64$, so approximately the RTM effect = 9.6 mg/dl, a seemingly large increase in the group's mean HDL cholesterol.

For some measurements the correlation ρ may be expected to decay over time. Equation (1) shows that as the correlation becomes smaller the RTM effect increases. Equation (1) also shows that the RTM effect is proportional to the population standard deviation, and that the effect increases as the value of $C(z)$ increases, which corresponds to a more extreme cut-off value (closer to either tail of the Normal distribution).

Some real-life examples of RTM

One of Galton's first examples was the average height of parents and their children. He found that tall parents had (on average) children who were smaller than them, and that short parents had (on average) children who were taller than them. In both cases the children with parents at the extreme ends of the distribution had heights closer to the population mean height.

In more recent studies RTM has been reported in birthweights,[8] blood pressure,[9] and cholesterol.[10] Nevertheless it continues to be missed by some public health researchers; for example, part of the 90% drop in cases of meningitis C in the UK after the introduction of an immunization programme could be due to a very bad year being likely to be followed by a better year.[11] Some other interesting examples are given by Bland and Altman.[12]

Identifying and dealing with RTM

Example: the Nambour Skin Cancer Prevention Trial

To illustrate the statistical methods used to detect and control for RTM we used a random subset of measurements of serum betacarotene from the Nambour Skin Cancer Prevention Trial.[13] This community-based randomized trial investigated the effect of a daily betacarotene supplement and daily application of sunscreen on skin cancer. The effect of the betacarotene supplement on serum levels was investigated in a random subsample of trial participants, who provided a blood sample at the start of the trial, in February 1992, and another blood sample at the end of the supplementation period in July 1996 (unpublished). The betacarotene measurements (μM/l) in this study were strongly positively skewed. For our purpose we therefore log-transformed the data to make them approximately Normally distributed. The data consist of n = 96 paired measurements, n = 52 from the treatment group (betacarotene supplement) and n = 44 from the placebo group. In the analyses presented here we are interested in whether the supplements increased betacarotene levels (i.e. a genuine treatment effect).

How to use graphs to help identify RTM

One should assume that RTM has taken place unless the data show otherwise. The initial examination of the data should include a scatterplot of change (follow-up minus baseline measurements) against baseline measurements, which can help identify the magnitude of the RTM effect. An example scatterplot is shown in Figure 3 for the log-transformed betacarotene data from the Nambour Skin Cancer Prevention Trial. The solid line represents perfect agreement (i.e. no change) between the follow-up and baseline values. The dotted lines were obtained by linear regression of the change values on baseline values including a group covariate; the higher line is for the treatment group and the distance between the regression lines indicates a possible treatment effect. Some RTM is apparent in the plots, as subjects whose baseline results were unusually low have tended to increase (so that change values are likely to be above the solid line), and subjects whose baseline results were unusually high have tended to decrease (so that change values are likely to be below the solid line). This pattern is clearer in the placebo group where there was less change in the group mean between the measurement times.

How to reduce the effects of RTM at the study design stage

The effect of RTM can be reduced by a good study design. We describe two such designs below. These designs can be combined to give even greater protection against RTM, and are described in detail by Yudkin & Stratton.[14]

1. Random allocation to comparison groups

If subjects are randomly allocated to comparison groups the responses from all groups should be equally affected by RTM. With two groups, placebo and treatment, the mean change in the placebo group provides an estimate of the change caused by RTM (plus any placebo effect). The difference between the mean change in the treatment group and the mean change in the placebo group is then the estimate of the treatment effect after adjusting for RTM.

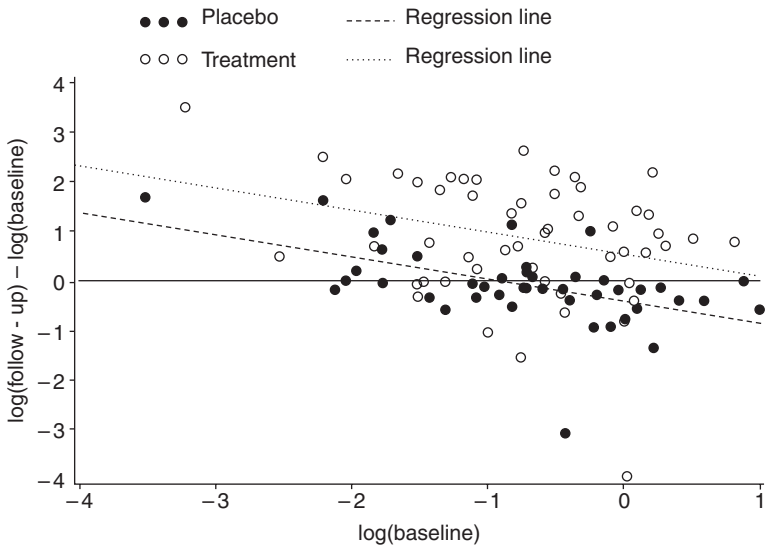

Figure 3 Scatter-plot of n = 96 paired and log-transformed betacarotene measurements showing change (log(follow-up) minus log(baseline)) against log(baseline) from the Nambour Skin Cancer Prevention Trial. The solid line represents perfect agreement (no change) and the dotted lines are fitted regression lines for the treatment and placebo groups

2. Selection of subjects based on multiple measurements

The effect of RTM increases with larger measurement variability (see Equation (1)). To reduce the variability we can select subjects using two or more baseline measurements. The study selection criterion (i.e. a cut-off) is then applied to either the mean of the multiple measurements, or the second (or later) measurement, assuming that the RTM effect has taken place between the first and second (or later) measurements. This method can be thought of as an attempt to get a better estimate of each subject's true mean before the intervention. The advantages of taking extra measurements are it gives better estimates of the mean and the within-subject variation.

With multiple baseline measurements the expected RTM effect[7,15] is,

$$\text{RTM effect} = \frac{\sigma_w^2/m}{\sqrt{(\sigma_w^2/m) + \sigma_b^2}} C(z), \qquad (2)$$

where m is the number of baseline measurements, and σ_w^2,σ_b^2, and $C(z)$ are as in Equation (1). We give SAS code to calculate Equations (1) and (2) on our web page.[16]

Figure 4 shows the reduction in the RTM effect due to increasing the number of baseline measurements using $z = 1.33$, $\sigma_t = 15$, and with $\rho = 0.1$ and $\rho = 0.5$. The reduction in the RTM effect is biggest between the first and second measurements; the benefit of extra baseline measurements decreases. As $m \to \infty$ the RTM effect tends to zero according to Equation (2), but in both examples here the RTM effect is still reasonably large even when $m = 5$. However, Johnson and George[15] use a more realistic model in which two measurements from the same subject are more similar than those from two different subjects and hence show that the reduction in RTM by taking repeated measurements does not tend to zero as $m \to \infty$.

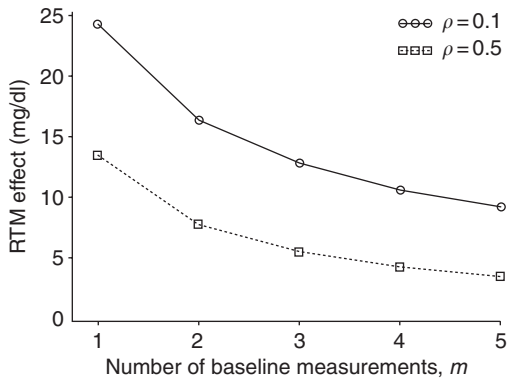

Figure 4 An example of the reduction in the regression to the mean (RTM) effect due to taking multiple baseline measurements and using each subject's mean as the selection variable. We use some values from the earlier high density lipoprotein (HDL) cholesterol example, $\sigma_t = 15$, $\mu = 60$, and $c = 40$ mg/dl, but we use two different values of ρ (the correlation)

How to deal with RTM in data analysis

Many different methods have been proposed to estimate the size of the RTM effect and to adjust observed measurements for RTM.[7,14,17,18] We give details of two methods below.

1. Correction using Equations (1) or (2)

If we know, or can estimate the mean and standard deviation of the population distribution and the within-subject standard deviation then we can estimate the RTM effect using Equation (1) or Equation (2) for multiple baseline measurements. This value can then be subtracted from the observed change to give an adjusted estimate.

Table 1 Analysis of change (follow-up result minus baseline) in log-transformed betacarotene measurements

Parameter	*n*	Mean change	95% CI change	*P*-value
a) No cut-off (n = 96)				
Placebo	44	−0.09	−0.33, 0.15	0.46[a]
Treatment	52	0.85	0.51, 1.20	<0.0001[a]
Difference (Treatment-Placebo)	96	0.94	0.51, 1.37	<0.0001
b) Baseline cut-off <0.5 μM/l (n = 49)				
Placebo	23	0.24	−0.04, 0.53	0.09[a]
Treatment	26	1.09	0.60, 1.57	0.0001[a]
Difference (Treatment–Placebo)	49	0.84	0.27, 1.41	0.004

[a] Using a paired *t*-test.

Table 2 Analysis of covariance (ANCOVA) of log-transformed follow-up betacarotene measurements

Coefficient	Mean	95% CI Mean	*P*-value
a) No cut-off (n = 96)			
Baseline	0.56	0.33, 0.78	<0.0001
Difference (Treatment–Placebo)	0.94	0.55, 1.33	<0.0001
b) Baseline cut-off <0.5 μM/l (n = 49)			
Baseline	0.36	−0.03, 0.76	0.072
Difference (Treatment–Placebo)	0.86	0.37, 1.36	0.0006

2. ANCOVA

An approach which is often more practical is to use analysis of covariance (ANCOVA) which has high statistical power and adjusts each subject's follow-up measurement according to their baseline measurement.[19] This approach can be summarized using a single regression equation:

$$\text{Follow-up} = \text{constant} + a \times (\text{baseline-baseline mean}) + b \times \text{group} + \text{error}, \quad (3)$$

where group = 1 for treatment group, and group = 0 for placebo group. The coefficient b is the estimated treatment effect adjusted for RTM. Other terms may be added to Equation (3) to account for confounders or other variables of interest.

ANCOVA can also be used with the change between baseline and follow-up as the outcome variable, although the only difference from Equation (3) is that the regression coefficient, a, for the centred baseline value is decreased by one unit.[20]

As ANCOVA is a special case of a general linear model it can be performed in most statistical software packages used in epidemiological research (e.g. SPSS, SAS, Stata). The commands to perform ANCOVA and check the model's adequacy in a number of statistical packages are given on our web site,[16] and can also be obtained by contacting the authors.

Example analyses

We use the betacarotene data from the Nambour Skin Cancer Prevention Trial to demonstrate the statistical methods of adjusting for RTM. We use the data set of $n = 96$ pairs of values, and a data set of $n = 49$ pairs based on a cut-off of <0.5 μM/l (−0.69 on the log-scale) for baseline betacarotene to highlight the impact of RTM when cut-offs are used. In this example, treatment allocation was random, and hence the study was protected against RTM at the design stage.

Table 1 shows an analysis of the serum betacarotene data from the example data set. Using the full data set (no cut-off), the results show a significant increase in betacarotene in the treatment group and no apparent change in the placebo group. The increase in the treatment group, compared with the placebo group, is 0.94 (95% CI: 0.51, 1.37) on the log scale. For the data with a baseline cut-off, there appears to be a possible increase in the placebo group of 0.24 (95% CI: −0.04, 0.53); unless there is a placebo effect, this could be due to RTM. Using Equation (1) with $\sigma_t = 0.86$, $\rho = 0.56$, $\mu = -0.75$, and $z = -0.07$, we estimate the RTM effect in the placebo group as 0.29. Hence we conclude that there was no real change in the betacarotene levels in the placebo group.

Table 2 shows the results using ANCOVA, the estimated treatment effect is similar to the paired *t*-test results, but with narrower confidence intervals, particularly from the subset of data using the cut-off. The narrower intervals are due to the baseline term explaining more of the variance in the outcome in the ANCOVA model.

Discussion

We have highlighted the problem of regression to the mean (RTM) using some simple biological examples where the variable was approximately Normally distributed. However, RTM is not restricted to biological variables. It will occur in any measurement (biological, psychometric, anthropometric, etc) that is observed with error. Also it is not restricted to distributions that are Normal, or even to distributions that are continuous. RTM can occur in binary data where it would cause subjects to change categories without any true change in their underlying response.

Using data from a study in which subjects were randomly allocated to groups *t*-tests and ANCOVA gave results that were the same when there was no baseline cut-off. When a cut-off was used, ANCOVA gave narrower confidence intervals for the treatment effect, and the paired *t*-test showed a change in the placebo group consistent with RTM. We recommend using ANCOVA in any situation where *t*-tests could be used.

RTM occurs in any variable that is subject to random error, and therefore it needs to be ruled out as a cause of an observed change before any other explanation is sought. It has already caught out many researchers[21]—we hope that people who read this article will avoid this mistake.

Acknowledgements

We thank Professor Adèle Green, Associate Professor Geoff Marks and Dr Philip Gaffney for providing the Nambour Skin Cancer Prevention Trial data. We also thank members of the Longitudinal Studies Unit, School of Population Health, The University of Queensland, for helpful discussions on this topic. This work was funded by the National Health and Medical Research Council of Australia (grant number: 252834) and carried out at The University of Queensland and The Queensland Institute of Medical Research.

KEY MESSAGES

- Reduce regression to the mean (RTM) at the design stage: (1) include a randomly allocated placebo group, (2) take multiple baseline measurements, although this is unlikely to completely eliminate the problem.
- Identify RTM at the analysis stage: (1) examine a scatterplot of change against baseline; is there more change at the tails of the baseline measurements?
- Deal with RTM at the analysis stage: (1) estimate the size of the RTM, (2) analyse the data using analysis of covariance.

References

[1] Marcovina SM, Gaur VP, Albers JA. Biological variability of cholesterol, triglyceride, low- and high-density lipoprotein cholesterol, lipoprotein(a), and apolipoproteins A-I and B. *Clin Chem* 1994;**40:**574–78.

[2] Stigler SM. Regression towards the mean, historically considered. *Statist Meth Med Res* 1997;**6:**103–14.

[3] Chesher A. Non-normal variation and regression to the mean. *Statist Meth Med Res* 1997;**6:**147–66.

[4] Bays HE, McKenney JM, Dujovine CA *et al*. Effectiveness and tolerability of a new lipid-altering agent, Gemcabene, in patients with low levels of high-density lipoprotein cholesterol. *Am J Cardiol* 2003;**92:**538–43.

[5] MacMahon S, Peto R, Cutler J *et al*. Blood pressure, stroke, and coronary heart disease. *Lancet* 1990;**335:**765–74.

[6] Gardner MJ, Heady JA. Some effects of within-person variability in epidemiological studies. *J Chron Dis* 1973;**26:**781–95.

[7] Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 1976;**104:**493–98.

[8] Wilcox MA, Chang AM, Johnson IR. The effects of parity on birthweight using successive pregnancies. *Acta Obstet Gynecol Scand* 1996;**75:**459–63.

[9] Kario K, Schwartz JE, Pickering TG. Changes of nocturnal blood pressure dipping status in hypertensives by nighttime dosing of α-adrenergic blocker, Doxazosin: Results from the HALT study. *Hypertension* 2000;**35:**787–94.

[10] Schectman G, Hoffmann RG. A history of hypercholesterolemia influences cholesterol measurements. *Arch Intern Med* 1988;**148:**1169–71.

[11] Wise J. Meningitis C rates show steep fall. *BMJ* 2001;**322:**70.

[12] Bland JM, Altman DG. Statistics notes: Some examples of regression towards the mean. *BMJ* 1994;**309:**780.

[13] Green A, Williams G, Neale R *et al*. Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin: a randomized controlled trial. *Lancet* 1999; **354:**723–29.

[14] Yudkin PL, Stratton IM. How to deal with regression to the mean in intervention studies. *Lancet* 1996;**347:**241–43.

[15] Johnson WD, George VT. Effect of regression to the mean in the presence of within-subject variability. *Stat Med* 1991;**10:**1295–302.

[16] Barnett AG. SAS code to calculate the effect of regression to the mean and an example analysis of covariance (ANCOVA), May 2004. http://hisdu.sph.uq.edu.au/lsu/adrian/rtmcode.htm.

[17] Chinn S, Heller RF. Some further results concerning regression to the mean. *Am J Epidemiol* 1981;**114:**902–05.

[18] Chapurlat RD, Blackwell T, Bauer DC, Cummings SR. Changes in biochemical markers of bone turnover in women treated with raloxifene: Influence of regression to the mean. *Osteoporosis Int* 2001;**12:**1006–14.

[19] Twisk JWR. *Applied Longitudinal Data Analysis for Epidemiology: a Practical Guide*. Cambridge University Press, 2003.

[20] Laird N. Further comparative analysis of pre-test post-test research designs. *Am Statistician* 1983;**37:**329–30.

[21] Senn S. Regression to the mean. *Statist Meth Med Res* 1997;**6:**99–102.