Centring in regression analyses: a strategy to prevent errors in statistical inference

HELENA C. KRAEMER,[1] CHRISTINE M. BLASEY[1]
1 Stanford University, USA

Abstract

Regression analyses are perhaps the most widely used statistical tools in medical research. Centring in regression analyses seldom appears to be covered in training and is not commonly reported in research papers. Centring is the process of selecting a reference value for each predictor and coding the data based on that reference value so that each regression coefficient that is estimated and tested is relevant to the research question. Using non-centred data in regression analysis, which refers to the common practice of entering predictors in their original score format, often leads to inconsistent and misleading results. There is very little cost to unnecessary centring, but the costs of not centring when it is necessary can be major. Thus, it would be better always to centre in regression analyses. We propose a simple default centring strategy: (1) code all binary independent variables +1/2; (2) code all ordinal independent variables as deviations from their median; (3) code all 'dummy variables' for categorical independent variables having m possible responses as 1 – 1/m and –1/m instead of 1 and 0; (4) compute interaction terms from centred predictors. Using this default strategy when there is no compelling evidence to centre protects against most errors in statistical inference and its routine use sensitizes users to centring issues.

Key words: regression, centring, multicollinearity

Centring in regression analyses: a strategy to prevent errors in statistical inference

Centring in regression analysis means making a decision before entering data, consistent with the goals and context of the research project, as to how to code the independent variables in order to ensure accurate, relevant interpretation of results. The advantages of appropriate centring and consequences of not centring have been known for decades and have been covered in a few textbooks. However, the concept still appears rarely discussed in research training or reported in research applications. Yet regression analyses are not rare in research reports. Over a 6-month period in 2003, approximately 29% (24/84) of new research articles in the *American Journal of Psychiatry* contained some form of regression analysis. Of these, only four (17%) used centred data. One present author (HCK) learned about centring some 40 years ago, from the late Professor Lee Cronbach who taught: 'In regression analysis, always center!' He cited two reasons. First, when data are not centred, the regression coefficients that are estimated and tested may be irrelevant and misleading. Second, centring, thoughtfully done, can diminish the almost inevitable multicollinearity problems in regression, thus increasing both the precision of parameter estimation and the power of statistical testing of those parameters. Finally, he pointed out that if centring is done unnecessarily the cost is minor.

Since then, a few texts have accurately described centring and the consequences of non-centred data (Aitken and West, 1991; Cohen et al., 2003) and these should be consulted for greater mathematical detail than is here presented However, in contrast to Cronbach's injunction, these authors and others (Glantz and Slinker, 2001; Kromrey and Foster-Johnson, 1998) take the stand that centring doesn't usually change the statistical results, is necessary only in certain circumstances, and can thus easily be

avoided. Perhaps as a result, centring seems to be done only in a minority of published papers using regression analyses, and errors may often result.

For example, in a randomized clinical trial (RCT) of treatment for Alzheimer patients where age at entry to the study is an independent variable, in absence of centring, the effect of treatment is evaluated for newborns with Alzheimer disease! With centring at the median, the effect of treatment is more reasonably evaluated for those at the median age of Alzheimer patients recruited into the RCT.

The evident disagreement among methodologists on this issue relates not to centring 'per se', but to the value of a preventive injunction: 'In regression analysis, always center!' to avoid contamination of statistical inference. Like an analogous preventive injunction to avoid food contamination: 'Wash your hands both before and after handling raw meat!' such an injunction is not necessary for those expert in the area. They know when and how to take precautions against contamination. As is true of hand washing, following this preventive injunction does not guarantee 100% certainty of the desired outcome, but it does substantially reduce the risk of an undesirable outcome. At the same time, not following this preventive injunction does not guarantee 100% certainly of contamination. As is true of hand washing, in many cases, the results will be the same whether or not the injunction is followed. However, it is difficult to ascertain when the risk of contamination is low enough to ignore safely the injunction, and the consequences, even if rare, are serious. Finally, like hand washing, to follow the recommendation takes only a little time and effort and serves as a constant reminder of potential problems.

Aiken, West and others recommend, however, that one centre only in the presence of interactions. However, there are circumstances when, even in absence of interactions, centring makes a difference to statistical inference (see below). Moreover, regression users often respond to this recommendation by omitting important interactions from the model in order to avoid dealing with issues of centring, which then creates even greater statistical problems (see below).

Some methodologists object to this injunction, arguing that the source of any discrepancies among statistical findings based on regression analyses in absence of centring is not mysterious; they can always be explained and resolved. This is true (see below). However, the basic concept of prevention is that it is more important to prevent problems than to understand their source after they occur, particularly when unwanted consequences may not be immediately recognized as such.

Here we argue for Cronbach's advice that one should always centre in regression analyses. Ideally, how to centre would be determined with expert consultation. However, to be realistic, such consultation is not always available. We therefore propose a simple default strategy to be used in order to avoid errors of statistical inference in the absence of such consultation. In what follows, we demonstrate some of the consequences of not centring and the effects of using this particular default centring strategy. We will first provide a brief reminder of the general principles of regression analyses. Then we will illustrate the consequences of not centring, introduce our default proposal for centring, and show what the protection to statistical inference would then be. Finally, lest it be thought we are trivializing Cronbach's advice by comparing it with hand washing, we will discuss how far-reaching and profound Cronbach's simple advice actually is.

General principles

A regression model is one in which the distribution of some outcome or dependent variable is assumed to be determined by a linear combination of m independent variables (X_1, X_2, etc.). For example, with two independent variables a linear regression model would be of the form:

$$Y = \beta_0 + \beta_1 (X_1 - X_1^*) + \beta_2(X_2 - X_2^*) + \beta_3(X_1 - X_1^*)(X_2 - X_2^*) + \varepsilon$$

where Y is the dependent variable and X_1 and X_2 the independent variables, β_0 'the intercept', and β_1, β_2 and β_3 the 'regression coefficients', with ε the error. For example, the dependent variable may indicate the probability of success of some intervention, or measure the severity of symptoms following intervention in a RCT, or time of onset of a disorder in a risk research study. Xs may indicate which of two treatments was used or the subject's gender (binary variables), the age or educational level of the subject (ordinal variables), ethnicity or site (categorical variables). In uncentred analysis, all X*s equal zero; in centred analyses, these are constants set by the researchers.

Regression coefficients estimated and tested in

regression analyses are population parameters; what matters to statistical inference is their meaning and interpretation in the population. Coefficients like β_1 and β_2 are 'simple' or 'main' effects, and coefficients attached to products of two or more independent variables are 'interaction' effects. A full (or saturated) linear model includes all interaction effects (two-way, three-way, four-way, . . . m-way, the number indicating how many independent variables are included in each product). However, researchers often choose to set some or all of the interaction βs to be zero and to proceed with a smaller model.

Each regression coefficient estimates the effect of the associated independent variable on the dependent variable when all other independent variables in the model are equal to their centred values (when X = X).*

In an uncentred analysis, when X* = 0 for all variables, this would mean the effect of the association when all other independent variables are zero. This statement indicates why choosing the X*s carefully (centring) is an important issue in using regression analyses. What is true only when certain higher order interactions are zero in the population (not merely set equal to zero in the model) is that the regression coefficient may estimate the effect of the associated independent variables on the dependent variable, not only at X = X* for all other independent variables in the model but more generally.

In an uncentred regression, if X is ordinal (for example, age, educational level) X* is zero, and if X is binary or categorical (for example, male/female, treatment/control), typically each response is coded 1 or 0. 'Centring', on the other hand, means *deliberately and thoughtfully* to select a reference value of X, X* for each ordinal value (which may or may not be zero), and value of a and b (which may or may not be 0 and 1) to code the responses of each binary or categorical independent variable, to make sure that each regression coefficient that is estimated and tested is both meaningful and relevant to the specific research questions of interest. Expert consultation would focus on what the specific research questions are and how best to centre in order that the regression coefficients address those specific research questions.

In data analysis, the population regression coefficient associated with each independent variable (up to and including the m-way interaction) is either (1) set to zero by the decision of the researcher or (2) estimated from the data. If the research question concerns only the linear combination as a whole (for example, the multiple correlation coefficient), or if the only coefficient of interest is the m-way interaction, it will not matter whether or not the independent variables are centred. The data analytic results are unaffected by centring, and thus the statistical inference is unaffected.

However, it is rare in medical applications that only the multiple correlation coefficient or the highest order interaction is of interest. In fact, the highest order interaction is the one most often set equal to zero. Generally, the individual coefficients are the primary focus of interest: for example, the treatment effect in a randomized controlled trial (RCT), or the potency of a risk factor, where other independent variables are considered. Thus, in general, for applications of regression models, centring will make a difference to the data analysis results, and thus to statistical inference. The methodological argument revolves only around the question of how often it will make a difference, when and how much, issues we will discuss using illustrations rather than statistical theory.

Regression with binary independent variables

Data for illustration are taken from the Infant Health and Development Program (1990), an eight-site randomized clinical trial for low birth-weight premature infants comparing an educational treatment against usual care, the dependent variable, the IQ at age 3. Consider first the situation in which there are two binary independent variables: treatment versus control, and advantaged versus disadvantaged status. This particular example is used for illustration because the model will fit the data perfectly. Consequently it can be readily seen exactly what each regression coefficient means, which is not always the case with more complex models.

Actually, were researchers faced with such data, they would often use a two-way analysis of variance (ANOVA), in which case, perhaps unknown to the users, the statistical program would centre the variables exactly as we will here recommend and obtain exactly the same results that would be obtained from following our recommendation. However, it is relevant to ask what would happen if, instead of allowing the experts who designed the ANOVA program to make the centring decisions, we made them ourselves.

Here, each of the independent variables can have its two responses coded in a variety of ways (for example, treatment = 1 and control = 0 or vice versa), and which

response is coded 1 and which 0 may differ from one researcher to another. Table 1a gives the means and sample sizes and the test statistics in these four cells of the IHDP sample. Table 1b also shows the regression parameter estimates for the four different ways of assigning 1 and 0 to the two independent variables.

We recommend that for binary independent variables (in absence of specific reason to choose otherwise) the values +1/2 and –1/2 be assigned to the two responses.

As is always true, the magnitude of the interaction effect (here the highest order term) remains the same regardless of how one centres. Here, the interaction effect is not statistically significant at the 5% level. However, the magnitudes of both the simple effects, as well as of the intercept, and their associated t-test statistics, change depending on how one chooses to code. Indeed the size of the treatment effect, which is of primary interest here, almost doubles between one way of coding and another.

Where each effect estimated in Table 1b comes from can readily be seen in Table 1a:

- For coding schemes I–IV, β_0 is always the mean response in the cell in which both independent variables are coded 0. For the coding scheme we recommend (V), β_0 is the average of the four cell means. In this particular example, testing whether β_0 is zero, for example, that an average IQ is zero, is of no interest: every IQ score is well above zero. However, if the outcome here were a change score, say a pre-post treatment difference, β_0, with centring as we recommend, would estimate the overall pre-post change (the main effect of time in a repeated measures ANOVA). Such an overall change score might even be a regression coefficient of primary research interest. This is one situation in which not centring makes a difference even in absence of interaction effects.

Table 1. Centring versus non-centring in a 2 × 2 design where X_1 is treatment (treatment versus control) and X_2 is an indicator of SES status (high versus low). Means and sample size are shown in 1a (in bold). The results (including regression coefficient estimates for the intercept (β_0), predictors (β_1 and β_2), and interaction (β_3); their standard error and the associated t-test statistics) of five examples of coding, use of ANOVA for testing with and without interaction are shown in 1b. Note that each parameter estimate in Table 1b can be located in Table 1a. (df = 844)

1a

	SES indicator (X_2)			
Treatment (X_1)	c = Low	d = High	Row means	Difference
a = Treatment	78.6 (33.2)	96.5 (180)	87.6	
b = Control	89.9 (236)	102.5 (100)	96.2	8.7
Column means	84.2	99.5	91.9	
Difference	15.2			

1b

Coding scheme	Coding values					Parameter estimates		
	a	B	c	d	β_0±se (t)	β_1±se (t)	β_2± (t)	β_3± (t)
I	1	0	1	0	102.5±1.8 (56.5)	–6.1±2.3 (–2.7)	–12.6±2.2 (–5.8)	–5.3±2.7 (–1.9)
II	1	0	0	1	89.9±1.2 (76.2)	–11.4±1.5 (–7.4)	+12.6±2.2 (+5.8)	+5.3±2.7 (+1.9)
III	0	1	1	0	96.5±1.4 (71.4)	+ 6.1±2.3 (+2.7)	–17.9±1.7 (–10.7)	+5.3±2.7 (+1.9)
IV	0	1	0	1	78.6±0.9 (78.9)	–11.4±1.5 (–7.4)	+17.9±1.7 (+10.7)	–5.3±2.7 (–1.9)
V	+1/2	–1/2	+1/2	–1/2	91.9±0.7 (134.3)	+8.7±1.4 (+6.4)	+15.2±1.4 (+11.1)	–5.3±2.7 (–1.9)
ANOVA	Automatic coding				(134.3)	(+6.4)	(+11.1)	(–1.9)
ANOVA	Automatic coding Omit interaction				(135.7)	(+7.6)	(+11.9)	Set to 0

- For coding schemes I–IV, β_1, the treatment effect is the difference between the treatment and control means (the treatment effect) *for whichever SES group was coded 0*. For coding scheme V, it is the average of the two treatment effects.
- For coding schemes I–IV, β_2, the SES effect, is the difference between the low and high SES means (the SES effect) for *whichever treatment group was coded 0*. For coding scheme V (–1/2 and +1/2), it is the average of the two SES effects.
- For all coding schemes, the interaction effect is the difference between the two treatment effects, or equivalently the difference between the two SES effects.

If one tested the null hypotheses of the above model using ANOVA rather than multiple regression analysis, coding scheme V (centred) is automatically used. Thus if one researcher used ANOVA and others used one or another of the uncentred regression analyses, their conclusions might differ, even though all are using exactly the same underlying linear model, exactly the same data, and all are computing correctly. All five of these solutions are technically correct (data analysis), but correct solutions to different research questions (statistical inference). Which question did the researcher intend to ask, and what question does the reader think is being answered? It is quite possible that the heterogeneity of results often seen across research reports results, at least in part, from inconsistent centring.

In fact, researchers are unlikely to design a RCT stratified on some factor (here SES) when they are interested in the treatment effect in only one of the strata, which is what they get from coding schemes I–IV. This is why ANOVA procedures automatically centre as they do, and why we also recommend that (in absence of strong reason to do otherwise) every binary independent variable be coded +1/2 and –1/2.

As shown here, inconsistent results between centred and non-centred data are not mysterious. Those related to the treatment effect arise because of the presence of the interaction (β_3), even though that interaction was *not* found to be statistically significant. The emphasis on the need for centring *only* in absence of interactions (Aiken and West, 1991; Cohen et al., 2003) has generated problems. It has led some researchers to omit important interactions from their model in order to avoid dealing with the centring issue, as if ignoring it in the model removes it from the population. Moreover, those who advocate omitting the interaction from the model often assert that in the presence of an interaction, none of the main effects are interpretable (Cohen, 1978). On the contrary, main effects are uninterpretable only because appropriate centring was not done. Even worse, some researchers test the null hypothesis of zero interaction and justify subsequently omitting it from the model if it is non-significant. As the sample size necessary for adequate power to detect an interaction effect is generally larger than to detect main effects, this can be a serious error. Non-statistically significant results do not 'prove' the null hypothesis.

When, for whatever reason, an existing population interaction is omitted from the model, at best the error variance is exaggerated, thus decreasing the precision of estimation and the power to detect clinically significant effects. At worst, some of the interaction effect is also remapped into the main effects, thus sacrificing accuracy of estimation as well. In short, by omitting an interaction in the model when it exists in the population, centring may not matter to the data analysis but the answers may lead to incorrect statistical inference. The right answers are obtained by both including any interaction for which there is 'a priori' rationale and justification and appropriate centring.

In Table 1, the interaction effect was not statistically significant ($p < 0.05$). However the estimated interaction effect was about five points on the IQ scale, a consequential difference. It can be seen that the effect of treatment was almost twice as high in the low SES group (11.4) as in the high SES group (6.1). Raising the IQ from 78.6 to 89.9 in the low SES group may be a far more important treatment effect from a clinical or policy standpoint than raising the IQ from 96.5 to 102.5 in the high SES group, even though the sample size in this study was not large enough to declare this difference statistically significant. Thus either summarily omitting the interaction effect or testing the interaction effect and omitting it if it is non-significant can seriously mislead statistical inferences about the population, when, as here, there is 'a priori' reason to believe that the effect of the treatment may be greatest for those most in need of treatment.

Regression with ordinal independent variables

Suppose now one were comparing the effects of treatment versus control (binary independent variable) and

an ordinal index of disadvantage on IQ. If there is reason to believe that the effect of treatment may be influenced by disadvantage, both the main effect of disadvantage (here measured on a 1 to 5 scale) and its interaction with treatment (binary) would be included as independent variables.

Based on the above, we would recommend coding the two responses to X_1(treatment) as +1/2 and –1/2. If we did not centre X_2, (leaving $X_2^* = 0$ by default), the effect of treatment would be evaluated for those with $X_2 = 0$, one step below the lowest possible value of the index of disadvantage, thus on non-existent individuals!

We recommend (in absence of strong reason to the contrary) that every ordinal independent variable be centred at its median.

Thus we here choose to centre X_2^*at 3. Then β_0, the intercept, would be the mean of T and C responses for the median subject. β_1 would be the difference between T and C responses for the median subject. β_2 would be the average of the T and C slopes of the dependent variable on the disadvantage score. β_3 would be the difference between those two slopes. It should again be here noted that, with centring, all main effects are meaningful and interpretable parameters of the population even with interactions included.

Summarized in Table 2, as always, both centred and uncentred approaches yield identical regression coefficients for the interaction term. However, with the uncentred approaches, the value of the intercept (51.8) is totally meaningless, representing the mean IQ in the control group of those with a disadvantage score one step below the minimum possible such score, that is, on non-existent subjects. With centring, it is 87.6, the overall mean IQ of low birth-weight premature children at median SES. The interpretation of simple effects is also compromised. Without centring, in Table 2, we see that the treatment effect is 16.9 in the uncentred analysis (evaluated on non-existent subjects), compared with 10.1 in the centred analysis (evaluated at median SES).

Yet, this model without centring is often used in RCTs to assess treatment effect 'controlling for' a baseline independent variable or in risk research to assess the effect of a risk factor 'controlling for' another (for example, age). This analysis is often implemented with analysis of covariance, where the centring of the binary independent variable would automatically be exactly as we recommend here, but the ordinal independent variable would not be centred ($X^* = 0$ by default), and the interaction would automatically be omitted.

With our proposed default rule, all effects are evaluated for the 'typical' subject – the one at the median of all ordinal independent variables. This protects against the most troublesome effect of not centring.

Regression with categorical independent variables

In many situations, independent variables are categorical (more than two non-ordered responses). For example there are RCTs in which patients are randomized to m (>2) treatments, or in which patients from m (>2) sites are randomized to treatment and control groups.

Often m 'indicator' or 'dummy' binary independent variables are used to code each of the m possible responses on the categorical independent variable. Every response is assigned a 1 on its designated dummy independent variable and 0 on all the others. Then one categorical response is selected as the reference group (often arbitrarily chosen), and its 'dummy' independent variable is omitted from the regression analysis.

For example, with four ethnic groups (white, African-American, Latino, other), one would have four dummy independent variables, one for each category. Then each subject's response on each dummy independent variable is 1 for the ethnic group to which that subject belongs, 0 otherwise. One of the four dummy independent variables is then omitted from the analysis, its associated group thus designated

Table 2. An example of multiple regression with a binary (treatment versus control) and an ordinal independent variable (index of disadvantage) of IQ at age 3, showing results (regression coefficient estimates, their standard errors, and the associated t-statistics), centred and non-centred (df = 844)

	Intercept	Treatment	Disadvantage	Interaction
Centred	+87.6 ± 0.6 (149.8)	+10.1 ± 1.2 (8.6)	+6.1 ± 0.4 (17.9)	–1.5 ± 0.7 (2.2)
Uncentred	+51.8 ± 2.2 (23.9)	+16.9 ± 3.4 (4.9)	+6.8 ± 0.4 (16.1)	–1.5 ± 0.7 (–2.2)

as the reference group (with ethnic groups, typically white).

We recommend (in absence of strong reason to the contrary) to centre as follows: as above, m 'dummy' variables are created, except that now, every categorical response would be coded 1 – 1/m instead of 1, and –1/m instead of 0. Again, one categorical response is omitted from the regression analysis.

Table 3 gives the uncentred results for the IHDP data comparing T and C (binary independent variable) at the eight sites (categorical independent variable), when independent variables are not centred and each site in turn is selected as the reference group. In these cases, the treatment effect is always the treatment-control difference *at the reference site*. Even in absence of a significant interaction effect (here $F(7,832) = 1.5$, $p = 0.161$) the effect of treatment will change depending on which site is used as the reference group. If Harvard were the reference group, the treatment effect would not even be statistically significant. If, Arkansas or Miami were the reference group, the treatment effect would be significant at the $p < 0.001$ level.

On the other hand, if centring is done as we recommend, the treatment effect does not change depending on which site is used as the reference group. The treatment effect is always the average treatment effect over the sites (One can check that by averaging the treatment effect from the first eight rows of Table 3 and comparing that with the treatment effect from the centred analysis.). Moreover, the results of testing in the centred analysis are exactly the same as would have been obtained had a 2 × 8 ANOVA been used, because ANOVA is generally programmed automatically to centre as we here recommend.

Table 3. Comparison of results on the treatment effect (treatment effect estimate, its standard error, and the associated t-statistic, df = 832) with a categorical independent variable (eight sites), using each site in turn as the reference category, non-centred and centred. Also presented is the 2 × 8 ANOVA test. (*p < 0.05, **p < 0.01, ***p < 0.001).

Reference site	Uncentred dummies
Arkansas	13.8 ± 3.5 (3.9)***
Einstein (NY)	10.2 ± 3.4 (3.0)**
Harvard	0.4 ± 3.4 (0.1)
Miami	13.6 ± 3.9 (3.4)***
Penn	11.8 ± 3.8 (3.1)**
Texas	7.7 ± 3.4 (2.3)*
Yale	9.2 ± 3.5 (2.7)**
Washington	10.3 ± 3.7 (2.8)**
Centred-all sites	9.6 ± 1.3 (7.6)***
8 × 2 ANOVA	(7.6)***

It is also very important to consider the ethnicity example here. If, in a RCT, independent variables include ethnicity and gender, with whites and males used respectively as the reference groups, with uncentred analysis, the treatment effect is estimated for white males only. Given recent emphasis on inclusion of both genders and all ethnicities in RCTs, it would certainly seem counterproductive, then, to use an uncentred analysis that estimates and tests exactly the treatment effect that would have been estimated, had only white males been included in the RCT!

Centring to help deal with multicollinearity problems

'Multicollinearity' means that there is a high correlation between one of the independent variables and some linear combination of the remaining ones (Glantz and Slinker, 2001), although many erroneously interpret multicollinearity as a high pairwise correlation between the independent variables. Multicollinearity always challenges interpretation of effects, but, of course, the stronger the multicollinearity the more difficult to meet the challenge.

To make matters worse, some regression users confuse 'multicollinearity' with 'interaction'. They argue, for example, that with random assignment to treatments in a RCT, there can be no multicollinearity of any baseline independent variables with treatment assignment (true), and therefore there can be no interaction (false). Multicollinearity refers to the correlational structure of the independent variables that exists in a population whatever the dependent variable. Interaction refers to the non-additivity of the effects of the independent variables that exists in a population for a particular dependent variable. With the same set of independent variables in the same population, there may be interaction for one dependent variable and not for another, but the multicollinearity is the same for all possible dependent variables. Thus multicollinearity may exist with or without interaction, and interaction may exist with or without multicollinearity. The two concepts are totally different, and their confusion generates both analytic and interpretive errors.

Multicollinearity seems ubiquitous in all but strictly experimental research with group sizes exactly balanced. In times past, when computers did not have the precision they do today, the major concern about multicollinearity was computational accuracy. Today multicollinearity is seldom extreme enough that the computer programs return error messages, but the multicollinearity warnings they do provide (such as 'Variance Inflation Factor') are not easily interpretable. Instead if the multicollinearity is too strong, the programs return unstable and imprecise estimates of regression coefficients that are hard to interpret correctly (Appelbaum and Cramer, 1974; Cramer and Appelbaum, 1980; McGee and Reed, 1984; Flack and Chang, 1987; Kramer et al., 2001, 2002). Thus while one can seldom avoid multicollinearity entirely, there is a great deal to be gained by minimizing its impact by appropriate research design and analysis.

Structural multicollinearity in regression models is well dealt with by centring (Glantz and Slinker, 2001). That is, if independent variables are, for example X, X^2, X^3 . . . – a polynomial regression – multicollinearity will be reduced if X is centred at its median before powers are computed (another example of a situation in which centring matters even in the absence of interactions). Also, with interactions in the model, say X_1, X_2, and their product X_1X_2, the effects of multicollinearity (especially loss of power) are likely to be reduced if both X_1 and X_2 are centred at their medians. This effect can be seen by noting in the tables that the standard error of the treatment effect is always least (and thus power greatest) for analyses that are centred.

In considering centring issues, one might choose to use a model based on independent variables that both reduces multicollinearity and increases informativeness of the results. For example, instead of using eight sites as a single categorical independent variable in Table 3, we might have chosen to code each site on three binary independent variables describing the sites: high/low SES, high/low representation of bilingual families, Northeast/ other geographical location (all coded +1/2). Then one could examine the effect within this sample of these particular site factors and their interactions. If there were site-by-treatment interactions, one might have more specific information about the source of such interactions. If this had been done with centring, the treatment effect would be exactly the same as that in the centred analysis of Table 3. Here none of the interaction effects would have been found statistically significant. However, the estimated three-way interactive effect of treatment-by-SES-by-bilingual sites would be 9.4 points on the IQ scale, almost the same magnitude as the overall treatment effect of 9.6. This provides a warning for future studies or policy decisions that there may be (as yet unproven) effects both of SES and bilingualism on treatment effect, and that the effect of site bilingualism may be different depending on whether the site was high SES (little effect as at Harvard) or low SES (large effect as at Miami).

How to centre using SPSS or SAS

We have emphasized that it is a quick and easy matter to centre, going so far as to compare it with routine hand washing. Most researchers today use statistical packages such as SPSS or SAS to process the data, where non-centring is the default. It would be preferable if such programs offered the default centring we here propose, and options to centre otherwise when indicated. This would remind users of the issue of centring with each regression analysis.

However, to show how quick and easy it is even now, once the idea is mastered, let us discuss how to convert data from uncentred to centred using either SPSS or SAS. In SPSS, the analyst can use either the compute or recode options, both under the transform menu. Binary variables can be recoded as –1/2 and +1/2 using the recode option, which prompts the analyst for 'old and new values'. For ordinal independent variables, one could simply compute the median and then create deviation scores using the compute function. Similarly, in SAS, medians can be generated using the windows selection for descriptives for independent variables of interest (for example, X1 and X2). Using a statement for DATA CENTER and specifying that X1 and X2 be subtracted from the mean 0 (or other designated reference point for centring), a new file is created that contains the median-centred data (for example, C_X1 and C_X2). It is imperative that interaction terms are computed from the centred rather than the original uncentred data. Once data are centred, the standard multiple regression model can be run in SPSS by writing a REGRESSION statement (Method = Enter) or by clicking on the linear regression option under the 'analyze' menu. In SAS, linear models are run using PROC REG or PROC GLM.

Summary and discussion

The above comments about centring generalize to all linear regression analyses – including logistic, and Cox proportional hazards models – and are often pertinent to use of non-linear regressions as well. Only two independent variables were used in the illustrations, but the principles and recommendations apply to as many independent variables as are used.

The crucial point is that every regression coefficient in a linear model reflects the effect of the associated independent variable on the dependent variable when all other independent variables in the models are equal to their centred value. Only if higher order interactions are zero in the population (not merely in the model) does it reflect that effect more generally.

To date, the most common approach to regression analyses has often been simply to submit the data, however entered, to the computer program and accept whatever results – uncentred analyses. How the data were entered is not usually reported in a research report, so readers may easily misinterpret results being reported. Not centring, after all, represents a *de facto* decision that all ordinal variables be centred at zero, that all binary and categorical independent variables be coded somewhat arbitrarily, 1 and 0, and that one category, also often arbitrarily chosen, be used as the reference category. As we have shown, this can lead to serious errors of statistical inference. Clearly the optimal approach would be to seek expert statistical advice in running each regression analysis, discussing what each coefficient means and how it corresponds to the research questions the study is meant to address. In the absence of such expert statistical guidance, we recommend the following alternative default approach:

- Every binary independent variable should be coded +1/2 and –1/2.
- Each ordinal independent variable should be centred with X* the median response.
- Categorical independent variables should be 'dummy coded' as usual, but instead of coding each response as +1 and 0, it is recommended it be coded 1 – 1/m and –1/m, where m is the number of categories. As in the usual situation, one categorical 'dummy' is omitted, but with the proposed centring it doesn't matter which one.

This would allow users of regression analyses to follow Cronbach's advice on centring with minimal effort. Since, under this proposal, one would always have to centre, researchers would not be tempted to exclude interactions from the model merely to ensure interpretability of the simple effects, or to ignore regression coefficients (like the intercept or simple effects in the presence of interactions) that are often informative. Since they would centre in any case, researchers might also be encouraged to consider what different ways of structuring their models best relate to their specific research questions, and how best to deal with multicollinearity, and they would be reminded to state how the data were entered in the regression analysis in publications. In this way, the overall quality of research findings could be improved for very little cost and effort.

Good statistical advice deals well with the target question for which advice is sought. Excellent statistical advice not only does that, but deals well with issues beyond the target question, and does not solve one problem only to spawn more serious problems elsewhere. In this sense, Cronbach's advice long ago on the issue of centring has proven to be excellent statistical advice.

First, Cronbach's advice draws a clear and necessary distinction between data analysis ('what went on in the sample?') and statistical inference ('what do we learn from the data analysis about the population represented by the sample?'). Centring and issues related to inclusion and exclusion of interactions matter little to data analysis, but are very pertinent to statistical inference. This explains why many data analysts remain unaware of the issues of centring in regression analyses.

Second, Cronbach's advice clearly acknowledges that statistical inference based on a mathematical model (such as linear regression) is always conditional on the truth of the assumed model. When one uses any linear regression model, one makes certain assumptions that, if false, may well make any conclusions based on the model false as well. Consequently, in regression models, one should be very reluctant to set any regression coefficient equal to zero when there is rationale and justification to believe that effect exists in the population. The statistical advice that focuses on centring only in the presence of interactions, or suggests that main effects are not interpretable in the presence of interactions has, in our consulting experience, led to omission of known interactions for reasons of convenience, often resulting in misleading statistical inferences.

There is growing awareness of the importance of interactions. For example, in a RCT, the treatment effect is always an average of the treatment effects over the subjects in the population. There may be subgroups, identifiable by their characteristics at baseline (moderators of treatment), who may have treatment effects much larger or much smaller than the overall treatment effect[11, 12]. To identify such moderators carries clinical and policy importance. The process of identifying such moderators involve detection of interactions of baseline independent variables with treatment. Also, how or why a treatment "works"(mediators of treatment), i.e., the search for the mechanisms of treatments, involves detection of interactions between independent variables related to events or changes during treatment, for a treatment might 'work' by changing the function, as well as the level, of an intervening variable.

In risk research, too, there is a growing awareness of the importance of interactions. It is generally conceded, for example, that the causes of many disorders are complex. It is very unlikely that a single gene, a single toxic agent, a single organism, will ever be found to be the cause of heart disease, many cancers, or mental illnesses of any type, or that the effects of multiple causal factors are necessarily additive. Instead, for example, genes may moderate the effects of environmental risk factors, or certain gene expressions mediate the effect of other genes or environmental risk factors. Discovery of such interaction effects may help to identify complex causal pathways to disorders and thus may facilitate prevention of such disorders.

Third, the distinction between statistical significance and clinical/policy significance is clearly acknowledged in Cronbach's approach. The emphasis is on the unbiased estimation of population parameters (regression coefficients) that are meaningful in terms of the research hypotheses, not merely on statistical significance. Generally statistical significance indicates that the data are sufficient to indicate that something non-random is going on – not that what is going on is of clinical, policy or practical significance. At the same time, an effect that is not statistically significant does not 'prove' the null hypothesis. Instead, the implication is that the data (sample size, reliability of measurement, and so forth) were not sufficient to detect any deviation from randomness. In the first case, estimation of relevant population parameters allows the researchers to evaluate the clinical or policy significance of their statistically significant findings; in the second case, to evaluate whether the effect of their non-statistically significant findings seems large enough to pursue further in later better designed studies. In any case, the focus on appropriate effect size estimation that results from centring is more consistent with the current emphasis on evaluation of practical, clinical and policy significance.

Finally, it should be noted, that, in a peculiar sense, one has no choice as to whether to follow Cronbach's advice or not. When one chooses *not* to centre, that represents a *de facto* decision to centre all ordinal values at zero and to code binary and categorical variables in some way as 1 and 0. Readers of papers based on regression analyses should always be informed exactly what was done so that they might properly interpret the results that are presented. In short, requiring that centring always be done merely asks that what is done implicitly anyway be done explicitly and thoughtfully, which would promote better application and understanding of the results of regression analyses.

References

Aiken LS, West SG. Multiple Regression: Testing and Interpreting Interactions. Newbury Park CA: Sage Publications, 1991.

Appelbaum MI, Cramer EM. Some problems in the nonorthogonal analysis of variance. Psychological Bulletin. 1974; 81: 335–43.

Cohen J. Partialed products are interactions; partialed powers are curve components. Psychological Bulletin 1978; 85: 858–66.

Cohen J, Cohen P, West S, Aiken L. Applied Multiple Regression/correlation Analysis for the Behavioral Sciences. Hillsdale NJ: Lawrence Erlbaum Associates, 2003.

Cramer EM, Appelbaum MI. Nonorthogonal analysis of variance-once again. Psychological Bulletin 1980; 87: 51–7.

Flack VF, Chang PC. Frequency of selecting noise variables in subset regression analysis: a simulation study. The American Statistician 1987; 41: 84–6.

Glantz SA, Slinker BK. Primer of Applied Regression and Analysis of Variance. New York: McGraw-Hill, 2001.

IHDP. Infant Health and Development Progam: enhancing the outcomes of low birth weight, premature infants: a multisite randomized trial. Journal of the American Medical Association 1990; 263: 3035–42.

Kraemer HC, Stice E, Kazdin A, Kupfer D. How do risk factors work together to produce an outcome?

Mediators, moderators, and independent, overlapping and proxy risk factors. The American Journal of Psychiatry 2001; 158: 848–56.

Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry 2002; 59: 877–83.

Kromrey JD, Foster-Johnson L. Mean centering in moderated multiple regression: much ado about nothing. Educational and Psychological Measurement 1998; 58: 42–68.

McGee D, Reed DYK. The results of logistic analyses when the variables are highly correlated: an empirical example using diet and CHD incidence. Journal of Chronic Diseases 1984; 37(9): 713–19.

Correspondence: Christine M. Blasey, 401 Quarry Road, Stanford CA 94305, USA.
Email: cblasey@stanford.edu.