STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2865–2873
Published online 24 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3107

Scaling regression inputs by dividing by two standard deviations

Andrew Gelman*, †

Department of Statistics and Department of Political Science, Columbia University, New York, NY, U.S.A.

SUMMARY

Interpretation of regression coefficients is sensitive to the scale of the inputs. One method often used to place input variables on a common scale is to divide each numeric variable by its standard deviation. Here we propose dividing each numeric variable by *two* times its standard deviation, so that the generic comparison is with inputs equal to the mean ±1 standard deviation. The resulting coefficients are then directly comparable for untransformed binary predictors. We have implemented the procedure as a function in R. We illustrate the method with two simple analyses that are typical of applied modeling: a linear regression of data from the National Election Study and a multilevel logistic regression of data on the prevalence of rodents in New York City apartments. We recommend our rescaling as a default option—an improvement upon the usual approach of including variables in whatever way they are coded in the data file—so that the magnitudes of coefficients can be directly compared as a matter of routine statistical practice. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: generalized linear models; linear regression; logistic regression; standardization; z-score

1. RESCALING INPUT VARIABLES TO MAKE REGRESSION COEFFICIENTS MORE DIRECTLY INTERPRETABLE

1.1. Background

A common trick in applied regression is to 'standardize' each input variable by subtracting its mean and dividing by its standard deviation. Subtracting the mean typically improves the interpretation of main effects in the presence of interactions, and dividing by the standard deviation puts all predictors on a common scale. Each coefficient in this standardized model is the expected difference in the outcome, comparing units that differ by one standard deviation in an input variable with all other inputs fixed at their average values.

*Correspondence to: Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, NY, U.S.A.
†E-mail: gelman@stat.columbia.edu, URL: http://www.stat.columbia.edu/∼gelman

Contract/grant sponsor: National Science Foundation
Contract/grant sponsor: New York City Department of Health
Contract/grant sponsor: Applied Statistics Center at Columbia University

Received 22 March 2007
Accepted 12 September 2007

Standardizing can create its own problems. For example, Bring [1] notes the incompatibility of scaling the inputs based on their marginal distributions and then interpreting regression coefficients conditionally. King [2] points out that comparisons of rescaled coefficients across data sets are problematic, because changing the range of a predictor will change its rescaled coefficient even if the regression model itself is unchanged. Blalock [3] notes the challenges of comparing the magnitudes of coefficients, rescaled or not, within a single regression. Greenland *et al.* [4] discuss challenges in causal interpretations of standardized regression coefficients.

While recognizing that standardizing does not solve the problems of causal inference and comparison of the importance of regression coefficients, we do believe that an automatic default standardization procedure can be helpful as a routine tool for understanding regressions.

1.2. Methods used for standardizing regression inputs

We first consider some standardization methods used in statistics and quantitative social science and then discuss our proposed method, which is to scale each input variable by dividing by two times its standard deviation.

A regression of the logarithm of men's earnings on height (in inches) from a national survey [5] yields a slope of 0.024, or 0.00096 if height is measured in millimeters, or 1549 if height is measured in miles. The coefficient is difficult to use if the scale is poorly chosen. Linear rescaling of predictors does not change the t-statistics or p-values but can aid or hinder the interpretation of coefficients.

Existing options for scaling include

1. Using round numbers (for example, height in inches or centimeters, age in tens of years, or income in tens of thousands of dollars).
2. Specifying lower and upper comparison points (for example, comparing people who are 5′6″ and 6′ tall, or comparing a 30 year-old to a 60 year-old, or persons with incomes in the 25th and 75th percentiles).
3. Subtracting the mean of each input variable and dividing by its standard deviation. (Strictly speaking, subtracting the mean is not necessary, but this step allows main effects to be more easily interpreted in the presence of interactions.)
4. Transforming nonlinearly, for example, using the logarithm. This can be effective in many cases but cannot be used automatically, for example, with variables that can have zero or negative values, or measurements such as Likert scales for which log transformations are typically inappropriate even if the variable is coded to be positive.

Each of these approaches has its strengths but also weaknesses. Rescaling using round numbers or comparison points is difficult to do automatically since additional information must be supplied. Logarithms and other nonlinear transformations should certainly be considered for many examples but, as noted above, they are inappropriate for many social science variables. Finally, dividing by the standard deviation is a convenient automatic method but leads to systematic problems in interpretation, as we discuss next.

1.3. Using binary inputs as a benchmark for rescaling

We shall understand rescaling by considering binary inputs—that is, variables x that can take on the value 0 or 1. At first this might seem silly, since the coefficient of a binary variable is directly interpretable as the comparison of the 0's to the 1's (with all other inputs held constant). But this is our point: we want to use this benchmark to interpret standardized coefficients more broadly.

A binary variable with equal probabilities has mean 0.5 and standard deviation 0.5. The usual standardized predictor (scaled by one standard deviation) then takes on the values ±1, and a 1-unit difference on this transformed scale corresponds to a difference of 0.5 on the original variable (for example, a comparison between $x=0.25$ and 0.75), which cannot be directly interpreted. To think of this another way, consider a regression with some binary predictors (for example, a male/female indicator) left intact, and some continuous predictors (for example, height) scaled by dividing by one standard deviation. The coefficients for the binary predictors correspond to a comparison of $x=0$ to $x=1$, or two standard deviations.

For these reasons, we recommend the general practice of scaling numeric inputs by dividing by *two* standard deviations, which allows the coefficients to be interpreted in the same way as with binary inputs. (We leave binary inputs unscaled because their coefficients can already be interpreted directly.)

To perform the rescaling automatically, we wrote a function `standardize()` as part of the `arm` package for applied regression and multilevel models in R [6, 7] to take arbitrary regression models and re-fit using standardized inputs. A key step in setting up this function is to identify the input variables. (The set of *input variables* is not, in general, the same as the set of *predictors*. For example, in a regression of earnings on height, sex, and their interaction, there are four predictors (the constant term, height, sex, and height × sex), but just two inputs: height and sex.) The function then transforms the inputs as specified and feeds them into the regression model. Input variables can be included in a regression nonlinearly or through interactions, and so it is not enough to fit the model and rescale the coefficients; the fitting procedure must be applied to the rescaled data.

Our procedure scales inputs to be comparable with binary variables that are roughly symmetric: if the probability falls between 0.3 and 0.7, then 2 standard deviations will be between 0.9 and 1. Highly skewed binary inputs still create difficulty in interpretation, however; for example, two standard deviations for a 90 per cent/10 per cent binary variable come to only 0.6. Thus, leaving this binary variable unscaled is not quite equivalent to dividing by two standard deviations. One might argue, however, that when considering rare subsets of the population, a full comparison from 0 to 1 could overstate the importance of the predictor in the regression, hence it might be reasonable to consider this two-standard-deviation comparison, which is less than the comparison of the extremes. Our main point, however, is that 2 standard deviations is a more reasonable scaling than 1—even if neither automatic approach solves all problems of interpretation.

2. EXAMPLES

2.1. Linear regression for party identification

We illustrate rescaling with a regression of party identification on sex, ethnicity, age, education, income, political ideology, and parents' party identification, using data from the National Election Study 1992 pre-election poll [8]. This example is intended to represent the sort of descriptive model fitting that is common in applied statistics, in which the researcher is interested in the contributions made by different variables in predicting some outcome of interest. This is also a good example to illustrate the method because our model includes binary, discrete numeric, and continuous numeric inputs, as well as nonlinearity for the age predictor and an interaction of income and ideology (an interaction that is of current interest in American politics [9, 10]).

DOI: 10.1002/sim

Variable	2 sd's
Female (1=female, 0=male)	1.0
Black (1=African American, 0=other)	0.7
Age (years)	34
Education (1=less than high school, …, 4=college graduate)	1.9
Income (1-5 scale)	2.2
Political ideology (1=very liberal, …, 7=very conservative)	2.9
Party identification (1=strong Democrat, …, 7=strong Republican)	4.2
Parents' party id (2=both Democrats, …, 6=both Republicans)	3.4
Binary (p=0.5)	1
Continuous uniform (-5, 5)	5.8
Continuous uniform (0, 100)	58
Discrete uniform (1, 2, 3, 4, 5)	2.8
Poisson (1)	2
Poisson (10)	6.3

Figure 1. The two-standard-deviation scale for some variables from the 1992 National Election Study. At the bottom of the table are some theoretical distributions for comparison.

Figure 1 lists the variables in the model, along with the scaling factor for each. Two standard deviations typically cover a wide range of the data, so the standardized coefficients, as we compute them, represent a comparison from low to high for each input.

Figure 2 shows a fitted regression, followed by a standardized regression, in which each numeric input has been mean centered and divided by two standard deviations. The binary inputs have been simply shifted to have mean zero but have not been rescaled. The coefficients in the new model can be more easily interpreted since they correspond to two-standard-deviation changes (roughly, from the low to the high end) of each numeric input, or the difference between the two conditions for binary inputs. The centering also improves the interpretation of the main effects of income and ideology in the presence of their interaction. The residual standard deviation and explained variance do not change under this linear reparameterization, but the coefficients become more comparable with each other. Most notably, on the raw scale, the coefficient for `black` is much larger (in absolute scale) than the coefficients for `parents.party` and for the `income:ideology`; after rescaling, however, this has changed dramatically.

An experienced practitioner might realize immediately the difficulty of interpreting the coefficients in the unscaled regression at the top of Figure 2; standardizing formalizes these intuitions and performs the computations automatically.

2.2. Multilevel logistic regression for prevalence of rodents

As a second example, we fit a multilevel logistic regression to predict the occurrence of rodents in New York City apartments, given physical factors (a count of defects in the apartment, its level above ground), social factors (a measure of the residents' poverty, indicators for ethnic groups), and geography (indicators for 55 city neighborhoods). The multilevel model includes ethnicity and neighborhood indicators as non-nested factors, each with its own group-level variance.

Figure 3 shows a possible display of the results, first using the parameterization in the raw data and then with standardized predictors. In the reparameterization, the varying coefficients are summarized by two standard deviations to be comparable with the numeric inputs. As with the

```
> M1 <- lm (partyid ~ female + black + age + I(age^2) + parents.party +
    education + income + ideology + income:ideology)
> display (M1)
lm(formula = partyid ~ female + black + age + I(age^2) + parents.party +
    education + income + ideology + income:ideology)
                coef.est coef.se
(Intercept)      0.99     0.64
female          -0.08     0.10
black           -0.98     0.17
age             -0.03     0.02
I(age^2)         0.00     0.00
parents.party    0.49     0.03
education        0.18     0.06
income          -0.43     0.15
ideology         0.20     0.11
income:ideology  0.15     0.03
  n = 989, k = 10
  residual sd = 1.58, R-Squared = 0.49
```

(a)

```
> M2 <- standardize (M1)
> display (M2)
lm(formula = partyid ~ c.female + c.black + z.age + I(z.age^2) +
    z.parents.party + z.education + z.income + z.ideology + z.income:z.ideology)
                    coef.est coef.se
(Intercept)          3.54     0.08
c.female            -0.08     0.10
c.black             -0.98     0.17
z.age               -0.15     0.12
I(z.age^2)           0.34     0.22
z.parents.party      1.66     0.11
z.education          0.34     0.12
z.income             0.41     0.12
z.ideology           1.84     0.10
z.income:z.ideology  0.94     0.22
  n = 989, k = 10
  residual sd = 1.58, R-Squared = 0.49
```

(b)

Figure 2. (a) A linear regression fit in R of individual party identification on several predictors (see Figure 1 for descriptions). The coefficients are difficult to interpret because different predictors are on different scales. The notation 'I(age^2)' represents age squared, and 'income:ideology' represents the interaction (that is, the product) of 'income' and 'ideology'. (b) The model fit to transformed inputs: the binary variables ('female' and 'black') have been centered by subtracting their mean in the data, and the numeric variables have been rescaled by subtracting the mean and dividing by two standard deviations. The new coefficients reflect the different scales. For example, the coefficient for the interaction of income and ideology is now higher than the coefficient for race. We display raw computer output in both cases to illustrate how these summaries are used in routine practice.

previous example, the standardized coefficients are directly comparable in a way that the raw coefficients are not. Most notably, the coefficients for the continuous predictors have all increased in absolute value to reflect the variation in these predictors in the data. The figure also illustrates that the results can easily be displayed on both scales for the convenience of the user.

Predictor	coef (s.e.)	standardized coef (s.e.)
(Intercept)	−2.25 (0.34)	−1.43 (0.27)
defects	0.49 (0.05)	1.47 (0.14)
poverty	0.12 (0.05)	0.37 (0.16)
floor	−0.01 (0.04)	−0.04 (0.16)
hispanic	0.51 (0.15)	0.51 (0.15)
black	0.36 (0.16)	0.36 (0.16)
asian	−0.17 (0.24)	−0.17 (0.24)
white	−0.56 (0.16)	−0.56 (0.16)
$\hat{\sigma}_{\text{ethnicity}}$	0.65	
$2\hat{\sigma}_{\text{ethnicity}}$		1.30
$\hat{\sigma}_{\text{neigbborhood}}$	0.47	
$2\hat{\sigma}_{\text{neigbborhood}}$		0.94

Figure 3. Multilevel logistic regression model predicting the occurrence of rodents in city apartments, given numeric predictors (representing physical defects in the apartment, poverty of the occupants, the floor of residence) and indicators for ethnicity and neighborhood. The two columns show the summaries using the direct and reparameterized input variables. The rescaled coefficients are directly interpretable as changes on the logit scale comparing each input variable at a low value to a high value: for the numeric predictors, this is the mean ±1 standard deviation, and for the indicators, this is each level compared with the mean.

3. DISCUSSION

3.1. Options in the rescaling of inputs

We are rescaling the *input variables*, not the *predictors*. For example, `age` is rescaled to `z.age`, and the new model includes `z.age` and its square as predictors. The 'age-squared' predictor is not itself standardized. Similarly, we standardize income and ideology, and interact these standardized inputs; we do not directly standardize the income × ideology interaction.

In Figure 2 we have used the default standardization (as can be seen in the function call `standardize (M1)`, which does not specify any options). Other choices are possible. For example, we might want to transform the outcome (`partyid`) as well, which can be done using the command,

```
M3 <- standardize (M1, standardize.y=TRUE)
```

Or we might want to leave the variable `black` unchanged (that is, on its original scale):

```
M4 <- standardize (M1, unchanged="black")
```

These options can also be combined; for example,

```
M5 <- standardize (M1, standardize.y=TRUE, unchanged=c("female","black"))
```

Finally, we could choose to rescale the binary inputs also; for example,

```
M6 <- standardize (M1, binary.inputs="full")
```

which rescales all the inputs, including the binary variables, `female` and `black`, by subtracting the mean and dividing by two standard deviations.

It is also helpful to have options when considering predictors on the logarithmic scale, in which case a change of 1 in a predictor corresponds to multiplying by a factor of $e=2.7\ldots$ (for the natural log) or 10 (for log base 10). We certainly do not want to subtract the mean and rescale an input variable before it has been log transformed. When inputs and outcome variables are on the log scale, the coefficients have the interpretation as 'elasticities' (relative change in y per relative change in x), and, again, rescaling would just muddy this clear picture. More challenging cases arise in which some inputs have been log transformed and others are not. We have no general solution here, but we would start by centering and rescaling the variables that have not been log transformed. It might also make sense to rescale the variables *after* the log transformation—for example, in Figure 1, if income had been coded as 'log (income in dollars),' we might still consider transforming it.

3.2. *Variance components and multilevel models*

To be consistent with the interpretation of coefficients as corresponding to a typical comparison for an input variable (0 to 1 for a binary input, or the mean ± 1 standard deviation for a numeric input), it makes sense to summarize variance parameters by twice their standard deviation. For example, fitting a multilevel version of the model shown in Figure 2, in which the intercepts vary by state, yields an estimated standard deviation of 1.57 for the individual-level errors and 0.16 for the state-level errors. To compare with the scaled regression coefficients, it would make sense to double them—thus, summarizing the scale of individual- and group-level variation by 3.14 and 0.32, respectively. This is slightly awkward but allows direct comparisons with the coefficients for binary predictors, which we believe is the most fundamental standard of reference.

More generally, a set of varying coefficients (random effects) can be considered as a single numerical predictor with latent (unobserved) continuous values. For example, the model in Figure 3 can be viewed as having a single continuous 'ethnicity' predictor that takes on the (estimated) values 0.51, 0.36, -0.17, or -0.56, depending on whether the respondent is hispanic, black, etc. As defined in this way, this predictor has a coefficient of 1 in the regression model, by definition, and its standardized coefficient is simply twice the standard deviation of the possible values it attains, which is approximately $2\hat{\sigma}_{\text{ethnicity}} = 1.30$ in this case.

4. CONCLUSIONS

Rescaling numeric regression inputs by dividing by two standard deviations is a reasonable automatic procedure that avoids conventional standardization's incompatibility with binary inputs. Standardizing does not solve issues of causality [4], conditioning [1], or comparison between fits with different data sets [2], but we believe it usefully contributes to the goal of understanding a model whose predictors are on different scales.

It can be a challenge to pick appropriate 'round numbers' for scaling regression predictors, and standardization, as we have defined it here, gives a general solution which is, at the very least, an interpretable starting point. We recommend it as an automatic adjunct to displaying coefficients on the original scale.

This does not stop us from keeping variables on some standard, well-understood scale (for example, in predicting election outcomes given unemployment rate, coefficients can be interpreted as percentage points of vote per percentage point change in unemployment), but we would use our standardization as a starting point. In general, we believe that our recommendations will generally lead to more understandable inferences than the current default, which is typically to include variables; however, they happen to have been coded in the data file. Our goal is for regression coefficients to be interpretable as changes from low to high values (for binary inputs or numeric inputs that have been scaled by two standard deviations).

DOI: 10.1002/sim

We also center each input variable to have a mean of zero so that interactions are more interpretable. Again, in some applications it can make sense for variables to be centered around some particular baseline value, but we believe our automatic procedure is better than the current default of using whatever value happens to be zero on the scale of the data, which all too commonly results in absurdities such as age = 0 years or party identification = 0 on a 1–7 scale. Even with such scaling, the correct interpretation of the model can be untangled from the regression by pulling out the right combination of coefficients (for example, evaluating interactions at different plausible values of age such as 20, 40, and 60); the advantage of our procedure is that the default outputs in the regression table can be compared and understood in a consistent way.

We also hope that these ideas could also be applied to predictive comparisons for logistic regression and other nonlinear models [11], and beyond that to multilevel models and nonlinear procedures such as generalized additive models [12]. Nonlinear models can best be summarized graphically, either compactly through summary methods such as graphs of coefficient estimates or nomograms [13–15], showing the (perhaps nonlinear) relationship between the expected outcome as each input is varied. But to the extent that numerical summaries are useful—and they certainly will be used—we would recommend, as a default starting point, evaluating at the mean ±1 standard deviation of each input variable. For linear models this reduces to the scaling presented in this paper.

Finally, one might dismiss the ideas in this paper with the claim that users of regressions should understand their predictors well enough to interpret all coefficients. Our response is that regression analysis is used routinely enough that it is useful to have a routine method of scaling. For example, scanning through recent issues of two leading journals in medicine and one in economics, we found

- Table 5 of Itani *et al.* [16], which reports odds ratios (exponentiated logistic regression coefficients) for a large set of predictors. Most of the predictors are binary or were dichotomized, with a few numeric predictors remaining, which were rescaled by dividing by one standard deviation. As argued in this paper, dividing by one (rather than two) standard deviation will lead the user to understate the importance of these continuous inputs.
- Table 2 of Murray *et al.* [17], which reports linear regression coefficients for log income and latitude; the latter has a wide range in the data set and so unsurprisingly has a coefficient estimate that is very small on the absolute scale.
- Table 4 of Adda and Cornaglia [18], which reports linear regression coefficients for some binary predictors and some numerical predictors. Unsurprisingly, the coefficients for predictors such as age and education (years), house size (number of bedrooms), and family size are much smaller in magnitude than those for indicators for sex, ethnicity, church attendance, and marital status.

We bring up these examples not to criticize these papers or their journals, but to point out that, even in the most professional applied work, standard practice yields coefficients for numeric predictors that are hard to interpret. Our proposal is a direct approach to improving this interpretability.

ACKNOWLEDGEMENTS

We thank Dimitris Rizopoulos and Gabor Grothendieck for help with R programming, Wendy McKelvey for the rodents example, Aleks Jakulin, Joe Bafumi, David Park, Hal Stern, Tobias Verbeke, John Londregan, Jeff Gill, Suzanna De Boef, Justin Gross, and two anonymous reviewers for comments, and the National Science Foundation, the New York City Department of Health, and the Applied Statistics Center at Columbia University for financial support.

REFERENCES

1. Bring J. How to standardize regression coefficients. *The American Statistician* 1994; **48**:209–213.
2. King G. How not to lie with statistics: avoiding common mistakes in quantitative political science. *American Journal of Political Science* 1986; **30**:666–687.
3. Blalock HM. Evaluating the relative importance of variables. *American Sociological Review* 1961; **26**:866–874.
4. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology* 1986; **123**:203–208.
5. Ross CE. *Work*, *Family*, *and Well-being in the United States*, 1990. Survey data available from Inter-university Consortium for Political and Social Research, Ann Arbor, MI.
6. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2007. www.R-project.org [accessed 23 August 2007].
7. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York, 2007.
8. Miller WE, Kinder DR, Rosenstone SJ. *American National Election Study*, 1992. Survey data available from Inter-university Consortium for Political and Social Research, Ann Arbor, MI.
9. McCarty N, Poole KT, Rosenthal H. *Polarized America*: *The Dance of Political Ideology and Unequal Riches*. MIT Press: Cambridge, MA, 2006.
10. Gelman A, Shor B, Bafumi J, Park D. Rich state, poor state, red state, blue state: what's the matter with Connecticut? *Quarterly Journal of Political Science* 2007, under revision.
11. Gelman A, Pardoe I. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology* 2007, in press.
12. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: New York, 1990.
13. Lubsen J, Pool J, van der Does E. A practical device for the application of a diagnostic or prognostic function. *Methods of Information in Medicine* 1978; **17**:127–129.
14. Harrell FE. *Regression Modeling Strategies*. Springer: New York, 2001.
15. Jakulin A, Mozina M, Demsar J, Bratko I, Zupan B. Nomograms for visualizing support vector machines. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Association for Computing Machinery: New York, 2005; 108–117.
16. Itani KMF, Wilson SE, Awad SS, Jensen EH, Finn TS, Abramson MA. Ertapenem versus cefotetan prophylaxis in elective colorectal surgery. *New England Journal of Medicine* 2006; **355**:2640–2651.
17. Murray CJL, Lopez AD, Chin B, Feehan D, Hill KH. Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis. *The Lancet* 2006; **368**:2211–2218.
18. Adda J, Cornaglia F. Taxes, cigarette consumption, and smoking intensity. *American Economic Review* 2006; **96**: 1013–1028.

DOI: 10.1002/sim