

Splitting a predictor at the upper quarter or third and the lower quarter or third*

Andrew Gelman[†] David K. Park[‡]

July 31, 2007

Abstract

A linear regression of y on x can be approximated by a simple difference: the average values of y corresponding to the highest quarter or third of x , minus the average values of y corresponding to the lowest quarter or third of x . A simple theoretical analysis shows this comparison performs reasonably well, with 80%–90% efficiency compared to the linear regression if the predictor is uniformly or normally distributed. Discretizing x into three categories claws back about half the efficiency lost by the commonly-used strategy of dichotomizing the predictor.

We illustrate with the example that motivated this research: an analysis of income and voting which we had originally performed for a scholarly journal but then wanted to communicate to a general audience.

Keywords: discretization, linear regression, statistical communication, trichotomizing

1 Introduction

Linear regression is perhaps the most commonly used tool in statistics and as such is familiar to a diverse range of students and researchers. But an even wider segment of the educated public does not understand regression or least squares and thus has to take many statistical results on faith, essentially accepting results that are labeled as statistically significant without being able to interpret their numerical values.

We would like to approximate the regression of y on x by a simple comparison. This interpretation is immediate for binary predictors, but more generally one can simplify the interpretation of a regression by discretizing. In common practice, variables are discretized into two categories (that is, the predictor x falling above or below some threshold).

However, as we show here, we can do better by discretizing x into three values and throwing away the middle category, thus comparing the average value of y for x in the

*We thank Boris Shor and Joseph Bafumi for collaboration with the original example, David Dunson, Ian McKeague, and John Carlin for helpful comments, and the National Science Foundation, the National Institutes of Health, and the Applied Statistics Center at Columbia University for financial support.

[†]Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/~gelman

[‡]Department of Political Science, George Washington University

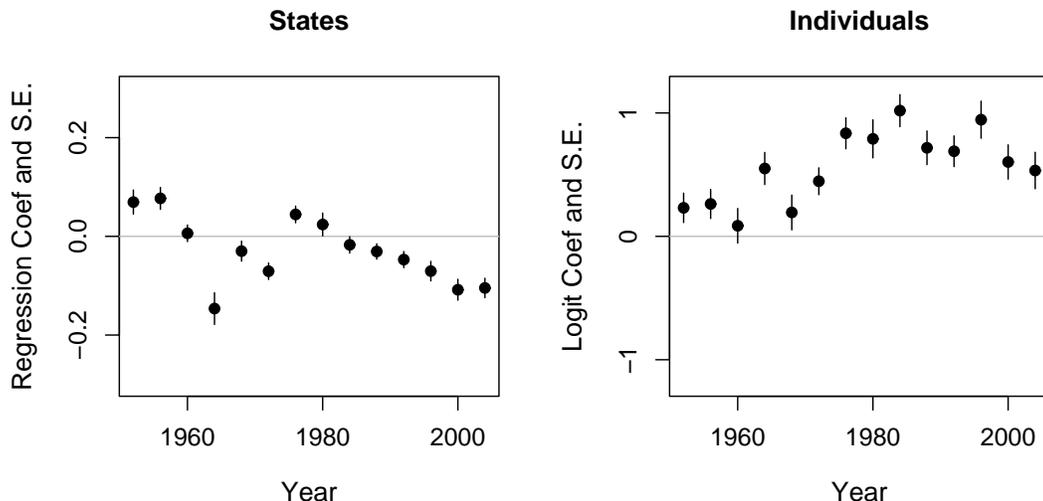


Figure 1: (left plot) Coefficients (± 1 standard error) for average state income in a regression predicting Republican vote share by state in a series of U.S. presidential elections. The model was fit separately for each election year. The negative coefficients in recent years indicate that richer *states* now tend to support the Democrats.

(right plot) Coefficients for income in logistic regressions of Republican vote, fit to individual survey data from each election year. The positive coefficients indicate that richer *voters* continue to support the Republicans.

These plots are clear to statistically sophisticated readers, but we would like more transparent data summaries for a general audience.

high category to the average value of y for x in the low category. After a study of the efficiency of this comparison, we make the general recommendation that the high and low categories each be set to contain 1/4 to 1/3 of the data, which results in comparisons with approximately 80% to 90% of the efficiency of linear regression if the predictor x follows a uniform or normal distribution.

A loss of 10% or 20% of efficiency is not minor, and so we do not recommend that the comparisons replace regressions but rather that they be considered as useful supplementary summaries, especially for the goal of communicating to a general audience.

1.1 Income and voting example

We shall illustrate with an example from our current research, a study of income and voting in United States presidential elections (Gelman et al., 2007). Figure 1 shows the graphs that begin that article; each displays a time series of estimated regression coefficients.

The first graph shows, for each election year, the coefficient of *average state income* on the Republican candidate's proportion of the vote in the state: in recent years, these

coefficients have become strongly negative, indicating that the Republicans are doing better in poor states than in rich states. This regression was estimated using election results and state-level income data.

The second graph shows coefficients for *individual income* from logistic regressions predicting individual vote (1 for Republicans, 0 for Democrats), estimated from national survey data from each election year. Here, the coefficients are positive, implying that the Republicans do better among rich voters than among poor voters.

We will not further discuss here the substantive concerns of our analyses (see, for example, Brooks and Brady, 1999, and McCarty, Poole, and Rosenthal, 2006, for more on the general topic of trends in income and voting in the United States), except to note that these results are of wide interest, not just to political scientists and election analysts, but also to the general public, which has been bombarded in recent elections with discussions of “red-state” and “blue-state” voters.

1.2 Goal of expressing regressions as comparisons that can be understood by the general reader

In order to present our results in a form that is understandable to a general audience, we would like to minimize the gap between the numerical results (for example, the regression coefficients shown in Figure 1) and the substantive conclusions (Republicans doing worse in rich states, and so forth). The goal is to bring the non-statistical reader closer to a direct engagement with our findings. Presenting regressions as simple differences is one step in this process.

Statisticians have come up with methods of summarizing logistic regressions and other nonlinear models using predictive comparisons (see Gelman and Pardoe, 2007), but even these summaries require an understanding of regression which is beyond many educated laypersons. For example, taking the difference between expected values of the outcome when a predictor is set to prechosen high or low values requires understanding the regression model itself. Correlations are another simple summary but, again, are not readily interpretable to the general reader.

At the other extreme, simple data summaries can be appealing—it is common to present electoral results as colored maps—but such displays are not structured enough for us, as they are awkward for understanding the relation between two variables (such as income and voting, in our example). Scatterplots are a good way of displaying the relation between variables, but it is also important to have numerical summaries, both for their own sake

and for comparisons such as the time series in Figure 1.

2 Method

2.1 Replacing a regression slope by a simple comparison of average values of y in the upper and lower quantiles of x

Consider a linear regression of y on x based on n data points, where the regression model is assumed to be true; thus, $y_i = \alpha + \beta x_i + \text{error}_i$, with errors that are normally distributed with equal variance and are independent of the predictor x . We shall compare the least-squares estimate $\hat{\beta}$ to a simple difference of the mean of data values y in the upper and lower quantiles of x .

More precisely, given a fraction f of data to be kept in the upper and lower range of x , we set thresholds x^{lower} and x^{upper} to be the $(fn)^{\text{th}}$ and $((1-f)n+1)^{\text{th}}$ order statistics of x in the data, respectively. The fraction f must be greater than 0 (so that at least some data are kept) and not exceed 0.5 (at which point we would be comparing the average values of y corresponding to the upper and lower half of x). We discretize the predictor based on the selected order statistics:

$$z = \begin{cases} -0.5 & \text{if } x \leq x^{\text{lower}} \\ 0 & \text{if } x^{\text{lower}} < x < x^{\text{upper}} \\ 0.5 & \text{if } x \geq x^{\text{upper}}. \end{cases} \quad (1)$$

We then summarize the linear relation of y given x by

$$\text{simple comparison: } \bar{y}_{z=0.5} - \bar{y}_{z=-0.5} = \frac{1}{fn} \left(\sum_{i: z_i=0.5} y_i - \sum_{i: z_i=-0.5} y_i \right) \quad (2)$$

in place of the estimated regression slope.

This comparison and the regression slope are not on the same scale, however, and so in comparing efficiencies we shall consider the ratio,

$$\hat{\beta}^{\text{simple}} = \frac{\bar{y}_{z=0.5} - \bar{y}_{z=-0.5}}{\bar{x}_{z=0.5} - \bar{x}_{z=-0.5}} = \frac{\sum_{i: z_i=0.5} y_i - \sum_{i: z_i=-0.5} y_i}{\sum_{i: z_i=0.5} x_i - \sum_{i: z_i=-0.5} x_i} \quad (3)$$

and compare this to the least-squares estimate. Both the comparison (2) and the ratio (3) depend through (1) on x^{lower} and x^{upper} , which themselves are functions of the fraction f of data kept in the upper and lower ranges of the data. Thus, we can determine the variance of the estimate $\hat{\beta}^{\text{simple}}$ as a function of f and optimize it (under various assumptions).

2.2 Identifying the estimated linear regression slope as a weighted average of all paired comparisons

Before getting to our main findings, we recall a simple algebraic identity that expresses the least-squares regression of y on x as a weighted average of all pairwise comparisons:

$$\hat{\beta}^{\text{ls}} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_{i,j} (y_i - y_j)(x_i - x_j)}{\sum_{i,j} (x_i - x_j)^2} = \frac{\sum_{i,j} \frac{y_i - y_j}{x_i - x_j} (x_i - x_j)^2}{\sum_{i,j} (x_i - x_j)^2}.$$

The estimated slope is thus equivalent to a weighted average of difference ratios, $\frac{y_i - y_j}{x_i - x_j}$, with each ratio weighted by $(x_i - x_j)^2$. This makes sense since the variance of a difference ratio is proportional to the squared difference of the predictors.

We shall not directly use this formula in our analysis, but it is a helpful starting point in reminding us that regressions can already be expressed as comparisons. Our goal here is to come up with a simpler and easier-to-understand difference of means which is still a reasonable approximation to the above expression.

2.3 Theoretical derivation of optimal fraction of data to keep

We shall work out the asymptotic variance of $\hat{\beta}^{\text{simple}}$ in (3) and check the results using simulations. Asymptotic results are fine here since we would not expect to be using this procedure with very small sample sizes. (For example, if $n = 7$, we would just display the regression along with all seven data points, if necessary. There would not be much of a point to comparing, for example, the mean of the values of y corresponding to the highest two values of x to the mean of the values of y corresponding to the lowest two values of y .)

The asymptotic variance of (3) is easily worked out using standard sampling-theory formulas for the variance of a ratio estimate (see, for example, Lohr, 1999):

$$\begin{aligned} \text{var}(\hat{\beta}^{\text{simple}}) &= \frac{\sigma^2}{n} \frac{\Pr(x \geq x^{\text{upper}}) + \Pr(x \leq x^{\text{lower}})}{(\mathbb{E}(x|x \geq x^{\text{upper}})\Pr(x \geq x^{\text{upper}}) - \mathbb{E}(x|x \leq x^{\text{lower}})\Pr(x \leq x^{\text{lower}}))^2} \\ &= \frac{\sigma^2}{n} \frac{2}{(\mathbb{E}(x|x \geq x^{\text{upper}}) - \mathbb{E}(x|x \leq x^{\text{lower}}))^2 f}, \end{aligned} \quad (4)$$

where f is the fraction of data kept at each extreme, and σ^2 is the residual variance of the regression of y on x . By comparison, the least-squares estimate has sampling variance

$$\text{var}(\hat{\beta}^{\text{ls}}) = \frac{\sigma^2}{n} \frac{1}{\text{var}(x)}. \quad (5)$$

The ratio of (5) to (4) is the efficiency of the simple comparison.

We shall determine the optimal fraction f by minimizing (4) for any particular distribution $p(x)$. It is most convenient to find the minimum by differentiating the logarithm of the variance:

$$\log \text{var}(\hat{\beta}^{\text{simple}}) = \log(2\sigma^2/n) + \log f - 2 \log \left(\int_{x^{\text{upper}}}^{\infty} xp(x)dx - \int_{-\infty}^{x^{\text{lower}}} xp(x)dx \right).$$

Differentiating with respect to f yields,

$$\frac{d}{df} \log \text{var}(\hat{\beta}^{\text{simple}}) = \frac{1}{f} - \frac{2}{f} \frac{x^{\text{upper}} - x^{\text{lower}}}{\text{E}(x|x \geq x^{\text{upper}}) - \text{E}(x|x \leq x^{\text{lower}})}. \quad (6)$$

Here we have used the chain rule when differentiating with respect to x^{upper} and x^{lower} , plugging in $df/d(x^{\text{upper}}) = -p(x^{\text{upper}})$ and $df/d(x^{\text{lower}}) = p(x^{\text{lower}})$.

Finally, setting the derivative (6) to zero and rearranging terms yields,

$$\text{at optimum } f: \frac{\text{E}(x|x \geq x^{\text{upper}}) - \text{E}(x|x \leq x^{\text{lower}})}{2(x^{\text{upper}} - x^{\text{lower}})} = 1. \quad (7)$$

2.4 Computation of the optimum

For any specific model, we can numerically solve (7) and thus compute the optimal f via simulation:

1. Simulate some large even number m (for example, 10,000) random draws from $p(x)$. Order these simulations from lowest to highest: $x_{(1)}, x_{(2)}, \dots, x_{(m)}$.
2. For each $f = \frac{1}{m}, \frac{2}{m}, \dots, \frac{m/2-1}{m}, \frac{1}{2}$, define $lower = fm$ and $upper = (1 - f)m + 1$ and then approximate the left side of (7) by

$$\frac{\frac{1}{fm} \sum_{i=upper}^m x_{(i)} - \frac{1}{fm} \sum_{i=1}^{lower} x_{(i)}}{2(x_{(upper)} - x_{(lower)})}. \quad (8)$$

3. The above expression should be less than 1 for small values of f and greater than 1 for large values of f . Compute the optimal f as that where the ratio (8) is closest to 1.

Figure 2 illustrates the simulation-based optimization for the uniform distribution, for which the optimal fraction f is $1/3$ (easily derived analytically) and the normal, whose optimal fraction is 0.27. As illustrated by these graphs, the curve of $\text{E}(x|x \geq x_{(upper)}) - \text{E}(x|x \leq x_{(lower)}) - 2(x_{(upper)} - x_{(lower)})$ will always cross zero, since this difference is negative at $f = 0$ (where the ratio (8) is 1) and positive at $f = 0.5$ (where $x_{(upper)} - x_{(lower)} = 0$). However, there can be some numerical instability for very heavy-tailed distributions, where extreme outliers can affect the calculation for small values of f .

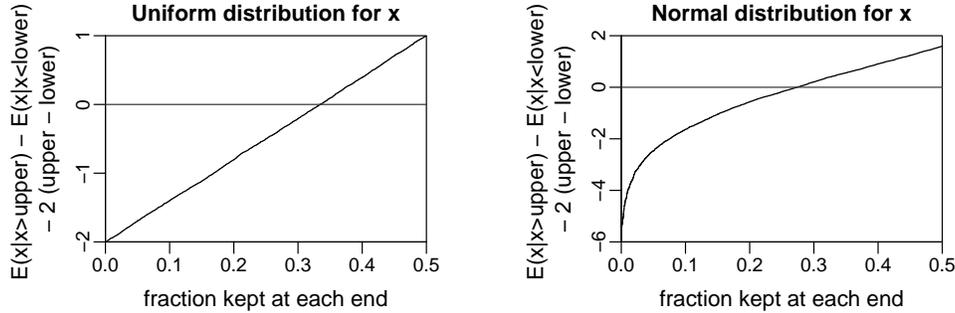


Figure 2: Results of computing the optimal fraction f for the uniform and normal distributions. For each model, we simulated $m = 10,000$ random draws and then, for each f between 0 and $1/2$, computed the difference in order statistics $x_{(upper)} - x_{(lower)}$ and the difference in expectations $E(x|x \geq x_{(upper)}) - E(x|x \leq x_{(lower)})$ as in (8). For each model, the horizontal line shows where the ratio of these equals 2, which comes at $f = 0.33$ when x is uniformly distributed and $f = 0.27$ when x is normally distributed.

2.5 Results for specific distributions

Having determined the optimal fraction to keep, it is helpful to simulate an example set of fake data from each of several models to see how the difference (2) compares to the regression line. The left column of Figure 3 displays a simple example for each of four models for x —two with short tails (the uniform and normal distributions) and two with long tails (the Laplace and t_4 distributions)—illustrating in each case the estimated regression line and the optimal comparison based on quantiles. The assumed distributions are symmetric, but data from any particular simulation will have some skewness, which is why the cutoff points for the quantiles are not exactly centered in the graphs.

The right column of Figure 3 shows the efficiencies of the comparisons under each of the assumed distributions for x (assuming large sample sizes, and assuming that the linear regression model is correct). For each model, we take our 10,000 simulations and compute the efficiency using the ratio of numerical estimates of (4) and (5) for each value of f .

These curves show that the fraction of data kept should not be too small or too large. A reasonable consensus value would appear to be $f = 0.25$, that is, comparing the upper and lower quartiles. However, if the distribution of the predictor is short-tailed (such as the uniform or normal), we might prefer $f = 0.33$, that is, comparing the upper and lower thirds of the data. Either of these simple rules would seem reasonable.

As can be seen from the right column of Figure 3, discretizing x into three categories claws back about half the efficiency lost by dichotomizing the predictor, while retaining the

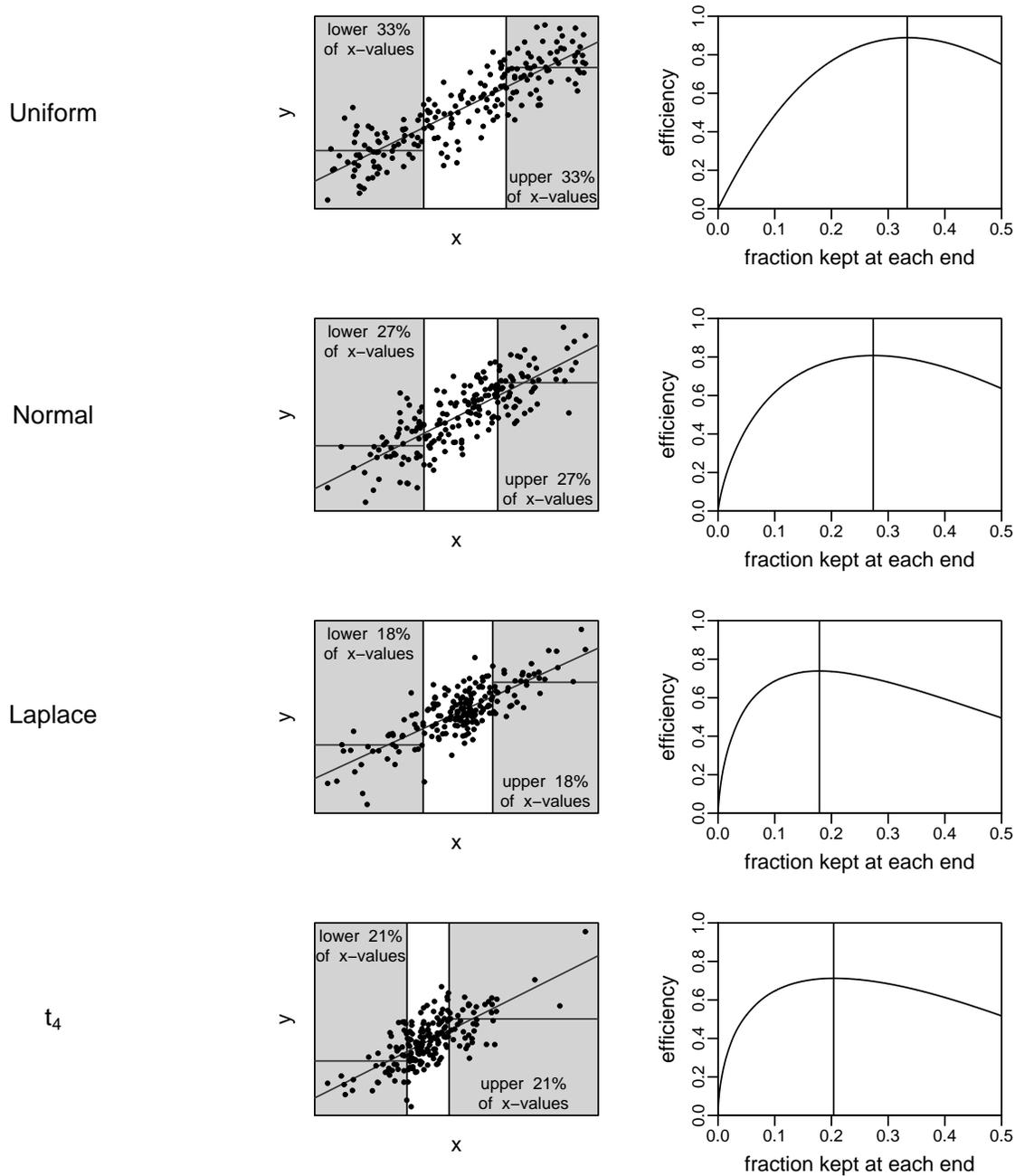


Figure 3: (left panel) Instances of simulated data from linear regression models where the predictor x is drawn from the uniform, normal, Laplace (folded-exponential), or t_4 distribution. Fitted regression lines and the optimal comparisons are shown.

(right panel) Efficiencies of comparisons (compared to linear regression), shown as a function of the fraction f kept at each end, so that $f \rightarrow 0$ corresponds to comparing the most extreme quantiles, and $f = 0.5$ corresponds to no trimming (i.e., comparing the upper half of the data to the lower half).

The optimal comparisons for the four scenarios have efficiencies of 89%, 81%, 74%, and 69%, respectively, compared to linear regression. (By comparison, simply dichotomizing x yields efficiencies of 75%, 63%, 49%, and 52%, respectively.)

simple interpretation as a high vs. low comparison.

2.6 Discrete predictors

We can use our simulation results to guide summaries for discrete predictors as well. If x takes on three values with approximately equal numbers of cases in each, we would compare the average values of y in the high and low categories of x (thus, $f = 0.33$); and if x takes on four approximately equally-populated values, we would again compare the highest and lowest categories (in this case, $f = 0.25$). If x takes on five equally-populated levels, we have the choice between comparing highest and lowest ($f = 0.2$), or the highest two versus the lowest two ($f = 0.4$). Based on the simulations, we would probably compare highest to lowest, which also has the advantage of a simpler interpretation. If the different levels have different numbers of cases, we recommend counting how many cases are in each category and aggregating to get approximately 1/4 to 1/3 of the data in the high and low categories.

2.7 Discrete outcomes

Logit and probit regressions can similarly be replaced by the difference of the proportion of successes in the high and low categories. This is a comparison of $\Pr(y = 1)$ or, equivalently, $E(y)$, so we can use the same comparison (2) as before. Compared to continuous data, binary data represent an even clearer candidates for simple comparisons, since logit and probit coefficients are themselves difficult to interpret on the scale of the data (see, for example, Gelman and Pardoe, 2007). Because of the nonlinearity of the model, however, it is not possible to work out the relative efficiency of the simple comparison as in Section 2.3—there is not a single parameter that the difference and the regression coefficient can both be considered to be estimating. One option is to compare the difference to the corresponding predicted difference, $E(y|x \geq x^{upper}) - E(y|x \leq x^{lower})$, with expectations evaluated under the logit or probit regression model and averaging over the empirical distribution of x in the data.

One could similarly summarize ordered logit or probit models by differences, but we do not generally recommend this approach when there is a risk of discarding information on non-monotonic patterns (for example, the frequency of a category in the middle of the scale that increases, then decreases, as a function of x). At some point when a model becomes complicated enough, you just have to bite the bullet and figure out how to summarize it, ideally graphically and then with numerical summaries that can be illustrated in an example graph and then be used in further comparisons.

2.8 Multiple regression

So far we have considered regression with a single predictor. Various extensions are possible to multiple regression. With two input variables, we can simply discretize each of them into three values as in (1) and then report differences for each variable, holding the other constant. With more than two, the best choice perhaps is to discretize the inputs of interest, then run a regression and express the estimated regression coefficients as differences between the upper and lower quartiles. (This is why we set the values of z to 0.5 and -0.5 , rather than 1 and -1 , in defining the discretized variable in (1), so that a regression coefficient on z corresponds to a change from the lower to the upper zone. See Gelman, 2007, for more on this issue.) Variables are often discretized before entering them into multiple regressions, so it is a small step to use three categories rather than two.

Another way to look at this is that, with a single predictor x , the simple difference (2) is also the estimated coefficient regressing y on the discretized predictor z defined in (1). Thus, if we add further predictors to the model, we can interpret the coefficient for this particular z as the average difference between high and low quantiles, after controlling for the other variables.

A useful point of comparison is to the common practice of dichotomizing predictors. Compared to dichotomizing, using three categories preserves more information (as shown in Section 2.5, regaining about half the information lost by dichotomizing) while preserving the simple interpretation as a comparison of high to low values. So, if regression inputs are to be discretized, we recommend three categories rather than two. Another option, as always, is to fit the full model with continuous predictors and then devote some effort into explaining the model and the coefficients.

3 Example

3.1 Income and voting

Returning to the example of Section 1.1, we redo Figure 1, this time comparing the average proportion of Republican vote for states in the upper and lower thirds of income, then comparing the proportion of Republican voters among voters in the upper and lower thirds of income. Figure 4 shows the results: the graphs look similar to those in Figure 1, but the numbers are much more directly understood and can be explained without reference to regression, correlation, or any statistical method more complicated than averaging. We calculate standard errors here just using the simple formula for a difference in means.

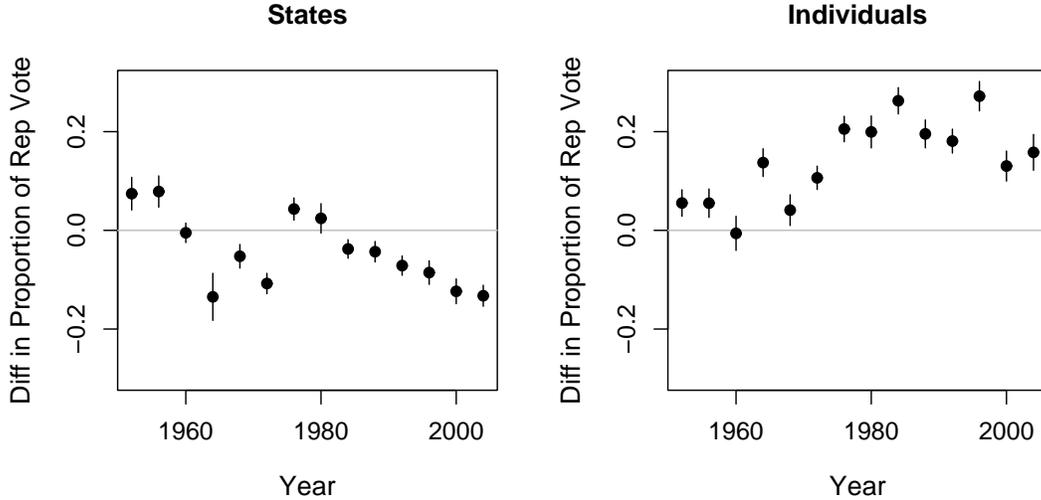


Figure 4: (left plot) For each presidential election year, difference in Republican vote share (± 1 standard error), comparing states in the upper third of income to the states in the lower third.

(right plot) For each year, difference in proportion of Republican vote, comparing voters in the upper third of income to voters in the lower third.

Compare to Figure 1, which shows similar results using regression coefficients. The results shown here can be interpreted more directly without reference to regression models.

In addition, the two analyses—continuous data at the state level and binary data at the individual level—can be interpreted on the common scale of vote proportions. By contrast, the linear and logistic regressions of Figure 1 are on different scales. They can be put on the same scale—quickly by dividing the logistic regression coefficients by 4, or more precisely by computing expected predictive differences—but that would represent another level of effort and explanation.

3.2 Income, religious attendance, and voting

We illustrate how our method can handle a second input variable by considering how religiosity as well as income predicts vote choice. The correlation of religious attendance with Republican voting in recent years is well known (see, for example, Glaeser and Ward, 2006), but it is not so well understood how this pattern interacts with income. Figure 5 shows the basic result from individual-data regressions: in recent years, the predictors have had a positive interaction—that is, religious attendance is a stronger predictor of Republican voting among higher-income Americans (and, conversely, income predicts better among religious attenders). We have also done state-level analyses but do not include them here.

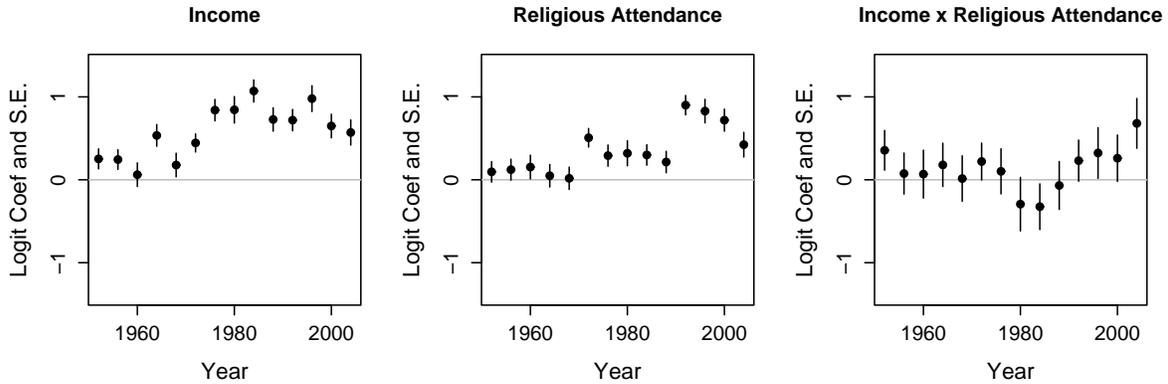


Figure 5: Coefficients of income, religious attendance, and their interaction, from a logistic regression of Republican vote preference fit to each presidential election year. Both inputs have been centered and scaled so that the main effects and interactions can all be interpreted on a common scale.

In the analysis leading to Figure 5, both variables have been centered to have mean zero and rescaled to have standard deviation 0.5 (Gelman, 2007), so we can interpret the main effects and the interaction directly as comparisons between high and low values of the predictors.

For an even more direct interpretation, however, that can be understood by nonstatisticians, we compare high income (upper third) to low income (lower third) and religious attendance once per week or more (from the data, the upper 36% in recent years) to religious attendance once per month or less (lower 49%). In this particular example, the discreteness of the religion scale made it difficult for us to pick categories that capture a quarter to a third of the data at each end.

Figure 6 shows the results, which are similar to the logistic regressions but can be immediately interpreted as differences in proportions. For example, rich people were almost 20% more likely than poor people to support George Bush in 2004, religious attenders were about 10% more likely than nonattenders to support Bush, and the difference between rich and poor is over 20% higher among the religious than the nonreligious. For a similar analysis in an international context, this time comparing low to middle income voters, see Huber and Stanig (2007).

4 Discussion

Discretization is not generally recommended when the goal is efficient inference, but it can be effective in aiding the communication of regression results. Comparing the average value

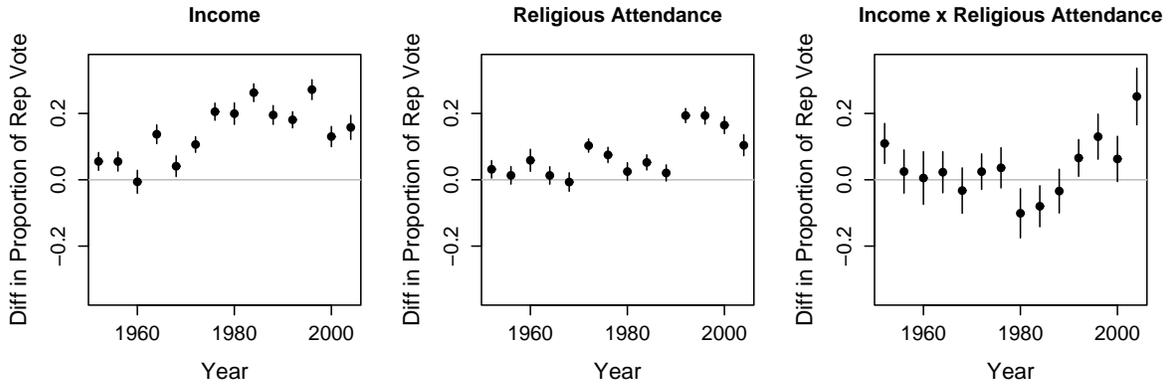


Figure 6: Difference in Republican vote between rich and poor, religious and non-religious, and their interaction (that is, the difference in differences), computed separately for each presidential election year. Compare to Figure 5, which shows similar results using regression coefficients.

of the outcome for the upper and lower third or quarter of the predictor is a quick and convenient summary that, as we have shown, loses little in efficiency compared to linear regression on the original continuous predictor. We recommend these simple differences for displays and summaries for general audiences, perhaps reserving the full regression results for appendixes or presentation in specialized journals. The ideas of this article should illuminate the connection between regression and simple differences and ultimately allow a greater understanding of the former in terms of the latter.

Finally, we performed our theoretical analysis in Section 2.3 under the assumption that the linear regression model was true. One could consider other models—for example, discretization could perform particularly well if the underlying regression were a step function, or particularly poorly if the regression slope increased sharply at the ends of the range of x . Our approach as described in this paper is most relevant for summarizing relationships that are monotonic and not far from linear—that is, the settings where linear regression would be routinely used. More generally, discretization can be used to capture nonlinear patterns, as discussed by O’Brien (2004).

References

- Brooks, C., and Brady, D. (1999). Income, economic voting, and long-term political change in the U.S., 1952–1996. *Social Forces* **77**, 1339–1374.
- Gelman, A. (2007). Scaling regression inputs by dividing by two standard deviations. Technical report, Department of Statistics, Columbia University.

- Gelman, A., and Pardoe, I. (2007). Average predictive comparisons for models with non-linearity, interactions, and variance components. *Sociological Methodology*, to appear.
- Gelman, A., Shor, B., Bafumi, J., and Park, D. (2007). Rich state, poor state, red state, blue state: what's the matter with Connecticut? Technical report, Department of Statistics, Columbia University.
- Glaeser, E. L., and Ward, B. A. (2006). Myths and realities of American political geography. Harvard Institute of Economic Research discussion paper.
- Huber, J. D., and Stanig, P. (2007). Why do the poor support right-wing parties? A cross-national analysis. Technical report, Department of Political Science, Columbia University.
- Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, Calif.: Duxbury.
- McCarty, N., Poole, K. T., and Rosenthal, H. (2006). *Polarized America: The Dance of Political Ideology and Unequal Riches*. Cambridge, Mass.: MIT Press.
- O'Brien, S. M. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics* **60**, 504–509.