Working with Ordinal Predictors*

Paul E. Johnson

March 30, 2009

Abstract

The use of categorical variables in regression modeling is discussed. Some pitfalls in the use of numerically scaled ordinal variables are considered.

*Prepared for delivery at the annual meeting of the Midwest Political Science Association, Chicago Illinois, April 2-5, 2009.

In the day-to-day practice of social science, nominal and ordinal variables are the predominant variable types. Survey data sets may contain the occasional numerical variable like "age" (in years), but that type is grossly outweighed by the categorical indicators. Categorical variables are usually subdivided into "nominal" and "ordinal" types. The use of ordinal variables is the main focus of this paper. The presentation is not aimed at methodologists, but rather at practitioners who work with these very common data types.

The basic idea that differentiates ordinal variables is that, in the eye of the person who uses the data, an observation's membership in one category can somehow be interpreted as "less" or "more" than another observation's membership in a different category. Ordinal variables may be based on judgment scales, so the ordering is not purely in the mind of the analyst. Survey respondents select "never", "sometimes", "often", or "frequently" to describe their perception of police brutality or how often they read the local paper. Some ordinal variables are constructed by grouping observations together when they might in fact have different observed values. A respondent's educational level may actually exist in days or years, but it is almost always scored in categories, such as "high school or less", "some college", and "college degree or higher".

In statistical analysis, it is relatively common to treat ordinal variables as if they were numerical scales. (I would hate to name names by listing specific articles, since the use of numerical scales as predictors is so frequent that it seems unfair to point the finger at any particular projects. I suppose it will be necessary to name some examples, eventually.) This practice dates back to the introduction of statistics in political science. In SPSS, the first widely used statistical package, variable "values" are numerical scores, while substantive meaning of these scores is recorded separately in "value labels." The values 0, 1, and 2, might represent a scale "none", "some" and "lots." The problem with conceptualizing the variable as a numerical score is that the assigned numbers lack intrinsic meaning–they might just as well be 0, 10, and 54. In the classic sense of "garbage in and garbage out," statistical procedures in many (most) programs will accept that variable, with either coding, at numerical "face value."

The R statistical program (R Development Core Team, 2008) answers commands in a language that is quite similar to S, which was developed at Bell Labs (Becker et al., 1988). In S and R, categorical variables are called "factors." One of the most notable characteristics of the R/S language is that it never allows us to forget that categorical variables need to be treated differently than numerical variables. One does not think of the "values" of a factor variable as numbers (1, 2, 3), but rather as qualitative levels like "none", "some" and "lots", or "male" and "female", or "Democrat" and "Republican". Statistical models in S and R are almost always designed to prevent us from making mistakes with categorical variables. If one tries to use a factor variable in a statistical model, R will either refuse to calculate estimates[1], or, where possible, it will construct a workaround so as to avoid placing any weight on the numerical values[2].

While the R/S approach to factors is uncomfortable and tedious for new users, I have grown to appreciate it. This has put me out of step with the social science crowd, however. The things we "usually" do seem rather questionable.

[1]If "sex" is a factor variable, mean(sex) in R returns "argument is not numeric or logical: returning NA"

[2]The workaround is the creation of a system of numerical "contrasts". The details will be discussed below.

The argument proceeds as follows. First, I begin with an example of a deeply flawed model that results from using a categorical variable as if its numerical scores were meaningful. Second, I offer some observations about the kinds of graphs that we should use when presenting categorical variables. Third, a small simulation exercise is presented which demonstrates some likely sources of error when factor variables are treated as numerical variables. Finally, some recommendations about finding the simplest workable model are described.

1 A Vignette on the Civic Virtue of Republicans

We begin with an example that uses survey responses collected by the General Social Survey in 2006 (Davis, 2007). Suppose one begins with the hypothesis that Democrats are bad people who don't vote as often as they should. Republicans, on the other hand, are good people who carry their fair share of the civic burden. The strength of a respondent's Republican attachment is represented by familiar 7 point scale that that ranges from strong Democrat to strong Republican. A logistic regression model to predict whether the respondent claimed to have voted in 2004 was estimated. The predicted probabilities from a model using party identification as a 7 point numerical scale are illustrated in a familiar way in Figure 1.

The parameter estimates seem to confirm our suspicions about Democrats. The positive coefficient on party ID indicates that the more strongly a person identifies with the Republican party, the more likely that person is to vote.

An alternative view of the situation can be found in Figure 2. The plot of voter participation in Figure 2 illustrates the V-shaped relationship between party identification and voter participation. This fitted model does not treat party identification as a 7 point numerical scale. Rather, in this is a categorical coding model. To make the relationship between the two sets of estimates as clear as possible, we begin by writing down the coding matrix. See Table 1. The various values of party identification are represented by "contrast variables" that are coded 0 or 1. While there are many different ways to code these contrasts to emphasize various insights, this model uses the standard approach that is called "treatment contrasts" in the R statistical program (R Development Core Team, 2008). The regression intercept represents the baseline category, which in this case is Strong Democrat. The predicted probability of voting in this model matches the observed participation rates exactly, of course, because each value of party identification is treated separately.

This second set of parameter estimates tells a rather different story. Strong Democrats and Strong Republicans are the most likely to vote, weaker identifiers are less likely, and Independents are the least likely to vote. This conclusion is, of course, quite consistent with the conventional view that strong partisans are more likely to vote.

The right hand side of the categorical model has the intercept (the baseline) and estimates for the six contrast effects:

$$\beta_0 + \zeta_1 C1_i + \zeta_2 C2_i + \zeta_3 C3_i + \zeta_4 C4_i + \zeta_5 C5_i + \zeta_6 C6_i \tag{1}$$

The model with the numerical coding would have this right hand side

$$\beta_0 + \beta\, party\, identification_i \tag{2}$$

Table 1: Coding Matrix

Observation	Numerical Score: $party\,identification$	Contrasts					
		C1	C2	C3	C4	C5	C6
Strong Dem.	1	0	0	0	0	0	0
Dem.	2	1	0	0	0	0	0
Indep. Lean Dem.	3	0	1	0	0	0	0
Indep	4	0	0	1	0	0	0
Indep. Lean Repub.	5	0	0	0	1	0	0
Repub.	6	0	0	0	0	1	0
Strong Repub.	7	0	0	0	0	0	1

The second model is a simplification of the first. The simplification is obtained by imposing a restriction on the parameters ζ_j:

$$\zeta_j = (1 + j) \cdot \beta \tag{3}$$

In other words, the numerical model places this simplifying structure on top of the categorical model

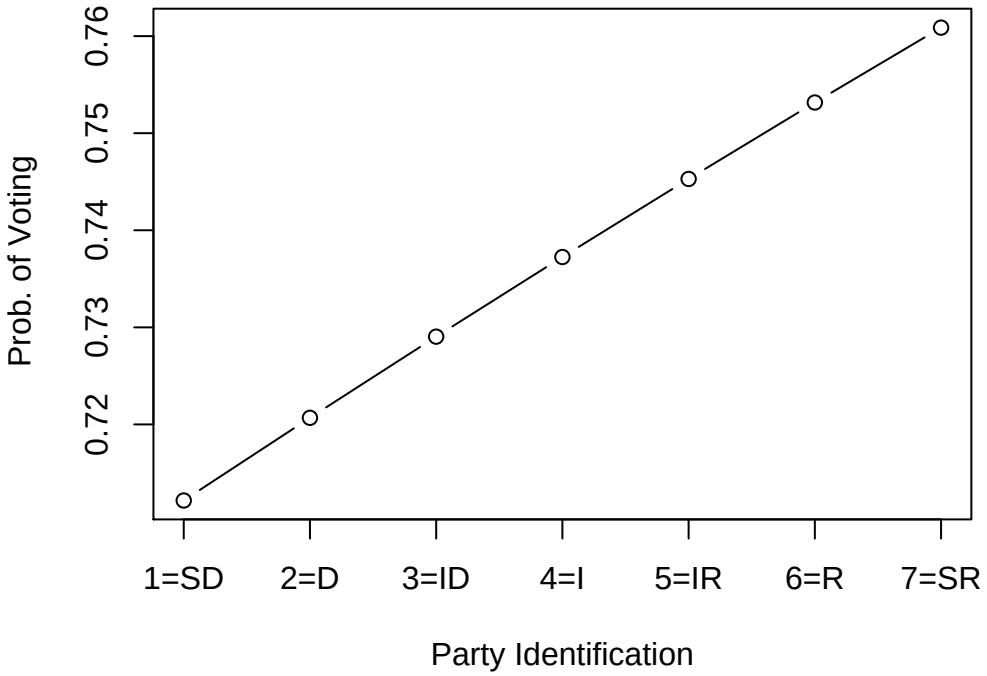
$$\beta_0 + 2\beta C1_i + 3\beta C2_i + 4\beta C3_i + 5\beta C4_i + 6\beta C5_i + 7\beta C6_i \tag{4}$$

Taking note of the fact that $(1 + j)Cj_i = party\,identification_i$, we see that equation 4 collapses into 2.

The numerical coding model is thus nested within the more general categorical coding model. As a result, a hypothesis test of the idea that the numerical model is "just as good" is easily constructed. With a continuous dependent variable in an OLS framework, we would use an F test. To compare maximum likelihood models, we use a likelihood ratio test. These tests measure the difference in the quality of the final fitted models. If the model fits are similar, that test statistic, which is distributed as a χ^2 variable, will be small. As one might expect, in this case the likelihood ratio test indicates that simpler numerical coding does indeed "throw away" a significant amount of explanatory power (χ^2 =-507.2, df=5,$p \leq 0.001$).

In this case, a likelihood ratio test may seem like overkill. After glancing at the graph in Figure 2, it is difficult to conceive of a reason why a researcher would prefer the numerical coding of party identification to the categorical coding. In a more complicated case involving many predictors, however, a figure may not reveal the problem so clearly.

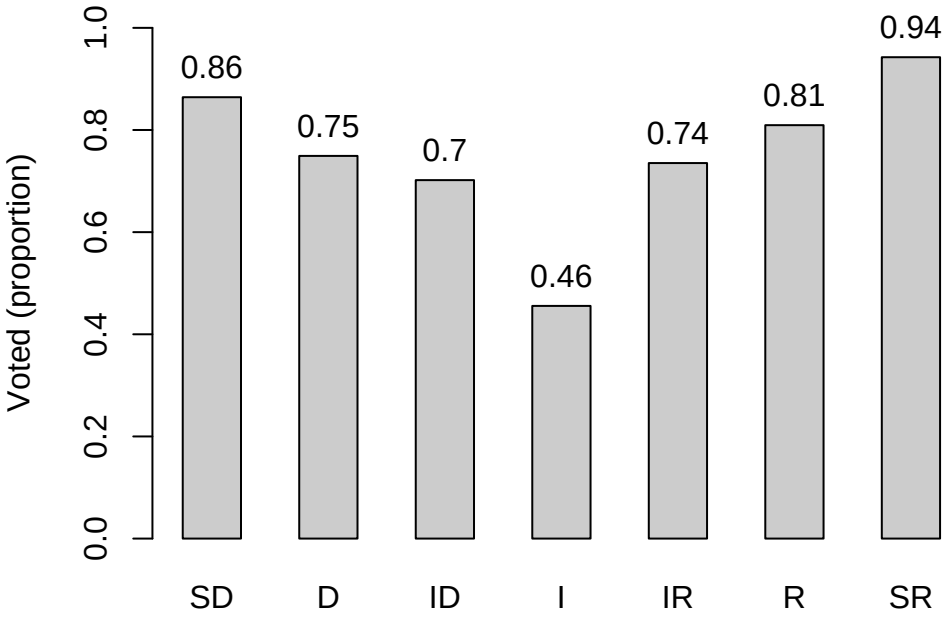
Figure 1: Do Republicans Have Superior Civic Virtue?



Voted in 2004	ML Estimate	(std. error)
Intercept	0.864^{***}	(0.075)
Party ID (1-7)	0.042^{*}	(0.018)
Nagelkerke R-sq.	0.002	
Likelihood-ratio (Model χ^2)	5.515	
p (χ^2)	0.019	
Log-likelihood	−2339.402	
Deviance	4678.804	
N	4052	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

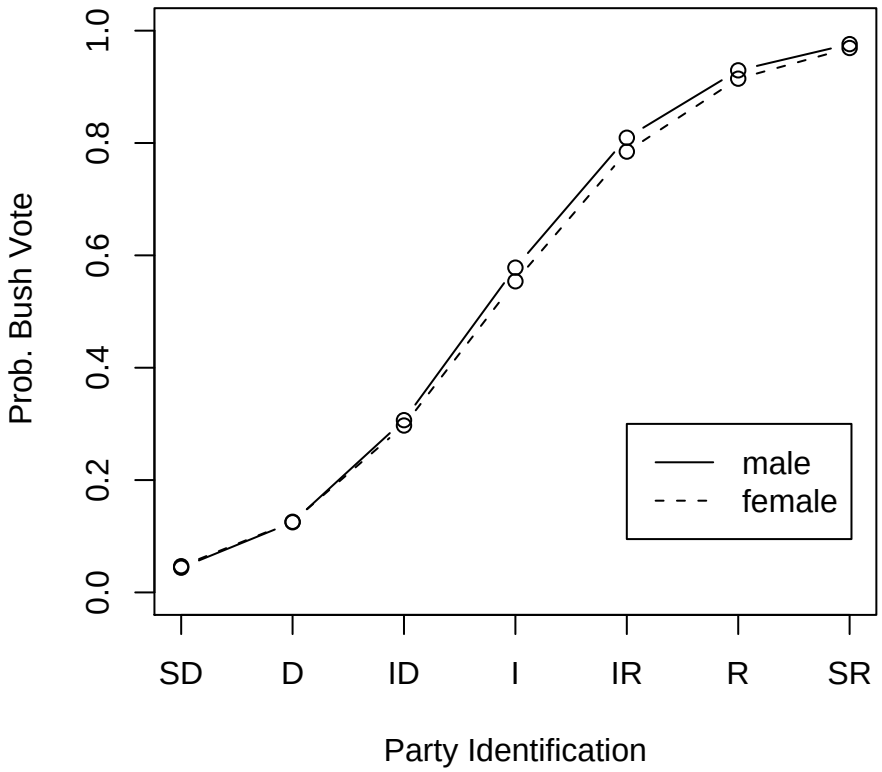
Figure 2: Voter Participation in 2004



Voted in 2004	ML Estimate	(std. error)
(Intercept)	1.850***	(0.113)
Democrat	−0.756***	(0.143)
Indep. Leaning Dem.	−0.994***	(0.150)
Independent	−2.029***	(0.133)
Indep. Leaning Rep.	−0.829***	(0.172)
Republican	−0.404**	(0.153)
Strong Repub.	0.946***	(0.225)
Nagelkerke R-sq.	0.173	
Likelihood-ratio (Model χ^2)	512.706	
p (χ^2)	0.000	
Log-likelihood	−2085.807	
Deviance	4171.613	
AIC	4185.613	
BIC	4229.762	
N	4052	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Figure 3: Predicted Probabilities for 2004



2 We Can Make Better Illustrations

A plot can reveal the mismatch between the linear model and the categorical data. But it seems that many scholars do not inspect these plots in the ordinary course of their research. If a dependent variable is more-or-less continuous, the mismatch is easier to spot, but even when the dependent variable is limited or qualitative, a graph may still help.

In research articles that use logistic regression, the most commonly presented graph presents the "predicted probabilities." A specimen is offered in Figure 5. Political party identification is used to predict the probability of voting for George Bush in the American election of 2004 (again with data from the General Social Survey of 2006). The logistic regression model predicts the respondent's choice between presidential candidates George Bush and John Kerry as a function of the respondent's position on the 7 point political party identification scale. The figure seems to illustrate the logistic "s-shaped" curve quite nicely. Men and women are not all that different, it seems, once we have taken into account their political party identification.

While this kind of figure is ubiquitous, it may be also be misleading. The first problem, of course, is that we have estimated the effect of party identification as if it were a meaningful

numerical scale, and this figure does nothing to warn us if that was a mistake. In this particular case, the numeric coding of the party identification variable does not appear to be hugely wrong. In Figure 4, two types of plots are presented. These estimates pool men and women together for illustrative purposes. The bar chart in 4a represents the proportion who voted for Bush by vertical bars, and the predicted probabilities of the two models are represented by lines. In Figure 4b, the same estimates are represented on a dot chart. I suspect that, in practice, the numeric coding would be "close enough" for most researchers. Nevertheless, the χ^2 test rejects the numerical coding ($\chi^2 = 70.99$, $df = 5$, $p = 6.3 \times 10^{-14}$).

There is a more fundamental problem in these graphs. The deception implicit in any line graph flows from the fact that it implicitly converts a categorical, integer-valued scale into a real numbered variable. The figure supposes not only that it is meaningful to count 1, 2, 3, but it also "synthesizes" predicted values for decimal positions between the integers. The lines that "connect the dots" create an illusion of a smoothly increasing curve. The input is measured only at 7 discrete values, but the figure output creates a continuum.

A more honest depiction of the relationship can be had with a bar plot (or a 'dotchart') or a step plot, as illustrated in Figure 5.

3 Will Diagnostics Help?

It is not always wrong to treat a seven point categorical scale as if it were an integer-valued numeric variable. If the effect of the categorical variable happens to be increasing in even steps, then estimation with a numerical scale might be useful. The main danger for the practical researcher is that regression estimates can be calculated whether or not the linear model is appropriate. Moreover, the summary statistics from the regression model do not provide much protection against making a mistake. Recall the misleading nature of the estimates of voter participation and the 7 point party identification scale (Figure 1).

The essence of the problem is easier to illustrate if we use a continuously distributed dependent variable. Simulated data is presented in Figure 6. The input variable, x_i is coded 1 through 7. The dependent variable, y_i, is equal to $1x_i$ plus a random error that is normally distributed with a standard deviation of 5. The "occular regression" test clearly indicates that the output variable is increasing in "steps". Most applied researchers would probably agree that the linear model is good enough.

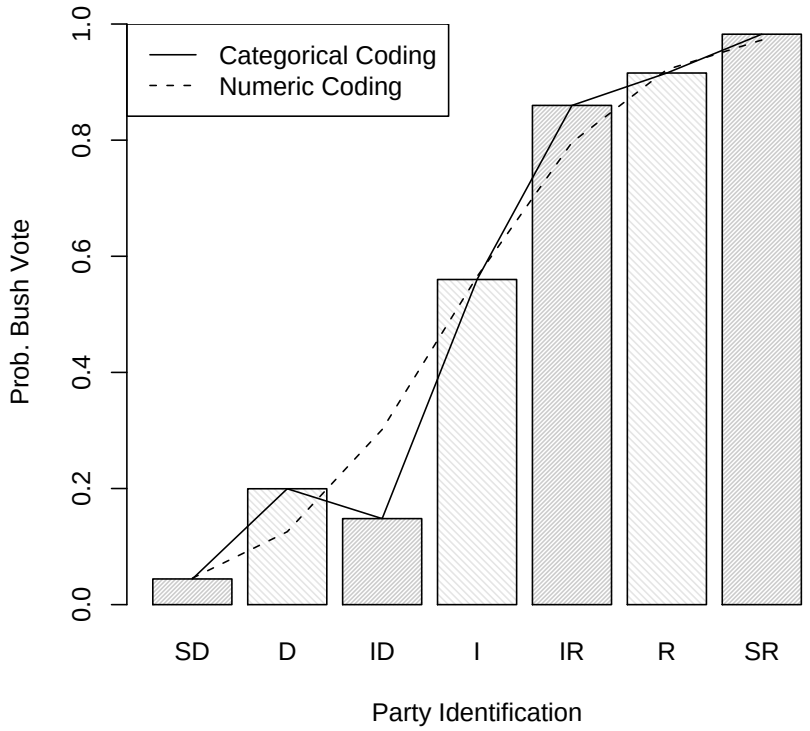
Figure 6, two types of graphs are compared. A scatterplot and a box plot are presented for the same data. The OLS regression line is superimposed. The box plot is preferred when the sample size is large. The weakness of the scatterplot is that plotted points overlap. The box plot avoids that problem by representing information differently. Each box represents the interquartile range of observed values.

On the other hand, it is easy to construct an example in which the linear regression seems wrong. Consider Figure 7. The expected values of the seven groups are ordered, but not in even steps:

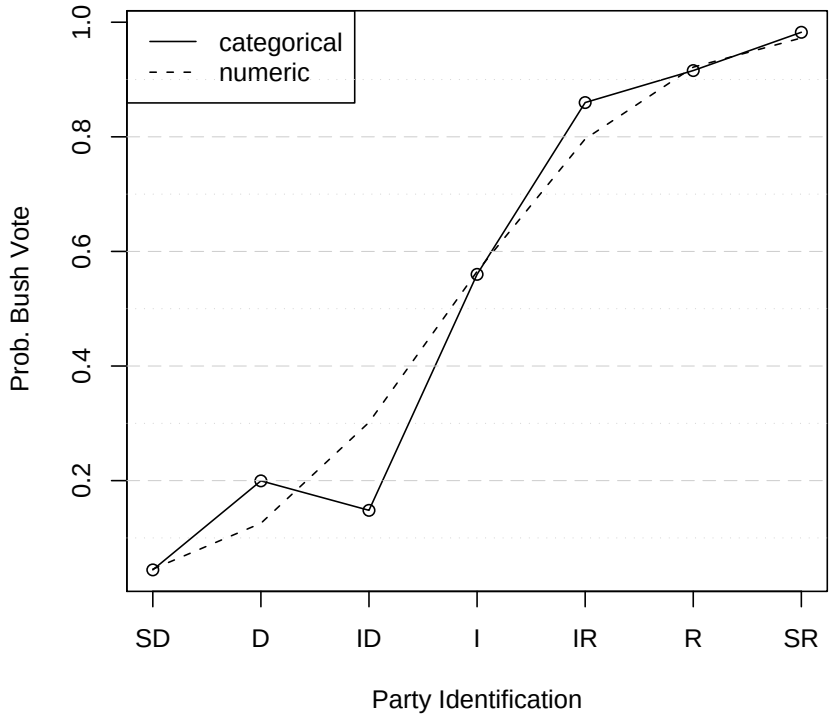
$$\{-40,\ 1,\ 2,\ 3,\ 4.8,\ 5,\ 6,\ 13\}$$

The OLS estimates do not offer a very strong warning about the apparent nonlinearity. The t values are "even better" than the other model and the R^2 is very impressive by usual standards. Almost half of the variance is "explained" by x_i treated as a numerical variable.

Figure 4: Observed Presidential Votes and Predictions



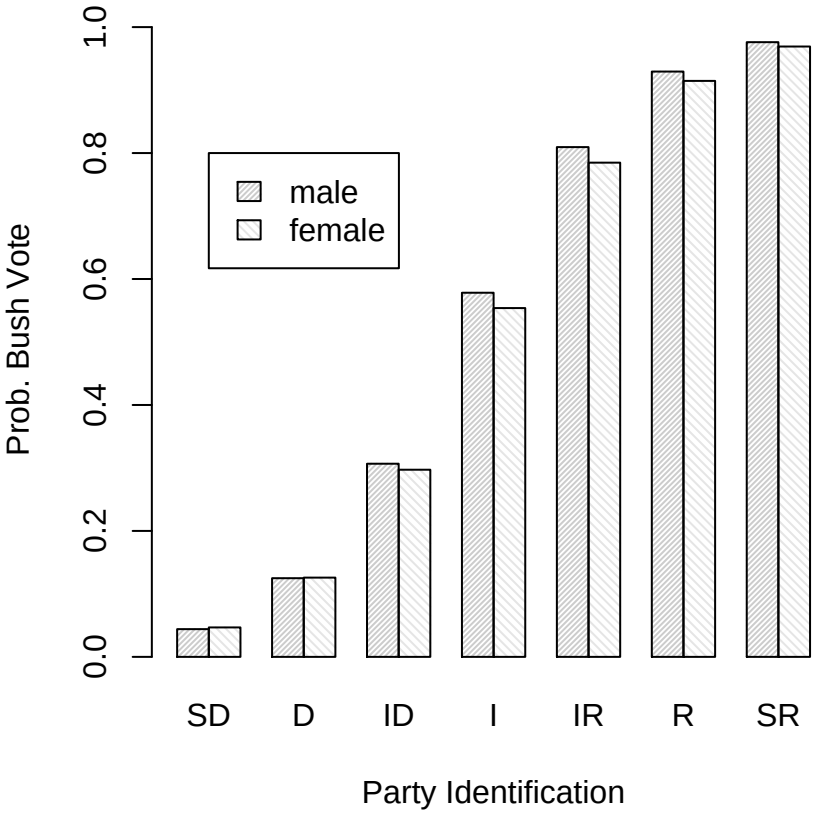
(a) Bar Chart



(b) Dot Chart

Figure 5: Predicted Probabilities In Other Types of Graphs

a) Bar plot



b) Step plot

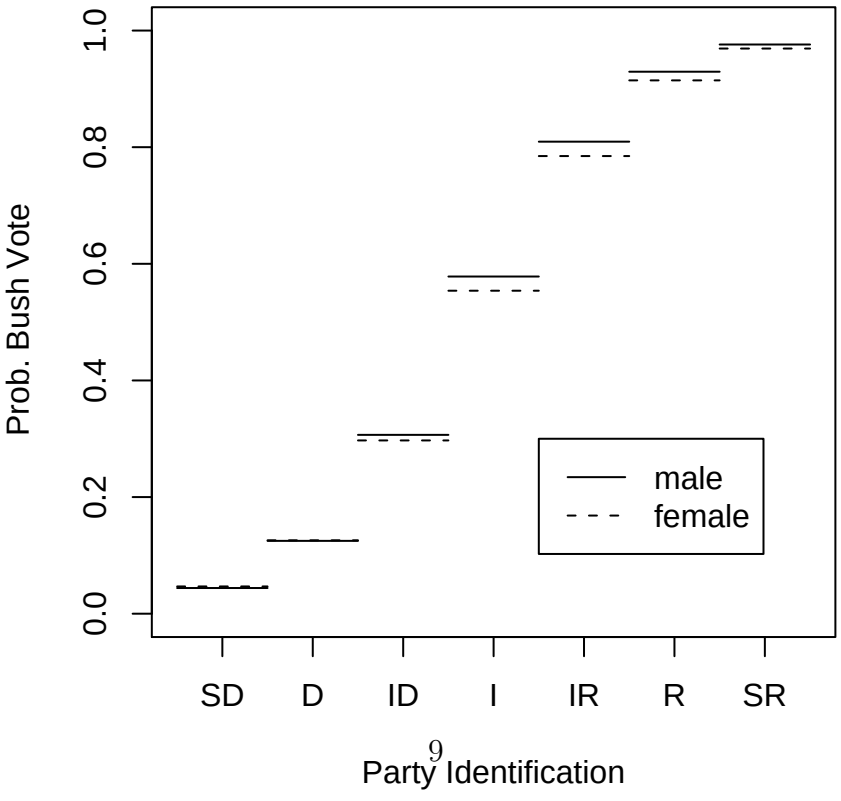
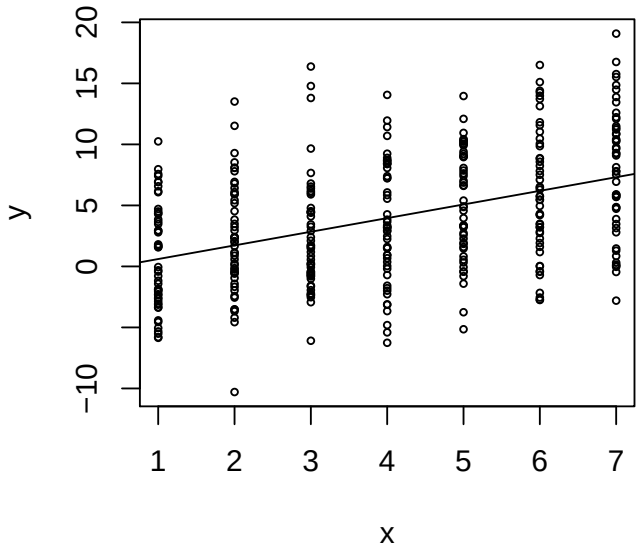
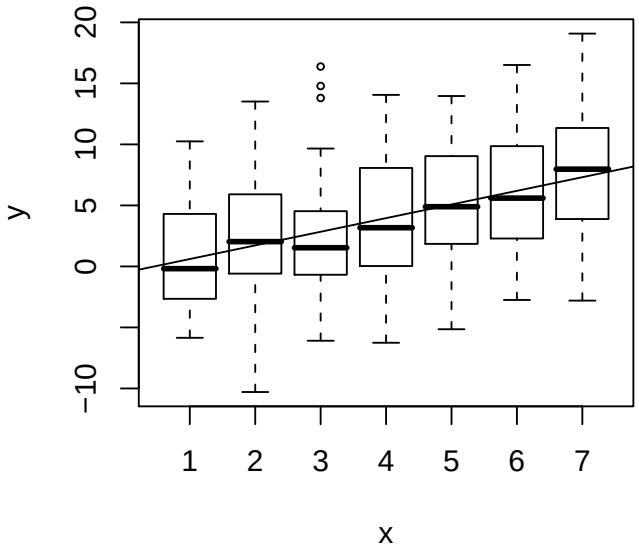


Figure 6: A Stepping Categorical Predictor

a) Scatter plot



b)Box plot



	Dependent Variable: y OLS Estimate (std. error)
(Intercept)	-0.513^{***}
	(0.565)
x	1.118^{***}
	(0.126)
R-squared	0.183
adj. R-squared	0.181
root MSE	4.730
N	350

A diagnostic test (either an F test or a likelihood ratio test) can be used to compare the regression based on the "numerically coded" predictor against the "categorically coded" predictor. In this case, the test indicates that the numerical coding does not fit as well. If applied researchers would at least conduct that one piece of post-hoc hypothesis testing, then we would have more confidence in regression results that are reported.

Those examples piqued my curiosity. How often will the "p" value in the linear regression be misleading? What are the chances that post hoc comparison of the models (by an F test or a likelihood-ratio test) will reject the linear regression?

A simulation exercise was conducted. Consider 7 groups of 50 observations. The expected outcome of each group is drawn from a normal distribution.

$$\delta_j \sim N(0, \sigma_\delta^2), j \in \{1, 2, \ldots, 7\}.$$

The observed outcome equals that expected value plus some "noise"

$$y_{ji} = \delta_j + e_{ji},\ j \in \{1, 2, \ldots, 7\}, i \in \{1, 2, \ldots, 50\},$$

where e_{ji} is a normally distributed random error with a mean of 0 and a standard deviation of σ_e. Let $x_{ji} = j$. That is, x_{ji} is an integer-valued 7 point scale. A linear ordinary least squares regression model that uses the numerical scale as the predictor is.

$$\hat{y}_{ji} = \hat{\alpha} + \hat{\beta} x_{ji} \tag{5}$$

I have created 1000 sets of δ_j coefficients and executed the simulation with several different values of σ_e^2. For each run, a regression model is calculated and a plot with summary information is created.

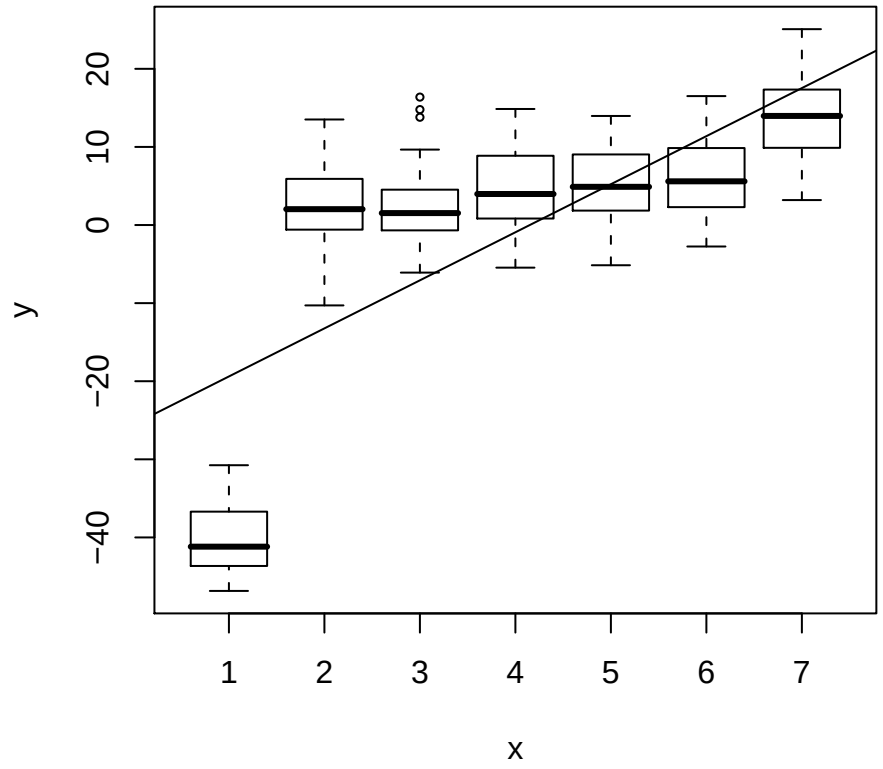
The extent to which the group coefficients δ_j fall into linear "steps" is measured by R_δ^2, the coefficient of determination for a regression of δ_j on the sequence from 1 through 7.

A case in which the group coefficients fall almost exactly on a straight line is illustrated in figure 8. (The legend indicates $R_\delta^2 = 0.93$). In this case, it appears that the regression model with the numerical predictor gives a fairly meaningful portrayal of the situation.

A more troublesome case is illustrated in Figure 9. There is no linear pattern among the group center points, as indicated by an R_δ^2 not distinguishable from 0. However, the OLS regression indicates there is a statistically significant pattern, in the sense that the estimated slope is negative and the t-ratio allows us to reject a null hypothesis of 0. One of the important insights to be found in this example is that "good" (statistically significant) regression results should not be interpreted as support for the specification of the model itself. People who claim that their estimates of the slopes are meaningful because they have big t-values are just plain wrong.

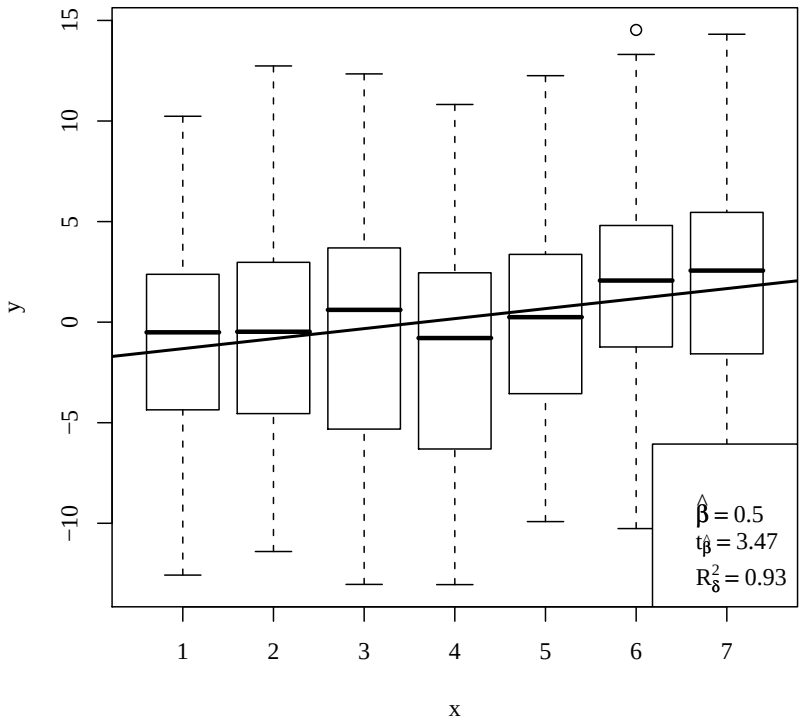
How often does it happen that there is no strong linear "step" pattern but the regression estimate indicates the presence of a statistically significant linear effect? And how often does the likelihood ratio test reject the linear model in favor of the linear coding? In the end, we find that the variance of individual level error plays a determinative role. When the error is increased, fewer of the t tests for the regression slope are significantly different

Figure 7: A Dubious Regression



Output	OLS Estimate (std. error)
(Intercept)	-25.542^{***}
	(1.425)
Input	6.154^{***}
	(0.319)
R-squared	0.517
adj. R-squared	0.516
root MSE	11.921
N	350

Figure 8: The Linear Model Almost Fits



$$\sigma_e^2 = 25$$

Table 2: Individual Level Error Obscures Nonlinearity

$\sigma_e = 0$			$\sigma_e = 1$			$\sigma_e = 5$		
	$t-test$			$t-test$			$t-test$	
LR test	$p \leq .05$	$p > .05$	LR test	$p \leq .05$	$p > .05$	LR test	$p \leq .05$	$p > .05$
$p \leq .05$	100%	100	$p \leq .05$	99.59%	99.27	$p \leq .05$	55%	59
$p > .05$	0	0	$p > .05$	0.41	0.79	$p > .05$	45	41
N	830	170	N	726	274	N	277	723

from 0. However, at the same time, the probability that the likelihood-ratio test will weed out probable mistakes is also reduced. Table 2 displays a cross tabulation of the results of the hypothesis test that the linear equation's slope is equal to 0 and the probability that the linear model is rejected by the likelihood-ratio test. When the standard deviation of the individual error is 0, the $t-test$ on the slope coefficient is statistically significant at the 0.05 standard in 830 of 1000 runs. In all 1000 runs, the likelihood ratio test rejects the linear model. When the standard deviation of the error increases, however, the pattern changes. When the standard deviation of the individual level random error is raised to 1, and then to 5, the t-test is less likely to reject the null hypothesis. However, the likelihood ratio test is also much less powerful.

I've attempted to piece together the many moving parts of this puzzle in Figure 10. The R^2_δ values are represented on the horizontal axis, so on the left side of the graph, the OLS regression would be doing us a favor if it reported that β is near 0. The observed estimates of β are presented on the vertical axis. The estimates of β that are not statistically significant according to the t-test are represented by small red x's, while the statistically significant ones are represented by small black circles. The figures indicate that even when R^2_δ is near 0, there are plenty of cases in which the estimate of β is significantly different from 0. When the individual level error is smaller, we are more likely to estimate a regression that claims there is a statistically significant relationship. As the bottom panel of the figure indicates, when the individual error is raised, then the OLS is less likely to find that there is a significant linear relationship.

4 Finding The Simplest Plausible Model

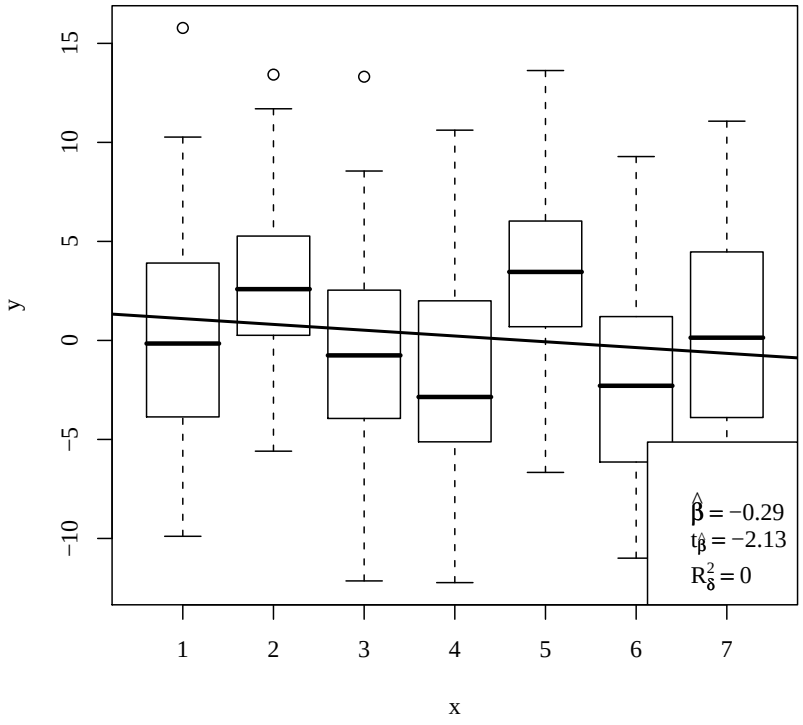
There is a clash between the simple, parsimonious regression model that we want and the reality of work with categorical predictors. Everybody who has taken a regression course wants to say, almost like a poem, "a one unit increase in x causes a β increase in y." If we fit a model like this:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

we can also fit a model like this:

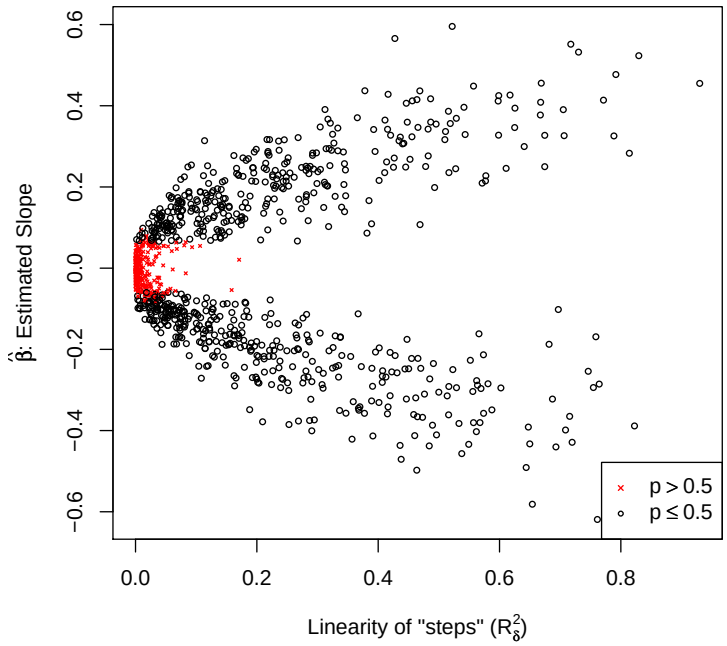
$$y_i = \beta_0 + \zeta_1 C1_i + \zeta_2 C2_i + \zeta_3 C3_i + \zeta_4 C4_i + \zeta_5 C5_i + \zeta_6 C6_i + e_i$$

Figure 9: Troublesome Case

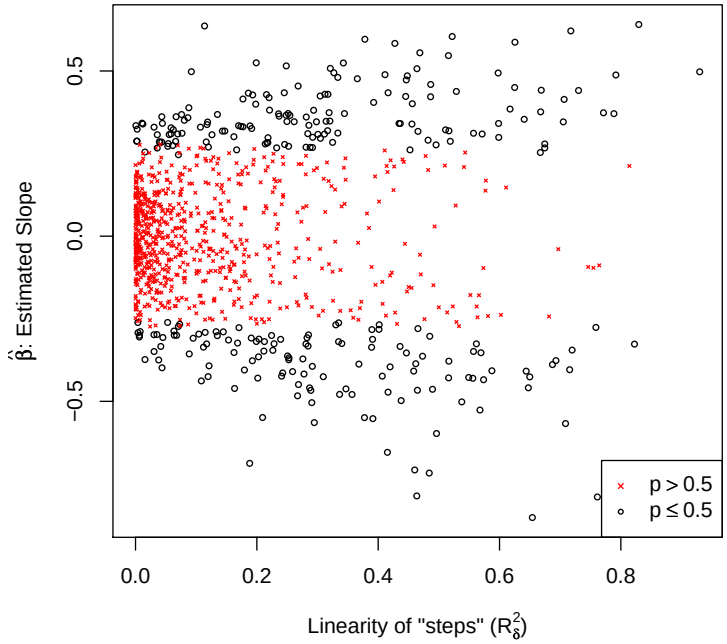


$\sigma_e^2 = 25$

Figure 10: β Estimates and Underlying Nonlinearity
a) Small individual level random error ($\sigma_e^2 = 1$)



b) High individual level random error ($\sigma_e^2 = 25$)



and the hypothesis test between the two models is straightforward (F or χ^2, depending on the context).

When it is apparent that a straight line model does not "fit," the first thought of many applied social scientists is "put in a squared term," but the options are considerably richer than that. These options are, for the most part, intended for numeric, rather than categorical variables. Putting regression trees aside for a moment, because they represent a very different strategy, I believe it is safe to say that the prominent approaches to nonlinear relationships will take one of two strategies:

- bend the data to fit the curve, or
- bend the curve to fit the data

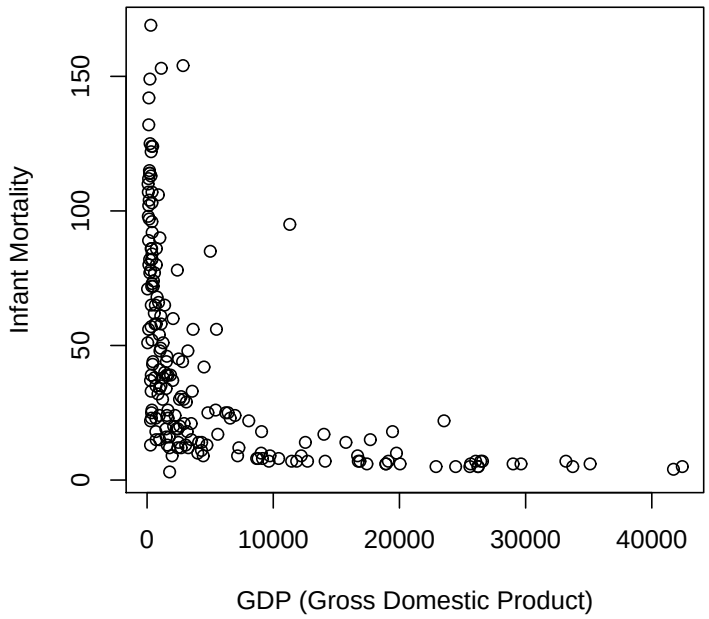
A common method of bending the data is the use of logged data in place of the original variables. Most students have seen the ubiquitous infant mortality data set. It seems as though there is a version of this data for every statistical program. Professor John Fox's version of that data (Fox, 2009) for R is presented in Figure 11. Logging the variables produces a second set of data points that appear more consistent with a linear model. We fit the model to the logged data, and then convert the results back into the scales of the original observations.

The use of logs is certainly not the end of the story. The log is a special case of another parametric transformation, the Box-Cox transform (Box and Cox, 1964) which can be estimated with the MASS package of routines that are distributed with Venables and Ripley (Venables and Ripley, 2002). The approach proposed by Tibshirani, 1988, which is implemented for R in "acepack" (Spector et al., 2009), takes another large step in the direction of abstraction. It will squeeze and stretch values of variables in a way that makes the data fit a straight line as well as possible and also make the fit as homoskedastic as possible.

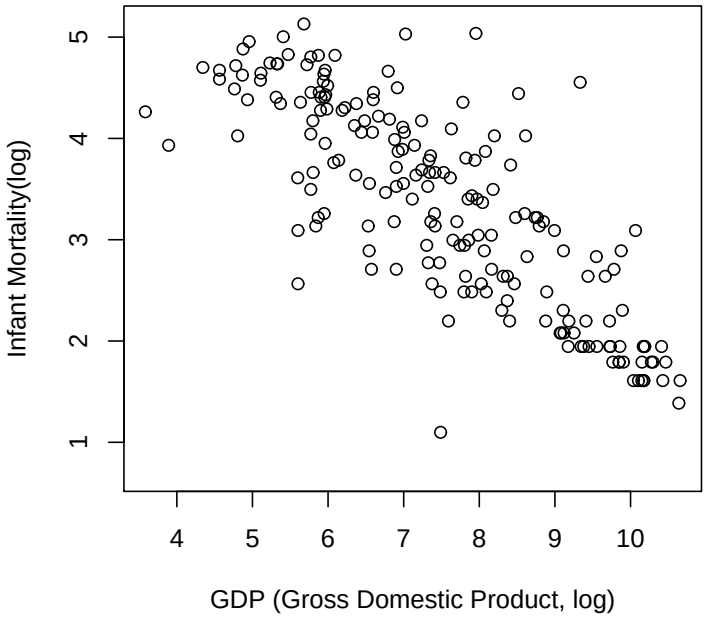
Approaching this from the other direction, we may bend the curve to fit the data. A parametric approach would add x_i^2 or other mathematically transformed terms into a regression model. If one proposes a predictive model that is inherently nonlinear, there are fitting tools like "nonlinear least squares" (Pinheiro et al., 2008). Alternative nonparametric line-bending approaches have rapidly grown in popularity. LOESS is a "locally weighted" regression model in which a predictive model is built for each value of x_i putting weight on the nearby observations (Cleveland and Devlin, 1988; for which software is also available (Cleveland et al., 1992)). A closely related type of local regression is "kernel smoothing"; programs that can estimate one can usually estimate the other (e.g,Loader, 2007; Hayfield and Racine, 2008). Models that use regression splines fit into this spectrum somewhere between the transformed parametric models of the first approach and the unstructured nonparametric approach. A regression spline model will divide the data into sections and then try to fit a curve within each section (Harrell, 2008). The sections of the curve are separated by break points called "knots" and various types of splines are differentiated according to the mathematical restrictions imposed within knots and where two smooth curves "join together" at the knots. There are several implementations (Wood, 2006; Hastie, 2008) of the so-called "generalized additive model" (Hastie and Tibshirani, 1990) takes an ordinary regression model

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + e_i \tag{6}$$

Figure 11: Infant Mortality Plots

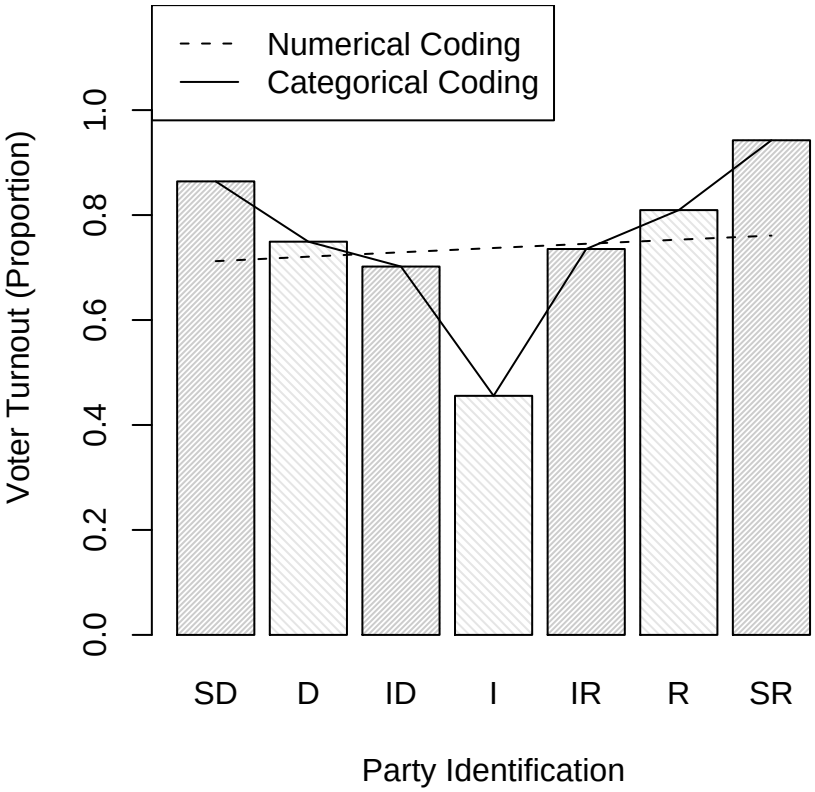


(a) Unlogged Data



(b) Logged Data

Figure 12: Participation: Bend the Curve



and replaces its predictors by one of the kinds of "smoothing" lines

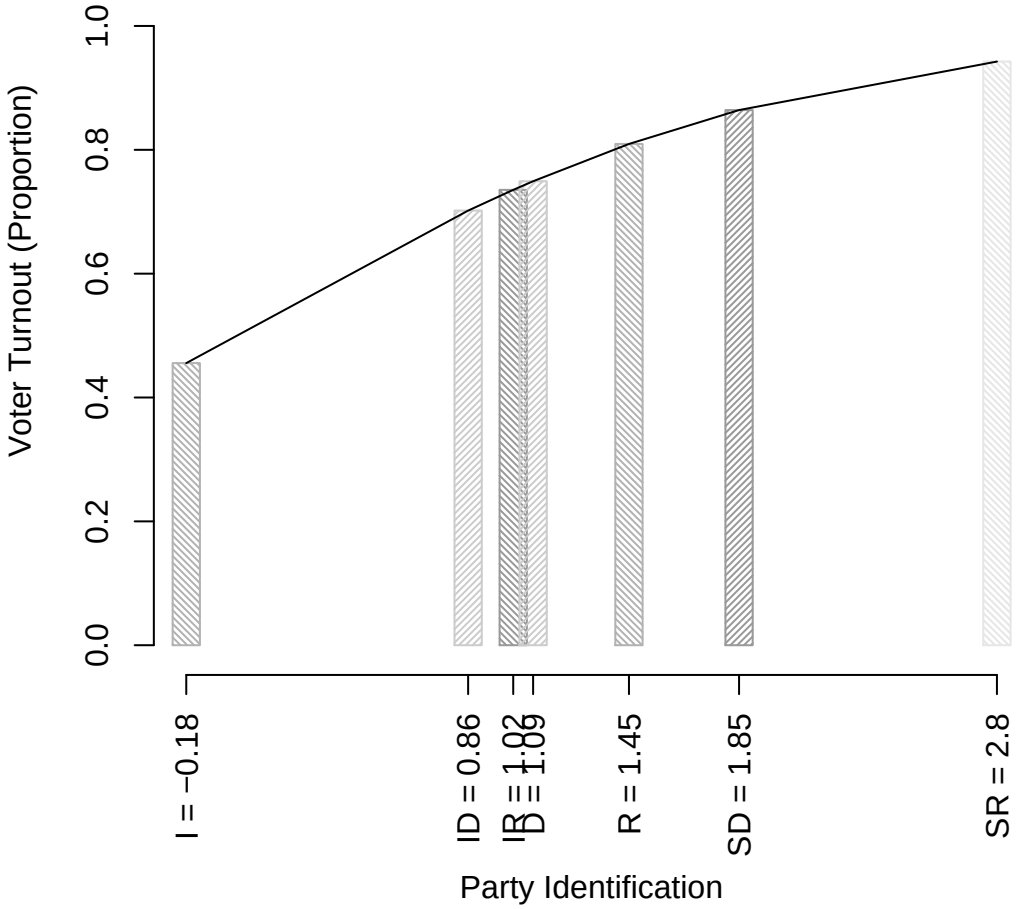
$$y_i = \beta_0 + \beta_1 s(x1_i) + \beta_2 s(x2_i) + e_i \tag{7}$$

One can even suppose that there is a general, nonparametric interaction between variables (Wood, 2003).

I'm not convinced that there is a great substantive difference between "bending the data" and "bending the regression line" when the inputs are numeric variables. I'm almost sure there is no difference when the predictors are categorical. Recall the categorical model equation introduces contrasts as predictors to replace a factor variable. See equation 1. The parameter estimates for that model were presented in Figure 2. The predicted values obtained from that model amount to "bending the line" to fit the observed data. The fit for the voter participation in the 2004 GSS is presented in Figure 12. The bars represent observed voter participation and the solid line connects the predicted values from a logistic regression on the categorical predictor, as in equation 1. The dotted line shows the predicted probability of voting from a logistic regression on the numerically coded party identification variable.

If we pursue the other strategy–bend the data to fit the curve–it is necessary to "shuffle" the party identification categories so that the Independent voters are on the left and the

Figure 13: Participation: Bend the Data



Voted in 2004	ML Estimate	(std. error)
(Intercept)	0.000	(0.057)
Party Identification	1.000***	(0.048)
Nagelkerke R-sq.	0.173	
Likelihood-ratio (Model χ^2)	512.706	
p (χ^2)	0.000	
Log-likelihood	−2085.807	
Deviance	4171.613	
AIC	4175.613	
BIC	4188.227	
N	4052	

partisans are on the right. Consider Figure 13, where we illustrate the predicted values of a logistic regression in which the party identification variable is recoded so that the observed proportions exactly match the predictions of the S-shaped curve. The "bend the data" approach requires us to think of the modeling problem in a very unfamiliar way. Instead of plotting observations and then sketching a line, we sketch the line and then figure out what the observations should be. I've not seen this kind of illustration in a publication, perhaps it is original (unlikely).

The substantive conclusion of the modeling process is the same, whether we take the data-bending or the curve-bending approach. Both models predict that the probability that an Independent will vote is 0.455 and the probability that a Strong Republican will vote is 0.942.

The data-bending illustration in Figure 13 does lay bare some of the central issues in regression modeling with categorical data. Perhaps most importantly, the usual distinction between "nominal" and "ordinal" input variables is probably mistaken. Ordinality is not a necessary attribute of an input variable, considered in isolation. Rather, ordinality is a property of the relationship between the input and the output. While the 7 point party identification scale appears to be an ordinal predictor of presidential preference (recall Figure 5), it does not appear that the 7 point party identification scale is an ordinal predictor of voter participation.

The positioning of the bars in Figure 13 is substantively important. If two bars are close, or even "on top" of one another, then we should suspect that an observation's membership in one or the other does not have a major effect on the outcome. One wonders if we might re-group the respondents, as Independents, Strong Partisans, and Everyone Else. A likelihood ratio test rejects that simplification, unfortunately. In this case, there is actually a statistically significant difference between Strong Democrats and Strong Republicans. It appears that, at least in 2004, Republicans were actually (statistically significantly) more virtuous that Democrats.

The curve-bending approach is not popular among practitioners because it produces regression models that are "hard to interpret" and "have too many parameters." The data-bending approach appears to give a significant simplification because the final fitted regression only includes 2 parameters, the intercept and the slope. Of course, the simplification is an illusion. The same complex part is in there, but it is concealed in the first step of the analysis.

We would like to have a systematic way to decide if it is necessary to estimate parameters separately for all of the different categories. Suppose we have a categorical variable that is coded

- Never
- Rarely
- Infrequently
- Sometimes
- Often

It is easy for me to imagine that respondents who place themselves into either of the first three categories are actually the same. We could re-code this as

- Never-Rarely-Infrequently
- Sometimes
- Often

It may be that "Sometimes" and "Often" have effects that cannot be distinguished, and so we could end up with a dichotomous variable

- Never-Rarely-Infrequently
- Sometimes-Often

It seems that it is impossible to avoid a two stage estimation process. We must first we estimate the general model (equation 1) and then pare away redundant categories. Recent developments in the calculation of optimal re-grouping of categories Boulesteix (2007) and grouping of scale variables (Hothorn and Zeileis, 2008), offer some hope that a rigorous approach to regression with ordinal predictors might be developed.

The two step approach engages us in a process of post hoc hypothesis testing, which should cause some alarm bells to sound. The alarm sounds because we may be making decisions by mistake because we are using p-values incorrectly. The p-values reported in a regression are premised on the idea that we make a single test. We are in fact making a series of comparisons. If we have an 0.05 chance of finding an illusory difference in one comparison, and then we make 20 comparisons, the chance of making a mistake across the whole exercise is $.95^{20} = .3584$. This high "experiment wise" error rate should be taken into account.

I have been studying several approaches to deal with the experiment wise error rate. Staying close to the current customs of regression modeling, we can estimate the unrestricted categorical model and then use properly calculated post hoc hypothesis tests to group together the redundant categories. Until recently, post hoc tests were considered only for ANOVA models (categorical predictor with numerical output). While I cannot say post hoc hypothesis testing is a completely "solved" problem, it seems safe to say that there are competing strategies that work in some cases. Hothorn et al. (2008) have proposed a method of post hoc hypothesis testing that can be applied to most regression models (anything that has parameter estimates that are asymptotically Normal, such as generalized linear models or maximum likelihood estimates). They demonstrate that it is possible to, for example, calculate comparisons for equality among the estimates for all categories of a nominal variable with an extension of Tukey's HSD test.

[Imagine a worked example here]

A more substantial break with current customs would use the regression (or classification) trees (Breiman et al., 1984) and probably the random forest (Breiman, 2001). In a nutshell, here is the idea. Subdivide a sample into two groups and for each group make a prediction for the output variable. Then search for a way to subdivide one of the groups that improves the overall match between the observed outcomes and the predictions. Continue separating

the subgroups (creating new branches in the "tree") until it is not possible to find more subdivisions. The process depends on user-specified parameters, especially the guidelines for specificity and homogeneity of subgroups. There are several implementations of the regression tree, including a commercial program CART and packages for R that rely on the same underlying algorithms and, in some cases, computer code (Therneau and Atkinson, 2009;Zeileis et al., 2008; Hothorn et al., 2006). These are called "recursive trees" because the algorithm that separates the observations considers one variable at a time, but it seems to me this is not intrinsic to the development of an end result: prediction from subgroups.

There is a danger of "over-fitting". If we trim down a model on the basis of a single sample, and then estimate the final model with the same sample, there is a good chance that the estimates will reflect sample-specific quirks rather than population characteristics. To deal with the problem of over-fitting, we may try some of the strategies for model validation (estimation on bootstrapped subsamples, e.g.), but it appears that the random forest approach is likely to win out over the long run (Breiman, 2001). This is easier to describe for a dichotomous output variable. Suppose we repeatedly draw random samples out of our data and create the "best" fitting model for each. Because we draw many samples, the quirks of an individual sample will not dominate when we turn to consider the whole ensemble of models. Each subsample's model makes a prediction for each observation, and our overall prediction for each case is the outcome of a 'majority vote' among the models. Researchers are actively engaged in the project of developing diagnostic tools for these trees, perhaps it is too soon for us to say for sure about the long run (Zeileis et al., 2008;Strobl et al., 2007). Nevertheless, this approach is producing useful results in "real life" research decisions in which important decisions are made. For example, this approach is used in bioinformatics to attempt to discern which genes are linked to diseases, for example.

[Please Imagine an illustration of a regression tree and a random forest regression at this point.]

5 Conclusion

Researchers should be more conscious about "throwing" categorical predictors with numeric codings into regression models. While it is not always "wrong" to do so, neither is it "right."

The best practice for model fitting should estimate two models, the numeric coding and the categorical coding, and then apply a formal hypothesis test to answer the question "is the numeric coding wrong?" If the numeric coding is wrong, then some kind of categorical coding scheme should be used. In the worst case scenario, it is necessary to estimate $M - 1$ parameters for an input variable that has M categories. It may be that several of the categories cannot be differentiated, and so grouping observations from those categories may offer considerable simplification. Post hoc hypothesis testing is now available for generalized linear models.

References

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The new {S} language: a programming environment for data analysis and graphics*. Wadsworth and Brooks/Cole Advanced Books & Software, Monterey, CA, USA.

Boulesteix, A.-L. (2007). *exactmaxsel: Maximally selected statistics for binary response variables - Exact methods*. R package version 1.0-2.

Box, G. E. P. and Cox, D. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26:211–252.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, New York.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-Weighted Fitting: An Approach to Fitting Analysis by Local Fitting. *Journal of the American Statistical Association*, 83:596–610.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. In Chambers, J. and Hastie, T., editors, *Statistical Models in S*, pages 309–376. Wadsworth & Brooks/Cole.

Davis, J. (2007). *General Social Surveys, 1972-2006*. Chicago: National Opinion Research Center.

Fox, J. (2009). *car: Companion to Applied Regression*. R package version 1.2-12.

Harrell, Frank E., J. (2008). *Design: Design Package*. R package version 2.1-2.

Hastie, T. (2008). *gam: Generalized Additive Models*. R package version 1.0.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hothorn, T. and Zeileis, A. (2008). Generally Maximally Selected Statistics. *Biometrics*, 64:1263–1269.

Loader, C. (2007). *locfit: Local Regression, Likelihood and Density Estimation.* R package version 1.5-4.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Core team (2008). *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-90.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Spector, P., Friedman, J., Tibshirani, R., and Lumley, T. (2009). *acepack: ace() and avas() for selecting regression transformations.* R package version 1.3-2.2.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8(25).

Therneau, T. M. and Atkinson, B. (2009). *rpart: Recursive Partitioning.* R Port by Brian Ripley, R package version 3.1-43.

Tibshirani, R. (1988). Estimating Transformations for Regression Via Additivity and Variance Stabilization. *Journal of the American Statistical Association*, 83(402):394–405.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wood, S. N. (2003). Thin-plate Regression Splines. *Journal of the Royal Statistical Society-B*, 65(1):95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.