Journal of Behavioral Data Science, 2023, *3* (2), 15–42.
DOI: https://doi.org/10.35566/jbds/v3n2/montoya

Lasso and Group Lasso with Categorical Predictors: Impact of Coding Strategy on Variable Selection and Prediction

Yihuan Huang, Tristan D. Tibbe[0000−0003−0684−8304], Amy Tang, and Amanda K. Montoya[0000−0001−9316−8184]

University of California, Los Angeles, USA
akmontoya@ucla.edu

Abstract. Machine learning methods are being increasingly adopted in behavioral research. Lasso regression performs variable selection and regularization, and is particularly appealing to behavioral researchers because of its connection to linear regression. Researchers may expect properties of linear regression to translate to lasso, but we demonstrate that this assumption is problematic for models with categorical predictors. Specifically, we demonstrate that while the coding strategy used for categorical predictors does not impact the performance of linear regression, it does impact lasso's performance. Group lasso is an alternative to lasso for models with categorical predictors. We investigate the discrepancy between lasso and group lasso models using a real data set: lasso performs different variable selection and has different prediction accuracy depending on the coding strategy, while group lasso performs consistent variable selection but has different prediction accuracy. Using a Monte Carlo simulation, we demonstrate a specific case where group lasso tends to include many variables when few are needed, leading to overfitting. We conclude with recommended solutions to this issue and future directions of exploration to improve the implementation of machine learning approaches in behavioral science. This project shows that when using lasso and group lasso with categorical predictors, the choice of coding strategy should not be ignored.

Keywords: Lasso regression · Categorical predictors · Regularization

1 Introduction

Many behavioral research questions involve categorical predictors, including education, ethnicity, religion, gender, or experimental conditions. Unlike numerical predictors, which typically have a natural scale, to be included in statistical models categorical predictors require researchers to select a method for encoding these variables (i.e., representing the categories using a numeric system). Thus, a single categorical predictor can be represented in a model using different sets of variables, each set embodying the same predictor but representing different contrasts of the categories. This special property of categorical predictors motivates our exploration of categorical predictors in the case of linear regression and two machine learning algorithms: least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) and group lasso regression (Yuan & Lin, 2006). We explore both variable selection and prediction accuracy for these models and how they are impacted by using different coding strategies for categorical predictors using a real-world data set.

We use a data set focusing on stress during COVID-19 as the primary outcome, measured in over 100,000 participants (Yamada et al., 2021). The *stress* score is an aggregated score from the Perceived Stress Scale (PSS-10) on a 1-5 scale. The data set includes categorical predictors, such as *Education*, *Gender* and *Marital Status*, and continuous predictors, such as *Age* and *Trust in the Country*. The overall goal is to predict participant's *Stress* using the available predictors.

In the remainder of this section, we introduce the three analytical approaches examined in this paper: linear regression, lasso regression, and group lasso regression. We focus on the application of these methods with a continuous outcome and one or more categorical predictors. After introducing these methods, we demonstrate their use with the applied example, exploring peculiar behavior of the machine learning approaches that does not occur with linear regression.

1.1 Linear Regression With Categorical Predictors

Categorical predictors need to be encoded into a set of variables to be included in regression models. Different coding strategies can be implemented, such as dummy, contrast, sequential, or Helmert coding. Tables 1–4 show different ways to encode a categorical variable, *Education*, with 7 categories (no education, up to 6 years of school, up to 9 years of school, up to 12 years of school, some college or equivalent, college degree, PhD/doctorate). Dummy coding uses only 0's and 1's to indicate category membership. One category is selected as the *reference category* (or *reference group*) and is assigned a score of 0 on all indicators. For other categories, only the indicator corresponding to the category is coded as 1 and all other indicators are set to 0 (Table 1). Contrast coding is similar to dummy coding, but the reference category which is coded as all 0 in dummy coding is now coded with all -1 instead, changing the interpretation of the intercept and slope coefficients (Table 2). Sequential coding compares each category to the previous category (Table 3), while Helmert coding examines how each category is compared to the average of all subsequent categories (Table 4). Note that if a categorical variable has k categories, $k-1$ indicators are needed, regardless of the coding strategies used. This type of design matrix is defined as nonsingular because the matrix is invertible. The design matrix has to be nonsingular for linear regression but this is not necessarily the case for lasso or group lasso. In Appendix B we discuss singular matrix options for lasso regression.

In linear regression, each coding scheme represents categories using a different numerical system, which leads to different interpretations of their coefficients. However, each coding scheme always predicts the category mean for each category (or adjusted means if covariates are included), and the explained variance is the same regardless of coding choice (Darlington & Hayes, 2016). Therefore, researchers can choose coding

Table 1: Dummy Coding

Education	D_1	D_2	D_3	D_4	D_5	D_6
1. no education	0	0	0	0	0	0
2. up to 6 years of school	1	0	0	0	0	0
3. up to 9 years of school	0	1	0	0	0	0
4. up to 12 years of school	0	0	1	0	0	0
5. some college or equivalent	0	0	0	1	0	0
6. college degree	0	0	0	0	1	0
7. PhD/doctorate	0	0	0	0	0	1

Note. No education is selected as the reference group (coded 0 on all indicators) and every other category scores 1 on a single indicator and 0 on all other indicators.

Table 2: Contrast Coding

Education	C_1	C_2	C_3	C_4	C_5	C_6
1. no education	1	0	0	0	0	0
2. up to 6 years of school	0	1	0	0	0	0
3. up to 9 years of school	0	0	1	0	0	0
4. up to 12 years of school	0	0	0	1	0	0
5. some college or equivalent	0	0	0	0	1	0
6. college degree	0	0	0	0	0	1
7. PhD/doctorate	-1	-1	-1	-1	-1	-1

Note. PhD/doctorate is selected as the omitted category (coded -1 on all indicators) and every other category scores 1 on a single indicator and 0 on all other indicators.

strategies among all these options according to their needs without concern about model performance. Dummy and contrast coding are often used for nominal categorical variables, while sequential and Helmert coding are particularly helpful when categories are ordered.

When using different coding strategies, the regression coefficients have different interpretations. For example, a researcher might want to know whether *Stress* during the COVID-19 pandemic can be predicted by *Education*. The seven categories within the variable *Education* are encoded by 6 indicators. Linear regression fits the following model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i, \tag{1}$$

where Y_i is the outcome value for the i^{th} observation (person), X_{ji} is the j^{th} variable to convey category membership for the i^{th} observation, and ε_i is the error term for the i^{th} observation. Equation 1 is the general equation for all coding strategies. If different coding strategies are used, the intercept β_0 and coefficients for different indicators, β_1 through β_6, have different meanings. For example, suppose the fitted linear regression model (with $\hat{Y}_i$ representing the predicted value for the i^{th} observation) is

$$\hat{Y}_i = 2 + 0.3X_{1i} + 1.5X_{2i} + 0.2X_{3i} + 0.5X_{4i} - 0.2X_{5i} - 0.4X_{6i}. \tag{2}$$

The interpretation of these coefficients would depend on which coding strategy was used. If dummy coding was used with no education as the reference group (as in Table

Table 3: Sequential Coding

Education	S_1	S_2	S_3	S_4	S_5	S_6
1. no education	0	0	0	0	0	0
2. up to 6 years of school	1	0	0	0	0	0
3. up to 9 years of school	1	1	0	0	0	0
4. up to 12 years of school	1	1	1	0	0	0
5. some college or equivalent	1	1	1	1	0	0
6. college degree	1	1	1	1	1	0
7. PhD/doctorate	1	1	1	1	1	1

Note. The lowest category scores 0 on all indicators. Each subsequent category scores 1 on one more indicator than the previous.

Table 4: Helmert Coding

Education	H_1	H_2	H_3	H_4	H_5	H_6
1. no education	-6/7	0	0	0	0	0
2. up to 6 years of school	1/7	-5/6	0	0	0	0
3. up to 9 years of school	1/7	1/6	-4/5	0	0	0
4. up to 12 years of school	1/7	1/6	1/5	-3/4	0	0
5. some college or equivalent	1/7	1/6	1/5	1/4	-2/3	0
6. college degree	1/7	1/6	1/5	1/4	1/3	-1/2
7. PhD/doctorate	1/7	1/6	1/5	1/4	1/3	1/2

Note. The lowest indicator scores $-(k-1)/k$ on the first indicator and 0 on all subsequent indicators. The next highest scores $1/k$ on the first indicator, $-(k-2)/(k-1)$ on the second indicator, and 0 on all subsequent indicators. The next highest scores $1/k$ on the first indicator, $1/(k-1)$ on the second indicator, $-(k-3)/(k-2)$ on the third indicator, and 0 on all subsequent indicators. And so on.

1), we would interpret the coefficient for X_4, 0.5, as the difference between the average stress score of individuals with no education and the average stress score with some college education. However, if contrast coding was used (as in Table 2), 0.5 would indicate the difference between the average stress score of individuals with up to 12 years of school and the average score of all categories. If sequential coding was used (as in Table 3), 0.5 would be interpreted as the difference between the average stress score of individuals with some college education and the average stress score of individuals with up to 12 years of school. If Helmert coding was used (as in Table 4), 0.5 would indicate that on average individuals with up to 12 years of school are 0.5 points less stressed than the average of those who have some college education, those who have a college degree and those who have a PhD/Doctorate. The interpretations of the coefficients are inseparable from the coding strategy used.

Different selections of reference categories in dummy and contrast coding and ordering of categories in Helmert and sequential coding can also produce coefficients with different meanings. For example, if no education is the reference category for dummy coding, β_0 represents the average stress score for people with no education and β_1 through β_6 will represent the difference between no education and the corresponding coded category. On the other hand, if up to 6 years of school is the reference category, β_0 represents the average stress for individuals with up to 6 years of school, and β_1 through β_6 will represent the difference between “up to 6 years of school” and the corresponding coded category.

Though different ways to code categorical variables produce different model coefficients, they do not affect the predictions/prediction accuracy of linear regression. To demonstrate that linear regression with a categorical predictor will predict the same category means for each coding scheme, we used *Education* to predict *Stress*. We randomly sampled 10,000 participants from the COVID-19 Stress Data (Yamada et al., 2021) to serve as our sample data set, and then we randomly split our sample into training (80%) and test (20%) data. Next, we fit linear regression on the training data set with four different coding strategies from Tables 1 - 4 applied to the variable *Education*. Table 5 contains the model coefficients.

Table 5: Linear Regression Example for Coding

Coefficient	**Dummy**	**Contrast**	**Sequential**	**Helmert**
β_0	2.852	2.955	2.852	2.955
β_1	0.031	-0.103	0.031	0.121
β_2	0.110	-0.072	0.079	0.107
β_3	0.145	0.007	0.035	0.036
β_4	0.161	0.041	0.016	0.001
β_5	0.138	0.058	-0.023	-0.022
β_6	0.139	0.035	0.001	0.001

Note. Each column of the table represents one coding strategy and rows represent the coefficients of the indicator X_j for each coding strategy.

Using the values of X_1–X_6 from Table 1–4 and the coefficient estimates from Table 5, we reconstruct the predicted score (i.e., category mean) for the “some college or equivalent” category for dummy, contrast, sequential, and Helmert coding respectively.

$$2.852 + 0.031(0) + 0.110(0) + 0.145(0) + 0.161(1) + 0.138(0) + 0.139(0) = 3.013 \qquad (Dummy)$$

$$2.955 - 0.103(0) - 0.072(0) + 0.007(0) + 0.041(0) + 0.058(1) + 0.035(0) = 3.013 \qquad (Contrast)$$

$$2.852 + 0.031(1) + 0.079(1) + 0.035(1) + 0.016(1) - 0.023(0) + 0.001(0) = 3.013 \qquad (Sequential)$$

$$2.955 + 0.121(\frac{1}{7}) + 0.107(\frac{1}{6}) + 0.036(\frac{1}{5}) + 0.001(\frac{1}{4}) - 0.022(-\frac{2}{3}) + 0.001(0) = 3.013 \qquad (Helmert)$$

The predicted score for “some college or equivalent” using dummy coding is the same as that for contrast, sequential, and Helmert coding. Following a similar procedure, it can be shown that all predicted scores match the category means for each coding strategy (Cohen, Cohen, West, & Aiken, 2003; Darlington & Hayes, 2016).

Since predicted scores are the same across coding strategies in linear regression, this means prediction accuracy is also the same across the different coding strategies. In our example data, prediction accuracy quantifies how far a model’s predicted *stress* scores are from the observed *stress* scores of participants in the test data. We use Mean Squared

Error (MSE) to measure the prediction accuracy. Mathematically, MSE is calculated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{3}$$

where n represents the number of observations in the test data; Y_i represents the observed outcome value of the i^{th} observation in the test data; and $\hat{Y}_i$ represents the predicted outcome value of the i^{th} observation from the model (which is generated using the training data). When we calculate the MSE of four linear regression models each fit using one of the four coding strategies mentioned previously, we find that all models have the exact same MSE of 0.13674. This illustrates that prediction accuracy is not affected by coding strategy when using linear regression.

While these results may seem trivial and require only a basic understanding of linear regression to understand, they stand in stark contrast to similar results we will examine in alternative regularized regression approaches. In summary, linear regression models with different coding strategies predict the same scores (i.e., category means) and give the same prediction accuracy, though they produce different coefficients. These properties persist when there are additional predictors (categorical and/or continuous) in the model, where the predicted scores (which are now *adjusted means*) are the same for all coding strategies, and thus prediction accuracy is always the same as well.

1.2 Lasso and Group Lasso Regression

In contrast to linear regression, lasso regression is useful when the proposed model involves many predictors, but only a few may be true predictors of the outcome (i.e., sparsity). Lasso is gaining popularity in behavioral science presumably because it shares many properties with linear regression, an already common statistical approach in the field (McNeish, 2015). For example, a lasso model fit to the COVID-19 data using *Education* to predict *Stress* would share the same equation as linear regression given in Equation 1. However, the values of the β_j coefficients would differ between the two methods because linear and lasso regression differ in the way they estimate the vector containing these regression coefficients, β. In linear regression, the estimated coefficient vector is calculated as follows,

$$\hat{\beta}_{linear} = \underset{\beta}{\operatorname{argmin}}(|Y - X\beta|_2^2), \tag{4}$$

where $|\cdot|_2$ is the notation for the L2 norm. Lasso, on the other hand, adds a penalty term governed by the penalty parameter λ to regulate the size of the coefficients:

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}}(|Y - X\beta|_2^2 + \lambda|\beta|_1), \tag{5}$$

where $|\cdot|_1$ is the notation for the L1 norm.[1] When λ is nonzero, nonzero values of β result in increases in $\lambda|\beta|_1$, and so Equation 5 reaches its minimum when both the

[1] Another alternative to lasso is ridge regression which is expressed by Equation 5 except with an L2 norm instead of an L1 norm for the regularization term. In Equation 5, the L1 norm

prediction error and the size of the elements of β are considered. A large λ value results in the coefficients in β being shrunk toward or equal to zero so fewer predictor variables are selected in the model (where "selected" means that the coefficient is nonzero in the final solution). A small λ value, on the other hand, results in less shrinkage so more predictor variables can be selected into the model. Linear regression is actually a special case of lasso regression when λ is set to zero.

While lasso has many benefits over linear regression (Hastie & Tibshirani, 2018; McNeish, 2015; Tibshirani, 1996), when applying lasso regression to models with categorical predictors, additional considerations must be made. Lasso regression models select variables based on the penalty parameter λ and the sizes of the entries in coefficient vector β. However, as we demonstrated with linear regression, using different coding strategies for a categorical predictor creates models with different coefficient vectors. This means that the choice of coding strategy may result in different variable selection in lasso regression models. The issue of coding strategies is related to the issue of variable scaling with continuous predictors, which also influences variable selection and prediction accuracy in lasso regression models. One common solution to this problem is to standardize all continuous predictors before applying lasso regression (Marquardt, 1980). In this way, the effect of scaling is excluded from the variable selection of lasso regression with continuous predictors. While dichotomous variables can be standardized, different coding strategies representing more than two categories do not result in the same standardized solution. Given this, there is reason to believe that the performance of lasso regression with categorical variables may be impacted by the choice of coding strategies for those variables.

A generalization of lasso regression which may also be impacted by coding strategy—but in different ways—is group lasso regression. Group lasso, as opposed to lasso, performs variable selection by selecting groups of variables rather than individual variables (Yuan & Lin, 2006). This is particularly valuable for the case of categorical predictors because the set of indicators for each variable forms a natural group. The mathematical formula for estimating the coefficient vector β in group lasso is

$$\hat{\beta}_{group} = \underset{\beta}{\operatorname{argmin}}(|Y - X\beta|_2^2 + \lambda \sum_{g=1}^{G} |\beta_{I_g}|_2) \tag{6}$$

where G represents the number of groups of variables, and β_{I_g} represents the coefficient vector of that corresponding group. Other notation is the same as Equation 5. Using the L2 norm within each group g is what allows group lasso to either select all or none of the variables within each group. Also, multiplying by λ after summing the L2 norms of all groups penalizes each group instead of each individual indicator variable. These differences provide group lasso with distinct properties: When all variables are considered one group, group lasso performs as ridge regression. On the other hand, when all the variables are their own group, group lasso performs as lasso regression.

penalizes the absolute value of the coefficients, used by lasso; while in ridge regression, the L2 norm penalizes the squares of all coefficients. Given this property, ridge regression is not as effective at penalizing parameters to zero compared to lasso regression (Tibshirani, 1996). Therefore, lasso regression is preferred for variable selection.

The advantage of group lasso is that when there are multiple groups of more than one variable, the result is a combination of within-group ridge regression and across-group lasso regression.

The group lasso has special properties with respect to variable selection. Within a group, group lasso typically includes or excludes all variables because of the within-group ridge regression. Given its unique properties with respect to variable selection, group lasso has been recommended as a useful alternative to lasso regression when dealing with models with categorical variables (Detmer, Cebral, & Slawski, 2020; McNeish, 2015); however, no prior research has explored the sensitivity of group lasso to different coding strategies. In group lasso, all indicators for a categorical variable are defined as a group, and the algorithm should either include all indicators associated with one categorical predictor or exclude all these indicators.

1.3 Motivation

With the increasing use of lasso techniques across scientific fields, but especially within the social and behavioral sciences, many researchers rely on their intuitions about the similarities between lasso and linear regression to understand, use, and interpret the results of lasso regression. This could be particularly problematic for models with categorical predictors. Prediction accuracy in linear regression is unaffected by the selection of coding strategy; however, lasso regression conducts regularization by minimizing regression coefficients, which differ across coding strategies. This may lead to different prediction accuracy and variable selection depending on the coding strategy used when using lasso. Since group lasso treats the variables in a group as a whole set, it seems less likely that its variable selection will be impacted by the choice of coding strategy. However, the prediction accuracy of group lasso may still be impacted by the coding strategy.

To explore the potential impacts of coding strategy on important characteristics of lasso and group lasso regression, we combine both real data analysis and simulation. First, using the COVID stress data set described previously, we demonstrate the use of lasso and group lasso regression with categorical variables, where different coding strategies of categorical variables impact two aspects of model performance: variable selection and prediction accuracy. Next, we use a Monte Carlo simulation to demonstrate a specific case where group lasso may tend to overfit the training data. In the last section, we explore other potential solutions, important future directions, and general conclusions.

2 Real Data Analysis with COVID Stress Data

We used the COVID stress data set with the same sample of 10,000 participants and the same training/test data sets used in Section 1.1 to explore how coding strategies affect models estimated by lasso and group lasso. In the models, we included six categorical predictors (where a predictor with k categories was represented by $k-1$ indicator variables): *Education* (7 categories), *Employment status* (6 categories), *Gender* (3 categories), *Isolation status* (4 categories), *Marital status* (4 categories), and *Mother's* *education* (7 categories). We also included seven continuous predictors in the models. Thus, after coding all categorical variables and adding the seven continuous variables, the models predicted the outcome *Stress* with $6+5+2+3+3+6+7=32$ variables. In total, we trained eight different models using lasso and group lasso with four coding strategies: dummy, contrast, sequential, and Helmert. We used 10-fold cross-validation on the training data to select the penalty parameter from the model with the best prediction accuracy, so the penalty parameter that was selected is different across models with different coding strategies.[2] We then examined if the variable selection and prediction accuracy of these lasso and group lasso models were affected by the choice of coding strategy.

2.1 Variable Selection

We first examined differences in the variable selections of the four lasso models. Results are shown in Table 6. Focusing on the *Education* variable, we illustrate how the use of different coding strategies can result in conflicting findings. Both the dummy coding model and the sequential coding model have a predictor which represents the difference between no education and 6 years of education. After applying lasso, the dummy coding model includes this predictor, whereas the sequential coding model excludes this predictor. Based on these results, using the dummy coded model, a researcher might conclude that COVID stress differs across the no education and 6 years of education groups, whereas using a sequential coded model, the opposite conclusion would be made.

Fitting similar dummy-, contrast-, sequential-, and Helmert-coded models with group lasso, we found that the results differed notably from the traditional lasso. While lasso's variable selection was affected by the choice of coding strategy (see Table 6), the group lasso's variable selection seemed stable across different coding strategies, with all predictor variables selected to remain in all four models. Thus, based on the applied data analysis, it seems that variable selection is not impacted by the coding strategy for group lasso, though this should be subject to additional investigation. This suggests that if researchers are interested in using lasso for variable selection and have categorical predictors, using group lasso could avoid the arbitrary choice of coding strategy. However, group lasso was not successful in reducing the set of potential predictors, and thus, it may suffer from a limitation of being overly inclusive. We explore this issue more in a simulation.

2.2 Prediction Accuracy

In this section, we investigate whether prediction accuracy is affected by the choice of coding strategy using both lasso and group lasso. We examined the prediction accuracy in two ways: predicted category scores and MSE of the model applied to the test data set.

[2] Note that even with the same penalty parameter, models with different coding strategies or reference categories will still have different variable selection and prediction accuracy.

Table 6: Variable Selection for Different Coding Strategies by Lasso

	Lasso Regression			
Variable	*Dummy*	*Contrast*	*Sequential*	*Helmert*
Education	6 years - no	no - Average	**6 years - no**	no - Average(6 years and more)
	9 years - no	6 years - Average	9 years - 6 years	6 years - Average(9 years and more)
	12 years - no	**9 years - Average**	12 years - 9 years	**9 years - Average(12 years and more)**
	some college - no	12 years - Average	some college - 12 years	12 years - Average(some college, college, PhD)
	college - no	some college - Average	college - some college	some college - Average(college + PhD)
	PhD - no	**college - Average**	PhD - college	college - PhD
Employment Status	part-time - no	no - Average	part-time - no	no - Average(part-time, self-employed, student, full-time, retired)
	self-employed - no	**part-time - Average**	self-employed - part-time	**part-time - Average(self-employed, student, full-time, retired)**
	student - no	self-employed - Average	student - self-employed	self-employed - Average(student, full-time, retired)
	full-time - no	student - Average	full-time - student	student - Average(full-time, retired)
	retired - no	full-time - Average	retired - full-time	full-time - retired
Gender	man - woman	woman - Average	man - woman	woman - Average(man, other)
	other - woman	man - Average	other - man	man - other
Isolation Status	minor changes - usual	usual - Average	minor changes - usual	usual - Average(minor changes, isolated, medical isolated)
	isolated - usual	**minor changes - Average**	isolated - minor changes	minor changes - Average (isolated, medical isolated)
	medical isolated - usual	isolated - Average	medical isolated - isolated	isolated - medical isolated
Marital Status	divorced - single	single - Average	divorced - single	single - Average(divorced, married, other)
	married - single	divorced - Average	married - divorced	divorced - Average(married, other)
	other - single	married - Average	other - married	married - other
Mom's Education	6 years - no	no - Average	6 years - no	no - Average(6 years and more)
	9 years - no	6 years - Average	9 years - 6 years	6 years - Average(9 years and more)
	12 years - no	9 years - Average	12 years - 9 years	9 years - Average(12 years and more)
	some college - no	12 years - Average	some college - 12 years	12 years - Average(some college, college, PhD)
	college - no	some college - Average	college - some college	some college - Average(college, PhD)
	PhD - no	college - Average	PhD - college	college - PhD

Note. Variables with a white background color were selected to be in the model, and variables with a grey background color were not selected.

Predicted Category Scores We first examined whether the predicted *stress* score for each *Education* group is the same with different coding strategies in lasso and group lasso models. In this section, We generated the predicted score for each category using a model with only *Education* as a predictor, so the models contained 6 indicator variables in total. While this model is oversimplified, it eases the direct comparison between the true means of each group and the predicted scores.

The predicted category scores for lasso models fit using the four different coding strategies are shown in Table 7, with the final column providing the actual category means for *Education* observed in the training data. First off, it is important to note that category scores shown in the table were rounded. Thus, some category scores that were very close to the actual category scores were rounded to the same value, but there were no lasso models where the predicted scores were exactly equal to the group means like they would have been in linear regression. Also, it is evident in the table that the predicted means often differ depending on the coding strategy used. For five of the seven categories, the dummy-coded model estimated the category mean most accurately among all models.

Table 7: Predicted Category Scores for Different Coding Strategies by Lasso

	Dummy	*Contrast*	*Sequential*	*Helmert*	Training Mean	Test Mean
None	2.864	2.879	**2.864**	2.872	2.852	2.912
6 years	**2.883**	2.897	2.888	2.896	2.883	2.824
9 years	**2.962**	2.961	2.965	2.973	2.962	2.857
12 years	2.997	2.995	2.999	**2.997**	2.997	3.038
Some college	**3.013**	3.012	3.008	3.008	3.013	3.009
College	**2.990**	2.990	2.991	2.991	2.990	2.999
PhD/Doctorate	**2.991**	2.992	2.991	2.991	2.991	3.008

Note. Rows represent *Education* categories, and the middle four columns give the model predicted values with different coding strategies. The last two columns give the actual mean of each category observed in the training and test data, respectively. The closest value to the training mean is bolded and the closest value to the test mean has a grey background color in each row.

The results of the predicted category scores for the four group lasso models, shown in Table 8, are very similar to the lasso models: Group lasso estimated each category score within a categorical variable differently depending on the coding strategy used. Thus, although variable selection is not impacted by the coding strategy used for group lasso, the predicted category score *is* impacted by the choice of coding strategy. Also, among all group lasso models, the dummy-coded model generated the most accurate category scores for four of the seven categories. Thus, regardless of whether lasso or group lasso was used, the dummy-coded model estimated the majority of the category means better than the other three models. It is unclear whether this finding would remain true with other data sets, however.

The results in Table 7 and 8 show that different coding strategies result in different predicted category scores. While this is an important finding, it is equally important

Table 8: Predicted Category Means for Different Coding Strategies by Group Lasso

	Dummy	*Contrast*	*Sequential*	*Helmert*	Training Mean	Test Mean
None	2.828	**2.863**	2.879	2.868	2.852	2.912
6 years	**2.886**	2.892	2.906	2.894	2.883	2.824
9 years	2.965	**2.962**	2.954	2.963	2.962	2.857
12 years	**2.997**	2.996	2.994	2.996	2.997	3.038
Some college	**3.013**	3.012	3.011	3.012	3.013	3.009
College	**2.990**	2.990	2.991	2.990	2.990	2.999
PhD/Doctorate	2.991	2.991	**2.991**	2.990	2.991	3.008

Note. Same as Table 7.

to understand why this occurs and whether the degree of difference is predictable and understandable rather than random variability due to estimation. A core aspect of lasso and group lasso models is shrinkage: different coding strategies will result in different model intercepts and coefficients, because the degree of shrinkage is different across coding strategies.

To visualize the shrinkage effect of each coding strategy, we plotted the predicted scores from each lasso model along with each model's intercept in Figure 1. In the dummy-coded model, the predicted scores are all shrunk slightly toward the no education category score (since it is the intercept in this model) relative to the contrast-coded model, where the scores are instead all pulled closer to the grand mean (i.e., the model's intercept). The predicted scores from the sequential-coded and Helmert-coded models, on the other hand, are shrunk closer towards each other more than those from the dummy-coded or contrast-coded models, reflecting the fact that shrinkage in sequential coding and Helmert coding relies not on the intercept, but on the differences between neighboring categories or the average of multiple neighboring categories. For example, the 9 years and the some college categories are shrunk closer to the college category or Phd/Doctorate category in sequential-coded and Helmert-coded models. In summary, models fit with different coding strategies have different shrinkage patterns, and so predicted scores differ across these models, leading to different prediction accuracy. These results suggest that one way to select a coding strategy is to consider the pattern of shrinkage which seems most reasonable.

Model Fit Next, we recorded MSEs calculated from models including all six categorical variables and all seven continuous variables, to the test data set (Table 9). Model fit (MSE) differs by coding strategy for both lasso and group lasso. Contrast-coded models yielded the best MSE for both lasso and group lasso regression. This exposes uncertainty regarding which coding strategy should be used when lasso or group lasso regression is applied. While some differences in MSE are expected due to the stochastic nature of procedures like cross-validation used to choose the penalty parameter (λ), it is notable that the MSEs were more variable for the group lasso models than they were for the lasso models, suggesting that choice of coding strategy could result in a much less optimal model (possibly worse than linear regression) when using group lasso. We explore this issue more in the Monte Carlo simulation. In Appendix A, we

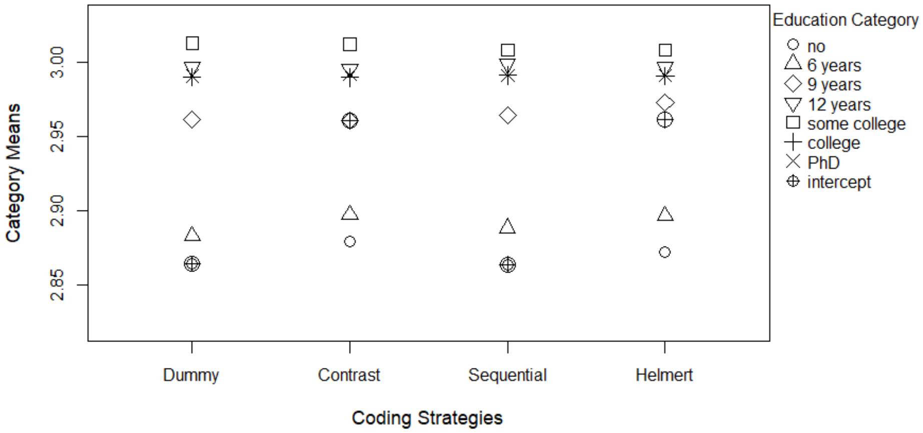


Figure 1: Graphical Presentation of Category Means for *Education* Recreated by Lasso Models with Different Coding Strategies. Intercept values are different across coding strategies. The intercept value is the estimated category mean for no education in dummy and sequential coding and the average of the category means in contrast and Helmert coding.

demonstrate similar issues with the choice of reference group or category order across different coding strategies, and in Appendix B we demonstrate that the use of singular design matrices (e.g., including dummy codes for all categories) does not ameliorate this issue.

Table 9: Model Fit (MSE) for Different Coding Strategies by Lasso and Group Lasso Regression

Coding strategies	*Dummy*	*Contrast*	*Sequential*	*Helmert*
Lasso Regression	0.13669	**0.13660**	0.13675	0.13677
Group Lasso Regression	0.13711	**0.13689**	0.13691	0.13695

Note. Rows represent different lasso methods, and columns represent models with different coding strategies. The lowest value (best prediction) in each row is bolded.

2.3 Summary

Choice of coding strategy has the potential to affect both variable selection and prediction accuracy in lasso regression models. As a result, depending on the coding strategy used, an analyst may end up with different variables included in their model, different predicted scores, and different prediction accuracy. With both the model's variable selection and predictive performance dependent on how categorical predictors are represented in the model, it is not a choice that should be taken lightly. Ideally, there would be a method which provides the same variable selection and the same predicted scores regardless of the coding strategy chosen.

Group lasso partly addresses the issues caused by the choice of different coding strategies in lasso regression, because group lasso's variable selection is not affected by the coding strategy used. Therefore, if researchers use group lasso to select which variables contribute to the outcome variable, they do not need to worry that different coding strategies may result in different conclusions. However, coding strategies still affect the prediction accuracy of group lasso models. Therefore, if researchers aim to predict the outcome variable by using group lasso regression, they need to be aware that different coding strategies can result in different prediction accuracy. In addition, because group lasso is selecting more variables into the model, the robustness of group lasso across coding strategies may come at the cost of prediction accuracy. Comparing the MSEs between the lasso models and group lasso models, the lasso models typically have lower MSE (i.e., better prediction accuracy) than group lasso.

This trade-off between prediction accuracy and robustness leads to some additional concerns about the group lasso. There seems to be a trade-off between including a *set* of predictors in a model, as compared to when a specific predictor. For example, if the average stress for all levels of education was the same except for those with PhDs, would group lasso still select the education set of variables into the model? Will the set of indicators for the categorical variables be selected if there is only one category that differs from the other categories within that variable? If this group is selected into the model, this means that many additional parameters would also be included to capture an effect that is only attributable to one indicator variable. Alternatively, if the group is not selected, then the predictive ability of the group lasso model may suffer. This problem does not occur with lasso, as it is able to include a single indicator variable to represent one category differing from the rest. Next, we explore this specific case and examine if group lasso's ability to include groups of variables leads to issues with overfitting.

3 Monte Carlo Simulation

In this section, we use a Monte Carlo simulation to explore a potential weakness of group lasso: overfitting. Group lasso may select more variables than necessary into the model, leading to larger variance and lower prediction accuracy. We explore a particularly extreme data generation case, where across all categories within one categorical variable, only one category differs from the rest. We call this category the *dominant* category and refer to all others as *non-predictive* categories. A non-predictive category is always used as the reference category in the analysis. While the simulation is much simpler than cases that would occur in real data analysis, it provides a clear demonstration of a pattern that is likely to occur and be problematic and hard to identify in more complex situations.

Simulation Method The data was generated such that the dominant category had a nonzero category mean, while non-predictive categories all had category means of zero.

All categorical variables were encoded using dummy coding. A second predictor variable was generated to follow a standard normal distribution. The outcome variable was created by adding the category mean, the value of the continuous variable, and a random error term drawn from a standard normal distribution. For optimal prediction, both the continuous predictor and the indicator variable which estimates the difference between the dominant category and other non-predictive categories should be included in the model, while the variables associated with non-predictive categories should not.

As previously mentioned, the number of categories within categorical predictors may affect how the coefficients are estimated and how the model selects predictors in group lasso. Therefore, we varied the number of non-predictive categories (2,3,4). To examine how the effect size would affect group lasso's prediction accuracy and variable selection, we also simulated different dominant category means (0.1, 0.2, 0.3). For each combination of number of categories and effect size, we randomly generated 500 data sets with a sample size of 1200.

For each data set, we first split the data set into training and test sets randomly based on an 8:2 ratio. Then we fit lasso and group lasso models with the same training data. We selected the penalty parameter using the same cross-validation methods used in previous sections. For each model, we calculated the MSE, whether the model included the dominant category, and whether the model included the non-predictive categories. We calculated the average prediction accuracy of each method as well as the proportion of models that included the dominant category and the proportion that included non-predictive categories across each condition. For group lasso, these two proportions were always the same because group lasso either includes or excludes all categories within the categorical predictor.

Simulation Results We first found that in all conditions lasso had a higher prediction accuracy than group lasso, indicated by lower MSEs (Table 10). Though the differences in MSE of lasso and group lasso were small, they were consistent across different conditions. Secondly, for both group lasso and lasso regression, when the number of non-predictive categories increased, the probability for models to include the dominant category decreased, but the probability for lasso was consistently greater than or equal to that for group lasso (Figure 3). This means that lasso is more likely to include the dominant category than group lasso across the number of non-predictive groups. Figure 2 shows that when the number of non-predictive categories stayed the same, the probability for group lasso to include non-predictive categories increased when the effect size increased, while the probability for lasso remained relatively flat. For both models, the probability of including non-predictive categories decreased as the number of non-predictive categories increased.

Returning to the potential issue of overfitting in group lasso, consider the case where the dominant group mean is large. Figure 2 shows that when the dominant group mean was 0.3, group lasso had a higher probability than lasso of including non-predictive categories. In this case, group lasso could overfit the data because group lasso was more likely to include categories that were not supposed to be in the model. This also explains group lasso's lower prediction accuracy than lasso in Table 10 when the dominant group mean was large.

Table 10: Differences in MSE of Lasso and Group Lasso Models for Monte Carlo Simulation

	Number of Non-predictive Categories		
Dominant Category Mean	2	3	4
0.1	0.0028	0.0029	0.0003
0.2	0.0020	0.004	0.0008
0.3	0.0029	0.0029	0.0030

Note. Values larger than zero mean that the MSE for group lasso is larger than the MSE for lasso.

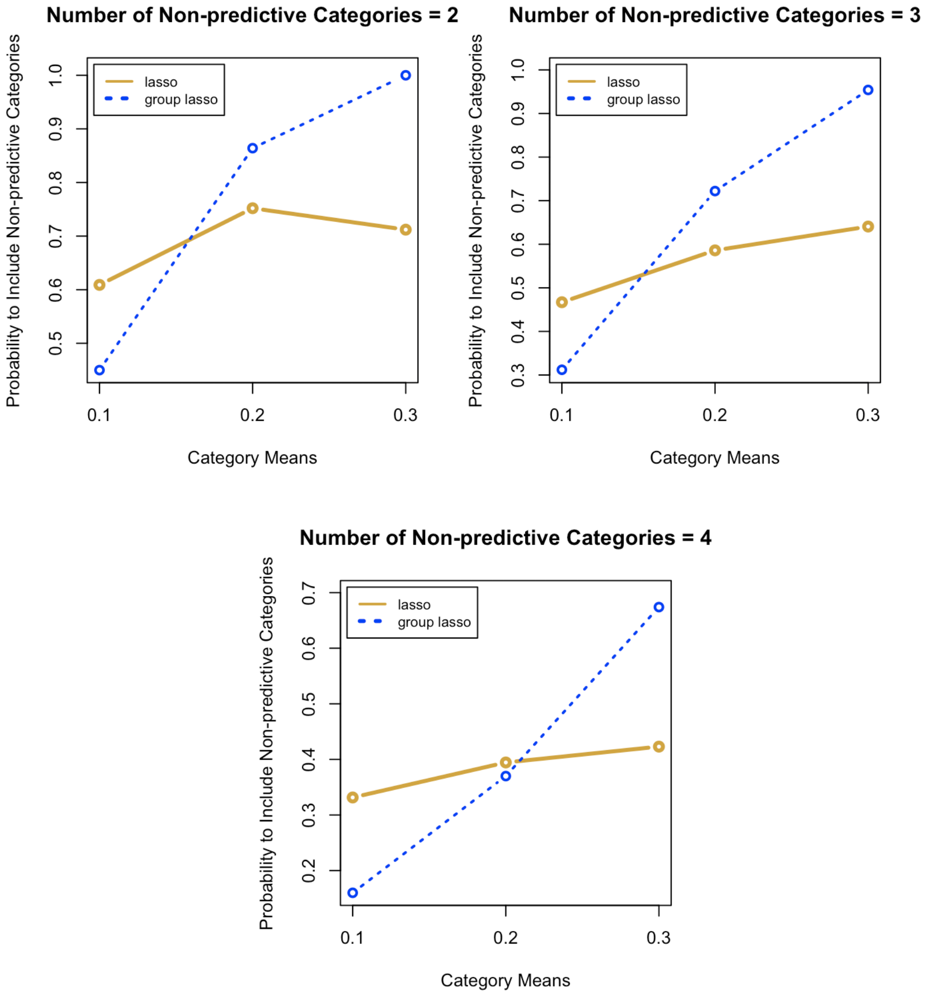
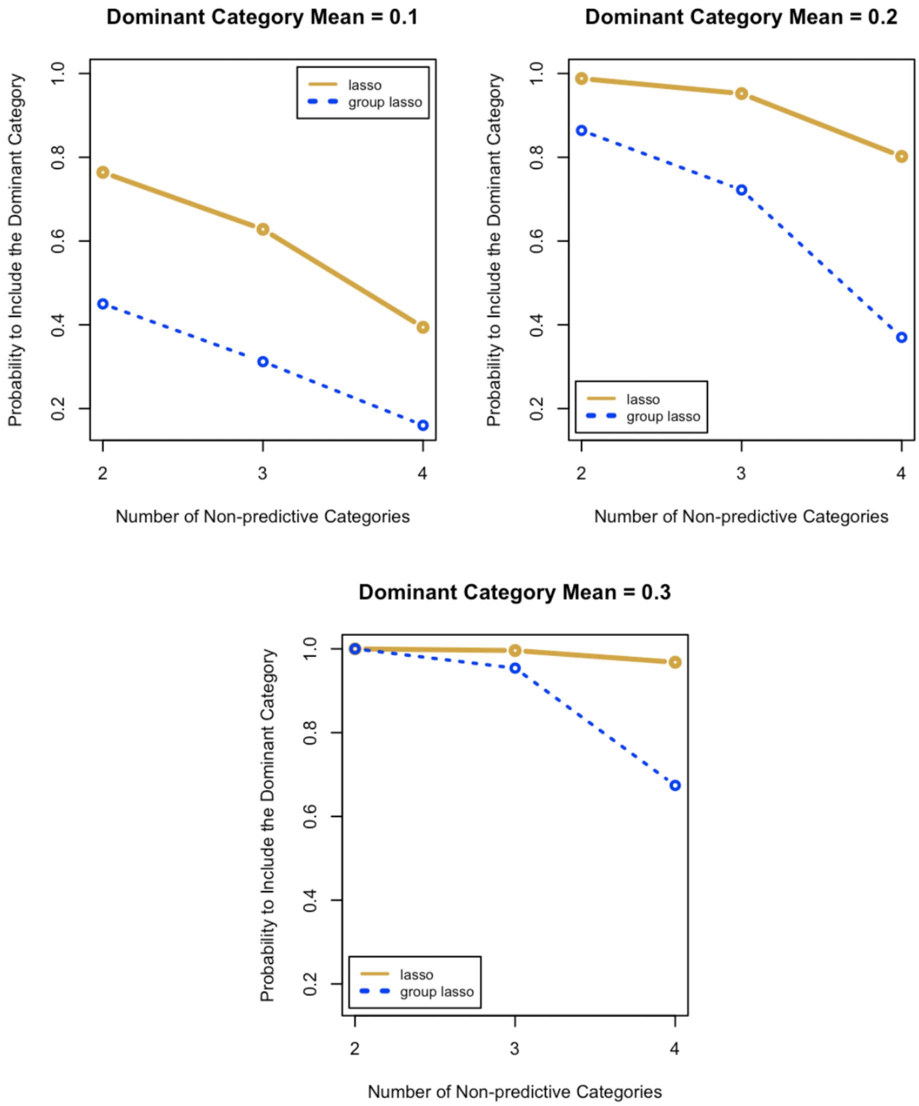


Figure 2: Comparison of Probabilities of Including Non-Predictive Categories under Different Numbers of Categories for Lasso and Group Lasso Models



Dominant Category Mean = 0.1

Dominant Category Mean = 0.2

Dominant Category Mean = 0.3

Figure 3: Comparison of Probabilities of Including the Dominant Category under Different Dominant Category Means for Lasso and Group Lasso Models

Simulation Summary Using Monte Carlo simulation, we demonstrated conditions under which group lasso may be likely to have issues with overfitting. When one or just a few categories differ from the rest, lasso may be more efficient with better prediction accuracy than group lasso. In these cases, group lasso is likely to include the categorical variable, including all non-predictive categories. Therefore, if researchers use group lasso to build predictive models, they may want to examine if one or two categories have relatively dominant means within categorical variables in advance, or if this pattern is hypothesized to occur they might prefer lasso. Looking for these effects may be particularly difficult in cases with many predictors where limited theoretical knowledge are driving the modeling, which is often the case when lasso is used. The differences must be conditional on all other variables in the data, not just examining the group means. If there are many categorical predictors in the model, exploratory analyses could be undertaken for each categorical variable, but this could a be tedious undertaking. Overall, this simulation demonstrates that there may be situations in which group lasso is not optimal for handling categorical predictors, especially if prediction accuracy is a high priority.

4 Discussion

In this paper, we demonstrate that lasso and group lasso models are sensitive to decisions about coding strategy for categorical predictors (e.g., dummy or sequential) and the choice of reference group/order of the categories (Appendix A). Linear regression does not have this problem, as the model fit and predicted values do not vary depending on the coding strategy. Group lasso presents a partial solution by having consistent variable selection across coding strategies. However, this consistency may come at a cost of reduced prediction accuracy. Ultimately, this leaves open the question of which coding strategy should be chosen. In the next section, we explore potential solutions to this issue with categorical predictors in lasso-based models.

4.1 Exploring Potential Solutions

Regardless of which of the following solutions researchers choose, one thing is always required: transparency. In searching the literature for examples of applications of lasso with categorical predictors, we found very few teams reported the coding strategy or order of categories used. Researchers using categorical variables in lasso or group lasso regression need to report how they coded the variables (both coding strategy and variable order/reference group) as this is imperative for reproducing or replicating their results. The following are a few proposed solutions, none of which seem satisfactory for all cases. As such, we weigh the pros and cons of each and consider cases when each approach might be most acceptable.

Prioritize Interpretability In cases where one coding strategy provides better interpretability of the model coefficients than another strategy, the most interpretable coding strategy could be chosen. This comes at the risk of having a worse predictive model, since the idea of interpretability is still very much rooted in the origins of inferential rather than predictive statistical models. In particular, because the coefficient estimates in lasso regression are biased, they should not be interpreted directly. Rather, after variable selection is completed, common recommendations are to fit a linear regression model that only includes the selected variables (Hastie, Robert, & Wainwright, 2015). It would be unusual to include a coding strategy in the follow-up linear regression that is different from the strategy used in the lasso regression. Thus, researchers should choose the coding strategy for each categorical variable that would be most interpretable if that variable was selected by a variable selection procedure to remain in the model. Coding schemes like Helmert coding require the presence of all predictors to have the intended interpretation, and should perhaps only be used in concert with group lasso (ensuring all predictors are selected in or out of the model) if interpretability is the top priority. Notably, machine learning approaches are often used in cases where there are many variables included in the analysis, and relatively little theory regarding which variables should be predicting the outcome. This could make it difficult for the researcher (or analyst) to decide which coding scheme would be “most interpretable,” especially considering the many possible combinations of coding schemes and variable orders/reference groups.

Prioritize Robust Variable Selection Based on the real data analysis and the simulation results, the group lasso is robust to coding strategy choices with respect to variable selection. Prediction accuracy is not necessarily optimized for the group lasso. However, when the goal is to select variables, and especially when it is conceptually useful to keep or drop all indicators for each categorical variable, group lasso seems to be an optimal choice. Nevertheless, this may come at a cost of prediction accuracy, particularly if categorical variables follow the dominant group pattern explored in the Monte Carlo simulation above, where one group is distinct from all other groups.

Prioritize Prediction Another option when estimating lasso or group lasso models would be to try many different coding strategies in order to select the one with the best overall prediction accuracy. This process should likely be completed using the training data so it does not influence the final prediction accuracy estimate acquired using an independent sample of the data. This approach can be very computationally intensive. With multiple categorical variables in the data set, trying different combinations of coding strategies would result in maximized prediction accuracy.

Notably, if prediction accuracy is of the highest priority, alternative machine learning approaches typically have higher prediction accuracy than lasso approaches, and many are robust to coding strategy. Techniques like classification and regression trees (CART) are unaffected by coding strategy because categorical predictors are treated as a single variable (Finch & Schneider, 2007). Realistically, researchers may be balancing their comfort with advanced analytic methods and their priority of prediction accuracy. CART methods do not provide the ”regression-like” estimates which many behavioral scientists rely on for interpreting their results.

4.2 Future Directions

There are several future directions we believe would be particularly beneficial for improving the state of research in the area of (group) lasso regression with categorical predictors. The first is the concept of intercept penalization. The typical practice within lasso is not to penalize the intercept (Wu & Lange, 2008), but the interpretation of the intercept varies greatly depending on which coding scheme is used. For example, when dummy coding is used, the intercept is the average of the reference group. Alternatively, when contrast coding is used, the intercept is the average of all groups. Ultimately, this means that different group means have differential penalization depending on the coding strategy used (as reflected in Figure 1). Thus, it is worth investigating whether penalizing the intercept may be appropriate in certain cases, and whether this would improve prediction accuracy (just as penalizing all other regression coefficients improves prediction accuracy in lasso). This question remains largely unexplored and would be informative to researchers who are interested in improving prediction accuracy.

Current defaults in software suggest that the field norm for coding strategy is dummy coding. The current research has demonstrated that dummy coding is a potentially risky choice as a default, as the choice of reference group can greatly impact the model, and the shrinkage is toward a group mean. Alternatively, contrast coding may make it an appealing default for researchers unsure about which coding strategy to use. Because the interpretation of the intercept for contrast coding is the average across all groups, the penalization of the groups is symmetric about this average. This means that when a coefficient is dropped from the model, the group that is indicated by this predictor is assumed to be equal to the grand mean. This method contrasts with dummy coding where all estimated group means are shrunk towards the reference group score. As a result, the selection of the reference group in contrast coding has less of an impact on parameter estimates than it does in dummy coding, because by selecting a reference group in dummy coding, that group's score is not at all penalized (if the intercept is not penalized). The interpretation of the intercept from contrast coding also aligns with how intercepts would be interpreted if there were no categorical variables in the model and all continuous predictors were standardized (i.e., sample average). Thus, contrast coding stands as a reasonable default if researchers are unsure of which coding strategy to choose; however, the use of contrast coding should be studied further in a variety of contexts to assess its appropriateness as a potential default.

Another observation our team made during this investigation was that group size mattered quite a lot with respect to how much predicted group scores varied across different coding strategies. In particular, in the COVID stress data, the no education group was particularly small ($N = 77$ out of 10,000 observations). This resulted in two problems that merit further investigation. The first is how group size can impact estimates and interact with the selection of coding strategy/reference group. Previous research by Choi, Park, and Seo (2012) has already shown that variability in the number of groups that categorical predictors contain can influence whether lasso or group lasso produces better prediction accuracy and recovery of model coefficients. As can be seen in Figure 1 and Table 7, the estimated means for the no education group in the COVID stress data were very unstable and varied more across coding strategies than any other group. Similarly, in Table 12 in Appendix A, we can see that the estimates of all of the *Education* group means have the greatest bias when no education is used as the reference group. Future research should examine how variability in the sizes of those groups can impact the fitting of lasso and group lasso models

A second issue brought up by having small groups is the difficulty of splitting test and training data sets. This may become particularly problematic when there are many categorical variables that include many groups. Previous researchers have resolved to combine groups that are particularly small (e.g., racial/ethnic minorities; Webb et al., 2019). It is unclear how this practice impacts estimates for these groups, however, and in general combining groups is actively discouraged for other analytic methods (Tarantola & Dellaportas, 2005). Methods for splitting the data such as block randomization may provide more accurate predictions for small groups if the groups can be evenly split across the training and test sets.

4.3 Conclusion

Overall, our findings suggest that researchers should be cautious and purposeful about selecting their coding strategies when using lasso or group lasso. These choices will impact both variable selection and prediction accuracy when using lasso and prediction accuracy when using group lasso. However, just because variable selection is not impacted in group lasso does not mean this method should always be preferred. In a simulation study, we demonstrated cases where group lasso may have lower prediction accuracy than lasso, particularly when there is a dominant group (one group that differs from all other groups). The choices of which method to use (lasso or group lasso), what coding strategy to use, and which group order/reference category to use should depend on the researcher's priorities. How categorical variables are represented in lasso or group lasso models must be transparently reported to maximize reproducibility and replicability. Future research should explore specific practices in this area such as penalization of the intercept, the use of contrast coding, and how small groups should be accounted for to optimize prediction accuracy for these groups.

Behavioral scientists are quickly adopting useful tools developed in statistics and computer science which fit under the broad area of machine learning and artificial intelligence. The use of these tools will likely improve the ability of behavioral researchers to predict out-of-sample data, which may be particularly important in clinical settings and precision medicine. However, it is important to acknowledge that these new tools do not necessarily perform in the same ways that many researchers expect based on their training, which is primarily in linear regression and ANOVA frameworks (Aiken, West, & Millsap, 2008). Ensuring that the differences between these more traditional statistical frameworks and the newly developed machine learning frameworks are clearly defined will improve the implementation of these new methods throughout the field of behavioral science.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of aiken, west, sechrest, and reno's (1990) survey of phd programs in north america. *American Psychologist*, *63*(1), 32 – 50. doi: https://doi.org/10.1037/0003-066X.63.1.32

Choi, Y., Park, R., & Seo, M. (2012). *Lasso on categorical data.* Retrieved from `http://cs229.stanford.edu/proj2012/ChoiParkSeo-LassoInCategoricalData.pdf`

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/-correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum. doi: https://doi.org/10.4324/9781410606266

Darlington, R., & Hayes, A. (2016). *Regression and linear models: Concepts, applications, and implementation*. New York: Guilford Press.

Detmer, F. J., Cebral, J., & Slawski, M. (2020). A note on coding and standardization of categorical variables in (sparse) group lasso regression. *Journal of Statistical Planning and Inference*, *206*, 1–11. doi: https://doi.org/10.1016/j.jspi.2019.08.003

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees. *Methodology*, *3*(2), 47-57. doi: https://doi.org/10.1027/1614-2241.3.2.47

Hastie, T., Robert, T., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press. doi: https://doi.org/10.1201/b18401

Hastie, T., & Tibshirani, R. (2018). Best subset, forward stepwise, or lasso? analysis and recommendations based on extensive comparisons.. doi: https://doi.org/10.1214/19-sts733

Marquardt, D. W. (1980). Comment: You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, *75*(369), 87-91. doi: https://doi.org/10.1080/01621459.1980.10477430

McNeish, D. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471 – 484. doi: https://doi.org/10.1080/00273171.2015.1036965

StataCorp. (2019). *Stata statistical software: Release 16.* College Station, TX: StataCorp LLC.

Tarantola, C., & Dellaportas, P. (2005). Model determination for categorical data with factor level merging. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67(2)*, 269 - 283. doi: https://doi.org/10.1111/j.1467-9868.2005.00501.x

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267 – 288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., ... Pizzagalli, D. A. (2019). Personalized prediction of antidepressant v. placebo response: evidence from the embarc study. *Psychological Medicine*, *49*(07), 1118-1127. doi: https://doi.org/10.1017/s0033291718001708

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, *1*, 224-244. doi: https://doi.org/10.1214/07-aoas147

Yamada, Y., Ćepulić, D.-B., Coll-Martín, T., Debove, S., Gautreau, G., Han, H., ... Lieberoth, A. (2021, 1). Covidistress global survey dataset on psychological and behavioural consequences of the covid-19 outbreak. *Scientific Data*, *8*(3). doi: https://doi.org/10.1038/s41597-020-00784-9

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, *68*, 49-67. doi: https://doi.org/10.1111/j.1467-9868.2005.00532.x

Appendix A Different reference categories

In addition to the analyses presented in the primary manuscript, we also examined how variable selection and prediction accuracy in lasso and group lasso models differ across choices within a specific coding strategy. These choices include reference categories (dummy and contrast coding) and the order of categories (sequential and Helmert coding). We tested whether the category chosen as the reference category in the dummy coding strategy matters for variable selection and prediction accuracy. Consider, for example, the dominant group case where all groups have the same mean except one group. If that one group is selected as the reference category, then all $k-1$ predictors should be selected into the model, because all other groups are different from the reference. If any other group is selected as the reference group, then only 1 predictor should be selected into the model (the indicator for the difference between the one deviant group and the reference). While the pattern of means is not different, the reference group may have a large impact on the size of the coefficients and the number of non-zero coefficients.

We fit lasso and group lasso models with all six dummy-coded categorical variables and seven continuous variables using the COVID stress data. To explore how choices of reference categories affect estimated coefficients, we fit seven models for each regression method with differences only in their choices of reference categories in the variable *Education*. The reference categories were chosen and fixed for all other categorical variables. Therefore, the differences between these models can only be attributed to the different choices of the reference category of the variable *Education*. While this example uses dummy coding, we believe the results would generalize to other coding strategies (e.g., choice of the reference group for contrast coding, order of groups for Helmert and sequential coding).

Appendix A.1 Variable Selection

Table 11 shows the coefficients of indicators for *Education*. The size of the coefficients varies depending on which group is the reference, which could pose a problem for lasso regression because coefficients and the penalty parameter decide whether the variable will be selected into the model, according to Equation 5. Different coefficients are not necessarily a problem by themselves; however, these results demonstrate certain asymmetries that are concerning. When coefficients vary from model to model, the variable selection can differ. For example, when “none” was the reference category, the college category was not selected into the model (i.e., the none and college categories are assumed to be equal). However, when “college” was chosen as the reference category, the none category *was* selected into the model (i.e., the none and college categories are treated differently). This marks a particularly concerning lack of symmetry between these lasso models.

Table 11: Model Coefficients for Different Reference Categories by Lasso

	Reference Category						
Variables	*None*	*6 years*	*9 years*	*12 years*	*Some college*	*College*	*PhD/Doctorate*
Intercept	2.637	2.574	2.649	2.668	2.657	2.641	2.649
None	.	-0.013	-0.076	-0.092	-0.083	-0.068	-0.075
6 years	-0.068	.	-0.079	-0.095	-0.086	-0.071	-0.078
9 years	0.010	0.065	.	-0.006	0	0.009	0.002
12 years	0.033	0.084	0.024	.	0.017	0.032	0.024
Some college	0.017	0.067	0.008	-0.006	.	0.016	0.008
College	0	0.049	-0.008	-0.024	-0.015	.	-0.008
PhD/Doctorate	0.005	0.057	0	-0.015	-0.006	0.005	.

Note. Each column represents one model, and each row represents the coefficients for *Education* produced by each model.“.” is the reference category for the corresponding model, and 0 means that lasso does not select the corresponding predictor to be included in the model.

Group lasso models included all categories within the variable *Education* when different categories were chosen as the reference categories, meaning that all categories were treated as different in all group lasso models. Group lasso ensures stable performance of variable selection across reference categories.

We also explored the effect of different reference categories in education on other predictors and found that choosing different reference categories affects the coefficients and variable selection of other predictors (categorical and continuous) in lasso models. Group lasso models, on the other hand, still performed consistent variable selection for predictors that did not have their reference categories changed. In our case, group lasso models always included all categories within the other five categorical predictors and all seven continuous predictors.

Appendix A.2 Prediction Accuracy

We examined the prediction accuracy from two aspects: predicted category scores and model fit, varying the reference group used in dummy coding education.

Predicted Category Scores Predicted values for each category were different in both lasso and group lasso models from Tables 12 and 13. For the no education category, lasso models with different reference categories predicted different values, ranging from 2.982 to 2.915. Group lasso models also predicted different values for the no education category, ranging from 2.983 to 2.991. This indicates that with different choices of reference categories, predicted values vary from model to model for both lasso and group lasso.

Table 12: Predicted Category Means and Prediction Accuracy for Different Reference Categories by Lasso

	Reference Category						
Category	*None*	*6 years*	*9 years*	*12 years*	*Some college*	*College*	*PhD/Doctorate*
None	2.982	2.920	2.915	2.915	2.915	2.915	2.915
6 years	2.914	2.933	2.912	2.912	2.912	2.912	2.912
9 years	2.992	2.998	2.990	3.001	2.998	2.992	2.992
12 years	3.015	3.017	3.014	3.006	3.014	3.014	3.014
Some college	2.999	3.001	2.998	3.000	2.998	2.998	2.998
College	2.982	2.983	2.982	2.982	2.982	2.983	2.982
PhD/Doctorate	2.988	2.990	2.990	2.991	2.991	2.987	2.990
MSE	0.13669	0.13678	0.13674	0.13684	0.13675	0.13674	0.13674

Note. Each column represents one model, and each row (besides the last) represents the predicted category means for *Education* produced by each model (with all continuous predictors set to their means and all other categorical variables set to their modes). The last row contains the MSE of the corresponding model.

Table 13: Predicted Category Means and Prediction Accuracy for Different Reference Categories by Group Lasso

	Reference Category						
Category	*None*	*6 years*	*9 years*	*12 years*	*Some college*	*College*	*PhD/Doctorate*
None	2.986	2.986	2.986	2.990	2.991	2.983	2.987
6 years	2.971	2.978	2.974	2.983	2.981	2.971	2.975
9 years	2.988	2.987	2.968	2.979	2.975	2.965	2.969
12 years	3.002	3.001	3.004	2.991	2.992	2.985	2.988
Some college	2.996	2.996	2.997	2.996	3.004	3.003	3.004
College	2.983	2.983	2.983	2.983	2.983	2.996	2.997
PhD/Doctorate	2.988	2.988	2.988	2.990	2.990	2.987	2.983
MSE	0.13711	0.13719	0.13709	0.13727	0.13709	0.13708	0.13710

Note. Same as Table 12.

Figure 4 visualizes the shrinkage effect when different reference categories were chosen in lasso models using *Education* to predict *Stress*. In this case, the intercept is the predicted category mean of each model's reference category because models are coded by dummy coding strategies. Similar to Figure 1, we can conclude that recreated category scores shrink towards the reference value for dummy coding.

Model Fit Model fit, measured by MSE, for both lasso and group lasso models are shown in Table 12 and 13. MSEs were generally different across reference categories. Note that MSEs in Table 12 and 13 were rounded. Although some MSEs were very close to each other and were rounded to the same value, they were not exactly the same, which would be the case if linear regression was used.

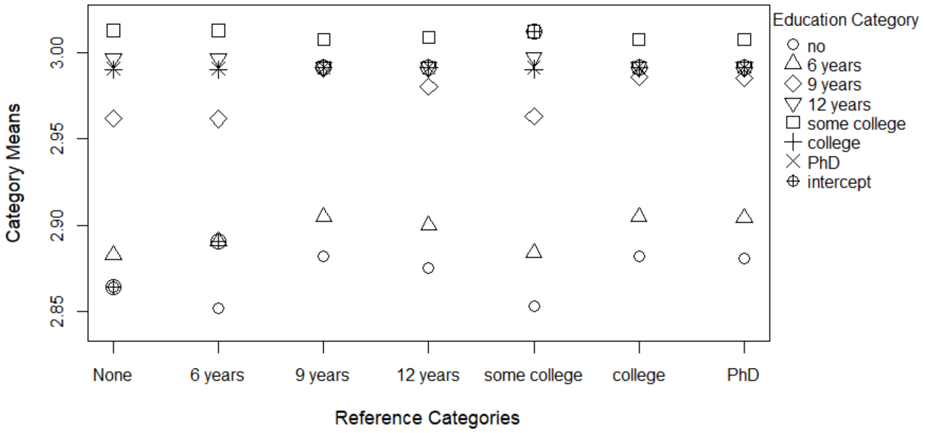


Figure 4: Graphical Presentation of Category Means for *Education* Recreated by Lasso Models with Different Reference Categories. Intercept values are different across reference categories. In dummy coding, the intercept value is the estimated category mean of the corresponding reference category.

Appendix B Singular Design Matrices

STATA is a commonly used statistical software that can implement lasso regression, and in STATA categorical predictors are handled by including a singular design matrix StataCorp (2019). In this section, we examine this alternative method for creating the design matrices for categorical variables. When we introduced categorical variables, we noted that for a variable with k categories, $k-1$ indicators are created for this variable. Different coding strategies use different matrices to represent the $k-1$ indicators and model coefficients represent differences between categories and the reference value, as this is common practice for linear regression. The researcher must then choose the reference category for analysis. However, there is another way to create the design matrix for categorical predictors where the researcher does not need to explicitly choose the reference category. Instead of using $k-1$ indicators for a categorical variable with k categories, we use k indicators. This design matrix allows lasso or group lasso to essentially select the reference values. Mathematically, this type of design matrix is defined as singular, because the matrix is not invertible. Singular design matrices cannot be used for linear regression, but lasso and group lasso regression can accommodate singular design matrices, making this a unique potential solution to the variable selection and prediction accuracy issue related to categorical variables in lasso and group lasso.

Can singular design matrices solve the inconsistency in lasso's variable selection and prediction accuracy or group lasso's prediction accuracy across coding strategies? To create singular design matrices, we appended a linearly independent column with only 1 in the first row to the matrices in Table 1 and 3, and a linearly independent column with only 1 in the last row to matrices in Table 2 and 4. If using singular design matrices solves the issues of variable selection and prediction accuracy, these two properties should be equivalent across these four design matrices. To test this, we used the same data set and applied the same process as before to fit lasso and group lasso models with *Education* serving as the only predictor variable. Table 14 shows the coefficients of the categorical variable *Education* in lasso models as an example of lasso's variable selection. Using a singular design matrix for categorical variables, different coding strategies still lead to different lasso model's variable selection. Contrastingly, group lasso selected all categories and performed the same variable selection. For example, the contrast-coded lasso model treated the 9 years of education and PhD categories as the same, while these two categories were always treated as different in the other three lasso models and the four group lasso models. In addition, lasso and group lasso models using different coding strategies led to different prediction accuracies, shown in Table 15 and 16. This means that using singular design matrices does not solve the inconsistent variable selection or prediction accuracy for lasso, nor does it solve the inconsistency in prediction accuracy for group lasso. There are infinitely many singular design matrices that could be used, and if they all result in different solutions, this does not provide strong evidence that the identity matrix system used by StataCorp (2019) would perform optimally.

Table 14: Model Coefficients Using Singular Design Matrix with Lasso

Coding strategies	*Dummy*	*Contrast*	*Sequential*	*Helmert*
Intercept	2.991	2.961	2.873	2.961
1. no	-0.109	-0.081	0.015	0.103
2. 6 years	-0.086	-0.063	0.076	0.095
3. 9 years	-0.006	0	0.034	0.023
4. 12 years	0	0.034	0.010	0
5. some college	0.016	0.051	-0.017	-0.017
6. college degree	0	0.028	0	0
7. PhD	0	0	-0.009	0

Note. Each column represents one model, and each row represents the coefficient for an indicator of *Education* produced by the corresponding model. A 0 means that lasso does not select the corresponding category into the model.

Table 15: Predicted Category Means and Prediction Accuracy for Different Coding Strategies using Singular Design Matrices with Lasso

	Coding Strategy				
Category	*Dummy*	*Contrast*	*Sequential*	*Helmert*	Observed Mean
None	2.882	2.881	2.864	2.873	2.852
6 years	2.905	2.898	2.888	2.897	2.883
9 years	2.986	2.961	2.964	2.973	2.962
12 years	2.991	2.995	2.999	2.997	2.997
Some college	3.007	3.012	3.008	3.008	3.013
College	2.991	2.990	2.991	2.991	2.990
PhD/Doctorate	2.991	2.992	2.991	2.991	2.991
MSE	0.15630	0.15620	0.15616	0.15621	/

Note. Each column (besides the last) represents one model, and each row (besides the last) represents the predicted category means for *Education* produced by each model. The last column contains the category means observed in the training data set. The last row contains the MSE of the corresponding model.

Table 16: Predicted Category Means and Prediction Accuracy for Different Coding Strategies Using Singular Design Matrices with Group Lasso

	Coding Strategy				
Category	*Dummy*	*Contrast*	*Sequential*	*Helmert*	Observed Mean
None	2.873	2.869	2.878	2.871	2.852
6 years	2.892	2.890	2.909	2.891	2.883
9 years	2.962	2.961	2.955	2.962	2.962
12 years	2.996	2.996	2.994	2.996	2.997
Some college	3.012	3.012	3.010	3.012	3.013
College	2.990	2.990	2.991	2.990	2.990
PhD/Doctorate	2.990	2.991	2.991	2.990	2.991
MSE	0.15619	0.15619	0.15622	0.15619	/

Note. Same as Table 15.