UCLA
Library Prize for Undergraduate Research

## Title

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

## Permalink

https://escholarship.org/uc/item/40b200z6

## Author

Huang, Yihuan

## Publication Date

2020

## Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, availalbe at https://creativecommons.org/licenses/by-nc-nd/4.0/

Undergraduate

eScholarship.org Powered by the California Digital Library
University of California

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

Yihuan Huang and Amanda Montoya

University of California, Los Angeles

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

## Introduction

Machine learning is becoming increasingly popular in social and behavioral sciences, and is frequently used by researchers in different scientific fields to solve practical problems with complex data (e.g., Leach, O'Connor, Simpson, Rifai, & Mama, 2016; Ma, Chang, & Cui, 2012; Steele, Denaxas, Shah, Hemingway, & Luscombe, 2018a). Specifically, psychologists have begun to utilize these algorithms to analyze underlying factors in psychological phenomenon (e.g., Sauer et al., 2018), guide improvements of current treatments, and use previous patients records to make data-driven decisions for incoming patients (e.g., Zilcha-Mano, Errázuriz, Yaffe-Herbst, German, & DeRubeis, 2019). For example, Leach et al. (2016) used a decision tree to determine environment characteristics contributing to the classification of African American women as at risk for cardiovascular disease and to predict future cardiovascular disease risk in African American women based on age. Bainter, McCauley, Wager, and Losin (2019) utilized a stochastic search variable selection to characterize the contributions of different psychological, sociocultural, and neurobiological factors of pain experiences, with which can then be used to predict pain.

The least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), a very popular machine learning algorithm, is useful when the data set involves many predictors and the outcome variable is continuous. Lasso is gaining popularity in psychology and one of the reasons is that it has many shared properties with linear regression, an already common statistical approach in the field. Models built by both linear and lasso regression can be expressed as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{N} \beta_j X_{ij} + \epsilon_i. \tag{1}$$

The above equation calculates the $i^{th}$ entry of the outcome vector $Y$, where $\beta_0$ is the intercept term; $\beta_j$ is the $j^{th}$ entry of the coefficient vector $\beta$; $X_{ij}$ is the entry in the $j^{th}$ column and $i^{th}$ row of the design matrix X; $\epsilon_i$ is the $i^{th}$ entry of the error vector $\epsilon$.

Linear and lasso regression differ in the way they estimate $\beta$. Linear regression aims to minimize the sum of squared of errors generated between predicted and observed values. The

coefficient vector is calculated as follows,

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

where $|\cdot|_2$ is the notation for the L2 norm, which is also known as the Euclidean norm. Lasso adds a penalty term, a new parameter $\lambda$, to regulate the size of the coefficients which can affect the number of predictors included in the model. The coefficient vector is calculated as follows,

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}}(|Y - X\beta|_2^2 + \lambda \sum_j |\beta_j|) \tag{3}$$

Linear regression is a special case of lasso regression. Equation 2 is equal to Equation 3 where $\lambda$ is set to zero. This means that linear regression models are built to maximally predict each outcome variable without taking into account the size of the coefficients. While, in lasso regression with the added penalty parameter, non-zero values of $\beta$ result in increases in $\lambda \sum_j |\beta_j|$ that need to be minimized simultaneously with the sum of squared errors in Equation 3. Therefore, $|Y - X\beta|_2^2 + \lambda \sum_j |\beta_j|$ reaches its minimum when both the prediction and the size of $\beta$ are taken into consideration. The magnitude of $\lambda$ determines the shrinkage of the elements of $\beta$. When the penalty parameter is large, the coefficients are shrunk toward zero and fewer predictors are selected in the model; while when the penalty parameter is small, the shrinkage is less extreme so more predictors can be selected in the model. Another alternative to lasso is ridge regression which is expressed by Equation 3 except with an L2 norm instead of an L1 norm for the regularization term. In Equation 3, L1 norm $\lambda \sum_j |\beta_j|$ penalizes the absolute value of the coefficients, used by lasso; while ridge regression uses L2 norm $\lambda \sum_j |\beta_j|^2$ in which the regularization term is the sum of squares of all coefficients. Therefore, ridge regression is not as good at penalizing parameters to zero as lasso regression.

Lasso can be used much more effectively than linear regression for the process of variable selection (Tibshirani, 1996). In linear regression, a model is built with all variables and statistical inference is typically used to determine which variable contributes significantly to the model. While in lasso, only predictors that make big enough contributions to explaining the outcome variable are selected in the model. Lasso's variable selection results in two particular advantages: reducing dimension of the design matrix and improving prediction accuracy (Tibshirani, 1996).

Lasso can be used as a dimension reduction method to select strong variables, which is particularly useful for data of high dimension because models with too many variables can be hard to interpret. After performing variable selection, lasso can help researchers make clearer interpretations of the results (Hastie, Robert, & Wainwright, 2015). In psychology, lasso is often used to select important features that can explain one specific behavior. For example, Ammerman, Jacobucci, Kleiman, and McCloskey (2018) used lasso to identify that number of non-suicidal self-injury (NSSI) methods was the most important correlate of NSSI frequency. After removing this variable, Ammerman et al. (2018) reran lasso regression and further determined that suicide plan and depressive symptomology were also strong correlates across methods. Therefore, the study not only confirms the relationship between NSSI frequency and NSSI methods but also identified the importance of suicide plans, an often-overlooked factor, and depression in NSSI severity.

Besides increasing the interpretability of models, lasso's dimension reduction can be used as an initial data preprocessing step. Most statistical models can not be applied to data of high dimension, especially if the number of variables exceed the number of total observations. The reduced dataset processed by lasso allows the application of many different statistical models. Burningham, Leng, Peters, and Huynh (2018) provided a good illustration of this method in psychology, where his primary goal was to identify aging Veterans with psychiatric disease in attempt to prevent psychiatric crises. Prior to logistic regression, Burningham et al. (2018) used lasso to filter out variables that are not closely related to geriatric psychiatric hospitalization. Then individual predicted probabilities were estimated using logistic regression.

Lasso regression can also be used to gain better prediction accuracy because the penalization term decreases the model's over-fitting (McNeish, 2015). Linear regression may fit a model which is better able to predict the sample data by including all variables. However, if the model fit by linear regression is used to predict out-of-sample observations, the prediction accuracy tends to be low because of over-fitting (large variance and unbiased estimates). This issue can be solved by lasso regression because only strong predictors will be selected into models, and the model will not be heavily influenced by some extreme data points (Steele, Denaxas, Shah, Hemingway, & Luscombe, 2018b). In other words, a small additional bias in the estimates is

introduced which decreases the variance of the predictions, and the prediction accuracy increases.

Lasso regression was developed with two advantages over the linear regression: clear variable selection and better prediction accuracy. These advantages have made lasso an attractive alternative to linear regression, particularly when fitting models with many variables. This method has been attractive to psychology researchers, because of the similarity between lasso and linear regression, allowing them to easily generalize their previous knowledge to a novel method. Researchers in psychology are now able to use lasso regression for variable selection in exploratory research and to create models with improved predictive power.

**Categorical Predictors in Linear Regression**

Given the advantages of lasso over linear regression, it is important to explore how lasso should be applied in common cases within psychological data analysis. Categorical variables are frequently used in psychological models, including variables like ethnicity, gender, experimental conditions, or religion. Unlike numerical predictors which typically have a natural scale, categorical variables require researchers to select a method for coding the variables (i.e., representing the categories using a numeric system). Categorical variables with more than two categories need to be encoded into a set of indicators in order to be considered in regression models. Different coding strategies can be chosen, such as dummy coding, contrast coding, or Helmert coding. Dummy coding uses only 0's and 1's to indicate the category membership. One category is selected as the *reference category* and is assigned a score of 0 on all indicators. For all other categories in dummy coding, only the corresponding indicator is coded as 1 and 0 for the rest of indicators (e.g., Table 1). Contrast coding is similar to Dummy coding, but the category which is coded as all 0 is now coded with all -1 instead, changing the interpretation of the intercept (e.g., Table 2). Helmert coding examines more complex comparisons where each category is compared to the average of all subsequent categories (e.g., Table 3).

While each coding scheme represents the categories using a different numerical system, ultimately they only differ in the interpretation of their coefficients. Each coding scheme always recreates the category mean for each category. In linear regression, coding strategies only vary in the way they convey the categorical data and they always generate the same predicted scores for individual cases. Therefore, researchers can choose coding strategies among all these options

according to their needs. Dummy coding or contrast strategies can be used for nominal categorical variables, while Helmert coding strategy is particularly helpful when groups within the categorical variables can be ordered relative to each other. For example, when a study has multiple experimental conditions and a single control condition, dummy coding can be used so that each regression coefficient provides an estimate of the difference between one experimental condition and the control. Alternatively, in cases when categories are ordered, for example level of education, a researcher may want to use Helmert coding. When Helmert coding is used, the researcher can learn about the difference between individuals with some high school education and no high school education. Then individuals who completed high school could be compared to the average of some and no high school. Individuals with some college experience could be compared to the average of those who completed and did not complete high school, and so on. Tables 1 - 3 show different ways to encode a categorical predictor, *Marital Status*, which includes 5 categories (single, married, widowed, divorced, and separated).

Regression coefficients produced by these coding strategies have different meanings. To explore the relationship between the the categorical variable *marital status* and the outcome variable *wage* (in thousands of dollars), the five categories within the variable *Marital Status* are encoded by 4 indicators. Linear regression fits the following model:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i4} + \epsilon_i, \tag{4}$$

where $X_{ij}$ is the $j^{th}$ indicator to convey category membership information from the category predictor for the $i^{th}$ person, and $Y_i$ is the outcome value for the $i^{th}$ person. The intercept $b_0$ and coefficients for different indicators, $b_1$ – $b_4$, have different meanings if different coding strategies are used. For example, suppose our linear regression model is

$$Y_i = 12 + 3X_{i1} + 4X_{i2} + 2X_{i3} + 1X_{i4} + \epsilon_i. \tag{5}$$

If dummy coding was used with single as the reference group (as in Table 1), we would interpret the coefficient for $X_2$, 4 means that the difference between the salary of single and married people is \$4000. However, if contrast coding was used (as in Table 2), 4 would indicate the difference between the salary of the married people and the average salary of all people is \$4000. If Helmert coding is used (as in Table 3), 4 means that the married people earn \$4000

more than the average salary of the widowed, divorced, and separated people. In other words, coefficients of dummy coding represent the differences between each category and the reference category; those of contrast coding quantify the differences between one category and the average of all categories; while those of Helmert coding quantify the differences between one category and the average of all subsequent categories. Apart from the differences in coefficients caused by choices of coding strategies, choices of reference categories in dummy coding can also produce different coefficients. For example, if single is the reference category, $b_0$ represents the average wage for the single people and $b_1$ through $b_4$ will represent the difference between single and the coded category. While if married is the reference category, $b_0$ represents the average wage for the married people and $b_1$ through $b_4$ will represent the difference between married and the coded category.

Different ways to code categorical variables do not affect the prediction accuracy of models fit by linear regression. Regardless of which coding strategy the model uses, linear regression always recreates category means from the data. We show this consistency using a data set with wages in 3,000 US MidAtlantic Men (James, Witten, Hastie, & Tibshirani, 2014). This wage data includes six categorical variables (*Race, Job Class, Health, Health Insurance, Marital Status*, and *Education*), and three continuous variables (*Year, Age,* and *Wage*). The overall goal is to predict *Wage* using the available predictors. To show the exact recreation of category means in linear regression, we used only one categorical predictor *Marital Status* to predict *Wage*. We used linear regression to predict wage by coding the variable *Marital Status* with three coding strategies from Tables 1 - 3 and three linear regression models were fit. Table 4 contains the coefficients of the three models.

For the dummy coding, using the values of $X_1$ – $X_4$ from Table 1 and the coefficient estimates from Table 4, the predicted mean for the Single category is

$$92.735 + 26.126 \times 0 + 6.804 \times 0 + 10.435 \times 0 + 8.481 \times 0 = 92.735.$$

For the contrast coding, using the values of $X_1$ – $X_4$ from Table 2 and the coefficient estimates from Table 4, the predicted mean for the Single category is

$$103.102 - 10.367 \times 1 + 15.759 \times 0 + -3.563 \times 0 + 0.058 \times 0 = 92.735.$$

Using the values of $X_1$ – $X_4$ from Table 3 and the coefficient estimates from Table 4, the predicted value is

$$103.102 + 12.959 \times (-\frac{4}{5}) + (-17.556) \times 0 + 2.649 \times 0 + (-\frac{4}{5}) \times 0 = 92.735.$$

The mean for the Single category recreated by the dummy coding is the same as that by the contrast and Helmert coding. It can be shown that all category means for each coding strategy and show that category means are the same throughout these three coding strategies. From this example, we can conclude that linear regression models with different coding strategies recreate same category means, though they produce different coefficients.

**Motivation**

With increasing use of lasso techniques across scientific fields, many researchers rely on the similarities between lasso and linear regression in order to understand, use, and interpret the results of lasso analysis. Researchers often use lasso in the same way as linear regression, including models with categorical variables. Heckman, Handorf, Darlow, Ritterband, and Manne (2017) used lasso to investigate intervention effects of UV4.me, an internet intervention that decreased ultraviolet radiation exposure and increased skin protection behaviors among young adults. The study used Helmert coding for the two categories of treatment (control and experiment). Heckman et al. (2017) found two specific modules that were most strongly associated with behavioral improvements were for UV exposure and four modules which best predicted improvements in skin protection. Though the researchers used Helmert coding, it is unclear if the findings would be the same if a different coding strategy had been used instead. Would the same predictors be identified as most associated with intervention effects? Ultimately, if coding strategy impacts the models fit using lasso regression, then two questions arise: First, is there a method for fitting lasso regressions which is not impacted by coding strategy choice, and second which coding strategy would allow the researchers to most accurately predict their outcome?

No research has yet explored the interplay between the way that categorical variables are typically used in linear models and how this practice impacts the results of lasso regression. Lasso regression models are frequently used for variable selection. The model selects variables based on the penalty parameter and the size of coefficient vector $\beta$. However, using different coding

strategies fits models with different coefficient vectors. Therefore, it is reasonable to expect that choice of coding strategies may result in a different selection of variables in lasso regression models. In other words, because of the impact that coding strategies can have, it is unclear if the same conclusions would be made based on models with the same variables which are coded in different ways. If indeed coding strategy does impact the results of a model (e.g., variable selection and/or prediction accuracy), the question remains: which coding strategy should researchers use when building models involving categorical predictors in their studies?

Ultimately, the issue of coding strategies is related to the issue of variable scaling with continuous predictors. The scaling of continuous predictors also influences variable selection and prediction accuracy in lasso regression models. For example, changing a variable from height in feet to height in inches would impact the coefficient for height and thus impact the variable selection approach. By changing the interpretation of a one unit change in the variable, researchers could change how large the impact of the variable will seem to be. Inconsistency in scaling practices can result in a lack of replicability of lasso models and potential misrepresentation of the relative contributions of the predictors in the model. One common solution is to standardize the values of all predictors before applying lasso regression (Marquardt, 1980). In this way, the effect of scaling is excluded from the variable selection of lasso regression. Dichotomous variables can always be standardized such that any other scaling would result in the same standardized variables. However this is not so with categorical variables with more than 2 categories: standardizing a dummy coded set of variables would still result in a different set of variables from standardizing a Helmert coded set of the same variables.

In order to explore the potential impacts of coding strategy on important characteristics of lasso regression we undertake a variety of steps using both real data analysis and simulation. First, using the wage dataset described above, we explore the use of lasso regression with categorical variables, where different coding strategies of categorical variables impact two aspects of lasso models: the variable selection and prediction accuracy. An alternative method of lasso, group lasso, is introduced in the next section. Group lasso is also applied to the wage dataset and both the variable selection and prediction accuracy of group lasso models are examined. We describe an over-fitting issue of group lasso using Monte Carlo Simulation in the next section. In

the last section, potential solutions, important future directions, and a summary are provided.

## Lasso with Categorical variables

We used the wage data to explore how coding strategies affect the models estimated by lasso. We used six categorical variables (*Race, Job Class, Health, Health Insurance, Marital Status, and Education*) and one continuous variable (*age*) to predict the outcome variable wage. Different from the continuous variable, which can be represented by one variable, each categorical variable is represented by $k-1$ indicators where $k$ is the number of categories. Different coding strategies represent the categorical variables in different ways. In the wage data, the variable *Marital Status* includes 5 categories (single, married, widowed, divorced, separated); *Education* includes 5 categories (less than high school education, high school education, some college, college education, advanced degree); *Race* includes 4 categories (White, Black, Asian, other); *Job Class* includes 2 categories (industrial and information); *Health* includes 2 categories (good or lower and very good or higher); and *Health Insurance* includes 2 categories (yes and no). Therefore, after coding all categorical variables and including *Age*, we estimated the wage with $4+4+3+1+1+1+1=15$ predictors. After data preprocessing, we examined the impact of coding strategy on the two primary purposes of lasso: variable selection and prediction accuracy.

### Different types of coding strategies

We explored how variable selection and prediction accuracy are affected by different types of coding strategies used to estimate lasso regression models. In order to measure the prediction accuracy, we randomly split data into training and testing parts with ratio 6:4. The training data includes information from 1800 males, while the testing data includes information from 1200 males. We trained three different lasso models using three coding strategies (dummy, contrast, and Helmert) on the same training data. We used cross validation on the training data set to select the penalty parameter from the model with the best prediction accuracy. It is worth to mention that the penalty parameter is different for models with different coding strategies, which means that each model is penalized differently. By examining the performance of these three lasso models we examined if variable selection and prediction accuracy of lasso models are affected by the choices of coding strategies. Note that because this is based on a single dataset, it is not valid to compare the prediction accuracy between models in order to determine which coding strategy

is "best". We leave this issue for future simulation research.

**Variable Selection.** We examined differences in the variable selection between three models. Results are show in Table 5. Take the variable *Marital Status* for example. The dummy-coded model includes all variables except the one representing the difference between Single and Widowed. It means that the dummy-coded model treats the mean of the Single category as the same as that of the Widowed category. The contrast-coded model include all variables. The Helmert-coded model excludes the variable representing the difference between the Widowed and the average of Divorced and Separated categories. It also excludes the variable representing the difference between the Divorced and Separated. Therefore, the Helmert coding model treats the Divorced the same as the Separated category, and the Widowed the same as the average of Divorced and Separated. In other words, the Widowed, Divorced, and Separated are treated equally in the Helmert-coded model. If the results of the dummy-coded model were to align with those of Helmert, the difference between Divorced and Single, Separated and Single, and Widowed and Single should all be the same. However, in dummy-coded model, the Widowed is the only group treated the same as the Single. After carefully examining the models shown in Figure 5, we can conclude that different coding strategies select different variables in the model. The result is problematic because models with different variable selection can produce different interpretations of the models. Which category within the variable marital status will be selected into the model depends on the chosen coding strategy. Researchers who use dummy coding will probably conclude that the Widowed people on average have the same wage as the Single people, while those using Helmert coding will probably interpret that the Widowed have the same wage as the Divorced and Separated. Recall that in the linear regression, all variables are selected into the model, but this example demonstrates that lasso models with different coding strategies select different variables in the model.

**Prediction Accuracy.** We calculated the predicted wage of each category within the variable *Marital Status* for each lasso model. In linear regression, the predicted value for each category is unaffected by coding strategy. Here we examined whether every category has the same value of predicted wage across models with different coding strategies using lasso. The three lasso regression models were further used to predict the wages of people in the testing data. Prediction

accuracy was calculated to determine the differences between the predicted wage and the actual wage for people in the testing data. We used Mean Squared Errors (MSE) as a measurement of the prediction accuracy. Mathematically, MSE is calculated as following:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{6}$$

The results of category means and MSEs for three models are shown in Table 6. Using the same method for recreating category means for the linear regression model, we recreated the category means for the variable *Marital Status*. The Helmert-coded model recreated same category mean (88.227) for the Widowed, Divorced, and Separated categories. While in other models, these three category means are different. Dummy-coded model recreated 69.988 for Widowed, 73.665 for Divorced, and 76.195 for Separated; and Contrast-coded model recreated 83.773 for Widowed, 85.772 for Divorced, and 89.907 for Separated. This validates our interpretation in previous sections. Besides differences in category means, MSEs are also different. It means that models with different coding strategies have different prediction accuracy. Recall that linear regression always recreates the same category mean regardless of the choices of coding strategies and the prediction accuracy stays the same. Therefore, researchers can choose the coding strategy only according to its interpretation. Nevertheless, from Table 6, we can see that different coding strategies estimate category means differently and result in different MSEs when lasso regression is applied. This exposes uncertainty regarding which coding strategy should be used when lasso regression is applied.

Next, we explored one reason why different coding strategies recreate category means differently and result in different prediction accuracy. As mentioned in the previous section where we first introduced three coding strategies, coefficients in models built with different coding strategies have different meanings. In the dummy coding strategy, the coefficient represents the difference between a category mean with the reference category mean (typically the first category, the Single in our case). For the contrast coding, the coefficient represents the difference between a category mean with the average of all category means, which is the same for Helmert coding. Shrinking coefficients to zero means that models aim to shrink category means to the corresponding model intercepts. Each coding strategy has different model intercept. Therefore,

the shrinkage effect is different across coding strategies. To visualize the shrinkage effect of each coding strategy, we plotted the category means recreated by lasso models with different coding strategies and the intercept for each coding strategy in Figure 1 (left column of each coding strategy). We can see that for each coding strategy, category means are shrunk towards the intercept. Take the Widowed category for instance. Three models recreated different category means for the Widowed people. The Widowed category mean in dummy-coded model is closer to the dummy-coded model's intercept than that in contrast and Helmert models. Models built with different coding strategies shrink category means toward their respective model intercepts, and because different coding strategies have different reference values, category means will be recreated differently and lead to different prediction accuracy. It is worth to note that the reference for contrast and Helmert coding strategies are the same (the average of all category means). However, category means are still created differently for these models, indicating that the choice of intercept is only one of the reasons leading to the difference in recreation of category means.

**Different reference categories**

It is clear from the previous results that coding strategies affect the variable selection and prediction accuracy of lasso regression models. Next, we examined how coefficients in lasso models change when different choices are made within a specific coding strategy. These include choices of reference categories (e.g., dummy and contrast coding) and the order of comparisons (e.g., Helmert coding). Specifically, we examined the impact of which categories within a categorical variable is chosen as the reference category in dummy coding strategy. We used dummy coding for the categorical variables *(Race, Job Class, Health, Health Insurance, Marital Status*, and *Education*) in wage data. We built lasso models with all these categorical variables and one continuous variable (*Age*). In order to explore how choices of reference categories affect model's coefficients, we built five models with differences only in their choices of reference categories in the variable *Marital Status*. The reference categories were chosen for all other categorical variables except the variable *Marital Status*. Therefore, the differences between these models should be caused by the different choices of the reference category of the variable *Marital Status*. Table 7 shows the coefficients of categories within the categorical variable *Marital Status*, which all differ due to choices of reference categories in our models. This is a problem for lasso

regression because coefficients and the penalty parameter decide whether the variable will be selected into the model, according to Equation 3. When coefficients vary from model to model, the variable selection also varies. When the Widowed is the reference category, the Divorced category is not selected into the model (i.e., the Widowed and Divorced are assumed to be equal). While, when the Divorced category is chosen as the reference category, all categories are selected into the model (i.e., no groups are assumed to be equal). Similarly, when the Single is the reference category, the Widowed category is excluded from the model, but when the Widowed is the reference category, the Single is still included in the model. There exists an inconsistency of variable selection from model to model. Moreover, after calculation we can see that group means are also recreated differently from Table 8. We take the widowed people for example. If single category is chosen as the reference one, widowed is treated the same as the single and the estimated wage is 69.988(k). If married category is chosen as the reference one, the widowed estimated wage is $87.344 - 15.133 = 72.211(k)$. Similarly, the estimated wage for the widowed from the other three models are 69.988, 72.227, and 71.985. Besides, from Table 7 and 8, we can conclude that with different choices of reference categories, coefficients and category means vary from model to model. Therefore, different reference categories within a categorical variable also influence the variable selection and prediction accuracy of lasso models.

**Singular Design Matrices**

In this section, we are going to explore an alternative method for creating the design matrices for categorical variables. When we introduced lasso with categorical variables, we noted that for a categorical 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 comparisons between categories and the reference value, as this is common practice for linear regression. In this case, the researcher must choose the reference group for the analysis. For example, we choose the category mean for the Single as the reference in the dummy coding strategy. However, there is another way to create the design matrix for categorical predictors where the researcher does not need to explicitly choose the reference value. Instead of using only the $k - 1$ indicators for a categorical variable with $k$ categories, we use $k$ indicators. This design matrix allows lasso to essentially select the reference values. Mathematically, this

type of design matrix is defined as singular, because it is not invertible. Singular design matrices cannot be used for linear regression, because solving the ordinary least squares solution requires taking the inverse of the design matrix. However, 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. In fact, using a singular design matrix with dummy variables, is the default method for fitting factors in lasso regression in the popular statistical packages STATA (StataCorp, 2019).

We explore whether singular design matrices solve the inconsistency in the variable selection and prediction accuracy across coding strategies. We appended a linearly independent column with only 1 in the last row to the matrices in Table 1 - 3 to create the singular design matrices. If using singular design matrices solves the issues of variable selection and prediction accuracy, then these two properties should be equivalent across these three design matrices. To test this, we used the same data set and applied the same process as before to build lasso models. We recreated the category means for the lasso models in Table 9. For lasso, different coding strategies have different variable selection, which can be seen from the Table 9. The helmert-coded model treats the Widowed and Divorced categories as the same, while the dummy-coded model treats the Widowed, Divorced, and Separated categories as the same. In addition, lasso models using different coding strategies lead to different prediction accuracy. This means that using singular design matrices does not solve the inconsistency for lasso in variable selection or prediction accuracy.

**Lasso Summary**

When conducting a lasso regression with categorical predictors, the analyst must choose two important characteristics for each categorical variable: the coding strategy (e.g., dummy vs. contrast vs. Helmert) and also the ordering of the categories, which involves which group is the reference (dummy and contrast) and the order of comparisons (Helmert). In the above sections, we examined how these choices affect variable selection and prediction accuracy in lasso regression models. For each choice of coding strategy, we obtained the model's variable selection, category means for the variable *Marital Status*, and the prediction accuracy. It is clear to see that lasso regression models select different sets of variables when different choices of coding strategies are

made. From the singular matrix design, we can also conclude that without choices of the ordering of the categories or the reference categories, lasso models still perform different variable selection and model fits.

Different variable selection can not only cause completely different interpretations, but also result in estimating different category means. Suppose when a coefficient is shrunk to zero by lasso regression, the corresponding variable is not selected in the model. For example, with dummy coding, the model would regard the category coded by that variable as no different from the reference category within same categorical variable on the outcome variable. Therefore, models will generate the same predicted outcome values for cases in the category coded by the excluded variable and the reference category. In this way, different variable selection can result in different category means. Additionally, even when the same variables are selected into the model, the degree of shrinkage will depend on the initial coefficient size, meaning that even when all variables are included in the model, the predicted values for specific categories may depend on which coding strategy gets used.

Consequently, differences in category means lead to different prediction accuracy. The mean for each category is estimated differently in each coding strategy. Therefore, different models will arrive at different estimations of the outcome variable with same variables' values. Ideally, there would be a method which would provide the same predicted scores regardless of coding strategy; however, if the method is not possible prediction accuracy could be used by researchers as a factor to determine which coding strategy to use to build the lasso regression model.

We can conclude that lasso models heavily depend on the choices of coding strategies (types of coding strategies and choices of reference categories) for categorical variables. With different coding strategies, lasso performs different variable selection and model fit. This raises a problem when lasso regression is applied to a real-world data set, like in psychology. Lasso is frequently used to explore the relationships among psychological phenomenon. As different coding strategies build lasso models with different variable selection, inferences may differ if coding strategies change. Based on this, lasso is clearly inappropriate to use when categorical variables are present in the model. This leads to the questions: Is there an adjustment to the lasso method that would always performs the same variable selection and prediction accuracy regardless of coding

strategy? In the next section, we will introduce a lesser known variant of lasso regression which may solve some of the issues with coding strategy.

## Group Lasso

Group lasso is a generalization of lasso for doing group-wise variable selection (Yuan & Lin, 2006). The group lasso algorithm was first introduced to allow predefined groups of predictors selected in or out of a model together. Similar to linear and lasso regression, models estimated by group lasso can be expressed using Equation 1. The mathematical formulae for calculating coefficients $\beta$ is the following:

$$\hat{\beta}_{group} = \underset{\beta}{\operatorname{argmin}}(|Y - X\beta|_2^2 + \lambda \sum_{g=1}^{G} |\beta_{I_g}|_2) \tag{7}$$

where $G$ represents the number of groups within the dataset, and $\beta_{I_g}$ represents the coefficient vector of that corresponding group. Other notations are the same as previous equations. From the Formulae 7 we can see that lasso and group lasso differ in what type of norm is used for the sum and how the penalty parameter is weighted. Lasso regression uses L1 norm to sum all coefficients before multiplying the penalty parameter. Instead, group lasso first uses L2 norm to sum the coefficients within each group and then sums across the groups, which is equivalent to taking the L1 norm of the L2 norms of the groups. Using the L2 norm within a group makes it more likely to either select all variables within the group or not. Moreover, multiplying the penalty parameter after summing the coefficients within groups penalizes each group instead of each variable. Therefore, the number of variables within a group can affect the evaluation of coefficients. These differences with regard to the regularization term (second term in equation 7) provide group lasso with distinct properties.

The group lasso has special properties with respect to variable selection. Within a group, group lasso typically either includes all or excludes all variables. Given this unique process of variable selection, we propose that group lasso may be useful as an alternative to lasso regression when dealing with models with categorical variables. One previous study has recommended the use of group lasso for accounting for categorical variables (Detmer & Slawski, 2018); however, the paper demonstrated that group lasso can be used to select categorical variables, but did not explore the role of different coding strategies in the actual fitting of the model. Additionally, we

were not able to find any applications within psychology which used this method, suggesting that additional dissemination may be required to improve adoption. As mentioned above, categorical variables need to be coded using different coding strategies when regression methods are applied. Specifically, we can define all indicators for a categorical variable as a group. In this way, the algorithm can either include all indicators associated with one categorical predictor or completely exclude these indicators. When all the variables are in one group, group lasso performs as ridge regression; while, when all the variables are their own group, group lasso performs as lasso. The advantage of group lasso is 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.

In our wage data, using our proposed application of group lasso, the algorithm either includes all indicators within a categorical variable (single, married, widowed, divorced, separated in the case of variable *Marital Status*) or excludes the set of indicators. This property of group lasso increases the ability to make omnibus claims about the predicting ability of the categorical variable (e.g., marital status predicts wage). Lasso does not take category membership information into consideration when doing variables selection. As was seen in previous sections, some categories within the categorical variable are selected, while others are left out. Take the variable *Marital Status* in our lasso model with dummy coding strategy for example. The lasso model regarded widowed participants to have same salary as single if other predictors are also same. The result is hard to interpret because nothing can be concluded about the omnibus predictive value of the variable *Marital Status*. However, in group lasso *Marital Status* is an important variable of the outcome variable wage if the group is selected into the model, or different marital status does not lead to important differences in the outcome variable if the group is excluded from the model. Based on these properties, group lasso seems like a promising alternative to lasso when dealing with categorical predictors. Because group lasso treats the variables in a group as a whole set, it seems less likely to be impacted by the coding strategy. Similar to lasso regression, we used the wage data to explore whether group lasso estimates different models with different types of coding strategies. Specifically, we explored the impact of coding strategy on the same outcomes we investigated with lasso: variable selection and out-of-sample prediction.

We estimated group lasso models in the same procedure that we estimated lasso models. We used six categorical variables (*Race, Job Class, Health, Health Insurance, Marital Status,* and *Education*) and one continuous variable (*Age*) to predict the outcome variable *Wage*. We trained three different group lasso models with three coding strategies (dummy, contrast, and Helmert). We encoded the categorical variables in the same way as we did in lasso regression. Each categorical variable is represented by $k-1$ predictors where $k$ is the number of categories. Therefore, we estimated the outcome wage variable using the same 15 predictors as lasso. The training and testing datasets that we used to estimate the models are the same as those we used for lasso. Not only do we examine the performance of group lasso in it's own right, but we also compare the variable selection and prediction accuracy between the lasso and group lasso models.

**Variable Selection.** Group lasso models perform the same variable selection even with different coding strategies. In our case, all variables were selected in all three models. Take the categorical variable *Marital Status* for instance. No two categories are treated the same in either of the three models in lasso models, while all variables are selected in the three group lasso models. The group lasso's property of variable selection is different from lasso's. Lasso's variable selection is affected by the coding strategies. However, the performance of variable selection for group lasso seems stable across different coding strategies.

**Prediction Accuracy.** We examined whether group lasso recreates the same means for each category within the categorical variable. We calculated the predicted wage of each category with the variable *Marital Status* for each of three group lasso models. The results are shown in Table 11. Similar to lasso, group lasso estimates each category mean within a categorical variable differently. So though variable selection is not impacted by coding strategy for group lasso, recreation of means is impacted by coding strategy in group lasso. Similarly, MSEs differ across the three models.

Similar to lasso, we plotted the category means recreated by group lasso models with different coding strategies and the reference values in Figure 1 (the right column of each coding strategy). Reference values remain the same as they are in lasso regression and are different across different coding strategies. Group lasso models using different coding strategies shrink all category means to their corresponding intercept. Therefore, category means are recreated

differently when different coding strategies are chosen, leading to different model fits and prediction accuracy. Comparing between lasso and group lasso, we can see that category means recreated by group lasso in general are closer to intercepts than those recreated by lasso. However, sometimes in lasso, a particular category mean is much closer to the model intercept than that in group lasso. Take the dummy-coded model for instance. The Widowed category mean in lasso model is closer to the true Single category mean than the Widowed category mean in the group lasso model. However, the category means for the Married, Separated, and Divorced are closer to the true Single category mean in the group lasso model than those in the lasso model. The differences in the shrinkage effect between lasso and group lasso can be explained by the differences in their penalty parameters (Equation 3 and 7), especially on categorical variables. Lasso adds the penalty parameter to the sum of the L1 norms of the category coefficients, while group lasso first uses L2 norm to sum the category coefficients within each variable and then adds penalty parameter to the sums across variables. In other words, each category coefficient is penalized in lasso model, but in group lasso model it is the sum of all category coefficients that gets penalized. Therefore, if one coefficient is shrunk to zero in the lasso model and the corresponding coefficient in group lasso is greater than zero, then the associated category mean is closer to the reference value in lasso model. However, if the group lasso model chooses to include one categorical variable, even none of the category coefficients are zero most of them will be close to zero due to the shrinkage effect. In our case, the coefficient of the Widowed category is shrank to zero in dummy-coded lasso model while the corresponding coefficient in dummy-coded group lasso model is greater than zero. Thus, the Widowed is treated the same as the Single category and the Widowed category mean is closer to the true Single category mean. In group lasso, as all categories within the variable *Marital Status* are included in group lasso models, coefficients are penalized on a group level and therefore most are smaller than the corresponding coefficients in lasso models.

**Group Lasso Summary.** For each coding strategy, we examined the group lasso's variable selection, calculated means for categories within the variable *Marital Status*, and the overall prediction accuracy. From Tables 11, we can conclude that group lasso partly solves the issues caused by choices of different coding strategies in lasso regression. Group lasso's variable

selection is not affected by the coding strategy. In other words, even when different coding strategies are used, group lasso models still perform the same variable selection. 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 it seems possible that the robustness of group lasso across coding strategies may come at a cost of prediction accuracy.

This trade off between prediction accuracy and robustness leads to some additional concerns about the group lasso. In particular, we are interested in when the set of indicators for a categorical variable will be selected into the model. Will the set of indicators for the categorical variables be selected or not if there are only a few categories with category means different from other categories within that variable? If that is the case, will group lasso's variable selection property lead to over-fitting issues because group lasso models may include several categories that are not good predictors of the outcome in order to include one category which is a good predictor of the outcome?

## Monte Carlo Simulation

In this section, we used Monte Carlo simulation to explore one of the potential weaknesses of the group lasso, over-fitting. The group lasso models may select more variables than necessary into the model, leading to large variance and low prediction accuracy. Monte Carlo simulation allows us to randomly generate and analyze data through repeated random sampling from a population with pre-specified characteristics. Using this method we can systematically fit group lasso models in order to find patterns across these models. The purpose of using Monte Carlo simulation is to investigate in what situations group lasso models have over-fitting issues and low prediction accuracy. We explore a particularly extreme case, where across all categories within one categorical variable, only one category differs from the rest. We call this category a *dominant* category and the others are referred to as *non-predictive*. The non-predictive category is always used as the reference category.

**Method.** In the simulation, we created a categorical variable with one dominant category and several non-predictive categories. The data set is designed in a way such that the dominant category has a different category mean than all other categories; while non-predictive categories have means which are all the same, equal to 0. Categorical variables are encoded by dummy coding strategy. Besides the categorical variable, we also included a continuous variable following a normal distribution with mean equal to 0 and variance equal to 1. The outcome variable is created by adding corresponding category means from the categorical variable, value of the continuous variable, and a random error following standard normal distribution. For optimal prediction, only the variable which estimates the difference between the dominant category and other non-predictive categories should be included in models built on the dataset. Those variables associated with non-predictive categories should have no effect on the outcome variable and not be selected in the model. In Equation 3, we see that the number of categories within categorical predictors may also effect how the $\beta$ coefficients are estimated and how the model selects predictors. The property is embedded in $\lambda \sum_{g=1}^{G} \left|\beta_{I_j}\right|_2$ within the formulae, which takes both the number of categories within the categorical variable and size of category coefficients into consideration. To explore the effect of the number of categories within the categorical variable, we made simulations with different numbers of non-predictive categories (1,2,3,4). Moreover, to figure out how the difference between the dominant category mean and non-predictive category means affects group lasso's prediction accuracy and predictor selection, we simulated different dominant category means (0.1, 0.2, 0.3). For each combination of number of categories and mean difference between dominant and non-predictive categories, we randomly generated 500 simulations of size 1200.

With each simulation, we applied lasso and group lasso regression. Specifically, we split each dataset into training and testing parts randomly according to the 8:2 ratio. Then we applied lasso and group lasso on the same training data. We selected the penalty parameter in the same way as we built lasso and group lasso models. For each data set and each method (group lasso and lasso) we calculated the MSE (mean squared error), which indicates the model's prediction accuracy and we recorded whether the model included the dominant category and whether the model included non-predictive categories. We calculated the average prediction accuracy of each

method by taking average of the 500 MSEs produced by the models in the same condition (number of categories and mean difference). For each condition we also calculated the proportion of models which included the dominant category and the proportion of models which included non-predictive categories. For group lasso, the two proportions are the same because group lasso either includes or excludes all categories within the categorical predictor.

**Results.** We first find that in all cases lasso has a higher prediction accuracy than group lasso, indicated by lower MSEs (Table 12). Though the differences in MSE of lasso and group lasso are small, they are consistent across different conditions. Secondly, for both group lasso and lasso regression, when the number of non-predictive categories increases, the probability for models to include the dominant category decreases, but this probability drops faster for group lasso models across effect sizes (Figure 3). For group lasso, when the difference between dominant category mean and non-predictive categories means is small, this probability drops faster than when the difference size is large. Specifically, when the effect size is small (dominant category mean = 0.1), the probability for group lasso to include the dominant category drops from 0.998 to 0.37 with the increase of number of categories, while the probability for lasso is relatively stable. When the effect size is big (dominant category mean = 0.3), the probability for group lasso to include the dominant category only drops from 1 to 0.67. Additionally from Figure 2, we can tell that when the number of non-predictive categories stays the same, the probability for group lasso to include non-predictive categories increases when the difference between the dominant category mean and non-predictive categories increases, while the probability for lasso is approximately the same. For example, when the difference between dominant category and non-predictive means is 0.1 and the number of categories equal to 5, the probability to include non-predictive categories for lasso is 0.332, and that for group lasso is 0.16. When the difference increases to 0.3, the probability for lasso is 0.645, and that for group lasso is 0.864. For both models, the probability to include non-predictive categories decreases when the difference between dominant category and non-predictive means stays the same, and the number of non-predictive categories increases.

To more closely examine potential over-fitting issues in group lasso, we focus on the case when the difference between dominant category and non-predictive categories is large. Figure 2 shows that when the difference is 0.3, group lasso always has higher probability than lasso to

include non-predictive categories. Recall that group lasso either includes the dominant category and non-predictive categories or excludes all categories. Large dominant category mean leads to group lasso's high possibility to include the dominant category and non-predictive predictors. In comparison with lasso, group lasso is more likely to include non-predictive categories when the dominant category mean is large. In this case, group lasso can over-fit the data because group lasso is more likely to include categories that are not supposed to be in the model. This also explains group lasso's lower prediction accuracy than lasso in Table 12. For example, when the difference between the dominant category and non-predictive categories is 0.3, the difference between group lasso's and lasso's MSE is bigger than those when the dominant category mean is smaller (see Table 12).

**Simulation Summary.** Using Monte Carlo simulation, we concluded that group lasso may over-fit data under certain conditions. Specifically, when few categories differ greatly from the other categories and the other categories contribute little to predicting the outcome variable, 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 through exploratory analysis in advance. Otherwise, they may need to use other regression methods because group lasso may over-fit issues. Looking for these effects may be particularly difficult in cases with many predictors and limited theoretical knowledge are driving the modeling, which is often 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 which could be very tedious.

## Discussion

Lasso has recently been adopted as a promising analytic method in psychological science due to its two major advantages over linear regression: variable selection and prediction accuracy. However, we have demonstrated that when there are categorical variables in the model, both of these qualities are sensitive to the coding strategy selected for the categorical variables. 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 method should be used? In the wage data example, lasso overall predicts better than group lasso, which predicts better than linear regression. However, there is no guarantee that these qualities will hold across other datasets. Researchers may want to balance the pros and cons of these methods above and beyond prediction accuracy. Which method should we choose when dealing with data with categorical predictors: linear regression, lasso, group lasso, or something else entirely? We explore potential solutions to this issue with categorical predictors in lasso based models.

**Exploring Potential Solutions**

Regardless of which of the following solutions researchers choose, one thing is required: transparency. 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 would be imperative for reproducing or replicating the research. 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. Each of the recommended approaches relies on the priorities of the researcher, in particular weighting the priorities of interpretability, best prediction, accurate variable selection, and robustness to coding decisions.

**Prioritize Interpretability.** In cases when a certain coding strategy provides increased interpretability of the coefficients in the model, the most interpretable coding strategy could be used. This comes at the risk of having a worse predictive model. This idea of interpretability is still very much rooting in the origins of linear regression, rather than machine learning. In particular, because the coefficient estimates in lasso regression are biased, they should not be interpreted directly. Rather, after variables selection is completed, common recommendations are to fit a linear regression model which only includes the selected variables (e.g., Hastie et al., 2015). However, it would seem odd to include a different coding strategy in the follow-up linear regression, as compared to the lasso regression. Thus it makes sense to use a coding strategy for each categorical variable which would be most interpretable, if the variables are selected in. With respect to the use of lasso, this would typically involve using coding strategies like dummy coding

or contrast coding where when individual predictors are dropped the interpretation of the remaining coefficients are unchanged. However, coding schemes like Helmert coding require the presence of all predictors in order 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).

A particular difficulty of this method is that oftentimes machine learning approaches are used in cases when 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 order or reference groups. Additionally, by prioritizing interpretability the researcher may be losing prediction accuracy, which is often one of the reasons that machine learning approaches are used.

**Prioritize Prediction.** One option in estimating lasso or group lasso models would be to try many different coding strategies in order to select the one with the most promise with regard to prediction accuracy. This process should likely be completed using the training data, so as not to influence the final estimates of prediction accuracy using an independent sample of the data. One issue with this method is that it may be very computationally intensive. There are technically an infinite number of coding strategies that one could use for any given variable. With multiple categorical variables in the dataset, one would want to try different combinations of coding strategies, as there is no reason to expect that using the same coding strategy for each variable would result in maximized prediction accuracy. Additionally, it is unclear the types of gains which could occur in prediction accuracy using this method, and for some researchers the benefits in prediction accuracy may not be worth the additional computational time. Indeed, with the wage data, the largest differences in MSE corresponded to an average difference in prediction of $763.54. Depending on the research aims, this may be a useful gain in prediction accuracy, and for other research aims this may seem menial.

Another alternative, if prediction accuracy is of highest priority, is to use alternative machine learning approaches which are robust to coding strategy. Alternative approaches like classification and regression trees (CART) are unaffected by coding strategy, because categorical predictors are treated as a single variable (Finch & Schneider, 2007). One downside to these types

of models is that they are often less interpretable, and they do not provide the "regression like" estimates which many researchers in psychology rely on for interpreting their results.

**Prioritize Robust Variable Selection.** Based on the simulation results, the group lasso is robust to coding strategy choices with respect to variable selection. In addition, the prediction accuracy seems to vary less when using group lasso compared to lasso; however, this does not necessarily mean that prediction is 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 groups within each categorical variable, group lasso seems to be an optimal choice. Nevertheless, this may come at a cost in 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).

**Field Norms.** Just as standardizing continuous variables has become a field norm, it may be possible for researchers within a field to agree on a single coding strategy throughout the field. However, additional research would be needed in order to proceed with a single recommended coding method. This may also be restricting to researchers who have clear reason to use a different coding strategy other than the field recommended norm. This may not ultimately be too problematic if researchers can be transparent about which method is being used for a given analysis, to ensure reproducibility. However, to a large extent the field norm seems to be dummy coding, as this is often a default in software, though it is not immediately clear whether dummy coding is optimal in most or any cases.

## Future Directions

This research opens many paths for future exploration of the intersection of lasso and group lasso regression with categorical predictors, and beyond. There are a few particular directions which we believe would be most beneficial for improving the state of research in this area.

First, while exploring the role of coding strategy in lasso and group lasso models, it became immediately clear that the intercept plays an important role in the interpretation of these models. The typical practice within lasso is not to penalize the intercept (Wu & Lange, 2008). However, 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 Figures 1). This brings about the question of whether it would be appropriate to penalize the intercept 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.

This issue of penalization of the intercept brings up an important characteristic of contrast coding which suggests itself as 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 coefficients are dropped from the model, the group that is indicated by this predictor is assumed to be equal to the group mean, rather than pulling the group directly toward another group. This means that the selection of the "reference" group is likely to have less of an impact on parameter estimates in comparison to dummy coding, because by selecting a reference group in dummy coding, that group's mean is then unpenalized (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 not categorical variables in the model and all continuous predictors were standardized (i.e., sample average). This presents an opportunity for contrast coding to be a reasonable default if researchers are unsure how to proceed with selecting an alternative coding strategy; however, the use of contrast coding should be studied in a variety of contexts more in-depth in order to assess it's 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 means varied across different coding strategies. In particular, in the wage data, the widowed group was particularly small ($N = 19$ out of 3000 observations). This resulted in two problems which merit further investigation: how group size can impact estimates and interact with selection of coding strategy and reference group and how training and testing data should be split in the presence of small groups. Each of these is discussed in turn.

First, the observed behavior of the widowed group in the wage data made it clear that the estimates for this group were very unstable and of any of the groups, most affected by coding strategy. Figure 1 show how much the widowed group predicted mean varies across different coding strategies, and that this variance is much larger than any of the other groups. This can also be seen in Table 6 where most of the predicted group means show a range of about 2.0 across the different coding strategies but the widowed group ranges by about 5.0. Similarly in Table 8 we can see that the estimates of all of the group means have the greatest bias when the widowed group is used as the reference category, and the lasso model with dummy coding and widowed as the reference group has the highest MSE. This suggests that there may be a particularly important interaction of group size and choice of coding strategy, where selecting a small group to be a reference group causes additional instability in the estimates, and should be avoided. However, future research should examine the role of group size in the fitting of lasso and group lasso models; in particular, it would be interesting to know if group lasso models are less sensitive to these issues.

A second issue brought up by having small groups is the difficulty of splitting testing and training data sets when groups are particularly small. This may become particularly problematic when there are many categorical variables which include many groups. Previous researchers have resolved to collapse groups that are particularly small (e.g., racial/ethnic minorities). It is unclear how this practice impacts estimates for these groups, and in general is not recommended in other analytic practices (e.g., Tarantola & Dellaportas, 2005). Throughout this project, there were certain cases, where the training-testing split of the data resulted in no cases from certain groups being selected into the training dataset. This made it impossible to fit a model in the testing set which provided a unique estimate for the missing group. Methods for splitting the data such as block randomization may provide more accurate estimates of means for small groups, if the groups can be evenly split across training and testing sets of data. However, this issue is compacted by methods which repeatedly split data, or split the data into smaller parts (e.g., K-fold cross-validation for selecting the tuning parameter), and it is important that future research explores alternative ways to estimate unique group means for small groups, rather than a priori collapsing them with other groups.

## Conclusion

Overall, our findings suggest that researchers should be aware that their coding strategies will likely impact both variable selection and prediction accuracy when using lasso regression and their prediction accuracy when using group lasso. We demonstrate cases when group lasso may have lower prediction accuracy than lasso, in particular when there is a dominant group (one group that differs from all other groups). The choice of what method to use (lasso or group lasso), what coding strategy to use, and which group order to use or reference category to choose, may all depend on the priorities of the researcher with respect to maximizing interpretability, prediction accuracy, variable selection, and robustness. It is important that the choices of the researcher or analyst in how categorical variables are included in lasso and group lasso models are transparently reported to improve the reproducibility and replicability of research in this area. Future research needs to explore specific practices in this area (e.g., penalization of the intercept, use of contrast coding) and how small groups should be accounted for in order to optimize prediction accuracy for these groups and avoid collapsing across groups.

Psychologists are quickly adopting the new and incredibly useful tools being 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 psychology researchers to predict out of sample data, which may be particularly important in clinical settings. 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, ANOVA, and structural equation modeling frameworks. 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 psychology.

References

Ammerman, B. A., Jacobucci, R., Kleiman, L. L., E. M.and Uyeji, & McCloskey, M. S. (2018). The relationship between nonsuicidal self-injury age of onset and severity of self-harm. *Suicide and Life-Threatening Behavior*, *48*(1).

Bainter, S., McCauley, T. G., Wager, T. D., & Losin, E. R. (2019, 02). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/j8t7s`

Burningham, Z., Leng, J., Peters, C. B., & Huynh, T. (2018). Predicting psychiatric hospitalizations among elderly veterans with a history of mental health disease. *PsyArXiv*, *6*(1), 1-9.

Detmer, F., & Slawski, M. (2018). A note on coding and standardization of categorical variables in (sparse) group lasso regression.

Finch, H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees: Three- and five-group cases. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3(2)*, 47-57.

Hastie, T., Robert, T., & Wainwright, M. (2015). Statistical learning with sparsity: The lasso and generalizations.

Heckman, C. J., Handorf, E. A., Darlow, S. D., Ritterband, L. M., & Manne, S. L. (2017). An online skin cancer risk-reduction intervention for young adults: Mechanisms of effects. *Health Psychology*, *36*(3), 215 – 225.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: with applications in r*.

Leach, H. J., O'Connor, D. P., Simpson, R. J., Rifai, H. S., & Mama, S. K. (2016). An exploratory decision tree analysis to predict cardiovascular disease risk in african american women. *Health Psychology*, *35*(4), 397 – 402.

Ma, H., Chang, W., & Cui, G. (2012). Ecological footprint model using the support vector machine technique. *PloS one*, *7*(1).

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.

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.

Sauer, S., Buettner, R., Heidenreich, T., Lemke, J., Berg, C., & Kurz, C. (2018). Mindful machine learning: Using machine learning algorithms to predict the practice of mindfulness. *European Journal of Psychological Assessment*, *34(1)*, 6-13.

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

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018a). Ai beats doctors at predicting heart disease deaths. *PloS one*, *13(8)*, 1-20.

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018b, 08). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE*, *13*, 1-20.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267 – 288.

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat. 2*, *1*, 224-244.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, *68*, 49-67.

Zilcha-Mano, S., Errázuriz, P., Yaffe-Herbst, L., German, R. E., & DeRubeis, R. J. (2019). Are there any robust predictors of "sudden gainers," and how is sustained improvement in treatment outcome achieved following a gain? *Journal of Consulting and Clinical Psychology*, *87(6)*, 491-500.

| Marital Status | $D_1$ | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|---|
| 1. Single | 0 | 0 | 0 | 0 |
| 2. Married | 1 | 0 | 0 | 0 |
| 3. Widowed | 0 | 1 | 0 | 0 |
| 4. Divorced | 0 | 0 | 1 | 0 |
| 5. Separated | 0 | 0 | 0 | 1 |

Table 1

*Dummy Coding*

| Marital Status | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| 1. Single | 1 | 0 | 0 | 0 |
| 2. Married | 0 | 1 | 0 | 0 |
| 3. Widowed | 0 | 0 | 1 | 0 |
| 4. Divorced | 0 | 0 | 0 | 1 |
| 5. Separated | -1 | -1 | -1 | -1 |

Table 2

*Contrast Coding*

| Marital Status | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| 1. Single | -4/5 | 0 | 0 | 0 |
| 2. Married | 1/5 | -3/4 | 0 | 0 |
| 3. Widowed | 1/5 | 1/4 | -2/3 | 0 |
| 4. Divorced | 1/5 | 1/4 | 1/3 | -1/2 |
| 5. Separated | 1/5 | 1/4 | 1/3 | 1/2 |

Table 3

*Helmert Coding*

| Marital Status | *Dummy* | *Contrast* | *Helmert* |
|---|---|---|---|
| 1. Intercept | 92.735 | 103.10172 | 103.1017 |
| 2. $X_1$ | 26.126 | -10.36707 | 12.959 |
| 3. $X_2$ | 6.804 | 15.75854 | -17.556 |
| 4. $X_3$ | 10.425 | -3.56307 | 2.649 |
| 5. $X_4$ | 8.481 | 0.05754 | -1.943 |

Table 4

*LR example for coding*

Each column of the table represents one coding strategy and row2 - row5 represent the coefficients of the indicator $X_i$ for each coding strategy.

| Group Lasso Regression | | |
|---|---|---|
| *Dummy* | *Effect* | *Helmert* |
| Married - Single | Single - Average | Single - Average(Married + Widowed + Divorced + Separated) |
| **Widowed - Single** | Married - Average | Married - Average(Widowed + Divorced + Separated) |
| Divorced - Single | Widowed - Average | **Widowed - Average(Divorced + Separated)** |
| Separated - Single | Divorced - Average | **Divorced - Separated** |
| Black - White | White - Average | White - Average(Black + Asian + Others) |
| Asian - White | Asian - Average | **Black - Average(Asian + Others)** |
| Others - White | Others - Average | Asian - Others |
| High School Grad - Less than High School Grad | High School Grad | Less than High School Grad - Average(High School Grad + Some College + College Grad + Advanced Degrees) |
| Some College - Less than High School Grad | Some College | High School Grad - Average(Some College + College Grad + Advanced Degrees) |
| College Grad - Less than High School Grad | College Grad | Some College - Average(College Grad + Advanced Degrees) |
| Advanced Degrees - Less than High School Grad | Advanced Degrees | College Grad - Advanced Degrees |
| Information - industrial | Job Class | Information - industrial |
| Very Good or Higher - Good or Lower | Health | Very Good or Higher - Good or Lower |
| No - Yes | Health Insurance | No - Yes |
| Age (Continuous Variable) | Age (Continuous Variable) | Age (Continuous Variable) |

Table 5

*Variable Selection for Different Coding Strategies by Lasso*

Variables with no background color are selected by all three models, and those with grey as the background color are only selected by the model represented by the column.

Table 6

*Prediction Accuracy for Different Coding Strategies by Lasso.*

Rows represent the categories within the variable, and the middle three columns represent models with different coding strategies. Last column is the actual category mean from the training data.

| Coding strategies | *Dummy* | *Contrast* | *Helmert* | Actual Category Mean |
|---|---|---|---|---|
| 1.*Single* | 69.988 | 68.983 | 70.074 | 68.096 |
| 2.*Married* | 87.115 | 86.251 | 87.089 | 85.593 |
| 3.*Widowed* | 69.988 | 71.604 | 74.565 | 69.409 |
| 4.*Divorced* | 73.665 | 73.594 | 74.565 | 72.392 |
| 5.*Separated* | 76.195 | 75.719 | 74.565 | 75.655 |
| MSE | 1200.803 | 1201.114 | 1201.386 | / |

| Variables | 1.*Single* | 2.*Married* | 3.*Widowed* | 4.*Divorced* | 5.*Separated* |
|---|---|---|---|---|---|
| Intercept | 69.988 | 87.344 | 74.410 | 74.447 | 76.062 |
| 1.*Single* | . | -17.322 | -4.225 | -4.404 | -6.214 |
| 2.*Married* | 17.127 | . | 13.161 | 12.978 | 11.200 |
| 3.*Widowed* | 0 | -15.133 | . | -2.220 | -4.077 |
| 4.*Divorced* | 3.678 | -12.733 | 0 | . | -1.623 |
| 5.*Separated* | 6.207 | -9.616 | 2.140 | 2.036 | . |
| MSE | 1200.803 | 1200.401 | 1201.084 | 1200.950 | 1201.069 |

Table 7

*Model Coefficients of Categorical Variable Marital Status for Different Reference categories.*

Each column represents one model, and each row represents coefficients of the predictor produced by five models. "." is the reference category for this model, and 0 means that the model does not select this category into the model.

Table 8

*Prediction Accuracy for Different Reference Categories by Lasso.*

Rows represent the categories within the variable, and the middle five columns represent models with different reference categories. Last column is the actual category mean from the training data.

| Coding strategies | 1.*Single* | 2.*Married* | 3.*Widowed* | 4.*Divorced* | 5.*Separated* | Actual Category Mean |
|---|---|---|---|---|---|---|
| 1.*Single* | 69.988 | 70.022 | 74.410 | 70.043 | 76.062 | 68.096 |
| 2.*Married* | 87.115 | 87.344 | 87.571 | 87.425 | 87.262 | 85.593 |
| 3.*Widowed* | 69.988 | 72.211 | 74.410 | 72.227 | 71.985 | 69.409 |
| 4.*Divorced* | 73.665 | 74.611 | 74.410 | 74.447 | 74.437 | 72.392 |
| 5.*Separated* | 76.195 | 77.728 | 76.550 | 76.483 | 76.602 | 75.655 |
| MSE | 1200.803 | 1200.401 | 1201.084 | 1200.950 | 1201.069 | / |

Table 9

*Prediction Accuracy for Different Coding Strategies using Singular Design Matrices by Lasso*

Rows represent the categories within the variable, and middle three columns represent models with different coding strategies. Last column is the actual category mean from the training data.

| Coding strategies | *Dummy* | *Contrast* | *Helmert* | Actual Category Mean |
|---|---|---|---|---|
| 1.*Single* | 70.702 | 69.653 | 69.817 | 68.096 |
| 2.*Married* | 87.737 | 86.885 | 87.020 | 85.593 |
| 3.*Widowed* | 75.135 | 73.393 | 74.247 | 69.409 |
| 4.*Divorced* | 75.135 | 74.538 | 74.247 | 72.392 |
| 5.*Separated* | 75.135 | 75.792 | 75.081 | 75.655 |
| MSE | 1204.915 | 1204.072 | 1203.446 | / |

Table 10

*Prediction Accuracy for Different Coding Strategies using Singular Design Matrices by Group Lasso*

Rows represent the categories within the variable, and middle three columns represent models with different coding strategies. Last column is the actual category mean from the training data.

| Coding strategies | *Dummy* | *Contrast* | *Helmert* | Actual Category Mean |
|---|---|---|---|---|
| 1.*Single* | 68.957 | 69.225 | 69.038 | 68.096 |
| 2.*Married* | 85.677 | 85.977 | 85.710 | 85.593 |
| 3.*Widowed* | 73.311 | 73.651 | 73.675 | 69.409 |
| 4.*Divorced* | 73.171 | 73.488 | 73.414 | 72.392 |
| 5.*Separated* | 75.468 | 75.868 | 75.463 | 75.655 |
| MSE | 1198.235 | 1199.726 | 1198.547 | / |

Table 11

*Prediction Accuracy for Different Coding Strategies by Group Lasso*

Rows represent the categories within the variable, and middle three columns represent models with different coding strategies. Last column is the actual category mean from the training data.

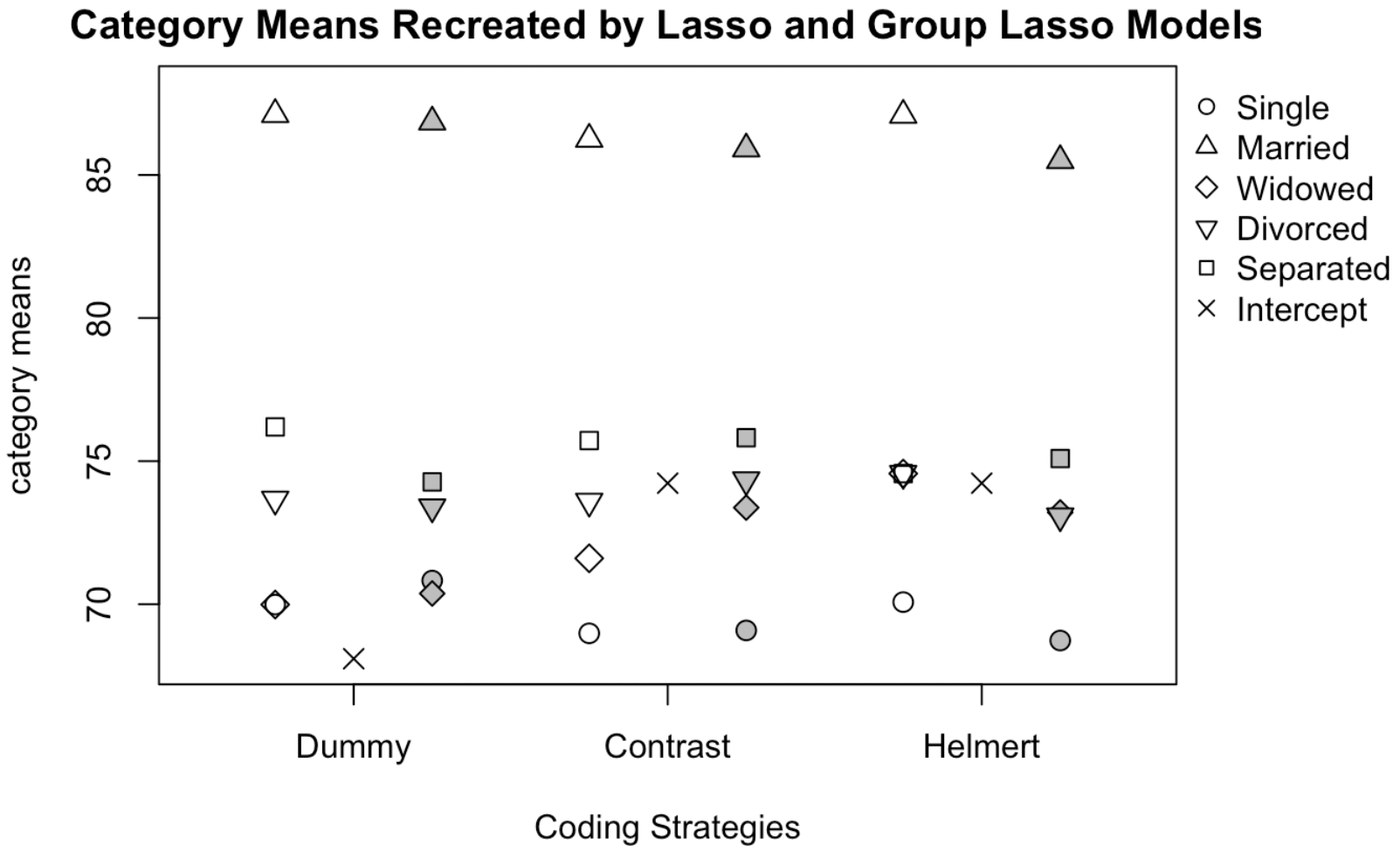
| Coding strategies | *Dummy* | *Contrast* | *Helmert* | Actual Category Mean |
|---|---|---|---|---|
| 1.*Single* | 70.827 | 69.084 | 68.731 | 68.096 |
| 2.*Married* | 86.852 | 85.913 | 85.506 | 85.593 |
| 3.*Widowed* | 70.374 | 73.376 | 73.207 | 69.409 |
| 4.*Divorced* | 73.395 | 74.343 | 73.081 | 72.392 |
| 5.*Separated* | 74.273 | 75.817 | 75.090 | 75.655 |
| MSE | 1199.496 | 1199.474 | 1197.668 | / |

Table 12

*Differences in MSE of Lasso of Group Lasso models for Monte Carlo Simulation. "Difference" means subtracting MSEs for lasso from MSEs for group lasso.*

| Dominant Category | Number of Categories | | | |
|---|---|---|---|---|
| mean | 2 | 3 | 4 | 5 |
| 0.1 | 0.0024 | 0.0028 | 0.0029 | 0.0003 |
| 0.2 | 0.0016 | 0.0020 | 0.004 | 0.0008 |
| 0.3 | 0.0045 | 0.0029 | 0.0029 | 0.0030 |

*Figure 1*. Graphical Presentation of Category Means Recreated by Lasso and Group Lasso Models with different Coding Strategies. The left column of each coding strategy represents the lasso model; and the right column represents the group lasso model. Reference values are different across coding strategies. Reference value is the true category mean for Single in the dummy coding; the average of true category means in the contrast coding; and the true category mean for Separated in the Helmert coding.



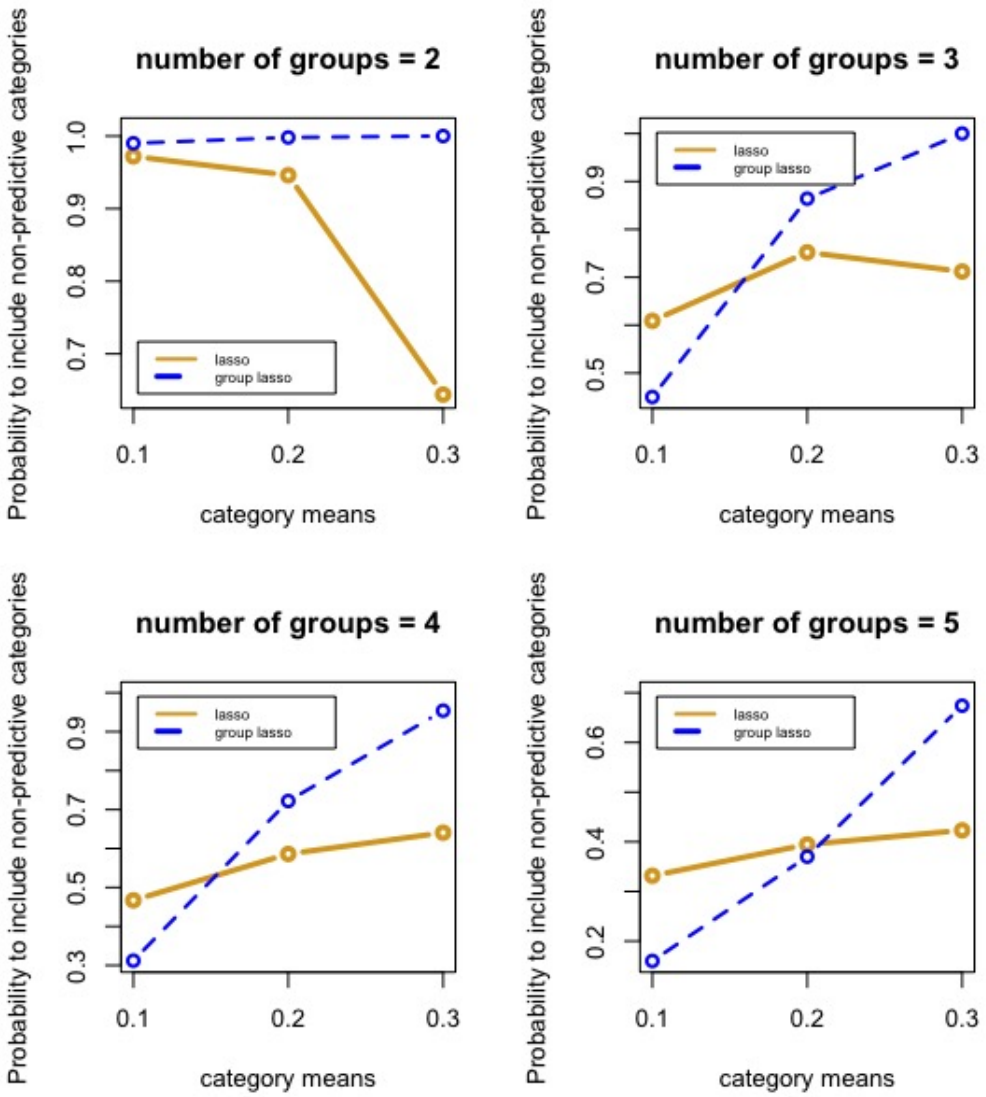



*Figure 2*. Comparison between probability of to include non-predictive categories. Simulations in the same plots have the same number of groups.

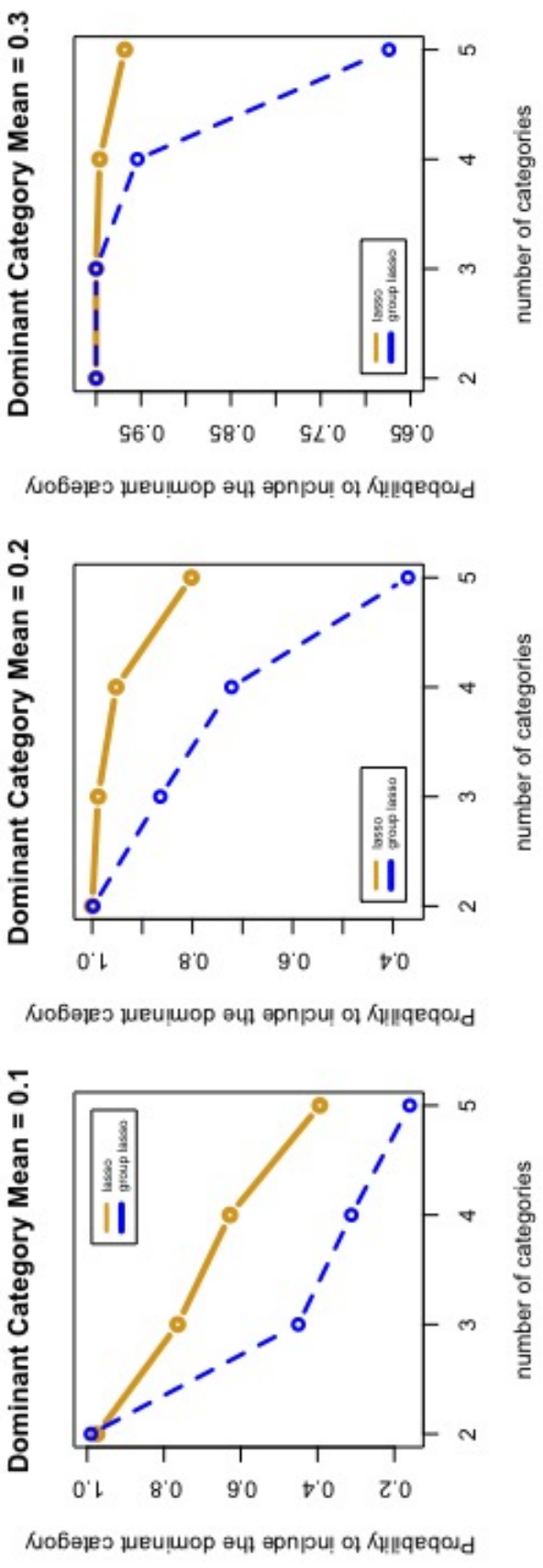



*Figure 3.* Comparison between probability of lasso and group lasso models to include the dominant category under different Dominant Category Means