Prevention Science (2019) 20:575–584
https://doi.org/10.1007/s11121-018-0963-9

Efficient Exploration of Many Variables and Interactions Using Regularized Regression

Tyson S. Barrett[1] · Ginger Lockhart[1]

Published online: 30 November 2018

Abstract

The prevention sciences often face several situations that can compromise the statistical power and validity of a study. Among these, research can (1) have data with many variables, sometimes with low sample sizes, (2) have highly correlated predictors, (3) have unclear theory or empirical evidence related to the research questions, and/or (4) have difficulty selecting the proper covariates in observational studies. Modeling in these situations is difficult—and at times impossible—with conventional methods. Fortunately, *regularized regression*—a machine learning technique—can aid in exploring datasets that are otherwise difficult to analyze, allowing researchers to draw insights from these data. Although many of these methods have existed for several decades, prevention researchers rarely use them. As a gentle introduction, we discuss the utility of regularized regression to the field of prevention science and apply the technique to a real dataset. The data ($n = 7979$) for the demonstration consisted of 76 variables (151 including the modeled interactions) from the Youth Risk-Behavior Surveillance System (YRBSS) from 2015. Overall, it is clear that regularized regression can be an important tool in analyzing and gaining insight from data in the prevention sciences.

Keywords Regularized regression · Adolescent development · Health · Drug/substance abuse

Introduction

Prevention science and epidemiology often confront several situations that can compromise the statistical power and validity of a study. Among these, research can (1) have data with many variables, sometimes with low sample sizes, (2) have highly correlated predictors, (3) have unclear theory or empirical evidence related to the research questions, and/or (4) have difficulty selecting the proper covariates in observational studies. Statistical modeling challenges in these situations are well-documented (Wooldridge 2013), such as misleading inference and results that cannot be adequately replicated. Many of the approaches to dealing with these situations are tedious and error-prone, are simply a work-around, or heavily rely on many null hypothesis significance tests (NHST; Cumming 2014), which can lead to "p-hacking"[1] or introduce high levels of researcher bias.

In general, theory and prior literature must act as guides in developing and building statistical models (Lockhart et al. 2011). Yet, in situations where theory and prior literature are vague, data can be misused or underutilized due to the limitations inherent in many common statistical techniques (e.g., structural equation modeling, generalized linear regressions, ANOVA family of methods). For example, these methods generate estimation problems in multi-collinearity (Vanderweele 2012; Wooldridge 2013). It can also be difficult to find a parsimonious model without arbitrarily removing variables or basing it on significance testing.

Prevention researchers often seek to uncover malleable risk-causing mechanisms and want to understand which populations (or subpopulations) are most at risk. Although machine learning approaches, in general, have been developed under a very different paradigm (i.e., prediction is more important than explanation), prevention research and machine learning should not be considered mutually exclusive. That is, many machine learning techniques can offer valuable

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11121-018-0963-9) contains supplementary material, which is available to authorized users.

✉ Tyson S. Barrett
tyson.barrett@usu.edu

[1] Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, UT 84322, USA

[1] Only reporting and discussing significant relationships when many were tested and incurs a high type-I error rate.

insights about patterns that are otherwise overlooked. In this paper, we discuss two well-developed machine learning techniques that are compatible with the predictive goals of prevention science: (1) *cross-validation* and (2) *regularized (penalized) regression*.

Cross-Validation

The central purpose of cross-validation is to determine the extent to which a predictive model generalizes across independent datasets. In this approach, the analyst uses a so-called *training set* to generate a predictive model and then validates this model on a subset of unknown data, called the *testing set.* In addition to assessing the generalizability of model estimates, the technique can help with the "over-fitting" problem often found in conventional regression and other regression-type models (McNeish 2015). Though many types of cross-validation techniques are in use, we focus on 10-fold cross-validation to highlight its utility on smaller samples and as a supplementary validation tool to use in conjunction with other machine learning techniques. In this approach, the data are split into 10 approximately equal random splits. The model is then fit to 9 of the 10 splits and then tested on the 10th. This is done 10 times so that each 10th of the data is used as the testing set. Unlike other cross-validation approaches, 10-fold cross-validation is useful when the data cannot be split into independent training and testing sets without losing too much information. Further, cross-validation is compatible with virtually any type of modeling approach, making it a flexible tool that can be applied across a variety of research questions.

Regularized Regression

Regularized regression, on the other hand, is a family of machine learning methods (Hastie et al. 2009; James et al. 2007) that is built on conventional regression with a penalty term added, as shown in the equation below (where $\lambda f(\beta_j)$ represents the penalty term).

$$\min_{\beta}\left[y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right]^2 + \lambda f\left(\beta_j\right)$$

This shows that the method performs least squares or maximum likelihood as in conventional regression but the β_js are penalized. Thus, the solution minimizes the residuals plus a function of the penalized βs simultaneously. As such, the estimated βs will no longer be "unbiased." However, this adjustment can add several attributes that are otherwise missing from other regression-based approaches:

a. The cross-validated prediction accuracy, and hence evidence for model generalization, is often higher than conventional methods. This suggests the model is better able to uncover important relationships in the data that predict the outcome.[2]
b. Correlated predictors are common in prevention research, where many risk factors are highly related. Fortunately, many regularized regressions can handle this in a much more theoretically appeasing way—correlated variables will have estimates more closely together without masking one or the other. For example, if two copies of the same variable are used (perfectly correlated) in the same regularized regression model, where individually they have an effect of 0.5, when they are both included in the model, their effects are reduced to approximately 0.25. This feature is particularly helpful in understanding causal mechanisms because one variable will not be arbitrarily masked.
c. Regularized regressions can handle many variables, even to the point where there are more predictors than there are observations (i.e., $p > n$). In many fields, this is already common (e.g., genetics, biology), but is likely becoming more common in the prevention sciences, particularly in exploratory work when researchers take a more holistic view, have low funding, and/or seek to explore risk among small subpopulations. Additionally, this allows one to test for important subpopulations particularly at risk without performing many individual analyses (e.g., all demographic, behavioral, and other important indicators can be included). Finally, many interactions can also be included, testing for moderating variables that can mask important relationships.
d. In exploratory settings, more information can be obtained per dataset that can help researchers design more well-informed confirmatory models.
e. Instead of relying on human judgment solely for variable selection, the methods can select the most important predictors of the outcome (e.g., Kessler et al. 2015). This helps remove researcher bias and can reduce type-I error because fewer tests are performed. Given the recommendation by the American Psychological Association of becoming less reliant on NHST (Cumming 2014), these methods are particularly useful since they do not use significance tests in selecting variables. The technique can also be used to select covariates to improve causal interpretation of the modeled effects.

Results from regularized regression are interpreted similarly to those of conventional regression. As stated previously, the estimates, due to the penalty terms, are biased. This bias

[2] It is critical to note that when using the term "important" in this context, it refers to importance in the model's predictive accuracy. In other words, a variable is important if, in conjunction with the other variables in the model, it is useful in predicting the outcome accurately. This is distinct from the usual discussion of significance and effect size in convention statistics.

helps reduce overfitting (McNeish 2015) and has better prediction on data not used in the model fitting procedure. However, the penalty component of the model introduces a challenging trade-off: it produces biased estimates that may ultimately hamper interpretability in order to construct a more predictive, flexible model. In response to this limitation, Bécu et al. (2015) and Hastie et al. (2009) recommend using regularized regression to select the model (and use it for any predictions) and then use that selected model in an un-penalized regression to obtain unbiased estimates (supposing all assumptions hold). This two-part approach capitalizes on the utility of the predictive power of the overall modeling procedure while still maintaining the interpretability of the individual parameters. Importantly, for this approach to work best, the unbiased estimation should be done on a separate data set (or a subset of the data set not used for the variable selection).

Three of the most common regularized regression techniques include *ridge* (Hoerl and Kennard 1970), *lasso* (least absolute shrinkage and selector operator; Tibshirani 2011), and *elastic net* (Zou and Hastie 2005). Ridge regression manages highly correlated predictors (multi-collinearity) well by pushing together the estimates of correlated variables and can handle more predictors than observations. Lasso regression, on the other hand, can select the most predictive variables from among many. It struggles, much like un-penalized regression, more with correlated predictors by arbitrarily selecting one of the collinear variables. Elastic net balances the strengths of these two techniques by integrating each of their penalties in the modeling procedure.

With the combined benefits of ridge regression and lasso, elastic net can handle many common data situations encountered in prevention science (e.g., multi-collinearity, low sample size, many variables). It has the ability to select the most important variables from among many, which can aid in important decisions regarding intervention target selection. These selections can include interactions, quadratic terms, and both continuous and categorical predictors. It can also handle more variables than observations ($p > n$) unlike nearly all conventional methods. Finally, it can adequately account for multi-collinearity by pushing the estimates of correlated predictors closer together.

The penalty on elastic net has the form:

$$\lambda\left[\sum_{j=1}^{p}\left(\alpha|\beta_j| + \frac{(1-\alpha)\beta_j^2}{2}\right)\right]$$

where α is a tuning parameter (range 0–1) that adjusts the balance between the lasso penalty of $|\beta_j|$ and the ridge penalty of β_j^2. When α is closer to 1, the elastic net model will focus more on selection; when the α is closer to 0, the model will handle multi-collinearity more strongly. Hence, elastic net has two tuning parameters: the α—which controls the balance of the penalties—and λ—which controls the magnitude of the penalty (e.g., smaller λ means the estimates will be closer to conventional regression). When $\alpha = 1$, the model is a lasso model; when $\alpha = 0$, the model is a ridge model. These tuning parameters are often selected based on cross-validated predictive accuracy (i.e., the accuracy of the model's prediction on data not used to fit the model). The best performing, theoretically meaningful model is often the selected model.

Before demonstrating the elastic net on prevention data, it is important to highlight a few nuances of elastic net and regularized regression in general. First, in many of the statistical packages that perform elastic net, the dummy coding of categorical variables is necessary. This means that, in some models, only a subset of the dummy-coded variables may be selected (e.g., only one or two of the four dummy codes for a five-level categorical variable may be selected). Obviously, this changes the immediate interpretation of what the dummy-coded coefficient means. In response to this issue, newer "group lasso" methods have been developed (Friedman et al. 2010; Simon et al. 2013; Simon et al. 2015). Group lasso is more restricted in its implementation at this point and, therefore, is not a focus of this paper.

Second, although cross-validation can reduce statistical power by reducing the sample size, elastic net and related methods often perform well under situations where conventional methods would struggle with low power (Hastie et al. 2009; Zou and Hastie 2005). For example, sample sizes at 40 would represent an issue for cross-validation with conventional techniques. Certainly, having only 36 subjects per 10-fold cross-validation can limit statistical power and introduce bias. However, special forms of regularized regression have performed well (appropriate variable selection, high predictive accuracy) under these circumstances (Zou and Zhang 2009).

General Steps to Perform Elastic Net

In the following section, we describe a series of steps for performing elastic net and follow with a demonstration on a real dataset. In this application, we specify a seven-step framework for using elastic net to discover the critical predictors and interactions of marijuana use and incorporate 10-fold cross-validation to assess the generalizability of the model. For a more comprehensive treatment of the range of options within regularized regression and cross-validation techniques, the reader is encouraged to consult *An Introduction to Statistical Learning* (James et al. 2007) and *The Elements of Statistical Learning* (Hastie et al. 2009).

Step 1. *Perform necessary exploratory data analyses to understand the distributions of important variables (especially dependent variables) and missing data patterns*. If many values are missing, the researcher can consider imputation techniques (e.g., Graham

et al. 2007), use only the observations available, or remove the variable(s) from the model. Based on the information discovered here, select the family of model to be fit (e.g., linear, logistic, cox). Again, the sample size can be lower, in general, than in most generalized linear models (e.g., even in cases where the number of predictors is higher than observations).

Step 2. *Decide which variables will be included in the elastic net model*. In general, specifying the model is similar to that done in un-penalized regression. The researcher must decide which variable is the dependent (outcome) variable and which are independent (predictors or correlates). The model can have hundreds or thousands of predictors (although the lower the "signal-to-noise" ratio—more noise variables and fewer "true" predictors—the worse the variable selection performs). We advocate including any variables that may influence the outcome and/or interact with other variables to influence the outcome.

Step 3. *Create dummy codes for any categorical variables included in the model*. In software such as the R statistical environment, there are functions that do this for the researcher.

Step 4. *Specify the elastic net model*. Based on information from step 2, include the outcome variable and all the predictors in the model. This can include continuous and categorical predictors, interactions, quadratic terms, and other variable transformations as can be done in conventional regression. This specified model will be fit to the data in step 6 using cross-validations.

Step 5. *Select the number of folds in cross-validation* (one, five, and ten folds are most common). This can be used to help researchers select the proper parameters for the elastic net, as is done automatically in many packages (e.g., the *caret* package in R [Kuhn 2008]). This cross-validation will be applied to the fitted models below.

Step 6. *Fit the elastic net models*. Fit the cross-validated elastic net models to various λ and α values. This produces several models to choose from. Generally, the "best" model is where the cross-validated error (i.e., the prediction error from the model on data the model was *not* trained on) is the lowest or within one standard error of the lowest (i.e., the "one-SE rule"). This selection process is straightforward in R (e.g., the *glmnet* and *caret* packages in R [Friedman et al. 2016; Kuhn 2008] have built-in functions for performing cross-validation). For example, in the *caret* package, both the λ and α are automatically selected based on the best cross-validated prediction accuracy. Therefore, this step selects the best parameters and, simultaneously, allows the elastic net model to select the most important predictors. If step 7 is not planned, report the final models, including their prediction accuracies, and highlight relationships via estimates and data visualization. Among these relationships, it is informative to assess the relative importance of the variables in the model. The relative importance is another way to see what variables are driving the predictive accuracy (i.e., the variables that appear to be most closely tied to the outcome).

Step 7. (Recommended). *Using the selected predictors, fit an un-penalized model (e.g., linear regression, generalized linear models) to obtain the unbiased estimates and their confidence intervals*. Assess the model's fit to the data and the model assumptions (e.g., normality, homoscedasticity). Report the findings explaining both the selection process and the final model. If possible, it is best if this final unbiased model uses different data than used in the elastic net models. However, this is often not possible with smaller samples. When using the same data for both elastic net and unbiased models, this should be stated, including the potential issues of selecting and inferring from the same data (e.g., higher type-I error).

These steps are also useful when performing ridge, lasso, and other related methods. Below, steps 1–7 are used to demonstrate the use of elastic net to understand the possible impact of asthma on marijuana use.

Demonstration

The following demonstration explores the impact of asthma on marijuana use via a series of elastic net models: (1) a base model with main effects ("Model 1") and (2) the base model plus interactions between all covariates and the asthma variable ("Model 2"). The analyses in the demonstration were performed in R version 3.4.4 using the *glmnet* and *caret* packages (Friedman et al. 2016; Kuhn 2008). All corresponding R code for this example is available at https://osf.io/kat82/.

Data from the Youth Risk-Behavior Surveillance System in 2015 was used to address our research question. The data contain self-reported information on hundreds of health, behavior, and demographic indicators. For the present analysis, predictors of current marijuana use among high school students ($n = 7979$) were investigated.

Our research questions are progressive, building from one to the other, to investigate the effect of asthma on the risk of having ever smoked marijuana. Adolescents with asthma have been shown to be at-risk for many behavioral problems (Pinquart and Shen 2011) and may be particularly at risk for marijuana

use. Our data agreed with this prior evidence, showing that those with asthma have meaningfully higher rates of currently using marijuana (26.0%) compared to those without asthma (22.0%; $\chi^2(4) = 16.8, p < .001$). Although this simple relationship is noteworthy on its own, it can be useful to determine whether asthma is one of the most important predictors among many variables, what some important covariates may be, and whether asthma appears to interact strongly with the covariates. From a prevention standpoint, understanding the relative importance of these influences can help interventionists, practitioners, and policymakers prioritize their efforts as they work to address public health concerns. Additionally, due to the flexibility of this modeling technique, the discovery of critical interactions in the model can yield valuable information on which individuals with asthma are most at risk for marijuana use.

Step 1. *Perform necessary exploratory data analyses to understand the distributions of important variables and missing data patterns*. Logistic elastic net regression (i.e., logistic regression with the elastic net penalty) was chosen because the outcome (marijuana use) is dichotomous. The amount and patterns of missingness in variables that were being considered for inclusion in the model were also evaluated. Due to missing values in the outcome (marijuana use), the main predictor (asthma), and other covariates, the total sample size for the analyses was 7909 adolescents. This sample size provides ample statistical power to select and estimate effects in both elastic net and logistic regression.

Step 2. *Decide which variables will be included in the elastic net model*. All variables that were of interest or could have been an important covariate in the models were included. Here, that included many demographic, health, and behavioral variables. We removed redundant variables (e.g., "never used cigarettes" and "used cigarettes") and variables with a high level of missing values (~ 10% missing or more). In this case, many of the variables removed due to missing values were redundant with variables included in the models. In total, 76 variables were included in Model 1 and 151 variables—including interactions—were included in Model 2. Summary statistics, stratified by the outcome variable (marijuana use), of the predictors of interest are shown in Table 1. The variable descriptions shown are worded similarly to that found in the *2015 YRBS Data User's Guide* (2016)

Step 3. *Create dummy codes for any categorical variables included in the model.* The majority of our predictors were categorical and were dummy coded using built-in R functions.

Step 4. *Specify the elastic net model.* The outcome of interest—ever having used marijuana—was included as the outcome variable along with all predictor variables that were potentially important, as discussed in step 2. In addition, Model 2 included interactions between all predictors and asthma to determine important interactions among the variables. Lastly, the sample was not randomly collected (i.e., a probability sample was used); therefore, the sample weights were included in the model as well.

Step 5. *Select the number of folds in cross-validation.* Ten-fold cross-validation was chosen, due to our sample size being sufficient for 10 splits and due to convention.

Step 6. *Fit the elastic net models.* To select both the α and λ values, we fit the model to many tuning values to find which has the lowest cross-validated error using the built-in functionality of the *caret* package. This package fits each parameter combination automatically and reports information regarding the best fitting model, simplifying the actual use of the elastic net model. Using this built-in approach of testing many combinations of α and λ values, the best cross-validated predictive accuracies resulted in $\alpha = .6$ and $\lambda = 0.002$ for Model 1 and in $\alpha = .6$ and $\lambda = 0.008$ for Model 2. Notably, both of these models predicted marijuana use with 82% accuracy.

One of the first things to assess here beyond the predictive accuracy is the relative importance of the variables included in the models. Figure 1 shows the 40 most important variables and their relative importance. Several main effects, along with three interactions, were among the most important predictors. The biggest drivers of the high predictive accuracies are generally risk behaviors (ever vaped, ever drank alcohol, ever had sexual intercourse). A few demographic characteristics showed up as important as well, including age and race. Further, other psychological/behavioral characteristics were important, including diet, suicidality, and perception of being overweight.

The three interactions found to be relatively important were tobacco use with asthma, ever had sex with asthma, and race (multiple Hispanic racial categories) with asthma. Because the research question of interest is in understanding these interactions, we will analyze their significance and meaning in step 7.

Step 7. *Using the selected predictors, fit an un-penalized model to obtain the unbiased estimates and their confidence intervals*. This step is a highly recommended approach in the prevention sciences due to its production of "unbiased" estimates. Here, the exploration of interactions with asthma were of

Table 1 Descriptive statistics of the main variables of interest

	Marijuana		
	Not used $n = 4851$	Used $n = 3128$	p value
Asthma	1066 (22%)	813 (26%)	< .001
Race/ethnicity			< .001
Asian	250 (5.2%)	62 (2%)	
American Indian	44 (0.9%)	38 (1.2%)	
Black	334 (6.9%)	282 (9%)	
Pacific Islander	22 (0.5%)	12 (0.4%)	
White	2446 (50.4%)	1375 (44%)	
Hispanic/Latino	708 (14.6%)	529 (16.9%)	
Multiple Hispanic	829 (17.1%)	672 (21.5%)	
Multiple	218 (4.5%)	158 (5.1%)	
Age			< .001
14 years or younger	603 (12.4%)	153 (4.9%)	
15 years	1317 (27.1%)	543 (17.4%)	
16 years	1250 (25.8%)	787 (25.2%)	
17 years	1101 (22.7%)	1045 (33.4%)	
18 years or older	580 (12%)	600 (19.2%)	
Female	2544 (52.4%)	1560 (49.9%)	0.026
Grade			< .001
9th grade	1421 (29.3%)	460 (14.7%)	
10th grade	1271 (26.2%)	703 (22.5%)	
11th grade	1137 (23.4%)	926 (29.6%)	
12th grade	1017 (21%)	1038 (33.2%)	
Ungraded or other grade	5 (0.1%)	1 (0%)	
Hispanic/Latino	1537 (31.7%)	1201 (38.4%)	< .001
Usually wear seatbelt	148 (3.1%)	207 (6.6%)	< .001
Threatened/injured by a weapon at school	134 (2.8%)	231 (7.4%)	< .001
Ever raped	146 (3%)	366 (11.7%)	< .001
Ever planned suicide	544 (11.2%)	662 (21.2%)	< .001
Ever used tobacco	122 (2.5%)	327 (10.5%)	< .001
Ever vaped	1106 (22.8%)	2471 (79%)	< .001
Ever drank alcohol	2249 (46.4%)	2857 (91.3%)	< .001
Ever drank 5+ consecutive drinks of alcohol	282 (5.8%)	1213 (38.8%)	< .001
Ever used cocaine	10 (0.2%)	336 (10.7%)	< .001
Ever used meth	6 (0.1%)	139 (4.4%)	< .001
Ever used ecstasy	10 (0.2%)	341 (10.9%)	< .001
Ever used Rx medications (without Rx)	285 (5.9%)	1045 (33.4%)	< .001
Ever had sexual intercourse	1147 (23.6%)	2185 (69.9%)	< .001
Perceive self as overweight	1575 (32.5%)	1080 (34.5%)	0.06
Eat fruit	436 (9%)	266 (8.5%)	0.481
Watch 3+ hours of TV per day	1179 (24.3%)	849 (27.1%)	0.005
Use computer 3+ hours per day	2113 (43.6%)	1397 (44.7%)	0.344
Ever had an HIV checkup	252 (5.2%)	472 (15.1%)	< .001
Obtained mostly A's and B's in school	3836 (79.1%)	1937 (61.9%)	< .001
Drink 3+ cans of soda per day	271 (5.6%)	279 (8.9%)	< .001

particular interest. The previous model uncovered three important interactions. Using logistic regression (while adjusting for the weighting of the probability sample), we can gain an understanding of the interactions' significance and meaning. To maintain proper weighting in the sample, we did not split the data into different sets for the selection of the variables and the estimation of the unbiased parameters. But to reduce the chance of type-I error, we adjusted the significance threshold to $p < .01$.

Table 2 presents the results of the unbiased models for each test. The first model, "Main Effects", just reports the main effects. Notably, asthma is not significant here when controlling for these other risk factors. The next models, termed "Interaction 1," "Interaction 2," and "Interaction 3," tested each individual interaction on its own. Finally, the "Full Model" shows all the interactions and main effects together. The full model shows that only one of the three important interactions (asthma × tobacco) was statistically significant at $p < .01$.

To understand the meaning of the important interactions, we further visualized the risk of marijuana use by interaction (see Fig. 2). For the interaction with tobacco use, there is no difference between having asthma and not when the individual has used tobacco, but those with asthma are more at risk of using marijuana when the individual has not used tobacco. For the individuals that have not had sexual intercourse, there is no difference between the asthma groups. However, for those who have had sexual intercourse, those with asthma are more likely to use marijuana (although this difference is not statistically significant here). Lastly, for the "White," "Black," and "Hispanic/Latino" racial ethnic groups, individuals with asthma have higher average risks of using marijuana than their non-asthmatic counterparts (although this difference is not statistically significant here).[3]

Discussion

Regularized regression techniques can aid researchers in understanding data in situations where it is otherwise difficult. Although in no way diminishing the great importance of theory and prior research in developing important research questions and study designs, elastic net can aid in situations where the researcher (1) has data with many variables and/or low sample sizes, (2) has highly correlated predictors, (3) has unclear theory or empirical evidence related to the research questions, and/or (4) has difficulty selecting the proper covariates in observational studies. Using regularized regression approaches, researchers can obtain actionable insights into the important variables, relationships, and interactions in the data that may play a role in designing more effective interventions and help understand for whom an intervention is most useful.

Further, it is notable that regularized regression can likely aid in causal inference, especially in observational, cross-sectional work. Because regression approaches assume that all variables that influence both the outcome and the predictor are in the model, regularized regression can be used to find those variables in the data that help satisfy this assumption (Urminsky et al. 2016). Although not touched on specifically

[3] All estimates from the elastic net and unbiased models are included in the Supplementary Table.

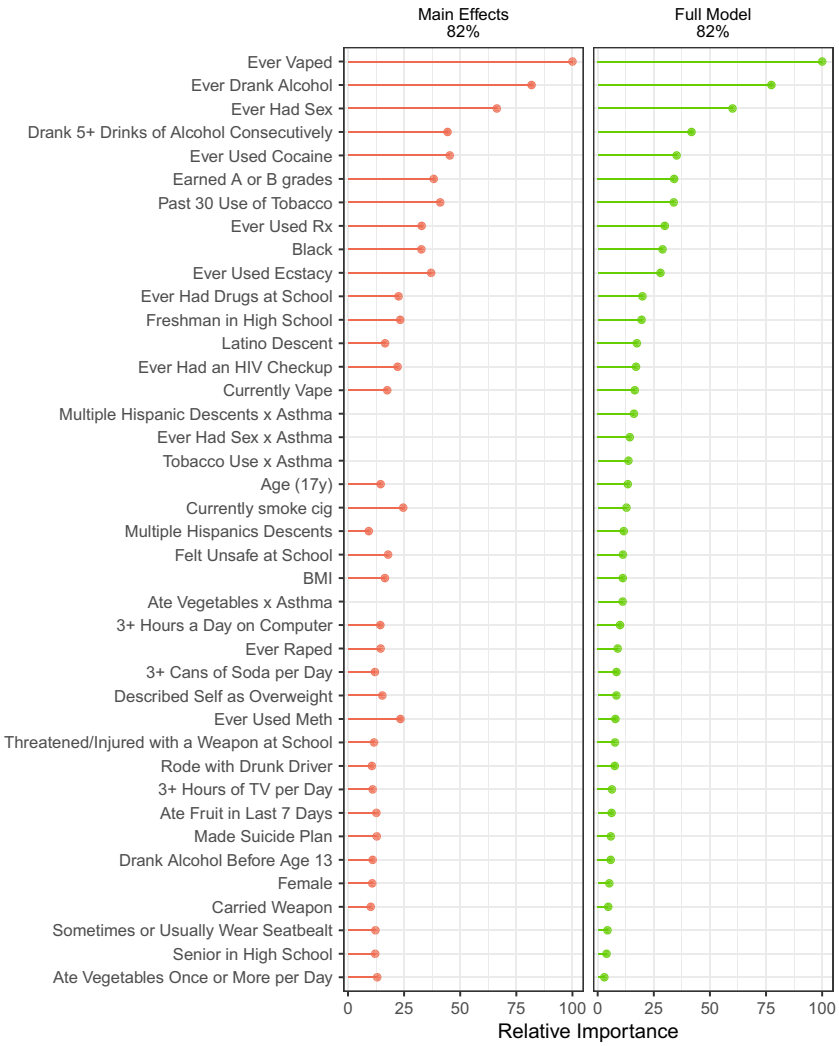

Fig. 1 The relative importance of the top 40 variables in the two models: Main Effects and Full Model

Table 2 Unbiased modeling results of the logistic regression models

	Marijuana use				
	Main effects	Interaction 1	Interaction 2	Interaction 3	Full Model
Ever had sex	1.88 ***	1.81 ***	1.89 ***	1.88 ***	1.81 ***
Ever used tobacco	2.37 ***	2.37 ***	2.20 ***	2.37 ***	2.22 ***
Race/ethnicity (Asian)					
American Indian	1.18 **	1.19 **	1.16 **	1.09 *	1.09 *
Black	1.02 ***	1.02 ***	1.00 ***	1.07 ***	1.07 ***
Pacific Islander	1.01 *	1.02 *	1.01 *	0.84	0.84
White	0.48 *	0.49 **	0.47 *	0.46 *	0.48 *
Hispanic/Latino	1.05 ***	1.05 ***	1.03 ***	1.01 ***	1.01 ***
Multiple Hispanic	1.06 ***	1.07 ***	1.05 ***	1.19 ***	1.19 ***
Multiple	0.4	0.4	0.39	0.39	0.41
Asthma	0.11	−0.05	0.04	0.16	−0.01
Interactions					
Ever had sex × asthma		0.31 *			0.33 *
Ever used tobacco × asthma			0.70 **		0.68 **
American Indian × asthma				0.4	0.44
Black × asthma				−0.2	−0.27
Pacific Islander × asthma				0.37	0.39
White × asthma				0.06	−0.03
Hispanic/Latino × asthma				0.16	0.14
Multiple Hispanic × asthma				−0.56	−0.61
Multiple × asthma				0.01	−0.11

Results are in log odds

*$p < .05$, **$p < .01$, ***$p < .001$

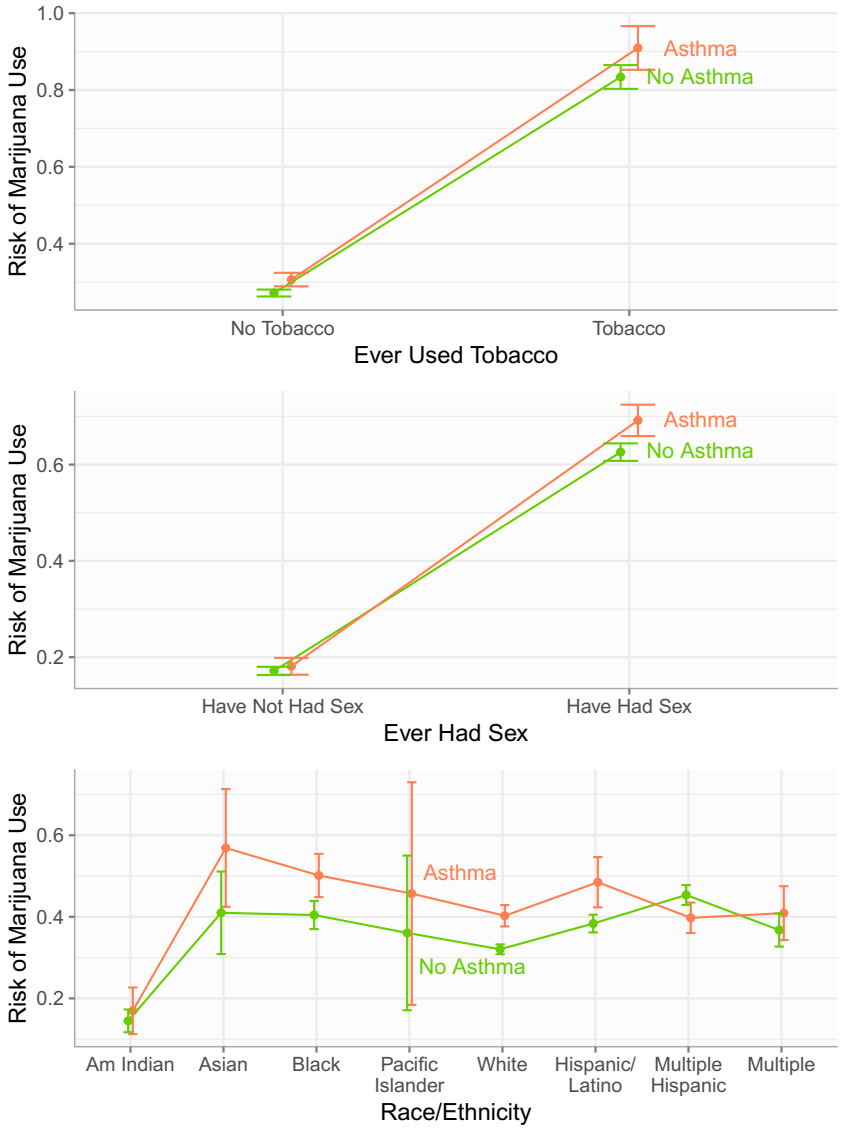

Fig. 2 The important interactions on the risk of marijuana use are shown, with ever used tobacco, ever had sexual intercourse, and race/ethnicity all interacting with asthma. The top panel (ever used tobacco) was statistically significant while the last two (ever had sexual intercourse and race/ethnicity) were not

in the example, regularized regression can be used to assess which variables are predictive of both the outcome and the predictor(s) of interest (Belloni et al. 2013). In conjunction with theory and prior literature, having this covariate selection tool can be beneficial. It is also notable that regularized regression can be used in longitudinal data as well, thus strengthening the exploration of causal mechanisms.

In this paper, we demonstrated how a researcher can understand more about an important relationship, in this case more about how asthma relates to the risk of marijuana use. Previous empirical work has shown asthma is a risk factor for poor behavioral outcomes (Pinquart and Shen 2011) but the interaction of asthma with other factors had, to our knowledge, not been previously explored. Using the information gleaned from the regularized regression models, we then assessed unbiased estimates of these interactions. Given the exploratory nature of these findings, further research is needed to investigate the mechanisms behind these relationships more thoroughly. Of particular benefit could be the use of longitudinal studies that can assess whether individuals with asthma, depending on tobacco use or sexual activity, present with more risk of marijuana use. This type of study could shed light on the temporal pattern, providing evidence for whom and when an intervention could be useful. Mediation analysis, in conjunction with longitudinal data and regularized regression, could shed light on the pathways of the effect (Zhao and Luo 2016).

Cautions and Limitations

Though regularized regression techniques offer many benefits to prevention researchers, some challenges and limitations must be acknowledged. First, the method, much like conventional regression, removes missing observations from the analyses. Imputation techniques can sometimes help here, but more research is needed on the effects of missing data mechanisms on

the robustness of the results across varying conditions. Second, confidence intervals are not naturally obtained from these approaches, and therefore, *p* values are not commonly computed either. Third, because elastic net, and machine learning in general, can handle hundreds (and even thousands) of predictors, it may be tempting to use it without giving the proper thought about the model specification. Since elastic net selects the predictors, one may be tempted to include all the variables in a data set into the model without proper understanding of the data, theory, or previous empirical work. In this way, the model can be highly misspecified, just as in conventional regression, producing misleading results. Fourth, prediction can be good even without any causal predictors (i.e., "correlation does not imply causation"). Thus, even with high predictive accuracy, we may not be able to derive any causal inference. Work in this area of causal inference with regularized regression, possibly under the framework discussed by Sauer et al. (2013), could be beneficial. Finally, without much more involved adjustments to the model (Zou 2006), the estimates are biased, generally toward zero.

Conclusions

Regularized regressions are powerful techniques that can be used to better understand important relationships among many variables, can adjust for multi-collinearity, and can help select important covariates. In conjunction with theory and prior research, elastic net can help researchers specify better informed models, including finding potentially important interaction effects. Researchers in prevention, and across the health and psychological sciences, should consider the benefits of utilizing regularized regression within their work.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approvals The present study uses data collected through the direction of the Centers for Disease Control and Prevention. As such, the proper ethical approvals were obtained through its supervision.

Informed Consent As in part three above, the Centers for Disease Control and Prevention were in charge of supervising the informed consent of each of the subjects.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

2015 YRBS Data User's Guide. (2016). https://doi.org/10.1016/j.jadohealth.2016.03.017.

Bécu, J.-M., Grandvalet, Y., Ambroise, C., & Dalmasso, C. (2015). Beyond support in two-stage variable Selection, 1–25. Retrieved from http://arxiv.org/abs/1505.07281. Accessed May 2017

Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies, 81*, 608–650. https://doi.org/10.1093/restud/rdt044.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. https://doi.org/10.1177/0956797613504966.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. ArXiv:1001.0736 [Math, Stat], 8. https://doi.org/10.1111/biom.12292

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). Package "glmnet": Lasso and elastic-net regularized general linear models. *R Package Version*, 23. Retrieved from https://www.jstatsoft.org/article/view/v033i01. Accessed May 2017

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206–213. https://doi.org/10.1007/s11121-007-0070-9.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Elements (Vol. 1). https://doi.org/10.1007/b94608

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2007). An introduction to statistical learning. Performance evaluation (Vol. 64). Springer US. https://doi.org/10.1016/j.peva.2007.06.006

Kessler, R., Warner, C., Ivany, C., Petukhova, M., Rose, S., Bromet, E. J., et al. (2015). Predicting suicides after psychiatric hospitalization in US Army soldiers: The Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). *JAMA Psychiatry, 72*, 49–57. Retrieved from. https://doi.org/10.1001/jamapsychiatry.2014.1754.

Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5).

Lockhart, G., Mackinnon, D. P., & Ohlrich, V. (2011). Mediation analysis in psychosomatic medicine research. *Psychosomatic Medicine, 73*, 29–43. https://doi.org/10.1097/PSY.0b013e318200a54b.Mediation.

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

Pinquart, M., & Shen, Y. (2011). Behavior problems in children and adolescents with chronic physical illness: A meta-analysis. *Journal of Pediatric Psychology, 36*, 375–384. https://doi.org/10.1093/jpepsy/jsq104.

Sauer, B., Brookhart, M. A., Roy, J. A., & VanderWeele, T. J. (2013). Covariate selection. In P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith, & M. M. Torchia (Eds.), *Developing a protocol for observational comparative effectiveness research: A user's guide* (pp. 93–108). Rockville, MD: Agency for Healthcare Research and Quality.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics, 22*, 231–245. https://doi.org/10.1080/10618600.2012.681250.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2015). Fit a GLM (or Cox model) with a combination of lasso and group lasso regularization.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 73*, 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Urminsky, O., Hansen, C., & Chernozhukov, V. (2016). Using double-lasso regression for principled variable selection. Available at SSRN 2733374, 1–70.

Vanderweele, T. J. (2012). Invited commentary: Structural equation models and epidemiologic analysis. *American Journal of Epidemiology, 176*, 608–612. https://doi.org/10.1093/aje/kws213.

Wooldridge, J. M. (2013). *Introductory econometrics* (4th ed.). Mason, OH: South-Western Cengage Learning. https://doi.org/10.1016/j.jconhyd.2010.08.009.

Zhao, Y., & Luo, X. (2016). Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators. Retrieved from http://arxiv.org/abs/1603.07749

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418–1429. https://doi.org/10.1198/016214506000000735.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, 67*, 301–320.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics, 37*, 1733–1751. https://doi.org/10.1214/08-AOS625.ON.