

CEHS Workshop:  
Reproducible Research via  
Lasso, Ridge, and Elastic Net

Outline

- 1 Why learn about regularized regression? And what is regularized regression?
- 2 How to apply them (with R code)
- 3 Some problems in R



Reproducibility is low in many fields

2 Aspects of Reproducible Research

**Methods:** computerized methods and code of data analysis  
**Results:** computerized methods and code of data analysis  
**Interpretation:** computerized methods and code of data analysis

What is a Good Problem?

- A challenging problem is one that requires selecting variables based on the p-value, rather than the best-fitting, multi-dimensional model (Greenberg, 2010; Platt et al., 2012).
- It is a problem that is not easily solved by high-dimensional data ("big data").

So what can we do?



**One Variable:** We will use the Cross Validation for model tuning to select the best fit.  
**An ensemble model:** The ensemble model is a set of models that are not highly generalizable. Adjust the cross-validation and assess the prediction accuracy.



Simple Steps to Use Elastic Net, Lasso or Ridge

- Step 1:** Understand the data. Summarize the data. Summarize the data. Summarize the data.
- Step 2:** Specify the model. Specify the model. Specify the model.
- Step 3:** Summarize the data. Summarize the data. Summarize the data.
- Step 4:** Number of models. Number of models. Number of models.
- Step 5:** Summarize the data. Summarize the data. Summarize the data.
- Step 6:** Summarize the data. Summarize the data. Summarize the data.

- Step 1:** Logistic Regression (The outcome is binary).  
**Step 2:**  $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1 \text{Predictor} + \epsilon$   
**Step 3:** Dummy coded variables
- Step 4:** We selected 10 folds for cross-validation.  
**Step 5:** Fit the cross-validated model.  
Check:  
• the cross-validated error  
• the selected variables between the folds and the 1-SE  
**Step 6:** We selected model at the 1-SE (usually more parsimonious than min).

Notes and Such

By removing a reliance on p-values, reducing researcher bias in variable selection, and providing replicable steps:

Reproducibility ↑ Generalizability ↑

Notes and Such

Theory and prior literature still act as guides but much of the arbitrary variable selection is gone.

Now in R  
Perform regularized regression in R using the cleaned data set I sent out to you.

If you didn't get it, visit:  
[github.com/CEHSWorkshop/RegularizedRegression/](https://github.com/CEHSWorkshop/RegularizedRegression/)

Tyson S. Barrett

@tysonstanley

tyson.barrett@usu.edu

tysonstanley.github.io

# CEHS Workshop: Reproducible Research via Lasso, Ridge, and Elastic Net

# CEHS Workshop:

## *Reproducible Research via Lasso, Ridge, and Elastic Net*

### *Outline:*

- 1 Why learn about regularized regression? And what is regularized regression?

Reproduc

3 As

## *Outline:*

- 1** Why learn about regularized regression? And what is regularized regression?
- 2** How to apply them (without R code)
- 3** Demonstration in R



Note: This is a workshop and so we want you to ask all the questions you have--don't hold back!

# Reproducibility is low in many fields

## 3 Aspects of Reproducible Research

**Methods:** can replicate the methods with same data and obtain same results

**Results:** can replicate the methods with independent data and obtain same results

**Inference:** can replicate the methods with independent data and obtain same inference

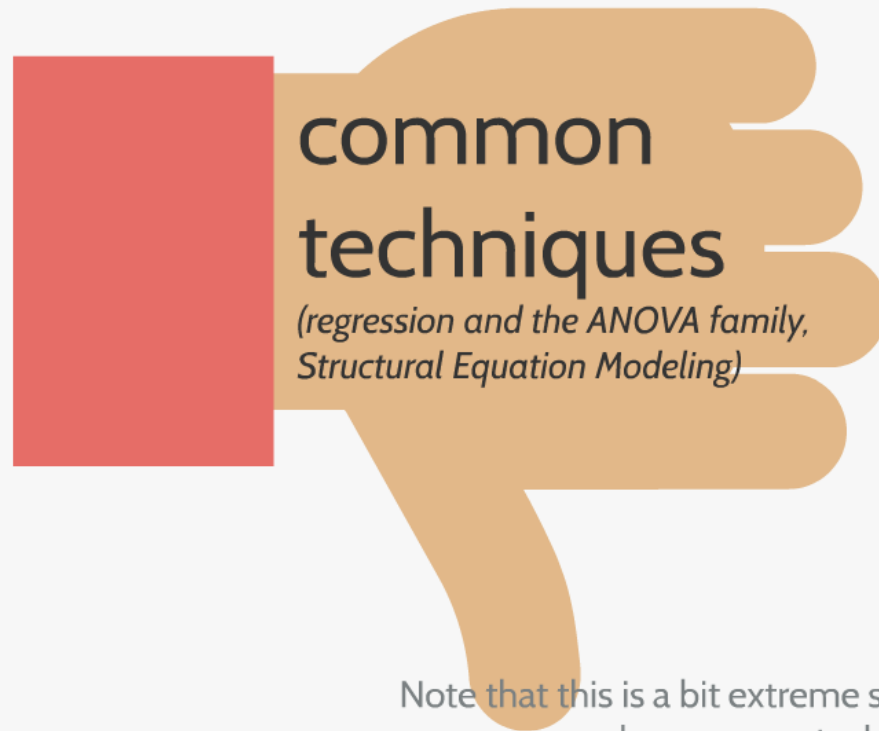
Goodman et al. (2015)

***What is Causing These Problems?***

## *What is Causing These Problems?*

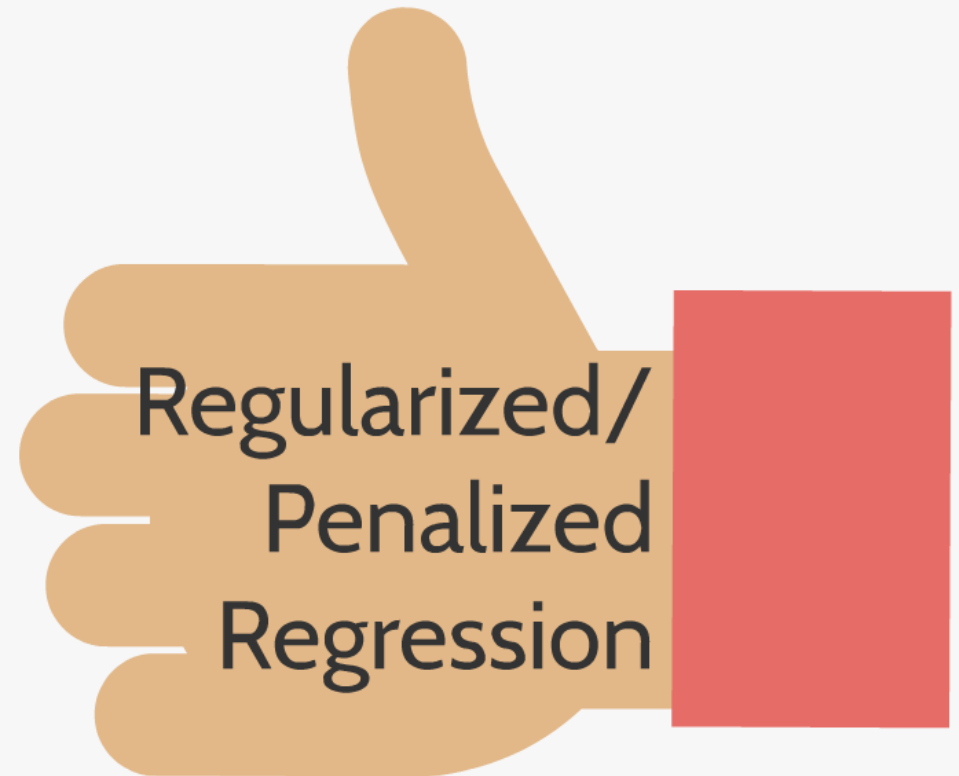
- P-hacking or even a reliance on p-values (selecting variables based on the p-value), researcher bias, over-fitting, multi-collinearity, among others (Cumming, 2014; Munafo et al., 2017)
- These problems are even worse in high-dimensional data ("big data")

# So what can we do?



**common  
techniques**

*(regression and the ANOVA family,  
Structural Equation Modeling)*



**Regularized/  
Penalized  
Regression**

Note that this is a bit extreme since there are great approaches to make common techniques great in big data

## Cross-Validation



Note that this is a bit extreme since there are great approaches to make common techniques great in big data

## Cross-Validation

We will also use Cross-Validation (for model tuning and to avoid overfitting)

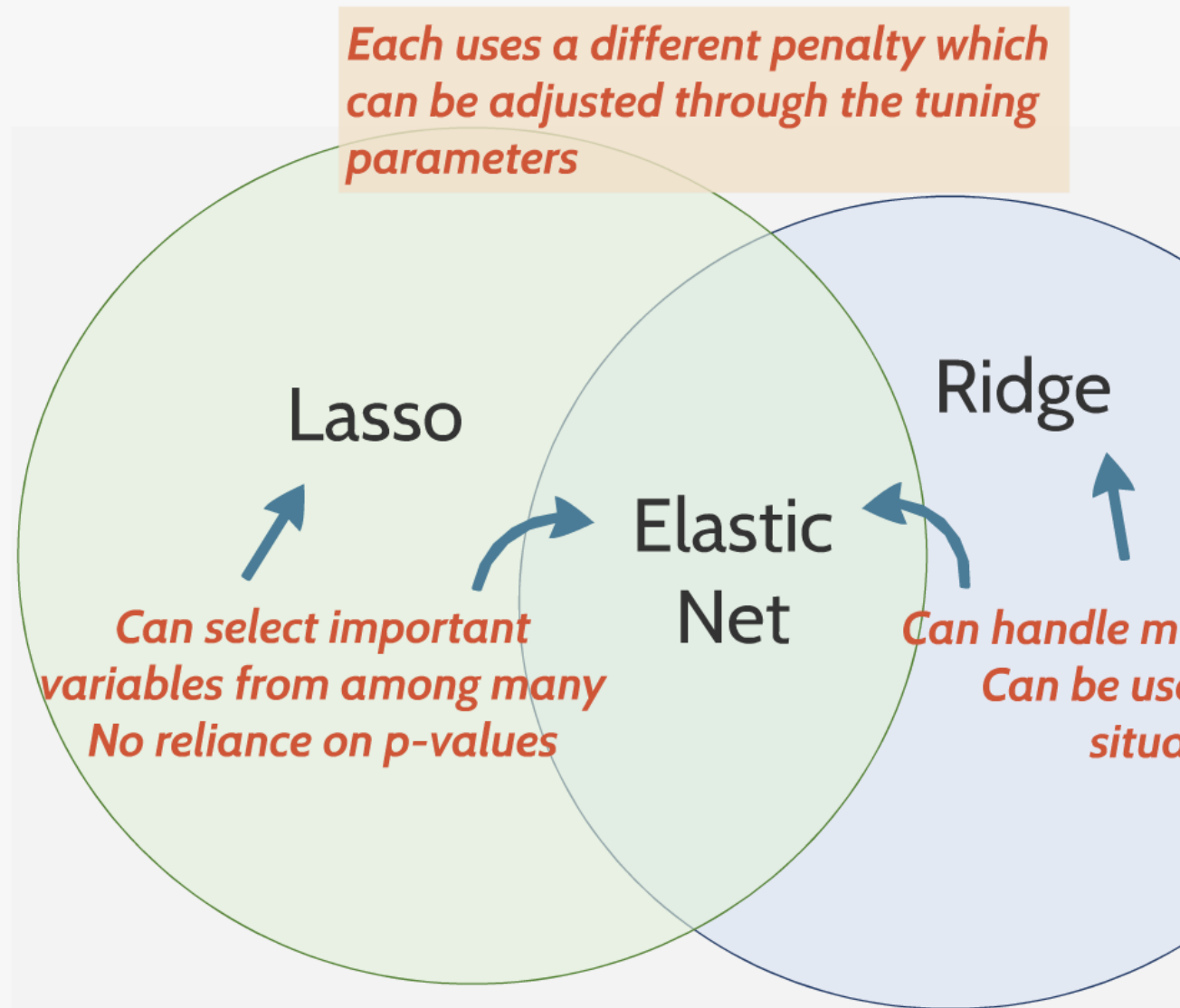
### *An overfit model*

It's results are technically unbiased but are not highly generalizable

Adjust: use cross-validation and assess the prediction accuracy

# Regularized Regressions

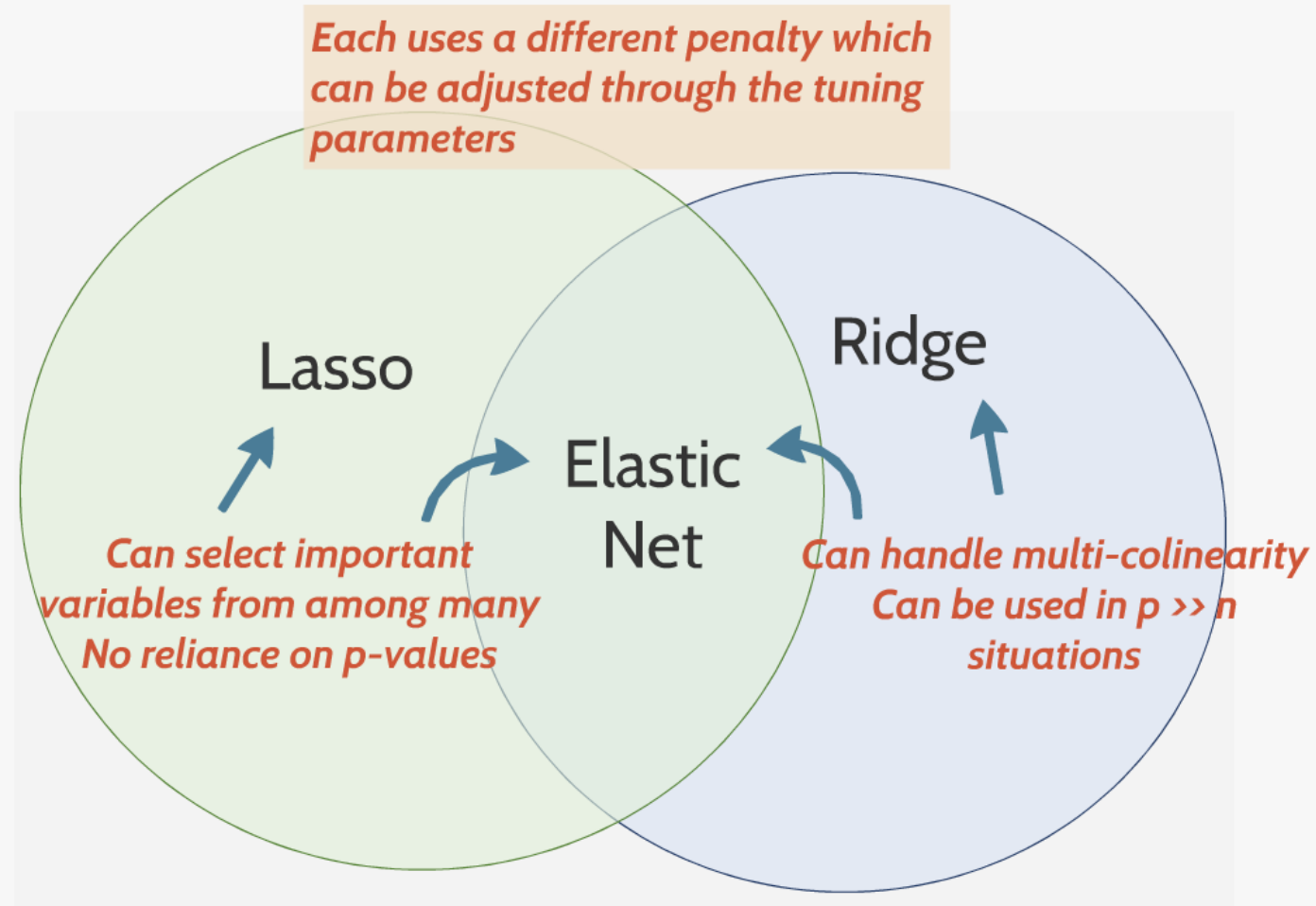
- 1 Are simply penalized versions of regression
- 2 Are interpretable (no "black box")
- 3 Can handle situations that are otherwise impossible to analyze
- 4 Often have higher prediction accuracy than other methods (more generalizable)





# Regularized Regressions

- 1 Are simply penalized versions of regression
- 2 Are interpretable (no "black box")
- 3 Can handle situations that are otherwise impossible to analyze
- 4 Often have higher prediction accuracy than other methods (more generalizable)

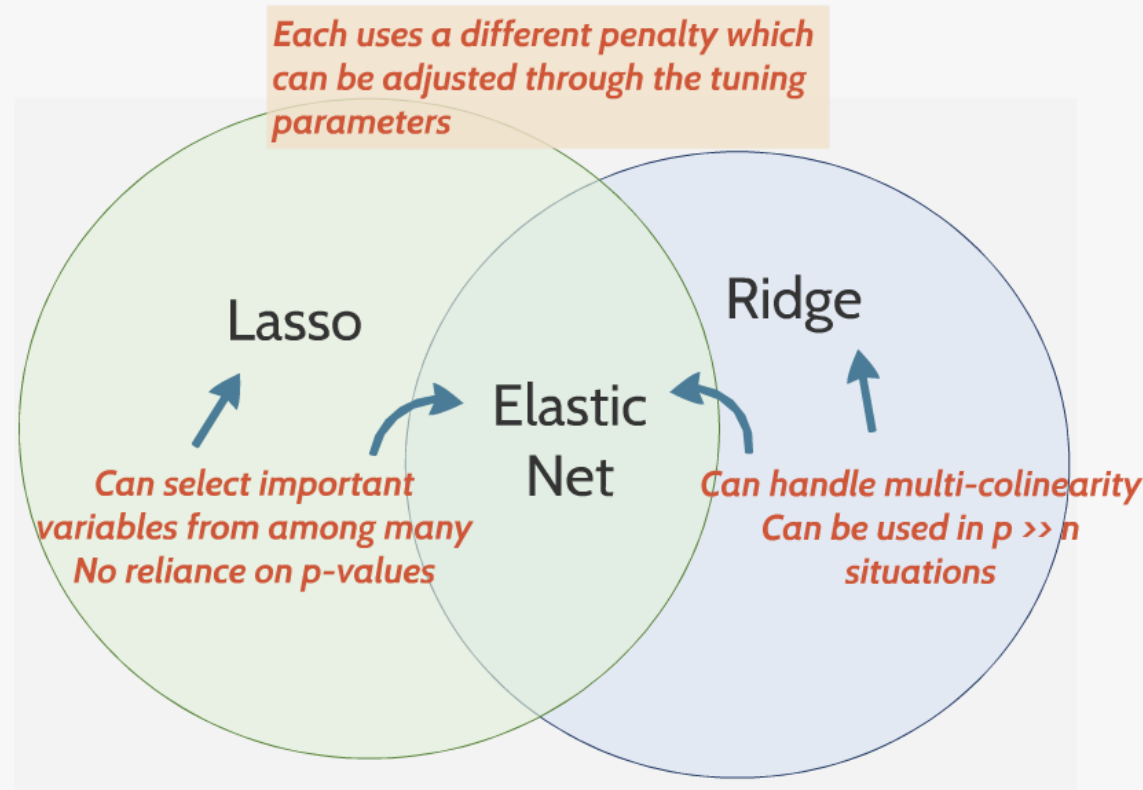


**R** `library(glmnet)`  
`library(elasticnet)`

**SAS** `proc GLMSELE`

# Regularized Regressions

- 1 Are simply penalized versions of regression
- 2 Are interpretable (no "black box")
- 3 Can handle situations that are otherwise impossible to analyze
- 4 Often have higher prediction accuracy than other methods (more generalizable)



**R** `library(glmnet)`  
`library(elasticnet)`

**SAS** `proc GLMSELECT`

# Simple Steps to Use Elastic Net, Lasso or Ridge

## Step 1

Understand your data  
Select type of model  
(linear, logistic, etc.)



## Step 2

Specify the model (just like  
specifying regression)



## Step 3

Dummy code  
categorical variables



## Step 4

Number of "folds" in  
cross-validation



## Step 5

Fit cross-validated model

## Step 6

Select model

Let's  
example  
using ci  
adolesce

# Simple Steps to Use Elastic Net, Lasso or Ridge

## Step 1

Understand your data  
Select type of model  
(linear, logistic, etc.)

## Step 2

Specify the model (just like  
specifying regression)

## Step 3

Dummy code  
categorical variables

## Step 4

Number of "folds" in  
cross-validation

## Step 5

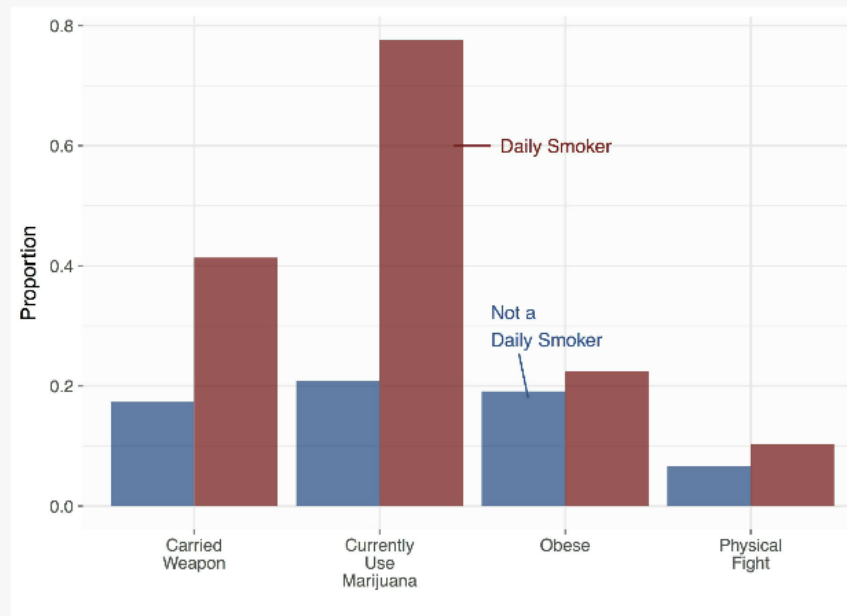
Fit cross-validated model  
to various tuning  
parameter values

## Step 6

Select model  
Fit un-penalized model  
with selected variables  
(optional)

Let's go through an  
example about the risk of  
using cigarettes among  
adolescents with asthma

## Step 1:



Logistic Regression  
(the outcome is binary)

## Step 2:

$$\text{Logit}(\text{Smoker}_i) = \beta_0 + \beta_1 \text{Grade}_i + \dots + \varepsilon_i$$

## Step 3:

Dummy coded variables

	Total Sample with Asthma n = 1856	Marijuana Users n = 803
Used Marijuana	43.2%	-
Currently Use Marijuana	22.6%	52.2%
Used Synthetic Marijuana	10.8%	24.9%
Female	52.9%	53.3%
School grade		
9th	26.4%	23.0%
10th	23.3%	16.7%
11th	25.2%	29.4%
12th	25.1%	30.9%
Rode with drinking driver	21.6%	32.0%
Carried a weapon	18.1%	24.8%
Weapon at school	4.9%	8.0%
Unsafe at school	5.3%	6.2%
Were bullied at school	23.3%	26.3%
Were electronically bullied	19.0%	23.3%
Made plan to commit suicide	19.9%	27.5%
Smoked cigarette before age 13	7.5%	14.8%
Used electronic vapor products	48.0%	80.0%
Drank alcohol before age 13 years	18.6%	29.4%
Drank five or more drinks of alcohol in a row	21.2%	41.2%
Ever used cocaine	5.1%	11.5%
Ever used inhalants	7.8%	12.6%
Ever used heroin	0.9%	2.0%
Ever used methamphetamines	2.3%	5.2%
Ever used ecstasy	5.6%	12.8%
Ever took prescription drugs (no prescription)	20.7%	38.2%
Ever injected any illegal drug	0.9%	1.9%
Offered/sold/given an illegal drug at school	24.8%	35.5%
Ever had sexual intercourse	45.8%	73.8%
Made mostly A's or B's in school	70.3%	61.6%

## Step 4:

## Step 3: Dummy coded variables

Ever had sexual intercourse	45.8%	73.8%
Made mostly A's or B's in school	70.3%	61.6%

## Step 4:

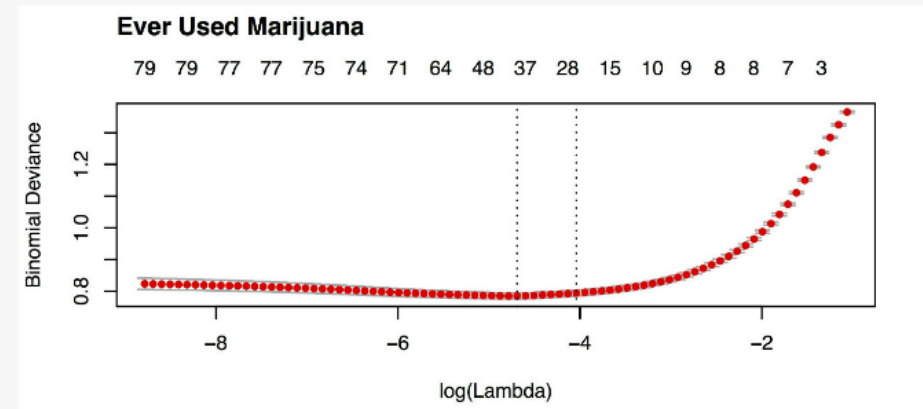
We selected 10 folds for cross-validation

## Step 5:

Fit the cross-validated model

Check:

- the cross-validated error
- the selected variables between the min and the 1-SE



## Step 6:

We selected model at the 1-SE  
(usually more parsimonious than min)

# Notes and Such

By removing a reliance on p-values, reducing researcher bias in variable selection, and providing replicable steps:

Reproducibility



Generalizability



# Notes and Such

Theory and prior literature still act as guides but much of the arbitrary variable selection is gone

## Now in R

Perform regularized regression in R using the cleaned data set I sent out to you

If you didn't get it, visit:

[github.com/CEHSworkshop/RegularizedRegression/](https://github.com/CEHSworkshop/RegularizedRegression/)



CEHS Workshop:  
Reproducible Research via  
Lasso, Ridge, and Elastic Net

Outline

- 1 Why learn about regularized regression? And what is regularized regression?
- 2 How to apply them (without R code)
- 3 Demonstration in R

Reproducibility is low in many fields

3 Aspects of Reproducible Research

- Methods:** can replicate the methods with same data and same analysis
- Results:** can replicate the methods with independent data and their own analysis
- Inferences:** can replicate the methods with independent data and their own interpretation

Goodman et al. (2015)

What is Causing These Problems?

- 1. Relying on over a reliance on p-values
- 2. Reducing variability based on the p-value
- 3. Overfitting, overfitting, multi-collinearity, among others (Gelman, 2014; Greenland et al., 2017)
- 4. These problems are even worse in high-dimensional data ("big data")

So what can we do?



Cross Validation

We will discuss Cross Validation (the model training and to avoid overfitting)

An overfit model

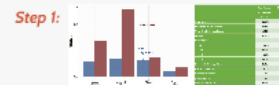
It usually has high accuracy on training data but is not highly generalizable. A plot can cross-validation and assess the predictive accuracy.



Simple Steps to Use Elastic Net, Lasso or Ridge

- Step 1: Load the data into R
- Step 2: Split the data into training and testing sets
- Step 3: Fit the model using the training data
- Step 4: Evaluate the model using the testing data
- Step 5: Cross-validate the model to assess its predictive accuracy
- Step 6: Select the best model based on the cross-validation results

Let's go through an example about the risk of using cigarettes among adolescents with asthma



Logistic Regression (the outcome is binary)

Step 2:  $\text{logit}(\text{Outcome}) = \beta_0 + \beta_1 \text{Cigarette} + \epsilon$

Step 3: Dummy coded variables

Step 4:

We selected 10 folds for cross-validation

Step 5:

Fit the cross-validated model (Check: the cross-validated error, the selected variables between the min and the 1-SE)

Step 6:

We selected model at the 1-SE (usually more parsimonious than min)

Notes and Such

By removing a reliance on p-values, reducing researcher bias in variable selection, and providing replicable steps.

Reproducibility ↑ Generalizability ↑

Notes and Such

Theory and prior literature still act as guides but much of the arbitrary variable selection is gone

Perform regularized regression in R using the cleaned data set I sent out to you

Now in R

If you didn't get it, visit:

[github.com/CEHSWorkshop/RegularizedRegression/](https://github.com/CEHSWorkshop/RegularizedRegression/)

Tyson S. Barrett

@tysonstanley

tyson.barrett@usu.edu

tysonstanley.github.io