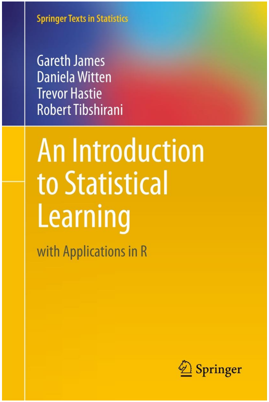
Shrinkage Methods: Ridge and Lasso

Jonathan Hersh[1]

Chapman University, Argyros School of Business
hersh@chapman.edu

February 27, 2019

Source material



- Introduction to Statistical Learning, Chapter 6

Plan

Shrinkage Methods

- Consider the case where we have many more variables than predictors

Shrinkage Methods

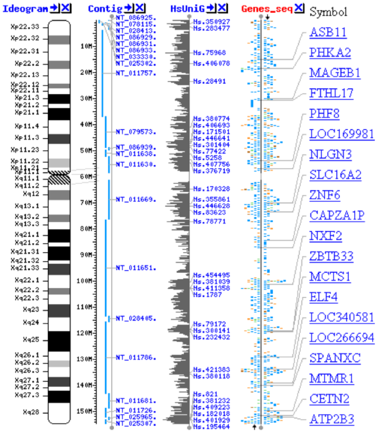
- Consider the case where we have many more variables than predictors
- ***Shrinkage* methods fit a model with all p predictors, but estimate coefficients are "shrunken" towards zero**

Shrinkage Methods

- Consider the case where we have many more variables than predictors
- ***Shrinkage* methods fit a model with all p predictors, but estimate coefficients are "shrunken" towards zero**
- In extreme case, $N < p \Rightarrow \beta = (X^{\intercal}X)^{-1}(X^{\intercal}Y)$ not full rank $\Rightarrow$ **cannot invert**

Shrinkage Methods

- Consider the case where we have many more variables than predictors
- ***Shrinkage*** **methods fit a model with all p predictors, but estimate coefficients are "shrunken" towards zero**
- In extreme case, $N < p \Rightarrow \beta = (X^\intercal X)^{-1}(X^\intercal Y)$ not full rank $\Rightarrow$ **cannot invert**
- Example: bioinformatics. Predict cancer cells (Y), by gene type (X)



Recall: bias-variance tradeoff

Prediction Error: $\mathbb{E}\left[\left(\underbrace{y_i}_{\text{data}} - \underbrace{\hat{f}(x_i)}_{\text{model}}\right)^2\right] = Var(y) + Var\left(\hat{f}\right) + \left(f - \mathbb{E}\left[\hat{f}\right.\right.$

Prediction Error

$$\mathbb{E}\left[\left(y_i - \hat{f}\right)\right] = \underbrace{\sigma_\varepsilon^2}_{\text{irreducible error}} + \underbrace{Var\left(\hat{f}\right)}_{\text{variance}} + \underbrace{\left(\left[f - \mathbb{E}\left[\hat{f}\right]\right]\right)^2}_{\text{bias}}$$

- Bias is minimized when $f = \mathbb{E}\left[\hat{f}\right]$
- **But total error (variance + bias) may be minimized by some other** $\hat{f}$

Recall: bias-variance tradeoff

Prediction Error: $$\mathbb{E}\left[\left(\underbrace{y_i}_{\text{data}} - \underbrace{\hat{f}\left(x_i\right)}_{\text{model}}\right)^2\right] = Var\left(y\right) + Var\left(\hat{f}\right) + \left(f - \mathbb{E}\left[\hat{f}\right.\right.$$

Prediction Error

$$\mathbb{E}\left[\left(y_i - \hat{f}\right)\right] = \underbrace{\sigma_\varepsilon^2}_{\text{irreducible error}} + \underbrace{Var\left(\hat{f}\right)}_{\text{variance}} + \underbrace{\left(\left[f - \mathbb{E}\left[\hat{f}\right]\right]\right)^2}_{\text{bias}}$$

- Bias is minimized when $f = \mathbb{E}\left[\hat{f}\right]$
- **But total error (variance + bias) may be minimized by some other $\hat{f}$**
- $\hat{f}(x)$ with smaller variance $\Rightarrow$ fewer variables, smaller magnitude coefficients

Plan

1. Intro and Background
   - Introduction
2. Ridge Regression
   - Example: Ridge & Multicollinearity
3. Lasso
4. Applications & Extensions
5. Conclusion

Ridge Regression

$$\textbf{OLS Sum of Squared Resids} \;=\; \underbrace{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Squared Sum of Residuals}}$$

Ridge Regression

$$\textbf{OLS Sum of Squared Resids} = \underbrace{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Squared Sum of Residuals}}$$

- To reduce prediction error: minimize $Var(\hat{f}(x)) = Var(X\beta)$
- One way: decrease β in absolute value

Ridge Regression

Definitions

Ridge estimator β^R is defined as

$$\beta_{ridge} = \underset{\beta}{\textbf{argmin}} \underbrace{\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda \cdot \sum_{j=1}^{p}\beta_j^2}_{\text{Shrinkage Factor}}$$

where $\lambda \geq 0$ is a tuning parameter (or hyper-parameter)

Ridge Continued

- The ridge estimator also wants to find coefficients that fits the data well, and reduces RSS
- **The second term, $\lambda \cdot \sum_{j=1}^{p} \beta_j^2$ ensures that it does so in a balanced way, so that bias isn't minimized at the expense of variance**

Ridge Continued

- The ridge estimator also wants to find coefficients that fits the data well, and reduces RSS
- **The second term, $\lambda \cdot \sum_{j=1}^{p} \beta_j^2$ ensures that it does so in a balanced way, so that bias isn't minimized at the expense of variance**

- The tuning parameter λ controls the relative impact of bias and variance
- Larger $\lambda \Rightarrow$ more bias
  - Note as $\lambda \to \infty \Rightarrow \beta^R \to \mathbf{0}$
  - Note as $\lambda \to 0 \Rightarrow \beta^R \to \beta^{OLS}$

Ridge Continued

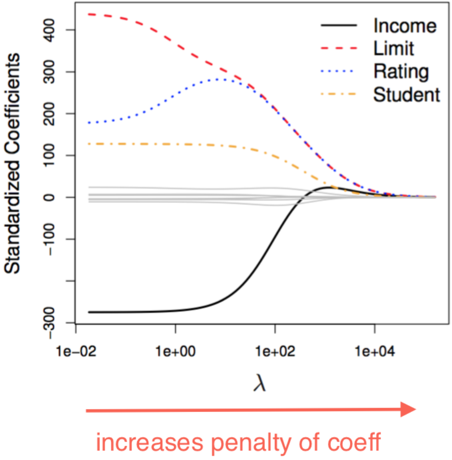
- In matrix form:
$$\beta^R = (X^\intercal X + \lambda I_K)^{-1}(X^\intercal Y)$$
- Note is positive definite even when $K > N$

Ridge Continued

- In matrix form:
$$\beta^R = (X^\intercal X + \lambda I_K)^{-1}(X^\intercal Y)$$
- Note is positive definite even when $K > N$
- Coefficients are shrunk smoothly towards zero.
- Bayesian interpretation: Laplace priors $\beta^R \sim \mathcal{N}\left(0, \tau^2 I_K\right)$ β^R is the posterior mean/mode/median

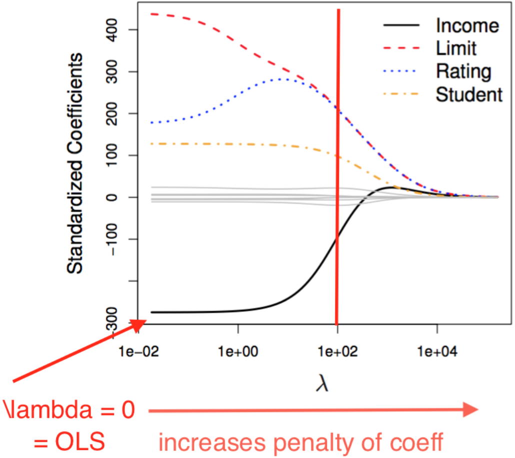
How to Choose λ?

- In practice we estimate a range of λ values and choose



How to Choose λ?

- In practice we estimate a range of λ values and choose



Root Mean Squared Error Across λ Values

6.4 Penalized Models 125

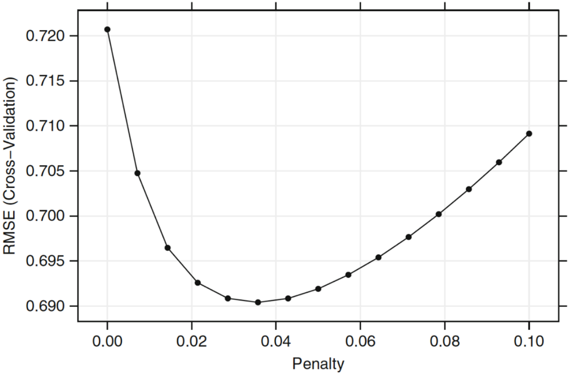


Fig. 6.16: The cross-validation profiles for a ridge regression model

Root Mean Squared Error Across λ Values

6.4 Penalized Models 125

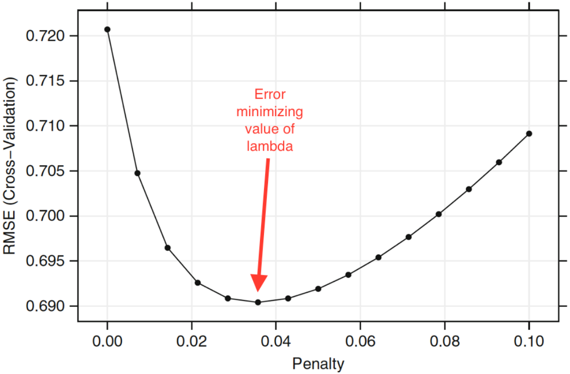


Fig. 6.16: The cross-validation profiles for a ridge regression model

Ridge Notes

- Small amount of shrinkage usually improves prediction performance
    - Paricularly when the number of variables is large and variance is likely high

Ridge Notes

- Small amount of shrinkage usually improves prediction performance
  - Paricularly when the number of variables is large and variance is likely high
- Variables are never completely shrunk to zero – but very small in absolute value
  - Works poorly for variable selection
  - Useful for when you have reason to suspect underlying DGP is non-sparse

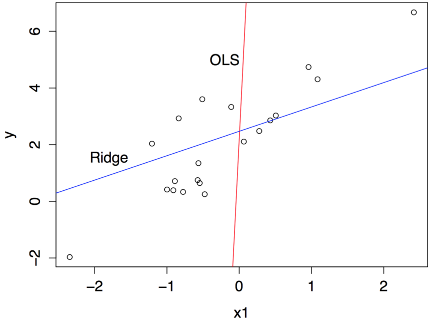
How can ridge help with multicollinearity?

- Quick example in R

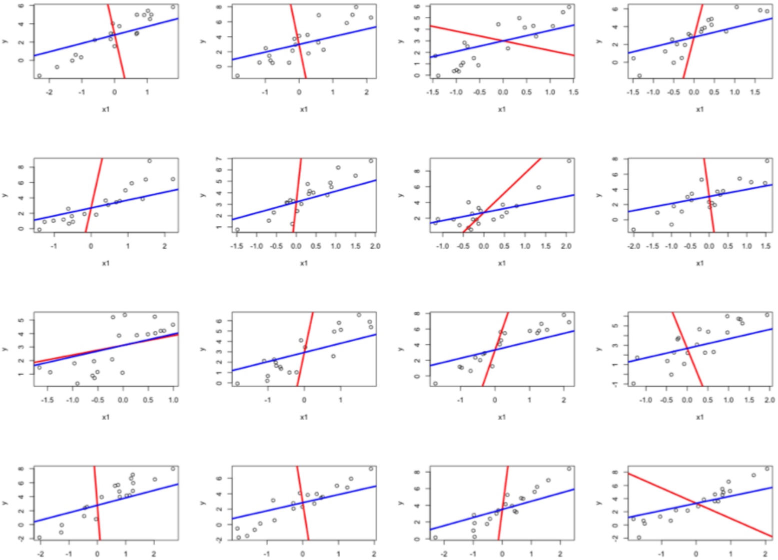
```
#Generate x1 and x2 that are highly colinear
x1 <- rnorm(20)
x2 <- rnorm(20,mean=x1,sd=.01)
y <- rnorm(20,mean=3+x1+x2)
# OLS Reg
OLSmod <- lm(y~x1+x2)
#Ridge Reg
RIDGEmod <- lm.ridge(y~x1+x2,lambda=1)
```

Multicollinearity Example

```
> OLSmod

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
      2.169       50.386      -48.784

> lm.ridge(y~x1+x2,lambda=1)
                 x1        x2
2.4710161 0.8605031 0.8062424
> vif(OLSmod)
     x1      x2
17687.3 17687.3
>
```



Red line = OLS, Blue = Ridge



Plan

LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996)

- Lasso Regression looks very similar to Ridge
- Lasso estimator β^{Lasso} will minimize the modified likelihood

$$\underbrace{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{Shrinkage Factor}}$$

where $\lambda \geq 0$ is a tuning parameter

LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996)

- Lasso Regression looks very similar to Ridge
- Lasso estimator β^{Lasso} will minimize the modified likelihood

$$\underbrace{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{Shrinkage Factor}}$$

  where $\lambda \geq 0$ is a tuning parameter

- Magnitude of variables is "penalized" with absolute value loss
- Because of absolute value, more efficient to "spend" only on useful variables
  - Acts as variable selection. Though will in addition get shrinkage of estimates towards zero

LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996)

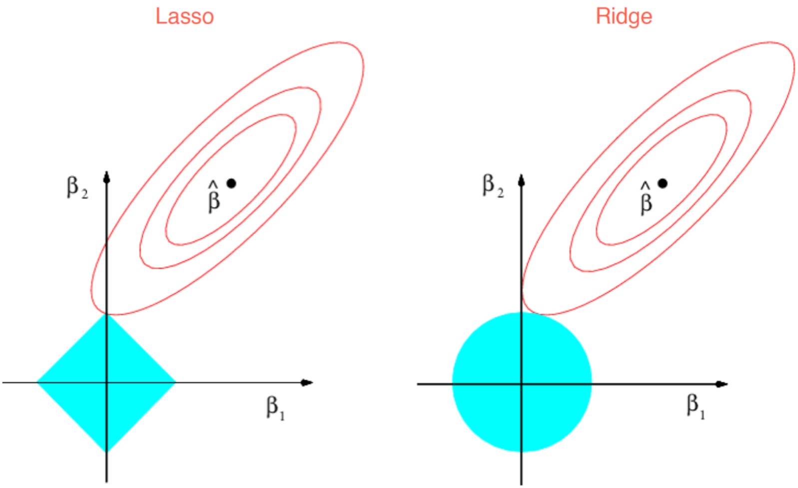
- Lasso Regression looks very similar to Ridge
- Lasso estimator β^{Lasso} will minimize the modified likelihood

$$\underbrace{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{Shrinkage Factor}}$$

where $\lambda \geq 0$ is a tuning parameter

- Magnitude of variables is “penalized” with absolute value loss
- Because of absolute value, more efficient to “spend” only on useful variables
  - Acts as variable selection. Though will in addition get shrinkage of estimates towards zero
- Again as $\lambda \to 0 \Rightarrow \beta^{Lasso} \to \beta^{OLS}$, as $\lambda \to \infty \Rightarrow \beta^{Lasso} \to \mathbf{0}$

Visualization Lasso, Ridge, and OLS Coefficients



Comparing Ridge and Lasso

- No analytic solution to Lasso, unlike ridge, but computationally very feasible with large datasets given the convex optimization problem.

Comparing Ridge and Lasso

- No analytic solution to Lasso, unlike ridge, but computationally very feasible with large datasets given the convex optimization problem.
- With large datasets, inverting $(X^\intercal X + \lambda I_K)$ is expensive

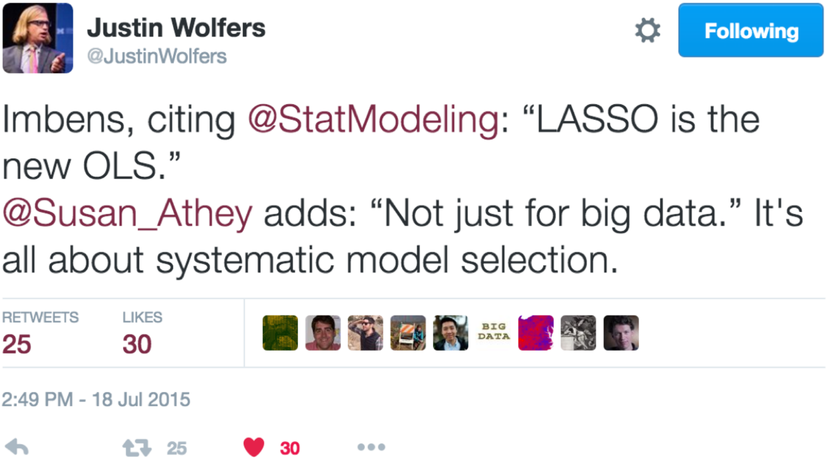
Comparing Ridge and Lasso

- No analytic solution to Lasso, unlike ridge, but computationally very feasible with large datasets given the convex optimization problem.
- With large datasets, inverting $(X^\intercal X + \lambda I_K)$ is expensive
- Lasso has favorable properties if the true model is sparse.

Comparing Ridge and Lasso

- No analytic solution to Lasso, unlike ridge, but computationally very feasible with large datasets given the convex optimization problem.
- With large datasets, inverting $(X^{\intercal}X + \lambda I_K)$ is expensive
- Lasso has favorable properties if the true model is sparse.
- If the distribution of coefficients is very thick tailed (few variables matter a lot) Lasso will do much better than ridge. If there are many moderate sized effects, ridge may do better

Social Scientists are Coming Around to Lasso



Bayesian Interpretation of Lasso

- Lasso coefficients are the mode of the posterior distribution, given a normal linear model with Laplace priors $p(\beta) \propto exp\left(\lambda \sum_{k=1} |\beta_k|\right)$
- Slightly odd that we're picking the mode rather than the mean from the posterior distribution
- Related: Spike and Slab prior

Related Estimators

- **Least Angle RegreSsion** (LARS) - The "S" here suggests stepwise.
  - A stagewise iterative procedure that iteratively selects regressors to be included in the regression function

Related Estimators

- **Least Angle RegreSsion** (LARS) - The "S" here suggests stepwise.
  - A stagewise iterative procedure that iteratively selects regressors to be included in the regression function
- **Dantzig Selector** (Candes & Tao, 2007)
  - Lasso type regularization, but minimizing the maximum correlation between residuals and covariates
  - Doesn't particularly work well.

Related Estimators

- **Least Angle RegreSsion** (LARS) - The "S" here suggests stepwise.
  - A stagewise iterative procedure that iteratively selects regressors to be included in the regression function
- **Dantzig Selector** (Candes & Tao, 2007)
  - Lasso type regularization, but minimizing the maximum correlation between residuals and covariates
  - Doesn't particularly work well.
- **Elastic-Net**

$$\min_{\beta}\left\{\sum_{i=1}^{N}(Y_i - X_i\beta)^2 + \lambda\left(\underbrace{\alpha\cdot\|\beta\|_1}_{\textbf{Lasso penalty}} + \underbrace{(1-\alpha)\cdot\|\beta\|_2^2}_{\textbf{Ridge penalty}}\right)\right\}$$

Related Estimators

- **Least Angle RegreSsion** (LARS) - The "S" here suggests stepwise.
  - A stagewise iterative procedure that iteratively selects regressors to be included in the regression function
- **Dantzig Selector** (Candes & Tao, 2007)
  - Lasso type regularization, but minimizing the maximum correlation between residuals and covariates
  - Doesn't particularly work well.
- **Elastic-Net**

$$\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \left(\underbrace{\alpha \cdot \|\beta\|_1}_{\textbf{Lasso penalty}} + \underbrace{(1-\alpha) \cdot \|\beta\|_2^2}_{\textbf{Ridge penalty}} \right) \right\}$$

  - α controls the weighting between ridge and Lasso, obtained through cross-validation

Oracle Property (Fan and Zhuo, 2001)

- If the true model is sparse -- so that there are few (say k^*) true non-zero coefficients – and many true zero coefficients $(K - k^*)$ **an estimator has the oracle property if inference is as if you knew the true model**, i.e. knew a priori exactly which coefficients were truly zero.

Oracle Property (Fan and Zhuo, 2001)

- If the true model is sparse -- so that there are few (say k^*) true non-zero coefficients – and many true zero coefficients $(K - k^*)$ **an estimator has the oracle property if inference is as if you knew the true model**, i.e. knew a priori exactly which coefficients were truly zero.
- Limitation: sample size needs to be large relative to k

Oracle Property (Fan and Zhuo, 2001)

- If the true model is sparse -- so that there are few (say k^*) true non-zero coefficients – and many true zero coefficients $(K - k^*)$ **an estimator has the oracle property if inference is as if you knew the true model**, i.e. knew a priori exactly which coefficients were truly zero.
- Limitation: sample size needs to be large relative to k
- What this means: you can ignore the selection of covariates in the calculation of the standard errors. Can just use regular OLS SEs

Estimating Lasso/Ridge Model in R

- Many packages, but glmnet is maintained by Tibshirani
- `cv.glmnet()` estimates a series of Lasso models for various levels of λ

```
lasso.mod <- cv.glmnet(x = X, y = Y, alpha = 1, nfolds = 10)
```

- `build.x()` and `build.y()` are helper functions for glmnet that build glmnet compatible X and Y matrices respectively.

```
Xvars <- build.x(formula, data = df)
Yvar <- build.y(formula, data = df)
```

Stata Implementation of Lasso: elasticregress

```
Title

    elasticregress -- Elastic net regression

    lassoregress -- LASSO regression

    ridgeregress -- Ridge regression


Syntax

        elasticregress depvar [indepvars] [if] [in] [weight] [, alpha(#) options]

        lassoregress depvar [indepvars] [if] [in] [weight] [, options]

        ridgeregress depvar [indepvars] [if] [in] [weight] [, options]

    options                Description
    ---------------------------------------------------------------------------------------------------------
    Main
      alpha                weight placed on the LASSO (L1) norm, one minus weight placed on the ridge (L2) norm - by default found by cross-validat
      lambda               penalty placed on larger coefficients - by default found by cross-validation;
      numfolds             number of folds used when cross-validating lambda or alpha - default is 10.
    ---------------------------------------------------------------------------------------------------------
    Options which only matter when alpha is found through cross-validation
      numalpha             number of alpha tested when alpha is found by cross-validation.
```

Automatic α Selection: Package `glmnetUtils`

Introduction to glmnetUtils

The glmnetUtils package provides a collection of tools to streamline the process of fitting elastic net models with glmnet. I wrote the package after a couple of projects where I found myself writing the same boilerplate code to convert a data frame into a predictor matrix and a response vector. In addition to providing a formula interface, it also features a function `cva.glmnet` to do crossvalidation for both α and λ, as well as some utility functions.

The formula interface

The interface that glmnetUtils provides is very much the same as for most modelling functions in R. To fit a model, you provide a formula and data frame. You can also provide any arguments that glmnet will accept. Here are some simple examples for different types of data:

```
# least squares regression
(mtcarsMod <- glmnet(mpg ~ cyl + disp + hp, data=mtcars))
```

```
## Call:
## glmnet.formula(formula = mpg ~ cyl + disp + hp, data = mtcars)
##
## Model fitting options:
##     Sparse model matrix: FALSE
##     Use model.frame: FALSE
##     Alpha: 1
##     Lambda summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03326 0.11690 0.41003 1.02839 1.44125 5.05505
```

Automatic α Selection: Function `cva.glmnet`

Crossvalidation for α

One piece missing from the standard glmnet package is a way of choosing α, the elastic net mixing parameter, similar to how `cv.glmnet` chooses λ, the shrinkage parameter. To fix this, glmnetUtils provides the `cva.glmnet` function, which uses crossvalidation to examine the impact on the model of changing α and λ. The interface is the same as for the other functions:

```
# Leukemia dataset from Trevor Hastie's website:
# http://web.stanford.edu/~hastie/glmnet/glmnetData/Leukemia.RData
leuk <- do.call(data.frame, Leukemia)

leukMod <- cva.glmnet(y ~ ., data=leuk, family="binomial")
leukMod
```

```
## Call:
## cva.glmnet.formula(formula = y ~ ., data = leuk, family = "binomial")
##
## Model fitting options:
##     Sparse model matrix: FALSE
##     Use model.frame: FALSE
##     Alpha values: 0 0.001 0.008 0.027 0.064 0.125 0.216 0.343 0.512 0.729 1
##     Number of crossvalidation folds for lambda: 10
```

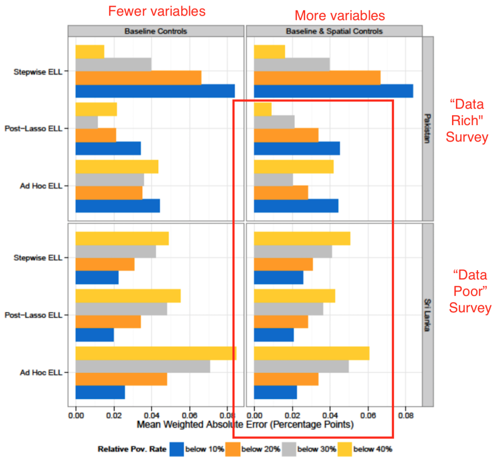
Caveats of Lasso for Poverty Modeling

1. Because $|\beta_{Lasso}| < |\beta_{OLS}| \Rightarrow |\hat{y}_{lasso}| < |\hat{y}_{OLS}|$
    1.1 Don't predict from Lasso, use Lasso for model selection, then do ELL
2. Lasso may drop variables with hierarchical relationships, e.g. *age* and age^2
    2.1 Use Sparse Group Lasso (SGL package) which specifies hierarchical structure (e.g. if select age^2, must select *age*).

Plan

Afzal, Hersh and Newhouse (2015)

- Lasso for model selection for poverty mapping in Sri Lanka and Pakistan [source]



Afzal, Hersh and Newhouse (2015) Continued

- **Lasso works well for model selection when # of candidate variables is large (100+)**
- No worse than stepwise when set of variables is small

Post-Model Selection Estimator: Belloni & Chernozhukov, 2013

- Belloni & Chernozhukov (2013) define the two step Post-Lasso estimator as

1. **Estimate a Lasso model** using full candidate set of variables ($X_{\text{candidate}}$)

Post-Model Selection Estimator: Belloni & Chernozhukov, 2013

- Belloni & Chernozhukov (2013) define the two step Post-Lasso estimator as

1. **Estimate a Lasso model** using full candidate set of variables ($X_{\text{candidate}}$)
2. **Use selected variables** (X_{selected}) to estimate final model using modeling strategy of choice

Post-Model Selection Estimator: Belloni & Chernozhukov, 2013

- Belloni & Chernozhukov (2013) define the two step Post-Lasso estimator as

1. **Estimate a Lasso model** using full candidate set of variables ($X_{\text{candidate}}$)
2. **Use selected variables** (X_{selected}) to estimate final model using modeling strategy of choice

- Because of oracle property of Lasso (Fan and Li, 2001) inference in the second stage using the reduced set of variables is consistent with inference with single stage estimation strategy using only the selected variables present in the true data-generating process

Heteroscedastic Robust Lasso

- See: Belloni, Chen, Chernozhukov, Hansen (Econometrica, 2012)

Double Selection Procedure for Estimating Treatment Effects

- Consider the problem of estimating the effect of treatment d_i on some outcome y_i in the presence of possibly confounding controls x_i

Double Selection Procedure for Estimating Treatment Effects

- Consider the problem of estimating the effect of treatment d_i on some outcome y_i in the presence of possibly confounding controls x_i

Double Selection Method:

1. Select via Lasso controls x_{ij} that predict y_i

Double Selection Procedure for Estimating Treatment Effects

- Consider the problem of estimating the effect of treatment d_i on some outcome y_i in the presence of possibly confounding controls x_i

Double Selection Method:

1. Select via Lasso controls x_{ij} that predict y_i
2. Select via Lasso controls x_{ij} that predict d_i

Double Selection Procedure for Estimating Treatment Effects

- Consider the problem of estimating the effect of treatment d_i on some outcome y_i in the presence of possibly confounding controls x_i

Double Selection Method:

1. Select via Lasso controls x_{ij} that predict y_i
2. Select via Lasso controls x_{ij} that predict d_i
3. Run OLS of y_i on d_i on the **union** of controls selected in steps 1 and 2

Double Selection Procedure for Estimating Treatment Effects

- Consider the problem of estimating the effect of treatment d_i on some outcome y_i in the presence of possibly confounding controls x_i

Double Selection Method:

1. Select via Lasso controls x_{ij} that predict y_i
2. Select via Lasso controls x_{ij} that predict d_i
3. Run OLS of y_i on d_i on the **union** of controls selected in steps 1 and 2

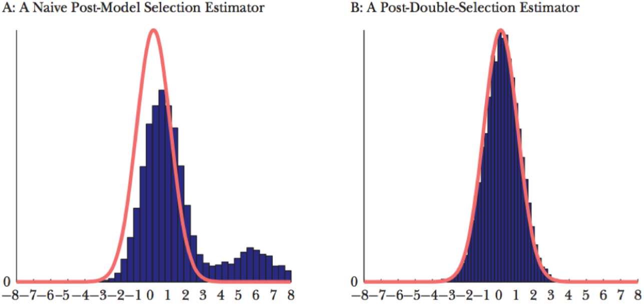
- Authors' claim: additional selection step controls the omitted variable bias

Double Selection vs Naive Approach

Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)



Source: Belloni, Chernozhukov, and Hansen (forthcoming).

Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

Replicating Donohue and Levitt

Table 1
Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Notes: This table reports results from estimating the effect of abortion on violent crime, property crime, and murder. The row labeled "First-difference" gives baseline first-difference estimates using the controls from Donohue and Levitt (2001). The row labeled "All controls" includes a broad set of controls meant to allow flexible trends that vary with state-level characteristics. The row labeled "Double selection" reports results based on the double selection method outlined in this paper and selecting among the variables used in the "All controls" results.

Replicating AJR (2001)

Table 2

Effect of Institutions on Output

	Latitude	*All controls*	*Double selection*
First stage	−0.5372	−0.2182	−0.5429
	(0.1545)	(0.2011)	(0.1719)
Second stage	0.9692	0.9891	0.7710
	(0.2128)	(0.8005)	(0.1971)

Notes: In an exercise that follows the work of Acemoglu, Johnson, and Robinson (2001), this table reports results from estimating the effect of institutions, using settler mortality as an instrument. The row "First Stage" gives the first-stage estimate of the coefficient on settler mortality obtained by regressing "$ProtectionfromExpropriation_i$" on "$SettlerMortality_i$" and the set of control variables indicated in the column heading. The row "Second stage" gives the estimate of the structural effect of institutions on log(GDP per capita) using "$SettlerMortality_i$" as the instrument and controlling for variables as indicated in the column heading (see text for details). Each column reports the results for different sets of control variables. The column "Latitude" controls linearly for distance from the equator. The column "All controls" includes 16 controls defined in the main text and in footnote 9, and the column "Double selection" uses the union of the set of controls selected by LASSO for predicting GDP per capita, for predicting institutions, and for predicting settler mortality. Standard errors are in parentheses.

Plan

Conclusion

- Use Lasso regression if you have reason to believe the true model is sparse
- Use Ridge otherwise

Conclusion

- Use Lasso regression if you have reason to believe the true model is sparse
- Use Ridge otherwise

- Lasso's sparsity offers disciplined method of variable selection
- But be careful predicting from Lasso. Do Lasso + ELL (Elbers, Lanjouw, Lanjouw)

Conclusion

- Use Lasso regression if you have reason to believe the true model is sparse
- Use Ridge otherwise

- Lasso's sparsity offers disciplined method of variable selection
- But be careful predicting from Lasso. Do Lasso + ELL (Elbers, Lanjouw, Lanjouw)
- Select λ using k-fold cross-valdiation
- Use test sample to approximate out of sample error