Shrinkage Estimators

William Brasic

The University of Arizona

Shrinkage Estimators

Definition 1: Shrinkage Estimators

Shrinkage estimators refer to techniques in linear regression models where parameter estimates are systematically reduced towards zero.

- Two prominent examples are:
  - Ridge Regression
  - Lasso Regression

Shrinkage Estimators

Question 1

Why would we want to shrink our parameter estimates towards zero?

Shrinkage Estimators

Question 1

Why would we want to shrink our parameter estimates towards zero?

Answer to Question 1

- The lower variance of shrinkage estimators relative to OLS offsets the increase in bias.
- Lower model complexity implies better generalization performance.
  - The model's parameters have less impact, reducing the likelihood of overfitting to the training data.

OLS Objective

Definition 2: OLS Objective

Recall the OLS objective is given by

$$\underset{\widehat{\boldsymbol{\beta}}}{\arg\min}\, SSR = \underset{\widehat{\boldsymbol{\beta}}}{\arg\min}\, \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\right).$$

Shrinkage Estimators

Question 2

How could we alter the OLS objective to force the parameter estimates towards zero?

Shrinkage Estimators

Question 2

How could we alter the OLS objective to force the parameter estimates towards zero?

Answer to Question 2

Simply add a penalization factor!

L2-Norm

Definition 3: L2-Norm

The L2-Norm of a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is given by

$$\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{k} \beta_j^2}.$$

and its squared L2-Norm is given by

$$\|\boldsymbol{\beta}\|_2^2 = \boldsymbol{\beta}'\boldsymbol{\beta} = \sum_{j=1}^{k} \beta_j^2.$$

- Measures a vector's magnitude (how "large" it is).
- Similar to the Euclidean distance formula.

Ridge Regression

Definition 4: Ridge Regression

Ridge regression is a regression technique that adds the squared L2-Norm to the OLS objective function.

- The objective function is forced to choose the parameters minimizing the SSR while also considering how large to make the parameter estimates.
- Typically don't penalize the intercept term.

Ridge Regression Objective

Definition 5: Ridge Regression Objective

The ridge regression objective is given by

$$\underset{\widehat{\boldsymbol{\beta}}_R}{\arg\min} \left[SSR + \lambda \sum_{j=1}^{k} \widehat{\beta}_{jR}^2 \right]$$

Ridge Regression Objective

Definition 5: Ridge Regression Objective

The ridge regression objective is given by

$$\begin{aligned}
&\arg\min_{\widehat{\boldsymbol{\beta}}_R} \left[SSR + \lambda \sum_{j=1}^{k} \widehat{\beta}_{jR}^2 \right] \\
&= \arg\min_{\widehat{\boldsymbol{\beta}}_R} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R\right)' \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R\right) + \lambda \left\| \widehat{\boldsymbol{\beta}}_R \right\|_2^2 \right].
\end{aligned}$$

Ridge Regression Objective

Definition 5: Ridge Regression Objective

The ridge regression objective can also be written as

$$\arg\min_{\widehat{\boldsymbol{\beta}}_R} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R\right)'\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R\right)\right] \quad \text{subject to} \quad \left\|\widehat{\boldsymbol{\beta}}_R\right\|_2^2 \leq s.$$

- This objective is identical to that on the prior slide.
- s is inversely related to λ.

Penalization Parameter λ

Definition 6: Penalization Parameter λ

The penalization parameter $\lambda \geq 0$ determines how much shrinkage we want:

- Higher λ implies higher bias, but likely lower variance and vice versa.
- Higher λ "penalizes" the model for higher parameter estimates.

- Use a method like k-Fold CV to "tune" λ to its optimal value.
- As $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}_R \to \widehat{\boldsymbol{\beta}}_{OLS}$.

Ridge Regression Solution

Theorem 1: Ridge Regression Solution

The ridge regression solution $\widehat{\boldsymbol{\beta}}_R$ is given by

$$\widehat{\boldsymbol{\beta}}_R = \left(X'X + \lambda I_{k+1}\right)^{-1} X'\boldsymbol{y}.$$

- Higher λ implies "higher" $\left(X'X + \lambda I_{k+1}\right)$, resulting in a "lower" $\left(X'X + \lambda I_{k+1}\right)^{-1}$, meaning $\widehat{\boldsymbol{\beta}}_R$ is shrunk.

Ridge Regression Solution

Proof 1: Ridge Regression Solution Part 1

The ridge regression objective is

$$\underset{\widehat{\boldsymbol{\beta}}_R}{\arg\min} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R \right)' \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_R \right) + \lambda \left\| \widehat{\boldsymbol{\beta}}_R \right\|_2^2 \right].$$

Differentiating this objective function with respect to $\widehat{\boldsymbol{\beta}}_R$ and taking the first order condition we get

$$-2X'\boldsymbol{y} + 2X'X\widehat{\boldsymbol{\beta}}_R + 2\lambda\widehat{\boldsymbol{\beta}}_R = 0.$$

Ridge Regression Solution

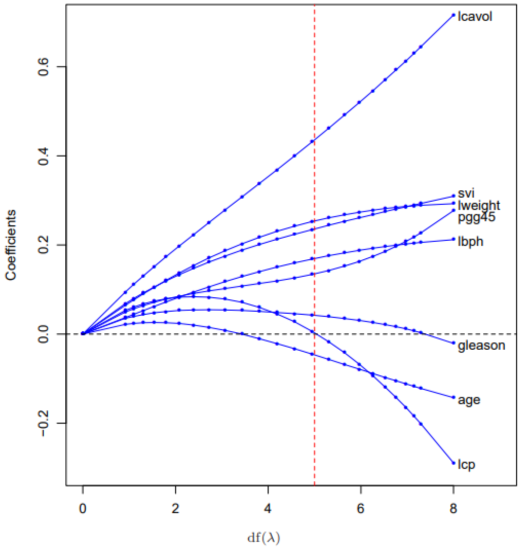
Proof 1: Ridge Regression Solution Part 2

Solving for $\widehat{\boldsymbol{\beta}}_R$ we find

$$X'X\widehat{\boldsymbol{\beta}}_R + \lambda\widehat{\boldsymbol{\beta}}_R = X'\boldsymbol{y} \iff \left(X'X + \lambda I_{k+1}\right)\widehat{\boldsymbol{\beta}}_R = X'\boldsymbol{y}$$
$$\iff \widehat{\boldsymbol{\beta}}_R = \left(X'X + \lambda I_{k+1}\right)^{-1} X'\boldsymbol{y}.$$

- We need the λI_{k+1} when factoring our $\widehat{\boldsymbol{\beta}}$ so we can add $X'X$ to it; without the I_k the sum inside the inverse would not be defined.

Ridge Regression



Ridge Regression Solution Always Produces a Unique Solution

Property 1: Ridge Regression Solution Always Produces a Unique Solution

The ridge regression estimator given by

$$\widehat{\boldsymbol{\beta}}_R = (X'X + \lambda I_k)^{-1} X'\boldsymbol{y}$$

always exists and is unique because $(X'X + \lambda I_k)$ is always invertible irrespective of the correlation among covariates.

- If we have high correlation between covariates, we likely want to use ridge regression over OLS to ensure:
  - Estimator is defined and is unique.
  - Lower standard errors.

Ridge Regression vs OLS MSE

Theorem 2: Ridge Regression vs OLS MSE

There *always* exists some value of λ for any given dataset such that the MSE of given by a model using $\widehat{\boldsymbol{\beta}}_R$ will be lower than that given by a model using $\widehat{\boldsymbol{\beta}}_{OLS}$.

- Thus, if we can fine tune λ properly, ridge regression will be a better predictor than OLS!

Ridge Regression Bias

Theorem 3: Ridge Regression Bias

Assuming $\mathbb{E}[\boldsymbol{\epsilon} \mid X] = \mathbf{0}$, the bias (in econometrics sense) of the ridge regression estimator is given by

$$\text{bias}\left[\widehat{\boldsymbol{\beta}}_R\right] = -\lambda \left(X'X + \lambda I_k\right)^{-1} X'\boldsymbol{\beta}.$$

- Since $\lambda \geq 0$, $-\lambda \leq 0$ meaning the bias is downwards because this is a shrinkage estimator.
- As $\lambda \to 0$, the $\widehat{\boldsymbol{\beta}}_R$ becomes less biased.
- When $\lambda = 0$, $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_R\right] = \boldsymbol{\beta}$.

Ridge Regression Variance

Theorem 4: Ridge Regression Variance

Given $\boldsymbol{\epsilon} \sim \mathbb{N}\left(\mathbf{0}, \sigma^2 I_n\right)$, the variance of the ridge regression estimator is given by

$$\mathbb{V}\left[\widehat{\boldsymbol{\beta}}_R\right] = \sigma^2 \left(X'X + \lambda I_k\right)^{-1} X'X \left(X'X + \lambda I_k\right)^{-1}.$$

- A large λ means a larger $(X'X + \lambda I_k)$, implying a smaller $(X'X + \lambda I_k)^{-1}$, which lowers $\mathbb{V}\left[\widehat{\boldsymbol{\beta}}_R\right]$.
  - Thus, a large λ shrinks the variance but raises the bias of $\widehat{\boldsymbol{\beta}}_R$ and vice versa.

L1-Norm

Definition 7: L1-Norm

The L1-Norm of a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is given by

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{k} |\beta_j| \,.$$

- Measures a vector's magnitude (how "large" it is).

Lasso Regression

Definition 8: Lasso Regression

Lasso regression is a regression technique that adds the L1-Norm to the OLS objective function.

- The objective function is forced to choose the parameters minimizing the SSR while also considering how large to make the parameter estimates.

Lasso Regression Objective

Definition 9: Lasso Regression Objective

The lasso regression objective is given by

$$= \underset{\widehat{\boldsymbol{\beta}}_L}{\arg\min} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_L \right)' \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_L \right) + \lambda \left\| \widehat{\boldsymbol{\beta}}_L \right\|_1 \right].$$

Lasso Regression Objective

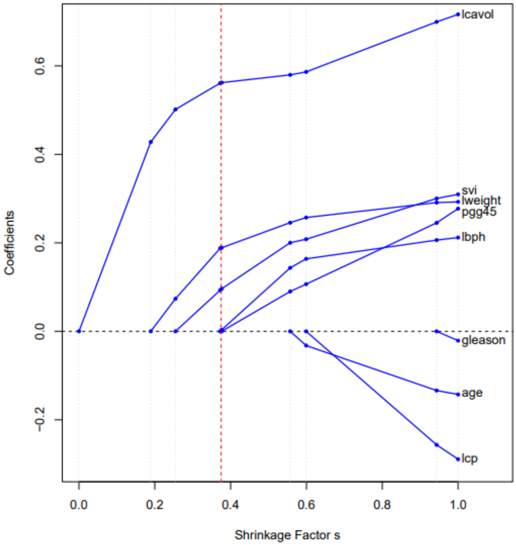
Definition 9: Lasso Regression Objective

The lasso regression objective can also be written as

$$\arg\min_{\widehat{\boldsymbol{\beta}}_L} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_L\right)'\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_L\right)\right] \quad \text{subject to} \quad \left\|\widehat{\boldsymbol{\beta}}_L\right\|_1 \leq s.$$

- This objective is identical to that on the prior slide.
- s is inversely related to λ.

Lasso Regression



Lasso Solution Properties

Property 2: Lasso Solution Properties

1. The lasso regression objective does not have a closed form solution because the L1-Norm is not differentiable at zero.
   - Need more advanced techniques such as subgradients to find the solution

Lasso Solution Properties

Property 2: Lasso Solution Properties

1. The lasso regression objective does not have a closed form solution because the L1-Norm is not differentiable at zero.
   - Need more advanced techniques such as subgradients to find the solution
2. The lasso solution always exists, but may not be unique.

Lasso Solution Properties

Property 2: Lasso Solution Properties

1. The lasso regression objective does not have a closed form solution because the L1-Norm is not differentiable at zero.
   - Need more advanced techniques such as subgradients to find the solution
2. The lasso solution always exists, but may not be unique.
3. Lasso does not just shrink estimates towards zero, but can *force* them to be zero.

Ridge vs. Lasso Regression

Property 3: Ridge vs. Lasso Regression

1. Ridge is more easily interpretable.

Ridge vs. Lasso Regression

Property 3: Ridge vs. Lasso Regression

1. Ridge is more easily interpretable.
2. Ridge is less computationally heavy due to not needing advanced optimization techniques like the lasso.

Ridge vs. Lasso Regression

Property 3: Ridge vs. Lasso Regression

1. Ridge is more easily interpretable.
2. Ridge is less computationally heavy due to not needing advanced optimization techniques like the lasso.
3. Ridge always has a unique solution, unlike OLS and the lasso.

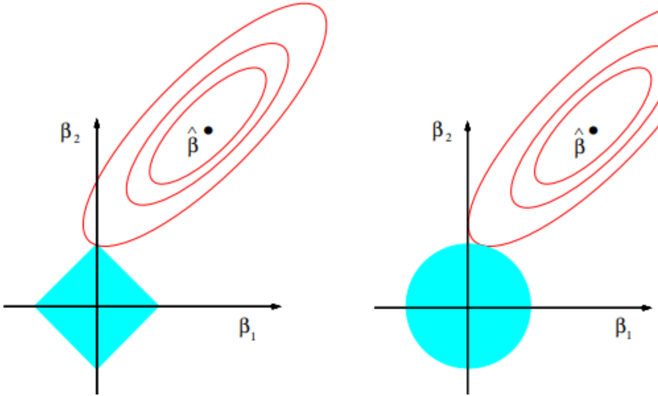
Ridge vs. Lasso Regression

Property 3: Ridge vs. Lasso Regression

1. Ridge is more easily interpretable.
2. Ridge is less computationally heavy due to not needing advanced optimization techniques like the lasso.
3. Ridge always has a unique solution, unlike OLS and the lasso.
4. Lasso forces parameter estimates towards zero unlike ridge.
    - Lasso is an example of a *model selection technique* while ridge is not.

- ML engineers generally like lasso over ridge and econometricians like ridge over lasso.

Ridge vs Lasso Regression



- Left side depicts lasso
- Right side depicts ridge

Elastic Net Regression

Definition 10: Elastic Net Regression

Elastic net regression is a regression technique that adds the L1-Norm and the L2-Norm to the OLS objective function.

- Essentially a mix of the ridge and lasso regressions.

Elastic Net Objective

Definition 11: Elastic Net Objective

The elastic net regression objective is given by

$$\arg\min_{\widehat{\boldsymbol{\beta}}_E} \left[\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_E \right)' \left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_E \right) + \lambda \left(\alpha \left\| \widehat{\boldsymbol{\beta}}_E \right\|_1 + (1-\alpha) \left\| \widehat{\boldsymbol{\beta}}_E \right\|_2^2 \right) \right].$$

- $\alpha \in [0, 1]$ is another hyperparameter that determines how much L1-penalization we want versus L2-penalization.
  - Downfall of elastic net is we have another hyperparameter to deal with.

BRM Shrinkage Methods

Definition 12: BRM Shrinkage Methods

A binary response model (BRM) shrinkage method simply adds the penalization term to the log-likelihood function and carries out MLE as usual.

BRM Log-Likelihood Function

Definition 13: BRM Log-Likelihood Function

Recall the log-likelihood function for a BRM is given by

$$\sum_{i=1}^{n} \left[y_i \ln G(\boldsymbol{x}_i'; \boldsymbol{\beta}) + (1 - y_i) \ln \left[1 - G(\boldsymbol{x}_i'; \boldsymbol{\beta})\right]\right].$$

Lasso BRM Log-Likelihood Function

Definition 14: Lasso BRM Log-Likelihood Function

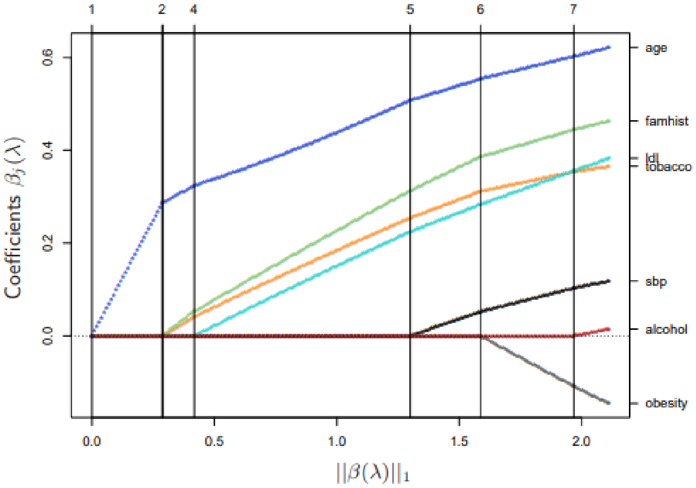
The log-likelihood function for a lasso BRM is given by

$$\sum_{i=1}^{n} \left[y_i \ln G(\boldsymbol{x}_i'; \boldsymbol{\beta}) + (1 - y_i) \ln \left[1 - G(\boldsymbol{x}_i'; \boldsymbol{\beta})\right]\right] - \left\|\widehat{\boldsymbol{\beta}}_L\right\|_1 .$$

where $G(\boldsymbol{x}_i'; \boldsymbol{\beta}) = \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i'; \boldsymbol{\beta})$.

- Maximize the log-likelihood to get Lasso BRM solution.
- We can easily extend shrinkage estimators to logit and probit models (which are really classification methods).
- Replacing the L1-Norm with the L2-Norm gives us a ridge BRM.

Lasso Logistic Regression



Thank You!