Lecture 5: Regularized Regression and Double Machine Learning

Isaiah Hull[1,2]

[1]BI Norwegian Business School
[2]CogniFrame

November 21, 2023

Lecture 5: Overview

1. Regularized Regression.
2. Double Machine Learning.
3. Applications.

1. Regularized Regression

Unsupervised Learning

Overview

- Based on James et al. (2023).
  - https://hastie.su.domains/ISLP/ISLP_website.pdf

Regularized Regression

Introduction

- Linear model fit with least squares:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

- Extend linear model: accommodate non-linear (but additive) relationships.
- Linear model advantages: inference and real-world competitiveness.
- Improve linear model and use alternative fitting procedures.

Regularized Regression

Advantage of Alternative Procedures

- Least squares: low bias when true relation is linear.
- If $n \gg p$: low variance, good test performance.
- If $p > n$: overfitting, poor test performance.
- Solution: constrain or shrink estimated coefficients.

Regularized Regression

Alternative Procedures

- Irrelevant variables add unnecessary complexity.
- Remove irrelevant variables by assigning zero coefficient estimates.

Regularized Regression

Alternatives to Least Squares

- Subset Selection:
  - Identify related predictors.
  - Fit model using least squares on reduced set.
- Shrinkage:
  - Fit with all predictors.
  - Coefficients shrunken towards zero.

Regularized Regression

Alternatives to Least Squares

- Subset Selection:
  - Identify related predictors.
  - Fit model using least squares on reduced set.
- Shrinkage:
  - Fit with all predictors.
  - Coefficients shrunken towards zero.

Alternatives to Least Squares

- Dimensionality Reduction:
  - Project p predictors into M-dimensional subspace ($M < p$).
  - Use M projections as predictors.

Regularized Regression

Subset Selection

- Best subset selection.
  - Fit least squares for each combination of p predictors.
  - Identify best model among 2^p possibilities.

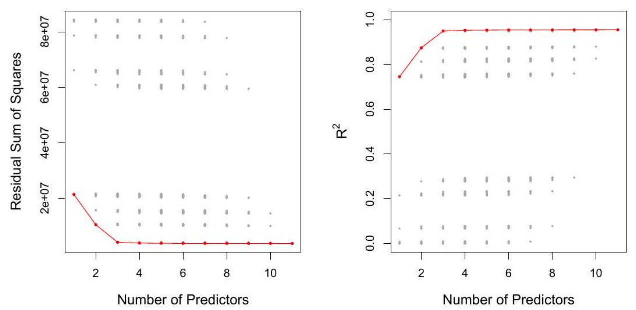
Regularized Regression

Algorithm: Best Subset Selection

1. Null model $\mathcal{M}_0$: Predict sample mean.
2. For $k = 1, 2, \ldots, p$:
   - Fit all $\binom{p}{k}$ models with k predictors.
   - Best model: smallest RSS (or largest R^2).
3. Select best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using validation, C_p, BIC, or cross-validation.

Regularized Regression

- Fit improves with more variables, but little improvement after 3 variables.



Source: James et al. (2023).

Regularized Regression

Best Subset Selection

- Applies to least squares regression, and others.
- Computationally intensive: 2^p models.
- Infeasible for large p.

Regularized Regression

Computational Considerations:

- $p = 20$ gives $1,048,576$ models for best subset.
- Alternative: forward stepwise algorithms only gives 211 models.

Regularized Regression

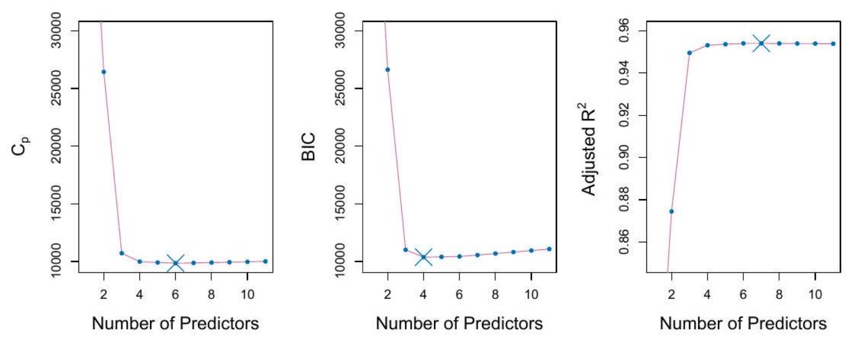
Choosing the Optimal Model

- Model with all predictors:
  - Smallest RSS and largest R^2
  - Training error $\neq$ Test error
- Need to estimate test error:
  1. Adjust training error for overfitting.
  2. Directly estimate test error (validation set or cross-validation).

Adjustment Methods

- Training set MSE: underestimate of test MSE.
- Can't use training set RSS and R^2 for models with different numbers of predictors.
- Adjusting techniques: AIC, BIC, and Adjusted R^2.

Regularized Regression



Source: James et al. (2023).

Regularized Regression

Shrinkage Methods vs. Subset Selection

- Subset selection uses least squares to fit a model with a subset of predictors.
- Shrinkage methods use all p predictors but shrinks coefficients towards zero.
- This reduces coefficient variance.
- Ridge regression and lasso.

Regularized Regression

Ridge Regression

- Extends least squares fitting.
- Minimizes:

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Regularized Regression

Ridge Regression Formula

- Ridge regression minimizes:

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Where:
  - $\lambda \geq 0$ is a tuning parameter.
  - $\lambda \sum_j \beta_j^2$ is the shrinkage penalty.

Regularized Regression

Role of the Tuning Parameter

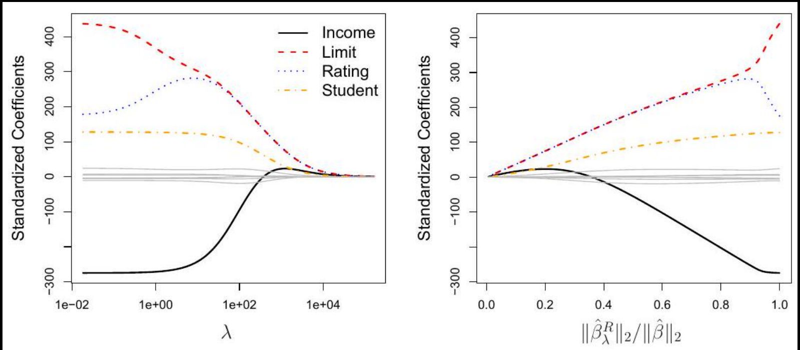
- When $\lambda = 0$, ridge regression yields least squares estimates.
- As $\lambda \to \infty$, coefficient estimates approach zero.
- Different coefficient estimates for each λ.
- λ selection is critical.

Shrinkage Penalty in Ridge Regression

- Applied to $\beta_1, \ldots, \beta_p$, not to β_0 (intercept).
- Shrink variable-response associations but not the intercept.
- If variables are centered (mean zero): $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i / n$.

Regularized Regression

Visualization of Ridge Regression



Source: James et al. (2023).

Application to Credit Data

- Figure shows ridge regression coefficients for credit data.
- Coefficients plotted as function of λ.
- At $\lambda = 0$, ridge coefficients = least squares estimates.

Regularized Regression

Application to Credit Data

- As λ increases, coefficients shrink towards zero.
- Variables with largest estimates: income, limit, rating, student.
- Right panel: same coefficients but with normalized ℓ_2 norm on the x-axis.

Regularized Regression

Application to the Credit Data

- Ridge regression coefficients depend on scaling of predictors.
- For interpretation: standardize predictors for ridge regression.
- Formula for standardization:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

Regularized Regression

Ridge Regression Advantages

- Rooted in bias-variance trade-off.
- Increased λ leads to decreased variance but increased bias.

Regularized Regression

Why Does Ridge Regression Work?

- Ridge regression advantages:
  - Computationally efficient.
  - Performs well even if $p > n$.

Regularized Regression

Lasso

- Alternative to ridge regression. Lasso coefficients:

$$\mathsf{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- ℓ_1 penalty forces coefficients to be exactly zero for large λ.
- Lasso performs variable selection, yielding sparse models.

Regularized Regression

Lasso

- At $\lambda = 0$, lasso gives least squares fit.
- At large λ, lasso gives null model.
- Ridge regression always includes all variables.
- Lasso model varies with λ value.

Regularized Regression

Lasso Formulation

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

- Relationship between λ and s.

Regularized Regression

Ridge Regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

- Relationship between λ and s.

Regularized Regression

Variable Selection with Lasso

- Lasso gives coefficient estimates equal to zero.
- In higher dimensions, lasso leads to feature selection.

Comparing Lasso and Ridge

- Lasso is simpler and more interpretable.
- Ridge shrinks each coefficient estimate proportionally.
- Lasso performs soft thresholding, shrinking coefficients toward zero by $\lambda/2$.
- Neither dominates universally; depends on data.
- Cross-validation can help decide the better approach.

Regularized Regression

Selecting the Tuning Parameter

- Need a method to select λ for ridge regression and the lasso.
- Use cross-validation: grid of λ values and compute CV error for each.
- Choose λ with smallest CV error.
- Refit model with all observations and selected λ.

Regularized Regression

Considerations in High Dimensions

- Traditional statistical techniques mostly for $n \gg p$.
- Most historical problems in statistics were low-dimensional.
- New technologies changed data collection methods.
- High-dimensional data: $p > n$ or $p \approx n$.

Why High Dimensions Matter

- Classical approaches like least squares not appropriate for $p > n$.
- High bias-variance trade-off; overfitting risks.
- Even when $p < n$, considerations apply.

Regularized Regression

What Goes Wrong in High Dimensions?

- Applying techniques not for high dimensions can be problematic.
- E.g., least squares: perfect fit with zero residuals when $p \geq n$.
- Perfect fits usually result in overfitting.
- Overfit models perform poorly on test sets.

2. Double Machine Learning

Double Machine Learning

Overview

- Based on Chernozhukov et al. (2018).
  - https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf

Double Machine Learning

Introduction

- Semiparametric problem of inference on a low-dimensional parameter θ_0.
- High-dimensional nuisance parameters η_0.
- Traditional assumptions (like Donsker properties) break down.
- Solution: Machine Learning (ML) methods for estimating η_0.

Double Machine Learning

Machine Learning in High Dimensions

- ML performs well in very high-dimensional problems.
- Regularization reduces variance.
- Regularization bias and overfitting cause bias in θ_0.
- Naive estimators fail to be $N^{-1/2}$ consistent.

Double Machine Learning

Addressing Bias

- Address bias by:
  - Using Neyman-orthogonal moments/scores.
  - Cross-fitting: an efficient data-splitting method.
- Result: Double or Debiased ML (DML).
- DML provides unbiased and approximately normally distributed estimators.

Double Machine Learning

Theoretical Basis of DML

- Elementary and requires weak theoretical prerequisites.
- Supports various ML methods for estimating nuisance parameters.
  - E.g., Random forests, lasso, ridge, neural nets, boosted trees.

DML Applications

- Regression parameters in partially linear regression.
- Coefficients on endogenous variables.
- Average treatment effects.
- Local average treatment effect in IV settings.

Double Machine Learning

Partially Linear Regression (PLR)

$$
\begin{aligned}
Y &= D\theta_0 + g_0(X) + U, \qquad & Ep[U|X, D] = 0, \\
D &= m_0(X) + V, \qquad & Ep[V|X] = 0,
\end{aligned}
$$

- Y: outcome, D: policy/treatment, X: controls.
- Infer θ_0, the treatment effect parameter.
- Treatment variable dependence on controls.
- High dimensional X.

Double Machine Learning

Regularization Bias

- Naive approach: ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for $D\theta_0 + g_0(X)$.
- Sample split: Main sample (size n) and auxiliary sample (size $N - n$).
- Estimator $\hat{\theta}_0$ often slower than $1/\sqrt{n}$ rate.
- Bias in learning g_0.

Double Machine Learning

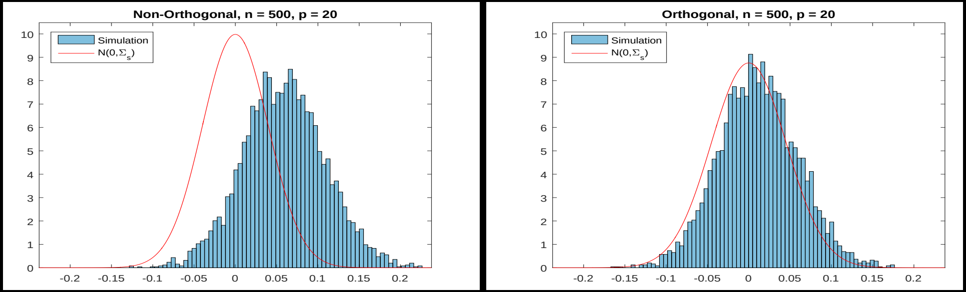
Overcoming Regularization Biases using Orthogonalization

- Use orthogonalized formulation: $\hat{V} = D - \hat{m}_0(X)$.
- Concept: Double prediction or "double machine learning."
- Debiased ML estimator for θ_0:

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i \in I} \hat{V}_i D_i\Big)^{-1} \frac{1}{n}\sum_{i \in I} \hat{V}_i (Y_i - \hat{g}_0(X_i)).$$

- Benefit: Removes regularization bias.
- Links to classical econometric literature and debiased lasso.

Double Machine Learning



Source: Chernozhukov et al. (2018).

- Left Panel: Non-orthogonal ML estimator.
- Right Panel: Orthogonal, DML estimator.

Double Machine Learning

Removing Bias with Sample Splitting

- Uses sample-splitting to ensure remainder terms vanish in probability.
- In the partially linear model, remainder contains terms like:

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i(\hat{g}_0(X_i) - g_0(X_i)) \tag{2}$$

Double Machine Learning

Removing Bias with Sample Splitting

- Sample splitting provides tight control of terms.
- Observations assumed to be independent.
- Variance of the term is of order:

$$\frac{1}{n}\sum_{i\in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \to_P 0 \tag{3}$$

- Term vanishes in probability by Chebyshev's inequality.

Double Machine Learning

Efficiency and Cross-fitting

- Direct application of sample splitting may lose efficiency.
- Swapping roles of main and auxiliary samples can regain efficiency.
- This procedure is known as *cross-fitting*.
- Cross-fitting offers an efficient averaging procedure.

Double Machine Learning

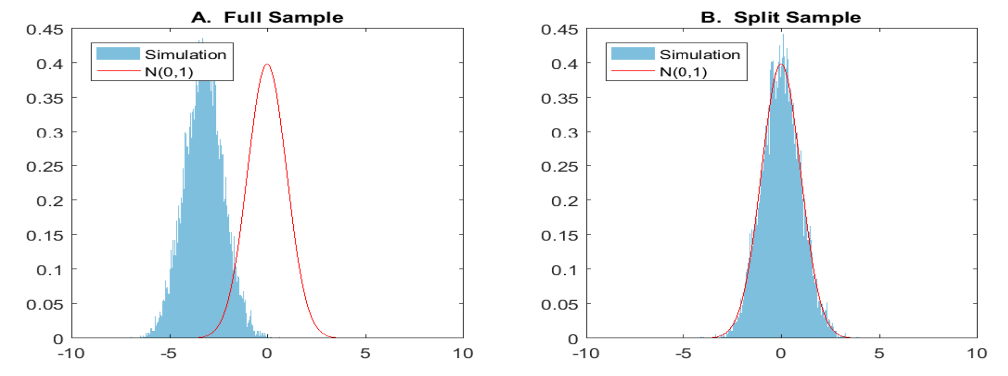
Potential Issues Without Sample Splitting

- Without sample splitting, terms may not vanish leading to poor estimator performance.
- Issues arise when data used in forming estimator.
- Poor performance even with favorable convergence rates.
- Overfitting example:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} V_i(\hat{g}_0(X_i) - g_0(X_i)) \propto N^{\epsilon} \to \infty \tag{4}$$

Double Machine Learning

Bias without sample splitting.



Source: Chernozhukov et al. (2018): Bias from overfitting in nuisance function estimation.

3. Applications

Introduction to TensorFlow

Colab Tutorial

- Regularized Regression
- Double Machine Learning

References I

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21 (1), C1–C68, 10.1111/ectj.12097.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2023) *An Introduction to Statistical Learning: with Applications in Python*, Springer Texts in Statistics: Springer Cham, 1st edition, XV, 60, https://doi.org/10.1007/978-3-031-38747-0, eBook ISBN: 978-3-031-38747-0; Softcover ISBN: 978-3-031-39189-7; Series ISSN: 1431-875X; Series E-ISSN: 2197-4136.