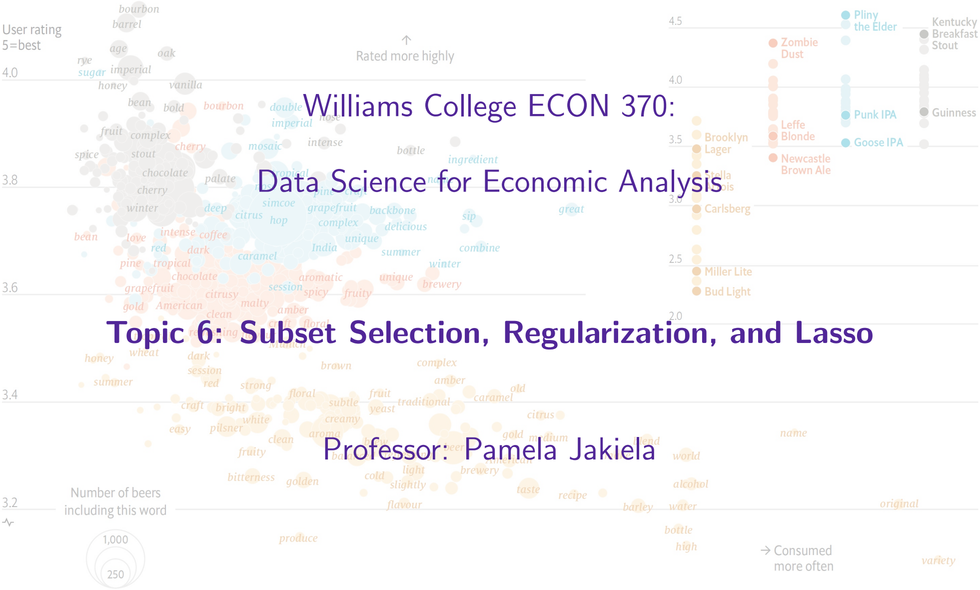


Williams College ECON 370:

Data Science for Economic Analysis

Topic 6: Subset Selection, Regularization, and Lasso

Professor: Pamela Jakiela

Outline

- Approaches to subset selection
- Regularization methods: ridge regression and lasso
- Choosing the penalty parameter

When Do Economists Care About Prediction? RCTs as a Case Study

THE CONVERSATION

Academic rigour, journalistic flair

How randomised trials became big in development economics

Published: December 9, 2019 8.19am GMT

Esther Duflo (L) and Abhijit Banerjee, who, along with Michael Kremer (not pictured), won the 2019 Nobel Prize in Economic Sciences. EPA/CJ Gunther

Source: The Conversation (2019)

When Do Economists Care About Prediction? RCTs as a Case Study

- Statistical power in a randomized trial depends on residual variance in the outcome (Power is the probability of of finding an impact – i.e. rejecting H_0 – if there is one.)
  - A typical RCT regression equation: $Y_{1,i} = \alpha + \beta D_i + \delta X_{0,i} + \gamma Y_{0,i} + \varepsilon_i$
    - $Y_{1,i}$ is the outcome, measure after the intervention
    - D_i is a dummy for being randomly assigned to the treatment group
    - $Y_{0,i}$ is the baseline value of the outcome
    - $X_{0,i}$ is a set of other baseline covariates that (one hopes) predict $Y_{1,i}$
  - The **minimum detectable effect** that a researcher can expect to measure through an RCT is proportional to the standard deviation of the residuals, i.e. to the unexplained variation in Y
- Economists running RCTs choose which covariates to measure, and pay for data collection
  - We want to measure Xs that predict Y, and we don't want to throw money away (by measuring a large number of baseline covariates that do not predict variation in Y)

Choosing Covariates in an RCT: The EMERGE Project

Choosing Covariates in an RCT: The EMERGE Project

EMERGE was a cluster-randomized evaluation of an early literacy program in rural Kenya

- Intervention involved mother tongue storybooks and parent education
- Key child development outcomes of interest: literacy and vocabulary
- Large research team including me, Prof. Ozier, and two public health collaborators
- We designed survey instruments, and had to choose which variables to measure at baseline

Child development and educational outcomes tend to have high serial correlation

- Individual ability at time $t-1$ is a strong predictor of individual ability at time t
- The right set of covariates can substantially increase effective sample size
- Measuring child development is costly in terms of time/money because each variable is constructed from multiple survey questions, and modules are administered one-on-one

Best Subset Selection

A **best subset selection** algorithm:

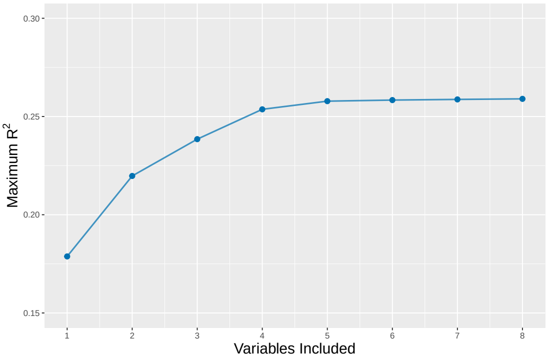
- For each number of possible covariates $k = 1, 2, \ldots, p$,
  - Fit all models containing exactly k covariates
  - Identify the "best" in terms of R^2
- Choose the best subset using cross-validation or an alternative approach
  - Need to address the fact that R^2 always increases with k

Best Subset Selection Example: EMERGE

Use $N = 1{,}000$ data set on child development outcomes from EMERGE project

- `literacy`: measure of early literacy based on Early Grade Reading Assessment
- `age_months`: child age in months at time of survey
- `male`: dummy for boys
- `haz`: height-for-age z-score, measure of nutritional status
- `receptive`: receptive vocabulary, i.e. the ability to understand words (z-score)
- `expressive`: expressive vocabulary, i.e. the ability to produce words (z-score)
- `fine_motor`: fine motor skills (z-score)
- `hh_size`: household size
- `mom_educ`: mother's years of schooling

R^2 Is Increasing in the Number of Covariates



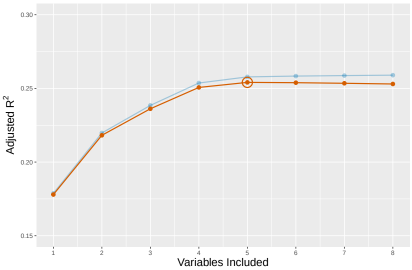
	Number of Covariates							
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
age_months						X	X	X
male		X	X	X	X	X	X	X
haz			X	X	X	X	X	X
receptive					X	X	X	X
expressive	X	X	X	X	X	X	X	X
fine_motor								X
hh_size							X	X
mom_educ				X	X	X	X	X

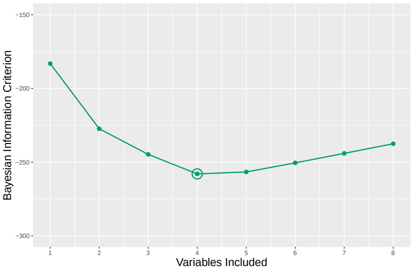
Choosing the Number of Covariates: Alternatives to Cross-Validation

Three alternatives to R^2 that adjust for the number of covariates in the specification, d

- Adjusted R^2: $1 - \frac{RSS(n-d-1)}{TSS(n-1)}$ (seek to maximize)
- Akaike Information Criterion (AIC): $\left(RSS + 2d\hat{\sigma}^2\right)/n$ (seek to minimize)
- Bayesian Information Criterion (BIC): $\left(RSS + \ln(d)\hat{\sigma}^2\right)/n$ (seek to minimize)

Choosing the Number of Covariates: Alternatives to Cross-Validation





Best Subset Selection Is an Extension to OLS

In OLS, we seek to minimize:

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

Best subset selection can be expressed as: choose β to minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p} I\left(\beta_j \neq 0\right) \leq s$$

where s is the number of regressors/predictors/features/covariates

Best Subset Selection Is Not Feasible with Many Covariates

Best subset selection is an extension to OLS that is solved algorithmically, not analytically

- When p is large, finding the best subset is computationally impossible ($2^p - 1$ regressions)
  - With 8 possible covariates: 255 regressions
  - With 20 possible covariates: over one million regressions
- Best subset selection makes sense when you can narrow the set of potential controls
  - Surveys often contain hundreds of questions
- Less computationally-intensive alternatives (forward and backward stepwise selection) exist but they are not robust to all patterns of correlation among potential covariates
  - Stepwise approaches involve adding (forward selection) or dropping (backward selection) the variable that gives the largest increase (forward) or smallest decrease (backward) in R^2

Shrinkage Operators: Machine Learning Extensions to OLS

Machine learning **shrinkage operators** (ridge regression, lasso) extend OLS to better predict Y

- Basic idea is to fully "kitchen sink" our regressions while proactively correcting for potential over-fitting, allowing us to leverage info from more covariates effectively

Lasso is attractive because it identifies a subset of Xs that are most effective predictors of Y

Can We Improve on OLS?

A standard linear model may not be the best way to predict Y:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Can we improve on OLS?

- When p is large relative to N, OLS is prone to over-fitting
- OLS explains both structural and spurious relationships in data

Like best subset selection, shrinkage operators minimize RSS subject to an additional constraint

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

Ridge Regression

Ridge regression solves the minimization problem:

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

or, equivalently,

$$\min_{\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

for some **tuning parameter** $\lambda \geq 0$

Ridge regression shrinks OLS coefficients toward zero

- Shrinkage is more or less proportional, so ridge regression does not identify a subset of regressors to include in the regression model (it just down-weights some relative to others)

Shrinkage Operators: What's in a Name?

Like OLS, ridge regression has an analytical solution, as we can see in the $p = 1$ case:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} x_i y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - N\bar{X}^2} > \frac{\sum_{i=1}^{n} x_i y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - N\bar{X}^2 + 2\lambda} = \hat{\beta}_{ridge}$$

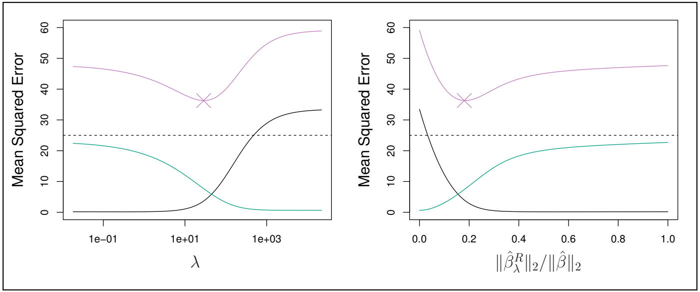
The (bivariate) ridge regression coefficient is smaller than the (bivariate) OLS coefficient

- When λ is close to 0, $\hat{\beta}_{ridge}$ is similar to $\hat{\beta}_{OLS}$
- $\hat{\beta}_{ridge}$ approaches 0 as λ gets large

With more than one independent variable, some ridge regression coefficients may be larger than OLS counterparts, and the coefficient on a specific X_k need not decline monotonically with λ

- Shrinkage is more or less proportional, so ridge regression does not identify a subset of regressors to include in the regression model (it just down-weights some relative to others)

OLS is BLUE, But Ridge Regression (Sometimes) Has Lower MSE



Source: James et al. (2021)

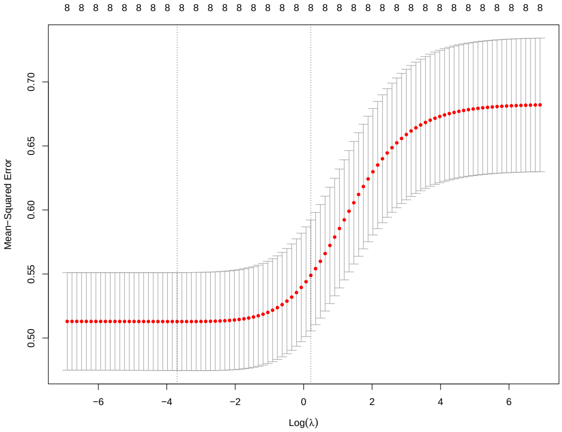
Gauss-Markov Theorem: OLS is the **best linear unbiased estimator** (BLUE) of Y

- Ridge regression is biased (black line), but has lower variance relative to the true underlying β (green line) and can therefore achieve lower MSE (pink line) for some λs

Ridge Regression in Practice

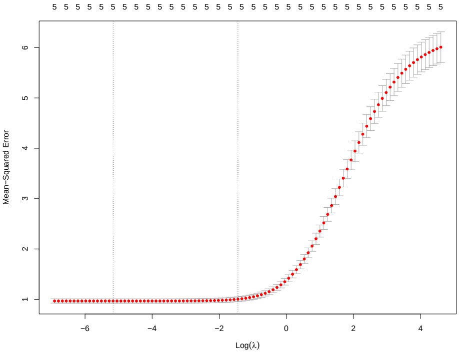
		Ridge Regression		
	OLS	$\lambda = 10^{-2}$	$\lambda = 10$	$\lambda = 10^4$
Variable	**(1)**	**(2)**	**(3)**	**(4)**
expressive	0.2543	0.2498	0.0247	0.0003
male	-0.3152	-0.3106	-0.0203	-0.0002
haz	0.0847	0.0844	0.0130	0.0002
mom_educ	0.0439	0.0436	0.0051	0.0001
receptive	0.0651	0.0671	0.0195	0.0002
age_months	0.0024	0.0024	0.0002	0.0000
hh_size	-0.0085	-0.0084	-0.0011	0.0000
fine_motor	0.0257	0.0269	0.0160	0.0002

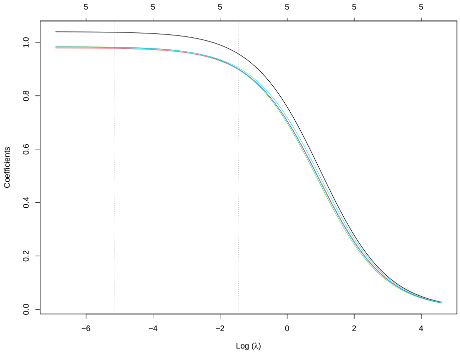
Choosing the Penalty Parameter to Minimize Test MSE: EMERGE Data



data source: EMERGE project (Jakiela, Ozier, Fernald, and Knauer 2021)

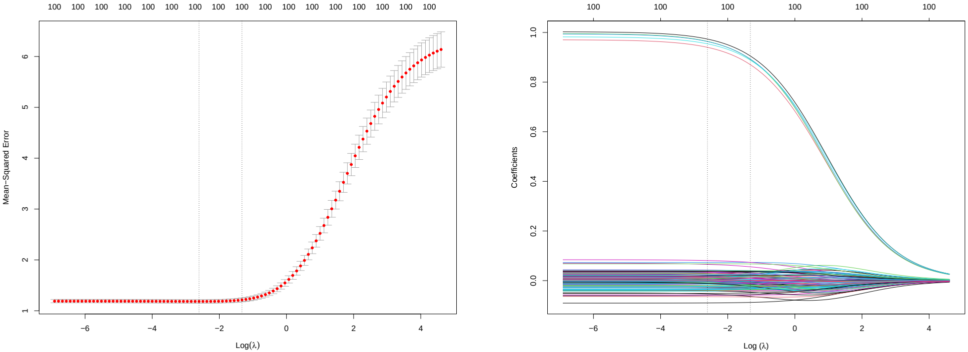
Ridge Regression in Simulated Data: $N = 1000$, $K = 5$





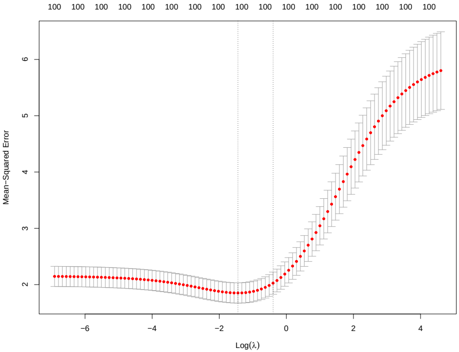
data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 5$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 5$

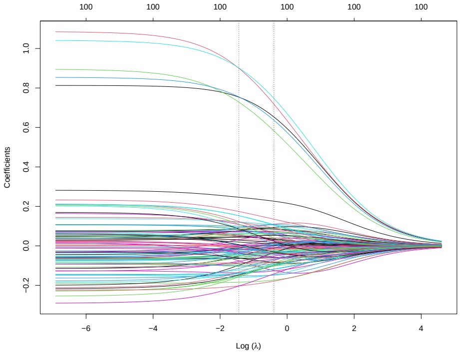
Ridge Regression in Simulated Data: $N = 1000$, $K = 100$



data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 100$

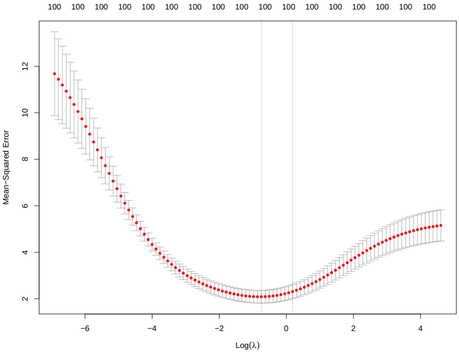
Ridge Regression in Simulated Data: $N = 200,\ K = 100$

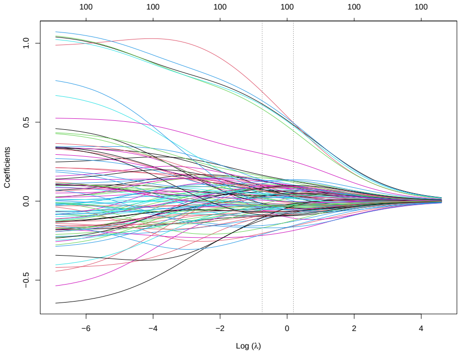




data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 200$, $K = 100$

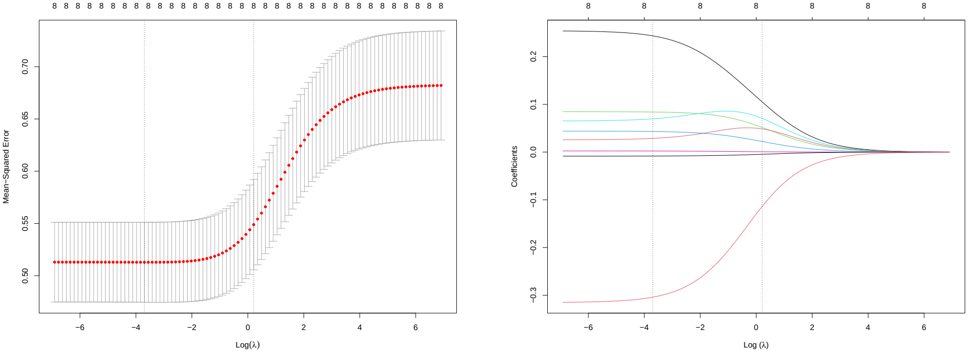
Ridge Regression in Simulated Data: $N = 120$, $K = 100$





data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 120$, $K = 100$

Ridge Regression in the EMERGE Data



data source: EMERGE project (Jakiela, Ozier, Fernald, and Knauer 2021)

Ridge Regression in Practice: Comparing MSEs in EMERGE Data

Splitting the data into a training data set and a test data set, we see that ridge reduces the MSE in the test data as expected, but not by much (relative to the SD of the outcome, 0.8258)

OLS	λ^*	λ^{1SE}
0.4928	0.4899	0.5930

λ^* is the λ that minimizes test MSE in cross-validation, λ^{1SE} is 1 SE higher than λ^*

Shrinkage Operators: Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) seeks to minimize:

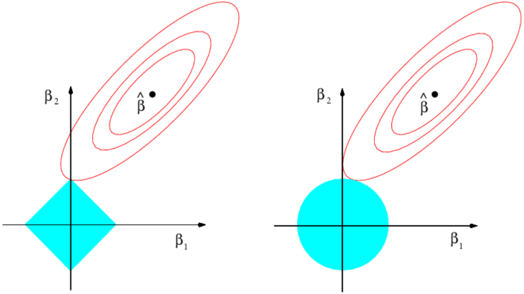
$$\min_{\beta} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

for some **tuning parameter** $\lambda \geq 0$

Lasso combines benefits of subset selection, ridge regression; useful for choosing covariates

- Less computationally intensive than subset selection
- Sets some coefficients to 0 $\rightarrow$ identifies parsimonious model

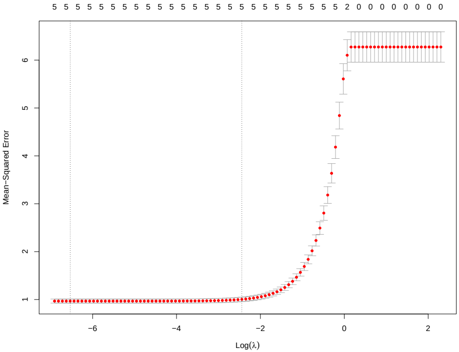
Lasso Sets Some Coefficients to Zero

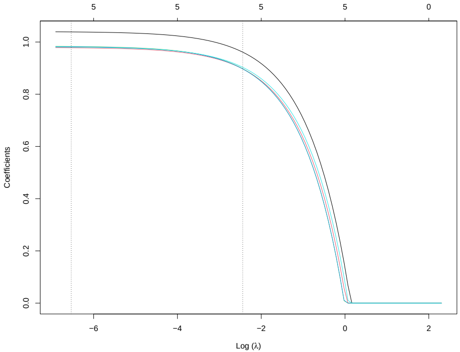


Source: James et al. (2021)

The lasso constraint region has sharp corners $\Rightarrow$ some coefficients set to 0

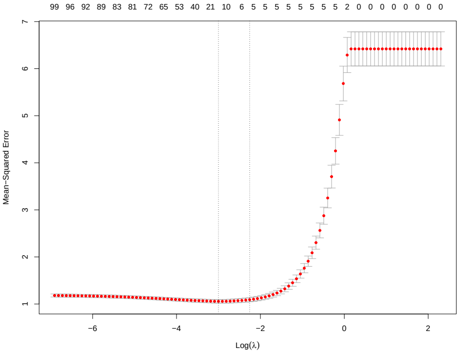
Lasso in Simulated Data: $N = 1000$, $K = 5$

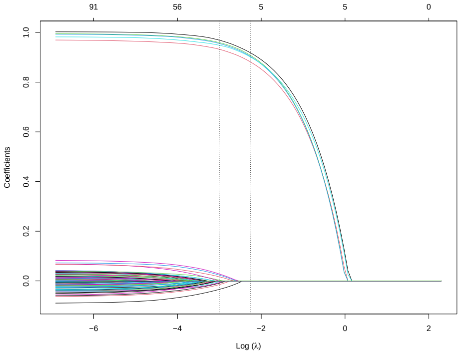




data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 5$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 5$

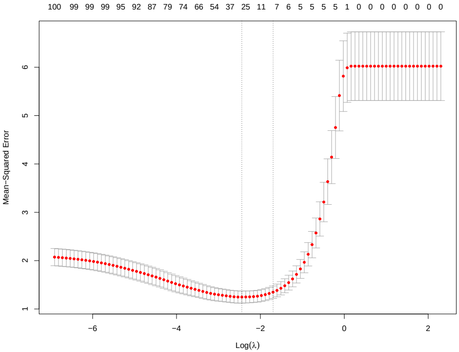
Lasso in Simulated Data: $N = 1000$, $K = 100$

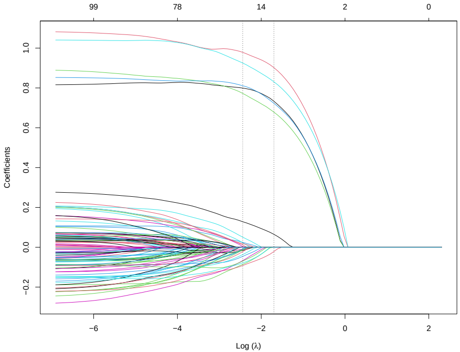




data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 1000$, $K = 100$

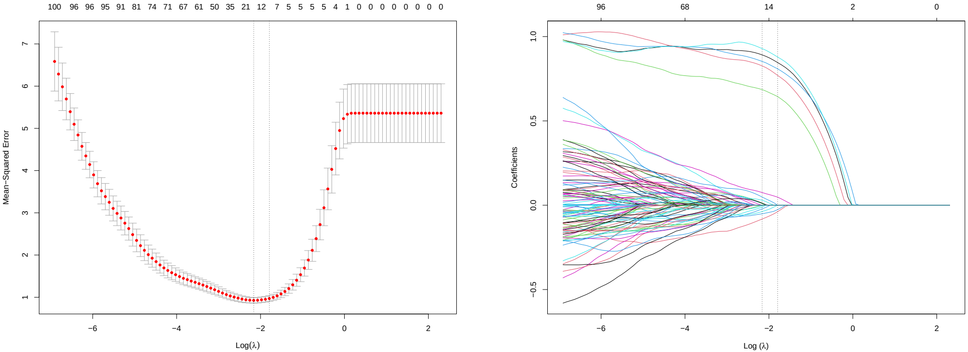
Lasso in Simulated Data: $N = 200$, $K = 100$





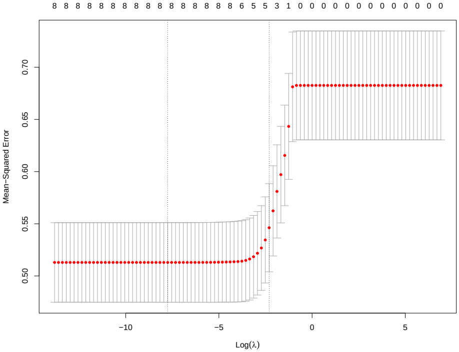
data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 200$, $K = 100$

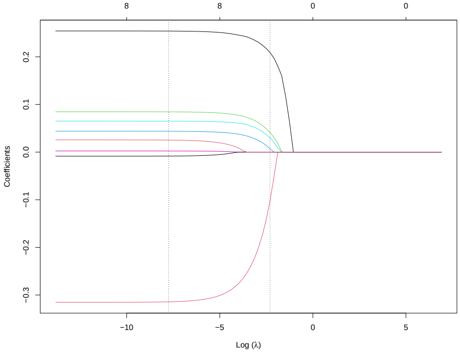
Lasso in Simulated Data: $N = 120$, $K = 100$



data-generating process: $Y = \sum_{k=1}^{5} X_k + \varepsilon$ where $X_k \sim N(0,1)$ for $k = 1, \ldots, 100$, $\varepsilon \sim N(0,1)$, $N = 120$, $K = 100$

Lasso in Practice: EMERGE Data





data source: EMERGE project (Jakiela, Ozier, Fernald, and Knauer 2021)

Alternative "Data-Driven" Approach to Choosing λ

Belloni and Chernozhukov (2011), Belloni et al. (2012): alternative approach to choosing λ

- Chooses λ iteratively based on data, penalties vary across variables
- Errs on the side of choosing fewer controls to avoid over-fitting
- Allows for heteroskedasticity
- Designed to allow for valid post-selection lasso estimation (within a single data set)

Approaches may generate different sets of controls

- Costs of too many/too few may vary across empirical contexts

In $N = 120$, $K = 100$ simulated data, data-driven lasso $\Rightarrow$ 9 X variables selected

Comparing Approaches to Choosing Covariates via Lasso

Variable	OLS	λ^*	λ^{1SE}	λ^{DD}
age_months	X	X		
male	X	X	X	X
haz	X	X	X	X
receptive	X	X	X	X
expressive	X	X	X	X
fine_motor	X	X		X
hh_size	X	X		
mom_educ	X	X	X	X

Summary

Best subset selection, ridge regression, and lasso are constrained extensions to OLS

- Ridge and lasso are regularized: coefficients are shrunk toward zero to reduce over-fitting
- Best subset selection and lasso are useful for model selection (i.e. choosing covariates)

Lasso is now widely used by economists to choose a subset of (many) controls to include in OLS

- Number of controls selected depends on the penalty (or tuning) parameter
    - Cross-validation is optimizing prediction, leads to the inclusion of more controls
    - Data-driven approach of Belloni et al. (2012) or 1 SE rule typically better heuristics
- Desired number of controls may also depend on the cost of adding/including a variable
    - Expressive vocabulary, male dummy both predict emergent literacy in EMERGE data, but measuring expressive vocabulary probably costs thousands of times more per child

Epilogue: Covariate Selection via Post-Double-Selection (PDS) Lasso

EMERGE example focuses on choosing covariates **in advance**, deciding what to measure

- The error terms in the covariate selection process are independent of eventual analysis

Model selection – choosing covariates using your analysis sample – is more complicated

- Conducting statistical inference after covariate selection within a sample is complicated
- Post-double selection lasso is a widely used approach that addresses inference concerns
    - Step 1: run lasso to choose covariates that predict outcome Y
    - Step 2: run lasso to choose covariates that predict treatment of interest T
    - Step 3: run OLS including all covariates identified in first two steps, conduct inference

Lab # 6

Objective: use lasso to choose covariates in the EMERGE data

- `ECON370-lab6-data.csv` contains 200 observations and 25 variables
- You will create dummies for survey enumerator and randomization strata

Two approaches to choosing tuning/penalty parameter: CV and "data-driven" approach

- CV possible in both R and Python, "data-driven" approach only possible in R (?)
- Use CV to identify both MSE-minimizing tuning parameter and 1 SE alternative
  - 1 SE rule: find the SE of the MSE (across k folds) at the minimum, then find largest value of the tuning parameter that yields a test MSE below the sum of the minimum MSE + the SE
- Set of selected covariates will depend (a lot!) on your choice of tuning parameter