# Model Selection & Regularization

Lecture 7

Termeh Shafie

## Recall: Linear Models and Least Squares

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j-1}^{p} \hat{\beta}_{j_1} x_{ij})^2$$

Model with all available predictor variables is commonly referred to as **the full model**

**Issues:**

- predictive accuracy
- model interpretability

**Solutions:**

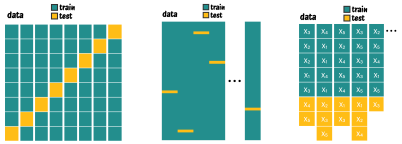
- select subset of predictors
- consider extension to the least squares solution of full model

## Part I - Variable Subset Selection

## Model Selection Criteria: Validation by Prediction Error

**Last week:** how to use cross validation to choose a set of predictors
by directly estimate prediction error using cross-validation techniques

e.g. $\text{MSE} = \frac{\text{RSS}}{n}$ $\text{RMSE} = \sqrt{\frac{\text{RSS}}{n}}$ $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$








**Now:** indirectly estimating test performance using an approximation

## Model Selection Criteria

### Four ways to estimate test performance using an approximation

Full model has $p$ predictors

RSS is the residual sum of squares for model with $d$ predictors

$\hat{\sigma}^2 = \text{RSS}_p/(n-p-1)$ is an estimate of the error variance for full model

**1. Mallow's $C_p$ criterion:**

For a given model with $d$ (out of the $p$ available) predictors

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

we are penalizing models of higher dimensionality (larger $d$, greater penalty)
$\Longrightarrow$ choose the model which has **minimum** $C_p$

## Model Selection Criteria

### Four ways to estimate test performance using an approximation

Full model has $p$ predictors

RSS is the residual sum of squares for model with $d$ predictors

$\hat{\sigma}^2 = \text{RSS}_p/(n-p-1)$ is an estimate of the error variance for full model

**2. Akaike Information Criterion (AIC)**

For linear models: equivalent to Mallow's $C_p$ (proportional to)

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

we are penalizing models of higher dimensionality (larger $d$, greater penalty)
$\Longrightarrow$ choose the model which has **minimum** $AIC$

## Model Selection Criteria

### Four ways to estimate test performance using an approximation

Full model has $p$ predictors

RSS is the residual sum of squares for model with $d$ predictors

$\hat{\sigma}^2 = \text{RSS}_p/(n-p-1)$ is an estimate of the error variance for full model

**3. Bayesian Information Criterion (BIC)**

For linear models: equivalent to Mallow's $C_p$ (proportional to)

$$BIC = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + \underbrace{\log(n)}_{\text{heavier penalty}} d\hat{\sigma}^2\right)$$

we are penalizing models of higher dimensionality (larger $d$, greater penalty)

$\Longrightarrow$ choose the model which has **minimum** $BIC$

## Model Selection Criteria

### Four ways to estimate test performance using an approximation

Full model has $p$ predictors

RSS is the residual sum of squares for model with $d$ predictors

$\hat{\sigma}^2 = \text{RSS}_p/(n-p-1)$ is an estimate of the error variance for full model

**4. Adjusted R-squared value**

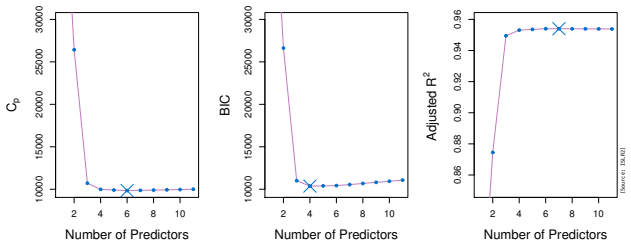
Adjust the regular $R^2$ by taking into account number of predictors

$$\text{Adjusted-}R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

$\Longrightarrow$ choose the model which has **maximum** Adjusted-$R^2$

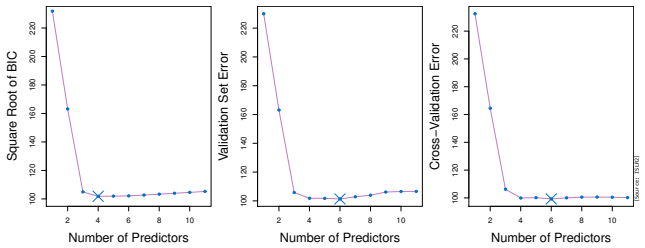
## Model Selection Criteria

### Four ways to estimate test performance using an approximation



# Model Selection Criteria

### ...and compared to cross validation




# Model Search Methods

### Best Subset Selection

1. Let $M_0$ denote null model which contains no predictors. This model simply predicts the of the response for each observation.
2. For $k = 1,2,\ldots,p$
   - Fit all $\binom{p}{k}$ models that contain exactly $p$ predictors
   - Pick the best among these $\binom{p}{k}$ models and call it $M_k$.

     Here, *best* is defined as having the smallest RSS or largest $R^2$
3. Select a single best model from among $M_0, M_1, \ldots, M_p$ using cross validated prediction error, $C_p$ ($AIC$), $BIC$, or Adjusted-$R^2$

requires training $2^p$ models

**Example**

$p = 3$

$M_0$ : intercept only (null)

$C_1$: $(X_1)$ $(X_2)$ $(X_3)$

lowest training RSS within $C_1$

$\Longrightarrow M_1$

$C_2$: $(X_1, X_2)$ $(X_1, X_3)$ $(X_2, X_3)$

lowest training RSS within $C_2$

$\Longrightarrow M_2$

$M_3$ : full model with $(X_1)$ $(X_2)$ $(X_3)$

# Model Search Methods

### Forward Stepwise Selection

1. Let $M_0$ denote null model which contains no predictors.
2. For $k = 1,2,\ldots,p-1$
   - Consider all $p-k$ models that augment the predictors in $M_k$ with one additional predictor
   - Choose the best among these $p-k$ models and call it $M_{k+1}$. Here, *best* is defined as having the smallest RSS or largest $R^2$
3. Select a single best model from among $M_0, M_1, \ldots, M_p$ using cross validated prediction error, $C_p$ ($AIC$), $BIC$, or Adjusted-$R^2$

requires training $1 + \frac{p(p+1)}{2}$ models

**Example**

$p = 3$

$M_0$ : intercept only (null)

$C_1$: $(X_1)$ $(X_2)$ $(X_3)$

lowest training RSS within $C_1$

$\Longrightarrow M_1$

$C_2$: $(X_1, X_2)$ $(X_2, X_3)$

lowest training RSS within $C_2$

$\Longrightarrow M_2$

$M_3$ : full model with $(X_1)$ $(X_2)$ $(X_3)$

# Model Search Methods

## Backward Stepwise Selection

1. Let $M_p$ denote full model which all predictors.
2. For $k = p, p-1, p-2, \ldots, 1$
   - Consider all $k$ models that contain all but one of the predictors in $M_k$, for a total of $k-1$ predictors
   - Choose the best among these $k$ models and call it $M_{k-1}$. Here, *best* is defined as having the smallest RSS or largest $R^2$
3. Select a single best model from among $M_0, M_1, \ldots, M_p$ using cross validated prediction error, $C_p$ ($AIC$), $BIC$, or Adjusted-$R^2$

requires training $1 + \frac{p(p+1)}{2}$ models

**Example**

$p = 3$

$M_3$ : full mode $X_1$ $X_2$ $X_3$

$C_2$: $X_1, X_2$ $X_1, X_3$ $X_2, X_3$

lowest training RSS within $C_1$

$\Longrightarrow M_2$

$C_1$: $X_1$ $X_2$

lowest training RSS within $C_2$

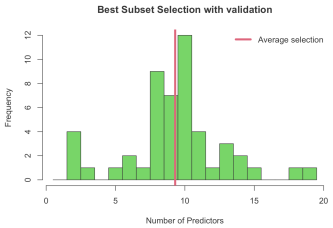
$\Longrightarrow M_1$

$M_0$ : intercept only (null)

# Model Search Methods

## Best Subset Selection

validation approach based on 50 different seeds and storing number of predictors in selected model each time




[plot is made based on the 'hitters' data se used in this week's practical in ISLR2]

# Part II - Shrinkage

## Shrinkage Methods

Before: **Discrete model search methods**

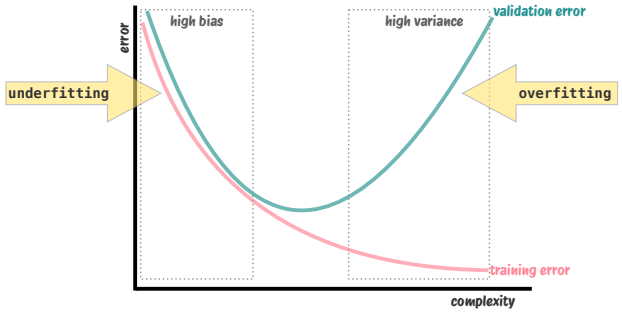
$$\underbrace{\text{model fit}}_{\text{RSS}} + \text{penalty on model dimensionality}$$

Now: **Continuous model search method** (also faster)

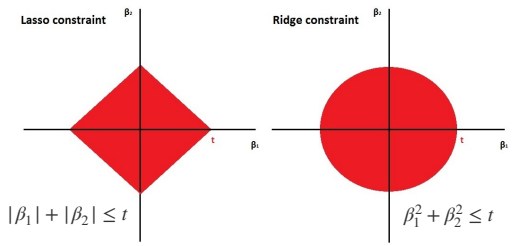
$$\underbrace{\text{model fit}}_{\text{RSS}} + \text{penalty on size of coefficients}$$

this is called **penalized** or **regularized regression**

## Bias Variance Trade-Off




## Ridge and Lasso Regression: The Constraints



# Ridge Regression

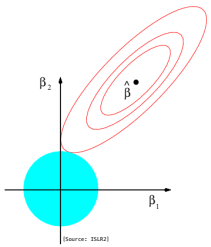
ridge uses $\ell_2$ penalty

Least Squares produces estimates by minimizing

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j-1}^{p}\hat{\beta}_{j_1}x_{ij})^2$$

Ridge regression instead minimizes

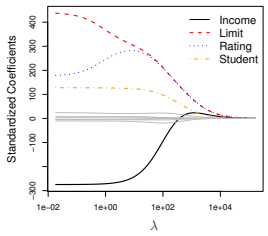
$$\underbrace{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j-1}^{p}\hat{\beta}_{j_1}x_{ij})^2}_{\text{model fit}} + \underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{penalty}} = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

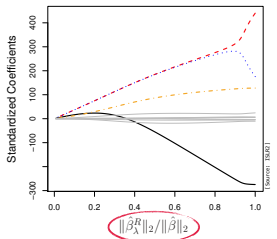



where $\lambda \geq 0$ is the tuning parameter controlling trade off between model fit and size of coefficients ($\lambda \to \infty, \hat{\beta}_j \to 0$)

# Ridge Regression

## Regularization Paths



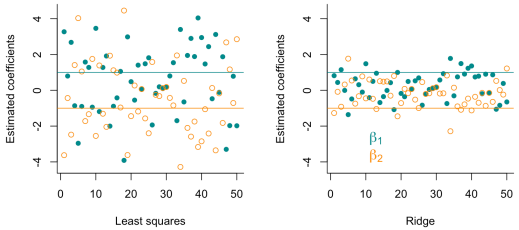




$\frac{\|\hat{\beta}_\lambda^R\|_2}{\|\hat{\beta}\|_2}$

$\ell_2$ norm $= \|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$

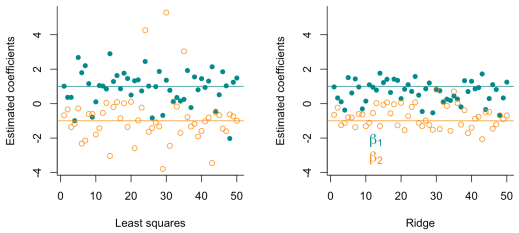
# Ridge Regression

## Advantage 1: Multicollinearity (a simulation study)



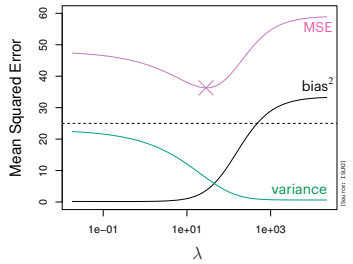
## Ridge Regression

**Advantage 2: When $p$ is close to $n$ (a simulation study)**




## Ridge Regression

**Bias-Variance Trade Off**




## Lasso Regression

**Least Absolute Shrinkage and Selection Operator**

Least Squares produces estimates by minimizing

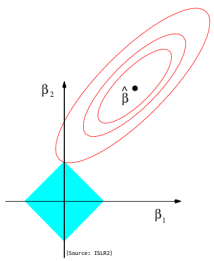
$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j-1}^{p}\hat{\beta}_{j_1}x_{ij})^2$$

Lasso regression instead minimizes

$$\underbrace{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j-1}^{p}\hat{\beta}_{j_1}x_{ij})^2}_{\text{model fit}} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{penalty}} = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

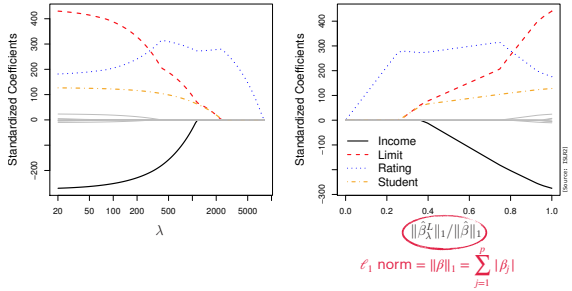
where $\lambda \geq 0$ is the tuning parameter controlling trade off between model fit and size of coefficients ($\lambda \to \infty, \hat{\beta}_j = 0$)

lasso uses $\ell_1$ penalty



## Lasso Regression

### Regularization Paths



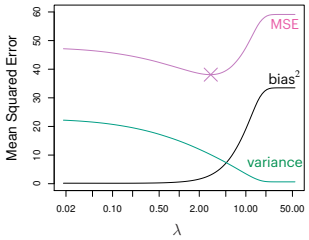


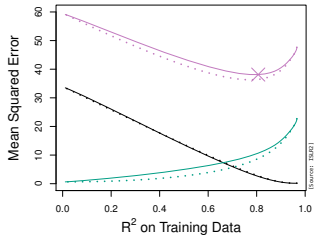


$\ell_1$ norm $= \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

## Lasso Regression

### Bias-Variance Trade Off




## Ridge vs. Lasso Regression

- Both ridge and lasso are convex optimization
- The ridge solution exists in closed form
- Lasso does not have closed form solution, but very efficient optimization algorithms exist

### When to choose which?

- When the actual data-generating mechanism is sparse lasso has the advantage
- When the actual data-generating mechanism is dense ridge has the advantage

Sparse mechanisms: Few predictors are relevant to the response → good setting for lasso regression
Dense mechanisms: A lot of predictors are relevant to the response → good setting for ridge regression

- Also depends on:
  - Signal strength (the magnitude of the effects of the relevant variables)
  - The correlation structure among predictors
  - Sample size $n$ vs. number of predictors $p$

## Ridge vs. Lasso Regression

**Ridge**

+ Reduces Multicollinearity
+ Continuous Shrinking
+ Stable Solutions
+ Computationally Efficient

- No variable selection
- Interpretability
- Sensitive to scale

**Lasso**

+ Variable selection
+ Sparse models
+ Improves interpretability
+ Particularly useful for when $p > n$

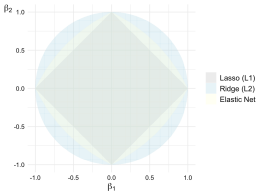
- Collinearity issues
- Bias in coefficients ($\ell_1$ penalty is harsher)
- Computationally intensive

## $\lambda$ Tuning


- *K*-fold Cross Validation
  1. Choose the number of folds *K*
  2. Split the data accordingly into training and testing sets.
  3. Define a grid of values for $\lambda$
  4. For each $\lambda$, calculate the validation MSE within each fold
  5. For each $\lambda$, calculate the overall cross-validation MSE
  6. Locate under which $\lambda$ cross-validation MSE is minimized, i.e. minimum_cv $\lambda$
- Packages such as will `glmnet` do this automatically

## Hybrid Approach: Elastic Nets

$$\text{RSS} + \underbrace{\lambda_1 \sum_{j=1}^{p} \beta_j^2}_{\text{“ridge”}} + \underbrace{\lambda_2 \sum_{j=1}^{p} |\beta_j|}_{\text{“lasso”}}$$




$\lambda_1$ and $\lambda_2$ are regularization parameters controlling the strength of the penalties

- Helps stabilize the solution when predictors are correlated
- Shrinks some coefficients to zero, enabling feature selection
- Particularly useful for high-dimensional datasets with correlated predictors

## Part III- Dimensionality Reduction

another strategy which aims to reduce dimensionality **before** applying LS
create $q$ transformed variables which are linear combinations of the original predictors ($q < p$)
we return to this during our PCA lecture...

## Part IV- Transformations: next week!

extensions to the regression model when the best straight line doesn't quite work!

## This Week's Practical

**Hands on discrete and continuous model search!**