Lecture 17: Model Selection Algorithms and Testing

Module 5: part 3

Spring 2025

Logistics

- Wrapping up Module 5 today (hopefully)
- Next module will be on generalized linear models (GLMs)

Recap: The Model Selection Problem

- Given a large set of potential predictors, how can we select an optimal model (i.e., a subset of covariates)?
- **Primary objective**: Select a model that generalizes well to unseen data
- We aim to minimize the *generalization error*, which requires:
- Avoiding the trap of simply evaluating performance on training data
- Finding the optimal complexity through the bias-variance trade-off:
  - Including relevant covariates reduces bias (improves prediction accuracy)
  - Too many covariates increases variance in parameter estimates $\hat{\beta}$
  - The goal is finding the "sweet spot" between underfitting and overfitting

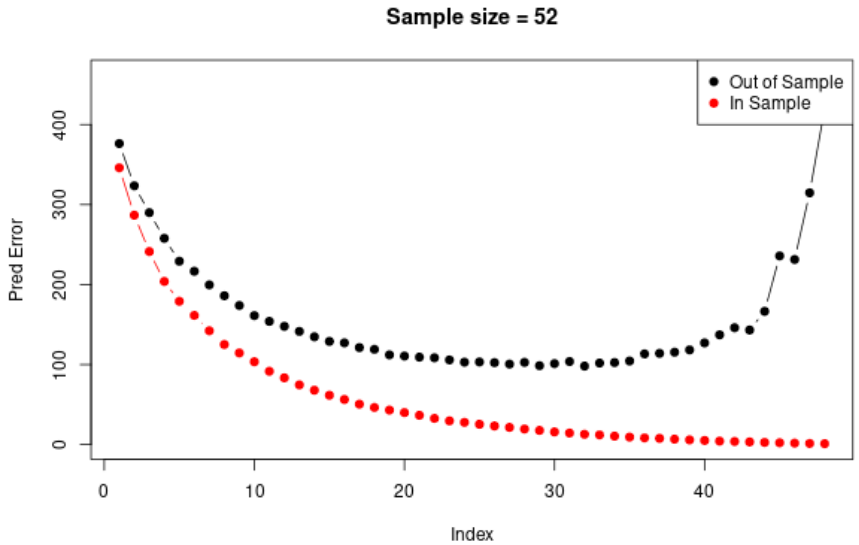

Figure: The U-shaped relationship between model complexity and prediction error

Model Evaluation Criteria

- Given a candidate model (a specific set of covariates), we can assess its generalization error using:
  - Cross-validation (CV)
  - Information-theoretic criteria:

$$\text{AIC} = -\frac{n}{2}\log(RSS/n) - (p+2)$$

$$\text{BIC} = -\frac{n}{2}\log(RSS/n) - \frac{\log(n)}{2}(p+2)$$

- **Key differences**:
  - AIC: Asymptotically efficient (minimizes prediction error)
  - BIC: More severe penalty for complexity ($\log(n)$ factor)
  - BIC: Consistent (selects true model with probability 1 as $n \to \infty$)
- **Computational challenge**: With p predictors, we have 2^p possible models to compare

Branch and Bound: Efficient Model Search

Instead of exhaustively fitting all 2^p models, branch and bound uses mathematical properties to eliminate entire groups of suboptimal models.

Theoretical Foundation

- Consider a set of models $\{M_1, M_2, \ldots, M_S\}$ which is a subset of all candidate models
- Let p_{min} be the minimum number of covariates in this subset, so $p_s \geq p_{\text{min}}$ for all s
- Define M_{sup} as the *supermodel* that contains every covariate appearing in at least one model from $\{M_1, M_2, \ldots, M_S\}$
- Key property: Each model M_s is nested within M_{sup}
- This implies: $RSS(M_{\text{sup}}) \leq RSS(M_s)$ for all s (the supermodel always fits at least as well)

$$\begin{aligned} AIC(M_s) &= -\frac{n}{2}\log(RSS(M_s)/n) - (p_s + 2) \\ &\leq -\frac{n}{2}\log(RSS(M_{\text{sup}})/n) - (p_{\text{min}} + 2) \end{aligned}$$

Branch and Bound: Implementation and Applications

- **Pruning the search tree**: If we find any model (outside our considered subset) with AIC better than our calculated upper bound, we can eliminate the entire group $\{M_1, M_2, \ldots, M_S\}$ without computing individual RSS values
- **Computational efficiency**:
  - Best case: Dramatic reduction in computation time
  - Worst case: Still requires fitting most models (similar to exhaustive search)
  - Practical limit: Generally feasible for problems with up to $30 - 35$ covariates
- **Extensions**:
  - Works equally well with BIC or other nested criteria
  - Can be adapted for generalized linear models
  - Implemented in R package `leaps` (function `regsubsets`)
  - Modern alternatives: `bestglm`, `glmulti` packages

Forward Selection: Algorithm

- **Forward selection** is a greedy stepwise procedure that builds a model sequentially
- **Algorithm**:
  1. Start with only an intercept (null model)
  2. Consider all models formed by adding a single covariate to the current model
  3. Select the model with the best criterion value (highest AIC/BIC or lowest CV error)
  4. If the best model improves the criterion, update the current model
  5. Repeat steps 2-4 until no additional covariate improves the criterion
- **Advantages**:
  - Computationally efficient: examines only $O(p^2)$ models vs. $O(2^p)$ for exhaustive search
  - Easy to implement and interpret
- **Limitations**:
  - Greedy algorithm: may miss optimal model due to variable interactions
  - Once a variable enters the model, it cannot be removed

Example: Forward Selection Process

Suppose we have 5 covariates $\{X_1, X_2, X_3, X_4, X_5\}$ under consideration

- **Step 1**: Start with intercept-only model, $AIC(\{b_0\}) = 2$
- **Step 2**: Evaluate all single-covariate additions
  - $AIC(\{b_0, X_1\}) = 1$ and $AIC(\{b_0, X_2\}) = 4$
  - $AIC(\{b_0, X_3\}) = 4$ and $AIC(\{b_0, X_4\}) = 6$ $\Rightarrow$ Highest AIC
  - $AIC(\{b_0, X_5\}) = 5$
- **Step 3**: Update current model to $\{b_0, X_4\}$ with $AIC = 6$
- **Step 4**: Evaluate all possible additions to $\{b_0, X_4\}$
  - $AIC(\{b_0, X_4, X_1\}) = 5$ and $AIC(\{b_0, X_4, X_2\}) = 8$
  - $AIC(\{b_0, X_4, X_3\}) = 7$ and $AIC(\{b_0, X_4, X_5\}) = 10$ $\Rightarrow$ Highest AIC
- **Step 5**: Update current model to $\{b_0, X_4, X_5\}$ with $AIC = 10$
- **Step 6**: Evaluate all possible additions to $\{b_0, X_4, X_5\}$
  - $AIC(\{b_0, X_4, X_5, X_1\}) = 9$
  - $AIC(\{b_0, X_4, X_5, X_2\}) = 9.5$
  - $AIC(\{b_0, X_4, X_5, X_3\}) = 9$
- **Step 7**: Terminate - No addition improves AIC
- **Final model selected**: $\{b_0, X_4, X_5\}$ with $AIC = 10$

Note: This example demonstrates how forward selection may not explore all possible interactions, potentially missing the global optimum.

Backward Elimination: Algorithm

- **Backward elimination** works in the opposite direction of forward selection
- **Algorithm**:
  1. Start with the full model (all covariates included)
  2. Consider all models formed by removing a single covariate from the current model
  3. Select the model with the best criterion value (highest AIC/BIC or lowest CV error)
  4. If the best model improves the criterion, update the current model
  5. Repeat steps 2-4 until no removal improves the criterion
- **Advantages**:
  - Considers interactions between variables from the beginning
  - Better when most variables are relevant but a few are noise
  - More computationally feasible than exhaustive search: examines $O(p^2)$ models
- **Limitations**:
  - Requires fitting the full model initially (problematic when $p > n$)
  - Still a greedy algorithm that may miss the optimal model

Example: Backward Elimination Process

Suppose we have 5 covariates $\{X_1, X_2, X_3, X_4, X_5\}$ under consideration

- **Step 1**: Start with full model, $AIC(\{b_0, X_1, X_2, X_3, X_4, X_5\}) = 6$
- **Step 2**: Evaluate all single-covariate removals
  - $AIC(\{b_0, X_2, X_3, X_4, X_5\}) = 6.5 \quad \Rightarrow$ Highest AIC
  - $AIC(\{b_0, X_1, X_3, X_4, X_5\}) = 6.2$ and $AIC(\{b_0, X_1, X_2, X_4, X_5\}) = 5$
  - $AIC(\{b_0, X_1, X_2, X_3, X_5\}) = 5.8$ and $AIC(\{b_0, X_1, X_2, X_3, X_4\}) = 4.9$
- **Step 3**: Update current model to $\{b_0, X_2, X_3, X_4, X_5\}$ with $AIC = 6.5$
- **Step 4**: Evaluate all possible removals from $\{b_0, X_2, X_3, X_4, X_5\}$
  - $AIC(\{b_0, X_3, X_4, X_5\}) = 6.7 \quad \Rightarrow$ Highest AIC
  - $AIC(\{b_0, X_2, X_4, X_5\}) = 5.5$ and $AIC(\{b_0, X_2, X_3, X_5\}) = 5.9$
  - $AIC(\{b_0, X_2, X_3, X_4\}) = 5.2$
- **Step 5**: Update current model to $\{b_0, X_3, X_4, X_5\}$ with $AIC = 6.7$
- **Step 6**: Evaluate all possible removals from $\{b_0, X_3, X_4, X_5\}$
  - $AIC(\{b_0, X_4, X_5\}) = 6.5$ and $AIC(\{b_0, X_3, X_5\}) = 6.1$
  - $AIC(\{b_0, X_3, X_4\}) = 5.5$
- **Step 7**: Terminate - No removal improves AIC
- **Final model selected**: $\{b_0, X_3, X_4, X_5\}$ with $AIC = 6.7$

Note: Compare with forward selection result $\{b_0, X_4, X_5\}$ - different algorithms may yield different models!

Stepwise Selection and Alternative Approaches

- **Stepwise selection**: Combines forward and backward approaches
  1. Start with an intercept-only model
  2. Add variables sequentially as in forward selection
  3. After each addition, check if removing any previously added variable improves the criterion
- **Comparison of stepwise methods**:
  - Forward selection: Works well when the final model has few variables (sparse)
  - Backward elimination: Better when most variables are relevant
  - Stepwise: More flexible but still not guaranteed to find the optimal model
  - All are computationally efficient but potentially suboptimal

Factors Affecting Model Selection Performance

When is model selection more challenging or more reliable?

- **Sample size** (n):
  - Larger sample sizes provide more information for accurate model selection
  - Small samples increase the risk of spurious correlations
- **Number of potential covariates** (p):
  - More covariates exponentially increase the search space (2^p possible models)
  - Higher dimensionality increases the chance of finding spurious predictors
- **Effect size** (magnitude of coefficients):
  - Larger effects are easier to detect
  - Small but important effects may be missed, especially in small samples
- **Correlation structure**:
  - Highly correlated predictors make variable selection more difficult
  - Multicollinearity can lead to unstable estimates and inconsistent selection

Simulation Study: Impact of Sample Size and Dimensionality

- Simulation setup:
  - Uncorrelated covariates: $X_j \sim N(0, 1)$
  - True model: $b_1 = b_2 = \ldots = b_5 = 0.25$, all others $b_6 = \ldots = b_p = 0$

AIC: Proportion Correct			
	p		
n	10	20	30
50	0.04	0.01	0.00
100	0.16	0.03	0.01
200	0.37	0.06	0.01
800	0.45	0.06	0.01

BIC: Proportion Correct			
	p		
n	10	20	30
50	0.01	0.00	0.00
100	0.07	0.05	0.04
200	0.47	0.38	0.28
800	0.95	0.85	0.76

Key observations:

- BIC outperforms AIC for model selection consistency (especially at larger sample sizes)
- Performance deteriorates as p increases (search space grows exponentially)
- Larger sample sizes dramatically improve selection accuracy

Simulation Study: Impact of Correlation Structure

- Simulation setup:
  - Comparing uncorrelated vs. correlated ($\text{cor}(X_j, X_k) = 0.25$) predictors
  - True model: $b_1 = b_2 = \ldots = b_5 = 0.25$, all others $b_6 = \ldots = b_p = 0$

Uncorrelated (BIC)			
	p		
n	10	20	30
50	0.01	0.00	0.00
100	0.07	0.05	0.04
200	0.47	0.38	0.28
800	0.95	0.85	0.76

Correlated (BIC)			
	p		
n	10	20	30
50	0.00	0.00	0.00
100	0.02	0.01	0.01
200	0.30	0.20	0.16
800	0.96	0.85	0.78

Key observations:

- Correlation among predictors makes model selection more challenging
- The effect is particularly pronounced at moderate sample sizes
- With very large samples, both scenarios perform similarly
- Multicollinearity remains a fundamental challenge for model selection

Simulation Study: Impact of Effect Size

- Simulation setup:
  - Uncorrelated covariates
  - True model: $b_1 = b_2 = \ldots = b_5 = \beta$, all others $b_6 = \ldots = b_p = 0$
  - Varying effect size: $\beta \in \{0.25, 0.5, 1.0\}$

$\beta = 0.25$			
	p		
n	10	20	30
50	0.01	0.00	0.00
100	0.07	0.05	0.04
200	0.47	0.38	0.28
800	0.95	0.85	0.76

$\beta = 0.5$			
	p		
n	10	20	30
50	0.45	0.18	0.08
100	0.81	0.53	0.32
200	0.89	0.67	0.53
800	0.95	0.86	0.80

$\beta = 1.0$			
	p		
n	10	20	30
50	0.70	0.32	0.13
100	0.81	0.57	0.35
200	0.88	0.69	0.54
800	0.95	0.86	0.78

Key observations:

- Larger effect sizes are much easier to detect, especially with small samples
- Even with small effects, large samples can achieve high accuracy
- Signal-to-noise ratio is a critical factor in model selection performance

Simulation Study: Model Size Selection

- Simulation setup:
  - Uncorrelated covariates
  - True model size = 5 ($b_1 = b_2 = \ldots = b_5 = 0.25$, all others zero)

AIC: Average Model Size			
	p		
n	10	20	30
50	4.21	7.07	11.77
100	5.15	7.22	9.67
200	5.74	7.52	9.52
800	5.76	7.35	9.14

BIC: Average Model Size			
	p		
n	10	20	30
50	2.41	3.44	5.16
100	3.22	3.79	4.13
200	4.53	4.76	5.03
800	5.05	5.16	5.28

Key observations:

- AIC tends to select larger models (overfitting)
- BIC is more conservative and closer to the true model size (especially at large n)
- With small samples, BIC tends to underfit (models too small)
- With large samples, BIC converges to the correct model size

Post-Selection Inference: The Challenge

- **The statistical issue**: Standard inference is invalid after model selection
  - Model selection procedures use the data to choose variables
  - This invalidates the sampling distributions of test statistics
  - Results in inflated significance and narrower confidence intervals than warranted
- **Analogy**: The "Instagram effect"
  - Your friend's Instagram feed shows only their best moments
  - Comparing your average life to their curated feed is an unfair comparison
  - Similarly, examining only the "selected" variables gives a biased view
- **Formal issue**: Conditioning on selection events changes the sampling distribution
  - Variables are selected because they appear significant in the sample
  - This creates a "winner's curse" - effects appear stronger than they truly are
  - Standard p-values and confidence intervals no longer have their claimed properties

Post-Selection Inference: Valid Approaches

- **Data splitting**: The simplest valid approach
  1. Split data into two independent parts: training and test sets (typically 50/50)
  2. Use training data exclusively for model selection (exploratory analysis, stepwise procedures, etc.)
  3. Use test data exclusively for inference (hypothesis tests, confidence intervals)
  4. Test set inference is valid because it's independent of selection process
- **Limitations of data splitting**:
  - Reduced power due to smaller sample size for both tasks
  - Confidence intervals are wider due to using only $n/2$ observations
  - Trade-off: Is eliminating irrelevant variables worth the cost of using half the data?
- **Advanced alternatives** (beyond scope of this lecture):
  - Selective inference methods that adjust for selection
  - Multiple testing corrections
  - Bootstrapping and subsampling approaches

Key Takeaways

- Model selection is fundamentally about finding the optimal trade-off between bias and variance
- Information criteria (AIC, BIC) provide theoretically justified approaches to this trade-off
- The computational challenge grows exponentially with the number of predictors (2^p possible models)
- Branch and bound offers an efficient exact search strategy when feasible
- For high-dimensional problems ($p \gg 30$), consider:
  - Stepwise procedures (computationally efficient but potentially suboptimal)
  - Regularization methods (lasso, elastic net)
  - Domain knowledge to reduce the initial variable set
- Remember: The "best" model depends on your objective (prediction vs. inference)

Key Takeaways

- **Model selection fundamentals**:
  - Finding the optimal trade-off between bias and variance
  - Balancing model complexity with generalization performance
- **Selection methods**:
  - Exhaustive search with branch and bound (when computationally feasible)
  - Forward/backward/stepwise procedures (efficient but potentially suboptimal)
  - Regularization methods (lasso, ridge, elastic net) for high-dimensional problems
- **Performance factors**:
  - Sample size, number of predictors, effect size, and correlation structure
  - BIC tends to select more parsimonious models than AIC
  - BIC is consistent for model selection with large samples
- **Valid inference**:
  - Standard inference is invalid after data-driven model selection
  - Data splitting provides valid inference at the cost of reduced power
  - The "best" approach depends on your objective (prediction vs. inference)