Lecture 17: Model Selection Algorithms

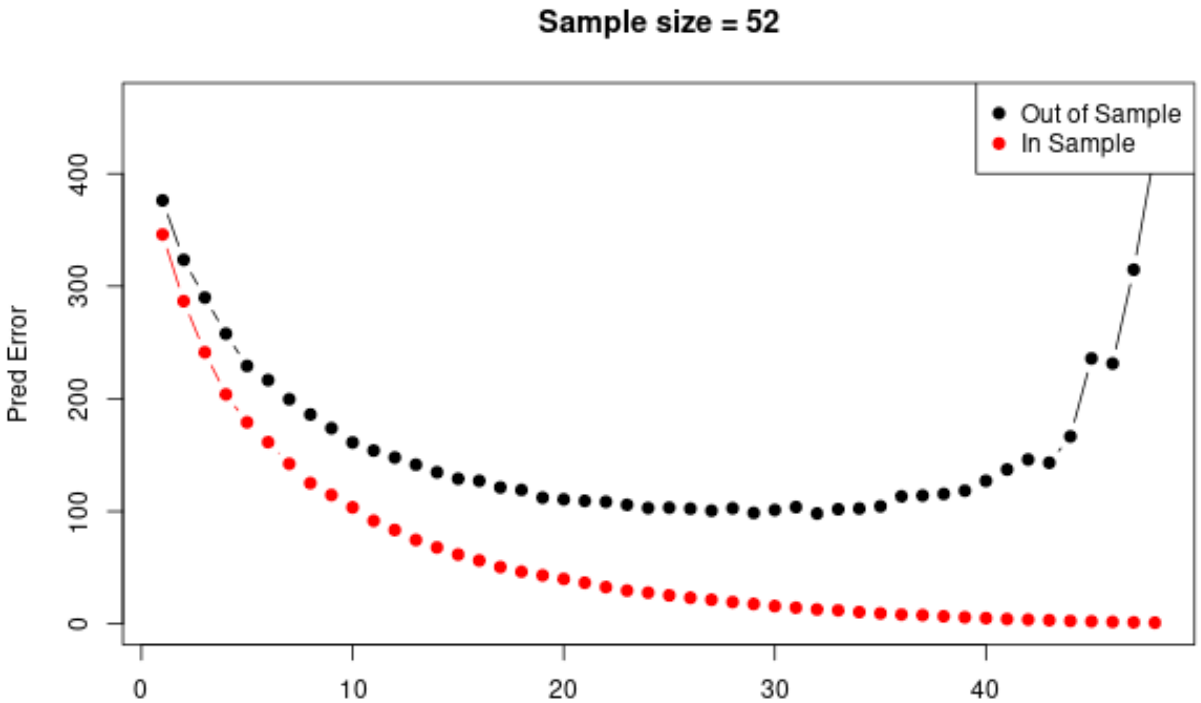
Module 5: part 2

Spring 2025

Model Selection Recap

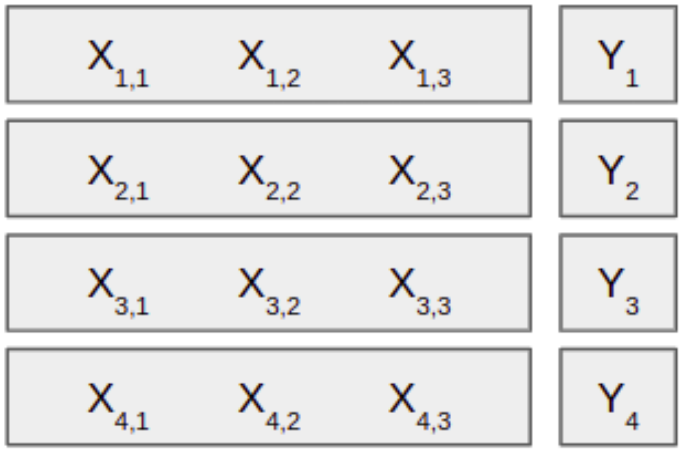
Recap

- When selecting a model which will perform well on new data, we can't just consider how well it performs on the data it is fit to
- Choosing the level of complexity involves a bias vs variance trade-off
- Including more "truly relevant" covariates can improve prediction (decrease bias)
- Including more covariates means $\hat{b}$ are generally not estimated as well (increase variance)



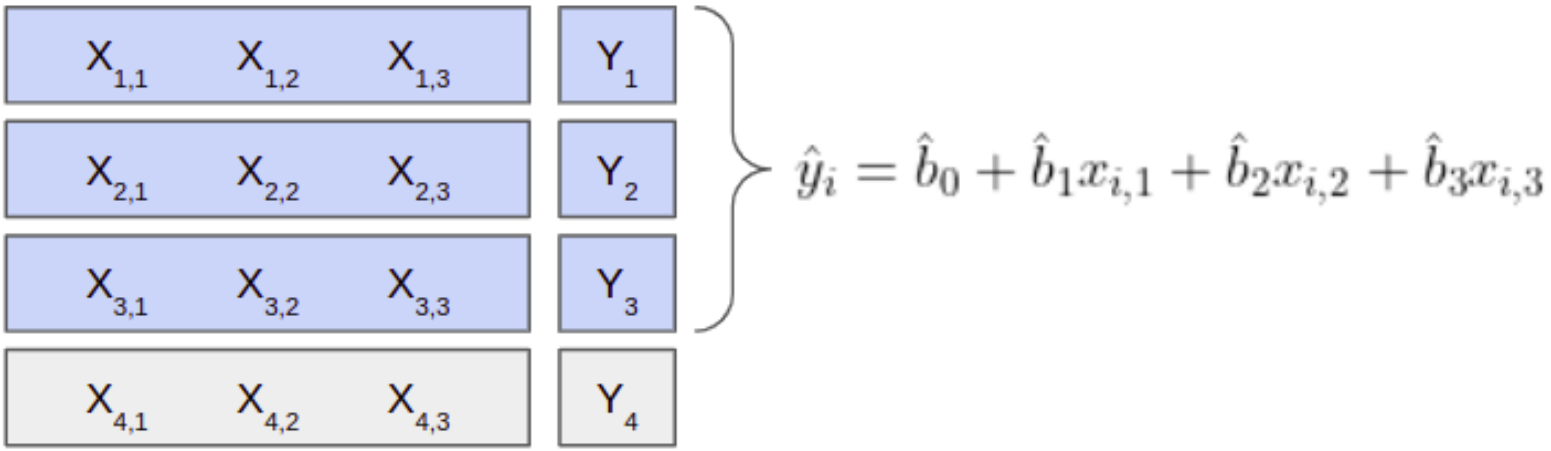
Leave one out cross validation

Cross validation gives us a way to estimate how a model will perform on "new data"



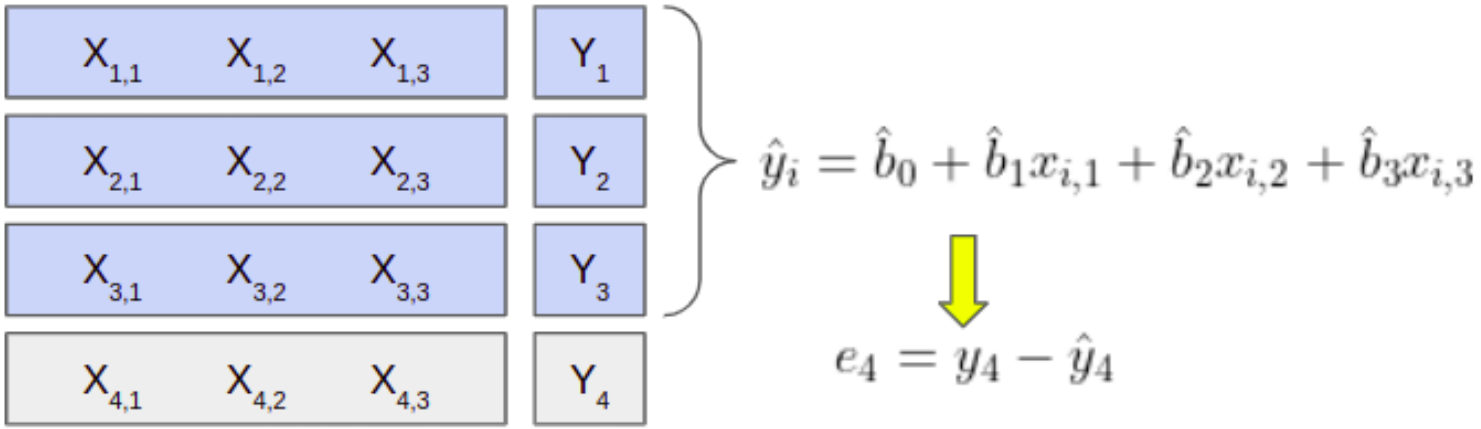
Leave one out cross validation

Cross validation gives us a way to estimate how a model will perform on "new data"



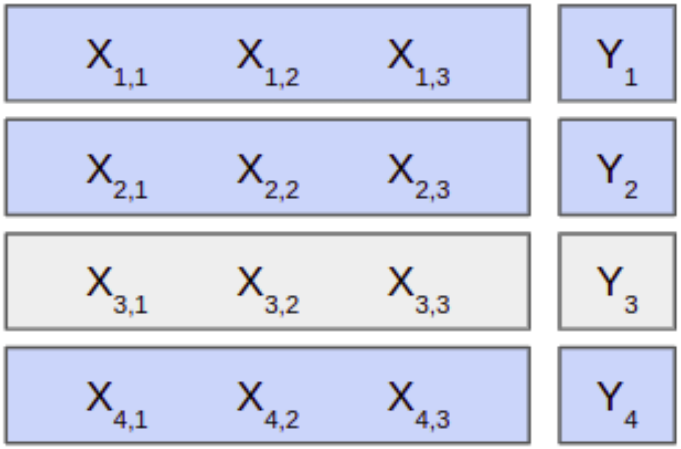
Leave one out cross validation

Cross validation gives us a way to estimate how a model will perform on "new data"



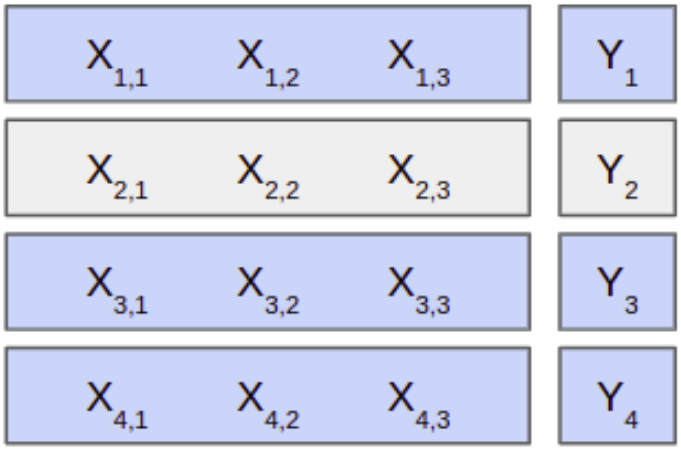
Leave one out cross validation

Cross validation gives us a way to estimate how a model will perform on "new data"



Leave one out cross validation

Cross validation gives us a way to estimate how a model will perform on "new data"



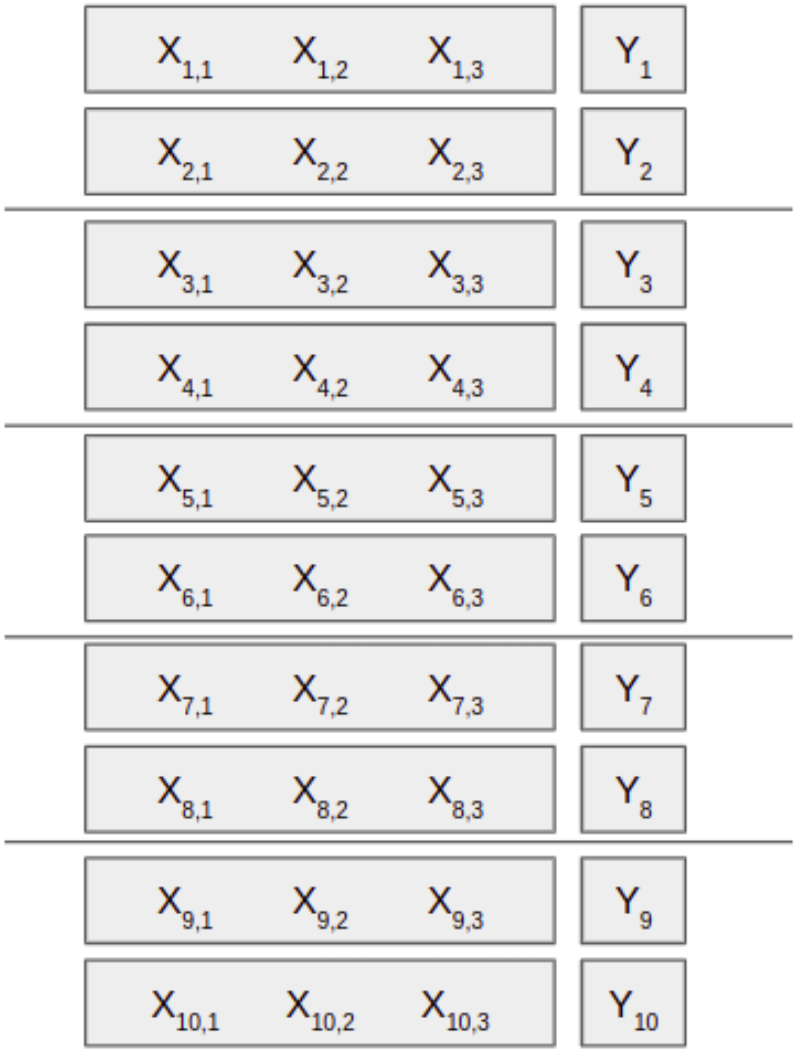
Leave one out

- This is called **leave one out cross-validation**
- The cross validation score is:

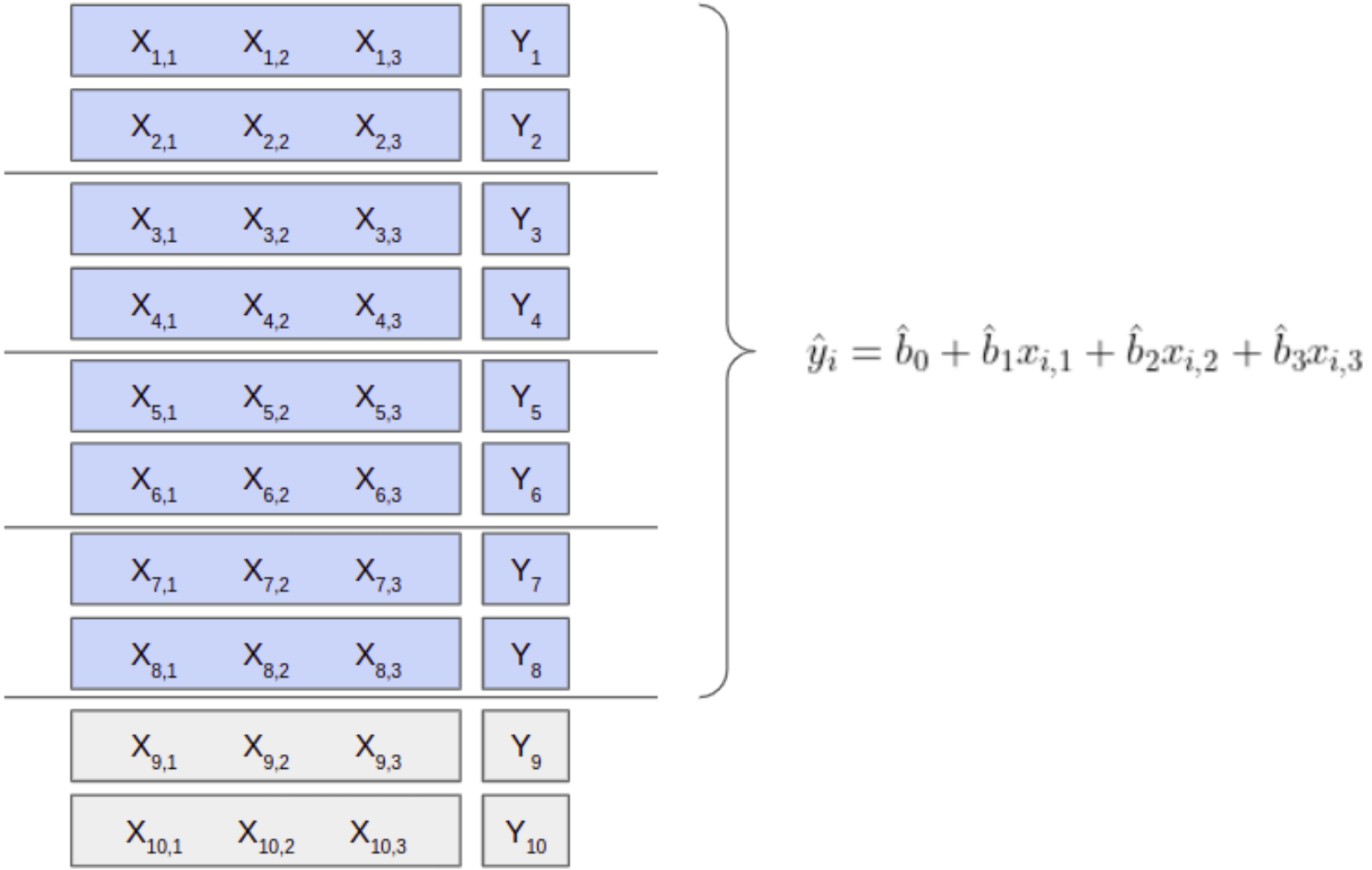
$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

- Not quite RSS since the model we use for $\hat{y}_i$ is different each time
- The model that produces $\hat{y}_i$ was fit without using the ith observation
- This is an unbiased estimator of generalization error; i.e., how well your model will generalize to new data
- Depending on how computationally intensive your model is, and how many data points you have, this may be computationally infeasible

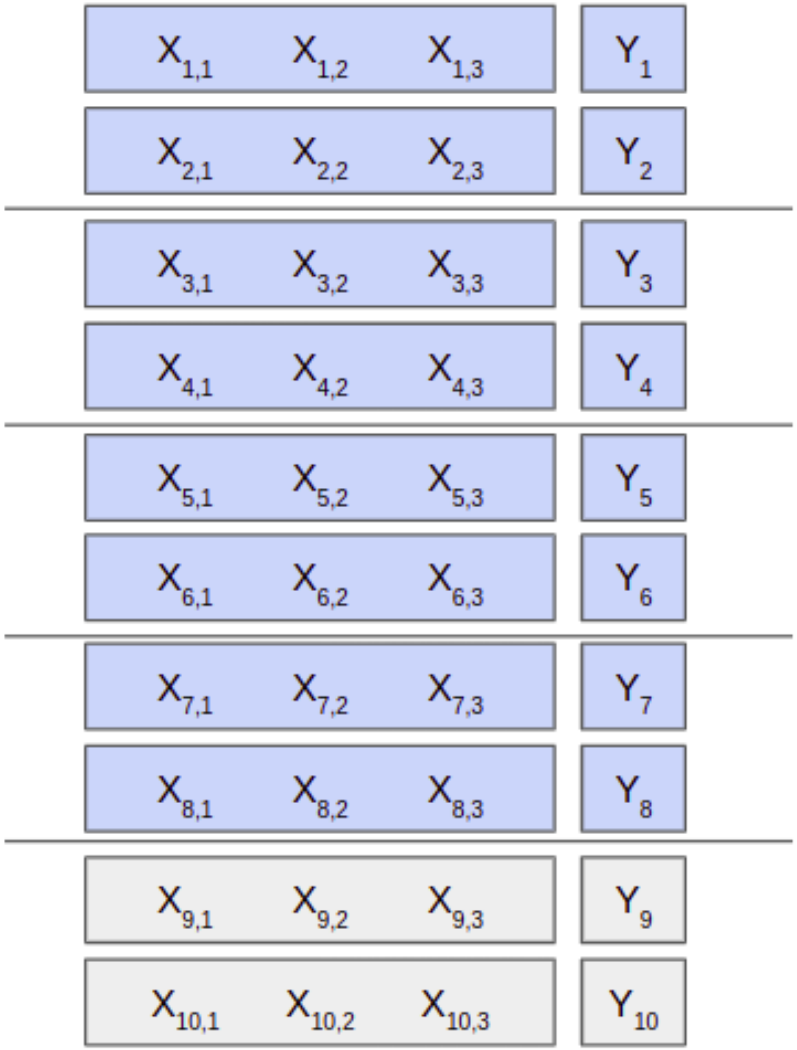
K-fold cross validation



K-fold cross validation



K-fold cross validation



$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + \hat{b}_3 x_{i,3}$$

$$e_9 = y_9 - \hat{y}_9$$
$$e_{10} = y_{10} - \hat{y}_{10}$$

Recap

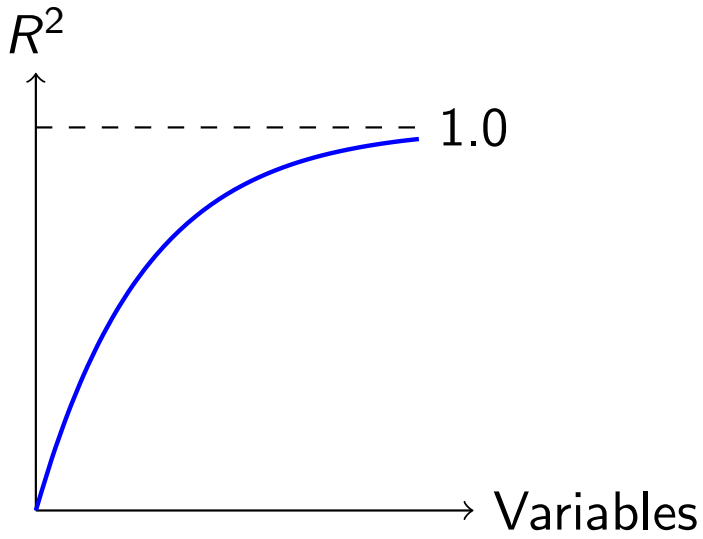
- Several ways to select a training set (data used to estimate parameters of model) and test set (data used to evaluate performance)
  - Leave One Out
  - K-fold CV
- Because the "best" model selection will also depend on the number of observations you have, Leave One Out will generally be best, but also has the highest computational cost
- Use approximations instead which only require fitting the model once
- Models with the lower cross validation scores are better

The Model Selection Problem

- **Core challenge:** Finding the optimal balance between model complexity and predictive performance
- **Overfitting:** Complex models may fit training data perfectly but perform poorly on new data
- **Underfitting:** Simple models may miss important patterns in the data

Penalized Scores for Model Selection

R^2 and Its Limitations



$$R^2 = 1 - \frac{\text{RSS}}{\sum_i (y_i - \bar{y})^2}$$

- Measures how well model fits *training data*
- **Key issue:** R^2 *always increases* with additional predictors
- This incentivizes unnecessarily complex models

Adjusted R^2: A First Approach

Adjusted R^2 Formula

$$R^2_{\text{adj}} = 1 - \frac{\frac{n}{n-p-1}\text{RSS}}{\sum_i (y_i - \bar{y})^2}$$

- **Improvement:** Penalizes for model complexity via $\frac{n}{n-p-1}$ factor
- p = number of predictors in the model
- **Theoretical limitation:** Penalty is too weak
- Often selects models that are still too complex
- Better than regular R^2, but inferior to AIC/BIC/CV for model selection

Akaike Information Criterion (AIC)

AIC for Linear Regression with Gaussian Errors

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

- p = number of coefficients (excluding intercept)
- **Interpretation:** Higher AIC values indicate better models
- **Note:** Different resources may multiply AIC by -2 or 2, changing interpretation
- **Theoretical foundation:** AIC is an unbiased estimator of expected out-of-sample prediction error
- **Computational advantage:** Fast approximation of cross-validation

AIC and Cross-Validation: Theoretical Properties

- As sample size $n \to \infty$:
  - Both AIC and CV select models with near-optimal generalization error
  - Both tend to select slightly more complex models than the theoretical optimum
  - This happens because more complex models have higher variability in their error estimates

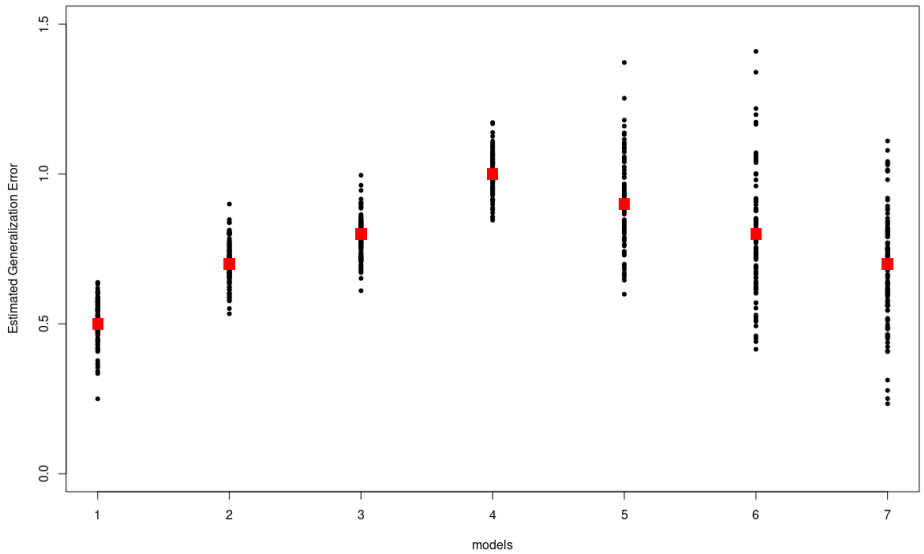


Figure: AIC vs. True Error

Bayesian Information Criterion (BIC)

BIC for Linear Regression with Gaussian Errors

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

- **Key difference from AIC:** Complexity penalty grows with sample size n
- **Comparison:**
  - AIC penalty: $(p+2)$
  - BIC penalty: $\frac{\log(n)}{2}(p+2)$
- **Consequence:** BIC favors simpler models than AIC, especially with large datasets
- **Theoretical property:** BIC is consistent - will select the true model as $n \to \infty$ (if true model exists in candidate set)

Comparing Model Selection Criteria

Criterion	Complexity Penalty	Consistency	Prediction	Computation
R^2	None	No	Poor	Fast
Adjusted R^2	Weak	No	Moderate	Fast
AIC	Moderate	No	Good	Fast
BIC	Strong	Yes	Good*	Fast
Cross-Validation	Flexible	No	Excellent	Slow

Table: *BIC may underperform AIC for prediction with finite samples

- **No perfect criterion** - choice depends on goals:
  - **Pure prediction:** Cross-validation or AIC
  - **Finding true model:** BIC (if you believe a true model exists)

Computational Strategies for Model Selection

The Computational Challenge

Number of Possible Models

With p potential predictors, we have 2^p possible models

Number of predictors	Possible models	Feasibility
5	$2^5 = 32$	Trivial
10	$2^{10} = 1,024$	Easy
20	$2^{20} = 1,048,576$	Challenging
30	$2^{30} \approx 1$ billion	Impractical
60	$2^{60} \approx$ number of sand grains on Earth	Impossible

We need efficient search strategies!

Branch and Bound Algorithm

Key Insight

We can eliminate entire groups of models without evaluating each one individually

- **Consider** a set of models $\{M_1, M_2, \ldots, M_S\}$
- Let M_{sup} be a model containing all covariates that appear in at least one model in our set
- Let p_{min} be the number of covariates in the smallest model in our set
- For all models M_s in our set: $\text{RSS}(M_s) \geq \text{RSS}(M_{\text{sup}})$

Upper Bound for AIC

$$\text{AIC}(M_s) \leq -\frac{n}{2} \log\left(\frac{\text{RSS}(M_{\text{sup}})}{n}\right) - (p_{\text{min}} + 2)$$

Branch and Bound: Practical Application

- **Key advantage:** If we find any model with AIC exceeding our upper bound, we can eliminate all models in our set without evaluating them
- **Effectiveness varies:**
  - Best case: Drastically reduces computation
  - Worst case: Similar to exhaustive search
- **Practical limit:** Up to 30 predictors
- **Software implementation:** `leaps` package in R
- **Applicability:** Works with both AIC and BIC
- **Limitation:** Specialized for linear regression (uses computational tricks specific to linear models)

Stepwise Selection Methods

When Exhaustive Search Is Infeasible

Stepwise methods provide heuristic approaches that:

- Are computationally efficient
- Work for various model types (not just linear regression)
- Usually find good (though not guaranteed optimal) solutions

Two Main Approaches

- **Forward selection:** Start simple, add variables
- **Backward selection:** Start complex, remove variables

Forward Selection Algorithm

How Forward Selection Works

A greedy algorithm that builds a model iteratively:

- Start with an empty model (no predictors)
- At each step, add the most significant variable
- Continue until no remaining variables improve the model

Algorithm Steps

1. Start with an intercept-only model
2. For each candidate variable not in the model:
   - Fit model with the variable added
   - Compute selection criterion (e.g., p-value, AIC, BIC)
3. Add variable that most improves the criterion (if significant)
4. Repeat steps 2-3 until no further improvement

Forward Selection Algorithm

Advantages & Limitations

Advantages

- Computationally efficient
- Easy to implement
- Interpretable process

Limitations

- May miss optimal subset
- Ignores multicollinearity
- Cannot remove variables once added

Backward Selection Algorithm

How Backward Selection Works

A pruning algorithm that simplifies a model iteratively:

- Start with the full model (all predictors)
- At each step, remove the least significant variable
- Continue until all remaining variables are significant

Algorithm Steps

1. Start with all variables in the model
2. For each variable currently in the model:
   - Compute significance measure (e.g., p-value, t-statistic)
   - Identify least significant variable
3. Remove least significant variable if below threshold
4. Repeat steps 2-3 until all variables meet significance criterion

Backward Selection Algorithm

Advantages & Limitations

Advantages

- Considers interactions between all variables initially
- May detect effects hidden by confounders
- Works well with many predictor variables

Limitations

- Requires fitting full model first
- Computationally intensive with many variables
- Cannot reconsider variables once removed

Computational Complexity of Stepwise Methods

- **Forward selection:**
  - At step i: evaluate $(p - i)$ models
  - Maximum steps: p
  - Total evaluations: $\mathcal{O}(p^2)$
- **Backward selection:**
  - At step i: evaluate i models
  - Maximum steps: p
  - Total evaluations: $\mathcal{O}(p^2)$
- **Compared to exhaustive:** $\mathcal{O}(p^2)$ vs. $\mathcal{O}(2^p)$
- **Important note:** Forward and backward selection may yield different models
- **Hybrid approaches:** Can allow both addition and removal at each step

Summary: Model Selection Best Practices

Selection Criteria

- **Avoid:** R^2 (no penalty)
- **Adequate:** Adjusted R^2 (weak penalty)
- **Better:** AIC/BIC (theoretically grounded)
- **Best:** Cross-validation (directly measures generalization)

Search Strategies

- **Small p (< 30):** Branch and bound
- **Large p:** Stepwise methods
- **Modern alternatives:** LASSO, Elastic Net (next lecture)

Final Recommendations

- Always validate your final model on holdout data
- Consider your goals: prediction vs. explanation
- Incorporate domain knowledge when possible
- Remember model selection introduces additional uncertainty