Lecture 16: Cross-validation

Module 5: part 1

Spring 2025

Logistics

- Module 4 assessment posted, due March 23rd 11:59
- Starting Module 5 on Model Selection
- How to assess model's predictive ability

Model Selection

Regression Analysis: Two Fundamental Purposes

1. **Describe Scientific Processes**
   - Uncover relationships between variables
   - Identify potential causal mechanisms
   - Test specific hypotheses
   - Focus: Understanding

2. **Build Predictive Models**
   - Forecast future outcomes
   - Identify associations without causation
   - Optimize predictive accuracy
   - Focus: Performance

Why Do We Fit Regressions?

Scientific Understanding

- Describe underlying mechanisms
  - *Example:* What factors drive a company's share price changes?
  - Which variables, if manipulated, would change share price?
  - Requires careful consideration of causal assumptions

Why Do We Fit Regressions?

Scientific Understanding

- Describe underlying mechanisms
  - *Example:* What factors drive a company's share price changes?
  - Which variables, if manipulated, would change share price?
  - Requires careful consideration of causal assumptions

Prediction

- Create models that forecast outcomes
  - *Example:* Will this specific company's share price rise?
  - No explicit causal model required
  - Identifies associations that improve prediction accuracy
  - Can be developed purely from observed data patterns

Regression for Scientific Understanding

When using regression to describe underlying processes:

- We assume data generation follows our specified model
- We aim to estimate specific model parameters
- Useful approximation even when reality is more complex
- Causal interpretations require strong assumptions:
  - No unmeasured confounders
  - Correct functional form
  - Proper temporal ordering

Regression for Scientific Understanding

When using regression to describe underlying processes:

- We assume data generation follows our specified model
- We aim to estimate specific model parameters
- Useful approximation even when reality is more complex
- Causal interpretations require strong assumptions:
  - No unmeasured confounders
  - Correct functional form
  - Proper temporal ordering

Model Selection for Scientific Understanding:

- Helps identify parsimonious explanatory models
- F-tests can compare nested models for specific hypotheses
- Information criteria (AIC, BIC) help approximate "true model"

Regression for Prediction

Core Principles:

- Focus on predicting *new* data, not fitting existing data
- No assumption of causality required
- Goal: minimize prediction error on unseen observations
- Trade-off: model complexity vs. generalizability

Regression for Prediction

Core Principles:

- Focus on predicting *new* data, not fitting existing data
- No assumption of causality required
- Goal: minimize prediction error on unseen observations
- Trade-off: model complexity vs. generalizability

Model Selection for Prediction:

- RSS alone is insufficient: $\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$
- Complex models always reduce RSS but may overfit
- Better approaches:
  - Cross-validation (train/test splits)
  - Regularization (Ridge, LASSO)
  - Information criteria with complexity penalties
- Optimal model complexity depends on sample size and noise

Bias-Variance Trade-off

- In statistics, the bias-variance trade-off is a fundamental constraint
- Two competing reasons for why our predictive model may not perform well:
  - **Bias**: We don't include the right covariates in our model, so the model we are fitting is inherently wrong (but may still be useful)
  - **Variance**: The model we are fitting is very complex relative to the amount of data we have, so our estimated parameters are not very precise

Expected Prediction Error:

$$\mathbb{E}[(Y - \hat{f}(X))^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

As model complexity increases:

- Bias decreases (better fit to training data)
- Variance increases (more sensitive to sampling)
- Total error follows U-shaped curve
- Goal: Find optimal complexity that minimizes total error

Bias-Variance Trade-off

Bias and variance in making cakes:

Figure: High Bias, Low Variance

Figure: Low Bias, High Variance

- **Betty Crocker**: Simple recipe, consistent results but not gourmet (high bias, low variance)
- **Home made**: Complex recipe, potentially amazing but more variable results (low bias, high variance)

Bias-Variance Example

Suppose data are generated by the model for $i = 1, \ldots, n$:

$$Y_i = b_0 + \sum_{k=1}^{50} b_k X_{i,k} + \varepsilon_i$$

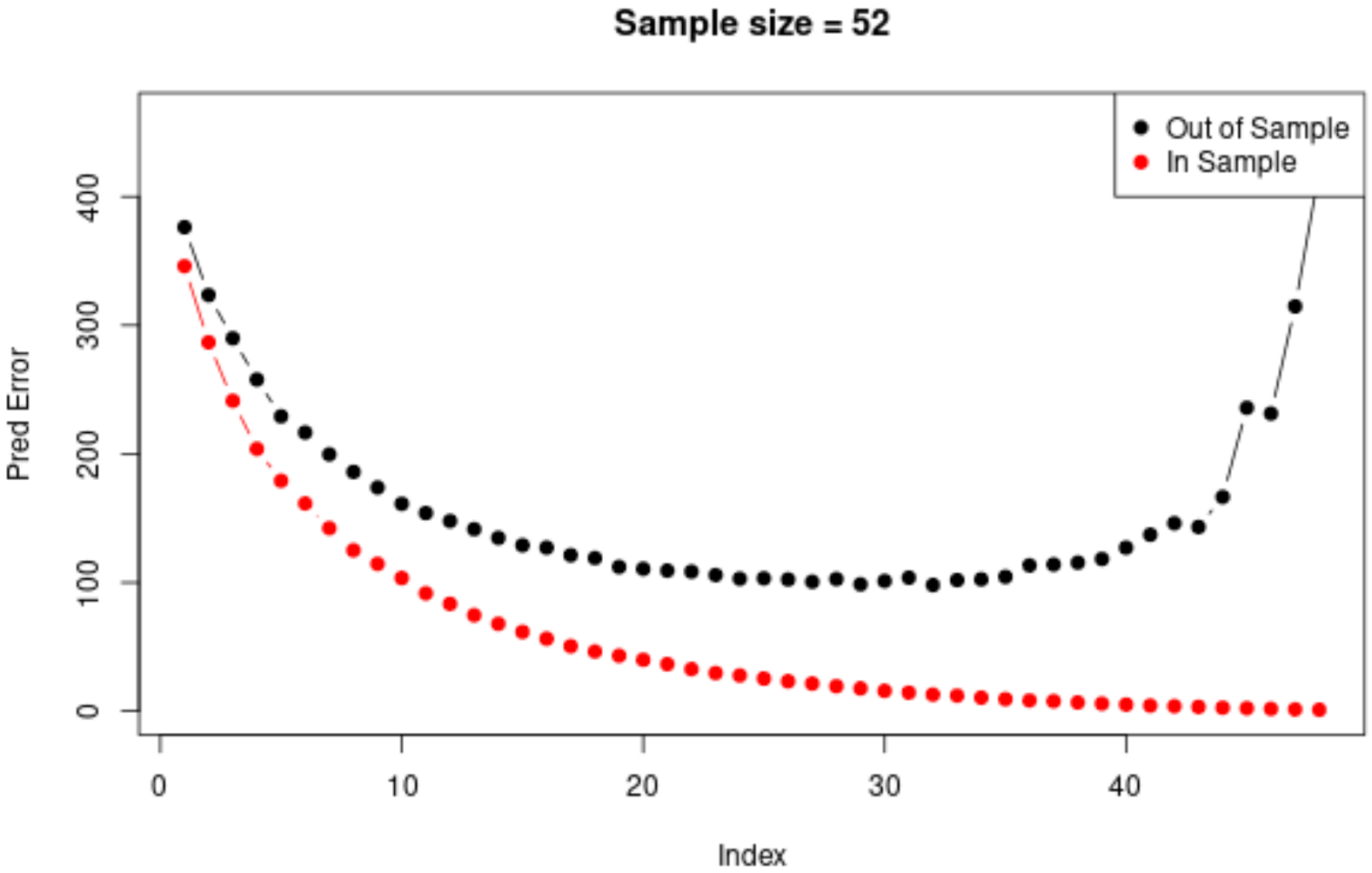
$$\mathrm{cov}(X_{i,k}, X_{i,l}) = 1/5 \qquad \mathrm{var}(X_{i,k}) = 1$$

$$\varepsilon_i \sim N(0, 16)$$

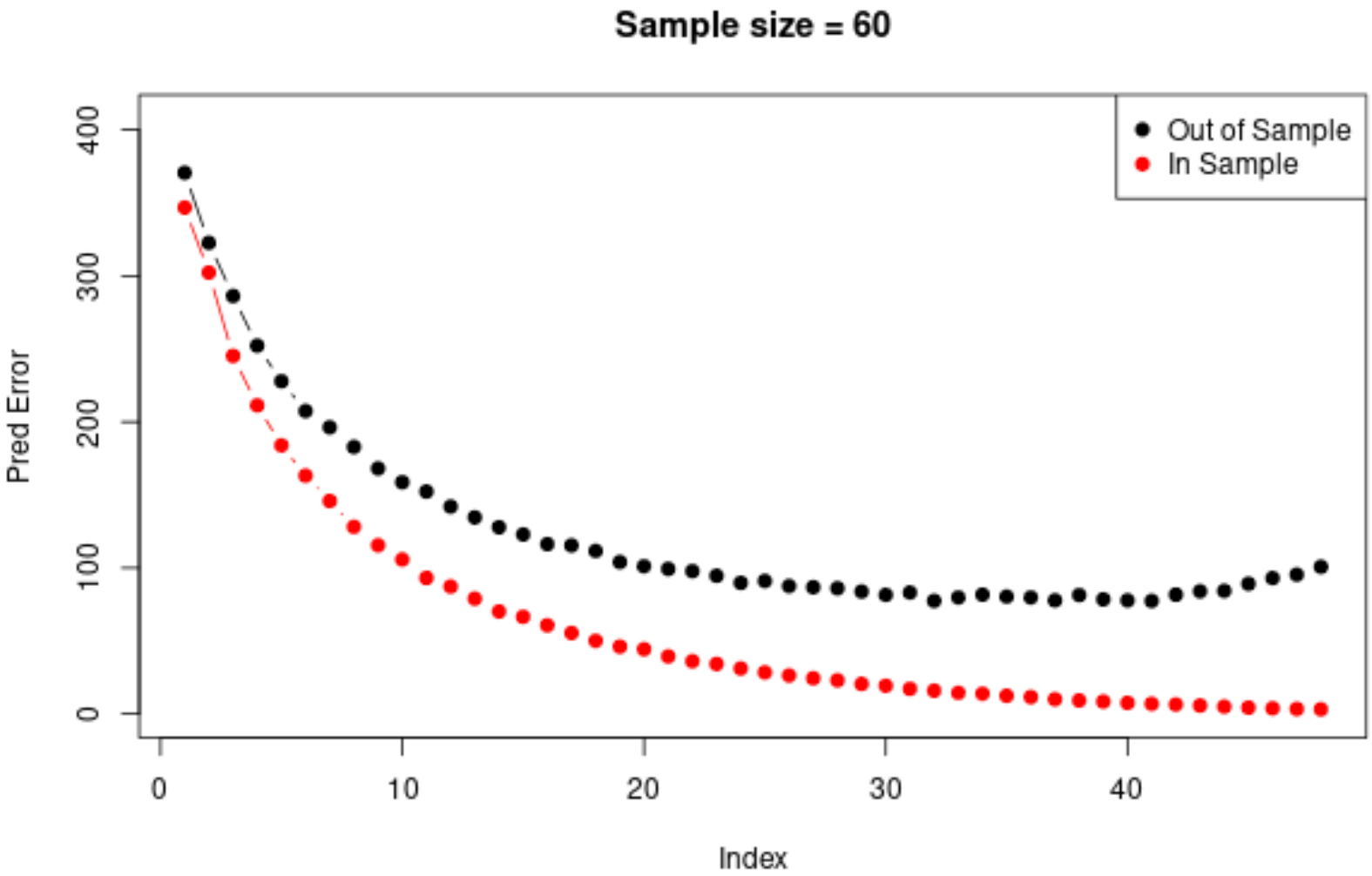
Experiment:

1. For $m = 1, 2, \ldots 50$
2. Fit a model using only the first m predictors (omitting the rest)
3. When m is smaller, the model we fit is "more biased"
4. Using estimated $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$, predict $\hat{y}_i$ for new data
5. Measure RSS $(y_i - \hat{y}_i)^2$ on **observed** data
6. Measure prediction error $(y_i - \hat{y}_i)^2$ on **new** data

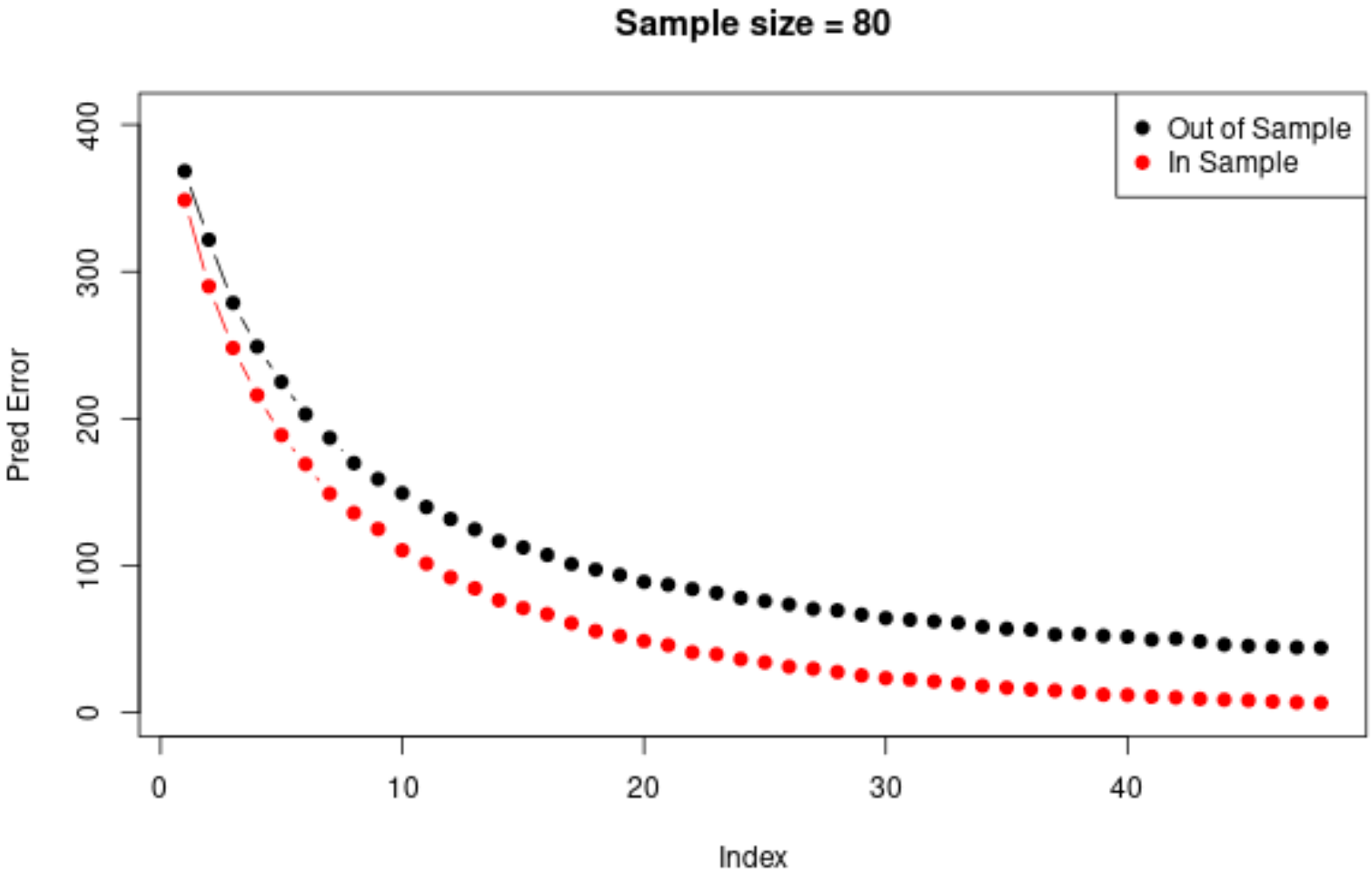
Bias-Variance Example Results



Bias-Variance Example Results



Bias-Variance Example Results



Models and Parameters

Structural Decisions

Model Selection: Which variables do we include?

- Linear vs. non-linear relationships
- Interaction terms
- Polynomial terms
- Variable transformations

Estimation

Parameter Estimation: What are the coefficients for included variables?

- Ordinary least squares (OLS)
- Maximum likelihood estimation (MLE)
- Regularized approaches (Ridge, LASSO)
- Robust estimators

Class Discussion

Scientific Understanding:

- What are examples in your work where you need a causal model?
- What assumptions are critical for your analysis?
- How do you validate these assumptions?

Prediction:

- What are examples where you need predictive models?
- How do you evaluate predictive performance?
- What techniques do you use to avoid overfitting?

Your Turn: Identify a research question where the distinction between causal and predictive modeling would be crucial.

Cross-Validation: Assessing Predictive Performance

The Challenge of Generalization

Key Question in Predictive Modeling:

How well will our model perform on **new, unseen data**?

- Training error is an overly optimistic measure of predictive performance
- We need methods to estimate **generalization error** accurately
- Cross-validation provides a principled approach to this problem

Generalization Error: The Goal of Predictive Modeling

We want to minimize the expected prediction error on **new observations**:

- Define a loss function $L(y_i, \hat{y}_i)$ that quantifies prediction error
  - For continuous outcomes: $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ (squared error)
  - For categorical outcomes: $L(y_i, \hat{y}_i) = \mathbf{1}_{\{y_i \neq \hat{y}_i\}}$ (misclassification)
- Our goal: Minimize $\mathbb{E}[L(y_{\text{new}}, \hat{y}_{\text{new}})]$

- The fundamental challenge: We don't have access to the **true distribution** of future data
- Solution: Use existing data to **simulate** the process of predicting new observations

Cross-Validation: Core Concept

The Central Idea of Cross-Validation

Repeatedly split data into training and testing sets to estimate how well models will generalize to new data.

The general procedure:

1. Set aside a portion of your data as a "test set"
2. Train your model on the remaining data (the "training set")
3. Evaluate model performance on the test set
4. Repeat this process with different train/test splits

Key benefit: Every observation serves as both training and test data across iterations

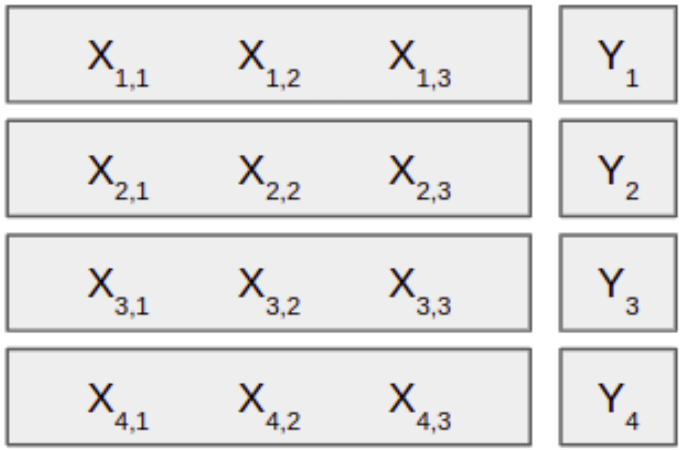
Leave-One-Out Cross-Validation (LOOCV)

Procedure:

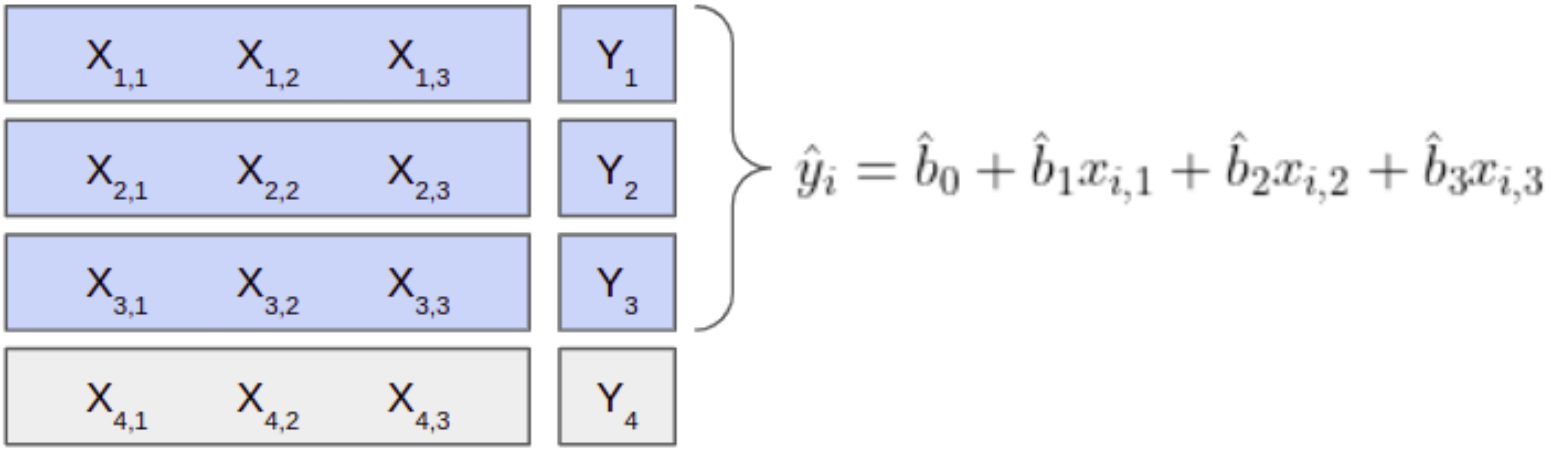
1. For each observation i in your dataset:
   - Train model on all observations *except* i
   - Predict $\hat{y}_i$ for the held-out observation
   - Calculate error: $e_i = y_i - \hat{y}_i$
2. The LOOCV estimate of prediction error is:

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

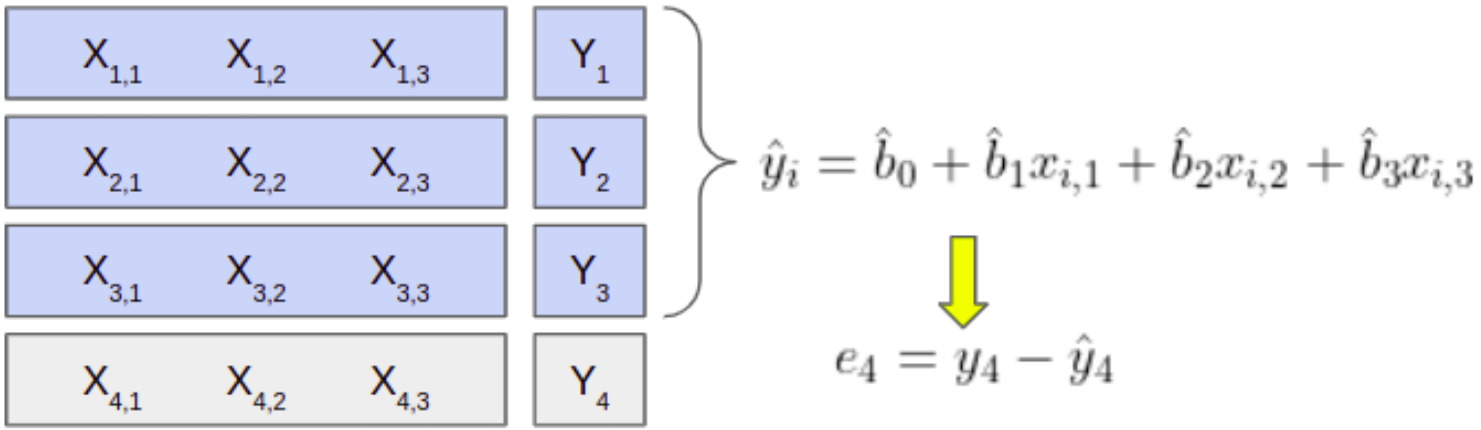
Leave-One-Out Cross-Validation: Visualization



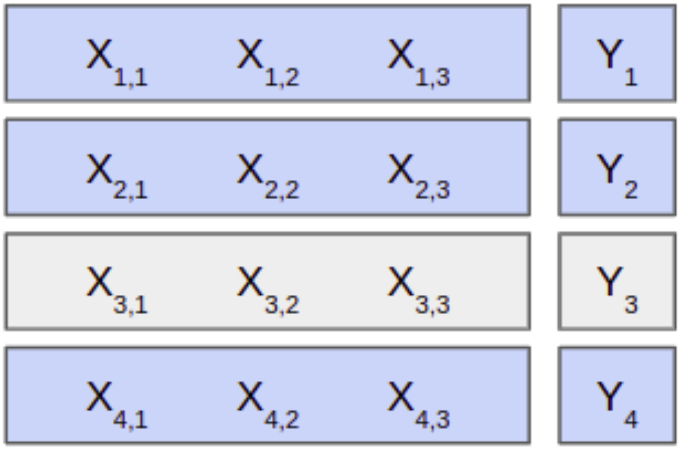
Leave-One-Out Cross-Validation: Visualization



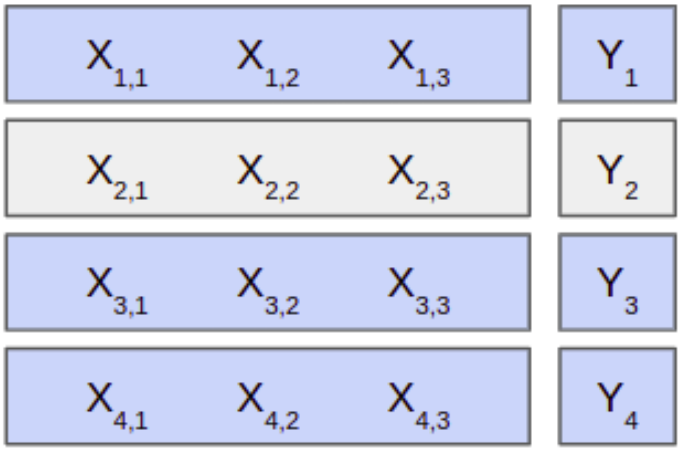
Leave-One-Out Cross-Validation: Visualization



Leave-One-Out Cross-Validation: Visualization



Leave-One-Out Cross-Validation: Visualization



Properties of LOOCV

Advantages:

- Nearly unbiased estimate of generalization error
- Maximizes the size of the training set
- Deterministic (no randomness in the splits)
- Especially useful for small datasets

Disadvantages:

- Computationally expensive (fit model n times)
- High variance in the estimate
- Test sets consist of single observations
- Less effective for model selection

- The LOOCV error differs from RSS because each $\hat{y}_i$ is computed from a different model
- For linear regression, there is a computational shortcut:

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

where h_{ii} is the ith diagonal element of the "hat matrix"

Train-Test Split: The Simplest Approach

Procedure:

1. Randomly split data into training set (e.g., 70-80%) and test set (20-30%)
2. Train model using only the training set
3. Evaluate performance on the held-out test set

Advantages:

- Computationally efficient (fit model only once)
- Simple to implement
- Mimics real-world application scenario

Disadvantages:

- Results depend on the specific random split
- Less efficient use of data (some observations never train the model)
- Higher variance in the error estimate

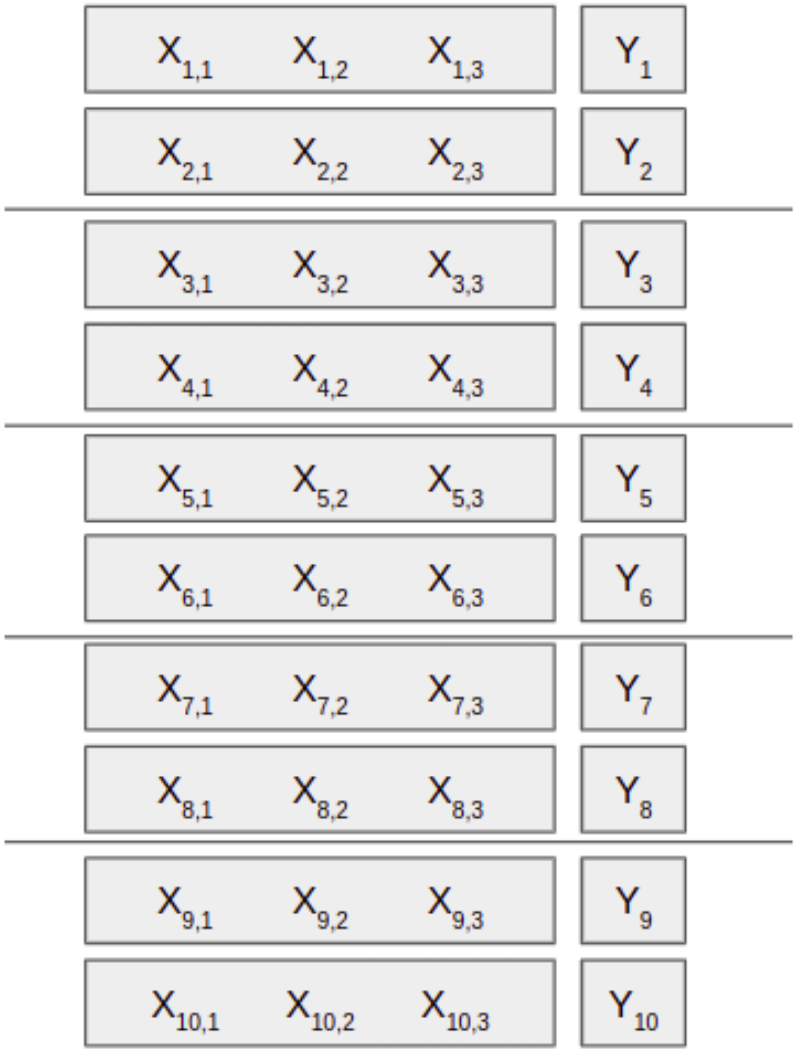
K-Fold Cross-Validation: The Practical Solution

Procedure:

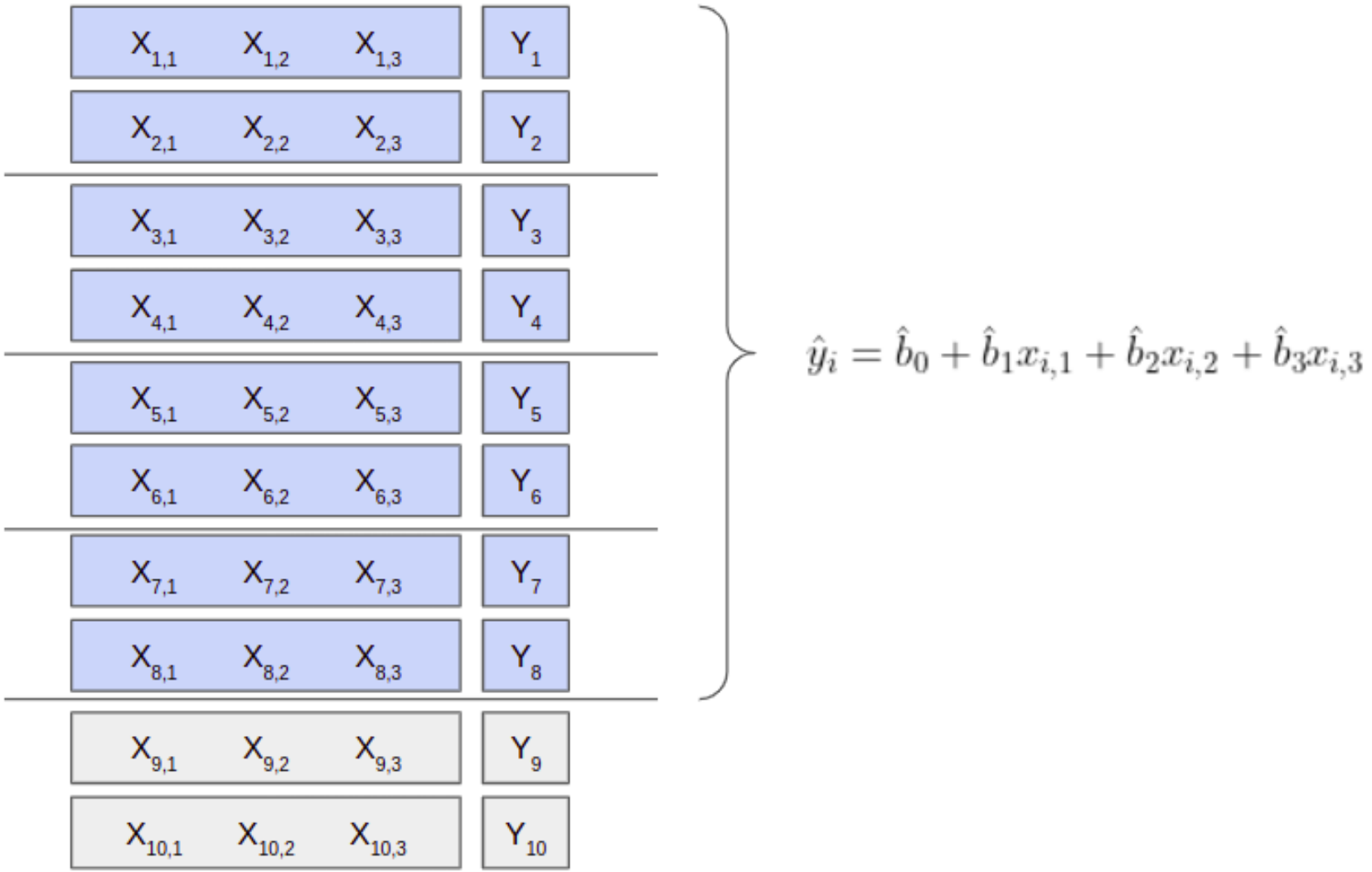
1. Randomly divide data into K equal-sized folds (typically $K = 5$ or $K = 10$)
2. For each fold $k = 1, 2, \ldots, K$:
   - Train model on all folds except fold k
   - Evaluate performance on fold k
3. Average the error across all K folds:

$$\text{CV}_{(K)} = \frac{1}{K} \sum_{k=1}^{K} \text{Error}_k$$

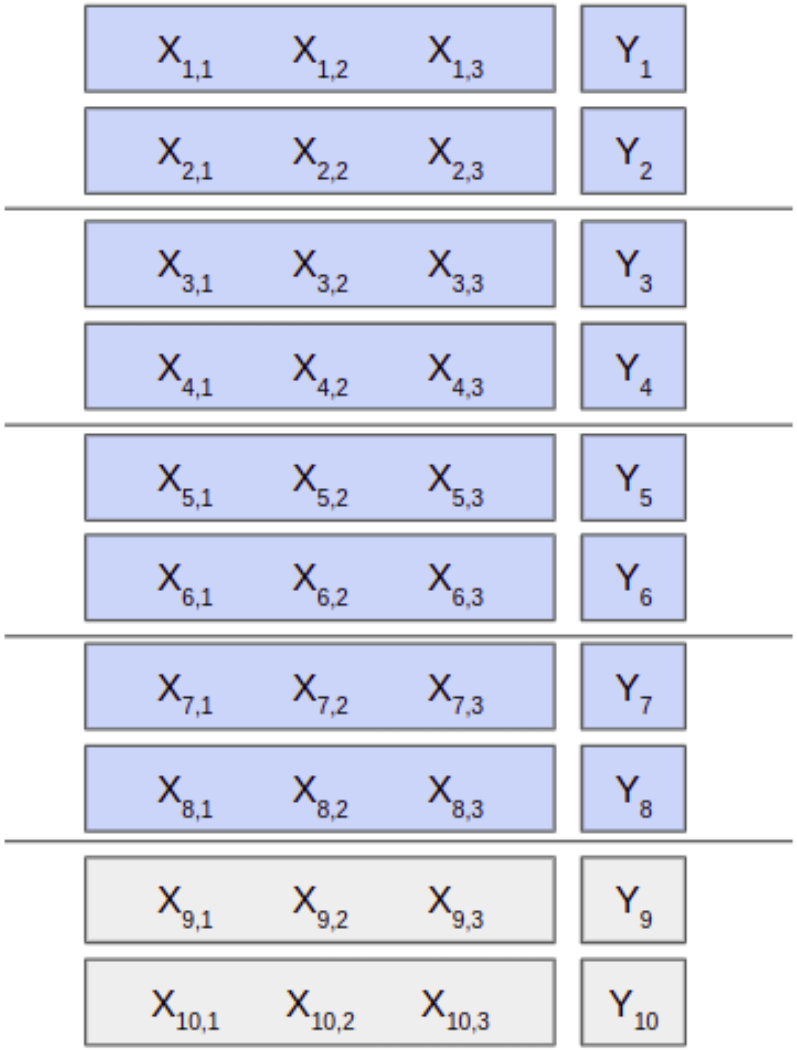
K-Fold Cross-Validation: Visualization



K-Fold Cross-Validation: Visualization



K-Fold Cross-Validation: Visualization



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

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

Comparing Cross-Validation Methods

Method	Bias	Variance	Computation
LOOCV	Lowest	Highest	Most expensive
K-fold CV	Low	Medium	Moderate
Train-Test Split	Highest	Lowest	Least expensive

K-fold CV is often the preferred method:

- Good balance between bias and variance
- More efficient use of data than simple train-test split
- Less computationally intensive than LOOCV
- $K = 5$ or $K = 10$ typically works well in practice
- For added stability, repeat K-fold CV multiple times with different random partitions

Cross-Validation for Model Selection

Cross-validation provides a principled approach to comparing different models:

1. For each candidate model:
   - Perform cross-validation
   - Calculate CV error estimate
2. Select model with lowest CV error
3. Refit selected model on full dataset

Important Note

When using CV for both model selection and performance estimation, use **nested cross-validation** to avoid selection bias in your final error estimate!

Putting It All Together: The Model Building Process

1. Define your prediction task and loss function
2. Select candidate models to compare
3. Use cross-validation to estimate generalization error for each model
4. Select the model with the lowest cross-validation error
5. Refit the selected model on the full dataset
6. If desired, evaluate final model on a completely held-out test set

Key Takeaways

- Cross-validation provides an objective way to assess predictive performance
- It helps manage the bias-variance tradeoff in model selection
- Different CV approaches offer tradeoffs between bias, variance, and computational cost
- The goal is always to find models that generalize well to new, unseen data