MA477: Data Science

Lesson 8 Outline — 27 January 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

---

1 Administrative

- Project 1 Discussion
- Student review
- Model Review Questions
- Model Review Lecture

2 Model Selection Lesson Objectives

- Understand the purpose and implementation of best-subsets and various stepwise approaches to regression modeling.
- Gain familiarity with various measures of model performance that are commonly used for model selection.

3 Student Review

3.1 Adjusted Error Measures for Model Selection

Consider a linear regression model with:

- n observations,
- d predictors (excluding the intercept),
- Residual Sum of Squares RSS,
- Error variance estimate $\hat{\sigma}^2$ (typically from the full model),
- Total Sum of Squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Mallows' C_p

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

Akaike Information Criterion (AIC) For least squares regression with Gaussian errors:

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

(up to an additive constant).

Bayesian Information Criterion (BIC)

$$BIC = \frac{1}{n}\left(RSS + \log(n)\,d\,\hat{\sigma}^2\right)$$

(up to an additive constant).

Adjusted R^2

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

Discussion Questions: From Least Squares to Regularization

1. Discuss some of the advantages of using linear model (Least Squares). What are some of the limitations? Least squares is simple, fast to compute, and easy to interpret, with good statistical properties when the sample size is large. Its main limitations are high variance, sensitivity to multicollinearity, and instability or non-uniqueness when the number of predictors is large relative to the number of observations.

2. As the number of predictors approaches the number of observations, what changes do you expect in the behavior of the fitted coefficients, even if the training error continues to decrease? The fitted coefficients tend to become highly variable and unstable, with small changes in the data leading to large changes in estimates, even though the training error keeps decreasing.

3. In a setting where the number of predictors exceeds the number of observations, what happens? There is no unique least squares solution; instead, there are infinitely many coefficient vectors that fit the training data perfectly.

4. If many different coefficient vectors produce zero training error, what criteria might you use to decide which solution is preferable for prediction or interpretation? We can prefer solutions with smaller coefficient magnitudes, greater stability, better generalization to new data, or improved interpretability.

5. Instead of selecting a single subset of predictors, what else can we do? We can prefer solutions with smaller coefficient magnitudes, greater stability, better generalization to new data, or improved interpretability.

4 Guided In-Class Flow: From Least Squares to Regularization

Phase 1: Establish comfort with least squares

Prompt: Suppose we're fitting a linear regression with least squares. When does this work really well?

Responses:

- When the sample size is large
- When predictors aren't too correlated
- When the model is simple

Follow-up: So far, so good. Now let's stress the system.

Phase 2: Push toward the edge case $p \approx n$

Question: What happens to the least squares solution as the number of predictors gets close to the number of observations?

Responses:

- Estimates get unstable
- Variance increases
- Overfitting

Follow-up: Okay. Now let's go one step further.

Phase 3: Trigger the realization $p > n$ (core insight)

Setup: Let's consider a tiny dataset.

Suppose we have:

- $n = 2$ observations
- $p = 3$ predictors

We write:

$$y_1 = \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$$
$$y_2 = \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23}$$

Question: How many equations do we have? How many unknowns?

Responses:

- Two equations
- Three unknowns

Follow-up: So… do we get a unique solution?

Realization: No — infinitely many solutions.

Takeaway: Least squares doesn't fail by giving bad answers — it fails by giving *too many* answers.

Concrete numeric example: $n < p$ implies infinitely many solutions

Take $n = 2$ observations and $p = 3$ predictors (no intercept for simplicity). Let the design matrix and response be

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \qquad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We want $\beta = (\beta_1, \beta_2, \beta_3)^\top$ such that $X\beta = y$, i.e.

$$\begin{cases} \beta_1 + \beta_3 = 1, \\ \beta_2 + \beta_3 = 1. \end{cases}$$

Solve by treating β_3 as a free parameter. Let $\beta_3 = t$ for any real number t. Then

$$\beta_1 = 1 - t, \qquad \beta_2 = 1 - t, \qquad \beta_3 = t.$$

So the set of solutions is

$$\beta(t) = \begin{bmatrix} 1 - t \\ 1 - t \\ t \end{bmatrix}, \qquad t \in \mathbb{R}.$$

Check:

$$X\beta(t) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 - t \\ 1 - t \\ t \end{bmatrix} = \begin{bmatrix} (1 - t) + t \\ (1 - t) + t \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = y.$$

Therefore, the residual sum of squares is

$$RSS(t) = \|y - X\beta(t)\|_2^2 = 0 \quad \text{for infinitely many distinct } \beta(t).$$

Concrete distinct solutions:

$$t = 0 \Rightarrow \beta = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad t = 1 \Rightarrow \beta = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad t = 2 \Rightarrow \beta = \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}.$$

All of these fit the training data perfectly, illustrating why when $p > n$ the least squares solution is not unique.

Phase 4: Why this is a real problem

Question: If all of these solutions fit the training data perfectly, how do we choose one?

Responses:

- Small coefficients
- Simpler model
- Fewer predictors

Follow-up: What happens if we just pick one arbitrarily?

Responses:

- Poor predictions
- High variance
- Unstable coefficients

Phase 5: Try model selection (and let it fail)

Prompt: Okay, what if we avoid the problem by choosing fewer predictors?

Response: Subset selection.

Follow-up: How would we do that in principle?

Response: Try all combinations.

Follow-up: What happens when p is large?

Expected student response: It becomes computationally infeasible.

Follow-up: And even if we approximate with forward or backward stepwise selection, do we still face instability when p is large?

Response: Yes.

$$\textbf{Best subset selection:} \quad \text{fits all } 2^p \text{ possible models} \quad \Rightarrow \quad \text{exponential time complexity } O(2^p).$$

$$\textbf{Stepwise selection:} \quad \text{fits about } 1 + \frac{p(p+1)}{2} \text{ models} \quad \Rightarrow \quad \text{polynomial time complexity } O(p^2).$$

Phase 6: Lead to regularization (the pivot)

Pivot question: Instead of picking one subset, what if we keep all predictors but control how much influence each one can have?

Rephrase: What if we say: among all models that fit the data well, we prefer the one with smaller coefficients?

Responses:

- More stable
- Less overfitting
- Better prediction

Phase 7: Name the idea after they invent it

What do we have?: What we've just described has a name.

Name it: This is called **regularization**.

Phase 8: Bridge to next topic

Closing time: Regularization doesn't ask: *Which predictors should I delete?* It asks: *How strongly should each predictor be allowed to matter?*