STA 35C: Statistical Data Science III

Lecture 15: Linear Model Selection – Subset Selection

Dogyoon Song

Spring 2025, UC Davis

Today's topics

- **Recall: Resampling methods**
  - Cross-validation: *estimate test performance* using training data
  - The bootstrap: *quantify uncertainty* by resampling from the given dataset

- **Model Selection (Today & Wed):** Identify relevant predictors among many
  - **Why?**
    - Improve prediction accuracy (avoid overfitting)
    - Improve model interpretability
  - **How?**
    - Subset selection (today)
    - Regularization (next lecture on Wed)
    - Dimension reduction (not covered in STA 35C; possibly in STA 142A)

Brief recap: Resampling methods

Given a single dataset & a single model, we often want to assess **model performance**

- **Test performance** (e.g., test MSE)
  - We care about performance on new (test) data, but only have a training dataset
  - **Key idea:** Hold out part of the data for validation
  - **Cross-validation:** Repeat data splits multiple times & aggregate results for a more reliable test performance estimate
- **Uncertainty quantification** (e.g., standard error)
  - We want to gauge variability in parameter estimates
  - If we could draw fresh samples from nature, we'd see how estimates vary
  - **The bootstrap:** Since we cannot acquire new data, we resample from our existing dataset (treating it as an empirical distribution)

(Linear) Model selection

Recall multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- In reality, we might have many predictors, unsure which are truly helpful
- **Example:** `Credit` **dataset**
  - Response: `balance`
  - Predictors: `income`, `limit`, `rating`, `cards`, `age`, `education`, `own`, `student`, `married`, `region`
- **Goal:** Choose a subset of *relevant* predictors

Why model selection?

Two main reasons:

- **Prediction accuracy**
  - Overfitting can occur if we use too many predictors
  - If $p > n$, we might not even get a unique least squares solution (variance $\to \infty$)
  - Reducing predictors can lower variance and improve generalization
- **Model interpretability**
  - Many of the available predictors might not be truly associated with the response
  - Including unnecessary predictors can mislead interpretation
  - Simpler models are easier to interpret and explain

How to do model selection?

Three key approaches for linear model selection:

- **Subset selection**
  - Identify a relevant subset of predictors, then fit via least squares
- **Regularization** (to be discussed on Wed)
  - Add a penalty term to least squares formulation that favors "simpler" models
- **Dimension reduction** (not covered in STA 35C)
  - Project the p predictors into a smaller set of $p' \ll p$ linear combinations

Today's focus: Subset selection

- Best subset selection
- How to choose the optimal model
- Stepwise selection (greedy approximation)

Best subset selection

Idea: Try *all* subsets of predictors, and pick the one that performs the best

- With p predictors, there are 2^p possible subsets
- Compare models of different sizes carefully (recall R^2 vs. R^2_{adj})

Procedure[1]**:**

- Let $\mathcal{M}_0$ be the null model (no predictors, just intercept)
- For $k = 1, \ldots, p$:
  - Fit all $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ models with exactly k predictors
  - Pick the best (lowest RSS or highest R^2) among them, call it $\mathcal{M}_k$
- Finally, select the best among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using a test-performance proxy
  - e.g., adjusted R^2 or cross-validation (more on this later)

[1] See [JWHT21, Chapter 6.5.1] for example codes

Best subset selection: Example ($n = 3$, $p = 2$)

Example

Dataset: 3 points with 2 predictors (X_1, X_2) and a response Y:

$$(X_1, X_2, Y) = (1, 2, 3),$$
$$(X_1, X_2, Y) = (2, 1, 4),$$
$$(X_1, X_2, Y) = (3, 3, 5).$$

Candidate subsets:

- $\mathcal{M}_0$: Null model (intercept only).
- $\mathcal{M}_1^{(X_1)}$: Use X_1 only.
- $\mathcal{M}_1^{(X_2)}$: Use X_2 only.
- $\mathcal{M}_2$: Use (X_1, X_2).

Best subset selection: Example (Step 1)

Example

Step 1: Fit each model and compute R^2.

1) $\mathcal{M}_0$: intercept only

- $\hat{\beta}_0 = \bar{Y} = \frac{3+4+5}{3} = 4$.
- $\mathrm{RSS}_0 = \sum (Y_i - 4)^2 = 1 + 0 + 1 = 2$.
- $TSS = 2$, and thus, $R_0^2 = 1 - \frac{2}{2} = 0$.

2) $\mathcal{M}_1^{(X_1)}$: one predictor X_1

- $(x_i, y_i) = \{(1,3), (2,4), (3,5)\}$; $\bar{x} = 2, \bar{y} = 4$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2} = \frac{(-1)\cdot(-1) + 0 \cdot 0 + 1 \cdot 1}{(-1)^2 + (0)^2 + (1)^2} = \frac{2}{2} = 1, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4 - 1 \cdot 2 = 2.$$

- Thus $\hat{Y} = 2 + 1X_1 \implies$ fitted values $(3, 4, 5)$.
- $\mathrm{RSS}_1 = 0 \implies R_1^2 = 1$.

Best subset selection: Example (Step 1, cont'd)

Example

...(continued from the previous slide)...

3) $\mathcal{M}_1^{(X_2)}$: one predictor X_2

- $(x_2, y) = \{(2,3), (1,4), (3,5)\}$, $\bar{x}_2 = 2$, $\bar{y} = 4$.

$$\hat{\beta}_1 = \frac{0 + 0 + 1 \cdot 1}{0 + (-1)^2 + 1^2} = \frac{1}{2} = 0.5, \quad \hat{\beta}_0 = 4 - 0.5 \cdot 2 = 3.$$

- $\hat{Y} = 3 + 0.5X_2 \implies \hat{Y} = \{4, 3.5, 4.5\}$.

$$\mathrm{RSS}_2 = (3-4)^2 + (4-3.5)^2 + (5-4.5)^2 = 1 + 0.25 + 0.25 = 1.5, \quad R_2^2 = 1 - \frac{1.5}{2} = 0.25.$$

4) $\mathcal{M}_2$: two predictors (X_1, X_2)

- With 3 points, a model with 2 predictors can fit perfectly if consistent.
- We get $\mathrm{RSS}_{1,2} = 0$, and thus, $R_{1,2}^2 = 1$.

Best subset selection: Example (Step 2)

Example

Step 2: Compare the four candidate models.

Choose the best by adjusted R^2 or a simpler-subset preference.

Recall $R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$:

$$
\begin{aligned}
&R^2_{\text{adj}}(\mathcal{M}_0) = 0,\\
&R^2_{\text{adj}}(\mathcal{M}_1^{(X_1)}) = 1,\\
&R^2_{\text{adj}}(\mathcal{M}_1^{(X_2)}) = -0.5,\\
&\mathcal{M}_2: \text{ undefined due to } n - p - 1 = 0.
\end{aligned}
$$

Therefore, we choose $\mathcal{M}_1^{(X_1)}$.

Best subset selection: Visualization

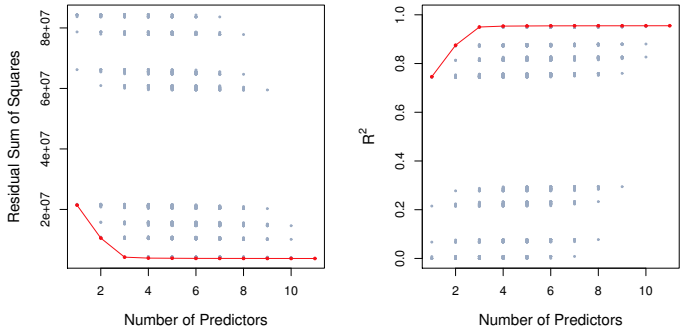

Figure: In the `Credit` dataset, RSS and R^2 are displayed for each subset of the ten predictors. The red frontier indicates the best model at each subset size. The x-axis goes from 1 to 11 because one categorical predictor (three levels) is split into two dummy variables [JWHT21, Figure 6.1].

- Pick the model with the lowest test MSE or best adjusted R^2
- If the improvement is marginal (e.g., within 1 SE of the best), pick a simpler subset

Evaluating models & criteria

Goal: Out of $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$, choose the model with the best *test* performance

- Training performance (e.g., RSS or R^2) alone can be misleading

Common criteria:

- **Adjusted** R^2:
  - $R^2_{\text{adj}} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$
  - Increases only if adding predictors significantly decreases RSS
- **Cross-validation**:
  - An empirical approach splitting/re-splitting data to estimate test error
- C_p**, AIC, BIC** (beyond the scope of this course):
  - Analytical formulas penalizing model size (p) under certain theoretical assumptions

Visualization of selection criteria

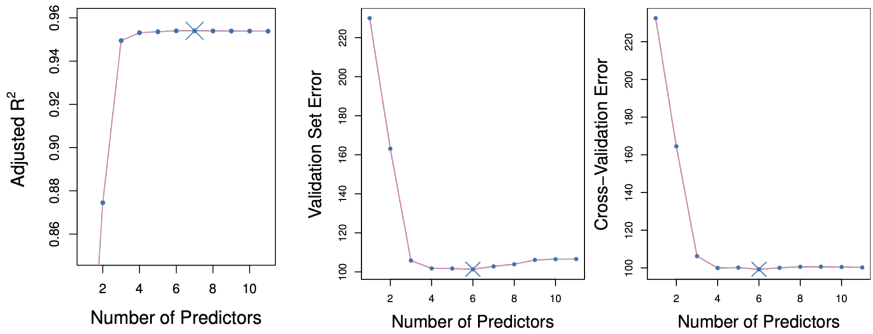

Figure: For the `Credit` dataset, adjusted R^2, validation error (single split), and cross-validation error are displayed for the best model containing k predictors, for k ranging from 1 to 11. The overall best model, based on each of these quantities, is shown as a blue cross [JWHT21, Figures 6.2 & 6.3, excerpted].

- These methods often choose similar models

Pop-up quiz #1: Best subset selection

Question: Which of the following statements about *best subset selection* is **TRUE**?

A) It considers only a single split of the training data.

B) It fits *all* possible models of each size k and picks the best for that k, and also best k.

C) It can handle extremely large p quickly by skipping some potential subsets.

D) Once a predictor is in the model, it cannot be removed at later steps.

Answer: (B) is correct. Best subset selection exhaustively checks every combination (subsets) of predictors of size k. The other options refer to other approaches or constraints.

Best subset selection: Summary & limitations

Key idea: Exhaustively explore 2^p subsets; pick the best by a test-performance criterion

- Useful when p is small
- $\{\mathcal{M}_k\}$ denotes the best k-predictor model; we choose among $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ using a test-performance measure
- Common performance metrics: adjusted R^2, C_p, AIC, BIC, cross-validation
- Straightforward, systematic approach for accuracy & interpretability

Limitation: 2^p grows rapidly (with p), often infeasible for large p

- e.g., $p = 10 \to 2^p \approx 10^3$; $p = 50 \to 2^p \approx 10^{15}$ (infeasible)

Forward stepwise selection

Idea: A *greedy*[2] approximation to best subset selection, adding one predictor at a time

Procedure[3]**:**

- $\mathcal{M}_0$: null model with intercept only
- **For** $k = 0, \ldots, p-1$:
  - Consider all $(p-k)$ models that add exactly 1 unused predictors to $\mathcal{M}_k$
  - Pick the best updated model, and call it $\mathcal{M}_{k+1}$
- Finally, compare $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ using adjusted R^2 or other test-based metrics

[2] At each step, pick the best addition via a *local* search

[3] See [JWHT21, Chapter 6.5.1] for example codes

Forward stepwise selection: Example (Overview)

Example

Dataset: 4 points with 3 predictors (X_1, X_2, X_3) and response Y:

$$(X_1, X_2, X_3, Y) = (1,\ 2,\ 2,\ 2.5),$$
$$(X_1, X_2, X_3, Y) = (2,\ 1,\ 1,\ 3.5),$$
$$(X_1, X_2, X_3, Y) = (3,\ 3,\ 2,\ 6),$$
$$(X_1, X_2, X_3, Y) = (4,\ 1,\ 3,\ 6.5).$$

- Step 0: $\mathcal{M}_0$ fits $Y = \beta_0$; compute $\mathrm{RSS}_0 \approx 11.19$.
- Step 1: Fit $\mathcal{M}_1^{(X_1)}, \mathcal{M}_1^{(X_2)}, \mathcal{M}_1^{(X_3)}$. Pick best single predictor (with largest R^2).
- Step 2: Add a second predictor from the remaining, forming $\mathcal{M}_2$. Check R^2, R^2_{adj}.
- Step 3: Possibly add the last predictor ($\mathcal{M}_3$ with all three predictors).
- Final selection: Compare $\mathcal{M}_0, \ldots, \mathcal{M}_3$ and choose the subset subset with best test performance.

Forward stepwise selection: Example (Step 0)

Example

Step 0: Null model. $\mathcal{M}_0 : \ Y = \beta_0$.

- $\hat{\beta}_0 = \bar{Y} = \frac{2.5+3.5+6+6.5}{4} = 4.625$.
- $\begin{aligned} \mathrm{RSS}_0 &= \sum (Y_i - 4.625)^2 = (2.5 - 4.625)^2 + (3.5 - 4.625)^2 + (6 - 4.625)^2 + (6.5 - 4.625)^2 \\ &= (-2.125)^2 + (-1.125)^2 + (1.375)^2 + (1.875)^2 \\ &\approx 4.51 + 1.27 + 1.89 + 3.52 \\ &= 11.19. \end{aligned}$
- $TSS = 11.19, \quad R_0^2 = 1 - \frac{11.19}{11.19} = 0.$

Step 1: Fit and compare each single-predictor model $X_1, \ X_2, \ X_3$.

- We now fit $\mathcal{M}_1^{(X_1)}$, $\mathcal{M}_1^{(X_2)}$, and $\mathcal{M}_1^{(X_3)}$.
- Then compute RSS and R^2 for each; See the next slide for sample calculation for $\mathcal{M}_1^{(X_1)}$.

Conclusion of Step 1: Whichever single predictor yields the highest R^2 (or lowest RSS) is $\mathcal{M}_1$.

Forward stepwise selection: Example (Step 1, further details)

Example

Illustration for X_1:

- $X_1 = \{1, 2, 3, 4\}$, $Y = \{2.5, 3.5, 6, 6.5\}$.
- Slope $\hat{\beta}_1, \hat{\beta}_0$ are obtained by least squares:

$$\hat{\beta}_1 = \frac{\sum(x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum(x_{1i} - \bar{x}_1)^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

  where $\bar{x_1} = 2.5$ and $\bar{y} = 4.625$. Eventually, we find

$$\hat{\beta}_1 \approx 1.2, \quad \hat{\beta}_0 \approx 3.025 \text{ (approx)}.$$

- Then $\mathrm{RSS}_1 \approx 2.55$, $\quad R_1^2 = 1 - \frac{2.55}{11.19} \approx 0.77$.

Similarly for X_2, X_3: $\mathrm{RSS}_2 \approx 4.12$, $R_2^2 \approx 1 - \frac{4.12}{11.19} = 0.63$, and $\mathrm{RSS}_3 \approx 3.20$, $R_3^2 \approx 1 - \frac{3.20}{11.19} = 0.71$.

Pick the best single predictor $= X_1$, whic yields highest R^2 or lowest RSS.

Forward stepwise selection: Example (Step 2)

Example

Step 2: Add a second predictor to $\mathcal{M}_1$

Now $k = 1$. Our model has X_1, so the unused are X_2 and X_3:

$$\mathcal{M}_2^{(X_1,X_2)}, \quad \mathcal{M}_2^{(X_1,X_3)}.$$

We fit each, compute RSS&R^2. For instance:

(1) $\mathcal{M}_2^{(X_1,X_2)}$: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. After fitting this model:

$$\mathrm{RSS}_{1,2} \approx 1.80, \quad R^2_{1,2} = 1 - \tfrac{1.80}{11.19} \approx 0.84.$$

(2) $\mathcal{M}_2^{(X_1,X_3)}$: Similarly, we get $\mathrm{RSS}_{1,3} \approx 1.40, \quad R^2_{1,3} = 1 - \tfrac{1.40}{11.19} \approx 0.875.$

Hence $\mathcal{M}_2 = \mathcal{M}_2^{(X_1,X_3)}$ (larger R^2).

Forward stepwise selection: Example (Step 3 & Selection)

Example

Step 3: Add the remaining predictor to $\mathcal{M}_2$

Now $k = 2$. Our model includes X_1 and X_3. The remaining predictor is X_2. So we consider:

$$\mathcal{M}_3 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

- Fit the full model with 3 predictors, compute $\mathrm{RSS}_{1,2,3}$, $R^2_{1,2,3}$
- Suppose $\mathrm{RSS}_{1,2,3} \approx 1.25$, $\quad R^2_{1,2,3} = 1 - \frac{1.25}{11.19} \approx 0.89$.

Finally, we might pick $\mathcal{M}_1$ or do cross-validation among the four models:

$$\begin{aligned}
&\mathcal{M}_0 : && R^2 = 0, && R^2_{\text{adj}} = 0, \\
&\mathcal{M}_1 = \{X_1\} : && R^2 \approx 0.77, && R^2_{\text{adj}} \approx 0.66, \\
&\mathcal{M}_2 = \{X_1, X_3\} : && R^2 \approx 0.875, && R^2_{\text{adj}} \approx 0.63, \\
&\mathcal{M}_3 = \{X_1, X_2, X_3\} : && R^2 \approx 0.89, && R^2_{\text{adj}} \text{ undefined } (\ n - p - 1 = 0\ ).
\end{aligned}$$

Pop-up quiz #2: Stepwise selection

Question: In *forward stepwise selection*, how do we choose a new predictor at each step?

A) By removing whichever predictor contributes least to the model's fit.

B) By guessing randomly among the remaining predictors.

C) By trying each unused predictor one at a time and selecting the one that yields the best improvement in the chosen metric (e.g., R^2).

D) By fitting all 2^p possible models and picking the global best.

Answer: (C) is correct. Forward stepwise selection tries each *remaining* predictor individually at each step, then adds the one that most improves the model.

Example: Stepwise selection may yield a different subset

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income` `student`, `limit`	`rating`, `income`, `student`, `limit`

Figure: The first four chosen models for best subset selection and forward stepwise selection on the `Credit` dataset. The first three models are identical, but the fourth differs [JWHT21, Table 6.1].

- Stepwise typically performs well and is computationally much cheaper:

$$1 + \sum_{k=0}^{p-1} (p-k) = 1 + \frac{p(p+1)}{2} \quad \ll \quad 2^p$$

- However, it may pick a different subset if the greedy path diverges

Backward stepwise selection

There is an alternative stepwise method reverses the order of search

Backward stepwise selection:

- Start with the full model; remove one predictor at a time
- Usually require $n > p$ so the full model can be fit initially

Comparison with forward stepwise:

- Both are *greedy* algorithms using local decisions
- Both drastically reduce the search space vs. best subset when p is large
- They can yield different subsets if they take different paths

Wrap-up & Takeaways

Model selection (Subset selection): Identify a subset of relevant predictors

- **Purposes:**
  - Improve prediction accuracy and avoid overfitting
  - Enhance model interpretability
- **Methods:**
  - **Best subset selection**:
    - Exhaustively checks all 2^p subsets (optimal but expensive)
    - Feasible only for small p (e.g., $p \lesssim 20$)
  - **Stepwise selection** (forward or backward):
    - Much fewer model fits needed: $1 + \frac{p(p+1)}{2}$ vs. 2^p
    - Often performs well in practice, but may miss the globally optimal subset
  - **Overall:**
    - Stepwise methods generally give good models but are *not guaranteed* to be optimal
    - For moderate or large p ($\gtrsim 50$), stepwise is typically the only feasible approach

References


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.