

STA 35C: Statistical Data Science III

Lecture 15: Linear Model Selection – Subset Selection

Dogyoon Song

Fall 2026, UC Davis

Announcements

Homework 4 is due tomorrow (Thu, Oct 29, 11:59 PM PT)

Extended office hours: Wednesdays, 1–3 PM (MSB 1143)

- The 2–3 PM hour is shared with STA 130A, whose students may receive priority

Agenda

- **Last time: Resampling methods**
 - Cross-validation: *estimate test performance* using training data
 - The bootstrap: *quantify uncertainty* by resampling from the given dataset
- **Model Selection (Today & Fri):** Identify relevant predictors among many
 - **Why?**
 - Improve prediction accuracy (avoid overfitting)
 - Improve model interpretability
 - **How?**
 - Subset selection: search over predictor subsets (today)
 - Regularization: penalize overly complex models (Fri)
 - Dimension reduction: replace predictors by lower-dimensional summaries

Recap: Resampling methods

Given one dataset, we often want to assess model performance or quantify uncertainty

- **Test performance estimation** (e.g., test MSE)
 - We care about performance on new (test) data, but only have a training dataset
 - **Validation set approach:** Hold out part of the data for validation
 - **Cross-validation:** Repeatedly hold out part of the data & 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:** Resample with replacement from the observed data (treating it as an empirical proxy) to approximate the sampling distribution of an estimator

(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
 - e.g., {**income, limit, age**} or {**rating, education**} or something else

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 $\rightarrow \infty$)
- 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**
 - Search over subsets of predictors to identify a relevant subset
- **Regularization** (next lecture)
 - 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 candidate predictors, there are 2^p possible subsets
- Compare models of different sizes carefully (recall R^2 vs. R^2_{adj})

Procedure¹:

- Let $\mathcal{M}_0$ be the null model (no predictors, just intercept)
- For $k = 1, \dots, p$:
 - 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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ using a model-selection criterion
 - e.g., adjusted R^2 or cross-validation (more on this later)

¹See [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

- $\hat{\beta}_0 = \bar{Y} = \frac{3+4+5}{3} = 4.$
- $RSS_0 = \sum (Y_i - 4)^2 = 1 + 0 + 1 = 2.$
- $TSS = 2$, 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 $(3, 4, 5).$
- $RSS_1 = 0 \implies R_1^2 = 1.$

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

Example

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

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

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

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

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

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

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

- The unique full-model fit is $\hat{Y} = 2 + X_1 + 0X_2$, giving fitted values $(3, 4, 5)$.
- We get $\text{RSS}_{1,2} = 0$, 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 for k predictors and $n > k + 1$, $R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$:

$$R_{\text{adj}}^2(\mathcal{M}_0) = 0,$$

$$R_{\text{adj}}^2(\mathcal{M}_1^{(X_1)}) = 1,$$

$$R_{\text{adj}}^2(\mathcal{M}_1^{(X_2)}) = -0.5,$$

$\mathcal{M}_2$: undefined due to $n - p - 1 = 0$.

Therefore, we choose $\mathcal{M}_1^{(X_1)}$ (the one with highest adjusted R^2).

- It has the highest defined adjusted R^2 , and adding X_2 does not improve the fit.

Best subset selection: Visualization

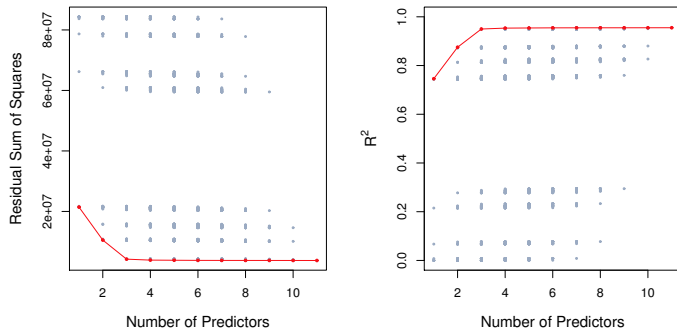


Figure: In the **Credit** dataset, RSS and R^2 are displayed for each subset of the ten predictors. The red frontier indicates the best model at each subset size. The x-axis goes from 1 to 11 because one categorical predictor (three levels) is split into two dummy variables [JWHT21, Figure 6.1].

- Choose model size using validation/CV error or adjusted R^2 , not training RSS alone
- If estimated predictive performance is similar, prefer a simpler model

Evaluating models & criteria

Goal: Out of $\{\mathcal{M}_0, \dots, \mathcal{M}_p\}$, 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 :**
 - $R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$ for k selected predictors, $n > k + 1$, and $\text{TSS} > 0$
 - Increases only if the RSS decrease is large enough to offset the extra predictor
- **Cross-validation:**
 - Repeat the subset search within each training fold, then assess on its held-out fold
- C_p , **AIC**, **BIC** (beyond the scope of this course):
 - Analytical formulas penalizing model size (k) under certain theoretical assumptions

Visualization of selection criteria

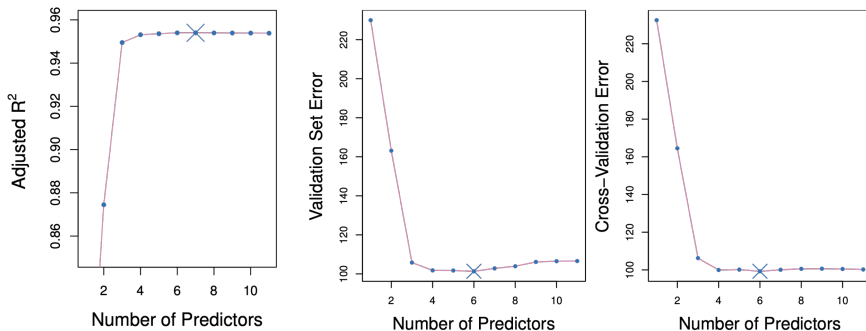


Figure: For the Credit dataset, adjusted R^2 , validation error (single split), and cross-validation error are displayed for the best model containing k predictors, for k ranging from 1 to 11. The overall best model, based on each of these quantities, is shown as a blue cross [JWHT21, Figures 6.2 & 6.3,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

- These methods often choose similar models

Pop-up quiz: Best subset selection

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

- A) It chooses the model with the smallest training RSS among all subset sizes.
- 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) It always improves test error as more predictors are added.

Answer: B. Best subset selection searches all subsets, keeps the best model of each size, and then uses a criterion beyond training RSS to choose the final size.

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 $\{\mathcal{M}_0, \dots, \mathcal{M}_p\}$ using a test-performance measure
- Common selection criteria: adjusted R^2 , validation error, cross-validation error, C_p , AIC, BIC
- Straightforward, systematic approach for accuracy & interpretability

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

- e.g., $p = 10 \Rightarrow 2^p = 1024$, while $p = 50 \Rightarrow 2^p \approx 10^{15}$ (infeasible)

Forward stepwise selection

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

Procedure³:

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

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

³See [JWHT21, Chapter 6.5.1] for example codes

Illustration: Best subset selection vs. forward stepwise

An illustrative example with $p = 4$:

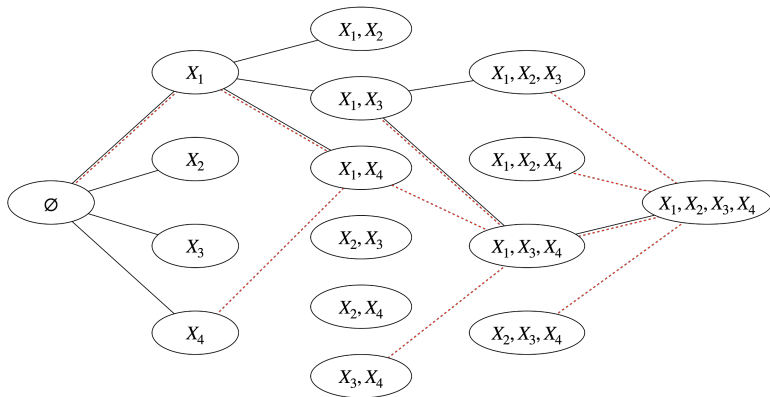


Figure: Illustration of forward stepwise (**black solid path**; successively adding $X_1 \rightarrow X_4 \rightarrow X_3 \rightarrow X_2$) and backward stepwise (**red dashed path**; removing $X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_1$). Best subset selection checks all 2^p possibilities; all three can yield different outcomes.

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:

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

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

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 $\hat{Y} = \bar{Y}$; compute $\text{RSS}_0 \approx 11.188$.
- 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: Compare the two possible additions by training RSS to choose $\mathcal{M}_2$.
- Step 3: Add the last predictor to obtain $\mathcal{M}_3$.
- Final selection: Compare adjusted R^2 of $\mathcal{M}_0, \dots, \mathcal{M}_3$ and choose the model with the highest value.

Forward stepwise selection: Example (Step 0)

Example

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

- $\hat{\beta}_0 = \bar{Y} = \frac{2.5+3.5+6+6.5}{4} = 4.625$.
- $\text{RSS}_0 = \sum (Y_i - 4.625)^2 = (2.5 - 4.625)^2 + (3.5 - 4.625)^2 + (6 - 4.625)^2 + (6.5 - 4.625)^2$
 $= (-2.125)^2 + (-1.125)^2 + (1.375)^2 + (1.875)^2$
 $= \frac{179}{16} \approx 11.188$.
- $\text{TSS} = \frac{179}{16}$, $R_0^2 = 1 - \frac{179/16}{179/16} = 0$.

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 :

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

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

where $\bar{x}_1 = 2.5$ and $\bar{y} = 4.625$. Eventually, we find

$$\hat{\beta}_1 = 1.45, \quad \hat{\beta}_0 = 1.$$

- Residuals $(0.05, -0.4, 0.65, -0.3)$ give $\text{RSS}_1 = 0.675$, $R_1^2 = 1 - \frac{0.675}{179/16} \approx 0.940$.

Similarly for X_2, X_3 : $\text{RSS}_2 = \frac{243}{22} \approx 11.045$, $R_2^2 \approx 0.013$, and $\text{RSS}_3 = \frac{107}{16} \approx 6.688$, $R_3^2 \approx 0.402$.

Pick the best single predictor = X_1 , which 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 :

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

(1) $\mathcal{M}_2^{(X_1, X_2)}$: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The fit $\hat{Y} = 1.5X_1 + 0.5X_2$ matches all four responses exactly:

$$\text{RSS}_{1,2} = 0, \quad R_{1,2}^2 = 1.$$

(2) $\mathcal{M}_2^{(X_1, X_3)}$: The fit is $\hat{Y} = \frac{11}{12} + \frac{17}{12}X_1 + \frac{1}{12}X_3$, giving

$$\text{RSS}_{1,3} = \frac{2}{3} \approx 0.667, \quad R_{1,3}^2 = 1 - \frac{2/3}{179/16} = \frac{505}{537} \approx 0.940.$$

Hence $\mathcal{M}_2 = \mathcal{M}_2^{(X_1, X_2)}$ (smaller RSS = 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_2 . The remaining predictor is X_3 . So we consider:

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

- The full-model fit is $\hat{Y} = 1.5X_1 + 0.5X_2 + 0X_3$ (intercept 0).
- Thus $\text{RSS}_{1,2,3} = 0$ and $R_{1,2,3}^2 = 1$: adding X_3 does not improve training fit.

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

$\mathcal{M}_0 :$	$R^2 = 0,$	$R_{\text{adj}}^2 = 0,$
$\mathcal{M}_1 = \{X_1\} :$	$R^2 \approx 0.940,$	$R_{\text{adj}}^2 \approx 0.909,$
$\mathcal{M}_2 = \{X_1, X_2\} :$	$R^2 = 1,$	$R_{\text{adj}}^2 = 1,$
$\mathcal{M}_3 = \{X_1, X_2, X_3\} :$	$R^2 = 1,$	R_{adj}^2 undefined ($n - k - 1 = 0$).

Choose $\mathcal{M}_2$ by this criterion. Perfect training fit does not guarantee good test performance.

Pop-up quiz: Best subset vs. forward stepwise

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. Forward stepwise selection tries each *remaining* predictor individually at each step, then adds the one that most improves the model.

Backward stepwise selection

Backward stepwise selection is similar to the forward selection, but reverses the direction of the 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
 - Improve prediction accuracy and avoid overfitting
 - Enhance model interpretability
- **Methods:**
 - **Best subset selection:**
 - Exhaustively checks all 2^p subsets (best training fit at each size, but expensive)
 - Exhaustive enumeration becomes costly as p grows; feasible only for small p
 - **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

Suggested reading: [JWHT21, Ch. 6.1; optional R lab: Ch. 6.5.1]

Next lecture: Ridge regression and an introduction to the lasso

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.