Stepwise Model Selection with Multiple Paths

Description for Final Project

1 Overview

This document describes a stepwise model selection procedure based on the Akaike Information Criterion (AIC). The main idea is to explore several promising model paths at once rather than following just a single "best" sequence of variables. At each step, we add new predictors that improve the model's AIC by a meaningful amount and keep all models that perform almost as well as the best one. Repeating this process creates a small collection of alternative models: different ways to explain the data that fit nearly equally well.

To check how stable these results are, we repeat the search on resampled versions of the data (for example, bootstrap samples) and record how often each variable is selected. Finally, we combine this information to identify the most reliable models: those that are both good in terms of AIC and made up of variables that tend to appear repeatedly across resamples.

This framework applies to both **linear** and **logistic** regression models and can be implemented as a compact R package.

2 Model Setup

We have data (x_i, y_i) for $i = 1, \ldots, n$, where each $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is a vector of p predictors and y_i is a response variable.

We consider:

- **Linear regression:** $y_i = x_i^\top \beta + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- **Logistic regression:** $\Pr(Y_i = 1 \mid x_i) = \dfrac{1}{1 + e^{-x_i^\top \beta}}$.

3 Method Outline

The method has three main parts:

1. Build several possible model paths by adding predictors step by step.
2. Check how stable the selections are when the data are slightly changed.
3. Keep only the models that are both good (low AIC) and stable (built from frequently selected variables).

3.1 Step 1: Building Multiple Model Paths

Instead of building just one model path (as in ordinary forward selection), we build several. At each step, we look at all predictors that could be added to the models from the previous step (parent models) to generate a complete set of models with one additonal predictor (children models) and choose the ones that improve AIC the most. If a few of them lead to nearly identical AIC values, we keep all of them for the next round.

Algorithm 1 Multiple-Path Forward Selection

1: **Input:** Data (X, y), model type (linear or logistic), number of steps K (maximum model size), minimum AIC improvement ε, tolerance for keeping near-ties δ, and optional limit L on how many models to keep per step.
2: Start with the empty model $S_0 = \varnothing$ (only the intercept).
3: **for** step $k = 1$ to K **do**
4: From every model in the previous step (parent model), add each variable (not already included in it) one at a time to obtain candidate (child) models.
5: For each candidate (child) model, compute its AIC and find the minimum (AIC_{min}) corresponding to the child's best AIC.
6: For each parent model, keep all children whose AIC is within δ of that parent's best AIC, as long as AIC_{min} decreases compared to the parent's best AIC by at least ε. If the last condition is not met, stop the loop.
7: Combine and remove duplicate models. If there are too many, keep only the best L by AIC.
8: **end for**
9: **Output:** A list of models at each step (a tree of alternative paths).

Intuition:

- If $\delta = 0$, you only keep the single best child per model: this reduces to normal forward selection.
- If $\delta > 0$, you keep near-ties, which allows exploring different combinations of predictors that perform almost equally well.
- The optional cap L helps keep the search from growing too large.

3.2 Step 2: Checking Stability with Resampling

We now repeat the multi-path search several times on resampled data sets. Each resample slightly perturbs the data (by resampling with or without replacement). If a predictor consistently appears in many of these resamples, it's a sign of stability.

Algorithm 2 Stability Estimation with Resampling

Input: Data (X, y), number of resamples B, resampling type (bootstrap or subsample), subsample size m if used, and multi-path parameters $(K, \varepsilon, \delta, L)$.
for $b = 1$ to B **do**
    Draw a bootstrap or subsample of the data.
    Run the multi-path search (Algortihm 1) on that resample.
    For each predictor j, compute the proportion of models it appears in (call the proportion $z_j^{(b)}$).
end for
Compute stability scores $\pi_j = \dfrac{1}{B} \sum_{b=1}^{B} z_j^{(b)}$ for all predictors.
Output: Stability vector $\pi = (\pi_1, \ldots, \pi_p)$.

Interpretation:

- π_j close to 1 means variable j is almost always selected.
- π_j near 0 means it's rarely used.

These proportions give a simple, intuitive picture of which predictors are most reliable.

3.3 Step 3: Selecting the Final Plausible Models

We now combine the results from Steps 1 and 2. We take all models built on the full data (Step 1) and keep those that are both:

1. Close to the best model in terms of AIC (within a tolerance Δ), and
2. Made up mostly of stable predictors (average stability above a threshold τ).

Algorithm 3 Selecting Plausible Models

Input: All models from the full-data search (Algorithm 1), their AIC values, and stability scores π (Algorithm 2); thresholds Δ and τ.
Find the lowest AIC value across all models.
Keep all models whose AIC is within Δ of this minimum.
For each retained model S, compute its average stability:

$$\overline{\pi}(S) = \frac{1}{|S|} \sum_{j \in S} \pi_j.$$

Keep only those models with $\overline{\pi}(S) \geq \tau$.
Output: The final set of plausible, stable models.

Interpretation: These "plausible" models represent alternative ways to explain the data that are both statistically sound (low AIC) and reliable (built from stable variables).

4 Putting Everything Together

Algorithm 4 Overall Multi-Path AIC Procedure

1: Run the multi-path search on the full data (Step 1).
2: Compute variable stability across resamples (Step 2).
3: Combine both results to filter for plausible, stable models (Step 3).
4: **Output:** Final set of plausible models and stability scores.

5 Typical Parameter Choices

Parameter	Description	Typical value
K	Maximum number of steps	$\min(p, 10)$
ε	Minimum AIC improvement to expand	10^{-6}
δ	AIC tolerance for keeping near-ties	0-2
L	Max number of models kept per level	25-100
B	Number of resamples for stability	50-100
m	Subsample size (if used)	$\lceil\sqrt{n}\rceil$
Δ	AIC tolerance for plausibility	2
τ	Minimum average stability	0.6

6 Summary of the Intuition

- **Why multiple paths?** Because different variables may provide similar fits; exploring several paths reveals alternative explanations.
- **Why resampling?** To see which predictors keep showing up, not just which happen to fit one dataset well.
- **Why the AIC window?** Models within a few AIC points are statistically indistinguishable; keeping them helps avoid overconfidence.
- **Why average stability?** It highlights models built from consistently useful predictors rather than those driven by random noise.