S610 Final Project: An Application of Forward Stepwise Model Selection

Junying Tong

Project Description

This project aims to explores the correlational relationship between fire weather conditions and the number of prescribed burn incidence in Georgia and Florida from 2010 to 2017, using historical weather data from ERA5 reanalysis and geospatial fire activity dataset from the Environmental Protection Agency. The forward stepwise model selection approach helps selecting only a set of meteorological parameters to be used for predicting the number of prescribed burns.

Implementation

I begin by loading a daily county-level fire weather dataset that links prescribed fire counts with a set of meteorological variables derived from ERA5 reanalysis data. I specify two benchmark models: a null model containing only an intercept, and a full model that includes all candidate predictors. These serve as reference points for the subsequent forward stepwise selection procedure.

The core of the analysis is a manually implemented forward stepwise selection algorithm based on the Akaike Information Criterion (AIC). Instead of relying on built-in routines, I write the selection procedure explicitly to maintain full transparency over each step of the algorithm. The procedure begins with the intercept-only model. At each iteration, I consider adding each one of remaining candidate predictors at a time, estimate the corresponding linear regression, and compute its AIC. The predictor that yields the largest improvement is added to the model, if it strictly improves upon the current best AIC. The algorithm terminates when no remaining predictor produces further AIC improvement. I apply this function to the observed fire weather dataset to obtain a selected regression model for the number of prescribed fires. The resulting model includes eight meteorological variables out of nine with statistically significant coefficients, reflecting strong associations between weather conditions and prescribed fire activity.

To evaluate the inferential properties of the model selection, I conduct a simulation under a known null data-generating process (DGP). The goal is not to improve model fitness, but to examine how standard statistical inference behaves after model selection. I keep the observed predictor matrix fixed and generate a synthetic outcome variable consisting purely of random noise. Specifically, I draw an error term from a normal distribution and replace the observed fire count with this noise, thereby imposing a true data-generating process in which all regression coefficients are zero. In each iteration, I rerun the same forward stepwise selection procedure. I then track whether a specific weather variable—daily mean 2-meter temperature ("*mean_t2m_F*")—is selected into the final model. Conditional on selection, I extract the usual OLS coefficient estimate and its corresponding p-value from the selected model. I repeat this process 500 times.

Tests

To ensure that the manually implemented forward stepwise selection algorithm behaves as intended, I constructed a small suite of unit tests using the "*testthat*" function. The goal of these tests is to verify that the mechanics of the algorithm are correct and behave as expected. My tests have two components: testing the type and structure of the linear model object, and testing the monotonic improvement relative to the null model as more predictors are added. First, I test that the selection function always returns a valid linear regression object. Using simulated data, I confirm that the output is of class "*lm*" and that the specified response variable appears correctly. Next, I test that the AIC-based selection rule is implemented correctly. Because the algorithm only adds predictors when doing so strictly improves the AIC, the final selected model should never have a higher AIC than the intercept-only model.

Results

The simulation results reveal two key findings. First, even when the true effect of temperature is zero, the stepwise procedure selects "*mean_t2m_F"* more than half of the time. Second, conditional on being selected, the standard OLS p-values for this variable are statistically significant. This provides clear evidence of post-selection inference failure: conditioning on selection fundamentally alters the distribution of the test statistic and the resulting p-values are not reliable.