S610 Final Project: An Application of Forward Stepwise Model Selection

Github: *https://github.com/Junying-T/S610-Toy-Repo/tree/main*

Junying Tong

Project Description

Prescribed burning is a land management strategy widely used in the United States. In the southeastern states, many local agencies employ a permitting system to regulate prescribed burn practices, with the primary policy goals of mitigating smoke dispersion and reducing the risk of escaped fires. In practice, prescribed burn permits are approved largely on the basis of fire weather conditions. In other words, prescribed burns are allowed only on days when meteorological conditions are deemed suitable for safe burning.

This project examines the correlational relationship between fire weather conditions—such as transport wind speed, relative humidity, dispersion index, and surface temperature—and the number of prescribed burn incidences in Georgia and Florida from 2010 to 2017. The analysis uses historical weather data from the ERA5 reanalysis dataset and geospatial fire activity data from the Environmental Protection Agency.

I employ a forward stepwise model selection approach to identify a subset of meteorological variables that best predict the number of prescribed burns. Through this approach, the project has two main objectives. First, I aim to identify the “best” combination of fire weather parameters for estimating prescribed burn incidence. Second, I examine the consequences of model selection for statistical inference, with particular attention to how forward stepwise selection affects the validity of conventional p-values and confidence intervals for the selected parameters.

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.

Code Validation and Testing

Because the forward stepwise selection procedure used in this project is manually implemented rather than relying on a built-in function, it is important to verify that the algorithm behaves exactly as intended. To address this concern, 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.

More specifically, the tests are designed to validate two core aspects of the forward stepwise selection procedure. First, they verify that the function reliably produces a well-defined linear regression model with the correct structure. Second, they confirm that the AIC-based selection rule is implemented correctly, such that predictors are added only when doing so improves model fit relative to a null benchmark. Together, these tests provide assurance that the behavior observed in the simulation study reflects genuine statistical properties of stepwise selection, rather than coding errors or unintended behavior.

Results

Although I only use surface temperature as a single demonstration for all other parameters, the simulation results reveal two key findings in common. First, even when the true effect of surface temperature is zero, the stepwise procedure selects "*mean_t2m_F"* more than half of the time in total repetitions. Which shows the selection of the variable actually does not render evidence of strong correlations between the parameter of interest and the dependent variable. Second, conditional on being selected, the standard OLS p-values for this variable are always statistically significant. This provides clear evidence of post-selection inference failure: conditioning on selection fundamentally alters the distribution of the test statistics and the resulting p-values are not reliable for a referential interpretation.