Prof. H. Bozdogan	FALL Semester: Oct. 23, 2025
Stat-575: Time Series Analysis	Due on or before Friday Nov. 4, 2025

PROJECT #2 TIME SERIES REGRESSION MODEL

INSTRUCTIONS:

- Please SHOW ALL YOUR WORK. Please submit your write up, computational modules and all your graphs with the codes in A ZIPPED FILE such as JOHN_DOE_STAT575_Proj#2.zip to CANVAS.

- You can use any computational software such as Matlab, R, JMP, Python, and SAS to be able to answer the following questions.

OBJECTIVE

The purpose of this project is to analyze a self-selected **time series regression dataset** and to perform **subset selection of predictor variables** using advanced covariance estimation methods. Students will estimate models using heteroskedasticity- and autocorrelation-consistent (HAC) covariance estimators and compare model performance using three information-theoretic criteria:

AIC, SBC (BIC), and CICOMP (Consistent Information Complexity).

Many time series data sets typically contains autocorrelation and/or heteroskedasticity of unknown form and for statistical inference to model such data sets it is essential to use ***covariance matrix estimators*** that can consistently estimate the covariance of the model parameters. For this reason, in the literature, there are several suitable ***heteroskedasticity consistent*** (***HC***) and ***heteroskedasticity and autocorrelation consistent*** (***HAC***) ***estimators*** have been proposed over the last 20 years.

Dataset Requirements

Each student must provide or construct their own dataset satisfying:

- The dataset must be in either `.csv` or `.xlsx` format.
- It should include a numeric response variable y_t (time series) and a model matrix $\mathbf{X}_t$ of predictor variables.
- The number of predictors must not exceed $k = 10$.
- The number of observations n should be sufficiently large ($n > k$) to estimate all candidate models.
- Variables should be labeled clearly and described in a summary table (see Table below).

Modeling Framework

You will fit the time series regression model

$$y_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{tj} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma^2),$$

and estimate the covariance of the OLS coefficient vector $\hat{\beta}$ using a HAC estimator:

$$\mathrm{Cov}_{\mathrm{HAC}} = \hat{\Sigma}_{\mathrm{HAC}}(w),$$

where $w(\cdot)$ is the kernel weight function:

$$\mathrm{Kernel} \in \{\mathrm{TR}, \mathrm{BT}, \mathrm{PZ}, \mathrm{TH}, \mathrm{QS}\}.$$

Here TR = Tukey, BT = Bartlett, PZ = Parzen, TH = Tukeyâ€“Hanning, and QS = Quadratic Spectral.

Implementation Requirement: Each student must use the provided MATLAB module:
`dr_TS_AllSubsets_HAC_TOP_EXPORT.m`
to analyze their own dataset. This script will:

- Enumerate all possible subsets of predictors (2^p models, $p \leq 10$);
- Fit each subset model via OLS and compute HAC covariance matrices;
- Calculate AIC, SBC, and CICOMP for each subset;
- Export a **`.csv`**, **`.tex`**, and **`.rtf`** summary of the best model(s);
- Generate comparative plots illustrating model complexity vs. information criterion.
- Which Kernel Covariance is the best choice for your dataset? Hint: You can compare the score of the criteria on all the variables, saturated model.

Information Criteria

For each subset model, compute the following:

$$AIC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2k,$$

$$SBC = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + k\log n,$$

$$CICOMP = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + k + 2\log(n)C_{1F},$$

where C_{1F} is the covariance complexity term derived from the eigenvalues $\lambda_1, \ldots, \lambda_k$ of the HAC covariance matrix of $\hat{\beta}$:

$$C_{1F} = \frac{1}{4\bar{\lambda}^2} \sum_{i=1}^{k} (\lambda_i - \bar{\lambda})^2, \qquad \bar{\lambda} = \frac{1}{k} \sum_{i=1}^{k} \lambda_i.$$

Deliverables

Each student must submit a concise technical report containing:

1. **Introduction and Data Description**
   - Dataset source and description;
   - Table of variables (see example below);
   - Summary statistics and exploratory plots.
2. **Modeling and Results**
   - Output from the module **`dr_TS_AllSubsets_HAC_TOP_EXPORT.m`**;
   - Tables of the best subsets under AIC, SBC, and CICOMP;
   - Visual comparison of each criterion across subset sizes.
3. **Discussion and Conclusions**
   - Comparison among AIC, SBC, and CICOMP results;
   - Discussion of model parsimony and covariance complexity;
   - Insights into the predictive and structural properties of the selected models.

Example Variable Table

Variable	Description	Type / Unit
y_t	Monthly retail sales	Thousands USD
x_{1t}	Consumer confidence index	Index (0â€“100)
x_{2t}	Average temperature	Â°C
x_{3t}	Unemployment rate	%
x_{4t}	Average gasoline price	USD per gallon

Submission Format

- Submit one compressed folder (**`.zip`**) containing:
  - Your dataset (**`.csv`** or **`.xlsx`**);
  - MATLAB code and output files from **`dr_TS_AllSubsets_HAC_TOP_EXPORT.m`**;
  - Exported **`.csv`**, **`.tex`**, and **`.rtf`** tables;
  - Final report in **`.pdf`** format (compiled from LaTeX or using MS Word).

- Include your name, course title, and date on the report title page.

Grading Rubric (100 points total)

Component	Description	Points
Data Description	Dataset quality, variable documentation	25
Modeling and Computation	Proper use of HAC module and criteria	40
Analysis and Interpretation	Discussion, model comparison, reasoning	25
Report Format	Clarity, tables, plots, and LaTeX quality	10
Total	□	**100**

This project applies Bozdogan's Information Complexity (ICOMP) framework to time series regression, illustrating the role of covariance regularization and model parsimony in modern model selection.

If you have any questions, please ask me and Haoqi our GTA.
If you need more time, please let me know.