Machine Learning Assignment CH3 slide33

Vincent Chen 陳昕佑 61375017H

A multi-model evaluation of probabilistic streamflow predictions via residual error modeling

Journal of Hydrology(2024)

Summary of Results:

1. **Performance Ranking of Models:**

   LSTM.lumped model was the best-performing model in deterministic (KGE) and probabilistic (CRPS) metrics across 141 basins. MESH.CLASS.Raven model consistently showed the poorest performance in both metrics.

   Intermediate models showed variable performance rankings between deterministic and probabilistic evaluations.

2. **Impact of Post-Processing:**

   The residual error model post-processing (REM-PP) did not significantly change the rankings of the best and worst models but altered rankings among middle-performing models. Post-processing could improve the CRPS performance for some deterministic

3. **General Findings:**

   Better deterministic models tend to have better probabilistic outcomes.

   REM-PP cannot fully compensate for poor deterministic model predictions.

   In basins with poor deterministic KGE, post-processing sometimes equalized performance across models, but not consistently.

Residual Error Modeling Approach Used:

- The authors employed a Residual Error Model Post-Processor (REM-PP) to convert deterministic streamflow predictions into probabilistic ones.
- The model used flow-dependent mean residuals ($\mu_t = \alpha + \beta \cdot z(q_t)$) and autocorrelated errors (η_t) modeled with the AR(1) process.
- The residual error model was calibrated using the Method-of-Moments (fast and no need for hydrological re-simulation).

The robust F-statistic as a test for weak instruments

Journal of Econometrics(2025)

Summary of Results:

1. **Generalization of Effective F-Statistic:**
   The paper extends the concept of the effective F-statistic (Montiel Olea and Pflueger, 2013) beyond 2SLS estimators to a wider class of linear GMM estimators.
2. **Robust F-statistic as Generalized Effective F-statistic:**
   The robust F-statistic is shown to belong to this class, and a new associated estimator called GMMf (GMM using first-stage residual-based weights) is introduced.
3. **Discrepancy between Fr and Feff:**
   The paper illustrates that a large robust F-statistic (Fr) does not imply strong instruments for 2SLS, but it does indicate strong instruments for GMMf.
4. **Bias-based Testing Framework:**
   The authors develop a generalized effective F-statistic-based weak instrument test for any GMM estimator, utilizing a Nagar bias approximation and proposing a benchmark harmonization using OLS bias.
5. **Critical Value Derivation:**
   Tests can be performed using Monte Carlo, Patnaik approximation, or a simplified conservative test for GMMf using chi-squared critical values.

Conclusion:

- Use Feff when interpreting 2SLS results (per MOP and Andrews et al. recommendations).
- Use Fr only if you're using the GMMf estimator, in which case Fr becomes a valid instrument strength diagnostic tool.
- In complex designs, GMMf with Fr-based testing is preferred over 2SLS with Feff-based testing due to better bias performance.