Discussion: Using Advanced AI methods to improve statistical estimation and official statistics at the IRS

Jae kwang Kim

Department of Statistics, Iowa State University

2024 Joint Statistical Meetings

Four presentations

1. A semi-supervised approach to anomaly detection for tax compliance (by Evan Schulz)
2. Estimating undetected income (by John Guyton)
3. Intermediate Outcomes (by Brandon Anderson)
4. Integrating probability and nonprobability data to estimate tax compliance: selection bias and measurement error (by Ishani Roy)

1. A semi-supervised approach to anomaly detection for tax compliance (by Evan Schulz)
2. Estimating undetected income (by John Guyton)
3. Intermediate Outcomes (by Brandon Anderson)
4. Integrating probability and nonprobability data to estimate tax compliance: selection bias and measurement error (by Ishani Roy)

1. Presentation by Evan Schulz

- **Problem:** IRS needs to detect anomalies in over 100 million tax returns efficiently and accurately
- **Approach:** Semi-supervised autoencoder model for anomaly detection in sparse tax form data
- **Methodology:**
  - Semi-supervised autoencoder with customized loss function
  - Two-model ensemble approach (M^+ and M^-)
  - Cube root transformation for data preprocessing

Key Findings

- Semi-supervised approach outperforms unsupervised and benchmark models
- Two-model ensemble often improves performance
- Cube root transformation enhances results
- Performance generalizes well across strata and datasets
- Relatively simple model architecture (5 hidden layers) is effective
- Good performance achieved with few epochs (5-20) and small η values

Discussion

- Data structure for semi-supervised learning

Sample	Type	X	Y
A	Unlabeled sample	✓	
B	Labeled sample	✓	✓

- Let

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise} \end{cases}$$

We assume that

$$Y \perp \delta \mid \mathbf{x}.$$

- The goal is to develop a model for $Y = -1$ (non-compliance): The parameter of interest θ is defined as the minimizer of

$$R(\theta) = E\{L(\theta; X, Y)\}$$

Discussion: A possible approach

- Estimate $\pi_i = P(\delta_i = 1 \mid \mathbf{x}_i)$ using a parametric working model.
- Construct the calibration weights for sample B by maximizing

$$Q(\omega) = \sum_{i \in B} \log(\omega_i)$$

subject to

$$\sum_{i \in B} \omega_i \mathbf{b}(\mathbf{x}_i) = \sum_{i \in U} \mathbf{b}(\mathbf{x}_i) \tag{1}$$

and

$$\sum_{i \in B} \omega_i \hat{\pi} = \sum_{i \in U} \hat{\pi}_i \tag{2}$$

where $U = A \cup B$ and $\mathbf{b}(\mathbf{x})$ is the control function for covariate balancing.

- The role of (1) is to incorporate the auxiliary information in the unlabeled data through covariance balancing in $\mathbf{b}(\mathbf{x})$.
- The role of (2) is to control the selection bias in the labeled sample.
- Once the final weight $\hat{\omega}_i$ are obtained from the calibration problem, we can estimate θ by minimizing

$$\sum_{i \in B} \hat{\omega}_i L\left(\theta; \mathbf{x}_i, y_i\right)$$

  with respect to θ.
- The optimal choice of control function is

$$\mathbf{b}^*(\mathbf{x}) = E\left\{ \frac{\partial}{\partial \theta} L(\theta; \mathbf{x}, Y) \mid \mathbf{x} \right\}.$$

1. A semi-supervised approach to anomaly detection for tax compliance (by Evan Schulz)
2. Estimating undetected income (by John Guyton)
3. Intermediate Outcomes (by Brandon Anderson)
4. Integrating probability and nonprobability data to estimate tax compliance: selection bias and measurement error (by Ishani Roy)

2. Paper by John Guyton: Relative rate model

- Relative detection rates can be estimated, but absolute rates are not identified
- Statistical Approaches
  1. Detection Controlled Estimation (DCE)
     - Currently used by IRS
  2. Relative Rate Model (New approach)
     - Goals: Interpretability, ease of fitting, Bayesian framework compatibility
     - Models return-level adjustment as function of covariates and examiner skill
     - Uses GLM with Tweedie distribution
     - Incorporates hierarchical prior for examiner-specific effect

Generalized linear model with random effects

- The new approach can be viewed as a two-level model
  - Level One Model: For the j-th case in examiner i,

$$\log E(A_{ij} \mid \mathbf{x}_{ij}, \gamma_i) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \gamma_i \tag{3}$$

    where $\mathbf{x}_{ij}$ is the case-specific covariates and γ_i is the examiner-specific effect for examiner i
  - Level Two model

$$\gamma_i \sim N\left(\mathbf{z}_i'\boldsymbol{\alpha}, \sigma^2\right)$$

    where $\mathbf{z}_i$ is the covariate for the examiners (e.g. years of experience, gender).
- As noted by the authors, a version of EM algorithm can be used to estimate the model parameters.

Discussion

- If Y_j is the true income for case j, then

$$A_{ij} = Y_j \times R_i$$

  where R_i is the average detection rate for examiner i
- Instead of using the two-level model on A_{ij}, we can consider the following models:
  - Model for Y_j: log-normal regression model

$$\log(Y_j) \sim N\left(\mathbf{x}_j'\boldsymbol{\beta}, \sigma_e^2\right)$$

  - Model for R_i: Beta regression model with covariate $\mathbf{z}_i$
  - Instead of observing (Y_j, R_i), we only observe $A_{ij} = Y_j \cdot R_i$.
  - As long as we have multiple observations in each interviewer, the model can be identified.

1. A semi-supervised approach to anomaly detection for tax compliance (by Evan Schulz)
2. Estimating undetected income (by John Guyton)
3. Intermediate Outcomes (by Brandon Anderson)
4. Integrating probability and nonprobability data to estimate tax compliance: selection bias and measurement error (by Ishani Roy)

3. Presentation by Brandon Anderson

- Key Concepts
  - Risk-Based Exam Selection
  - Feedback Loop
  - Time-to-Insight
  - Concept Drift
  - Intermediate Outcomes
- Challenge
  - Long time-to-insight for examination outcomes
  - Models making selections without accounting for concept drift
  - Need for improved feedback loop

Proposed Approaches

1. Imputation
   - Synthesizing labels for open cases
   - Incorporating into training data
2. Meta-Learning
   - Multi-target model predicting all outcomes
   - Considers time elapsed and closure status

Discussion

- Adaptive nature of sampling based on updated model is also called active learning.
- When applying the active learning, the main idea is to sample the element with minimum model variance under the current model.
- In the context of tax audits, the goal seems to be selecting the element with the higest risk.
- How to compromise the conflicting goal will be an important area of research.

1. A semi-supervised approach to anomaly detection for tax compliance (by Evan Schulz)
2. Estimating undetected income (by John Guyton)
3. Intermediate Outcomes (by Brandon Anderson)
4. Integrating probability and nonprobability data to estimate tax compliance: selection bias and measurement error (by Ishani Roy)

4. Presentation by Ishani Roy

- Wish to combine information from two sources.
  - NRP data: probability sample data, measures X and Y, expensive to measure, relatively small size
  - OP data: non-probability sample data, measure X and Y^*, cheap, large scale data
- Two problems with OP data
  1. Selection bias (with unknown selection probability): need a model assumption
  2. Measurement error in Y_i^*: may assume

$$Y_i^* = Y_i + u_i \tag{4}$$

where $u_i \sim (\alpha, \sigma_u^2)$.

- The goal is to estimate θ in the outcome model

$$Y_i = g(X_i; \theta) + e_i \tag{5}$$

  where $e_i \sim (0, \sigma_e^2)$.
- If we can estimate the selection probability for OP data and construct the pseudo weights $\hat{\pi}_{i,\mathrm{OP}}^{-1}$ in the optimization for estimating θ.
- A novel use of ABC method is developed to reflect the uncertainty in estimating $\hat{\theta}$

Discussion

- Under model (4) and (5), we can estimate θ by finding the minimizer of

$$\begin{aligned} Q(\theta, \alpha) &= \sum_{i \in S_{\mathrm{NRP}}} \frac{1}{\pi_{i,\mathrm{NRP}}} \left\{ y_i - g(x_i; \theta) \right\}^2 \frac{1}{\sigma_e^2} \\ &\quad + \sum_{i \in S_{\mathrm{OP}}} \frac{1}{\hat{\pi}_{i,\mathrm{OP}}} \left\{ y_i - \alpha - g(x_i; \theta) \right\}^2 \frac{1}{\sigma_e^2 + \sigma_u^2} \end{aligned}$$

- Wiring $\tau = \sigma_u^2 / \sigma_e^2$, we can minimize

$$\begin{aligned} Q^*(\theta, \alpha \mid \tau) &= \sum_{i \in S_{\mathrm{NRP}}} \frac{1}{\pi_{i,\mathrm{NRP}}} \left\{ y_i - g(x_i; \theta) \right\}^2 \\ &\quad + \sum_{i \in S_{\mathrm{OP}}} \frac{1}{\hat{\pi}_{i,\mathrm{OP}}} \left\{ y_i - \alpha - g(x_i; \theta) \right\}^2 \frac{1}{1 + \tau} \end{aligned}$$

- May use 10-fold cross validation to determine $\tau > 0$.