Variable selection in small area level models subject to sampling errors

Jiraphan Suntornchost
Department of Mathematics and Computer Science,
Faculty of Science, Chulalongkorn University

Based on joint work with
Partha Lahiri, University of Maryland, College Park
Nuchjaree Lenghoe, Chulalongkorn University
Annop Angkunsit, Chulalongkorn University

Workshop on Statistical Data Integration
August 7, 2019
Outline

- Introduction to Area Level Models
  - Fay-Herriot Model
  - Panel extension of Fay-Herriot Model
  - Multivariate Fay-Herriot Model
- Variable Selection for Area Level Models
  - Effect of Sampling Errors to Variable Selection Criteria
  - A proposed methodology
  - Simulation study and Data analysis
  - R functions for the proposed method
The traditional methods of direct estimation are done based on survey data.

The small area model is used as an alternative to the direct estimation by auxiliary information through the model.

The advantage of small area model is strongly based on the choice of auxiliary variables included in the model.

The models in small area estimation are classified into two broad types: Unit Level model and Area Level Model.
  - The unit level models require micro data that has information at individual level.
  - The area level models require information at the area level.

The are level models are widely used in applications when the micro data are not available due to confidentiality.
The response variable is usually collected from a small survey when sampling error occurs.

While the auxiliary variables come from either census, big survey or administrative records.

The goal is to predict response variable using auxiliary information from other sources.

Many area level models have been considered in literature:

- Fay-Herriot Model
- Panel extension of Fay-Herriot Model
- Multivariate Fay-Herriot Model
Basic Area Level Model

**Linking model**

\[ \theta_i = x_i^T \beta + v_i \ (i = 1, \ldots, m), \]

where

- \( \theta_i \) is the unobserved true mean for small area \( i \);
- \( v_i \) is the random small area specific error assumed to be iid with mean 0 and variance \( A \).

**Sampling Model**

\[ y_i = \theta_i + e_i \ (i = 1, \ldots, m), \]

where \( \{e_i \ , i = 1, \ldots, m\} \) are independent with means 0’s and *known* sampling variances \( D_i \)’s.
Under normality, this two-level model reduces to the well-known Fay-Herriot model (see Fay and Herriot, 1979) who used it to estimate the per-capita income of small geographical areas using complex survey and administrative records.

The assumption of known sampling variances $D_i$'s often follows from the asymptotic variances of transformed direct estimates (Efron and Morris 1975; Carter and Rolph 1974) and/or from empirical variance modeling (Fay and Herriot 1979).
The Fay-Herriot Models can be extended to accommodate multiple year data as follows.

\[ \theta_{it} = \alpha_i + x_{it}' \beta + u_{it}, \quad i = 1, \ldots, N, \]
\[ t = 1, \ldots, T, \]

where

- \( \theta_{it} \) is the response variable for the individual \( i \) at time \( t \),
- \( x_{it}' = (x_{it1}, \ldots, x_{itk}) \) is a \( k \times 1 \) vector of known auxiliary variables,
- \( \alpha_i \) is an individual effect,
- \( \beta = (\beta_1, \beta_2, \ldots, \beta_k)' \) is a \( k \times 1 \) vector of regression coefficients,
- \( u_{it} \) is the error term.
Linking model

\[ \theta_{it} = \alpha_i + \chi_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \]
\[ t = 1, \ldots, T, \]

where \( \theta_{it} \) is the response variable of interest for unit \( i \) at time \( t \).

Sampling Model

\[ y_{it} = \theta_{it} + e_{it}, \quad i = 1, \ldots, N, \]
\[ t = 1, \ldots, T, \]

where \( e_{it} \) is independent over \( i \) and \( t \), with mean zero and known sampling variances \( D_{it} \).
The Fay-Herriot Model can be extended to multi-variate model to take into account of correlations between several variables.

The Multivariate Fay-Herriot was introduced in small area estimation such as Datta et al. (1996), Gonzalez-Manteiga et al (2008) and Benavent and Morales (2016).

For example, Ubaidillah et al. (2019) applied the Multivariate Fay-Herriot to model household consumption per capita of food and non-food items.
Multivariate Fay-Herriot Model

**Linking model**

\[ \theta = X\beta + V, \]

where

- \( \theta = [\theta_1, \theta_2, \ldots, \theta_q] \) is an \( m \times q \) matrix of \( q \) unobserved response variables of \( m \) individuals
- \( V = [v_1, v_2, \ldots, v_q] \) is a matrix of error terms assumed that \( v_i \sim N_m(0, \sigma^2I) \) \((i = 1, \ldots, q)\) are random vector.

**Sampling model**

\[ Y = \theta + E, \]

where \( E = (e_{ij}) \) is an \( m \times q \) v of sampling errors assumed that \( e_{ij} \sim N(0, D_{ij}) \) \((i = 1, \ldots, m, j = 1, \ldots, q)\) are independent over \( i, j \).
Variable Selection for linear regression models
### Some commonly use model selection criteria

- **Root Mean Square Error (RMSE):**
  \[
  RMSE = \sqrt{MSE};
  \]

- **Adjusted \( R^2 \):**
  \[
  R^2 = 1 - \frac{MSE}{MST};
  \]

- **AIC:**
  \[
  AIC = 2p + m \cdot \log \left( \frac{SSE}{m} \right);
  \]

- **BIC:**
  \[
  BIC = p \log(m) + m \cdot \log \left( \frac{SSE}{m} \right);
  \]

where

- \( p \) is the number of auxiliary variables and \( m \) is the sample size,
- \( SSE = y^T[I - P]y, \ P = X(X^TX)^{-1}X^T \).
- \( MSE = \frac{SSE}{m-p} \)
- \( SST = y^T[I - m^{-1}11^T]y, \ MST = \frac{SST}{m-1} \).

### Remark

*Note that each of these model selection criteria can be expressed as \( f(MSE, MST) \), a smooth function of MSE and MST.*
Effect of sampling errors on standard regression model selection criteria

- investigate the extent of error in approximating $f(MSE_\theta, MST_\theta)$ by $f(MSE, MST)$ in presence of sampling errors.

- study the properties conditional on $\theta$. 
First note that

\[
E[MSE \mid \theta] = \frac{1}{m-p} E \left[ y' (I - X(X'X)^{-1}X') y \mid \theta \right]
\]

\[
= \frac{1}{m-p} E \left[ (\theta + e)' (I - X(X'X)^{-1}X') (\theta + e) \mid \theta \right]
\]

\[
= \frac{1}{m-p} \left[ SSE_\theta - 2 \theta' (I - X(X'X)^{-1}X') E[e] + E \left[ e' (I - X(X'X)^{-1}X') e \right] \right]
\]

\[
= MSE_\theta + \frac{1}{m-p} E \left[ e' (I - X(X'X)^{-1}X') e \right]
\]

\[
= MSE_\theta + \frac{1}{m-p} \text{tr} \left\{ (I - X(X'X)^{-1}X') \Sigma_e \right\}
\]

\[
= MSE_\theta + \frac{\sum_{i=1}^{m}(1 - h_{ii}) D_i}{m-p}
\]

\[
= MSE_\theta + D_{w1}.
\]

Therefore, MSE is a biased estimator of \( MSE_\theta \).
Similarly, we can show that $MST$ is a biased estimator of $MST_\theta$. That is

$$E(MST|\theta) = MST_\theta + D_{w2},$$

where

- $D_{w1} = \frac{\sum_{i=1}^{m}(1-h_{ii})D_i}{m-p},$
- $h_{ii} = x_i^T(X^TX)^{-1}x_i,$ and
- $D_{w2} = m^{-1}\sum_{i=1}^{m}D_i.$
Since under standard regularity conditions

\[ \text{var}(MSE|\theta) = O(m^{-1}), \]
\[ \text{var}(MST|\theta) = O(m^{-1}), \]

we have

\[ MSE - [MSE_\theta + D_{w1}] = o_p(1), \]
\[ MST - [MST_\theta + D_{w2}] = o_p(1). \]

Thus,

\[
\begin{align*}
&f(MSE, MST) - f(MSE_\theta, MST_\theta) \\
&= f(MSE_\theta + D_{w1}, MST_\theta + D_{w2}) - f(MSE_\theta, MST_\theta) + o_p(1).
\end{align*}
\]

Even for large \( m \), the leading term does not even tend to zero if \( D_i \)'s are uniformly bounded.
Example: Adjusted $R^2$

$$R_{adj; y}^2 - R_{adj; \theta}^2 = \frac{MSE_\theta}{MST_\theta} - \frac{MSE_\theta + D_{w1}}{MST_\theta + D_{w2}} + o_p(1),$$
Example: Adjusted $R^2$

Figure 1: Plot of $\text{Adj}R^2$ for case $m = 775$.

Remark

The biases are caused from sampling errors. Therefore, we propose to replace MSE and MST in variable selection criteria statistics by unbiased estimators of $\text{MSE}_\theta$ and $\text{MST}_\theta$, respectively.
Example: Adjusted $R^2$

**Figure 1:** Plot of $Adj R^2$ for case $m = 775$.

**Remark**

The biases are caused from sampling errors. Therefore, we propose to replace $MSE$ and $MST$ in variable selection criteria statistics by unbiased estimators of $MSE_\theta$ and $MST_\theta$, respectively.
Inference on the proposed estimates.
Theorem (Lahiri and Suntornchost, 2015)

The unbiased and consistent estimators for \( \text{MSE}_\theta \) and \( \text{MST}_\theta \) are

\[
\begin{align*}
\text{MSE}_\theta &= \text{MSE} - D_{w1}, \\
\text{MST}_\theta &= \text{MST} - D_{w2},
\end{align*}
\]
Then we propose $f(\hat{\text{MSE}}_\theta, \hat{\text{MST}}_\theta)$ as a new model selection criterion for the linking model with unobserved dependent variable.

Then,

$$f(\hat{\text{MSE}}_\theta, \hat{\text{MST}}_\theta) - f(\text{MSE}_\theta, \text{MST}_\theta) = o_p(1),$$

under certain regularity conditions, for large $m$. 
• True $R^2 = 1 - \frac{MSE_{\theta}}{MST_{\theta}}$, the true adjusted $R^2$;

• Naive $R^2 = 1 - \frac{MSE}{MST}$, the standard adjusted $R^2$ that ignores the measurement errors in $y$;

• Proposed $R^2 = 1 - \frac{\hat{MSE}_{\theta}}{MST_{\theta}}$, an adjustment to naive $R^2$. 
Figure 2: Plot of $\text{Adj} R^2$ for case $m = 775$. 

Example: Adjusted $R^2$
Recall that $\widehat{\text{MSE}}_\theta$ and/or $\widehat{\text{MST}}_\theta$ could be negative, especially for small $m$.

$\widehat{\text{MSE}}_\theta = \text{MSE} - D_{w1}$,

$\widehat{\text{MST}}_\theta = \text{MST} - D_{w2}$.

Therefore, $f(\widehat{\text{MSE}}_\theta, \widehat{\text{MST}}_\theta)$ may go out of admissible range.

For example, the proposed adjusted $R^2$ may go out of the interval $(-\frac{p-1}{m-p}, 1)$ in which adjusted $R^2$ belongs.
To make adjustment to the estimates, we can apply the following two positive approximations of $f(x, y) = x - y$ derived based on a Taylor-Series expansion.

- The $h$-function (Chatterjee and Lahiri, 2007) defined as

$$h(x, y) = \frac{2x}{1 + \exp \left( \frac{2y}{x} \right)}.$$

- The $g$-function (Angkunsit and Suntornchost, 2018) defined as

$$g(x, y) = x + \frac{2x^3 \left( 1 - \exp \left( \left( \frac{y}{x} \right)^3 \right) \right)}{y^2 \left( 1 + \exp \left( \left( \frac{y}{x} \right)^3 \right) \right)}.$$ 

**Remark**

Angkunsit and Suntornchost (2018) showed that the $g$ function has smaller approximation error if $y < \sqrt{2}x$ while the $h$ function has smaller approximation error if $x\sqrt{2} < y$. 
Consequently, we consider the following alternative approximations of $MSE_\theta$ and $MST_\theta$:

\[
\widehat{MSE}_{\theta, gh} = \begin{cases} 
MSE_\theta & \text{if } MSE > D_{w1} \\
g(MSE, D_{w1}) & \text{if } D_{w1} \leq \sqrt{2}MSE \\
h(MSE, D_{w1}) & \text{otherwise.}
\end{cases}
\]  \hspace{1cm} (1)

\[
\widehat{MST}_{\theta, gh} = \begin{cases} 
MST_\theta & \text{if } MST > D_{w2} \\
g(MST, D_{w2}) & \text{if } D_{w2} \leq \sqrt{2}MST \\
h(MST, D_{w2}) & \text{otherwise.}
\end{cases}
\]  \hspace{1cm} (2)
We use the public-use data for the 775 U.S. largest counties from the 2005 Small Area Income and Poverty Estimates (SAIPE) program of the U.S. Census Bureau in order to compare different approximations to the true adjusted $R^2$.

For details on SAIPE, the readers are referred to Bell (1999) and the website:
First obtain the estimates of the regression coefficients $\beta$ and the model variance $A$ of the Fay-Herriot model.

Using these estimates and real $D_i$ and $x_i$, we generate 1000 samples using the Fay-Herriot model.
Simulation Study: Notations

In the tables and figures, we use the following notations:

- \( AdjR^2_{\text{true}} = 1 - \frac{MSE_\theta}{MST_\theta} \), the true adjusted \( R^2 \);
- \( AdjR^2_{\text{naive}} = 1 - \frac{MSE}{MST} \), the standard adjusted \( R^2 \) that ignores the sampling errors in \( y \);
- \( AdjR^2_{\text{unbiased}} = 1 - \frac{\widehat{MSE}_\theta}{\widehat{MST}_\theta} \), based on the unbiased estimator of \( MSE_\theta \) and \( MST_\theta \);
- \( AdjR^2_h = 1 - \frac{\widehat{MSE}_{\theta,h}}{\widehat{MST}_{\theta,h}} \), an adjustment to naive \( R^2 \) by the \( h \) function (Lahiri and Suntornchost, 2015)

\[
\widehat{MSE}_{\theta,h} = \begin{cases} 
\widehat{MSE}_\theta & \text{if } MSE > D_{w1} \\
h(MSE, D_{w1}) & \text{otherwise,}
\end{cases}
\]

and

\[
\widehat{MST}_{\theta,h} = \begin{cases} 
\widehat{MST}_\theta & \text{if } MST > D_{w2} \\
h(MST, D_{w2}) & \text{otherwise,}
\end{cases}
\]
\( \text{Adj} R^2_g = 1 - \frac{\hat{\text{MSE}}_{\theta,g}}{\text{MST}_{\theta,g}}, \) an adjustment to naive \( R^2 \) by the \( g \) function (Angkunsit and Suntornchost, 2018);

\[
\hat{\text{MSE}}_{\theta,g} = \begin{cases} 
\text{MSE}_\theta & \text{if } \text{MSE} > D_{w1} \\
g(\text{MSE}, D_{w1}) & \text{otherwise},
\end{cases}
\]

and

\[
\hat{\text{MST}}_{\theta,g} = \begin{cases} 
\text{MST}_\theta & \text{if } \text{MST} > D_{w2} \\
g(\text{MST}, D_{w2}) & \text{otherwise},
\end{cases}
\]

\( \text{Adj} R^2_{gh} = 1 - \frac{\hat{\text{MSE}}_{\theta,gh}}{\text{MST}_{\theta,gh}}, \) an adjustment to naive \( R^2 \) by the \( gh \) function;
A Case study: Case 1

Figure 3: Plot of $\text{Adj} R^2$ for simulated data when $A = 0.0045$
## A Case study: Case 1

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Adj} R^2_{\text{true}}$</td>
<td>0.9709</td>
<td>0.9744</td>
<td>0.9752</td>
<td>0.9760</td>
<td>0.9768</td>
<td>0.9794</td>
</tr>
<tr>
<td>$\text{Adj} R^2_{\text{naive}}$</td>
<td>0.8936</td>
<td>0.9030</td>
<td>0.8979</td>
<td>0.9003</td>
<td>0.9032</td>
<td>0.9155</td>
</tr>
<tr>
<td>$\text{Adj} R^2_{\text{unbiased}}$</td>
<td>0.9681</td>
<td>0.9809</td>
<td>0.9729</td>
<td>0.9742</td>
<td>0.9799</td>
<td>0.9907</td>
</tr>
<tr>
<td>$\text{Adj} R^2_{h}$</td>
<td>0.9681</td>
<td>0.9809</td>
<td>0.9729</td>
<td>0.9742</td>
<td>0.9799</td>
<td>0.9907</td>
</tr>
<tr>
<td>$\text{Adj} R^2_{g}$</td>
<td>0.9681</td>
<td>0.9809</td>
<td>0.9729</td>
<td>0.9742</td>
<td>0.9799</td>
<td>0.9907</td>
</tr>
<tr>
<td>$\text{Adj} R^2_{gh}$</td>
<td>0.9681</td>
<td>0.9809</td>
<td>0.9729</td>
<td>0.9742</td>
<td>0.9799</td>
<td>0.9907</td>
</tr>
</tbody>
</table>
Case study: Case 2

![Plot of Adj R-squared: Case A = min(D)](image)

**Figure 4:** Plot of $\text{Adj}R^2$ for simulated data when $A = \text{min}(D)$
# Case study: Case 2

## Table 1: Table of $\text{Adj}R^2$ for SAIPE 2005 data

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Adj}R^2_{\text{true}}$</td>
<td>0.8058</td>
<td>0.8229</td>
<td>0.8283</td>
<td>0.8340</td>
<td>0.8398</td>
<td>0.8608</td>
</tr>
<tr>
<td>$\text{Adj}R^2_{\text{naive}}$</td>
<td>0.7619</td>
<td>0.7890</td>
<td>0.7697</td>
<td>0.7730</td>
<td>0.7820</td>
<td>0.7989</td>
</tr>
<tr>
<td>$\text{Adj}R^2_{\text{unbiased}}$</td>
<td>0.8230</td>
<td>0.8433</td>
<td>0.8279</td>
<td>0.8311</td>
<td>0.8368</td>
<td>0.8588</td>
</tr>
<tr>
<td>$\text{Adj}R^2_{h}$</td>
<td>0.8230</td>
<td>0.8433</td>
<td>0.8279</td>
<td>0.8311</td>
<td>0.8368</td>
<td>0.8588</td>
</tr>
<tr>
<td>$\text{Adj}R^2_{g}$</td>
<td>0.8230</td>
<td>0.8433</td>
<td>0.8279</td>
<td>0.8311</td>
<td>0.8368</td>
<td>0.8588</td>
</tr>
<tr>
<td>$\text{Adj}R^2_{gh}$</td>
<td>0.8230</td>
<td>0.8433</td>
<td>0.8279</td>
<td>0.8311</td>
<td>0.8368</td>
<td>0.8588</td>
</tr>
</tbody>
</table>
Numerical results for BIC:

\[ \text{BIC} = p \log(m) + m \cdot \log \left( \frac{\text{SSE}}{m} \right); \]

Figure 5: Plot of \( \text{AdjR}^2 \) for simulated data when \( A = \min(D) \)
<table>
<thead>
<tr>
<th>Percentiles</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>-6397</td>
<td>-6314</td>
<td>-6291</td>
<td>-6264</td>
<td>-6238</td>
<td>-6156</td>
</tr>
<tr>
<td>NAIVE</td>
<td>-3185</td>
<td>-3188</td>
<td>-3202</td>
<td>-3250</td>
<td>-3238</td>
<td>-3118</td>
</tr>
<tr>
<td>UNBIASED</td>
<td>-7784</td>
<td>-4300</td>
<td>-4341</td>
<td>-4477</td>
<td>-4445</td>
<td>-5032</td>
</tr>
<tr>
<td>HFUNC</td>
<td>-7784</td>
<td>-5193</td>
<td>-5309</td>
<td>-5707</td>
<td>-5613</td>
<td>-5032</td>
</tr>
<tr>
<td>GFUNC</td>
<td>-7784</td>
<td>-5193</td>
<td>-5309</td>
<td>-5707</td>
<td>-5613</td>
<td>-5032</td>
</tr>
<tr>
<td>GHFUNC</td>
<td>-7784</td>
<td>-5193</td>
<td>-5309</td>
<td>-5707</td>
<td>-5613</td>
<td>-5032</td>
</tr>
</tbody>
</table>
Figure 6: Plot of BICs for SAIPE 2005 data
Remark

- The naive Adj $R^2$ always underestimates the true value.
- The naive BIC overestimates the true value.
- The cases where the unbiased estimates are negative occur when the sampling variances are large.
- When such cases occur, the positive approximation by the gh-function seems to be the best option.
We applied the new adjusted $R^2$ to SAIPE data on U.S. poverty produced by Bell and Franco (2017).

- The response variable is direct CPS estimated poverty rate.
- Following analysis studied in Erciulescu, Franco and Lahiri (2014), we consider the following auxiliary variables:
- The auxiliary variables considered in the model are as follows:
  - Pseudo-poverty rate tabulated from IRS tax data
  - Tax nonfiler rate
  - Food stamp participation proportion
  - Final GVF estimate of std. error of pseudo-poverty rate after it is updated iteratively in conjunction with REML estimation of the model.
The following table shows comparisons between the naive estimates of $\text{Adj } R^2$ and the proposed estimates.

<table>
<thead>
<tr>
<th>Year</th>
<th>Naive</th>
<th>GHFUNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0.6013</td>
<td>0.7817</td>
</tr>
<tr>
<td>1996</td>
<td>0.6724</td>
<td>0.8654</td>
</tr>
<tr>
<td>1997</td>
<td>0.5575</td>
<td>0.8287</td>
</tr>
<tr>
<td>1998</td>
<td>0.6824</td>
<td>0.9026</td>
</tr>
<tr>
<td>1999</td>
<td>0.4269</td>
<td>0.6269</td>
</tr>
<tr>
<td>2000</td>
<td>0.7381</td>
<td>0.9421</td>
</tr>
<tr>
<td>2001</td>
<td>0.7378</td>
<td>0.8850</td>
</tr>
<tr>
<td>2002</td>
<td>0.7517</td>
<td>0.9184</td>
</tr>
<tr>
<td>2003</td>
<td>0.7843</td>
<td>0.9498</td>
</tr>
<tr>
<td>2004</td>
<td>0.6799</td>
<td>0.8692</td>
</tr>
</tbody>
</table>
The following table shows comparisons between the naive estimates of BICs and the proposed estimates.

<table>
<thead>
<tr>
<th>Year</th>
<th>Naive</th>
<th>GHFUNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>160.2744</td>
<td>116.1010</td>
</tr>
<tr>
<td>1996</td>
<td>150.7533</td>
<td>92.37293</td>
</tr>
<tr>
<td>1997</td>
<td>147.5103</td>
<td>78.6694</td>
</tr>
<tr>
<td>1998</td>
<td>143.7648</td>
<td>69.0016</td>
</tr>
<tr>
<td>1999</td>
<td>156.6999</td>
<td>114.7383</td>
</tr>
<tr>
<td>2000</td>
<td>101.7486</td>
<td>12.09174</td>
</tr>
<tr>
<td>2001</td>
<td>116.0069</td>
<td>64.55192</td>
</tr>
<tr>
<td>2002</td>
<td>110.9827</td>
<td>43.8247</td>
</tr>
<tr>
<td>2003</td>
<td>105.0051</td>
<td>20.6914</td>
</tr>
<tr>
<td>2004</td>
<td>111.9652</td>
<td>53.5891</td>
</tr>
</tbody>
</table>
Variable Selection for Panel Linear Models
Panel data Model

- The same calculation can be applied to linear panel data to obtain new variable selection criteria.
- Recall that the panel Fay-Herriot model consists of
  - Linking Model: $\theta_{it} = \alpha_i + x_{it}^\prime \beta + u_{it}, \ i = 1, \ldots, N$
  - Sampling Model: $y_{it} = \theta_{it} + e_{it}, \ i = 1, \ldots, N$
- The model can be expressed as
  $$
  \begin{align*}
  \theta &= (I_N \otimes j) \alpha + X \beta + v \\
  y &= \theta + e,
  \end{align*}
  $$

  where
  - $\theta$ is an $NT \times 1$ vector of unobserved response variables,
  - $v$ is an $NT \times 1$ vector of error terms with mean zero and covariance matrix $\text{cov}(v) = \sigma^2 V$ when $\sigma^2$ is unknown and $V(= PP')$ is a known positive definite matrix and
  - $e$ is an $NT \times 1$ vector of sampling errors with sampling covariance matrix $\Sigma_e$.  

To obtain parameter estimations, we consider the transformed model

\[ Qy = Q(I_N \otimes j)\alpha + QX\beta + Qu, \]

where \( Q = I_N \otimes \left(I_T - \frac{1}{T} J_T\right) \).
The least square estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are

\[ \hat{\beta} = (X'QX)^{-1}X'Qy \]

where $Q = I_N \otimes \left( I_T - \frac{1}{T} J_T \right)$;

\[ \hat{\alpha} = \frac{1}{T} (I_N \otimes j)'(y - X\hat{\beta}) ; \]

\[ \hat{\sigma}^2 = \frac{SSE}{N(T - 1) - k} , \text{ where } SSE = y'(Q - QPQ)y. \]
Theorem (Lenghoe and Suntornchost, 2019)

The unbiased and consistent estimators for $\text{MSE}_\theta$ and $\text{MST}_\theta$ are defined as:

$$\widehat{\text{MSE}}_\theta = \text{MSE} - D_{s1},$$
$$\widehat{\text{MST}}_\theta = \text{MST} - D_{s2},$$

respectively, where $D_{s1} = \frac{1}{m - p} \text{tr} \left\{ (I - X(X'X)^{-1}X')\Sigma_e \right\},$

$D_{s2} = \frac{1}{m - 1} \text{tr} \left\{ (-m^{-1}(P^{-1})'JP^{-1}) \Sigma_e \right\},$ and

$\Sigma_e = \text{diag}(D_1, D_2, \ldots, D_m).$
Next, consider conditional variance,

\[
\text{Var}[\widehat{\text{MSE}}_{\theta} - \text{MSE}_{\theta} | \theta] = \text{Var}[\text{MSE} - D_{s1} | \theta]
\]

\[= \text{Var}[\text{MSE} | \theta]
\]

\[= \frac{1}{(NT - N - k)^2} \text{Var}[y'(Q - QPQ)y | \theta]
\]

\[= \frac{\sigma^4}{NT - N - k}
\]

\[= O((NT)^{-1}),
\]

and

\[
\text{Var}[\widehat{\text{MST}}_{\theta} - \text{MST}_{\theta} | \theta] = \text{Var}[\text{MST} - D_{s2} | \theta]
\]

\[= \text{Var}[\text{MST} | \theta]
\]

\[= \frac{1}{(NT - 1)^2} \text{Var}[y' \left( I_{NT} - \frac{1}{NT} J_{NT} \right) y | \theta]
\]

\[= O((NT)^{-1}).
\]
From conditional expectation and variance of statistics \( \hat{MSE}_\theta \) and \( \hat{MST}_\theta \), we can show that

\[
\hat{MSE}_\theta - MSE_\theta \xrightarrow{p} 0, \\
\hat{MST}_\theta - MST_\theta \xrightarrow{p} 0.
\]

We obtain,

\[
f(\hat{MSE}_\theta, \hat{MST}_\theta) - f(MSE_\theta, MST_\theta) \xrightarrow{p} 0.
\]
From the definition, the function \( f(MSE_\theta, MST_\theta) \) could produce negative \( \text{Adj} R^2 \) or undefined BIC if the sampling errors are too large.

Therefore, we can consider the following alternative approximations of \( MSE_\theta \) and \( MST_\theta \):

\[
\hat{MSE}_{\theta, gh} = \begin{cases} 
MSE_\theta & \text{if } MSE > D_{s1} \\
g(MSE, D_{s1}) & \text{if } D_{s1} \leq \sqrt{2}MSE \\
h(MSE, D_{s1}) & \text{otherwise}
\end{cases}
\]

\[
\hat{MST}_{\theta, gh} = \begin{cases} 
MST_\theta & \text{if } MST > D_{s2} \\
g(MST, D_{s2}) & \text{if } D_{s1} \leq \sqrt{2}MSE \\
h(MST, D_{s2}) & \text{otherwise}
\end{cases}
\]
In our simulation studies, we simulate data from SAIPE data on U.S. poverty produced by Bell and Franco (2017).

The simulation is done for a simple case where the regression covariance matrix is $\sigma^2 I$.

First obtain the estimates of the regression coefficients $\beta$ and the estimates $\hat{\sigma}^2$ of the model based on SAIPE data for 51 states in years 1995-2004.

Using these estimates and real sampling variances $D_{it}$ and auxiliary variables $X$, we generate 1000 samples using the panel Fay-Herriot model.
# Case study: Case 1

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AdjR^2_{true}$</td>
<td>0.9030</td>
<td>0.9132</td>
<td>0.9163</td>
<td>0.9205</td>
<td>0.9238</td>
<td>0.9353</td>
</tr>
<tr>
<td>$AdjR^2_{naive}$</td>
<td>0.7272</td>
<td>0.7083</td>
<td>0.6944</td>
<td>0.7309</td>
<td>0.7421</td>
<td>0.7276</td>
</tr>
<tr>
<td>$AdjR^2_{unbiased}$</td>
<td>0.9131</td>
<td>0.8815</td>
<td>0.8729</td>
<td>0.9099</td>
<td>0.9164</td>
<td>0.9108</td>
</tr>
<tr>
<td>$AdjR^2_h$</td>
<td>0.9131</td>
<td>0.8815</td>
<td>0.8729</td>
<td>0.9099</td>
<td>0.9164</td>
<td>0.9108</td>
</tr>
<tr>
<td>$AdjR^2_g$</td>
<td>0.9131</td>
<td>0.8815</td>
<td>0.8729</td>
<td>0.9099</td>
<td>0.9164</td>
<td>0.9108</td>
</tr>
<tr>
<td>$AdjR^2_{gh}$</td>
<td>0.9131</td>
<td>0.8815</td>
<td>0.8729</td>
<td>0.9099</td>
<td>0.9164</td>
<td>0.9108</td>
</tr>
</tbody>
</table>
Figure 7: Plot of $AdjR^2$ for simulated data when $A = \min(D)$
## Case study: Case 1

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>319</td>
<td>382</td>
<td>406</td>
<td>429</td>
<td>453</td>
<td>530</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1111</td>
<td>1100</td>
<td>1193</td>
<td>1177</td>
<td>1136</td>
<td>1158</td>
</tr>
<tr>
<td>ADJ</td>
<td>240</td>
<td>174</td>
<td>585</td>
<td>529</td>
<td>363</td>
<td>459</td>
</tr>
<tr>
<td>HFUNC</td>
<td>240</td>
<td>174</td>
<td>585</td>
<td>529</td>
<td>363</td>
<td>459</td>
</tr>
<tr>
<td>GFUNC</td>
<td>240</td>
<td>174</td>
<td>585</td>
<td>529</td>
<td>363</td>
<td>459</td>
</tr>
<tr>
<td>GHFUNC</td>
<td>240</td>
<td>174</td>
<td>585</td>
<td>529</td>
<td>363</td>
<td>459</td>
</tr>
</tbody>
</table>
Figure 8: Plot of BICs for simulated data.
Figure 9: Plot of $\text{Adj} R^2$ for simulated data when $A = \text{min}(D)$
Figure 10: Plot of BICs for simulated data when $A = \min(D)$
Variable Selection for Multivariate Linear Models
Recall that the multivariate regression considered here is in the form

\[ \theta = X\beta + V, \]

where \( \theta = [\theta_1, \theta_2, \ldots, \theta_q] \).

From the model, we can express the sum of squared errors and cross product matrix \( \text{SSE} \) and the total sum of squares and cross product matrix \( \text{SST} \) as quadratic forms in \( Y \) as follows

\[ \text{SSE} = Y'(I - X(X'X)^{-1}X')Y, \]
\[ \text{SST} = Y'(I - m^{-1}J)Y, \]

where \( J \) is an \( m \times m \) matrix of ones.
Following Barrett and Gray (1994), variable selection criteria based on the matrices of $MSE$ and $MST$ can be done by three approaches:

- Trace Criterion,
- Determinant Criterion,
- Minimum Largest eigenvalue Criterion,

In our study, we consider the simplest criterion, the trace criterion.

For example,

$$Adj R^2 = 1 - \frac{\text{tr}(MSE)}{\text{tr}(MST)} = 1 - \frac{\text{tr}(SSE)/(q(m-p))}{\text{tr}(SST)/(q(m-1))},$$

where $m$ is the sample size, $q$ is the number of response variables and $p$ is the number of covariates.
Theorem (Angkunsit and Suntornchost, 2018)

Define \( \text{tr}(\widehat{\text{MSE}}_\theta) \) and \( \text{tr}(\widehat{\text{MST}}_\theta) \) as

\[
\text{tr}(\widehat{\text{MSE}}_\theta) = \text{tr}(\text{MSE}) - \text{tr}(D_{w1})
\]

and

\[
\text{tr}(\widehat{\text{MST}}_\theta) = \text{tr}(\text{MST}) - \text{tr}(D_{w2}),
\]

respectively, where

\[
\text{tr}(D_{w1}) = \frac{1}{q(m-p)} \sum_{j=1}^{q} \sum_{i=1}^{m} (1 - x_i'(X'X)^{-1}x_i)D_{ij}
\]

and

\[
\text{tr}(D_{w2}) = \frac{1}{qm} \sum_{j=1}^{q} \sum_{i=1}^{m} D_{ij}.
\]

Then

1) \( \text{tr}(\widehat{\text{MSE}}_\theta) \) is an unbiased and consistent estimator of \( \text{tr}(\text{MSE}_\theta) \), and

2) \( \text{tr}(\widehat{\text{MST}}_\theta) \) is an unbiased and consistent estimator of \( \text{tr}(\text{MST}_\theta) \).
Figure 11: Plot of $\text{Adj} R^2$ for simulated data when $\sigma = 0.1874 \times 0.1$
Figure 12: Plot of BIC for simulated data when $\sigma = 0.1874 \times 0.1$
There are three functions in this file:

- **Var.Sel** produces different versions of Adjusted $R^2$ and BICs statistics for Fay-Herriot Model.

- **Var.Sel.Panel** produces different versions of Adjusted $R^2$ and BICs statistics for a panel extension of the Fay-Herriot Model.

- **Var.Sel.MLM** produces different versions of Adjusted $R^2$ and BICs statistics for multivariate Fay-Herriot Model.

To access the file, click [R functions for variable selection](#).
Usage of R-Functions

Fay-Herriot Model

\texttt{Var.Sel}(\texttt{y, X,D})

**Input:**

- \texttt{y} is the vector of direct estimates,
- \texttt{X} is the matrix of auxiliary variables,
- \texttt{D} is the vector of sampling variances.

**Output:** Adj $R^2$ and BICs computed from 5 methods: the naive estimate, unbiased estimate, the $h$-function approximation, the $g$ function, and the $gh$ approximation.

- **Naive**: the variable selection criteria computed from the naive estimates of $MSE$ and $MST$,
- **Unbiased**: the variable selection criteria computed from unbiased estimates of $MSE$ and $MST$,
- **Hfunc**: the approximation by the $h$ function,
- **Gfunc**: the approximation by the $g$ function,
- **GHfunc**: the approximation by the $gh$ function,
The output contains Adj $R^2$ and BICs computed from 5 methods: the naive estimate, unbiased estimate, the $h$-function approximation, the $g$ function, and the $gh$ approximation.

```
> Var.Sel(y, x, D)
$AdjR2
    Naive  Unbiased   Hfunc   Gfunc   GHfunc
1  0.6995753  0.8750141  0.8750141  0.8750141  0.8750141

$BIC
    Naive  Unbiased   Hfunc   Gfunc   GHfunc
1  160.7932  104.5916  104.5916  104.5916  104.5916
```
This presentation is based on:


THANK YOU

Jiraphan Suntornchost
Department of Mathematics and Computer Science,
Faculty of Science, Chulalongkorn University,
Bangkok, Thailand
Email: jiraphan.S@chula.ac.th