UStat Package Manual

July 21, 2025

1 Overview

This document provides some more detailed documentation for the functions to compute the $U-$statistic estimators of the variance-covariance and its sampling covariance proposed by Rose, Schellenberg, and Shem-Tov (2022). This package is available on PyPI here (link forthcoming).

1.1 Empirical setup

Consider a population of students indexed by i assigned to one of J possible teachers in time t. Also assume that teachers effects are constant across students. In this setup, define "observational" teacher effects on outcome A for student i as:

$$Y_{it}^A = \sum_j \alpha_j^A D_{ijt} + X_{it}'\Gamma + u_{it} \tag{1}$$

where $D_{ijt} = 1$ when student i is assigned to teacher j in time t. Using this, we can define the teacher-year level mean residual as:

$$\bar{Y}_{jt}^A = \frac{1}{n_{jt}^A} \sum_{i|j(i,t)=j} (Y_{it}^A - X_{it}'\hat{\Gamma}) = \alpha_j^A + \bar{v}_{jt} \tag{2}$$

We assume that the $\bar{v}_{jt}$ are uncorrelated across years, i.e. $E[\bar{v}_{jt}\bar{v}_{jt'}] = 0$ when $t \neq t'$, and that $E[\bar{v}_{jt}] = 0$. We can use this setup and definitions to derive estimators of $\text{Cov}(\alpha_j^A, \alpha_j^B)$ and its sampling variance, which are the main objects of interest and the focus of the functions below. Rose, Schellenberg, and Shem-Tov (2022) contains the complete setup and the required assumptions for the following estimators to estimate the variance of causal teacher effects.

2 Functions

2.1 varcovar

The 'ustat.varcovar(A, C)' function computes the unbiased covariance between two datasets A and C which contain the residuals $\bar{Y}_{jt}^A$ and $\bar{Y}_{jt}^C$. The function also supports weighted variance calculations (where each weight corresponds to a row of A and C) and weighting by year. Specifically, the function calculates any of the following:

(1) Unweighted:

$$\hat{C}_{unweighted} = \Big(\frac{J-1}{J}\Big)\frac{1}{J}\sum_{j=1}^{J}\binom{T_j}{2}^{-1}\sum_{t=1}^{T_j-1}\sum_{k=t+1}^{T_j}\bar{Y}_{jt}^A\bar{Y}_{jk}^C - \frac{2}{J^2}\sum_{j=1}^{J-1}\sum_{k>j}^{J}\bar{Y_j^A}\bar{Y_k^C} \tag{3}$$

(2) Weighting each individual

$$\hat{C}_w = \sum_{j=1}^{J}\binom{T_j}{2}^{-1}\tilde{w}_j(1-\tilde{w}_j)\sum_{t=1}^{T_j-1}\sum_{k=t+1}^{T_j}\bar{Y}_{jt}^A\bar{Y}_{jk}^C - 2\sum_{j=1}^{J-1}\sum_{k>j}^{J}\tilde{w}_j\bar{Y_j}^A\tilde{w}_k\bar{Y_k}^C \tag{4}$$

(2) Weighting each individual by years observed

$$\hat{C}_y = \sum_{j=1}^{J} \frac{\tilde{T}_j^{A\wedge C} - \tilde{T}_j^A \tilde{T}_j^C}{|T_j^{A\wedge C}|(|T_j^{A\wedge C}|-1)} \sum_{t\in T_j^{A\wedge C}} \sum_{k\neq t}^{k\in T_j^{A\wedge C}} \bar{Y}_{jt}^A \bar{Y}_{jk}^C - 2\sum_{j=1}^{J-1}\sum_{k>j}^{J} \tilde{w}_j \bar{Y_j}^A \tilde{w}_k \bar{Y_k}^C \tag{5}$$

where $\tilde{w}_j = w_j / \sum_{j=1}^J w_j$, $\tilde{T}_j^A = |T_j^A| \sum_{j=1} |T_j^A|$, and $|T_j^A|$ represents the number of time periods individual j is observed for outcome A.

Notes:

1. this function can yield negative variance estimates due to the debiasing procedure. Negative variance estimates occur when the variance of teacher means is close to 0.

2.1.1 Arguments

ustat_var.varcovar(A, C, w, yearWeighted=False, quiet=True)

1. A, C = two J-by-T arrays between which you want to calculate the variance-covariance. A, C can contain missing values (in the form of a Nan), and each row of A and C can have missings in different spots.

2. w = an array of length J containing weights for the rows of A, C. Used to compute a weighted variance-covariance.

3. yearWeighted = option to compute weights based on the number of time periods each row is observed. Supports missing values in the same way as A, C.

4. quiet = True/false on whether to report to user what type of variance was calculated and whether the panels were balanaced/unbalanced. Reporting messages suppressed by default.

2.1.2 Usage

```
import ustat_var as ustat
import numpy as np

# Data and weights
np.random.seed(48912)
n_teachers, n_time = 50, 10
X, Y = ustat.generate_test_data.generate_unique_nan_arrays(n_rows=n_teachers, n_cols=n_time,
    n_arrays=2, min_int=1, max_int=9, nan_prob=0.25, seed = 48912, balanced = False)

weights = np.random.exponential(size = n_teachers)

# Variance-covariance
ustat.varcovar(X, X) # Var(X)
ustat.varcovar(X, Y) # Cov(X, Y)

ustat.varcovar(X, X, w = weights) # weighted Var(X)
ustat.varcovar(X, Y, w = weights) # weighted Cov(X, Y)

ustat.varcovar(X, X, yearWeighted = True) # year weighted Var(X)
ustat.varcovar(X, Y, yearWeighted = True) # year weighted Cov(X, Y)
```

2.2 ustat_samp_covar

The 'ustat.ustat_samp_covar(A, B, C, D)' function computes the sampling covariance of $\text{Cov}(A, B)$ and $\text{Cov}(C, D)$. Note that we do not impose any logical cap on the sampling variance, meaning this function can yield sampling covariances-variances which imply correlations exceeding 1. Specifically, the function computes an estimator for:

$$
\begin{aligned}
&\mathrm{Cov}\left(\hat{\mathrm{Cov}}(a_j^A,a_j^B)-\mathrm{Cov}(a_j^A,a_j^B),\hat{\mathrm{Cov}}(a_j^C,a_j^D)-\mathrm{Cov}(a_j^C,a_j^D)\right)= \\
&\quad\sum_i \sigma_i^{AC}\Big(\sum_{k\neq i} C_{ik}^{AB}a_{j(k)}^{B}\Big)\Big(\sum_{k\neq i} C_{ik}^{CD}a_{j(k)}^{D}\Big)+\sum_i \sigma_i^{AD}\Big(\sum_{k\neq i} C_{ik}^{AB}a_{j(k)}^{B}\Big)\Big(\sum_{k\neq i} C_{ik}^{DC}a_{j(k)}^{C}\Big) \\
&\quad+\sum_i \sigma_i^{BC}\Big(\sum_{k\neq i} C_{ik}^{BA}a_{j(k)}^{A}\Big)\Big(\sum_{k\neq i} C_{ik}^{CD}a_{j(k)}^{D}\Big)+\sum_i \sigma_i^{BD}\Big(\sum_{k\neq i} C_{ik}^{BA}a_{j(k)}^{A}\Big)\Big(\sum_{k\neq i} C_{ik}^{DC}a_{j(k)}^{C}\Big)+ \\
&\qquad\qquad\sum_i \sigma_i^{AD}\sum_{k\neq i} C_{ik}^{AB}C_{ik}^{DC}\sigma_k^{BC}+\sum_i \sigma_i^{AC}\sum_{k\neq i} C_{ik}^{AB}C_{ik}^{CD}\sigma_k^{BD} \qquad (6)
\end{aligned}
$$

where σ_i^{AC} represents the covariance between A and C and

$$
C_{ik}^{AC}=\begin{cases}\frac{J-1}{J^2}\frac{1}{|T_j^A||T_j^C|-|T_j^A\cap T_j^C|} & \text{if } j(i)=j(k)\\ \frac{1}{J^2}\frac{-1}{|T_{j(i)}^A||T_{j(k)}^C|} & \text{if } j(i)\neq j(k)\end{cases}
$$

Notes:

1. the function computes *unbiased* estimators of the product-sums $\left(\sum_{k\neq i} C_{ik}^{AB}a_{j(k)}^{B}\right)\left(\sum_{k\neq i} C_{ik}^{CD}a_{j(k)}^{D}\right)$. As with the variance-covariance estimator embodied in varcovar(), this means estimated sampling variances *can* be negative, though this does not happen often.

2. the function accepts row-weights to compute teacher/individual level weighted sampling covariances. This enters the function through the C_{ik}^{AC} coefficients. Instead of being pre-multiplied by $(J-1)/J^2$ and $1/J^2$, they are pre-multiplied instead by $\tilde{w}_j(1-\tilde{w}_j)$ and $\tilde{w}_j{}^2$, where $\tilde{w}_j = w_j/\sum_{j=1}^T w_j$ and w_j represents the weight given to row/individual/teacher j.

2.2.1 Arguments

ustat_var.ustat_samp_covar(A, B, C, D)

1. A, B, C, D = four J-by-max(T_j) arrays. Each can contain missing values (in the form of a Nan), and each row of each array can contain missing values in different spots.

2. w = a J-by-1 array containing the weights to be applied to each row/individual/teacher of A, B, C, D.

2.2.2 Usage

```
import ustat_var as ustat
import numpy as np

# Data and weights
np.random.seed(48912)
n_teachers, n_time = 50, 10
A, B, C, D = ustat.generate_test_data.generate_unique_nan_arrays(n_rows=n_teachers,
    n_cols=n_time, n_arrays=4, min_int=1, max_int=9, nan_prob=0.25, seed = 48912, balanced =
    False)

# Compute
ustat.ustat_samp_covar(A, A, A, A) # Var(Var(A))
ustat.ustat_samp_covar(A, B, A, B) # Var(Cov(A, B))
ustat.ustat_samp_covar(A, B, C, D) # Cov(Cov(A, B), Cov(C, D))
```