

## **Supplemental Material: Replication Instructions**

Reference: “Analysis of Testing-Based Forward Model Selection”

Damian Kozbur

March 26, 2020

This document provides replication instructions for calculations in the main paper, “Analysis of Testing-Based Forward Model Selection.” This document contains the following sections.

- List of files provided
- Description of files
- Replication instructions
- References

### **List of files provided**

The following files are included in the replication directory.

```
cleandata.dta
Export_Peru_to_Matlab.do
fsel_hetero.m
hetero_se.m
lasso_CV.m
lasso_hetero.m
peru_matlab_export.csv
Simulation_FSel_2020_Controls.m
Simulation_FSel_2020_Prediction_Ext.m
stepwisefit_hetero.m
TBFMS_PERU_ANALYSIS.m
TBFMS_PERU_MASTER.m
```

### **File descriptions**

cleandata.dta

Dataset for Table 1. This file was originally used and created in the reference Hanna and Olken (2018). The data in this file were downloaded from the source : Instituto Nacional de Estadística e Informática (INEI), Encuesta Nacional de Hogares (ENAHOG). Additional details about creating cleandata.dta from the original data source can be found in the replication files made available for the Hanna and Olken (2018) reference.

`Export_Peru_to_Matlab.do`

This script creates a Matlab readable file called `peru_matlab_export.csv` from the input file `cleandata.dta`.

`fsel_hetero.m`

This script is the main TBFMS script for the simulation studies.

`hetero_se.m`

This script calculates heteroskedasticity robust standard errors.

`lasso_CV.m`

This script runs a version of Lasso with cross-validated penalty level.

`lasso_hetero.m`

This file runs a version of Lasso for heteroskedastic data.

`peru_matlab_export.csv`

This file is the output of `Export_Peru_to_Matlab.do`. This file is used as when running `TBFMS_PERU_MASTER.m`. It is a Matlab-readable version of the data in `cleandata.dta`.

`Simulation_FSel_2020_Controls.m`

This script is performs Simulation Study II in the main paper and produces Table 4.

`Simulation_FSel_2020_Prediction_Ext.m`

This script is performs Simulation Study I in the main paper and produces Table 3.

`stepwisefit_hetero.m`

This script is a modified version of the Matlab native function `stepwisefit`. It has been modified to incorporate heteroskedasticity robust standard errors as described in the main paper.

`TBFMS_PERU_ANALYSIS.m`

This script is a supporting script for producing Table 1.

`TBFMS_PERU_MASTER.m`

This script is the main script for producing Table 1.

### **Replication instructions**

To Replicate Table 1, run in Matlab `TBFMS_PERU_MASTER.m`.

To replicate Table 3, run in Matlab `Simulation_FSel_2020_Prediction_Ext.m`.

To replicate Table 4, run in Matlab `Simulation_FSel_2020_Controls.m`.

To create `preu_matlab_export.csv`, run in Stata `Export_Peru_to_Matlab.do`.

To create `cleandata.dta`, follow instructions in Hanna and Olken (2018).

All calculations reported in the main paper were performed in Matlab2018b and Stata/MP 14.1. All files above should be included in the working directory. For information on which scripts require which inputs, see file descriptions above, or more detailed comments within the files.

### **References**

R. Hanna and B. A. Olken. Universal basic incomes versus targeted transfers: Antipoverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–26, 2018.

Instituto Nacional de Estadística e Informática (INEI), Encuesta Nacional de Hogares (ENAHOG)