README for the Reproducibility Package of the analytical output of the manuscript "The Exposure of Workers to Artificial Intelligence in Low- and Middle-Income Countries"

Authors: Gabriel Demombynes, Jörg Langbein, and Michael Weber

Overview

This reproducibility package (RP) reproduces the analytical output (estimators, tables, and graphs) of the manuscript "The Exposure of Workers to Artificial Intelligence in Low- and Middle-Income Countries". Most of the data used in this work comes from the Global Labor Database (GLD). This is a harmonized database of surveys with a labor force module. GLD data is stored on a server managed by the GLD team. The team aims to use data sources we can share at least with World Bank colleagues whenever possible. For other researcher outside the World Bank it is possible to reproduce the datasets using the published do-files.[1] In addition to the GLD data, we make use of data for one country, the US, from the International Income Distribution Database (I2D2). I2D2 is another harmonization effort within the World Bank. Access to it is managed by different GPs/Bank units within the World Bank and data availability can be assessed via datalibweb. The data is only available to World Bank Group staff. Lastly, the datasets are combined with information from the AI occupation exposure index developed by Felten et al. (2021). Their data is publicly available at the journal webpage. This reproducible package can be rerun by adapting the path in the Master do-file.

Requirements

The code is run in Stata 18. Extra packages required for running the code are:

- iscogen - a package to translate between different ISCO codes
- Grstyle - a package for graph customization
- Wbopendata a package to retrieve World Bank data
- Estout - a package to export regression results
- Grc1leg – a package to multiple graphs

[1] For more information see the GLD Github: Welcome to the Global Labor Database Repository — Global Labor Database

The packages are installed in the master do-file before calling the different, specific do-files.

Data sources

Country	Data name	Data source	Date access
Bangladesh	BGD_2016_QLFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Bolivia	BOL_2021_ECE_V01_M_V02_A_GLD_ALL	GLD	Feb 2024
Brazil	BRA_2022_PNADC_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Chile	CHL_2017_CASEN_V01_M_V06_A_GLD_ALL	GLD	Feb 2024
Egypt	EGY_2019_LFS_V01_M_V02_A_GLD_ALL	GLD	Feb 2024
Ethiopia	ETH_2021_LFS_V01_M_V03_A_GLD_ALL	GLD	Feb 2024
Georgia	GEO_2022_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Gambia	GMB_2023_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Indonesia	IDN_2015_SAKERNAS_v01_M_v06_A_GLD_ALL	GLD	Feb 2024
India	IND_2022_PLFS_V01_M_V02_A_GLD_ALL	GLD	Feb 2024
Sri Lanka	LKA_2021_LFS_V01_M_V03_A_GLD_ALL	GLD	Feb 2024
Mexico	MEX_2020_ENOE_V01_M_V06_A_GLD_ALL	GLD	Feb 2024
Mongolia	MNG_2022_LFS_V01_M_V02_A_GLD_ALL	GLD	Feb 2024
Nepal	NPL_2017_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Pakistan	PAK_2020_LFS_V01_M_V04_A_GLD_ALL	GLD	Feb 2024
Philippines	PHL_2022_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Rwanda	RWA_2021_LFS_v01_M_v01_A_GLD_ALL	GLD	Feb 2024
Sierra Leone	SLE_2014_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Thailand	THA_2021_LFS-Q3_V01_M_V03_A_GLD_ALL	GLD	Feb 2024
Turkey	TUR_2019_HLFS_V01_M_V03_A_GLD_ALL	GLD	Feb 2024
Tanzania	TZA_2020_ILFS_V01_M_V03_A_GLD_ALL	GLD	Feb 2024
United States of America	USA_2018_I2D2_CPS	I2D2	Feb 2024
South Africa	ZAF_2020_QLFS_v02_M_v05_A_GLD_ALL	GLD	Feb 2024
Zambia	ZMB_2022_LFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024
Zimbabwe	ZWE_2022_QLFS_V01_M_V01_A_GLD_ALL	GLD	Feb 2024

Note: The GLD datasets from Armenia, Colombia, Tunisia are originally in the data but dropped during the process of data cleaning due to missing information, mostly on the occupation ISCO level.

Note that for the imputation of the electricity variable we also add the following, reduced form, I2D2 datasets:

Country	Original data name	Data source	Date access
Pakistan	PAK_2015_I2D2_PSLM_v01_M_v01_A	I2D2	Sep 2024
Sierra Leone	SLE_2011_I2D2_SLIHS	I2D2	Sep 2024
South Africa	ZAF_2014_I2D2_LCS	I2D2	Sep 2024

Zambia	ZMB_2015_I2D2_LCMS	I2D2	Sep 2024
Zimbabwe	ZWE_2017_I2D2_ICES	I2D2	Sep 2024

In addition to the I2D2 and GLD datasets, there are two other datasets used in this manuscript:

- Excel that provides a crosswalk from SOC to ISCO. It is named "ISCO SOC Crosswalk" and provided by the US Bureau of Labor Statistics.
- Dataset named "SOCAIOE". This is the Stata dataset that contains the published AIOE database from Felten et al. (2021). It contains information on SOC, job title as well as AIOE. This information is downloaded from their GitHub

Instructions for replication

1. Datasets are available for World Bank staff (except for GLD Egypt, which is restricted). They can be downloaded from the GLD server or datalibweb, together with the I2D2 datasets. Copy the GLD datasets in the folder: Data/GLD-Add. The I2D2 data can be found on datalibweb. Store this in the folder I2D2. The AIOE can be downloaded from the Felten et al. (2021) GitHub and the Excel from the US Bureau of Labor Statistics – see also Data sources.
2. Alternatively if you are a World Bank staff, feel free to contact any of the authors of this paper to get direct access to the datasets.
3. Use the master do-file that is saved I the "Code" folder to run the code. After specifying the path in the master do-file everything should run on its own. Runtime takes about 20 minutes.
4. All called do-files are stored in the folder "Code" and any used or created dataset can be found in the folder "Data"
5. Outputs are stored in the following two folders:
    a. Tables – Table 1 and Table 2 are in Sheet "All", Table 3, the rergression results, are in sheet "income level AIOE"
    b. Figures – This folder contains all figures

Literature

Felten, E., M. Raj, and Seamans, R. 2021. Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses. Strategic Management Journal 42 (12): 2195– 217.