

README — Predicting Debt Distress in Low-Income Countries (GHKR)

Contents

1. [Overview](#)
 2. [Data Availability](#)
 3. [Instructions for Replicators](#)
 4. [List of Exhibits](#)
 5. [Requirements](#)
 6. [Code Description](#)
 7. [Folder Structure](#)
-

Overview

This package reproduces the exhibits in "Predicting Debt Distress in Low-Income Countries" (Graf von Luckner, Horn, Kraay, Ramalho). It combines a Stata pipeline (data cleaning, merging, descriptive tables/figures, Tables A1/A2/C1) with two Python/Jupyter notebooks (`model_analyses _Mar2026.ipynb` for main-text Tables 1–4 and Figures 3–6; `appendix_exercises _Mar2026.ipynb` for the Appendix B/C tables and figures), plus a cluster-only notebook (`j_k_fold CV for model selection _Mar2026.ipynb`) whose output is provided and which need not be rerun.

Confidential data — please read. The empirical analysis relies in part on confidential data compiled by the IMF and the World Bank. In particular, coverage gaps in a number of otherwise-public series are filled from the confidential 2017 LIC DSF review database (see `do-files/data cleaning and preparation/merge all variables.do`), so the merged analysis database `databases/licdsf_final_database` contains confidential data points, as does `data input/DSF_2016_external_8May2017.dta`. We do not have approval to redistribute these sources.

Consequently, as in the 2024 reproducibility seal (RR_WLD_2024_110), the **code is written to run on the full data**, but the **public deposit ships the non-confidential inputs only**. `data input/` is included, less the fourteen confidential or restricted files listed under [Data Availability](#) — the raw confidential sources and the cleaned intermediates built from them; `databases/` ships the non-confidential debt-distress extract only. The public data reproduces the subset of exhibits that build on the provided debt-distress indicators and estimated-model result files; it does not reproduce exhibits that require the confidential predictor database. Full reproduction — verified internally by the reproducibility team — runs on the complete data.

Exhibit numbering has been aligned with the current manuscript throughout the package (July 2026): do-file names, header comments, notebook markdown/labels, and output filenames now carry the current exhibit numbers (e.g. table A2 _ ..., table C1 _ ..., table C3 _ ...). The [List of Exhibits](#) below is cross-checked against the current manuscript.

Data Availability

Our analysis partly draws on confidential data compiled by the IMF and the World Bank, so a full replication package cannot be made publicly available. This statement lists all raw data sources used and describes how the data have been adjusted (winsorized/normalized, gap-filled).

Rights

- ☐ **[CONFIRM]** The authors have legitimate access to and permission to use all data used in this manuscript.
- ☐ **[CONFIRM]** The authors have documented permission to redistribute/publish the data included in the *public* deposit — the non-confidential raw inputs and cleaned intermediates in `data input/`, the debt-distress indicators, and the estimated-model result files. The confidential inputs (2017 LIC DSF review database and its derivatives, the IMF domestic debt and Fund-arrangements files, and CPIA) are **not** redistributed.

Confidential vs. public data

Item	Status
<code>data input/DSF_2016_external_8May2017.dta</code> (2017 LIC DSF review database)	Confidential — not redistributed. Vintage 8 May 2017 (per filename); internal WB–IMF product — see the confidential-data statement above
<code>data input/dsf_2016_extract.dta</code> , <code>data input/2017dsfpredsamplerdata.dta</code>	Confidential — not redistributed (extracts of the 2017 review database)
<code>databases/licdsf_final_database.*</code> , <code>databases/licdsf_alldata.dta</code>	Confidential — not redistributed (contain <code>_old</code> gap-fills from the 2017 review database)
<code>data input/imf _ domestic debt database.dta</code> (IMF domestic debt database)	Confidential — not redistributed
<code>data input/Fund Arrangements 1952-present.xlsx</code> (IMF Fund Arrangements / MONA extract)	Confidential — not redistributed

Item	Status
data input/cpia_WDI.dta (World Bank CPIA)	Confidential — not redistributed
data input/FX_Reinhart_and_Rogoff.dta, data input/FGR_SovereignDefaults.xlsx, data input/DDR CR dataset_updated.xlsx	Not redistributed here. The authors will share these raw files with replicators on request — please contact us; they can also be obtained from the original providers listed in the source list below
All other raw sources in data input/ (WDI, WEO, IFS, PWT, Reinhart–Rogoff public debt, Abbas, Mauro, IDS, DPI, FRED, WGI, etc.)	Included in the public deposit
Cleaned intermediates built from the withheld files above — data input/domesticdebt.dta, depvar_imf.dta, fx_rates.dta, default_fgr.dta, depvar_domrestructurings.dta	Not redistributed. These are cleaned country-year copies of the restricted sources, so they are withheld on the same terms as the raw files they derive from. The corresponding clean *.do scripts are included, and regenerate each of them once the raw file is obtained
databases/non_confidential_data_extract.dta (debt-distress dependent variables, non-confidential extract)	Included in the public deposit. Some of its binary signals are <i>derived from</i> restricted sources — imf22 from the Fund-arrangements file, pd22 from the domestic-restructuring dataset, and the composites ed_event22, ed_event22_nomissing and depvar22 from those plus the FGR default dummy. They are 0/1 country-year indicators computed under the definitions in the Variables subsection below, not extracts of the underlying datasets, and the same country-year signal map is published in Figure A2 of the paper. The restricted sources themselves are withheld, as set out above

Item	Status
output/RHS_Selector_result_filtered_t1.csv, _t12345.csv (estimated models over all 559,872 specifications)	Included in the public deposit (aggregated model outputs; no raw data points)

Raw source list (for the file-by-file mapping the review requested)

Series codes and construction are in the "Variables" subsection below. Access dates are not listed.

Files marked  below are **not** shipped in the public deposit; → marks the three the authors will supply on request. See the confidential/public table above.

File(s) in data input/	Source
WDI series (via wdi_raw;  cpia_WDI.dta)	World Bank, World Development Indicators — https://databank.worldbank.org/source/world-development-indicators
WEO*.xlsx/.csv, dsi.dta, weo_publicdata.dta, WEO_RevGrants.xlsx	IMF World Economic Outlook (multiple vintages; dsi = WEO-derived debt-service-to-interest series)
IFS_full_annual.csv, IFS_fulldatabase.csv	IMF International Financial Statistics (incl. HPG_XDR, HPQ_XDR)
pwt1001.dta	Penn World Table 10.01 — Feenstra, Inklaar & Timmer (2015)
Inflation_annual_Ohnsorge et al.dta	Ha, Kose & Ohnsorge (2021), Global Database of Inflation (hcpi_a)
PCTOT_*.csv	Gruss & Kebhaj (2019), Commodity Terms of Trade, IMF WP 2019/021
FGR_SovereignDefaults.xlsx	Farah-Yacoub et al. (2023), sovereign default database
boc_boe_default_data.xlsx	Beers, Ndukwe & Charron (2024), BoC–BoE Sovereign Default Database
abbas _ historical public debt.xlsx, abbas _ public debt.dta	Abbas et al. (2010), IMF Historical Public Debt Database
Public Finances in Modern History Dec 2023.dta, mauro _ public finances in modern history.dta	Mauro et al. (2015), Public Finances in Modern History (2023 update)
RR_publicdebt_*.xlsx, → FX_Reinhart_and_Rogoff.dta	Reinhart & Rogoff public debt / exchange-rate database
HRT_HiddenDefaults_ReplicationFile.xlsx	Horn, Reinhart & Trebesch, Hidden Defaults

File(s) in data input/	Source
dpi2020_thin.dta	Cruz et al. (2021), Database of Political Institutions 2020 (yrsoffc, yrscurnt)
IDS_ArrearsData_Dec22.xlsx, Projected debt service _ from IDS.xlsx, ids _ debt data raw.dta, r _ interest on new commitments _ ids.xlsx	World Bank International Debt Statistics (IDS)
FRED _ US 10 year yield.xlsx	FRED, 10-Year Treasury constant maturity (DGS10)
Fund Arrangements 1952-present.xlsx	IMF Fund Arrangements / MONA
imf _ domestic debt database.dta, → DDR CR dataset_updated.xlsx	IMF (2021a), domestic debt / domestic debt restructuring database (67 LIC restructurings)
NPV data series 2015-2021 ...xlsx	World Bank IDS net-present-value series (5% discount rate)
wgidataset.dta	Worldwide Governance Indicators, World Bank — used by clean wgi government effectiveness.do in the cleaning pipeline

Variables (sources, series codes, construction)

Distress-series inputs

- **Outright default on private external creditors:** core source is Farah-Yacoub et al. (2023), tracking defaults on private external creditors across 200 countries and 200 years, with comprehensive LIC coverage from each country's year of independence; used to build default histories as predictors.
- **Arrears:** signal = 1 if total outstanding arrears exceed 5% of the total public and publicly guaranteed (PPG) external debt stock for at least three consecutive years. From World Bank IDS: arrears to official and private creditors on missed principal (DT.AXA.DLXF.CD) and missed interest (DT.IXA.DLXF.CD), divided by total PPG external debt (DT.DOD.DPPG.CD). HIPC treatment: all arrears to official creditors are deemed resolved once a country reaches its HIPC completion point, so completed-HIPC countries re-enter the estimation sample.
- **Large IMF balance-of-payments assistance:** signal = 1 if a sovereign draws more than 30% of quota in the first six months of a non-concessional IMF program. Program timing from Reinhart & Trebesch (2016), matched to monthly IFS net increase in fund credit from the general resource account (HPG_XDR), scaled by quota at program start (HPQ_XDR). Excludes concessional programs and Rapid Financing Instrument disbursements.

- **Domestic debt distress events:** from an IMF dataset of 67 domestic debt restructurings in LICs, 1970–2020 (IMF 2021a); mapped to defaults assuming (per Erce, Mallucci & Picarelli 2022) the default occurs one year prior to the restructuring.

Predictor variables

- **External debt stocks:** PPG external debt from World Bank IDS (DT.DOD.DPPG.DT), scaled by nominal GDP (WDI ny_gdp_mkt_p_cd, WEO NGDPD) and by exports of goods and services (WDI ne_exp_gnfs_cd, WEO BXS_BP6).
- **Debt service due:** interest + principal on PPG external debt due in year t ; from an internal World Bank IDS variable that projects year- $t+1$ service from the year- t outstanding stock (enters with a lead). Scaled by exports (WDI ne_exp_gnfs_cd, WEO BXGS_BP6) and by total general government revenue excluding grants (WEO GGR, GGRG).
- **Net present value of debt:** NPV of future interest and principal on disbursed and outstanding PPG external debt, from World Bank IDS (constant 5% discount rate); scaled by nominal GDP (ny_gdp_mkt_p_cd, NGDPD) and exports (ne_exp_gnfs_cd, BXS_BP6).
- **Total public debt:** starting from general government debt to GDP (WEO GGXWDG_NGDP), supplemented for pre-1990s with Abbas et al. (2010) and Reinhart & Rogoff (2011); scaled by nominal GDP (ny_gdp_mkt_p_cd, NGDPD).
- **Public debt interest payments:** from the 2023 update of Mauro et al. (2015), Public Finances in Modern History; coverage gaps filled from the IMF WEO (series dsi); scaled by exports.
- **CPIA:** World Bank Country Policy and Institutional Assessment (institutional quality).
- **GDP per capita:** log of nominal GDP / midyear population, WDI (NY.GDP.PCAP.CD).
- **Openness:** (exports + imports) / GDP; WDI (NE.IMP.GNFS.CD, NE.EXP.GNFS.CD), WEO (BMGS_BP6, BXGS_BP6); GDP from WDI/WEO (NY_GDP_MKTP_CD, NGDPD).
- **Credit history I:** years since the last external debt distress event (built 1960–2021; pre-1970 reflects only private-creditor defaults and large IMF assistance).
- **Credit history II:** following IMF (2021b), a unit impulse at each external distress event decaying geometrically at 10% (built 1960–2021, from year of independence). *(This is the second credit-history measure; in code it is credit_hist3.)*
- **GDP growth:** annual % growth of constant-LCU GDP; WEO (NGDP_RPCH), WDI (ny_gdp_mkt_p_kd).
- **Inflation:** annual % change in CPI (headline), from Ha et al. (2021) Global Database of Inflation (hapi_a).
- **World growth:** annual real GDP growth for the WEO "World" aggregate (NGDP_RPCH).
- **US interest rates:** market yield on 10-year constant-maturity US Treasuries, St. Louis Fed (DGS10).

- **Remittances:** personal remittances received / GDP; WDI (BX.TRF.PWKR.CD.DT), scaled by nominal GDP (NY.GDP.MKTP.CD).
- **Foreign exchange reserves:** total reserves / imports; WDI (FI.RES.TOTL.CD), scaled by imports (WDI.NE.IMP.GNFS.CD, WEO BMGS_BP6).
- **Foreign exchange rate:** annual % change in nominal FX (LCU per USD, period average); WEO (ENDA).
- **Foreign direct investment:** net FDI inflows / GDP; WDI (BX.KLT.DINV.WD.GD.ZS), WEO (BFD_GDP).
- **Current account balance:** current account / GDP; WDI (BN.CAB.XOKA.GD.ZS).
- **Commodities terms of trade:** country-specific ToT indices over 45 commodities; Gruss & Kebhaj (2019).
- **Years in office:** years the chief executive has been in office; Database of Political Institutions (Cruz et al. 2021, yrsoffc), 1975–2020, linearly extrapolated to early sample years.
- **Years left in current term:** years remaining in the chief executive's term; DPI (Cruz et al. 2021, yrscurnt), 1975–2020, extended under stated assumptions.

Algorithmic data adjustments

- **Filling data gaps from old vintages:** to maximize country/time coverage, data points from earlier vintages are used where later vintages drop them (e.g., IDS removes graduated countries). Gaps are filled using IMF and World Bank (2017) and Horn et al. (2023). *(These old-vintage/2017-review fills are the source of the confidentiality constraint above.)*
- **Winsorizing:** all predictor variables are winsorized at the 1st and 99th percentile unless naturally bounded. Exceptions: CPIA, GDP per capita, world growth, the US 10-year yield, and the credit-history variables are **not** winsorized; remittances, nominal GDP, and the debt stock/flow variables (naturally bounded below at zero) are winsorized only at the 99th percentile.

Instructions for Replicators

The package combines Stata and Python, each with its own entry point.

1. **Stata (do-files/).** Open do-files/master do file.do, set the single global main path to your local copy of this folder, and run. Three local toggles near the top — clean_all_files, merge_and_prepare, graphs_and_tables (defaults 0/0/1) — control what runs; leave the first two at 0 for a standard run (uses the provided intermediate databases). Required Stata packages must be installed first (see [Requirements](#)).
2. **model_analyses _Mar2026.ipynb.** Set BASE_PATH at the top (the only path to change).

- depvar selects the horizon: 't1' (one-year-ahead, default) or 't12345' (within the next five years). Output filenames carry the suffix so the two runs do not overwrite each other.
 - **apply_filter** selects which table the shared model-table cell produces: **0** → **Table 1** (top models, unconstrained), **1** → **Table 2** (top models under the sensibility filter). With depvar='t12345' the same cell produces **Table 3**.
 - RUN_HEAVY (default False) loads the precomputed result files; leave False for a standard run.
 - Reproduces Figures 3–6 and Tables 1–4.
3. **appendix_exercises_Mar2026.ipynb**. Same BASE_PATH/RUN_HEAVY pattern; run after model_analyses. Reproduces the Appendix B/C tables and Figures B1–B2, and writes the Excel workbooks in output/appendix_tables_excel/.
 4. **j_k_fold CV for model selection_Mar2026.ipynb** (cluster-only; input data confidential). Not rerun for a standard replication — its output (output/RHS_Selector_result_filtered_t1.csv, _t12345.csv) is provided. Included to document the model-selection algorithm.

Public deposit. The deposit ships data input/ less the fourteen withheld files (see [Data Availability](#)) plus the distress indicators and the result files. Two consequences. First, the exhibits that depend on the confidential predictor database do not run — this is expected; full reproduction requires the complete data. Second, clean_all_files = 1 and merge_and_prepare = 1 will stop at the first withheld input; leave both at 0, which is the default and which uses the provided intermediates. Replicators who want to rebuild the cleaning steps that depend on Reinhart–Rogoff FX, Farah-Yacoub et al., or the IMF domestic-restructuring dataset should contact the authors for those three raw files.

List of Exhibits

Reflects the current manuscript numbering and the current state of the code (including this round's fixes). "Needs full data" marks exhibits that require the confidential predictor database and therefore do not reproduce from the public deposit.

Exhibit	Generating script / cell	Note
Table 1	model_analyses, model-table cell, apply_filter = 0	Needs full data
Table 2	same cell, apply_filter = 1	Needs full data
Table 3	same cell, depvar = 't12345' + apply_filter = 1	Verified — reproduces the manuscript to the digit (N=821; manuscript shows n=2–7 and 10, omitting the code's n=8/9 columns). Needs both five-year

Exhibit	Generating script / cell	Note
		result files, RHS_Selector_result_filtered_t12345.csv and all_models_with_coeffs_t12345.csv (both now included). Known edge case: with this toggle combination the <i>Appendix-B1 diagnostics cell below Table 3</i> stops with a division-by-zero (no filtered five-year model retains the remittances trio); Table 3 itself prints above it. The plain depvar='t12345' run (default apply_filter=0) executes end-to-end. Needs full data for the coefficient re-estimation
Table 4	model_analyses, confusion-matrix cell	Needs full data
Figures 1, 2	do-files/analysis/figure 1 ...do, figure 2 ...do	Built on the distress indicators
Figures 3, 4, 5, 6	model_analyses	3/4/6 build on the result files; 5 (re-estimation in larger sample) needs full data
Table A1	do-files/analysis/table a1 _ country and time coverage.do	Sample coverage — exports output/Table A1 _ Sample coverage.xlsx (80 countries incl. Kiribati, which has zero observed years and is therefore omitted from the printed table's 79 rows; export added July 2026, previously in-memory only)
Table A2	do-files/analysis/table A2 _ incidence of debt distress episodes by signal.do	Distress signals and episodes (renamed from table 1 _ ...)
Figures A1–A7	do-files/analysis/	Descriptive
Table B1	appendix_exercises, "Appendix table B1" cell	"Credit History 2" row = second credit-history measure (credit_hist3) — see DAS
Table B2	appendix_exercises, Appendix Table B2 cell	Explicit cell added this round ; reproduces manuscript p.68 (N=821)
Tables B3–B9	appendix_exercises	B7–B9 read pre-computed .tex; regeneration needs RUN_HEAVY=True
Table B4	appendix_exercises, Table B4 cell	Reconstruction — the original bespoke code was not preserved; the cell reproduces the published table to the digit
Table B5	appendix_exercises, Table B5 cell	Reproduces the manuscript's Table B5 (July 2026, p.72) to the digit — every coefficient, star,

Exhibit	Generating script / cell	Note
		loss/error rate, N=1,575. (The manuscript table was updated in July 2026 to these regenerated values, superseding an earlier 2025 draft version whose generating session was not preserved.)
Figure B1	do-files/analysis/Appendix B _ Investigate gdp_pc.do	
Figure B2	appendix_exercises, LASSO cell	Corrected this round — the grey "Preferred" curve now reads the sensibility-filtered best-per-size set (top_models_with_coeffs), reproducing the published figure to the digit. Note: model_analyses also writes a reduced 9-row best-per-size version of top_models_with_coeffs_t1.csv; the appendix notebook's bootstrap cell regenerates the full filtered set from all_models_with_coeffs_t1.csv before this figure cell runs, so run the appendix top-to-bottom
Table C1	do-files/analysis/table C1 _ reproduce old probit regressions.do (+ Python cell)	Renamed from the stale table 4 _ ...filename. Needs full data
Table C2	—	Definitional (non-computational)
Table C3	appendix_exercises, Table C3 cell	Explicit cell added this round — 2017 LIC DSF sample confusion matrix (N=391), port of Stata table C3 _ confusion matrix BPM vs LIC DSF 2017 sample.do (formerly table 9.do). Needs the 2017 confidential sample

Requirements

Software

- **Stata 18 MP** (tested by the reviewers on StataNow 19.5 MP). Install first:

```

ssc install estout
ssc install binscatter
ssc install carryforward
ssc install labutil          // provides labmask
net install grc1leg2         // grc1leg is no longer installable via ssc

```
- **Python 3.12** (developed and run on 3.12.4) — see requirements.txt (pip) or environment.yml (conda), which pin the versions the code is actually developed

and run on: pandas 2.2.2 (< **3.0** — pandas 3.x breaks a `literal_eval` column-assignment step used throughout), numpy 1.26.4, scipy 1.13.1, statsmodels 0.14.2, scikit-learn 1.4.2, seaborn 0.13.2, matplotlib 3.8.4, lxml 5.2.1, openpyxl 3.1.2, multiprocessing 1.1.5, requests. (The reviewers verified reproduction on a newer stack — pandas 2.3.3 / numpy 2.5.1 — so both work; we pin the lower, author-tested versions.)

Computational

- The Stata master do-file (using the provided intermediate databases) and both analysis notebooks (`RUN_HEAVY = False`) each complete in a few minutes. The `RUN_HEAVY = True` branches and the `j_k_fold` notebook require a many-core machine and hours, and are not needed for a standard replication.
- Path separators in all do-files use forward slashes (converted July 2026), so the Stata pipeline runs on Windows, macOS, and Linux.
- **TIFF exports on macOS:** run the Stata master in batch via the GUI binary (`.../StataMP.app/Contents/MacOS/StataMP -e do ...`) — the console binary (`stata-mp`) lacks the `Graph2tif` translator needed for Figure 1/2's journal-resolution `.tif` exports (PNG/PDF exports work in both).

Code Description

- `do-files/master do file.do` — single Stata entry point; sets the path global and (per its toggles) cleans, merges, and produces the descriptive figures/tables and Tables A1/A2/C1.
 - `do-files/data cleaning and preparation/` — one script per raw input; merge all `variables.do` performs the merge, including the old-vintage/2017-review gap-fills.
 - `do-files/analysis/` — one script per exhibit (some filenames reflect old numbering; see List of Exhibits).
 - `python code/model_analyses _Mar2026.ipynb` — Tables 1–4, Figures 3–6.
 - `python code/appendix_exercises _Mar2026.ipynb` — Appendix B/C tables and figures; Excel workbooks.
 - `python code/j_k_fold CV for model selection _Mar2026.ipynb` — cluster-only model search; output provided.
 - `From confidential to public data version.do` — extracts the non-confidential debt-distress indicators from the merged database (`non_confidential_data_extract.dta`), i.e. the public data shipped in the deposit. Paths are `$main-relative` (updated July 2026).
-

Folder Structure

```
gkhr_replication folder ... /
├── README.md                ← this file
├── data input/              ← raw inputs and cleaned intermediates; the
public deposit ships all but the fourteen withheld files (see Data Availabili
ty)
├── databases/               ← merged/final analysis databases (confident
ial; public deposit ships non_confidential_data_extract.dta only)
├── do-files/
│   ├── master do file.do
│   ├── data cleaning and preparation/
│   └── analysis/
├── python code/             ← the three notebooks
└── output/                  ← figures, tables, appendix_tables_excel/, f
igures_pdf_tif/, and the RHS_Selector result files
```