BOTTOM CODING PROJECT - README

Date: February 19, 2026

TABLE OF CONTENTS

1. OVERVIEW

This document describes the structure and workflow of the Bottom Coding research project. The project estimates consumption/income floors using multiple methodologies and analyzes their impact on poverty and inequality measures.

Key Features:
- Multiple floor estimation methodologies (Ravallion, Beta, TPLN, Dagum III, LIS)
- Sensitivity analysis through DropX variants (dropping observations below $0.50 and $1.00 thresholds)
- Comprehensive poverty and inequality comparisons
- 34 analysis files generating 16 tables and 5 figures
- Fully automated workflow via master do file
- Total runtime: ~24 hours on Stata 18 MP

Data Sources:
- See Section 4 (Raw Data Files) for detail; some data is restricted
- Global Monitoring Database 2017 (microdata - primary source)
- Country classification data (regional/income groups)
- Harmonized national poverty lines
- World Development Indicators

- Food Prices for Nutrition database

Outputs:
- 16 main and appendix tables (Excel)
- 5 publication-ready figures (PNG)
- 14 intermediate datasets for analysis

Reproducing Results:
To reproduce all results in the paper:
1. Ensure you have Stata 18 MP (or compatible version) installed
2. Update the global path in 01-do/00_Master.do (line 10) to match your
   local directory structure
3. Run: do "01-do\00_Master.do"

The master file will automatically install all required SSC packages, execute
all 34 analysis files in the correct sequence, and generate all intermediate
data, tables, and figures. All outputs will be saved to the 04-output/ folder.
Total runtime is approximately 24 hours (see Section 8: WORKFLOW SEQUENCE for
recommended parallel execution to reduce runtime).

==
2. SYSTEM SPECIFICATIONS
==

Operating System: Microsoft Windows 11 Enterprise
System Type: x64-based PC
Processor: 13th Gen Intel(R) Core(TM) i7-1365U
    - Physical Cores: 10
    - Logical Processors: 12
Memory (RAM):   32.1 GB (32,100 MB)
Stata Version: Stata 18 MP (64-bit)

==
3. FOLDER STRUCTURE
==

01-do/   Stata do and ado files for data processing and analysis
02-rawdata/   Raw input data files
03-intermediatedata/   Intermediate processed datasets
04-output/   Final tables and figures

==
4. RAW DATA FILES (02-rawdata/)
==

Dataset: Global Monitoring Database 2017 – all surveys (microdata)
Filename: GMD_all_2017.dta
Citation: World Bank, Global Monitoring Database (GMD) (version September 2024).
Accessed September 2024.
Access Instructions:

    * World Bank staff can access most surveys in the GMD collection through the datalibweb Stata package
        using syntax such as: dlw, country(alb) year(2012) type(gmd) surveyid(LSMS) mod(all) cpivin(v09)
    * Installation instructions and download guidelines for the datalibweb package are available on the World Bank Datalibweb website.
        The file BottomCoding-SurveyList.xlsx lists all surveys used in the paper. The do-file 01-PrepData.do documents how the dataset
        GMD_all_2017.dta is constructed once access to the underlying surveys is granted.
        See the Supplementary files folder in this package.
    * As many surveys require permission and approval, users may need to request access through Datalibweb. Alternatively, access
        to the constructed file GMD_all_2017.dta can be requested directly from the authors.

CLASS.dta
    Description: Country classification data including region, income group
    Variables: code, region, year_data, incgroup_historical
    Source: Data downloaded January 15, 2025. https://github.com/GPID-WB/Class/tree/master/OutputData

harmonized_npl.dta
    Description: Harmonized national poverty lines
    Source: Samuel
    Variables: country_code, welfare_time, harm_npl, welfare_type
    Source: Jolliffe ⓡ, D., Mahler ⓡ, D.G., Lakner ⓡ, C., Atamanov ⓡ, A., Tetteh-Baah, S.K., 2024.
                Poverty and Prices: Assessing the Impact of the 2017 PPPs on the International Poverty
                Line and Global Poverty. The World Bank Economic Review lhae035. https://doi.org/10.1093/wber/lhae035.
            Filename: data/01_raw/harm_npl_all_2017ppp.dta

https://datacatalog.worldbank.org/int/data/dataset/0066850/harmonized_national_poverty_lines_2017_ppp_dollars
            Data downloaded January 15, 2025

P_Data_Extract_From_WDI_Database_Archives.xlsx
    Description: World Development Indicators database archives (March 2024 vintage)
    Usage: GDP per capita data using 2017 PPP
    Source: WDI archives, downloaded 05 Feb 2026, https://databank.worldbank.org/source/wdi-database-archives
            Selection:
                Country: All
                Series: GDP per capita, PPP (constant 2011 international $)
                Time: 2017
                Version: Mar 2024

P_Data_Extract_From_Food_Prices_for_Nutrition.xlsx

    Description: Food Prices for Nutrition 2.1 (2017 data)
    Usage: Cost of energy sufficient diet data
    Source: WDI Databank, downloaded 12 Feb 2024,
https://databank.worldbank.org/source/food-prices-for-nutrition#
            Selection:
                Country: All
                Classification: Food Prices for Nutrition 1.0
                Series: Cost of energy sufficient diet in PPP dollars
                Time: 2017

==
5. DO FILES (01-do/)
==

MASTER FILE

00_Master.do
    Purpose: Master do file that:
             1) Sets global paths for data and output folders
             2) Installs required SSC packages (ineqdeco, apoverty, schemepack,
                robstat, moremata)
             3) Executes all 33 analysis do files in the correct sequence
    Execution: Running this single file will reproduce all results in the paper
    Outputs: Generates all intermediate data, tables, and figures
    Run time: Approximately 24 hours
    Note: Run the do files xx-IntroStats-benin15awi.do and xx-IntroStats-lso02awi.do
            for additional statistics reported in the Introduction of the paper.

    Required SSC Packages (auto-installed):
    - ineqdeco: Inequality decomposition measures
    - apoverty: Poverty indices calculation
    - schemepack: Graph color schemes
    - robstat: Robust statistical measures
    - moremata: Extended Mata functions (dependency for robstat)

    Custom Ado Files (included in 01-do/ folder):
    - lognfit3.ado: Three-parameter lognormal maximum likelihood estimation
    - lognfit3_ll.ado: Log-likelihood function for lognfit3
    - dagumIII.ado: Dagum Type III distribution fitting

    Note: Custom ado files are loaded automatically within their respective
          do files (01d-TPLNFloor_LMLE.do and 01e-TPLN_LMLE_DropX.do use
          lognfit3; 01g-DagumFloor.do and 01h-Dagum_DropX.do use dagumIII).
          No manual installation required.

DATA PREPARATION AND FLOOR ESTIMATION FILES (01a-01i)

These files prepare the main dataset and estimate various consumption/income
floors using different methodologies:

```
01a-MainData.do
    Purpose: Prepares MainData.dta by extracting only necessary variables
             from GMD microdata (excludes imputed, binned, and grouped data)
    Input:   02-rawdata/GMD_all_2017.dta
    Output:  03-intermediatedata/MainData.dta
    Variables kept: code, ref_year, survname, datatype, surveyid, year,
                    welfare_dec, weight
    Run time: <5 minutes

01b-RavallionFloor.do
    Purpose: Calculates Ravallion-type floors using various alpha values
             (1-4) and different poverty lines (IPL, NPL, SPL, quantiles)
             Also computes FGT poverty measures
    Input:   02-rawdata/harmonized_npl.dta, 03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/RavallionFloor.dta
    Run time: <5 minutes

01c-BetaFloor.do
    Purpose: Estimates beta distribution floors using different poverty lines
    Input:   03-intermediatedata/RavallionFloor.dta
    Output:  03-intermediatedata/BetaFloor.dta
    Run time: 1 minutes

01d-TPLNFloor_LMLE.do
    Purpose: Estimates 3-parameter lognormal (TPLN) floors using maximum
             likelihood estimation (LMLE)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/TPLNFloor_LMLE.dta
    Notes:   Uses custom ado file lognfit3.ado
    Run time: ~3.5 hrs

01e-TPLN_LMLE_DropX.do
    Purpose: Estimates TPLN floors using LMLE after dropping observations
             below $0.50 and $1.00 per day thresholds
    Input:   03-intermediatedata/TPLNFloor_LMLE.dta,
03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/TPLNFloor_LMLE_DropX.dta
    Notes:   Uses custom ado file lognfit3.ado; analyzes sensitivity to
             bottom observations
    Run time: ~30 mins

01f-TPLNFloor_ME.do
    Purpose: Estimates 3-parameter lognormal (TPLN) floors using method of
             moments (Cohen & Whitten 1980, eq. 5.3)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/TPLNFloor_MME.dta
    Run time: <5 mins

01g-DagumFloor.do
```

```
    Purpose: Estimates Dagum Type III distribution floors
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/DagumFloor.dta
    Notes:   Uses custom ado file dagumIII.ado
    Run time: ~13 hrs

01h-Dagum_DropX.do
    Purpose: Estimates Dagum III floors after dropping observations below
             $0.50 and $1.00 per day thresholds
    Input:   03-intermediatedata/DagumFloor.dta, 03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/DagumFloor_DropX.dta
    Notes:   Uses custom ado file dagumIII.ado; analyzes sensitivity to
             bottom observations
    Run time: ~3.5 hours

01i-LISfloors.do
    Purpose: Calculates LIS-methodology floors using Q1 - x*IQR of log
             income (x = 2, 3, 4)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/LISFloor.dta
    Run time: ~1 hour


ANALYSIS FILES (02a-02d)
------------------------
02a-InfluenceFunction.do
    Purpose: Generates data for influence function analysis of poverty gap
             by decile (used in Table 9)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/Influence_PG_byDecile.dta
    Run time: ~30 mins

02b-PovertyComparisons.do
    Purpose: Compares poverty indicators with and without $0.25/day floor
             for top 10 surveys with highest share below $0.25/day
             (used in Table 10)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/ComparingIndicators_withFloor_poverty.dta
    Run time: <5 mins

02c-InequalityComparisons.do
    Purpose: Compares inequality indicators with and without $0.25/day floor
             for top 10 surveys with highest share below $0.25/day
             (used in Table 11)
    Input:   03-intermediatedata/MainData.dta
    Output:  03-intermediatedata/ComparingIndicators_withFloor_inequality.dta
    Run time: <5 mins

02d-SurveyStats.do
    Purpose: Calculates robust survey statistics (mean, median, MAD, Qn, Sn,
```

```
              SD, IQR) for log welfare by survey (used in Table D4)
    Input:    03-intermediatedata/MainData.dta
    Output:   03-intermediatedata/SurveyStats.dta
    Notes:    Uses robstat package for robust scale estimators
    Run time: ~1.5 hours


TABLE GENERATION FILES (03a-03k)
--------------------------------
Main tables for the paper:

03a-tab1.do
    Purpose: Table 1 - Distribution of surveys by region and data type
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/MainData.dta
    Output:  04-output/Tables.xlsx (sheet: tab1)

03b-tab2.do
    Purpose: Table 2 - Analysis of negative welfare values and zeros by
             data type
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Tables.xlsx (sheet: tab2)

03c-tab3.do
    Purpose: Table 3 - Ravallion floors
    Input:   03-intermediatedata/RavallionFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tab3)

03d-tab4.do
    Purpose: Table 4 - Beta floors
    Input:   03-intermediatedata/BetaFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tab4)

03e-tab5.do
    Purpose: Table 5 - Three Parameter Log Normal floors (LMLE, LMLE DropX,
             and ME) with sensitivity analysis
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/TPLNFloor_LMLE.dta,
             03-intermediatedata/TPLNFloor_LMLE_DropX.dta,
             03-intermediatedata/TPLNFloor_MME.dta
    Output:  04-output/Tables.xlsx (sheet: tab5)

03f-tab6.do
    Purpose: Table 6 - Dagum III floors with sensitivity analysis
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/DagumFloor.dta,
             03-intermediatedata/DagumFloor_DropX.dta
    Output:  04-output/Tables.xlsx (sheet: tab6)

03g-tab7.do
    Purpose: Table 7 - LIS-type floor distribution statistics
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/LISFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tab7)
```

```
03h-tab8.do
    Purpose: Table 8 - [Floor estimates by survey characteristics]
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/RavallionFloor.dta,
             03-intermediatedata/BetaFloor.dta,
             03-intermediatedata/TPLNFloor_LMLE.dta,
             03-intermediatedata/DagumFloor.dta,
             03-intermediatedata/LISFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tab8)

03i-tab9.do
    Purpose: Table 9 - Influence function analysis
    Input:   03-intermediatedata/Influence_PG_byDecile.dta
    Output:  04-output/Tables.xlsx (sheet: tab9)

03j-tab10.do
    Purpose: Table 10 - [Additional analysis]
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Tables.xlsx (sheet: tab10)

03k-tab11.do
    Purpose: Table 11 - [Summary statistics or comparisons]
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Tables.xlsx (sheet: tab11)


APPENDIX TABLE GENERATION FILES (03l-03p)
-----------------------------------------
Appendix tables for supplementary analysis:

03l-tab_d1.do
    Purpose: Appendix Table D1 - Ravallion floor distribution statistics
    Input:   03-intermediatedata/RavallionFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tabd1)

03m-tab_d2.do
    Purpose: Appendix Table D2 - Beta floor distribution statistics
    Input:   03-intermediatedata/BetaFloor.dta
    Output:  04-output/Tables.xlsx (sheet: tabd2)

03n-tab_d3.do
    Purpose: Appendix Table D3 - Distribution statistics for TPLN (LMLE,
             LMLE DropX, ME), Dagum III (full and DropX), and LIS floors
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/MainData.dta,
             03-intermediatedata/TPLNFloor_LMLE.dta,
             03-intermediatedata/TPLNFloor_LMLE_DropX.dta,
             03-intermediatedata/TPLNFloor_MME.dta,
             03-intermediatedata/DagumFloor.dta,
             03-intermediatedata/DagumFloor_DropX.dta,
             03-intermediatedata/LISFloor.dta
```

```
    Output:  04-output/Tables.xlsx (sheet: tabd3)

03o-tab_d4.do
    Purpose: Appendix Table D4 - Distribution of Belotti et al. (2024) floors
             (LIS floors and z-score based robust statistics)
    Input:   03-intermediatedata/MainData.dta, 02-rawdata/CLASS.dta,
             03-intermediatedata/SurveyStats.dta
    Output:  04-output/Tables.xlsx (sheet: tabd4)

03p-tab_d5.do
    Purpose: Appendix Table D5 - Dagum floor distribution statistics
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Tables.xlsx (sheet: tabd5)


FIGURE GENERATION FILES (04a-04e)
---------------------------------
04a-Fig1.do
    Purpose: Figure 1 - Frequency of surveys by welfare type and year
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Fig1.png

04b-Fig2.do
    Purpose: Figure 2 - Ravallion floor chart with IPL and consumption
             floors by income group and decade
    Input:   02-rawdata/CLASS.dta, 03-intermediatedata/RavallionFloor.dta
    Output:  04-output/Fig2.png

04c-Fig3.do
    Purpose: Figure 3 - Cost of energy sufficiency diet (2,300 kcal)
    Input:   02-rawdata/P_Data_Extract_From_WDI_Database_Archives.xlsx,
             02-rawdata/P_Data_Extract_From_Food_Prices_for_Nutrition.xlsx,
             02-rawdata/CLASS.dta
    Output:  04-output/Fig3.png

04d-Fig4.do
    Purpose: Figure 4 - Share of population living on less than x% relative
             to median welfare
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Fig4.png

04e-Fig5.do
    Purpose: Figure 5 - Distribution of population below $0.25
    Input:   03-intermediatedata/MainData.dta
    Output:  04-output/Fig5.png


ADDITIONAL STATISTICS REPORTED IN INTRODUCTION
----------------------------------------------
xx-IntroStats-benin15awi.do
```

    Purpose: Estimates Atkinson welfare index reported in the Introduction.

xx-IntroStats-lso02awi.do
    Purpose: Estiamtes Atkinson welfare index for LSO2002 and 17 surveys
                reported in Introduction.

CUSTOM ADO FILES
----
lognfit3.ado & lognfit3_ll.ado
    Purpose: Maximum likelihood estimation for 3-parameter lognormal
             distribution

dagumIII.ado
    Purpose: Dagum Type III distribution fitting routines

==
6. INTERMEDIATE DATA FILES (03-intermediatedata/)
==

These files are created by the do files in 01-do/:

MainData.dta
    Created by: 01a-MainData.do
    Description: Preprocessed microdata containing only necessary variables
                 from GMD_all_2017.dta (excludes imputed, binned, and
                 grouped data); used as input for most analysis files
    Variables: code, ref_year, survname, datatype, surveyid, year,
               welfare_dec, weight
    Access: Derivative of restricted data. (RESTRICTED DATASET)
             See access instruction for the "Global Monitoring Database 2017 – all
surveys (microdata)" in raw data files.

RavallionFloor.dta
    Created by: 01b-RavallionFloor.do
    Description: Ravallion-type floors with alpha = 1-4, multiple poverty
                 lines, FGT measures

BetaFloor.dta
    Created by: 01c-BetaFloor.do
    Description: Beta distribution floor estimates

TPLNFloor_LMLE.dta
    Created by: 01d-TPLNFloor_LMLE.do
    Description: 3-parameter lognormal floors using maximum likelihood

TPLNFloor_MME.dta
    Created by: 01f-TPLNFloor_ME.do
    Description: 3-parameter lognormal floors using method of moments

TPLNFloor_LMLE_DropX.dta
    Created by: 01e-TPLN_LMLE_DropX.do
    Description: 3-parameter lognormal floors using LMLE after dropping
                 observations below $0.50 and $1.00 thresholds; includes
                 gamma, xbar, sigma estimates for each threshold

DagumFloor.dta
    Created by: 01g-DagumFloor.do
    Description: Dagum Type III distribution floor estimates

DagumFloor_DropX.dta
    Created by: 01h-Dagum_DropX.do
    Description: Dagum Type III distribution floors after dropping
                 observations below $0.50 and $1.00 thresholds; includes
                 a, b, p, x0 parameter estimates and share below threshold

LISFloor.dta
    Created by: 01i-LISfloors.do
    Description: LIS-methodology floors (Q1 - x*IQR)

Influence_PG_byDecile.dta
    Created by: 02a-InfluenceFunction.do
    Description: Influence function data for poverty gap by decile

ComparingIndicators_withFloor_poverty.dta
    Created by: 02b-PovertyComparisons.do
    Description: Comparison of poverty indicators (headcount, poverty gap,
                 FGT(2), Watts) with and without $0.25/day floor for top 10
                 surveys with highest share below $0.25/day

ComparingIndicators_withFloor_inequality.dta
    Created by: 02c-InequalityComparisons.do
    Description: Comparison of inequality indicators (Gini, MLD, Atkinson(2),
                 Pigou-Dalton) with and without $0.25/day floor for top 10
                 surveys with highest share below $0.25/day

SurveyStats.dta
    Created by: 02d-SurveyStats.do
    Description: Survey-level robust statistics (mean, median, standard
                 deviation, IQR, MAD, Sn, Qn) of log welfare for use in
                 calculating Belotti et al. (2024) z-score based floors

==
7. OUTPUT FILES (04-output/)
==

Tables.xlsx
    Description: Main output file containing all tables
    Sheets: tab1, tab2, tab3, tab4, tab5, tab6, tab7, tab8, tab9, tab10,

            tab11, tabd1, tabd2, tabd3, tabd4, tabd5
    Created by: Table generation do files (03a-03p)

Fig1.png through Fig5.png
    Description: Figures for paper
    Created by: Figure generation do files (04a-04e)

==
8. WORKFLOW SEQUENCE
==

AUTOMATED EXECUTION (Recommended):

To reproduce all results automatically, simply run:
    do "01-do\00_Master.do"

This will execute all 34 analysis files in the correct order and generate
all tables and figures. Total runtime: approximately 24 hours.

MANUAL EXECUTION (Alternative):

If you prefer to run files individually, follow this sequence:

1. 00_Master.do   (Set paths and install packages)
2. 01a-MainData.do   (Prepare MainData.dta from GMD microdata)
3. 01b-RavallionFloor.do   (Independent, uses MainData.dta)
4. 01c-BetaFloor.do   (Uses output from 01b)
5. 01d-TPLNFloor_LMLE.do   (Independent, uses MainData.dta)
6. 01e-TPLN_LMLE_DropX.do   (Uses output from 01d)
7. 01f-TPLNFloor_ME.do   (Independent, uses MainData.dta)
8. 01g-DagumFloor.do   (Independent, uses MainData.dta)
9. 01h-Dagum_DropX.do   (Uses output from 01g)
10. 01i-LISfloors.do   (Independent, uses MainData.dta)
11. 02a-InfluenceFunction.do (Independent, uses MainData.dta, needed for Table 9)
12. 02b-PovertyComparisons.do (Independent, uses MainData.dta, needed for Table 10)
13. 02c-InequalityComparisons.do (Independent, uses MainData.dta, needed for Table
11)
14. 02d-SurveyStats.do   (Independent, uses MainData.dta, needed for Table D4)
15. 03a-03p (Tables)   (Use raw/intermediate data)
16. 04a-04e (Figures)   (Use raw/intermediate data)

Note: After completing steps 1-2, files 3-14 can be run in parallel except
      that 01c requires 01b, 01e requires 01d, and 01h requires 01g.

==
9. NOTES
==

- All data files use 2017 PPP conversion
- Microdata excludes imputed, binned, and grouped data
- Floor estimates vary by methodology and are compared across tables
- DropX analyses (01e, 01h) test sensitivity by excluding observations below $0.50 and $1.00 thresholds to assess impact of extreme bottom values
- Income group classifications are historical (FY25 for regional groups)
- All monetary values are in daily per capita terms unless noted otherwise

==========