Replication Package for "Default Risk and Pricing in the U.S. Credit Card Market"

Authors: Kyle Dempsey and Felicia Ionescu

This package contains the code and data necessary to replicate all results in the paper. It also contains information on how to obtain the data and software from the various sources. This document is divided into four sections:

1. data availability statements
2. computational requirements (software, memory, and runtime)
3. layout of the replication package
4. details of each replication step

1 Data Availability Statements

The supervisory confidential Federal Reserve Y-14M data, collected monthly by the Federal Reserve Board as part of the Comprehensive Capital Analysis and Review, used in the regressions and the empirical results in the paper cannot be shared publicly to protect the confidentiality of the banks that submitted their information to the Federal Reserve for the purposes of stress tests. Even within the Federal Reserve, only those with an approved data need can see this dataset. For this reason, we do not include any of the raw (or even more aggregated) data in this replication package. Therefore, many of the codes will not actually reproduce the results in the paper since the data required is not included. The codes we include, however, replicate all the results with the proper data inputs.

2 Computational Requirements

2.1 Software requirements

In order to fully replicate the results in the paper, the following pieces of software are needed:

- **Microsoft Excel** (I used Version 2411)
- Generating the figures and certain empirical moments requires **Matlab 2021a** or newer.
- All the main programs for solving the model are written in **Fortran 90** and were solved on Ohio State's Unity computer cluster**.** The programs are written using OpenMP parallelization, which allows for parallelization across several cores on the same node.[1]
- Much of the empirical analysis is done using stata **Stata 18.5 MP** (for regressions) and **R 4.4.1** (for creating variables and computing summary statistics).

2.2 Memory and runtime requirements

- For generating tables / data moments and figures, **Matlab** and **Excel** are extremely fast (no more than seconds) on standard laptop and desktop computers.
- The main programs in **Fortran**, on the other hand, can only be run feasibly using parallel processing on a compute node with many cores. Starting from arbitrary initial conditions for a given set of parameters, it can take many hours to solve for an equilibrium of the full model. For this reason, to ease the replication, I supply a set of results which can be used as initial conditions. Starting from these initial conditions, each version of the model solves in approximately 2-3 minutes on a single Intel Xeon node with 40 cores and 192 GB of RAM.[2]

[1] https://www.openmp.org/

[2] For details of the computing specifications, see https://osu.teamdynamix.com/TDClient/1929/ASC/KB/ArticleDet?ID=65574#Family_ancho and look at the "Exclusive (Econ)" nodes.

- The empirical analysis in **Stata** and **R** runs in a matter of minutes. The costly step for the empirics is querying the full database to generate the subsamples we use for our analyses.

3 Layout of the Replication Package

The replication package consists of the following subfolders:

- `Fortran\`: contains all of the Fortran codes (.f90 files) for solving the model.
- `FortranOutput\`: contains output of all the Fortran codes in the folder above that are needed to recreate the figures in the paper. This contains the following subfolders, one for each of the model variants analyzed in Section 5 of the paper: `Baseline\`, `BKOnly\`, `LowIntCost\`, and `ShortTerm\`.
- `Matlab\`: contains all Matlab codes required to generate the figures and tables in the paper given either the raw data or the output of the Fortran programs.
- `Stata\`: contains all Stata codes required for the emprical analysis given the raw data.
- `R\`: contains all R codes for aggregating data after analysis in Stata.
- `Figures\`: contains all figures in the paper (in .eps format)

Note that by using the model output in the `FortranOutput\` folder, one can reproduce all of the Figures in the paper without running any of the costly Fortran programs.

4 Details of the Replication

This section proceeds in the following steps. In the first subsection, I describe how to replicate all of the displays in the paper (tables and figures) given the output of the main Fortran programs. In the second subsection, I describe how to generate the output of the main programs.

4.1 Replication of tables and figures

4.1.1 Empirical analysis

There are four sets of codes used for the empirical results which use proprietary Y-14M data (schedule D):

1. `Replication_DempseyIonescu_Summarystats.R` contains the codes used to compute the summary statistics in Tables 1 and 2. The code describes all variables used in the analysis, specifying which ones are used directly from the data and which ones are computed. For the latter category the codes for their calculation are specified.
2. `Replication_DempseyIonescu_ProbitRegression.do` contains the regression codes used to construct Probit regression results Tables 3 and A.1. This code needs to run first as the output is used in the OLS regressions.
3. `Replication_DempseyIonescu_OLSRegression.do` contains the regression codes used to construct OLS regression results Tables 4, 5, A.2, and A.3. This code needs to run after the probit regression codes as the OLS regressions use the output from the probit regressions.
4. `Replication_DempseyIonescu_Figures.R` contains the codes used to construct all the empirical variables in all figures. The codes to draw the actual figures using aggregated data are also provided separately in Matlab, as some of the figgures contain both empirical data and model simulated data.

All the variables used in the paper are computed using R and all the regressions in the paper are run using STATA for random samples drawn from the Y-14M dataset. Please refer to Section 2 and Appendix A for our data construction and sample selection methodology. Our main analyses use a 0.3% random sample for credit card accounts originated after 1995, which were open during 2018 and either open or closed during

2019. Overall, 14 banks are used in this analysis. Table 1 and Table 2 contain summary statistics with definitions and data sources for all variables used in the analysis. All codes are anonymized to ensure that we do not expose information on any individual banks, loans, or customers, or Federal Reserve file paths.

4.1.2 Model and aggregated empirical output

This section describes how to replicate Figures 1 – 5 in the main text, Figures A.1 – A.4 and C.1 – C.2 in the Appendix, and Tables 6 – 8 in the main text. Tables 1 – 5 in the main text and Tables A.1 – A.3 in the Appendix are based solely on proprietary data and are described elsewhere. In order to facilitate working directly with the output of the model (solved in Fortran), I include a full set of results from all Fortran programs in the replication package.

Running the Matlab program `DI_JPEMacro_Replicate.m` generates the following figures and tables. To run this program (given the Fortran output structured as in the replication package or as described in Section 4.2 below), simply go to line 21 and change the variable `RootPath` to the correct folder for your working environment.

- Figure 1 "Predicted vs realized default rates and the distribution of borrowers"
- Figure 2 "Interest rate spreads by default probability"
- Figure 3 "Experiments to highlight the role of key model features"
- Figure 4 "Pricing schedules by income and across model variants"
- Figure 5 "Interest rate spreads by default probability: model vs. data"
- Figure A.1 "Interest rate spreads by default probability and income quartile"
- Figure A.2 "Interest rate spreads by default probability and revolver status"
- Figure A.3 "Predicted vs realized default rates and the distribution of borrowers for broader default measure"
- Figure A.4 "Interest rate spreads by default probability for broader default measure"
- Figure C.1 "Decisions without extreme value shocks"
- Figure C.2 "Ex post vs ex ante default probability across model variants"
- Table 7 "Determinants of spreads: model vs. data"

Tables 6 and 8 are produced directly from the Fortran programs, as described below.

4.2 Running the main Fortran programs

The majority of the results in the paper are produced by running `SS_JPEM_Main.f90` using the supplied initial conditions in the folder `FortranOutput\`. This program uses the following source codes:

- `SS_JPEM_Main.f90`: main program. Solves steady state equilibrium for each model variant presented in the paper.
- `SS_JPEM_Module.f90`: main module files containing all subroutines and variables specific to the analysis of this paper.
- `misc_utilities.f90`: miscellaneous standard objects such as sorting, regressions, reading / writing, etc.
- `submit_SS_JPEM.sh`: shell script for compiling the various source codes into an executable and executing the file.

To run this program, just compile the source codes using the shell script, modified for your computing environment but with the results in `FortranOutput\` file stored in a directory labeled `input\` and the remaining paths laid out as in the replication package. This program produces output in the folder corresponding to the indicated model version. To specify the version of the model, go to line 205 in `SS_JPEM_Module.f90` and set the variable `ModelVersion` to be equal to the desired version (0 = baseline, 1 = short term debt, 2 = low intermediation cost, 4 = bankruptcy only). Then collect the output from each run into the `FortranOutput\` folder.