# Replication data for: Systematic Bias and Nontransparency in US Social Security Administration Forecasts

Konstantin Kashin, Gary King, and Samir Soneji

March 2015

This document provides information needed to replicate the results in Kashin, King and Soneji (2015*a*) and Kashin, King and Soneji (2015*b*). All the figures in the body of the paper are generated by the code in `analysis.Rnw`. This is a Sweave file that can be compiled using the `knitr` package in R. `knitr` runs the `R` code, outputs the figures into a `figures` subdirectory, and creates the `analysis.tex` file. This file can then be compiled using `pdflatex`. All the figures in the online Appendix can be similarly compiled using the `analysis_appendix.Rnw`. The workflow is in the `Makefile` and can be run by typing `make` in the command line. To compile just the main analysis or the Appendix, type `make paper` or `make appendix`, respectively. In addition to having LaTeX and the base version of `R` installed, you need to have the following `R` packages: `knitr`, `ggplot2`, `reshape2`, `gridExtra`, and `simpleboot`.

All data needed for replication is available in the `data` subdirectory. The `helpers` subdirectory contains auxiliary scripts that create some of the data files used as inputs into the analysis (described in more detail below).

## Data

### Observed Data

The subdirectory `data/observed` contains 2 `RData` files: `obs.ex.RData` and `obs.tfr.RData`. `obs.ex` is a 4-dimensional array that contains the observed life expectancy from 2 different sources: HMD and SSA. The dimensions are year (1933-2010), sex (Male and Female), age (0 and 65), and source (HMD and SSA). `obs.tfr` is a matrix containing total fertility rate (TFR) values. The rows are years (1933-2010) and the columns are source (HFD and SSA). The subdirectory also contains the raw period life tables and population from HMD (`hmd_fltper_1x1.txt`, `hmd_mltper_1x1.txt`, `hmd_population.txt`).

**Sources:** We obtain observed period life expectancy data for 1982–2010 from the Human Mortality Database (HMD). We obtain observed total fertility rate (TFR) data for 1982–2010 from the Human Fertility Database (HFD). Observed SSA values are gathered from the 2014 Trustees Report.

*Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on June 1, 2014).

*Human Fertility Database*. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org (data downloaded on June 1, 2014).

## ssa_demog.RData

This file contains two data frames: `ex.ssa` and `tfr.ssa`.

The variables in `ex.ssa` are:

**TR:** year of Trustees Report

**forecast.year:** year forecast

**sex:** sex (Male or Female)

**age:** age (0 or 65)

**forecast:** SSA's life expectancy forecast for intermediate cost scenario ("best guess" scenario)

**high:** SSA's life expectancy forecast for high cost scenario

**low:** SSA's life expectancy forecast for low cost scenario

**hmd_observed:** Observed life expectancy (HMD)

**ssa_observed:** Observed life expectancy (SSA)

**hmd_residual:** residual based on HMD observed values (`forecast-hmd_observed`)

**ssa_residual:** residual based on SSA observed values (`forecast-ssa_observed`)

**lower:** lower bound of SSA "confidence interval"

**upper** : upper bound of SSA "confidence interval"

**ci_width:** width of SSA's "confidence interval" (`upper-lower`)

**hmd_coverage:** coverage of SSA's "confidence interval" based on HMD observed value (`TRUE` if HMD observed value falls inside the SSA "confidence interval")

`ssa_coverage`: coverage of SSA's "confidence interval" based on SSA observed value (`TRUE` if SSA observed value falls inside the SSA "confidence interval")

The variables in `tfr.ssa` are:

`TR`: year of Trustees Report

`forecast.year`: year forecast

`forecast`: SSA's total fertility rate forecast for intermediate cost scenario ("best guess" scenario)

`high`: SSA's total fertility rate forecast for high cost scenario

`low`: SSA's total fertility rate forecast for low cost scenario

`hfd_observed`: Observed total fertility rate (HFD)

`ssa_observed`: Observed total fertility rate (SSA)

`hfd_residual`: residual based on HFD observed values (`forecast-hfd_observed`)

`ssa_residual`: residual based on SSA observed values (`forecast-ssa_observed`)

`lower`: lower bound of SSA "confidence interval"

`upper` : upper bound of SSA "confidence interval"

`ci_width`: width of SSA's "confidence interval" (`upper-lower`)

`hfd_coverage`: coverage of SSA's "confidence interval" based on HFD observed value (`TRUE` if HFD observed value falls inside the SSA "confidence interval")

`ssa_coverage`: coverage of SSA's "confidence interval" based on SSA observed value (`TRUE` if SSA observed value falls inside the SSA "confidence interval")

**Sources:** We collected all life expectancy forecasts published in the annual Trustees Reports 1982–2010. In reports prior to 2001, SSA published life expectancy at birth and at age 65 forecasts for males and females projected in 5-year intervals for a total of 75 years into the future. Post-2001, supplementary single-year tables are included online. Our sources are Table 11 of Trustees Reports 1982-1991, Table II.D.2 of Trustees Reports 1992-2000, and Table V.A3 of the supplemental single-year tables of Trustees Reports 2001-2010. We collected all TFR forecasts published in the annual Trustees Reports 1982–2010. In reports prior to 2001, SSA published TFR forecasts in 5-year intervals for a total of 75 years into the future in the same tables as life expectancy (see preceding paragraph). For 2001 and onward, supplementary single-year Table V.A1 of each Trustees Report contains TFR forecasts.

## demog_loess_boot.RData

This file contains two data frames, created using the script in the `helpers` subdirectory: `demog_loess_boot` and `demog_loess_ssa_boot`. These data frames contain the confidence intervals around the LOESS lines fit in Figures 4, 13, 15, and 16 of `analysis.pdf`, calculated using bootstrapping. Each dataframe has the following variables:

**TR:** year of Trustees Report

**fit:** fitted value from LOESS

**lb:** lower bound of LOESS confidence interval

**ub:** upper bound of LOESS confidence interval

**type:** type of forecast

## ssa_expend.csv

Observed SSA expenditures from 2000-2013. The following variables are available (where [PROGRAM] can be `oasdi` or `oasi`):

**[PROGRAM]_total_expend:** total expenditures in millions of current $ (http://j.mp/OASDIexpend; http://j.mp/OASIexpend)

**[PROGRAM]_oasdi_benif_payments:** benefit payments in millions of current $ (http://j.mp/OASDIexpend; http://j.mp/OASIexpend)

**[PROGRAM]_oasdi_admin_expend:** administrative expenses in millions of current $ (http://j.mp/OASDIexpend; http://j.mp/OASIexpend)

**[PROGRAM]_transfer_to_rr:** transfers to Railroad Retirement program in millions of current $ (http://j.mp/OASDIexpend;http://j.mp/OASIexpend)

**total_benif:** total number of beneficiaries receiving benefits on December 31 (http://j.mp/OASIbenif)

**oasi_total_benif:** total number of OASI beneficiaries receiving benefits on December 31 (http://j.mp/OASIbenif)

**oasi_ret_workers:** total number of retired workers and dependents receiving benefits on December 31 (http://j.mp/OASIbenif)

**oasi_surv:** total number of survivors receiving benefits on December 31 (http://j.mp/OASIbenif)

**ssdi_dis_workers:** total number of disabled workers and dependents receiving benefits on December 31 (http://j.mp/OASIbenif)

**CPI-U:** CPI-U (http://bls.gov/data)

### ssa_expend_forecasts.csv

Forecasts of SSA OASDI expenditures from 2000-2013. The following variables are available:

**TR:** year of Trustees Report

**year forecast:** year forecast

**oasdi total cost:** SSA's total cost forecast (in billions of $)

**Sources:** Forecasts of total OASDI expenditures in billions of current dollars for 2000-2010 Trustees Reports are found in Table III.B3 of 2000 TR, Table VI.E9 of 2001-2002 TR, Table VI.F9 of 2003-2004 TR, and Table VI.F8 of 2005-2010 TR.

### unexpected_benif.RData

This file contains a matrix called `ub` with the estimated number of unanticipated OASI beneficiaries for a given Trustees Report and projection year from 2000-2010. The rows denote the Trustees Report and the columns denote the year forecast. This matrix is calculated using the script `unexpect_benif.R` in the `helpers` subdirectory.

In order to estimate the number of unanticipated beneficiaries, we determine the proportionate change in age-specific mortality rates that would correspond to the forecast error in life expectancy for each Trustees Report and projection year. For each age 65 years and older, we multiply the age-specific population count by the difference between the counterfactual age-specific mortality rate necessary to achieve the projected life expectancy and the observed age-specific mortality rate. This product represents the number of unanticipated beneficiaries for a given age and sex. We then sum across all ages for males and females to estimate the total number of unanticipated beneficiaries. For example, the 2005 Trustees Report projected year 2010 life expectancy at age 65 for males to be 16.6 years while observed life expectancy in 2010 equaled 17.9 years. This forecast error of 1.3 years corresponds to a 18.75% increase in observed age-specific mortality rates, or 150,844 unanticipated male beneficiaries.

### ssa_finance.RData

This file contains two data frames: `tf.balance.pred` and `tf.ratio.pred`.

The variables in `tf.balance.pred` (where `[metric]` is a placeholder for `income`, `cost`, or `balance`) are:

**TR:** year of Trustees Report

**forecast.year:** year forecast

**income.rate:** forecast income rate for intermediate cost scenario ("best guess" scenario)

**cost.rate :** forecast cost rate for intermediate cost scenario ("best guess" scenario)

`balance:` forecast balance for intermediate cost scenario ("best guess" scenario)

`[metric].lower :` lower bound of forecast scenarios

`[metric].upper :` upper bound of forecast scenarios

`[metric].I:` forecast for low-cost scenario / alternative I

`[metric].III:` forecast for high-cost scenario / alternative III

`[metric].IIA:` forecast for second intermediate scenario / alternative II-A

`[metric].truth:` observed value

The variables in `tf.ratio.pred` are:

`TR:` year of Trustees Report

`forecast.year:` year forecast

`tfratio:` forecast trust fund ratio for intermediate cost scenario ("best guess" scenario)

`tfratio.lower :` lower bound of trust fund ratio forecast scenarios

`tfratio.upper :` upper bound of trust fund ratio forecast scenarios

`tfratio.I:` forecast of trust fund ratio for low-cost scenario / alternative I

`tfratio.III:` forecast of trust fund ratio for high-cost scenario / alternative III

`tfratio.IIA:` forecast of trust fund ratio for second intermediate scenario / alternative II-A

`tfratio.truth:` observed value of trust fund ratio

**Notes and sources:** For 1978–2012, we take the observed cost rate and balance from Table IV.B1 and the observed trust fund ratio from Table IV.B3 of the 2013 Trustees Report (http://j.mp/2013tables). We calculate residuals as the difference between SSA's "best guess" projection (intermediate-cost scenario / alternative II) and the historic statistics from SSA. Note that for 1981–1990, Trustees Reports have two intermediate-cost scenarios: alternative II-A and II-B. For these years, we follow subsequent Trustees Reports and use II-B as the "best guess" projection. To calculate the uncertainty interval, we use the minimum and maximum values of projected life expectancy across the three scenarios (four for 1981-1990).

Sources for SSA projections published in the annual Trustees Reports (TR) from 1978 to 2012:

**Cost rate:** Table 26 of TR 1978, Table 27 of TR 1980-1980, Table 28 of TR 1982, Table 29 of TR 1982-1983, Table 30 of TR 1984-1985, Table 28 of TR 1986, Table 26 of TR 1987-1991, Table II.F.13 of TR 1992-2000, Table IV.B1 of the supplemental single-year tables of TR 2001-2012.

**Balance:** For Trustees Reports from 1986, balance projections are available from the same tables as the cost rate projections. In 1978-1982 TR, the Trustees project the same scheduled tax rate across the cost scenarios (available in Table 25 for TR 1978, Table 26 for TR 1979-1980, Table 28 for TR 1981, and Table 29 for TR 1982). We subtract the cost rates across the scenarios from the scheduled income rate to obtain the range of balance projections. For 1983-1984 TR, the income tax rate varies slightly across the cost scenarios and is only published for the two intermediate projections (Table 27 for TR 1983 Table 28 for TR 1984). We are thus unable to evaluate coverage of SSA's uncertainty intervals for projections made in these two reports.

**Trust fund ratio:** Table 28 of TR 1978, Table 29 of TR 1979-1980, Table 31 of TR 1981, Table 32 of TR 1982-1983, Table 33 of TR 1984-1985, Table 31 of TR 1986, Table 29 of TR 1987, Table 31 of TR 1988-1990, Table 32 of TR 1991, Table II.F.19 of TR 1992-1994, Table II.F20 of TR 1995-2000, Table IV.B3 of the supplemental single-year tables of TR 2001-2012.

## `ssa_immigration.RData`

This file contains a data frame called `immigration` with forecasts of net legal immigration made by the SSA in the 2000 to 2010 Trustees Reports. The years forecast are 2005 and 2010.

The variables in `immigration` are:

`TR`: year of Trustees Report

`forecast.year`: year forecast

`forecast`: forecast net legal immigration under intermediate cost scenario ("best guess" scenario) (1000s of people)

`high`: forecast net legal immigration under high cost scenario (1000s of people)

`low`: forecast net legal immigration under low cost scenario (1000s of people)

`truth`: observed net legal immigration (1000s of people), as reported by SSA

`residual`: forecast net legal immigration under intermediate cost scenario ("best guess" scenario) minus observed net legal immigration (1000s of people)

`coverage`: coverage of SSA's "confidence interval" (`TRUE` if observed value falls inside the SSA "confidence interval")

**Sources:** Observed net legal immigration is available in Table V.A1 of the 2014 Trustees Report. Forecasts of net legal immigration are available on page 62 of the 2000 Trustees Report and in Table V.A1 of the 2001–2010 Trustees Reports.

# References

Kashin, Konstantin, Gary King and Samir Soneji. 2015*a*. "Systematic Bias and Nontransparency in U.S. Social Security Administration Forecasts." *Journal of Economic Perspectives* . In press. 1

Kashin, Konstantin, Gary King and Samir Soneji. 2015*b*. "Systematic Bias and Nontransparency in U.S. Social Security Administration Forecasts: Online Appendix." *Journal of Economic Perspectives* . In press. 1