



# Outline

- History & Design Decisions
- Multiple Imputation Approach
- SSB data sources
- Production Overview
- SSB Research
- How to use the SSB
- Future Plans



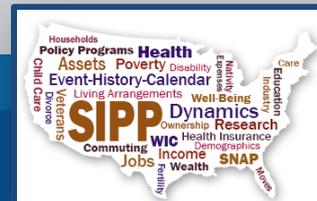
# SSB History

- 1999-2000: Census, SSA, and IRS engage in negotiations about the use of administrative data by Census
- February 2001: Federal Regulation published authorizing sharing of data items from W-2 tax forms
  - Data arrives at Census shortly thereafter
- Goal: Use these data to improve the SIPP

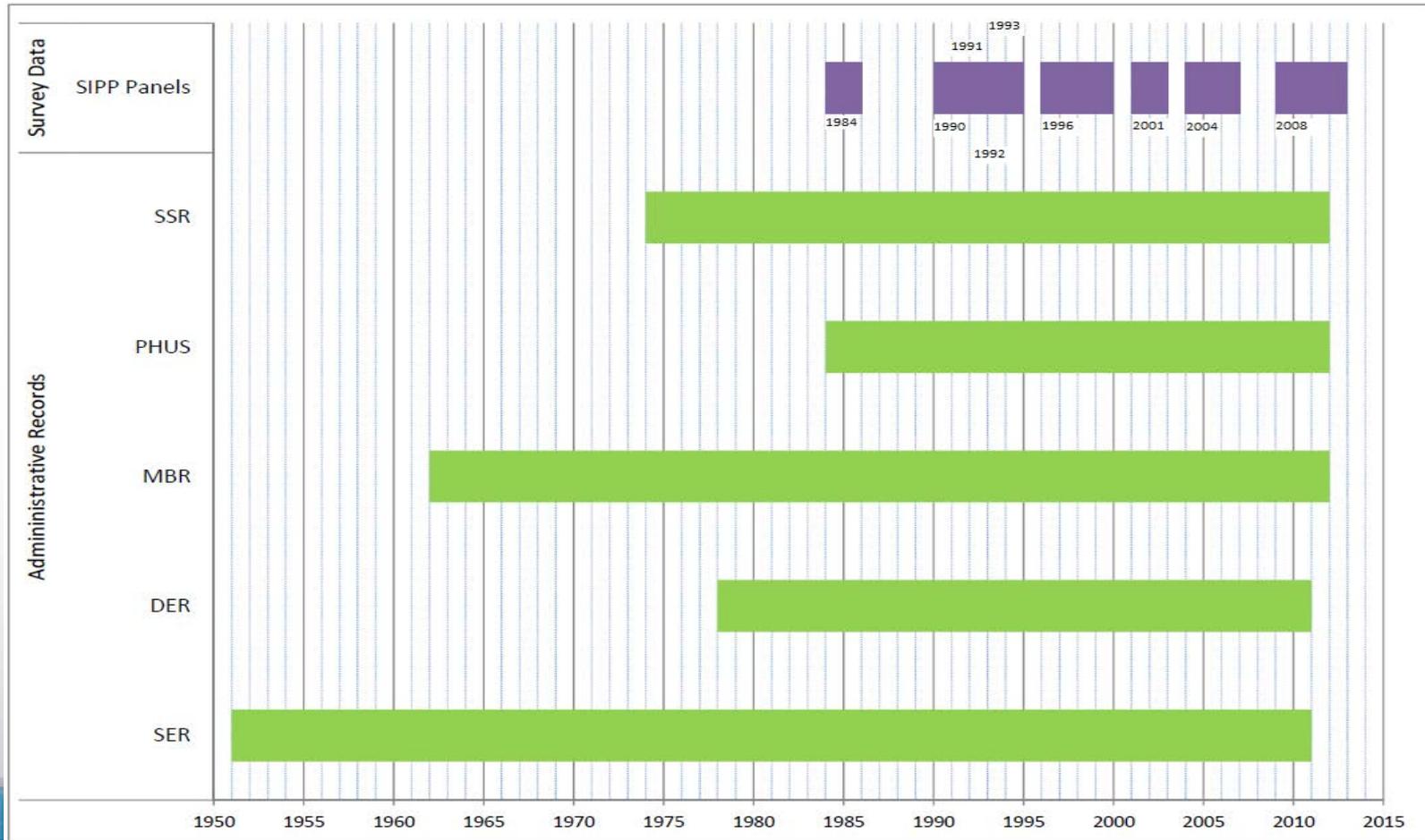


# SSB History

- Committee convenes with representatives from IRS, Census, SSA, and CBO
- Decision to pursue a new SIPP public use file that contains the administrative data linked to the survey data
- Challenge: Confidentiality
- Solution: Data Synthesis



# Design Decision: Data Sources











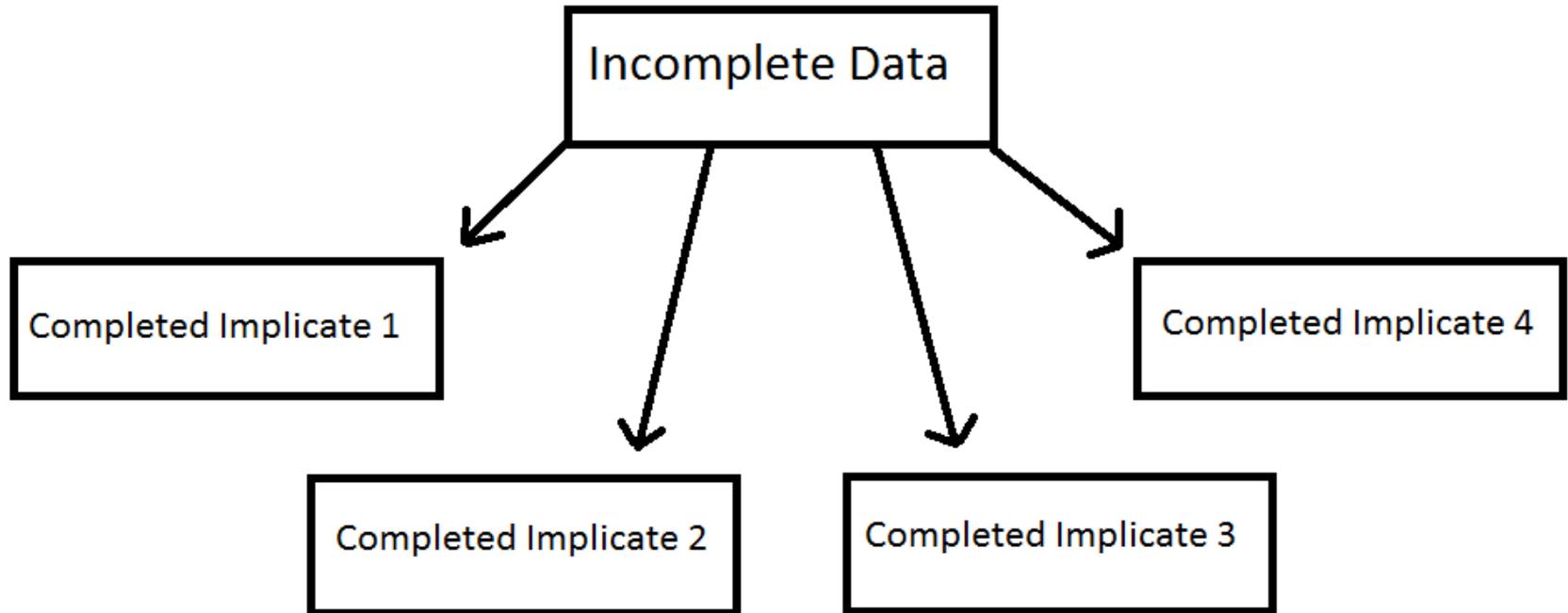








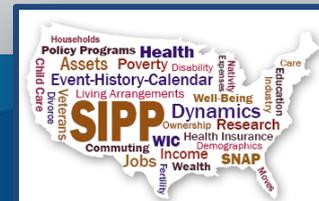
# Multiple Completed Implicates





# Estimate $p(x_4|x_1, \dots, x_3)$ & impute

Var1	Var2	Var3	Var4
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

































# Process Overview: Synthesize & Test

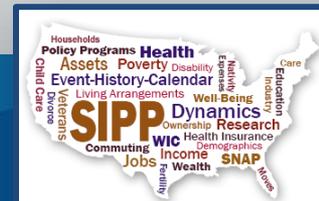
- Synthesis is basically the same as completion
  - Now replace all values with draws from estimated distributions
  - Non-iterative; once through list of variables
- Test synthetic data to see how close it lies to the completed data
- Test for disclosure risk
  - Very conservative approach
  - Public-use SIPP files already available





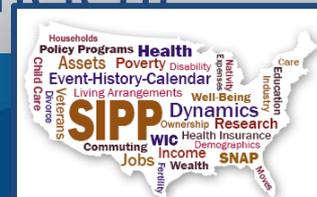
# Using the SSB: Requesting Validation of Synthetic Results

- Census will run your programs on the confidential data and release these results to you
- Summary of steps
  - Write code that works on one data set
  - Only piece of your code that will be changed by Census staff is file location and file base name; Have this place clearly marked
  - Test this code by running on each synthetic data set
  - Save code in special folder on the SDS server
  - Prepare memo asking for disclosure of results
  - Email memo to Census SSB staff and request validation
- Can take as little as 2-3 weeks
- Only works as well as your code



# Disclosure Review

- Census reviewer needs to see a person-level observation count for every number you produce
  - Regression – need sample size
  - Summary Statistics – need sample size of each sub-group for each variable
  - If probit or logistic regression, need to see cross-tab of any independent variables with dependent variable
  - Person-year observation counts are not sufficient

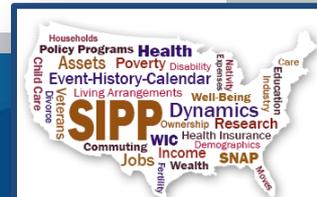
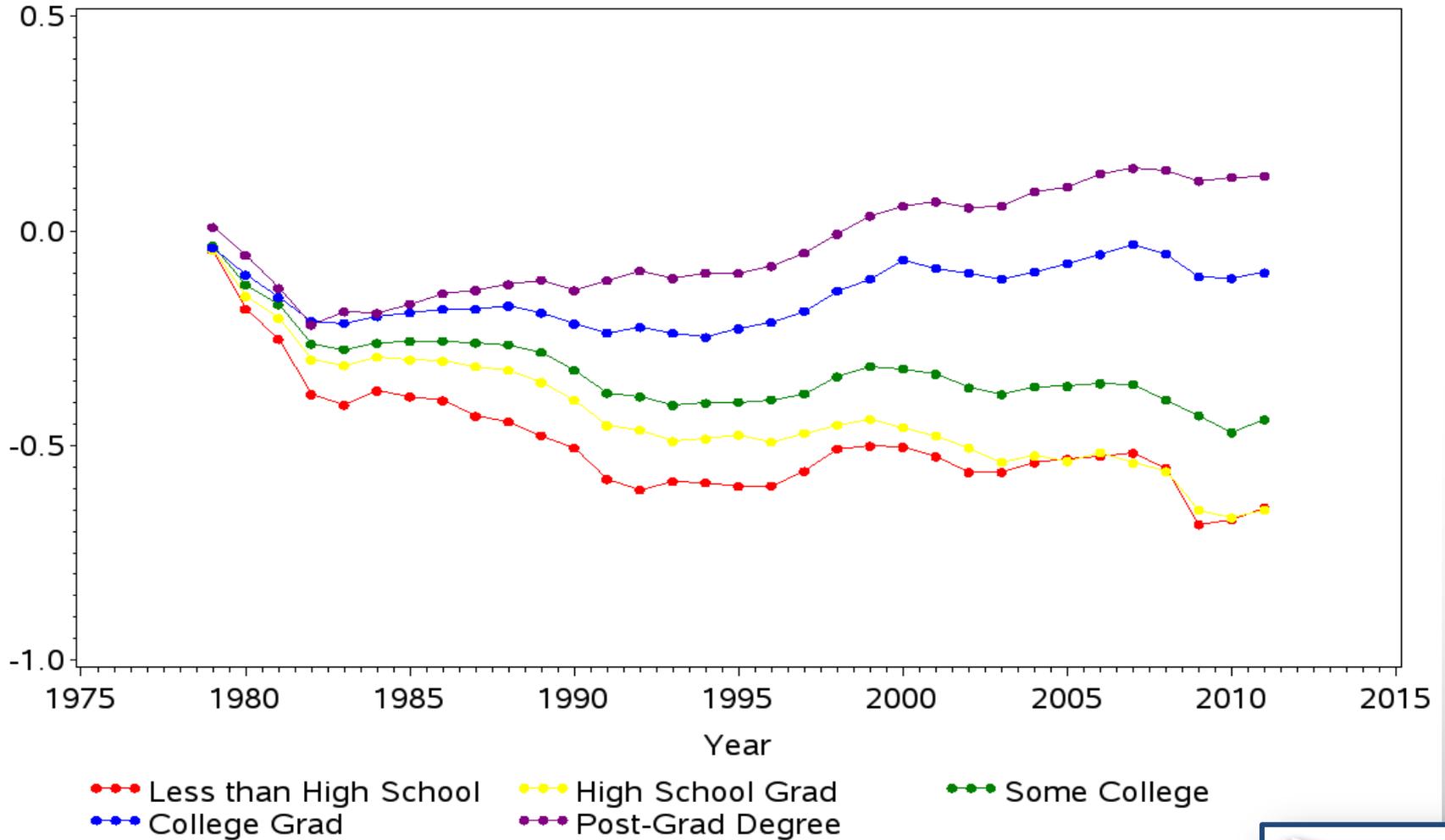






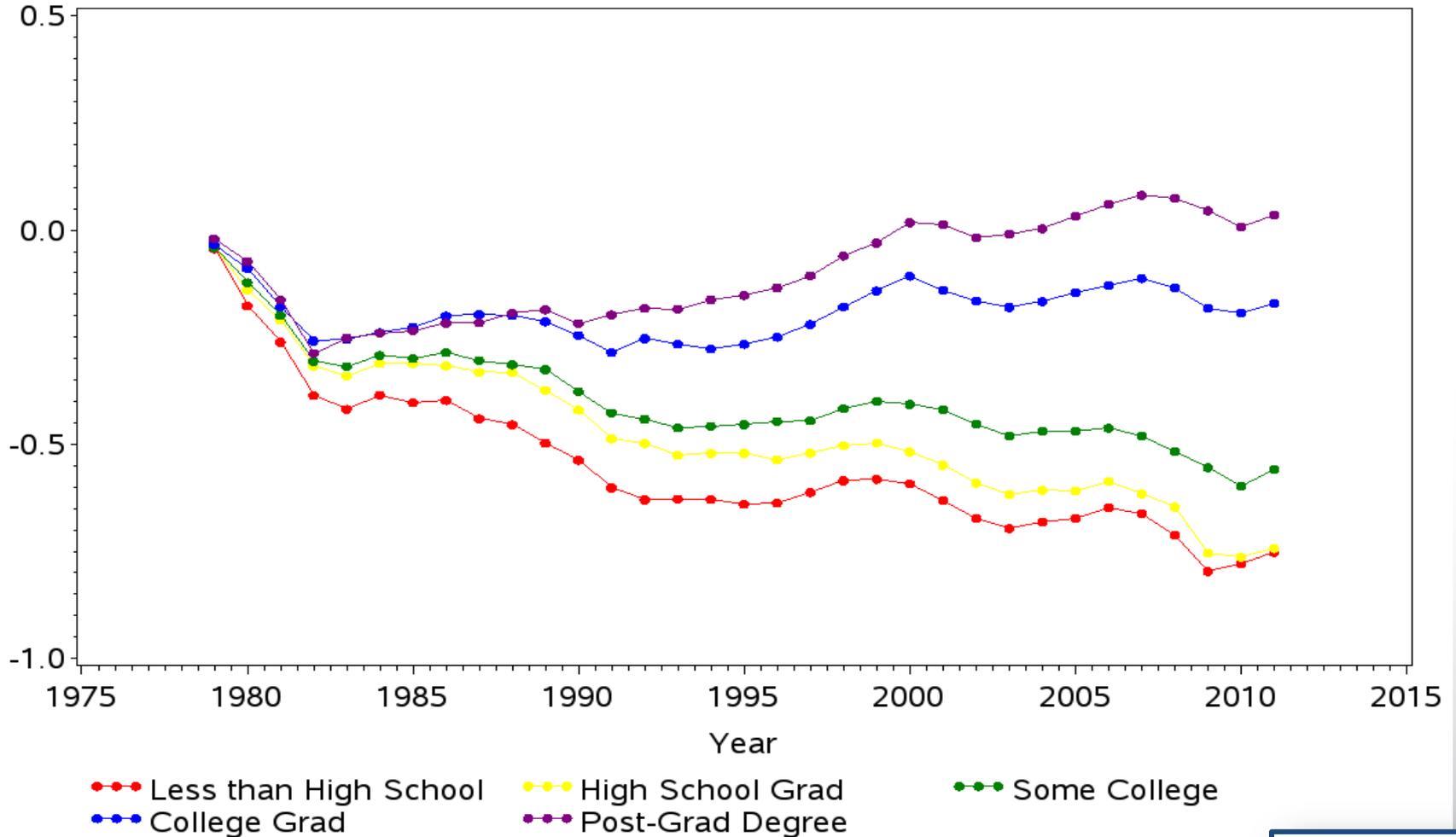
# Log Earnings Relative to 1978 for Males

Completed Data



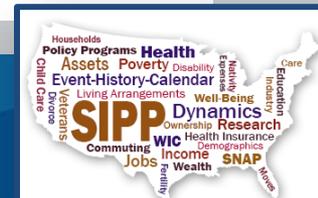
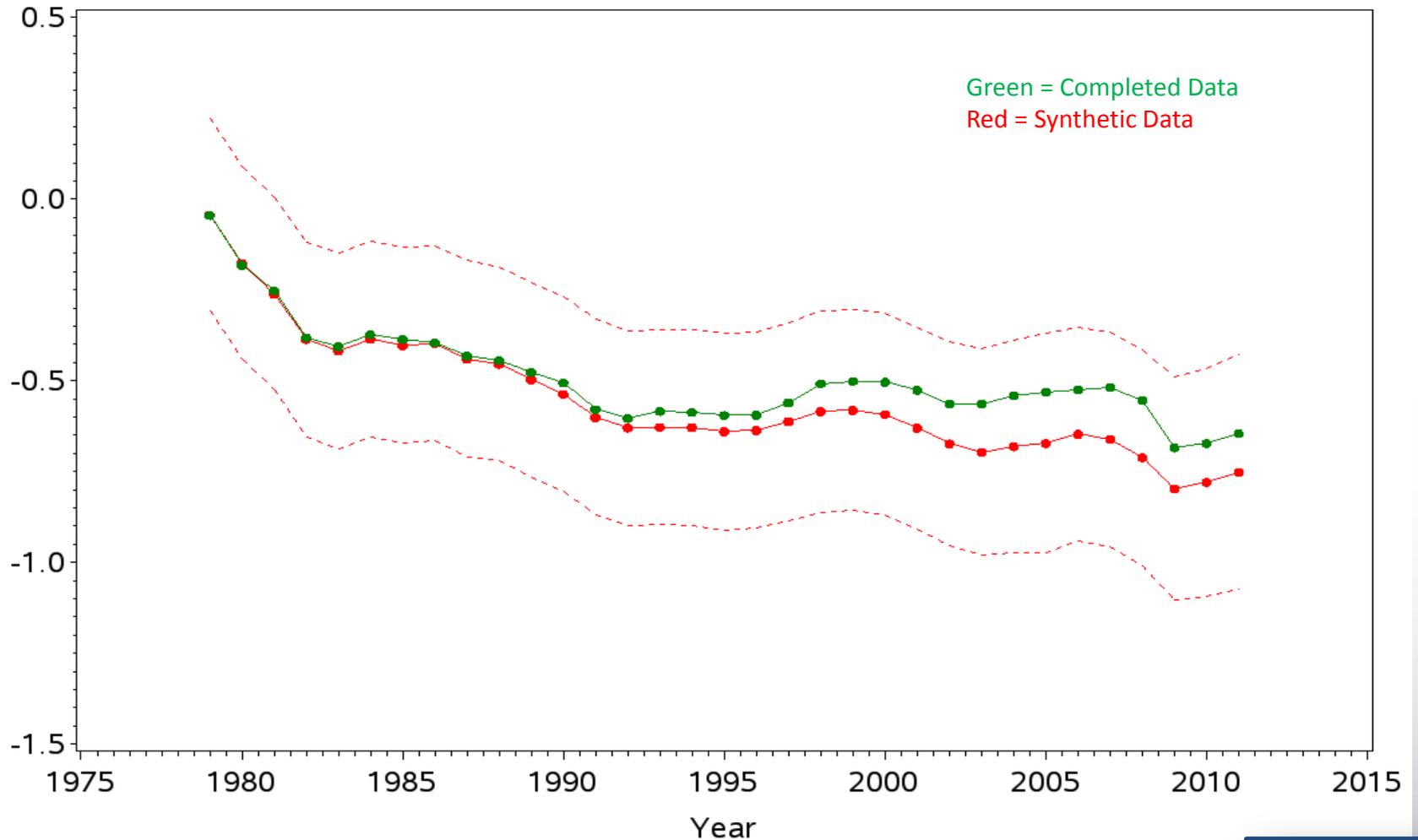
# Log Earnings Relative to 1978 for Males

Synthetic Data



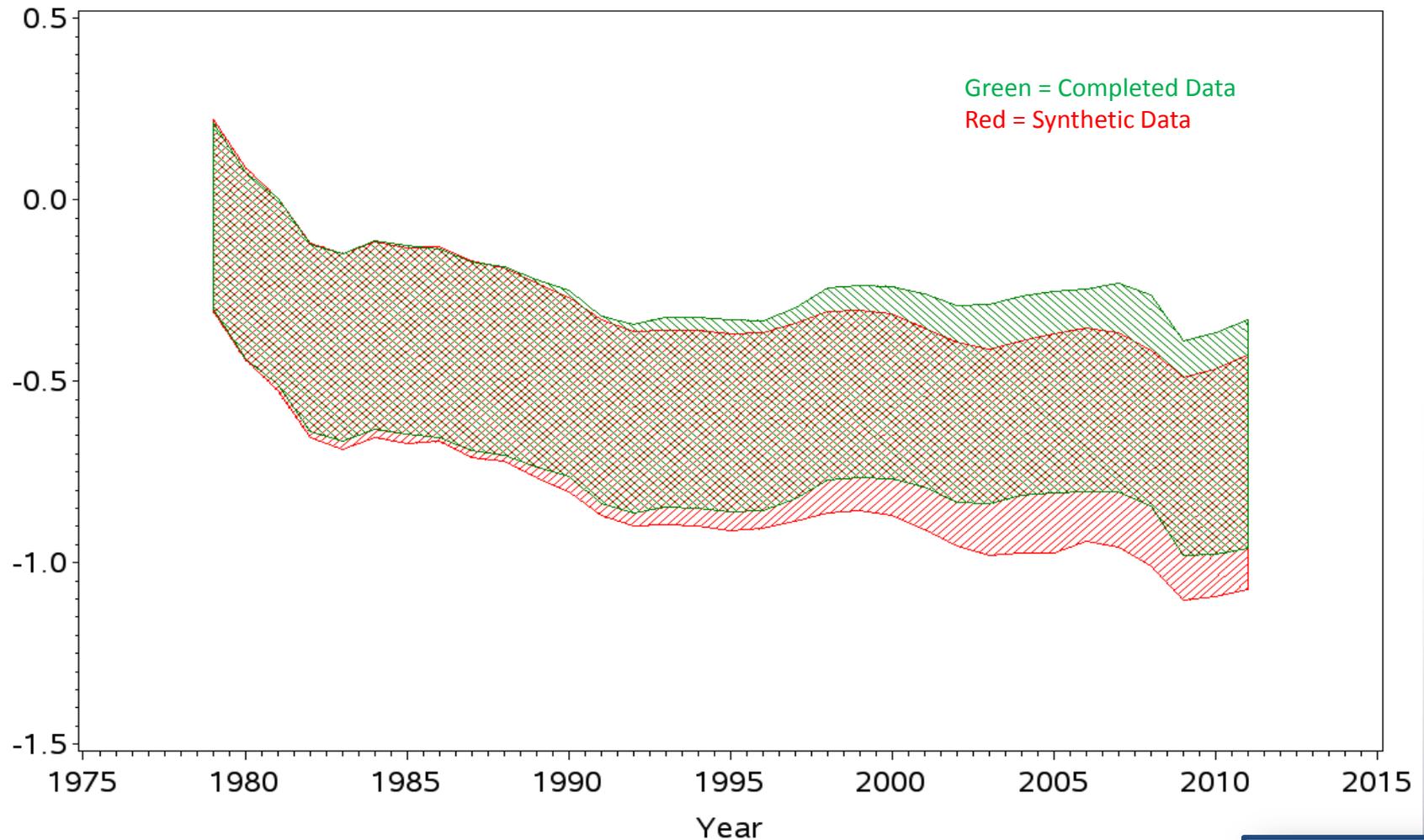
# Log Earnings Relative to 1978 for Males Without H.S. Diploma

## Comparison of Completed and Synthetic Data



# Log Earnings Relative to 1978 for Males Without H.S. Diploma

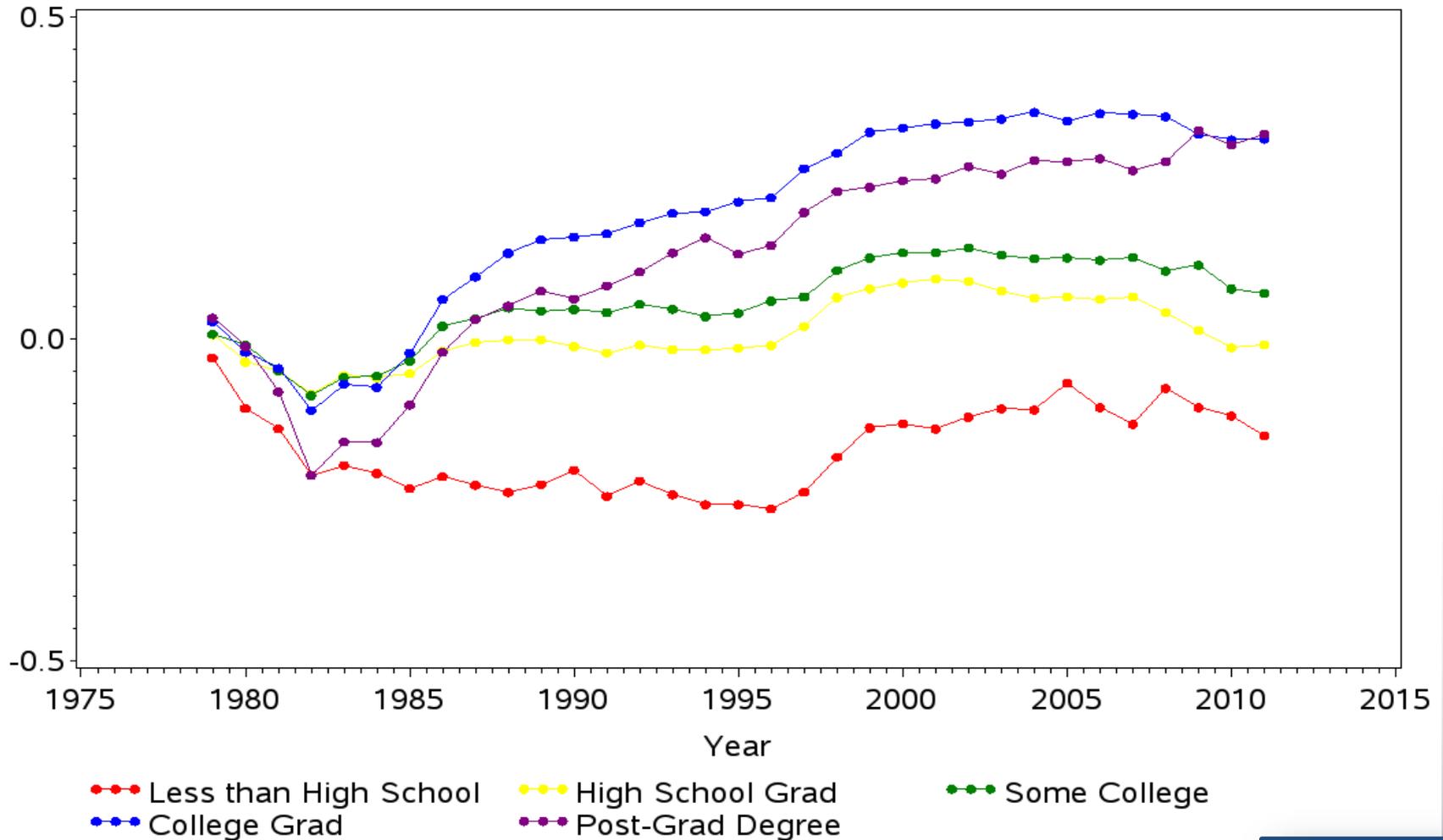
## Comparison of Completed and Synthetic Data





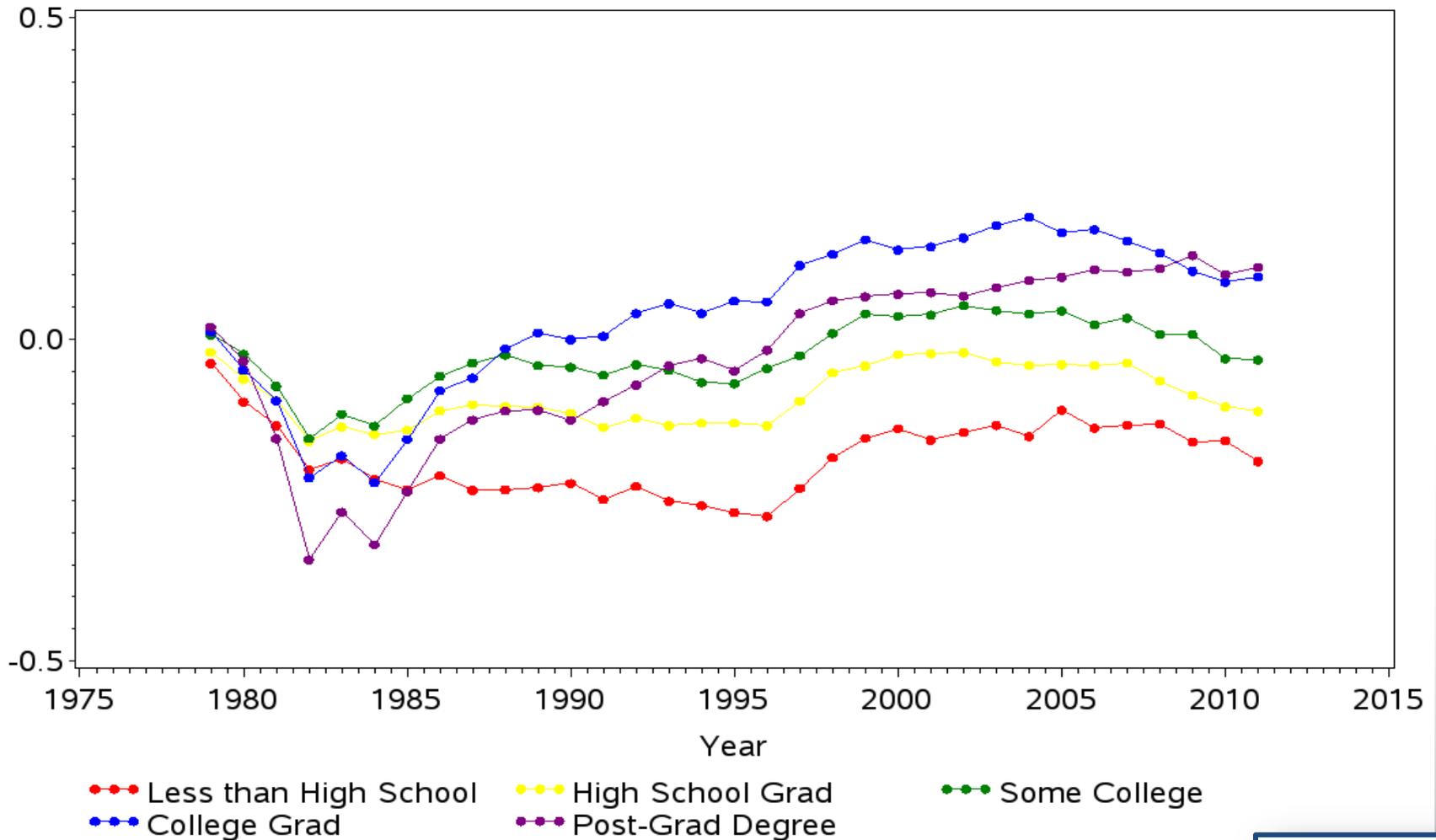
# Log Earnings Relative to 1978 for Females

Completed Data



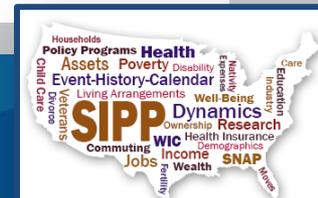
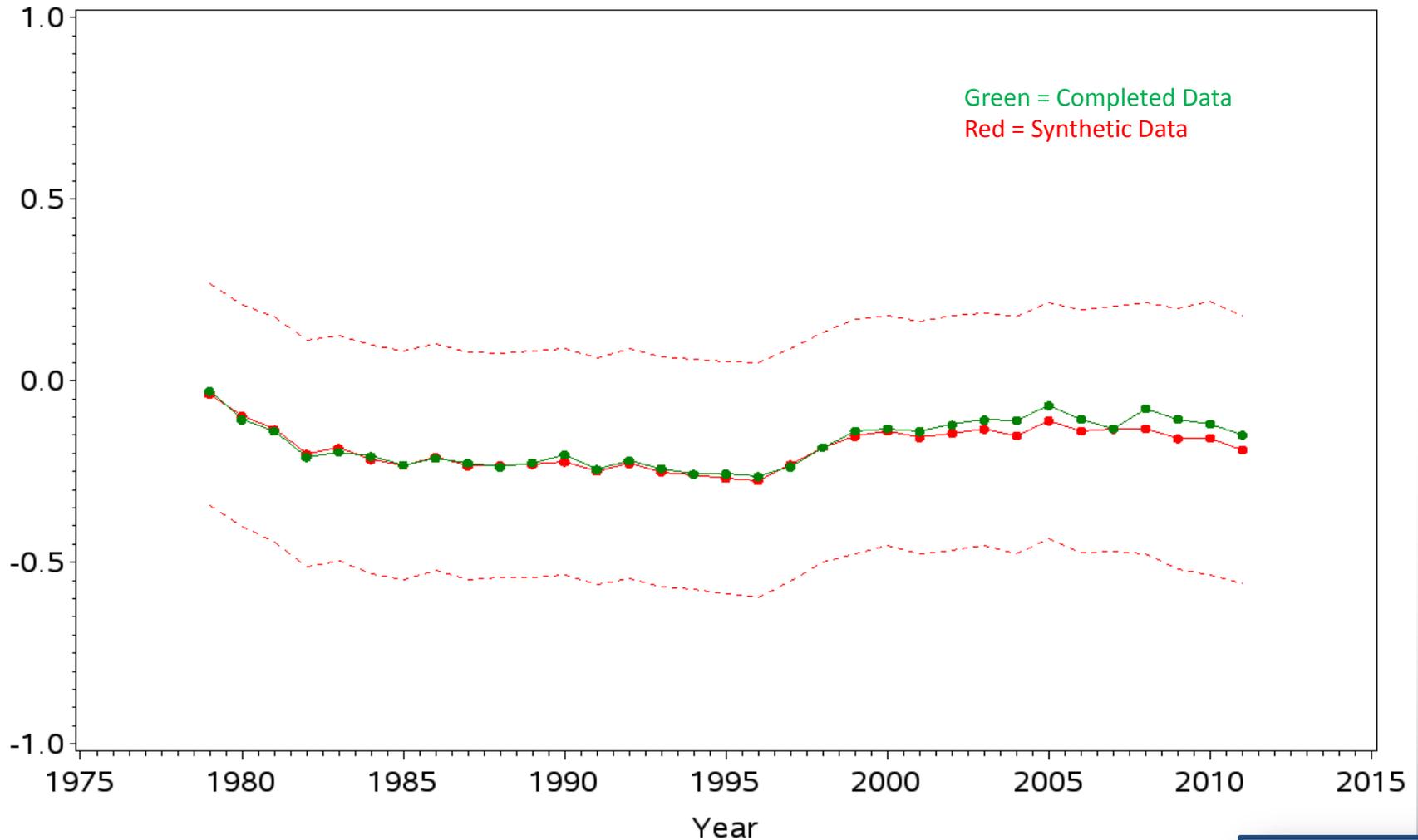
# Log Earnings Relative to 1978 for Females

Synthetic Data



# Log Earnings Relative to 1978 for Females Without H.S. Diploma

## Comparison of Completed and Synthetic Data



# Log Earnings Relative to 1978 for Females Without H.S. Diploma

## Comparison of Completed and Synthetic Data

