DUKE UNIVERSITY
OPM SYNTHETIC DATA PROJECT

Comparing Synthetic Human Capital Data to OPM Source Data

Conference on Analyzing Federal Personnel Data | June 2-3, 2016
Center for Institutional and Organizational Performance
Duke University | Durham, North Carolina
thomas.balmat@duke.edu

PURPOSE

- Develop systematic procedures for assessing homogeneity of synthesized CPDF data to source OPM data
- Emphasis on utility assessment (does a researcher reach the same conclusions using synthetic data as she would using source OPM data?)

EXAMPLE RESEARCH QUESTIONS USED IN COMPARISON

- What is the distribution of federal employees by sex, agency, occupation?
- Are sex, race, age, and education functional predictors of federal employee pay?
- Is probability of promotion a function of sex, race, age, and education?
- Do gender proportions by occupation change through time?

HUMAN CAPITAL CPDF DATA ASSETS

Requestor	Req Type	Req Year	Contents		Period	Files	Records	Code Tables
de Fig	FOIA	2012	DOD	Dynamic	1988-2011	92	58,996,920	OPM GDS 1999-2014
de Fig	FOIA	2012	DOD	Status	1988-2011	24	18,732,308	OPM GDS 1999-2014
de Fig	FOIA	2012	Non-DOD	Dynamic	1988-2011	92	84,108,247	OPM GDS 1999-2014
de Fig	FOIA	2012	Non-DOD	Status	1988-2011	24	28,257,629	OPM GDS 1999-2014
de Fig	FOIA	2014	DOD	Status	1988-2011	24	18,732,308	OPM GDS 1999-2014
de Fig	FOIA	2014	Non-DOD	Status	1988-2011	24	28,257,629	OPM GDS 1999-2014
Tim J	FOIA	2008	Non-DOD	Status	1973-2007	35	39,317,925	OPM SCT 1973-2007
Tim J	NARA	2009	DOD Non-DOD	Status	1973-1997	25	52,532,129	NAR SCT 1973-1982
Tim J	NARA	2009	Non-DOD	Survey	1979-1981 1983 1986 1988-1989 1991-1993 1996-1997 1999-2000 2005	26	683,326	Individual by Survey
Tim J	NARA	2009	TVA	HRIS	1992-1995 2000	5	222,644	TVA HR Codes 1992-2003
Tim J	FOIA	2012	Non-DOD	Dynamic	Mar-Sep 1979-2013 Jan-Jun 1973-2013	152	135,067,802	OPM SCT 1973-2013
Tim J	FOIA	2012	Non-DOD	Status	1973-2013	41	47,333,165	OPM SCT 1973-2013
de Fig	FedScope	2016	DOD Non-DOD	Status	Sep 1998 – Sep 2015	18	34,720,487	
						582	546,962,519	

SYNTHETIC DATA ASSETS

- Current version 5, May 2016
- 28,412,573 non-DOD status observations
- Variables synthesized:
  - Pseudo ID
  - Fiscal Year
  - Agency
  - Sex
  - Race
  - ERI Bridge
  - Education Level
  - Age Range
  - Years Since Degree Range
  - Instructional Program
  - Occupational Category
  - Occupation
  - Functional Class
  - FLSA
  - Type Appointment
  - Political Appointee Type
  - Position Occupied
  - Tenure
  - Supervisory Status
  - Pay Plan
  - Grade
  - Step Rate
  - Pay Basis
  - Pay Rate Determinant
  - Work Schedule
  - Basic Pay

COMPARISON DATA

- CPDF: Non-DOD status observations supplied by OPM to Duke University as a result of John de Figueiredo's 2012 FOIA request (referred to as **JdF-2012**) n = 28,257,629

- Synthetic data: Duke generated version 0.5 (referred to as **DIBBS-2016**) n = 28,412,573

- Variables in study:

  - Pseudo ID
  - Fiscal Year
  - Agency
  - Sex
  - Race
  - ERI Bridge
  - Education Level
  - Age Range
  - Occupational Category
  - Occupation
  - Pay Plan
  - Grade
  - Step Rate
  - Pay Basis
  - Work Schedule
  - Basic Pay

ADDITIONAL COMPARISON DATA

- FedScope 1998-2011 September CPDF data (source: www.opm.gov/data)
- Pseudo ID unavailable (needed for longitudinal studies, promotion, etc.)
- Education Level unavailable for 1998-2003
- Race unavailable
- FedScope Salary = CPDF Adjusted Basic Pay (not synthesized)

DATA ANALYSIS TECHNOLOGIES AND METHODS

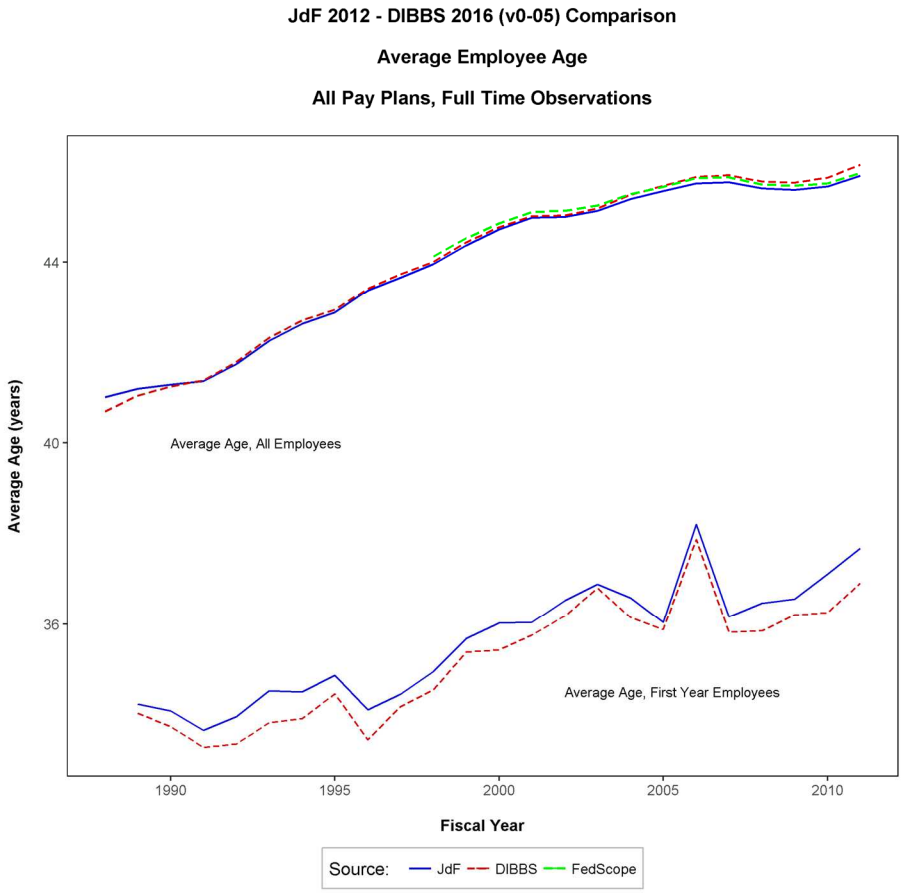
- SQL for synthetic data versioning and comparative data set generation
- R for statistical, graphical, and numerical operations
- Observation frequency, proportion, cumulative density comparison, χ^2 tests
- Regression: ordinary least squares, logistic regression, quantile regression
- Parallel processing: X'X indicator parsing, sparse matrix solutions
- n

QUESTIONS?

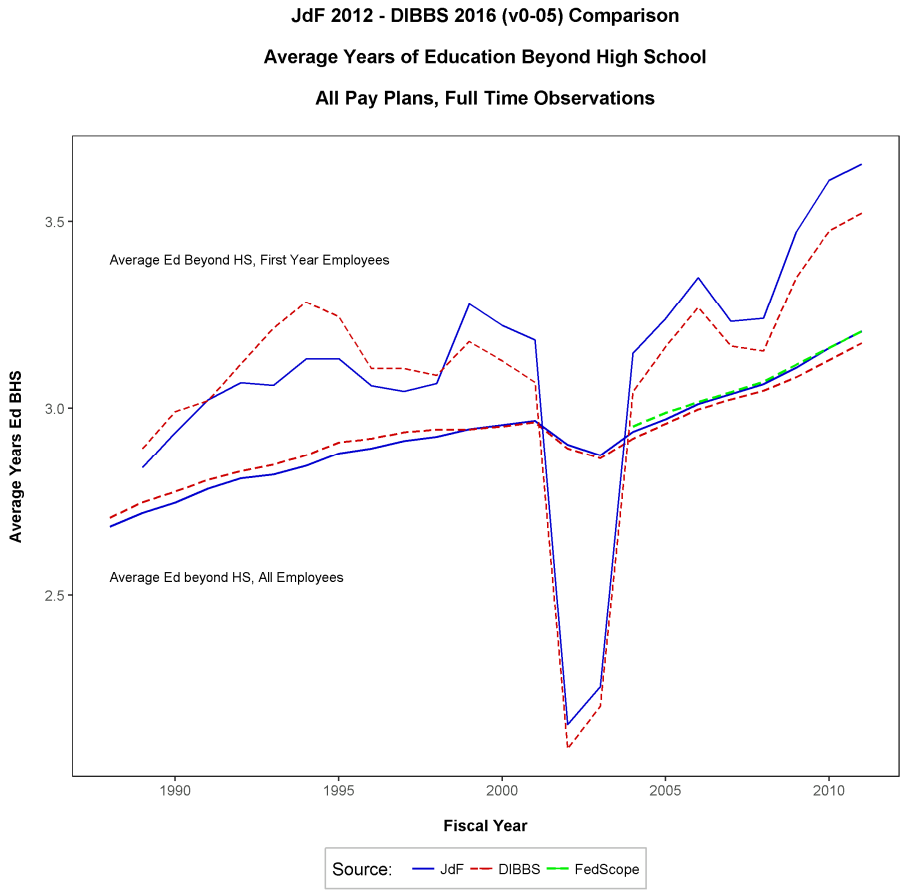
SUMMARY COMPARISONS

From Grade Inflation Research

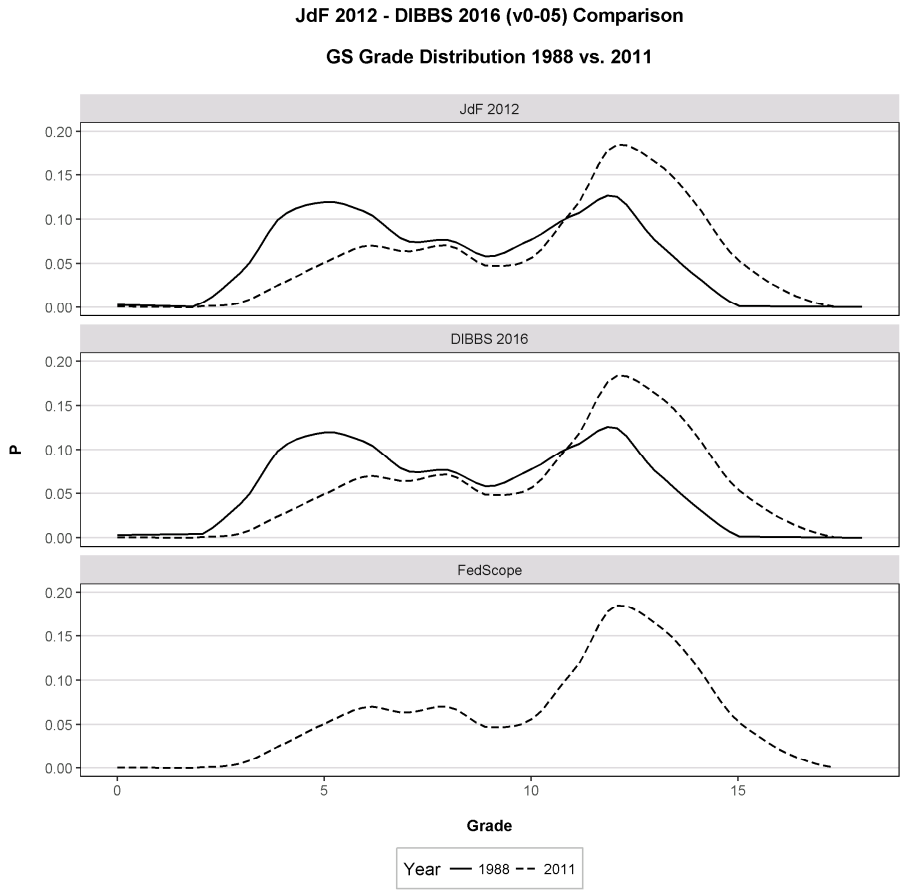
AVERAGE FEDERAL EMPLOYEE AGE, 1988-2011



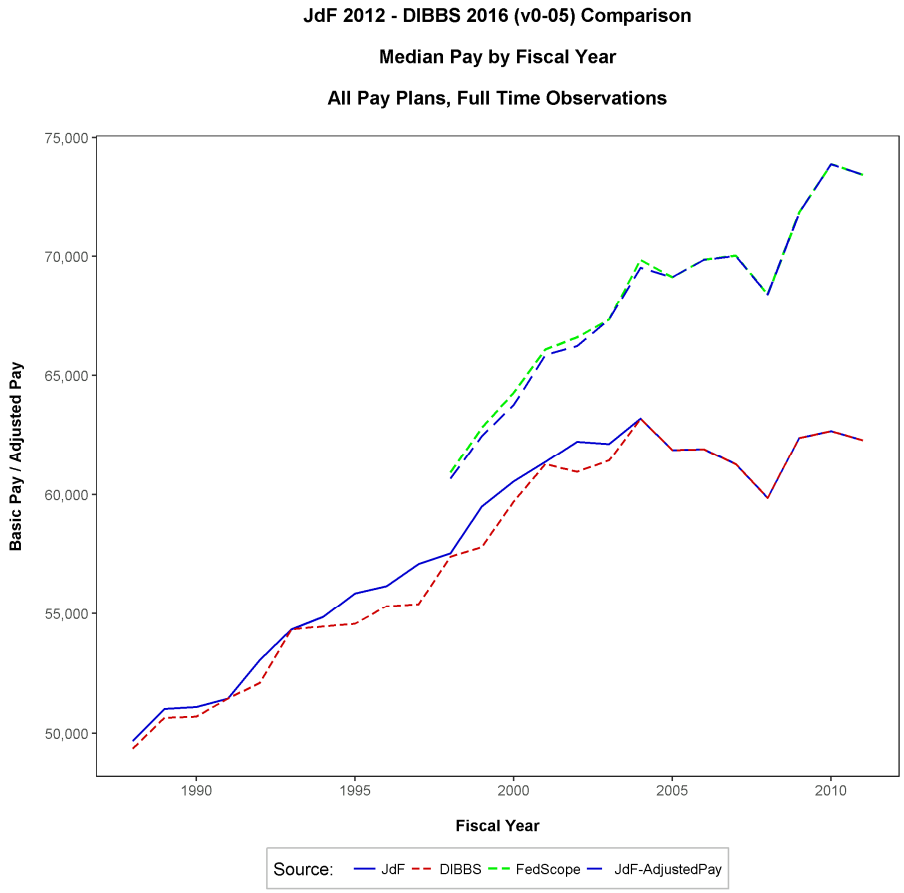
AVERAGE FEDERAL EMPLOYEE EDUCATION, 1988-2011



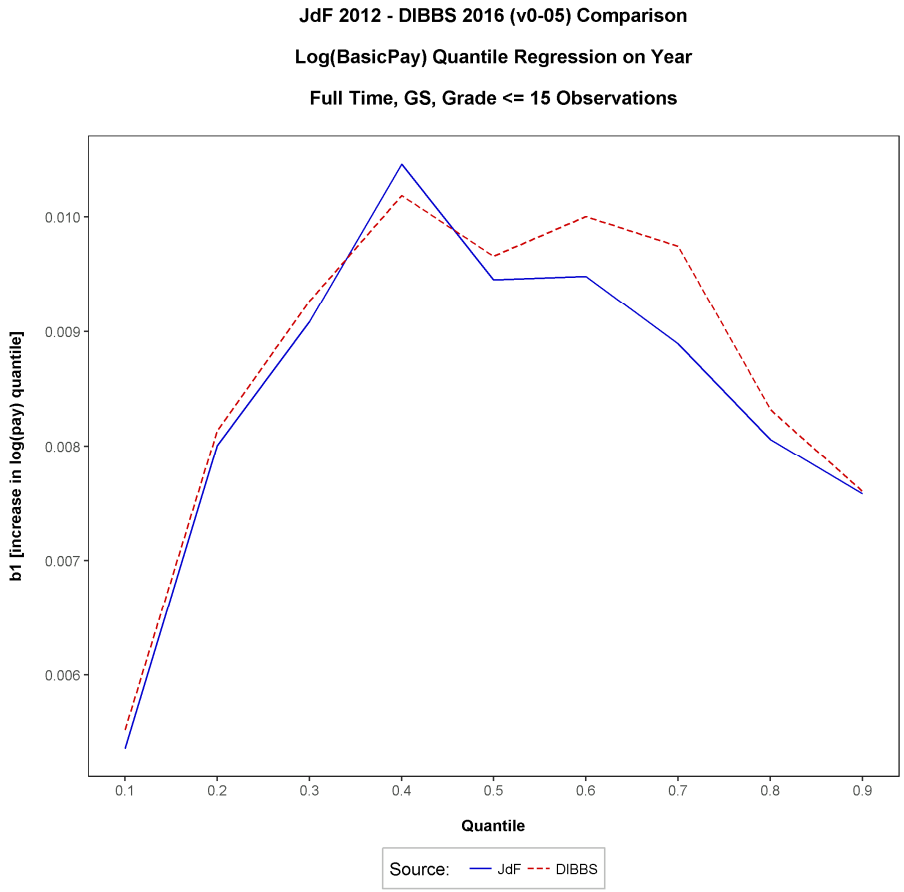
DISTRIBUTION OF GS EMPLOYEES BY GRADE, 2011 VS. 1998



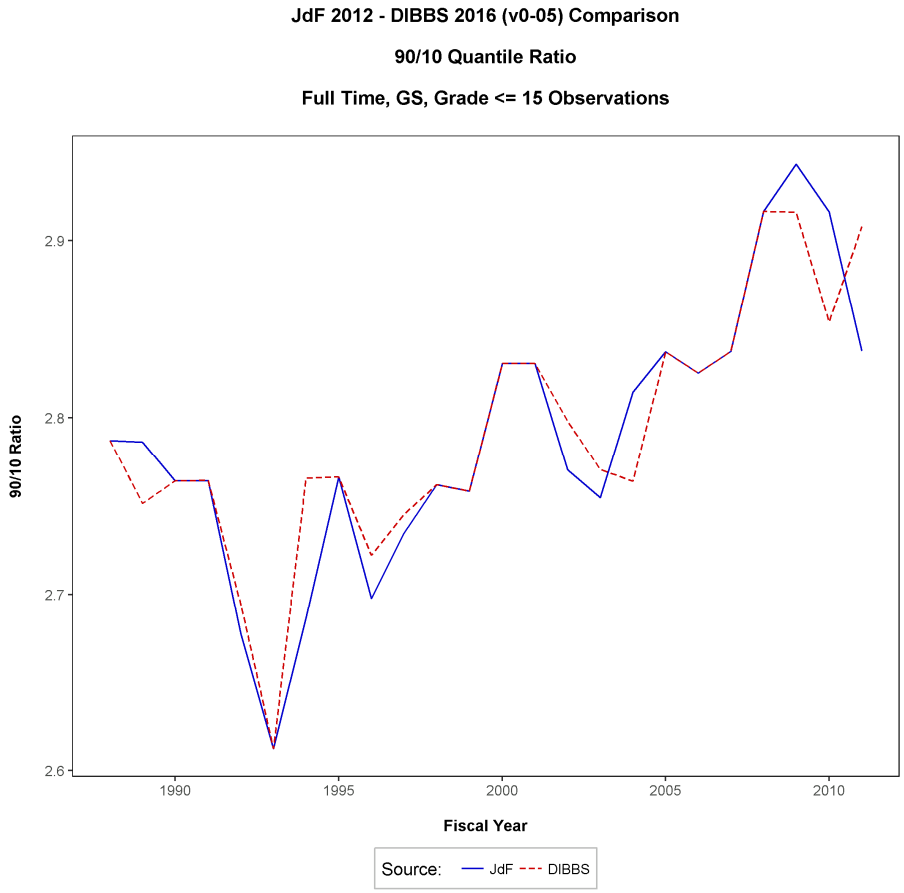
MEDIAN FEDERAL EMPLOYEE PAY, 1988-2011



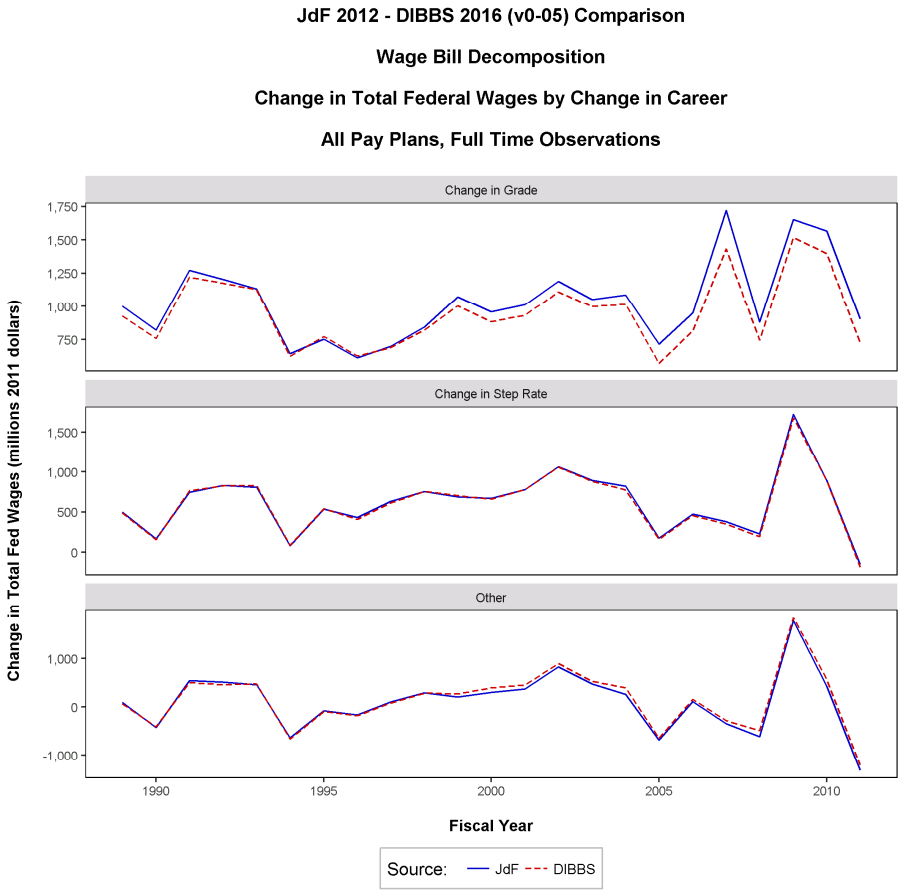
ANNUAL RATE OF CHANGE IN PAY PERCENTILES, GS PAY PLAN, 1988-2011



RATIO OF 90TH TO 10TH PAY PERCENTILES, GS PAY PLAN, 1988-2011



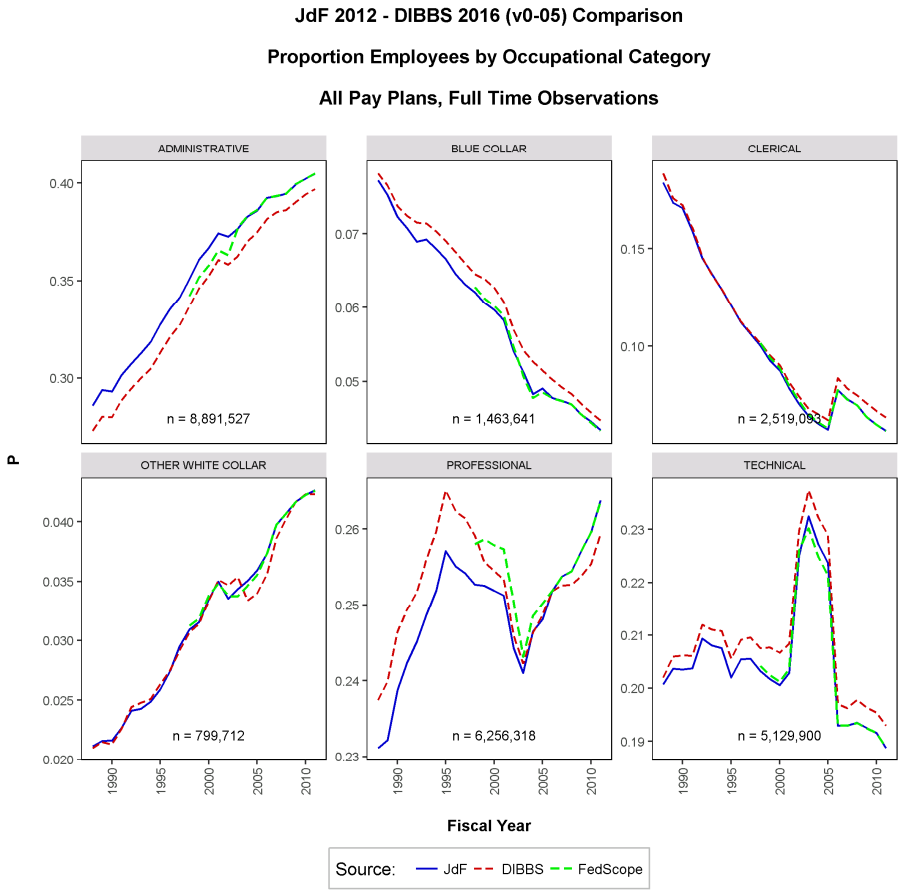
ANNUAL CHANGE IN FEDERAL PAYROLL DUE TO INCREASES, 1989-2011



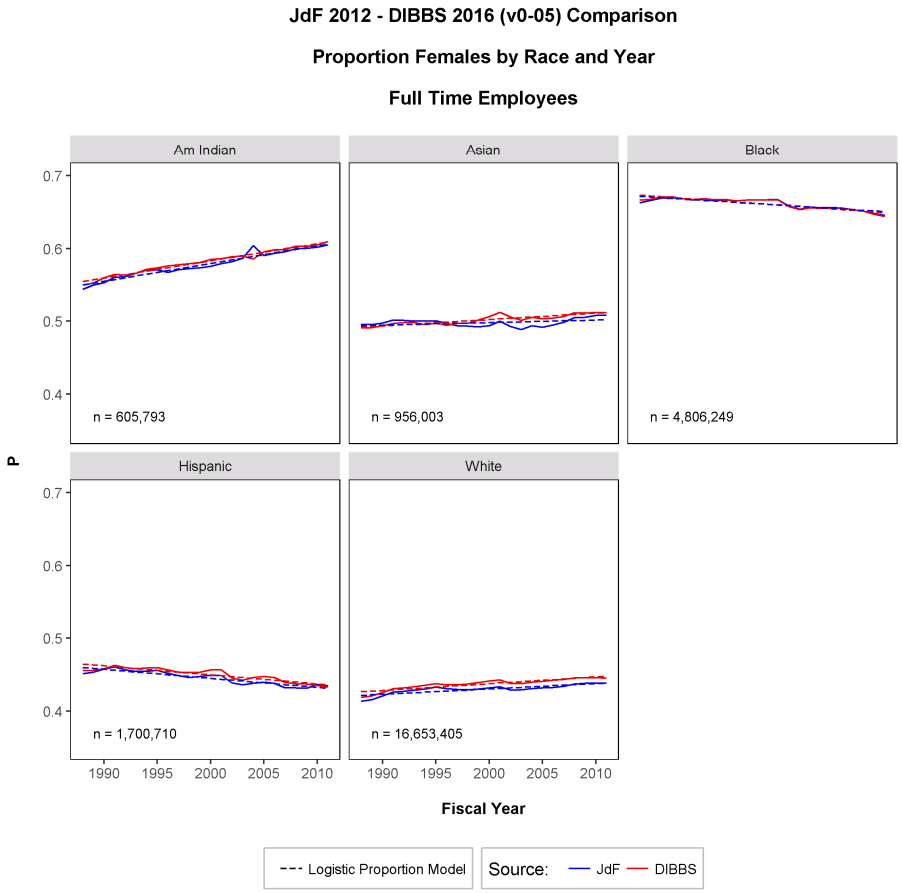
DISTRIBUTION MODELS

From Gender Pay Differential Research

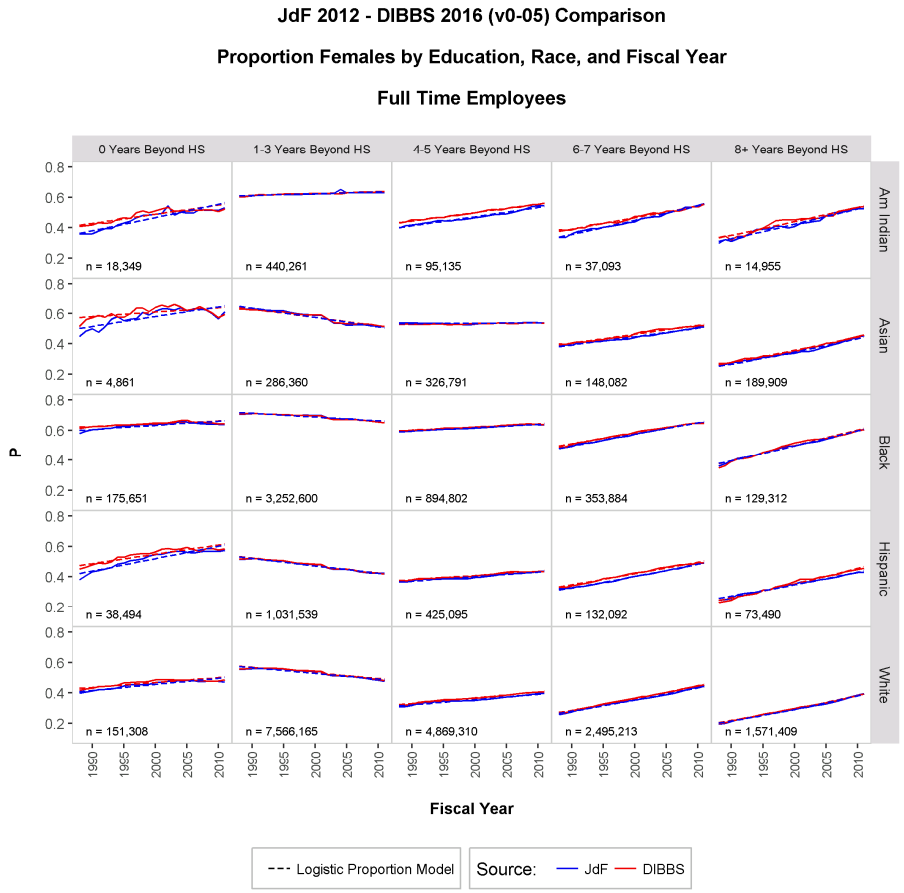
DISTRIBUTION OF FEDERAL EMPLOYEES BY OCCUPATIONAL CATEGORY, 1988-2011



PROPORTION FEMALES BY RACE, 1988-2011



PROPORTION FEMALES BY RACE AND EDUCATION, 1988-2011



REGRESSION MODELS

From Grade Inflation and Gender Pay Differential Research

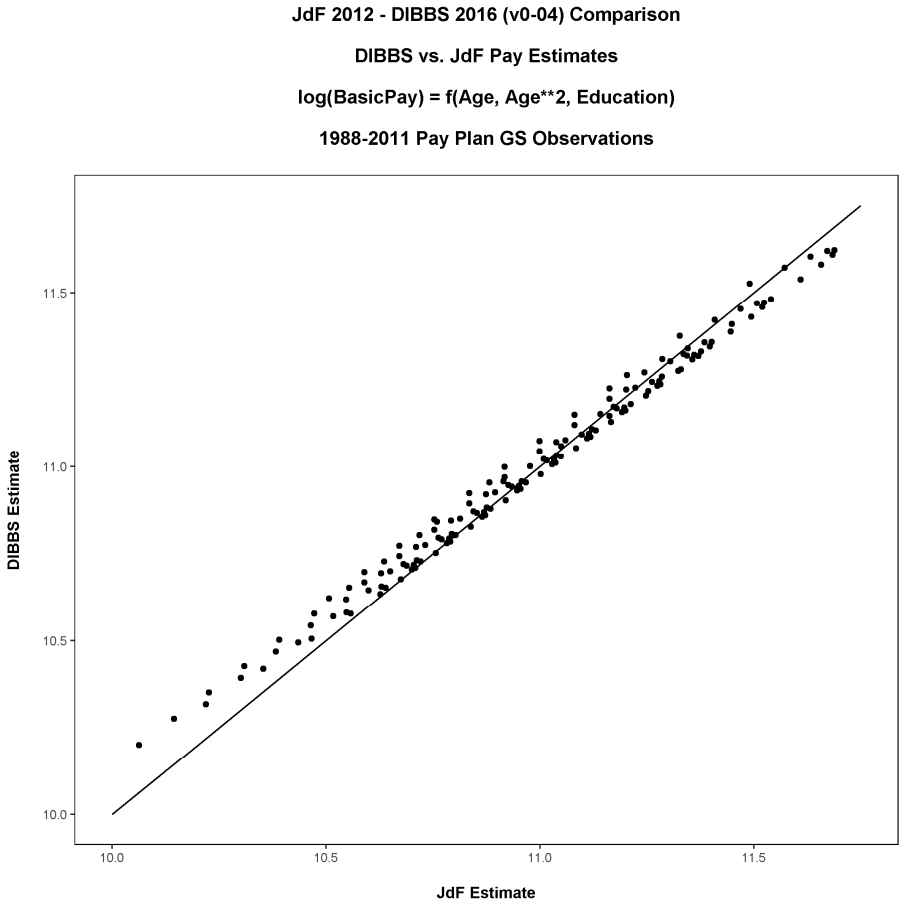
SIMPLE OLS PAY MODEL

Age and Education Predictors

Duke OPM Synthetic OLS Model 1 Analysis Model: ln(BasicPay) = Age + Age^2 + Education				
	JdF		DIBBS	
Coefficient	Estimate	SE	Estimate	SE
Intercept	9.3700	0.0010	9.6951	0.0010
Age	0.0481	4.0E-05	0.0348	4.0E-05
Age^2	-0.0004	5.0E-07	-0.0003	5.0E-05
Education	0.0819	3.0E-05	0.0753	3.0E-05

SIMPLE OLS PAY MODEL

Synthetic Predicted Pay vs. Predicted Pay from JdF Model



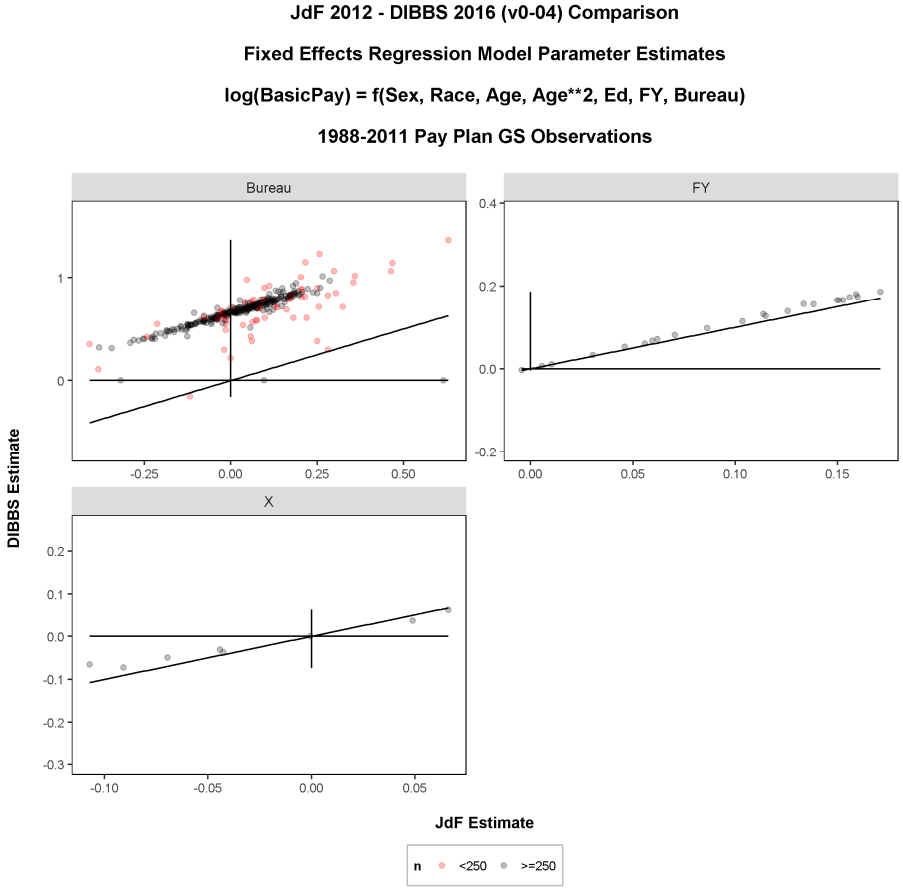
FIXED EFFECTS OLS PAY MODEL

Age and Education Predictors; Sex, Race, Agency, and Year Fixed Effects

Duke OPM Synthetic OLS Model 2 Analysis Model: ln(BasicPay) = Age + Age^2 + Education + Agency + Year				
	JdF		DIBBS	
Coefficient	Estimate	SE	Estimate	SE
Intercept	9.5025	0.0734	9.119	0.0414
Age	0.0489	4.1E-05	0.0365	3.9E-05
Age^2	-0.0004	4.6E-07	-0.0003	4.4E-07
Education	0.0661	3.0E-05	0.0617	3.1E-05
Sex	-0.0910	0.0001	-0.0723	0.0001
Race	√		√	
Agency	√		√	
Year	√		√	

OLS PAY MODEL WITH SEX, AGENCY, AND YEAR FIXED EFFECTS

Parameter Estimate Homogeneity



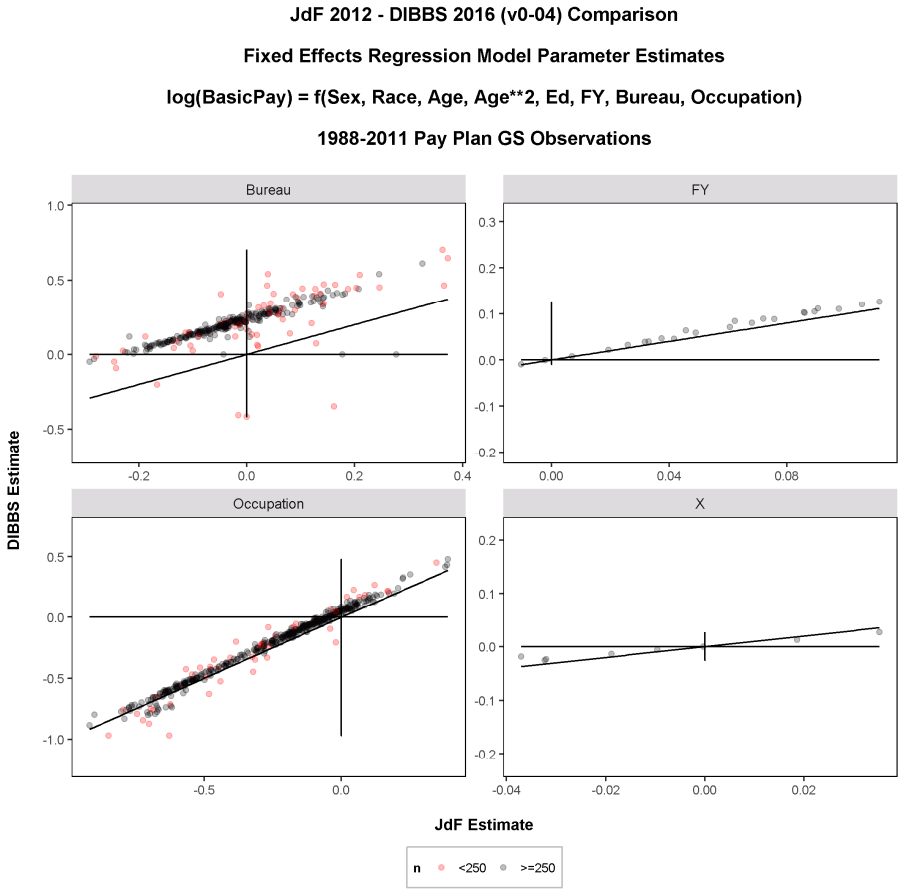
FIXED EFFECTS OLS PAY MODEL

Age and Education Predictors; Sex, Race, Agency, Year, and Occupation Fixed Effects

Duke OPM Synthetic OLS Model 3 Analysis Model: ln(BasicPay) = Age + Age^2 + Education + Agency + Year + Occupation				
	JdF		DIBBS	
Coefficient	Estimate	SE	Estimate	SE
Intercept	10.2621	0.0520	10.2127	0.0298
Age	0.0352	2.9E-05	0.0266	2.9E-05
Age^2	-0.0003	3.3E-07	-0.0002	3.3E-07
Education	0.0186	2.7E-05	0.0125	2.8E-05
Sex	-0.0320	0.00011	-0.0234	0.0001
Race	√		√	
Agency	√		√	
Year	√		√	
Occupation	√		√	

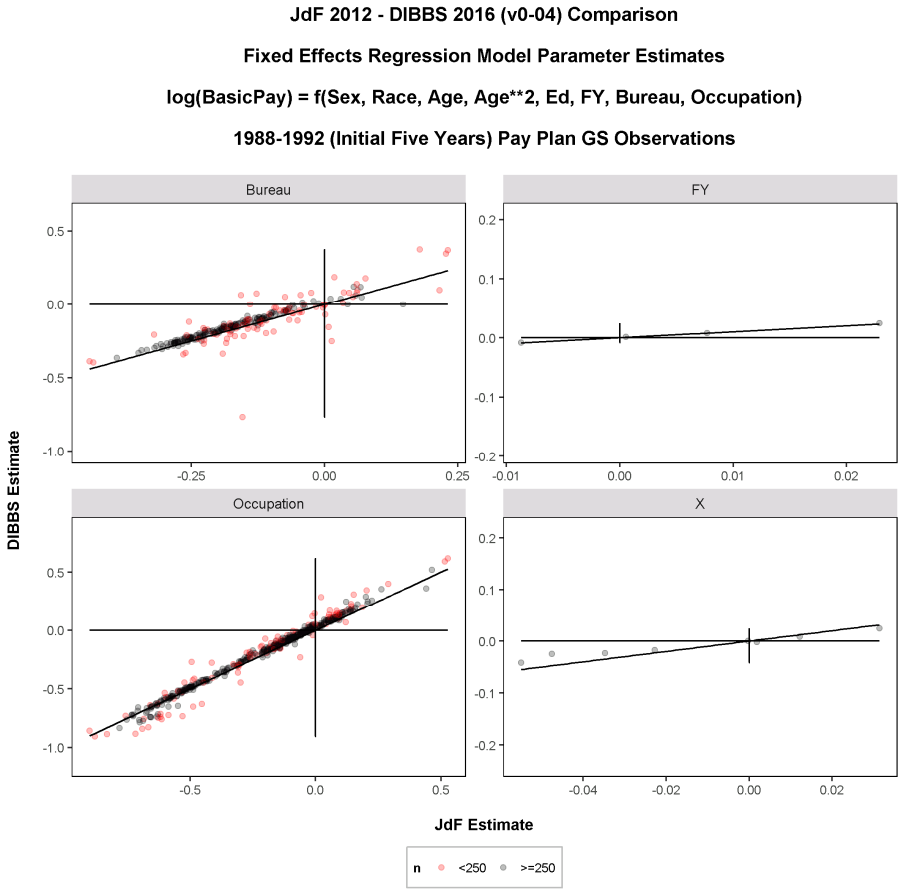
OLS PAY MODEL WITH SEX, RACE, AGENCY, YEAR, AND OCCUPATION FIXED EFFECTS

Parameter Estimate Homogeneity



OLS PAY MODEL WITH SEX, RACE, AGENCY, YEAR, AND OCCUPATION FIXED EFFECTS

Initial Five Years of Study Data

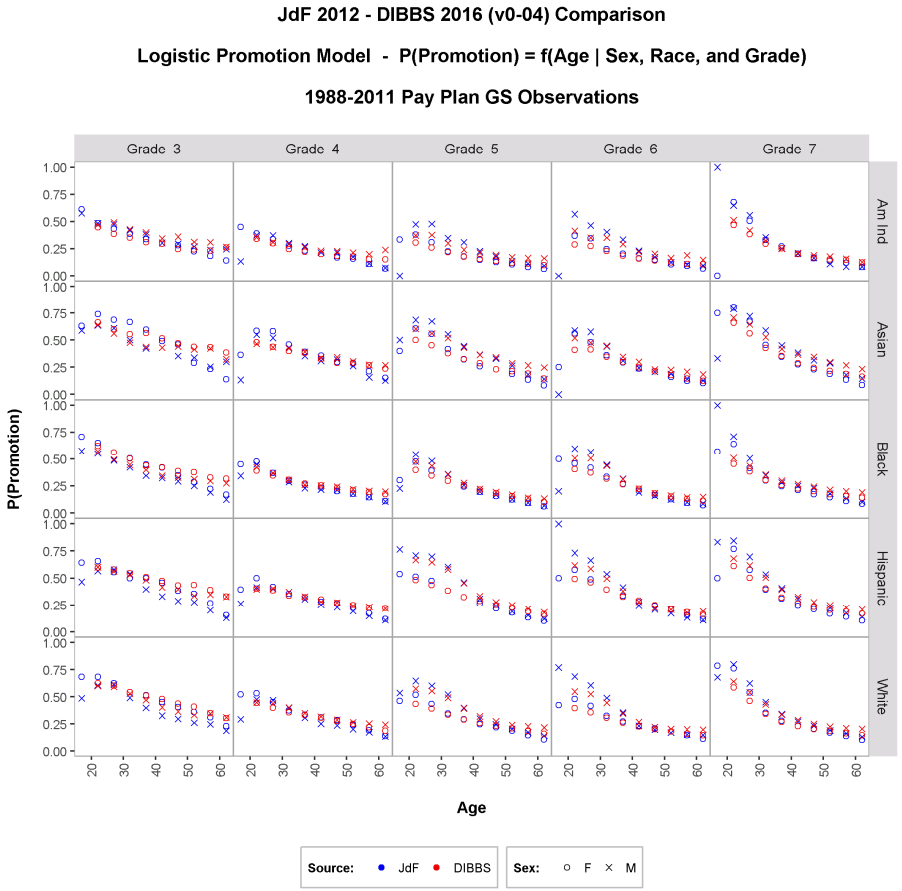


PROMOTION MODELS

From Gender Pay Differential Research

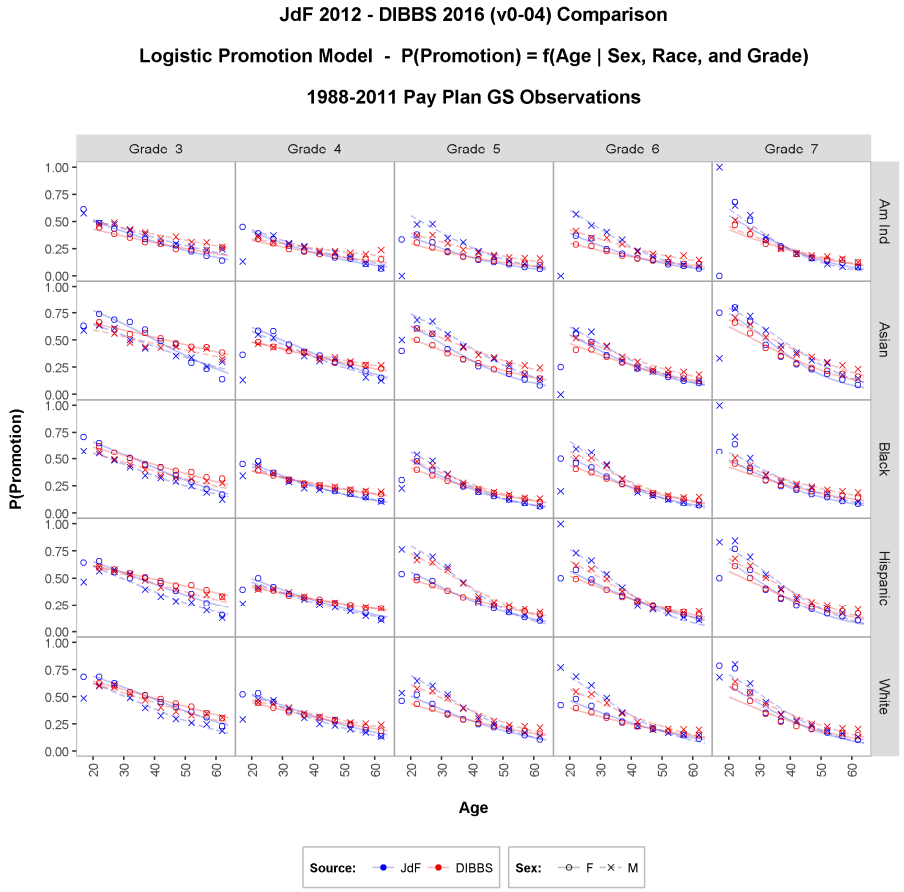
PROPORTION EMPLOYEES PROMOTED BY SEX, RACE, AGE, AND GRADE

Synthetic Proportions Compared to Actual



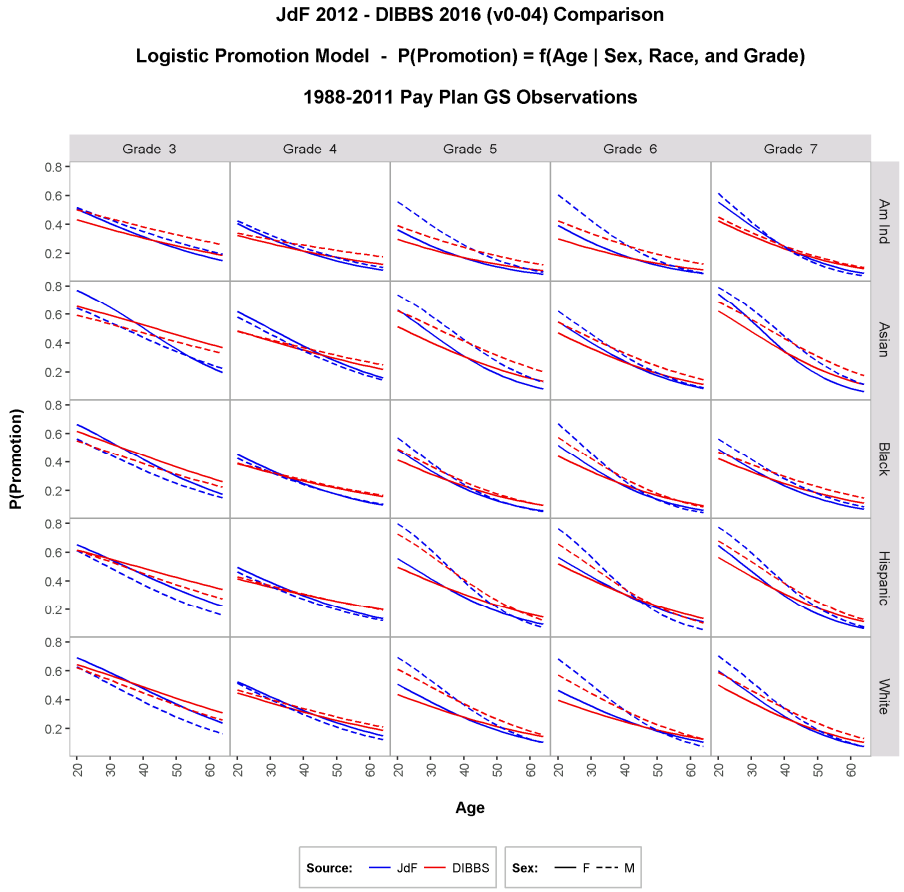
PROPORTION EMPLOYEES PROMOTED WITH LOGISTIC REGRESSION MODEL

Probability of Promotion = f(Sex, Race, Age, Grade)



COMPARISON OF MODELS DERIVED FROM SYNTHETIC AND ACTUAL DATA

Probability of Promotion = f(Sex, Race, Age, Grade)



COMING ATTRACTIONS

- Gender pay quantile regression model
- GS -> SES promotion proportional hazards regression model

CONCLUSION

- Key summary statistics and distributions (mean age by year, grade distribution, gender proportions) derived from synthetic data strongly agree with those observed in the source OPM data

- Important research models (pay, promotion, OLS, logistic regression) based on synthetic data provide meaningful estimates with strong to moderate agreement with those derived from source OPM data

- Demanding models (gender differential pay quantile regression, occupational category switching by age/education regression) are being assessed

- The iterative process of data synthesis → model and fit analysis → synthesis model adjustment → data synthesis yields improved utility with each iteration (currently at version 5)

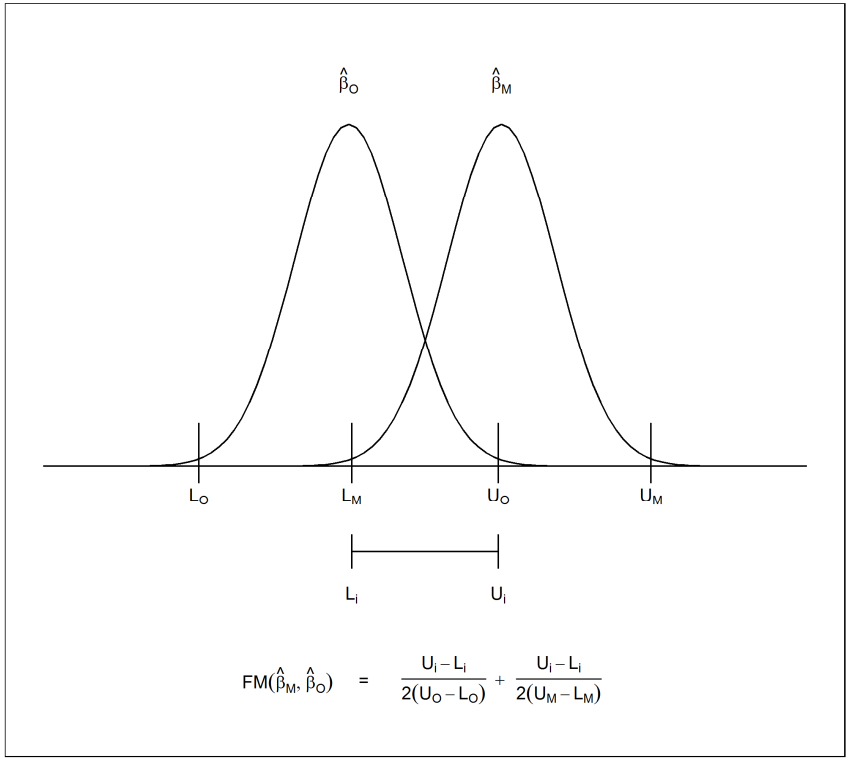
QUESTIONS?

ADDENDUM: VERIFICATION SERVER FIDELITY MEASURE EXAMPLE

Comparing Estimates from Synthetic (M, or masked) and Original (O) Data (taken from Reiter, 2009)

Synthetic Data Verification Server

Fidelity Measure



$$FM(\hat{\beta}_M, \hat{\beta}_O) = \frac{U_i - L_i}{2(U_O - L_O)} + \frac{U_i - L_i}{2(U_M - L_M)}$$