The Earnings Impact of Graduating from College during a Recession*

Preliminary draft prepared for submission to 2011 SOLE Meetings

Gary Benedetto, Graton Gathright†, and Martha Stinson

US Census Bureau

October 2010

Abstract

We test the prediction that labor market conditions upon initial labor force participation have a persistent impact on earnings. Using data from seven panels of the Survey of Income and Program Participation linked to earnings records for 1978 to 2006 from the Social Security Administration's Master Earnings File, we estimate the impact of the unemployment rate faced upon college graduation on male college graduates' earnings for up to 15 years after graduation. We find no evidence of a persistent impact of the graduation-year unemployment rate on earnings beyond the graduation year. We repeat our analysis on Version 5.0 of the SIPP Synthetic Beta (SSB), a recently released public-use synthetic version of these linked data, as a test of the analytical validity of the SSB. We obtain estimates from the SSB that are similar in sign and magnitude to estimates from the underlying confidential data.

JEL Classification: J31, J64

1 Introduction

Two recent papers present evidence that young men graduating from college in periods of high unemployment continue to suffer negative labor market consequences even after labor market conditions

*We would like to thank Judy Eargle and David Johnson for helpful comments and support. This paper is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological, technical and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. Some of the data used in this paper are confidential.

†Submitting author (graton.m.gathright@census.gov).

improve. Using Canadian administrative earnings and university data, Oreopoulos, von Wachter, and Heisz (2008) find that the unemployment rate faced upon college graduation has a negative impact on the earnings of male college graduates and that this impact persists for five or more years. Kahn (2010) finds that, even fifteen years after graduation, male college graduates in the National Longitudinal Survey of Youth 1979 (NLSY79) continue to experience a negative wage impact from graduating in a period of high employment.

In this paper, we perform an analysis similar to Oreopoulos, von Wachter, and Heisz (2008) using the administrative earnings for a large sample of college graduates from the Survey and Income and Program Participation (SIPP). In these data, we find no evidence of persistence in the earnings impact of entering the labor force during a period of high unemployment. We estimate a variety of empirical specifications to assess the robustness of our (lack of) findings.

Economic theory suggests that labor market conditions upon entering the labor force may have a lasting impact on earnings since fewer job opportunities might lead to lower quality employee-employer matches in which the workers develop less (or less valuable) human capital. The focus of our analysis is to test this theoretical prediction.

The data that we analyze come from the SIPP Gold Standard file, a restricted-use data product of the US Census Bureau consisting of data from multiple SIPP panels linked to administrative records from the US Social Security Administration. We use reported educational attainment and year of college graduation from SIPP to match the individuals in our sample to the national unemployment rate for the year of their college graduation.

We estimate models of log real annual earnings for college graduates as a linear function of the interaction of the graduation-year unemployment rate with indicators for the years of potential post-college work experience as well as additional conditioning variables. Following Kahn (2010), we address the possibly endogenous timing of graduation by instrumenting for the graduation-year unemployment rate with the unemployment rate in the twenty-second year of age for each sample

person.[1]

We repeat some of our estimations on Version 5.0 of the SIPP Synthetic Beta, a recently released public-use synthetic version of the SIPP Gold Standard. We find substantial similarity in the results across the two datasets and are encouraged about the usefulness of this new synthetic data product.

In the next section, we describe the samples and the analysis variables used in this study. In Section 3, we describe our empirical specifications. We discuss our results in Section 4 and conclude in Section 5.

2 Data

In this section we describe our samples and analysis variables. We create three analysis datasets, one from each of three sources: the SIPP Gold Standard (SGS), the *completed* SGS, and the SIPP Synthetic Beta (SSB). The SIPP Gold Standard is a confidential Census Bureau dataset that consists of data from the 1990 to 2004 panels of SIPP[2] linked to administrative records from the Social Security Administration. The SIPP Synthetic Beta (SSB) is a public-use synthetic version of these data designed to provide broad access to some of the advantages of linked survey and administrative data. As an intermediate step in the creation of the SSB, missing data in the SGS is multiply-imputed resulting in a set of files referred to as the completed SGS. The creation of the SGS, completed SGS, and SSB is documented in more detail in the Appendix. The structure of the SSB mirrors the structure of the underlying SIPP Gold Standard so, except where otherwise indicated, the discussion in this section about sample selection and analysis variables applies to all three data sources equally.

For comparability with the literature, we restrict our attention in this analysis to males. The sex of each sample person in the SGS is taken from the SIPP. Since sample persons in SIPP are interviewed every four months (in *waves*), there are multiple reports for each person on many (typically)

[1] Oreopoulos, von Wachter, and Heisz (2008) instrument for graduation-year unemployment rate with the unemployment rate in the year of predicted graduation based on college start date and field of study.

[2] This year range includes seven SIPP panels: 1990,1991,1992,1993,1996,2001, and 2004.

time-invariant characteristics such as sex. For details on how a single report is selected for each sample person, see the SIPP Gold Standard File Codebook[3]. The sex of each sample person is left unsynthesized in the SSB file.

We also include in our sample only college graduates for whom the (SIPP-reported) highest level of completed education is a four-year college degree. The SGS includes two measures of educational attainment, one from the *core* or wavely SIPP questions and another from an education history topical module administered in only one wave of each panel. We take as our measure of educational attainment the SGS measure based on the topical module because the topical module is the only source for the dates of the start of post high school education and of bachelor's degree completion.

In order to be reasonably confident that we have observed the graduate school matriculation decision of most sample persons, we restrict attention to sample persons who were at least 30 years of age at the beginning of the SIPP panel in which they were sampled. The age of sample persons in SGS is taken from SSA administrative records.

In our analysis sample, we also drop sample persons whose age at the beginning of post high school education is less than seventeen or greater than twenty. We also exclude any sample person whose reported bachelor's degree completion date is more than eight years after their reported start of post high school education.

Each person-earnings year observation in our sample is coded as representing a particular (potential) experience year for the sample person. Experience years is simply the number of years since the year of college graduation. In Table 1, we present the size of our SGS sample by cohort and experience years.

In most of our specifications, we include additional demographic variables: indicators for three categories of race (White, Black, and Other), an indicator for Hispanic ethnicity, and an indicator for whether the sample person was born outside of the United States. These SGS variables are all taken from the SIPP.

[3]Codebooks for the SIPP Gold Standard and SIPP Synthetic Beta files are available from the authors.

The SGS contains earnings records from the SSA's Detailed Earnings Record in four categories: non-deferred FICA earnings, deferred FICA earnings, non-deferred non-FICA earnings, and non-deferred FICA earnings. For our analysis, we aggregate all four categories of earnings to create an annual total earnings measure for each sample person. We convert annual total earnings to 2006 dollars using the Consumer Price Index(All Urban Consumers, 1982-84=100). We recode zero earnings to \$1 in order to work with the log of real earnings without losing the zero earners from our sample.

As our measure of the unemployment rate in each year we use the national unemployment rate in June of each year as calculated from the Current Population Survey by the Bureau of Labor Statistics.

3 Empirical Specifications

We estimate empirical specifications of the form

$$log(real\_earnings_{it}) = \alpha + \sum_{v=0}^{V}(\beta_v * grad\_yr\_urate_i * exp_v + \gamma_v * exp_v) + \mathbf{x_{it}}' * \psi + \epsilon_{\mathbf{it}} \qquad (1)$$

where i indexes sample persons and v indexes years of potential experience. The maximum number of included experience years is V. The variable exp_v is an indicator for whether the annual earnings observation corresponds to experience year s for the sample person. The column vectors $\mathbf{x_{it}}$ and ψ contain conditioning variables specific to each specification and corresponding coefficients, respectively. The vector of β_v's are the parameters of interest which we interpret as the variation in earnings-year-specific log real earnings that is explained by the graduation-year unemployment rate.

We employ six variations of the specification in Equation 1 in the ordinary least squares estimates that we present using the SGS. These specifications are also extended to the instrumental variables for estimation by two-stage least squares. We report robust standard errors clustered at the person level.

We repeat our OLS estimations on the completed SGS and SSB. Since these data are created using multiple impuation and partial data synthesis techniques, inference based on these estimations is based on appropriate aggregation of results across completed or synthetic implicates according to the approach presented in Reiter (2004).

4 Results

Table 2 presents results for several linear specifications estimated using least squares using our SGS sample. The estimates labeled "Prelim" (in Column 1) are from a regression that includes no additional conditioning variables beyond those specified in Equation 1. The estimated coefficient on graduation-year unemployment rate for experience year zero is negative and statistically significant. The coefficient for experience year one is less negative and less statistically significant. For no other experience year is the coefficient on the interaction statistically significant.

In a second regression presented as "Base" (in Column 2 of Table 2), we include additional conditioning variables intended to explain variation in log real earnings and therefore increase the precision of our estimates of the parameters of interest. These conditioning variables are indicators for three categories of reported race, an indicator for whether the sample person was reported to be born outside of the United States, and an indicator for whether the reported origin of the sample person was Hispanic. The estimated coefficients for graduation-year unemployment rate in years zero and one are slightly smaller and more statistically significant. The coefficients on the interaction for years beyond year one remain statistically insignificant.

In columns 3 and 4 of Table 2 we augment the "Base" specification with a linear or quadratic time trend specific to pairs of adjacent (in time) graduation cohorts.[4] Conditioning on these time trends allows us to isolate the power of the graduation-year unemployment rate to explain variation experience-year log real earnings from other cohort-specific earning trends. In these results,

[4]Time trends are specified for adjacent cohorts using cohort-specific time trends appeared to introduce multicollinearity to our estimation.

the graduation-year unemployment rate has explanatory power for log real earnings only in the graduation year.

In the last two columns of Table 2, we present results of regressions which include earnings year fixed effects (column 5) or the earnings-year unemployment rate. Once we condition on either of these (sets of) variables, none of the variation in log real earnings is explained by the graduation-year unemployment rate.

In Table 3 we present results from estimating the same array of empirical specifications using the unemployment rate in the sample person's twenty-second year as an instrument for the graduation-year unemployment rate. We estimate these specifications by Two-staged Least Squares. This approach addresses the possibility that college students respond to the labor market conditions in choosing when to graduate.

Our two-stage least squares results are very similar to our OLS results. With the exception of the "Quadratic" specification, the pattern of signs, magnitudes, and statistical significance are very similar across the two approaches. In the "Quadratic" specification, the pattern of statistical significance of is the same as in the OLS results, but we note that the (statistically insignificant) coefficients decrease dramatically with progressive potential experience. This particular estimation may suffer from multi-collinearity.

We also repeat our OLS estimations using a sub-sample of our data that more closely matches the cohorts and earnings years available to Oreopoulos et al in their data on Canadian college graduates. The sample used is described in Table 4 and the results are presented in Table 5. These results are very similar those from full sample estimation only with slightly lower statistical significance.

In Table 6 we compare results for two of our specifications using the SGS, completed SGS, and SSB data. The imputation of missing data reflected in the completed SGS do not change our substantive findings. Coefficients from SSB data also have similar magnitude and sign as from the completed data but the statistical significance is muted.

5 Conclusion

For a large sample of SIPP sample persons, we find no evidence that the unemployment rate faced upon college graduation explains variation in log real earnings within experience years. These results suggest that poor labor market conditions upon initial entrance to the labor force may not have persistent impact on earnings. We continue to explore several possible explanations for the differences between our results and those of Oreopoulos, von Wachter, and Heisz (2008).

We also confirm that similar estimated coefficients can be obtained from the SIPP Synthetic Beta. We are encouraged about the usefulness of these data for allowing wider access to some of the benefits of these linked survey and administrative data.

References

Kahn, L. B. (2010): "The long-term labor market consequences of graduating from college in a bad economy," *Labour Economics*, 17(2), 303–316.

Oreopoulos, P., T. von Wachter, and A. Heisz (2008): "The Short-and Long-Term Career Effects of Graduating in a Recession: Hysteresis and Heterogeneity in the Market for College Graduates," Discussion Paper 3578, IZA.

Reiter, J. P. (2004): "Simultaneous use of multiple imputation for missing data and disclosure limitation," *Survey Methodology*, 30(235), l242.

Rubin, D. B. (1996): "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association*, 91(434), 473–489.

Woodcock, S. D., and G. Benedetto (2006): "Distribution-preserving statistical disclosure limitation," LEHD technical paper tp-2006-04, U.S. Census Bureau, http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-04.pdf (October 31, 2006).

Table 1: Sample Size by Graduation Cohort (1978 to 1999-2001) and Years of Potential Experience (0 to 15)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
1978	393	393	393	393	393	393	393	393	393	393	393	393	393	393	393	393	6288
1979	371	371	371	371	371	371	371	371	371	371	371	371	371	371	371	371	5936
1980	392	392	392	392	392	392	392	392	392	392	392	392	392	392	392	392	6272
1981	379	379	379	379	379	379	379	379	379	379	379	379	379	379	379	379	6064
1982	398	398	398	398	398	398	398	398	398	398	398	398	398	398	398	398	6368
1983	355	355	355	355	355	355	355	355	355	355	355	355	355	355	355	355	5680
1984	323	323	323	323	323	323	323	323	323	323	323	323	323	323	323	323	5168
1985	347	347	347	347	347	347	347	347	347	347	347	347	347	347	347	347	5552
1986	240	240	240	240	240	240	240	240	240	240	240	240	240	240	240	240	3840
1987	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	257	4112
1988	222	222	222	222	222	222	222	222	222	222	222	222	222	222	222	222	3552
1989	182	182	182	182	182	182	182	182	182	182	182	182	182	182	182	182	2912
1990	177	177	177	177	177	177	177	177	177	177	177	177	177	177	177	177	2832
1991	143	143	143	143	143	143	143	143	143	143	143	143	143	143	143	143	2288
1992	158	158	158	158	158	158	158	158	158	158	158	158	158	158	158	0	2370
1993	139	139	139	139	139	139	139	139	139	139	139	139	139	139	0	0	1946
1994	135	135	135	135	135	135	135	135	135	135	135	135	135	0	0	0	1755
1995	86	86	86	86	86	86	86	86	86	86	86	86	0	0	0	0	1032
1996	85	85	85	85	85	85	85	85	85	85	85	0	0	0	0	0	935
1997	57	57	57	57	57	57	57	57	57	57	0	0	0	0	0	0	570
1998	30	30	30	30	30	30	30	30	30	0	0	0	0	0	0	0	270
1999-2001	22	22	22	22	22	22	21	13	0	0	0	0	0	0	0	0	166
Total	4891	4891	4891	4891	4891	4891	4890	4882	4869	4839	4782	4697	4611	4476	4337	4179	75908

Each cell contains the number of person-experience year observations for the college graduation-year cohort and number of years of potential post-college experience. The cell sample sizes for the 1999 to 2001 cohorts are collapsed together only for the purpose of this table, not in the analysis.

Table 2: Ordinary Least Squares Estimation Using Full Sample

	Prelim	Base	Linear	Quadratic	Year FE	UE
(exp==0)*ugrad	-0.0942**	-0.112***	-0.112***	-0.112***	-0.00846	0.00150
	(0.0295)	(0.0272)	(0.0272)	(0.0272)	(0.0315)	(0.0305)
(exp==1)*ugrad	-0.0565*	-0.0705**	-0.0643*	-0.0630	0.0193	0.00626
	(0.0274)	(0.0244)	(0.0252)	(0.0322)	(0.0281)	(0.0264)
(exp==2)*ugrad	-0.0276	-0.0424	-0.0301	-0.0286	0.0171	-0.00722
	(0.0263)	(0.0236)	(0.0274)	(0.0461)	(0.0264)	(0.0240)
(exp==3)*ugrad	-0.0230	-0.0376	-0.0190	-0.0186	-0.0194	-0.0256
	(0.0258)	(0.0229)	(0.0314)	(0.0596)	(0.0253)	(0.0229)
(exp==4)*ugrad	0.000149	-0.0147	0.0100	0.00826	-0.0171	-0.0384
	(0.0249)	(0.0222)	(0.0369)	(0.0713)	(0.0245)	(0.0224)
(exp==5)*ugrad	0.0242	0.00922	0.0401	0.0350	-0.00340	-0.0173
	(0.0249)	(0.0228)	(0.0431)	(0.0799)	(0.0247)	(0.0230)
(exp==6)*ugrad	0.0292	0.0140	0.0511	0.0413	0.0105	-0.00407
	(0.0236)	(0.0217)	(0.0491)	(0.0856)	(0.0238)	(0.0218)
(exp==7)*ugrad	0.0151	-0.00246	0.0419	0.0251	0.0128	-0.0209
	(0.0235)	(0.0224)	(0.0565)	(0.0895)	(0.0242)	(0.0225)
(exp==8)*ugrad	0.0101	-0.00649	0.0457	0.0231	0.0300	0.00115
	(0.0231)	(0.0221)	(0.0638)	(0.0911)	(0.0238)	(0.0221)
(exp==9)*ugrad	-0.0322	-0.0468*	0.0156	-0.00739	0.00977	0.00396
	(0.0222)	(0.0212)	(0.0700)	(0.0902)	(0.0231)	(0.0222)
(exp==10)*ugrad	-0.0262	-0.0407	0.0225	0.00279	0.0280	0.0258
	(0.0227)	(0.0221)	(0.0771)	(0.0889)	(0.0238)	(0.0236)
(exp==11)*ugrad	-0.0154	-0.0265	0.0334	0.0311	0.0403	0.0289
	(0.0232)	(0.0231)	(0.0847)	(0.0881)	(0.0249)	(0.0243)
(exp==12)*ugrad	0.00613	-0.00583	0.0573	0.0646	0.0451	0.0164
	(0.0231)	(0.0233)	(0.0925)	(0.0897)	(0.0253)	(0.0236)
(exp==13)*ugrad	0.0262	0.0151	0.0801	0.102	0.0380	-0.00282
	(0.0237)	(0.0242)	(0.0993)	(0.0942)	(0.0262)	(0.0241)
(exp==14)*ugrad	0.00487	-0.00748	0.0620	0.0963	-0.0110	-0.0432
	(0.0250)	(0.0255)	(0.108)	(0.106)	(0.0280)	(0.0257)
(exp==15)*ugrad	0.00435	-0.00764	0.0697	0.116	-0.0289	-0.0444
	(0.0254)	(0.0260)	(0.116)	(0.123)	(0.0289)	(0.0263)
N	75908	74972	74972	74972	74972	74972

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The columns present selected coefficients from regressions of log real annual earnings on the college-graduation-year unemployment rate interacted with indicators for experience years 0 to 15, experience year fixed effects, and (except for "Prelim") other conditioning variables: indicators for race, birth outside of US, and Hispanic ethnicity. Reported standard errors are clustered at the person level. The sample consists of males in a 1990 to 2004 SIPP panel who were at least 30 at the beginning of their SIPP panel, began college between ages 17 and 20 and completed a bachelor's degree in no more than eight years, and reported no post-graduate education. Where indicated by column title, a (linear or quadratic) time trend trend specific to pairs of adjacent-year cohorts was included. Year fixed effects were included in the estimation reported under the column title "Year FE." Earnings year unemployment rate was included in the estimation reported under the column title "UE." See Section 2 of the text for details on the sample and variables and Section 3 for more details on the empirical specification.

Table 3: Two-stage Least Squares Estimation Using Full Sample

	Prelim	Base	Linear	Quadratic	Year FE	UE
(exp==0)*ugrad	-0.0916	-0.0999*	-0.0994*	-0.0993*	0.0242	0.00956
	(0.0545)	(0.0499)	(0.0499)	(0.0498)	(0.0694)	(0.0579)
(exp==1)*ugrad	-0.0794	-0.0825	-0.215	-0.454	-0.00379	-0.0225
	(0.0514)	(0.0460)	(0.111)	(0.289)	(0.0631)	(0.0500)
(exp==2)*ugrad	-0.0270	-0.0357	-0.302	-0.736	-0.00428	-0.0247
	(0.0503)	(0.0460)	(0.200)	(0.519)	(0.0581)	(0.0464)
(exp==3)*ugrad	-0.00133	-0.00383	-0.404	-0.989	-0.00936	-0.0237
	(0.0488)	(0.0440)	(0.290)	(0.711)	(0.0526)	(0.0431)
(exp==4)*ugrad	0.00724	0.00439	-0.529	-1.223	-0.0124	-0.0325
	(0.0460)	(0.0411)	(0.381)	(0.867)	(0.0471)	(0.0401)
(exp==5)*ugrad	-0.0206	-0.0226	-0.690	-1.448	-0.0284	-0.0512
	(0.0469)	(0.0431)	(0.475)	(0.988)	(0.0486)	(0.0425)
(exp==6)*ugrad	-0.0391	-0.0334	-0.834	-1.614	-0.0103	-0.0429
	(0.0455)	(0.0417)	(0.565)	(1.071)	(0.0466)	(0.0415)
(exp==7)*ugrad	-0.0799	-0.0792	-1.012	-1.771	-0.0229	-0.0624
	(0.0430)	(0.0410)	(0.657)	(1.124)	(0.0459)	(0.0413)
(exp==8)*ugrad	-0.0805	-0.0775	-1.142	-1.835	0.00465	-0.0235
	(0.0414)	(0.0398)	(0.748)	(1.146)	(0.0452)	(0.0417)
(exp==9)*ugrad	-0.109**	-0.102**	-1.293	-1.865	-0.00763	-0.0216
	(0.0396)	(0.0376)	(0.838)	(1.144)	(0.0447)	(0.0420)
(exp==10)*ugrad	-0.0944*	-0.0890*	-1.422	-1.827	-0.00436	-0.0165
	(0.0427)	(0.0420)	(0.933)	(1.128)	(0.0510)	(0.0461)
(exp==11)*ugrad	-0.0833	-0.0744	-1.547	-1.722	-0.0203	-0.0379
	(0.0426)	(0.0423)	(1.023)	(1.099)	(0.0530)	(0.0437)
(exp==12)*ugrad	-0.0687	-0.0654	-1.680	-1.591	-0.0539	-0.0745
	(0.0424)	(0.0432)	(1.118)	(1.083)	(0.0549)	(0.0430)
(exp==13)*ugrad	-0.0123	0.00139	-1.745	-1.343	-0.0275	-0.0442
	(0.0446)	(0.0449)	(1.206)	(1.090)	(0.0559)	(0.0446)
(exp==14)*ugrad	-0.0773	-0.0656	-1.956	-1.193	-0.122*	-0.122**
	(0.0458)	(0.0464)	(1.303)	(1.154)	(0.0545)	(0.0467)
(exp==15)*ugrad	-0.0400	-0.0295	-2.046	-0.884	-0.0970	-0.0794
	(0.0453)	(0.0462)	(1.382)	(1.271)	(0.0498)	(0.0464)
N	75908	74972	74972	74972	74972	74972

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The columns present selected coefficients from two-stage least squares regressions of log real annual earnings on the college-graduation-year unemployment rate interacted with indicators for experience years 0 to 15, experience year fixed effects, and (except for "Prelim") other conditioning variables: indicators for race, birth outside of US, and Hispanic ethnicity. Unemployment rate in the sample person's twenty-second year of age is used as an instrument for the college-graduation-year unemployment rate. Reported standard errors are clustered at the person level. The sample consists of males in a 1990 to 2004 SIPP panel who were at least 30 at the beginning of their SIPP panel, began college between ages 17 and 20 and completed a bachelor's degree in no more than eight years, and reported no post-graduate education. Where indicated by column title, a (linear or quadratic) time trend trend specific to pairs of adjacent-year cohorts was included. Year fixed effects were included in the estimation reported under the column title "Year FE." Earnings year unemployment rate was included in the estimation reported under the column title "UE." See Section 2 of the text for details on the sample and variables and Section 3 for more details on the empirical specification.

Table 4: Sample Size by Graduation Cohort (1978 to 1995) and Years of Potential Experience (1 to 10)

	0	1	2	3	4	5	6	7	8	9	10	Total
1978	0	0	0	0	393	393	393	393	393	393	393	2751
1979	0	0	0	371	371	371	371	371	371	371	371	2968
1980	0	0	392	392	392	392	392	392	392	392	392	3528
1981	0	379	379	379	379	379	379	379	379	379	379	3790
1982	398	398	398	398	398	398	398	398	398	398	398	4378
1983	355	355	355	355	355	355	355	355	355	355	355	3905
1984	323	323	323	323	323	323	323	323	323	323	323	3553
1985	347	347	347	347	347	347	347	347	347	347	347	3817
1986	240	240	240	240	240	240	240	240	240	240	240	2640
1987	257	257	257	257	257	257	257	257	257	257	257	2827
1988	222	222	222	222	222	222	222	222	222	222	222	2442
1989	182	182	182	182	182	182	182	182	182	182	182	2002
1990	177	177	177	177	177	177	177	177	177	177	0	1770
1991	143	143	143	143	143	143	143	143	143	0	0	1287
1992	158	158	158	158	158	158	158	158	0	0	0	1264
1993	139	139	139	139	139	139	139	0	0	0	0	973
1994	135	135	135	135	135	135	0	0	0	0	0	810
1995	86	86	86	86	86	0	0	0	0	0	0	430
Total	3162	3541	3933	4304	4697	4611	4476	4337	4179	4036	3859	45135

Each cell contains the number of person-experience year observations for the college graduation-year cohort and number of years of potential post-college experience. The included cohorts and experience years are selected to more closely match the cohorts and experience years in Oreopoulos, von Wachter, and Heisz (2008).

Table 5: OLS Estimation Using Cohorts and Earnings Years from Oreopoulos et al (2008)

	Prelim	Base	Linear	Quadratic	Year FE	UE
(exp==0)*ugrad	-0.0429	-0.0693*	-0.0694*	-0.0694*	0.0207	0.0211
	(0.0345)	(0.0312)	(0.0312)	(0.0312)	(0.0338)	(0.0330)
(exp==1)*ugrad	-0.0465	-0.0687*	-0.0599*	-0.0727	0.0257	-0.00626
	(0.0315)	(0.0272)	(0.0301)	(0.0432)	(0.0300)	(0.0278)
(exp==2)*ugrad	-0.0223	-0.0457	-0.0276	-0.0471	0.0246	-0.0161
	(0.0299)	(0.0261)	(0.0371)	(0.0652)	(0.0284)	(0.0257)
(exp==3)*ugrad	-0.00702	-0.0196	-0.00215	-0.0293	-0.00763	-0.0160
	(0.0284)	(0.0243)	(0.0471)	(0.0828)	(0.0270)	(0.0242)
(exp==4)*ugrad	0.0120	0.000616	0.0159	-0.0140	-0.0193	-0.0280
	(0.0264)	(0.0226)	(0.0591)	(0.0950)	(0.0250)	(0.0231)
(exp==5)*ugrad	0.0416	0.0292	0.0455	0.0205	-0.0112	-0.00215
	(0.0268)	(0.0238)	(0.0720)	(0.101)	(0.0256)	(0.0242)
(exp==6)*ugrad	0.0444	0.0326	0.0480	0.0320	-0.00711	0.00967
	(0.0254)	(0.0226)	(0.0847)	(0.102)	(0.0245)	(0.0227)
(exp==7)*ugrad	0.0118	-0.000647	0.0167	0.0151	-0.0183	-0.0215
	(0.0246)	(0.0232)	(0.0985)	(0.102)	(0.0251)	(0.0234)
(exp==8)*ugrad	0.00243	-0.00913	0.0139	0.0321	0.00547	-0.00260
	(0.0240)	(0.0231)	(0.113)	(0.105)	(0.0253)	(0.0231)
(exp==9)*ugrad	-0.0373	-0.0508*	-0.0262	0.0161	-0.00935	-0.00724
	(0.0233)	(0.0224)	(0.125)	(0.115)	(0.0253)	(0.0235)
(exp==10)*ugrad	-0.0167	-0.0312	-0.0176	0.0509	0.0186	0.0208
	(0.0247)	(0.0241)	(0.141)	(0.141)	(0.0274)	(0.0257)
N	45135	44559	44559	44559	44559	44559

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The columns present selected coefficients from regressions of log real annual earnings on the college-graduation-year unemployment rate interacted with indicators for experience years 0 to 15, experience year fixed effects, and (except for "Prelim") other conditioning variables: indicators for race, birth outside of US, and Hispanic ethnicity. Reported standard errors are clustered at the person level. The sample consists of males in a 1990 to 2004 SIPP panel who were at least 30 at the beginning of their SIPP panel, began college between ages 17 and 20 and completed a bachelor's degree in no more than eight years, and reported no post-graduate education. Where indicated by column title, a (linear or quadratic) time trend trend specific to pairs of adjacent-year cohorts was included. Year fixed effects were included in the estimation reported under the column title "Year FE." Earnings year unemployment rate was included in the estimation reported under the column title "UE." See Section 2 of the text for details on the sample and variables and Section 3 for more details on the empirical specification.

Table 6: Comparison of results from SGS, Completed SGS, and SSB

	Base (GS)	Base (CMP)	Base (SSB)	Quad (GS)	Quad (CMP)	Quad (SSB)
(exp==0)*ugrad	-0.112***	-0.124***	-0.126*	-0.112***	-0.124***	-0.137***
	(0.0272)	(0.0332)	(0.0587)	(0.0272)	(0.0332)	(0.0359)
(exp==1)*ugrad	-0.0705**	-0.0917**	-0.105	-0.0630	-0.0837*	-0.110*
	(0.0244)	(0.0284)	(0.0746)	(0.0322)	(0.0420)	(0.0471)
(exp==2)*ugrad	-0.0424	-0.0662*	-0.0809	-0.0286	-0.0519	-0.102
	(0.0236)	(0.0335)	(0.0745)	(0.0461)	(0.0693)	(0.0682)
(exp==3)*ugrad	-0.0376	-0.0451	-0.0583	-0.0186	-0.0265	-0.0875
	(0.0229)	(0.0306)	(0.0786)	(0.0596)	(0.0669)	(0.0888)
(exp==4)*ugrad	-0.0147	-0.0159	-0.0392	0.00826	0.00530	-0.0556
	(0.0222)	(0.0235)	(0.0565)	(0.0713)	(0.0931)	(0.106)
(exp==5)*ugrad	0.00922	0.00135	-0.0302	0.0350	0.0233	-0.0702
	(0.0228)	(0.0220)	(0.0531)	(0.0799)	(0.100)	(0.120)
(exp==6)*ugrad	0.0140	0.0153	-0.0266	0.0413	0.0361	-0.0632
	(0.0217)	(0.0285)	(0.0479)	(0.0856)	(0.108)	(0.130)
(exp==7)*ugrad	-0.00246	0.00533	-0.0192	0.0251	0.0224	-0.0503
	(0.0224)	(0.0311)	(0.0409)	(0.0895)	(0.111)	(0.136)
(exp==8)*ugrad	-0.00649	0.00856	-0.0287	0.0231	0.0233	-0.0742
	(0.0221)	(0.0291)	(0.0327)	(0.0911)	(0.111)	(0.138)
(exp==9)*ugrad	-0.0468*	-0.0233	-0.0396	-0.00739	-0.00500	-0.117
	(0.0212)	(0.0223)	(0.0349)	(0.0902)	(0.108)	(0.137)
_IexpXugra_10	-0.0407	-0.0171	-0.0463	0.00279	-0.000606	-0.142
	(0.0221)	(0.0245)	(0.0389)	(0.0889)	(0.106)	(0.132)
_IexpXugra_11	-0.0265	-0.0142	-0.0416	0.0311	0.00818	-0.0993
	(0.0231)	(0.0284)	(0.0336)	(0.0881)	(0.106)	(0.125)
_IexpXugra_12	-0.00583	-0.00107	-0.0261	0.0646	0.0267	-0.0660
	(0.0233)	(0.0238)	(0.0278)	(0.0897)	(0.0992)	(0.117)
_IexpXugra_13	0.0151	0.0107	-0.0174	0.102	0.0472	-0.0259
	(0.0242)	(0.0255)	(0.0241)	(0.0942)	(0.0962)	(0.109)
_IexpXugra_14	-0.00748	-0.00634	-0.00691	0.0963	0.0380	0.00849
	(0.0255)	(0.0247)	(0.0290)	(0.106)	(0.105)	(0.104)
_IexpXugra_15	-0.00764	-0.000938	-0.00000704	0.116	0.0507	0.0334
	(0.0260)	(0.0290)	(0.0368)	(0.123)	(0.121)	(0.106)
2.race	-0.510***	-0.566***	-0.635*	-0.488***	-0.556***	-0.774***
	(0.124)	(0.120)	(0.260)	(0.124)	(0.122)	(0.222)
3.race	-0.443*	-0.851***	-0.288*	-0.462*	-0.863***	-0.334
	(0.191)	(0.179)	(0.128)	(0.190)	(0.177)	(0.184)
hispanic	-0.124	-0.253	-0.354*	-0.136	-0.267	-0.306
	(0.166)	(0.198)	(0.175)	(0.165)	(0.201)	(0.195)
foreign_born	-2.906***	-1.809***	-0.0380	-2.914***	-1.817***	-0.000280
	(0.191)	(0.136)	(0.119)	(0.190)	(0.134)	(0.130)
N	74972	104442	84577	74972	104442	84577

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The columns present selected coefficients from two regressions repeated on three datasets, one based on each of three sources: SGS, completed SGS, and SSB. The "Base" estimations regress log real annual earnings on the college-graduation-year unemployment rate interacted with indicators for experience years 0 to 15, experience year fixed effects and other conditioning variables: indicators for race, birth outside of US, and Hispanic ethnicity. The "Quad" specifications augment the "Base" specifications with a quadratic time trend trend specific to pairs of adjacent-year cohorts. Reported results from the completed SGS and SSB are based on aggregation of results from multiple implicates according to the procedures in Reiter (2004). See Section 2 of the text for details on the sample and variables and Section 3 for more details on the empirical specification.

A Description of SIPP Gold Standard and SIPP Synthetic Beta

The SIPP Gold Standard (SGS) was created by standardizing a basic set of variables across seven panels of SIPP and then merging administrative records from the Internal Revenue Service (IRS) and the Social Security Administration (SSA) to these SIPP records. Missing data in the SGS is multiply-imputed to create the *completed* SGS. The SGS and completed SGS contain confidential data are available only from within the Census Bureau secure network to Census staff and Special Sworn Status researchers on approved projects. In order to make some of the benefits of these linked data more widely available, a synthetic version of the completed SGS, called SIPP Synthetic Beta (SSB) is produced and released publicly. The synthesis process is designed to preserve relationships amongst variables in the completed SGS data while protecting the identity of individual respondents. Due to the experimental nature of the SSB, the Census Bureau is actively engaged in evaluating the validity of the synthetic data by assessing to what extent similar results are obtained from the SSB and the underlying confidential data.

A.1 Gold Standard Creation

We chose a basic set of variables that described demographic characteristics of each respondent as well as marital, fertility, education, and immigration histories and labor supply, income, wealth, and limited program participation information for the time period covered by the SIPP panel. These variables were cleaned and standardized across panels and then individuals were stacked to form a file called the Gold Standard.

The Census Bureau has conducted the Survey of Income and Program Participation (SIPP) for 25

years and during this time, the survey instrument has undergone significant changes in format and content. The first four SIPP panels of the 1990s were overlapping and lasted 32-36 months. Data was collected in waves, with the reference period for a wave being four months. In 1996, a major redesign was undertaken, the household sample size was approximately doubled, and the survey was extended to cover 48 months. The 2001 and 2004 panels covered 36 and 48 months respectively and follow a format very similar to the 1996 panel, with some additional content added.

Administrative records are matched to SIPP respondents in the Gold Standard using a validated Social Security Number. The 1990s SIPP panels collected the SSN from respondents. Using name, birth date, gender, and race information, SSA validated these self-reports against the SSA Numident file. If the demographic variables in the SIPP and Numident for a given SSN were identical, then the SSN was declared valid.[5] If the demographic variables did not match, an alternative SSN was sought. For individuals who reported that they did not know their SSN, an SSN was sought in the Numident file based on these demographic variables. For individuals who refused to provide an SSN, no match was sought in the PCF and we did not receive administrative records for these individuals.

Beginning with the 2001 panel and continuing through the present, Census has undertaken the role of validating SSNs using its Person Identification Validation System (PVS). The system performs a probabilistic comparison between fields from the Census Person Characteristics File (PCF) and the corresponding data elements collected by the survey with the major enhancement that it additionally incorporates physical address. Survey SSNs are valid if they match a record in the PCF with the same SSN with a probability above a certain threshold value. Again, for individuals who refuse to provide SSN, no SSN assignment is made and therefore no administrative data are received.

[5]Prior to 2003, the process of validating an SSN was performed by a clerical edit using the same information.

Beginning with the 2004 panel, respondents were no longer asked SSN but were instead asked to opt-out if they did not want their survey data to be linked. The PVS system is then used to find SSNs for these individuals. Only those who opted-out or who could not be found in the PCF with a reasonable degree of certainty have missing SSNs and no administrative data. Overall the gold standard contains 27% of people with missing administrative data due to no valid SSN. This number is substantially higher than previous versions of the gold standard because of the inclusion of the 2001 and 2004 SIPP panels. Individuals refused to provide SSNs at much higher rates in these panels than in the 1990s.

We matched two main types of administrative records: earnings from W-2 forms and OASDI and SSI benefit information from SSA master data files. The administrative data are different from the survey data in several ways. First, the administrative data are either entirely missing or completely present. There are no item-missing variables. Second, the time period covered by these data is the same for all SIPP respondents and includes many years outside the range of the survey. For example the Gold Standard contains capped earnings from 1951-2006. Third, the administrative data are not self-reports but rather information maintained by government agencies which permanently record employer reports of earnings or agency reports of benefits awarded and paid. In combining survey data with government records, one must think carefully about the time periods covered by each. For all individuals except those in the 2004 SIPP panel[6], the earnings and benefits histories cover the periods before, during, and after the survey. Because the SIPP collects a significant amount of retrospective history through the use of topical modules in the first and second waves, it is possible to know something about an individual's life before the survey period and to correlate this with the lengthy past earnings history. However, although earnings continue after the survey ends, we lose

[6]The last year of administrative earnings is 2006 and the last year of benefits is 2007. Hence for 2004 panel respondents, survey and administrative data end at similar times.

all information on marital status, birth of children, and other important demographic changes for this time period. Hence for many analyses, including the one presented in this paper, years after the survey cannot be included.

A.2 Creation of Completed and Synthetic Data

After the Gold Standard is created, we then impute all missing data, using Bayesian multiple imputation techniques. This process creates a set of files (implicates) that we refer to as the completed data. The completed data is identical to the Gold Standard except for cases where there were missing values. The completed data then serve as the input files for the data synthesis. Each completed implicate spawns multiple synthetic implicates. The synthetic files contain values that are draws from the joint posterior predictive distribution of the underlying variables conditional on the existing confidential data.

Since the late 1970's, the theory and techniques for multiple imputation in order to fill missing data have been developed and refined (Rubin 1996). These methods offer an analytically useful set of completed data that allows the analyst to measure the noise introduced through imputation and properly take that into account in estimating statistics and their measures of uncertainty. Adapting Rubin's notation to our missing data problem, the data can be expressed as Y where Y is a matrix of variables at least some of which contain some missing values. Y can be expressed as (Y_{obs}, Y_{miss}) where Y_{obs} represents the observed values of Y and Y_{miss} represents the missing values of Y. The inclusion indicator, I, is a structure equivalent in size to Y with elements equal to 1 where Y is non-missing and 0 otherwise. The database can then be expressed by the joint distribution, $p(Y, I, \theta)$, where θ are unknown parameters. In this case, the missing data mechanism is said to be missing at

random if

$$p(I|Y) = p(I|Y_{obs}). \tag{2}$$

Draws are taken from the posterior predictive distribution

$$p(\widetilde{Y}|Y_{obs}) = p(\widetilde{Y}|\theta)p(\theta|Y_{obs})d\theta \tag{3}$$

to produce M multiply-imputed completed data files Y^m where $Y^m = (Y_{obs}, \widetilde{Y}^m)$ for $m = 1, ..., M$. The resulting M data files are individually referred to as completed implicates.

In practice, it is very difficult to estimate the joint likelihood $p(\mathbf{Y}|\theta)$, especially in a case such as ours where there are many confidential variables, both continuous and discrete, that relate to each other in very complex ways. As a result, we chose to estimate a sequence of univariate conditional models. Letting $\mathbf{Y} = [\mathbf{y}_0 \ \ \mathbf{y}_1 \ \ \mathbf{y}_2 \ \ \cdots \ \ \mathbf{y}_K]$ (where $\mathbf{y}_0$ is a subset of columns of Y containing no missing data and each $\mathbf{y}_k$ for $k > 0$ has non-missing elements $\mathbf{y}_{k,obs}$ and missing elements $\mathbf{y}_{k,miss}$), and $\theta = [\theta_1 \ \ \theta_2 \ \ \cdots \ \ \theta_K]$, the joint likelihood can be factorized as:

$$p(\mathbf{Y}|\theta) = p_1(\mathbf{y}_1|\mathbf{y}_0, \theta_1)\, p_2(\mathbf{y}_2|\mathbf{y}_0, \mathbf{y}_1, \theta_2) \cdots p_K(\mathbf{y}_K|\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{K-1}, \theta_K) \tag{4}$$

Estimating a univariate conditional model for each $\mathbf{y}_k$ permits sequential generation of completed values $\mathbf{Y}^c = [\mathbf{y}_0 \ \ \mathbf{y}_1^c \ \ \mathbf{y}_2^c \ \ \cdots \ \ \mathbf{y}_K^c]$ (where $\mathbf{y}_k^c$ is the same as $\mathbf{y}_k$ except $\mathbf{y}_{k,miss}$ has been replaced by draws $\tilde{\mathbf{y}}_k$). That is, sample $\tilde{\mathbf{y}}_1$ from the posterior predictive distribution of $\mathbf{y}_1$ given $\mathbf{y}_0$**,** then $\tilde{\mathbf{y}}_2$ from the posterior predictive distribution of $\mathbf{y}_2$ given $\mathbf{X}$ and $\tilde{\mathbf{y}}_1$, etc. Doing this independently M times results in completed implicates, $\mathbf{Y}^{c1}, ..., \mathbf{Y}^{cM}$.

Once the data has been completed, we use similar techniques to generate R synthetic implicates from

each of the M completed implicates, resulting in $M \times R$ synthetic data files. Denoting $\mathbf{Y}^{cm} = [X \;\; Z] = [X \;\; z_1 \;\; z_2 \;\; ... \;\; z_{K_s}]$ (where X is the subset of columns of $\mathbf{Y}^{cm}$ that are not confidential and will not be synthesized), the joint likelihood can now be factorized as:

$$p\left(\mathbf{Y}^{cm}|X,\theta\right) = p_1\left(\mathbf{z}_1|X,\theta_1\right) p_2\left(\mathbf{z}_2|X,\mathbf{z}_1,\theta_2\right) \cdots p_{K_s}\left(\mathbf{z}_{K_s}|X,\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_{K_s-1},\theta_{K_s}\right) \tag{5}$$

Estimating a univariate conditional model for each $\mathbf{z}_k$ permits sequential generation of synthetic values $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1 \;\; \tilde{\mathbf{z}}_2 \;\; \cdots \;\; \tilde{\mathbf{z}}_K]$, resulting in $\tilde{\mathbf{Y}}^m = [X \;\; \tilde{\mathbf{Z}}]$. Once again, we sample $\tilde{\mathbf{z}}_1$ from the posterior predictive distribution of $\mathbf{z}_1$ given X**,** then $\tilde{\mathbf{z}}_2$ from the posterior predictive distribution of $\mathbf{z}_2$ given $\mathbf{X}$ and $\tilde{\mathbf{z}}_1$, etc. Doing this independently R times for each completed implicate results in synthetic implicates, $\tilde{\mathbf{Y}}^{m,1}, ..., \tilde{\mathbf{Y}}^{m,R}$, for $m = 1, ..., M$.

Every variable except gender, the first spouse-link observed in the SIPP, and the type of SSA benefit information was synthesized. We used three different techniques to estimate univariate conditional models for the variables in our data. For nearly continuous variables such as dates and dollar amounts we used linear regression. For binary variables (indicators for positive earnings, death) or certain discrete variables that could be split into sensible trees of binary variables (education category, race), we used logistic regression. For some discrete variables that were especially difficult to model, we used Bayes' Boostraps independently across a very fine stratification of the data.

We also stratified the data for the regression methods, although the stratification had to be much coarser so that coefficients could be efficiently estimated for a large number of explanatory variables. When cells were too small to estimate a reliable regression, those cells would be stacked up and split up again according to a coarser stratification. In an attempt to reduce some of our influence on the models, Bayes Information Criterion (BIC) was used to reduce the variable list by eliminating

right-hand-side variables that had a posterior odds ratio below a pre-specified level.

Many of the variables modeled using linear regression have very irregular univariate distributions, such as massive spikes, cliffs, multiple modes, and/or distant outliers. Moreover, the extremely large number of variables in our dataset made it impractical to find appropriate transforms to use for each variable, so we used a general transform method where these variables with irregular distributions on both sides of the equation were independently transformed to be approximately standard normal using kernel density estimators (KDE)[7]. This helped to preserve many of the irregular univariate features in the imputed data.

As mentioned earlier, our data contains many very complex, and often exact, relationships between variables that required very careful "book-keeping." Some variables are exact functions of other variables, in which case, those other variables would be modeled sequentially after which the variable in question could be calculated. Sometimes, variables seem like they ought to have exact relationships but clearly do not in the underlying confidential data. In such cases, we could either (1) clean the underlying confidential data prior to completion and synthesis and then enforce the intuitive rules during imputation or (2) allow the imputed values to violate the intuition but try to control the models so that the violations are not exacerbated in the imputed data. These decisions were made weighing the competing incentives of wanting to change the original data as little as possible and wanting to make certain very important variables as palatable as possible to the expected users of the data. Birthdate, deathdate, and the earnings histories were forced to be consistent with each other. We also forced the dates of educational achievement to be internally consistent. On the other hand, a variable such as the categorical time of arrival to the USA for foreign born individuals was not edited despite occasionally contradicting birthdate or labor force participation in the gold standard

[7]See Woodcock and Benedetto (2006) for details.

file.

Some variables are only in-scope to have values in certain situations. For these variables we would create "parent-child" relationships, where the parent variable takes on certain values that determine whether the child variable is in-scope. Then we would estimate the univariate conditional model for the parent and sample from its posterior predictive distribution. Once the parent was completed/synthesized, we would then estimate the univariate conditional model for the child and sample from its posterior predictive distribution only for the cases where the parent was imputed to be in the correct range for the child to be in-scope. Sometimes these parent-child trees would go several layers deep (e.g. the marital histories – one can not have a second marriage unless one has had a first marriage and that first marriage has ended). Every variable modeled with linear regression had a parent variable so that there would not be massive spikes at 0. For instance, an indicator for whether an individual worked in a given year would be modeled with logistic regression before proceeding to the amount of earnings received in that year.

Other times a variable (let's call it variable A) does not affect whether another variable (variable B) is in-scope, but does affect the range of values B can take on. Once again, variable A would be modeled and imputed first, and then variable B would be modeled and imputed. However, in this situation, when imputing variable B, we keep sampling from its posterior predictive until a draw falls in the acceptable range determined by the imputed value of A. For instance, the dates and ages of educational achievement restricted each other. If someone was imputed to have attained a post-graduate degree, then the date of completion for that degree was restricted to be greater than the date of attaining a bachelor's degree, which in turn was restricted to be greater than the date he/she finished high school.