Nature or Nurture?
Data and Estimation Appendix

Alessandra Fogli
University of Minnesota
and CEPR

Laura Veldkamp
NYU Stern School of Business
and NBER

March 11, 2010

This appendix contains details about the construction of our county level dataset, summary statistics for all variables, survey data about changing attitudes towards female labor force participation in US, international evidence about labor force participation, and details about the results of the dynamic panel estimation reported in Table 2 of the paper.

1 Data Description

1.1 County-level data

Our county level dataset has information on a vast array of economic and socio-demographic variables for 3074 US counties over the period 1940-2000 for each decade. Most of the information comes from Census data, and in particular from a dataset called "Historical, Demographic, Economic and Social Data: The United States, 1790-2000", ICPSR, Study No. 2896. However, we integrated this dataset using several others, including the Census of Population and Housing, the County and City Data Book, the Census 2000 Summary Files, and IPUMS to obtain the most complete and homogeneous information at the county level for this span of time. Sources and details about the construction of each single variable are presented in Table 1. Table 2 reports summary statistics for each variable decade by decade.

1.2 Survey data

The survey data from GSS begin only in 1972. However, the increasing speed of female entry in the labor force (start of the S) precedes that date. To establish the contemporaneous S-shaped evolution of beliefs, it is vital to have more historical data. We have one measure of beliefs that is collected infrequently, since the 1930's. This data are from IPOLL databank, maintained by the Roper Center for Public Opinion Research. Unfortunately, the phrasing of the questions differs slightly over time. We describe below the questions and the replies.

August 1936 The Gallup Poll asked: "Should a married woman earn money if she has a husband capable of supporting her?" 18% said yes, 82% no. No uncertain or no response entries were allowed.

October 1938 The Gallup Poll asked: "Do you approve of a married woman earning money in business or industry if she has a husband capable of supporting her?" 22% approve, 78% disapprove.

November 1945 The Gallup Poll (AIPO) asked: "Do you approve or disapprove of a married woman holding a job in business and industry if her husband is able to support her?" 62% disapprove, 18% approve. The rest of the replies are miscellaneous open answers (e.g., if she has a good job, if she has no children, etc.).

June 1970 The Gallup Poll asked: "Do you approve of a married woman earning money in business or industry if she has a husband capable of supporting her?" 60% approve, 36% disapprove, 4% do not know.

From 1977 on, data come from http://webapp.icpsr.umich.edu/GSS/. The question is: *Do you agree with the following statement: A preschool child is likely to suffer if his or her mother works.* (Strongly agree=1, agree=2, diagree=3, strongly disagree=4, don't know=8, no answer=9, na=0). The only modification we make is to treat "don't know" and "na" replies as missing observations. There are 14 observations, one in 1977, and then at least every two years from 1995-2004. There are between 890 and 2,344 responses per year, totalling 19,005 observations. The average reply ranges from 2.2 in 1977 to 2.6 in 2004.

Merging the two data series: From the Roper data, there are 3 observations available before 1967 and then regular observations starting in 1970. For each of the pre-1977 observations, we compute the growth rate from one data point to the next. Then, we apply these same growth rates to project our preschool data back from 1977 to the earlier observations. We believe that using one series to infer another is a reasonably accurate procedure because for years in which both survey questions are asked, the correlation in the replies is 0.75.

1.3 Cross-country data

The key moments of the data that the model seeks to explain are the rise and fall of the dispersion in female participation rates and the S-shaped increase in the level. Both of these patterns are not unique to the U.S.. The same patterns show up in European country data as well.

We use data from ILO, Economically Active Population, 1950-2010, (Geneva, 1997) to describe this fact. The data set covers Denmark, Finland, Sweden, UK, Ireland, Belgium, France, Netherlands, Greece, Italy, Portugal, Spain, Austria, Germany. We do not have local data within each country. However, we can treat each country like a region and compute the moments across countries. We computed the equally-weighted mean and cross-country standard deviation of female labor force participation rates in each decade. The results are reported in figures 1 and 2.

Not only is the shape of the participation and dispersion graphs similar in Europe, the timing is similar as well. As in the U.S., participation takes off in the 1970's and 80's. And as in our model, the dispersion of participation rates peaks around 1980. The major difference is that in Europe, dispersion decreases slightly in the 1950's and 60's, before taking off again in the 1970's.

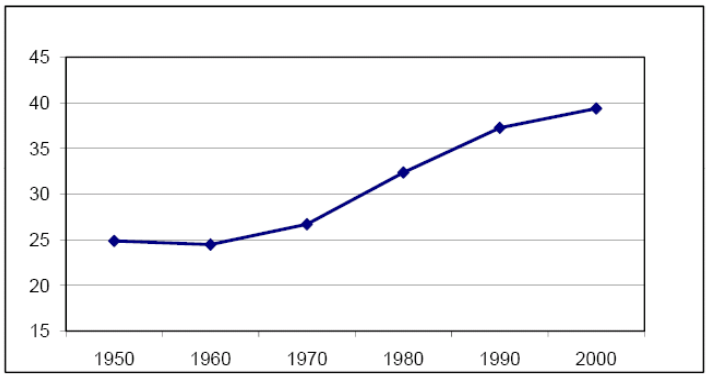

Figure 1: Average female labor force participation across European countries.

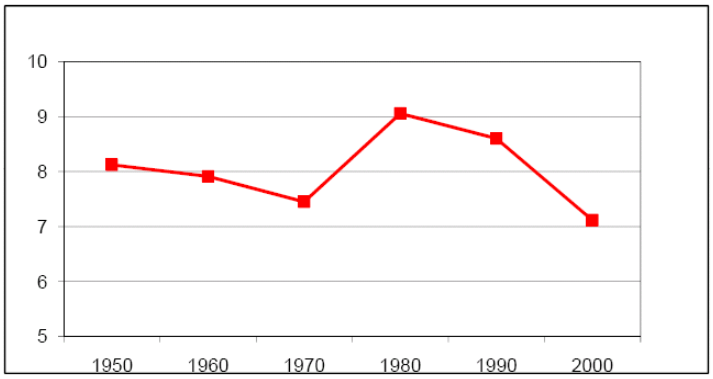

Figure 2: Dispersion of female labor force participation rates across European countries.

2 Panel Data Estimation Procedure

In order to gauge the statistical strength of the relationship between neighboring counties' LFP, we estimate the coefficients of equation (11) in the main text, which we reproduce here for convenience:

$$LFP_{it} = \rho LFP_{i(t-1)} + \beta \bar{L}_{i(t-1)} + \gamma_t + \phi_i x_{it} + \alpha_i + \epsilon_{it}. \tag{1}$$

The term $\bar{L}_{i(t-1)}$ is distance-weighted sum of other counties' participation rates, where the distance is one for counties that share a common border with the region of interest and is zero otherwise. We construct the contiguity matrix from latitude and longitude of the centroid of each county using the function" xy2cont" in Pace and Barry's Spatial Statistical Toolbox for MATLAB. The spatial weight matrix is row-standardized.

The exogenous county-level control variables x_{it} are listed in Table 3.

In the discussion that follows, we start with simple estimation procedures, point out the econometric problems that they may suffer from, and show how we address each problem. In each specification, we find that the coefficient on $\bar{L}_{i(t-1)}$, which captures the geographic relationship our model predicts, is statistically and economically significant. Furthermore, the estimates that come

from the data are similar to those that emerge when we apply the same estimation procedure to the simulation output from the model. Thus, the results are consistent with the prediction of a model based on local learning.

2.1 Ordinary Least Squares estimation.

The first row of Table 3 reports OLS estimates of equation (1). This estimation raises two causes for concern. The first issue, typical of dynamic panels, is that the lagged variable is correlated with the individual fixed effects (μ_i) and therefore with the error term. This makes the OLS estimator biased and inconsistent, even if the errors are not serially correlated. The same problem applies to the lagged spatial variable, which is a linear combination of the y_{it}s and therefore also a function of the individual effects. The second issue is that, in the presence of serial correlation in the error term, again both the lagged variable and the lagged spatial variable would be correlated with the error term.[1]

2.2 Instrumental Variables

We first-difference (1) to eliminate fixed effects:

$$LFP_{it} - LFP_{i(t-1)} = \rho(LFP_{i(t-1)} - LFP_{i(t-2)}) + \beta(\bar{L}_{i(t-1)} - \bar{L}_{i(t-2)}) + \gamma_t + \phi_i(x_{it} - x_{i(t-1)}) + \tilde{\epsilon}_{it}. \quad (2)$$

The remaining problem is that $(LFP_{i(t-1)} - LFP_{i(t-2)})$ is correlated with $\tilde{\epsilon}_{it} \equiv \epsilon_{it} - \epsilon_{i(t-1)}$. Therefore, we use $LFP_{i(t-2)}$ as an instrument for $(LFP_{i(t-1)} - LFP_{i(t-2)})$. Because the spatial lag term may have similar problems, we use $\bar{L}_{i(t-2)}$ as an instrument for $\bar{L}_{i(t-1)} - \bar{L}_{i(t-2)}$.

Also, since US counties may differ not just because of individual fixed effects in the levels, but also in the growth rates, in the second column of Table 3 we report estimates of equation (2) with fixed effects. This specification is controlling for time effects, individual fixed effects in levels and individual fixed effects in growth rates while instrumenting differences with lagged levels and still finds that the lagged labor force participation of contiguous counties is an important determinant of a county's female labor force participation rate.

As long as the error ϵ_{it} are serially uncorrelated, our instruments are valid. The drawback of this approach is that it is not efficient because it does not take into account all the possible moment restrictions. The next procedure remedies this problem.

2.3 Arellano Bond (1991) estimator

Arellano and Bond (1991) point out that all of the lags of the dependent variable are valid instruments, as are the additional independent explanatory variables. Including these variables as instruments improves efficiency, as long as they are correlated with the regressor they are instrumenting for.

Therefore, we use three lags: $LFP_{i(t-2)}, LFP_{i(t-3)}$, and $LFP_{i(t-4)}$ as instruments for $(LFP_{i(t-1)} - LFP_{i(t-2)})$, and $\bar{L}_{i(t-2)}, \bar{L}_{i(t-3)}$ and $\bar{L}_{i(t-4)}$ as instruments for $\bar{L}_{i(t-1)}$. In addition, we use the entire time series of all the exogenous regressors x_{it}.

[1] Static spatial panel data models have been successfully estimated using maximum likelihood (See Elhorst 2003). This approach is not directly implementable in our context since we have an explicitly dynamic model where the lagged value of the spatial lag appears on the right hand side.

The results are reported in the last two columns of Table 3. These are two-step estimates with heteroskedasticity consistent standard errors. While the estimates in the last column uses three lags as instruments for the dependent variable, the specification reported in the previous column uses only two lags and finds similar results. In both cases, the geographic variable is statistically and economically significant.

Whereas the previous IV approach was just identified, this system has more instruments than regressors and is therefore over-identified. Therefore, we can use the Sargan statistic to test the validity of the over-identifying restrictions and the validity of our instruments. The null hypothesis is that the instruments are not correlated with the residuals. For the model estimated in the fourth column, we obtain a $\chi^2(\mathbf{3})$ =1.94 and the null hypothesis cannot be rejected with a p-value of 0.58. The results of the Sargan test for the last specification are similar and indicate that the model is correctly specified.

The GMM estimator is consistent if there is no second-order serial correlation in the error term of the first-differenced equation. The test statistic m_2 is the Arellano-Bond test for second order serial correlation in the errors: the null hypothesis is that of no second order serial correlation which cannot be rejected by the data (p-values in parenthesis).

References

[1] Arellano, M., Bond S., 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, vol. 58(2), pages 277-97.

[2] Elhorst J.P., 2001. Dynamic models in space and time, *Geographical Analysis,* 33, pp. 119–140.

[3] Hsiao, C.,1986. Analysis of Panel Data, Cambridge University Press.

Table 1: Data Sources

Variables	1940	1950	1960	1970
Female labor force participation[1] %	DS32: F14, FL4LF	DS35: FL4PLUS, FL4LF	DS39: FTOT, F0_4, F5_9, 10_14 DS74: VAR34, VAR36	DS41: FTOT, F04, F56, F79, F1013, F14, F15. DS76: VAR35
Urban population %	DS71: VAR95	DS73: VAR6	DS74: VAR6	DS76: VAR8
Rural farm population %	DS70: VAR12, VAR3	DS72: VAR9,VAR2	DS74: VAR7	DS76: VAR168, VAR169, VAR3
White population %	DS32: NWTOT, FBWTOT, TOTPOP	DS35: NWMTOT, FBWMTOT, NWFTOT, FBWFTOT, TOTPOP	DS38: WHTOT, TOTPOP	DS41: WPOP, TOTPOP
Black population %	DS32: NEGTOT, TOTPOP	DS35: NEGMTOT, NEGFTOT, TOTPOP	DS38: NEGMTOT, NEGFTOT, TOTPOP	DS41: NEGTOT, TOTPOP
Education[2]	DS32: MESCHF25, MESCHM25	DS35: MEDSCH25	DS75: VAR19	DS76: VAR24
Density (persons per sq. mile)	DS70: VAR7	DS72: VAR6	DS74: VAR1, VAR3	DS76: VAR4
Wholesales establishments[3] %	DS70: VAR78 (1939)	DS72: VAR74 (1948)	DS74: VAR113 (1958)	DS76: VAR159 (1967)
Service establishments %	DS70: VAR80 (1939)	DS72: VAR77 (1948)	DS74: VAR120 (1958)	DS76: VAR149 (1967)
Manufacturing establishments %	DS70: VAR65 (1939)	DS72: VAR81 (1947)	DS74: VAR86 (1958)	DS76: VAR121 (1967)
Retail establishments %	DS70: VAR73 (1939)	DS72: VAR66 (1948)	DS74: VAR98 (1958)	DS76: VAR132 (1967)
Manufacturing wages[4]	DS70: VAR67, VAR66 (1939)	DS73: VAR73, VAR72 (1954)	DS75: VAR65, VAR64 (1963)	DS77: VAR185, VAR184 (1972)

Note: unless otherwise specified, data are from ICPSR, Study No. 2896, "Historical, Demographic, Economic, and Social Data: The United States, 1790-2000".
[1]Female labor force participation refers to female population 14 years of age and over in 1940, 1950, and 1960. In the other years, it refers to female population 16 years and over.
[2]Median school years completed by population 25 years and over. In 1980, 1990, and 2000, total population by educational attainment is weighted by average years of education.
[3]All the establishments' variables are computed as percentages of the total number of establishments.
[4]In the panel, wages are average deflated annual manufacturing wages, 1982-84=100. In 2000, it refers to median earnings.

Table 1: (Cont.)

Variables	1980	1990	2000
Female labor force participation[1] %	DS78: VAR110, Census of Population and Housing, 1980, ICPSR 8108, Var. 3,18-3,77	DS80: VAR131X, VAR133X	Census 2000 Summary File 3, Table P43
Urban population %	DS78: VAR6, VAR3	DS83: PO51090D, VAR026X	Census 2000 Summary File 3, Table P5
Rural farm population %	DS78: VAR205, VAR3	DS80: PO54090D , VAR026X	Census 2000 Summary File 3, Table P5
White population %	DS78: VAR7, VAR3	DS80: VAR9, VAR5	DS81: B2_POP06 and "County and City Data Book: 2000", Table A-2 from CENSUS
Black population %	DS78: VAR8, VAR3	DS80: VAR10, VAR5	DS81: B2_POP08 and "County and City Data Book: 2000", Table A-2 from CENSUS
Education[2]	DS78: VAR97, VAR98, VAR99, and EDUC from CENSUS IPUMS (1980)	DS80: VAR69, VAR70, VAR71, and EDUC from CENSUS IPUMS (1990)	Census 2000 Summary File 3, Table P37, and EDUC from CENSUS IPUMS (2000)
Density (persons per sq. mile)	DS78: VAR5	DS80: VAR004	DS81: B1_POP05
Wholesales establishments[3] %	DS78: VAR183 (1977)	DS80: VAR176 (1987)	DS81: B11_WHS01 (1997)
Service establishments %	DS78: VAR188 (1977)	DS80: VAR186 (1987)	DS80: VAR186 (1987)
Manufacturing establishments %	DS78: VAR165 (1977)	DS80: VAR167 (1987)	DS81: B9_MAN01 (1997)
Retail establishments %	DS78: VAR177 (1977)	DS80: VAR181 (1987)	DS81: B11_RTL01 (1997)
Manufacturing wages[4]	DS79: VAR133, VAR131	DS81: B9_MAN05, B9_MAN04	Census 2000 Summary File 3, Table P85

Note: unless otherwise specified, data are from ICPSR, Study No. 2896, "Historical, Demographic, Economic, and Social Data: The United States, 1790-2000".
[1]Female labor force participation refers to female population 14 years of age and over in 1940, 1950, and 1960. In the other years, it refers to female population 16 years and over.
[2]Median school years completed by population 25 years and over. In 1980, 1990, and 2000, total population by educational attainment is weighted by average years of education.
[3]All the establishments' variables are computed as percentages of the total number of establishments.
[4]In the panel, wages are average deflated annual manufacturing wages, 1982-84=100. In 2000, it refers to median earnings.

Table 2: Summary Statistics – County Dataset

1940	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	18.49	6.66	4.56	47.90
Urban population %	3074	23.23	25.36	0	100
Rural farm population %	3074	45.79	21.97	0	93.75
Rural non-farm population %	3074	30.99	16.94	0	100
White population %	3074	88.58	17.90	14.44	100
Black population %	3074	10.69	17.83	0	85.51
Other population %	3074	0.73	3.86	0	77.36
Education	3073	8	1.16	1.85	12.25
Density (persons per sq. mile)	3074	189.71	1979.78	0.20	85905.64
Wholesales establishments %	2954	6.77	4.23	0	29.71
Service establishments %	2954	20.64	4.83	2.74	50.82
Manufacturing establishments %	2954	4.67	2.721	0.30	26.77
Retail establishments %	2954	67.92	6.03	38	87.5
Manufacturing wages	2248	5774.12	1614.10	1640.87	11118.12

1950	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	22.47	6.49	4.58	46.56
Urban population %	3074	28.25	27.027	0	100
Rural farm population %	3074	35.77	19.78	0	93.67
Rural non-farm population %	3074	35.98	17.89	0	100
White population %	3074	89.17	17.02	15.63	100
Black population %	3074	10.079	16.86	0	84.33
Other population %	3074	0.75	3.98	0	84.05
Education	3067	8.78	1.37	0	12.7
Density (persons per sq. mile)	3074	202.37	2038.58	0.17	89096
Wholesales establishments %	3074	6.21	3.45	0	44
Service establishments %	3074	29.15	6.75	0	65
Manufacturing establishments %	3074	7.14	5.03	0	50
Retail establishments %	3074	57.50	6.92	28.11	100
Manufacturing wages	2501	8362.90	2434.15	2334.02	16100.45

1960	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	30.09	6.38	7.87	61.26
Urban population %	3074	32.02	28.28	0	100
Rural farm population %	3074	22.69	16.19	0	86.6
Rural non-farm population %	3074	45.29	21.77	0	100
White population %	3074	89.34	16.44	15.92	100
Black population %	3074	9.82	16.26	0	83.42
Other population %	3074	0.02	0.06	0	1.54
Education	3074	9.64	1.46	4.2	12.8
Density (persons per sq. mile)	3074	203.56	1838.31	0.17	77194.59
Wholesales establishments %	3074	7.46	3.81	0	41.67
Service establishments %	3074	22.04	5.91	0	55
Manufacturing establishments %	3074	7.58	4.86	0	61.54
Retail establishments %	3074	62.92	6.76	29.10	100
Manufacturing wages	2568	11731.28	3716.23	750.75	23437.07

Table 2: (Cont.)

1970	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	36.53	6.47	8.24	65.28
Urban population %	3074	34.72	29.02	0	100
Rural farm population %	3074	14.93	13.35	0	82.35
Rural non-farm population %	3074	50.36	24.47	0	100
White population %	3074	89.62	15.23	13.50	100
Black population %	3074	9.22	14.96	0	80.11
Other population %	3074	1.15	4.52	0	86.40
Education	3074	10.90	1.38	5.3	14.4
Density (persons per sq. mile)	3074	210.58	1730.21	0.18	66923
Wholesales establishments %	3074	6.92	3.32	0	29.51
Service establishments %	3074	30.34	5.73	0	55.24
Manufacturing establishments %	3074	7.23	4.82	0	53.19
Retail establishments %	3074	55.50	6.09	27.13	100
Manufacturing wages	2289	13498.61	15139.14	1030.93	27384.02

1980	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	44.59	6.94	18.45	79.99
Urban population %	3074	35.96	29.10	0	100
Rural farm population %	3074	9.56	9.88	0	64.82
Rural non-farm population %	3074	54.47	25.72	0	100
White population %	3074	88.48	14.98	6.05	100
Black population %	3074	8.61	14.41	0	84.16
Other population %	3074	2.90	6.48	0	93.84
Education	3074	11.96	0.79	9.88	15.01
Density (persons per sq. mile)	3074	206.60	1570.39	0.2	64395.2
Wholesales establishments %	3074	7.99	3.67	0	31.58
Service establishments %	3074	36.39	5.95	0	63.57
Manufacturing establishments %	3074	7.17	4.11	0	39.02
Retail establishments %	3074	48.45	6.01	22.47	100
Manufacturing wages	2360	12816.09	3600.33	3640.78	44902.91

1990	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	51.856	7.06	25.8	84.1
Urban population %	3074	36.19	29.60	0	100
Rural farm population %	3074	6.56	7.38	0	68.41
Rural non-farm population %	3074	57.25	26.92	0	100
White population %	3074	87.53	15.30	5.04	99.95
Black population %	3074	8.61	14.36	0	86.23
Other population %	3074	3.86	7.55	0	94.91
Education	3074	12.66	0.70	10.42	15.15
Density (persons per sq. mile)	3074	209.01	1434.32	0.312	53126.29
Wholesales establishments %	3074	8.53	3.85	0	36.36
Service establishments %	3074	24.11	6.92	0	54.03
Manufacturing establishments %	3074	7.17	3.78	0	33.33
Retail establishments %	3074	60.18	7.77	29.02	100
Manufacturing wages	2334	14664.19	4296.08	3060.44	30305.86

Table 2: (Cont.)

2000	N	Mean	Std. Dev.	Min	Max
Female labor force participation %	3074	54.69	6.51	26.62	80.86
Urban population %	3074	39.80	30.66	0	100
Rural farm population %	3074	4.91	5.78	0	43.94
Rural non-farm population %	3074	55.28	28.07	0	100
White population %	3074	84.87	15.97	4.5	99.7
Black population %	3074	8.80	14.54	0	86.5
Other population %	3074	6.32	8.79	0.3	95.4
Education	3074	12.85	0.69	10.63	15.84
Density (persons per sq. mile)	3074	232.02	1665.90	0.3	66834.6
Wholesales establishments %	2113	13.47	4.89	1.96	38.39
Service establishments %	2113	21.36	5.38	3.12	50.55
Manufacturing establishments %	2113	14.86	5.29	3.07	43.48
Retail establishments %	2113	50.30	6.52	26.09	71.43
Manufacturing wages	1965	16562.77	4231.06	6430.60	35959.49

Table 3. Dynamic Panel with Spatial Lag Estimation Results

Dependent variable: Labor Force Participation at time t

	OLS	IV DIF FE	GMM DIF (2L)	GMM DIF (3L)
Labor Force Participation$_{t-1}$	0.664***	0.305***	0.887***	0.916***
	(0.010)	(0.052)	(0.064)	(0.062)
Labor Force Participation Spatial Lag $_{t-1}$	0.195***	0.577***	0.522***	0.570***
	(0.011)	(0.125)	(0.107)	(0.103)
Density (thousands persons per sq. mile)	-0.063	0.051	-0.504*	-0.589*
	(0.032)	(0.072)	(0.226)	(0.255)
Urban population (percentage)	0.015***	0.013	-0.022	-0.010
	(0.002)	(0.007)	(0.026)	(0.026)
Rural farm population (percentage)	0.007*	-0.012	-0.108***	-0.098***
	(0.003)	(0.023)	(0.026)	(0.026)
Education (average years)	0.643***	-0.176	-1.120	-0.975
	(0.036)	(0.120)	(0.604)	(0.587)
Wages	-0.041	-0.015	4.224	3.093
	(0.031)	(0.017)	(1.835)	(1.790)
m1	2.59	-10.85	-1.7	-2.36
m2	4.30	-1.44	-0.27	0.03
Sargan			0.585	0.349

Year dummies included in all specifications. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
Robust standard errors in parentheses are clustered at county level.
m1 and m2 are tests for first order and second order serial correlation.
GMM results are two-step estimates with heteroskedasticity consistent standard errors.
Sargan is a test of the overidentifying restrictions for the GMM estimators. P value is reported.