An Application of Panel Regression to Pseudo Panel Data

Jeffrey E. Russell **John W. Fraas**
Ashland University

This article illustrates how, in the absence of true panel data, multivariate regression analysis can be used in conjunction with a pseudo panel data set to identify variables that were related to the increase in the proportion of two-income spouses in the United States between 1940 and 2000. We present the procedures used to form the pseudo panel data set, construct and estimate the various models used to analyze the pseudo panel data, and interpret the results produced by those models. Our analysis revealed an inverse relationship between the proportion of two-income spouses and the presence of young children as well as an increasing trend across generations in the proportion of two-income spouses.

This article provides an illustration of how researchers can apply panel regression analysis techniques to pseudo panel data when true panel data are not available. The analysis conducted in this study was designed to identify variables that were related to the dramatic increase in the proportion of two-income, married couples in the United States between 1940 and 2000. Ben-Porath (1973) suggested that such investigations are best addressed through the use of panel data, and Baltagi (1995) described the substantial advantages provided by panel data analysis relative to cross-section or time-series data. However, due to the extensive time period covered by the data in our study, we were unable to form an appropriate panel data set. Thus, we constructed a pseudo panel data set to act as a substitute for a true panel data set.

The remaining sections of this article present the techniques we employed to construct and analyze this pseudo panel data set. Specifically, the first section of this article shows how a pseudo panel data set was constructed using cross-section data from the Decennial United States Census collected from 1940 to 2000. The second section describes the specifications of the various models used to analyze the pseudo panel data. The third section discusses the procedures followed to interpret the results produced by the various models. The final section summarizes the findings and the procedures used to produce those findings.

Construction of a Pseudo Panel Data Set

Panel data and pseudo panel data sets are obtained by pooling comparable cross-section data collected repeatedly over time. To maintain comparability, both true panel data and pseudo panel data should be based on responses to similar questions collected in a similar manner. True panel data also needs to be repeatedly collected from the same individuals across time to ensure comparability. The formation of a true panel data set is usually not a significant problem if individuals are defined to be a relatively small number of entities such as the member countries of the United Nations Security Council, and the questions are unambiguous (e.g., What is the population of each country?). In these situations, panel data covering an extended period of time may be constructed and pseudo panel data are usually not needed as an alternative.

Comparability over time becomes a more significant issue for true panel data if an individual is defined to be *individual* people or households and the number of individuals is very large. The Panel Study of Income Dynamics (PSID) from the Survey Research Center at the University of Michigan, and the National Longitudinal Surveys of Labor Market Experience (NLS) from the Center for Human Resource Research at The Ohio State University are two examples of this type of large, individual panel data. These high-quality data sets are very careful to pose consistent questions to the same individuals across time. Nonetheless, continued comparability becomes increasingly difficult over time as data are lost. A loss of data can occur due to individuals (a) failing to answer some questions in one or more time periods, (b) failing to respond at all in some years, or (c) dropping out of the data set because of death, migration, or deciding to no longer participate in the survey.

When the loss of data is non-random, researchers are faced with potential problems of bias that become increasingly problematic over time, even in top quality panel data. Since the likelihood of non-random data loss increases as the time period covered by the panel data increases, large panel data sets usually cover a relatively short period of time. To answer long-term individual behavioral questions, such as the ones we are addressing in this article, pseudo panel data can be used as a substitute for the unavailable true panel data.

Deaton (1985) demonstrated that a pseudo panel data set has the advantage of a less stringent requirement. That is, the data can be repeatedly collected from random samples drawn from the same time-stable cohort of individuals rather than repeatedly from the same specific individuals. Pseudo panel data are constructed by first defining cohorts using individual characteristics that are stable over time. If the size of each cohort is sufficiently large, successive surveys will generate successive random samples of individuals from each of the cohorts. For every cohort, the mean value for each variable is then calculated for each time period. These mean values become the observations in the pseudo panel data. As noted by Deaton, this procedure allows pseudo panel data to be constructed from any series of cross-section data that includes variables that can be used to identify stable cohorts.

In addition to filling gaps in the availability of true panel data, Deaton (1985) identified four additional advantages of pseudo panel data. First, data from different sources can be combined into a single set of pseudo panel data if comparable cohorts can be defined in each source. Second, attrition problems often found in true panel data are minimized. Third, the problem of the individuals' response errors is smoothed by the use of cohort means and can be explicitly controlled by using errors-in-variables methods. Fourth, inconsistencies between micro and macro analysis can be analyzed by moving from individual data to ever larger cohorts to one macro cohort.

Source of Data and Data Issues

As previously discussed, it was necessary to identify a source of successive surveys for the 1940–2000 time period. For our analysis, the successive surveys were the one percent public use microdata samples available through the United States Census Bureau for the seven census years beginning with 1940 and ending with 2000 (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King, & Ronnander, 2004). Prior to forming our pseudo panel data set from this information, three data issues were addressed: (a) cohort stability over time, (b) measurement error bias, and (c) differentiation between age, period, and cohort effects.

Establishing the stability of cohorts over time. Even if awkward variables result, time-constant cohort definitions must be used with pseudo panel data. We defined cohorts using race, gender, and generation to prevent the movement of individuals between cohorts over time. Because we were investigating the work behavior of married couples, we first considered marital status as an additional cohort definition. While this would allow the straightforward calculation of the proportion of married couples that are two-income couples, marital status cannot be used as a cohort definition because individual marital status is not necessarily constant over time. To create a dependent variable of interest while maintaining cohort stability we calculated the proportion of the generation-race-gender cohort that consisted of working individuals married to working individuals. While the proportion of the cohort that consists of working individuals married to working individuals is not as straightforward as the simple proportion of married couples that are two-income couples, this type of variable definition was necessary to maintain cohort stability.

Addressing errors in measurement. As with true panel data, observations that are measured with a systematic error may need to be eliminated from the data to avoid biased results. The possible gains in obtaining more interpretable results produced by this technique must be weighed against the potential for bias due to a systematic elimination of a non-random group of individuals. For example, our study uses data from questions posed to individuals by the United States Census Bureau regarding their work behavior. The questions are not designed to reflect a farmer's work pattern. Consequently, a high degree of error in the responses of individuals engaged in farming exists. To eliminate this source of error in the data, we followed the practice suggested by Coleman and Pencavel (1993) and eliminated all observations from individuals living on farms prior to the calculation of the pseudo panel data cell means. This decision limits the applicability of the results to non-farmers. More importantly, this decision implicitly assumes the migration between census years of individuals off the farm and into the population used to calculate the cohort means was a random event and introduced no systematic bias. Immigrants to the United States were also eliminated from the data to maintain cohort stability.

Differentiating between age, period, and cohort effects. When data contain observations on many individuals over an extended period of time, observed variance can be attributed to three functionally related effects: (a) differences between cohorts, which are labeled the cohort effect; (b) differences

associated with different points in the life cycle, which are labeled the age effect; and/or (c) differences associated with different periods, which are labeled the period effect. The problem that must be addressed regarding these three effects is that they cannot be simultaneously identified because only one time dimension and one individual or cohort dimension exists. More specifically, the functional relationship between all three effects causes perfect colinearity when all three effects are fully specified (Fienberg & Mason, 1985; Ryder, 1965). For example in our data set, if a regression model includes a cohort variable of 1910 for the 1906-1915 birth cohort, and a mean age variable of 40 for that cohort using the 1950 census, the 1950 period variable cannot be specified because it is already defined by the cohort and age variables (e.g., 1910 + 40 = 1950).

The question of how best to solve this identification problem has generated controversy, especially among sociologists (Rodgers, 1982; Smith, Mason, & Fienberg, 1982). If a linear restriction is imposed on any pair of age, period, or cohort variables (e.g., the membership in the cohort born 1906-1915 is no different from membership in the 1916-1925 cohort, thereby restricting the cohort variables to be equal for this pair), then the results are identifiable. However, Rodgers shows that such a restriction must be made on strong a priori grounds, and the researcher should know the restriction can easily distort the results.

An alternative solution is to recognize that the three accounting variables are proxies for substantive characteristics associated with age, period, and cohort. If one of the accounting measures can be replaced with a direct measure of a characteristic, the identification problem is solved (Feinberg, et al., 1985). For example, if the accounting variable for period is replaced with a substantive measure of the unemployment rate for each year, the age and cohort effects can be identified. The weakness of this strategy, however, is the inherent assumption that the substantive measure fully captures all aspects of the effect. In other words, the use of the unemployment rate implicitly assumes there are no other substantive period effects such as military conflicts or high rates of inflation.

We addressed the age, period, and cohort identification problem by using a linear restriction that all period effects are equal and are included in the constant term. This assumption allows a set of mean age dummy and cohort dummy variables to exactly identify the cohort and age effect in the regressions. An assumption that the period effect is a linear trend would also solve the identification problem. It should be noted, however, that the regression results produced by using a trend specification are more difficult to interpret because the cohort and age coefficients would then measure deviations from the trend.

Formation of the Cohorts

Stable cohorts were defined by race, gender, and generation (i.e., year of birth). The race characteristic was restricted to Caucasians and African-Americans only to maintain sufficient sample size for each of the two cohorts. The gender characteristic consisted of a male cohort and a female cohort. And the generation characteristic consisted of seven cohorts with each cohort representing a ten-year span. The first and seventh generation cohorts contained individuals born between 1906 and 1915, and between 1966 and 1975, respectively.

The race (2), gender (2), and generation (7) cohort definitions describe 28 potential (2*2*7 = 28) cohorts. Repeated over the seven census years, there was a potential of 196 cells of cohort mean data. However, to reduce the impact of schooling and retirement on the decision to work, individuals younger than 25 or older than 64 were excluded from the data. Consequently, beginning with the generation born in the years from 1946 to 1955, complete working age life-cycle data were not available because individuals in these later generations were less than 55 in the 2000 census. Similarly, as the oldest cohorts reach the age of 65, they no longer contribute data. As a result, the number of cells with data was reduced to 88 cells of sample mean data drawn from 28 distinct cohorts. Table 1 lists the actual set of 88 data cells that define the 88 cohort observations. The numbers listed in the cells indicate the number of individuals contained in the cohorts each census year. A review of Table 1 reveals the secular movement of younger cohorts into the data set and older cohorts out of the data set which reduced the number of useable cells to 88.

Table 1. Total Number of Observations in Each of the 88 Cohort Cells

	Census years						
	1940	1950	1960	1970	1980	1990	2000
White male born:							
1906-1915	21,943	23,208	80,242	72,874			
1916-1925		26,066	94,786	92,855	86,610		
1926-1935			89,373	93,129	92,353	89,550	
1936-1945				101,325	105,428	107,078	97,284
1946-1955					152,274	155,418	146,964
1956-1965						170,077	164,808
1966-1975							131,788
Black male born:							
1906-1915	1,897	2,299	7,865	6,957			
1916-1925		2,463	9,446	9,144	8,338		
1926-1935			9,539	9,926	9,684	7,516	
1936-1945				10,805	11,685	9,648	9,511
1946-1955					18,202	15,272	15,986
1956-1965						18,152	20,151
1966-1975							17,398
White female born:							
1906-1915	23,373	25,246	84,353	82,503			
1916-1925		29,139	100,706	99,122			
1926-1935			94,274	97,827	98,293	99,286	
1936-1945				106,583	108,219	111,627	105,490
1946-1955					155,083	159,867	151,485
1956-1965						174,432	169,314
1966-1975							134,886
Black female born:							
1906-1915	2,459	2,678	8,886	8,379			
1916-1925		3,240	11,338	10,799	10,138		
1926-1935			12,208	12,660	12,431	10,221	
1936-1945				14,065	14,852	12,304	12,079
1946-1955					22,416	19,071	20,126
1956-1965						23,611	25,833
1966-1975							23,256
Mean	12,418	14,292	50,251	51,907	62,821	73,946	77,897

Variable Formation

Three different types of variables can be used to represent the various characteristics of the pseudo panel data cohorts. A given characteristic can be represented by (a) a continuous variable, (b) one or more dummy variables, or (c) one or more proportional variables. The type of variable or variables formed to represent a given characteristic is, for the most part, dictated by the type of individual information collected in the surveys and its relationship to the cohort definitions.

Continuous variable. Some of the information used to form a pseudo panel data set may reflect a continuous type of measurement, such as income. Information of this nature would be used to form a continuous variable in the pseudo panel data set. For example, a continuous income variable in the pseudo panel data set would be formed by calculating the mean income for the individuals in each cell

(i.e., cohort). It would have been possible for us to form an income variable in this manner, but we did not because income and educational level variables were highly correlated. We included only educational level variables in our analysis. One reason for selecting educational levels rather than income was to avoid the problems caused by truncated income data for individuals who were not working because their income was too low. Thus, it should be noted that we did not form any continuous variables to represent cohort characteristics.

Dummy variables. Other types of information used to form a pseudo panel data set could simply reflect the presence or absence of a specific characteristic for a given person. For certain cells, a characteristic was possessed by everyone in the cell or by no one in the cell. Variables formed from this type of information are true dummy variables. That is, these variables contain only values of zero or one. Dummy variables were used to represent four characteristics identified in our pseudo panel data set: (a) gender, (b) race, (c) generation, and (d) age.

Since gender and race consisted of only two categories, only one dummy variable was required to represent each of these characteristics. A value of zero was assigned to every cell that contained only males, while a value of one was assigned to every cell that contained only females. For example, the cohort of Caucasian males who were born between 1906 and 1915 and who responded in 1940 contained only males. Thus, the value for the gender variable was set equal to zero for this cohort. In the same fashion, a value of zero was assigned to every cohort that contained only Caucasians, while a value of one was assigned to every cohort that contained only African Americans. To facilitate the interpretations of the models used to analyze the pseudo panel data the gender and race variables were named for the groups assigned the value of one. Thus, the gender and race variables were labeled female and African American, respectively.

The generation characteristic specified whether an individual was or was not a member of a given generation. Unlike the gender and race characteristics, however, the generation characteristic consisted of more than two categories or levels. The generation characteristic consisted of the following seven levels: (a) Born 1906-1915, (b) Born 1916-1925, (c) Born 1926-1935, (d) Born 1936-1945, (e) Born 1946-1955 (f) Born 1956-1965, and (g) Born 1966-1975. Thus, seven dummy variables, with names corresponding to the cohort labels, were constructed to represent this generation characteristic.

While the information related to age would have allowed the calculation of a continuous mean age variable, we instead used four dummy mean age variables defined by the cohort's birth year and the census year to represent four levels of age. The four age levels were (a) Mean Age 30, (b) Mean Age 40, (c) Mean Age 50, and (d) Mean Age 60. Each variable contained a zero or a one value. We used this set of dummy variables with names corresponding to the cohort labels rather than a single, continuous age variable to avoid a linear restriction on the impact of age on the proportion of two-income spouses. Another reason for not using a continuous variable is that the calculation of a mean age every ten years for a group that is evenly distributed over ten possible birth-years results in very discontinuous mean age values that cluster tightly around the mean ages.

Proportional variables. Even though some information may indicate the presence or absence of a specific characteristic for each person in a cohort, the cohort will not be uniform regarding that characteristic. That is, the cohort will contain both individuals with the characteristic and individuals without the characteristic. For such variables, which are called proportional variables, a value equal to the proportion of individuals in the cohort with the characteristic was assigned to that cell for the variable. Our pseudo panel data contained four characteristics that required the formation of one or more proportional variables.

The dependent variable for this study, which was named two-income spouse, was a proportional variable. The 88 values formed for this variable were equal to the proportion of individuals identified as working and married to a working spouse in each of the 88 cells. To illustrate, since 40% of the individuals in the cohort containing male Caucasians born between 1906 and 1915 who responded to the survey in 1940 had a working spouse, the value for that cell in the dependent variable was 0.40. This value indicates the probability is 0.40 that an individual in that cohort will be a two-income spouse.

Proportional variables were constructed for three additional characteristics: (a) young children, (b) marital status, and (c) education level. The proportional variables formed for these characteristics were identified as independent variables. One proportional variable was constructed for the young children characteristic. Each value contained in this variable, which was named young child present, indicated the

proportion of the individuals in a given cohort who had at least one child less than five years of age. The marital characteristic was also represented by one proportional variable. Each value recorded for this variable, which was labeled married, represented the proportion of individuals in a given cohort who were married. The educational characteristic reflected four levels of education: (a) less than high school, (b) high school graduate, (c) more than high school but less than four years of college, and (d) four or more years of college. Four proportional variables formed to represent these four education levels were named (a) less than HS, (b) HS graduate, (c) more than HS, and (d) four or more years of college. Every value for each of these variables was the proportion in the cohort with that level of education. For example, the 0.60 value recorded for the cohort containing male Caucasians born between 1906 and 1915 who responded to the 1940 survey indicated that 60% of these individuals had an education level less than high school.

Specification and Estimation of the Pseudo Panel Linear Regression Models

Since panel data can vary over both time and individuals, variables in a panel data regression model typically have a double subscript as follows:

$$y_{it} = \alpha + \beta x_{it} + u_{it} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{1}$$

where i represents the cross-section dimension (e.g., individuals, households, firms, countries, etc.) and t represents the time series dimension. α is a scalar, β is a vector of K explanatory variables, and x_{it} is the ith observation from time t on K explanatory variables.

Most panel data analyses use the following *one-way* error component model:

$$u_{it} = \mu_i + \nu_{it} \tag{2}$$

where μ_i represents unobservable, individual specific effects that do not change over time and ν_{it} represents the remaining unobserved effects that vary over both individuals and time. Combining Equations 1 and 2, the one-way model is fully described as follows:

$$y_{it} = \alpha + \beta x_{it} + \mu_i + \nu_{it} \qquad i = 1,\ldots,N \quad t = 1,\ldots,T \tag{3}$$

The term *one way* refers to the decomposition of the error component in only the one dimension of time-constant, individual specific unobserved effects. The following *two-way* error component model is also possible:

$$u_{it} = \mu_i + \lambda_t + \nu_{it} \tag{4}$$

where the symbol λ_t represents unobservable, time-specific effects that do not change over individuals. An example of this type of effect would be different levels of funding in different years for a school district that impact all individual students in a similar, yet unobservable manner.

As with true panel data, a set of T independent cross sections represented by Equation 3 is pooled in pseudo panel data. Unlike true panel data, however, with pseudo panel data, N is a new, and most likely different set of individuals sampled in each census. To construct pseudo panel data, a set of C cohorts is defined such that any individual i sampled from the population will always be in the same, unique cohort every year. For example, in the data used in our analysis, an African-American male born in 1930 would be included in the African-American, male, 1926 -1935 cohort if that person was sampled in the 1960 census, and that person would be included in the same cohort if that person happened to be included in the 1980 sample.

Taking the mean value of each cohort's sample in each time period results in:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\mu}_{ct} + \bar{\nu}_{ct} \qquad c = 1, \ldots, C \quad t = 1, \ldots, T \tag{5}$$

In this equation $\bar{y}_{ct}$ is the average of y_{it} over all individuals belonging to cohort c at time t. Unlike μ_i obtained from the true panel data Equation 3, $\bar{\mu}_{ct}$ retains the t subscript to indicate that each period's cohort mean is calculated from a new, and most likely different set of individuals. This results in a potentially different $\bar{\mu}_{ct}$ value for each period. In practice, if the number of individuals in each cell is

large, as is the case for the data used in this article, the assumption is made that ($\bar{\mu}_{ct} = \bar{\mu}_c$) for every *t* and the fixed cohort effect ($\bar{\mu}_c$) is treated like a fixed individual effect (μ_i), resulting in the basic pseudo panel equation:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\mu}_c + \bar{\nu}_{ct} \qquad c = 1, \ldots, C \quad t = 1, \ldots, T \tag{6}$$

Additionally, if the cell size is large, random individual fixed effects will tend to be eliminated in the process of estimating the cell mean, leaving only the cohort fixed effect.

Much of Deaton's (1985) seminal work on pseudo panel data focuses on the availability of variances and covariance obtained in the construction of the cohorts' sample means which can then be used to weight the analysis of the pseudo-panel data using an "errors-in-variables" technique. Baltagi (1995) notes that as the average cohort size (number of cohorts/sample size) tends to infinity, measurement errors as well as their estimates tend to approach zero. Consequently, as is the practice followed by many applied researchers (e.g., Pencavel, 1998), the analyses presented in this article ignore the measurement error problem and simply weight the analysis using cell-size to address heteroscedasticity arising from the different levels of precision for cell means with different numbers of observations.

The Random Effects Model

The analysis of the pseudo panel data set begins with the estimation of the Random Effects Model using Equation 6. This model assumes that the $\bar{\mu}_c$ error term, which represents possible bias from unobserved, fixed cohort heterogeneity, is identically and independently distributed (IID) with a mean of zero (Baltagi, 1995). Baltagi also notes that this assumption allows the Random Effects Model to support inference for the population, assuming the sample is representative of the underlying population. Consequently, the Random Effects Model is preferred when analyzing either panel or pseudo panel data sets.

In our Random Effects Model the vector of cohort variables, x_{ct}, included the following:

1. The gender and race characteristics were represented by dummy variables named female and African American, respectively.

2. The characteristics of whether individuals had at least one child less than 5 and their marital status were represented by proportional dummy variables named young child present and married, respectively.

3. Since the four dummy variables used to represent the four levels of the age characteristic were linearly dependent, only three of the variables were included in the model: (a) mean age 40, (b) mean age 50, and (c) mean age 60. The Mean Age 30 age level served as the reference group for the coefficients estimated for these three variables.

4. Since the six dummy variables used to represent the generation characteristic were linearly dependent, only five of the variables were included in the model: (a) born 1916-1925, (b) born 1926-1935, (c) born 1936-1945, (d) born 1946-1955, (e) born 1956-1965, and (e) born 1966-1975. The Born 1906-1915 cohort level served as the reference group for the coefficients estimated for these five variables.

5. Since the four dummy variables used to represent the four levels of the education characteristic were linearly dependent, only three of the variables were included in the model: (a) less than HS, (b) more than HS, and (c) four or more years of college. The HS Graduate education level served as the reference group for the coefficients estimated for these three variables.

As previously mentioned, the use of the Random Effects Model relies on the assumption that $\bar{\mu}_c$ is IID with a mean of zero, that is, significant fixed effects do not exist. Thus before we begin to interpret the results of the Random Effects Model we must determine if significant fixed effects do, in fact, exist. The first step in this testing procedure is to construct and estimate a Fixed Effects Model.

The Fixed Effects Model and Testing for Fixed Effects.

The Fixed Effects Model, which is also called a *Least Squares Dummy Variable* (LSDV) model (Green, 1993), is estimated as follows when using pseudo panel data:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\mu}_{c} + \bar{\nu}_{ct} \quad (7)$$

This is identical to Equation 6. However, x_{ct} now includes a set of C cohort dummy variables. This assumes the impact of each cohort contains an estimable component that is fixed across time and these cohort components are significantly different. This is in contrast to the Random Effects Model which assumes these cohort components are IID and are simply included in the $\bar{\mu}_{c}$ error term. Unlike the Random Effects Model, inference from the results of the Fixed Effects Model is limited to the type of cohorts included in the analysis.

The Fixed Effect Model includes the same variables used to represent the characteristics of age, young children, and education as those included in the Random Effects Model. However, the dummy variables used to represent the gender, race, and generation characteristics are fully defined by the 28 cohort dummy variables. Consequently, the variables used to represent these characteristics, which are included in the Random Effects Model, are not specified in the Fixed Effects Model.

As previously mentioned, the preferred Random Effects Model can only be used if there are no significant fixed effects. To test for significant fixed effects the random effects estimation is compared to the fixed effects estimation. Specifically, a test for joint significance of the individual fixed effect dummy variables is calculated as follows (Baltagi, 1995):

$$F_0 = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT - N - K)} \sim F_{N-1,\, N(T-1)-K}$$

In this equation RRSS is the restricted residual sum of squares obtained from the random effects estimation and URSS is the unrestricted residual sum of squares obtained from the fixed effects estimation. N represents the total number of individuals, ($N = C = 28$ cohorts for our analysis), while T is the number of time periods (7 census years for our analysis). K represents the number of independent (non-cohort) variables in the x_{ct} vector of the Fixed Effects Model, (K = 8 for our analysis). If the panel is balanced, $C*T$ will result in the total number of observations used in the regressions.

When the data do not contain information on all cohorts in all time periods, as is the case for our pseudo panel data set, CT overstates the number of observations and the associated degrees of freedom. For example, in our data set, $CT = 28*7 = 196$. However, only 88 observations are actually available due to the life-cycle nature of the data. Consequently, when calculating the F value n for our pseudo panel data set, the NT is equal to 88 and the denominator's degrees of freedom becomes 52 (88 - 28 - 8 = 52).

Data Transformation Models

If significant fixed effects exit, the Random Effects Model cannot be used. One alternative to using the Random Effects Model is to use the Fixed Effects Model. The Fixed Effects Model, however, may result in an undesirable loss of degrees of freedom due to the addition of a large number of cohort dummy variables. In these situations, the Within and the First-Differenced Models, which use transformed data, provide attractive alternative techniques to eliminate fixed effects without a large decrease in degrees of freedom. It should be noted that even if no significant fixed effects are present, the Within Transformation Model and the First-Differenced Model, along with the Between Transformation Model, provide additional insight into the core results of the Random Effects Model.

Researchers should be aware that when the Fixed Effects Model or the Within Transformation Model and First-Differenced Model are used, they do not eliminate bias from unobserved cohort heterogeneity that changes over time. In addition, the data transformations we employed for the Within and First-Differenced Models eliminate all observed as well as unobserved time-constant variables from the regressions. Despite these limitations, transformed panel data can offer a powerful rebuttal to criticisms that conclusions based on observed variables are actually just the result of correlation with unobserved variables.

The Within Transformation Model. The within transformation allows estimation of an equation where the bias from unobserved fixed cohort effects is *swept* from the equation, along with observed fixed effects. The within transformation controls for cohort fixed effects by calculating each variable's mean value across time for each cohort, then subtracting that mean from all observations. With pseudo panel data, this transformation first requires the calculation of each cohort's time mean values using the set of the cohort's mean values found in the data cells. Specifically, time mean values for the equation to be estimated are calculated as follows:

$$\bar{y}_c = \alpha + \beta \bar{x}_c + \mu_c + \bar{v}_c \tag{8}$$

Equation 8 is identical to Equation 6, except the *t* subscript has been eliminated to indicate a mean value across time as well as across cohorts. The μ_c error term represents the unobserved fixed cohort effect and consequently is unchanged between Equations 6 and 8. The within transformation is obtained by subtracting Equation 8 from Equation 6 as follows:

$$y_{ct} - \bar{y}_c = \alpha - \alpha + \beta(x_{ct} - \bar{x}_c) + \mu_c - \mu_c + v_{ct} - \bar{v}_c \tag{9}$$

The intercept term (α), as well as the cohort fixed effect (μ_c), do not change over time. Consequently, they are already time means by definition. If μ_c is assumed to sum to 0 across all cohorts, the within transformation is estimated as follows (Baltagi, 1995):

$$y_{ct} - \bar{y}_c = \beta(x_{ct} - \bar{x}_c) + (v_{ct} - \bar{v}_c) \tag{10}$$

Data for the four race-gender cohorts from the youngest generation were also eliminated from the estimation of the Within Transformation Model. This was necessary because the youngest generation had observations in only the 2000 Census. Consequently, the time mean equaled the actual 2000 Census observations and the within transformation resulted in a set of zero values for all variables. Thus the number of observations for the Within Transformation Model was reduced to 84.

Variables used to represent characteristics that do not change over time (e.g., gender, race, and generation) are not included in the within transformation data set because the transformation of the values contained in these variables caused them to equal zero. The same set of proportional variables contained in the Random and Fixed Effect Models are also included in the Within Transformation Model.

The within transformations of true dummy variables that vary with time (e.g., the dummy variables for the age characteristic) cause a problem with the interpretation of the results produced by the Within Transformation Model. To allow us to interpret the coefficients for such variables the values that are generated by the within transformation are replaced in the transformed data set by their original 0 and 1 values.

If true panel data are used, the residual sum of squares (RSS) for the Within Transformation Model will be identical to the RSS for the Fixed Effects model. This relationship allows researchers to test for significant fixed effects in large data sets (e.g., the PSID) where software limitations on matrix size preclude estimation of a Fixed Effects Model with thousands of dummy variables. Unfortunately, when the Within Transformation Model is used with pseudo panel data the RSS produced by the model is not identical to the Fixed Effects RSS. This inconsistency is caused by the cell-size weighting used in pseudo panel estimations. As a result, when the transformed pseudo panel data are analyzed with the Within Transformation Model its RSS cannot be used to test for fixed effects.

Fortunately, the potential problem of too many dummy variables in the Fixed Effects Model can be addressed with pseudo panel data by deciding how narrowly to define the cohorts. For example, if we had defined generations on a one-year basis rather than a ten-year basis, an unmanageable 280 cohorts, and thus 280 additional dummy variables, would be required. In that case we would not have been able to practically test for the presence of significant fixed effects. Since our pseudo panel used ten-year generations, only 28 additional dummy variables were needed to estimate the Fixed Effects Model.

The Between Transformation Model. We also estimated the Between Transformation Model, which used data transformed by Equation 8. It should be noted that the application of this transformation procedure to the pseudo panel data set produces values for the set of dummy variables used to represent the age characteristic that do not vary. Hence that set of variables cannot be included in the Between Transformation Model. With the exception of the dummy variables used to represent the age characteristic, the Between Transformation Model includes the same set of variables used in the Random Effects Model. Rather than eliminating unobserved fixed cohort heterogeneity, the Between

Transformation Model isolates this heterogeneity and provides useful insights in the interpretation of the results of the Random Effects Model.

The First-Differenced Model. If a hypothesis involving a trend is to be tested and significant fixed effects from unobserved cohort heterogeneity is a concern, a first-differencing transformation can be used to sweep away the fixed effect and retain the trend. Calculating the first differences results in fewer degrees of freedom in the First-Differenced Model than exist in the Random Effects Model. This difference is due to the loss of the oldest observations for all 28 cohorts. As a result, the transformed pseudo panel data set used in conjunction with the First- Differenced Model contains 60 observations rather than 88. First-differenced data are obtained by subtracting each cohort's variable values from the prior year's values as follows:

$$y_{ct} - y_{c,t-1} = \alpha - \alpha + \beta(x_{ct} - x_{c,t-1}) + \mu_c - \mu_c + v_{ct} - v_{c,t-1} \quad (11)$$

The First Differenced model is estimated as follows:

$$y_{ct} - y_{c,t-1} = \beta(x_{ct} - x_{c,t-1}) + v_{ct} - v_{c,t-1} \quad (12)$$

The unobserved cohort fixed effect (μ_c) is removed from the data along with any observed variable that does not change over time (e.g., gender, race, and generation). Calculating the first difference for age characteristic dummy variables resulted in three possible values (i.e., -1, 0 or 1). These age characteristic dummy variables were dropped from the First-Differenced Model because the change in actual mean age for the cohorts between censuses was a constant value of ten. Thus the only variables contained in the First-Differenced Model are the proportional variables representing the young children, marital status, and education characteristics.

Interpretation of the Regression Results

Our analysis of the pseudo panel data set began by estimating the Random Effects and Fixed Effects Models. The results produced for these models are listed in Table 2. Once these models were estimated the following F test was conducted to determine whether the fixed effects were statistically significant:

$$F = \frac{(0.0769 - 0.0548)/27}{(0.0548/52)} = 0.773$$

where: (a) RRSS=.0769, (b) URSS=.0548, (c) N-1=28-1=27, and (d) *NT-N-K* = 88 - 28 – 8 = 52.

The probability value corresponding to the *F* value of .773 (p = .76) indicates the fixed effects were not statistically significant. This result indicates the impacts of the gender, race, and generation characteristics were sufficiently consistent across all 28 cohorts such that controlling for all the interactions of these characteristics in the Fixed Effects Model does not significantly improve the fit of the regression. Thus we are able to use the Random Effects Model as the foundation of the analysis due to our finding of no significant fixed cohort effect.

To assist in assessing the relationship of each characteristic to the dependent variable, the Within Transformation, Between Transformation, and First-Differenced Models were also estimated. The results produced for all five models are contained in Table 2.

An Analysis of the Independent Variables

Gender. In the Random Effects Model the coefficient for the female variable (-0.0022) was not significant at the 0.05 level, suggesting no significant difference between the proportion of two-income spouses was found in male cohorts compared to female cohorts. In addition, the coefficient for the female (-.0120) was not significant at the 0.05 level in the Between Transformation Model.

Additional information regarding the relationship between the gender characteristic and the proportion of two-income spouses is not produced by the Fixed Effects, the Within Transformation and the First-Differenced Models. In the Fixed Effects Model the single dummy variable for gender was not estimated because it was interacted with the race and generation variables to produce the 28 cohort dummy variables. Because gender is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, the impact of gender cannot be estimated in those models.

Race. In the Random Effects Model the coefficient for the African American variable (-0.0431) was not significant at the 0.05 level, suggesting no significant difference between the proportion of two-income spouses found in the African-American cohorts compared to the Caucasian

cohorts. In addition, the coefficient for the African American variable (-0.0583) was not significant at the 0.05 level in the Between Transformation Model. Additional information regarding the relationship between the race characteristic and the proportion of two-income spouses is not produced by the Fixed Effects, the Within Transformation and the First-Differenced Models. In the Fixed Effects Model the single dummy variable representing the race characteristic is not estimated because it was interacted with the gender and generation variables to produce the 28 cohort dummy variables. Because race is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, the impact of race cannot be estimated in those models.

Age. The amount of variation in the dependent variables accounted for by the three dummy variables used to represent the age characteristic in the Random Effects Model was significant at the 0.01 level. To understand the life-cycle work pattern we compared and tested coefficients of adjacent age cohorts. The test results of those comparisons revealed that as people aged there was a significant increase in the proportion of two-income spouses until they reached the age category of Mean Age 60. At that point in time the proportion declined to a level that was not significantly different from the proportion estimated for the Mean Age 30 cohort.

Before we drew any conclusions regarding the relationship between the age characteristics variables and the dependent variable, we compared the results of the Fixed Effects Model to the results of the Random Effects Model to assess the robustness of the relationship. The amount of variation in the dependent variables accounted for by the three dummy variables used to represent the age characteristic in the Fixed Effects Model was also significant at the 0.01 level. Comparisons of the adjacent age cohort coefficients verified the life-cycle work pattern estimated by the Random Effects Model. The statistical tests of those adjacent coefficients, however, were not significant except for the decline in the coefficient values from the Mean Age 50 cohort to the Mean Age 60 cohort.

Because the Within Transformation Model is an alternative method of controlling for fixed effects, we anticipated it would produce similar results. As expected, the amount of variation in the dependent variable accounted for by the three dummy variables used to represent the age characteristic in the Within Transformation Model was also significant at the 0.01 level. Comparisons of the adjacent age cohort coefficients verified the life-cycle work pattern and statistical significance estimated by the Fixed Effects Model.

Additional information regarding the relationship between the age characteristic and the proportion of two-income spouses was not produced by the Between Transformation and the First-Differenced Models because any variance in the age characteristic is eliminated by the transformations. Thus, the impact of age cannot be estimated in those models.

Generation. In the Random Effects Model, the amount of variation in the dependent variable accounted for by the six dummy variables used to represent the generation characteristic in the Random Effects Model was significant at the 0.01 level. Once again the adjacent coefficient of these variables were compared and tested. The tests of adjacent coefficients were significant, which suggests that the proportion of two-income spouses increases with each generation.

The Fixed Effects Model contained 27 of the possible 28 three-way interaction variables created from the various levels of the gender, race, and generation characteristics. The variable not included in the series of linearly dependent variables represented the cohort labeled Caucasian, Male, Born 1906-1915. The amount of variation in the dependent variable accounted for by these 27 dummy variables in the Fixed Effects Model was significant at the 0.01 level. To determine whether the generational work pattern revealed by the Random Effects Model also existed for four race-gender subsets (African-American Female, African-American Male, Caucasian Female, and Caucasian Male), we compared and statistically tested the coefficients of adjacent generations within these subsets. These test results indicated each generation had a higher proportion of two-income spouses than its preceding generation. The tests of adjacent coefficients were significant, except for the two youngest Caucasian, female cohorts. Thus the increasing generational work pattern revealed by the Random Effects Model was also found for each of the four race-gender subsets.

Table 2. Regression Results of the Random Effects, Fixed Effects, Within Transformation, Between Transformation, and First-Differenced Models

	Type of Model				
Independent variables	Random Effects[a]	Fixed Effects[b]	Within Transformation[c]	Between Transformation[d]	First Differenced[e]
Female	-0.0022 (0.0112)	n/a	n/a	-0.0120 (0.0079)	n/a
African-American	-0.0431 (0.0384)	n/a	n/a	-0.0583 (0.0320)	n/a
Mean age 40	0.0599* (0.0269)	0.0180 (0.0332)	0.0186 (0.0353)	n/a	n/a
Mean age 50	0.1076** (0.0423)	0.0527 (0.0502)	0.0326 (0.0561)	n/a	n/a
Mean age 60	0.0117 (0.0457)	-0.0207 (0.0534)	-0.1010 (0.0581)	n/a	n/a
Born 1916-1925	0.0939** (0.0258)	n/a	n/a	0.0809** (0.0168)	n/a
Born 1926-1935	0.1989** (0.0386)	n/a	n/a	0.1708** (0.0271)	n/a
Born 1936-1945	0.3684** (0.0550)	n/a	n/a	0.3283** (0.0319)	n/a
Born 1946-1955	0.5369** (0.0702)	n/a	n/a	0.4783** (0.0368)	n/a
Years of college ≥ 4	0.1707 (0.3303)	1.4238* (0.6658)	0.2385 (0.3827)	-0.2464 (0.3240)	1.3633 (0.8716)
Caucasian	**Male**				
Born 1916-1925	n/a	0.2459** (0.0560)	n/a	n/a	n/a
Born 1926-1935	n/a	0.3883** (0.0858)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6449** (0.1244)	n/a	n/a	n/a
Born 1946-1955	n/a	0.8108** (0.1494)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0027** (0.1473)	n/a	n/a	n/a
Born 1966-1975	n/a	1.0945** (0.1586)	n/a	n/a	n/a
African	**American**	**Male**			
Born 1906-1915	n/a	-0.2811** (0.1047)	n/a	n/a	n/a
Born 1916-1925	n/a	-0.0679 (0.0720)	n/a	n/a	n/a
Born 1926-1935	n/a	0.2086** (0.0513)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6149** (0.0831)	n/a	n/a	n/a
Born 1946-1955	n/a	0.8923** -0.1223	n/a	n/a	n/a
Born 1956-1965	n/a	1.1036** (0.1445)	n/a	n/a	n/a
Born 1966-1975	n/a	1.2912** (0.1621)	n/a	n/a	n/a

Table 2 (Continued).

Type of Model					
Independent variables	Random Effects[a]	Fixed Effects[b]	Within Transformation[c]	Between Transformation[d]	First Differenced[e]
			Coefficient		
Caucasian	**Female**				
Born 1906-1915	n/a	0.0944** (0.0338)	n/a	n/a	n/a
Born 1916-1925	n/a	0.3896** (0.0731)	n/a	n/a	n/a
Born 1926-1935	n/a	0.5656** (0.0977)	n/a	n/a	n/a
Born 1936-1945	n/a	0.7635** (0.1213)	n/a	n/a	n/a
Born 1946-1955	n/a	0.9081** (0.1435)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0138** (0.1486)	n/a	n/a	n/a
Born 1966-1975	n/a	1.0597** (0.1658)	n/a	n/a	n/a
African	**American**	**Female**			
Born 1906-1915	n/a	-0.2707** (0.1051)	n/a	n/a	n/a
Born 1916-1925	n/a	-0.0428 (0.0731)	n/a	n/a	n/a
Born 1926-1935	n/a	0.2465** (0.0597)	n/a	n/a	n/a
Born 1936-1945	n/a	0.6207** (0.0909)	n/a	n/a	n/a
Born 1946-1955	n/a	0.9140** (0.1271)	n/a	n/a	n/a
Born 1956-1965	n/a	1.0842** (0.1436)	n/a	n/a	n/a
Born 1966-1975	n/a	1.2552** (0.1550)	n/a	n/a	n/a
Constant	-0.0177 (0.1076)	-0.9773** (0.2549)	0.2627** (0.0507)	0.0783 (0.0758)	-0.0195 (0.0207)

*significant at the 5% level
**significant at the 1% level
[a]Random Effects Model: N=88; F(16,71)=157.1**; R^2=.97; Root MSE=.033; RSS=.0769
[b]Fixed Effects Model: N=88; R^2=.98; F(35,52)=74.5**; Root MSE=.033; RSS=.0548
[c]Within Transformation Model: N=84; R^2=.8237; F(8,75)=43.8**; Root MSE=.047;RSS=.1665
[d]Between Transformation Model: N=28; R^2=.99; F(13,14)=521.0**; Root MSE=.011; RSS=.0016
[e]First-Differenced Model: N=60; R^2=.7331; F(5,54)=29.7**; Root MSE=.0610; RSS=.2012
[f]The standard errors are enclosed in parentheses

In the Between Transformation Model, the amount of unique variation in the dependent variable accounted for by the six dummy variables used to represent the generation characteristic was significant at the 0.01 level. The tests of adjacent coefficients found the same pattern of a significant increase in the proportion of two-income spouses with each generation that was found in the Random Effects and Fixed Effects Models, further supporting the robustness of the generational pattern.

Since generation is a time-constant variable, it was eliminated from the Within Transformation and First-Differenced Models. Thus, these models could not provide additional information regarding the relationship between the generation characteristic and the proportion of two-income spouses.

Young children. In the Random Effects Model the coefficient for the young child present variable (-0.2873) was significant at the 0.01 level. Because this is a proportional variable, interpreting the coefficient for a one-unit change in the variable is the method of interpretation. A more realistic interpretation is that a 0.10 increase in the proportion of a cohort with at least one young child present was associated with a 0.02873 decrease in the proportion of two-income spouses. The Fixed Effects Model confirmed this same, significant relationship.

In the Within Transformation Model the coefficient for the young child present variable (-0.2778) was also significant. This suggests that over the course of a cohort's life-cycle, the timing of when couples choose to have children significantly impacts changes in the proportion of two-income spouses within that cohort. The coefficient for the young child present variable was not significant in the Between Transformation model at the 0.05 level. This suggests that cohorts which *average* a higher proportion of individuals with young children present do not have a significantly lower proportion of two-income spouses. Finally, in the First-Differenced Model, the coefficient for the young child present variable (-0.6781) was significant at the 0.01 level. This provided further confirmation of the inverse relationship between the proportion of two-income spouses and the proportion of individuals with young children that was found in the Random Effects, Fixed Effects, and Within Transformation Models.

Marital status. To control for changes in a cohort's proportion of married individuals, all five models included a married variable that measured the proportion of a cohort that was married. This variable was not significant at a 0.05 level in any of the models.

Education. In all five models, the amount of unique variation in the dependent variable accounted for by the three dummy variables used to represent the education characteristic was significant at the 0.01 level. The coefficient for the less than HS variable in the Random Effects Model (0.4774) was significant at $\alpha = 0.01$. Apparently, individuals with the least amount of education are more likely to be two-income spouses relative to individuals with a high school education. Assuming education and income are correlated, this finding may be due to greater pressure for both spouses to work if income is low. The coefficient for less than HS variable was also significant and positive in the Fixed Effects and Between Transformation Models, which supports the pattern suggested by the Random Effects Model. The coefficient in the Within Transformation Model was also positive, but it was not significant at $\alpha = 0.05$.

The coefficient for less than HS variable in the First-Differenced Model (3.0199) was significant at the 0.01 level. Because our data set excludes individuals younger than age 25, this indicates that a significant number of individuals completed their high school education after age 25. It should be noted that this coefficient has the largest of any of the proportional variables used to represent the educational characteristics. This coefficient indicated that a 0.10 increase between censuses in the proportion of a cohort with at least a high school education was associated with a 0.30199 decrease in the proportion of two-income spouses. We compared the coefficients for each level of education beyond high school with the coefficient for the adjacent level of education and found no significant difference in the coefficients for any of the five models at the 0.05 level. This suggests that the incentive for both spouses to work to earn the higher salary available with increased education is approximately offset by the decreased need for both spouses to work as higher education allows one spouse to contribute more income to the household.

Summary

In an attempt to identify characteristics that are related to the increase in the proportion of two-income spouses we constructed a pseudo panel data set from the Decennial United States Census collected from 1940 to 2000. A comparison between the Random Effects Model and the Fixed Effects Model revealed that the fixed effects were not statistically significant. Consequently, the Random Effects

Model formed the core of our analysis. Additional insights regarding the relationships of the various characteristics and the proportion of two-income spouses were provided by the results produced by the Within Transformation, Between Transformation, and First-Differenced Models, which used transformed data sets. The Random Effects Model revealed a significant life-cycle pattern of increasing probability of two-income spouses as the cohorts age. The Random Effects, the Fixed Effects, and the Between Transformation Models all showed a significant increase across generation cohorts in the proportion of two-income spouses.

Our findings consistently support the hypothesis that within a cohort, the presence of young children reduces the probability of two-income spouses. Moving from less than a high school education to a high school level of education was significantly associated with a decrease in the proportion of two-income spouses, possibly due to decreased pressure to work as incomes increased with education. The insignificant coefficients for education levels beyond high school appear to reflect offsetting incentives for spouses to work to take advantage of higher potential income and the reduced need to work if household income is higher.

We have sought to provide a reference or guide to researchers who encounter questions that can be addressed with the use of panel data, yet find no true panel data set is available. This paper demonstrated how pseudo panel data can be constructed to address the lack of a true panel data set with special attention given to some of the nuances inherent in its construction. We also described and illustrated how the Random Effects, Fixed Effects, Within Transformation, Between Transformation, and First-Differenced Models were constructed and interpreted when applied to a pseudo panel data set. It is our hope this article will encourage researchers to investigate questions that may have been left unanswered due to a lack of panel data.

References

Baltagi, B. H. (1995). *Econometric analysis of panel data.* New York: John Wiley and Sons, Inc.

Ben-Porath, Y. (1973). Labor force participation rates and the supply of labor. *Journal of Political Economy*, 81, 697-704.

Coleman, M. T. & Pencavel, J. (1993). Changes in work hours of male employees, 1940-1988. *Industrial and Labor Relations Review, 46*(1), 262-283.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, *30*, 109-126.

Fienberg, S. E. & Mason, W. M. (1985). Specification and implementation of age, period, and cohort models. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem.* New York: Springer-Verlag.

Greene, W. H. (1993). *Econometric analysis.* New Jersey: Prentice-Hall, Inc.

Pencavel, J. (1998). The market work behavior and wages of women. *The Journal of Human Resources*, *33* (4), 771-804.

Rodgers, W. L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review*, *47*, 774-787.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C. A., Goeken, R., Hall, P.K., King, M., & Ronnander, C. (2004). *Integrated public use microdata series: Version 3.0* [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor]. http://www.ipums.org

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30*, 843-861.

Smith, H. L., Mason, W. M. & Fienberg, S. E. (1982). More chimeras of the age-period-cohort accounting framework: Comment on Rodgers. *American Sociological Review, 47*, 787-793.

Send correspondence to: Jeffrey E. Russell, Ph.D.
Ashland University
Email: jrussell@ashland.edu