Model Specification and the Robustness of Empirical Results from Migration Models

by

Brian Cushing

RESEARCH PAPER 9605

Brian Cushing
Department of Economics and
Regional Research Institute
West Virginia University

Paper presented at the Southern Regional Science Association Conference,
Baltimore, MD, April 11-13, 1996

ABSTRACT: A characteristic of the empirical literature on internal population migration is widely varying results and often conflicting conclusions regarding relative importance of explanatory factors. There are a number of possible explanations for these conflicting findings, some of which have been noted in the literature. Model specification issues have received relatively little detailed attention. This paper focuses on three key issues: type of migration model, functional form, and consideration of spatial variables. First, I examine the extent to which these three considerations influence empirical estimates. Then I investigate how these specification issues affect the robustness of empirical results in the presence of other specification problems, focusing on omitted variable bias.

ACKNOWLEDGMENT: Besides my department, I wish to acknowledge the research support of the Regional Research Institute at West Virginia University and the conference support of the College of Business and Economics.

I. Introduction

One characteristic of the vast empirical literature on internal population migration is the diversity of conclusions regarding the statistical significance and relative importance of most explanatory variables that have been considered. For any given explanatory factor, some models are likely to have a variable with a statistically significant coefficient and the expected sign, while others will yield either a statistically insignificant estimated coefficient, or even a statistically significant coefficient with an unexpected sign. Even those models that have a statistically significant coefficient with the expected sign vary widely with respect to estimates of the variable's relative importance, as indicated by the magnitude of estimated elasticities or standardized beta coefficients. Sometimes the disparities within the literature have become the focus of the literature itself, such as with the literature on the relative importance of "economic" versus "amenity" variables, the literature on the effect of welfare benefits, and some literature attempting to explain the positive estimated coefficient of the destination unemployment rate on inmigration.

There are a number of obvious reasons for the conflicting findings in the empirical literature. First, researchers often use different series of migration data. Data from the decennial Census of Population and Housing, the Panel Survey of Income Dynamics, the Internal Revenue Service, the National Longitudinal Survey, and other sources are likely to yield somewhat different results due to factors such as different coverage of the population and different lengths of the migration period (e.g., one-year versus five-year). Related to differences in data series is the selected geographical unit of analysis for migration, i.e., region, state, metropolitan, or county. Often the selection of the data source determines this. A third source of differences in results is the selection of the specific point in time of the analysis, such as 1955-60 versus 1965-70 versus 1975-80. Relationships, such as the willingness to trade-off between economic factors and amenities, may change over time. The level of analysis (individual versus aggregate) also can greatly affect results. Increasingly, studies of U.S. migration are relying on microdata with details on individuals. Even after many years, however, there is still uncertainty regarding the most effective way to integrate contextual (place) variables with individual characteristics, and how best to present and interpret the coefficients of these models. Finally, the empirical

literature varies widely with respect to the number and types of explanatory variables included in analyses. Many empirical migration models have included only a small number of explanatory variables, especially the models of aggregate migration flows. Omitted variable bias has the potential to substantially alter empirical results and conclusions.

Even controlling for all of the above, much diversity in empirical results is apt to persist on account of some less obvious, but important, model specification issues. One such issue is functional form. Most models of aggregate migration use either the linear or double-log functional form. These are generally chosen out of convenience (ease of interpretation of coefficients), without much thought given to what is appropriate based on theory. Choice of functional form almost always affects estimates of relative importance, such as elasticities, and often affects the sign and statistical significance of some estimated coefficients. In practice, a mixed functional form model is most appropriate, with the functional relationship of each explanatory variable vis-à-vis the dependent variable consistent with theory. Even though they do not fully explore the implications of alternative functional forms, Goss and Chang (1983) raise the issue of functional form in migration models and begin to demonstrate how it can influence empirical results.

A second important specification issue for migration models is inclusion of spatial variables. Whether microdata or aggregate models, most migration models are devoid of spatial variables. For models using metropolitan areas or counties as the unit of analysis, omission of spatial variables is a practical matter. Modeling spatial relationships would be quite cumbersome. State-level models are more likely to account for space, but, as Cushing (1986) notes, the specification of spatial relationships in these models is generally rudimentary.

The type of migration model used influences choice of functional form and inclusion of spatial relationships. For example, consider a model of aggregate interstate migration as indicated by the rate of migration. The dependent variable could focus only on aggregate flows, such as the aggregate rate of inmigration, outmigration, or net migration. Alternatively, the model could focus on place-to-place migration flows, using either the rate of inmigration to each destination state from each origin state or an allocation rate, i.e., the proportion of migrants from each origin state who select each of the destination states. Each of these choices of dependent variable could also be expressed in terms of number of migrants rather than a migration rate.

Theoretically, different specifications of the dependent variable imply different functional forms. For example, a model of net migration is not consistent with a double-log form since the dependent variable can have negative values. The choice of dependent variable will also influence the specification of space in the model. By nature, the aggregate flow models have much less information on spatial relationships than the place-to-place models.

The remainder of this paper focuses on this group of "less obvious" specification issues in the context of an interstate migration model based on aggregate migration data. First I investigate the extent to which type of model, functional form, and inclusion of spatial variables influences empirical estimates and conclusions. My working hypothesis is that all of these can substantially alter the conclusions drawn from a model. Next, I investigate how these specification issues affect the robustness of empirical results in the presence of other specification problems, focusing especially on omitted variable bias, but also considering how type of model, functional form, and inclusion of spatial variables interact to increase or decrease potential specification biases.

II. The Data and Empirical Model

For illustrative purposes, I use relatively simple models and choose data to fit in the context of the largest number of migration studies. The models are of aggregate interstate migration flows and use 1965-70 migration data from the *1970 Census of Population*. I employ two types of models and three different functional forms. One model type is of place-to-place migration flows with an allocation rate as the dependent variable. The other is a model of aggregate migration flows, with the inmigration rate as the dependent variable. I estimate both models for linear, double-log, and mixed functional forms. In the mixed functional form model, the functional form of specific variables is determined primarily based on theory, considering the numerical range of each variable, the behavior of the model as the value of the variable approaches extremes such as zero or infinity, properties of the slope, and properties of the elasticities.[1] Box-Cox/Box-Tidwell analysis has provided some empirical support for the

[1] Cushing (1988) has a detailed discussion of selection of functional form in migration models.

functional forms selected in the mixed functional form model.[2] The six estimated equations (two model types and three functional forms) are the basis for investigating robustness of results with respect to model type and functional form. The study checks the two types of models and three functional forms for robustness to omitted variable bias by dropping the climate variable from each of the models. Finally, the analysis checks the place-to-place flow models for robustness with respect to inclusion or exclusion of spatial variables by adding two spatial variables to each of the models.

The dependent variable used for the inmigration rate model (INRATE) is the number of inmigrants to state j during the five-year migration period as a percentage of the 1965 base population of state j, where the base population is the 1970 population aged five years and over for state j, plus the number of outmigrants less the number of inmigrants. The allocation rate model is less typical, but is similar in nature to models by Cushing (1989), Greenwood (1969), Goss and Chang (1983), Kau and Sirmans (1976), and Wadycki (1974). The dependent variable is an allocation rate of migration (ALLRATE): the percentage of all outmigrants from state i who chose state j as the destination. The allocation rate is easily derived from a gravity-type model of migration. Because it focuses on those who migrated without regard for those who did not migrate, origin characteristics need not be considered [see Cushing (1989)]. Since both models consider aggregate (as opposed to individual) migration, they rely on the assumption that areas with characteristics that are generally associated with higher levels of welfare disproportionately attract migrants. Such characteristics include better economic opportunities and better amenities. High costs of migration may mute this attraction.

For simplicity, the econometric models include only a small number of explanatory variables. Both types of models include the following:

JOBGROW — percentage change in nonagricultural employment (by place of work) in the destination state between 1960 and 1968 -- Employment and Earnings, States and Areas, 1939-78.

[2] See Box and Cox (1964) and Box and Tidwell (1961).

INCOME	1965 per capita personal income for the destination state -- State Personal Income, 1929-82.
UNEMPLOY	mean annual average unemployment rate for the destination state, 1964-68 -- Manpower Report of the President, 1973.
POPULATION	1965 base population of the destination state -- computed based on U.S. Census of Population, 1970.
MILITARY	percent of the population of the destination state, aged 16 and over, that was employed in the armed forces, 1970 -- U.S. Census of Population, 1970.
TEMPJAN	mean temperature of the destination state during the month of January (degrees Fahrenheit) -- County and City Databook, 1977.[3]

In addition, some of the allocation rate models include two spatial variables:

DISTANCE	highway mileage between the principal city of origin state i and that of destination state j - Official Table of Distances, 1979.
DSTATE	unity if the origin and destination states are adjacent, equals zero otherwise.

Destinations with more rapid employment growth (JOBGROW), higher incomes (INCOME), and lower unemployment rates (UNEMPLOY) should attract migrants due to greater expectations of economic well-being. Larger populations (POPULATION) provide more and better economic, social, and cultural opportunities, ceteris paribus. In addition, large populations increase information flows regarding a potential destination, thus reducing migration costs. Large populations, however, may also indicate more disamenities such as congestion and pollution. Therefore larger populations may be associated with either greater or less inmigration depending on which effects dominate. Since the rate of mobility is higher among those in the military (including those entering or leaving the military), states with relatively greater military employment (MILITARY) should experience greater inmigration, ceteris paribus. Many studies have shown that, on average, cold climates are unattractive for migrants. Thus, cold January

[3] Computed as a weighted average of all metropolitan areas with a population exceeding 100,000 for which data was available. For states with no metropolitan areas in this size range, an average of the principal cities was used.

temperature (TEMPJAN) should be associated with less inmigration. I omit TEMPJAN from some models to check for robustness in the face of omitted variable bias.

Some of the allocation rate models, which use place-to-place migration flows, include two variables to capture the spatial relationship between the origin and destination states. Greater spatial distance (DISTANCE) between the origin and destination should yield less migration. A simple measure of distance, however, inadequately captures spatial relationships when large areal units, such as states, are used. As in Cushing (1986), the allocation rate models include a dummy variable for adjacent states (DSTATE), with adjacent states expected to have relatively large migration flows, all else equal.

The econometric estimation uses migration flows for the lower 48 states. For the inmigration rate models, this means that the sample size is 48. The observations for the allocation rate consider migration from each of the lower 48 states to each of the other 47 states, thus yielding a sample size of 2,256. Given the choice of the explanatory variables, including the time period covered, simultaneity bias should be minimal, so the estimation employs ordinary least squares (OLS). In the mixed functional form specifications, the dependent variable is in log form, as are POPULATION, MILITARY, and DISTANCE. INCOME is in reciprocal form. All other variables are in linear form, i.e., untransformed.

III. Empirical Results

A. General Comparison and Robustness across Functional Forms

Table 1 shows the empirical results for the basic inmigration rate and allocation rate models, for all three functional forms. For comparability, the allocation rate models exclude the spatial variables, initially. The numbers in the table are elasticities at the mean.[4]

With the exception of UNEMPLOY, all estimated coefficients for the allocation rate models are statistically significant at the one percent level. In the double-log and linear models, the coefficients of UNEMPLOY are only significant at the five percent and ten percent level, respectively. The elasticities appear to be fairly robust to differences in functional form,

[4]For the double-log functional form, the estimated coefficients of the equation are constant elasticities.

Table 1: Comparison of Allocation Rate and Inmigration Rate Models by Functional Form

	Allocation Rate			Inmigration Rate		
Explanatory Variable	Mixed	Double-Log	Linear	Mixed	Double-Log	Linear
JOBGROW	**0.254**	**0.242**	**0.232**	**0.261**	**0.286**	**0.355**
INCOME	**0.817**	**0.705**	**0.520**	**0.835**	**0.742**	**1.177**
UNEMPLOY	**-0.259**	***-0.222***	*-0.231*	***-0.248***	*-0.203*	0.020
POPULATION	**0.807**	**0.817**	**0.687**	**-0.310**	**-0.305**	**-0.262**
MILITARY	**0.152**	**0.175**	**0.166**	*0.050*	***0.068***	*0.079*
TEMPJAN	**0.422**	**0.268**	**0.524**	**0.557**	**0.398**	***0.398***
R-squared	0.49	0.49	0.25	0.81	0.78	0.65

Bold equals significant at the one percent level

Bold/Italic equals significant at the five percent level

Italics equals significant at the ten percent level

especially between the mixed and double-log forms, both of which use the log form of the dependent variable. The temperature variable is the most sensitive to functional form.

Several coefficients have a lower level of statistical significance in the inmigration rate models compared with the allocation rate models, with the coefficient of UNEMPLOY insignificant in the linear model. Given the much smaller sample size for the inmigration rate models, this is not a surprising result. With a more complete model specification (additional potentially relevant explanatory variables), these differences in statistical significance would likely be substantial, which confers a distinct advantage to the place-to-place models due to the larger sample size. Similar to the allocation rate models, the elasticities are reasonably consistent between the mixed and double-log forms of the inmigration rate model, with the exception of the climate variable. Compared with the allocation rate models, the linear form of the inmigration rate model exhibits much wider swings in elasticities relative to the other functional forms.

With the exception of the population variable, the elasticities for the mixed and double-log allocation rate models are remarkably close to their counterparts for the inmigration rate models. The basic conclusions regarding relative importance of explanatory variables would be the same. This is not the case for the linear models which lead to some diverse conclusions, primarily due to the greater instability of the elasticities in the inmigration rate model.

The POPULATION variable provides an interesting illustration of the need for caution when interpreting results from different formulations of a model. The allocation rate models suggest that larger population has a net positive (attractive) effect, while the inmigration rate models suggest just the opposite. Clearly, the general formulation of the empirical model is behind the difference in results, since these are otherwise identical models (same explanatory variables). In this case, the negative coefficients of the inmigration models reflect the disamenity effect of larger populations, which suggests *relatively* less inmigration to larger places. The positive coefficients in the allocation rate models reflect a proportionality effect -- even if there is *relatively less* migration to large places, there is still *absolutely more* migration to them so that a greater proportion of a state's outmigrants are still liable to end up in larger places.[5]

Finally, a comparison of the R-squared statistics for the allocation rate and inmigration rate models also provides a useful illustration, this time regarding the need for caution in using the R-squared statistic. For example, in comparing the mixed functional form models, many people might prefer the inmigration rate model on the basis of its seemingly stronger explanatory power. Yet, the estimated elasticities are generally very close in these models -- nearly identical for several of the variables. A reasonable interpretation is that this small model (few explanatory variables) is able to explain very aggregated flows more completely than it can detailed place-to-place flows. Another way of stating this is that if you do not want to know very much, a small, simple model may suffice, but detailed information about migration requires a more comprehensive model for completeness.

[5] Cushing (1989) gives other examples of misinterpretation of the allocation rate model.

B. Robustness to Omitted Variables

Empirical estimates for the same six models, but excluding the temperature variable, are in Table 2. Omitting the temperature variable has a noticeable effect on the elasticities of all variables of all six equations. In terms of model conclusions, the most important changes resulting from the omitted variable are the statistical insignificance of the unemployment rate in all equations and the small magnitude and statistical insignificance of the income variable in the linear form of the allocation rate model. In terms of proportional changes in elasticities, the mixed form and double-log form of the allocation rate model are more robust to the omitted variable bias than their counterparts of the inmigration rate model, although many of the elasticities remain reasonably close to each other. The linear form of the allocation rate model is less stable in the presence of the omitted variable than the linear form of the inmigration rate model. By far, it is the most unstable equation of the set with respect to elasticities and the conclusions that would be drawn from the elasticities. This may simply mean that in the presence of a couple of serious econometric problems, i.e., a very inappropriate functional form

Table 2: Allocation Rate and Inmigration Rate Models Omitting the Climate Variable

	Allocation Rate			Inmigration Rate		
Explanatory Variable	Mixed	Double-Log	Linear	Mixed	Double-Log	Linear
JOBGROW	**0.396**	**0.307**	**0.449**	**0.448**	**0.382**	**0.520**
INCOME	**0.580**	**0.582**	0.115	**0.522**	**0.559**	**0.869**
UNEMPLOY	-0.116	-0.105	-0.093	-0.060	-0.030	0.124
POPULATION	**0.872**	**0.861**	**0.762**	**-0.224**	**-0.240**	**-0.206**
MILITARY	**0.199**	**0.210**	**0.220**	**0.112**	**0.120**	***0.120***
R-squared	0.49	0.48	0.25	0.73	0.69	0.61

Bold equals significant at the one percent level

Bold/Italic equals significant at the five percent level

Italics equals significant at the ten percent level

and omitted variable bias, even the increased information in the place-to-place model cannot salvage the empirical results, in which case choice of model type may not really matter. Notably, the R-squared statistic indicates that the goodness of fit of the linear forms of the allocation rate model is very low compared with all other models, which is consistent with its instability. Interestingly, the omitted variable has virtually no effect on the R-squared statistics of all three forms of the allocation rate model, unlike the inmigration rate models.

C. Influence of Spatial Variables

Table 3 presents estimated elasticities for the complete versions (includes the temperature variable) of the three allocation rate models, with and without the two spatial variables, DISTANCE and DSTATE. The coefficient of the DSTATE variable is the proportional change in the allocation rate when the variable takes a value of unity (adjacent state). On average, the

Table 3: The Allocation Rate Model with and Without Spatial Variables

	Nonspatial			Spatial		
Explanatory Variable	Mixed	Double-Log	Linear	Mixed	Double-Log	Linear
JOBGROW	**0.254**	**0.242**	**0.232**	**0.317**	**0.349**	**0.260**
INCOME	**0.817**	**0.705**	**0.520**	**1.146**	**0.977**	**0.928**
UNEMPLOY	**-0.259**	***-0.222***	*-0.231*	-0.002	0.071	0.026
POPULATION	**0.807**	**0.817**	**0.687**	**0.746**	**0.766**	**0.660**
MILITARY	**0.152**	**0.175**	**0.166**	**0.168**	**0.196**	**0.202**
TEMPJAN	**0.422**	**0.268**	**0.524**	**0.549**	**0.296**	**0.605**
DISTANCE				**-0.577**	**-0.560**	**-0.406**
DSTATE				**1.334**	**1.355**	**2.768**
R-squared	0.49	0.49	0.25	0.77	0.77	0.57

Bold equals significant at the one percent level

Bold/Italic equals significant at the five percent level

Italics equals significant at the ten percent level

allocation rate more than doubles when states are adjacent, all else equal. In all three models, inclusion of the spatial variables increases the R-squared statistic substantially, and noticeably changes all elasticities, increasing most of them. The most dramatic change is the loss of explanatory power of the unemployment rate. With the spatial variables included, the allocation rate models yield results that are noticeably different from those of the inmigration rate models (Table 1).

Like the nonspatial versions of the allocation rate model, the mixed and double-log forms of the spatial versions yield nearly identical empirical results, with the exception of the climate variable. The estimated elasticities are somewhat different for the linear model, but, for the model with spatial variables, the results are qualitatively similar to those of the mixed and double-log forms. The one glaring difference across functional forms is the effect of being adjacent, which is more than twice as large in the linear model than in the other two forms.

Table 4 shows empirical results for the two sets of allocation rate models with temperature excluded. Comparing these to the results in Table 3 reveals that, for the most part, the spatial and nonspatial versions of the allocation rate models are equally robust to the omitted variable bias. The one notable exception is the income variable in the linear form. With spatial variables in the model, omitting TEMPJAN substantially decreases the elasticity of income, but does not completely take away its explanatory power. Since the spatial variables, DISTANCE and DSTATE, are defined for origin-destination combinations, they are highly robust to omission of a destination characteristic.

IV. Conclusions and Implications

This paper has focused on sources of instability in empirical estimates of internal population migration models, paying particular attention to three specification issues: type of model, functional form, and inclusion of spatial variables. Since there is potential for serious omitted variable bias in most migration models, the analysis considers how these three specification problems interact to increase or decrease the bias from an omitted variable. While I consider these results to be very rough and preliminary, they suggest some important

Table 4: Nonspatial and Spatial Allocation Rate, Omitting Climate

	Nonspatial			Spatial		
Explanatory Variable	Mixed	Double-Log	Linear	Mixed	Double-Log	Linear
JOBGROW	**0.396**	**0.307**	**0.449**	**0.500**	**0.419**	**0.508**
INCOME	**0.580**	**0.582**	0.115	**0.833**	**0.840**	**0.453**
UNEMPLOY	-0.116	-0.105	-0.093	*0.178*	*0.198*	*0.179*
POPULATION	**0.872**	**0.861**	**0.762**	**0.832**	**0.814**	**0.745**
MILITARY	**0.199**	**0.210**	**0.220**	**0.229**	**0.235**	**0.264**
DISTANCE				**-0.565**	**-0.555**	**-0.394**
DSTATE				**1.346**	**1.363**	**2.773**
R-squared	0.49	0.48	0.25	0.77	0.77	0.57

Bold equals significant at the one percent level

Bold/Italic equals significant at the five percent level

Italics equals significant at the ten percent level

implications for specification of migration models.

Functional form, type of model structure, and inclusion of spatial detail all influence the empirical estimates from migration models. While the exact choice of functional form always makes some quantitative difference, choice of the proper functional form for the dependent variable makes the greatest qualitative difference (i.e., effect on model conclusions). Proper form for the dependent variable also appears to improve robustness of model results in the face of omitted variable bias. Given a reasonable choice of functional form (mixed or even double-log in this case), place-to-place models, with detailed information on migration flows appear to be somewhat more robust to specification errors than aggregate flow models. With the exception of the interpretation of the population variable, however, the qualitative results of the two types of models are very close, given proper functional form.

The great advantage of place-to-place models comes with inclusion of spatial variables. The spatial variables make an appreciable quantitative and qualitative difference in the empirical

results, yield greater robustness to inappropriate functional form, and at least marginally improve robustness to omitted variable bias in the preceding analysis. The spatial variables themselves are highly significant and greatly increase the overall explanatory power of the allocation rate model.

On a theoretical level, migration models with a spatial dimension are clearly superior to nonspatial models. Probably due to simplicity and ease of use, nonspatial aggregate flow models have dominated the empirical migration research. The analysis here suggests that the results from these nonspatial aggregate flow models all suffer from potentially important omitted variable bias. The results of the aggregate flow inmigration rate models are generally comparable to those from the nonspatial (incomplete) versions of the allocation rate models. The elasticities are nearly identical, despite the deceptively different R-squared statistics (at least for the mixed functional form and double-log versions). In this paper, the qualitative differences in results between the spatial and nonspatial models are small other than the obvious importance of the spatial variables themselves and the change in importance of the unemployment rate variable. Further analysis, with more complete models, will determine if there are generally large qualitative differences in empirical results between spatial and nonspatial migration models, thus invalidating results from the latter. Such a finding would have very serious implications since most migration models continue to be developed as nonspatial models, mostly focusing on intermetropolitan or intercounty migration. Adding spatial variables to these models may be cumbersome but may be a necessity to claim any validity for conclusions drawn from the empirical work.

References

Box, G. and D. Cox. (1964). An Analysis of Transformations, *Journal of the Royal Statistical Society*, Series B 26, 211-243.

Box, G. and P. Tidwell. (1961). Transformation of the Independent Variables, *Technometrics* 4, 531-550.

Cushing, Brian. (1986). Accounting for Spatial Relationships in Models of Interstate Population Migration, *Annals of Regional Science* 20, #2, 66-73.

Cushing, Brian. (1988). Analysis of Functional Form in a Model of Population Migration, mimeo.

Cushing, Brian. (1989). Use and Misuse of the Allocation Rate in Models of Population Migration, *Annals of Regional Science* 23, 51-58.

Ernest Goss and H. Chang. (1983). Changes in Elasticities of Interstate Migration: Implication of Alternative Functional Forms, *Journal of Regional Science*, 23, 223-232.

Michael Greenwood. (1969). An Analysis of the Determinants of Geographic Labor Mobility in the United States, *Review of Economics and Statistics*, 51, 189-194.

James Kau and C.F. Sirmans. (1976). New, Repeat, and Return Migration: A Study of Migrant Types, *Southern Economic Journal*, 1144-1148.

Walter Wadycki. (1974). Alternative Opportunities and Interstate Migration: Some Additional Results, *Review of Economics and Statistics*, 6, 254-257.