



universität
wien

DISSERTATION

Titel der Dissertation

„A statistical model of annual migration counts for rural municipalities“

Verfasser

Mag. Johannes Klotz, Bakk.

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2014

Studienkennzahl lt. Studienblatt:

A 094 136

Dissertationsgebiet lt. Studienblatt:

Statistik und Operations Research

Betreuer:

Univ.-Prof. Dr. Wilfried Grossmann

A statistical model of annual migration counts for rural municipalities

Johannes Klotz

Preface

When I finished my master thesis in Statistics at the University of Vienna in April 2008, it was my intention to quit with statistical research for some years and concentrate on my job as a demographer at Statistics Austria (the Austrian national statistical institute), as well as on my private life. However, my absence from statistical research did not last long. In late 2008, I received a phone call from a professor of demography in Rostock, Germany, who asked me about my intentions to continue my studies. Following some discussions with statisticians and demographers in Austria, I enrolled in the Abraham Wald PhD program of statistics and operations research at the University of Vienna in the summer term of 2009.

At that time I had only a vague idea of what precisely would be the value added of my PhD thesis, although the general topic was very clear from the beginning. For some years, I had already tried to promote stochastic approaches in demography in situations where the data was not sample data in a genuine sense. A problem with that was that it was often unclear which statistical models would be appropriate for demographic outcomes, even for basic statistics like migration counts. It also seemed to me that although several demographers, both in Austria and abroad, regularly applied stochastic ideas in their work, they were often unaware of their theoretical foundations, by which I mean substantive reasons to apply statistical approaches in general as well as the choice of particular models. I hope that my thesis sheds some light on these two issues which I consider very basic for any statistical application.

My original idea was to develop a conditional variance formula that would be applicable for period outcomes of any demographic process. Digging deeper and deeper in the existing literature, I discovered that some of my thoughts had already been developed by others. Moreover, I realized early on that the existing literature on statistical models for demographic data is effectively divided into a vital events strand and a migration strand. In my opinion this is an unfortunate division because the basic challenges in modeling demographic outcomes are the same in both strands. Whereas the vital statistics strand seems to have made greater progress in handling simultaneity between demographic counts and person-years lived, the migration strand seems to have made greater progress in regression model specification and empirical evaluation of different models. Clearly, both strands would benefit from a more unified view.

During the years of my study, I discovered ever more pitfalls and caveats which would have made it more and more difficult to develop a general conditional variance formula, so I decided to narrow down the topic. Compared to my initial intentions, the following four major changes have taken place: (1) The model is now exclusively designed for gross migration flows (in-migration and out-migration, respectively), although many ideas can readily be applied to vital events as well. (2) The approach is now more directed towards regression models,

meaning that explanatory variables are available, although non-regression models can of course easily be formulated as special instances of regression models. (3) I have limited the scope of applications to rural areas, meaning that the investigation area is not dominated by a large single city. This decision was mainly driven by statistical considerations. (4) I have devoted an entire chapter of my thesis to simultaneity between migration count and person-years lived and developed an (instrumental-variable) method to account for it, which to the best of my knowledge has not been used so far by demographers.

In the last chapter of my thesis, I mention some possible extensions of my model which could be starting points for further research. An obvious extension would be a bivariate model of in-migration and out-migration. From a statistical viewpoint, this would allow for more efficient estimation and cross-equation hypotheses testing. From a demographic point of view, it would allow for calculation of the conditional variance of net migration. Another promising extension of my approach would be to allow for dependence between observational units, e.g. spatial correlation between the errors of adjacent municipalities.

To illustrate my thoughts, I have included an empirical application of my approach with data from a rural Austrian federal province. Initially this was merely meant as a means to compare my new approach with existing methods. However, when presenting parts of my thesis on several occasions, I have come to understand that many people—even statisticians—were more interested in the empirical application than in my theoretical derivations. Taking this into account, I decided to refine my application example by first finding a large number of candidate explanatory variables based on various migration theories and then applying an appropriate model-selection algorithm.

It appears that the application is now no longer merely an illustration of theoretical thoughts, but has become an added value of my thesis in its own right. Nevertheless, I hope I have made in the subsequent chapters the necessary distinction between the generic model and the application. In particular, the results of my application are relevant only for the municipalities considered there; for different rural regions, one will in general have to consider other regressors. A basic fact of aggregate migration analyses is that there are always many regressors which are not measureable or simply unknown. The purpose of my work was to develop a statistical tool that is able to handle this in a general way.

Many thanks go to all those who discussed and criticized parts of my work on various occasions. In particular, I would like to express my gratitude to my supervisor, Prof. Wilfried Grossmann, who always took time for me when I asked and who gave me some critical advice which I had not considered otherwise. I am also indebted to my parents, family members and friends who showed interest in the progress of my study and supported me to continue at any time. Special thanks go to my wife Katharina, who has shown an abundance of patience and support not only, but especially during the last five years.

Vienna, in the summer of 2014

Johannes Klotz

Contents

1	Background	1
1.1	Introduction	1
1.2	Randomness in demographic data	3
1.3	Established models of demographic count data	6
1.3.1	Vital events	6
1.3.2	Migration	7
1.4	Likelihood-based regression models for count data	8
1.4.1	The Poisson regression model	8
1.4.2	Overdispersion	11
1.4.3	Other maximum likelihood estimators	11
1.4.4	Pseudo maximum likelihood	12
1.4.5	Quasi maximum likelihood	15
1.4.6	Quasi-generalized pseudo maximum likelihood	16
2	Analyzing aggregate migration flows	19
2.1	Defining and measuring aggregate migration	19
2.2	Explaining migration	21
2.3	Further model specification issues	28
3	The basic model	33
3.1	The Poisson process on the nonnegative real line	33
3.2	Violations of the Poisson process regarding migration flows	34
3.2.1	Clustering by household migration	34
3.2.2	Unobserved heterogeneity	35
3.2.3	Contagion	35
3.3	A variance function with two parameters	36
3.3.1	Mean and variance of the household count	36
3.3.2	Mean and variance of the person count	38
3.4	Parameter estimation	40
3.4.1	Mean function parameters	40
3.4.2	Variance function parameters	41

4	Application	45
4.1	The dataset	45
4.1.1	Investigation area	45
4.1.2	Dependent variables	45
4.1.3	Candidate explanatory variables	47
4.1.4	Model selection	50
4.1.5	Descriptive statistics	52
4.2	Estimation results	54
4.3	Model evaluation	58
4.3.1	Mean function	58
4.3.2	Variance function	59
4.3.3	Sensitivity analysis: Alternative PML point estimates	62
5	The rate model	67
5.1	Simultaneity of demographic rates	67
5.1.1	Endogenous regressors	67
5.1.2	Demographic rates	68
5.1.3	The rate regression model	70
5.1.4	Traditional demographic approaches to simultaneity	72
5.2	An new instrumental-variable approach	73
5.2.1	Two-stage residual inclusion	73
5.2.2	Adjustment of the variance-covariance matrix	75
5.3	Application	78
5.3.1	The data	78
5.3.2	First-stage estimation	78
5.3.3	Second-stage QGPML with 2SRI	81
6	Conclusions and outlook	85
6.1	Summary	85
6.2	Possible extensions of the model	88
6.2.1	Extensions of the univariate model	88
6.2.2	Joint analysis of in- and out-migration	89
A	References	93
B	Abstract	102
C	Abstract in German	103
D	Curriculum Vitae	104

List of Tables

4.1	Descriptive statistics of variables on the original scale	53
4.2	Descriptive statistics of regressors on the logarithmic scale	53
4.3	Correlation matrix of regressors	54
4.4	QGPML regression results for the basic model	55
4.5	Estimated variance-to-mean ratios	58
4.6	Alternative PML estimates for the in-migration model	64
4.7	Alternative PML estimates for the out-migration model	65
5.1	First-stage GLS regression results	81
5.2	Second-stage QGPML regression results for the rate model	83
5.3	Second-stage QGPML t-ratios	84

List of Figures

4.1	Municipality map of Burgenland	46
4.2	In-migration residual plots	60
4.3	Out-migration residual plots	61
4.4	Graphical evaluation of the variance function	63
5.1	First-stage residuals: OLS vs. GLS	80

Abbreviations

2SPS	Two-stage predictor substitution
2SRI	Two-stage residual inclusion
AIC	Akaike information criterion
AICc	Corrected Akaike information criterion
GLM	Generalized linear model
GLS	Generalized least squares
LEF	Linear exponential family
LEFN	Linear exponential family with nuisance parameter
LTS	Least trimmed squares
ML	Maximum likelihood
NB	Negative binomial (Note: “NB2” refers to a particular parameterization)
OLS	Ordinary least squares
PML	Pseudo maximum likelihood
QAIC	Quasi Akaike information criterion
QAICc	Corrected quasi Akaike information criterion
QGPML	Quasi-generalized pseudo maximum likelihood
QML	Quasi maximum likelihood
RS	Robust sandwich
VC	Variance-covariance matrix
VF	Variance function

An E following an abbreviation (e.g., MLE) means estimation, estimator or estimate. NUTS (in Chapter 4) is a statistical system of territorial units in the European Union.

Symbols

A	Matrix, component of a VC
B	Matrix, component of a VC (basic model)
C	Random variable (household size, number of persons in a household)
D	Matrix, component of a VC (rate model)
E	Expected value
F	Matrix, component of a VC (rate model)
G	Matrix, component of a VC (rate model)
K	Population count on 1 January of the year of interest
M	Random variable (household migration count)
P	Person-years lived during the year of interest
T	An estimate divided by its estimated standard error
U	Matrix, component of a VC (rate model)
Y	Random variable (person migration count)
Var	Variance
a	Function, part of a LEF(N) density
b	Function, part of a LEF(N) density
c	Function, part of a LEF(N) density
e	Error term in the (second-stage) regression equation
h	Vector, component of a VC (rate model)
i	Index of observations ($i = 0$ means an out-of-sample observation)
k	Nuisance parameter, NB2 shape parameter
l	Index of iterations (for iterative estimation algorithms)
n	Sample size
p	Number of (second-stage) regressors
q	LTS trimming parameter
r	Demographic rate (event count divided by person-years lived)
s	Score vector of (second-stage) regression parameters, component of a VC
t	Score vector of first-stage regression parameters, component of a VC (rate model)
u	Error term in the first-stage regression equation, additional second-stage regressor
v	Assumed variance function, corresponding to a nominal LEF distribution
w	Vector of first-stage regressors (including the instrument)
w^+	First-stage instrument
x	Vector of (second-stage) regressors

x_e	Endogenous component of x
x_o	Exogenous components of x
x_u	An index of unobserved regressors
y	Dependent variable (person migration count)
z	Vector of unobserved (second-stage) regressors
α	Vector of first-stage regression parameters
β	Vector of (second-stage) regression parameters
β^*	The same as β , except for the intercept parameter
β_e	The component of β regarding the endogenous regressor
β_o	The components of β regarding the exogenous regressors
β_u	The component of β regarding the additional regressor u
γ	Vector of variance function parameters
δ	Vector of regression parameters of unobserved (second-stage) regressors
ϵ	Disturbance term in the estimable variance function model
λ	Rate of the Poisson process
μ	Expected value of the (second-stage) dependent variable
ξ	An index of unobserved regressors
σ^2	Scalar (second-stage) dispersion parameter
ϕ^2	Scalar first-stage dispersion parameter

All vectors are column vectors. Components of parameter vectors are denoted by γ_0 , γ_1 , and the like. Estimators or estimates of model parameters and their functions are denoted by $\hat{\beta}$, $\tilde{\alpha}$, $\widehat{VC}$, and the like. “Second-stage” applies to the rate model in Chapter 5.

1 Background

1.1 Introduction

This thesis deals with the statistical analysis of migration on an aggregate level. We are modeling period counts, i.e. annual totals of in-migration and out-migration. The observational units that we have in mind are municipalities in rural areas. Aggregate migration data for rural municipalities come with certain particularities: The populations are quite small and not dominated by a single large city (Gorton et al. 1998), which allows for statistical independence assumptions. The annual flows may be zero or small positive numbers requiring methods for count data. Moreover, the empirical findings for rural areas are often essentially different from urban areas (Greenwood 1975).

Even within rather small investigation areas, one usually finds considerable variation in migration rates between different municipalities. A key function of statistical analysis is to disaggregate that variation into systematic influences caused by structural differences (such as demographic and economic disparities) between the places and unexplained random variation. The classic approach to that is regression analysis, i.e. modeling the expected value of a migration count (or rate) conditional on explanatory variables.

Though this is not a formal requirement of our approach, our model is designed for situations where migration is measured from administrative sources like registration of residence data. Compared to sample survey data, administrative data naturally come with many observations, but relatively few variables, for they are not primarily collected for statistical analyses. Thus, one has little information on the migrants, and in regression models unobserved heterogeneity is a central issue.

The “rural migration turnaround” There is considerable interest in small-area migration differences related to regional planning policy. During the last decades, migration has become an increasingly important part of population change in rural areas. It was previously noted by Ravenstein (1885) that most migration is short-distance, which implies that migration becomes the more important the finer the geographical disaggregation is. However, variation in migration rates has become particularly important for rural areas in the last decades because of the virtual disappearance of natural population increase (i.e., live births minus deaths).

The long term shift of the rural economy from agriculture to manufacturing and services has resulted in a substantial decline of rural fertility. The once large gap between rural and urban fertility rates has largely disappeared and in some instances even reversed. Simultaneously, the long term customary net migration loss of rural areas has overall declined or even reversed too

(Champion 1998, Cadwallader 1992), which has been called the “rural migration turnaround”. Popular explanations are the increasing importance of commuting to large cities, industry-specific in-migration to rural areas (tourism, mining), and non-labor market related phenomena such as retirement migration and dropouts.

In the 1970’s, some researchers believed in a “rural renaissance” (Boyle et al. 1998). As we know today, this has not come true, at least not as a general finding. First, the temporal pattern of rural migration showed considerable variation in the late 20th century, as described e.g. by Cadwallader (1992) or by Boyle et al. (1998). In-migration waves were followed by drain periods and vice versa. Second, the last decades have come along with increasing variability of migration patterns within rural areas (Boyle et al. 1998, Hugo and Bell 1998), often taking the form of local population concentration (Champion 1998). Moreover, international immigration has become an important component of population change in recent decades and although this disproportionately is with regards to urban areas, it is also visible for rural areas.

Migration influences population growth not only directly, but also indirectly via its influence on natural increase, most notably by increasing or decreasing the population at reproductive ages. In general, the migrants are a selective part of the total population and even small migration rates may have a serious impact on the demographic makeup of a community. For some rural places, highly selective forms of migration like relocation to institutional households are particularly important.

Given the great importance of migration as a component of demographic change in rural areas, it is important to understand its determinants. In the 1980’s and 1990’s, the increasing availability of individual-level micro data induced a temporal shift in scientific interest from aggregate to individual migration analysis. However, it was already predicted by Mueser and White (1987: p. 19) that “there will always be concern with explaining aggregate migration, if only because aggregate statistics are relevant for local governments and policy makers.” Indeed, there has been considerable interest in aggregate migration models in recent years (Cushing and Poot 2004, Etzo 2008, Lamonica and Zagaglia 2013, Pendakur and Young 2013). One reason is that micro level approaches suffer from specific problems, e.g. they focus largely on the “push factors” but often ignore the “pull factors” of migration (Maier and Weiss 1991). For some applications it may be beneficial to use hybrid approaches which combine micro and macro models (Tervo 2001).

Outlook on the thesis This thesis is divided into six chapters. In the remainder of the present chapter we first discuss some philosophical issues regarding the application of statistical methods in demography, for demographic data is usually based on “complete enumerations”. Next, we review established statistical models for demographic counts and then for count data in general. In Chapter 2, we summarize important issues in the statistical analysis of aggregate migration counts such as definitions, theory-based choice of explanatory variables, and challenges in model specification. In Chapter 3, we derive our basic semi-parametric model of period migration flows for rural municipalities and discuss consistent and asymptotically

efficient estimation. Our derivation is based on violations of the Poisson process. In Chapter 4, our basic model is applied to the 2012 annual totals of in-migration and out-migration of the 171 municipalities of Burgenland, a rural federal province of Austria. Chapter 5 extends the basic model to the case of endogenous regressors, in particular the total of person-years lived in migration rate models. Finally, some possible further extensions of the basic model, in particular bivariate models for joint analyses of in-migration and out-migration, are discussed in Chapter 6.

The value added of this thesis is as follows: First, a particular migration model, namely the quasi negative binomial distribution, is embedded into a sound theory of stochastic processes. We show that it naturally arises if the Poisson process is violated in two ways, namely clustering of persons by household migration and unobserved heterogeneity between places. Second, based on the finding that population sizes are typically right-skewed, we propose a robust estimator for the shape parameter of the negative binomial distribution, namely the least trimmed squares estimator. Third, we handle simultaneity between migration count and person-years lived by two-stage residual inclusion, an instrumental-variable technique which to the best of the author's knowledge has not been applied so far in demographic rate models. Finally, there is also value added in the application example. Based on various migration theories, we have developed a large set of migration regressors that are available for Austrian municipalities and some of our substantive results are likely to be of interest to policymakers as well as scientists in various ways.

1.2 Randomness in demographic data

A particularity of demographic data is that it is usually obtained by “complete enumerations” such as censuses or administrative data like registration of residence records. This means that in demographic analyses one does not usually work with “samples” in a sense that only a fraction of a finite target population were observed. Traditional survey sampling complexities such as clustering, subsampling or nonresponse are virtually unknown to demographers, at least regarding aggregate population outcomes. There exists a philosophical divide among demographers as to what extent statistical approaches are appropriate for such outcomes.

On the one hand, it is sometimes argued that for statistics based on complete enumerations stochastic approaches such as variance estimation or hypothesis testing are unnecessary, or even not allowed, for the statistics were already “the true values”. This deterministic approach is particularly popular in official statistics, where demographic data is traditionally analyzed from a bookkeeping perspective. Under such a paradigm, there would be no difference between the work of statisticians and “those of accountants, cashiers, and bankers” (Kish 1986). The stochastic approach, on the other hand, interprets demographic outcomes as realizations of random variables.

Stochastic approaches have become quite common in demography during the last 30 years or so and nowadays they are applied not only in academia, but also in official statistics. One explanation for this is that the appropriateness of stochastic approaches depends not only on

the data source, but also on the purpose of the analysis and there are indeed many applications of “complete-enumeration” data where statistical techniques clearly make sense. As quoted by Hoem (1986: p. 718), it was realized already in the 19th century that “there can be stochastic variation in vital statistics even when the data do not come from a sample survey”. Deming and Stephan (1941: p. 45) stated that “as a basis for scientific generalizations and decisions for action, a census is only a sample”. Udry et al. (1979) discussed fluctuations in annual birth rates and concluded that they “although not sampling error, may be thought of as being generated by a random process operating in the population”. Chiang (1984: p. 78), with respect to age-specific death probabilities in life tables, considered human life to be a random experiment and its outcome, survival or death, subject to chance. Brillinger (1986) asked when “two death rates differ by more than some level of natural fluctuations”. He also clarified that “the data are virtually complete (...) the purpose of this study is the stochastic conceptualization and formulation of the natural variability of vital statistics to support their use in comparisons and other analyses.”

Applications of the stochastic approach Let us now discuss the precise applications in demography for which the stochastic approach has proved beneficial. First, statistical methods provide an objective criterion to identify “small” populations. An age-specific death rate of 0.001 may be one death in a population of 1,000 person-years lived, or 100 deaths in a population of 100,000. Although numerically the same, most demographers might agree that the two rates should be treated differently. In general, summary statistics like rates, which hide the size of the underlying populations, are problematic for small populations (see also Scherbov and Ediev 2011). As Manton et al. (1981: p. 266) put it regarding cross-tabulation of rates, “the occurrence of even one death in a cell may produce a very high cell death rate”. In short, demographic rates based on small exposures are not meaningful. But how small is small? A natural way to assess this is to interpret the rate as an outcome of a random variable and estimate its reliability.

A second application of the stochastic approach is to find appropriate weights in a regression analysis when the observational units are populations in different areas, as it is frequently the case in epidemiology (Pocock et al. 1981, Manton et al. 1981). The deterministic approach would imply that the values of each area have to be weighted equally. This is unsatisfactory when there is, as usual, great variation in population sizes between places. On the contrary, in a stochastic framework the inverse of the variance of the outcome is a natural choice for the weight in a regression analysis.

In general, as soon as the demographer is dealing not merely with data analysis, but with regression models it is almost inevitable to introduce stochastic elements. In the social sciences, relying on observational data, one never observes causal relationships between variables. Measured regression coefficients are subject to countless confounders, most of them unobserved. Moreover, many of the observable regressors are only more or less precise estimates or proxies of the quantities of theoretical interest.

Then, at a closer look many demographic analyses are in fact not based on complete enu-

merations. For many applications, the true target population in fact exceeds the observations at hand. This is certainly the case when statistics serve as a basis for political decisions, for the data describe a population in the past, whereas the political decisions will come into force for a population in the future. In that sense, a census may be interpreted as a picture taken randomly in time resulting in one of the many population stocks that might have resulted under essentially the same conditions (Deming and Stephan 1941; see also the references given in Brillinger 1986). A similar assumption is often made in macroeconomic analyses, when annual observations of an economy are interpreted as “sampled” observations (e.g., Barro 1977). Regarding survey sampling, Kish (1986) distinguishes two steps of statistical inference. First, an “objective” step from the sample to the sampling frame and second, a “subjective” step from the sampling frame to the inferential population, which may be under, above, behind or beyond the frame population.¹ Bessant and MacPherson (2002) draw the distinction between mathematical and statistical reasoning by the former eventuates a single immutable law, whereas the latter means deducing group trends and tendencies that may not hold for a given individual.

Another application of the stochastic approach is plausibility checking and error detection. Between primary collection of demographic data, usually by government agencies, and final aggregate statistics publication, usually by national statistical institutes, there are data inspection and correction steps, both on the micro and macro level. An important issue regarding aggregate outcomes is to assess whether they lie “far” from what can be expected. The inspection limits may be defined by statistical criteria like standardized residuals. An example is given by Brillinger (1986: p. 707), who discovered by regression analysis an unexpectedly high age-specific mortality rate for Canadian females in 1951, which turned out to be a misprint.

Finally, statistical methods are also a strategy to account for uncertainty in population projections. Uncertainty naturally stems from the fact that the future is unknown and there exist various approaches how to deal with it. A common strategy in official statistics is to project different scenarios, each based on specific assumptions. An alternative way is to calculate prediction intervals for future outcomes, based on statistical models. It should be mentioned that the last years have seen an upswing in “probabilistic” projections, although most of them are based on a Bayesian understanding of probability and thus not the best instances of the stochastic approach that we have in mind here.

A basic fact in aggregate demographic analyses is that the “random experiments” cannot be repeated like in the natural sciences. The randomization is not under the statistician’s control, so he or she has to look for alternative ways than repeated measurements to estimate the variability of aggregate demographic outcomes. Essentially two approaches do exist, often in combination: The first approach is to construct pseudo-replicates such as different areas in a period or observations from different periods (or a combination of both, see Udry et al. 1979). The second approach is to base variability estimation on theoretical considerations. An instance of this is given in Chapter 3.

¹In an analysis of labor mobility based on the German socio-economic panel, the authors reckon that “the period is so long that macro effects can be expected to play a minor role on average, so that the observed individual mobility patterns may be indicative also for different time periods with other macroeconomic conditions” (Jung and Winkelmann 1993: p. 549).

1.3 Established models of demographic count data

Here we briefly summarize the history of statistical models in demography. Precisely, we look at demographic flow data—as opposed to demographic stock data, which would be e.g. census data. In accordance with the aim of this thesis, we concentrate on models of aggregate population outcomes. Also we focus on cross-sectional analyses rather than time series.

An interesting finding in the history of statistical models in demography is that models for flow data have emerged in two rather isolated strands, namely vital statistics (live births, deaths) on the one hand and migration on the other hand. It therefore seems reasonable to separate the summary accordingly.

1.3.1 Vital events

Concerning vital statistics, early approaches for demographic rates used the binomial or the Poisson distribution for the number of events (Spiegelman 1968, Udry et al. 1979 and the references given there, Chiang 1984). The risk population was typically treated as a fixed quantity, although there was some debate how the risk population should actually be defined. For example, regarding birth rates, it was realized that the risk population is in fact the female population of reproductive ages, even if the rate refers to the total population (Udry et al. 1979). The primary focus of these early stochastic approaches was to estimate standard errors of rates and subsequently more complicated vital statistics such as age-standardized death rates and life expectancies. Regression models were rarely used, except in epidemiology (Pocock et al. 1981, Manton et al. 1981).

A step forward was made by the 1986 Biometrics paper by Brillinger with a primary focus on age-specific death rates and included an extensive discussion of leading statisticians of its day. Although Brillinger's results were based on rather restrictive assumptions such as a closed population (which is for itself not suitable for migration analyses), Poisson birthtimes,² and estimation of person-years lived by mid-year population, his paper made progress in several respects. First, it gave an explicit reference to stochastic point processes in the Lexis diagram. The key statistical feature of the Lexis diagram, namely piecewise constant intensities, had in fact been assumed implicitly in earlier works, but he was the first to formulate explicitly a stochastic planar point process. Second, Brillinger acknowledged the simultaneity between event count and person-years lived, so “randomness” was extended to the denominator of a demographic rate. He also clarified that in period analyses the binomial distribution is not a suitable model for age-specific mortality rates, for the exposure times of the individuals are in general not the same during the year of interest. Finally, the paper stressed the general importance of accounting for unobserved heterogeneity in risk. What remains confusing is that

²The assumption of a Poisson process for birthtimes is insofar unsuitable as populations reproduce themselves, meaning that fertility is not exogenous and thus the Poisson process assumption of stationary and independent increments untenable. However, this is not a problem in a cross-sectional analysis conditional on what happened before the observational period. Interestingly, Hoem (1986) mentions that a Poisson process was sometimes assumed for immigration times, but unfortunately he does not give any references.

Brillinger limited the discussion of unobserved heterogeneity to the regression case, whereas it should be clear that unobserved heterogeneity is even more important in the non-regression case. Under idealized conditions, the numerator and denominator of an age-specific mortality rate were shown to follow a bivariate Poisson distribution (in the sense of Kocherlakota and Kocherlakota 1992: Ch. 4).

Much of the following work on statistical models for vital events referred to Brillinger's 1986 paper. Stochastic processes in the Lexis diagram were further elaborated by Keiding (1990), who discussed nonparametric alternatives. Kegler (2007) considered clustering of events with regard to violent deaths such as multiple homicides or homicide followed by suicide, leading to a generalized (stopped-sum) Poisson distribution. It seems that although the presence of "overdispersion" (compared to the mean-variance-equality of the Poisson distribution) was often realized, few attempts have been made to model it directly, particularly in non-regression contexts (although Manton 1986 mentioned the negative binomial distribution). In regression contexts, the 1981 papers by Pocock et al. and by Manton et al. considered different sources of random variation like unobserved individual "frailty" in mortality analyses. Altogether the main focus of stochastic approaches in vital statistics is still not regression, but variance estimation for sophisticated nonlinear demographic measures.

1.3.2 Migration

Most aggregate migration analyses refer to origin-to-destination specific migration flows within a closed system. The usual starting point for modeling such flows is the gravity model (Flowerdew and Aitkin 1982, Adam 2009). In analogy to the Newtonian gravity model in physics, the expected migration flow is positively associated with the population sizes of two areas and negatively with the distance between them. Note that unlike population size, the meaning of "distance" is ambiguous for areas and is one of the challenges in statistical applications. The principal ideas behind the gravity model were already mentioned in Ravenstein's famous 1885 paper on the "laws of migration".

How exactly the expected migration flow is related to population sizes and distance in general depends on parameters to be estimated. Often additional regressors such as regional wage levels and unemployment rates are incorporated, leading to the modified gravity model (Mueser and White 1987, Adam 2009). So in contrast to vital events, statistical models for migration flows have primarily emerged in the regression context. Standard errors for sophisticated nonlinear measures have been of little interest, because there are not many such measures in migration analyses.

Historically, applications of the (modified) gravity model were based on the lognormal distribution. This changed after the seminal 1982 paper by Flowerdew and Aitkin, who placed emphasis on the superiority of the Poisson distribution compared to the lognormal. It was however realized immediately that the mean-variance-equality of the Poisson distribution was a too restrictive assumption for migration counts and that realistic models should allow for overdispersion. Two important classes of models emerged, one based on the negative binomial

distribution, the other based on semiparametric extensions of the Poisson distribution like Poisson quasi-likelihood. Details are given by Congdon (1993).

Models that account for simultaneity between migration count and person-years lived have been rarely applied in migration analyses, although the general problem of endogenous regressors has often been realized (Greenwood 1975, Cadwallader 1992). Compared to vital statistics, a lot of material has been published on explanatory variables and regression specification. Details are given by Congdon (1992). Since migration is by definition a spatial phenomenon, applying the techniques of spatial econometrics (LeSage and Pace 2004) would be self-evident in migration modeling, but at present applications of spatial econometrics are still rare in the migration literature (Cushing and Poot 2004).

An issue that is particularly important in migration modeling is clustering of events by household and family migration. It was early recognized that such clustering was a major reason (although not the only one) for extra-Poisson variability (Flowerdew and Aitkin 1982). An attempt to model clustering directly was made by Flowerdew and Boyle (1995), who applied a generalized Poisson distribution where the number of migrating households is Poisson distributed and generalized by an empirical household size distribution. Though their approach has some appeal, it is limited to situations where households can be identified in the data and that is rarely the case.

So although models for vital events and migration have emerged in rather different strands, a closer look reveals several issues that have been considered relevant in both strands. A certain dichotomy nevertheless remains. To give an example, Hyndman and Booth (2007) developed a stochastic population projection framework for Australia. For mortality and fertility rates, they assume a Poisson process. For net migration, they do not make any parametric assumptions, mentioning that its “law is unknown”. The author hopes that this thesis contributes to a more unified view on stochastic approaches in demography, because the basic issues are largely the same for any demographic process and clearly both strands would benefit from an exchange of experience.

1.4 Likelihood-based regression models for count data

1.4.1 The Poisson regression model

In Section 1.3 we have seen the key role played by Poisson regression in modeling demographic count data. Indeed, the Poisson regression model is the general starting point to analyze any count data (Cameron and Trivedi 1998, Winkelmann 2010, Hilbe 2011). Let us denote by Y an integer-valued random variable and by y a particular realization of Y .³ Denote by x a $p \times 1$ vector of explanatory variables, for convenience assumed to be nonstochastic, and by β the corresponding $p \times 1$ vector of regression parameters. The Poisson regression model is then

³At this introductory stage we distinguish explicitly between random variables and observations. Later we relax that distinction when there is no risk of confusion.

defined by the conditional mean (conditional expectation, regression) function

$$\mu = E(Y|x) = \exp(x'\beta) \quad (1.1)$$

and the probability mass function

$$\Pr(Y = y|\mu) = \exp(-\mu) \frac{\mu^y}{y!}. \quad (1.2)$$

Note that the Poisson regression model contains an implicit error

$$e = y - \mu, \quad (1.3)$$

which is helpful in deriving properties of estimators (see also Wooldridge 2010: ch. 12).

Exponential mean function In this thesis we use the exponential mean function $\exp(x'\beta)$ to relate the linear predictor (or single index) $x'\beta$ to μ , which is the conditional expectation of the dependent variable. This functional form is not a formal requirement of the Poisson regression model; one could instead use e.g. the linear mean function. The exponential mean function has however several important advantages: First, it guarantees non-negative predictions for any value of x . Then, it allows for meaningful parameter interpretation. The coefficients of continuous regressors are semi-elasticities, indicating the effect in proportion to the unobserved expected value (King 1989). If continuous regressors enter in logged form, then parameters are elasticities. For binary predictors, the exponentiated parameters measure multiplicative effects, which are to be expected in many count data applications and called incidence rate ratios (Hilbe 2011: p. 109). For the Poisson regression model it happens that the exponential is in Generalized Linear Model (GLM) parlance also the “canonical” mean function, although this is usually not a criterion in applications (McCullagh and Nelder 1989: Ch. 2).

There is of course no guarantee in applications that the exponential mean function is correct. By a correct functional form we understand that there exists some parameter value β in the parameter space such that μ is, for any x , actually the conditional expectation of Y .⁴ If there is exactly one such parameter value, then the model is said to be identifiable or estimable (Greene 2003: Ch. 17, Wooldridge 2010: Ch. 12).

Maximum likelihood estimation Assume we have a sample of n independent observations, indexed by $i = 1, \dots, n$. The Poisson regression model parameters β may be estimated by maximum likelihood (ML). ML estimation is an instance of a general class of estimators, the so-called M-estimators, which are obtained by minimizing a function (Gourieroux et al. 1984a, Greene 2003: Ch. 16, Wooldridge 2010: Ch. 12). If the function to be minimized is sufficiently differentiable and the value of interest lies in the interior of the parameter space,

⁴This means that $E(e|x) = 0$.

then an M-estimator may equivalently be defined as the solution of estimating equations. In this framework the Poisson MLE $\hat{\beta}_{\text{ML}}$ is defined as the solution of the p estimating equations

$$s_n(\hat{\beta}_{\text{ML}}) = -\frac{1}{n} \sum_{i=1}^n [y_i - \exp(x_i' \hat{\beta}_{\text{ML}})] x_i := 0. \quad (1.4)$$

The estimating equations are of course the negative average partial derivatives of the Poisson loglikelihood with respect to β (the negative average “scores”) with $\hat{\beta}_{\text{ML}}$ inserted. In analogy to the implicit error we may define the implicit residual

$$\hat{e}_i = y_i - \hat{\mu}_i = y_i - \exp(x_i' \hat{\beta}_{\text{ML}}), \quad (1.5)$$

so the estimator can also be defined as a set of orthogonality conditions between the residuals and the regressors.

The variance-covariance matrix (VC) of the Poisson MLE may be estimated in the usual way based on the inverse of the negative expected Hessian matrix, with $\hat{\beta}_{\text{ML}}$ inserted:

$$\widehat{\text{VC}}(\hat{\beta}_{\text{ML}}) = \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n \exp(x_i' \hat{\beta}_{\text{ML}}) x_i x_i' \right)^{-1}. \quad (1.6)$$

There exist alternative VC estimators, such as the outer product of the score estimator, the robust sandwich estimator, and re-sampling estimators such as Jackknife and Bootstrap estimators (Cameron and Trivedi 1998: Ch. 3, Hardin et al. 2003).

MLEs are particularly attractive because of their asymptotic properties.⁵ Count data models are usually nonlinear, so asymptotic properties are the essential criteria of the goodness of an estimator. MLEs are consistent, asymptotically Gaussian distributed and efficient, in the sense that they reach the asymptotic Cramer-Rao lower bound. This means that in large samples the VC of any other consistent and Gaussian distributed estimator equals the VC of the MLE plus a nonnegative definite matrix (Greene 2003: Ch. 17). MLEs also meet the invariance principle, which is important for predictions and for transformed parameters like incidence rate ratios. Moreover, many MLEs can be computed by standard statistical software packages such as R, SAS, or SPSS.

The major drawback of MLE is that it requires complete specification of the conditional distribution of the dependent variable up to the regression parameters to be estimated. In macro demographic applications it is barely possible to specify an entire conditional distribution correctly. In particular, there is usually strong evidence that the observations are not Poisson distributed, so estimation methods other than Poisson MLE have to be applied.

⁵Such statements are always subject to regularity conditions. A detailed discussion of regularity conditions for likelihood-based inference and for M-estimation in general is beyond the scope of this thesis. We refer to McCullagh (1983), Gourieroux et al. (1984a), Newey and McFadden (1994), Wooldridge (1997, 2010: Ch. 12), Stefanski and Boos (2002), Greene (2003: Chs. 9, 10, 17), Ziegler (2011), and Kuan (2013). In this thesis we assume that the usual regularity conditions are fulfilled in the relevant situations, if not otherwise stated.

1.4.2 Overdispersion

A key feature of the Poisson distribution is the equivalence of mean and variance. In a regression model this is called equidispersion, meaning equivalence of conditional mean and conditional variance:

$$\text{Var}(Y|x) = E(Y|x). \quad (1.7)$$

Thus, unlike in the linear Gaussian regression model, conditional mean and variance are not allowed to vary independently in the Poisson regression model, but they are equal. Conversely, if the data are not equidispersed, then the Poisson regression model is inappropriate.

The usual situation in demographic applications (in fact, in most count data applications) is overdispersion, i.e.

$$\text{Var}(Y|x) > E(Y|x). \quad (1.8)$$

It may also happen that the conditional mean is smaller than the conditional variance and called underdispersion, although such a situation hardly occurs in aggregate demography.

A detailed treatise on overdispersion is given by Hinde and Demétrio (1998). Overdispersion can be detected in a variety of ways (Winkelmann 2010: Ch. 3) that do not concern us here. What is important in case of overdispersion is that the asymptotic optimality properties of the Poisson MLE do not hold.

1.4.3 Other maximum likelihood estimators

Negative binomial (NB2) model If demographic count data indicates overdispersion, then the Poisson MLE is clearly inappropriate. A first idea is to look for other MLEs. The main alternative is the negative binomial (NB) distribution (Winkelmann 2010: p. 21, Hilbe 2011). The NB probability mass function depends not only on the expected value μ , but on an additional shape parameter k and is given by

$$\Pr(Y = y|\mu, k) = \frac{\Gamma(k+y)}{\Gamma(k)y!} \left[\frac{k}{k+\mu} \right]^k \left[\frac{\mu}{k+\mu} \right]^y. \quad (1.9)$$

In a regression model the conditional mean is as before

$$\mu = E(Y|x) = \exp(x'\beta). \quad (1.10)$$

This regression model is in fact a special instance of NB regression models (Hilbe 2011), namely the “NB2” parameterization, meaning that the shape parameter k is constant across observations. In applications one has to estimate, besides the p -dimensional mean parameter vector β , in general also the scalar shape parameter k . The geometric distribution is the special case of the NB2 with $k = 1$.

In the NB2 regression model, the conditional variance is

$$\text{Var}(Y|x) = \mu + \mu^2/k. \quad (1.11)$$

The Poisson conditional variance is the limit for $k \rightarrow \infty$.

Although the NB2 model is clearly more flexible than the Poisson model regarding the conditional variance, it is restrictive at the end of the day. First, NB2 models can be overdispersed too (Hilbe 2011: p. 180). Then, NB2 MLE is not robust to distributional misspecifications (see below). Finally, unlike the Poisson, the NB2 distribution is not closed under convolutions, which may cause theoretical problems in some applications. So although the NB2 model may be an attractive tool for some count data regression problems, it cannot be recommended as a general approach.

Other count data distributions Other fully parametric models for count data are discussed by Cameron and Trivedi (1998: Ch. 4), Winkelmann (2010: pp. 45-49) and Hilbe (2011: Ch. 10). Most of those models have been designed for specific applications.⁶ The MLEs may not only be inconsistent under distributional misspecification, but also computationally difficult to obtain, for at present they are in general not supported by standard statistical software packages.

1.4.4 Pseudo maximum likelihood

Instead of specifying a complete conditional distribution, it is often beneficial in count data problems to make fewer assumptions and apply semiparametric techniques.

M-estimators We have already mentioned that MLE is an instance of M-estimation, defined by solving estimating equations. Detailed references are Stefanski and Boos (2002) and Wooldridge (2010: Ch. 12). Let us denote by $\hat{\beta}_M$ a generic M-estimator. Then, under appropriate regularity conditions,

$$\sqrt{n} (\hat{\beta}_M - \beta) \rightarrow A^{-1}(\beta) s(\beta). \quad (1.12)$$

This representation is called the “influence function representation” of an M-estimator. Here s denotes the random gradient vector of the function to be minimized under the true model parameters, which has an expectation of zero. The sample analogue of s is denoted by s_n ; this is the quantity that is set to zero in applications to get $\hat{\beta}_M$. Furthermore, A is the $p \times p$ matrix given by

$$A = E \left(\frac{\partial s}{\partial \beta'} \right). \quad (1.13)$$

The matrix A^{-1} acts as a linear transformation in M-estimation. If s is the negative score vector of a correctly specified loglikelihood, then A is of course the Fisher information matrix.

Under standard regularity conditions, M-estimators are consistent and asymptotically Gaussian distributed. An important feature of M-estimators is that their asymptotic variance-covariance matrices have a “sandwich” form:

$$n \text{VC}(\hat{\beta}_M) \rightarrow A^{-1}(\beta) B(\beta) A^{-1}(\beta), \quad (1.14)$$

⁶An interesting example regarding variance components is the Generalized Waring distribution, touched on by Hilbe (2010: pp. 328-333) and elaborated in detail by Xekalaki (1983).

where B is the expected outer product of the gradient at the true parameter value, i.e.

$$B = E[s(\beta) s'(\beta)]. \quad (1.15)$$

If s is the negative score vector of a correctly specified loglikelihood, then the “information matrix equality” means that $A = B$ and consequently the asymptotic n -fold VC of $\hat{\beta}_M$ reduces to A^{-1} (McCullagh and Nelder: p. 469). This simplified form is the basis of the VC estimator of the Poisson MLE that we noted before.

Poisson pseudo maximum likelihood estimator Suppose we estimate a regression model by solving the Poisson ML estimating equations although the conditional distribution of the dependent variable is not Poisson. Gourieroux et al. (1984a) call this an instance of a pseudo maximum likelihood (PML) estimator. That is, regardless of the true conditional distribution of Y we estimate the model parameters β by solving the p equations

$$s_n(\hat{\beta}_{PML}) = -\frac{1}{n} \sum_{i=1}^n [y_i - \exp(x_i' \hat{\beta}_{PML})] x_i := 0. \quad (1.16)$$

This Poisson PML estimator $\hat{\beta}_{PML}$ is of course in general different from the ML estimator regarding to the true data generating process. An important property of the Poisson PMLE is that for independent observations, as long as the conditional mean function is correctly specified, it is consistent, though in general inefficient, for the true β . The reason for consistency is that the Poisson distribution is a member of a special class of distributions, namely the linear exponential families (LEFs).

Linear exponential families Distributions belonging to the class of LEFs have loglikelihoods that can be written in the form of

$$a(\mu) + b(y) + c(\mu) y, \quad (1.17)$$

with μ as usual the expected value of the random variable and a and c functions such that

$$\mu = - \frac{\partial a}{\partial \mu} \bigg/ \frac{\partial c}{\partial \mu}, \quad (1.18)$$

$$\text{Var}(Y) = \left[\frac{\partial c}{\partial \mu} \right]^{-1}. \quad (1.19)$$

The b function guarantees that the probabilities integrate to 1 (Gourieroux et al. 1984a, Cameron and Trivedi 1998: Ch. 2).

Many classic families of distributions are LEFs, such as the Gaussian with known variance, the multivariate Gaussian with known VC, the Poisson, the NB2 with known shape parameter, and also the Gamma with known shape parameter. It should be noted that their densities or probability mass functions may look uncommon when written in LEF form.

Within the class of LEFs, a family of distributions is characterized by its variance function (VF), which relates the variance of the random variable—up to a multiplicative constant—to its expected value (Jorgensen 1987),

$$\text{Var}(Y) = \sigma^2 v(\mu), \quad (1.20)$$

with $v(\mu)$ denoting the VF. For the Gaussian the VF is the unit function, for the Poisson the identity function (and there $\sigma^2 = 1$), and for the Gamma the quadratic function.

Let us return to regression models and data. Assume that μ_i is specified via some conditional mean function $\mu_i = \mu(x_i, \beta)$. It can then easily be shown that PML estimating equations based on a LEF are of the form

$$s_n(\hat{\beta}_{\text{PML}}) = -\frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{v_i} [y_i - \mu(x_i, \beta)] \frac{\partial \mu_i}{\partial \beta} \right\} \Big|_{\hat{\beta}_{\text{PML}}} := 0. \quad (1.21)$$

Details can be found in McCullagh and Nelder (1989: Ch. 9), Cameron and Trivedi (1998: Ch. 2), and Ziegler (2011).⁷

Suppose now that we estimate a model by solving such a set of equations, although the true conditional distribution is not necessarily a member of the assumed LEF of distributions. In particular, we assume that in general $\text{Var}(Y) \neq \sigma^2 v(\mu)$, where $v(\mu)$ refers to the nominal VF, i.e. the VF of the assumed LEF. Of course, the assumed LEF must conform to the data type; for example, one cannot use the Poisson PMLE if there are negative observations. Consistency of a PMLE essentially requires the population analogue of the estimating equations above to hold (Cameron and Trivedi 1998: p. 31). This is however the case if $E(Y|x) = \mu(x, \beta)$, i.e. if the conditional mean is correctly specified.

To give an example, it is not necessary for consistency of the Poisson PMLE that the data is actually Poisson distributed, as long as the conditional mean is correctly specified (and the observations are independent). Another well-known example is that the Gaussian distribution is not necessary for consistency of the ordinary least squares (OLS) estimator in the linear model. As long as the conditional mean function is correctly specified, any PMLE based on a LEF is consistent for β . Instances of misspecification of the conditional mean function are if the functional form is incorrect (e.g., linear instead of exponential) or if regressors are endogenous to the dependent variable.

Sandwich VC estimator As long as we are interested only in consistent point estimation, for count data we can always use the Poisson PMLE or any other appropriate PMLE based on an assumed LEF. However, the information matrix equality does in general not hold in a PML setting. A consistent estimator of the VC of a PMLE based on a LEF is instead the sandwich estimator:

$$\widehat{\text{VC}}(\hat{\beta}_{\text{PML}}) = \frac{1}{n} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}, \quad (1.22)$$

⁷The PMLE based on the LEF is the simplest form of PML estimation, sometimes called PML1. More complex forms are PMLE based on the quadratic exponential family (PML2) and on the quartic exponential family (PML4). See Gourieroux et al. (1984a), Lejeune (1997), Holly et al. (2008), and Ziegler (2011).

with $\hat{A}$ and $\hat{B}$ such that

$$\hat{A} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\hat{v}_i} \frac{\partial \mu_i}{\partial \beta} \frac{\partial \mu_i}{\partial \beta'} \bigg|_{\hat{\beta}_{\text{PML}}}, \quad (1.23)$$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n \frac{\widetilde{\text{Var}}(Y_i|x_i)}{\hat{v}_i^2} \frac{\partial \mu_i}{\partial \beta} \frac{\partial \mu_i}{\partial \beta'} \bigg|_{\hat{\beta}_{\text{PML}}}, \quad (1.24)$$

with $\hat{v}_i = v[\mu(x_i, \hat{\beta}_{\text{PML}})]$.

The estimator of the true conditional variance, $\widetilde{\text{Var}}(Y_i|x_i)$, is an auxiliary estimator, i.e. not part of the PML estimating equations. This is illustrated by using a tilde instead of a hat. How exactly it looks depends on the assumptions that one is willing to make on the conditional variance. The most general estimator uses

$$\widetilde{\text{Var}}(Y|x) = (y_i - \hat{\mu}_i)^2. \quad (1.25)$$

Such a VC estimator is called a “robust sandwich” (RS) estimator because it does not make any assumptions on the conditional variance. Although the individual residual squares do not converge to their expectations, one can show that the matrix $\hat{B}$ is consistent (references are given by Cameron and Trivedi 1998: pp. 65-66). The RS estimator is consistent, but as discussed below, it can be very inefficient. Furthermore, it may have substantial bias in finite samples. Methods to improve the finite-sample behavior of the sandwich estimator are discussed by Dean et al. (1993) and by Ziegler (2011: pp. 74-77).

1.4.5 Quasi maximum likelihood

We have seen that consistency of a PMLE based on a LEF is robust to misspecification of higher moments than the mean, which is the major advantage of this technique. Its major drawback is that PMLE can be very inefficient. Thus we look for ways to improve the efficiency of PML estimation without ruining its consistency.

Although in aggregate demographic applications it is rarely possible to specify correctly the entire conditional distribution of a count variate, in many instances one will have some idea on a few moments other than the mean, in particular the conditional variance. Suppose now that we are able to specify correctly not only the conditional mean, but also the functional form of the conditional variance:

$$\text{Var}(Y|x) = \sigma^2 v(\mu), \quad (1.26)$$

with σ^2 a constant dispersion parameter. Other moments of the distribution may be misspecified. This situation is called quasi maximum likelihood (QML) estimation.

Let us now discuss the statistical properties of QML estimators. First, it is apparent that the constant dispersion parameter σ^2 has no influence on the mean parameter estimates, for it does not change the relative weights of the observations in the estimating equations. For point estimation, it is therefore sufficient to specify $v(\mu)$. Second, compared to PML estimation,

the VC simplifies for QML estimation. The reason is that under the assumptions of QML the “generalized information matrix equality” holds (Wooldridge 2010: Ch. 12), meaning that $B = \sigma^2 A$. It follows that

$$n \text{VC}(\hat{\beta}_{\text{QML}}) \rightarrow \sigma^2 A^{-1}(\beta). \quad (1.27)$$

It can be shown that within the class of PML estimators there exists a lower bound for the asymptotic VC, which is reached by the QML estimator (Gourieroux et al. 1984a).

In a QML setting the VC estimator simplifies to

$$\widehat{\text{VC}}(\hat{\beta}_{\text{QML}}) = \frac{1}{n} \tilde{\sigma}^2 \hat{A}^{-1} \quad (1.28)$$

with $\hat{A}$ analogous to the PML term and $\tilde{\sigma}^2$ an auxiliary estimator, meaning that it is estimated outside the QML estimating equations. Typical estimators equate either the generalized chi-square statistic or the quasi-deviance to its expected value.

In principle, one could still use the RS VC estimator in a QML setting. Such an approach is however not recommended, as the RSE is then inefficient. An intriguing example for the simple linear model is given by Kauermann and Carroll (2001). Other even more extreme cases and a formal proof of inefficiency of the RSE are given by Ziegler (2011: pp. 69-74). The robustness of the RSE to VF misspecification has a price, namely its inefficiency if the true VF is known. Thus, the robust sandwich VC estimator (1.22) should be used only when the true VF is unknown, otherwise (1.28).

1.4.6 Quasi-generalized pseudo maximum likelihood

We have mentioned above that some families of distributions are LEFs under certain specifications. For instance, the Gaussian is a LEF if the variance is known and the NB2 is a LEF if the shape parameter is known. This led Gourieroux et al. (1984) to introduce the class of LEFs with nuisance parameter (LEFN). Members of the LEFN have loglikelihoods of the form

$$a(\mu, k) + b(y, k) + c(\mu, k) y, \quad (1.29)$$

where k is such that for given mean μ there is a one-to-one relationship between k and the variance of the distribution.

We know that within the class of PMLEs, asymptotic optimality is attained by the QMLE, i.e. when the functional form of the VF is correctly specified. For a LEFN, this functional form depends on the k parameter, what we may denote by $v(\mu, k)$. Now in general the true k is of course unknown, so PMLE based on an assumed value is in general not QMLE, even if besides k the VF is correctly specified. The exception occurs if the specified k is in fact a preliminary estimate obtained by a $\sqrt{n}$ -consistent estimator $\tilde{k}$ (meaning that $\sqrt{n}(\tilde{k} - k)$ is bounded with probability 1, see Ziegler 2011: p. 33). It can be shown that then the asymptotic distribution of the PMLE is the same as of the QMLE with known k . Gourieroux et al. (1984a) call such a two-step estimator a quasi-generalized pseudo maximum likelihood (QGPMLE) estimator. In terminology of Stefanski and Boos (2002), the QGPMLE is a partial M-estimator.

It is important to note that the estimator $\tilde{k}$ does not need to be efficient for the true k in order to guarantee efficient estimation of the true β , but merely $\sqrt{n}$ -consistent. Also, that the $\tilde{k}$ estimate enters the PML estimating equations as a “pre-specified” value does not mean that it has to be estimated from different data—what matters is that it is not estimated simultaneously with β . What may happen is that the nuisance parameter k and the QML dispersion parameter σ^2 are estimated simultaneously. One consequence of this is that in QGPML estimation in general neither the generalized chi-square statistic nor the quasi-deviance is equal to its expected value (Condgon 1993).

It might be helpful to mention the analogy to the heteroskedastic linear model. It is a well known result that for independent observations model parameters can be estimated consistently, albeit not efficiently, by OLS, whose VC has then a sandwich form. This corresponds to the PMLE based on the Gaussian distribution. Alternatively, if the heteroskedasticity is known up to a constant dispersion parameter, then more efficient generalized least squares (GLS) estimation is possible. This corresponds to the QMLE. Finally, the heteroskedasticity may be subject to unknown parameters which are not part of the conditional mean function. These parameters may be replaced by $\sqrt{n}$ -consistent estimators and the mean function parameters then estimated via feasible generalized least squares, which is asymptotically as efficient as GLS. This corresponds to the QGPML.

In the context of count data regression, the QGPML approach is particularly useful for the NB2 model variance $\mu + \mu^2/k$. The true shape parameter k is in general unknown. One could specify a certain value, e.g. based on past experience with similar observations and then have a consistent PMLE, but there is no guarantee that the assumed k is the true k and hence estimation efficient. Joint estimation of β and k via NB2 MLE is inconsistent if the data is not actually NB2 distributed, as the NB2 with unknown k is not a LEF. The most promising approach is therefore the NB2 QGPML which replaces the shape parameter by a $\sqrt{n}$ -consistent preliminary estimator $\tilde{k}$. The NB2 QGPML has the attractive property of being consistent and asymptotically efficient within the class of PMLEs if the NB2 variance assumption holds up to a multiplicative dispersion parameter. If the data is actually NB2 distributed, then the NB2 QGPML is for β as efficient as the NB2 MLE, although the preliminary $\tilde{k}$ is not necessarily efficient for k (Cameron and Trivedi 1998: p. 73).

2 Analyzing aggregate migration flows

2.1 Defining and measuring aggregate migration

In this thesis we consider aggregate migration flows. If an area is partitioned spatial units like municipalities, then by migration is meant an individual's change of residence which crosses the border of one of the units. Thus, we are not interested in changes of residence within the same unit. From the perspective of a particular place, in-migration occurs if a person relocates its residence to the unit and out-migration occurs if a person leaves the unit. In principle, migration may also be defined for households instead of individuals. In the following we concentrate on migration of persons.

Aggregate outcomes of demographic processes can be analyzed in several ways, e.g. period or time series data. In this thesis we take a period perspective. What we have in mind is a set of municipalities for which annual totals of in-migration and out-migration are available for one calendar year. In econometrics parlance, we analyze cross-sectional data.

We assume that the data allow for the same person migrating several times during the period. Thus the total of migrations is greater than or equal to the total of migrants. This implies observations of individuals in continuous time. Migration can also be defined based on observations of individuals in discrete time, by comparing residences at the beginning and end of the period; if they coincide, a person is defined to be a stayer, otherwise a mover. In their introductory chapter, Stillwell and Congdon (1991) call this a measurement of "transitions" in discrete time, as opposed to a measurement of "moves" in continuous time. Boden et al. (1992) clarify that the discrete time approach is in fact a cohort rather than period analysis. The discrete time approach has also many disadvantages from an analytical point of view. Return migrants are classified as stayers, multiple migrants cannot be distinguished from single migrants, and for those born, died or emigrated during the period migration cannot be measured. Essentially the discrete time approach is useful only when observations in continuous time are not available.

For a given place, the period totals of migrations are called the in-migration and the out-migration flow (or count), respectively. It is these period totals that we are modeling in this thesis. Other aggregate measures of migration are the migration volume (or turnover), which is the sum of in- and out-migration, then net migration, which is the difference between in- and out-migration, and the migration efficiency ratio, which is the ratio of net migration to migration volume. It is important to note that net migration and the migration efficiency ratio are genuine aggregate system measures and have no meaning on the micro level.

Net migration is an important variable from a policy perspective because, as Smith and

Swanson (1998: p. 254) put it, “policymakers are often more interested in the overall outcome of the migration process than its individual components.” Yet in regression analysis different predictors may be relevant for in-migration and out-migration, so even if one’s ultimate interest lies in net migration, it might still be useful to model the actual flows. A comparable application in economics is Mayer and Chappell (1992).

Change of residence In general, the definition of a “change of residence” varies from application to application. Compared to vital events such as life births and deaths, migration cannot be measured by biological outcomes, so when exactly a change of residence is counted as migration is sensitive to definitions. As Boyle et al. (1998: p. 38) point out, the choice of the definition “is not always determined by the researcher, as they may be forced to use datasets where the definition of migration has already been made for them”. Migration data is usually compiled and provided by national statistical institutes, based on national definitions. International comparability of migration figures is in general poor, however within the European Union some steps forward have been made in recent years by standardization of definitions.⁸

Some conventions in the definition of migration do exist. First, it is common to deal only with changes of principal residences. Changes of secondary residence are not of interest, for they are not (directly) related to the population size and demographic structure of a place. Second, in many instances a minimum duration criterion is applied to distinguish temporary from permanent changes of residence. Statistics Austria for example counts international in-(out-)migration only if the duration of continuous stay in Austria (abroad) exceeds 90 days (not necessarily at the same address and not necessarily within a calendar year).

Level of analysis Depending on the goal of the analysis, different aggregate migration quantities can be analyzed. First, one can be interested only in in-migration or out-migration. In such cases the target variable is obvious. Another situation occurs if both in- and out-migration are simultaneously of interest. Then the simplest approach is to analyze net migration. It may be appropriate if migration is of interest primarily as a component of population change as is often the case in policy applications for which only the increase or decrease of total population matters (e.g., apportionment of election districts or allocation of tax money).

Taking a more demographic point of view, migration influences not only the population size, but also the demographic structure of place, such as its age and sex distribution. Zero net migration may be associated with substantial changes in demographic structure, if large underlying flows are equal numerically, but differ structurally. So from a demographic point of view it is in general beneficial to analyze the two actual flows instead of their mere difference. An issue is then whether the two counts are analyzed separately or jointly. One technique for joint analysis is bivariate regression, which is brought up in Chapter 6.

An approach that is popular among geographers is directional analysis, i.e. analyzing origin-to-destination-specific migration, meaning that the quantities to be analyzed can be tabularized in a migration matrix. A major advantage of directional analysis is that the distance between

⁸Regulation (EC) No 862/2007.

places can be taken into account. The starting point for directional regression models is the gravity model touched in Section 1.3.

A serious drawback of directional analysis is that it is applicable only for a closed system. For example, when the investigation area is a country, then international migration flows have to be excluded from the analysis or analyzed in a separate way. This restriction is particularly relevant when the investigation area is small and flows within it make up only a small proportion of all migrations, making the relevance of directional analysis questionable. Then, directional flows are usually not of great interest to policymakers. Also, in directional analyses the observed counts can be very small, which is likely to devalue certain goodness-of-fit statistics (Flowerdew and Boyle 1995).

2.2 Explaining migration

Why do people move? “Despite the many different approaches to the analysis of migration, there is consensus about one point. All researchers seem to agree that people or households migrate to improve their situation or, more precisely, to be better off in the new location in the future than they would be in the old” (Maier and Weiss 1991: p. 21). Most people might immediately agree that migration is a social act in order to improve one’s situation. Cushing and Poot (2004) interpret migration imbalances between regions as indicators of quality of life differences.

Despite this general view, finding appropriate regressors in a statistical model of migration is an ambitious task. In the social sciences relationships between variables are less clear than e.g. in physics. For migration, an enormous number of possible explanatory variables exist. Any choice of regressors for a statistical model is subject to restrictions. The most obvious restriction is the availability of data, but other features of the model like the level of analysis (micro vs. macro) also matter. Several micro-level predictors, such as age or household size, have no direct meaning on the macro level and must therefore be appropriately summarized for aggregate analyses. Cadwallader (1992) argues that in time series analyses one might prefer other regressors than in cross-sectional studies. In the spirit of Gordon (1991), different regressors may be relevant for long- vs. short-distance moves. In directional analyses, some researchers have applied two-stage models, with a first-stage model for the total flow (or the decision to migrate, if on a micro level) and a second-stage model for its distribution among destinations. The explanatory variables are likely to differ between the two stages.

Regression models in the social sciences are often guided by social theories. Many theories of migration exist, with detailed treatises given by Cadwallader (1992) and Boyle et al. (1998). Next, we will summarize some theoretical approaches to migration that seem useful for aggregate cross-sectional observations. Our discussion focuses on the regressors implied by theory. Most citations refer to migration models for English-speaking countries, but also references to European countries like Germany are included.

Labor market equilibration: Neoclassical theory A well-known theory of migration stems from neoclassical economics. The basic idea is that workers intend to maximize their wages and therefore move from low wage to high wage places. Given this postulate, in-migration increases and out-migration decreases on average with increasing place wage level, which is usually measured by the median or mean wage. A refined approach would include besides the wage level also some measure of the probability of working, such as the unemployment rate or the labor force participation rate. Sandefur et al. (1991) use the county unemployment rate as an explanatory variable and the national unemployment rate as a control variable. Cadwallader (1992) mentions that income variables might be in general more relevant for in-migration than out-migration. Frees (1992) mentions also variation in social welfare benefits.

The major drawback of the neoclassical theory of migration lies in its unrealistic assumptions such as fully flexible wages, unlimited new housing supply, and a homogeneous workforce (Congdon 1991). It is also assumed that migration is speculative rather than contracted, i.e. workers first move and then look for a job. Another limitation of the neoclassical migration theory is that it covers only labor migration (Boyle et al. 1998). For certain groups of the population like pensioners, one can reasonably expect that they seek places with low wages (Sjastaad 1962). In other instances, the decision to move may simply be unrelated to wage levels; Flowerdew and Aitkin (1982) and Frees (1992) report very unique migration patterns among military personnel.

That migration is more than just an equilibrator in the labor market can be seen from empirical results. If the primary mechanism of migration were to equalize wage differences between places, then in-migration and out-migration would be negatively correlated. The typical case in applications is however that in-migration and out-migration are positively correlated, even for large units like United States metropolitan areas (Galle et al. 1993). Thus, although the neoclassical theory may be appropriate to explain some aspects of migration, it is insufficient as a general theory. Extensions of the neoclassical theory as well as alternative theories are presented next.

Real wages and the housing market An important point in migration analyses is that one must not compare money wages, but real wages (Cadwallader 1992). This means that besides wage differentials, also cost of living differentials have to be taken into account. The most obvious cost of living variable is housing costs. A higher nominal wage at a potential destination does not guarantee a higher standard of living, if the wage differential is offset by higher housing costs. A prominent example was the lack of substantial migration from Northern to Southern England in the 1980's, despite large disparities in income levels and unemployment rates. Instead, it was long-distance commuting from the North to the South which boomed during that decade. According to Gordon (1992: p. 131) this might have been an "efficient response" in terms of real wages (see also Hamnett 1992). Hugo and Bell (1998) found for Australian low earners disproportionate migration to rural areas due to affordable housing there. According to Owen and Green (1992), spatial immobility may be rational behavior for

the unemployed, because they trade off insecurity in the labor market against security in the housing market.

In general, housing supply is an inelastic variable, since houses cannot move as quickly as people. The scarcity of affordable housing is a major reason why economically strong places do not necessarily have large in-migration. A different situation occurs of course if large-scale new housing is provided, because then this additional housing supply causes in-migration. Congdon (1993) suggests to include in a model both housing cost and new housing.

Often housing market variables are endogenous to migration. As Forrest and Murie (1992) pointed out, it is still unclear in what way the housing market precisely influences migration. Researchers have tried to overcome endogeneity problems by replacing actual housing market variables with meaningful proxies. One choice is dummies for groups of places. Such dummies could represent larger regions or classes of population size. Also population density is sometimes used as a proxy for housing cost and the potential for new housing.

A finding confirmed in many studies is that owners are less mobile than tenants. Cooke (2013) uses owner vs. renter as a regressor in an American study. In a British analysis (Robinson 1992), by far the highest migration rate was recorded for people living in furnished rented apartments. Of course, tenure status may be endogenous to migration too, if people decide a certain tenure status as a consequence of future migration aspirations.

Place utility and non-monetary amenities A theory that refers exclusively to labor migration cannot explain e.g. student or retirement migration. As Flowerdew (1992) pointed out, movers outside the labor force should not be neglected because for certain places they may be a large group. But also to labor migration, there is often more than simply a response in wage differentials (Boyle et al. 1998). Therefore, it has been proposed to replace the migration goal of income maximization by a broader concept of “place utility” maximization.

Non-monetary variables that influence place utility are often called amenity variables (Milne 1991). Cushing and Poot (2004) distinguish between natural amenities (landscape, climate) and social amenities (infrastructure, health care, public safety). Cadwallader (1992) mentions smog, temperature, and geographic altitude. Fielding (1998) notes that pensioners’ decisions to migrate from urban to rural areas are often driven by expectations of lower crime levels and more quiet neighborhoods. He also mentions that net out-migration from Northern Ireland is not caused by particularly high out-migration, but extremely low in-migration, which can be explained by cultural conflicts acting as an entry barrier. Adam (2009: p. 7) mentions language. Pendakur and Young (2013) mention cultural attachments to place, the role of arts and culture, and the family situation. Fielding (1992) considers place identity. In an American micro-level study, Cooke (2013) uses the binary regressor “previous migration history” which is zero if a person lives in the same state in which either parent grew up. This variable is meant to summarize previous migration experience and potential personal and family ties to the state. In an aggregate analysis for Finland, Tervo (2001) considers the proportion of the population that changed residence during the preceding year.

Although the concept of place utility might be intuitively appealing, in applications it is not

free of problems. First, the more sophisticated a theoretical construct of place utility is, the less measurable it becomes. Consider for example the local degree of ethnic discrimination or hostility against newcomers—clearly important drivers of migration, but hardly available in statistical analyses. Second, many amenity variables are in fact highly correlated with income (Etzo 2008). Finally, what exactly defines place utility varies a lot between different segments of the population and even between individuals. Ethnic diversity for example has been shown to be a push factor in some instances (Cadwallader 1992), but a pull factor in other cases (Pendakur and Young 2013). Aggregate migration models incorporating non-monetary amenity variables may therefore not provide very meaningful results, if individual preferences statistically even out in larger populations.

Population size In modeling aggregate out-migration, it is clear that the population size of a place has to be taken into account, as the resident population is the population at risk of out-migration. Regarding in-migration, the situation is formally different because at risk is precisely the non-resident population. Thus, the in-migration rate (in-migration count divided by person-years lived) is not a rate in a probabilistic sense, but a measure of per capita incidence (Galle et al. 1993).

However, even if the resident population is not at risk of in-migration, there are good reasons to include it as a predictor in in-migration regression analyses. Neglecting population size at the destination implies that the motive to move to a destination is independent of the population living there, which is clearly absurd. Typically people move to houses, jobs, schools, partners or social sceneries, not to abstract places. Population size is roughly proportional to the number of jobs, houses etc. available and can therefore be understood as a measure of attractiveness and opportunities (Cadwallader 1992). According to Stillwell (1991), the attractiveness of a place can be measured by population size, the number of dwellings, or alternatively some composite measure. Pendakur and Young (2013) point out that population size can also be understood as a cultural amenity variable. For example, certain cultural lifestyle communities in large cities could probably not even exist in small villages.

Heterogeneous populations: The human capital theory A major limitation of the neoclassical theory of migration is its assumption of a homogeneous workforce (Hugo and Bell 1998). In reality, the same wage differential matters more for a 20-year-old person than for someone close to retirement. Moreover, today workers vary substantially with regard to education and occupational skills. A useful theoretical approach to account for such diversities is the human capital theory, which essentially explains why people make decisions that can be interpreted as investments. Sjastaad (1962) mentions migration as an investment aimed at increasing the productivity of human resources. Interpreting migration as an investment means that people migrate if the expected discounted returns from migration exceed the expected discounted costs (Milne 1991, Dong 2009: p. 7). Population heterogeneity may affect both expected returns and costs, and returns and costs may be both monetary and non-monetary.

The human capital approach readily allows for explaining selective migration by age. The

older an individual is, the shorter his or her remaining working life and hence the lower the expected lifetime wage gain from moving (Dong 2009). Decreasing migration rates with age represent also increasing costs like greater family ties to the existing area of residence. In a similar fashion, Gordon (1991) mentions that higher migration rates of the “cosmopolitan groups” are partly caused by lower costs, resulting from better access to informational networks. Those with higher education tend to use formal communication channels and power structures for career advancement, whereas the poorly educated tend to use family and friends and informal local networks for information and support (Fielding 1992). A lower cost of moving for the higher social classes may also exist in that they are more welcome at potential destinations (Fielding 1992). Sandefur et al. (1991) explain low migration rates among African Americans by the high costs of moving due to segregation in the housing market.

An existing network at a potential destination reduces the cost of migration, which is particularly important for international migration (Etzo 2008). Thus, the proportion of foreigners present in an area can be understood as a proxy for the cost of migration to their fellow countrymen. Cadwallader (1992) mentions the percentage of migrants at the destination as a proxy for the informational flow to outsiders.

Although basically a theory for micro-level behavior, the human capital theory can be applied to migration flows on a macro level when appropriately aggregated over individuals. This fits what Cadwallader (1992) calls “methodological individualism”, namely the view that all social behavior should be traced back to its roots in individual behavior. However, one must be cautious not to draw conclusions on individual behavior based on aggregate models, because such an approach may result in what is known as “ecological fallacy” (Cadwallader 1992) or “aggregation bias”. An example is given by Gordon (1992), who discovered that on the one hand unemployed household heads had a higher propensity to move than employed household heads on an individual level, but at the same time the unemployment rate was negatively correlated with the migration rate on an aggregate level. Thus, one has to be careful how exactly the human capital theory may explain aggregate migration.

The household context Several theories of migration assume implicitly that it is individuals who move. In fact, this is often untrue, for migration is largely a household act. Incorporating the household context into migration theories often increases their explanatory power. In neoclassical theory, for example, on a family level the goal of income maximization may be augmented by risk diversification. As mentioned by Adam (2009: p. 22), Southern Italian families were found to encourage migration of some of their members in order to diversify the risk of unemployment, even if the wage level at the destination was not substantially higher than at the origin. Sandell (1977) applies an economic model to family migration. The basic assumption is that the family aims at utility maximization, where utility is operationalized as labor income and leisure time. DaVanzo (1976) developed a model based on the human capital approach adapted to the family. Treating the household as the decision-making unit, the expected discounted returns and costs have to be totaled for all household members (Mincer 1978).

As Cushing and Poot (2004: p. 321) put it, “migration is a family decision, but within the family there may be winners and losers”. So it happens that individuals do not move as family members although they would if they were single. This is called “tied staying”. The opposite, moving with the family although one would not if he or she were alone, is called “tied migration”. Though some economies of scale do exist regarding migration, the diseconomies are likely to predominate, so members of larger households are more likely to experience tied staying than tied migration. This implies that on average larger households will be less migratory than smaller households, a finding which has been confirmed several times in the literature (as mentioned by Flowerdew and Boyle 1995). Conflicting interests within a household are likely to regard primarily the labor market; regarding amenity variables some positive correlation can be expected among family members, particularly couples. It has long been assumed that women are more frequently subjected to tied migration, but Cooke (2013) reports that in the recent past in the United States tied migration was equally experienced by men and women and moreover confirms that tied staying is in general more frequent than tied migration.

It is interesting that many seminal papers on family migration were published in the 1970's, a decade when the labor force participation of married women increased substantially and at the same time internal migration rates declined in the United States. Some researchers believed the emergence of the dual-earner household to be a major reason for falling migration rates. The argument given e.g. by Sandell (1977) was that both partners may be in conflict regarding potential wage changes after migration, implying an additional cost of migration. However, reality is more complicated. First, participation in the labor force induces economic independence of women, which in turn increases marital instability and consequently migration. Second, for many dual-earner households (as opposed to dual-career households, see Green 1992: p. 108) the costs of migration might not be particularly high, given that traditional female jobs in industries such as nursing, retail trade, or secretarial work show little geographic concentration. Third, two working family members increase the family income and thus the financial capabilities to move. Mincer (1978) found that two-earner families have, compared to one-earner families, higher migration intensity over small distances, but lower migration intensity over large distances. According to Pendakur and Young (2013) it might be beneficial in analyses to account for the “income loading”, which is the proportion of household income earned by the main breadwinner.

To summarize, the family dimension of migration is rather complex. A universally accepted fact is that the presence of school-aged children inhibits migration (Mincer 1978, Lamonica and Zagaglia 2013). Sandell (1977) mentions the termination of friendships as a non-monetary cost of migration to children. Cooke (2013) suggests to enter both younger children and elderly children as separate regressors.

Life cycle stages The ideas on family migration are closely related to the concept of the life cycle (Grundy 1992). The assumption is that one of the main functions of geographic mobility is the fulfillment of needs emerging from changes in the life cycle, which means

either cost reduction or overcrowding elimination. As Davies (1991: p. 208) puts it, “it is evident that either surplus or shortage of space resulting from life cycle changes may increase the probability of a residential move.” Events typically understood as life cycle changes are completion of schooling, university enrollment, entry into the labor force, family formation, divorce, retirement, widowhood, and final relocation to institutional households (Mincer 1978, Rogers 1990, Warnes 1992, Pendakur and Young 2013). In what way these events might be related to housing needs and aspirations is given by Warnes (1992: p. 177).

Life cycle changes are primarily associated with short-distance migration. There is also population heterogeneity: Cadwallader (1992) argues that households in single-family dwelling units are less likely to migrate as a response to changes in life cycle, because they are better able to adapt their current housing.

The life cycle approach has been criticized for its implicit assumption of a “regular” life course, when in fact life courses have become diversified in recent decades. Moreover, Warnes (1992) argues that the life cycle theory of migration actually refers to the middle classes and is a poor theory to explain the migration behavior of African Americans or poor Europeans that are dependent on public housing.

The life cycle is a concept on the micro level. Aggregate migration models can account for it only in a restricted way. The only life-course related variable that is universally accepted as a migration predictor is age (Stillwell 1991: p. 42). Life course model migration schedules were developed to reflect age-specific migration intensities (Rogers and Castro 1981, Boyle et al. 1998: p. 105). Typically such model migration schedules exhibit four local maxima: at very young ages when children move with parents to new family homes, at young adult ages when student migration, entry into the labor force and family formation are reflected in the global maximum, around retirement ages, although this is much more pronounced in English-speaking than in European societies, and at very old ages, when the onset of physical or social dependence induces final relocation to institutional households, particularly among females (Warnes 1992).

Competing strategies Finally, a key part of neoclassical economics is the prediction of long-run equilibria. Now even if one agrees that social behavior is on balance directed towards equilibrium in the labor market, this need not necessarily be achieved by migration. As Gordon (1992) points out, migration is never the only possible means through which an individual may seek to improve his or her situation, but the migration option is embedded within a wider range of choices. There exist alternative strategies by which people or households react to income and utility differentials between places.

The most important rival of labor migration is commuting. Adam (2009: p. 5) mentions papers on the decision between migration and commuting. He argues that commuting is one of the reasons why migration may not act as an equilibrator in the labor market. In discussing the 1970’s “rural migration turnaround”, Champion (1998) argued that the middle classes have always had some preference for rural life, but this preference was not realized on a large scale until corresponding infrastructure became available in rural areas. Transportation links between

rural municipalities and large cities are therefore likely to be important predictors of migration.

Though this is not often acknowledged, an implicit assumption underlying many migration models is that every person has a single residence. This is in fact often untrue (Boyle et al. 1998). For certain parts of the population alternatives to migration in utility maximization are temporary relocation to secondary residences (Hugo and Bell 1998) and switching between multiple residences. A well-known example is the seasonal migration of many elderly people in Northern America. Maier and Weiss (1991) mention that secondary residences are particularly important for people in education and around retirement age. A related phenomenon is that migration may occur in stages. In particular, leaving the parental home is often a process rather than a clear-cut event. As Gordon (1992: p. 121) expresses it, “recent migrants in particular may choose as an insurance to be only half absent from the old location and only half present in the new location”.

2.3 Further model specification issues

International vs. internal migration In the past migration analysis was often divided into internal migration (within a country) and international migration (between countries). In demographic textbooks the two of them were sometimes treated entirely different and in separate chapters. For regression analysis, this would imply two different models, possibly related e.g. by correlation between the error terms. Two models are of course more complicated than one model and there are also other objections against this distinction of migration flows.

Cushing and Poot (2004: p. 317) mention that “the two research programs would benefit from a unified framework”. Salt and Kitching (1992: p. 160) note that “migration is a phenomenon that has been artificially divided (...) between internal and international migration. Our conclusion is that this distinction is in many respects counterproductive”.

First of all, the meaning of “internal” migration is ambiguous when the investigation area is not a country (Milne 1991). Cushing and Poot (2004: p. 328) mention a study where migration from Australia to New Zealand is interpreted as internal migration. In some applications it might be of greater interest to distinguish between moves within a functional area like a metropolitan area and external moves (Cadwallader 1992). Second, both internal and international migration are themselves heterogeneous. To give an example, international migration of a German student to Austria is not subject to the same driving forces as international migration of an African asylum seeker. Third, within the European Union legal restrictions to international migration have virtually disappeared. Of course, despite the abandonment of legal restrictions there may still exist social restrictions such as different languages, values or customs (Champion 1992), but such social restrictions may also exist to migration within a country. Finally, internal and international migration are related to each other. International companies have created internal labor markets (Salt and Kitching 1992) and employers have in general become less reluctant to substitute domestic workers with foreign workers (Champion 1992). A relationship between international and internal migration also clearly exists in the housing market. Foreign immigration is often correlated with out-migration of natives, particularly in urban areas.

Other disaggregations of the dependent variable Some researchers have argued that one should disaggregate migration flows by classes of streams other than internal vs. international. Gordon (1991) differentiates between job-related migration, migration for family and environmental reasons, and migration related to housing and marriage. In their introductory chapter, Stillwell and Congdon (1991) mention approaches to distinguish between home-based and job-based moves. A problem with such distinctions is that migration data (especially from administrative sources) hardly provide the necessary information and one has to rely on proxies like the distance moved. In surveys some attempts have been made to ask migrants ex-post about the main reason for moving, but as Cooke (2013) argues, such data are often of poor validity because ex-post stated migration intentions are likely to be endogenous to the migration outcome.

Pendakur and Young (2013) distinguish between short-distance moves within a region and long-distance moves between regions, for the latter are more likely to be associated with a change not only in residence, but also in lifestyle. Owen and Green (1992) also distinguish inter-regional from intra-regional moves. One of their findings is that variation of migration rates by occupational class is larger for inter-regional moves and is explained by the greater importance of job-related migration among the higher social classes. Regarding labor migration, Gordon (1992) considers contracted vs. speculative moves, the former meaning that a new job is contracted before the worker changes residence. He mentions a study finding that speculative labor migration is more important for large countries like the United States than for small countries like the Netherlands, because of a lack of nationwide information in large countries. Salt and Kitching (1992) argue that labor market migration should be disaggregated by criteria such as same vs. different company or with vs. without assistance. From a lifetime perspective, referring to the region of birth, one can distinguish between primary, secondary (onward) and return migration (Boyle et al. 1998).

Stratified analysis Another issue in migration models is stratification by explanatory variables. In a regression framework, this means either to estimate separate models for sub-populations or to include a sufficient number of interaction terms. Cushing and Poot (2004) mention stratification as a technique to achieve an intermediate approach between micro and macro data analysis. The explanatory variables may or may not be the same for each stratum.

Congdon (1992) estimates separate regression equations for age groups and social classes. Greenwood (1975) considers different models for metropolitan and non-metropolitan areas. Pendakur and Young (2013) stratify their model by community size. Their smallest class contains places up to 10,000 inhabitants and is largely the size of communities that we have in mind when we speak of rural areas. Haurin and Haurin (1991) disaggregate by sex, finding that females are more responsive than males to regional differentials in wage levels. In American studies, a key disaggregation in almost any social sciences application is by race/ethnicity. In Europe, a distinction is often made between nationals and foreigners (Lamonica and Zagaglia 2013).

Congdon (1992) defines the “migration field” as the set of all possible destinations for a

particular sub-population. The migration field is assumed to depend on age as well as on occupational and educational skills. A key difference with education is that the highly educated participate in a national labor market, whereas the poorly educated participate in a local labor market. For those with intermediate educational level, the labor market is also national, but compared to the highly educated they are to a greater extent subject to housing constraints, meaning that their migration field is smaller. Stratification by educational or occupational group is particularly useful for models based on the human capital theory. As mentioned before, in aggregate analyses one has to summarize micro-level regressors, meaning that certain individuals behave “wrongly” in terms of the human capital theory. Disaggregation is a means to reduce that aggregation bias.

Endogenous regressors An important issue in aggregate migration models (but also in micro models, see Congdon 1992 and Dong 2009) is endogeneity of regressors. Endogeneity of a regressor may arise for different reasons (see Chapter 5). We already mentioned endogeneity in housing market variables and some possible remedies like using regional dummies. Adam (2009: p. 26) gives a German paper where unemployment and wage level are assumed to be endogenous to migration. Stillwell and Congdon (1991) mention that a strong economy may induce both high in-migration and a high employment rate of the residential population, so the employment rate is likely to be an endogenous regressor in in-migration models.

Congdon (1992) classifies endogeneity as one of the general problems of migration regression models. Informally speaking, it means that ordinary parameter estimates do not converge to the causal effects of interest. To what extent this calls for special estimators is likely to depend on the length of the observation period. If the period is short, then endogeneity is a smaller problem than if it is long (Cadwallader 1992).

A frequent strategy to overcome contemporaneous feedback from the dependent variable to a regressor is to use measurements taken at the beginning of the study period or even earlier. This is thought to map as best as possible the direction of cause and effect (Cadwallader 1992). Greenwood (1975) investigated the sensitivity of regression results when using predictor measurements taken at the end of period compared to at the beginning.

Lagged regressors Taking measurements of explanatory variables from previous periods is not only a promising approach in handling endogeneity, but may be required to estimate the actual effects of interest. As a response to signals e.g. in the labor market, migration can be expected to act with some delay. Maier and Weiss (1991) attribute this to diffusion of information, formation of expectations, and the time between decision and execution. In many cases the decision to migrate is made on the household rather than the individual level, probably increasing the time-lag between cause and effect. It may therefore be beneficial to consider lagged regressors. To the best of our knowledge however, lagged regressors have so far rarely been applied in aggregate migration analysis for reasons other than handling endogeneity.

Correlated regressors Most migration data from administrative sources provide only a limited amount of explanatory variables on the micro level—typically basic demographic variables such as age and sex. In macro models however, when the units of analysis are administrative areas like municipalities, one can often construct hybrid datasets containing many predictors obtained from different sources. For such datasets correlation between regressors is a major issue. Multicollinearity is problematic in numerical analysis and it also distorts a meaningful substantive interpretation of regression coefficients. Therefore, it has been recommended in aggregate migration modeling to first analyze the joint distribution of regressors (Flowerdew 1991).

Stillwell (1991) considers place attractiveness and suggests to combine various attractiveness variables into a composite measure. One technique to obtain such composite measures which is particularly useful when the variables are highly correlated is principal component analysis. An application, regarding internal migration in Italy, is given by Lamonica and Zagaglia (2013). They extracted two principal components, one summarizing economic and the other demographic characteristics of regions, with the former accounting for a greater proportion of variance. Cadwallader (1992) found strong correlations between family type and housing type variables as well as between housing quality and socio-economic status variables and therefore also applied principal component analysis.

A second strategy to deal with correlated regressors is model selection. Flowerdew (1991) mentions three criteria for the inclusion of a regressor out of a set of candidates in a model: First, the deviance when comparing blocks of regressors; second, the t-ratio for an individual regressor; and third (and arguably most important), the theoretical relevance of a regressor. If the model is not based on a completely specified density, then measures comparable to the deviance often exist. A model selection criterion which is based on information theory is the Akaike information criterion (AIC). A variant of the AIC is used for model selection in our application in Chapter 4.

Functional form Cushing and Poot (2004: p. 324) note that “the aggregate migration literature still seems to ignore the problem of finding the best functional form for the migration model”. As we discussed in Section 1.4, an exponential mean (or log-link) function is typically applied in macro models. The advantages of the exponential mean function are the interpretation of parameters as semi-elasticities (or elasticities, if regressors are logged) for continuous regressors, the multiplicative effects for binary regressors, and the non-negative predictions. The exponential mean function is an instance of the power family of transformations (Greene 2003: Ch. 9, Wooldridge 1997), which is indexed by a parameter. This parameter may be estimated jointly with the genuine regression parameters (Wooldridge 1997).

Transformation of regressors is an important issue when regressors have high variances (Frees 1992). For right-skewed predictors, taking the logarithm is a natural choice. Logged regressors are also attractive from an interpretational point of view because they allow parameters to be understood as elasticities. This is particularly useful when a regressor is some measure of exposure, like the person-years lived in a model of out-migration. Greenwood (1975)

takes logarithms of all regressors except the regional dummies. Congdon (1992) generally recommends to take logarithms. A problem with the logarithm is that it is not defined for zero.

Correlated errors A final point in regression model specification is the assumed correlation between the errors of different areas. In this thesis we work with the independence assumption implying uncorrelatedness. Our justification is that our model is designed for a study area with many small municipalities, so any correlation between two errors can be expected to be small. For other applications the independence assumption is more problematic. A reason for dependence is that in-migrants from outside the investigation area pick one out of several candidate destinations in the area. Also, any out-migration from a place is potential in-migration to another place. In Chapter 6 we mention some possible extensions of our basic model that allow for correlated errors.

3 The basic model

3.1 The Poisson process on the nonnegative real line

In Chapter 1 we discussed the exceptional importance of the Poisson distribution as a statistical model for demographic counts and for count data in general. This outstanding role can be explained in various ways (Cameron and Trivedi 1998: pp. 5-8). One of them is a powerful theoretical result, comparable to the Central Limit Theorem for the Gaussian distribution. It is called “the inevitability of the Poisson distribution” (Kingman 1993: p. 9) and arises from underlying assumptions of statistical independence.⁹ We will now sketch this result, based on the Poisson process. Although Poisson processes can be constructed in fairly general spaces (Kingman 1993: Ch. 2), for our purpose it is sufficient to consider the Poisson process on the nonnegative real line.

Given $(Y_t)_{t \geq 0}$ a stochastic count process in continuous time, suppose that the probability of occurrence of an event in $(t, t + h]$ is independent of Y_t . This is called the independent increments property. Suppose furthermore that the probabilities of zero and one occurrences in $(t, t + h]$ are, respectively,

$$\Pr(Y_{t+h} - Y_t = 0) = 1 - \lambda h + o(h), \quad (3.1)$$

$$\Pr(Y_{t+h} - Y_t = 1) = \lambda h + o(h). \quad (3.2)$$

This is called the stationary increments property, implying that the expected number of events in any interval depends only on the length of the interval. Then the number of events in any interval is “inevitably” Poisson distributed, with expectation equal to the rate parameter λ times the length of the interval. Proofs based on differential equations are given by Winkelmann (2010: pp. 11-13) for the nonnegative real line and by Kingman (1993: pp. 9-10) for more general spaces.

Thus, for a given rate λ , the Poisson distribution arises naturally from “complete randomness” or “pure chance” (Xekalaki 1983). For a regression model this means that all systematic variation in the number of events is covered by the conditional expectation, whereas the error arises purely from the randomness inherent in the Poisson process. Two outcomes with the same conditional expectation have the same conditional distribution.

Another key property of the Poisson distribution is its additivity. Given Y_1 and Y_2 two independent Poisson variates, in general with unequal expectations, define $Y := Y_1 + Y_2$. Then

⁹It should be mentioned that according to Flowerdew (1991), the use of the Poisson regression model in migration analyses did not originate from independence assumptions, but from the high frequency of zeros and small integers.

Y is again Poisson distributed, so the Poisson distribution is “closed under convolutions”. An easy proof based on probability generating functions is given by Winkelmann (2010: p. 9).

For aggregate migration regression models—we do not distinguish between in- and out-migration when our considerations are generic—the given properties have the following implications. Assume that conditional on observed explanatory variables x people migrate at constant rates, i.e. all systematic variation in the expected migration frequency is covered by a known function of x . Assume then that people migrate independent of each other and independent of previous events (as far as they are not part of x). Then, under such idealized conditions, the annual total of migrations is Poisson distributed.

Of course, the Poisson process is based on assumptions which may or may not hold in applications. Let us next discuss apparent violations of the Poisson process regarding aggregate migration counts.

3.2 Violations of the Poisson process regarding migration flows

3.2.1 Clustering by household migration

The assumption of people moving independently of each other is clearly violated, for migration usually happens in a household (or family) context. From a statistical viewpoint, this means that events happen in clusters. Clustering is an instance of “generalization” of a distribution (Gurland 1957, Winkelmann 2010: pp. 36-39). The cluster (household) sizes are positive integers, which implies overdispersion in the person migration count (McCullagh and Nelder 1989: Ch. 6). It should be noted that the household size distribution of the migrating households is in general not equal to the household size distribution of the staying households.

Clustering is an important issue for migration flows, much more than for vital events. Cooke (2013) mentions that more than 90 percent of all inter-county migration in the United States are among individuals who are either family members or non-family members who move for “family reasons”, which consists of change in marital status, establishment of own household, and other family reasons.

Households as unit of analysis It is sometimes asked whether the violation of the Poisson process caused by clustering of persons by household migration can be handled by modeling the number of migrating households instead of persons (Flowerdew and Aitkin 1982, Flowerdew and Boyle 1995). Though this argument is for itself correct, modeling the number of migrating households might not be the best choice for other reasons. First, in many instances households simply cannot be identified in the data. This is particularly often the case when the migration data is based on administrative records. Second, for most applications of migration research what is relevant is actually the number of migrating persons, so a household count model alone is insufficient. And third, even when migrating households are analyzed, the Poisson distribution can be violated in other ways than clustering. The typical situation in applications

is that clustering explains most, but by no means all overdispersion. As Congdon (1992: p. 145) observed, “the Poisson compounding effect due to the tendency for migrants to move as part of household units is not the only source of extravariation in the dataset concerned”.

3.2.2 Unobserved heterogeneity

Suppose that the expectation of a variate depends on two sets of regressors, x and z , of which x is observed and z is unobserved. Then “unobserved heterogeneity” means the model specification error that arises from observing only x .

Unobserved heterogeneity is an important issue in count data modeling because the Poisson regression model makes no allowance for it. Unlike the Gaussian distribution, the Poisson distribution does not contain a separate dispersion parameter; deviations of a random variable from its expectation arise purely from the randomness inherent in the Poisson process. For a regression model this implies that all systematic influences on migration frequency are included in x . On the contrary, if relevant regressors are unobserved, then the assumption of conditional Poisson distribution is violated. It can be shown that unobserved heterogeneity always results in overdispersion (Cameron and Trivedi 1998: Ch. 3, Winkelmann 2010: Ch. 5).

What is precisely meant with unobserved heterogeneity varies between applications. The most obvious case is when certain predictors (e.g., amenity variables) are not observed at all.¹⁰ A more subtle case is measurement errors in the observed predictors (Cameron and Trivedi 1998: Ch. 10). For example, in a demographic rate model (see Chapter 5), the number of counts is related to the number of person-years lived. Usually the person-years lived are however unknown and have to be approximated by some weighting of population counts. The deviation of the measured form the true person-years lived is an instance of unobserved heterogeneity. A third case of unobserved heterogeneity which is particularly relevant for macro data is aggregation bias. Many predictors of migration like income are genuine micro level variables. When the units of analysis are places, then such predictors can only be included by summary statistics like the median income. Aggregation bias also matters for the distance between two areas in modeling origin-to-destination-specific migration flows. Another source of unobserved heterogeneity, which is particularly important for small places, is the spatial concentration of institutional households such as old-age homes, prisons, or refugee camps (insofar as not covered the observed regressors).

3.2.3 Contagion

A third class of violations of the Poisson process occurs when the assumption of independence of the migration rate of previous events does not hold. This is called “contagion” and may occur in two ways (Congdon 1991, Winkelmann 2010: pp. 18-19). Occurrence-dependence, on the one hand, means that the migration rate increases or decreases erratically as a consequence

¹⁰In micro level analyses of mortality data, unobserved heterogeneity is usually interpreted as the “frailty” of individuals.

of a migration event. For instance, the migration rate of a person may be higher after a move than before, because of that move (positive occurrence-dependence). Duration-dependence, on the other hand, means that the migration rate is a function of the time elapsed since the last event. For instance, the migration rate of a person may be lower the longer he or she has resided at a place (negative duration-dependence).

It should be noted that not all kinds of non-stationarity of the migration rate necessarily violate the Poisson distribution. For example, seasonal patterns in the migration rate are compatible with the Poisson regression model (Winkelmann 2010: p. 14). The unique feature of contagion is that the migration rate depends on realized events. Contagion can result in overdispersion as well as underdispersion, but on balance overdispersion is more likely for migration counts. Most analyses found positive occurrence-dependence as well as negative duration-dependence (Cadwallader 1992, Fielding 1992: p. 206, Tervo 2001), both resulting in overdispersion. A notable exception is a lower migration rate after return migration, which is an instance of negative occurrence-dependence (DaVanzo 1976).

By its nature, contagion is a key issue in longitudinal micro analyses. In aggregate cross-sectional studies, where migration flows are observed for a short period like a year for an entire population, contagion is however of smaller relevance. The reason is that contagion in its proper sense covers only dependence on those events which happen during the observational period. In contrast, dependence of the migration rate on events in previous periods, which is more important, is covered either by the observed regressors or by unobserved heterogeneity. Therefore, one can expect that in cross-sectional analyses contagion in its proper sense is relatively unimportant. In Section 3.3 we neglect contagion as a separate influence on the conditional variance of a migration flow and concentrate on the other two given violations of the Poisson process, namely clustering of persons by household migration and unobserved heterogeneity between places.

3.3 A variance function with two parameters

Given the theoretical derivation of the Poisson distribution based on the Poisson process and the discussion of its violations in Sections 3.1 and 3.2, we will now derive a semiparametric model for aggregate cross-sectional migration counts. Precisely, we will derive the conditional mean and variance of a migration flow given its observed predictors. We start with an idealized situation resulting in a Poisson distribution and hence equidispersion. Then we relax assumptions by allowing first for unobserved heterogeneity between places and second for clustering of persons by household migration.

3.3.1 Mean and variance of the household count

Let us denote the (in general unobserved) household migration count by the random variable M . Clearly M is nonnegative integer-valued, so its probability mass lies on $\{0, 1, 2, \dots\}$. We

assume that its conditional expectation depends via the exponential mean function on a vector of observed regressors x , as well as on unobserved regressors z :

$$E(M|x, z) = \exp(x'\beta^* + z'\delta), \quad (3.3)$$

where β^* and δ are parameter vectors. The unobserved regressors z may be summarized by a scalar index $\xi = \exp(z'\delta)$, so we may re-write the conditional expectation as

$$E(M|x, \xi) = \exp(x'\beta^*) \xi. \quad (3.4)$$

Assuming that households move independently of each other (Flowerdew and Boyle 1995) and that contagion can be neglected in cross-sectional analyses, it is plausible to assume a Poisson distribution for M conditional on x and ξ . From that follows equidispersion:

$$E(M|x, \xi) = \text{Var}(M|x, \xi). \quad (3.5)$$

To obtain mean and variance of the household flow conditional on the observed regressors x only, we have to integrate out the unobserved ξ . In that respect we assume that ξ is mean-independent of x , i.e. $E(\xi|x) = E(\xi)$. This assumption is discussed in detail by Winkelmann (2010: Ch. 4). In principle, one could assume complete independence between x and ξ , however such an assumption would often be questionable on substantive grounds, e.g. when ξ contains measurement error in x . As long as the observed regressors x include a constant, there is no loss of generality in assuming $E(\xi) = 1$. Applying the law of total expectation, we get

$$\begin{aligned} E(M|x) &= E_\xi[E(M|x, \xi)] \\ &= E_\xi[\exp(x'\beta^*) \xi] \\ &= \exp(x'\beta^*) \times 1 \\ &= \exp(x'\beta^*). \end{aligned} \quad (3.6)$$

Furthermore we assume that $\text{Var}(\xi) := \gamma_1 < \infty$. Applying to law of total variance, we get

$$\begin{aligned} \text{Var}(M|x) &= E_\xi[\text{Var}(M|x, \xi)] + \text{Var}_\xi[E(M|x, \xi)] \\ &= E_\xi[E(M|x, \xi)] + \text{Var}_\xi[E(M|x, \xi)] \\ &= E_\xi[\exp(x'\beta^*) \xi] + \text{Var}_\xi[\exp(x'\beta^*) \xi] \\ &= E(M|x) + [E(M|x)]^2 \times \gamma_1. \end{aligned} \quad (3.7)$$

We see that the household flow M conditional on the observed regressors x is overdispersed and that its variance-to-mean ratio, which is unity for the Poisson distribution, equals

$$\frac{\text{Var}(M|x)}{E(M|x)} = 1 + \gamma_1 E(M|x) \quad (3.8)$$

If there is no unobserved heterogeneity, then $\text{Var}(\xi) = 0$ and the household migration count is equidispersed. The greater the amount of unobserved heterogeneity, measured by γ_1 , the greater the overdispersion.

3.3.2 Mean and variance of the person count

Suppose now that each migrating household contains a random variable C of migrating persons. Clearly, the probability mass of C lies on $\{1, 2, 3, \dots\}$, so $E(C) \geq 1$. We assume that the C 's for different households are identically distributed, with finite first two moments, and that they are independent of the household flow M . The person migration count, which is what we are ultimately interested in and denote by Y , is then a generalized (stopped-sum) variate:

$$Y := C_1 + C_2 + \dots + C_M, \quad (3.9)$$

where it is understood that a null sum is zero, i.e. $M = 0 \Rightarrow Y = 0$.

The expectation of the person migration count is then

$$\begin{aligned} E(Y) &= E_M[E(Y|M)] \\ &= E_M[E(C_1 + C_2 + \dots + C_M|M)] \\ &= E_M[M \times E(C)] \\ &= E(C) \times E(M) := \mu, \end{aligned} \quad (3.10)$$

and for the variance we get

$$\begin{aligned} \text{Var}(Y) &= E_M[\text{Var}(Y|M)] + \text{Var}_M[E(Y|M)] \\ &= E_M[\text{Var}(C_1 \dots + C_M|M)] + \text{Var}_M[E(C_1 \dots + C_M|M)] \\ &= E_M[M \times \text{Var}(C)] + \text{Var}_M[M \times E(C)] \\ &= \text{Var}(C) \times E(M) + [E(C)]^2 \times \text{Var}(M) \\ &= \text{Var}(C)E(M) + [E(C)]^2 \{E(M) + \gamma_1[E(M)]^2\} \\ &= E(M) \{ \text{Var}(C) + [E(C)]^2 \} + \gamma_1[E(C)E(M)]^2 \\ &= E(M)E(C^2) + \gamma_1\mu^2 \\ &= E(M)E(C) \frac{E(C^2)}{E(C)} + \gamma_1\mu^2 \\ &= \gamma_0\mu + \gamma_1\mu^2, \end{aligned} \quad (3.11)$$

where $\gamma_0 := [E(C^2)]/[E(C)]$ summarizes the household size distribution.

If all households are single person households, then $\gamma_0 = 1$ and the variance function has the functional form of (3.7). The actual value of γ_0 and hence the additional overdispersion introduced by clustering depends on the mean and the variance of the household size distribution. Note that γ_0 refers to the migrating households, which have in general a different size distribution than the staying households (Flowerdew and Boyle 1995).

Dependence of the person migration flow Y on the observed regressors x is introduced via dependence of the household flow, i.e.

$$Y|x = C_1 + C_2 + \dots + C_{M|x}, \quad (3.12)$$

so we assume that the household size distribution is independent of x . Effectively this means that the household size distribution is the same for each place. Flowerdew and Boyle (1995) discuss a relaxation of this assumption.

The mean household size $E(C)$ is absorbed by the intercept regression parameter,

$$E(Y|x) = E(C)E(M|x) = E(C) \exp(x' \beta^*) = \exp(x' \beta), \quad (3.13)$$

with β the same as β^* , except for the intercept component.

We see that the person migration count is overdispersed. Its variance-to-mean ratio equals

$$\frac{\text{Var}(Y|x)}{E(Y|x)} = \gamma_0 + \gamma_1 E(Y|x), \quad (3.14)$$

so overdispersion compared to the Poisson variance results from two components, namely clustering of persons by household migration, which is indexed by $\gamma_0 \geq 1$, and unobserved heterogeneity between places, which is indexed by $\gamma_1 \geq 0$. This two-parameter variance function is in accordance with Congdon (1992: p. 145) that “the Poisson compounding effect due to the tendency for migrants to move as part of household units is not the only source of extravariation.”

Special cases of the two-parameter variance function are the quasi Poisson variance function with $\gamma_1 = 0$ and the NB2 variance function with $\gamma_0 = 1$ (Wooldridge 1997, Cameron and Trivedi 1998: Ch. 3, Winkelmann 2010: Ch. 3). The two-parameter variance function has already been used by statisticians on several occasions (Cameron and Trivedi 1985: p. 26, McCullagh and Nelder 1989: pp. 352-353, Thall and Vail 1990). The only application in the migration literature that we are aware of is Congdon (1993), who used it as a generalization of more restrictive one-parameter variance functions.¹¹

To the best of the author’s knowledge, we are the first to give a theoretical derivation of this two-parameter variance function in the migration context. Our derivation is based on two violations of the Poisson process which have been found to be important in cross-sectional migration analyses, namely clustering of persons by household migration and unobserved heterogeneity between places. Part of the value added of this thesis is therefore that we provide a theoretical foundation of the two-parameter variance function.

Quasi negative binomial representation Let us now re-parameterize

$$\sigma^2 := \gamma_0, \quad k := \gamma_0 / \gamma_1, \quad (3.15)$$

provided that $\gamma_1 \neq 0$. So the conditional variance of the person migration count is

$$\text{Var}(Y|x) = \sigma^2(\mu + \mu^2/k), \quad (3.16)$$

which is for specified k the quasi negative binomial variance, with σ^2 the QML dispersion parameter and $\mu + \mu^2/k$ the variance function of the nominal NB2 distribution.

¹¹Congdon (1993) mentions another way of obtaining a (different) two-parameter variance function for migration flows, namely when in a hierarchical model the destinations are nested within the origins.

3.4 Parameter estimation

3.4.1 Mean function parameters

Suppose now that a sample of n observations is available. For the i 'th observation in the sample we assume that the conditional mean is correctly specified as

$$\mu_i = E(Y_i|x_i) = \exp(x_i' \beta) \quad (3.17)$$

and that the conditional variance is

$$\text{Var}(Y_i|x_i) = \gamma_0 \mu_i + \gamma_1 \mu_i^2 \quad (3.18)$$

or equivalently

$$\text{Var}(Y_i|x_i) = \sigma^2(\mu_i + \mu_i^2/k), \quad (3.19)$$

provided $\gamma_1 \neq 0$. Based on the considerations in Section 1.4, a consistent and asymptotically efficient estimation of the mean parameter vector β is now straightforward.

Infeasible QML Consider first the case when k is known. We know that for given k the negative binomial distribution is a LEF. The constant dispersion parameter σ^2 does not influence the relative variances of the observations in the sample and β may be estimated by solving the quasi NB2 estimating equations:

$$s_n(\hat{\beta}_{\text{QML}}) = -\frac{1}{n} \sum_{i=1}^n \frac{y_i - \exp(x_i' \hat{\beta}_{\text{QML}})}{1 + \exp(x_i' \hat{\beta}_{\text{QML}})/k} x_i := 0. \quad (3.20)$$

Compared to the Poisson QMLE, the observations with larger x values are relatively down-weighted by the NB2 QMLE, for their relative variances are larger in the NB2 than in the Poisson model. The Poisson QMLE is the limit of the NB2 QMLE for $k \rightarrow \infty$.

The VC of the NB2 QMLE may be estimated by

$$\widehat{\text{VC}}(\hat{\beta}_{\text{QML}}) = \frac{1}{n} \tilde{\sigma}^2 \left(\frac{1}{n} \sum_{i=1}^n \frac{\exp(x_i' \hat{\beta}_{\text{QML}})}{1 + \exp(x_i' \hat{\beta}_{\text{QML}})/k} x_i x_i' \right)^{-1}, \quad (3.21)$$

with $\tilde{\sigma}^2$ estimated auxiliary, i.e. outside the QML estimating equations. Typically, $\tilde{\sigma}^2$ is obtained by equating either the quasi-deviance or the generalized chi-square statistic to its expected value (McCullagh and Nelder 1989: Ch. 9, Congdon 1992).

Feasible QGPML Although the NB2 QML estimator has attractive properties—it is consistent and asymptotically efficient within the class of PML estimators—it is usually not a feasible estimator because k is in general unknown. Several approaches are possible in that situation. First, one could specify a value of k based on substantive grounds, e.g. from previous experience with similar data. Such an estimator is consistent, as for specified k the NB2 is

a LEF and the estimator therefore a PML estimator. However, such a PMLE is inefficient if the specified k is different from the true k . Another option is to estimate β and k jointly by NB2 ML. Such an estimator is however inconsistent even if conditional mean and variance are correctly specified if the observations are not truly negative binomial distributed. The most promising choice is the NB2 QGPML estimator which replaces the unknown k by a $\sqrt{n}$ -consistent estimator.

Suppose that a $\sqrt{n}$ -consistent $\tilde{k}$ is available. The NB2 QGPML estimating equations are then

$$s_n(\hat{\beta}_{\text{QGPML}}) = -\frac{1}{n} \sum_{i=1}^n \frac{y_i - \exp(x_i' \hat{\beta}_{\text{QGPML}})}{1 + \exp(x_i' \hat{\beta}_{\text{QGPML}}) / \tilde{k}} x_i := 0 \quad (3.22)$$

and can be solved by standard statistical software packages such as R, SAS or SPSS.

The VC of the NB2 QGPML estimator may be estimated by

$$\widehat{\text{VC}}(\hat{\beta}_{\text{QGPML}}) = \frac{1}{n} \tilde{\sigma}^2 \left(\frac{1}{n} \sum_{i=1}^n \frac{\exp(x_i' \hat{\beta}_{\text{QGPML}})}{1 + \exp(x_i' \hat{\beta}_{\text{QGPML}}) / \tilde{k}} x_i x_i' \right)^{-1}. \quad (3.23)$$

Often the auxiliary estimate $\tilde{\sigma}^2$ is obtained jointly with $\tilde{k}$. In large samples the NB2 QGPML estimator has the same properties as the NB2 QML estimator. This holds even if the auxiliary $\tilde{k}$ is not efficient for k , as long as it is $\sqrt{n}$ -consistent.

3.4.2 Variance function parameters

Although the QML dispersion parameter σ^2 (which is in our original derivation γ_0) can be set to unity in the mean function parameter estimator $\hat{\beta}_{\text{QGPML}}$ without altering its consistency, this is in general not true for variance function parameter estimation (Wooldridge 1997: p. 380). It is therefore recommended to estimate the two parameters of the variance function simultaneously. A consequence is that in QGPML estimation the generalized chi-square statistic is not necessarily equal to its expectation (Congdon 1993).

As long as we do not make any further assumptions on the data generating process, there is little alternative to basic regression or moment estimators. Suppose for a moment that the true expectations μ_i were known. Then an obvious choice would be to estimate the model

$$\frac{(y_i - \mu_i)^2}{\mu_i} = \gamma_0 + \gamma_1 \mu_i + \text{disturbance} \quad (3.24)$$

by a standard technique like OLS estimation. This is of course not a feasible estimator because the true μ_i 's are unknown. In applications they are replaced by $\sqrt{n}$ -consistent preliminary estimators $\tilde{\mu}_i$, which may be obtained e.g. by a PML estimator of β .¹² The estimable regression

¹²It is important to point again to the regularity conditions for likelihood-based inference. In particular, we assume implicitly that in large samples the difference between an expectation and its prediction goes to zero for any observation. This implies that the data are not dominated by a single observation (in our case a single large city). For rural municipalities, which are typically many and all of them rather small in size, this assumption seems reasonable.

model is then

$$\frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\mu}_i} = \gamma_0 + \gamma_1 \tilde{\mu}_i + \epsilon_i. \quad (3.25)$$

Note that the error term ϵ_i covers both the genuine disturbance and the sampling error of $\tilde{\mu}_i$. This model was estimated in a migration context by OLS (Congdon 1993). Estimates $\tilde{\sigma}^2$ and $\tilde{k}$ are then obtained by the analogy principle:

$$\tilde{\sigma}^2 := \tilde{\gamma}_0, \quad \tilde{k} := \tilde{\gamma}_0 / \tilde{\gamma}_1. \quad (3.26)$$

Consistency of $\tilde{\sigma}^2$ and $\tilde{k}$ follows from basic large sample theory. Also the rate of convergence of $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ translates to $\tilde{\sigma}^2$ and $\tilde{k}$ (Greene 2003: p. 904).

Robust estimation The ultimate justifications of QGPML estimation are asymptotic and since OLS estimation is $\sqrt{n}$ -consistent, there is no problem with OLS estimation of the variance function parameters regarding the large sample efficiency of $\hat{\beta}_{\text{QGPML}}$. However, since actual samples are always finite, one can ask whether problems are to be expected with OLS estimation in finite samples. In general, applied statisticians should focus on several criteria when comparing estimators. Rivers and Vuong (1988) mention asymptotic efficiency, small-sample properties, usefulness for testing, and computational convenience.

One deficiency of OLS estimation is that it is rather sensitive to points which deviate from the majority of the data (“outliers” and “leverage points”, respectively). This can be formalized by the breakdown point, which is a measure of robustness of a statistical estimator (Rousseeuw and Leroy 1987: pp. 9-10). The breakdown point indicates the minimum fraction of observations that have to be arbitrarily contaminated in order to introduce unlimited bias in estimation. It can easily be shown that for OLS estimation the breakdown point is $1/n$, which is very small.

Historically, “contamination” was primarily used to indicate actual typing or recording errors. Today, for migration data based on administrative sources this might not be a big problem. It may however be the case that certain observations have a high leverage in OLS estimation. The reason is that the expected flow μ_i is usually highly correlated with the population size of a place and population size distributions are well known to be right-skewed. For large cities, they usually follow a Pareto distribution and for smaller places a lognormal distribution, which can be formally justified by Gibrat’s law of proportionate effects (or proportionate growth in that case) (Eeckhout 2004). When using OLS estimation, one should therefore be aware that in finite samples the results can be heavily influenced by a few places. Therefore, we propose robust estimation.

A very robust estimator is the least median of squares estimator (Rousseeuw and Leroy 1987: Ch. 2), which has a breakdown point of roughly $1/2$. However, it cannot be used for our purposes because its rate of convergence is only $\sqrt[3]{n}$ instead of the required $\sqrt{n}$. An adequate choice is instead the least trimmed squares (LTS) estimator, which minimizes the sum of the smallest $n \times q$ squared residuals, with q the trimming parameter. Note that the residuals are first squared and then ordered. Common specifications of q are close to 1 (e.g., 0.90 or 0.95). Clearly OLS is the special case with $q = 1$.

Alternative estimation approaches Instead of modeling the variance-to-mean ratio, one could also model the variance directly. The equation for known expectations would be

$$(y_i - \mu_i)^2 = \gamma_0 \mu_i + \gamma_1 \mu_i^2 + \text{disturbance}, \quad (3.27)$$

and feasible estimation requires the μ_i 's to be replaced by consistent estimators. Regression estimators of that kind have been applied by Fleiss et al. (2003: p. 363) and by Winkelmann (2010: p. 93). Another method based on residuals is given by Bosquet and Boulhol (2010).

In a general semiparametric framework, Davidian and Carroll (1987) discuss alternative approaches such as accounting for leverage in the residuals, using absolute or logged-squared residuals, concentrated ML, and extended QML (Thaler 2009). They also mention sample standard deviations when replications are available. In another paper, Davidian and Carroll (1988) compare extended QML with methods based on squared residuals and recommend the latter.

Iterative estimation Another technique that is thought to improve the finite-sample performance of QGPML estimators is iterative estimation. It is recommended e.g. by Ziegler (2011: p. 85) and means to sequentially update estimates $\tilde{k}^{(l)}$ and $\hat{\beta}^{(l)}$, with $l = 1, \dots$, until either a convergence criterion is met or the number of iterations exceeds a pre-defined limit. The usual convergence criterion is that a distance between two successive estimates lies below a pre-defined threshold.

Iterative estimation is intuitively appealing, although it should be noted that there is no guarantee that it is superior in finite samples (see also Amemiya 1973). Furthermore, estimation becomes more sophisticated because one has to define the iteration criteria.

4 Application

4.1 The dataset

4.1.1 Investigation area

Burgenland is an Austrian federal province (in German “Bundesland”). Its NUTS2 code is AT11. It is located at the Eastern border of Austria, adjacent to Lower Austria, Styria, Slovakia (in the very North), Hungary, and Slovenia (in the very South). Its total population is about 280,000.

For historical reasons, Burgenland does not contain any large city and before the 1960’s it was only little industrialized. Traditionally it was an area of out-migration, particularly international emigration to the United States, but since the 1980’s net migration was overall positive.

Burgenland is partitioned into 171 municipalities. A municipality map is given in Figure 4.1. The largest municipality, Eisenstadt, is also the capital and has about 13,000 inhabitants. The smallest municipality, Tschanigraben, has less than 100 inhabitants. In the 1970’s many small municipalities were forced by law to unite to larger municipalities. Since then, the administrative subdivision has remained fairly stable.

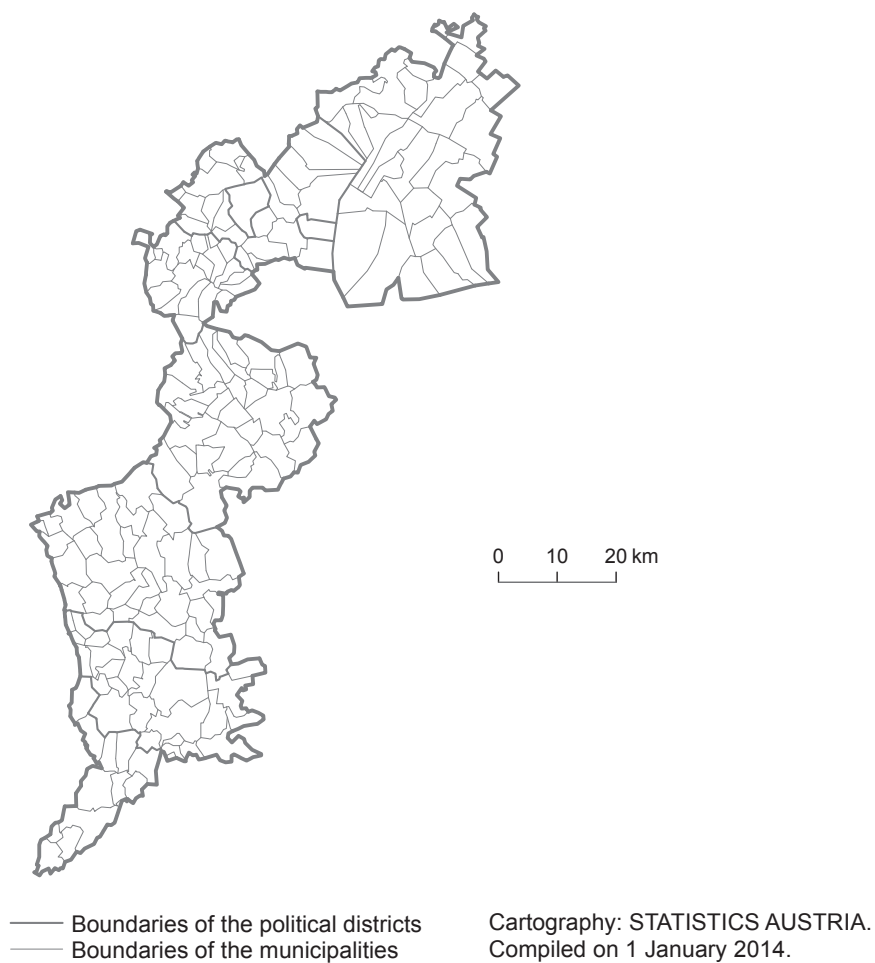
The landscape of Burgenland is flat and hilly, lacking high mountains. The predominant settlement structure is nucleated village and dispersed settlement is uncommon. Traditionally, a municipality was, with a few exceptions, identical with a village. Because of the mentioned forced administrative unions in the 1970’s, today many municipalities contain several villages.

4.1.2 Dependent variables

We separately analyze annual totals of in- and out-migration in 2012. Migration is defined as a change of residence in the Austrian population register which crosses municipal borders. The same person may migrate several times a year. In accordance with United Nations recommendations, international migration is counted only when the duration of continuous stay in Austria or abroad exceeds 90 days (not necessarily at the same address and not necessarily within a calendar year). In contrast, no minimum duration criterion is applied to migrations within Austria. For all of Burgenland, the following annual migration totals were observed in 2012:

Migration between municipalities of Burgenland	6,387
Arrivals from other municipalities within Austria	5,836
Departures to other municipalities within Austria	5,133

Figure 4.1: Municipality map of Burgenland
Municipalities of Burgenland, as of 1 January 2014



International immigration	3,455
International emigration	2,159

In Burgenland in 2012, there were also 6,355 changes of residence within municipal borders, which we do not count as migrations.

We see that migration to and from other parts of Austria is very important and also international migration accounts for a substantial amount of all flows. So any model covering only migration within the federal province would be of questionable relevance.

We see that overall net migration is positive for Burgenland, regarding both international flows and flows within Austria. Out of the 171 municipalities, 106 had positive, 60 negative and 5 zero net migration in 2012.

4.1.3 Candidate explanatory variables

In Section 2.2 we discussed various theories of migration. If one sets up a migration model based exclusively on one of those theories then the explanatory variables are given by the respective theory. In many situations, however, one is reluctant to choose one particular theory to explain a complex social phenomenon like migration. Indeed, it is often understood that we are living in a “post-modern” era where any form of all-encompassing theory is viewed with skepticism. Today most demographers might agree that although several migration theories do exist that properly explain some aspects of migration, finding an “ultimate truth” is a hopeless task. Moreover, even if explanatory variables are theoretically important, there is no guarantee in applications that measurements or useful proxies are actually available. To give an example, administrative data sources of migration often lack any information on the household context.

Since the model developed in the Chapter 3 is generic, one would, just for the purpose of empirical illustration, be relatively free in choosing regressors. Moreover, it is essentially up to domain experts and not primarily the statistician’s task to specify the predictors in a regression model. Nevertheless, the author has come to understand that concrete models often attract a lot of attention for themselves, even if they are intended just to illustrate theoretical derivations.

The approach in our application was as follows. First, a set of abstract regressors was developed based on the union of theoretical explanations for migration given in Section 2.2. Next we investigated for which of these abstract regressors measurements or useful proxies were available on a municipality level. We considered basic geographic and administrative variables, demographic and socio-economic characteristics, and variables covering potential substitutes for migration. Several theoretically important regressors had to be excluded because there was either no information available for municipalities (e.g., housing cost) or the available information was outdated (e.g., new housing). The remaining candidate regressors are given below.

NORTH Indicates municipalities located in the northern part of Burgenland (NUTS3 code AT112), as opposed to those located in the central or southern part (NUTS3 codes AT111, AT113). This variable is an aggregation of political districts and is meant to measure

housing cost, in the absence of any direct measurement. NORTH is binary with the following frequencies:

Northern part (NORTH = 1)	71 municipalities
Central or Southern part (NORTH = 0)	100 municipalities

SEAT Indicates administrative seats of political districts, comparable to county seats in the United States. Such municipalities are typically larger than average in population. They host administrative agencies and often also facilities such as secondary schools or hospitals. Altogether SEAT indicates local centers and is meant to index housing cost, location of institutional households, and certain amenities. SEAT has the following frequencies:

Administrative seat (SEAT = 1)	7 municipalities
Not an administrative seat (SEAT = 0)	164 municipalities

ELEVAT The geographic altitude of the municipality in meters above the sea level. The value is based on the “main township” of a municipality, mostly defined by the location of the church, town-hall or municipal office. ELEVAT is assumed to cover amenity variables. For Austria, considerable variation in demographic outcomes by geographic elevation has often been observed. Due to its right-skewed distribution, ELEVAT is logarithmized.

K Population size on 1 January 2012 (number of inhabitants). Due to its right-skewed distribution, and also to allow for useful interpretation of the regression coefficient, K is logarithmized.

AGE2034 The percentage of the population which was between 20 and 34 years old on 1 January 2012. Young adults are typically found to be a particularly migratory part of the population, which may be explained by the human capital theory as well as the life cycle theory. AGE2034 is logarithmized too, although its distribution on the original scale is rather symmetric.

FOREIGNB The percentage of the population which was foreign-born on 1 January 2012 (regardless of citizenship). The foreign-born population is usually found to be particularly migratory, which may be explained by weaker ties to the place of residence. Regarding in-migration, it may also represent existing networks of and hence the cost of immigration to foreign newcomers. FOREIGNB is strongly right-skewed and logarithmized.

DENSITY The ratio of the population on 1 January 2012 (see K) to the total of square kilometers of the settlement area. “Settlement area” includes building and industrial area, green space and cultivated land, but excludes forest and water area (Wonka 2008). DENSITY is heavily right-skewed and logarithmized. A squared term is added.

INCOME The median annual net personal income in 2010 in thousand euros taken from the Austrian integrated wage and income tax statistics (Milz 2013). “Income” means wages and salaries, income from self-employment, additional income from pensions, and certain social welfare benefits such as unemployment benefits and family allowances. So the median personal income may be understood as an index of wage level, unemployment rate, and the business situation of the self-employed. “Net” income means gross income minus social security contributions and taxes on personal income. Persons whose primary source of income is pension income are excluded from the calculation. At the time of writing, the 2010 data was the most recent available. INCOME is logarithmized, although its distribution on the original scale is rather symmetric.

EDUCAT The percentage of the population having attained at least higher secondary education which allows for university entrance (“Matura”), taken from the Austrian Census on 31 October 2011. The population under 15 years of age is excluded from the calculation. EDUCAT measures general human capital (as opposed to firm-specific human capital) (Jung and Winkelmann 1993) and is logarithmized.

OWNER The percentage of the population living in owner-occupied housing taken from the Austrian Census on 31 October 2011. The population not living in dwellings (living in institutions, hotels or caravans, or homeless) is excluded from the calculation.¹³ Theory suggests that those living in owner-occupied housing are less migratory because of higher cost of changing residence. Given the rural structure of Burgenland, the percentage of owner-occupied housing is very high, making up almost 80 percent for the entire federal province. In contrast to most other regressors, the distribution of OWNER is left-skewed on the original scale. OWNER is first ordinally categorized and then in the models treated as pseudo-metric. The following four levels are distinguished:

Class 0	69.9 percent or below	11 municipalities
Class 1	70.0-79.9 percent	24 municipalities
Class 2	80.0-89.9 percent	96 municipalities
Class 3	90.0 percent or above	40 municipalities

SINGLEDW The percentage of the population living in single-family dwelling units taken from the Austrian Census on 31 October 2011. Like for OWNER, the population not living in dwellings is excluded from the calculation. Theory suggests that those living in single-family dwelling units are less migratory because they find it easier to adapt their present housing to changes in the life cycle. In contrast to most other regressors, the distribution of SINGLEDW is left-skewed (though more symmetric than OWNER) on the original scale. SINGLEDW is first ordinally categorized and then in the models treated as pseudo-metric. The following four levels are distinguished:

¹³The excluded population makes up 1.3 percent of the total Burgenland population and no more than 7 percent in any municipality.

Class 0	69.9 percent or below	28 municipalities
Class 1	70.0-79.9 percent	51 municipalities
Class 2	80.0-89.9 percent	75 municipalities
Class 3	90.0 percent or above	17 municipalities

COMMUTE The average commuting time to Vienna (first municipal district) in minutes, provided by Statistics Austria. According to the Austrian Census 2011, a full 37 percent of the employed population of Burgenland is working in other federal provinces of Austria, with Vienna being the single most important workplace. As discussed in Section 2.2, commuting is often an alternative to migration. COMMUTE is logarithmized.

SECRES The ratio of secondary residences to 100 principal residences taken from the Austrian Census on 31 October 2011. Note that SECRES is not a proportion like AGE2034, but a ratio of two disjoint totals. The same person may have several secondary residences. SECRES is logarithmized.

4.1.4 Model selection

In total we have selected 15 candidate regressors (including the constant and the squared term of DENSITY) based on various migration theories and data availability. Given the 171 observations, this is a rather large number of explanatory variables. It is quite likely that some variables are included which, conditional on the remaining regressors, have very little explanatory power. Inclusion of such irrelevant regressors has several unpleasant consequences. First, it reduces the precision of the parameter estimates of the relevant regressors. Then, the model becomes less interpretable in substantive terms, especially when regressors are highly correlated. Finally, in some applications a large number of regressors may also cause numerical problems. Therefore it seems reasonable to apply a statistical model selection algorithm to our candidate regressors—all the more as our choice of candidate regressors was not based on a particular theory of migration, but on a union of different theories.

Quasi Akaike information criterion Statisticians have developed several model selection techniques. A very useful criterion is the Akaike information criterion (AIC) which is based on information theory. Basically, the idea is to minimize the expected Kullback-Leibler discrepancy¹⁴ between the unknown true data generating process and a statistical model over a set of candidate models (Lebreton et al. 1992, Anderson et al. 1994, Pan 2001). The true data generating process is irrelevant for comparison of models. In a fully parametric maximum likelihood application, one selects out of a set of candidate models, each indexed by a finite dimensional parameter vector, the one which minimizes the Akaike Information Criterion:

$$AIC = -2 \times \log\text{likelihood} + 2p. \quad (4.1)$$

¹⁴Some authors use the term Kullback-Leibler “distance”, yet “discrepancy” is a more adequate expression because of asymmetry (Anderson et al. 1994).

The number of regressors p is included as a correction term to account for the fact that the maximized loglikelihood is an upwardly biased estimator of the expected loglikelihood under the true model parameters (Pan 2001: p. 122). A small-sample second-order bias adjustment has been proposed if for the largest model $n/p < 40$ (as it is the case in our application). The corresponding criterion is the corrected AIC (AICc),

$$\text{AICc} = -2 \times \text{loglikelihood} + 2p + \frac{2p(p+1)}{n-p-1}. \quad (4.2)$$

Usually, one finds several candidate models with AIC(c) values very close to the minimum value—as a rule of thumb, models which exceed the minimum AIC(c) by less than 2 have substantial support if observations are independent (Lebreton et al. 1992, Anderson et al. 1994). So, for final model selection, it may be beneficial to give attention to several models with low AIC values and choose one of them based on substantive considerations.

Compared to other model selection criteria, an advantage of information-theory based criteria like the AIC is that they can be used to compare non-nested models. Moreover, they can be readily extended to quasi-likelihood models. The corresponding criteria are the quasi AIC (QAIC) and its corrected version (QAICc) in case of many regressors:

$$\text{QAIC} = -2 \times \text{quasi-loglikelihood} / \sigma^2 + 2p, \quad (4.3)$$

$$\text{QAICc} = -2 \times \text{quasi-loglikelihood} / \sigma^2 + 2p + \frac{2p(p+1)}{n-p-1}. \quad (4.4)$$

The only difference to the ordinary AIC(c) is the division of the quasi-loglikelihood by the dispersion parameter σ^2 . It has been suggested to estimate the dispersion parameter for the largest model and then enter this estimate $\tilde{\sigma}^2$ as a constant in model comparison. Since our model is not a QML, but a QGPML approach, we also had to estimate the LEF nuisance parameter k . In accordance with the suggestion just given, we estimated it for the largest model and then entered the estimate as a constant in what followed.¹⁵

Results All results were computed using R, version 2.14.1. We started with the full model, containing all 15 regressors. Then predictors (except for the intercept) were sequentially excluded in a “greedy” way. At each step of model selection we eliminated the predictor which resulted in the smallest local QAICc statistic. The procedure stopped when the QAICc could not be further reduced. Regarding in-migration, the following regressors were sequentially excluded:

	QAICc
Full model	566.91
– SECRES	564.52
– ELEVAT	562.24

¹⁵In our model selection σ^2 was taken to be 3.244 for in-migration and 2.162 for out-migration and k was taken to be 341.2 for in-migration and 512.5 for out-migration.

– OWNER	560.07
– SINGLEDW	557.89
– COMMUTE	556.17
– SEAT	554.51
– EDUCAT	553.72

Thus, besides the intercept the model for in-migration would contain as predictors NORTH, K, AGE2034, FOREIGNB, DENSITY, DENSITY², and INCOME. Regarding out-migration, model selection was as follows:

	QAICc
Full model	702.81
– NORTH	700.41
– SECRES	698.04
– DENSITY ²	695.74
– DENSITY	693.49
– OWNER	691.68
– SEAT	689.83
– EDUCAT	689.03

Thus, besides the intercept the model for out-migration would contain as predictors ELEVAT, K, AGE2034, FOREIGNB, INCOME, SINGLEDW, and COMMUTE.

For reasons of comparability, we decided to use the same final model for both counts, taking the union of all selected predictors. Thus, besides the intercept our final models contain as predictors NORTH, ELEVAT, K, AGE2034, FOREIGNB, DENSITY, DENSITY², INCOME, SINGLEDW, and COMMUTE.

Altogether, model selection supports some of the migration theories presented in Section 2.2 more than others. Interestingly, educational level (EDUCAT) was selected for neither flow. This makes Cadwallader's (1992) suggestion to always include an education variable look questionable. It could be that here educational level has actually little explanatory power because the models already contain income level or that it is poorly measured on an aggregate level. Also, the proportion of owner-occupied housing and the prevalence of secondary residences seem to be unimportant for our rural investigation area. It could be the case that these predictors work better for intra-urban migration. Moreover, some predictors seem to be relevant only for one of the two flows, such as COMMUTE for out-migration and NORTH for in-migration. This indicates that migration theories should be specific on which stream they actually refer to.

4.1.5 Descriptive statistics

Here we give some basic statistics of the variables included in our final models, except for the categorical variables NORTH and SINGLEDW for which the distribution was given above.

Table 4.1: **Descriptive statistics of variables on the original scale**

	Minimum	1 st quartile	Median	Mean	3 rd quartile	Maximum
y_1 (In-Migration)	2	39	60	91.7	96	1204
y_2 (Out-Migration)	5	38	55	80.0	85	958
ELEVAT	117	192.5	248	250.0	304.5	617
K	72	891	1287	1671	2006	13112
AGE2034	10.5	15.3	16.4	16.4	17.5	23.2
FOREIGNB	2.1	5.4	7.2	7.9	9.7	27.4
DENSITY	22	76.5	119	145.7	164.5	1139
INCOME	17.5	19.4	20.1	20.1	20.6	23.1
COMMUTE	29	41.5	60	64.3	84	121

Table 4.2: **Descriptive statistics of regressors on the logarithmic scale**

	Minimum	1 st quartile	Median	Mean	3 rd quartile	Maximum
ELEVAT	4.76	5.26	5.51	5.46	5.72	6.42
K	4.28	6.79	7.16	7.17	7.60	9.48
AGE2034	2.35	2.73	2.80	2.79	2.86	3.14
FOREIGNB	0.74	1.68	1.97	1.96	2.27	3.31
DENSITY	3.09	4.34	4.78	4.74	5.10	7.04
INCOME	2.86	2.97	3.00	3.00	3.03	3.14
COMMUTE	3.37	3.73	4.09	4.09	4.43	4.80

Table 4.1 summarizes the distributions on the original scale and Table 4.2 on the logarithmic scale, which is how the respective regressors enter the models.

We also give the correlation matrix for explanatory variables on the scale in which they enter the models (Table 4.3). It appears that multicollinearity is not a particular problem in our application; most correlations are rather weak. Only four out of the 36 Pearson correlation coefficients are greater than or equal to 0.50 in absolute value: between Northern location and commuting time to Vienna (negative), between Northern location and geographic elevation (negative), between population size and prevalence of single-family dwelling units (negative), and between income level and commuting time to Vienna (also negative). All of them are clearly understandable on substantive grounds.

Table 4.3: Correlation matrix of regressors

	NORTH	ELEVAT	K	AGE2034	FOREIGNB	DENSITY	INCOME	SINGLEDW	COMMUTE
NORTH	1.00								
ELEVAT	-.68	1.00							
K	.33	-.21	1.00						
AGE2034	.10	.09	.25	1.00					
FOREIGNB	.32	-.21	.28	.02	1.00				
DENSITY	.15	.29	.43	.34	.15	1.00			
INCOME	.45	-.20	.11	.00	.29	.11	1.00		
SINGLEDW	-.28	.16	-.60	-.26	-.17	-.44	.02	1.00	
COMMUTE	-.84	.47	-.38	-.19	-.32	-.29	-.58	.34	1.00

4.2 Estimation results

Following Chapter 3, we applied the NB2 QGPML estimator with exponential mean function. For the QGPML nuisance parameter—in this case the shape parameter of the quasi NB2 distribution—an initial value of $\tilde{k}^{(0)} = 50$ was specified. Given a specific value $\tilde{k}^{(l)}$, the estimate $\hat{\beta}^{(l)}$ was obtained by solving the NB2 QGPML estimation equations (3.22). Given a specific value $\hat{\beta}^{(l)}$, the estimates $\tilde{\gamma}^{(l+1)}$ and consequently $\tilde{\sigma}^{2(l+1)}$ and $\tilde{k}^{(l+1)}$ were obtained by the LTS estimator mentioned in Subsection 3.4.2. The trimming parameter q was chosen as

0.98. The procedure was iterated until

$$\frac{|\tilde{k}^{(l+1)} - \tilde{k}^{(l)}|}{\tilde{k}^{(l)}} < 0.01, \quad (4.5)$$

i.e. until two subsequent $\tilde{k}$ estimates did not differ by more than 1 percent in absolute value. In our application this meant 8 iterations for in-migration and 5 iterations for out-migration. If this criterion had not been met then the procedure would have stopped after 30 iterations. The estimated QGPML regression coefficients for both equations plus their t-ratios (estimate divided by estimated asymptotic standard error) and classes of the corresponding p-values are given in Table 4.4. The numbers were obtained by the R function *glm* with a nominal negative binomial specification with arguments *theta* ($\tilde{k}$) and *dispersion* ($\tilde{\sigma}^2$).

Table 4.4: **QGPML regression results for the basic model**

	In-migration			Out-migration		
	Estimate	t-ratio	p-value	Estimate	t-ratio	p-value
CONSTANT	-9.790			-12.205		
NORTH	-0.214	-2.48	*	-0.013	-0.19	
ELEVAT	0.010	0.11		0.169	2.57	*
K	0.982	26.40	***	0.954	32.30	***
AGE2034	0.694	3.85	***	0.892	5.82	***
FOREIGNB	0.476	10.56	***	0.413	11.08	***
DENSITY	0.359	1.36		0.070	0.33	
DENSITY ²	-0.032	-1.22		-0.005	-0.26	
INCOME	1.054	2.01	*	1.219	2.86	**
SINGLEDW	0.008	0.29		-0.036	-1.58	
COMMUTE	-0.011	-0.10		0.343	3.92	***
γ_0	1.807			2.179		
γ_1	0.0273			0.0044		
σ^2	1.807			2.179		
k	66.2			493.4		

* P<0.05 ** P<0.01 *** P<0.001

For both in- and out-migration, the regression coefficients for the proportions of the population 20-34 years old and foreign-born are positive and highly significant. The same holds, albeit

with a smaller t-ratio, for the income level. Clearly these variables indicate migration-prone populations. To give an example, a 1 percent increase in the median net personal income (that is a difference in annual income of around 200 euros) means a 1.1 percent increase in the expected in-migration count and a 1.2 percent increase in the expected out-migration count, respectively, if all other predictors remain the same. The regression coefficients of logged 1 January population count are highly significant compared to a null value of zero. A more interesting test is whether these coefficients are significantly different from 1. The corresponding t-ratios are -0.49 for in-migration and -1.55 for out-migration, so given the remaining regressors the null hypothesis of unit elasticity of migration flow to 1 January population count cannot be rejected.

That several regressors influence in- and out-migration in a similar way has also been found by Galle et al. (1993) who concluded that often predictors intended to explain a specific migration stream or net migration effectively explain the migration volume. Inspired by Mayer and Chappell (1992), we classify a predictor, according to sign and significance of its regression parameters, as follows:

	In-migration	Out-migration
Predictor of migration volume	+/-	+/-
Predictor of net migration	+/-	-/+
Predictor of in-migration	+/-	insignificant
Predictor of out-migration	insignificant	+/-

According to this classification, the proportion of young adults and of foreign-born population as well as personal income level are predictors of migration volume. This supports Galle et al. (1993) that the assumption that factors influencing migration are either “push” or “pull” factors has little empirical basis.

Interestingly, no predictor of net migration can be found in our application. Some variables are insignificant in either equation, namely population density (both the linear and the squared term) and the population living in single-family dwelling units. Looking at the parameter point estimates, the estimated relationship between expected in-migration and population density is inversely U-shaped and a greater proportion of the population living in single-family dwelling units reduces out-migration.

From the author’s point of view, the most interesting findings regard those variables which are significant in only one of the two equations. First, for Northern municipalities we observe a significantly negative coefficient in the in-migration model, whereas the parameter estimate for out-migration is statistically zero. Holding other factors constant, Northern municipalities have on average 19 percent less in-migration than Central or Southern municipalities.¹⁶ Given that NORTH is meant as a proxy for housing cost, and that housing cost matters most for potential in-migrants, this finding seems plausible.

¹⁶This percentage was calculated by $\exp(-0.214) = 0.81$. Note that for a binary predictor it does not make sense to interpret the regression parameter as semi-elasticity. Instead, we interpret the exponentiated parameter as incidence rate ratio.

Then, a significant influence on out-migration, but not on in-migration is found for geographic elevation. Holding other factors constant, a 40 percent increase in altitude (that corresponds for example to 350 vs. 250 meters above the sea level) increases the expected out-migration flow by around 6 percent. Although Burgenland does not contain high mountains, it seems that already the population living in hilly areas is somewhat disadvantaged (e.g., regarding infrastructure).

Finally, commuting time to Vienna is a highly significant predictor in the out-migration model, but statistically unrelated to in-migration. Holding other factors constant, a municipality has on average 9 percent less out-migration when commuting time is 45 minutes instead of 60 minutes (thus 25 percent shorter). This points to the importance of commuting as a substitute for out-migration. For those living in municipalities close to Vienna, commuting is a natural response to job and study opportunities in the Vienna area, because the costs of commuting may be more than balanced by lower housing costs and family ties to the municipality or residence. For those living in remote municipalities however, commuting is less attractive and they are better off with out-migration.¹⁷

Variance function parameters Although in our setting the variance function parameters are primarily seen as nuisance parameters, it might nevertheless be interesting to compare the estimated values with our theoretical considerations in Section 3.3. We see that for both in- and out-migration flow the parameter estimates support the theoretical relations $\gamma_0 \geq 1$ and $\gamma_1 \geq 0$. Comparing the estimates across the two regression equations, we see that $\tilde{\gamma}_0$ is larger for out-migration (2.2 vs. 1.8). In the context of our derivation of the variance function, this means that the household size distribution is statistically larger for out-migrating than in-migrating households. On the contrary, the parameter estimate $\tilde{\gamma}_1$ is much larger for in-migration. In our derivation γ_1 denoted the variance of ξ , the index of unobserved heterogeneity with expectation 1. The square root of $\tilde{\gamma}_1$ is therefore the estimated coefficient of variation of ξ . We find a coefficient of variation of around 17 percent for in-migration but only around 6 percent for out-migration, so it seems that overall the observed heterogeneity (i.e., the included regressors) accounts for a greater share of variation in out-migration than in-migration.

Variance-to-mean ratios Finally, we give estimated variance-to-mean ratios for some hypothetical mean values (Table 4.5). For the smallest municipalities in our sample with around 100 inhabitants, the expected annual migration flows are around 5. For such municipalities the conditional variance is around two times as large as the conditional mean, for both flows. Except for the capital Eisenstadt, the largest municipalities in our sample have a little more than 7,000 inhabitants, implying expected annual migration flows of close to 500. For such municipalities the variance-to-mean ratio is already around 4 for out-migration and even 15 for in-migration.

¹⁷In this thesis we do not care about the destinations of out-migrants. It is however evident from other sources that out-migration from Burgenland is extensively directed towards the Vienna area.

Table 4.5: **Estimated variance-to-mean ratios**

Mean	Variance-to-mean-ratio	
	In-migration	Out-migration
5	1.9	2.2
10	2.1	2.2
20	2.4	2.3
50	3.2	2.4
100	4.5	2.6
200	7.3	3.1
500	15.5	4.4

4.3 Model evaluation

4.3.1 Mean function

First, we evaluate the exponential mean function. We have to distinguish between systematic errors and isolated departures (McCullagh and Nelder 1989: Ch. 12). A systematic error arises for example if we assume an exponential mean function when the true mean function is linear. Isolated departures means that for most observations the model is indeed correctly specified, but some observations which in fact belong to different target populations have incidentally slipped into the sample. A typical example of an isolated departure in migration applications would be if the sample included a single large city to which specific social laws of migration are relevant which do not apply to the remaining smaller municipalities.

A misspecified mean function can best be detected graphically via residual plots. This means that residuals are plotted against either particular regressors or the linear predictor (here the logged expectations). A major advantage of residual plots is that they allow for detection of both systematic errors and isolated departures. Here we use for evaluation the Pearson residual, which is defined as

$$(y_i - \hat{\mu}_i) / \sqrt{\tilde{\gamma}_0 \hat{\mu}_i + \tilde{\gamma}_1 \hat{\mu}_i^2}. \quad (4.6)$$

Alternative residuals are discussed by McCullagh and Nelder (1989: Ch. 2) and by Cameron and Trivedi (1998: Ch. 5).

The Pearson residual plots for in-migration (Figure 4.2) do not indicate any systematic error. There is however an isolated departure, namely a particularly large in-migration residual for the municipality with the highest proportion of foreign-born population. This is the municipality Kittsee, which is located at the Slovakian border next to the Slovakian capital Bratislava. A detailed inspection of that municipality revealed that its population increased by about 10 percent during the year 2012. Moreover, similar population increases were recorded during the years before, with almost the entire population growth attributable to non-Austrian nationals.

According to domain experts, this special case is driven by international suburbanization of Bratislava.

Regarding out-migration (Figure 4.3), no isolated departure can be found. Yet there seems to be some systematic error in that the municipalities with the smallest populations have rather small Pearson residuals. This indicates some misspecification of the variance function, which we will discuss next.

4.3.2 Variance function

At first glance, one can evaluate also the assumed variance function (VF) graphically via the residual plots given before (McCullagh and Nelder 1989: pp. 331-332). A general problem with that approach is that for the human eye a deviation from a nonlinear function is difficult to detect. We therefore decided to evaluate the VF in an alternative way, namely by out-of-sample prediction intervals (Davidian and Carroll 1987).¹⁸

Consider an observation (x_0, y_0) not included in the sample. The predicted value for this observation is

$$\hat{y}_0 = \hat{\mu}_0 = \exp(x_0' \hat{\beta}), \quad (4.7)$$

and the out-of-sample prediction error is

$$\hat{e}_0 = y_0 - \hat{y}_0 = y_0 - \hat{\mu}_0. \quad (4.8)$$

If the assumed VF is correct, then the prediction variance may be consistently estimated by

$$\widehat{\text{Var}}(\hat{e}_0) = \tilde{\gamma}_0 \hat{\mu}_0 + \tilde{\gamma}_1 \hat{\mu}_0^2 + x_0' \widehat{\text{VC}}(\hat{\beta}) x_0 \hat{\mu}_0^2 \quad (4.9)$$

(Cameron and Trivedi 1998: p. 84), where the first two terms reflect the distribution of y_0 around its expectation and the third term reflects the prediction uncertainty of $\hat{\mu}_0$ caused by variation in $\hat{\beta}$. The $\hat{\mu}_0^2$ at the very right results from the delta method applied to $\hat{\beta}$ in the exponential mean function (Wolter 1985: Ch. 6).

The out-of sample prediction error can be normalized by the T_0 -statistic

$$T_0 = \frac{|\hat{e}_0|}{\sqrt{\widehat{\text{Var}}(\hat{e}_0)}}. \quad (4.10)$$

¹⁸Some authors (Delgado and Kniesner 1997, Wooldridge 1997) suggest to evaluate the VF by comparing the simplified QML VC estimate with the robust sandwich (RS) VC estimate. The reasoning is that the RS estimator is always consistent, so in case of a correctly specified VF there should not be a large difference between the two estimates and conversely a large difference between the two estimates is believed to indicate a misspecified VF. However, a comparison of VC estimates is problematic for two reasons: First, comparing large matrices is rather complicated, so one would effectively have to reduce the comparison to summary measures which may hide actual discrepancies. Second, as we have noted in Section 1.4, the RS estimator can be very inefficient (Kauermann and Carroll 2001, Ziegler 2011), so deviations between the estimates may simply reflect a lack of precision of the RS estimate. In fact, if the RS estimator were a reliable measure for evaluation, then there would be no need at all to consider simplified VC estimators.

Figure 4.2: In-migration residual plots

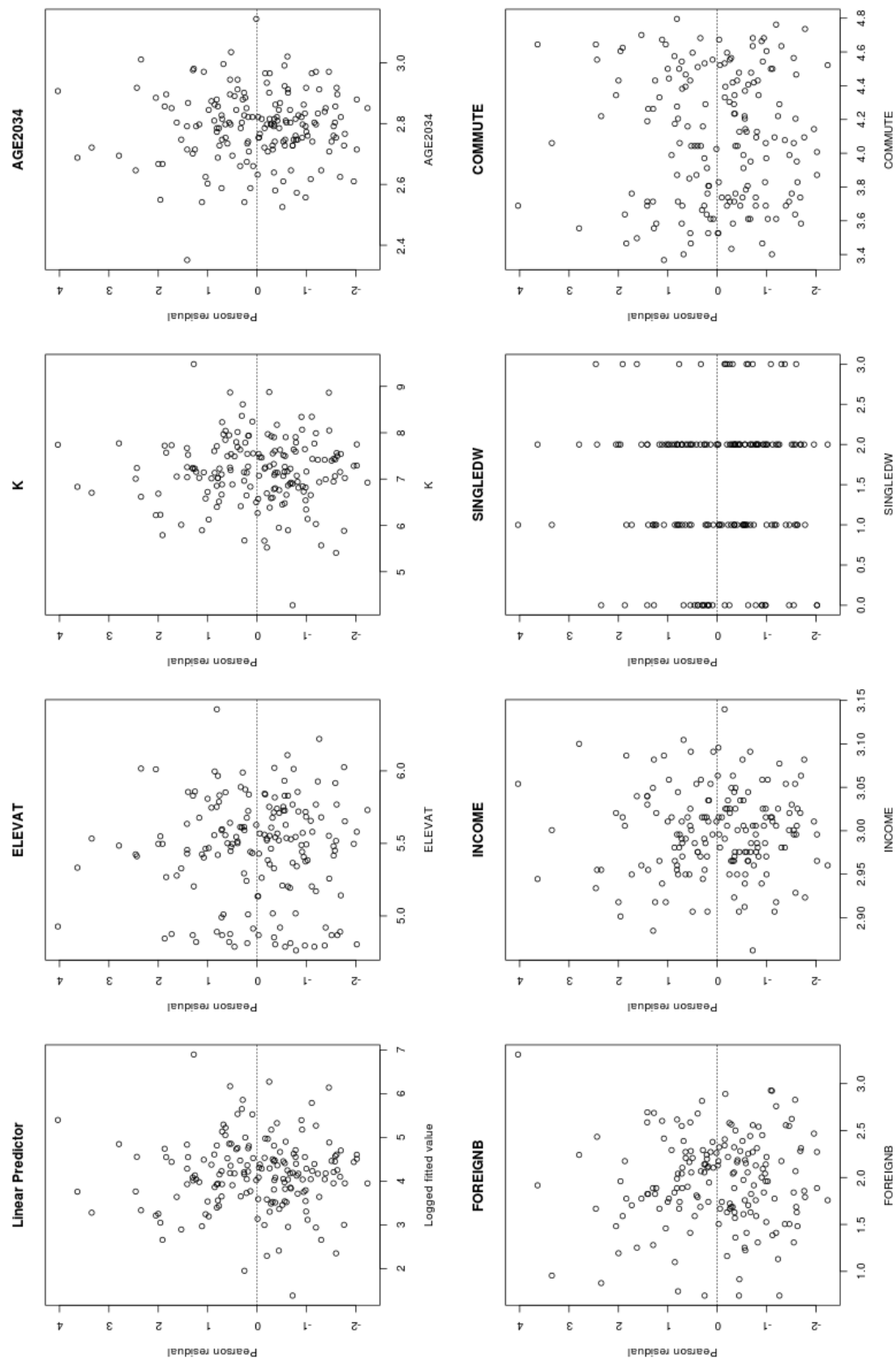
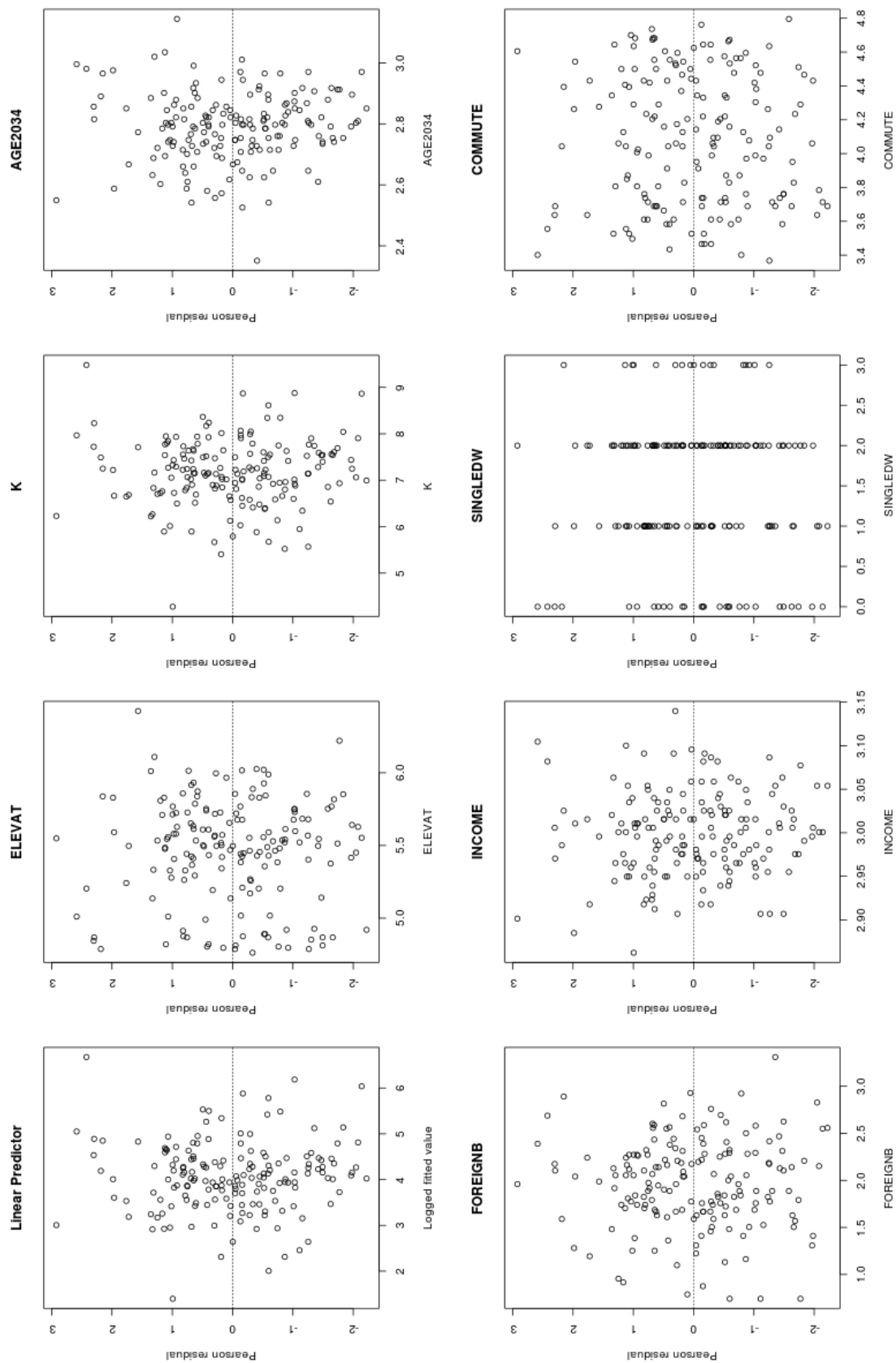


Figure 4.3: Out-migration residual plots



If the assumed VF is correct, then the central tendency of this T_0 -statistic does not depend on the expectation μ_0 . On the contrary, if the assumed VF is incorrect, then the mean value of T_0 depends on μ_0 , for then the true variance is overestimated for some expectations and underestimated for other expectations.

For each observation in the sample, an out-of-resample T_0 statistic was computed via Jack-knife resampling. This means that for the i 'th observation we used the resample containing the remaining $n - 1$ observations to estimate the model parameters and then compute the out-of-resample predicted value, prediction error and T_0 statistic. The 171 T_0 values are plotted against the logged out-of-resample predictions in Figure 4.4. A nonparametric Nadaraya-Watson kernel regression estimate (Härdle et al. 2004: Ch. 4) was fitted to the scatterplot data via the R function *ksmooth*. We used a Gaussian kernel with three different bandwidths: 1 (thin line, too rough), 2 (thin line, too smooth), and 1.5 (thick line, appears best).

We see that for in-migration the estimated kernel regression functions on the whole run horizontally, with no general upward or downward trend. Concerning out-migration, it seems that the T_0 average increases from around $\log \hat{\mu}_0 = 5$ onwards. That limit corresponds to a predicted annual out-migration count of around 150 implying a municipality population of roughly 3,000. It seems that our model variance underestimates the true variance of out-migration for the largest municipalities in our sample.

4.3.3 Sensitivity analysis: Alternative PML point estimates

Finally, let us look what mean function parameter estimates we would have obtained if we had applied alternative estimators. For that purpose we compare the point estimates across the following PML model distributions, all with exponential mean function: Gaussian (constant variance), Poisson (constant variance-to-mean ratio), negative binomial with specified $k = 50$, and Gamma (constant coefficient of variation). The Gaussian PMLE equals the homoskedastic nonlinear least squares estimator (Greene 2003: Ch. 9). The Poisson PMLE is a standard choice for count data (Wooldridge 1997, Winkelmann 2010: Ch. 3). The NB2 PMLE with $k = 50$ is the initial specification in our iterative QGPML procedure. The Gamma PMLE together with the exponential mean function implies that all municipalities receive the same GLM weight in the estimation algorithm (McCullagh and Nelder 1989: p. 40).

For comparative reasons we repeat in the tables our NB2 QGPML point estimates from before. We also give the asymptotic NB2 QGPML standard errors (S.E.s), meant as a device to roughly assess the difference of point estimates between models. It should be mentioned that if one were interested in PML standard errors then one would not be allowed to use the ordinary output of R's *glm* function, but had to apply either sandwich estimators or estimators based on resampling.

For in-migration (Table 4.6) we see that the coefficients of NORTH are all negative and of similar magnitude. Regarding logged 1 January population all four estimates are very close to 1. Positive coefficients are obtained across all models for the demographic variables regarding young adults and foreign-born population. This holds also for income level, albeit there the variation in coefficients between the models is relatively large. With respect to population

Figure 4.4: Graphical evaluation of the variance function

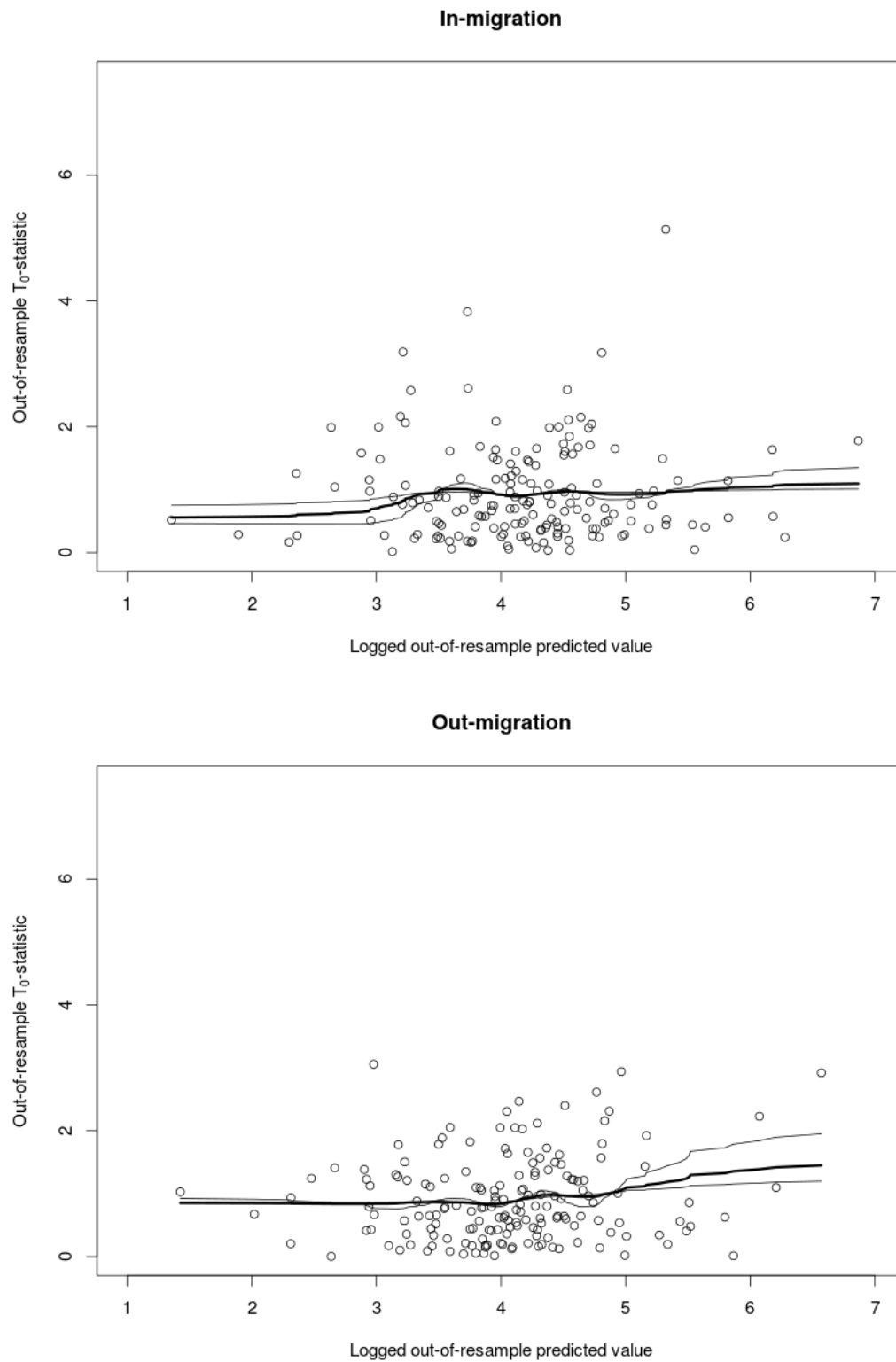


Table 4.6: **Alternative PML estimates for the in-migration model**

	NB2 QGPML		PML point estimates for various LEFs			
	Estimate	(S.E.)	Gaussian	Poisson	NB ($k=50$)	Gamma
CONSTANT	-9.790	(2.092)	-17.076	-13.090	-9.500	-8.583
NORTH	-0.214	(0.086)	-0.328	-0.227	-0.218	-0.252
ELEVAT	0.010	(0.084)	-0.216	-0.061	0.011	0.004
K	0.982	(0.037)	0.973	0.985	0.983	1.017
AGE2034	0.694	(0.180)	0.681	0.816	0.677	0.539
FOREIGNB	0.476	(0.045)	0.597	0.506	0.471	0.425
DENSITY	0.359	(0.264)	0.567	0.428	0.343	0.184
DENSITY ²	-0.032	(0.027)	-0.047	-0.037	-0.031	-0.016
INCOME	1.054	(0.524)	3.394	1.958	0.997	0.957
SINGLEDW	0.008	(0.028)	0.017	0.011	0.008	0.001
COMMUTE	-0.011	(0.108)	0.159	0.072	-0.019	-0.041

density, we observe a positive sign for the linear and a negative sign for the quadratic term, so altogether an inverse U-shaped relationship, in each model. The estimates regarding single-family dwelling units and commuting time to Vienna are small in any model. As regards geographic elevation, the pseudo Gaussian model provides a somewhat negative estimate.

For out-migration (Table 4.7) the findings are essentially similar. Coefficients of AGE2034, FOREIGNB and COMMUTE are robust to nominal LEF specification. A relatively large variation in estimates is found for INCOME. Generally, small effect sizes are estimated for DENSITY and SINGLEDW. Regarding the population size variable K, unit elasticity seems doubtful for the pseudo Gamma model estimate. In the pseudo Gaussian model, the estimated coefficient of ELEVAT is smaller than in our NB2 QGPML model.

Table 4.7: **Alternative PML estimates for the out-migration model**

	NB2 QGPML		PML point estimates for various LEFs			
	Estimate	(S.E.)	Gaussian	Poisson	NB ($k=50$)	Gamma
CONSTANT	-12.205	(1.688)	-17.494	-13.335	-10.770	-9.481
NORTH	-0.013	(0.070)	-0.141	-0.030	0.006	0.014
ELEVAT	0.169	(0.066)	0.055	0.142	0.202	0.228
K	0.954	(0.030)	0.984	0.965	0.946	0.937
AGE2034	0.892	(0.153)	0.955	0.929	0.806	0.728
FOREIGNB	0.413	(0.037)	0.355	0.400	0.430	0.461
DENSITY	0.070	(0.212)	-0.170	0.022	0.110	0.133
DENSITY ²	-0.005	(0.021)	0.020	0.000	-0.010	-0.014
INCOME	1.219	(0.426)	3.252	1.611	0.763	0.388
SINGLEDW	-0.036	(0.023)	-0.026	-0.033	-0.039	-0.039
COMMUTE	0.343	(0.088)	0.368	0.355	0.329	0.303

5 The rate model

5.1 Simultaneity of demographic rates

5.1.1 Endogenous regressors

So far we have assumed—at least implicitly—that our observed regressors x are exogenous. Informally this means that the regressors are either fixed quantities in repeated samples (like NORTH in our application), or if they are stochastic, then the data generating process of x can be separated from the data generating process of y conditional on x . A regressor which is not exogenous is called endogenous.

The formal definition of exogeneity vs. endogeneity is a topic of discussion among applied statisticians and econometricians (Greene 2003: pp. 381-382, Winkelmann 2010: Ch. 5). For our purposes the most promising definition relies on misspecification of the conditional mean function. Remember that our generic model contains an implicit error term

$$e = y - \mu = y - \exp(x'\beta). \quad (5.1)$$

The explanatory variables contain one or more endogenous components if

$$E(e|x) \neq 0, \quad (5.2)$$

i.e. if the error term is not mean-independent of x . In that case it follows that

$$E(Y|x) = \exp(x'\beta) + E(e|x) \neq \exp(x'\beta), \quad (5.3)$$

so the conditional mean is misspecified. Of course, this does not at all rule out that unconditionally $E(e) = 0$. An alternative definition of endogeneity, which we do not apply in this thesis, is based on parameter ancillarity (Cameron and Trivedi 1998: Ch. 11, Winkelmann 2010: Ch. 5).

Endogeneity of regressors is a key issue in any social sciences application, because they rely on observational data rather than controlled experiments, so correlation does not for itself indicate causation. As Blundell and Powell (2001) put it, “the analysis of data with endogenous regressors—that is, observable explanatory variables that are correlated with unobservable error terms—is arguably the main contribution of econometrics to statistical science.” Some reasons for endogeneity are (Cameron and Trivedi 1998: Ch. 11, Hilbe 2011: Ch. 13):

- In a model with unobserved heterogeneity, an observed regressor is correlated with the index of the unobserved predictors.

- A regressor is subject to measurement error and this measurement error is correlated with the dependent variable.
- There is contemporaneous feedback from the dependent variable to a regressor, so both the dependent variable and the regressor are simultaneously determined.

Regarding measurement error, it should be mentioned that “exogenous” measurement errors, meaning that they are uncorrelated with the dependent variable, can be treated as unobserved heterogeneity and need no special endogeneity handling (Cameron and Trivedi 1998: Ch. 10).

The consequences of endogeneity are fundamental. Since $E(e|x) \neq 0$, it follows that

$$E(Y|x) \neq \exp(x'\beta). \quad (5.4)$$

In other words, endogeneity means that the conditional mean function is not correctly specified. But an incorrectly specified conditional mean function implies that consistency of PML estimators breaks down (see Section 1.4). Even if a PMLE converges to some element of the parameter space, this element is not the parameter value that we are interested in because it mixes up the exogenous with the endogenous effect. So in case of endogenous explanatory variables, the NB2 QGPML estimator (3.22) loses its most important property.

Now a general problem in the social sciences is that it is often unclear whether a regressor is exogenous or endogenous. In a multiple regression model with many explanatory variables, like the one we have applied in Chapter 4, there is often a number of candidates for endogeneity. Regarding migration, Adam (2009) argues that housing cost is endogenous in particular to in-migration, and one strategy to overcome this is to replace the housing cost variables with nonstochastic regional dummies. He also mentions (p. 26) a German paper where unemployment and wage level are assumed to be endogenous to migration and handled with lagged (past-year) values as instruments. Whether a regressor is exogenous or endogenous may also differ between cross-sectional and dynamic models. It is therefore difficult to give any general recommendations on handling endogeneity.

Here we focus on one particular endogenous regressor, namely the total of person-years lived in a demographic rate model. We first will discuss the concept of a demographic rate and how this can be transformed to a count data regression model. Then we will show that there is simultaneity between the demographic count and the person-years lived and that two-stage residual inclusion (2SRI) is a proper method to handle it. We believe that 2SRI, which has to the best of the author’s knowledge not been applied so far to demographic rate models, is in some respects superior to other methods of handling simultaneity that have been traditionally applied by demographers.

5.1.2 Demographic rates

A demographic rate is the ratio of a demographic event count (e.g., migrations) to the total of person-years lived in the same population during a period. For example, the annual crude out-migration rate of a municipality is the annual total of out-migrations divided by the annual

total of person-years lived. If, as usual, the period equals one calendar year then the total of person-years lived is equivalent to the annual average population size and we sometimes use “population size” as an alternative expression. It is important to understand that a demographic rate is not the same as a relative frequency, for its numerator is not a subset of its denominator. Detailed treatises on demographic rates and their proper interpretation are given by Elandt-Johnson and Johnson (1980: Ch. 2) and by Chiang (1984: Ch. 4).

A demographic rate is called a crude rate if it is calculated for an entire population without disaggregation by age group or other variables which are relevant for the demographic process of interest. On the contrary, specific rates, e.g. age-specific death rates, restrict both the number of events and the total of person-years lived to certain subpopulations. In this thesis we are interested in crude migration rates. For any municipality, we may divide the annual migration counts by the total of person-years lived to get the crude in- and out-migration rate. In many demographic applications, such as population projections or comparative demography, it is these rates rather than the counts themselves that are of primary interest.

Person-years lived In theory, one calculates the total of person-years by adding the times lived in that municipality for all individuals who ever resided there during the year of interest. These times would for any individual in question be greater than zero and at most one year. In practice, individual life histories are usually unknown to demographers, so person-years lived have to be estimated. Typically this happens by taking some average of known population counts. In English-speaking countries, the usual choice is the mid-year (1 July) population. In European countries, it is more common to take the arithmetic mean of the population counts on 1 January and 31 December. Statistics Austria applies an even more sophisticated estimator based on quarterly population counts, namely $(\text{population on 1 January} + 2 \times \text{population on 1 April} + 2 \times \text{population on 1 July} + 2 \times \text{population on 1 October} + \text{population on 31 December})$ divided by 8. Lamonica and Zagaglia (2013) use the geometric mean of the population counts on 1 January and 31 December.

It is obvious that the total of person-years lived depends on the number of migrations during the year of interest. As soon as a person in-migrates to a municipality, he or she starts contributing to the person-years lived. On the other hand, a person who out-migrates stops contributing to the person-years lived. The migration counts and the person-years lived are therefore simultaneously determined. This implies that if the total of person-years lived is used as a regressor in a migration count model, it is endogenous.

In what way feedback from the migration count to the estimated person-years lived occurs depends on the estimator of person-years lived. For example, if the mean of population counts on 1 January and 31 December is used, then half of the annual migration count is part the total of person-years lived. This can easily be checked by the demographic bookkeeping equation, which states that the population on 31 December equals the population on 1 January plus the annual totals of births and in-migrations minus the annual totals of deaths and out-migrations.¹⁹

¹⁹In practice, there are also statistical adjustments.

If the 1 July population is used, then all migration during the first half of the year contributes to the total of person-years lived.

When modeling demographic rates, the total of person-years lived is sometimes interpreted as an “exposure” variable meant to measure the population at risk of the demographic event. One has to be careful that this is not necessarily the case. For example, regarding the in-migration rate, the person-years lived in a municipality cannot be an exposure variable because the population at risk of in-migrating is precisely the population not residing in the municipality. Deviations of the person-years lived from the population at risk also occur for demographic events other than in-migration. Regarding the crude birth rate, for example, the population at risk is only a subset of all person-years lived, namely those lived by women at reproductive ages.

5.1.3 The rate regression model

Let us denote the observed migration count by y and the total of person-years lived by P . The observed migration rate r is then by definition

$$r := y/P. \quad (5.5)$$

As we have discussed, both the numerator and the denominator of r are subject to randomness and moreover they are simultaneously determined. Specifying an appropriate data generating process for r is therefore not an easy task.

Separating y we get the identity

$$y = rP, \quad (5.6)$$

so modeling a migration rate can be done by modeling the migration count and taking the total of person-years lived as an endogenous explanatory variable. There exists a substantial body of literature regarding count data models with endogenous regressors, see Wooldridge (1997), Cameron and Trivedi (1998: Ch. 11), Winkelmann (2010: Ch. 5), and Hilbe (2011: Ch. 13). Remember that in a regression model one specifies the expected value of the migration count random variable, conditional on explanatory variables x . In a generic sense we have used

$$E(Y|x) = \exp(x'\beta). \quad (5.7)$$

The proportionality assumption Suppose that we want to model a migration rate by modeling the migration count and using the average population size P as an explanatory variable. The simplest approach is to assume proportionality

$$E(Y|x) = P \times \exp(x'_o \beta_o), \quad (5.8)$$

where x_o is the vector of the remaining explanatory variables, which we assume to be exogenous, and β_o is the corresponding vector of regression parameters. The proportionality assumption means that in a municipality with 2,000 person-years lived we expect twice as many migrations

as in a municipality with 1,000 person-years lived, if all other regressors are the same. Note that the proportionality assumption may be re-written as

$$E(Y|x) = \exp(x'_o \beta_o + \log P \times 1), \quad (5.9)$$

which means that $\log P$ is used as an “offset” variable in the exponential mean function, i.e. it has a fixed regression coefficient of 1. Such an attempt is made by Brillinger (1986) in the context of mortality rates to approximate more complicated models. It should be noted that the proportionality assumption does not simply mean that population size is an exogenous regressor, but more so that its regression coefficient is known in advance.

Although the proportionality assumption is often applied in demographics, it should be avoided in migration analyses because it is not plausible in general. On the contrary, it is a well known fact that for many demographic outcomes the expected rate is related to the size of the place (Greenwood 1975, Udry et al. 1979). For example, international immigration is a phenomenon that affects primarily large cities, not only in absolute numbers, but also in rates. This is because the population size itself influences many features of a place which in turn influence immigration rates, such as demand for certain jobs, location of higher education institutions, availability of immigrant networks, and hostility against newcomers. Also for rural municipalities the expected migration rate may depend on population size, e.g. if the economic structure favors population concentration within rural areas.²⁰ Assuming proportionality may therefore result in a misspecified mean function and consequently in an inconsistent QGPML estimator. Lamonica and Zagaglia (2013) mention empirical reasons to avoid a fixed regression coefficient. Tervo (2001) found for Finland higher in-migration and lower out-migration rates for larger regions, what he interprets to reflect “the concentration process of population”.

We recommend, at least in migration analyses, to relax the proportionality assumption:

$$E(Y|x) = \exp(x'_o \beta_o + \log P \times \beta_e). \quad (5.10)$$

Such a conditional mean function may still be misspecified in functional form if the dependence of the expected rate on population size is more complicated, e.g. U-shaped. However, the risk of misspecification is here greatly reduced compared to the proportionality assumption. Note that the proportionality assumption is the special case with $\beta_e = 1$. This hypothetical value can be tested by an asymptotic t-test. Applications in migration analyses are Greenwood (1975) and Lamonica and Zagaglia (2013). The general idea is described by Cameron and Trivedi (1998: p. 304).

The author has come to realize that some researchers, for whatever reasons, are reluctant to the idea of testing hypothetical values other than 0. In such situations one may write the conditional mean function as

$$E(Y|x) = \exp(x'_o \beta_o + \log P \times 1 + \log P \times \beta_e) \quad (5.11)$$

²⁰Eckhout (2004) fitted for U.S. places nonparametric regression of growth rate on population size. For the intermediate 90 percent, the average growth rate was found to be approximately the same as for the very large communities, but for the very small communities it was clearly below.

and then test the hypothesis $\beta_e = 0$ (Winkelmann 2010: pp. 74-76). Although formally the same as the equation just before, it seems that the latter formulation is more popular among non-statisticians. A similar formulation was used in a migration analysis by Galle et al. (1993).

However, even when accounting for non-proportionality, the conditional mean function of a rate model is still likely to be misspecified. The reason is the contemporaneous feedback from y to P . If the migration count y is extraordinarily large in a certain year, then this influences also $\log P$. The error term is therefore not mean-independent of this regressor and consequently the QGPML estimator does not converge to the parameter value of interest.

It might be helpful in this respect to consider the interpretation of a regression coefficient. In our setting with an exponential mean function, the coefficient of a logged continuous regressor has the interpretation of elasticity. It denotes by how many percent the expected value of the dependent variable changes when the regressor changes by 1 percent with everything else remaining unchanged. But if the dependent variable and the regressor are simultaneously determined, an ordinary parameter estimate cannot tell the exogenous variation in the regressor from the endogenous variation that is caused by the feedback from the dependent variable. For example, if the exogenous elasticity of migration count to person-years lived is 1, then the estimator (3.22) converges to a value greater than 1 in an in-migration model and smaller than 1 in an out-migration model. In short, we have to account for the endogeneity of person-years lived to achieve consistency.

5.1.4 Traditional demographic approaches to simultaneity

That the total of person-years lived is endogenous to a demographic count is not a new finding, but has been realized by the demographic community many times. An early reference on migration is Greenwood (1975), who considered endogeneity in migration models in general (see also Cadwallader 1992 and Adam 2009). A special treatise on the dependence between the numerator and the denominator of an age-specific mortality rate was given by Brillinger (1986). Yet his paper was restricted to the case when the total of person-years lived is estimated by the 1 July population count and therefore integer-valued.

The traditional approaches in demography to handle endogenous average population size can be classified into three general classes.²¹ The first—and by far the most common—approach is to ignore endogeneity in the case of “rare” events. Often the numerator or a demographic rate is much smaller than the denominator. For example, crude birth rates and crude death rates in developed countries are usually of the order of 0.01. This implies that even if feedback from events to person-years lived exists, it is so small that it is effectively irrelevant, so for all practical purposes population size can be handled as if it were an exogenous regressor. Such an approach is however less justified in migration studies, particularly for small places, where annual rates are of the order of 0.05–0.10.

A second strategy, which has frequently been applied in migration studies, is to use an estimator of person-years lived that is not influenced by the migration flow during the year

²¹Besides these three classes, special approaches are given by Chiang (1986), Keiding (1986) and Manton (1986).

of interest. An obvious choice is the 1 January population count. In that respect, our basic model application in Chapter 4 can be interpreted as a rate model with 1 January population used to approximate the person-years lived. Similar examples are given by Condgon (1992, 1993) and Galle et al. (1993). Greenwood (1975) considered even earlier population counts, e.g. the population count in 1960 when analyzing migration in 1965-1970. His model also accounted for population growth. However, taking the population count on 1 January is not free of problems. It ignores not only the migration flow but any component of population change during the year of interest, including births, deaths, and the opposite migration flow. Such events are however often related to migration decisions. Also, if the period of analysis is longer than one year, then the population count at the beginning of the period can be very outdated at the end of the period, particularly for small places.

The third strategy which has been applied to overcome endogeneity is to model the joint distribution of event count and person-years lived, rather than the conditional distribution of the event count given the person-years lived. From a statistical viewpoint, this is the most challenging task, which may explain why it has been applied only a handful of times. Brillinger (1986) showed that under certain (not very realistic) assumptions the numerator and the denominator of a mortality rate are bivariate Poisson distributed (Kochlakota and Kocherlakota 1992: Ch. 4, Winkelmann 2010: Ch. 7). With such a fully parametric specification, full-information maximum likelihood estimation is in principle possible, although at present hardly supported by standard statistical software packages. In any case, working with the joint distribution requires additional assumptions on the data generating process of the person-years lived. Moreover, the joint distribution is usually of little interest to demographers. What matters is the conditional distribution of the event count.

5.2 An new instrumental-variable approach

5.2.1 Two-stage residual inclusion

In this thesis we propose an alternative method to account for simultaneity between migration count and person-years lived in migration rate models. Our method is an instance of two-stage residual inclusion (2SRI) estimation, which is itself an instance of instrumental-variable estimation. The basic idea of instrumental-variable estimation is to account not only for observable, but also for unobservable predictors in the conditional mean function. Important general references on 2SRI are Wooldridge (1997) and Terza et al. (2008).

2SRI is a technique that is frequently applied by health econometricians. Shea et al. (2007) use 2SRI to disentangle selection bias and moral hazard in a study on the effect of health insurance coverage on prescription drug fills. A similar reference is Shin and Moon (2006). DeSimone (2002) uses 2SRI to model the influence of drug consumption on employment. To the best of the author's knowledge, 2SRI has not been applied so far in demographic rate modeling.

The following explanations draw heavily from Wooldridge (1997: pp. 382-386) and Terza

et al. (2008). Suppose we have given an exponential regression model with one or more exogenous regressors x_o and an endogenous regressor x_e :

$$y = \exp(x_o' \beta_o + x_e \beta_e) + e. \quad (5.12)$$

The assumption of a scalar x_e is here purely for simplicity reasons and the approach can easily be generalized to a vector-valued x_e .

Due to endogeneity of x_e we have

$$E(Y | x_o, x_e) \neq \exp(x_o' \beta_o + x_e \beta_e). \quad (5.13)$$

Suppose then that there exists an unobserved scalar x_u such that

$$E(Y | x_o, x_e, x_u) = \exp(x_o' \beta_o + x_e \beta_e + x_u), \quad (5.14)$$

i.e. the endogeneity of x_e ceases to exist conditional on x_u . This implies some statistical relationship between x_e and x_u .

Let us assume that one or more variables w^+ are available which fulfill the following three conditions:

- They are sufficiently correlated with the endogenous regressor x_e .
- Besides that correlation, they are uncorrelated with the error term e .
- They have no direct influence on the dependent variable y , i.e. they are not part of the conditional mean function.

In other words, the components of w^+ do influence the outcome y , but only via their influence on x_e . Such variables are called instruments and are used to model x_e in a first stage. Outcomes of this first-stage model are then used in modeling y in a second stage.

Finding suitable instruments is not always easy. If the correlation with the endogenous regressor is weak, then parameter estimation becomes unreliable. On the contrary, for an instrument highly correlated with the endogenous regressor it may be difficult to argue that it is truly uncorrelated with the error term (Greene 2003: pp. 74-80). For example, Dong (2009: p. 7) discusses endogeneity of labor income in migration models and regarding instruments concludes that “anything that affects wages may also affect expected wage gains and hence the decision to move”.

First-stage reduced form Let us denote by w the union of second-stage exogenous regressors x_o and first-stage instruments w^+ . The correlation between x_e and w^+ can be introduced by a linear regression model

$$x_e = w' \alpha + u. \quad (5.15)$$

An extension to non-linear reduced forms is discussed by Terza et al. (2008). The error term u may be understood as unexpected deviation in x_e given w . We assume that $E(u) = 0$.

Let us now assume that the unobserved x_u fulfills

$$x_u = u\beta_u + \text{error}, \quad (5.16)$$

effectively modeling the relation between x_e and x_u . As long as x_o contains a constant, there is no loss of generality in assuming $E[\exp(\text{error})] = 1$. Given independence assumptions discussed in detail by Wooldridge (1997: p. 383), it follows that

$$E(Y | x_o, x_e, u) = \exp(x_o'\beta_o + x_e\beta_e + u\beta_u). \quad (5.17)$$

So if u were observed, we could proceed in the ordinary way by including it as a regressor. The union of all regressors x_o , x_e , and u may then be denoted by x , and everything works just in the same way as we have developed it in Chapter 3. In particular, the assumptions on the unobserved regressors z are the same as in Section 3.3, i.e.

$$E(Y | x_o, x_e, u, z) = \exp(x_o'\beta_o + x_e\beta_e + u\beta_u + z'\delta), \quad (5.18)$$

with $\xi = \exp(z'\delta)$ such that it is mean-independent of the observed regressors and its expectation equals 1.

The drawback of this approach is of course that u is unobserved. However, due to the statistical relation $x_e = w'\alpha + u$, it is estimable by the residual in a first-stage regression of x_e on w . So the feasible estimator of β replaces the unknown u by the residual $\tilde{u}$. This makes it clear why that technique is called two-stage residual inclusion.

It should be noted that inclusion of residuals as regressors in a model comes in different ways and not all of them are 2SRI in the sense given above (e.g., Barro 1977). In particular, when we speak of 2SRI we mean that all exogenous regressors in the second stage are used as predictors in the first stage too.

It can be shown by standard arguments that 2SRI allows for consistent estimation of β if $\tilde{\alpha}$ is consistent. Indeed, 2SRI is an instance of the general class of two-stage estimators. Formal references on the topic are Newey and McFadden (1994), Terza (2008), and Wooldridge (2010: Ch. 12). A benefit of 2SRI is that it allows for testing for endogeneity of x_e by standard regression tests (Wooldridge 1997: p. 384).

A special case occurs if both stages are linear. In that case 2SRI has the same effect as two-stage predictor substitution (2SPS) meaning that the endogenous x_e is replaced by its predicted value from the first-stage regression. It is important to note that in nonlinear models—the common case in demographic applications—2SRI and 2SPS are not equivalent. In particular, 2SPS does in general not imply consistent estimation of β , even if the first-stage $\tilde{\alpha}$ is consistent. This has been elegantly proven in a formal way and illustrated by simulation studies by Terza et al. (2008). So in a regression model with an exponential mean function, the required technique is 2SRI, not 2SPS.

5.2.2 Adjustment of the variance-covariance matrix

Suppose one estimates the parameter vector β of a migration rate regression model by NB2 QGPMLE and handles endogeneity of person-years lived by 2SRI. Then, as we have just

discussed, the QGPMLE is consistent. Regarding its VC, however, one must not use the ordinary estimator (3.23). The ordinary VC estimator is consistent only if the residual $\tilde{u}$ is the error u , implying that the estimate $\tilde{\alpha}$ is the true parameter value α . In contrast, a proper VC estimator of $\hat{\beta}$ must take the sampling variation of $\tilde{\alpha}$ into account.

General comments To derive a consistent estimator of $\hat{\beta}$ in the 2SRI case, let us look back on the discussion of M-estimation in Section 1.4. We saw there that under regularity conditions, a “one-stage” M-estimator is consistent and satisfies

$$\sqrt{n} (\hat{\beta} - \beta) \rightarrow A^{-1}(\beta) s(\beta), \quad (5.19)$$

with $s(\beta)$ the random gradient vector and $A(\beta)$ the expected Hessian matrix of the function to be minimized, both evaluated at the true parameter values. This “influence function representation” of an M-estimator implies that its VC satisfies

$$n \text{VC}(\hat{\beta}) \rightarrow A^{-1} B A^{-1}, \quad (5.20)$$

with

$$B = E[s(\beta) s'(\beta)]. \quad (5.21)$$

As we discussed, in case of QML estimation $B = \sigma^2 A$ and consequently the n -fold VC simplifies to $\sigma^2 A^{-1}$.

We now discuss the case when an M-estimator depends on an additional parameter α , in general vector-valued, that is estimated outside the estimating equations of $\hat{\beta}$. Let us define a two-step (or two-stage) M-estimator²² generically as the solution of

$$s_n(\hat{\beta} | \tilde{\alpha}) = \frac{1}{n} \sum_{i=1}^n s_i(\hat{\beta} | \tilde{\alpha}) := 0, \quad (5.22)$$

with $s(\cdot | \tilde{\alpha})$ the gradient vector of the function to be minimized. In this function $\tilde{\alpha}$ is simply an additional argument.

It should be mentioned that the NB2 QGPMLE (3.22) is already a two-step M-estimator. We use the auxiliary estimate $\tilde{k}$ in the quasi NB2 estimating equations providing $\hat{\beta}$. Remember that using a $\sqrt{n}$ -consistent estimator $\tilde{k}$ instead of the true k does not influence the asymptotic VC of the NB2 QGPMLE. The reason is that if the conditional mean is correctly specified, then

$$E \left(\frac{\partial s}{\partial k} \right) \propto E[y - \exp(x'\beta)] = 0, \quad (5.23)$$

which is a sufficient condition for irrelevance of the variation in $\tilde{k}$ to the asymptotic VC of $\hat{\beta}$. A proof is given by Ziegler (2011: pp. 82-85). This condition holds not only for the

²²Regularity conditions for two-step M-estimators are discussed by Murphy and Topel (1985: pp. 373-374), Newey and McFadden (1994: p. 2178), and Wooldridge (2010: Ch. 12).

shape parameter of the NB2, but for any nuisance parameter of a LEFN making QGPMLE an attractive class of estimation.

One could ask whether the same holds for the first-stage residual $\tilde{u}$ that we just discussed. Unfortunately, the answer is no; the variation in $\tilde{\alpha}$ influences the VC of $\hat{\beta}$ even in large samples.²³

In the remainder of this thesis we shall make some assumptions on $\tilde{\alpha}$. More general cases of two-step M-estimation are discussed by Newey and McFadden (1994) and by Wooldridge (2010: Ch. 12). Here we assume that $\tilde{\alpha}$ is $\sqrt{n}$ -consistent for the true value α and satisfies

$$\sqrt{n}(\tilde{\alpha} - \alpha) \rightarrow G^{-1}(\alpha) t(\alpha). \quad (5.24)$$

If $\tilde{\alpha}$ is itself an M-estimator, then G and t regarding α are defined analogously to A and s regarding β . We furthermore assume that

$$n \text{VC}(\tilde{\alpha}) \rightarrow \phi^2 G^{-1}. \quad (5.25)$$

If $\tilde{\alpha}$ is a PML estimator, then this assumption implies that the first-step conditional variance function is correctly specified and $\tilde{\alpha}$ is hence a QML estimator (possibly a QGPML estimator). As derived extensively by Wooldridge (2010: Ch. 12), the VC of $\hat{\beta}$ then satisfies

$$n \text{VC}(\hat{\beta}; \tilde{\alpha}) \rightarrow A(\beta|\alpha)^{-1} D(\alpha, \beta) A(\beta|\alpha)^{-1}, \quad (5.26)$$

with A defined analogously to the one-step case with α as an additional argument and D given by

$$D(\alpha, \beta) = E[h(\alpha, \beta) h'(\alpha, \beta)] \quad (5.27)$$

with

$$h(\alpha, \beta) = s(\beta|\alpha) + F(\alpha, \beta) G^{-1}(\alpha) t(\alpha) \quad (5.28)$$

and

$$F(\alpha, \beta) = E\left(\frac{\partial s}{\partial \alpha'}\right). \quad (5.29)$$

Defining $U(\alpha, \beta) = E[s(\beta) t'(\alpha)]$ and assuming that $\hat{\beta}$ is a QML or QGPML²⁴ estimator, the VC of $\hat{\beta}$ then satisfies

$$n \text{VC}(\hat{\beta}; \tilde{\alpha}) \rightarrow \sigma^2 A^{-1} + A^{-1} (F G^{-1} U' + U G^{-1} F' + \phi^2 F G^{-1} F') A^{-1}. \quad (5.30)$$

In applications, the quantities ϕ^2 , F , G and U are of course unknown and have to be estimated from sample data, just like σ^2 and A have to be estimated. It becomes clear from the formula why the first-stage instruments should be highly correlated with the endogenous regressor. If the instruments are weak, then the unexplained error variance in the first stage ϕ^2 is large and so $\hat{\beta}$ is unreliable.

²³It is important to note that the robust sandwich VC estimator that we discussed in Section 1.4 is not robust to variation in a first-stage parameter, but to misspecification of the second-stage conditional distribution. Thus, the ordinary robust sandwich estimator cannot be applied in two-stage estimation. There exists a special robust sandwich estimator for two-stage models which is based on stacked estimating equations of the two stages (Hardin 2002).

²⁴It should be mentioned that for the NB2 QGPMLE (3.22) $\sqrt{n}$ -consistency of $\tilde{k}$ remains if it depends on a $\sqrt{n}$ -consistent $\tilde{\alpha}$. This follows from standard asymptotic arguments.

Special cases A special case of the two-stage VC given above occurs if $F = 0$. One immediately sees that then

$$n \text{VC}(\hat{\beta}; \tilde{\alpha}) \rightarrow \sigma^2 A^{-1}. \quad (5.31)$$

Another special case occurs if the scores of the two stages, evaluated at the true parameter values, are uncorrelated, i.e. if $E[s(\beta) t'(\alpha)] = 0$. Such an assumption seems natural if α and β are estimated from different data, although this rarely happens in demographic applications. The more important practical case is when the parameters are estimated from the same data, but uncorrelatedness is implied by the structure of the model. Barro (1977) gives an example, which is also discussed in the seminal paper by Murphy and Topel (1985). In that case

$$n \text{VC}(\hat{\beta}; \tilde{\alpha}) \rightarrow \sigma^2 A^{-1} + \phi^2 A^{-1} F G^{-1} F' A^{-1}. \quad (5.32)$$

5.3 Application

5.3.1 The data

We are now applying the theory we have just developed to a migration rate model. We use the same data as in Chapter 4, except that we now use P , the annual average population size, instead of K , the 1 January population count, as a regressor. Compared to Chapter 4, we now use italics because both K and P have only one character and are contained in several formulae that follow. The P variable was provided by Statistics Austria and is defined as (population on 1 January + 2 × population on 1 April + 2 × population on 1 July + 2 × population on 1 October + population on 31 December) divided by 8. Contrary to K and also to some other estimates of person-years lived, our P variable is in general not integer-valued.

As to be expected, the distribution of P is similar to the distribution of K . It ranges from 69.5 to 13193.875, with a mean of 1673.4. The in-migration rate (count divided by P) ranges from 0.014 to 0.162, with an average of 0.051. The out-migration rates are a bit more evenly distributed, ranging from 0.016 to 0.122, with a mean of 0.046.

A natural instrument for the endogenous $\log P$ is $\log K$, the logged population size on 1 January. Clearly, $\log K$ is highly correlated with $\log P$ and in our cross-sectional setting it is plausible to assume that the 1 January population influences the annual migration count only via its influence on the annual average population. Besides $\log K$, one could also look for additional instruments for $\log P$. Here we have chosen only $\log K$ because it is already a very good predictor for $\log P$ and we wanted to keep the application as simple as possible.

5.3.2 First-stage estimation

We specify a linear reduced form for $\log P$:

$$x_{ei} = \log P_i = w_i' \alpha + u_i = \alpha_K \log K_i + \alpha_o' x_{oi} + u_i. \quad (5.33)$$

The subscripts α_K and α_o indicate the first-stage parameters for the genuine instrument $w_i^+ = \log K_i$ and the exogenous second-stage regressors x_{oi} , respectively. Note that we are modeling logged population sizes, so the regression coefficient of $\log K$ is a measure of relative population growth during the period of interest.

The error term u_i can be interpreted as “unexpected deviation in logged population” during the period of interest (see also Barro 1977). It is the stochastic part of a municipality’s $\log P_i$ that cannot be explained by $\log K_i$ and x_{oi} . Clearly we expect u_i to be correlated with this municipality’s migration count during the year of interest, namely positively correlated with in-migration and negatively correlated with out-migration.

The first-stage parameters α were first estimated by OLS, i.e. assuming

$$\text{Var}(u_i) = \phi_{\text{OLS}}^2. \quad (5.34)$$

A standardized residual plot (upper panel of Figure 5.1) revealed considerable heteroskedasticity, with smaller residual variance for larger predictions. A heuristic explanation can be given as follows. The difference $P - K$ is a measure of the absolute population growth during the year of interest. This absolute population growth consists of live births, deaths, in- and out-migration, and statistical adjustments. If the variance of $P - K$ is roughly proportional to K , then the variance of P/K , a measure of relative population growth, is roughly proportional to $1/K$. This intuitively makes sense because for small municipalities we expect greater variation in relative growth than for large municipalities. Given that P and K are approximately of the same magnitude, the variance of $\log(P/K)$ is then also roughly proportional to $1/K$. But $\log(P/K)$ is essentially what we model in our first stage. This implies that the assumption underlying OLS, namely constant error variance, is violated. A misspecified error variance means not only inefficiency of the OLS estimator, but also that the standard estimator of the VC of $\tilde{\alpha}_{\text{OLS}}$ is inconsistent.

It is therefore necessary to relax the assumption of constant error variance and instead assume

$$\text{Var}(u_i) = \phi^2 / K_i. \quad (5.35)$$

The standardized residual plot in the lower panel of Figure 5.2 clearly supports this assumption. The first-stage parameters are then efficiently estimated by GLS, with K chosen as the *weight* in R’s *lm* command. From a substantive point of view this means that the larger municipalities get greater weights in our first-stage GLS estimation than in OLS estimation. The GLS parameter estimates are given in Table 5.1. As expected, logged 1 January population is an excellent predictor of logged annual average population. Also FOREIGNB seems to have some influence.

Note that under the assumption of a correctly specified conditional variance function GLS estimation is, regarding $\tilde{\alpha}$, equivalent to heteroskedastic Gaussian QML estimation. The only difference between GLS and the latter lies in the estimation of $\tilde{\phi}^2$. In QML estimation, which is based on asymptotic arguments, the weighted residual sum of squares is usually divided just by n , whereas in GLS estimation, motivated by finite sample optimality properties, a degrees-of-freedom correction is usually applied.

Figure 5.1: **First-stage residuals: OLS vs. GLS**

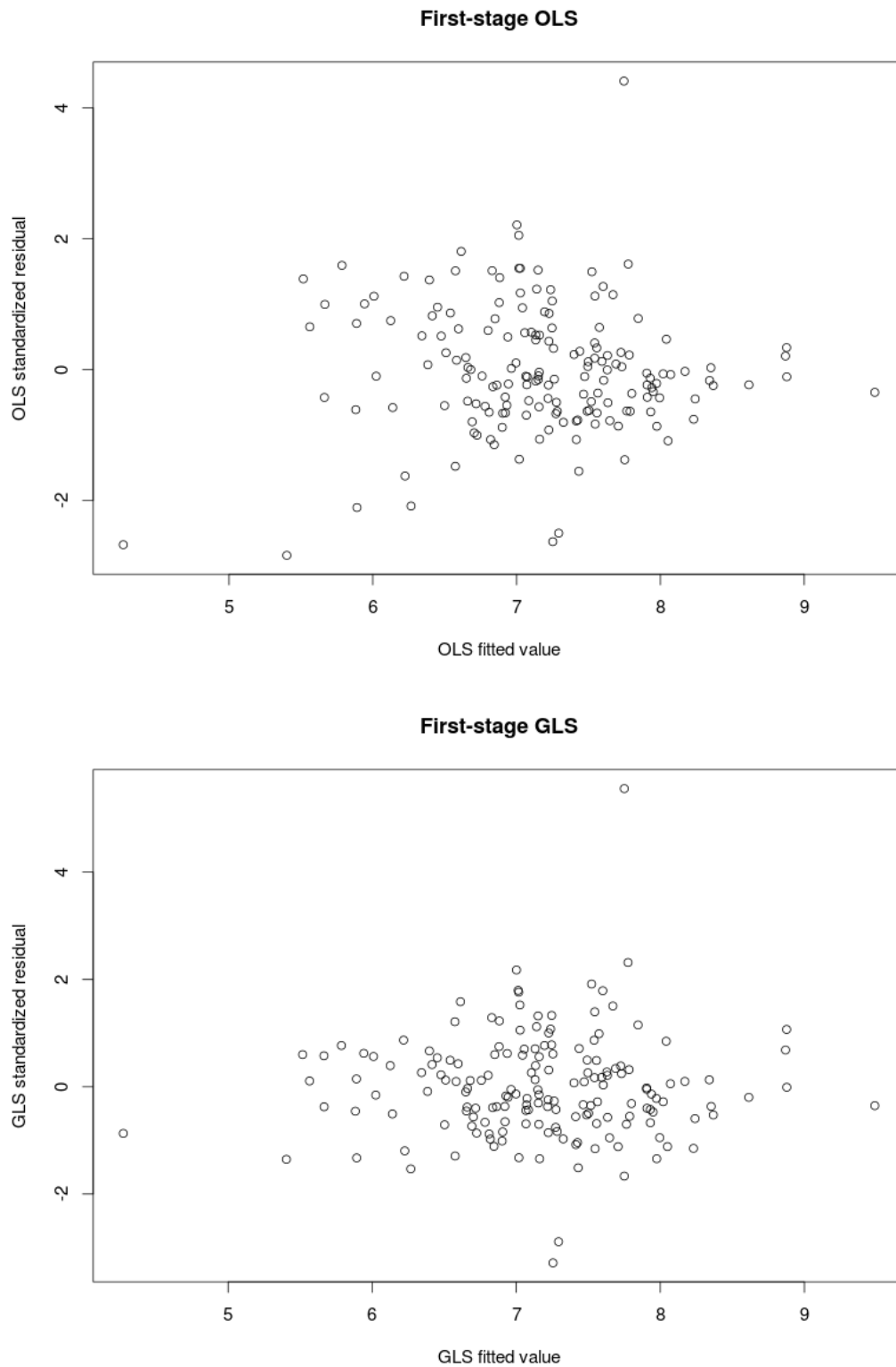


Table 5.1: **First-stage GLS regression results**

	Estimate	t-ratio	p-value
K	1.0002	835.90	***
CONSTANT	-0.0243	-0.36	
NORTH	-0.0036	-1.31	
ELEVAT	-0.0037	-1.40	
AGE2034	0.0035	0.56	
FOREIGNB	0.0034	2.23	*
DENSITY	0.0073	0.86	
DENSITY ²	-0.0007	-0.79	
INCOME	0.0120	0.72	
SINGLEDW	0.0003	0.35	
COMMUTE	-0.0068	-1.90	

* P<0.05 ** P<0.01 *** P<0.001

The first-stage GLS residuals, which are included in the second stage as additional regressors, range from -0.031 to 0.035, indicating unexpected deviations in person-years lived of up to almost 4 percent. The largest positive residual regards the municipality Kittsee that we already discussed in Section 4.3. The largest negative residual regards the smallest municipality in our sample, Tschanigraben.

5.3.3 Second-stage QGPML with 2SRI

The NB2 QGPML parameter estimates for the rate models are presented in Table 5.2. The t-ratios and p-values given are based on the asymptotically correct standard errors. The VC of the NB2 QGPML was estimated by

$$\widehat{VC}(\hat{\beta}; \tilde{\alpha}) = \frac{1}{n} \left[\tilde{\sigma}^2 \hat{A}^{-1} + \hat{A}^{-1} (\hat{F} \tilde{G}^{-1} \hat{U}' + \hat{U} \tilde{G}^{-1} \hat{F}' + \tilde{\phi}^2 \hat{F} \tilde{G}^{-1} \hat{F}') \hat{A}^{-1} \right], \quad (5.36)$$

with the matrix estimators given by

$$\hat{A} = \frac{1}{n} \sum_{i=1}^n \frac{\exp(x_i' \hat{\beta})}{1 + \exp(x_i' \hat{\beta}) / \tilde{k}} x_i x_i', \quad (5.37)$$

$$\hat{F} = \frac{1}{n} \sum_{i=1}^n \frac{-\hat{\beta}_u \exp(x_i' \hat{\beta})}{1 + \exp(x_i' \hat{\beta}) / \tilde{k}} x_i w_i', \quad (5.38)$$

$$\tilde{G} = \frac{1}{n} \sum_{i=1}^n K_i w_i w_i', \quad (5.39)$$

$$\hat{U} = \frac{1}{n} \sum_{i=1}^n \frac{[y_i - \exp(x_i' \hat{\beta})] [K_i (x_{ei} - w_i' \tilde{\alpha})]}{1 + \exp(x_i' \hat{\beta}) / \tilde{k}} x_i w_i'. \quad (5.40)$$

Here the regressors x_i contain the first-stage residual $\tilde{u}_i$. The endogenous x_{ei} is $\log P_i$, and K_i is the weight in the first-stage GLS estimator.

Altogether the effect sizes are similar to the basic models, particularly for out-migration. Regarding person-years lived, the null hypotheses of unit elasticity cannot be rejected at any reasonable level of significance (the respective t-ratios are -0.21 for in-migration and -1.37 for out-migration). As expected, the coefficient of the first-stage residual $\tilde{u}$ is positive for in-migration and negative for out-migration.

A particularity of the in-migration rate model is that the iterative parameter estimates did not converge (not even when the maximum number of iterations was increased from 30 to 100). Apparently successive estimates $\tilde{k}$ jumped between a few regions of the parameter space. Given this unsteadiness, the estimated variance function parameters should be treated with caution. In this thesis we did not pay special attention to numerical analysis, but used standard algorithms implemented in R functions. It could be that with a refined optimization method one would achieve convergence of the in-migration rate model parameters too. Thus one possible extension of the application presented here is to evaluate different optimization algorithms for our problem in order to improve the numerical solution. The in-migration parameter estimates in Table 5.2 refer to the 30th iteration. The out-migration parameter estimates fulfilled the convergence criterion after 5 iterations.

Regarding the out-migration variance function parameters, we see that $\tilde{\gamma}_0$ is similar to the estimate for the basic model in Table 4.4, whereas $\tilde{\gamma}_1$ is around 20 percent smaller. A heuristic explanation is that accounting for the entire person-years lived during the period plus the first-stage residual instead of merely the 1 January population count reduces the amount of unobserved heterogeneity, whereas it has no influence on the household size distribution.

The effect of the adjusted VC Finally, let us investigate what effect the necessary adjustment of the VC of $\hat{\beta}$ has in this two-step model. We compare the asymptotically correct t-ratios corresponding to (5.30) with the t-ratios that are output by R's *glm* function and implicitly

Table 5.2: **Second-stage QGPML regression results for the rate model**

	In-migration			Out-migration		
	Estimate	t-ratio	p-value	Estimate	t-ratio	p-value
CONSTANT	-11.985			-12.149		
NORTH	-0.232	-2.73	**	-0.008	-0.12	
ELEVAT	-0.016	-0.19		0.168	2.68	**
P	0.992	27.96	***	0.959	32.16	***
$\tilde{u}$	16.522	9.95	***	-9.218	-5.25	***
AGE2034	0.804	4.22	***	0.865	5.77	***
FOREIGNB	0.461	8.99	***	0.406	9.52	***
DENSITY	0.284	1.19		0.069	0.29	
DENSITY ²	-0.023	-0.98		-0.006	-0.24	
INCOME	1.687	3.12	**	1.218	3.16	**
SINGLEDW	0.014	0.41		-0.040	-1.86	
COMMUTE	0.041	0.39		0.346	3.97	***
γ_0	2.591			2.006		
γ_1	0.0019			0.0035		
σ^2	2.591			2.006		
k	1342.0			567.3		
	* P<0.05	** P<0.01	*** P<0.001			

assume that $\tilde{u} = u$ (“Naive I”, corresponding to (5.31)) and the t-ratios which would apply if the random variation in $\tilde{u}$ were accounted for but the random scores of first- and second-stage were uncorrelated (“Naive II”, corresponding to (5.32)). Results are given in Table 5.3.

Table 5.3: **Second-stage QGPML t-ratios**

	In-migration t-ratio			Out-migration t-ratio		
	Actual	Naive I	Naive II	Actual	Naive I	Naive II
NORTH	-2.73	-3.38	-2.81	-0.12	-0.12	-0.11
ELEVAT	-0.19	-0.25	-0.21	2.68	2.72	2.53
P	27.96	34.90	28.66	32.16	34.27	31.89
$\tilde{u}$	9.95	9.96	9.96	-5.25	-5.26	-5.26
AGE2034	4.22	5.29	4.38	5.77	5.96	5.54
FOREIGNB	8.99	12.76	10.48	9.52	11.35	10.57
DENSITY	1.19	1.40	1.15	0.29	0.34	0.32
DENSITY ²	-0.98	-1.17	-0.96	-0.24	-0.29	-0.27
INCOME	3.12	4.06	3.37	3.16	3.03	2.83
SINGLEDW	0.41	0.59	0.49	-1.86	-1.85	-1.71
COMMUTE	0.39	0.48	0.39	3.97	4.18	3.88

Naive I assumes that $\tilde{u}$ is the true u . Naive II assumes that the scores of first and second stage are uncorrelated.

In general the actual t-ratios are lower than the Naive I t-ratios, particularly for in-migration. This means that the actual standard errors of the parameter estimates are higher than the ordinary output of standard statistical software packages. Not accounting for the sampling variability in $\tilde{u}$ results in biased statistical inference, namely overestimation of the precision of second-stage parameter estimates. To give an example, in the in-migration rate model the coefficients of NORTH and INCOME are actually not significant at the 0.001 level, but they appear to be so when looking at the Naive I standard errors. Altogether most differences between actual and Naive I t-ratios are not very large. This can be explained by the fact that the logged 1 January population is a very good instrument for the logged person-years lived, implying that the first-stage residuals $\tilde{u}$ are close to the unobserved true errors u . In most cases the Naive II t-ratios lie somewhere between the actual and the Naive I t-ratios. Clearly, correlation between the scores of the two stages cannot be neglected in this application.

6 Conclusions and outlook

6.1 Summary

Background In this thesis we have developed a statistical model of annual migration counts for rural municipalities. For rural areas, migration—both international and within countries—has become a major component of population change during the last decades, for the traditional rural excess of live births over deaths has largely disappeared. This has come along with an increase in variability in population change within rural areas.

Migration count data is often based on complete enumerations like registration of residence data. There exists a philosophical divide among demographers as to what extent stochastic approaches are appropriate when analyzing such data. We have argued that, at least for certain applications, stochastic approaches are beneficial, in particular when one wants to explain relationships between a demographic outcome and other variables via regression analysis.

Stochastic approaches in demography have emerged in two rather separated strands, namely vital events such as live births and deaths on the one hand and migration on the other hand. Compared to the vital events strand, the migration strand is more directed towards regression. Despite their rather separate evolution, both strands have put emphasis on the Poisson regression model and its semiparametric extensions like Poisson QML.

The Poisson regression model, usually specified with an exponential mean function, is a fully parametric model and can thus be estimated by ML. Regarding aggregate demographic counts however, the Poisson regression model is often not an ideal choice because of its restrictive assumption of equivalence of conditional mean and variance. Instead, models that allow for overdispersion are preferred. Some fully parametric alternatives to the Poisson do exist, however more flexible is the semi-parametric class of PML estimators based on the LEF. This class of models requires only correct specification of the conditional mean for consistent parameter estimation. Efficiency can be gained by correctly specifying the conditional variance, which defines the subclass of QML estimators. Some distributions like the NB2 are LEF only when a nuisance parameter which relates the mean to the variance is known. In practice one replaces the nuisance parameter by a $\sqrt{n}$ -consistent preliminary estimator. An attractive property is that such a QGPML estimator has the same asymptotic VC as the corresponding QML estimator. In general, asymptotic properties of likelihood-based estimators can be obtained from the class of M-estimators.

Analyzing aggregate migration flows Compared to vital events, the frequency of migration is sensitive to its definition. In this thesis, migration is a change of residence that crosses

a municipal border, resulting in either in-migration to or out-migration from a municipality. We focus on annual totals of migration counts and assume that the data is such that the same person may migrate several times a year.

Many theories exist to explain migration. A key assumption of any theory is that migration is a social act to improve one's (an individual's or a family's) situation. In neoclassical economics, migration acts as an equilibrator of spatial wage level differentials. This theory is based on several unrealistic assumptions that are overcome by alternative or complementary theories. Comparing real instead of money wages points to the housing market. Generalizing wage differentials to place utility differentials allows for inclusion of non-monetary amenities such as climate, culture, and population size. The human capital theory interprets migration as an investment, which readily allows for explaining differential migration by age and by educational or occupational class. A drawback of many theories is that they are designed for individuals, whereas migration is often undertaken by households. This can be accounted for by extending the human capital theory to the household context and by introducing the life cycle theory. Moreover, migration may be substituted by competing events like commuting. A general challenge in aggregate demographic models is to properly summarize explanatory variables that are basically predictors on the individual level, e.g. age.

Migration has often been divided into international and "internal" (within-country) migration, but for several reasons this dichotomization is unsatisfactory. Alternative disaggregations, e.g. by migration motive, could be useful in some applications, provided the necessary information is available in the data. More important seems disaggregation by explanatory variables to account for specific migration patterns and/or interaction effects. If many regressors are available then it is likely that several of them are highly correlated, calling for model selection or principal component extraction. Moreover, it happens quite often that regressors are endogenous to migration.

The basic model Our derivation of the conditional variance function was based on violations of the Poisson process. The Poisson process is a stochastic count process that can be motivated from a powerful theoretical result relying on statistical independence assumptions. For migration they mean that if people migrate independently of each other and independently of past events at a constant rate, then the annual total of migrations in a population is "inevitably" Poisson distributed. The constant rate assumption implies that in a regression model any systematic variation in the migration intensity between places is covered by the observed regressors.

Yet the assumptions underlying the Poisson process are not realistic for migration counts. Obvious violations are clustering of persons by household migration and unobserved heterogeneity between places, which may be caused by unobserved predictors, measurement errors, or aggregation bias. Both clustering and unobserved heterogeneity result in overdispersion. A third violation of the Poisson process is contagion of events, which is however less important in cross-sectional studies because there it can effectively be incorporated into unobserved heterogeneity.

We derived a relationship between conditional mean and variance of aggregate migration counts as follows. First, we assumed that conditional on all relevant regressors the number of migrating households is Poisson distributed and hence equidispersed. Since in fact only a subset of all relevant regressors is observed, the number of migrating households conditional on the observed predictors only is overdispersed with a quadratic component, provided the index of unobserved heterogeneity is mean-independent of the observed regressors. When every migrating household comes with a random number of persons, independent between households and independent of the household flow, then the conditional variance of the person count has two parameters. The first parameter, linear in the mean, reflects clustering of persons by household migration and the second parameter, quadratic in the mean, reflects unobserved heterogeneity between places. That two-parameter variance function has been used before, albeit without our theoretical derivation. Under suitable re-parameterization it equals the quasi NB2 variance.

Consistent and asymptotically efficient mean function parameter estimation is obtained by QGMPL, meaning that the unknown NB2 shape parameter is replaced by a $\sqrt{n}$ -consistent estimator. Given the usual distribution of population size between places and the strong relationship between population size and expected migration flow, we recommend to estimate the shape parameter by a robust estimator like LTS. Although this has no asymptotic advantages over OLS estimation, we believe that in general it improves the accuracy of estimates in finite samples. Also iterative estimation may be beneficial.

Application We applied our basic model to the annual migration counts (in- and out-migration separately) of the 171 municipalities of the Austrian federal province Burgenland in 2012. Based on various migration theories and data availability, 15 candidate regressors were prepared. After model selection using the QAICc, we estimated models with 11 regressors. Most regressors entered logarithmically, implying parameter interpretation as elasticities.

Parameter estimates were obtained by NB2 QGPML. Aside from population size, significant coefficients were obtained in both equations for age structure, proportion of foreign-born, and income level, indicating migration-prone populations. Northern location, meant to proxy housing cost, was significant only for in-migration, whereas geographic altitude and in particular commuting time to Vienna were significant only for out-migration. The estimates of the variance function parameters supported their theoretical relations.

A sensitivity analysis revealed that using alternative PML estimators would have resulted in qualitatively similar conclusions. Residual plots revealed one isolated departure for in-migration. The variance function seemed to be by and large correctly specified, apart from some deviations for the largest municipalities regarding out-migration.

The rate model Often demographers are interested in rates rather than event counts. A demographic rate is the respective demographic event count divided by the total of person-years lived in the same population during the year of interest. The total of person-years lived equals

the annual average population and is for most, but not all outcomes, the population at risk of the event.

A fundamental problem in regression analysis of demographic rates is simultaneity between numerator and denominator. In general the demographic events contribute to the total of person-years lived, meaning that the person-years lived is endogenous to the event count. Endogeneity of a regressor implies a misspecified conditional mean function and hence an inconsistent QGPML estimator. Several approaches to handle endogeneity in demographic rate models have been developed, but in our opinion neither of them is very appealing.

We propose a new method to handle endogenous person-years lived in migration rate models based on an instrumental-variable approach. Precisely, logged 1 January population count serves as an instrument for logged person-years lived in a linear first-stage model. The first-stage residual indicates unexpected deviation in logged person-years lived during the year of interest; it is correlated positively with in-migration and negatively with out-migration. This first-stage residual is then used in the second-stage migration count model as an additional regressor that guarantees consistency of the QGPML estimator.

Because the first-stage residual is a function of the first-stage parameter estimates, its sampling variation has to be accounted for in the VC of the QGPML. The necessary adaptations are derived from the general properties of two-stage M-estimators. In our application the random variation in the first-stage parameter estimates by and large increased the variability of the NB2 QGPML.

6.2 Possible extensions of the model

6.2.1 Extensions of the univariate model

Non-independence of observations Our model can be extended in a variety of ways. First, a basic assumption in our work so far has been independence of the errors between municipalities. This is clearly a discussable assumption. In-migrants from outside of the study area may often choose one out of several candidate municipalities. Also neighboring municipalities may be subject to unobserved spatial characteristics influencing migration. Spatial correlation may be visualized by residual maps. As Cushing and Poot (2004: p. 324) mention it, “No one has yet seriously exploited the potential of spatial econometrics in the migration literature. This would seem to be a natural extension for migration research with potentially greater importance at greater levels of geographic disaggregation”. An obvious generalization of our model would therefore be to allow for dependence in the regression errors. Methods that allow for correlated errors range from nonlinear seemingly unrelated regression models (Cameron and Trivedi 1998: pp. 262-263) to specific spatial econometrics models (LeSage and Pace 2004).

Non-nuisance of variance function parameters In this thesis we were primarily concerned with estimating the conditional mean function parameters. The variance function

parameters were treated as nuisance parameters, which can be seen by the fact that we did not consider asymptotic efficiency of their estimator. Of course, it may happen that in certain applications, e.g. migration forecasts, the variance function parameters are equally important or even more important than the mean function parameters.

One strategy to obtain efficient variance function parameter estimates is to specify higher order moments of the migration count and apply an estimator based on the quartic exponential family. This method is discussed in detail by Holly et al. (2008). It might however be difficult to justify assumptions on higher order moments of migration counts on substantive grounds. An alternative technique is simultaneous estimation of mean and variance parameters based on the quadratic exponential family (Gourieroux et al. 1984a, Lejeune 1997).

Varying variance function parameters Sometimes it might be beneficial to allow for greater flexibility in the variance function, e.g. by allowing for variation in its parameters between observations. This can be implemented by an iterative algorithm of mean and dispersion submodels (Smyth 1989) or by extended quasi-likelihood (Nelder and Pregibon 1987, Thaler 2009). There our σ^2 would be replaced by a σ_i^2 . The benefit is that if the dispersion parameter is in fact not constant between places then an appropriate re-weighting (i.e., downweighting of observations with relatively large σ_i^2) improves the efficiency of the mean function parameter estimator.

A theoretical justification for variation in the dispersion parameter is unequal household size distributions between places. In this context, the only paper that the author is aware of is Flowerdew and Boyle (1995) who compared a place-specific household size distribution with a unique distribution for all places in their study area. They found the impact on the deviance of the model to be negligible.

6.2.2 Joint analysis of in- and out-migration

In this thesis we have developed a generic model of an aggregate migration count, either in- or out-migration. In our application we investigated both flows, but via separate regression equations. It may—quite often—happen that one is in fact interested in both flows. Although separate models can in principle be used for this, a bivariate model might be more promising for two reasons. First, it may come with greater efficiency than separate models by taking advantage of the correlation between the two flows. Second, a bivariate model allows for cross-equation parameter hypotheses testing (King 1989). Yet these advantages come with costs. Specifying a bivariate distribution is more complicated than specifying univariate distributions and may be subject to particular misspecifications. Some bivariate count data models have been developed by the statistical community (see below), but the scientific literature is much scarcer than on univariate models. Also, bivariate count data models are at present hardly supported by standard statistical software packages and one would possibly be reliant on general bivariate models that do not explicitly refer to count data (e.g., Srivastava and Giles 1987).

Correlation between in- and out-migration One of the most durable findings in applied demography is a positive correlation between in- and out-migration. At a first glance this might be counter-intuitive as out-migration is often associated with “push” factors and in-migration with “pull” factors, which suggests negative correlation. Yet the standard empirical finding is positive correlation, even when explanatory variables are taken into account. This holds not only for small places: Galle et al. (1993) found positive correlation even for United States metropolitan areas.

Explanations for positive correlation between in- and out-migration abound in the migration literature (Greenwood 1975, Mueser and White 1987, Cadwallader 1992, Galle et al. 1993). Some common explanations are:

- **Compositional effect:** Persons who have a migration history are more likely to move again than those without a migration history (see also our discussion of “contagion” in Section 3.2). Thus places with many in-migrants have disproportionately “migration-prone” populations inducing disproportionate out-migration.
- **Return and circulatory migration:** Those who in-migrate to a place sometimes find their expectations about the new place not met. Newcomers have in general rather weak ties to their destination, the cost of moving back or on is low for them, encouraging out-migration. This may be seen as a special variant of the compositional effect just mentioned.
- **Boundary location:** Some places have a substantial proportion of their population living close to borders to other places. Such groups of places may be separated administratively but belong together functionally. Many small distance movements crossing just formal borders are then counted as migrations.
- **Economic factors:** Where economic conditions are favorable, people are more willing and more likely to change jobs. Places that are economically buoyant may both gain and lose migrants who are moving from positions of strength (Boyle et al. 1998: p. 55). Owen and Green (1992: p. 22) note that “rates of out-migration were highest in the most prosperous parts of Britain”.
- **Unobserved heterogeneity:** The previous two items can be seen as a special instance of a more general argument, namely that “apparent” correlation between in-migration and out-migration is caused by unobserved heterogeneity between places (Boyle et al. 1998). Sjastaad (1962) mentions heterogeneity in occupational and industry structure.
- **Vacancy chain:** Given scarcity of dwellings, jobs and the like, out-migration causes vacant positions which are then to some extent filled by newcomers. Note that in this case out-migration is the cause and in-migration the effect, whereas for the compositional effect and return migration it is the other way round.

- Endogenous housing cost: Large-scale in-migration increases relative housing costs which induces out-migration. By the same token, large-scale out-migration reduces relative housing costs making an area more attractive to newcomers.
- Cultural factors: The established population may be hostile to particular in-migrants, e.g. ethnic minorities or poor people, and may react to their inflow by leaving the place.

Trivariate reduction methods There are many ways to incorporate positive correlation between in- and out-migration into statistical modeling. A class of models is based on the “trivariate reduction method”, meaning that in- and out-migration are assumed to share a common additive factor. Formally, if Y_1 denotes in-migration and Y_2 denotes out-migration, then the trivariate reduction method means $Y_1 = Y'_1 + Y'_{12}$ and $Y_2 = Y'_2 + Y'_{12}$. Regression models may specify either the two marginal expectations and a correlation structure (King 1989, Jung and Winkelmann 1993, Winkelmann 2000) or the expectations of the three underlying unobserved factors (Karlis and Ntzoufras 2003). In the context of migration models, one should be careful to choose a model which allows for marginal overdispersion.

Correlated unobserved heterogeneity A different class of models introduces correlation via unobserved heterogeneity. In the simplest case of a one-factor model, it is assumed that the index of unobserved regressors ξ is the same for in- and out-migration. Clearly this leads to positive correlation conditional on the observed regressors. For mathematical convenience a Gamma distribution is often assumed for ξ , although this has little substantive justification (Hausman et al. 1984, Riphahn et al. 2003, Winkelmann 2010: Ch. 7).

From a domain point of view it often seems questionable to assume a common ξ for both flows. A more flexible approach which allows also for negative correlation is the two-factor model. There ξ_1 and ξ_2 are specific factors for in- and out-migration related by a correlation coefficient. An instance is the bivariate Gaussian distribution for logged values of ξ_1 and ξ_2 ; an application is Riphahn et al. (2003). One drawback of this approach is that the density does not have a closed-form expression. Another deficiency, elaborated by Greene (2007), is the questionable substantive meaning of the “correlation coefficient” which refers to correlation between the logged conditional expectations rather than the random variables. Except for the sign, it is in general unclear how these two correlations are related to each other.

Hybrid and alternative approaches Mayer and Chappell (1992) were particularly interested in the difference between two correlated counts, both of them depending on specific regressors. They constructed a hybrid model where correlation stems from both a common additive factor and correlated unobserved heterogeneity. QGPML estimation is based on the bivariate Gaussian distribution which is for specified VC a LEF (Gourieroux et al. 1984a, Ziegler 2011). A similar, yet more general model, is given by Gourieroux et al. (1984b).

Cadwallader (1992) mentions conditional models where one migration flows serves as an explanatory variable for the other flow. The principal weakness of this approach is the arbitrary

ordering of the flows. Even if in-migration is an excellent predictor of out-migration, it remains unclear what causes in-migration. Other bivariate models that could be applied to our setting are given by Srivastava and Giles (1987), Iwasaki and Tsubaki (2005), and Özel (2011).

A References

Adam F (2009): Modelling aggregate migration and cross-border worker flows. Statec Working Paper, Economie et Statistiques, No. 35/2009, Luxembourg.

Amemiya T (1973): Regression analysis when the variance of the dependent variable is proportional to the square of its expectation. *Journal of the American Statistical Association*, 68(344): 928-934.

Anderson DR, Burnham KP, White GC (1994): AIC model selection in overdispersed capture-recapture data. *Ecology*, 75(6): 1780-1793.

Barro RJ (1977): Unanticipated money growth and unemployment in the United States. *The American Economic Review*, 67(2): 101-115.

Bessant KC, MacPherson ED (2002): Thoughts on the origins, concepts, and pedagogy of statistics as a “separate discipline”. *The American Statistician*, 56(1): 22-28.

Blundell R, Powell JL (2001): Endogeneity in nonparametric and semiparametric regression models. CEMMAP Working Paper CWP09/01, The Institute of Fiscal Studies, Department of Economics, University College London.

Boden P, Stillwell J, Rees P (1992): How good are the NHSCR data? In: Stillwell et al. (1992): 13-27.

Bosquet C, Boulhol H (2010): Scale-dependence of the negative binomial pseudo-maximum likelihood estimator. Greqam, Université Aix-Marseille, Working Paper.

Boyle P, Halfacree K (eds.) (1998): Migration into rural areas. Theories and issues. John Wiley & Sons, Chichester/England.

Boyle P, Halfacree K, Robinson V (1998): Exploring contemporary migration. Addison Wesley Longman Ltd., Harlow-New York.

Brillinger D (1986): A Biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. *Biometrics*, 42(4): 693-734 (with discussion).

Cadwallader M (1992): Migration and residential mobility. Macro and micro approaches. The University of Wisconsin Press, Madison.

Cameron AC, Trivedi PK (1985): Regression based tests for overdispersion. Technical Report No. 9, Econometric Workshop, Fourth Floor, Encina Hall, Stanford University.

- Cameron AC, Trivedi PK (1998): Regression analysis of count data. Cambridge University Press.
- Champion T (1992): Migration in Britain: research challenges and prospects. In: Champion and Fielding (1992): 215-226.
- Champion T (1998): Studying counterurbanisation and the rural population turnaround. In: Boyle and Halfacree (1998): 21-40.
- Champion T, Fielding T (eds.) (1992): Migration processes and patterns. Volume 1. Research progress and prospects. Belhaven Press, London-New York.
- Chiang CL (1984): The life table and its applications. Krieger Publishing Company, Malabar/Florida.
- Chiang CL (1986): Discussion on the paper by David R. Brillinger. In: Brillinger (1986): 714-717.
- Congdon P (1991): An application of general linear modelling to migration in London and South East England. In: Stillwell and Congdon (1991): 113-136.
- Congdon P (1992): Aspects of general linear modelling of migration. *The Statistician*, 41(2): 133-153.
- Congdon P (1993): Approaches to modelling overdispersion in the analysis of migration. *Environment and Planning A*, 25(10): 1481-1510.
- Cooke TJ (2013): All tied up: Tied staying and tied migration within the United States, 1997 to 2007. *Demographic Research*, 29: 817-836.
- Cushing B, Poot J (2004): Crossing boundaries and borders: Regional science advances in migration modelling. *Papers in Regional Science*, 83(1): 317-338.
- DaVanzo J (1976): Why families move: A model of the geographic mobility of married couples. United States Employment and Training Administration, Department of Labor, Rand Corporation.
- Davidian M, Carroll RJ (1987): Variance function estimation. *Journal of the American Statistical Association*, 82(400): 1079-1091.
- Davidian M, Carroll RJ (1988): A note on extended quasi-likelihood. Technical Report # 7, Texas A&M University, College Station/Texas.
- Davies R (1991): The analysis of housing and migration careers. In: Stillwell and Congdon (1991): 207-227.
- Dean CB, Eaves DM, Martinez CJ (1993): A comment on the use of empirical covariance matrices in the analysis of count data. *Journal of Statistical Planning in Inference*, 48(2):

197-205.

Delgado MA, Kniesner TJ (1997): Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism. *The Review of Economics and Statistics*, 79(1): 41-49.

Deming WE, Stephan FF (1941): On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36(213): 45-49.

DeSimone J (2002): Illegal drug use and employment. *Journal of Labor Economics*, 20(4): 952-977.

Dong Y (2009): Endogenous regressor binary choice models without instruments, with an application to migration. Working Paper, Department of Economics, Boston College.

Eeckhout J (2004): Gibrat's law for (all) cities. *The American Economic Review*, 94(5): 1429-1451.

Elandt-Johnson RC, Johnson NL (1980): *Survival models and data analysis*. John Wiley & Sons, New York.

Etzo I (2008): Internal migration: A review of the literature. Working Paper, University of Cagliari.

Fielding T (1992): Migration and culture. In: Champion and Fielding (1992): 201-214.

Fielding T (1998): Counterurbanisation and social class. In: Boyle and Halfacree (1998): 41-60.

Fleiss JL, Levin B, Paik MC (2003): *Statistical methods for rates and proportions*. Third edition. John Wiley & Sons, Hoboken/New Jersey.

Flowerdew R (1991): Poisson regression modelling of migration. In: Stillwell and Congdon (1991): 92-112.

Flowerdew R (1992): Labour market operation and geographical mobility. In: Champion and Fielding (1992): 135-147.

Flowerdew R, Aitkin M (1982): A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, 22(2): 191-202.

Flowerdew R, Boyle PJ (1995): Migration models incorporating interdependence of movers. *Environment and Planning A*, 27(9): 1493-1502.

Forrest R, Murie A (1992): Housing as a barrier to the geographical mobility of labour. In: Champion and Fielding (1992): 77-102.

Frees EW (1992): Forecasting state-to-state migration rates. *Journal of Business & Economic Statistics*, 10(2): 153-167.

- Galle OR, Burr JA, Potter LB (1993): Rethinking measures of migration: On the decomposition of net migration. *Social Indicators Research*, 28(2): 157-171.
- Gordon I (1991): Multi-stream migration modelling. In: Stillwell and Congdon (1991): 73-91.
- Gordon I (1992): Modelling approaches to migration and the labour market. In: Champion and Fielding (1992): 119-134.
- Gorton M, White J, Chaston I (1998): Counterurbanisation, fragmentation and the paradox of the rural idyll. In: Boyle and Halfacree (1998): 215-235.
- Gourieroux C, Montfort A, Trognon A (1984a): Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3): 681-700.
- Gourieroux C, Montfort A, Trognon A (1984b): Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica*, 52(3): 701-720.
- Green A (1992): Changing labour processes and internal migration. In: Champion and Fielding (1992): 105-118.
- Greene WH (2003): *Econometric analyses. Fifth Edition (International Edition)*. Prentice Hall, Pearson Education, Upper Saddle River/New Jersey.
- Greene WH (2007): Correlation in the bivariate Poisson regression model. NYU Working Paper, 2451/26045.
- Greenwood MJ (1975): Simultaneity bias in migration models: An empirical examination. *Demography*, 12(3): 519-536.
- Grundy E (1992): The housing dimension in migration research. In: Champion and Fielding (1992): 165-174.
- Gurland J (1957): Some interrelations among compound and generalized distributions. *Biometrika*, 44(1-2): 265-268.
- Hamnett C (1992): House price differentials, housing wealth and migration. In: Champion and Fielding (1992): 55-64.
- Härdle W, Müller M, Sperlich S, Werwatz A (2004): *Nonparametric and semiparametric models*. Springer-Verlag, Berlin-Heidelberg.
- Hardin JW (2002): The robust variance estimator for two-stage models. *The Stata Journal*, 2(3): 253-266.
- Hardin JW, Schmiedliche H, Carroll RJ (2003): Instrumental variables, bootstrapping, and generalized linear models. *The Stata Journal*, 3(4): 351-360.
- Haurin D, Haurin J (1991): Youth migration in the United States: Analysis of a deindustrializing region. In: Stillwell and Congdon (1991): 168-186.

- Hausman J, Hall BH, Griliches Z (1984): Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, 52(4): 909-938.
- Hilbe JM (2011): Negative binomial regression. Second Edition. Cambridge University Press.
- Hinde J, Demétrio CGB (1998): Overdispersion: Models and estimation. A short course for SINAPE 1998 (13th Brazilian Symposium of Probability and Statistics, Caxambu, Minas Gerais, July 1998).
- Hoem JM (1986): Discussion on the paper by David R. Brillinger. In: Brillinger (1986): 717-719.
- Holly A, Montfort A, Rockinger M (2008): Fourth order pseudo maximum likelihood methods. Working Paper No. 08/02, Institute of Health Economics and Management, University of Lausanne, Faculty of Business and Economics.
- Hugo G, Bell M (1998): The hypothesis of welfare-led migration to rural areas: The Australian case. In: Boyle and Halfacree (1998): 107-133.
- Hyndman RJ, Booth H (2007): Stochastic population forecasts using functional data models for mortality, fertility and migration. Monash University, Department of Econometrics and Business Statistics, Working Paper 14/06.
- Iwasaki M, Tsubaki H (2005): A bivariate generalized linear model with an application to meteorological data analysis. *Statistical Methodology*, 2(3): 175-190.
- Jorgensen B (1987): Exponential dispersion models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 49(2): 127-162 (with discussion).
- Jung RC, Winkelmann R (1993): Two aspects of labor mobility: A bivariate Poisson regression approach. *Empirical Economics*, 18(3): 543-556.
- Karlis D, Ntzoufras I (2003): Analysis of sports data using bivariate Poisson models. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 52(3): 381-393.
- Kauermann G, Carroll RJ (2001): A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456): 1387-1396.
- Kegler SR (2007): Applying the compound Poisson process model to the reporting of injury-related mortality rates. *Epidemiological Perspectives & Innovations*, 4:1.
- Keiding N (1986): Discussion on the paper by David R. Brillinger. In: Brillinger (1986): 721-724.
- Keiding N (1990): Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 332(1627): 487-509.
- King G (1989): A seemingly unrelated Poisson regression model. *Sociological Methods and*

Research, 17(3): 235-255.

Kingman JFC (1993): Poisson processes. Oxford Science Publications, Clarendon Press, Oxford.

Kish L (1986): Discussion on the paper by David R. Brillinger. In: Brillinger (1986): 724-725.

Kuan C-M (2013): National Taiwan University Fall 2013 Lecture Notes: Department of Finance, Econometric Theory I, and Department of Economics and International Business, Econometric Theory III.

Kocherlakota S, Kocherlakota K (1992): Bivariate discrete distributions. Marcel Dekker, New York.

Lamonica GR, Zagaglia B (2013): The determinants of internal mobility in Italy, 1995-2006: A comparison of Italians and resident foreigners. Demographic Research, 29: 407-440.

Lebreton J-D, Burnham KP, Clobert J, Anderson DR (1992): Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs, 62(1): 67-118.

Lejeune B (1997): Second-order pseudo-maximum likelihood estimation and conditional variance misspecification. CORE Discussion Paper 9779, Center for Operations Research & Econometrics, Université Catholique de Louvain.

LeSage JP, Pace RK (eds.) (2004): Spatial and spatiotemporal econometrics. Elsevier, Oxford.

Maier G, Weiss P (1991): The discrete choice approach to migration modelling. In: Stillwell and Congdon (1991): 17-33.

Manton KG (1986): Discussion on the paper by David R. Brillinger. In: Brillinger (1986): 725-729.

Manton KG, Woodbury MA, Stallard E (1981): A variance components approach to categorical data models with heterogeneous cell populations: Analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. Biometrics, 37(2): 259-269.

Mayer WJ, Chappell WF (1992): Determinants of entry and exit: An application of the compounded bivariate Poisson distribution to U.S. industries, 1972-1977. Southern Economic Journal, 58(3): 770-778.

McCullagh P (1983): Quasi-likelihood functions. The Annals of Statistics, 11(1): 59-67.

McCullagh P, Nelder JA (1989): Generalized linear models. Second Edition. Chapman&Hall/CRC, London.

Milne W (1991): The human capital model and its econometric estimation. In: Stillwell and Congdon (1991): 137-151.

- Milz J (2013): Integrierte Lohn- und Einkommensteuerstatistik für das Jahr 2010. *Statistische Nachrichten*, 68(7): 599-612.
- Mincer J (1978): Family migration decisions. *Journal of Political Economy*, 86(5): 749-773.
- Mueser PR, White MJ (1987): Explaining the association between rates of in-migration and out-migration. Research Report, Department of Economics, University of Missouri-Columbia & The Urban Institute, Washington DC.
- Murphy KM, Topel RH (1985): Estimation and inference in two-step econometric models. *Journal of Business & Economics Statistics*, 3(4): 370-379.
- Nelder JA, Pregibon D (1987): An extended quasi-likelihood function. *Biometrika*, 74(2): 221-232.
- Newey WK, McFadden D (1994): Large sample estimation and hypothesis testing. In: Engle DR, McFadden, D (eds.): *Handbook of econometrics*. Elsevier, Amsterdam: 2111-2245.
- Owen D, Green A (1992): Migration patterns and trends. In: Champion and Fielding (1992): 17-40.
- Özel G (2011): A bivariate compound Poisson model for the occurrence of foreshock and aftershock sequences in Turkey. *Environmetrics*, 22(7): 847-856.
- Pan W (2001): Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1): 120-125.
- Pendakur R, Young N (2013): Putting on the moves: Individual, household, and community-level determinants of residential mobility in Canada. *Demographic Research*, 29: 767-796.
- Pocock SJ, Cook DG, Beresford SAA (1981): Regression of area mortality rates on explanatory variables: What weighting is appropriate? *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 30(3): 286-295.
- Ravenstein EG (1885): The laws of migration. *Journal of the Statistical Society of London*, 48(2): 167-235.
- Riphahn RT, Wambach A, Million A (2003): Incentive effects in the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics*, 18(4): 387-405.
- Rivers D, Vuong QH (1988): Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3): 347-366.
- Robinson V (1992): Move on up: The mobility of Britain's Afro-Caribbean and Asian populations. In: Stillwell et al. (1992): 271-291.
- Rogers A (1990): Requiem for the net migrant. *Geographical Analysis*, 22(4): 283-300.
- Rogers A, Castro LJ (1981): Model migration schedules. *International Institute for Applied*

System Analysis (IIASA) Research Report RR-81-030, Laxenburg/Austria.

Rousseeuw PJ, Leroy AM (1987): Robust regression and outlier detection. John Wiley, New York.

Salt J, Kitching R (1992): The relationship between international and internal labour migration. In: Champion and Fielding (1992): 148-162.

Sandefur G, Tuma N, Kephart G (1991): Race, local labour markets and migration in the United States, 1975-83. In: Stillwell and Congdon (1991): 187-206.

Sandell SH (1977): Women and the economics of migration. *The Review of Economics and Statistics*, 59(4): 404-414.

Scherbov S, Ediev D (2011): Significance of life table estimates for small populations: Simulation-based study of standard errors. *Demographic Research*, 24: 527-550.

Shea DG, Terza JV, Stuart BC, Briesacher B (2007): Estimating the effects of prescription drug coverage for Medicare beneficiaries. *Health Services Research*, 42 (3:I): 933-949.

Shin J, Moon S (2007): Do HMO plans reduce health care expenditure in the private sector? *Economic Inquiry*, 45(1): 82-99.

Sjastaad LA (1962): The costs and returns of human migration. *The Journal of Political Economy*, 70(5:2): 80-93.

Smith SK, Swanson DA (1998): In defense of the net migrant. *Journal of Economic and Social Management*, 24(3-4): 249-264.

Smyth GK (1989): Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B (Methodological)*, 51(1): 47-60.

Spiegelman M (1968): Introduction to demography. Revised Edition. Harvard University Press, Cambridge/Massachusetts.

Srivastava VK, Giles DEA (1987): Seemingly unrelated regression equations models. Estimation and inference. Marcel Dekker, New York.

Stefanski LA, Boos DD (2002): The calculus of M-estimation. *The American Statistician*, 56(1): 29-38.

Stillwell J (1991): Spatial interaction models and the propensity to migrate over distance. In: Stillwell and Congdon (1991): 34-56.

Stillwell J, Congdon P (eds.) (1991): Migration models: macro and micro approaches. Belhaven Press, London-New York.

Stillwell J, Rees P, Boden P (eds.) (1992): Migration processes and patterns. Volume 2. Population redistribution in the United Kingdom. Belhaven Press, London-New York.

- Tervo H (2001): Does the compositional effect explain the association between rates of in-migration and out-migration? 41st European Regional Science Association Congress, Zagreb, August 29 – September 1, 2001.
- Terza JV, Basu A, Rathouz PJ (2008): Two-stage residual inclusion estimation: Addressing endogeneity in health econometrics modelling. *Journal of Health Economics*, 27(3): 531-543.
- Thaler T (2009): Die Extended-Quasi-Likelihood-Funktion in Generalisierten Linearen Modellen. Masterarbeit, Institut für Statistik der Technischen Universität Graz.
- Thall PF, Vail SC (1990): Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46(3): 657-671.
- Udry JR, Teddie C, Suchindran CM (1979): The random variation in rates based on total enumeration of events. *Population Studies*, 33(2): 353-364.
- Warnes T (1992): Migration and the life course. In: Champion and Fielding (1992): 175-187.
- Winkelmann R (2000): Seemingly unrelated negative binomial regression. *Oxford Bulletin of Economics and Statistics*, 62(4): 553-560.
- Winkelmann R (2010): *Econometric analysis of count data*. Fifth Edition. Springer-Verlag, Berlin-Heidelberg.
- Wolter KM (1985): *Introduction to variance estimation*, Springer-Verlag, New York.
- Wonka E (2008): Neuabgrenzung des Dauersiedlungsraums. *Statistische Nachrichten*, 63(5): 432-442.
- Wooldridge JM (1997): Quasi-likelihood methods for count data. In: Pesaran MH, Schmidt P (eds.): *Handbook of applied econometrics*. Vol. II, Microeconometrics. Blackwell, Oxford-Malden: 353-406.
- Wooldridge JM (2010): *Econometric analysis of cross section and panel data*. Second edition. The MIT Press, Cambridge-London.
- Xekalaki E (1983): The univariate generalized Waring distribution in relation to accident theory: Proneness, spells or contagion? *Biometrics*, 39(4): 887-895.
- Ziegler A (2011): *Generalized estimating equations*. Springer, New York-Heidelberg-Dordrecht-London.

B Abstract

The past years have seen an increasing interest in modeling aggregate demographic counts as realizations of random variables. A basic model assumption is the conditional variance of a count. Established models for migration counts have often applied the Poisson regression model or semiparametric extensions like Poisson quasi likelihood.

We review the assumptions of the Poisson process and their apparent violations in migration flows. For annual migration counts of rural municipalities we propose a two-parameter variance function which represents clustering of persons by household migration and unobserved heterogeneity between municipalities. The variance function has an interpretation as quasi negative binomial variance with specified shape parameter. Provided the conditional mean of the migration count is correctly specified, consistent and asymptotically efficient parameter estimation is achieved by quasi-generalized pseudo maximum likelihood using a $\sqrt{n}$ -consistent estimator (ideally a robust estimator) of the shape parameter.

A possible misspecification of the conditional mean arises from simultaneity between migration count and person-years lived in a migration rate model. Instead of traditional demographic approaches to this endogeneity problem, we propose a method based on instrumental variables, namely two-stage residual inclusion. Our model is illustrated by applying it to the 2012 annual totals of in- and out-migration of the 171 municipalities of Burgenland, a rural federal province of Austria.

C Abstract in German

In den vergangenen Jahren haben stochastische Modelle von Zählraten in der Demographie an Popularität gewonnen. Eine wesentliche Annahme dieser Modelle ist die bedingte Varianz der Ereignishäufigkeit. Etablierte Ansätze für Migrationsströme wenden zumeist die Poisson-Regression oder semiparametrische Erweiterungen wie Poisson-Quasi-Likelihood an.

In der vorliegenden Dissertation werden die dem Poisson-Prozess zu Grunde liegenden Annahmen und ihre offensichtlichen Verletzungen bei Migrationsströmen besprochen. Für Jahres-Migrationshäufigkeiten von Gemeinden in ländlichen Regionen wird eine zweiparametrische Varianzfunktion hergeleitet, deren Parameter einerseits das Clustering von Personen durch Haushaltsmigration und andererseits unbeobachtete Heterogenität zwischen den Gemeinden repräsentieren. Diese Varianzfunktion kann als Varianz der Quasi-Negativen Binomialverteilung mit bekanntem Formparameter aufgefasst werden. Sofern der bedingte Erwartungswert der Migrationshäufigkeit korrekt spezifiziert ist, wird mit dem Quasi-Generalized-Pseudo-Likelihood-Schätzer eine konsistente und asymptotisch effiziente Schätzung erzielt, wobei für den Formparameter ein $\sqrt{n}$ -konsistenter (idealerweise robuster) Schätzer eingesetzt wird.

Eine mögliche Fehlspezifikation des bedingten Erwartungswerts ergibt sich aus der Simultanität zwischen der Migrationshäufigkeit und den durchlebten Personenjahren in einem Migrationsratenmodell. Anstelle der bisher in der Demographie üblichen Verfahren zur Behandlung dieses Endogenitätsproblems wird ein Ansatz basierend auf Instrumentalvariablen vorgeschlagen, konkret ein zweistufiges Modell mit Residueninklusion. Das Modell wird durch eine Anwendung auf die Migrationsströme des Jahres 2012 der 171 Gemeinden des Burgenlandes illustriert.

D Curriculum Vitae

Personal information

Name	Johannes Klotz
Date of birth	05 September 1981
Place of birth	Innsbruck, Austria
Marital status	Married
Citizenship	Austrian

Education

2009-	PhD program of statistics and operations research, University of Vienna
2007-2008	Master studies of statistics, University of Vienna Thesis: Parametric (lognormal) estimation of the at-risk-of-poverty rate
2002-2006	Bachelor studies of statistics, University of Vienna
2000	High school diploma

Employment

2011-	University of Vienna Department of Statistics and Operations Research Department of Geography and Regional Research
2001-	Statistics Austria Directorate Social Statistics
2000-2001	Compulsory military service

Research interests

- Demographic indicators, measurement of demographic processes
- Stochastic approaches in aggregate demography
- Regression models of count data
- Differential mortality and morbidity
- Demography of rural areas
- Small-area estimation