ECOLOGICAL INFERENCE AND AGGREGATE ANALYSIS OF ELECTIONS

by

Won-ho Park

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Political Science)
in The University of Michigan
2008

Doctoral Committee:

Professor Christopher H. Achen, Co-Chair, Princeton University
Professor Kenneth W. Kollman, Co-Chair
Professor John E. Jackson
Professor Michael W. Traugott

To my parents, Yong-Tae Park and Kyung-ja Lee

ACKNOWLEDGEMENTS

I thank my advisors, Chris Achen, Ken Kollman, John Jackson, and Mike Traugott for their support and guidances over the years. I am also indebted to JungHwa Lee, Mike Hanmer, Sang-jung Han, Corrine McConnaughy, Ismail White, Clint Peinhardt, and Kwang-Il Yoon for many insights and lasting friendships. The project was supported by the Institute of International Education (Fulbright Graduate Degree Study Awards), the Rotary International (Rotary Ambassadorial Scholarship), the American National Election Studies (Pre-Doctorate Research Award), the Horrace H. Rackham Graduate School (Dissertation Fellowship), and the Department of Political Science at the University of Michigan.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 The Study of Elections and Voting

Elections are aggregation processes, and a majority of electoral analysis is bound to focus upon looking at electoral returns as a starting point at the least. Electoral outcomes are determined when the votes are tallied up in a given electoral unit, and such an outcome of elections constitutes *the* natural unit of analysis. For example, it could be either how the electoral fate of an incumbent party is swayed by the economic conditions over time nationwide, or how a given electoral district's peculiar configuration explains any electoral outcomes unique to that district. Strongly embedded in our language are such phrases as "how the country decided ...," or "how the district has chosen ...," and studying elections with aggregate information makes sense.

Voting is also an individual behavior. It is neither the country nor the districts that decide, but the voters that make choices. Analyzing elections and explaining the outcomes would necessarily involve a certain theory of voters.

Particularly meaningful in this context are the successes and achievements that survey methodology brought into the study of elections since the end of the Second World War. Epitomized in the American National Election Studies (ANES)

and followed by many survey studies in various democracies, this research method addresses questions on the psychological factors and rational calculus that work within voters.

Survey research combined with powerful statistical tools has produced perhaps the most enduring successes in quantitative electoral studies. Thanks to the ingenuity in the design and subsequent analyses of the collective body of the ANES, we now know more about the American voter than we did half a century ago. Also, many concepts and methodology developed through the ANES experience provide an excellent benchmark for the study of voters in other democracies as well, as is evidenced in the electoral studies conducted in many countries. For example, the Comparative Study of Electoral Systems (CSES) includes election survey studies from more than thirty countries in the world, with an eye towards unifying the study of voting behavior in the comparative politics context.

Survey measures are by no means perfect. Respondents are known to lie, forget, or misinterpret the questions when they face an interviewer or when they fill in a questionnaire. Yet, the simple virtue of directly measuring the voter's behavior, intention, and characteristics enables the researcher to examine the direct linkage between them. This exactly is the reason why survey studies have dominated electoral studies in the past fifty years with so many successes.

1.2 Limitations of Survey Studies in Studying Electoral Politics

Even though survey studies have opened up opportunities to investigate the psyche and attributes of the voters, there are inherent limitations that come with the approach. These are not simple problems—sometimes the problems would justify the usage of aggregate data in electoral studies over survey data.

First of all, survey data are limited in availability. Most importantly, since the collection of survey data at a large scale started relatively recently, it is impossible to address questions directed at individual-level relationships on elections that predate the Second World War. If we move outside of the United States and look into other countries with shorter history of election survey studies, the researcher, when solely dependent on survey data, is bound to only a few very recent elections.

Including the canonical examples of how and why German voters in the Third Reich supported the National Socialist Party, or how the Republican Party rose to power in the US, studies on many elections of historical importance would just have to resort to aggregate data: there simply is no individual-level data regarding these elections.

Secondly, in addition to the simple availability issue, survey data are subject to questions on their general reliability. To study elections under undemocratic or semi-democratic regimes would be difficult since survey data on these elections are scarce, and even when they are available, reliable data are even scarcer. Take the example of South Korean elections: under the authoritarian regime in the early 1980s, the conduct of asking and publishing the electoral intention of the voters before the election was pronounced illegal since it could "skew the electoral results." Moreover, we do not know well how much of their true political preferences the respondents would be willing to reveal under such a political climate.

In fact, the reliability issue is more of a problem in new democracies where the history of survey study runs shallow. Without the benefits of accumulated skills and experience, many problems plague electoral studies in different countries. Questionnaires are merely blind translations of ANES sometimes, featuring ques-

tions on "foreign" issues such as abortion in native languages; to make matters worse, they change often and drastically.

Election survey studies in such countries will continue to improve and help us understand the voters in new or semi-democracies. However, at the same time, there is no reason to exclude aggregate level analysis to at least complement the study of voters through survey data.

Aggregate data face fewer challenges in terms of availability and reliability problems. Exactly because elections are determined by the tallied votes, aggregate results of elections in the past are well archived, if not always in an electronic form that is readily available. The availability issue is not so much of a question under authoritarian regimes or in new democracies; barring election fraud, the reliability issue seems like a minor concern as well with aggregate data. Campaign processes may be questionable, but counting votes is usually not.

In fact, it is usually the case that efficient bureaucracies under authoritarian regimes meticulously collect and archive detailed electoral returns—often conscious of the legitimacy issue within the electoral process—as well as detailed census information of the voters. It would be unwise to disregard the entire body of rich and available information that aggregate data provide when survey studies are plagued with availability and reliability problems.

Thirdly, additional to the availability and reliability issues, there are inherent limitations to the survey approach addressing the dynamic aspect of elections (Achen and Shively 1995). For example, there is the question of whether surveys can detect electoral realignment of voters over time. By design, surveys cannot directly address the question of how electoral processes develop over time.

Surveys will usually give a clear depiction of the cross-sectional picture of the

electorate at a given time or in one election where the researcher will look into the static relationship that shapes the electoral choice of the voter. It would be hard to capture the impact of the past on current elections, since the past can not be measured or is difficult to measure in the cross-sectional snapshots surveys usually provide.

Measurements regarding past elections in the survey context can be achieved by adding questions asking the voters to recall back a few elections, or can be implemented in the context of panel studies. However, there is a limit to how long a time span can be measured looking at the history of the voters, and even when possible, there are additional problems related to panel attrition (Jennings and Markus 1984). The accuracy of recalling the past is also a problem.

Except in rare cases, survey studies will not allow the researcher to investigate the long-term dynamics of elections. This problem will prevent researchers, if they rely solely upon survey studies, from investigating elections from a historical standpoint of view, where changes are not always immediate but subtle.

In contrast, aggregate data can provide the researcher with opportunities to study the electoral dynamics directly. Regardless of the levels of aggregation, temporal variation of electoral returns can be put into a long time-series, which will constitute the major dependent variable of interest.[1] Explaining it with past elections and other temporally varying characteristics of the aggregate unit directly matches the data structure.

So far, I have discussed the problems that individual level survey data present, and highlighted the comparative advantage of aggregate data in terms of their availability, reliability, and ability to deal with electoral dynamics. However, it

[1] It should be noted that at certain levels of aggregation, the problem of boundary changes, such as electoral redistricting, can be a problem in this regard.

is important to note that studying individual level relationships using aggregate data—ecological inference—comes with many problems, too. The discussion in the next section considers these difficulties.

1.3 The Ecological Inference Problem

In previous sections, I have spelled out several simple points: that individual level relationships in the study of elections are important and interesting; that survey data directly address this relationship and have greatly contributed to the scholarship on elections; and that under certain circumstances, it is usually the case that survey data are simply not available and the aggregate level information is all there is left for us to analyze. In this section, I will discuss the problems related to using aggregate data to infer individual-level relationships: this is known as the ecological inference problem.

Consider an individual-level relationship that the researcher is studying, namely how an individual's past electoral choice, x, is related to her current vote, y. One would write a model such as

$$y_i = f(x_i, \cdot) \tag{1.1}$$

and would study the particular function f that describes the voter transition from one election to another.

However, this relationship, or the function f, will not hold at the aggregate level, since both variables will be transformed by unknown forces—"nature"—that determine, for instance, why and how certain people live close to one another. It is possible to think of an aggregating function, G, that sorts and selects individuals into different groups. For example, $Y_j = G(y_{ij})$ and $X_j = G(x_{ij})$, where X_j and Y_j would denote the aggregation of individual-level variables, $x_{ij}$and y_{ij}, in

district j. Plugging these into equation (1.1), we would get

$$G^{-1}(Y_j) = f\left[G^{-1}(X_j), \cdot\right]. \tag{1.2}$$

Then we can see that the relationship that we are mostly interested in, f, will be hard to trace back with aggregate information. Recovering the original micro-level relationship, f, using X and Y by adding assumptions about how G works, would be the definition of ecological inference. What might be a simple connection with data on individuals becomes a difficult inferential question with grouped data. Admittedly, this is a tough challenge to overcome, yet at the same time, as was argued above, it may be the only investigation tool left for the researcher.

Another difficulty with equation (1.2) is that it is challenging and perhaps fruitless to look for a *general* solution to ecological inference problems. Since the key piece of information in the ecological inference process is how the grouping function G aggregates the variables, more specific context of how the key variables of our interest, x and y are aggregated by G becomes vital information.

1.4 The Voter Transition model

In studying electoral dynamics, what is called the voter transition model parsimoniously describes the change in party support across elections. The model provides an important insight into the stability and volatility of elections over time. At the same time, the model is versatile enough to be modified to fit the researchers' need. The model's standard focus is on the partisan shift of the voters over time, but it can be also employed to investigate the dynamics of turnout and the correlates of it, or the straight-ticket voting patterns within elections. Such various applications of the voter transition model approaches are provided with

examples in later chapters.

To be sure, applying the model on aggregate data raises typical ecological inference problems. Exactly because of the secrecy principle of voting, it is only the aggregate marginals that are known to the researcher, and the task is to learn about the proportions of individual voter movement from one party to another across elections—namely, the loyalty and defection rates of the voters that describe how consistently the voters behave across time.

It turns out that the transition model comes with several inherent problems that make the ecological inference process more difficult than usual. First of all, the model usually comes with severe aggregation bias problems. As was described earlier, aggregation problems generally arise when individuals with similar attributes that are being studied are more likely to be grouped into same aggregate units. For example, if loyal Democrats live closer to one another and the same holds true for loyal Republican voters, aggregation bias is introduced.

Non-linearity is also a defining trait of the voter transition model. Characteristic forces of electoral choice such as uniform swings and emergence of minor parties typically complicate existing linear ecological inference strategies.

Accommodating multipartism is also another serious challenge when using the voter transition model. Political studies on real-world party systems would usually require incorporating more than three parties into the analysis. The current ecological inference literature has yet to overcome the problem of filling in the two-by-two table, when there are more than two parties, largely because of the fundamental limitations that come with the aggregate data. However, I argue we can do better.

Finally, it should be mentioned that the voter transition model by itself is a

limited description of elections. The electoral movement of the voters from one choice to another over time is an interesting and important matter in studying political dynamics. However, while individual-level data would typically tease out answers to *who* changes electoral support and *why* the movement occurs, these questions are most challenging for ecological analysis on aggregate data. In a later chapter, I develop an extension to the voter transition model that seeks to overcome this challenge.

1.5 Outline of Chapters

1.5.1 Current Strategies

The following chapters investigate the ecological inference problem focusing on the analysis of aggregate election data and the voter transition model.

In the second chapter, I analyze existing ecological inference techniques in the context of the voter transition model and put them under test comparing the ecological estimates against individual-level results. In a setting where two and only two parties compete in two elections, which is the simplest possible voter transition setup, the chapter provides ground work for later extensions to the voter transition model.

As it turns out, conventional linear approaches such as the Goodman regression (1959) and Gary King's EI technique (1997) routinely fail, and I show that the ecological inference strategies break down when facing challenges such as severe aggregation problems and nonlinearity in the data.

Thomsen's model (1987) is particularly attractive in the sense that it is based upon a consistent micro-level model of voter preferences and their electoral choices. Using data sets from South Korean elections and British parliamentary elections,

I show that the suggested technique produces successful ecological estimates that conform closely to panel or survey results.

I also provide a new criterion to evaluate ecological estimators, namely, disaggregation consistency, which requires ecological estimators to produce better answers with finer-grained data. Both theory and empirical results suggest that the Thomsen estimator is disaggregation-consistent, while linear ecological estimators are not.

1.5.2 Extensions to Multiparty Systems

The third chapter examines the possibility of an extension to the ecological voter transition model in multiparty systems. As mentioned, elections in most democratic political regimes in the world will include competition among more than two parties or candidates. Moreover, if the researcher decides to include abstention as a separate political choice, it becomes the case that the simplest voter transition model cannot describe the simplest two-party system that exists, for example, in the US.

Political methodologist have yet to develop more reliable techniques to model multipartism. Even with individual-level data, multinomial choice models are complex and require serious computing power. A typical multinomial probit estimation would usually require the estimation of the cross-choice covariance structures, with the necessary number of parameters to estimate rapidly increasing with more choice categories. Whether and how individual level models on multinomial choices can provide direct insight into the analysis of aggregate data is a subject that calls for further investigation.

In the ecological inference literature, approximations by adjusting first-round estimates remain to be the principal strategy in dealing with multiple choice prob-

lems. However, I show that some existing approximations, including Thomsen's method, are only moderately convincing, if serviceable.

In the chapter, I provide an application of the iterative proportional fitting (IPF)—which is also known as "raking"—to the context of adjusting ecological inference estimates. Essentially, IPF is a process that "rakes" the first-round ecological estimates in iteration to conform to the given population marginal distributions in the two elections, with the least amount of total change in the first-round estimates.

Theoretically, it makes sense to emphasize the information of population total marginals, since the information is discarded or transformed when the first-round ecological estimates are computed based upon merged categories. Empirically, I provide results that show IPF outperforms other alternatives when applied to multiparty voter transition problems. The evaluation of the method uses examples from South Korean elections and the Miami-Dade Florida ballot image data from the 2000 election in the US.

Several additional attractive features of applying IPF in the multiparty context should be highlighted. First of all, the IPF solution is a unique solution to the problem, and it eliminates the arbitrary choice of "reference parties," or which parties to merge first to conduct a two-by-two estimation. Second, it is independent of the ecological inference technique that produces the first-round estimates, and can generally be incorporated into other ecological inference strategies.

1.5.3 The Covariate Model

In the fourth chapter, I discuss and develop an extension of the Thomsen model to accommodate additional covariates into the voter transition model. It is an attempt to equip the voter transition model with an analytical mechanism that will

enable the researcher to study the correlates of voter movements across parties.

In a way, the question that the voter transition model asks can be characterized to be descriptive and somewhat coarse, even though the correct answer to the question is sufficiently hard to come across with aggregate data. The voter transition model typically studies the relative size of the movement of the voters, but it would usually be the case that the researcher will want to go a step further and ask *who* switched parties and *why* the transition occurred.

To this end, the Thomsen model is reinterpreted as a method of moment approach, and the model is extended to include covariates—simple fractions of people who belong to a certain demographic group, say, the proportion of workers, in a district—which will help the researcher to look into i) separate levels of party support in the demographic groups (different levels of support among workers and non-workers), and ii) separate sets of transition rates in the demographic groups (different loyalty and defection rates among workers and non-workers). This will enable the researcher to approach the more detailed mechanism that lies beneath the voter transitions: for example, it opens up a new way to study electoral realignments.

The example provided at the end of the chapter is a detailed picture of how South Korean voters responded to the democratic reform of the country in 1987. To study the impact of the major political transformation of the country on the voters, I look into the change in the turnout pattern of South Korean voters employing the developed model. More specifically, I examine whether formerly alienated voters started to come to the polling booth after the democratization; at the same time, I examine whether there would be any sign of decrease in turnout among formerly mobilized voters in post-democratization elections.

The foregoing investigation is to test whether and to what degree voter efficacy and mobilization played a role in subduing democratic demands of the electorate in pre-democratization elections; and whether and to what degree such political forces changed and shaped the electoral landscape in the post-democratization period. Answers to these questions will have further implications for the civil society thesis, which states that the ripening of civil societal groups was the main factor in bringing about the democratization of South Korea, not unlike in other new democracies that emerged and were labeled as the "Third Wave of Democracy." (Huntington 1991) Proponents of this thesis have had to rely only upon qualitative analyses or anecdotal descriptions to answer substantively important key questions. The chapter shows that ecological inference techniques can open up the possibility of answering such questions with quantitative evidence.

CHAPTER II

Voter Transition Rates and Ecological Inference

2.1 Introduction

Ecological inference can be stated in a general econometric context as a problem of statistical under-identification. The data obtained at the aggregate level are not sufficient to determine the micro-level data generating process. In other words, the inference requires the addition of strong assumptions or external information, upon which the validity and accuracy of the estimators heavily rely.

Gathering individual level data, such as those from survey studies, would be a more direct way to grapple with individual relationships: individual data require milder assumptions and can produce more desirable estimates. However, as every ecological inference study claims, the cost or availability of such data makes the employment of ecological inference often inevitable. For example, researchers bearing historical questions of quantitative nature would more often than not have to deal with aggregate data. Moreover, in most non-Western countries, researchers are faced by survey data problems that are more complicated than their mere availability. Without the benefits of accumulated skills and experience of conducting survey studies, the sampling process could be dubious; how the respondents in these countries react when asked to reveal their political pref-

erences is not well understood. Compared to survey data, aggregate records such as electoral outcomes or census studies are more accurate, extensive, and tabled down to relatively small geographical units. This is true especially in centralized bureaucratic governments. Thus, in a sense, ecological inference can be useful for comparative political studies of the non-Western world, if only as a complement to individual level studies.

Perhaps it is such promise that spurred the recent discussions and advances in the field of ecological inference (for example, Achen and Shively 1995; King 1997, Rosen *et al*. 2000). However, the techniques are still far from satisfactory—sometimes the suggested techniques work, sometimes they fail. And moreover, "better" aggregate data sets, such as fine-grained data sets, do not always produce better answers. I focus on the voter transition model, a typical ecological inference technique which nevertheless has failed at times. In the following section, I briefly summarize the voter transition model and derive the answers from generic approaches suggested by Goodman and King. I also provide empirical examples where such approaches fail by comparing ecological estimates and individual-level survey results. In Section 3, I discuss problems of aggregation and non-linearity that are native to the voter transition model and argue that such problems cause bias and inconsistency. Section 4 critically summarizes an alternative non-linear model originally suggested by Thomsen (1987), and interprets it in terms of a simple voter utility model. I derive a generalized assumption of the estimator that implies unbiasedness and consistency. Section 5 provides empirical results. The last section briefly concludes.

2.2 Voter Transition Rates and the Failure of Generic Ecological Estimators

2.2.1 The Baseline Model: Ecological Regression

Estimation of voter transition rates from aggregate data is a typical ecological inference problem. The electoral support for n contending parties for two consecutive elections (marginal probability) is known for each observation unit, while there is no information on the individual counts for all the possible n^2 cells (that is, the joint probabilities). The objective of the ecological inference here is to infer the individual voting choice based on the aggregate information, that is, marginal vote fractions.[1]

Suppose that we want to estimate the loyalty and defection rates of voters in two consecutive elections where two and only two parties compete against each other, with no option to abstain and no replacement of any voter. The loyalty and defection rates can be defined as conditional probabilities of voting for a party, say, Democratic, in the second election based on voters' previous choices. For example, suppose we have a random sample of the population of voters, with their voting record for two consecutive elections. Then the following table can be filled in with the conditional probabilities that a voter will either stay loyal or defect from the party for which she voted in election 1.

		Time 1 Vote	
		Democrat	Republican
Time 2 Vote	Democrat	p	q
	Republican	$1-p$	$1-q$

Figure 2.1: Voter Transition Rates in Two Successive Elections

These rates render an individual-level interpretation as well: For voter i, ($i =$

[1] By "voter transition models," I refer to a setup that exclusively deals with two, usually consecutive, sets of electoral returns. Researchers sometimes use the term to refer to relationships between social variables, such as race, as independent variables, and electoral returns as dependent variables, which is quite different questions from what I deal here, both substantively and statistically. I make a distinction between the two, and show the former is a more difficult problem than the later.

$1, 2, ..., N$), let $d_{2i} = 1$ if i voted Democratic, 0 if Republican in the second election, and define d_{1i} in the same fashion for the first election. Then the following relationship holds for each i:

$$Prob(d_{2i} = 1) = p_i d_{1i} + q_i(1 - d_{1i}) \tag{2.1}$$

If one voted Democratic last time ($d_{1i} = 1$), her probability to vote for the Democrats this time is p_i; while if one voted Republican last time ($d_{1i} = 0$), her probability to defect to the Democratic party is q_i. The p and q in Figure 2.1 are simply the mean of p_i's and q_i's and can be expressed as a linear regression at the individual level.

$$\begin{aligned} d_{2i} = pd_{1i} + q(1 - d_{1i}) + u_i &= q + (p - q)d_{1i} + u_i \\ \text{where} \qquad & u_i = (p_i - p)d_{1i} + (q_i - q)(1 - d_{1i}) \end{aligned} \tag{2.2}$$

Despite the dichotomous dependent and independent variables, the OLS estimation will exactly produce the parameters in Figure 2.1 and the estimates will have the usual desirable properties since $E(u_i) = 0$ and $Cov(d_{1i}, u_i) = 0$. In other words, with individual level data, we would be able to describe the exact loyalty and defection rates of the voters.

However, all the information about votes in our data is aggregated and tabulated by precinct, county, or other geographical units. For each unit j, let D_{1j} and D_{2j} denote the fraction of aggregated individual votes for the two consecutive elections shown in continuous forms. Since the relationship (2.2) will hold for any geographical unit, we may average both sides of the equation for unit j with n_j voters and write:

$$D_{2j} = p_j D_{1j} + q_j(1 - D_{1j}) \tag{2.3}$$
$$\text{where} \quad D_{1j} = \sum_i \frac{d_{1ij}}{n_j} \quad \text{and } D_{2j} = \sum_i \frac{d_{2ij}}{n_j}$$

Obviously the equation is not identified. Goodman (1953) suggests if we can reasonably assume that the parameters p_j's and q_j's are constant across districts, or at least mean independent of D_{1j}, we may obtain an aggregate regression relationship

$$D_{2j} = q + (p - q)D_{1j} + U_j \tag{2.4}$$
$$\text{where} \quad U_j = q_j - q + (p_j - p - q_j + q)D_{1j}$$

which is a simple OLS setup. If the assumption holds, the OLS estimates will be unbiased since $E(U_j) = 0$ and $Cov(D_{1j}, U_j) = 0$. Standard errors for the estimated probabilities can be obtained in an obvious way, although they may be subject to heteroscedastic corrections.[2]

2.2.2 King's Estimation Procedure for Ecological Inference

Among the attempts to improve Goodman's method, King (1997) focuses on the "out-of-bounds" problem in the Goodman estimators. The most important insight of King's setup is to take advantage of information that can be provided by the method of bounds. To this end, he adopts a two-stage procedure, first deriving a truncated bivariate normal distribution that underlies the local parameters and then using the distribution to calculate local estimates, which he then combines to create aggregate level estimates. King's estimation process in the context of the

[2] As Achen and Shively (1995) point out, sampling errors are less significant than actual disturbance variances compared to other sources of heteroscedasticity, such as incorrect assumptions. For example, since Goodman assumptions never hold in practice, the disturbance variance introduced by varying parameters always overwhelms the sampling error.

voter transition model can be described in more detail as follows: First, rearrange Equation (2.3), the district accounting identity, into a relationship between p_j and q_j. Then, our data points, (D_{1j}, D_{2j}) can be mapped into lines:

$$q_j = \frac{D_{2j}}{1 - D_{1j}} - \frac{D_{1j}}{1 - D_{1j}} p_j \tag{2.5}$$

Second, on a $[0, 1] \times [0, 1]$ parameter space, draw the lines for all j's. The possible ranges of p_j and q_j for each district exactly represent the possible ranges that can be calculated by the method of bounds. Now, if the Goodman assumption of constant parameters holds, all the lines should intersect at a single point, (p^*, q^*). If the parameters vary randomly across districts around the true point, the line segments will pass near the point. An example of such lines and their smoothed density is shown in Figure 2.

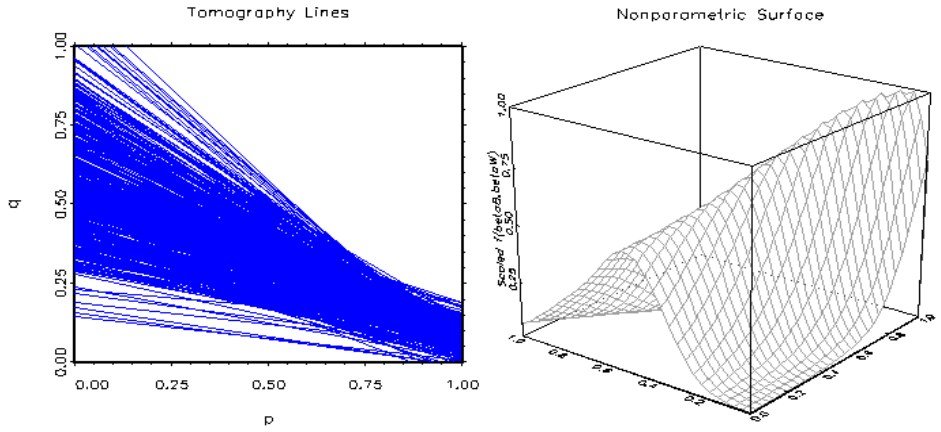


Figure 2.2: Parameter Bounds and Their Density

The lines in the left figure represent possible combinations of (p_j, q_j) for each district. And as is true with any real-world aggregate data, they do not meet at one point. King postulates, then, the density should tell us something about (p_j, q_j), the distribution of the parameters. For example, in Figure 2, (p_j, q_j) should lie somewhere around the lower-right corner with a high value of p and a low q.

According to King, it can further be assumed that the probability density of (p_j, q_j) follows a truncated bivariate normal distribution[3] with correlation ρ_{pq}. The five parameters of the distribution, *i.e.*, the means $(\bar{p}, \bar{q})$, variances (σ_p^2, σ_q^2), and their correlation, can be estimated by maximizing the following likelihood function (p. 134):

$$\begin{aligned} \mathcal{L} &= \prod_j \left[\text{Normal}\left(D_{2j} | \bar{q} + (\bar{p} - \bar{q}) D_{1j}, \sigma^2 \right) \times (\text{Truncation})_j \right] \\ &\text{where } \sigma_j^2 = \sigma_p^2 D_{1j}^2 + \sigma_q^2 (1 - D_{1j})^2 + 2\sigma_{pq}^2 D_{1j}(1 - D_{1j}) \end{aligned} \tag{2.6}$$

which corresponds to a GLS for the Goodman setup, except the truncation part and the construction of the variance term[4].

Next, simulate unit level parameters, p_j's and q's for all j's, from the derived posterior distribution, conditional on equation (2.5). Retrieve their means, $\widetilde{p}_j$'s and $\widetilde{q}_j$'s. Then, the weighted averages of the means represent King's estimates for the aggregate level quantities.

2.2.3 Empirical Examples

It has been reported that Goodman's regression, especially when applied to voter transition rate problems, tends to give logically impossible answers that are above 100% or below 0% (Gosnell 1942; Achen and Shively 1997, Ch. 3), or, as King (1997, p. 58) observes, outside of the bounds that can deterministically be narrowed down by the data. Usually, the estimates are known to be far from the "true" values when available, or from their corresponding survey estimates.

[3] A more recent version of this approach (King *et al.* 1999) assumes a more flexible beta distribution, but the basic concept is the same.

[4] Constructing a convolution of binomial distributions results in a different specification of the variance. In a binomial setup, the variance of the disturbance is $\sigma_j^2 = \left[D_{1j} p(1-p) + (1 - D_{1j}) q(1-q) \right] / n_j$ where n_j denotes the number of total votes in district j. Note that it is the regression residuals of D_{2j}'s that are assumed to be normally distributed in Goodman's model, while King explicitly assumes the bivariate-normality of (p_j, q_j) and tries to derive the distribution.

Worse, such ecological estimates usually come with near-perfect fits. In a word, the linear regression model has continuously failed the test of voter transition problems.

Contrary to its acclaimed success, however, neither does King's estimator work well with voter transition problems. Table 2.1 reproduces estimates of voter transition rates in 1964–1966 British Parliamentary elections which was originally used by Stokes (1969) to show the failure of ecological regression. Additional to Achen and Shively (1995, p. 125)'s re-estimation, I provide estimates from a constrained regression[5], and King's *EI* estimator.

Parameter	Panel	Ecol. Reg.	Const. Reg.	*EI*
Tory-to-Tory (p)	.87	.9488	.9036	.9047
(s.e.)	(.022)	(.0082)	(.0038)	(.0028)
Labour-to-Tory (q)	.03	-.0255	.010	.0054
(s.e.)	(.010)	(.0058)	—	(.0020)
		Aggregate Sample Size: $N = 145$		

Table 2.1: Voter Transition Estimates in British Parliamentary Elections 1964–1966, Straight-Fight Seats.
Note: Data Source: Achen and Shively (1995)

As is common with ecological regressions, the estimator produces a logically impossible estimate for the defection rate that is below zero. Also, the estimate for the loyalty rate is more than three standard errors above the panel survey estimate. Estimates from a constrained regression which sets the q parameter to be at an arbitrary minimum of .01 is also reported. King's EI estimator does not fare any better than the constrained regression, although the estimated loyalty rate is within the 95% confidence interval of the panel estimate.

Table 2.2 below is a parallel comparison of the estimators for the 1992–1997 presidential elections in South Korea, where the variables are the vote share of

[5]To estimate a constrained regression, I simply "fix" the out-of-bounds coefficients from the plain ecological regression to either .01 or .99, and run a second round of regression on the residuals.

Parameter	survey	Ecol. Reg.	Const. Reg	*EI*
DP-to-DP (p)	.8676	1.0352	.9900	.9843
(s.e.)	(.0117)	(.0022)	—	(.0015)
Others-to-DP (q)	.1984	.0641	.0777	.0858
(s.e.)	(.0138)	(.0015)	(.0014)	(.0008)
			Aggregate Sample Size: $N = 3380$	

Data Sources: Korean Social Science Data Center, *National Survey of the Fifteenth Presidential Election* (1997); Korean National Election Commission, *The Presidential Election*, `http://www.nec.go.kr`.

Table 2.2: Voter Transition Estimates in South Korean Presidential Elections, 1992–1997.

the Democratic Party—one of the major parties in Korea—in the two elections. It should be noted that the unit of aggregate observation is township, with the total N being over 3,300. As can be seen, the results show essentially a similar pattern where the aggregate estimators fail to produce reliable estimates of voter transition rates. With a larger sample size and smaller unit of observations, the estimates are even worse than those from the British elections.

Strategies such as adding aggregate covariates into the model can certainly provide help (Hanushek *et al.*, 1974). However, for the purpose of retrieving voter transition rates, the process is more complicated than merely "controlling" for demographic variables with individual data. Even assuming that the micro causal relationship is perfectly accurate, we do not know exactly what the "controlled" coefficients for voter transition rates exactly pick up from the aggregate data. In an extreme case, loyalty rates can come close to zero when control variables and combinations of them predict the dependent variable well. In other words, what the voter transition model tries to estimate is not the accurate causal relationship, but a quantity that just *describes* voters' conditional probability to remain loyal or defect from a given party. Thus, additional variables should be chosen in a way that *only* reduces aggregation bias: ecological cross-sectional regression is unbi-

ased only when the micro-level specification makes the correlation between geographical location and the parameters zero (Achen and Shively 1995, pp. 101–106). This seems to be hopelessly hard, if not impossible, with aggregated demographic information. Also, as a practical matter, not as many aggregate control variables are readily available as their micro-counterparts, in addition to the fact that their variances are usually very small compared to survey data sets. These problems make dealing with the aggregation bias more difficult.

2.3 Aggregation Bias, Non-linearity and Disaggregation Consistency

2.3.1 The Direction of Aggregation Bias in VTR Setup

The pattern shown in the previous section is quite general in ecological voter transition estimates. More specifically, when ecological regression is applied to voter transition models, the estimates are biased: loyalty rates are too high and defection rates are too low. The bias is introduced when voters with the same attributes are grouped into the same aggregate units. Although it is not straightforward to derive the functional form of the aggregation bias, it is possible to show the condition under which ecological regression estimates will be biased in the typical directions.

First, rewrite Equation (2.2) as $d_{2i} = q + \beta d_{1i} + u_i$. Further, suppose there exist individual-level data with N observations, and represent the variables into vectors, $\mathbf{d_1}$ and $\mathbf{d_2}$. Now, define a $N \times J$ indicator matrix $\mathbf{G}$ that "groups" N individual observations into J groups. Then, for example, $\mathbf{G'd_1}$ is the data matrix with sums of votes for each aggregate unit $j\,(= 1, 2, ..., J)$. Then we may write the aggregate variables as $D_1 = \mathbf{Ad_1}$and $D_2 = \mathbf{Ad_2}$ where we also define an "aggregation matrix" $\mathbf{A} = (\mathbf{G'G})^{-1}\mathbf{G'}$, which transforms individual votes into fractions

of aggregated votes for each unit j.

Rewriting the variables in the individual relationship shown above into vectors and pre-multiplying them by $\mathbf{A}$ yield

$$\mathbf{Ad_2} \quad = \quad \mathbf{q} + \mathbf{Ad_1}\boldsymbol{\beta} + \mathbf{Au}$$

which corresponds to the aggregate-level ecological regression relationship shown before, $D_{2j} = q + \beta D_{1j} + U_j$.

The aggregate slope estimate, $\hat{\beta}$, is biased when the regressor, D_{1j} , is correlated with the disturbance term. More specifically,

$$E(\hat{\beta}) - \beta \quad = \quad \frac{Cov(D_{1j},\, U_j)}{Var(D_{1j})} = \frac{Cov(\mathbf{Ad_1},\, \mathbf{Au})}{Var(\mathbf{Ad_1})} \tag{2.7}$$

If the aggregation by $\mathbf{A}$ occurs independently of the vote choices, the ecological estimates are unbiased. However, if the aggregation matrix disproportionately sorts voters with regard to d_{1i} and d_{2i}, high values of D_{1j}'s will be correlated with positive errors ($u = 1 - p$ when $d_1 = 1$ and $d_2 = 1$) while low D_{1j}'s will more likely be related to negative errors ($u = -q$ when $d_1 = 0$ and $d_2 = 0$).[6] For example, when repeating Democratic voters live closer to one another and the same holds true for repeating Republican voters, aggregation bias is introduced. Note that this is another way to say that the parameters vary and are correlated with the regressor: high D_{1j}'s are correlated with high loyalty rates and low D_{1j}'s

[6]Note, at the individual-level, there are four types of possible error values:

$$u_i \quad = \quad \begin{cases} 1-p & \text{when } d_{1i} = 1 \text{ and } d_{2i} = 1 \\ -p & \text{when } d_{1i} = 1 \text{ and } d_{2i} = 0 \\ 1-q & \text{when } d_{1i} = 0 \text{ and } d_{2i} = 1 \\ -q & \text{when } d_{1i} = 0 \text{ and } d_{2i} = 0 \end{cases}$$

with low defection rates. Quite typically, such aggregate data will be dispersed over a wide range on both the dependent and independent variables, and display a clear correlation pattern. If this is the case, the aggregate slope estimate contains a positive bias.

When the positive bias in the estimated slope exists, we may derive the direction of biases in the aggregate estimates assuming equal size of units, $\bar{d}_{1i} = \bar{D}_{1j}$ and $\bar{d}_{2i} = \bar{D}_{2j}$:

$$\begin{aligned} E(\hat{q}) - q &= (\bar{D}_{2j} - \hat{\beta}\bar{D}_{1j}) - (\bar{d}_{2i} - \beta\bar{d}_{1i}) = (\beta - \hat{\beta})\bar{D}_{1j} < 0 \\ E(\hat{p}) - p &= (\hat{\beta} + \hat{q}) - (\beta + q) = (\hat{\beta} - \beta) + (\hat{q} - q) = (\hat{\beta} - \beta)(1 - \bar{D}_{1j}) > 0 \end{aligned}$$

which complies with the typical directions of bias in aggregate ecological regressions applied to the voter transition model.

King's estimator is faced by the same problem since it is based on the same model. As King notes, the estimator does not directly deal with aggregation bias (1997, pp. 159–161) *per se*, although it is claimed that the method is more robust to such problems (King 1997, Chapter 11: "Robustness to Aggregation Bias"). Still, from the empirical results in the previous section, we also notice that King's estimator shows a certain improvement, which is due to the built-in truncation process.[7] Somewhat generally, *EI*'s truncation affects the fit in the opposite direction of the usual aggregation bias. Consider a bivariate-normal distribution of (p_j, q_j) with a high mean $\overline{p}$ and a low mean $\overline{q}$, which is the usual case in a voter transition model setup. If this is true, p_j would be more heavily truncated in large values while q_j would be more heavily truncated in small values resulting in

[7] Technically, there is no logical relationship between truncation and aggregation bias: truncation is meant to deal with the problem where the estimates are out of their theoretical bounds.

$$\overline{p} < p_{LS} \text{ and } \overline{q} > q_{LS}.$$

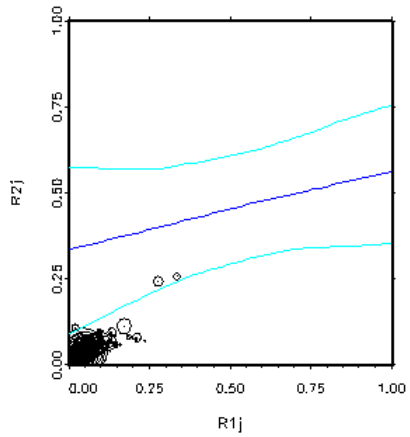


Figure 2.3: An "Overcorrection" by King MLE

In other words, King's truncation "corrects" the Goodman coefficients to the opposite directions of their biases, although the truncation is not meant to deal directly with aggregation problems. It works in a way that is quite similar to the constrained regression that suppresses out-of-bound estimates. Then, it is logically possible that King's estimates may be biased in the opposite directions of ecological regression estimates. In Figure 3 shown above, the vote shares of the Korean Democratic Party in one region—a subset of the observations used in the estimation before—are plotted against each other. The fit is poor, especially considering the intercept which is q: with heavy truncation occurring at small values of q_j's, it is "over-corrected" to the opposite direction of the bound.

To be exact, the normality assumption of the parameters does not hold. To satisfy the bivariate normality assumption, the estimation procedure critically assumes the unimodality of (p_j, q_j). In an extreme case which is shown below, where votes from two opposing parties in consecutive elections are plotted against each

other, a violation of the unimodality assumption, plus heavy truncations and a flat likelihood function, the results are at best unstable. As can be seen from the right sub-figure of Figure 2.4, a more reasonable modal point seems to be located around the upper-left corner with a negative p and high q values, which would have fit the data better. However, with a "wrong" starting value, the maximizing sequence could stop at a local maximum, which is located at the lower right hand side of the right figure as indicated. This would result in a disastrous fit that is shown in the left figure. This ML estimation can be "re-corrected" at the simulation stage, but still there is no guarantee that the final estimates yield reliable results.

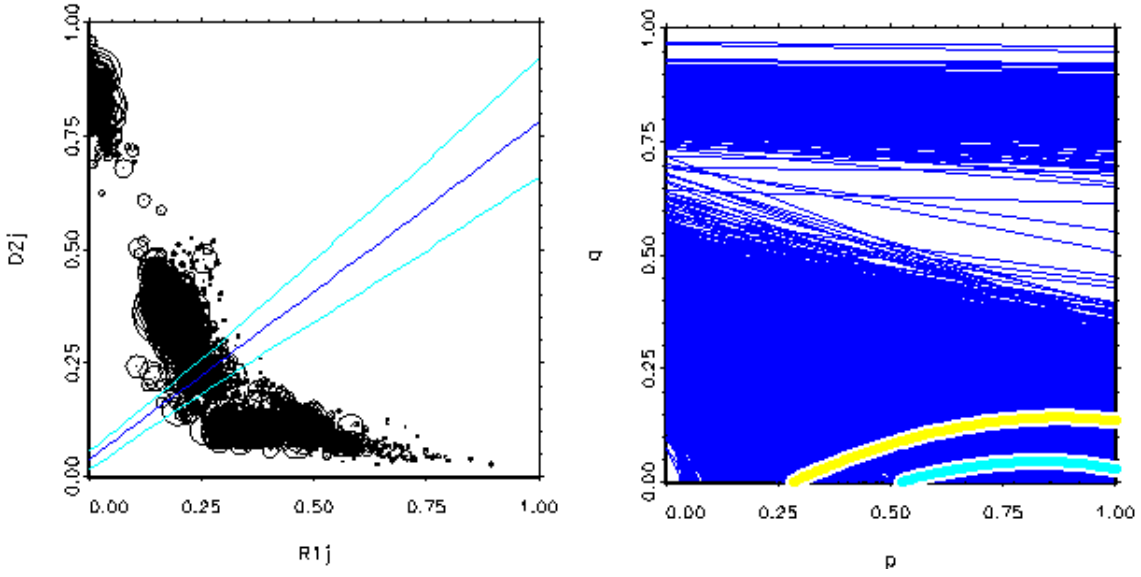


Figure 2.4: When Things Go Wrong

2.3.2 Non-Linearity Problems

Another obvious challenge for linear approaches is the fact that voter transition data usually display non-linearity when plotted, especially when the range of variables is wide. To be sure, the non-linearity problem cannot clearly be sep-

arated, at least observationally, from aggregation problems: it can be caused by aggregation (correlated parameters), or just simply by variable scales. However, the point here is that there are certain characteristic forces of electoral choices that make the non-linearity problem more severe in VTR setups than in usual aggregate data sets.

A uniform swing, a commonly observed feature of electoral systems, will usually create non-linearity in the data, since the variables are bound between zero and one. Another common cause of non-linearity is multi-partism. In a "near-two-party" system, where a minor party performs either better or worse in the second election than it did in the first, we would usually observe a significant change in the middle range, of which districts are not heavily loyal to either of the major parties.[8] In addition, minor parties may choose to field candidates mainly in such districts which will create non-linearity.

Figure 2.5 below is a plot of vote returns for a presidential candidate from the Grand National Party – another major South Korean party — which lost quite a few supporters to a minor-party candidate. The shift of the voters around the middle is quite obvious, making the non-linearity apparent, and as can be seen, the superimposed nonlinear fit[9] is better than the linear fit.

Such non-linearity problems are aggravated by the fact that variables of aggregate party support typically display large variance over a wide range. Under certain settings where the non-linearity problem is less severe, linear ecological estimators might be able to produce reliable answers, but the usage would have to be justified before applying the technique by checking if the data and substantive

[8] Of course, an exact opposite effect is also possible in congressional elections due to strategic voting, since minor parties are less likely to gain votes in competitive districts, which are clustered in the middle along the two axes. The point is that VTR models are prone to systematic non-linearities.

[9] The curve represents a linear fit of the two probit-transformed variables, which is implied by the Thomsen model that will be discussed in the next section.

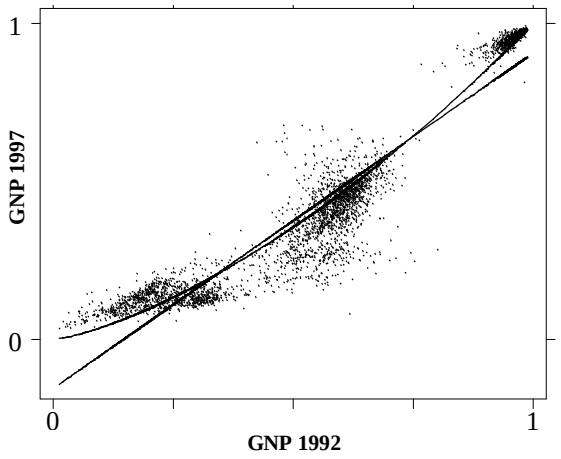


Figure 2.5: Non-Linearity in VTR Models

knowledge all suggest that the assumptions of the techniques will hold.[10] But if it is true that non-linearity is apparent in the data (which is usually the case for VTR models) with aforementioned problems, there is little reason to continue using linear models for voter transition problems.[11]

2.3.3 Do "Better" Data Always Help?: Disaggregation Consistency

The problems described above also imply that better data, or fine-grained information, do not necessarily produce better answers for linear approaches. Generally, we may consider "lower-level" aggregate data with smaller or less populous units of observations better for the following reasons: first, the aggregation problem may be less severe in smaller units, and second, more observations and larger variance will guarantee better precision of the estimates. In the extreme case where the the aggregate unit is a single person, we know that the aggregation bias will disappear, since there is no aggregation. Although different from the notion of consistency in the usual econometrics context, we may call this de-

[10] Certain demographic variables, turnout rates, or even vote fraction within small geographic regions may qualify. However, note that such variables usually tend to have small variances with high inflation factors (Palmquist 1993).

[11] Achen and Shively (1995, pp. 133-142.) prove that adding a quadratic term always reduces the bias in Goodman regression.

sirable property of ecological estimators as "disaggregation consistency," in the sense that we expect better answers when we have "better" data, or finer-grained data.

Precinct electoral returns make better data than county-, district-, or state-level records: but will linear models produce better answers? The discussion in the previous sections leads us to believe that the conventional way of thinking aggregate data does not always hold for linear VTR models.

First, note that the aggregation bias shown in Equation (2.7), $\frac{Cov(\mathbf{Ad_1}, \mathbf{Au})}{Var(\mathbf{Ad_1})}$ will be reduced in fine-grained data sets. More specifically, the denominator, the variance of the independent variable, will increase as we go down the ladder of aggregation with smaller aggregate unit sizes.[12] However, the numerator, the covariance between the independent variable and the aggregate error term, will not increase as fast as the denominator. Thus, the aggregation bias defined in such fashion will be reduced in "better" data sets, although it will still be present.[13]

However, recall the fact that non-linearity in the data will become more apparent when the variance and range of the variables increase. In other words, the bias introduced by non-linearity may become greater for linear models in "better" data sets. When the underlying true relationship is non-linear, test statistics or diagnostic measures are meaningless. The estimates will be biased with smaller standard errors; statistics such as Palmquist (1993)'s inflation factor[14] may work

[12]Define two aggregation matrices $\mathbf{A}_{P(recint)}$, and $\mathbf{A}_{D(istrict)}$. It is straightforward to show that $Var(\mathbf{A}_P\mathbf{d}) > Var(\mathbf{A}_D\mathbf{d})$, since the observations will regress to the mean as the aggregate unit becomes larger.

[13]Proof:

Define the variables as $V_P = Var(\mathbf{A}_P\mathbf{d})$, $V_D = Var(\mathbf{A}_D\mathbf{d})$, $C_P = Cov(\mathbf{A}_P\mathbf{d}, \mathbf{Au})$ and $C_D = Cov(\mathbf{A}_D\mathbf{d}, \mathbf{Au})$. Then, from above, $V_P > V_D$. Also, $V_P > C_P$, $V_D > C_D$, and $C_P > C_D$ will generally be true. The question is the relative size of aggregation biases in the two data sets.

Since the difference in aggregation biases of district data and precinct data are

$$\frac{C_D}{V_D} - \frac{C_P}{V_P} = \frac{C_D V_P - C_P V_D}{V_D V_P} > \frac{C_D V_P - C_P V_P}{V_D V_P} = \frac{V_P(C_D - C_P)}{V_D V_P} > 0$$

the precinct level aggregation bias will generally be smaller than the district-level bias. Q.E.D.

[14]The inflation factor can be defined as the inverse of the between-versus-total variance ratio, and is the ratio of how much the bias is inflated. Essentially, the measure suggests that data with smaller aggregate units will produce better

to the opposite direction.

The observations above suggest the following: first, linear approaches to the VTR model are prone to bias due to problems of aggregation and non-linearity. Second, although it is hard to remedy the aggregation problem, ecological estimators would benefit from modeling the linkage between individual and aggregate relationship. Third, when non-linearity is apparent in the data, there is little reason to continue using linear models. Especially, even when "better" data are available, linear models will not necessarily produce better answers, while non-linear models may take advantage of them.

Non-linear approaches, of course, come with a price: identification is hard to achieve, estimated parameters cannot be straightforwardly interpreted, and the results may be unstable depending on the assumptions. Among the non-linear approaches, a method suggested by Thomsen (1987) has been successful in a number of applications to various countries. The method is interesting in the sense that it tries to directly model a nonlinear specification into the voter transition model at the individual-level. The next section describes the model, examines its assumptions, and provides reasoning for the frequent success of the estimator.

2.4 Thomsen's Nonlinear Model

2.4.1 Modeling Partisanship

Going back to the voter transition model, Thomsen argues the regression approach that treats two consecutive elections asymmetrically (that is, imposing a causal relationship) is prone to problems such as misspecification or multicollinearity. Instead, he suggests treating them symmetrically, the two election outcomes being a result of a common latent factor (Thomsen 1987, pp. 46–45). Whether we

estimates.

call this unobservable variable true partisanship or policy position, once we establish the relationship between this variable and observed election outcomes, we would be able to model the relationship between the two elections.

Denote d_i^* as a latent long-term partisanship or policy position that determines i's vote for two consecutive elections, d_{1i} and d_{2i}. Define $\Phi(\cdot)$ as the cumulative distribution function of the standard normal distribution[15]. At its simplest form, we may write within unit j, i's probability to vote Democratic is

$$Prob(d_{tij} = 1) = \Phi(\alpha_t + \beta_t d_{ij}^* + e_{tj}) \tag{2.8}$$

by assuming the following: first, in a given election, voter i's underlying partisanship affects her probability to vote for a given party, probit-linearly. The linear term — the inverse of the equation — can be interpreted as the utility difference from voting for a given party instead of the alternative: if the utility is larger than zero, one would vote for party D, otherwise, one would vote for another. Factors not correlated to partisanship are assumed to be random in their effects on votes, which is the simplest possible setup and at the same time a fair approximation. Second, the coefficients for the partisanship factor, α_t's and β_t's, are different for each election, although they are assumed constant across voters and thus districts. Equation (2.8) implies that the inverse-probit transformed probabilities, or the utilities to vote for a given party in the two elections, are linearly related to each other.

The aggregate outcome D_{tj} we observe from a given district j, should correspond to the expected vote fractions in the district using equation (2.8). Assum-

[15] In actual estimation, Thomsen's *ECOL* implements logit specification and inversed tetrachoric correlation, which results in slightly different estimates from the steps shown in this chapter. A probit specification along with a direct integration is used for a more intuitive interpretation, and a corresponding algorithm is written and used for this chapter. Stata and Matlab routines that produced the results are available from `http://www.umich.edu/~wpark/progs`.

ing with Thomsen that the underlying dimension d^*_{ij} is normally distributed with mean D^*_j and a constant variance σ^2, find the average of both sides of equation (2.8):

$$E(d_{tij}) = D_{tj} = \int_{-\infty}^{\infty} \Phi(\alpha_t + \beta_t d^*_{ij} + e_{tij}) \phi(d^*_{ij} | D^*_j, \sigma^2) \partial d^*_{ij} \tag{2.9}$$

By carrying out the integral, we obtain (Achen and Shively 1995, p. 184.):

$$E(d_{tij}) = D_{tj} = \Phi \left(\frac{\alpha_t + \beta_t D^*_j}{\sqrt{1 + \beta_t^2 \sigma^2}} \right) \tag{2.10}$$

Since we established that the inverse-probit of D_{1j} and D_{2j} are normally distributed and linearly related to each other, they are bivariate normally distributed, with correlation ρ_{agg}. For identification, Thomsen substitutes this aggregate correlation for its individual counter-part: the correlation between the individual utilities to vote for a given party in two elections as seen in equation (2.8).[16] This does not identify all the parameters but is sufficient to estimate the voter transition rates. Recall that what the voter transition model is intended for is *not* α_t's and β_t's that concern the relationship between the unobservable variable and electoral outcome, but to infer the transition rates between the two electoral returns, D_{1j} and D_{2j}.

Since the individual utilities specified in equation (2.8) , $\Phi^{-1}\left[Pr(d_{1ij} = 1)\right]$ and $\Phi^{-1}\left[Pr(d_{2ij} = 1)\right]$, are bivariate normally distributed with known means—that is, the probit transformed national fraction of the party—and an estimated correlation, a double integration of the bivariate-normal function would yield the joint probability to vote for a given party two consecutive times. More specifically, denote x_1 and x_2 as random variables for the probit transformed values of

[16] This identification condition is discussed later in more detail.

vote fractions of each district, r as their correlation, and μ_1 and μ_2 as the probit-transformed national vote fractions. Then

$$p_{DD} = \int_0^\infty \int_0^\infty \phi(\mu_1, \mu_2; r) dx_2 dx_1 \tag{2.11}$$

would yield the estimated fraction of the voters in the entire electorate. So, dividing this quantity with the marginal proportion of voters from the first election would yield the estimate for the loyalty rate. The defection rate can be derived in a similar fashion, where the fraction of the voters who voted non-Democratic in the first election and converted Democratic can be expressed as:

$$p_{RD} = \int_{-\infty}^0 \int_0^\infty \phi(\mu_1, \mu_2; r) dx_2 dx_1 \tag{2.12}$$

2.4.2 Bias in the Aggregate Correlation

Thomsen's estimator achieves identification by equating individual (utility) correlation with aggregate (transformed vote fractions) correlation. To see when this postulate works, let us formulate the implied correlations at individual and aggregate levels and derive the bias.

First, decompose the variance structure of individual utilities. There can be two sources of variances: the dimensional and the non-dimensional that pertain to the latent partisanship dimension. For the entire population, the variation occurs both within and across the observed units, while the aggregate data would only reveal the across-unit variances in ecological data. Thus, the total individual variance can be decomposed into:

The individual correlation would include all four of the variances, while the

ecological estimation of the correlation would only involve the across-unit variances, the second column. To find the individual correlation, invert equation (2.8), and define the left-hand-side variable as voter i's utility (difference) to vote for a given party over the other party. Denote d_{ij}^* as the underlying individual partisanship for i, D_j^* as its mean, u_{ij} as d_{ij}^*'s deviation from D_j^*. Then we may write within district j,

$$\begin{aligned} U_{tij} &= \Phi^{-1}\left[\Pr(d_{tij}=1)\right] = \alpha_t + \beta_t d_j^* + e_{tij} \\ &\text{where} \quad Var(d_{ij}^*) = Var(D_j^* + u_{ij}) = 0 + \sigma^2 \text{ and } Var(e_{tij}) = 1 \end{aligned}$$

The within-district correlation can then be written as

$$\rho_{within}(U_{1ij}, U_{2ij}) = \frac{Cov(U_{1ij}, U_{2ij})}{\sqrt{Var(U_{1ij})}\sqrt{Var(U_{2ij})}} = \frac{\beta_1\beta_2\sigma^2}{\sqrt{\beta_1\sigma^2+1}\sqrt{\beta_2\sigma^2+1}} \quad (2.13)$$

For the entire nation, variances include the second column, the across-unit variances:

$$\begin{aligned} U_{tij} &= \alpha_t + \beta_t d_j^* + \xi_{tij} \\ &\text{where} \quad Var(d_{ij}^*) = Var(D_j^* + u_{ij}) = \omega^2 + \sigma^2 \text{ and } Var(\xi_{tij}) = \tau_t^2 + 1 \end{aligned}$$

Then the individual correlation for the entire nation is

Variances	Within-Unit	Across-Unit
Dimensional	σ^2	ω^2
Non-Dimensional	1	τ_t^2

Figure 2.6: Decomposition of Variances

$$\begin{aligned}\rho_{ind}(U_{1ij}, U_{2ij}) &= \frac{Cov(U_{1ij}, U_{2ij})}{\sqrt{Var(U_{1ij})}\sqrt{Var(U_{2ij})}} \\ &= \frac{\beta_1\beta_2(\omega^2+\sigma^2)}{\sqrt{\beta_1(\omega^2+\sigma^2)+(\tau_1^2+1)}\sqrt{\beta_2(\omega^2+\sigma^2)+(\tau_2^2+1)}} \quad (2.14)\end{aligned}$$

However, (2.14) cannot be computed directly since we do not observe individual utilities. Thus, we have to estimate it using the ecological observations. To derive the ecological correlation, invert equation (2.10) and write

$$\begin{aligned}\Phi^{-1}(D_{tj}) &= \alpha'_t + \beta'_t D^*_{tj} + \varepsilon'_{tj} \\ \text{where} \quad \alpha'_t &= \frac{\alpha_t}{\sqrt{1+\beta_t^2\sigma^2}},\ \beta'_t = \frac{\beta_t}{\sqrt{1+\beta_t^2\sigma^2}} \text{ and } \varepsilon'_{tj} = \frac{\varepsilon_{tj}}{\sqrt{1+\beta_t^2\sigma^2}}\end{aligned}$$

with across-unit variances only: $Var(D^*_j) = \omega^2$ and $Var(\varepsilon_{tij}) = \tau_t^2$.

Then the aggregate correlation we obtain is

$$\begin{aligned}\rho_{agg}\left[\Phi^{-1}(D_{1j}), \Phi^{-1}(D_{2j})\right] &= \frac{Cov\left[\Phi^{-1}(D_{1j}), \Phi^{-1}(D_{2j})\right]}{\sqrt{Var\left[\Phi^{-1}(D_{1j})\right]}\sqrt{Var\left[\Phi^{-1}(D_{2j})\right]}} \\ &= \frac{\beta_1\beta_2\omega^2}{\sqrt{\beta_1\omega^2+\tau_1^2}\sqrt{\beta_2\omega^2+\tau_2^2}} \quad (2.15)\end{aligned}$$

Now the task is to find under what conditions (2.15) approximates (2.14). By assuming $\tau_1 = \tau_2$, the difference between the squared aggregate and individual correlation can be factored as

$$\begin{aligned}\rho^2_{agg} - \rho^2_{ind} &= [\text{Equation}(2.15)]^2 - [\text{Equation}(2.14)]^2 \\ &= G\left(\omega^2 - \tau^2\sigma^2\right) \quad (2.16) \\ \text{where} \quad G &= \frac{\beta_1^2\beta_2^2\omega^4\left[\omega^2\left(\omega^2+\sigma^2\right)(\beta_1+\beta_2)+\tau^2\sigma^2+2\omega^2\tau^2+\omega^2\right]}{(\beta_1\sigma^2+1)(\beta_2\sigma^2+1)[\beta_1(\omega^2+\sigma^2)+(\tau^2+1)][\beta_2(\omega^2+\sigma^2)+(\tau^2+1)]}\end{aligned}$$

Equation (2.16) implies that the aggregate correlation equals the individual correlation when the ratios of the variances shown in Figure 2.6 are constant — $\sigma^2 : \omega^2 = 1 : \tau^2$. In other words, if the proportion of the aggregate variance to total variance is the same in both systematic and non-systematic variances, the ecological correlation is an unbiased estimate of the individual correlation, and thus, the Thomsen estimator is identified (Achen 2000).

$$\text{Identification Condition:} \quad \frac{\omega^2}{\omega^2 + \sigma^2} = \frac{\tau_t^2}{\tau_t^2 + 1} \tag{2.17}$$

A couple of substantive interpretations are implied by the identification condition. First, without spatial heterogeneity, the ecological correlation is unbiased. This is equivalent to Thomsen's original "homogeneous district assumption" that districts are (almost) perfectly similar dimensionally. When there are no cross-district variances, $\tau_t^2 = \omega^2 = 0$, and the condition above is trivially satisfied. Moreover, in such cases, the expression G in (2.16) becomes zero, eliminating the bias. However, the assumption is still troublesome in the sense that it assumes away any cross-unit variances, which is, in fact, assuming away of the aggregation problem. In this case, equation (2.15) can take any value between zero and 1 based on the sampling error. The only possible implication of Thomsen's district homogeneity assumption is that the total individual correlation (2.15) is equivalent to estimating a within-district correlation (2.13), the entire nation being one large district. Then, for (2.15) to approximate this quantity, our observations should be something like a randomly sampled individual probability to vote for a party, rather then the expectation of it in each district.

Second, with spatial heterogeneity, the aggregation bias of the ecological correlation can be reduced by fine-grained data. Note that the quantities in (2.17) will increase and get closer to unity as the between-unit variances increase and the

within-unit variances decrease. Instead of assuming away the aggregation problem, we may allow that voters of similar preferences are more likely to live in the same unit, which is a more realistic assumption, especially in the case of a voter transition setup. In such cases, the smaller the unit size is, ω^2 and τ_t^2 will become larger relative to their counter parts, making the quantities in (2.17) closer to unity, and satisfying the identification condition. Additionally, a smaller aggregate unit size will more likely satisfy the constant within-unit variance assumption. In other words, the estimator is disaggregation-consistent.

2.5 Empirical Results

British Parliamentary Elections 1964–1966

Table 2.3 reproduces Table 2.1 that was shown before, with the Thomsen estimates added to them. As can be seen, the Thomsen estimates are significantly closer to the survey estimates than any other estimates, both the loyalty and defection parameters well within the survey sampling error. Still, we cannot quite decisively say the performance is better than other estimators since the 95% confidence interval of the survey estimates includes some other estimates as well. Moreover, it is subject to questions whether the constant variance assumption of Thomsen's estimator is satisfied in the England data—which means that Scottish voters are assumed to have a similar preference structure to London voters. One possible way to solve the problem is to partition the data set into provinces or regions, where presumably, the voters can be seen as more homogeneous. However, because of the limited number of observations, the strategy is not feasible here.

Parameter	Survey	Ecol. Reg.	Const. Reg.	*EI*	Thomsen
Tory-to-Tory (p)	.87	.9488	.9036	.9047	.8861
(s.e.)	(.022)	(.0082)	(.0038)	(.0028)	(.0084)
Labour-to-Tory (q)	.03	-.0255	.010	.0054	.0185
(s.e.)	(.010)	(.0058)	—	(.0020)	(.0059)
			Aggregate Sample Size: $N = 145$		

Table 2.3: Voter Transition Estimates in British Parliamentary Elections 1964–1966, Straight-Fight Seats.

South Korean Presidential Elections 1992–1997

Table 2.4 reports the results from the Korean election data, but this time partitioning the dataset into eight regional areas to satisfy the constant variance assumption[17] of the Thomsen estimator. The figures report the average of regional estimates weighted by their respective population size. The pattern is essentially the same: large estimated loyalty rates and small defection rates from all the estimators. However, the Thomsen estimates outperform other estimators by a sizable margin. All the other estimates are missing the survey estimates by at least 10%, which is large inaccuracy considering the standard errors. The estimated defection rate of the Thomsen estimator is somewhat off the target, but a significant improvement from the other estimators.

Parameter	Survey	Ecol. Reg.	Const. Reg.	*EI*	Thomsen
DP-to-DP (p)	.8676	.9651	.9648	.9608	.8788
(s.e.)	(.0117)	(.0111)	(.0103)	(.0127)	(.0126)
Others-to-DP (q)	.1984	.0984	.0986	.0992	.1328
(s.e.)	(.0138)	(.0061)	(.0058)	(.0055)	(.0071)
			Aggregate Sample Size: $N = 3380$		

Table 2.4: Voter Transition Estimates in South Korean Presidential Elections, 1992–1997

It still requires caution to state that the assumptions for the Thomsen estimator are satisfied in the estimation. The assumptions for the estimator to work mostly concern the variance structure of the voter preference, and are extremely hard

[17] Predictably, the Thomsen estimates using the non-partitioned dataset did not fare well with estimates, $p = .95$ and $q = .09$.

to verify. More specifically, because the assumptions pertain to individual-level properties (which we do not observe in ecological data) and as well as to something that measures the latent utility of supporting a party, checking the validity of these assumptions will not be an easy task. Perhaps external information such as survey studies or prior knowledge, could be put to the task. However, seeing that the estimator gives reliable results in the table above, certain hypothetical explanations can be provided.

First, the constant within-unit dimensional variance (σ^2) assumption does not hold without partitioning the dataset in the Korean case. Arguably, the partisan variance in a small rural township or district is far smaller than in any of the Seoul districts. Within a region, σ_j^2's would become relatively similar to each other, at least more similar than those from out-of-region districts, making the data fit Thomsen's assumption. Also, this assumption is more likely to be satisfied when the unit of analysis is smaller. The results support this explanation.

Second, although the Korean case does not meet Thomsen's district homogeneity assumption, it might satisfy a weaker, or a generalized version of it, that was provided in the previous section. More specifically, the model can be identified when the cross-district variances overwhelm the within-district variances, which does fit the profiles of Korean voters who can be characterized as homogeneous within their "neighborhoods" and more heterogeneous across the borders.

South Korean Presidential Elections 1992–1997: Disaggregation Consistency

Now I provide comparable results using data sets from different levels of aggregation to examine whether the performance of ecological estimators change when "better" and "worse" data sets are used alternatively. Additional to the township-level data set that was used for the estimations shown above, I use

a more disaggregated data set recorded at the precinct-level and a more aggregated district-level data set. The number of observations for each level of aggregation can be found in the last row of Table 2.5 below. For example, since the total number of valid votes was roughly around twenty million, each precinct includes roughly two thousand voters.

	Survey	Ecological Regression			EI			Thomsen		
		Prcnt	Tnshp	Dist	Prcnt	Tnshp	Dist	Prcnt	Tnshp	Dist
p	.8676	.9454	.9651	.9840	.9812	.9608	.9813	.8609	.8788	.9384
(s.e.)	(.0117)	(.0081)	(.0111)	(.0363)	(.0071)	(.0127)	(.0280)	(.0094)	(.0126)	(.0343)
q	.1984	.1209	.0984	.1009	.1174	.0992	.1041	.1777	.1328	.1125
(s.e.)	(.0138)	(.0041)	(.0061)	(.0239)	(.0036)	(.0055)	(.0145)	(.0048)	(.0071)	(.0178)
N	840	12,016	3,380	302	12,016	3,380	302	12,005	3,380	302

Table 2.5: Comparison of Ecological Estimates at Different Levels of Aggregation, South Korean Presidential Elections, 1992–1997

As is expected, it is only the Thomsen estimates that improve when applied to increasingly less aggregated data sets. Both estimated parameters, especially the estimated defection rate, improve significantly moving from district-level to precinct level. The estimator is disaggregation-consistent in the sense that it is taking advantage of the additional information in less aggregated data sets. Other estimators do not show such disaggregation consistency. Aside from the fact that the estimates are significantly biased across the board, not all the estimates improve while moving from higher- to lower-levels of aggregation. For example, the loyalty rate estimated by EI from the precinct-level data is more biased than that from the township data, with the usual decrease in the standard error.

2.6 Conclusion

The ecological inference problem is in a sense a problem with deficient data. Using external information and/or reliable assumptions, one would have to model

the data generating process that underlies what we actually observe. Cross-level inference is only possible when such cross-level assumptions are correct. The difficulty with ecological problems, then, is to formulate reasonable and justifiable assumptions. The task of ecological inference should focus more on careful examination and substantive rumination of the model and the data.

Although the structure of the voter transition model is similar to other ecological inference models, problem-specific difficulties such as severe aggregation bias and systematic non-linearity make it hard for linear ecological inference strategies to produce reliable estimates. Meanwhile, a more specialized approach, the Thomsen estimator, which is based on micro-modeling of voter utility and party choice, produces more accurate and consistent estimates.

Perhaps comparing the Thomsen estimator with other ecological estimators in the first place might not be fair, since it is specifically designed to estimate voter transition rates, while the ecological regression and the King model are designed for more generic or sociological problems. For example, race or gender is just observed, not governed by any underlying dimension: thus, the Thomsen model cannot directly be applied to problems with such variables without substantial modification. However, the success of the Thomsen model in this chapter stresses the point that there is no general solution to the ecological inference problem but only data- and problem-specific solutions.

Although ecological inference is prone to a number of problems, it can be "solved" by constructing reasonable micro-relationships and assumptions about the data. When such assumptions plausibly represent the real data-generating process, ecological estimators would produce reliable answers. The Thomsen estimator shows a good example where such agreement between theory and reality

results in a successful ecological inference process.

From a substantive point of view, the voter transition model I have discussed so far is perhaps the simplest possible way to look at the movement of voters across two elections. To deal with more complex and sophisticated research questions, an apparent next step is to extend the simple ecological inference strategies to represent more realistic and substantively interesting relationships. The following chapters take up several of these tasks.

CHAPTER III

Ecological Inference in Multiparty Systems

3.1 Introduction

In the previous chapter, I discussed the overall setup of the voter transition model and how it can be approached from the viewpoint of aggregate data analysis. I demonstrated that traditional ecological inference techniques have usually failed while a model implied by Thomsen (1987) recovers the quantity of interest—the voter transition rates—reasonably well from aggregate data. Nevertheless, the model developed so far is restrictive in the following senses: first, it will not allow the researcher to study elections with more than two parties or candidates—or equivalently, force the researcher to simplify the electoral contest into two opposing parties; second, the model does not allow the researcher to incorporate information of the covariates into the model.

This chapter takes up the question on how to extend the ecological inference techniques to estimate voter transition rates in multiparty systems. First, I will define the problem at hand, and look into the existing techniques for multinomial category problems. I will also highlight some of the difficulties of suggested techniques and provide a set of possible solutions that can be implemented in practice. More specifically, I focus on the interative proportional fitting (IPF) process to ad-

just crude initial estimates using the information from aggregate marginals.

Second, the IPF process is then examined closely in a bootstrap simulation using ballot image data from the 2000 US presidential election. Most importantly, I compare the mean squared errors of the IPF estimates against Thomsen's multiparty extension. The simulation will provide a good opportunity to examine a complex computation involving the IPF adjustment.

Finally, I will empirically test the suggested techniques applying them to estimating the voter transition in South Korean presidential elections between 1992 and 1997. It will provide a useful foundation to test and examine different classes of ecological estimators in a multiparty setup.

3.2 Voter Transition Rates in Multiparty Systems: Current Methods

An immediate problem with the simple bivariate model in the voter transition context is that there usually are more than two parties. Many electoral systems in the world consist of multiple parties, and even two-party systems will pose a problem if we count non-voters ("party of abstainers") as a separate category. The usual practice has been either to drop non-significant categories, or collapse similar parties into bivariate categories, effectively reducing the task into estimating a 2×2 table that was shown in the previous chapter. But more often than not, researchers are forced to model the multidimensional relationship into the model. In this section, I critically review existing methods that extend ecological inference techniques into the multivariate context.

As it turns out, the Goodman ecological regression has a natural multivariate extension. One would just write a system of equations with all the possible time-2 party shares explained by all the time-1 party supports as independent variables.

Each equation can be estimated separately or estimated by seemingly unrelated regression (SUR) to gain full efficiency. For example, Zellner (1962) proposed a general form of an estimator for multivariate regression equations based on a feasible generalized least square technique. In any case, within the context of the voter transition setup, SUR coefficient estimates are the same as those from linear regression—that is, the Goodman estimates—in expectation, and the model only corrects the standard errors. For an SUR approach to explain multiparty election result with non-electoral covariates, see Jackson (2002).

For the particular purpose of retrieving individual-level parameter from aggregate data in the voter transition setup, there is no reason to believe a multivariate approach will come through where bivariate estimation falls short. Most importantly, the usual aggregation bias will remain a problem, and at the same time nonlinear relationships involving small parties will present serious challenges to the linear specification. More likely than not, the estimation will still yield out-of-bounds estimates on certain sets of coefficients, as will be demonstrated in a later section.

3.2.1 Multivariate Extension of the Constrained Regression

In a way, the constrained regression technique provides some insight into the matter on how to cope with the out-of-bounds estimates from a Goodman setup in a multiparty context. A constrained regression would be an estimation process where one would "fix" the first-round estimates that are greater than 1 or smaller than zero. In a bivariate relationship, the estimation is straightforward: one would just set the out-of-bounds coefficient to a fixed value (say, the loyalty rate to .99 if it initially exceeds unity) and estimate the remaining coefficient (say, the defection rate). However, how to implement this in a multiparty context is not

obvious, since fixing one coefficient in one equation will have an impact on *all* the coefficients in the entire system of equations. Here I propose a simple solution to the problem.

I shall start with a simple bivariate setup and provide the extension to a multi-party case. Suppose a simple two party case where we estimate

$$D_{2j} = pD_{1j} + qR_{1j} \tag{3.1}$$

$$R_{2j} = (1-p)D_{1j} + (1-q)R_{1j} \tag{3.2}$$

The second equation is implied by the first, so in practice, the second equation is not estimated separately. Now suppose that the least square estimate of p is larger than unity. Since the estimate exceeds the logically possible value of 1, we may choose to constrain the estimate to be under 1. If we fix the coefficient to be an arbitrarily large value of .99, this will have an impact on the remaining coefficient.

The new constrained coefficient can be retrieved by generating a new variable $D_{2j} - .99D_{1j}$ and regress it on R_{1j} without the constant term, that is

$$D_{2j} - .99D_{1j} = q'R_{1j} + U_j \tag{3.3}$$

where $U_j = (p - .99)D_{1j} + (q - q')R_{1j}$. Following the Goodman assumption from Chapter 2 that the parameters and regressor are not correlated, $Cov(R_{1j}, U_j) = 0$, then q' is an unbiased estimator of the constrained defection rate. The resulting pair, $[p, q] = [.99, q']$ would be the constrained regression estimates.

The key point here is how to work with the second equation to adjust to the impact of the constraint. Adding the residuals U_j to the second equation guarantees the coefficients add up to 1 since

$$\begin{aligned} R_{2j} + U_j &= (1-p)D_{1j} + (1-q)R_{1j} + \left[(p-.99)D_{1j} + (q-q')R_{1j}\right] \\ &= (1-.99)D_{1j} + (1-q')R_{1j} \end{aligned} \tag{3.4}$$

This is equivalent to transferring the impact of the constraint to the other equation and adjusting the unconstrained estimates.

In a bivariate setup, such adjustment is trivial in the sense that equation (3.3) is implied by (3.4). However, in a multiparty case, the approach above can be useful. Consider a case where three parties compete against one another in two consecutive elections. The simple Goodman setup would be

$$\begin{aligned} D_{2j} &= p_{DD}D_{1j} + p_{RD}R_{1j} + p_{SD}S_{1j} \\ R_{2j} &= p_{DR}D_{1j} + p_{RR}R_{1j} + p_{SR}S_{1j} \\ S_{2j} &= p_{DS}D_{1j} + p_{RS}R_{1j} + p_{SS}S_{1j} \end{aligned}$$

Now suppose the least squares estimate of p_{DD} is larger than 1 and we want to constrain it to be 99%. We would generate a new dependent variable $D'_{2j} = D_{2j} - .99R_{1j}$ and regress it on R_{1j} and S_{1j} without a constant term:

$$(D_2 - .99D_1) = p'_{RD}R_{1j} + p'_{SD}S_{1j} + U_j$$

where

$$U_j = (p_{DD} - .99)D_{1j} + (p_{RD} - p'_{RD})R_{1j} + (p_{SD} - p'_{SD})S_{1j}$$

Dividing the residuals generated from the first equation and adding them to the remaining two equations guarantees the coefficients to add up to unity. Any

possible allocation of the residuals to the remaining equations, $\left[rU_j, (1-r)U_j\right]$, where r is a fraction, and adding them up to the remaining dependent variables, will work for the purpose. Then the essential question is to determine the ratio r.

A theoretically consistent value of r should be the ratio of cross-equation correlations. In a two party system, $R_{2j} = (1 - D_{1j})$, thus, with the correlation being minus one, $r = (-1)/(-1) = 1$. In the example of three party system shown above, the division ration r could be written as

$$r = \frac{\rho_{D,R}}{|\rho_{D,R} + \rho_{D,S}|}$$

where ρ indicates the cross-equation correlations between the dependent variables. This will make the new dependent variables in the second round as $R_{2j} - \frac{\rho_{D,R}}{|\rho_{D,R}+\rho_{D,S}|}U_j$ and $S_{2j} - \frac{\rho_{D,S}}{|\rho_{D,R}+\rho_{D,S}|}U_j$ respectively.

This will also guarantee the coefficients from the new round of estimation add up to unity. In other words, the effect of fixing a coefficient in one equation is transferred to the remaining equations proportional to their correlation to the fixed equation. It still is an *ad hoc* fix, since when the two correlations take different signs and their absolute values are similar, the denominator for r could become arbitrarily small.

Alternatively, using covariances instead of correlations could be a reasonable alternative, where r in fact corresponds to (the negative of) regression coefficients between dependent variables. For example,

$$r = \frac{Cov(D_{2j}, R_{2j})}{\left|Cov(D_{2j}, R_{2j}) + Cov(D_{2j}, S_{2j})\right|} = \frac{-Cov(D_{2j}, R_{2j})}{Var(D_{2j})}$$

$$1 - r = \frac{Cov(D_{2j}, S_{2j})}{\left|Cov(D_{2j}, R_{2j}) + Cov(D_{2j}, S_{2j})\right|} = \frac{-Cov(D_{2j}, S_{2j})}{Var(D_{2j})}.$$

Extensions to electoral systems with more than three parties are straightforward. This formula is used in the next section to retrieve results using constrained regressions and is compared with other ecological estimators.

3.2.2 The King Estimator in A Multiparty Setup: Imputation

Extending King's method to a multiparty setup is not simple. Generally, when there are k parties in the first election and l in the second, the number of parameters to estimate is $k(l-1)$. Thus, in a two-party system, the number of parameters is just 2 (p and q), as in the running example, while in a three-party system, the number of parameters increases to 6. While Goodman's regression can simply be extended by adding independent variables on more rounds of equations, King's method does not allow a direct estimation on such setups. However, he provides an algorithm that computes a 3 by 2 table, where a third party has newly emerged in the second election. Conceptually, since the marginal of a 2 by 2 sub-table excluding the third party is not known, this involves an imputation process.

More specifically, consider the following table:

		Election 1	
		D_{1j}	R_{1j}
Election 2			
	D_{2j}	p_{DD}	p_{RD}
	R_{2j}	p_{DR}	p_{RR}
	S_{2j}	$1-p_{DD}-p_{DR}$	$1-p_{RD}-p_{RR}$

Figure 3.1: Coefficients in a Three Party System: King's Approach

Collapse the two major parties (M) versus the "small (S)" party and write

$$M_{2j} = p_{DM}D_{1j} + p_{RM}R_{1j} \tag{3.5}$$

Obviously, $p_{DM} = p_{DD} + p_{DR}$ and $p_{RM} = p_{RD} + p_{RR}$. They can be estimated by

constructing the new variable and applying the bivariate method to the observations. Now in the next step, estimate the four coefficients for the major parties in the upper side of the figure above. To do so, since we do not know the marginals excluding the small party, the marginals have to be imputed. Thus, we may write

$$D_{2j} = \frac{p_{DD}}{p_{DM}}\hat{D}_{1j} + \frac{p_{RD}}{p_{RM}}\hat{R}_{1j};\ R_{2j} = \frac{p_{DR}}{p_{DM}}\hat{D}_{1j} + \frac{p_{RR}}{p_{RM}}\hat{R}_{1j} \tag{3.6}$$

where $\hat{D}_{1j}$ and $\hat{R}_{1j}$ are estimated quantities from $\hat{p}_{DM}D_{1j}$ and $\hat{p}_{RM}R_{1j}$, respectively.[1]

Conceptually, this is equivalent to a round of estimation after deflating the independent variables by p_{DM} and p_{RM} respectively. In other words, the cross-equation correlations are ignored. An additional assumption is necessary to justify the independence between multiple choices, which will hold only when S_{2j} is equally correlated with D_{2j} and R_{2j}. Moreover, this imputation method cannot be extended to the independent variables when the number of parties in the first election is larger than 2.

As King suggests, along with Thomsen, such an extension would require multiple rounds of estimation of binary choices. Just to sketch the method, suppose a 2 by 3 table where a third party was included in the first election, and disappeared in the second; label the party S, along with the usual notation for major parties D and R. Our task is to estimate the conditional probabilities from the following equations:

$$\begin{aligned} D_{2j} &= p_{DD}D_{1j} + q_{SD}S_{1j} + q_{RD}R_{1j} \\ R_{2j} &= (1-p_{DD})D_{1j} + (1-q_{SD})S_{1j} + (1-q_{RD})R_{1j} \end{aligned} \tag{3.7}$$

[1]Note that in King's notation (1997, p. 69, Equation 4.5.), the first-round variable (M_{2j}) is included in the denominator because his objective is to compute the conditional fraction ($p|M_2$), which is different from the objective of the voter transition model.

The second equation is implied by the first, so estimating the first equation is enough as has been the case for 2 by 2 setup. First, collapse *any* two fractions from three parties and estimate the parameters. For example, choose to form a "liberal" block excluding party R, say, the Republican Party; call it L_{1j} and estimate

$$D_{2j} = p_{LD}L_{1j} + q_{RD}R_{1j} \tag{3.8}$$

Now choose another possible combination to form a "conservative" block, C_{1j}= R_{1j}+S_{1j}, then another round of estimation would be

$$D_{2j} = p_{DD}D_{1j} + q_{CD}C_{1j} \tag{3.9}$$

Since we have "fine estimates" of q_{RD} and p_{DD} from Equations (3.8) and (3.9), the only parameter left for estimation, q_{SD}, can be determined by substituting the results into equation (3.7) and solving (King 1997, pp. 265 –266) for:

$$\tilde{\tilde{q}}_{SD} = \frac{D_{2j} - \left(\tilde{p}_{DD}D_{1j} + \tilde{q}_{RD}R_{1j}\right)}{S_{1j}}. \tag{3.10}$$

As King admits, this procedure is not without problems. The results depend on the decision and sequence of combining the parties: for example, we can start by combining the major two parties excluding party S, and can come up with an estimate of $\tilde{q}_{SD}$ which may be different from $\tilde{\tilde{q}}_{SD}$ in Equation (3.10). As can be seen later, sometimes Equation (3.10) can yield negative values. In fact, this entire procedure is based on the assumption that the fractions of the three parties are uncorrelated with each other, which is a stronger assumption than is used in Equation (3.6). More recently, King provides a more sophisticated method with multinomial choices (Rosen *et al*. 2001) which implements Markov chain Monte Carlo (MCMC) methods on an assumed Dirichlet distribution: however, it still remains to be seen how robust and effective the method is against the voter transi-

tion setup in multiparty systems which necessarily will involve heavy truncation of possible distributions of parameters.

3.2.3 The Thomsen Estimator in Multiparty Setup

The multiparty extension of the Thomsen estimator is not straightforward, although he provides a reasonable algorithm that can be generalized to deal with an arbitrarily large number of parties (Thomsen 1987). Suppose we observe electoral support for the parties in time 1 ($X_1, \ldots, X_M$) and time 2 ($Y_1, \ldots, Y_M$). Along with Thomsen, let us define the joint proportion of voters that voted for party k at time 1 and party l at time 2 as p_{kl}, where $k = 1, ..., M$ and $l = 1, ..., N$. Note that the proportion parameters are fractions of the entire electorate, not the (conditional) loyalty rates. The goal is to fill in the table of coefficients shown in Figure 3.2.

		Election 1					
		X_1	X_2	$\cdots$	X_k	$\cdots$	X_M
Election 2	Y_1	p_{11}	p_{21}	$\cdots$	p_{k1}	$\cdots$	p_{M1}
	Y_2	p_{12}	p_{22}	$\cdots$	p_{k2}	$\cdots$	p_{M2}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	Y_l	p_{1l}	p_{2l}	$\cdots$	p_{kl}	$\cdots$	p_{Ml}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	Y_N	p_{1N}	p_{2N}	$\cdots$	p_{kN}	$\cdots$	p_{MN}

Figure 3.2: Coefficients in a Multiparty System: Thomsen's Approach

First, choose a "reference party" whose fraction of votes is not correlated or at least equally correlated to those of other parties. Thomsen (1987, p. 74) reports that choosing a neutral "party," such as an "abstention" category, often yields better estimation. From Figure 3.2, let Party M and N be the reference parties at time 1 and time 2 respectively. Second, choose a pair of parties against the

reference parties and compute the "crude-binary" correlation of their logits (or probits):

$$\hat{\rho}_{kl} = Corr(\ln \frac{X_l}{X_M}, \ln \frac{Y_k}{Y_N}).$$

Using this correlation, compute $\hat{p}_{kl,j}$, $\hat{p}_{kN,j}$, $\hat{p}_{Ml,j}$, and $\hat{p}_{ML,j}$ for each district using the inverse of Yule's tetrachoric correlation. This will enable one to impute $\hat{X}_{lj}$ and $\hat{Y}_{kj}$. Third, using these as "data," estimate the parameters, $\hat{p}_{kl}$, $\hat{p}_{kN}$, $\hat{p}_{Ml}$, and $\hat{p}_{ML}$. Fourth, compute the marginals with $\hat{\hat{Y}}_l = \sum_k \hat{p}_{kl}$and $\hat{\hat{X}}_k = \sum_l \hat{p}_{kl}$. Note that these estimated marginals will not necessarily equal the observed variables, and may not add up to unity. Then, finally, update the parameters with

$$\hat{\hat{p}}_{kN} = \hat{p}_{kN} \frac{X_k}{\hat{\hat{X}}_k} \quad \text{and} \quad \hat{\hat{p}}_{Ml} = \hat{p}_{Ml} \frac{Y_k}{\hat{\hat{Y}}_k}$$

and similarly for other coefficients. Go back to step three and iterate until the coefficients add up to unity, column by column.

Essentially, this iteration process tries to rescale the estimated coefficients by the amount of the "error" made in the initial binary estimation stage. As will be shown in the next section, this strategy produces reasonable estimates with real data. However, the results are sensitive to the choice of the "reference party" which will necessarily involve an arbitrary decision of the researcher. An ideal reference party would have vote proportions that are not related or are equally related to other choices: moreover, to guarantee stable results, the reference party would have to be sizable compared to other choices. In this vein, a category of "no voting" seems like a better candidate for the reference party—Thomsen's own suggestion. But this is not always the case. Sometimes, information on non-voters is simply not available; or sometimes, non-voting is strongly associated with certain political blocs, making it no better than using other parties as the reference

category.

To summarize, existing methods on adjusting multivariate coefficients have their own flaws. In the next section, I suggest a new alternative method of iterative proportional fitting and investigate its properties.

3.3 Iterative Proportional Fitting (IPF)

It should be noted that all the general strategies reviewed in the previous section are structured in two separate steps. First, split the categories into binary combinations and retrieve the ecological estimates; second, adjust the coefficients to match the known marginal proportions later. In fact, it is possible to think separately about the bivariate estimators and their multivariate extensions.

In this section, I propose an alternative way to combine such collections of bivariate ecological estimates into a multivariate context using the iterative proportional fitting (IPF) procedure. IPF has been used in many fields. But most importantly, it has been used to generate disaggregated spatial information from aggregated data in census studies. It is a mathematical scaling procedure to adjust a matrix of any dimension to converge to some pre-defined row and column totals, where the the constraining row and column totals are obtained from alternative sources. In a way, it acts as a weighting system whereby the original table values are gradually adjusted through repeated calculations to fit the row and column constraints.

Deming and Stephan (1940) first suggested using IPF in the context of adjusting sample cell proportions of a contingency table to external constraints given as marginal proportions. They argued that the technique, as often referred as "raking" estimates, can be used as a smoothing algorithm to adjust the cell frequencies

or proportions from survey estimates to conform to the marginal distribution of the same variables available from a census.

Even though IPF has never been used in the ecological inference literature, the theoretical underpinnings of IPF blend well with the objective of this section. The task here is to implement IPF into the process of adjusting estimated bivariate coefficients to obey the constraints, which are the aggregate marginal proportions. A description on how to implement IPF follows.

Following the running example of the multiparty voter transition problem, suppose we have crude estimates of entries in a two-by-two classification of electoral choices, X and Y, and call them p_{kl}. From Figure 3.2, suppose we retrieved the bivariate binary estimates by focusing upon one party against the rest of the parties. More specifically, we may ignore the estimated defection rates, that is, the transition *from* party i to other parties ($\sim i$), and just retrieve the loyalty rates. The first-round estimate can be written as

$$\hat{p}_{kl} = G(Y_{lj}, X_{kj})$$

where the function G indicates the ecological estimation process from aggregate data. For each permutation of time 1 and time 2 parties, this will yield a $M \times N$ matrix with first estimates, $\mathbf{P} = \{\hat{p}_{kl}\}$.

Let π_{kl} denote the true probability that $X = X_k$ and $Y = Y_l$ in the population of interest, and let π_{k+} and π_{+l} denote the known marginal probabilities. The problem of interest is to find estimates, $\hat{\pi}_{kl}$, of π_{kl} by adjusting the proportions of $\mathbf{P}$ to the known marginal probabilities so that

$$\begin{aligned} \sum_k \hat{\pi}_{kl} &= \pi_{k+}; \\ \sum_l \hat{\pi}_{kl} &= \pi_{+l}. \end{aligned} \tag{3.11}$$

Deming and Stephan (1940) proposed the following algorithm to retrieve the estimates $\{\hat{\pi}_{kl}\}$ that minimizes the discrepancy between them and $\{p_{kl}\}$. Let $\left\{\hat{\pi}_{kl}^{(t)}\right\}$ be the estimates of $\{\pi_{kl}\}$ at the tth iteration, and initially let $\pi_{kl}^{(0)} = p_{kl}$ for all k and l. The algorithm proceeds by row and column adjustments, such as the following at iteration t:

$$\begin{aligned} \hat{\pi}_{kl}^{(t)} &= \hat{\pi}_{kl}^{(t-1)} \frac{\pi_{k+}}{\hat{\pi}_{k+}^{(t-1)}} \quad \text{if } t \text{ is odd} \\ \hat{\pi}_{kl}^{(t)} &= \hat{\pi}_{kl}^{(t-1)} \frac{\pi_{+l}}{\hat{\pi}_{+l}^{(t-1)}} \quad \text{if } t \text{ is even} \end{aligned} \tag{3.12}$$

where "$+$" denotes summation over the corresponding subscript. Repeat this until the difference between the adjusted marginals and true marginals diminish under a reasonable tolerance level.

Before proceeding to look into the properties of the estimator in more detail, I provide a simple example that illustrates the IPF algorithm. Consider Table 3.1 where we have "crude" initial ecological estimates $\{p_{kl}\}$ and the target marginals $\{\pi_{k+}\}$ and $\{\pi_{+l}\}$. Of course, the first step binary ecological estimates do not normally produce marginals as inaccurate as it is shown in the example, but recall that the IPF algorithm was originally introduced to adjust non-representative survey estimates to conform to the census marginals. Thus, we may think of the left-hand-side table, the "initial estimates" as any general starting values for the iteration process. Here, the task is to adjust estimates of the joint probabilities in the cell to conform to the aggregate marginals.

Table 3.2 shows the iteration process applied to the given example. The first iteration consists of forcing the column marginals to conform to the target marginals,

	Initial Estimates (p_{kl})					Target Marginals (π_{k+} and π_{+l})			
	Dem.	Indep.	Rep.	Total		Dem.	Indep.	Rep.	Total
Dem.	.35	.03	.12	.50	Dem.				.3
Indep.	.03	.05	.02	.10	Indep.				.3
Rep.	.02	.07	.31	.40	Rep.				.4
Total	.40	.15	.45	1.0	Total	.5	.2	.3	1.0

Table 3.1: Implementing the IPF Algorithm: An Example

multiplying each row of the initial estimates by the target marginal proportion, and dividing it with the original marginal proportion. We can see that the row marginals are adjusted to the target fractions π_{k+}, that are .3, .3, and .4 for the Democratic, Independent, and the Republican parties, respectively. It should be noted that the estimated column marginals have changed but do not conform to the target marginals at this round of iteration.

The second step involves the adjustment of the cells to the column marginals this time, where estimates of the previous round gets adjusted to correspond to the target marginals, (.5, .2, .3). The row marginals that were forced to match the target marginals are changed back, but are now closer to the target marginals. The last table shows the final converged set of estimates where the initial estimates are transformed to satisfy the target marginals.

Several remarks are in order. First, this process does not necessarily require the entries to be probability distributions—it will work with frequency counts as well where one would simply use marginal count totals instead of the proportions. This would enable us to deal with situations where the initial estimates from the binary ecological relationships do not add up to unity. Second, as can be seen from the algorithm, the adjustments are aimed at finding estimates that transform p_{kl} minimally, based upon the constraints in Equation (3.11) above.

First Iteration				
	Dem.	Indep.	Rep.	Total
Dem.	.21	.02	.07	.30
	$(.35 \times .3/.5)$	$(.03 \times .3/.5)$	$(.12 \times .3/.5)$	
Indep.	.09	.15	.06	.30
	$(.03 \times .3/.1)$	$(.05 \times .3/.1)$	$(.02 \times .3/.1)$	
Rep.	.02	.07	.31	.40
	$(.02 \times .4/.4)$	$(.07 \times .4/.4)$	$(.31 \times .4/.4)$	
Total	.32	.24	.44	1.0

Second Iteration				
	Dem.	Indep.	Rep.	Total
Dem.	.33	.02	.05	.39
	$(.21 \times .5/.32)$	$(.02 \times .2/.24)$	$(.07 \times .3/.44)$	
Indep.	.14	.13	.04	.31
	$(.09 \times .5/.32)$	$(.15 \times .2/.24)$	$(.06 \times .3/.44)$	
Rep.	.03	.06	.21	.30
	$(.02 \times .5/.32)$	$(.07 \times .2/.24)$	$(.31 \times .3/.44)$	
Total	.50	.20	.30	1.0

$\vdots$

Final Convergence				
	Dem.	Indep.	Rep.	Total
Dem.	.27	.01	.02	.3
Indep.	.17	.11	.03	.3
Rep.	.06	.08	.25	.4
Total	.5	.2	.3	1.0

Table 3.2: IPF Iterations: Example Continued

3.4 Simulation using Ballot Images The 2000 US Presidential Election in Miami-Dade, FL

In this section, I provide simulation results using the 2000 US Presidential Election ballot image data from Miami-Dade, Florida, to compare the IPF estimator against the Thomsen estimator. Other aforementioned ecological inference strategies, for example, the regression approach or the King estimator are not examined in this section, since we have already established that the crude binary estimates, that are, the initial values, are plagued by serious biases. They are also the very different from the Thomsen estimates making the comparison almost meaningless. The two estimators that are compared here, namely, the Thomsen estimator

and the IPF extension of it, share the same identical initial estimates, enabling us to isolate and evaluate the IPF adjustment. The comparison will focus upon the biasedness and efficiency of the two estimators.

Before I proceed, it should be noted that there are some difficulties in simulating aggregate data in general. Simulations will have to mimic the data generating process closely while carefully introducing the random error component. However, researchers usually have limited information and tools on how the aggregation process works in general.[2]

3.4.1 Point Estimates of IPF

The ballot image data from the 2000 US presidential election in Florida provide an excellent opportunity to study these issues. After the electoral crisis that went on in the election concerning the examination of the contested ballots in Florida, the ballot image data were collected documenting the vote choices of individual voters in a few problematic counties. For example, we should be able to compute the true proportion of straight-ticket voters that chose to vote for both the Republican presidential candidate, Bush, and the Republican Senate candidate, McCollum. This exactly is the definition of the loyalty rate or the "transition" rate of Republican voters across two elections.

Aggregation of the ballot images by precincts will exactly produce ecological data that would just embody any unobserved aggregation problems. The first question is to see how well ecological estimators recover the true parameter from individual data. Table 3.3 produces individual-level estimates from the ballot image data. As can be seen in the table, about 40% of the voters were straight-ticket

[2]One exception would be Tom Cho and Anselin (2002) where they try to simulate the aggregation process by using geographical proximity using GIS.

		Senate		
		Rep.	Dem.	Others
Presidential	Bush	0.404	0.036	0.012
	Gore	0.034	0.479	0.024
	Others	0.003	0.006	0.002
$N = 538,025$				

Table 3.3: Distribution of Voters in Presidential and Senate Contests, Ballot Image Estimates: 2004 General Election, Miami-Dade, Florida

		IPF				Thomsen		
		Senate				Senate		
		Rep.	Dem.	Others		Rep.	Dem.	Others
Presidential	Bush	0.412	0.035	0.007	Bush	0.419	0.016	0.007
	Gore	0.017	0.483	0.035	Gore	0.021	0.490	0.004
	Others	0.007	0.005	0.000	Others	0.014	0.028	0.000
J (Number of Precincts) $= 613$								

Table 3.4: Distribution of Voters in Presidential and Senate Contests, Ecological Estimates: 2004 General Election, Miami-Dade, Florida

Republican voters, while 48% voted Democratic in both presidential and senate contests. To produce these estimates, more than half a million ballot image observations were used, after excluding invalid votes—both "over-votes" and "under-votes"—and dropping absentee ballots. In any case, treating these as the true parameters of interest, the ecological estimators are put to test.

Table 3.4 compares ecological estimates retrieved by Thomsen and the IPF extension of it. The data set was generated by aggregating ballot images into 613 precincts—which exactly should be in the same format as any aggregate data set. As can be seen, both estimators perform well, especially on the sizable parameters that represent the straight-ticket voters. The off-diagonal coefficients, however, namely the fraction of voters who represent the "defection rates," were somewhat off the target in the estimates. Among the six off-diagonal estimates, four of them produced by IPF were clearly closer to the ballot image estimates while only one

of Thomsen estimates—the Senate Republican voters who also voted for Gore—was closer to the ballot image estimate. Of course, without any discussion of the standard error, this comparison is less than clear, however, and I address this issue in the next section.[3]

As mentioned, another problem with Thomsen's multiparty model is that it requires an arbitrary decision of the "reference party" which will work as the baseline category for all the possible contrast sets. The estimation above was implemented based upon using vote shares of Republican candidates as the "reference party": choosing Democrats instead slightly changes the estimates, making them further away from the true parameters, while using the "Others" category as the reference produces results that are quite erratic. Thomsen asserts that choosing a "neutral" and sizable category, such as abstention produces better estimates, a piece of information that is not always reliably available in many electoral settings. In any case, an arbitrary choice of the reference party remains to be a problem in the Thomsen estimator, while it is not a problem with the IPF implementation.

[3]It is still possible to consider the possibility of applying the IPF algorithm to linear estimators. However, at least for the Goodman regression, IPF does not apply since the multivariate estimates from a corresponding seemingly unrelated regression (SUR) would exactly conform to the population marginals. In other words, there is noting to "rake." The following table provides estimates from SUR.

			Senate	
		Rep.	Dem.	Others
Presidential	Bush	0.460	-0.009	0.006
	Gore	-0.020	0.511	0.040
	Others	0.006	0.011	-0.005
		$J = 613$		

Distribution of Voters in Presidential and Senate Contests, Goodman Regression (SUR):
2004 General Election, Miami-Dade, Florida

3.4.2 Standard Errors and MSE

As was mentioned and implemented in the previous chapter, Achen (2000) suggested the standard errors of the binary Thomsen estimator using Fisher's z-transformation. He suggested computing the 95% confidence intervals of the estimated aggregate correlation coefficient by

$$r \pm 1.96 S.E.(r) \quad = \quad \tanh(\frac{1}{2} \ln \frac{1+r}{1-r} \pm \frac{1.96}{\sqrt{n-2.5}}).$$

This will enable us to compute the upper and lower bounds of the estimates: dividing the difference by 2 or 1.96, provided that the bounds are symmetric around the point estimates, will approximate the standard error of the estimates. For example, it approximately translates to a standard error of .002 in the crude binary estimate of the Bush-Republican Senate cell, while the standard error of the crude binary estimate of Others-Bush is around .0035. Applying them as the standard errors of the above estimates for hypothesis tests, we realize that not many of the true parameters are within two standard errors of the estimates.

It is possible that the additional IPF transformation of the original crude binary estimates introduces additional sources of error. However, the complex multivariate transformation will make it prohibitively difficult to trace out such addition of noise. Here, I design a simulation using the ballot image data to estimate the standard errors of the estimates.

A straightforward way to proceed is to collect bootstrap standard errors. More specifically, denote $\hat{\theta}_n$ as an estimate of a parameter vector based on a sample with n observations. We try to approximate the statistical properties of $\hat{\theta}_n$ by collecting a sample of bootstrap estimators, $\hat{\theta}_m^{(b)}$ where $b = 1, \ldots, B$, obtained by sampling m observations with replacement and computing $\hat{\theta}$ with each sample, B times. Then,

	IPF			Thomsen		
	Coefficient	Standard Error	Root-MSE	Coefficient	Standard Error	Root-MSE
Bush-Sen. Rep	0.412	0.0114	0.0145	0.419	0.0108	0.0194
Bush-Sen. Dem	0.035	0.0019	0.0021	0.021	0.0022	0.0149
Bush-Sen. Others	0.007	0.0006	0.0043	0.014	0.0012	0.0029
Gore-Sen. Rep	0.017	0.0020	0.0175	0.016	0.0017	0.0187
Gore-Sen. Dem	0.483	0.0110	0.0115	0.490	0.0106	0.0153
Gore-Sen. Others	0.035	0.0011	0.0102	0.028	0.0017	0.0037
Others-Sen. Rep	0.007	0.0003	0.0033	0.007	0.0004	0.0038
Others-Sen. Dem	0.005	0.0003	0.0010	0.004	0.0004	0.0013
Others-Sen. Others	0.000	0.0000	0.0022	0.000	0.0000	0.0023

Table 3.5: Bootstrap Estimates of Coefficients and Their Precision

the estimated asymptotic covariance matrix would be (Greene, 2005):

$$Var(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left[\hat{\theta}_m^{(b)} - \hat{\theta}_n\right]\left[\hat{\theta}_m^{(b)} - \hat{\theta}_n\right]'.$$

The key statistic that enables us to study the covariance structure is the difference between the bootstrap estimator $\hat{\theta}_m^{(b)}$ and the usual estimator on the full sample, $\hat{\theta}_n$. What is especially nice about the particular feature of the data set is that it enables us to get estimates of the mean squared error terms, which is another important criterion of the estimators here. Table 3.5 reports results of such estimates.

The results are from $B = 500$ repetitions of bootstrap simulation, although estimates of the standard errors converge fairly quickly with around $B = 50$. First of all, the estimated standard errors are comparable between the two estimators, indicating an equivalent level of efficiency. The standard errors of the larger entries such as the straight-ticket voters are both at around the 1% point mark, putting the true values of the coefficients within the confidence intervals. In a way, this just confirms the conventional understanding that ecological estimators come with

sizable biases and inadequately small and underestimated standard errors.

The mean squared errors are computed by the usual formula, $MSE = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$, and it is evident that IPF implementation of Thomsen has smaller MSE than that of the Thomsen estimator. This mostly is due to the fact that the IPF estimator is at least as efficient as that of Thomsen's, but comes with less amount of bias.

The complexity of the IPF adjustment as well as most of the multiparty estimation strategies I reviewed here makes it difficult to derive analytically various measures of uncertainty of the estimators. In addition, applying the IPF procedure in this particular context of adjusting the ecological estimation introduces more complexity into the analytical process.[4] In this section, I have provided results from a bootstrap simulation that supports the usage of the IPF estimator over the Thomsen estimator in multiparty system.

3.5 Empirical Test: Voter Transition in South Korean Presidential Elections 1992–1997

Here, I test different ecological estimation strategies as defined previously on the voter transition in South Korean presidential elections between 1992 and 1997. Setting aside some methodological considerations for a while, there are several substantively interesting issues involving the voter transition in these two elections.

First of all, the voter transition in the two elections provides a good starting point by letting us test and calibrate ecological inference techniques before we can move onto other elections. The 1992 and 1997 presidential elections featured

[4]For example, Little and Wu (1991) derive the standard error of the IPF estimator in the context of raking survey estimates to match census results, where they specify it to be only applicable to the cases where two sets of estimates are not correlated to each other. In the particular context of adjusting the first-round ecological estimates, this condition exactly is violated and their results do not apply.

similar configuration of parties, including a major candidate—Kim Dae Jung—running in both of them. Arguably, the five years between the two elections stands as a politically stationary period, which is a rarity in South Korean politics. With the availability of the survey data that can provide corresponding individual-level estimates for voter transition, we have a very good setup to check the accuracy and adequacy of ecological estimators on Korean elections.

In a way, the movement of voters is an important issue in studying a political system where parties are yet to be institutionalized. Suppose we are interested in finding out how capable Korean voters are to make consistent electoral choices. Given the fluid structure of the party system and frequent changes in party labels, it more often than not is the candidates (or importantly, political bosses) that provide the sense of continuity in the South Korean party system. The voter transition rates model will offer an important insight into the evolution of South Korean politics—how strong the politics of personalities is; how relevant party labels are; and whether "old politics" is in the process of being phased out by more institutionalized sets of party politics.

If individual-level data are available, we should be able to trace this back to the start of the country and learn the dynamic transformation in the calculus of the Korean voters. However, survey studies are not available for elections before 1987, and the only data available are aggregate-level data: thus, there are the a typical ecological inference problems. The elections 1992–1997 provide interesting leverage in the sense that they enable us to test whether and which ecological inference strategies provide reasonable answers for the voter transition problem in South Korean elections: by comparing ecological estimates to survey results.

Moreover, the transitions of voters between the two elections have interesting

1987	1992	1997
Roh Tae Woo (36%)	Kim Young Sam (40%)	Lee Hoe Chang (39%)
Kim Young Sam (27%)	Kim Dae Jung (32%)	Kim Dae Jung (40%)
Kim Dae Jung (26%)	Chung Ju Young (16%)	Lee In Je (19%)

Table 3.6: Candidates in Presidential Elections, 1987–1997

implications *per se*. As noted, the 1997 election was the first time in South Korean history when a presidential candidate from the "opposite" party, namely, Kim Dae Jung, won the election, after his three unsuccessful bid for the presidency. It would be interesting to find out how loyal his former supporters were and what constitutes his new support base.

3.5.1 Background and Data

Before going into the details of the estimation, a quick review of the "data-generating process" is in order. Especially when engaging in an ecological inference estimation, the researcher should provide some relevant "external information" that could help our inference.

The year 1987 witnessed the first democratic presidential election in South Korea since the start of the series of authoritarian regimes in 1972. Although it is still a matter of debate how meaningful elections were before 1987, it is safe to say that presidential elections became competitive and politically significant after the 1987 election[5]. The three elections shown in Table 3.6 were marked by regionally divided close races. None of the winners won the contest by more than an 8% margin of the total votes cast. None of the winners were able to gain a national majority support.

[5]Several presidential elections (1952, 1956, 1963, 1967, and 1971) that took place before the authoritarian regime were arguably competitive, although the governing bloc always won the competition by fairly comfortable margins.

Roh Tae Woo, the candidate from the ruling party in the authoritarian government before the democratization, came out as the winner of the 1987 election. Kim Young Sam, a former opposition leader, won the 1992 election after joining the governing party. In 1997, Kim Dae Jung, who lost in the two previous elections, became the first opposition party candidate to defeat a ruling party opponent. Many interesting and substantive questions can be asked based on voter transition behavior: for example, how was this first alternation of power between the competing parties (Huntington 1991: 266–267) achieved? Are Korean voters loyal to parties or the party bosses?

Even though the Korean party system is not stable and exhibits frequent changes in party labels, there is a significant degree of consistency in political support that is comparable to party identification in established democracies. In other words, voters manage to find, identify, and vote for "their" parties and candidates, even when party names change quite frequently. This fact is sufficient to justify the application of the voter transition model to the Korean electorate and clarify the meaning of "loyalty" or "defection" rates in this chapter.

Two data sets of Korean presidential elections are used in this chapter: an aggregate level election data set and one survey study. To build the aggregate data set, election results of the 1992 and 1997 Korean presidential elections were matched to each other. The official election results are reported at the precinct level, the numbers of which add up to around 15,300 and 16,400 for the two elections. The observations had to be aggregated to an administrative unit ("*Dong*") in the matching process, resulting in about 3,500 observations. The number of units is further reduced to 3,380 after dropping those with the population change ratio

over 100% and below 50%.[6]

A survey data set is used to compare the results from the aggregate estimation of voter transition rates. The data set is from the *National Survey of the Fifteenth Presidential Election* conducted by the Korean Social Science Data Center. One thousand and two hundred respondents were randomly sampled from the national voter registration list: the sampling quotas were assigned by region, gender, and age, based on the corresponding census proportions record. The survey started on the next day after the 1997 election and ended on the fourth day. Among the questionnaires in the original data set, two key variables are taken to form a cross tabulation: who the respondents voted for in the 1997 election and their recall of whom they had voted for in the previous presidential election, 5 years ago. Cross tabulations of the survey are used as references for the reliability of the ecological estimates in the next section.

It is a well-established fact that survey results are by no means perfect indicators of true voting behavior: respondents often forget or lie. For example, one of the sure predictions in survey research is the over-reporting of turnout: more respondents will claim to have voted in post-election interviews than actually cast ballots (Traugott and Katosh 1979; Duff et al. 2007). Also, recalls of previous votes are, at best, not without noise. For example, Blair Weir (1974) finds a post-election "bandwagon effect" toward the winner from 1956-58-60 Survey Research Center panel data of US voters.

Table 3.7 compares the survey estimates to the actual outcome (from the aggre-

[6] A potential problem stemming from these dropped cases should be mentioned. The units with the highest population increase are mostly the suburban areas of Seoul and Pusan ("bed-town areas"), where the residents tend to be more educated, younger, and employed middle class. Significant population decrease tends to occur mostly in rural areas where the residents are on average older, less educated and agriculture-based people with lower income. No study yet has clarified the relationship between these variables and partisan voting, which is an important substantive issue itself, but here, we just assume that the deletions are not systematic.

gate data used) of the two elections. We can see a number of sizable discrepancies between the corresponding results of the two data sets. First, in both elections, there are around 11% of over-report of turnout in the survey data. Second, the responses regarding the 1997 election, right after which the survey was conducted, show relatively less discrepancy with the actual outcome, except for the turnout. However, a serious amount of disagreement exists between the aggregate and survey data sets for Kim Young Sam and Chung Joo Young's support in 1992.

Although the sources of the problems are unknown, several hypothetical explanations can be provided. First, the over-reporting of Kim YS support in the survey is consistent with the "bandwagon effect." Some respondents could have forgotten whom they actually had voted for 5 years ago, and could have answered that they voted for Kim YS who won the election and had been serving as the President since. Second, the under-reporting of Chung's support in the survey could be due to the fact that the candidate's party was made *ad hoc* right before the election and vanished right after Chungs defeat in the 1992 election. This ephemeral nature of the party could have encouraged the respondents to forget or to hide their previous support for Chung.

The survey estimate of the defection rate from Chung JY to Kim DJ will also likely to have a positive bias since Chung's previous supporters are under-reporting while Kim DJ's 1997 support is inflated. In both cases, there are good reasons to believe that the true loyalty rate is smaller than the survey estimate, and any ecological estimate larger than the survey estimate is likely to be biased—more biased than the survey estimate.

However, despite these problems, the three party contest captured by the survey is somewhat workable. The point here is that the survey estimates are not

	Candidates	Aggregate	Survey
1992 Election	Turnout	83%	94%
	Kim Young Sam	45%	52%
	Kim Dae Jung	36%	35%
	Chung Joo Young	18%	13%
1997 Election	Turnout	81%	92%
	Lee Hoe Chang	39%	39%
	Kim Dae Jung	41%	43%
	Lee In Je	20%	18%

Note: The percentages for "Abstention" were calculated based on the total number of eligible voters in both data sets and both elections. Because of the large over-reporting of turnout, entries for each candidate represent percentages of the three party vote only.

Table 3.7: National Support for Candidates: Survey vs Aggregate

perfect but the best we have and there are still certain ways to make better use of them as points of reference. Even though survey data will not provide the perfect "truth" that we can use as a reference for our ecological estimation, it should be sufficient for our purpose where we test and compare different classes of ecological estimators.

3.5.2 Results

As was shown, three major candidates competed in the two elections. The party system in Korea can be characterized as a "two-and-a-half" party system: two major parties compete against each other, while numbers of parties and presidential candidates have always taken a distant but still noticeable third place. In 1992 and 1997, none of the third candidates gained less then 16% of the total national vote; and those two third-party candidates from both elections, Chung Joo Young and Lee In Je, are not directly linked to each other. As was described in the previous section, multiparty estimations require even more sets of strong assumptions and accordingly, the results are unstable. Thus the setup is in a sense a severe test to check the reliability and consistency of the estimators. How we

denote the candidates and the parameters of interest is shown in Figure 3.3.

1992 1997	Kim Young Sam	Kim Dae Jung	Chung Joo Young	Total
Lee Hoe Chang	p_{RR}	p_{DR}	p_{SR}	R_2
Kim Dae Jung	p_{RR}	p_{DD}	p_{SD}	D_2
Lee In Je	$1 - p_{RR} - p_{RD}$	$1 - p_{DR} - p_{DD}$	$1 - p_{SR} - p_{SD}$	S_2
Total	R_1	D_1	S_1	1

Figure 3.3: Variables and Parameters of Voter Transition Rates: South Korean Presidential Elections, 1992–1997

Table 3.8 provides the estimated voter transition rates of various estimators. "Survey" refers to the estimates from the cross-tabulation of reported votes in the two elections. Thus, the coefficients would mean the probability to vote for the time-2 candidate given s/he voted for a particular time-1 candidate. The survey estimates raise some interesting points, although most of the estimates look reasonable.

One picture that emerges from the survey estimates on transition rates is that Kim Dae Jung was able to draw sizable support from those who voted against him—especially from previous Kim Young Sam voters—even though he lost some of his previous supporters. In fact, this is consistent with Kim Dae Jung's central electoral strategy where he tried to change his leftist image and form a new conservative coalition, which also is known as the "New DJ Plan."

Meanwhile, some would argue that the loyalty rate of Kim Dae Jung at 87% is less than impressive given the fact that he had enjoyed a monolithic electoral support from those who are from *Chunla* region; and from those who are younger, liberal, and who feel strongly against the traditional ruling bloc that used be in charge of the authoritarian regime before the democratization of the country. In other words, it was his ever-loyal supporters that won Kim the 1997 election. His

supporters were in strong contrast to the GNP supporters many of whom changed their party of choice and supported Lee In Je.

Was it the monolithic supporters that won Kim the election and the presidency? Or was it the "New DJ Plan" that tapped new political support for him? In any case, these are empirical questions, and it is possible to see which storyline the aggregate estimates confirm.

The sub-table in Table 3.8 labeled "Goodman" shows results from linear regressions on the system of equations

$$\begin{aligned} R_2 &= p_{RR}R_1 + p_{DR}D_1 + p_{SR}S_1 \\ D_2 &= p_{RD}R_1 + p_{DD}D_1 + p_{SD}S_1 \end{aligned}$$

where the variables and parameters are defined in Figure 3.3. As was the case with binary estimations, we still observe out-of-bounds estimates; in this case, coefficients larger than unity and smaller than zero at the same time. However, the Goodman results seem to support the monolithic-supporter hypothesis, where the estimation shows that Kim (more than) perfectly retained his electoral support from the previous election, and that was almost enough for him to win the election, where Lee In Je took about 30% of the former GNP votes. In any case, the negative transition rates are biased, and so is the loyalty rate for Kim.

The overestimation of the Kim Dae Jung loyalty rate continues to be the problem from the bivariate estimation. As was argued above, if the survey estimate of 86.7% is already an overestimate, we can be quite confident that the Goodman estimate is biased, and more likely than not biased even from the true value of the coefficient.

Survey	KimYS92	KimDJ92	Chung92
LeeHC97	0.6311	0.0430	0.3429
KimDJ97	0.1748	0.8674	0.3571
LeeIJ97	0.1942	0.0896	0.3000

Goodman	KimYS92	KimDJ92	Chung92
LeeHC97	0.7171	-0.0101	0.2970
KimDJ97	-0.0033	1.0196	0.2758
LeeIJ97	0.2863	-0.0095	0.4272

Constrained	KimYS92	KimDJ92	Chung92
LeeHC97	0.7079	0.0100	0.2926
KimDJ97	0.0100	0.9800	0.2939
LeeIJ97	0.2821	0.0100	0.4135

King	KimYS92	KimDJ92	Chung92
LeeHC97	0.7720	0.0060	0.2328
KimDJ97	0.0477	0.9608	0.2112
LeeIJ97	0.1803	0.0332	0.5560

King-MLE	KimYS92	KimDJ92	Chung92
LeeHC97	0.7280	0.0065	0.3441
KimDJ97	0.0577	0.9542	0.1990
LeeIJ97	0.2143	0.0393	0.4570

Thomsen	KimYS92	KimDJ92	Chung92
LeeHC97	0.6211	0.0720	0.4815
KimDJ97	0.1446	0.8167	0.2682
LeeIJ97	0.2343	0.1112	0.2503

Thomsen-IPF	KimYS92	KimDJ92	Chung92
LeeHC97	0.6124	0.1260	0.3936
KimDJ97	0.1268	0.8549	0.2339
LeeIJ97	0.2608	0.0192	0.3726

Table 3.8: Ecological Estimates from a Three Party System: South Korean Elections 1992–1997

Note that the next sub-table produces estimates from constrained regressions, where certain adjustments are applied to the Goodman regression to make sure the estimates of each column add up to 1 in a way that is consistent to the cross-equation correlations that was developed in the previous section. Following the details of the previous section on how the adjustment for constrained regressions was applied, the Goodman coefficients are "corrected." For example, since all the loyalty and defection rates for Kim DJ are out of bounds, we set them to be .98, .01, and .01 before going into the second round of estimation. The results are not drastically different from the Goodman regression.

For King's model, this chapter adopts Equation (3.10) where Chung Joo Young's vote was used as the "neutral" party that is collapsed to the other major parties to retrieve binary estimates. The fact that he nor the party did not last until the next election justifies this selection. Moreover, Chung support is the least correlated with other candidates. I also report the result of using the same procedure on King's first-round estimates (the MLE estimates before simulation) generated by *EI* and label it as "King-MLE."

The results are quite interesting considering the pattern that we have seen from the previous chapter. Substantively, it does not differ from the linear regression results, but it should be noted that the estimate of GNP loyalty rate (KimYS92→LeeHC97) is at 77%, which is larger than the survey estimate and worse than the Goodman estimate. It appears that all the loyalty rates, or the main-diagonal entries, are inflated while the rest are underestimated—at least compared to the survey estimates.

The table also provides estimates implementing Thomsen's method into the multiparty context, as was described in detail in the previous section and was re-

trieved by his program ECOL. The third candidates in the two elections, namely, Chung Joo Young in 1992 and Lee In Je in 1997 elections were used as the "reference categories" under the reasoning that they are the more "neutral" candidates.

Finally, I also report the iterative proportional estimate applied to the probit version of Thomsen's estimator at the bottom of the table. As seen earlier, the difference between the two sets of estimates is that the former implements logit specification and tetrachoric correlation while the later uses probit specification and IPF, as was documented in the previous chapter and in the previous section.

The results approximate the survey estimates better than any other ecological estimators, supporting the "New DJ Plan" argument, which attributes Kim Dae Jung's success to his newly-found support basis. More specifically, it highlights the fact that Kim lost some of his support from the previous election, while managed to gain some political ground from previous GNP supporters. Another noticeable fact is that the transition rates from Chung Joo Young (the last column) are quite different from the survey results—where survey estimates are the least reliable due to the response problems we have seen in Table 3.7.

The difference between the two versions of the Thomsen estimator mainly involves the transition rates from Kim Dae Jung support in 1992. The loyalty rate from the IPF version approximates closer to the survey estimate than Thomsen's own ECOL estimate. Among the two, the question on which estimate is better is still inconclusive, mostly due to the possibility that the survey estimate could be an overestimation, as noted above. Also, both Thomsen estimators highlight the fact that Kim Dae Jung in 1997 did not do any better among the previous Chung supporters, which contradicts the survey estimates. However, noting that the survey estimate for this may be biased upwards due to the under-report of Chung

supporters, we suspect the survey estimate (36%) may be inflated and can reasonably argue the ecological estimate may represent the truth better.

3.6 Remarks

Continuing the trend in the previous chapter, the approach suggested by Thomsen, and especially the IPF implementation of it, produce better estimates. Here, in the multiparty implementation of the ecological estimators, it appears that estimations from the initial binary configurations dominate the final result. If the initial step gives results erroneous by a sizable margin, getting a reliable final result is almost impossible. Although none of the initial estimation is passed on to the next stage unchanged, any erroneous first estimate would accordingly inflate or deflate other first step estimates—even if they are correct.

In this regard, getting a good first-order estimate is important. It is possible to think of adjusting any first-round two-by-two estimates with the IPF algorithm. However, since IPF tries to minimize the change of estimates from the initial values in the adjustment process, the bias introduced in the initial stage will carry over to the converged final IPF estimate. Moreover, the Goodman regression or SUR will always produce estimates that exactly conform to the target marginals, thus IPF adjustment is not applicable to them.

There is still more work to be done investigating the properties and possibilities of the multiparty extension of various ecological estimators. In this chapter, I introduced the use of the iterative proportional fitting process in the context of adjusting first-round binary estimates. After looking into the IPF adjustment to the Thomsen estimator, I have shown some evidence that it may be the best ecological estimator we have so far studying multiparty voter transition rates.

CHAPTER IV

Ecological Inference with Covariates

4.1 Introduction

The previous chapter discussed an important extension to the ecological inference techniques: its application to multinomial choices. In this chapter, I will propose another extension to the basic model: how to incorporate covariates. The extension I develop here mainly focuses upon Thomsen's model which was shown to be the best estimator so far in studying the voter transition problem. This will prove to be the critical step for ecological inference techniques in analyzing the underlying forces that govern the dynamic electoral process.

To be sure, the voter transition model in itself has merits in understanding electoral dynamics. However, usually a researcher will want to move a step further than a mere description of loyalty and defection rates, and to understand the mechanism behind the movement of the voters. For example, in the previous chapter, with the help of ecological inference techniques, it was possible to estimate the proportion of straight-ticket voters in Miami-Dade, Florida to be around 90%. But who are they and what characterizes the 10% of those who split their electoral support to candidates from different parties? Or, as was shown in the South Korean example, it was possible to recover the retention rate of Kim Dae

Jung supporters—but the more substantively interesting task would be to identify the driving forces of such (non)movement of the voters. In the framework of the voter transition model discussed so far, it is not possible to ask such questions yet.

In this chapter, I first develop an extension to the Thomsen estimator and spell out the logic behind it. Later, I apply the extension to study the impact of democratization in South Korean elections.

4.2 The Model

4.2.1 The Voter Transition Setup with Covariates

Essentially, we are interested in the difference in the voter transition rates among different demographic groups. The question would be equivalent to modeling the loyalty and defection rates as functions of the covariate. Suppose we are interested in the transition rates of workers, where a binary variable z_i would indicate whether an individual is a worker or not. If we had individual-level survey data, we would be able to estimate the parameters by a three-way cross tabulation, as is depicted by Figure 4.1, where x_i indicates the vote choice of individual i at time 1, and y_i indicates that at time 2.

	Worker ($z_i = 1$)		Non-Worker ($z_i = 0$)	
	$x_i = 1$	$x_i = 0$	$x_i = 1$	$x_i = 0$
$y_i = 1$	p_1	q_1	p_0	q_0
$y_i = 0$	$1 - p_1$	$1 - q_1$	$1 - p_0$	$1 - q_0$

Figure 4.1: Voter Transition with a Covariate

In the same manner shown in the previous chapter, we should be able to write a regression relationship at the individual level as

$$\begin{aligned} \Pr(y_i = 1) &= p_i x_i + q_i(1 - x_i) \\ \text{where } p_i &= p_1 z_i + p_0(1 - z_i) \\ q_i &= q_1 z_i + q_0(1 - z_i) \end{aligned}$$

Or equivalently, we may write

$$y_i = z_i \left[p_1 x_i + q_1(1 - x_i)\right] + (1 - z_i) \left[p_0 x_i + q_0(1 - x_i)\right] + u_i \tag{4.1}$$

which is a fully saturated interactive model between the electoral choice and a covariate. Reorganizing the equation above yields

$$y_i = q_0 + (q_1 - q_0)z_i + (p_0 - q_0)x_i + \left[(p_1 - q_1) - (p_0 - q_0)\right] x_i z_i + u_i. \tag{4.2}$$

which is displayed in the usual OLS form of $y_i = \alpha + \gamma z_i + \beta x_i + \delta x_i z_i + u_i$. Loyalty and defection rate parameters in the two groups are all identified since

$$\begin{aligned} q_0 &= \alpha \\ q_1 &= \alpha + \gamma \\ p_0 &= \alpha + \beta \\ p_1 &= \alpha + \beta + \gamma + \delta. \end{aligned}$$

In other words, with individual-level data, OLS will exactly produce the parameters in Figure 4.1, and the estimate will have the usual desirable properties since $E(u_i) = 0$ and $Cov(u_i, x_i) = Cov(u_i, z_i) = 0$.

But what about the aggregate relationship? It is not straightforward to derive the aggregate equivalent of the equation analytically in the same fashion that

was demonstrated in Chapter 2, where the Goodman regression was derived. For example, by averaging both sides of Equation (4.2), we may try to derive the relationship between the aggregate variables, Y_j, X_j, and Z_j, namely, party support in the two elections and the proportion of workers in district j, in the following manner. Averaging both sides of Equation (4.2) in district j, with n_j total voters, would yield

$$Y_j = q_0 + (q_1 - q_0)Z_j + (p_0 - q_0)X_j + [(p_1 - q_1) - (p_0 - q_0)] \frac{\sum_i x_i z_i}{n_j}. \quad (4.3)$$

If x_i and z_i are not correlated, then $Cov(x_i, z_i) = \frac{1}{n_j} \sum_i x_i z_i - X_j Z_j = 0$, in which case we should have an aggregate level regression relationship that corresponds to Equation (4.2). This implies that even when the Goodman assumption of the constant parameter holds, a Goodman regression does not exist when it comes to estimating the impact of covariates that are correlated with the main independent variable.

4.2.2 Revisiting the Thomsen Estimator

Extending the Thomsen model to have covariates is not straightforward. In this section, I will reinterpret the Thomsen estimator to be a method of moment estimator and develop an extension of it that allows the inclusion of covariates into the model.

In Chapter 2, it was shown that the Thomsen model is based upon the individual-level utility of voting for a given party in two elections which are assumed to distribute joint-bivariate normal in the population, with means μ_x and μ_y and correlation ρ between them:

$$f(x_i^*, y_i^*) = \mathcal{N}^2(\mu_x, \mu_y; \rho). \quad (4.4)$$

Voter decisions are simply a function of these random variables. If the utilities to vote for a given party in the two elections, $x_i^*$and y_i^*, exceed a threshold, which can be set at zero for identification, the voter will cast her ballot for the party; otherwise, she will vote for the opposing party. This is a usual probit setup where voting for a given party in elections can be denoted as indicator variables, say, x_i for election 1 and y_i for election 2.

$$\begin{aligned} x_i = 1 \text{ if } x_i^* > 0, \quad \text{otherwise,} \quad x_i = 0 \\ y_i = 1 \text{ if } y_i^* > 0, \quad \text{otherwise,} \quad y_i = 0. \end{aligned} \tag{4.5}$$

Then, quite simply, we can see that our quantity of interest can be retrieved by evaluating

$$\begin{aligned} \Pr(x_i = y_i = 1) &= \Pr(x_i^* > 0 \text{ and } y_i^* > 0) \\ &= \int_0^\infty \int_0^\infty \mathcal{N}^2(x_i^*, y_i^* | \mu_x, \mu_y, \rho) dx_i^* dy_i^* \end{aligned} \tag{4.6}$$

where the double integration would estimate the probability that a voter supports a given party in both elections. The task then would be to estimate the three parameters of the bivariate-normal distribution, namely, the means of the two utility distributions, μ_x, μ_y, and their correlation, ρ.

First of all, since the μ_x is the mean of the normally distributed x_i^*, we may write from Equation (4.5) as

$$\begin{aligned} E\left[\Pr(x_i = 1)\right] &= \int_0^\infty \mathcal{N}(x_i^* | \mu_x, 1) dx_i^* \\ &= \int_{-\infty}^{\mu_x} \phi(z) dz \\ &= \Phi(\mu_x) \end{aligned}$$

where $\phi(z)$ denotes the standard normal density function and $\Phi(z)$ is the cumulative standard normal function. In a similar fashion, we may define $E\left[\Pr(y_i = 1)\right] =$

$\Phi(\mu_y)$.

It is straightforward to retrieve these two population parameters from the sample moments, since $E\left[\Pr(x_i = 1)\right]$ and $E\left[\Pr(y_i = 1)\right]$ are the respective first moments of the two binary variable, that is, the sample means. Note that this quantity can directly be retrieved by the weighted averages of the aggregate data:

$$\begin{aligned} \Phi(\mu_x) = E\left[\Pr(x_i = 1)\right] &= \frac{\sum_i^N x_i}{N} = \frac{\sum_j X_j n_j}{N} \\ \Phi(\mu_y) = E\left[\Pr(y_i = 1)\right] &= \frac{\sum_i^N y_i}{N} = \frac{\sum_j Y_j n_j}{N} \end{aligned} \tag{4.7}$$

where X_j and Y_j denote the proportion of party supporters in district j and n_j is the number of total voters in the district. N is the total number of national votes, so $\sum_j n_j = N$. The parameters of interest, $\mu_x$and μ_y can be retrieved by inverting the probit function. Note that even though the actual estimates will be computed from the aggregate data, the quantities exactly are the first moments of individual level samples. In short, there is no cross-level inference, yet.

The third parameter, the correlation coefficient, can be written in terms of the sample moments as well, although inestimable because they are not observed directly:

$$\rho = Corr\left[x_i^*, y_i^*\right]. \tag{4.8}$$

First, add a subscript j to the variables to indicate the district to which the voter belongs: then our task is to derive $Corr\left[x_{ij}^*, y_{ij}^*\right]$. Note that the aggregate-level information we have is in the form of

$$\begin{aligned} X_j = E(x_{ij}|j) &= \frac{\sum_i^{n_j} x_{ij}}{n_j} \\ Y_j = E(y_{ij}|j) &= \frac{\sum_i^{n_j} y_{ij}}{n_j} \end{aligned}$$

where each $\Phi^{-1}(X_j)$ and $\Phi^{-1}(Y_j)$ are the first sample moments of x_{ij}^* and y_{ij}^*

Individual		Aggregate	
x_i and y_i	Binary variables for vote choices in time 1 and time 2	X_j and Y_j	Corresponding aggregate supports in District j
z_i	Binary variable indicating an individual attribute	Z_j	Corresponding aggregate fraction of such voters in District j
x_i^* and y_i^*	Underlying individual utilities to vote for a given party that determines x_i and y_i. $\Pr(x_i = 1) = \Phi(x_i^*)$ and $\Pr(y_i = 1) = \Phi(y_i^*)$.	X_j^* and Y_j^*	Corresponding aggregate utilities of voters in district j to vote for a given party. Estimated by $\Phi^{-1}(X_j)$ and $\Phi^{-1}(Y_j)$.

Table 4.1: Key Variables in the Extended Thomsen Model with a Covariate

within district j, respectively. According to Thomsen, these can be thought of as means of district-level distributions of individual utilities to vote for a given party.

The correlation between the two aggregate variables, the probit transformed $Corr\left[X_j^*, Y_j^*\right]$ or $Corr\left[\Phi^{-1}(X_j), \Phi^{-1}(Y_j)\right]$, can replace equation (4.8), when the ratio of the systematic versus the non-systematic parts of x_i^* and y_i^* equal the same ratio in X_j^* and Y_j^* respectively. A more detailed discussion on the assumption can be found in the earlier section that derives the condition in Equation (2.17). If we can establish that the probit transformed variables distribute normal, standard estimates from its sample moments and correlations will be consistent, converging in probability (Greene 2005).

Before I proceed, I will define key variables and parameters involved in the extended Thomsen model with covariates. Table 4.1 provides summary of the variables at both individual and aggregate levels: in general, individual manifest variables, such as x_i and y_i, are binary variables, while their counterparts are the aggregated fraction of them at district j and are conventionally written in upper case.

4.2.3 Extending the Thomsen Estimator

A. Two Joint-Bivariate Normal Distributions

As was shown in the previous section, the Thomsen estimator assumes that the two key latent variables that underlie the electoral choice is joint-bivariate normally distributed. To add a covariate, we can think about the variables, (x_i^*, y_i^*), as draws from a mixture of two joint-bivariate distributions. More specifically, the probability density function of the two latent variables that capture voters' utility to support a given party in the two elections can be defined as:

$$f(x_i^*, y_i^* | z_i) \quad = \quad z_i \mathcal{N}_1^2 \left[\mu_{x,1}, \mu_{y,1}, \rho_1\right] + (1 - z_i)\mathcal{N}_0^2 \left[\mu_{x,0}, \mu_{y,0}, \rho_0\right]. \qquad (4.9)$$

In other words, if a voter belongs to a particular group of interest—say, if she is a worker—, then her latent utilities to vote for the party in the two elections is drawn from the first bivariate normal distribution, $\mathcal{N}_1^2$; otherwise, they are drawn from the second distribution, $\mathcal{N}_2^2$. There are six parameters to estimate to compute the transition rates in the two groups, three from each distribution.

One important point to make here is that such demographic variables, say, proportion of workers or the gender distribution, has a radically different theoretical interpretation from the vote shares within the context of the Thomsen model. As was argued in Chapter 2, and sketched in the previous section, voting for a given party in elections is set up as an outcome of an underlying partisanship of a voter. In fact, neither y_i or x_i are exogenous to each other—Thomsen is only interested in studying the correlation between the two. On the contrary, many types of covariates, such as demographic variables, are truly exogenous variables that are measured directly. In the model above, they are accordingly set up as strictly exogenous factors that determine to which (joint-normal) distribution the sample

belongs. Our task then is to study how different the two distributions are and use them to find out the transition rates in the two different demographic groups.

The key variables x_i^* and y_i^* are latent utility variables and if they were observed at the individual-level, we should be able to estimate directly the population parameters with the sample moments. However, since we do not have that information, it is necessary to take a detour and estimate auxiliary parameters that will help us study the two distributions. Since the voter preferences are distributed joint-bivariate normal in each group, it is possible to write the following regression relationships that accompany the two distribution functions:

$$y_i^* = \alpha_1 + \beta_1 x_i^* + u_{i,1} \qquad \text{if } z_i = 1$$

$$y_i^* = \alpha_0 + \beta_0 x_i^* + u_{i,0} \qquad \text{if } z_i = 0.$$

In a similar fashion that was shown in Equation (4.7), we may equate the first population moments to the sample means of probabilities to vote:

$$\Phi(\hat{\mu}_{y,1}) = E\left[\Pr(y_i = 1 | z_i = 1)\right] = E\left[\Phi\left(\alpha_1 + \beta_1 x_i^* + u_{i,1}\right)\right] \tag{4.10}$$

$$\Phi(\hat{\mu}_{y,0}) = E\left[\Pr(y_i = 1 | z_i = 0)\right] = E\left[\Phi\left(\alpha_0 + \beta_0 x_i^* + u_{i,0}\right)\right] \tag{4.11}$$

The significance of the above expression is that once we estimate the auxiliary parameters, β's and α's, we should be able to get estimates of $\mu_{y,1}$ and $\mu_{y,0}$ in terms of the function of x^* and x^* only. Most importantly, as can be seen in the right hand side of the above equations, we do not have to worry about the covariate, z_i, any more once we collect the the auxiliary parameters.

In a symmetric fashion, write the reversed regression relationship with x_i^* as the dependent variable and y^* as the independent variable such as

$$x_i^* = \delta_1 + \gamma_1 y_i^* + e_{i,1} \qquad \text{if } z_i = 1$$

$$x_i^* = \delta_0 + \gamma_0 y_i^* + e_{i,0} \qquad \text{if } z_i = 0$$

This may look counter-intuitive to have an earlier election as the dependent variable and a later election as the independent variable, but since the model does not claim any direct causal relationship between the two electoral variables—and they are only linked by a correlation—there is nothing that prohibits this. A different way to understand this is to say that it is the joint-distribution parameters of x_i^* and y_i^* that matter for our purpose, and regression coefficients are just auxiliary parameters to help us study the distributions. Thus, we write

$$\Phi(\hat{\mu}_{x,1}) = E\left[\Pr(x_i = 1|z_i = 1)\right] = E\left[\Phi\left(\delta_1 + \gamma_1 y_i^* + e_{i,1}\right)\right] \tag{4.12}$$

$$\Phi(\hat{\mu}_{x,0}) = E\left[\Pr(x_i = 1|z_i = 0)\right] = E\left[\Phi\left(\delta_0 + \gamma_0 y_i^* + e_{i,0}\right)\right] \tag{4.13}$$

The question now is how to estimate the four μ parameters with aggregate data. Take, the example of Equation (4.10). Once we have estimates of the auxiliary parameters, the expected value of the right-hand-side expression can be written as the following and can be estimated by aggregate data.

$$\begin{aligned}
\Phi(\hat{\mu}_{y,1}) &= E\left[\Phi\left(\hat{\alpha}_1 + \hat{\beta}_1 x_i^* + \hat{u}_{i,1}\right)\right] \\
&= E\left[\Phi\left(\hat{\alpha}_1 + \hat{\beta}_1 \Phi^{-1}\left[\overline{\Phi(x_i^*)}\right]\right)\right] \\
&= \frac{\sum_j \Phi\left(\hat{\alpha}_1 + \hat{\beta}_1 X_j^*\right) Z_j n_j}{\sum_j Z_j n_j}
\end{aligned} \tag{4.14}$$

The idea is to replace the individual utilities, x_i^* in district j, with X_j^*, the probit-transformed district vote shares of the party which are assumed to be the district means of aggregate utilities. Even though this is the solution I have at the current stage, it is not the perfect solution: the estimates will be slightly biased since the probit of means ($\Phi^{-1}\left[(\overline{\Phi(x_i^*)}\right]$) is not always the mean of probit-transformed variables. ($\overline{x_i^*}$). [1] However, it is not an unreasonable solution, as similar replace-

[1] There also is another source of bias in the equation. Since the expected value of the error term inside the cumulative normal function will not be zero in general, the estimates will slightly be biased.

ment was done in the simple Thomsen model.

The last expression shows that this can be computed by the weighted average of the aggregate sample. Note that expected turnout in district j is weighted by the number of the population group in the district, $Z_j n_j$. Following similar lines of logic, the rest of the necessary mean parameters of the two joint-bivariate normal distributions can be derived as:

$$\Phi(\hat{\mu}_{y,0}) = \frac{\sum_j \Phi\left(\hat{\alpha}_0 + \hat{\beta}_0 X_j^*\right)(1 - Z_j)n_j}{\sum_j Z_j n_j} \tag{4.15}$$

$$\Phi(\hat{\mu}_{x,1}) = \frac{\sum_j \Phi\left(\hat{\delta}_1 + \hat{\gamma}_1 Y_j^*\right) Z_j n_j}{\sum_j Z_j n_j} \tag{4.16}$$

$$\Phi(\hat{\mu}_{x,0}) = \frac{\sum_j \Phi\left(\hat{\delta}_0 + \hat{\gamma}_0 Y_j^*\right)(1 - Z_j)n_j}{\sum_j Z_j n_j} \tag{4.17}$$

Now to estimate the correlation parameter for the two joint distributions, we may write:

$$\hat{\rho}_1^2 = \hat{\beta}_1 \cdot \hat{\gamma}_1 \tag{4.18}$$

$$\hat{\rho}_0^2 = \hat{\beta}_0 \cdot \hat{\gamma}_0 \tag{4.19}$$

noting that the product of two reversed bivariate slope regression coefficients take the form of $\frac{Cov(x,y)}{Var(x)} \times \frac{Cov(x,y)}{Var(y)} = Corr^2(x,y)$. These complete the full two sets of parameters that are necessary to evaluate the two bivariate normal distributions. Quandt and Ramsey (1978) analyzed the problem of estimating parameters of such mixtures of normal distributions, and showed that the method of moment estimates are consistent. Estimates of the transition rates in the two different population groups are continuous functions of those consistent estimators, and hence consistent.

So far, I outlined how to estimate parameters of the two joint bivariate normal distributions from which the two groups of voters are modeled to be sampled. The

six equations, (4.14)-(4.19), provide formulas to estimate the parameters, using auxiliary regression parameters. Now the task left is to show how to estimate them.

B. Estimating the Auxiliary Parameters

The auxiliary regression models can be incorporated into the following two sets of nonlinear equations:

$$\Pr(y_i = 1|z_i) = z_i\Phi\left[\alpha_1 + \beta_1 x_i^* + u_{i,1}\right] + (1 - z_i)\Phi\left[\alpha_0 + \beta_0 x_i^* + u_{i,0}\right] \quad (4.20)$$

$$\Pr(x_i = 1|z_i) = z_i\Phi\left[\delta_1 + \gamma_1 y_i^* + e_{i,1}\right] + (1 - z_i)\Phi\left[\delta_0 + \gamma_0 y_i^* + e_{i,0}\right] \quad (4.21)$$

These equations are not estimable directly, since we do not have individual-level information, and the latent variables are unobservable. A good starting point is to sum the equations using the observed exogenous variable, z_i.

Now suppose within district j, the probability to draw a voter belonging to a demographic group of, say, the workers, will equal Z_j, then we may replace z_i with Z_j, and write

$$\Pr(y_i = 1|j) = Z_j\Phi\left[\alpha_1 + \beta_1 x_i^* + u_{i,1}\right] + (1 - Z_j)\Phi\left[\alpha_0 + \beta_0 x_i^* + u_{i,0}\right] \quad (4.22)$$

$$\Pr(x_i = 1|j) = Z_j\Phi\left[\delta_1 + \gamma_1 y_i^* + e_{i,1}\right] + (1 - Z_j)\Phi\left[\delta_0 + \gamma_0 y_i^* + e_{i,0}\right] \quad (4.23)$$

Taking the expectations within district j, we have

$$Y_j = Z_j E\left(\Phi\left[\alpha_1 + \beta_1 x_i^* + u_{i,1}\right]\right) + (1 - Z_j)E\left(\Phi\left[\alpha_0 + \beta_0 x_i^* + u_{i,0}\right]\right)$$

$$X_j = Z_j E\left(\Phi\left[\delta_1 + \gamma_1 y_i^* + e_{i,1}\right]\right) + (1 - Z_j)E\left(\Phi\left[\delta_0 + \gamma_0 y_i^* + e_{i,0}\right]\right)$$

Resorting to the same assumption that expectations of the cumulative normal

functions can be approximated by using district means, we may write

$$Y_j = Z_j\Phi\left[\alpha_1 + \beta_1 X_j^*\right] + (1 - Z_j)\Phi\left[\alpha_0 + \beta_0 X_j^*\right] + \varepsilon_{Y,j} \quad (4.24)$$

$$X_j = Z_j\Phi\left[\delta_1 + \gamma_1 Y_j^*\right] + (1 - Z_j)\Phi\left[\delta_0 + \gamma_0 Y_j^*\right] + \varepsilon_{X,j}. \quad (4.25)$$

Both equations can be estimated by non-linear least squares or maximum likelihood with the identifying conditions: $\alpha_0 \neq \alpha_1$ or $\beta_0 \neq \beta_1$ for Equation (4.24) and $\delta_0 \neq \delta_1$ or $\gamma_0 \neq \gamma_1$ for Equation (4.25). In the trivially special case where Z_j is not continuous and is a binary variable indicating whether the unit is, say, urban ($Z_j = 1$) or rural ($Z_j = 0$), the estimation is equivalent to collecting two sets of parameters in the simple transition problem without covariates, from two separate groups of geographic units.

With these results, it is now possible to estimate the voter transition model with covariates. In the following section, I demonstrate how the technique developed in this section can be applied to a real voter transition problem.

4.2.4 Estimation: The Thomsen Estimator with Covariates

In this section, I will provide a step-by-step example of the model developed in the previous section, using the example of the turnout rates in the South Korean elections in 1981-1985. The covariate here is the fraction of the more educated voters with college degrees or higher education. Using the model developed in the previous section, I will demonstrate that it is possible to estimate i) the turnout rates in the two different demographic groups and ii) the transition rates between the two elections in the two different groups.

1. (**Definition of Variables**) First, definitions of the variables are as follows:

   X_j: Turnout in the 1981 Election in District j;

X_j^*: The inverse probit transformation of X_j, that is $\Phi^{-1}(X_j)$;

Y_j: Turnout in the 1985 Election in District j;

Y_j^*: The inverse probit transformation of Y_j, that is $\Phi^{-1}(Y_j)$;

Z_j: Proportion of more educated voters, with in District j; and

n_j: The total number of eligible voters in district j.

2. (**NLS ESTIMATION**) Estimate the nonlinear regression relationships for the auxiliary parameters:

$$Y_j = Z_j\Phi\left[\alpha_1 + \beta_1 X_j^*\right] + (1 - Z_j)\Phi\left[\alpha_0 + \beta_0 X_j^*\right] + \varepsilon_{Y,j}$$

The estimation can be carried out by maximum likelihood or nonlinear least squares. Here, I employ nonlinear least squares and the estimated equation is:

$$Y_j = Z_j\Phi\left[.671 + .281 X_j^*\right] + (1 - Z_j)\Phi\left[.564 + .496 X_j^*\right] + \varepsilon_{Y,j}$$

3. (**PROJECTED TURNOUTS OF THE TWO GROUPS IN THE SECOND ELECTION**) The next step is to compute the predicted probabilities to turnout in the two groups.

$$E(Y_j|z_i = 1) = \frac{\sum_j \Phi\left[.671 + .281 X_j^*\right] Z_j n_j}{\sum_j Z_j n_j} = .850$$

$$E(Y_j|z_i = 0) = \frac{\sum_j \Phi\left[.564 + .496 X_j^*\right] (1 - Z_j) n_j}{\sum_j (1 - Z_j) n_j} = .810$$

Note that the predicted probabilities in the two population groups are weighted by the group sizes. The estimates indicate that around 85% of the more educated voters turned out in the 1985 election, while the turnout rate among other voters is estimated to be around 81%.

These two estimates constitute the first set of parameters that define the means of the two joint-bivariate normal distributions

$$\begin{aligned}\hat{\mu}_{y,1} &= \Phi^{-1}(.850) = 1.036 \\ \hat{\mu}_{y,0} &= \Phi^{-1}(.810) = .878.\end{aligned}$$

4. (**REVERSED NLS ESTIMATION**) The next parameters to estimate are the corresponding population means of the first election. Estimate the reversed nonlinear regression such as

$$X_j = Z_j\Phi\left[\delta_1 + \gamma_1 Y_j^*\right] + (1 - Z_j)\Phi\left[\delta_0 + \gamma_0 Y_j^*\right] + \varepsilon_{X,j}$$

which yields the following results:

$$X_j = Z_j\Phi\left[.055 + .043Y_j^*\right] + (1 - Z_j)\Phi\left[-.135 + 1.064Y_j^*\right] + \varepsilon_{X,j}$$

5. (**PROJECTED TURNOUTS OF TWO GROUPS IN THE FIRST ELECTION**) Again, compute the estimated projection of turnout in the two groups in the *first* election:

$$\begin{aligned}E(X_j|z_i = 1) &= \frac{\sum_j \Phi\left[.055 + .043Y_j^*\right] Z_j n_j}{\sum_j Z_j n_j} = .611 \\ E(X_j|z_i = 0) &= \frac{\sum_j \Phi\left[-.135 + 1.064Y_j^*\right] (1 - Z_j) n_j}{\sum_j (1 - Z_j) n_j} = .793\end{aligned}$$

Note that the turnout in the group that consists of more educated voters are significantly lower than that in the other group. With these results, we may estimate the second set of parameters,

$$\begin{aligned}\hat{\mu}_{x,1} &= \Phi^{-1}(.611) = .282 \\ \hat{\mu}_{x,0} &= \Phi^{-1}(.793) = .817.\end{aligned}$$

6. (**CORRELATION COEFFICIENTS**) The final parameters are the correlations in the two distributions, which can be retrieved by the multiplication of the two slope coefficients, β's and γ's as discussed before:

$$\begin{aligned}\hat{\rho}_1 &= \sqrt{\hat{\beta}_1\hat{\gamma}_1} = \sqrt{.281 \times .043} = .165 \\ \hat{\rho}_0 &= \sqrt{\hat{\beta}_0\hat{\gamma}_0} = \sqrt{.496 \times 1.064} = .726.\end{aligned}$$

As we can see, the correlation is much smaller for the group with higher education which indicates that more electoral change is happening in the group.

7. (**SUMMING UP THE TWO JOINT DISTRIBUTIONS**) This finishes estimating necessary parameters from the two bivariate-normal distributions as:

$$\begin{aligned}\text{Educated Voters:} \quad \mathcal{N}_1^2\left[\hat{\mu}_{x,1}, \hat{\mu}_{y,1}; \hat{\rho}_1\right] &= \mathcal{N}_1^2\left[.282,\ 1.036;\ .165\right] \\ \text{The Rest:} \quad \mathcal{N}_0^2\left[\hat{\mu}_{x,0}, \hat{\mu}_{y,0}; \hat{\rho}_0\right] &= \mathcal{N}_0^2\left[.817,\ .878,\ .726\right]\end{aligned}$$

8. (**FRACTION THAT VOTED IN BOTH ELECTIONS**) Now the task left is to find the fraction of voters in the two groups that voted in both elections. We can compute the double integrals of the estimated distributions to find the fraction of voters whose utilities exceed the threshold of zero in both elections.

$$\begin{aligned}\int_0^\infty \int_0^\infty \mathcal{N}_1^2\left[x^*,\ y^* | .282,\ 1.036;\ .165\right] dx^* dy^* &= .529 \\ \int_0^\infty \int_0^\infty \mathcal{N}_0^2\left[x^*,\ y^* | .817,\ .878,\ .726\right] dx^* dy^* &= .718.\end{aligned}$$

To reiterate, 53% of educated voters in South Korea voted in both elections, while 72% of the other voters turned out in both elections.

9. (**FINISHING UP THE TABLE**) Now fill in the fraction of the voters in the following three-way table, using the estimated total turnouts in each group.

		Highly Educated The 1981 Election			The Rest The 1981 Election		
		Voted	Not Voted	Total	Voted	Not Voted	Total
The 1985 Election	Voted	***0.529***	0.321	***0.850***	***0.718***	0.092	***0.809***
	Not Voted	0.082	0.068	0.150	0.075	0.115	0.191
		0.611	0.389		***0.793***	0.207	

Table 4.2: Estimated Distribution of Voters across Elections in Different Education Groups: South Korean Elections 1981–1985

Note that the estimated quantities we already have are emphasized in the table. For example, we have estimated turnout rates in the two groups in the two election from Steps 3 and 4, which will enable us to define the marginal probabilities in the table. Also, since we have estimated the fraction of two-time voters in Step 8, the table is identified, and we may fill in the table in an obvious way.

10. (**Loyalty and Defection Rates: The Conditional Probabilities**) Since loyalty rates and defection rates as defined in previous chapters are conditional probability terms, that is, the probability that a first time voter will return to the polling booth in the next election, the fractions can trivially be transformed into such rates. By dividing the fractions with the marginal turnout rates in the 1981 election, the voter transition rates can be retrieved and are shown in Table 4.3:

Among other things, Table 4.3 makes an interesting point on the estimated rate of more educated voters newly entering the election in 1985. It tells that around 83% of more educated non-voters in the previous election newly turned out in the 1985 election, perhaps because they were more responsive to the electoral climate than the rest of the voters. The rest of the chapter looks into the question of how

		Highly Educated The 1981 Election		The Rest The 1981 Election	
		Voted	Not Voted	Voted	Not Voted
The 1985 Election	Voted	0.866	0.825	0.905	0.444
	Not Voted	0.134	0.175	0.095	0.556

Table 4.3: Estimated Transition Rates in Different Education Groups: South Korean Elections 1981–1985

the democratization movement in South Korea made an impact on the electoral participation of the voters by examining aggregate data and applying the model I have developed so far.

4.3 Application: The Impact of Democratization on Voter Turnout

4.3.1 Introduction

The year 1987 marks an important point in time in South Korean political history. After a series of authoritarian governments that ruled the country from the 1960s, the government conceded to the massive protest and wide-spread demand of the citizens for democratic changes, including restoration of civil rights, amendment of the constitution to ensure more democratic elections, and other political reforms (Han 1988; Oh 1999). The presidential election that was held at the end of the year was the first direct, competitive contest since the 1972 election.

If there ever was an election in South Korea that can be called a "critical" election in the same sense as V. O. Key (1955) defines, it is the 1987 election. Elections afterward showed voting patterns that were not existent before: including massive levels of regional voting pattern and high levels of partisan votes. In short, the democratization induced a critical electoral realignment.

Realignment, simply defined, is a systematic and enduring change in electoral

preferences of the electorate over time. To study realignments, the researcher should naturally look into the change in partisanship or electoral support of voters. However, an equally, if not more, important aspect of electoral realignments is whether and why new voters are mobilized into the political arena and who they are. Change in the electoral environment or an issue that newly becomes salient may attract non-voters to the polling both, and if such new voters are systematically favoring one party over other parties, they will constitute a central component of the electoral realignment.

In this vein, the dynamics of voter participation takes on an important meaning in understanding the electoral politics in South Korea. The democratic reform of the country—a system-level change with a profound impact on the political life of the voters and politicians—should entirely alter how people perceive politics and how political parties mobilize them: that is, who starts to vote and whom they vote for.

Authoritarian regimes will alienate certain groups of citizens, and as the literature on participation and political efficacy suggests, we may expect a systematic influx of such voters into the electoral arena when such regimes reach their demise. Democratic reform however may also trigger the vanishing of traditional mobilization mechanisms. This will necessarily result in the decrease of turnout, and more importantly, the weakening of support for candidates and parties who have benefited from such mobilization networks and connections of the authoritarian government.

This section focuses on how the democratic reform in South Korea attracted voters that were formerly uninterested in the authoritarian regime, and how such voters are aligning into the new party system. Of course, it is not to say that such

influx of new voters is the one and only cause of the emergence of the new party system, but I argue that it constitutes a major part of the electoral realignment. In this chapter, I simply try to answer the following questions: first, did the democratization of the country attract new voters into the political arena? If so, who are they, and what are the consequences of such new voters? And finally, what caused such systematic movement in the electorate?

The following section sets up the background on the dynamics of turnout in South Korea. A simple adaptation of the voter transition setup will be applied to model the entrance and exit rates of the voters.

4.3.2 Background: The Dynamics of Voter Turnout in South Korea

The dynamics of voter turnout in South Korea, especially focusing upon the elections around 1987, shows that there are several complex story lines. Figure 4.2 shows the overall trend of national voter turnout in presidential, legislative, and local elections.

First, it is quite obvious that there is a short-term effect of democratization on turnout rates: the turnout rate in the 1987 presidential election, which is around 89%, is higher than any previous election since the 1970s; turnout in the National Assembly elections peaks in 1985—a widely contested election that triggered the democratization movement. However, it should be noted that turnout decreased by about 10 percentage points in the 1988 National Assembly election that took place a year after democratization. There is also a long-term trend of decline in turnout after the democratization of the country.

As the literature on political participation finds a positive relationship between

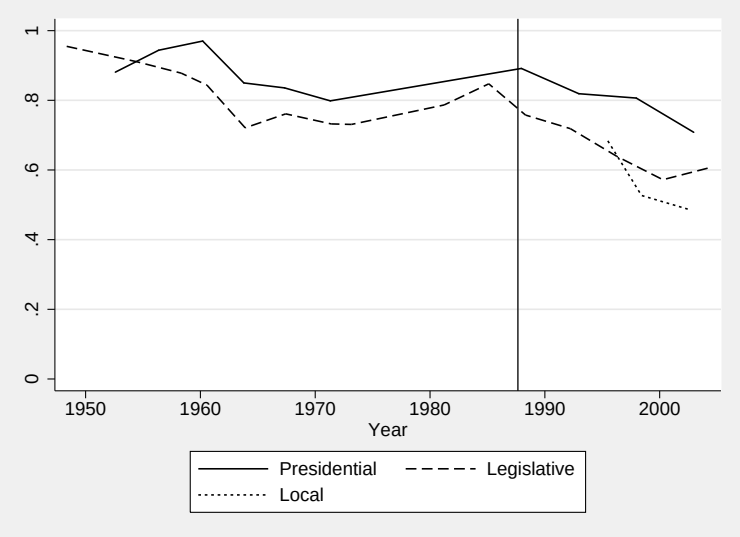


Figure 4.2: Turnout in South Korean Elections

political efficacy and participation, the democratization of a political system should facilitate higher turnout. The literature is quite clear on the relationship between participation and political efficacy: alienated voters are less likely to vote. For example, Luttbeg and Gant (1995: 134–136) distinguish between internal and external efficacies, and highlight that both strongly contribute to the voter turnout. The case in point here is that we may expect an increase in turnout in a more democratic regime (external efficacy), and especially from internally efficacious voters with more political resources.

Meanwhile, as the literature on voter mobilization and participation claims, there are reasons to expect a decrease in participation after the democratization of the country, since the ruling party lost its edge in mobilizing voters. (For example, see Rosenstone and Hansen 2002.) More specifically, under the authoritarian regime, there was a strong connection between local wards of governing parties and local governmental offices, which enjoyed uncontested and effective voter mobilization, usually in the form of strong turnout drives. Abolition of such prac-

tices, or guarantee of the neutrality of local governmental offices, was one of the main focuses in the discussion of democratic reform, and by the 1987 presidential election, old mechanisms of voter mobilization would not work for the ruling Democratic Justice Party.[2]

In short, there is no simple trend in turnout rates after democratization that we may expect at the national level. It is fair to summarize that the change in the mobilization patterns and the level of political efficacy affects turnout in opposite directions and cannot easily be separated although we may postulate that there is a short term increase and a long-term decline. And of course, there is no individual-level survey data from this period to look into this question.

4.3.3 Examining Entrances and Exits

Figure 4.3 depicts the possible consequences of changes in voter turnout. In times of changes we would observe higher rates of voter replacement, which could be defined by the proportion of voters who newly enter or do not anymore come to the polling booth. The apparent questions are: who exactly are "entering" and "exiting" the polling booth? ; and what impact does this have on the ensuing party system?

		BEFORE (Time 1)	
		Vote	Not Vote
AFTER (Time 2)	Vote		*"Enter"*
	Note Vote	*"Exit"*	

Figure 4.3: Replacement of Voters

[2]See *Donga Ilbo* [*Donga Daily*] (Feb. 1985) on the turnout drive lead by the government. Due to the accuracy of the citizen census register that the government holds, it was possible to identify exactly who voted and who did not. Also, see the discussion on the change in the organization of local party wards.

A direct adaptation of the voter transition model will allow us to investigate the turnout rates in the two elections before and after 1987. We may rewrite the Thomsen ecological estimation function as

$$T_{1987,j} = \Phi\left[\alpha + \beta\Phi^{-1}(T_{1985,j})\right]$$

where T_{1987} and T_{1985} indicate turnouts in the two consecutive elections. The "entrance" rate will be the usual defection rate and the "exit" rate will be one minus the usual loyalty rate.

		1985 Election	
		Vote	Not Vote
1987 Election	Vote	0.878	0.782
		(0.875, 0.881)	(0.768, 0.795)
	Not Vote	0.122	0.218
		(0.119, 0.125)	(0.205, 0.232)
			$N = 3922$

Note: Aggregate turnout rates in the two elections were 81% (1985) and 86% (1987).

Table 4.4: Voter Transition Rates Around Democratization

Table 4.4 shows the estimated entrance and exit rates retrieved by the Thomsen estimator. Entries in the parentheses represent the 95% confidence interval of the estimates, as was defined in Achen (2000) and discussed in the previous chapter. We see that about 12% of the previous voters did not vote in the 1987 election, while 78% of the previous non-voters came to the polling booth in the election. If the causes that systematically determine who votes and who does not are working in both elections, we should see a clear pattern of consistently high coefficients on the diagonal entries: voters will continue to vote, and non-voters will stay that way. Of course, since this is a time when a major electoral earthquake is taking place, we see a sizable proportion of the previous non-voters coming to vote in the first election after democratization.

Table 4.5 reports such entrance and exit rates in multiple elections in South Korea over time. The table provides an overlook of the composition of the voter turnout in the elections and will be a baseline estimation for the analysis in the next section.

First of all, consider the entrance rates for legislative elections over time. It becomes apparent that there is a significant number of new voters coming into the electoral system in the mid-1980s and into the early 1990s. Also, we note that the exit rates are kept quite small until the 1988 election, meaning that once the voter turns out, she mostly stayed a voter. This trend changes somewhat in the later elections where we see a significant number of voters consistently exiting the elections.

Another important point we observe is that the initial transformation in the pattern of who votes and who does not first started in the 1985 election, not after 1987. This is consistent with what was shown in Figure 4.2 and with what we know about the 1985 election, which featured a strong opposition party fielding competitive candidates that was a rarity in South Korean elections for many years. In a way, the results show that the democratization movement in 1987 was not a political episode that suddenly erupted, but there already was an underlying movement in the electorate that showed a surge in turnout.

Transition rates between presidential elections are less subtle. In the first two elections right after 1987, the majority of previous non-voters turned out in the elections. As it reaches down to more recent years, the entrance rate dropped down while the exit rate climbed up. In any case, the presidential elections picture a more volatile electorate where moving in and out of the polling booth seems more frequent.

Legislative			Presidential		
Year	Enter	Exit	Year	Enter	Exit
1978	0.439	0.131			
1981	0.341	0.102			
1985	0.504	0.077	1987	0.782	0.121
1988	0.321	0.171			
1992	0.385	0.167	1992	0.617	0.143
1996	0.219	0.210	1997	0.514	0.134
2000	0.192	0.206			

Note: Estimates for the 1987 Election was retrieved in comparison to the 1985 Election.

Table 4.5: Entrances and Exits from the Polling Booth

4.3.4 Unpacking the Entrances and Exits

A. Mobilization and the Turnout of Urban and Rural Voters

Based on our setup, we are ready to examine the entrance and exit rates in different demographic groups. Since we are interested in "identifying" the voters who are newly coming into and leaving from the electoral arena, it requires us to employ the technique developed earlier in this chapter.

Here I first start out by comparing the entrance and exit rates of the rural and urban voters. It should be noted that this round of estimation was carried out on two separate geographical samples—thus we may infer the rates as those of the urban voters and those of the rural voters. The important point to consider before examining the results in the table is the possible impact of the change in mobilization patterns. It can be assumed that former mobilization mechanisms were more concentrated in the rural areas, and we may see how the turnout patterns change over time there. Most importantly, we should observe lower exit rates and higher entrance rates in rural areas before 1987. After democratization, we can expect abrupt changes in such patterns—where we would start to observe higher exit rates and possibly lower entrance rates of the rural voter.

Table 4.6 presents estimated entrance and exit rates of the rural and urban vot-

	Year	Enter			Exit		
		Urban	Rural	Diff	Urban	Rural	Diff
Legislative	1978	0.448	0.614	-0.167	0.170	0.143	0.027
Elections	1981	0.366	0.576	-0.210	0.144	0.139	0.005
	1985	0.580	0.501	0.078	0.101	0.084	0.017
	1988	0.321	0.323	-0.001	0.187	0.159	0.028
	1992	0.496	0.401	0.095	0.237	0.112	0.125
	1996	0.535	0.309	0.226	0.379	0.188	0.190
	2000	0.372	0.236	0.136	0.342	0.176	0.166
Presidential	1987	0.793	0.710	0.083	0.124	0.116	0.008
Elections	1993	0.557	0.645	-0.088	0.140	0.145	-0.005
	1997	0.485	0.515	-0.030	0.123	0.141	-0.018

Table 4.6: Entrance and Exit Rates in Urban and Rural Districts

ers in several legislative presidential elections. Among many things in the table, we first observe larger entrance rates of the rural voters than their urban counterparts before the 1985 election—something that is consistent with the strong mobilization in the rural areas. This changes rapidly while we move down to later legislative elections where we see the entrance rates are almost always (except the case of the 1988 election) larger in urban areas. We observe that the rural entrance rate goes down slowly and never recovers to the level above the urban rates in legislative elections, after which it stabilizes. In other words, the recruitment of new voters is now far less effective in the rural areas than it once was.

The exit rates in legislative elections tell a slightly different story. Arguably, the exit rates in the two areas were comparable until the 1992 election, where the urban exit rates took off. Contrary to what we speculated—that the lack of mobilization would result in an immediate release of rural voters—it stays well below the level of urban exit rates. Still, we see a gradual increase in exit rates in rural areas, which may be the long-run effect of the changes in the mobilization pattern. In any case, the extraordinarily large exit rates in the more recent elections are directly related to the rapid decline of overall turnout in recent elections that was shown in Figure 4.2. The results above indicate that the decline in turnout is

heavily driven by the urban voters and less by rural voters.

To summarize, it seems to be the case that the impact of democratic reform—here, essentially the cut-off of previous mobilization structures—had an immediate impact on how new voters were recruited in the rural areas. Democratization did not immediately drive away the voters: but gradually and in a couple of elections voters were moving away from the polling booth at an astonishing rate, especially in urban areas. An overall comparison between urban and rural voters in legislative elections reveal that the urban turnout is always much more volatile then that of the rural voters. Especially in post-democratization elections, both the entrance rates and exit rates are larger in urban areas, indicating more urban citizens are coming in and out of the polling booth, making them less consistent voters.

Presidential elections tell a different story. In both areas, the movement of the voters are volatile with larger entrance rates, and there are not many distinguishable patterns in the three elections, perhaps except the fact that the urban-rural difference diminishes in these elections. Again, this could be due to the fact that we only have three observation points, all of which being "high-profile" elections. But we may learn from a non-finding here: when stakes are high, urban and rural voters behave similarly. Now we look into other demographic attributes.

B. Estimated Turnout Among Different Groups

While the previous section was able to address the question of political mobilization by dividing the observations into different types of geographic units, the question of political efficacy will be more complicated. For example, it is possible to argue that the particular type of mobilization under the authoritarian regimes in South Korea was targeted at specific types of geographic areas, while attributes

such as voter efficacy are truly individual-level characteristics. It would require a full-fledged individual level election study that investigates deep into the psyche of the voters to get a good sense of the interaction between voter efficacy and turnout in South Korea. Additionally, since we are interested in the dynamic aspect of the relationship around the democratization period, a perfect data set to address the question would consist of at least several panels around 1987. In this section, we approach the problem with aggregate demographic information and try to depict the impact of democratization on turnout through voter efficacy.

Ecological estimates shown in this section require even stronger assumptions to hold true, and sometimes, the estimations are less than stable. As was sketched and discussed in previous sections, there is still more work to be added for the improvement of the estimator. That said, the following analysis will show that the ecological inference strategy developed in this chapter can be used here to address the question of the impacts of covariates (demographic information) on turnout, and more specifically, on entrance and exit rates.

Figure 4.4 is a good place to start. Entries are estimated turnout rates of selected demographic groups, which was retrieved by the sample predictions shown in equation (4.14) after estimating equation (4.24) with nonlinear least squares. The exact figures are available in the appendix at the end of this chapter. The covariates (Z_j) here would indicate the demographic composition of the unit, say, proportion of young voters. Essentially, the estimates are predicted turnouts of homogeneous groups—for example, young voters—and will provide insights into how one can look at the dynamics of turnout in South Korean legislative elections over time.

The overall turnout rate is similar to what was shown in Figure 4.2, even though

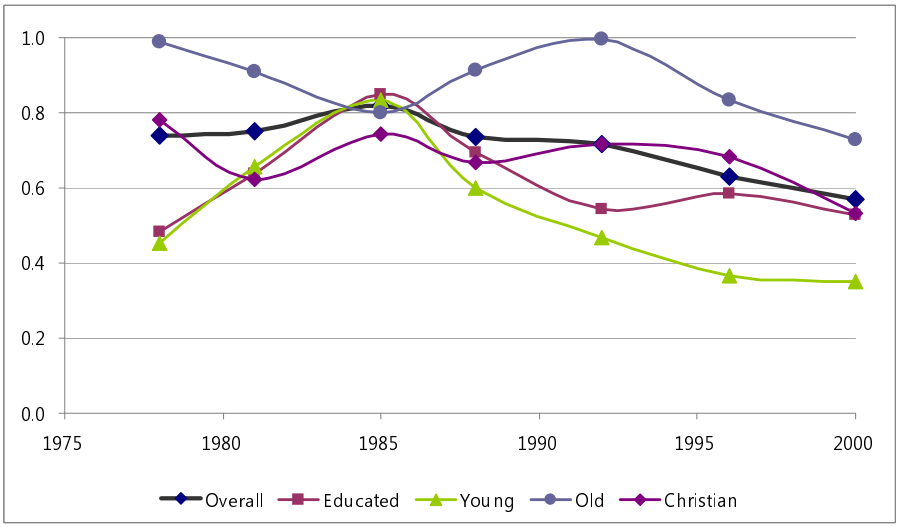


Figure 4.4: Estimated Turnouts in Legislative Elections, Selected Demographic Groups

the turnouts here are computed from the data directly—after the matching of the units over time and joining them to census units, the data had to go through a process of dropping some problematic units[3]. The rate hovers around mid 70% before the 1980s, and peaks at the 1985 election. After 1985, we see a continuous and rapid decline in turnout rates. The task here is to decompose the turnout in different demographic groups.

The demographic variables I investigate here are: 1) the percentage of young voters in their 20s and 30s at the time of the election; 2) the percentage of older voters who are in their 50s and above; 3) the percentage of voters with higher education defined by those with college or above education; and 4) the percentage of Christians.

The estimated turnout of high-education voters starts out with a fairly low level of estimated turnout—at around 48% which is about 25% lower than the national

[3]Joining geographic units from different sources of data sets will always create problems. For example, the boundaries can change over time or maybe are just defined differently. Merging adjacent units and aggregating them sometimes will help—cases that I was not able to salvage through this process were dropped. Also, sometimes it is the case that certain geographic units go through drastic population changes. Units that more than doubled in population and that lost more than half of their population were also dropped in the cleaning process.

turnout. At the time of the important 1985 election, the estimated turnout of voters with high education reaches its peak at around 85%: in fact, the estimated turnout is higher than that of any other demographic groups we have in the analysis. The turnout afterward declines and stays in the mid-50%.

A similar pattern can be found in the young voter category: it starts at the mid 40% level; peaks at the 1985 election; and rapidly dies down. The difference from the previous group of people is that the projected turnouts drop down even more radically where it hits the mid-30% mark by the 2000 election.

As was argued previously, these two groups of voters would have been the most alienated and/or uninterested voters under the authoritarian regime, and we can understand why they display the lowest estimated level of turnout before 1985. It is interesting to note both groups reach their peaks in the 1985 election: not just compared with their turnout level in other elections, but it marks the only time when the estimated turnout among young and highly educated voters record a higher turnout than the national average.

The fast decline—faster than other groups—in the estimated turnout also reveals an interesting point. In the short term, the 1988 legislative election which took place after the 1987 presidential would display a sharp decline of estimated turnouts among the young and highly educated voters. This could be due to the after-effect of the 1987 presidential election where the candidate from the traditional ruling bloc, Roh Tae Woo, would finally win the electoral contest flattening down all the heightened expectations. In any case, we see that the "critical" legislative election was the 1985 election, not the 1988 election.

In the long run, we see the continuous decline of turnout over time, especially among the young voters, since 1985. We note that the overall decline in turnout

is mainly driven by the low turnout of young voters which has consistently been dropping. The estimated turnout for the highly educated holds at around the mid-50%, which distinguishes it from that of the young voters. We will see the difference between the two demographic groups when it comes to looking at their respective entrance and exit rates.

An interesting pattern emerges when we look at the estimated turnout of the "old" voters. It essentially starts at around 98% in the 1973 election and remains to be the highest estimated turnout group except in the 1985 election. It reaches another peak in the 1992 election before showing some decline in later elections.

C. Voter Efficacy and The Dynamics of Turnout

So far, we have described the projected turnout rates among different groups of voters. In this section, we conclude our investigation into the impact of democratization on the turnout of South Korean voters by looking at the estimated exit and entrances rates of voters from different demographic groups.

As was constructed earlier, we want to study the entrance and exit rates among different demographic groups, and see how differently they reacted to democratization. As theory would suggest, we expect to observe larger entrance rates around democratization among those who who would have been politically alienated and who would feel politically more efficacious in the post-democratization era. Particularly, those with higher education (political resources) and the young voters (democratic values) would be a good example.[4]

Figure 4.5 plots the estimated entrance and exit rates of the three different demographic groups, namely, those with higher education, those in their 20s and

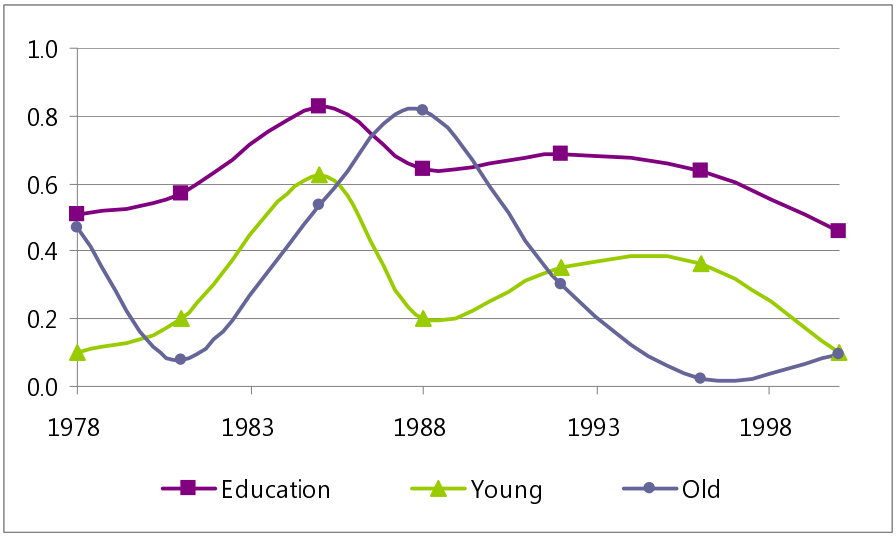
[4]See J. H. Rhee (2001) for the political efficacy arguments; Jung (2000) has an interesting discussion on the "Democratization Generation."

30s, and those who are over 50. The plots partially confirm what we expect, and at the same time provide some new insights. More detailed figures with other demographic variables are in the appendix that can be found at the end of this chapter.

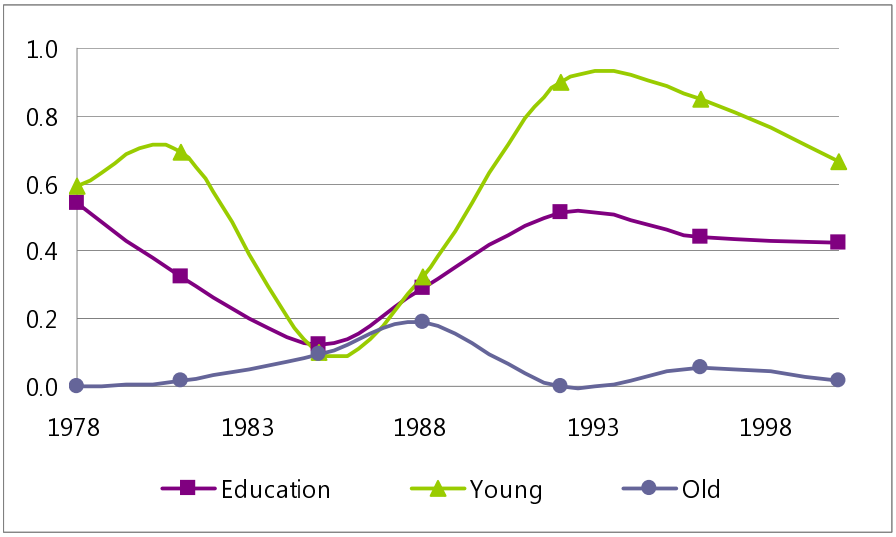
First of all, there is no question about whether there was a "shock" to the electoral system in 1985 that had a profound impact—at least a short-term effect. We observe surges of entrance rates in both the young voter group and the high-education group at the time; exit rates are the lowest for both groups as well.

Secondly, we are left to wonder whether this "shock" that was created by democratization had any lasting impact. Both groups of voters immediately go back to their pre-democratization levels of entrance and exit rates. One noticeable long-run trend is the drop in entrances and surge of the exit rates, especially among the young voters, but we cannot be sure that this is the direct impact of democratization. In any case, we note that the overall decline of turnout among the young voters that we have seen in Figure 4.4 in fact is a function of both the drop of the entrance rate and the surge of the exit rate in elections after the 1985 election.

Most interestingly, the impact of democratization can be seen in the older voter group as well. Generally, this is a group of people with very stationary turnout patterns, as can be seen by its minuscule exit rates in the second sub-figure of Figure 4.4: elderly voters consistently turn out to vote. The only time this is perturbed is in 1985 and 1988 which indicates that a significant amount of these voters stopped coming into the polling booth. Combine this with the sudden surge in the entrance rate of the "old" group around 1988. This indicates that a new batch of older voters was absorbed into the system while a significant chunk of them left in the two elections around democratization.



(a) Estimated Entrance Rates



(b) Estimated Exit Rates

Figure 4.5: Estimated Entrance and Exit Rates by Education and Age

4.3.5 Discussion

The above analysis on South Korean elections highlights several important points around the democratization period that have not been empirically investigated before. First, it was the 1985 National Assembly election, not the 1987 presidential election, that was the critical election to define the post-democratization realignment. In fact, it was centered around a certain group of voters that initiated the start of such immense political transformation through an election that took place two years before democratization.

More specifically, the group of voters with more political resources to participate—the more educated, the younger, those who live in urban areas, and most importantly, those who chose not to participate in previous elections under the authoritarian regime—suddenly decided to participate and were able to create a solid opposition against the government. An apparent explanation is that there were underlying political demands for more democracy all along, and it somehow solidified as a political force in the 1985 election and was instrumental in the democratization movement.

Second, consistent with the literature on electoral mobilization, this movement was later sustained by the institutional reforms in 1987. As seen in the analysis above, a long-term drop of turnout in rural areas with high exit rates is likely to be due to the demise of the local wards of the governing party and is a continuing trend.

Third, the initial impact around democratization wears off among the voters with more political resources, but they continue to show higher turnout rates and voice their opinions in presidential elections more effectively than their counterparts.

Finally, the observations are consistent with the comparative democratization literature on the relevance of the civil society in its role in democratic transformation. As seen in the section above, we see a clear pattern where an underlying demand for political reform is first manifested by a group of middle class voters in an election. As the social structure went through a gradual change in the 1970s, the middle class gained political weight in South Korea. Their political demand was expressed, among other channels, in the 1985 election, which later would become a precursor to the democratic reform of the country.

Appendix: Estimation Group Turnouts, Entrance and Exit Rates

Table 4.7: (Appendix) Estimated Turnout Rates by Different Demographic Groups in South Korean Elections

	Year	Overall	Educated	Old	Young	Christian
Legislative Elections	1978	0.741	0.482	0.988	0.455	0.781
	1981	0.753	0.639	0.908	0.656	0.624
	1985	0.819	0.850	0.798	0.833	0.743
	1988	0.737	0.695	0.914	0.599	0.668
	1992	0.716	0.543	0.998	0.470	0.719
	1996	0.628	0.584	0.835	0.367	0.681
	2000	0.569	0.529	0.727	0.351	0.534
Presidential Elections	1987	0.861	0.917	0.790	0.933	0.873
	1992	0.823	0.794	0.909	0.719	0.760
	1997	0.797	0.848	0.781	0.814	0.909

Table 4.8: (Appendix) Estimated Entrance Rates of Different Demographic Groups in South Korean Elections

	Year	Overall	Education	Young	Old	Christian
Legislative Elections	1978	0.439	0.511	0.100	0.468	0.414
	1981	0.341	0.568	0.200	0.080	0.580
	1985	0.504	0.825	0.624	0.534	0.664
	1988	0.321	0.644	0.200	0.817	0.360
	1992	0.385	0.686	0.350	0.300	0.785
	1996	0.219	0.637	0.363	0.020	0.560
	2000	0.192	0.457	0.100	0.092	0.252
Presidential Elections	1987	0.782	0.074	0.946	0.799	0.109
	1992	0.617	0.201	0.339	0.797	0.230
	1997	0.514	0.143	0.342	0.754	0.058

Table 4.9: (Appendix) Estimated Exit Rates of Different Demographic Groups in South Korean Elections

	Year	Education	Young	Old	Christian	Overall
Legislative Elections	1978	0.541	0.592	0.000	0.000	0.131
	1981	0.322	0.694	0.019	0.365	0.102
	1985	0.125	0.100	0.097	0.220	0.077
	1988	0.292	0.322	0.191	0.228	0.171
	1992	0.515	0.900	0.001	0.323	0.167
	1996	0.442	0.850	0.056	0.245	0.210
	2000	0.427	0.662	0.018	0.306	0.206
Presidential Elections	1987	0.894	0.000	0.160	0.841	0.121
	1992	0.756	0.379	0.111	0.698	0.143
	1997	0.811	0.123	0.175	0.832	0.134

CHAPTER V

Conclusion

One of the most important themes of this dissertation is on testing ecological inference techniques. As sketched out earlier, ecological inference strategies are bound to make certain assumptions on the aggregation process that generated the aggregate data. To be sure, these assumptions are not something that can be tested directly, since the aggregation process is the hidden force that we do not observe. Moreover, exactly due to this reason, it is hard to design a simulation study to generate data that truly mimics real aggregation processes.

In this vein, testing ecological inference techniques to help the researcher find or model the "right" estimator for the given problem is important. Since aggregation processes are different, unique, and dependent upon the particular context of the question and environment of the data, there is no general solution to ecological inference problems, but only problem- and data- specific solutions. Hence, it is important to put ecological estimation strategies to test by applying them to similar questions with data where the truth is known or where corresponding individual level analysis exists.

For the specific task of estimating voter transition rates from aggregate electoral records, it is fair to conclude the estimator suggested by Thomsen outper-

forms other ecological estimation strategies. For example, as I have illustrated in the first chapter, the voter transition problem comes with difficulties such as bigger-than-usual aggregation bias and severe nonlinearities that are generated by features specific to electoral dynamics. In short, the results suggest that, when it comes to the task of estimating voter transition rates, the Thomsen estimator produces estimates that are far superior to those of conventional and more generalized approaches, such as Goodman's or King's. This was clearly seen when the estimator is applied to voter transition problems in several different contexts of elections held in South Korea, Great Britain, and Florida.

It remains to be true that theory matters. Among the ecological estimators, I find the micro-modeling of voter choices that underlie the Thomsen estimator to fit the question and data the best, and accordingly is the best ecological voter transition estimator we have, both in theory and practice.

It is in fact the theoretical strength of the Thomsen estimator that enables the researcher to extend the voter transition question into more sophisticated problems. In Chapter 4, I proposed an extension of the Thomsen model that enables the estimator to address questions on the correlates of voter transition, where the researcher would be able to analyze heterogeneous transition rates across different demographic groups. In the chapter, I provided examples regarding the turnout of South Korean voters over time among different demographic groups. The extension is consistent with the micro-modeling of voters that the basic Thomsen model outlined, with more possibilities of further extensions.

There are two future avenues to explore when it comes to the covariate model. First, it is possible to imagine a more generalized extension of the model with multinomial covariates. What was illustrated in the chapter typically contrasts

transition rates of the voters in one demographic group against those from the rest of the population. A slight extension of the logic seen in the chapter should enable researchers to look at covariates that include more than two categories, for example, several occupation groups. Barring instability that comes with small-sized demographic groups, this definitely is an immediate possible extension of the covariate model.

Another possible further extension is to consider an additional covariate that taps a different dimension of the electorate other than the first covariate. For example, it is possible to envision voter transitions across elections as a function of class and age, where the question would involve comparing the transition rates among, say, young workers and older workers. Conceptually, this will involve a joint-mixture of the bivariate transition distributions, but implementing the model would require further theoretical work.

Chapter 3 proposed an application of the iterated proportional fit (IPF) extension that enables the estimation of transition rates in a multiparty system. As it turns out, the method is simple, general, and provides estimates that are more reliable than those from existing approaches. More theoretical work on how IPF is related with the literature on multinomial choices is in order, which will enable us to develop the model into a more generalizable solution.

Still, at the current stage, the technique is simply an adjustment process applied after the first-round estimation that provides initial values of IPF. It remains to be seen whether it is possible—or even desirable—to modify basic ecological estimators and embed them in the constraints of IPF, to use the full information in a simultaneous estimation. Or, perhaps it is the case that estimation of multiparty transition rates will have be built from the micro-models of multinomial

choices and *ad hoc* solutions including IPF will have to be replaced by such theoretical models. Still, knowing whether and how individual models of choices can offer insight into the ecological inference process would require more ingenuity and further advances on our understanding of elections and voters.

The key motivation of both extensions shown in the dissertation is to learn whether and how we can address substantively meaningful and interesting questions by overcoming the limitations that aggregate data present. For example, the discussion on the realignment of South Korean voters shows that political theories on the relationship between the civil society and democratization can be empirically examined with aggregate data. Armed with the ecological inference techniques developed here and applying them to more widely available aggregate electoral data sets on various new democracies around democratic transitions, we may conceive a new cross-national analysis agenda on comparative democratization.

Even though there is a multitude of difficulties surrounding the ecological inference process, the literature has advanced, if in small steps, toward providing better instruments and analytic eyes to study topics in time and places where important questions abound but data are scarce. This dissertation is an attempt to contribute to the scholarship in such way.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] ACHEN, CHRISTOPHER. 2000. "The Thomsen Estimator for Ecological Inference.", unpublished Manuscript.

[2] ACHEN, CHRISTOPHER and PHILLIPS SHIVELY. 1995. *Cross-Level Inference*. Chicago, IL: The University of Chicago Press.

[3] ANSELIN, LUC and WENDY K. TAM CHO. 2002. "Spatial Effects and Ecological Inference." *Political Analysis*, 10: 276–297.

[4] BERGLUND, STEN and SØREN THOMSEN. 1990. *Modern Political Ecological Analysis.* Abo, Finland: Abo Academy Press.

[5] CHHIBBER, PRADEEP and KEN KOLLMAN. 1998. "Party Aggregation and the Number of Parties in India and United States." *American Political Science Review*, 92: 329–342.

[6] CRAMER, ERHARD. 1998. "Conditional Iterative Proportional Fitting for Gaussian Distributions." *Journal of Multivariate Analysis*, 65 (2): 261–276.

[7] DEMING, W. E. and F. F. STEPHAN. 1940. "On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Tables Are Known." *The Annals of Mathematical Statistics*, 11.

[8] DIAMOND, LARRY JAY, JUAN J LINZ, and SEYMOUR MARTIN LIPSET. 1995. *Politics in Developing Countries: Comparing Experiences with Democracy*. Boulder, CO: L. Rienner Publishers, 2nd ed ed.

[9] DIAMOND, LARRY JAY and DOH CHULL SHIN. 2000. *Institutional Reform and Democratic Consolidation in Korea*. Stanford, Calif: Hoover Institution Press, Stanford University.

[10] DUFF, BRIAN, MICHAEL J. HANMER, WON-HO PARK, and ISMAIL K. WHITE. 2007. "Good Excuses: Understanding Who Votes With An Improved Turnout Question." *Public Opinion Quarterly*, 71: 67–90.

[11] FERREE, KAREN E. 2004. "Iterative Approaches to R×C Ecological Inference Problems: Where They Can Go Wrong and One Quick Fix." *Political Analysis*, 12: 143–159.

[12] GOODMAN, LEO. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology*, 64: 610–625.

[13] GOSNELL, HAROLD F. 1942. *Grass Roots Politics*. Washington, D.C.: American Council on Public Affairs.

[14] GREENE, WILLIAM. 2003. *Econometric analysis*. Upper Saddle River N.J.: Prentice Hall, 5th ed. ed.

[15] GROFMAN, BERNARD, ed. 1999. *Elections in Japan, Korea, and Taiwan Under the Single Non-Transferable Vote: The Comparative Study of an Embedded Institution*. Ann Arbor, MI: University of Michigan Press.

[16] HAN, SUNG-JOO. 1988. "South Korea in 1987: The Politics of Democratization." *Asian Survey*, 28 (1): 52–61.

[17] HANUSHEK, JOHN E. JACKSON, ERIC and JOHN F. KAIN. 1974. "Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy." *Political Methodology*, 1: 89–107.

[18] IMREY, PETER B., GARY G. KOCH, and MAURA E. STOKES. 1981. "Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression. Part I: Historical and Methodological Overview." *International Statistical Review / Revue Internationale de Statistique*, 49 (3): 265–283.

[19] JACKSON, JOHN E. 2002. "A Seemingly Unrelated Regression Model for Analyzing Multiparty Elections." *Political Analysis*, 10 (1): 49–65.

[20] JENNINGS, M. KENT and GREGORY B. MARKUS. 1984. "Partisan Orientations over the Long Haul: Results from the Three-Wave Political Socialization Panel Study." *The American Political Science Review*, 78 (4): 1000–1018.

[21] JOHNSTON, R. J. and C. J. PATTIE. 1993. "Entropy-Maximizing and the Iterative Proportional Fitting Procedure." *The Professional Geographer*, 45 (3): 317–322.

[22] KEY, V. O. JR. 1955. "A Theory of Critical Elections." *The Journal of Politics*, 17 (1): 3–18.

[23] KIL, SOONG HOOM and CHUNG-IN MOON. 2001. *Understanding Korean Politics: An Introduction (Suny Series in Korean Studies)*. State University of New York Press.

[24] KIM, CHONG LIM. 1980. *Political Participation in Korea: Democracy, Mobilization, and Stability*. Studies in international and comparative politics, Santa Barbara, CA: Clio Books.

[25] KING, GARY. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

[26] KING, GARY, ORI ROSEN, and MARTIN TANNER. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research*, 28: 61–90.

[27] KING, GARY, ORI ROSEN, and MARTIN ABBA TANNER. 2004. *Ecological Inference: New Methodological Strategies*. Analytical methods for social research, Cambridge, UK: Cambridge University Press.

[28] KOO, HAGEN, ed. 1993. *State and Society in Contemporary Korea*. Ithaca, NY: Cornell University Press.

[29] KOUSSER, J MORGAN. 2001. "Ecological Inference from Goodman to King." *Historical Methods*, 34: 101–127.

[30] LANGBEIN, LAURA IRWIN and ALLAN J LICHTMAN. 1978. *Ecological Inference*. Beverly Hills, Calif: Sage Publications.

[31] LEE, JUNHAN. 2002, "Economic Voting in South Korea." Paper presented at the Annual Meeting of the Midwest Political Science Association.

[32] LEWIS, JEFFREY B. 2001. "Understanding King's Ecological Inference Model: A Method-of-Moments Approach." *Historical Methods*, 34: 170–88.

[33] LITTLE, RODERICK J. A. and MEI-MIAU WU. 1991. "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ." *Journal of the American Statistical Association*, 86 (413): 87–95.

[34] LUTTBEG, NORMAN R and MICHAEL M GANT. 1995. *American Electoral Behavior, 1952-1992*. Itasca, Ill: F.E. Peacock Publishers, 2nd ed.

[35] OH, JOHN KIE-CHIANG. 1999. *Korean Politics: The Quest for Democratization and Economic Development*. Cornell University Press.

[36] PALMQUIST, BRADLEY. 1993. "Ecological Inference, Aggregate Data Analysis of U.S. Elections, and the Socialist Party of America." Ph.D. thesis, University of California, Berkeley.

[37] QUANDT, R. and J RAMSEY. 1978. "Estimating Mixtures of Normal Distributions and Switching Regressions." *Journal of American Statistical Association*, 73: 730–738.

[38] ROBINSON, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review*, 15: 351–357.

[39] ROSEN, ORI, WENXIN JIANG, GARY KING, and MARTIN TANNER. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The R $\times$ C Case." *Statistical Neerlandica*, 55: 134–156.

[40] ROSENSTONE, STEVEN J and JOHN MARK HANSEN. 1993. *Mobilization, Participation, and Democracy in America*. New topics in politics, New York: Macmillan Pub. Co.

[41] RUSCHENDORF, LUDGER. 1995. "Convergence of the Iterative Proportional Fitting Procedure." *The Annals of Statistics*, 23 (4): 1160–1174.

[42] TAM CHO, WENDY K. 1998. "Iff the Assumption Fits ... A Comment on the King Ecological Inference Solution." *Political Analysis*, 7: 143–163.

[43] THEIL, HENRI. 1971. *Principles of Econometrics*. Wiley.

[44] THEIL, HENRY. 1967. *Economics and Information Theory*. Amsterdam, Holland: North Holland.

[45] THOMSEN, SØREN. 1987. *Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference*. Arhus, Denmark: Politica.

[46] TRAUGOTT, MICHAEL W and JOHN P KATOSH. 1979. "Response Validity in Surveys of Voting Behavior." *The Public Opinion Quarterly*, 43: 359–377.

[47] WEIR, BLAIR T. 1975. "The Distortion of Voter Recall." *American Journal of Political Science*, 19: 53–62.

[48] WELLHOFER, E SPENCER. 2001. "Party Realignment and Voter Transition in Italy, 1987-1996." *Comparative Political Studies*, 34: 156–186.

[49] WONG, DAVID W. S. 1992. "The Reliability of Using the Iterative Proportional Fitting Procedure." *The Professional Geographer*, 44 (3): 340–348.

[50] ZELLNER, ARNOLD. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, 57 (298): 348–368.