New Approaches to Discrete Choice and Time-Series Cross-Section Methodology for Political Research

by
Jonathan Kropko

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Political Science.

Chapel Hill
2011

Approved by:

George Rabinowitz, Advisor

John Aldrich, Committee Member

Thomas Carsey, Committee Member

Skyler Cranmer, Committee Member

Justin Gross, Committee Member

Michael Mackuen, Committee Member

Abstract

JONATHAN KROPKO: New Approaches to Discrete Choice and Time-Series Cross-Section Methodology for Political Research. (Under the direction of George Rabinowitz.)

This dissertation consists of three projects which focus on methods, with direct applications to American and comparative politics, and are intended to either advance researchers' understanding of existing methods or to help develop new methods for political science research. These works focus generally on two methodological areas: discrete choice modeling and time-series cross-section (TSCS) methodology. In chapter 1 I conduct Monte Carlo simulations to compare multinomial logit (MNL), multinomial probit (MNP), and mixed/random parameters logit (MXL) on 7 criteria, including the accuracy of coefficient point estimates. Simulated data represent a range of violations of the independence of irrelevant alternatives (IIA) assumption, and show that MNL nearly always provides more accurate coefficients than MNP and MXL, even when the IIA assumption is severely violated. In chapter 2 a new method, called the between effects estimation routine (BEER), is developed to maximize information from TSCS data to model the cross-sectional effects while allowing these effects to change over time. This method is applied to two examples. First, it is used to analyze the variation in regional authority and federalism in 21 countries over 54 years. Second, it is used to reconsider the effect of income on state voting in US presidential elections. In chapter 3 I develop a method to estimate a non-linear logistic regression with survey weights which uses Markov-Chain Monte Carlo (MCMC) estimation. To demonstrate the utility of the method, I consider a voting model for U.S. presidential general elections which is non-linear, and addresses

a long-standing theoretical debate in the study of American elections regarding the moderating role of personal importance on the effect of issue evaluations on the vote.

Dedicated to George Rabinowitz

Acknowledgments

I would especially like to thank George Rabinowitz for all of his help and guidance. As my advisor, he pushed me to the point where I could see the finish line, but he was not able to cross it with me. If the research contained in this dissertation is in any way impressive, much of the credit belongs to George. If the research is unimpressive, the credit is mine.

This dissertation would not have been possible without the support of family and friends: Marv and Olga Kropko, Josh Kropko, Tia Kropko, Tamara Mittman, Jessica Brandes, Emanuel Coman, Ian Conlon, Shaina Korman-Houston, Caitlin LeCroy, Rachel Meyerson, Zach Robbins, Samantha Snow, and Sarah Turner.

I owe a great deal to the consultants at the Odum Institute for Research in the Social Sciences: Augustus Anderson, Ryan Bakker, Kimberly Coffey, Tab Combs, Anne K. Hunter, Heather Krull, Gerald Lackey, James E. Monogan, Evan Parker-Stephen, Jessica Pearlman, Daniel Serrano, Brian Stucky, Chris Wiesen, Bev Wilson, and Cathy Zimmer.

An earlier version of chapter 1 was presented for the annual meeting of The Midwest Political Science Association, Palmer House Hilton, Chicago, Illinois, April 5, 2008. I would like to thank John Aldrich, Fredrick J. Boehmke, Kenneth A. Bollen, Thomas Carsey, Skyler Cranmer, David Drukker, Justin E. Esarey, Garrett Glasgow, Justin Gross, Wendy Gross, Jon Krosnick, Stuart Macdonald, Michael Mackuen, Marco Steenbergen, and Georg Vanberg for helpful comments on earlier versions of this research.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

- **ANES**: American National Election Study
- **BE**: between estimator
- **BEER**: the between effects estimation routine
- **CDF**: cumulative density function
- **DGP**: data generating process
- **DUE**: decomposed unit effect
- **FE**: fixed effects
- **FEVD**: fixed effects with vector decomposition
- **GDP**: gross domestic product
- **GEE**: generalized estimating equations
- **GHK**: Geweke-Hajivassilou-Keane algorithm
- **GMM**: generalized method of moments
- **IIA**: independence of irrelevant alternatives
- **LL**: log-likelihood
- **LOWESS**: locally weighted scatterplot smoothing
- **MCMC**: Markov-chain Monte Carlo
- **ML, MLE**: maximum likelihood, maximum likelihood estimation
- **MNL**: multinomial logit
- **MNP**: multinomial probit
- **MSE**: mean squared error
- **MXL**: mixed logit, or random parameters logit
- **NL**: non-linear
- **NLS**: non-linear least squares
- **OLS**: ordinary least squares
- **PCSE**: panel corrected standard errors

- **PDF**: probability density function
- **RAI**: regional authority index
- **RE**: random effects
- **Reg**: individual regressions for each time point
- **SD**: standard deviation
- **SDP**: Social Democratic Party
- **SE**: standard error
- **TE**: time effects
- **TE1**: time effects interacted with linear time
- **TE2**: time effects interacted with quadratic time
- **TE3**: time effects interacted with cubic time
- **TSCS**: time-series cross-section

Chapter 1

A Comparison of Three Discrete Choice Estimators

1.1 Summary

Political researchers are often confronted with unordered categorical variables, such as the vote-choice of a particular voter in a multiparty election. In such situations, researchers must choose an empirical model with an appropriate estimator to analyze the data. Three well-known estimators are multinomial logit (MNL), multinomial probit (MNP), and mixed/random parameters logit (MXL). MNL is simpler, but also makes the often erroneous independence of irrelevant alternatives (IIA) assumption. Little is known, however, about the effect of violations of IIA on the quality of MNL coefficient estimates. MNP and MXL are more general estimators which do not assume IIA. In this paper, I conduct Monte Carlo simulations to compare the three estimators on 7 criteria, including the accuracy of coefficient point estimates. Simulated data represent a range of violations of the IIA assumption, and show that MNL nearly always provides more accurate coefficients than MNP and MXL, even when the IIA assumption is severely violated.

1.2 Introduction

Researchers who model discrete choices, such as the choice of a voter in a multiparty election, must choose between competing empirical estimators for the model. Several estimators for discrete choice models are now easily implemented in statistical software packages, and three options available to researchers are multinomial logit (MNL), multinomial probit (MNP), and mixed logit (MXL) which is also called random parameters logit. Technically, these estimators are very similar: they differ only in the distribution of the error terms. MNL has errors which are independent and identically distributed, and MNP and MXL use more general distributions which allow errors to be correlated.[1]

The independent errors of MNL force an assumption called independence of irrelevant alternatives (IIA). Essentially, IIA requires that an individual's evaluation of an alternative relative to another alternative should not change if a third (irrelevant) alternative is added to or dropped from the analysis. When IIA is violated, MNL cannot produce accurate estimates of substitution patterns, which are marginal effects of covariates when an alternative is hypothetically considered to have dropped out of the analysis. MNL is an inappropriate estimator for researchers who are interested in the perceived similarity between choices, or the proclivity of individuals to substitute one alternative for another. MNP and MXL do not assume IIA: although they do not estimate the error correlations, they do estimate parameters which account for these correlations. Many researchers use MNP and MXL as better theoretical models for the data.

Most of the work to describe the problems caused by IIA focuses on the inability of MNL to estimate substitution patterns. But it is important to note that MNL incorrectly specifies a model when IIA is violated, and therefore coefficient estimates are inconsistent. The nature of this inconsistency is not well understood: the severity of the bias,

[1]The following abbreviations will be used frequently throughout the article: MNL refers to multinomial logit; MNP refers to multinomial probit; MXL refers to mixed logit, which is also known as random parameters logit; IIA refers to the independence of irrelevant alternatives assumption.

and whether the bias becomes more severe as IIA becomes a less tenable assumption, are unclear. As Dow and Endersby (2004) have previously discussed, the principle concern of researchers who must choose a discrete choice estimator is not accurate estimation of substitution patterns, but rather the ability of the estimator to return accurate coefficients and inferences, and to do so reliably.

Other researchers have used MNP and MXL to produce estimates of the residual correlation between choices. However, as discussed in section 1.3.3 and as described previously by Bolduc (1999) and Keane (1992), while MNP and MXL estimate parameters to account for residual correlation, the individual correlations between residuals cannot be separately identified by MNP or MXL.

The goal of the analysis presented here is to inform the decisions of researchers in the field who must choose between MNL, MNP, and MXL. I consider the relative performances of the three estimators on 7 criteria: (1) the accuracy of coefficient point estimates, (2) the rate of correct signs and (3) correct inferences for coefficient estimates, (4) the frequency with which each estimator fails to converge to meaningful results, (5) the time it takes for each estimator to converge, (6) the accuracy of predicted probabilities, and (7) estimates of parameters to account for error correlations. The comparisons are conducted using Monte Carlo simulations.

The simulations suggest that, contrary to the focus of the literature, the validity of the IIA assumption should not be a major concern for researchers in choosing between the three estimators. The results indicate that, in most situations, MNL provides more accurate point estimates than MXL or MNP even when the IIA assumption is severely violated. If the goal is to estimate choice probabilities, then the simulations suggest that MNP provides an improvement over MNL and MXL, but at the expense of the coefficient estimates.

Monte Carlo simulation is a standard technique for researchers who wish to study the

properties of a statistical model or estimator. "True" parameter values are specified beforehand, and the results are analyzed in each iteration to see how accurately they return these parameter values. A comparison of discrete choice estimators, however, presents unique challenges. The binary logit and probit models differ in their specifications in that a logit model assumes a logistic distribution for the residuals and probit uses a standard normal distribution for the residuals. These distributions only vary significantly at the tails. More importantly, they assume different variances, leading to coefficient estimates which are not directly comparable to the true parameter values because of a difference in scaling. For the same reason, MNL and MNP coefficients for the same model are not directly comparable to each other.

One important challenge for this research is the development of a technique to compare MNL, MXL, and MNP coefficients to true parameter values and to each other in a way which removes the differences between the coefficients due to the scaling of each estimator. This technique is described in detail in section 1.5.1. With the scaling-differences removed, the models still differ in the functional form of their likelihood functions. MNL and MXL use a multivariate logistic link function, and MNP uses a multivariate normal distribution. In the simulations, the data are generated using a multivariate normal distribution; so MNP has something of a "home field advantage." Despite this advantage, MNL and MXL consistently outperform MNP in returning accurate coefficient point estimates.

This discussion proceeds as follows: the statistical derivations and current uses of MNL, MNP, and MXL in political science are discussed in section 1.3; the simulations are described in detail in section 1.4; results are presented and discussed in section 1.5; and in section 1.6 I summarize the results and offer a recommendation to researchers who use discrete choice models.

1.3 Background

There has been a great deal of work which compares MNL and MNP on theoretical or empirical grounds, and this project builds on that work in two ways. First, since MXL is included, three well-known estimators are compared rather than two. Second, the estimators are compared on the accuracy of their coefficient point estimates. Much of the previous work on choosing between MNL and MNP focuses on whether MNL coefficients change when a candidate is added or dropped from the analysis rather than on the consistency of MNL coefficients when IIA is violated in the full choice-set.[2]

In political science, some comparisons have focused on the question of whether IIA is a tenable theoretical assumption for a specific election. Alvarez and Nagler (1998) and Quinn et al (1999) compare relative suitability of MNL and MNP for elections in Britain and in the Netherlands. Other comparisons have been empirical. Dow and Endersby (2004) use MNL and MNP to estimate the same models of voter choice in the 1992 U.S. presidential election and 1995 French presidential election and find little difference between predictions whether MNL or MNP is used. They argue that given the concerns of the identification of MNP discussed by Keane (1992), researchers should be more confident in the MNL results. In their 1994 paper, Alvarez and Nagler conduct simple Monte Carlo simulation experiments to demonstrate that on a number of criteria MNP outperforms models which assume IIA. They generate data with a known covariance structure in the choice errors, and alter the correlations as an experimental treatment. A similar methodology is adopted for the comparisons made in this paper.

[2]On this basis, tests for IIA violation were developed to indicate whether MNP should be used instead of MNL (Hausman and Wise 1978; McFadden and Hausman 1984), although Fry and Harris (1996) and Cheng and Long (2007) report that these tests are generally unreliable.

1.3.1 Statistical Framework

All three estimators model discrete choices by using the framework of a random utility model, which derives latent utilities for each individual for choosing each alternative. An individual's choice is the alternative with the highest utility, and the probability of this choice is the probability that the associated utility is the highest among the alternatives. For example, a voter in the 1987 British election chose between the Conservative party, the Labour party, and the Social Democratic Party-Liberal Party alliance. For a particular voter, a random utility model uses three separate equations to estimate the voter's evaluation of the Conservative and Labour parties and the SDP-Liberal alliance. Formally, individual i evaluates alternative j according to the equation

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \tag{1.1}$$

where V_{ij} represents the deterministic part of the evaluation and ε_{ij} is the residual, stochastic part of this evaluation. V_{ij} may be modeled linearly, containing covariates and estimated coefficients. A model which assumes that $\text{corr}(\varepsilon_j, \varepsilon_k) = 0$ for any two alternatives j and k makes the IIA assumption. For all three estimators, the likelihood function for a model with N individuals choosing between J alternatives is

$$L = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y_i = j)^{I_{ij}}, \tag{1.2}$$

where $I_{ij} = 1$ if individual i chooses alternative j and is 0 otherwise. MNL, MXL, and MNP use different functions for the choice probability $P(y_i = j)$.

For all examples presented in this analysis, the number of alternatives J in the model is 3. The case of three alternatives is the simplest on which to compare the three methods, and is also the quickest and least problematic situation to estimate a model using MNP

or MXL. For notational ease, let

$$\eta_2 = \varepsilon_2 - \varepsilon_1, \text{ and } \eta_3 = \varepsilon_3 - \varepsilon_1. \tag{1.3}$$

The probability of choosing alternative 1 is the probability that alternative 1 is the most highly evaluated, so that U_1 is greater than both U_2 and U_3:

$$\begin{aligned} P(y_i = 1) &= P(U_{i1} > U_{i2} \text{ and } U_{i1} > U_{i3}) \\ &= P(V_{i1} + \varepsilon_{i1} > V_{i2} + \varepsilon_{i2} \text{ and } V_{i1} + \varepsilon_{i1} > V_{i3} + \varepsilon_{i3}) \\ &= P(\eta_{i2} < V_{i1} - V_{i2} \text{ and } \eta_{i3} < V_{i1} - V_{i3}) \\ &= \int_{-\infty}^{V_{i1}-V_{i2}} \int_{-\infty}^{V_{i1}-V_{i3}} f(\eta_2, \eta_3)\ d\eta_3\ d\eta_2, \end{aligned} \tag{1.4}$$

where $f(\eta_2, \eta_3)$ is the joint PDF of η_2 and η_3. MNL, MXL, and MNP differ only in their assumptions about this distribution. MNL uses a distribution which assumes IIA, but has an analytic integral, so that estimates are relatively easy to compute even when there are a large number of alternatives. MNL and the IIA assumption are discussed in more detail in section 1.3.2. MNP and MXL are more general than MNL, and do not assume IIA. However neither one uses choice probabilities which can be integrated analytically, so numerical methods must be used to approximate the integral in equation 1.4.[3] These two estimators are discussed in section 1.3.3.

[3]MNP uses the Geweke-Hajivassilou-Keane (GHK) simulation algorithm for approximating probabilities from multivariate normal CDFs (Geweke 1991; Geweke et al 1994; Keane and Wolpin 1994; Hajivassiliou et al 1996; Hajivassiliou and McFadden 1998). MXL uses a simulated maximum likelihood routine which is similar to accept-reject sampling (Train 2003).

1.3.2 MNL and the IIA Assumption

Presently, most political science articles that estimate a discrete choice model use MNL. Computation of MNL is highly efficient because the choice probability has an analytic solution:

$$P(y_i = 1) = \frac{e^{V_{i1}}}{e^{V_{i1}} + e^{V_{i2}} + e^{V_{i3}}}. \tag{1.5}$$

This convenient form for the MNL choice probabilities depends on two assumptions. The first assumption is that error terms in equation 1.1 are distributed by the type-1 extreme value distribution. The second assumption is that the error terms are also independent, which is the IIA assumption.[4]

The substantive implication of IIA that is often inappropriate is that the odds ratio between two alternatives depends only on information about those two alternatives, and no information about any other alternative will change this ratio. In modeling a multiparty election, IIA requires that a voter's relative evaluation of two parties must not change, even when a third party enters the race, leaves the race, or changes positions during the race. IIA is violated, for example, when a voter's relative evaluation of the Conservative and Labour parties in Britain may depend on how much of an issue the Liberal Democrat party is making out of taxes. Still, it is important to note that questions about substitution patterns across alternatives are hypothetical. In the full choice-set, IIA is a problem only in the residuals of the model, and only causes problems if the appropriate variables to explain similarity between choices are excluded from the model.

Little attention has been paid to other possible detriments of IIA. When IIA is false,

[4]The arguments of the distribution in equation 1.4 are differences of residuals, and the difference between two independent type-1 extreme value random variables is distributed logistically (Cameron and Trivedi 2005, p.486).

parameter estimates from MNL are inconsistent since the likelihood function being maximized is now incorrectly specified. Although MNL's inability to accurately model substitution effects has been well documented, the effects of IIA violation on MNL parameter estimates have not been widely reported in the literature. Whether and to what extent this bias increases as IIA is violated more severely is an open question and will be in part addressed here.

1.3.3 MNP and MXL

MNP and MXL do not make the IIA assumption. MNP assumes that the errors are distributed by a multivariate normal distribution:

$$\begin{bmatrix} \eta_2 \\ \eta_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & \rho \\ \rho & \sigma^2_{\eta_3} \end{bmatrix} \right). \tag{1.6}$$

For identification, the variance of η_2 is set to a fixed value.[5] For three alternatives, MNP estimates two extra parameters: $\sigma^2_{\eta_3}$, the variance of the second difference in residuals, and ρ, the correlation between the two differences of residuals. It is important to note that these parameters do not provide estimates of any individual elements from the error covariance matrix. The estimates for $\sigma^2_{\eta_3}$ and ρ relate to the choice errors $\varepsilon_1, \varepsilon_2$, and ε_3 as follows:

$$\sigma^2_{\eta_3} = \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_3} - 2\rho_{\varepsilon_1,\varepsilon_3}\sigma_{\varepsilon_1}\sigma_{\varepsilon_3}, \tag{1.7}$$

and

$$\rho = \frac{\rho_{\varepsilon_2,\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} - \rho_{\varepsilon_1,\varepsilon_2}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} - \rho_{\varepsilon_1,\varepsilon_3}\sigma_{\varepsilon_1}\sigma_{\varepsilon_3} + \sigma^2_{\varepsilon_1}}{\sqrt{\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2} - 2\rho_{\varepsilon_1,\varepsilon_2}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}}\sqrt{\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_3} - 2\rho_{\varepsilon_1,\varepsilon_3}\sigma_{\varepsilon_1}\sigma_{\varepsilon_3}}}. \tag{1.8}$$

[5] "asmprobit" and "mixlogit" in Stata 11 behave as if the variance of both the first and second choice are 1, and every correlation involving the first choice is zero, arriving at a fixed value for $\sigma^2_{\eta_2}$ of 2. See Bolduc (1999) for a more detailed description of variance normalization and simulated maximum likelihood for the MNP model.

The estimate for ρ, in particular, involves all three choice correlations, but none of the individual correlations are separately identified. The estimated correlation described in equation 1.8 accounts for the individual error correlations, but does not actually estimate them.

MNP is a popular alternative to MNL. It was introduced to political science in the mid and late 1990s through a series of articles that demonstrated applications to elections and the spatial model of voting (Alvarez and Nagler 1995, 2000; Schofield et al 1998; Alvarez et al 2000), and is now a standard topic in econometric textbooks, usually presented following a discussion of MNL and the IIA assumption (Greene 2003; Cameron and Trivedi 2005; Long and Freeze 2005). Since 2007, articles including a discrete choice model estimated through some type of MNP have appeared in a number of political science journals, including *Public Opinion Quarterly* (Fullerton et al 2007; Campbell and Monson 2008), *Political Behavior* (Kam 2007; Wilson 2008), *Electoral Studies* (Alvarez and Katz 2009; Blais and Rheault 2010; Fisher and Hobolt 2010), *the European Journal of Political Economy* (Van Groezen et al 2009), *Comparative Political Studies* (Ivarsflaten 2008), and *the American Journal of Political Science* (Humphreys and Weinstein 2008). Results in these articles appear to be robust, without any obvious signs of model non-convergence.

A second alternative to MNL, although less widely used in political science, is MXL, which is sometimes referred to as random parameters logit. MXL allows some coefficients to be modeled as random. This estimator incorporates characteristics of both MNL and MNP: logistic probabilities are weighted by the joint PDF of the random coefficients, $g(\beta)$, and it is typical, though not required, for $g(\beta)$ to be a multivariate normal distribution. The form of MXL used in these simulations is the "error components" specification (Train 2003), in which the choice-specific intercepts are the only random coefficients, and are specified to be correlated. The distribution of these random intercepts is the same as the

distribution used by MNP in equation 1.6. Like MNP, MXL estimates $\sigma^2_{\eta_3}$ and ρ. Again, as defined in equations 1.7 and 1.8, these parameters account for error correlation but do not provide estimates of individual error correlations.

MXL has desirable properties as a theoretical model when coefficients are believed to actually follow a distribution. For example, many applications of MXL have been in microeconomic studies of consumer choice in specific markets such as automobiles (Brownstone and Train 1999) and household appliances with different efficiency levels (Revelt and Train 1998). In these contexts, the random coefficients represent "random taste variation" (Train 2003). MXL does not assume IIA; covariates with random coefficients are treated as random effects for the residuals, and these random effects correlate across choices. Advocates of MXL note that it is a flexible estimator, and have proven that any random utility model has choice probabilities which can be modeled as closely as desired by an appropriately specified version of MXL (McFadden and Train 2000). Train (2003) and Cameron and Trivedi (2005) list several advantages that MXL has over MNP in both flexibility and computational ease. The application of MXL to the study of multiparty elections was introduced to political science by Glasgow (2001), and has recently been used in a piece in *Electoral Studies* (Clarke et al 2010), but is generally used much less frequently in political science than MNL and MNP. Still, MXL has been increasingly used in economics, so it is useful to gage the relative performance of MXL to MNP as well as to MNL.

1.4 Methods

I use Monte Carlo simulations to compare MNL, MXL, and MNP. Following the approach of Alvarez and Nagler (1994), this research compares the three estimation methods in a laboratory setting in which the correlations between the choice errors are chosen as an experimental treatment before running any models. The models considered here include

predictors that vary across alternatives, which I refer to as choice-variant predictors, and others that are fixed across alternatives, which I refer to as choice-fixed predictors.[6] For example, in election data, a choice-variant predictor is the policy distance between a voter and each party, and a choice-fixed predictor is the voter's age. Policy distance varies by individuals and by parties, since each party stands at a different distance from the voter's own ideal point. Age varies across voters, but does not depend on the party being considered. It may, however, have a different impact on the evaluations of each of the parties.

In order to be appropriate for a discrete choice model, data must have a unique observation for each combination of case and alternative. This shape allows for choice-variant and choice-fixed variables. Suppose that a discrete choice model is used to examine the effect of policy distance and voters' ages on their vote choices. In table 1.1, a row is defined by the unique combination of "Voter ID" and "Party." The dependent variable "Vote" is 1 if the voter chooses to vote for the party being considered on that row, and is 0 otherwise. "Policy Distance" is choice-variant, but "Age" is choice-fixed.

Table 1.1: Example Discrete Choice Data.

Voter ID	Party	Vote	Policy Distance	Age
1	1	0	5	37
1	2	0	7	37
1	3	1	2	37
2	1	1	3	29
2	2	0	4	29
2	3	0	8	29
3	1	0	4	64
3	2	1	4	64
3	3	0	8	64

The simulations generate data that resemble voter-choice data as illustrated in table

[6]The type of MNL used here is a hybrid of conditional logit, which only considers choice-variant predictors, and a basic version of MNL, which only considers choice-fixed predictors.

1.1. I use two types of data generation processes. The first type, which I label "basic," uses a minimal number of covariates. Only 1 choice-variant covariate and only 1 choice-fixed covariate are used. As described in section 1.4.1, these covariates, as well as their coefficients, are drawn from uniform distributions. The second type, which I label "Britain," follows the recommendation of Macdonald et al (2007) that, to the greatest extent possible, simulated data should resemble real data. Covariates are taken from the 1987 British Election Study, which has been used in a number of important political science papers on discrete choice methodology (Whitten and Palmer 1996; Alvarez and Nagler 1998; Quinn et al 1999; Alvarez et al 2000), and coefficients are drawn from the linear regression of party affect on these covariates. The Britain models are described in section 1.4.2. Both the basic and the Britain models simulate a dependent variable with 3 alternatives.

1.4.1 "Basic" Data Generation

The data generating processes are modeled on the random utility framework described in section 1.3. Each individual in the simulated data has a utility U_{ij} for each alternative which has a deterministic part V_{ij} and a stochastic part ε_{ij}. The deterministic part is a linear combination of choice-variant and choice-fixed predictors. V_{ij} also includes an alternative-specific constant. In order to identify each model, one alternative is chosen to be the base alternative for which the coefficients for choice-fixed predictors and the alternative-specific constant are set to zero.

For the basic models, the evaluation of individual i of alternative $j \in \{1, 2, 3\}$ is

$$U_{ij} = \lambda z_{ij} + \beta_{j,1} x_i + \beta_{j,0} + \varepsilon_{ij}, \tag{1.9}$$

where candidate 1 is the base choice, requiring that $\beta_{1,0} = \beta_{1,1} = 0$. The choice-variant data, z, are independently drawn from a uniform distribution from 0 to 1. x is also drawn

from a uniform distribution from 0 to 1, but is held constant across the alternatives for each individual. Five "true" coefficients (λ, $\beta_{2,0}$, $\beta_{2,1}$, $\beta_{3,0}$, and $\beta_{3,1}$) are independently drawn from a uniform distribution from -1 to 1.

The stochastic part of U_{ij} consists of choice errors, $\varepsilon_1, \varepsilon_2$, and ε_3, which are drawn from a trivariate normal distribution:

$$\begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \end{bmatrix}' \sim N\left(\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}', \Sigma \right). \tag{1.10}$$

The structure of the variance-covariance matrix of the choice errors, denoted by Σ, is crucial to the theoretical goals of the simulations. In these basic models, the variances in Σ are all set at one. I use 11 different correlation structures to represent a range of violations of the IIA assumption. Each structure implies a different matrix for Σ in equation 1.10. I present these error structures in section 1.4.3. The errors are drawn independently for individuals, but are correlated across the alternatives.

After all of the covariates, coefficients, and errors have been drawn, the simulated choice of an individual is the alternative with the highest latent utility for that individual.

During each iteration of the simulation, a simulated data set is generated, then MNL, MXL, and MNP are run using the simulated choice as the dependent variable, z as a choice-variant predictor, and x as a choice-fixed predictor.[7] The models are specified to include alternative-specific constants. The coefficients and standard errors from each model are saved, as well as the time for each model to converge, indicators for model non-convergence, predicted probabilities, and estimates of the parameters that account for error correlation. This information is used to assess the relative performance of each

[7]Multinomial logit is implemented by the "asclogit" command in Stata 11, and multinomial probit is implemented by the "asmprobit" command. These versions of MNL and MNP allow for covariates which vary across choices as with conditional logit, and also allow covariates which are fixed across choices. Mixed logit is implemented by the "mixlogit" command (package st0133_1, Hole 2007), which is not part of Stata's base language and needs to be downloaded. The simulations are run on a cluster of Linux compute nodes available for researchers at UNC Chapel Hill.

estimator. The exact criteria on which the estimators are compared are described and results are reported in section 1.5. Each simulation is iterated 300 times for each of the 11 error correlation structures listed in section 1.4.3.

1.4.2 "Britain" Data Generation

For the Britain models I use the data from the 1987 British election study (Heath et al 1987). Indices for the alternatives are $\{C, L, A\}$, where C refers to the Conservative party, L refers to the Labour party, and A refers to the Liberal/Social Democrat party alliance. For individual i, the deterministic part of the evaluation of alternative $j \in \{\mathrm{C}, \mathrm{L}, \mathrm{A}\}$ is a linear combination of 7 choice-variant predictors, 10 choice-fixed predictors, and alternative-specific constants:

$$U_{ij} = \sum_{k=1}^{7} \lambda_k z_{ijk} + \sum_{l=1}^{10} \beta_{jl} x_{il} + \beta_{j,0} + \varepsilon_{ij}. \tag{1.11}$$

To obtain realistic coefficients for λ and β in equation 1.11, I draw coefficients from a regression of the affect a respondent reports for each party, and the results from this regression are listed in table A.1 in Appendix A. The Conservative party is treated as the base alternative, so $\beta_{C,0} = \ldots = \beta_{C,10} = 0$. The choice-variant z variables are ideological distances between the respondents and the parties on defense, unemployment, taxation, nationalization, income redistribution, crime, and welfare. The choice-fixed x variables are characteristics of the voters which include the respondent's age, gender, income, geographic location, union membership, and homeowner status. More specific information about the variables included in the analysis are listed in Appendix A. The simulated choice of individual i is again the alternative with the highest associated utility.

The stochastic part of U_{ij} contains choice errors ε_C, ε_L, and ε_A. Again, the choice errors are randomly generated from a trivariate normal distribution with means equal to zero and a predefined variance-covariance structure Σ derived from one of the 11

correlation structures listed in section 1.4.3. The variances of the errors are also derived from the regression in table A.1 in Appendix A. The residuals from this regression are calculated and separated by the choice being considered. The variance-covariance matrix for the three vectors of residuals is

$$\Sigma = \begin{bmatrix} 1.133 & . & . \\ -0.406 & 1.127 & . \\ -0.083 & -0.039 & 0.604 \end{bmatrix}. \tag{1.12}$$

This matrix yields the correlation matrix

$$\chi = \begin{bmatrix} 1 & . & . \\ -0.359 & 1 & . \\ -0.100 & -0.047 & 1 \end{bmatrix}. \tag{1.13}$$

The variances in 1.12 are used for the error variances in all of the Britain models, and the correlations in 1.13 are used as model K in the error correlation structures described in section 1.4.3, which are applied to both the Britain and basic models.

The Britain simulations are also iterated 300 times for each error correlation structure listed in section 1.4.3. During each iteration, MNL, MXL, and MNP are run using the simulated choice as the dependent variable, the z variables as choice-variant predictors, and the x variables as choice-fixed predictors. As before, the models are specified to include alternative-specific constants. The same information is saved as in the basic simulations, and is also assessed as described and reported in section 1.5.

1.4.3 Error Correlation Structures

Correlation structures are chosen to represent a range of violations of the IIA assumption, with differing degrees of severity. For the basic models, all three errors are assumed to

have variances equal to 1. For the Britain models, errors have the variances listed in equation 1.12. I consider 11 correlation models, which I call models A through K.

$$\chi_A = \begin{bmatrix} 1 & . & . \\ 0 & 1 & . \\ 0 & 0 & 1 \end{bmatrix} \quad \chi_B = \begin{bmatrix} 1 & . & . \\ .10 & 1 & . \\ .10 & .10 & 1 \end{bmatrix} \quad \chi_C = \begin{bmatrix} 1 & . & . \\ .25 & 1 & . \\ .25 & .25 & 1 \end{bmatrix}$$

$$\chi_D = \begin{bmatrix} 1 & . & . \\ .50 & 1 & . \\ .50 & .50 & 1 \end{bmatrix} \quad \chi_E = \begin{bmatrix} 1 & . & . \\ .75 & 1 & . \\ .75 & .75 & 1 \end{bmatrix} \quad \chi_F = \begin{bmatrix} 1 & . & . \\ 0 & 1 & . \\ .80 & 0 & 1 \end{bmatrix}$$

$$\chi_G = \begin{bmatrix} 1 & . & . \\ 0 & 1 & . \\ -.80 & 0 & 1 \end{bmatrix} \quad \chi_H = \begin{bmatrix} 1 & . & . \\ 0 & 1 & . \\ .50 & .80 & 1 \end{bmatrix} \quad \chi_I = \begin{bmatrix} 1 & . & . \\ 0 & 1 & . \\ -.50 & .80 & 1 \end{bmatrix}$$

$$\chi_J = \begin{bmatrix} 1 & . & . \\ -.20 & 1 & . \\ -.50 & .80 & 1 \end{bmatrix} \quad \chi_K = \begin{bmatrix} 1 & . & . \\ -0.359 & 1 & . \\ -0.100 & -0.047 & 1 \end{bmatrix}$$

Models A, F, G, H, I, and J were considered by Alvarez and Nagler (1994). Models E through J set correlations at high levels in order to observe the behavior of MNL, MXL, and MNP in the case of extreme violation of IIA. The correlations in model K, listed in equation 1.13 come directly from real data, and are probably the most directly applicable to applied research. The basic simulation model and the Britain simulation model are run on each of the 11 correlation structures, for 22 simulations in total.

1.5 Results and Discussion

The primary focus of this analysis is to determine which estimator produces the best coefficients. I analyze the performances of MNL, MXL, and MNP in returning accurate coefficient point estimates with the correct signs. These results are listed in section 1.5.1. In section 1.5.2, I compare the three estimators on several other criteria which are of interest to researchers in the field: inferences, the rate at which each estimator fails to converge to meaningful results, the time it takes for each estimator to converge, and the accuracy of predicted probabilities and parameters which account for error correlations.

It can be argued that the results presented here are idiosyncratic to Stata, or to the particular estimation technique through which MNP is operationalized in Stata. To address these concerns, the simulations are rerun in R version 2.6.1 on exactly the same data.[8] These results are reported in Appendix B. MNP is implemented in R within the "MNP" library, written by Imai and van Dyk (2005), and uses a Gibbs Sampler rather than simulated MLE to estimate choice probabilities. In general, Bayesian MNP in R is less stable than MNP in Stata, and becomes more likely to produce an error and terminate as the number of draws in the Gibbs Sampler increases. When the estimator does converge, the results are very similar to the ones presented in section 1.5.1.

1.5.1 Comparison of Coefficient Estimates

MNL, MNP, and MXL are estimators of the probability that, for each individual i and each alternative j, individual i selects alternative j. However, these probabilities are not the first concern for most researchers who employ discrete choice models. As with most multivariate empirical models, researchers are mainly concerned with the magnitude and direction of coefficient estimates returned by the model. The fit of model probabilities

[8]Only MNL and MNP are compared in R. I was unable to find a reliable implementation of MXL in R, although a few are now in development.

is a secondary concern which illuminates the robustness of coefficient estimates. For example, an analyst of an election will be primarily concerned with how covariates are estimated to be associated with vote choice. The analyst may test whether voters with greater income are more likely to vote for a right-wing party, and if this effect is larger than the effect of regular church attendance. In choosing between MNL, MNP, and MXL, most researchers will want to choose the method that most reliably returns coefficients.

Comparing the accuracy of coefficient estimates in discrete choice estimators is difficult because each estimator makes assumptions about the error variances which scale coefficients by an unknown factor. A straight comparison of point estimates to the true parameter values used in the generation of the data is impossible, since the actual error cannot be separated from the deviation due to scaling. In order to assess the error in coefficient point estimates in this analysis, a statistic which is similar to a Pearson correlation is calculated between the vector of true coefficient values and the vector of point estimates from each of MNL, MXL, and MNP.

Formally, the derivation of this comparative statistic is as follows. Each estimator returns a vector of K non-constrained coefficient estimates. In the basic simulation models, $K = 5$ since we estimate one effect for a choice-variant X variable, one constant for each non-base alternative, and one effect for a choice-fixed X variable for each non-base alternative. In the Britain models, $K = 29$ since there are 7 choice-variant X variables, and a constant and 10 choice-fixed X variables for both non-base alternatives. For each model, the vector of true parameter values $\{\beta_1, \beta_2, \ldots, \beta_K\}$ is saved, and the vectors of point-estimates from each estimator $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \bar{\hat{\delta}}_K\}$ are saved. It is assumed that the returned coefficient estimates are scaled by the fixed standard deviation of each estimator, denoted σ. Therefore we can represent the returned estimate as a quotient of an estimate which is comparable to the true effect and the standard deviation:

$$\hat{\delta} = \frac{\hat{\beta}}{\sigma}. \tag{1.14}$$

In order to eliminate σ from these estimates, we calculate the mean across the estimates for each model, $\bar{\hat{\delta}}$, and compute

$$\frac{\sum_{i=1}^{K}(\hat{\delta}_i - \bar{\hat{\delta}})(\beta_i - \bar{\beta})}{\sqrt{\sum_{j=1}^{K}(\hat{\delta}_i - \bar{\hat{\delta}})^2}\sqrt{\sum_{j=1}^{K}(\beta_i - \bar{\beta})^2}} = \frac{\sum_{i=1}^{K}\left(\frac{\hat{\beta}_i}{\sigma} - \frac{\bar{\hat{\beta}}}{\sigma}\right)(\beta_i - \bar{\beta})}{\sqrt{\sum_{j=1}^{K}\left(\frac{\hat{\beta}_i}{\sigma} - \frac{\bar{\hat{\beta}}}{\sigma}\right)^2}\sqrt{\sum_{j=1}^{K}(\beta_i - \bar{\beta})^2}}$$

$$= \frac{\frac{1}{\sigma}\sum_{i=1}^{K}(\hat{\beta}_i - \bar{\hat{\beta}})(\beta_i - \bar{\beta})}{\frac{1}{\sigma}\sqrt{\sum_{j=1}^{K}(\hat{\beta}_i - \bar{\hat{\beta}})^2}\sqrt{\sum_{j=1}^{K}(\beta_i - \bar{\beta})^2}} = \frac{\sum_{i=1}^{K}(\hat{\beta}_i - \bar{\hat{\beta}})(\beta_i - \bar{\beta})}{\sqrt{\sum_{j=1}^{K}(\hat{\beta}_i - \bar{\hat{\beta}})^2}\sqrt{\sum_{j=1}^{K}(\beta_i - \bar{\beta})^2}}. \quad (1.15)$$

The standard deviation cancels from the top and bottom of this fraction. This correlation statistic, reported in table 1.2, removes the constant scaling parameter and provides the desired information: if estimates differ from true parameter values only because of the scale, the correlation is 1. Lower correlations represent correspondingly less accurate coefficient estimates.[9] Unlike the point estimates, the signs and inferences of the coefficient estimates are naturally meaningful since they do not depend on an assumed variance. The average percent of coefficients for which each estimator returns the correct sign is reported in table 1.3.

In tables 1.2 and 1.3, the evaluative statistic is reported for MNL, MXL, and MNP, for the basic and the Britain models, and for each of the 11 error correlation structures. The standard error of the mean for each measurement is reported, and two-tailed t tests are conducted. Stars indicate that one estimator performs significantly better than both competitors.

MNL returns more accurate coefficients than MXL and MNP in the majority of cases. In the basic models, which contain a minimum number of covariates, point estimates from

[9]Two other measures were used to compare coefficient estimates while removing the scale. The first involved normalizing each vector of coefficients by dividing each coefficient by the standard deviation of the vector. The second divided each coefficient by the sum of absolute elements of the vector. Both measures produce results which are extremely similar to the ones reported in table 1.2.

Table 1.2: Correlations Between Coefficient Estimates from MNL, MXL, and MNP and the True Coefficients.

	Basic Simulations			Britain Simulations		
Model	MNL	MXL	MNP	MNL	MXL	MNP
A	0.98(.00)***	0.93(.01)	0.92(.01)	0.51(.01)***	0.45(.01)	0.40(.01)
B	0.98(.00)***	0.94(.01)	0.91(.01)	0.54(.01)***	0.49(.01)	0.44(.01)
C	0.98(.00)***	0.95(.01)	0.93(.01)	0.53(.01)***	0.48(.01)	0.43(.01)
D	0.98(.00)***	0.95(.01)	0.93(.01)	0.55(.01)***	0.50(.01)	0.46(.01)
E	0.98(.00)***	0.97(.00)	0.96(.01)	0.56(.02)***	0.51(.02)	0.49(.02)
F	0.88(.01)	0.94(.01)*	0.93(.01)	0.41(.01)	0.48(.01)	0.48(.01)**
G	0.94(.01)**	0.92(.01)	0.90(.01)	0.54(.01)***	0.44(.01)	0.41(.01)
H	0.88(.01)	0.93(.01)	0.94(.01)	0.59(.02)***	0.52(.02)	0.48(.02)
I	0.86(.01)	0.93(.01)	0.93(.01)	0.58(.03)***	0.48(.03)	0.44(.04)
J	0.84(.01)	0.93(.01)	0.92(.01)	0.58(.02)***	0.48(.02)	0.44(.02)
K	0.97(.00)***	0.93(.01)	0.91(.01)	0.53(.01)***	0.47(.01)	0.41(.01)

Note: stars indicate that the marked correlation is significantly greater than the next highest correlation.

Two-tailed t tests: *** $p < .01$, ** $p < .05$, * $p < .1$

Table 1.3: Percent Correct Signs of Coefficient Estimates from MNL, MXL, and MNP.

	Basic Simulations			Britain Simulations		
Model	MNL	MXL	MNP	MNL	MXL	MNP
A	97.4(.45)***	94.1(.64)	93.8(.71)	82.3(.44)	82.4(.45)	81.6(.46)
B	96.2(.53)***	93.5(.70)	93.0(.76)	83.1(.46)	83.6(.45)*	82.4(.47)
C	95.6(.57)*	94.5(.59)	93.7(.67)	84.0(.48)	83.9(.48)	83.2(.48)
D	97.1(.45)***	95.1(.61)	94.6(.65)	85.1(.47)	85.0(.48)	84.4(.45)
E	97.7(.40)***	96.2(.51)	95.6(.56)	87.6(.56)	87.2(.60)	86.7(.63)
F	90.8(.80)	94.5(.67)	94.4(.69)	83.4(.44)	85.8(.42)	86.0(.41)
G	92.4(.73)**	90.8(.83)	90.3(.84)	81.0(.51)	80.9(.51)	79.9(.52)
H	91.6(.78)	92.9(.80)	93.2(.85)	86.7(.75)***	85.3(.76)	84.7(.75)
I	90.5(.84)	91.8(.79)	92.5(.82)	86.7(1.6)	86.1(1.4)	84.1(1.3)
J	88.5(.95)	93.5(.76)	93.1(.81)	82.6(1.0)*	81.5(.93)	80.5(1.0)
K	95.2(.61)***	91.9(.75)	91.6(.75)	83.2(.46)	83.0(.46)	81.7(.47)

Note: stars indicate that the marked value is significantly greater than the next highest value.

Two-tailed t tests: *** $p < .01$, ** $p < .05$, * $p < .1$

MNL have significantly more accurate signs and magnitudes than MXL and MNP for the cases of error independence (model A), equal correlations (models B through E), and a large and negative correlation (model G). Importantly, MNL is more accurate for

the most realistic correlation structure (model K). MXL and MNP have more accurate coefficients when a pair of errors are highly and positively correlated (models F, H, I, and J), and there is little difference between MXL and MNP in these cases.

However, in the Britain models MNL is more accurate in estimating the magnitude of coefficients for every correlation structure except F, which represents the unlikely situation that two alternatives are correlated at .8 while being independent from the third alternative. Likewise, with the exception of model F, MNL returns the correct signs more often than MNP, although there is little difference between MNL and MXL in these cases.

Violations of the IIA assumption have the expected effect on MNL estimates for the basic models: coefficient estimates are less accurate for the error correlation structures with large, positive correlations. Neither MXL or MNP are affected by IIA violations in such a systematic way. Comparing the basic model results to the Britain model results, however, it appears that the three estimators are hurt by the complexity of the Britain models, and MXL and MNP are hurt more by this complexity than MNL. Strangely, violation of IIA does not damage MNL estimates in the Britain models as it does in the basic models. In fact, MNL performs somewhat better with the highly correlated structures in the Britain models.

These results may seem surprising, and an argument can be made that the performance of MXL and MNP would improve if the number of draws used to approximate probability are increased. The default number of Halton draws for the maximum likelihood simulator used by MXL in Stata is 50, with 15 initial draws thrown out as a burn-in. The default number of Hammersley draws for MNP in Stata is 50 times the number of alternatives. For the three-choice model used here, the default number of draws is 150. I increased the number of draws for MXL to 200 and 500, and for MNP to 300 and 1000, and recalculated the comparative statistics from this section to see if the performance of

the estimators improves. I tested the increase in draws using correlation model K, as the most realistic correlation model. Increasing the number of draws results in an increase in the time it takes each estimator to converge, so it was only feasible to run a large number of iterations for the basic models. I also ran 10 iterations for the Britain models. These results are reported in table 1.4.

Table 1.4: Evaluation Statistics for the Three Estimators, Increasing the Number of Draws.

Statistic	MNL	MXL			MNP		
		50 draws	200 draws	500 draws	150 draws	300 draws	1000 draws
Basic Simulation, Model "K". N=266							
Correlation	0.96(.00)	0.93(.01)	0.92(.01)	0.92(.01)	0.89(.01)	0.90(.01)	0.89(.01)
% Correct Sign	93.8(.68)	92.7(.72)	92.5(.78)	92.4(.76)	90.7(.86)	90.6(.84)	90.3(.88)
Britain Simulation, Model "K". N=10							
Correlation	0.51(.06)	0.48(.07)	0.52(.06)	0.48(.07)	0.42(.08)	0.42(.08)	0.42(.08)
% Correct Sign	82.4(2.5)	81.0(2.4)	81.2(2.4)	81.0(2.2)	80.0(2.4)	80.0(2.4)	80.0(2.4)

The estimators do not consistently improve with more draws. In fact, as MXL moves from 200 to 500 draws and MNP moves from 300 to 1000 draws they both perform slightly less well, though not significantly so. These results suggest that the fact that MNL outperforms MXL and MNP is not due to the number of draws used to estimate MXL and MNP.

1.5.2 Additional Comparisons

In order to examine the performances of the three estimators as thoroughly as possible, MNL, MXL, and MNP are compared on five additional criteria which are of interest to researchers in the field. First, the ability of each estimator to return accurate inferences about coefficient estimates is considered. Next, the estimators are compared on the

frequency with which they fail to converge to meaningful results, and the time it takes each estimator to converge. Finally, since much of the work to compare MNL and MNP has focused on model fit and substitution patterns, the estimators are compared in the accuracy of their predicted probabilities and estimates of parameters to account for error correlation.

Inferences.

Inferences are difficult to gage using simulated data since the true parameters are specified with no variance or uncertainty. In order to generate a baseline against which to compare the inferences of MNL, MXL, and MNP coefficient estimates, I use the inferences from the regression on party affect from the 1987 British election study listed in table A.1 in Appendix A. Party affect is conceptually similar to the latent utilities which are used to generate the simulated dependent variables in the Britain models, described in section 1.4.2, and these models use the same predictors for the simulated vote choice as the ones that party affect is regressed on. As an illustrative exercise, it is reasonable to expect the estimators to return the same signs and inferences for the covariates as reported in the regression.

Inferences from the regression are matched to the inferences for the same covariates from MNL, MXL, and MNP on the simulated choice. An inference is coded as correct if the coefficient estimate has the same sign as the true parameter, and the same conclusion at the .1 level. The basic models draw true parameter values with no variance, therefore there is no basis on which to compare inference. Table 1.5 therefore only reports the percent of correct inferences for the Britain models. Clearly this measure is not perfect, and does not provide definitive evidence. However, the results listed in table 1.5 are illustrative of the general performance of each estimator, and are in accord with the results in tables 1.2 and 1.3.

Table 1.5: Percent Correct Inferences on Coefficient Estimates from MNL, MXL, and MNP.

Model	MNL	MXL	MNP
A	60.7(.40)***	57.1(.47)	56.8(.42)
B	61.7(.38)***	58.5(.43)	57.5(.39)
C	61.6(.39)***	58.5(.49)	58.1(.41)
D	64.0(.44)***	60.8(.50)	59.3(.45)
E	67.4(.67)***	62.7(.74)	52.6(1.3)
F	60.1(.42)	63.4(.48)	65.7(.41)***
G	59.4(.38)***	54.6(.49)	53.4(.41)
H	64.6(.72)***	59.6(.85)	56.4(1.0)
I	60.7(1.4)***	55.2(1.6)	49.3(1.7)
J	61.4(.80)***	56.3(.97)	48.5(.93)
K	60.5(.41)***	57.1(.50)	56.8(.46)

Note: stars indicate that the marked value is significantly greater than the next highest value. Two-tailed t tests:
*** $p < .01$, ** $p < .05$, * $p < .1$

Failure Rates and Time to Convergence.

Concerns about the consistency of parameter or probability estimates are moot if an estimator fails to produce usable results, or if the model cannot converge in a practical amount of time. The failure rate and convergence time for each estimator are not definitive criteria on which to choose between MNL, MXP and MNP. However, they must be considered while comparing the estimators in other ways. The failure rates for the three estimators as well as the average convergence time of each estimator, omitting failed attempts, are reported in table 1.6 for each of the 22 simulations.

The failure rate computes the number of failed convergences out of the total number of iterations run in each simulation. Criteria for identifying a failure are obvious signs of non-convergence, such as missing coefficient estimates, blown up or missing standard

Table 1.6: Failure Rates and Average Time to Successfully Converge for MNL, MXL, and MNP.

Model	MNL		MXL		MNP	
	Fail Rate	Time	Fail Rate	Time	Fail Rate	Time
Basic Simulations						
A	0%	.5 sec.	14.0%	1 min., 53 sec.	2.3%	50 sec.
B	0%	.5 sec.	12.0%	1 min., 59 sec.	3.0%	57 sec.
C	0%	.4 sec.	13.0%	1 min., 53 sec.	3.0%	1 min., 9 sec.
D	0%	.5 sec.	8.0%	1 min., 51 sec.	0.6%	47 sec.
E	0.3%	.7 sec.	8.3%	2 min., 15 sec.	3.3%	1 min, 4 sec.
F	0%	.5 sec.	17.0%	2 min., 15 sec.	1.7%	1 min., 7 sec.
G	0%	.4 sec.	18.3%	2 min., 17 sec.	3.3%	50 sec.
H	4.7%	.8 sec.	27.3%	3 min., 8 sec.	5.3%	1 min., 49 sec.
I	3.3%	.7 sec.	23.0%	3 min., 10 sec.	6.0%	1 min., 31 sec.
J	2.0%	.7 sec.	22.0%	3 min., 8 sec.	5.0%	1 min., 38 sec.
K	0%	.7 sec.	15.0%	3 min., 9 sec.	3.0%	1 min., 12 sec.
Britain Simulations						
A	0%	2 sec.	4.7%	16 min., 36 sec.	0%	3 min., 46 sec.
B	0%	2 sec.	4.7%	14 min., 33 sec.	0.3%	3 min., 54 sec.
C	0%	3 sec.	4.3%	14 min., 51 sec.	0%	8 min., 40 sec.
D	2.0%	2 sec.	12.0%	14 min., 31 sec.	0.3%	3 min., 48 sec.
E	34.2%	3 sec.	49.0%	13 min., 48 sec.	22.8%	3 min., 31 sec.
F	0%	1 sec.	13.0%	14 min., 14 sec.	0%	2 min., 21 sec.
G	0%	1 sec.	8.0%	10 min., 1 sec.	0%	2 min., 8 sec.
H	53.7%	1 sec.	60.7%	10 min., 11 sec.	34.0%	2 min., 21 sec.
I	59.6%	2 sec.	74.1%	19 min., 41 sec.	41.6%	3 min., 27 sec.
J	51.3%	2 sec.	63.0%	13 min., 11 sec.	46.7%	2 min., 19 sec.
K	0%	3 sec.	7.0%	14 min., 41 sec.	0%	3 min., 59 sec.

errors,[10] and too many iterations without improvement in the log-likelihood function.[11] In addition, each estimator is marked to have failed if it produces an error and terminates without producing results.

[10] A run of an estimator is marked as failed if the trace of the variance matrix of coefficient estimates is greater than 100. This statistic captures any instance in which a standard error is infeasibly large for the simulated data.

[11] MNP is coded as failed if it takes more than 16,000 iterations to converge. MXL tends to converge with fewer than 15 iterations, so MXL is coded as failed once it uses 50 iterations.

MNL should converge more quickly than the other two estimators since choice probabilities do not need to be approximated numerically. The results in table 1.6 demonstrate that, as expected, MNL is much quicker, although none of the estimators ever take more than 20 minutes to converge on average. MXL consistently has a higher failure rate than MNP, which tends to fail more often than MNL. However, the three estimators share the cases in which their failure rates are highest, and these situations involve higher error correlations (models E, H, I, and J). These models reflect weaker identification of error covariance elements, and cause MNL, which is normally very stable, to fail at high rates just like MNP and MXL. Since these models represent overly severe violations of IIA, the fail rates do not provide conclusive evidence to use one estimator over another in the three choice case.

Predicted Probabilities.

The fit of a discrete choice model can be assessed by the magnitude of probability estimates for the alternatives which are actually chosen in the data. A standard way to choose between models is to select the one with the best fit. However, there is a danger that a model may overfit the data, and provide probability estimates which are more certain than is appropriate. In the simulations used in this analysis, uncertainty is built into the simulated choice, therefore "correct" choice probabilities can be computed.[12] The predicted probabilities from MNL, MXL, and MNP are calculated for each run, and are assessed for their mean squared error with the true choice probabilities. The average errors for each estimator and each correlation structure are reported in table 1.7.

MNP is consistently estimating the choice probabilities with less error than MNL or

[12]Given the known values of coefficient and error covariance parameters, the correct values for the predicted choice probabilities are given by the CDF of the multivariate normal distribution. The choice probability for choice 1 in these simulations is the bivariate normal CDF with arguments η_2/σ_{η_2} and η_3/σ_{η_3} as defined in equations 1.3 and 1.7, and correlation ρ as defined in equation 1.8. The bivariate normal CDF is easily calculated in Stata with the "binormal" command. Probabilities in higher dimension spaces are more cumbersome to calculate.

Table 1.7: Errors in Predicted Probabilities from MNL, MXL, and MNP.

Model	MNL	MXL	MNP	(MNL-MNP)
Basic Simulations				
A	0.1031(.005)	0.1031(.005)	0.1030(.005)	0.0000
B	0.1299(.005)	0.1299(.005)	0.1298(.005)***	0.0001
C	0.1369(.006)	0.1369(.006)	0.1368(.006)***	0.0001
D	0.1815(.007)	0.1814(.007)	0.1813(.007)***	0.0002
E	0.2778(.008)	0.2775(.008)	0.2774(.008)***	0.0004
F	0.1655(.006)	0.1655(.006)	0.1655(.006)	0.0000
G	0.0953(.006)	0.0954(.006)	0.0953(.006)*	0.0000
H	0.1472(.006)	0.1471(.006)	0.1471(.006)	0.0001
I	0.1336(.005)	0.1336(.005)	0.1335(.005)	0.0001
J	0.1206(.005)	0.1206(.005)	0.1205(.005)***	0.0001
K	0.1002(.006)	0.1003(.006)	0.1002(.006)***	0.0000
Britain Simulations				
A	0.2728 (.000)	0.2715(.000)	0.2714(.000)***	0.0014
B	0.2908 (.000)	0.2894(.000)	0.2893(.000)***	0.0016
C	0.3238 (.000)	0.3224(.000)	0.3222(.000)***	0.0016
D	0.3960 (.000)	0.3946(.000)	0.3944(.000)***	0.0016
E	0.0765 (.000)	0.0746(.000)	0.0743(.000)***	0.0022
F	0.2755 (.000)***	0.2794(.000)	0.2795(.000)	-0.0039
G	0.2384 (.000)	0.2365(.000)	0.2364(.000)***	0.0020
H	0.3891 (.001)	0.3876(.001)	0.3877(.001)	0.0014
I	0.2848 (.001)	0.2837(.001)	0.2837(.001)	0.0011
J	0.2741 (.001)	0.2731(.001)	0.2731(.001)	0.0010
K	0.2457 (.000)	0.2448(.000)	0.2447(.000)***	0.0011

Note: stars indicate that the value is significantly lower than the next lowest value.
Two-tailed t tests: *** $p < .01$, ** $p < .05$, * $p < .1$

MXL, and for most error structures, MNP is significantly more accurate than the other two estimators. Note, however, that the differences in these errors are always very small. In fact, in the basic models the errors do not differ in the first three decimal places and in the Britain models the errors do not differ in the first two decimal places. Neither the complexity of the model, nor the error correlation structure seem to drastically effect the relative performances of the estimators in returning probabilities. Note, however, that the fit of MNP is helped by the fact that the errors in the simulations used here are generated from a multivariate normal distribution, which is the assumption of MNP. Moreover, MNP and MXL estimate two parameters that MNL does not, and extra parameters

generally result in a better fit.[13]

Parameters Which Account for Error Correlation.

One important reason why researchers may prefer MXL or MNP to MNL is that the more general MXL and MNP models allow estimation of substitution effects. That is, individuals may consider some alternatives to be more suitable replacements than others for an alternative that drops out of the analysis for reasons which are not specified in the model. These substitution patterns can be analyzed by deriving the marginal effect of covariates on the predicted probabilities when the utility for the "irrelevant" alternative is normalized to 0. However, these patterns are only reliable if the model can accurately estimate the correlation structure of the error. As noted in equation 1.8, neither MXL nor MNP can return individual elements of the error correlation matrix, but they can account for the correlation by estimating the correlation of residual differences. The true correlation values in table 1.8 are derived by evaluating equation 1.8 using the error covariance elements assumed by each model. Stars in this table indicate that the estimator produces a correlation point estimate which is on average significantly different from the correct value.

MNP appears, in general, to estimate the correlation parameter more accurately than MXL. In the Britain models, which are more complicated, MXL becomes much less reliable but MNP does not seem to suffer any worse than in the basic models. Note however that correlation estimates from MNP are biased downwards in every model. It may be the case that MNP correlation estimates are biased downwards in general.

[13]A good way to test whether MNP really does produce more accurate estimates of predicted probabilities is to analyze the out-of-sample performance of MNP, which would penalize MNP for overfitting the probabilities to the sample. Such an analysis is beyond the scope of this project, but is part of a future research agenda.

Table 1.8: Comparison of Parameter Estimates Which Account for Error Correlation.

	Basic Simulations			Britain Simulations		
Model	True Value	MXL	MNP	True Value	MXL	MNP
A	0.50	0.30(.05)***	0.45(.03)	0.57	-0.02(.05)***	0.54(.02)*
B	0.50	0.33(.04)***	0.44(.03)**	0.57	-0.01(.04)***	0.53(.02)**
C	0.50	0.36(.04)***	0.42(.03)***	0.57	0.11(.05)***	0.55(.02)
D	0.50	0.33(.04)***	0.44(.02)**	0.56	0.06(.04)***	0.53(.02)
E	0.50	0.32(.04)***	0.45(.02)**	0.54	0.14(.09)***	0.45(.04)*
F	0.22	-0.00(.04)***	0.18(.03)	0.51	0.25(.05)***	0.49(.03)
G	0.67	0.52(.05)***	0.60(.03)***	0.68	0.39(.06)***	0.58(.03)***
H	0.92	0.90(.02)	0.86(.02)***	0.96	0.65(.07)***	0.80(.03)***
I	0.94	0.75(.04)***	0.83(.02)***	0.92	0.54(.12)***	0.75(.06)***
J	0.93	0.83(.03)***	0.85(.02)***	0.92	0.68(.12)*	0.85(.03)**
K	0.58	0.43(.05)***	0.52(.03)*	0.65	0.14(.04)***	0.62(.02)

Note: stars indicate that the correlation is significantly different from the correct value.
Two-tailed t tests: *** $p < .01$, ** $p < .05$, * $p < .1$

1.6 Conclusion

The choice between MNL, MXL, and MNP should depend on the goal of the researcher. If the goal is to interpret the effects of covariates on the discrete choice, then the simulations presented here indicate that, in most situations, MNL provides more accurate point estimates than MXL or MNP even when the IIA assumption is severely violated. If the goal is to estimate choice probabilities, then the simulations suggest that MNP provides an improvement over MNL and MXL, but at the expense of the coefficient estimates. The error-components version of MXL used here rarely outperforms MNL and MNP in these simulations, but if the goal is to model the effects of covariates as random with a known distribution, then a random-slope version of MXL may well be the best option.

The simulations suggest that the validity of the IIA assumption should not be a major concern for researchers in choosing between the three estimators. Parameter estimates from MNL do not become more biased in the realistic Britain models as violation of the IIA assumption becomes more severe. In addition, the advantage of MNP over MNL in estimating accurate choice probabilities does not become greater as IIA violation increases. If IIA is still a concern, then it is important to note that the correlation in

the errors modeled by MNP and MXL is correlation between random variables which represent noise. Therefore, by definition, we do not have any theory to explain error correlation, or else we should have included that information in the deterministic part of the model. A thoughtful model which includes variables that capture perceived similarity between choices will be less likely to violate IIA.

In general, since MNL produces more accurate coefficient estimates and signs, and since MNP estimates probabilities only marginally more accurately than MNL, the simulations suggest that for the most common applications researchers in the field should prefer MNL to MNP.

Chapter 2

Drawing Accurate Inferences About the Differences Between Cases in Time-Series Cross-Section Data

2.1 Summary

Researchers with time-series cross-section (TSCS) data should be aware that different methods to analyze TSCS data are designed to produce inferences about different aspects of the data. Many commonly used methods consider only the variation over time. Some consider only the variation across cases, and others draw inferences by averaging the two dimensions of variance. A new method, called the between effects estimation routine (BEER), is developed to maximize information from the TSCS data to model the cross-sectional effects while allowing these effects to change over time. Individual regressions are run, and the results are combined using a Bayesian averaging process that places larger weights on time points which are similar and proximate to the time point under consideration. This method is applied to two examples. First, it is used to analyze the variation in regional authority and federalism in 21 countries over 54 years. Second, it is used to reconsider the effect of income on state voting in US presidential elections.

Simulations demonstrate that BEER is more accurate than the other commonly used TSCS methods when the goal of the researcher is to model cross-sectional effects that change in a smooth way over time.

2.2 Introduction

Time-series cross-section (TSCS) data contain a cross-section of N cases, measured at each of T time points, and have two variances: a variance across the cases, and a variance across time. Studies that use TSCS data are becoming increasingly prevalent in political science. In American politics, many studies consider the 50 states over a time span. In comparative politics and international relations, these data take the form of countries and years. While a cross-sectional dataset may have a sample size of N, and a time-series can contain T observations of one case, TSCS data can have as many as $N \times T$ observations. This increase in power is part of the reason why TSCS methods have become so popular, but different TSCS methods make different generalizations within the data; TSCS methods allow us to generalize across cases, across time, or both if it is appropriate to do so. Political researchers need to be aware of the assumptions being made by their methods, and they need to be more careful to choose a method which makes theoretically appropriate assumptions.[1]

In choosing an appropriate method, researchers should first be aware that different

[1]Choosing a method can be difficult. A great deal of confusion arises with TSCS methodology because different disciplines use inconsistent and contradictory terminology. For example, in econometrics, a dummy variable for a case is called a fixed effect and a slope estimate for an independent variable is called a coefficient. In psychometrics, a dummy variable for a case is called a random effect and a slope estimate for an independent variable is called a fixed effect. Political scientists experience more grief from the terminology because the field is interdisciplinary, and researchers commonly receive methods training in both economics and psychology departments. Furthermore, the treatment of TSCS data is a subject which is of concern to researchers in many disciplines, but discussion across disciplines has been limited. This lack of collaboration is partly due to different focuses for the use of TSCS data, but it is partly because this kind of data structure has many different names: TSCS, pooled time-series, longitudinal, repeated measures, and wave study data. All of these data structures are special cases of panel or pooled data, which are special cases of multilevel or hierarchical data.

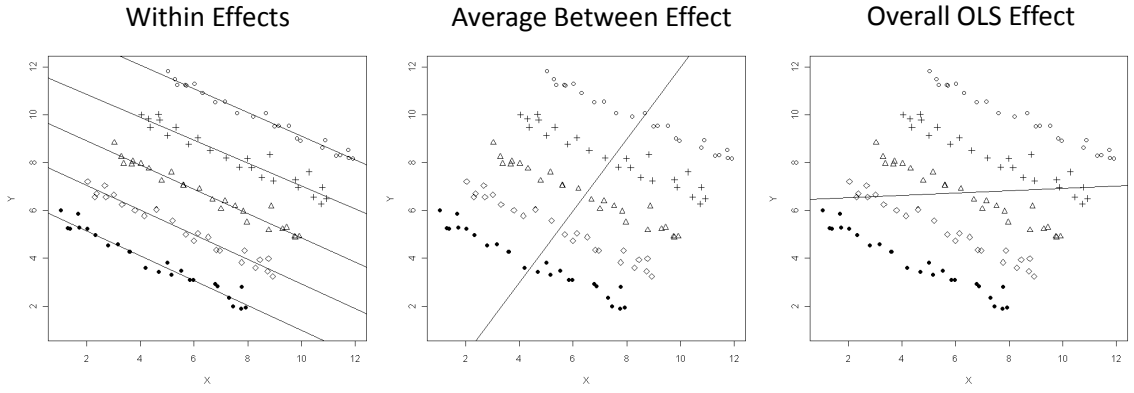
TSCS methods are designed to produce inferences about different aspects of the data. Many commonly used methods consider only the variation over time. Some consider only the variation across cases, and others draw inferences by averaging the two dimensions of variance. A method which analyzes the cross-sections should not in general be expected to produce the same results as a method which analyzes time variation. Since inferences on the cross-sectional and time variation are different, an average of the two will not typically provide accurate results on either dimension.

Let X refer generally to an independent variable in a linear model, and let Y refer to the dependent variable. In this discussion I will use the following terminology:

- Between Effect: the effect of X on cross-sectional differences in Y;
- Within Effect: the effect of X on changes in Y over time.

There is no reason why these two effects should necessarily be equal, and there are many situations in fact when the two effects might be expected to be in opposite directions. Figure 2.1 contains an illustration of hypothetical TSCS data in which the between and

Figure 2.1: Within Effects, the Between Effect of Case-Level Averages, and the Overall OLS Effect.



Note: The graphs are scatterplots of hypothetical data which include five cases, measured at 25 time points each. The 25 observations of each case are denoted by the solid dots, diamonds, triangles, crosses, and open dots respectively.

within effects have different signs. Each graph is a scatterplot with an overlaid best-fit regression line. In the graph on the left, the within effects are drawn. These slopes capture the fact that within each case, increases in X are associated with decreases in Y. A between effect is drawn on the graph in the middle which is derived from the average values of X and Y within each case. This line illustrates the fact that at any point in time, a case with a higher value of X than another case also tends to have a higher value of Y. The graph on the right contains the standard OLS best-fit line, which is an average of the between and within slopes, and fails to accurately depict the variation along either dimension.

In this paper, I develop a new method for researchers who are interested in estimating the between effects in TSCS data. Commonly used TSCS methods are discussed in section 2.3, including a few which estimate between effects. The new method discussed here, however, allows a researcher to model something that the other methods do not: the between effects are allowed to vary over time.

Political theory should suggest a model for how data are generated. This goal of this research is to develop a TSCS method to estimate a model in which the between effects are estimated and are allowed to change over time in a smooth fashion. Formally, this data-generating model can be represented as follows:

$$y_{it} = \alpha_t + x_{it}\beta_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_t^2), \tag{2.1}$$

$$Cov(\beta_t, \beta_{t-1}) > 0, \qquad \forall t > t_1,$$

where $i \in \{1, \ldots, N\}$ denotes cases, $t \in \{t_1, \ldots, T\}$ denotes time points, y_{it} is the dependent variable, x_{it} is a column vector of independent variables, α_t and σ_t^2 are scalar, and β_t is a column vector of between effects. The parameters α_t, β_t, and σ_t^2 are all allowed to vary over time, and the between effects are assumed to be positively correlated in

adjacent time points. The within variation is entirely accounted for by the time-specific constants α_t so that the regression coefficients only provide estimates of between effects. The method developed here to estimate this model is called the between effects estimation routine (BEER), and it is designed to estimate the between effects in a particular time point while maximizing the amount of information in the TSCS data used to make these estimates. The algorithm focuses solely on the between effects of predictors by "de-pooling" the data, running OLS on individual cross-sections, and using conjugate Bayesian priors to average results across the cross-sections. The formulation of BEER is discussed in detail in section 2.4.

The data generating model described in equation 2.1 applies to many political research topics. Two examples are presented in this paper: one from comparative politics and one from American politics. First, in section 2.5, BEER and other common TSCS methods are used to estimate a model of regional authority in 21 countries from 1950 to 2006. Regional authority is a measurement of the strength and autonomy of regional governments within a country relative to the federal government (Marks et al 2008). The principal research questions are concerned with the differences in regional authority across countries. Belgium, for example, exhibits a high level of regional authority across the time span, and Denmark exhibits a low level of regional authority. One important difference between Belgium and Denmark is that Belgium has two large regionally-based ethnic groups and Denmark to a greater extent is ethnically homogenous. If these ethnic differences become politically salient, then an electorate may push for greater decentralization of power. However, the time span of the data consists of 56 years, and it is unreasonable to assume that the effect of ethnic fragmentation is the same in 1950 as it is in 2006. In regards to ethnicity, Western democracies have increasingly focused on tolerance, so ethnic fragmentation may be a less important determinant of decentralization at the end of the time span than it had been at the beginning of the time span.

In the second example, described in section 2.6, BEER is used to analyze the effect of median state income on the state's votes in U.S. presidential elections from 1964 to 2004. Andrew Gelman et al (2008) find that Democrats receive higher percentages of the vote in richer states relative to poorer states. They also show, however, that this between effect has only emerged in the last 20 years. The growing importance of income as a predictor of state-level voting concords with other theories about political polarization in the American electorate. Income, perhaps, is an indicator of diverging cultures between the states, which would imply that these differences are much more pronounced now than in 1964.

In both of these examples, the research question calls for the analysis of between effects, and theory requires that these effects be allowed to vary over time in a smooth fashion. BEER assumes this exact model, and therefore should return more accurate results than other TSCS methods which do not assume this model. In order to test the robustness of BEER against alternative TSCS methods in a more general setting, simulations are run. These simulations, presented in section 2.7, generate artificial data using four different models, including the one described in equation 2.1. The simulations demonstrate that BEER is the most accurate alternative when the model in equation 2.1 is the true data generating process.

2.3 Background

This section focuses on demonstrating how various estimators of models for TSCS data produce different results, and why. A summary of the of how these approaches have been used in the literature is not the focus, although Wilson and Butler (2007) have provided an extensive literature review of how TSCS methods have been used in political science.

All models for TSCS data use the following equation:

$$y_{it} = g(\alpha_{it}, \beta_{it}, x_{it}) + u_i + v_t + e_{it}, \tag{2.2}$$

where $g(.)$ is a function of the predictors x_{it}, an intercept α_{it}, and coefficients β_{it}. When the intercept and coefficients are fixed across time and across cases, and when $g(.)$ is a linear function of x_{it}, this model is called the pooled model (Cameron and Trivedi 2005, p. 699). There are $\sum_{i=1}^{N} T_i$ total observations. The overall residual for the model is $\varepsilon_{it} = u_i + v_t + e_{it}$, where u_i is the unobserved variance of y_{it} that varies across cases but is fixed across time, and v_t is the unobserved variance of y_{it} that varies across time but is fixed across cases. These values are also called unit effects, or unobserved heterogeneity. e_{it} is the idiosyncratic error, which varies over time and across panels. In the discussion in this section, I assume that the residuals are independent of the regressors. This assumption would imply that $\text{Corr}(x_{it}, u_i) = 0$, which may be controversial, but only if endogeneity is a concern for an OLS regression within an individual cross-section in the data.

There are many options for researchers who have TSCS data and want to choose an estimator for the model in equation 2.2. Before discussing what distinguishes the different methods for TSCS data, however, it will be useful to define between and within variables. Any variable x_{it} that varies across cases and over time can be separated into parts which capture only the cross-sectional variation and only the temporal variation of x_{it}. The cross-sectional part of x_{it} is called a between variable, and its effect in a regression model is strictly a between effect as defined in section 2.2. One way to create a between variable is to take the average value of the variable over time for each case:

$$x_i^B = \frac{\sum_{t=1}^{T_i} x_{it}}{T_i}. \tag{2.3}$$

These variables are fixed over time, but have a variance between the cases. Note that the between effects are defined to be averages over the measured time frame. The temporal part of x_{it} is called a within variable, and its effect in a regression is strictly a within effect. One way to construct a within variable is to subtract the between variable from the whole variable:

$$x_{it}^{W} = x_{it} - x_{i}^{B}. \tag{2.4}$$

The within variable centers x_{it} around its mean for each case, so that average differences between cases are removed. The differences over time, however, are preserved. Notice that, by these definitions, any variable which varies over time and cross-sectionally is the sum of its between and within parts:

$$x_{it} = x_{i}^{B} + x_{it}^{W}. \tag{2.5}$$

Table 2.1 provides an example of how to calculate these between and within parts of a TSCS variable. Suppose we have data on the population of two towns from 2006 to 2010. Both towns are growing, but town 2 is always bigger than town 1. The between version of population is the average population of each town over the five year period. This variable captures the fact that town 2 is bigger than town 1. The within version of population is the difference between the populations and each town's average population. This variable preserves year to year differences within each town – the population of town 1 grew by 200 people between 2006 and 2007, and so did the within version of population – but eliminates meaningful differences between town 1 and town 2.

Generally, commonly used TSCS methods can be classified into one four groups depending on how they treat the between and within effects. The first group averages the between and within effects together. This category includes pooled OLS, pooled OLS with panel corrected standard errors (Beck and Katz 1995), and random effects. The

Table 2.1: Example of Between and Within Parts of a Variable in TSCS Data

Town ID	Year	Population	Pop. Between	Pop. Within
1	2006	2500	3200	-700
1	2007	2700	3200	-500
1	2008	3200	3200	0
1	2009	3400	3200	200
1	2010	3800	3200	600
2	2006	6850	7480	-630
2	2007	7100	7480	-380
2	2008	7350	7480	-130
2	2009	7800	7480	320
2	2010	8300	7480	820

second category refers to methods which estimate between and within effects separately in the same model, and includes fixed effects with vector decomposition (Plumper and Troeger 2007), and the approach described by Yair Mundlak (1978). The third category estimates the within effects only, and ignores the between effects entirely. Much of the recent work in the development of TSCS methods in political science falls within this category. Fixed effects estimators are classified in this group, as well as several variants such as error correction models (Keele and DeBouf 2008). Finally, the last category includes models which estimate the between effects only, ignoring the within effects: the time effects estimator and the between estimator are the two most prominent examples in this group.

In addition to this taxonomy of the methods, it will be important to observe that almost all of these methods assume that their coefficient point estimates, whether between effects, within effects, or averages of the two, are fixed across cases and time. There are three notable exceptions. First, a particular specification of random effects allows the coefficients to vary across cases according to a normal distribution (Rabe-Hesketh and Skrondal 2008, p. 141-173). Second, change point models are fixed effects models in which the within effect is assumed to change at a particular point in time (Western and

Kleykamp 2004, Spirling 2004, Park 2009). Third, a time effects model in which the regressor is interacted with time allows a between effect to change over time, although in a highly constrained way.

The following discussion will consider each of these methods in greater detail, focusing on what each one actually estimates in the data.

2.3.1 Methods Which Average the Between and Within Effects Together

Pooled OLS

Pooled OLS is simply an OLS regression on the TSCS data, with no adjustment for the time or repeated cross-sectional nature of the data. Like all standard OLS models, the variance of the dependent variable is assumed to be a scalar value, rather than a quantity with separate time and cross-sectional dimensions. The model does not distinguish between the unit effects and the overall residual, so the model which is estimated is

$$y_{it} = \beta_0 + \beta x_{it} + \varepsilon_{it}. \tag{2.6}$$

If substitute for x_{it} using equation 2.5, the model becomes

$$y_{it} = \alpha + \beta(x_i^B + x_{it}^W) + \varepsilon_{it}$$

$$= \alpha + \beta x_i^B + \beta x_{it}^W + \varepsilon_{it}. \tag{2.7}$$

The common econometric objection to pooled OLS is that it returns biased standard errors by assuming that residuals within a case are uncorrelated (Cameron and Trivedi 2005, p. 702). But there is another problem to consider. The effects of the between and

within parts of x_{it} on y_{it} are constrained to be equal. As a result, β_1 is a weighted average of the between and within effects of x_{it}. The weights depend on the amount of cross-sectional variance in y_{it} compared to the time variance. If the between and within effects of x_{it} should be different for substantive reasons, then it is unclear from the pooled OLS results alone which effect β_1 most resembles, and whether the average has any meaningful substantive interpretation.[2]

Pooled OLS with panel corrected standard errors

One recommendation in political science to correct the estimates from pooled OLS is to use panel corrected standard errors, or PCSEs (Beck and Katz 1995). It is important to note that PCSEs only change the estimates of standard errors, and do not change the coefficient point estimates from pooled OLS. PCSEs correct the standard errors for heteroskedasticity and contemporaneous error correlation, which is correlation between the errors for two different panels at the same point in time. This approach does not address any bias that may be present in the pooled OLS coefficient point estimates. Note that in the example in section 2.5 the coefficient estimates using panel corrected standard errors are the same as the point estimates for pooled OLS, but the standard errors are different. If we suspect that coefficients are biased for the pooled OLS model, then they must be biased for the PCSE model as well.

Random effects

Random effects treat the case-to-case differences u_i as a separate part of the residual,

[2]Furthermore, we lose the ability to accurately control for mediating variables for both the between and within effects, since the mediation effects are constrained.

where:

$$u_i \sim N(0, \sigma_u^2). \tag{2.8}$$

The unit effects u_i approximate an intercept which varies across the cases. Random effects estimators will compute the between variance, σ_u^2, the within variance, σ_e^2, and the ratio between the two. Like pooled OLS, effects are weighted averages of between and within effects, although the weights are calculated through a slightly different formula. Random effects are commonly used by researchers who want to account for the unit effects in the data while allowing the effects of time-invariant predictors to be estimated. However, in general, random effects place more weight on the within effects of each covariate than pooled OLS does. A more general specification of random effects allows the estimated coefficients to vary across the cases (Rabe-Hesketh and Skrondal 2008 p. 155), but these case-specific effects are assumed to vary normally around the effect derived in a standard random effects model, which remains an average of the between and within effects.

2.3.2 Methods Which Estimate the Between and Within Effects Separately and Within the Same Model

Mundlak's approach of separating the between and within parts of each independent variable

Mundlak (1978) recommended running a random effects model in which each predictor is broken into its between and within parts:

$$y_{it} = \alpha + \beta_1 x_i^B + \beta_2 x_{it}^W + u_i + \varepsilon_{it}, \quad \text{where} \quad u_i \sim N(0, \sigma_u^2). \tag{2.9}$$

Mundlak criticizes the pooled OLS and random effects estimators for averaging these effects, stating that "any matrix combination of the 'within' and 'between' estimates

is generally biased." The between effects provided a test of whether significant cross-sectional effects exist, and of whether the between and within effects are significantly different from each other, but the between coefficients themselves were not supposed to be interpreted. Two political scientists however, Zorn (2001) and Bartels (2009), have used Mundlak's approach and have explained the substantive interpretation of the between variables. The effect of the between variables can be interpreted as the effect of a one-unit change in x between cases at a typical time point in the data.

Fixed effects with vector decomposition

Fixed effects with vector decomposition (FEVD) is a method designed by Plumper and Troeger (2007) which estimates a model in three steps. First, a fixed effects model is run and the case-level averages, which are estimates of the unit effects, are saved. Variables which are fixed across time or have little variance over time are excluded from this first step. Second, these unit effect estimates are regressed on the independent variables which are time-invariant or rarely changing over time. The residual, which is called a decomposed unit effect (DUE), is saved from this regression. Finally, a pooled OLS model is run which includes all of the time-variant, time-invariant, and rarely changing predictors, as well as the DUE.

The idea behind FEVD is that the controversial assumption of random effects, that the unit effects are uncorrelated with all of the predictors, can be circumvented by separating the time-fixed variables from the unit effects. The unit effect estimates produced by the fixed effects model represent the unmodeled cross-sectional variance in the data. By regressing the unit effects on the time-invariant predictors, the residual DUE by definition is uncorrelated with the time-invariant predictors. In the final model, the DUE theoretically contains all cross-sectional variance outside of the time-invariant predictors.

For the variables which are specified as time-invariant or rarely change, FEVD estimates an effect from a random effects model, which theoretically resembles a between effect since there is little or no within variation. For every other variable, FEVD produces a within estimate only. Ideally, they should return the same values as the estimates from the fixed effects model. These estimates, however, are not exact since the DUE is derived from a regression on estimated unit effects. During each of the three steps, estimation error can throw final estimates slightly off. More importantly, FEVD does not estimate between effects for variables which vary over time and across cases.

2.3.3 Methods Which Estimate the Within Effects Only

Fixed effects

Like Mundlak's method, fixed effects estimators disaggregate the between and within variation. The interpretation of results from a fixed effects model are unambiguous, because the between variance is eliminated from a fixed effects model entirely: unit effects are directly subtracted out of the model. This behavior of fixed effects is not seen as a disadvantage, but as its best characteristic. Econometricians are always concerned with minimizing the variance of the residuals in a linear model, and fixed effects remove a significant portion of this variance by design (Cameron and Trivedi 2005, p. 704).

The dependent variable in a fixed effects model is constrained to be the within version of y_{it}, and this constraint is mathematically equivalent to including a dummy variable for each case. These dummy variables control for, but do not explain, the between variation. Since the between variation drops out of the model, all time-invariant predictors drop out as well. Fixed effects models can only answer questions about how changes in an independent variable over time associate with changes in a dependent variable over time, and because time-invariant predictors never change over time, they have no explanatory

value by construction.

Proofs that fixed effects are consistent should not be taken as a directive to use fixed effects by political scientists whose research questions consider the differences between cases. Fixed effects are consistent only for within estimates, and say nothing about the between-panel effects. Various reformulations of the fixed effects model, such as error correction models which account for autocorrelation and time trends within the residuals (Keele and DeBouf 2008) are still fixed effects models, and therefore continue to have no bearing on questions regarding the differences between panels.

Change point models

Generally, the question of when the effect of an independent variable on a dependent variable changes over time is the focus of a large statistical literature on Bayesian change-point models (Chib 1997). Whether an effect changes at a particular time point can be tested in a time series by interacting the independent variable with an indicator for observations after the time point, as in the following regression equation:

$$y_t = \alpha + \beta_1 x_t + \beta_2 x_t \times I(t > 1980) + \varepsilon_t. \tag{2.10}$$

In this equation, the function $I(t > 1980)$ is 1 if the year under consideration is later than 1980, and 0 otherwise. The effect of x on y is β_1 for years before 1980, and $\beta_1 + \beta_2$ for years after 1980. When β_2 is significant, then 1980 can be called a change point. Bayesian change-point models try to determine the probability that any of the time points in the data are change points, and change point models for TSCS data in political science have been discussed by Western and Kleykamp (2004), Spirling (2007) and Park (2009). It is important to note, however, that all of these models, including the applications to TSCS data, posit change points on within effects only. These models are placing interactions

on the coefficients of fixed effects models, and are treating the between effects like noise.

2.3.4 Methods Which Estimate Between Effects Only

The Between Estimator

The between estimator is designed to analyze between effects only, but does so in a way that results in a dramatic loss of degrees of freedom. Every variable is averaged within cases over the time span, then the regression is performed. For example, if the data contain information about 10 cases over 10 time points each for a total of 100 observations, then the between estimator averages every variable across the 10 time points, and leaves only 10 observations. While this estimator provides unambiguous between effects, it is too inefficient to be of much practical value. It should be noted, however, that the between variables are constructed in the same way as in Mundlak's method. Each resulting between effect returned by Mundlak's estimator or the between estimator can be thought of as the "effect of an averaged between variable."

Time Effects

Time effects work in the same way as fixed effects, only dummy variables are included for the time points rather than for the cases. The interpretation of time effects is precisely analogous to the interpretation of fixed effects: the dummy variables completely account for, but do not explain, the within variation. The remaining effects are between effects only. We can think of the time effects as removing the average of the dependent variable within each time point, thereby subtracting out a global trend from the data. Like the coefficients from the between estimator, the coefficients derived using time effects are unambiguously between effects. However, unlike the between estimator, time effects do not

result in a huge loss of degrees of freedom. Cross-sectional effects are considered within each time point, and these effects are then averaged to produce one result. Where a between coefficient under Mundlak's method and the between estimator is the "effect of an averaged between variable," the coefficient produced under time effects can be thought of as the "average of the effects of the between variable."

Note that both the between estimator and time effects return between effects which are fixed over time. If we had run a model involving a time frame from 1950 through 2010, these approaches would have yielded the same result for 1950 as for 2010. In order to dispel this assumption in a time effects model, the regressor can be interacted with some specification of time. However, the way in which the between effect will change over time will be highly constrained by the functional form of the interaction. If the regressor is interacted with linear time, the between effect is assumed to change linearly over time, If the regressor is interacted with quadratic or cubic time, the between effects are allowed to change in more flexible ways, but degrees of freedom are sacrificed. These models are the closest relatives to BEER, and are tested against BEER in the simulations in section 2.7.

It is not the case that these methods are all approximations of the same correct answer, and it is not the case that the results from one method are likely to resemble the results from the others. These are methods which estimate fundamentally different parameters in the data. In table 2.3 in section 2.5.2 each of the methods described in this section are used to estimate an example regression, and the results are strikingly different. The point estimates have very different magnitudes and inferences, and in some cases, different signs. Researchers need to pay very close attention to the methods they choose to utilize.

The above taxonomy can provide guidance to researchers who must choose a method

to analyze TSCS data. However, the methods described above do not cover every situation a researcher with TSCS data is likely to encounter. What's missing? None of these approaches provide satisfying results for a researcher who is concerned with the differences between panels at recent, or at particular points in time, or who want to model a between effect that changes over time, or who are simply not comfortable with the assumption that the effect of a cross-sectional difference is the same at the beginning of the time period as at the end of it. In the next section, I develop a methodology which is designed to analyze between effects directly, and allows these effects to change over the course of the time period measured in the data.

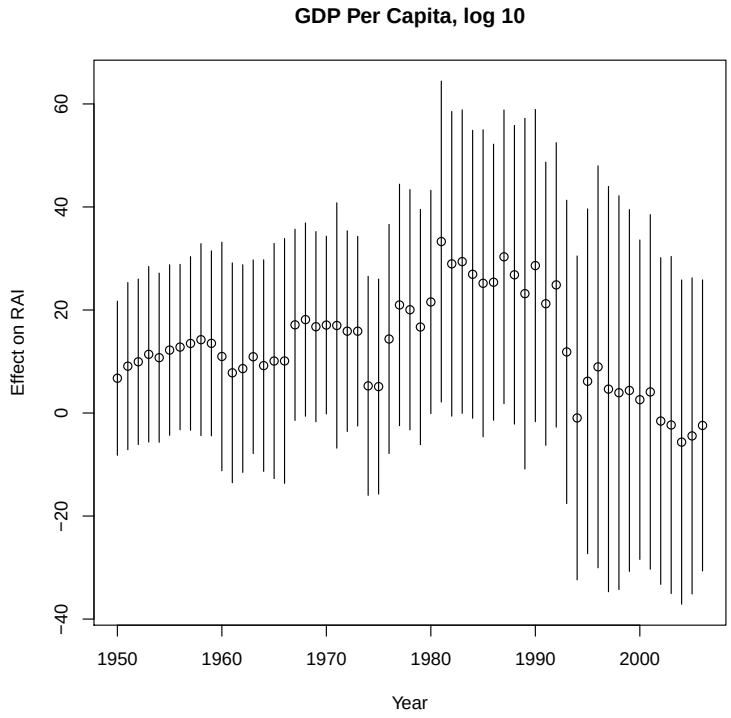
2.4 Methodology

The easiest way to obtain between effects is to perform a simple OLS regression on a cross-section. If we consider only one time point in a TSCS dataset, then time is guaranteed to be controlled in the model, and therefore there is no risk of confounding the between effects and the within effects. These regressions will often suffer from the inefficiencies of a small sample size and the results will not be satisfying: not because variables of interest necessarily fail to achieve significance, but because we have so much more information than is contained in one time point's sample size.

If we repeat this analysis for each time point in the data, we may see similar results, but we may see changing results. What is guaranteed is that, no matter what the results are, the standard errors will be higher than they should be given the total amount of data we have. Even if we find a significant effect, we should be able to be even more conclusive about the magnitude of that effect. Figure 2.2 demonstrates this point. The data used to create this graph comes from a dataset describing the authority of regional governments within 21 countries from 1950 to 2006. These data are described in detail in section 2.5. One of the predictors of regional authority is the GDP per capita of the

country. The dots in the graph in figure 2.2 represent the coefficient point estimate for logged GDP per capita from a regression of regional authority, using only the data points from a particular year. Separate estimates are plotted for each of the 56 years, and the vertical lines through each point represent the 95% confidence interval for each of these estimates.

Figure 2.2: OLS Coefficient Point Estimates for GDP Per Capita, with 95% Confidence Intervals, From 1950 to 2006



The intuition suggested by figure 2.2 is that there is valuable information which can be gained by looking at the effects over the entire timespan, and that there is a way to leverage this information to improve our inferences about the between effects. Until about 1992, figure 2.2 illustrates a relationship between wealth and federalism which is very strong and steadily becoming stronger, even though few individual cross-sectional results are significantly different from 0. The figure also shows a drastic change after 1992 which researchers will want to examine closely, although these effects individually

are not significantly different from the stronger ones in the 1980s.

The method formulated here is called the between effects estimation routine. BEER depends upon two observations about between effects: first, repetition of an effect over time provides more evidence about the size of the effect than any one time point alone can provide; second, between effects may change over time, and therefore between effects need to be calculated with respect to a particular time point.

We must make a decision about what information is appropriate to use in drawing an inference about a particular time point. In figure 2.2, we might decide that each year contains unique information which is unrelated to the information in any other year. If that is the case, individual regressions are the only appropriate method for analyzing the data, and we must accept the large standard errors as true reflections of our uncertainty. If we make an assumption at the other extreme, we can decide that every year is equally informative about an underlying causal process which is fixed across this time span. In this case, we draw one inference by considering every time point equally: the story in 1950 is the same as the story in 2006. Time effects treat the data this way, and would obscure the changes which occurred in the 1990s.

In both of these situations, we are dealing with extremes. It is not likely that the results in each time point are completely unrelated to each other, nor is it likely that the causal process is fixed over time. BEER is a method which captures the more realistic middle ground: that the between effect of a variable at a particular time point is related to the effect of the variable at proximate and at similar time points, but is less related to the effect of the variable at time points which are more dissimilar and less proximate. In order to inform the estimate of a between effect at a particular time point, the results from different time points are weighted and are averaged using these weights. The weight for a particular time point depends on three considerations:

1. Similarity: time points which are empirically similar to the time point of interest

should be weighted more heavily than time points which are less similar.

2. Proximity: time points which are proximate to the time point under consideration should be weighted more heavily than time points which are further away.

3. Sample size: time points with larger cross-sectional sample sizes should be weighted more heavily than time points with smaller sample sizes.

Weights take the form $w_t n_t$, where n_t is the cross-sectional sample size in a time point t, and $w_t \in [0, 1]$ is a discount factor which depends on the similarity and proximity of time point t to the time point under consideration in the analysis. Using these weights, we can calculate an "updated" sample size. If $w_t = 1$ for every time point t, then every time point is weighted only according to its corresponding n_t, and the updated sample size is the total number of observations in the data. If $0 < w_t < 1$ for at least some time points t, then the updated sample size is bigger than the sample size within the time point under analysis, but is less than the total number of observations in the data. A time effects model is a special case of BEER in which $w_t = 1$ for every time point t, and individual regressions are another special case of BEER in which $w_r = 1$ for the time point r under consideration, and $w_s = 0$ for every other time point $s \neq r$.

The use of weighted averages is not arbitrary. These averages are implementations of a Bayesian learning algorithm which uses conjugate priors, as described in texts by Jeff Gill (2008, p.81-84), Andrew Gelman et al (2004, p. 87-88), and Christian P. Robert (2001, p. 189-190). Conjugacy makes it possible for these averages to be computed quickly and exactly, without relying on computationally intensive simulation techniques. The prior and posterior distributions used to compute the averaged point estimates and standard errors in a time point are listed in appendix C. A more general discussion of how BEER operates is listed here.

To calculate similarity, an independent-samples t-test allowing unequal variances and sample sizes is used which tests the null hypothesis that a point estimate in time point

j, $\hat{\beta}_j$, with estimated variance $\hat{\sigma}_j^2$ and sample size n_j is equal to the point estimate from the time point k, $\hat{\beta}_k$, with variance $\hat{\sigma}_k^2$ and sample size n_k. The test statistic is

$$t_{j,k} = \frac{\hat{\beta}_j - \hat{\beta}_k}{\sqrt{\frac{(n_j-1)\hat{\sigma}_j^2+(n_k-1)\hat{\sigma}_k^2}{n_j+n_k-2}}\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}}, \tag{2.11}$$

which is distributed by the student's T distribution with $n_j + n_k - 2$ degrees of freedom. To make the weights less sensitive to small differences, the t statistic is divided by 10, and the p-value is calculated. This calculation is performed for each covariate in the analysis, and the mean of all of the p-values is taken as the measure of the similarity between two time points.

Proximity is now included in the calculation by taking the absolute difference between each time point and the time point under consideration, in terms of the unit of time being used. For example, the proximity score for the comparison of years 2000 and 2010 is 10 years if our data are annual, and 2 if we have taken measurements at 5 year intervals. This proximity score is set as the exponent of the similarity score, so that the weights are lower as years become less proximate.

The total updated sample size can be easily calculated after computing similarity and proximity. For example, in figure 2.3, data from 1950 through 1954 are used, and 1952 is the year under consideration. The years 1950-1954 each include a sample size of 10, and have similarity scores with 1952 of 0.8, 0.6, 1, 0.85, and 0.5 respectively. To calculate an updated sample size, each year's similarity score is raised to its proximity to 1952. The contribution of 1950 to the updated sample size for 1952 is $10 \times (0.8)^2 = 6.4$. 1951 contributes $10 \times (0.6) = 6$, 1952 contributes its entire sample size of 10, 1953 contributes $10 \times (0.85) = 8.5$, and 1954 contributes $10 \times (0.5)^2 = 2.5$, for a total updated sample size of 33.4.

Coefficient point estimates at time point t_r are calculated by taking a weighted average

Figure 2.3: An Example of How to Calculate the "Updated" N for OLS Results in 1952.

1950 → 1951 → **1952** ← 1953 ← 1954

N	10	10	10	10	10
Similarity to 1952	0.8	0.6	1	0.85	0.5
Years away from 1952	2	1	0	1	2
Weight	$(0.8)^2$ =0.64	$(0.6)^1$ =0.6	$(1)^0$ =1	$(0.85)^1$ =0.85	$(0.5)^2$ =0.25

Total "updated" N = 0.64(10) + 0.6(10) + 1(10) + 0.85(10) + 0.25(10) = 33.4

of the OLS coefficient point estimates from every time point in the set $\{t_1, \ldots, t_T\}$:

$$\beta_{t_r} = \frac{\sum_{j=1}^{T} \hat{\beta}_{t_j} w_{t_r,t_j} n_{t_j}}{\sum_{j=1}^{T} w_{t_r,t_j} n_{t_j}}, \tag{2.12}$$

where n_{t_j} is the sample size in time point t_j, $\hat{\beta}_{t_j}$ is the vector of OLS coefficient point estimates from time point t_j, and

$$w_{t_r,t_j} = {s_{t_r,t_j}}^{p_{t_r,t_j}}, \tag{2.13}$$

where s_{t_r,t_j} and p_{t_r,t_j} are respectively the similarity and proximity scores between time points t_r and j. A more detailed discussion of this formula and the calculation of standard errors is presented in appendix C.

By averaging, both the coefficient standard errors and the point estimates will be different from the OLS results within a single time point. If many time points yield similar results to the time point of interest, then the average has the effect of increasing

the model sample size, thereby lowering the standard errors. In other words, we will become more certain about the size of an effect if we see it duplicated again and again. If proximate time points yield different results than the time point under consideration, then the point estimates are adjusted to better reflect a current trend. If one time point is an aberration, then we want to be cautious about accepting evidence from that time point as the truth if things return to "normal" later. If we observe wildly different effects over time, then we should accept less evidence from the other time points.[3]

Change points for the between effects

BEER can also provide evidence to help researchers find the time points in which the between effects change.[4] Similarity scores are computed for every pair of distinct time points, and a $T \times T$ matrix which contains these similarity scores is built. Each score is subtracted from 1 to produce a dissimilarity matrix for the data, which can be analyzed using many different statistical measurement techniques. Latent dimensions of the variance among time points can be estimated, or latent groups of time points can be identified.

Problems with BEER

As with any method, researchers need to be aware of some problems that arise when

[3]There are several conceptual similarities between BEER and Locally Weighted Scatterplot Smoothing (LOWESS), which is a non-parametric technique for estimating a linear regression model (Cleveland and Devlin 1988). Both methods are best illustrated pictorially: BEER on a plot of cross-sectional coefficient estimates over time, and LOWESS on a scatterplot. Both also utilize a weighted average of proximate points to derive an estimate. However, the mathematical formulations of each method are similar only to the extent that both are weighted averages. BEER is derived from the parametric Bayesian literature, while LOWESS is not Bayesian and is non-parametric.

[4]When we derive change points for between effects, we have an informational advantage over change point models for within effects. Models that rely on fixed effects consider variables that are known and have no standard error. In contrast, the similarity score introduced in equation 2.11 takes the differences in point estimates into account, but also considers the standard errors of each estimate.

using BEER. First, since the method depends on a series of cross-sectional OLS models, potentially with very small sample sizes, the number of covariates must be less than the minimum number of observations in any one year. This requirement may present severe difficulties especially when some data are missing and are treated with listwise deletion. Additionally, due to collinearity with the constant, any variable that is fixed across cases but varies over time will drop out. Variables which are constant across cases in even one time point, for example an indicator variable for EU membership before the EU exists, will not be estimated at time points in which there is no cross-sectional variation. Sharp changes in point estimates may be due to the addition or loss of cases rather than a more significant change in the underlying data generating process. Finally, although alternative OLS specifications such as robust or cluster-robust standard errors are currently allowed, the method has not yet been formulated for limited dependent variables and general linear models.

In section 2.5, BEER and the other methods are used to examine the cross-sectional determinants of regional authority. In section 2.6, BEER is used to derive the effect of a state's median income on its voting behavior in U.S. presidential elections. In both examples, BEER is used to examine the results for potential change points. In section 2.7, BEER and a set of competing estimators are run on simulated data, and the results from each method are compared for accuracy.

2.5 Example 1: Regional Authority in 21 Countries, 1950-2006

This example demonstrates the use of BEER in comparative politics, and focuses on federalism in a comparative context. In order to demonstrate the characteristics of each

commonly used estimator for TSCS data, the same linear model is estimated using each of these common methods as well as BEER. The different estimators provide shockingly discordant results, which can be confusing unless a researcher pays close attention to what each method actually estimates in the data.

2.5.1 Data and Model

The dependent variable is the regional authority index (RAI) compiled by Gary Marks, Liesbet Hooghe and Arjan H. Schakel (2008), which measures the strength of regional governments within a country relative to the federal government. The index is provided for 21 countries, measured annually, beginning in 1950 and ending in 2006. No countries enter or drop out of the analysis during this time span.[5] Table 2.2 contains a list of the countries included in the analysis and the over-time average value of the RAI for each country.

Table 2.2: Countries in the Sample and Average Regional Authority Index

Country	RAI	Country	RAI	Country	RAI
Australia	18.3	Greece	3.1	New Zealand	5.6
Belgium	22.3	Iceland	0.0	Norway	7.4
Canada	22.6	Ireland	1.5	Sweden	11.3
Denmark	7.9	Italy	13.6	Switzerland	19.5
Finland	2.6	Japan	8.2	Turkey	4.2
France	10.8	Luxembourg	0.0	United Kingdom	9.1
Germany	29.1	Netherlands	13.7	United States	23.1

There are many theories of why some countries adopt greater levels of regional authority than others. Federalism is expensive to run successfully, so we can expect that countries with higher levels of RAI are also wealthier. Some countries may have higher

[5]The number of countries is kept constant across time here so that changes in between effects over time cannot be due to cases entering or dropping out of the analysis. All of the methods discussed and presented in this paper, however, can be used on unbalanced panels as well as balanced panels.

levels of regional authority because strong regions are needed for administrative purposes: countries which are geographically large have more land to administer, and countries which have large populations have more people to serve. Therefore countries with greater areas and larger populations should also have higher levels of regional authority. A country may also have greater regional authority because segments of the population have a desire for strong regional governance. Specifically, a country which has regions where a majority of people are part of an ethnic minority, such as Quebec in Canada, should have greater regional authority. In addition, regional-based political parties have an incentive to institute greater regional authority to increase their political power. Finally, authoritarian regimes will be less willing to cede power to regional governments, so we expect countries which are more autocratic to have lower levels of RAI.

Note that these hypotheses speak only about the differences between countries. Theories regarding changes within countries are different questions. We can hypothesize about the effects when a country becomes richer over time, or larger, or more populous, ethnically fragmented, politically regionalized, or democratic. However, it is not at all clear that these effects should be necessarily equal, similar, or even in the same direction as the between effects suggested above.

For the analyses presented here, RAI is regressed on six independent variables:

1. the country's GDP per capita in each year, measured in 2006 U.S dollars,
2. the country's population in each year, in thousands of people,
3. the country's area, in units of 1000 sq. km.,
4. a measure of the country's ethnic fragmentation which expresses the probability that two people chosen at random from the country belong to different ethnic groups,

5. the proportion of seats in the country's federal legislature in each year which are controlled by regional political parties, and

6. the country's Freedom House democracy score in each year.[6]

These data were collected by Gary Marks and Liesbet Hooghe, and are used here with their permission. The measure of ethnic fragmentation was compiled by Fearon (2003), which considers one time point for each country. To account for the skewness of population, area, and GDP per capita across countries, the common (base 10) logarithm of these three variables is used. The interpretation of the effects of these variables now changes: a one-unit increase in a $\log_{10}$ transformed variable is equivalent to a 10-fold increase in the untransformed variable. Since data on ethnic fragmentation is not measured over time, this variable is treated as a time-invariant predictor in the analysis.

2.5.2 Results Using Commonly Used TSCS Methods

The regression described in section 2.5.1 is run using each of the following TSCS estimators: pooled OLS, pooled OLS with panel-corrected standard errors, random effects, fixed effects, fixed effects with vector decomposition, Mundlak's method, the between estimator, and time effects. The results are listed in table 2.3. In the table, separate rows are provided for estimates of between and within effects. When the effects are averaged, the results appear in the middle of the two rows. When the results only estimate the within effects, they appear on the within row, and when they only estimate the between effects they appear on the between row. The results from methods which return both between and within effects are listed on both the between and within rows.

Table 2.3, as a whole, provides an illustration of the behavior of each estimator. Pooled OLS and random effects are different weighted averages, typically - but not always

[6]Freedom House uses a coding scheme that awards lower scores to the most democratic countries, but this scale is reversed for this study.

Table 2.3: Results for the Regional Authority Regression.

Variable		POLS	PCSE	RE	FE	FEVD	Mundlak	BE	TE
GDP p.c., $\log_{10}$	B	2.08(.32)***	2.08(.18)***	2.78(.20)***	-	-	17.8(13.3)	17.81(15.8)	14.79(1.1)***
	W				3.62(.24)***	3.62(.13)***	3.62(.24)***	-	-
Pop., $\log_{10}$	B	6.02(.26)***	6.02(.09)***	-1.74(1.2)	-	-	6.15(1.71)***	6.15(2.04)***	6.27(.25)***
	W				-7.57(1.6)***	-7.57(.19)***	-7.57(1.6)***	-	-
Area, $\log_{10}$	B	1.09(.23)***	1.09(.08)***	2.88(1.4)**	-	7.65(.12)***	1.58(1.6)	1.58(1.8)	1.24(.22)***
	W				-5.67(3.2)*	-	-5.67(3.2)*	-	-
Ethnic frag.	B	11.8(.90)***	11.8(.51)***	16.4(6.5)**	-	15.2(.36)***	6.85(6.6)	6.85(7.8)	9.10(.90)***
Reg. party seats	B	30.3(2.1)***	30.3(1.5)***	18.6(1.3)***	-	-	45.5(19.4)**	45.51(23.1)*	33.82(2.04)***
	W				17.5(1.3)***	17.5(.84)***	17.5(1.3)***	-	-
Freedom House	B	1.79(.12)***	1.79(.12)***	0.23(.07)***	-	-	0.63(2.0)	0.63(2.4)	0.82(.15)***
	W				0.16(.07)**	0.16(.05)***	0.16(.07)**	-	-
Constant		-65.8(2.3)***	-65.8(1.8)***	-1.08(7.7)	62.0(12.4)***	26.5(1.5)***	-113.7(33.1)***	-113.75(39.5)**	-93.21(3.4)***
Groups		-	21	21	21	21	21	21	21
$N \times T$		1197	1197	1197	1197	1197	1197	21	1197
σ_u		-	-	5.21	15.7	-	4.36	-	5.04
σ_e		-	-	2.05	2.05	-	2.05	-	6.40
ρ		-	-	.866	.983	-	.818	-	.382
DUE		-	-	-	-	1(.01)***		-	-
R^2 between		-	-	.290	.220	.738	-	.738	.676
R^2 within		-	-	.409	.420	.420	-	.361	.959
R^2 overall		.645	.645	.301	.164	.946	.708	.404	.446

Note: FEVD treats area, which is rarely changing, and ethnic fragmentation as invariant.
* $p < .1$, ** $p < .05$, *** $p < .01$

- falling between the fixed effects and time effects estimates. Note that the coefficient estimates using PCSE are the same as the point estimates for pooled OLS, but that standard errors are different. If we suspect that coefficients are biased for the pooled OLS model, then they must be biased for the PCSE model as well. The coefficients and standard errors for random effects, however, are different from the results from the pooled OLS model since it expresses a different weighted average. The estimates from fixed effects are consistent, but only for the within effects. Note that ethnic fragmentation, which is time-invariant, drops out of the fixed effects model.

The FEVD model used here specifies that the between effects for both ethnic fragmentation, which is time-invariant, and area, which is rarely changing, be estimated. These estimates might suffer from omitted variable bias because the between effects for the other variables remain excluded. The effects for all of the other variables in this implementation of FEVD are within effects, and are very similar, though not exactly equal, to the fixed effects estimates. The standard errors for these estimates are also very low compared to the fixed effects estimates, and the R^2 returned by the model is much higher.

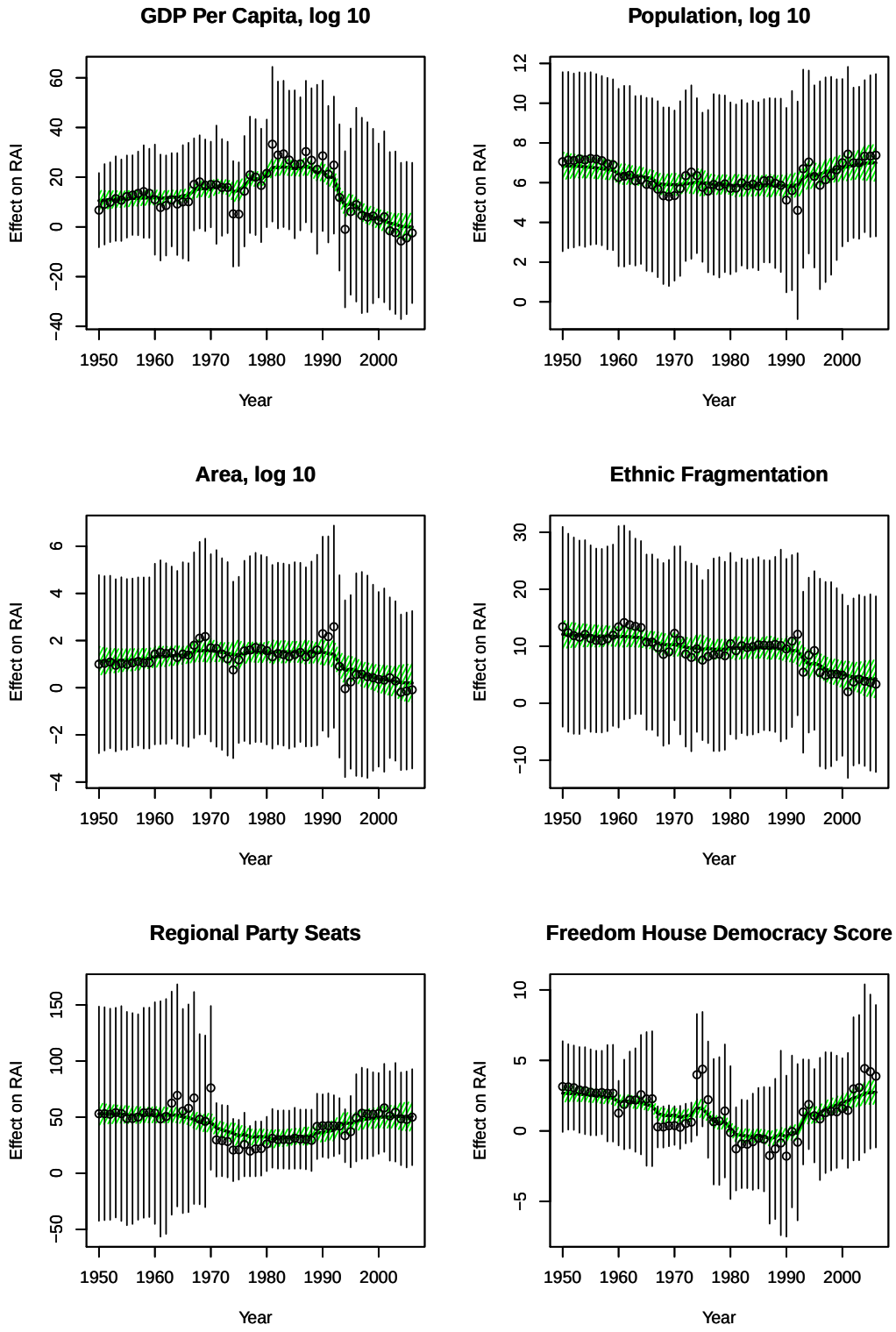
In Mundlak's model, note that the coefficient point estimates and standard errors for the within variables are nearly identical to the results from the fixed effects model. Also note that the Mundlak results for the between effects are nearly identical to the results from the between estimator, although the standard errors are slightly smaller. The between estimator and time effects both estimate between effects only, and while the two methods yield similar point estimates, notice that the standard errors for the time effects are much smaller than the standard errors for the between estimator. The between estimator takes the average of each variable across the 56 time points, yielding estimates from only 21 cases.

2.5.3 Results Using BEER

The regional authority index is regressed on the common logarithms of GDP per capita, population, and area, as well as the measures of ethnic fragmentation, seats controlled by regional parties, and the Freedom House democracy score within each of the 56 years in the data. BEER uses these OLS estimates to compute between effects in every year from 1950 through 2006. The dots in the graphs in figure 2.4 represent the OLS estimates in every year, and the vertical lines represent the 95% confidence intervals for each point estimate. The between effect estimates with updated sample sizes are overlaid as a line on the OLS results, and the 95% confidence regions for the estimates are shaded. These graphs show how the between effects use more information than the OLS estimates, and therefore have lower standard errors. But they also show how the between effects change over time. It is immediately clear by looking at the graphs that the effects of GDP per capita, area, and ethnic fragmentation diminish over time. It is also clear that the between effects of regional party seats and the Freedom House score are weakest in the middle to late 1980s, but recover by the 2000s.

For the most part, the between effects confirm the hypotheses laid out in section 2.5.1. In the 1960s and 1970s, all of the variables have effects in the hypothesized direction and achieve significance. It appears that population is a powerful logistic reason for regional authority, while area does not exhibit as strong of an effect when controlling for population. It also appears as though ethnic fragmentation is becoming a less important determinant of regional authority, especially when compared to the consistently strong effect of regional parties. This may indicate that ethnic conflicts are increasingly becoming institutionalized through political parties. Finally, the effect of the Freedom House democracy score is the most difficult one to understand, but the disappearance of the effect in the 1980s and 1990s may be due to a lack of variation in the quality of democracy across the 21 cases during this time.

Figure 2.4: Between Effect Coefficient Estimates, From 1950 to 2006



One important observation about the trends illustrated in figure 2.4 is that the effects for each independent variable do not move independently of one another; rather they tend to move together. The effects are basically stable for the first 15 years or so, and thereafter changes in the effects of different independent variables seem to occur around the same time. It is possible that these changes are a result of a changing underlying process which generates the observed distribution in regional authority. It seems as if there are eras in the data, and a measurement model can be used to gain some insight into the locations of these eras. As discussed in section 2.4, we can compute similarity scores for each pair of years in the data and build a similarity (or dissimilarity) matrix for the years. If the years can be aligned on a latent structure, then a method such as factor analysis or multidimensional scaling can provide an estimate of the dimensionality of the latent space and of how the years are aligned on this space. Cluster analysis can provide an estimate of how the years are grouped on this space.

It is important to note that these similarity measures depend only on the similarity of OLS results between two years. The proximity of one year to another is not considered in this calculation. If two consecutive years are very similar, it is strictly because the data in one year resemble the data in the next year.

Table 2.4: Groups from Average Linkage Cluster Analysis on Year Dissimilarity, Six Group Solution

Group	Years
1	1950-1966
2	1967-1973, 1976-1980
3	1974, 1975
4	1981-1992
5	1994, 1995
6	1993, 1996-2006

For the regional authority data, similarity scores are calculated for every pair of years between 1950 and 2006. These scores are converted to dissimilarity scores and

are assembled in a matrix. To illustrate the capabilities of this approach, I perform an average linkage cluster analysis on this matrix. If we take a dissimilarity score greater than .1 to indicate a unique group, then there are 6 unique groups among the years in the data. These groups are listed in table 2.4. Note that although proximity is not specified as a criterion for similarity, the years are grouped for the most part in chronological order. Intuitively, these results make sense, and they accord with the trends we see in the graphs in figure 2.4. These groups should not be taken as a strict definition of the eras in the data. They can, however, inform additional theory by pointing to particular time points at which the underlying data generating process for the cross-section changes. The next step in the analysis of these data can be an investigation of special circumstances in 1966-1967, for instance, which lead to the changes we observe in the between effects.

2.6 Example 2: The Effect of Median Income on State-Level Voting in U.S. Presidential Elections, 1964-2004

This example demonstrates the use of BEER in American politics, where TSCS data often consist of the 50 states over a period of time. Specifically, the findings presented here confirm the conclusion of Andrew Gelman et al (2008) regarding the role of income on a state's voting behavior in Presidential elections. Gelman and his coauthors find that, while higher individual income is a significant predictor of an individual vote for a Republican, richer states are more likely to vote for the Democrat. Gelman et al point out that the distribution of income across the states has not changed very much since 1960, and even less since 1980 (p. 66). The interesting variation at the state level is between variation: we are comparing the rich states to the poor states rather than considering the effect of changes in state income. Furthermore, they find that the "systematic differences

between rich and poor states have largely arisen in the past twenty years" (p. 74). In order to accurately model the effect of income on state-level voting with a linear model, we require a method which considers the between effects - not the within effects, and not an average of between and within effects - and allows the between effects to change over time. Of the methods discussed here, only individual regressions for each time point and BEER handle this specification.

2.6.1 Data and Model

For each of the 50 states and the District of Columbia, the Democratic share of the two-party state-level vote is measured in each Presidential election from 1964 through 2004.[7] Vote totals are reported as percents, ranging from 0 to 100. The primary independent variable of interest is each state's income, measured as the median individual income in each state in each year, as published in the "State-Level Data on Income Inequality" dataset by Joshua Guetzkow et al (2007). The effect of income on state-level voting is conditional on income inequality and on the industrial makeup of the state. Inequality is operationalized as the GINI coefficient for each state in each year, as reported by Guetzkow et al (2007). To control for the industrial makeup of each state, the shares of state GDP in the manufacturing, education, and government sectors as measured by the U.S. Bureau of Economic Analysis (2011) are included.

In order to capture the between effect of income on state-voting while allowing this effect to change over time, the model is estimated using BEER and using individual regressions for each time point. The results are presented in section 2.6.2.

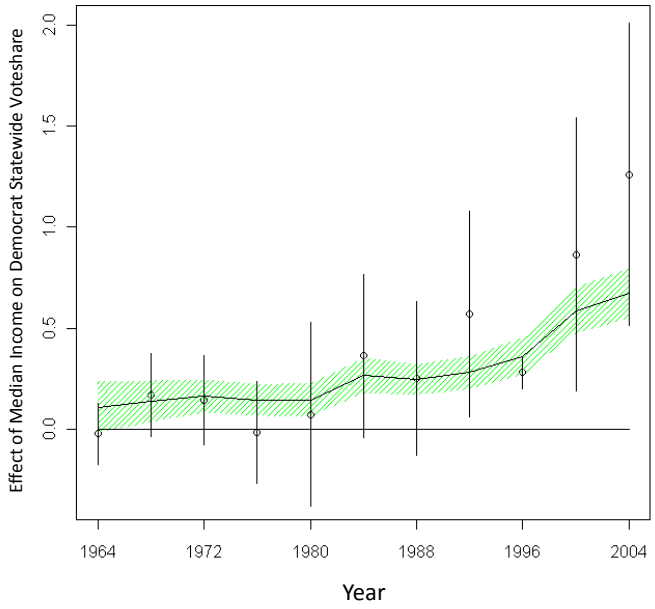
---

[7]The state-level voting data is acquired from Dave Leip's Atlas of U.S. Presidential Elections (2011). With one exception, all votes for third party candidates are removed from the data, and the percent of votes for the Democrat is normalized over the total number of votes for the Republican or the Democrat. In 1968, the votes for George Wallace are added to the votes for the Republican, Richard Nixon.

2.6.2 Results

Figure 2.5 contains the estimated effect of income on state-level voting in each election. The open dots represent OLS coefficient point estimates in a particular year, and the vertical line through each point represents the 95% confidence interval for the estimate. The moving solid line represents the point estimates produced by BEER, and the shaded region indicates the 95% confidence intervals for these estimates.

Figure 2.5: OLS and BEER Estimates of the Between Effect of Median State Income.



The individual regressions confirm the finding of Gelman et al that no significant effect of income is visible until later in the time series. Considered individually, the regressions do not demonstrate a significant result until the 1992 election. However, taken together, the regression results describe a situation in which income moves from a very small effect in the 1960s and 1970s to a much larger positive effect by the 2004 election. Since 1980, the between effect of income has grown steadily larger. The individual regressions are too conservative, since they do not individually confirm the larger story that their point estimates listed consecutively describe. BEER leverages this information, and

shows a significant and positive effect of income dating back to the 1968 election. This effect is estimated by BEER to have grown between the 1980 and 1984 elections, and monotonically since 1988.

If we apply a cluster analysis to the similarity matrix derived by BEER for the time points, and find a five group solution, then the clusters are defined as listed in table 2.5. As was the case with the regional authority example in section 2.5, even though the

Table 2.5: Groups from Average Linkage Cluster Analysis on Year Dissimilarity, Five Group Solution

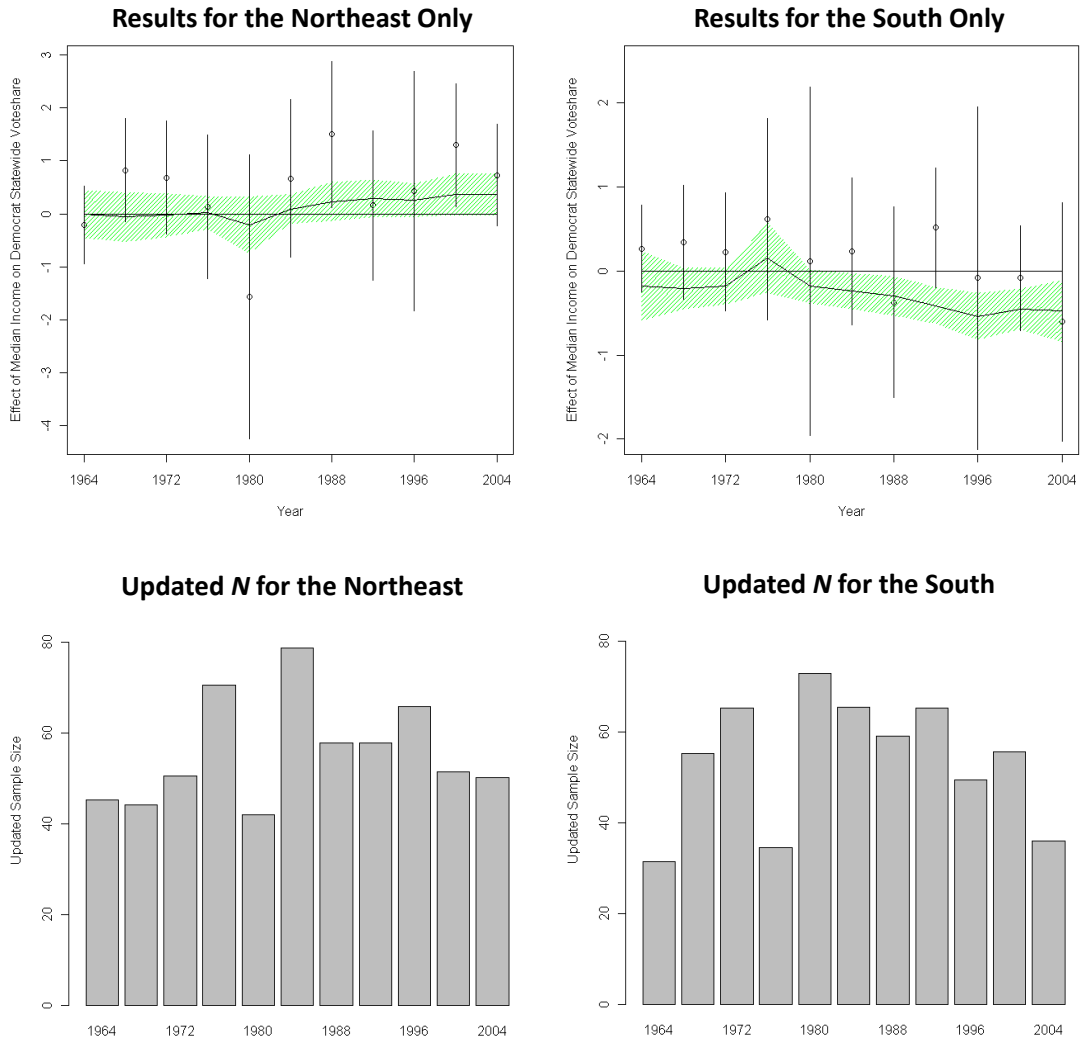
Group	Years
1	1964, 1968
2	1976, 1980
3	1984
4	1972, 1988, 1992, 1996
5	2000, 2004

similarity scores do not consider how proximate two time points are when calculating their similarity, the groups derived in the data largely contain consecutive time points. The only exception is 1972, which aligns more closely with 1988, 1992, and 1996. The fact that consecutive time points are also the most similar is evidence for the theory that the causal process which translates states' differences in income into differences in voting is changing over this time frame, slowly and deterministically. In other words, income is becoming more important over time as a factor which explains how the states vote differently.

The effect of income on state-level voting within regions

A subsequent research question may ask whether the relationship between income and voting really exists between states, or if this effect is manifest more accurately between regions. There might be a big difference between the South and the Northeast in regards

Figure 2.6: OLS and BEER Estimates of the Between Effect of Median State Income and Updated Sample Sizes.



Note: the non-updated sample size for the Northeast is 9, including CT, ME, MA, NH, NJ, NY, PA, RI, and VT. The non-updated sample size for the South is 10, including AL, AR, GA, KY, LA, MS, NC, SC, TN, and VA.

to both income and voting, but there may not be a real relationship between income and voting within these regions. We can begin to account for this possibility by including fixed effects for regions in the model. In that case, the model is now asking whether

richer states vote for the Democrat in greater proportions than the average for the region. Some regions, however, have very few states. California makes up most of what we usually think of as the west coast region, with Oregon, Washington, and maybe Arizona included. With so few observations in the regions, the available degrees of freedom are quickly used up. Taken to a greater extreme, we might want to interact the income effect with the regions if we posit that income has a greater effect in some regions than others. We may even want to run the regression selectively only on the observations within a particular region. In figure 2.6, the results for two subsequent analyses are illustrated. On the left, the model is run just for the 9 states in the Northeast, and on the right the model is run just for the 10 states in the South.

Within each time point, there is a sample size of 9 for the Northeast and 10 for the South. These sample sizes are not big enough to be able to distinguish between a real effect and a null effect in most cases, so individual regressions are not very informative. BEER, however, uses information from proximate and similar time points to update the N within each time point. The individual OLS point estimates and 95% confidence intervals are listed in the top two graphs of figure 2.6, along with the BEER point estimates and 95% confidence intervals. Notice how BEER derives a smaller confidence interval, but fails in either region to return the same positive and significant result that was derived for all 50 states and DC. In fact, BEER estimates the effect of income on voting in the South to be slightly but significantly negative since 1988. Since the theory is not supported within these regions while using more of the available information, we can be more confident that the relationship actually exists between regions rather than states. It is less likely that the null results are due simply to a lack of data.

The bottom two graphs in figure 2.6 show how BEER generates more power for the estimates by leveraging information from other years in the data. For the Northeast, although each cross-section has only 9 observations, updated sample sizes range from

approximately 40 to 80. In the South, the 10 observations are updated to sample sizes ranging from about 30 to about 70. These graphs demonstrate how BEER allows researchers with a small number of cases over time to derive between effects despite the small sample size in each cross-section.

2.7 Simulation

In this section, simulations are conducted to compare the properties of BEER to other linear models for TSCS data. Artificial data are generated according to one of four theoretical models. In each case, the data have known population parameters. BEER is run on the artificial data along with a set of commonly used TSCS methods to see how accurately each one returns the known parameters. The various specifications for the data are defined in section 2.7.1 and the methods through which data are generated are described in section 2.7.2. The competing models are listed in section 2.7.3, the measures by which they are evaluated are described in section 2.7.4, and a priori expectations for the simulations are laid out in section 2.7.5. The results are presented in section 2.7.6 and are discussed in section 2.7.7.

2.7.1 Data Generating Processes

The accuracy of a linear model depends on whether the assumptions it makes about the underlying data generating process are correct. Here, four data generating processes (DGPs) are specified, each one conforming to the assumptions made by a class of TSCS methods.

DGP 1: Between and within effects are equal; between effects (and within effects) are fixed over time.

Methods which take pooled approaches to the data posit that each independent variable has one underlying effect on the dependent variable, and that this effect is the same for variation over time and variation across cases. This DGP can be formally represented as follows:

$$y_{it} = \alpha + x_{it}\beta + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma^2), \tag{2.14}$$

where $i \in \{1, \ldots, N\}$ denotes cases, $t \in \{t_1, \ldots, T\}$ denotes time points, y_{it} is the dependent variable, x_{it} is a column vector of independent variables, α and σ^2 are scalar, and β is a column vector. Note that α and β are fixed across time points and cases. Homoskedasticity is assumed since σ^2 is also fixed across time points and cases. This DGP is the model for the data assumed by methods which average the between and within effects, such as pooled OLS. Similar methods, such as panel-corrected standard errors and random effects, make this assumption of causal homogeneity for the coefficients, but allow the residual variance to be non-constant and correlated within time points.

DGP 2: Between and within effects are different; between effects are fixed over time.

The second DGP which will be considered relaxes the assumption that the between effects and within effects are equal, but still requires between effects to be fixed over time. This DGP is expressed as

$$y_{it} = \alpha_t + x_{it}\beta + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma^2). \tag{2.15}$$

DGP 2 differs from DGP 1 only in that the constant is allowed to vary over time. If separate constants are estimated for each time point, then they act as fixed effects for the time points; this is the approach of the time-effects estimator. All temporal variation is

contained within these constants, so that the coefficients are strictly between estimates. However, these estimates remain constrained so that the effect is the same at all points in the time period.

DGP 3: Between and within effects are different; between effects exhibit smooth change over time.

The third DGP allows the between effects to vary over time in a smooth fashion. Coefficients are assumed to be positively correlated with their estimates from adjacent time points, so that changes from one time point to the next are relatively smooth. Formally, this model is

$$y_{it} = \alpha_t + x_{it}\beta_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_t^2),$$

$$Cov(\beta_t, \beta_{t-1}) > 0, \qquad \forall t > t_1. \tag{2.16}$$

The residual variance is also allowed to vary across time points. This model is the formulation of the data assumed by BEER.

DGP 4: Between and within effects are different; between effects change randomly over time.

Finally, if the assumption that consecutive between effects have positive correlation is relaxed, then it is possible that the between effects are independently drawn in each time point. In this case, none of the time points can provide any information about any other

time point. This model is written as

$$y_{it} = \alpha_t + x_{it}\beta_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_t^2),$$

$$Cov(\beta_j, \beta_k) = 0, \qquad \forall j, k \in \{t_1, \ldots, T\}. \tag{2.17}$$

Only individual regressions for each time point can analyze these data without making false assumptions which will lead to inconsistent estimates.

2.7.2 Simulated Data

The most important step in generating data is the creation of a dependent variable that has known properties. The four DGPs described in section 2.7.1 differ in their assumptions about the construction of the dependent variable; they all, however, describe the dependent variable as the sum of a deterministic function of independent variables and noise. In a simulation, the coefficients in the deterministic part of the DGP are specified first. An accurate method will provide close estimates to these "true" values. This section describes the methods through which the true parameters are chosen.

The dependent variable is simulated, but following the recommendation of Macdonald, Rabinowitz, and Listhaug (2007), all other aspects of the artificial data are drawn from real political data. These simulations are modeled on the state median income and Presidential voting data presented in section 2.6, which contains information about each of the 50 states and the District of Columbia in each Presidential election between 1964 and 2004. The dependent variable to be generated is the Democratic share of the two-party vote in each state. Values for the independent variable - median state income - and the controls - state income inequality (GINI), the share of each state's GDP in the

manufacturing, education, and government sectors - are taken from the actual data.[8]

The real Democratic vote share is regressed on median state income and the control variables. Two versions of this regression are performed: the regression is run once on the pooled data including all of the time points; and the regression is run once for each year, using only the data within that year. Four sets of simulations are conducted, one for each DGP described in the previous section. The parameter estimates from the pooled regression are used to build DGP 1 and part of DGP 2. The regressions within each year are used to build part of DGP 2 and all of DGP 3. The simulations use the regression's estimates of the coefficients, constant, and residual variance. Errors are randomly drawn from a normal distribution with a mean of 0 and a variance equal to the residual variance pulled from the data.

For DGP 1, the coefficients, constant, and residual variance are all fixed over time, so parameters are drawn from a regression on the pooled data. The parameters used to generate the dependent variable are listed in table 2.6.

Table 2.6: Simulation parameters: DGP 1.

Time point	Effect of Med. Income	Constant	Root MSE	N
All time points (pooled)	0.03	41.3	9.72	561

For DGP 2, the constant varies over time, so the year-specific constant is used along with the pooled coefficient and root MSE. For DGP 3, the year-specific values are used for the coefficient, constant, and root MSE. These values are listed in table 2.7. For DGP 4, the constant and residual variance vary in the same way as in DGP 3, but the coefficient

[8]These control variables are included to use the same model for the simulations as the one formulated for the example in section 2.6. The methods, however, will only be evaluated on how accurately they estimate the effect of median state income. The effects of the controls are set to be fixed over time, even when the effect of median state income is allowed to vary over time.

Table 2.7: Simulation Parameters: DGP 2 and 3.

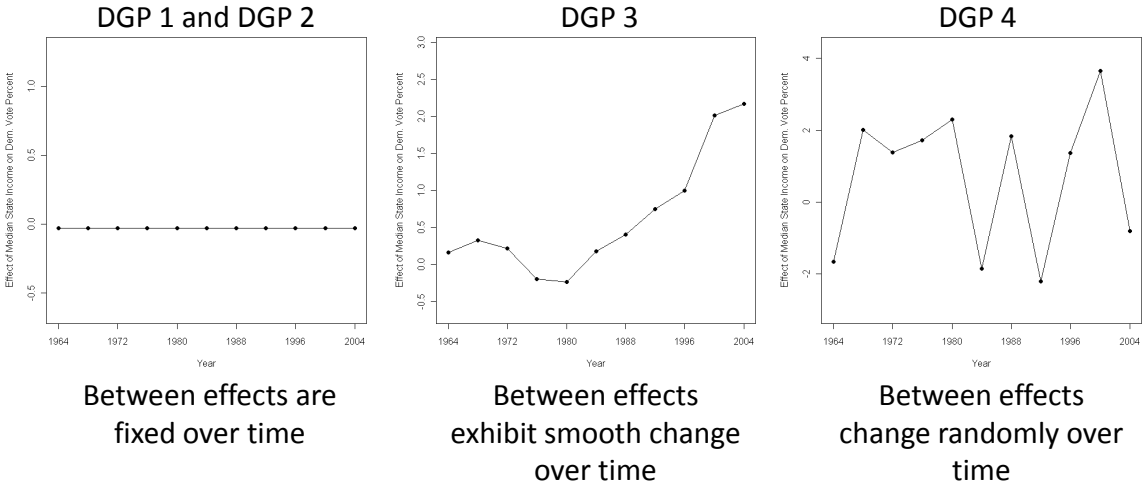
Time point	Effect of Med. Income	Constant	Root MSE	N
1964	0.12	64.1	9.31	51
1968	0.24	35.2	7.48	51
1972	0.30	-3.25	6.66	51
1976	0.08	-30.3	5.94	51
1980	-0.02	-18.2	8.13	51
1984	0.45	-62.8	5.77	51
1988	0.17	1.28	5.90	51
1992	0.38	6.68	5.45	51
1996	0.12	30.7	6.10	51
2000	0.76	-37.5	6.21	51
2004	1.19	-37.8	6.33	51

on median income is specified to vary randomly.[9] The time-invariant coefficients in DGP 1 and 2, the smoothly changing coefficients in DGP 3, and one example of random coefficients in DGP 4 are depicted in figure 2.7. In these graphs, the true between effect is represented on the y-axis, and the election year is plotted on the x-axis. Each point represents the true parameter value in a particular year used to generate the data.

After the dependent variable is generated, several TSCS methods are used to estimate the effects. The estimates are compared to the true estimates, and an evaluation statistic is computed for each method. To overcome sampling error, the simulation is iterated

[9]The coefficient in each year is drawn from a uniform distribution with a range three times larger than the range of coefficients listed in table 2.7. The minimum of the effects listed in table 2.7 is -0.18, and the maximum is 1.18. The range is 1.36. So the random coefficients are drawn from the Uniform$[-1.54, 2.54]$ distribution. New coefficients are drawn for each year in every iteration of the simulation.

Figure 2.7: The "True" Between Effects Generated by Each Data Generation Model.



10,000 times for each DGP, and the mean evaluation for each method is compared.

2.7.3 Competing Methods

The following TSCS methods are used to regress the simulated Democratic vote-share on median state income and the control variables, using each of the four DGPs to generate the Democratic vote-share:

1. Fixed effects (FE)
2. Pooled OLS with panel-corrected standard errors (PCSE)
3. Random Effects (RE)
4. Time-effects (TE)
5. Time-effects and an interaction with linear time (TE1)

   In addition to a dummy variable for each time point $(t_1, t_2, \ldots, t_T)$, a time-sequence variable τ is computed (e.g. 1964 is 1, 1968 is 2, etc.) and interacted with median

state income. Formally, the model can be written as

$$y_{it} = \alpha + \beta_1 x_{it} + \beta_2 x_{it}\tau + z_{it}\gamma + \sum_{k=1}^{T} t_k + \varepsilon_{i,t}, \tag{2.18}$$

where $y_{i,t}$ is the Democratic vote share in state i during year t, x_{it} is state i's median income in year t, and z_{it} represents the control variables. The marginal effect of x on y now depends linearly on time:

$$\frac{\partial y_{it}}{\partial x_{it}} = \beta_1 + \beta_2\tau. \tag{2.19}$$

This model is designed to capture a changing between effect when the rate of change is constant over time.

6. Time-effects and an interaction with quadratic time (TE2)

   In this model, median state income is interacted with both time and time-squared. The model can be written as:

$$y_{it} = \alpha + \beta_1 x_{it} + \beta_2 x_{it}\tau + \beta_3 x_{it}\tau^2 + z_{it}\gamma + \sum_{k=1}^{T} t_k + \varepsilon_{i,t}, \tag{2.20}$$

   which yields the following marginal effect:

$$\frac{\partial y_{it}}{\partial x_{it}} = \beta_1 + \beta_2\tau + \beta_3\tau^2. \tag{2.21}$$

   This model is designed to capture a between effect which changes over time in a curvilinear fashion.

7. Time-effects and an interaction with cubic time (TE3)

In this model, median state income is interacted with time, time-squared, and time-cubed. The model can be written as:

$$y_{it} = \alpha + \beta_1 x_{it} + \beta_2 x_{it}\tau + \beta_3 x_{it}\tau^2 + \beta_4 x_{it}\tau^3 + z_{it}\gamma + \sum_{k=1}^{T} t_k + \varepsilon_{i,t}, \tag{2.22}$$

which yields the following marginal effect:

$$\frac{\partial y_{it}}{\partial x_{it}} = \beta_1 + \beta_2\tau + \beta_3\tau^2 + \beta_4\tau^3. \tag{2.23}$$

This model is designed to capture a between effect which changes over time in a way that is more general than a linear or quadratic function.

8. Individual regressions within each time point (Reg)

9. BEER

For each iteration, the marginal effect of median state income is derived for each year and compared to the known population effect in that year. The results returned by each model are evaluated with a mean squared error measurement, as described in section 2.7.4.

2.7.4 Evaluation

Each TSCS method estimates the marginal effect of median state income on the Democratic vote share in each year. Some methods - FE, PCSE, RE, TE - produce the same estimate for each year. The others produce different estimates for each year. The point estimate and standard error produced by each model at each year are saved and are compared to the generating parameters listed in tables 2.6 and 2.7.

In order to judge the accuracy of each estimator against the true parameters, the mean squared error is derived for each estimator across 10000 simulation iterations. Ideally,

a method should produce estimates which are both consistent and efficient, and mean squared error is a statistic is which takes into account both the bias and the variance of the estimator.

In order to examine bias and inefficiency separately, two additional evaluative statistics are used. First, the simple absolute deviation between the point estimate and the true parameter is computed for each method using each DGP. This statistic does not consider efficiency, but also does not obscure the evaluation for biased estimates with measurements of variance. Second, the number of times each estimator produces a 95% confidence interval which contains the true parameter is counted, and reported as a percentage of the total number of estimates across every iteration of the simulation. This coverage percentage is in itself insufficient to evaluate the methods because an estimate with a large standard error will often include the true parameter in a confidence interval, even when the point estimate is far from the correct value. These additional statistics are reported in table 2.9 in section 2.7.6.

2.7.5 Expectations

As described in section 2.7.1, the four data generation models are designed to conform to the assumptions made by a class of TSCS methods. Therefore we should expect that the best performing methods will be the ones that make the assumptions that correspond to the DGP. DGP 1 assumes that the between effects and within effects are equal and that the between effect is fixed over time. Fixed effects estimates only the within effects, so it can only provide an accurate estimate of between effects when the between effects are equal to the within effects. In addition, random effects and pooled OLS with panel-corrected standard errors average the between and within effects, and can only provide an accurate estimate of either one when they are equal. We can expect FE, RE, and PCSE to perform very well in DGP 1.

FE, RE, and PCSE should not do as well in DGP 2 because the between effects are no longer equal to the within effects. FE is now estimating effects on the wrong variance, and RE and PCSE are averaging the between effects with estimates for wrong variance. Time effects, however, estimate between effects only, and assumes these effects are fixed over time. Since in DGP 2 the between effects are still assumed to be fixed over time, TE should perform very well.

In DGP 3, between and within effects are unequal, and the between effects exhibit smooth change over time. TE should not perform well because it does not accommodate changing between effects. However, time effects which include an interaction between the independent variable and time do model a changing between effect. There is a tradeoff between degrees of freedom and the flexibility of the model to capture the between effect at a particular point in time. TE1 has the most estimation power, but only captures between effects which change linearly over time. TE2 and TE3 are more flexible. In addition, BEER is designed to model this situation exactly without parameterizing how the between effects change. TE1, TE2, TE3, and BEER should all perform well in this situation, but BEER should generally be the most accurate.

Under DGP 4, however, BEER will fail because it incorrectly assumes that information from proximate time points are informative for the estimate at a particular time point. In this case, between effects are drawn randomly in each time point, and are not correlated with the between effects in other time points. Furthermore, the randomness of this DGP ensures that the between effects will only ever resemble linear, quadratic, or cubic growth due to chance, so TE1, TE2, and TE3 should also perform less well. The only method which avoids making false assumptions about the data in this case is the method of running individual regressions for each time point. Reg should be the most accurate method under DGP4.

2.7.6 Results

The mean squared errors and standard deviations of the evaluation statistics over 10,000 iterations are reported in table 2.8 for each method and for each DGP.[10] Lower values indicate greater levels of accuracy, and the most favorable evaluations are shaded. All shaded cells are significantly different from all non-shaded results in the column at the .01 level using a paired two-tailed T-test. These results should be viewed with the two

Table 2.8: Mean Squared Error Measures for Each Method Using Each of the Four Data Generating Processes.

	DGP 1	DGP 2	DGP 3	DGP 4
FE	0.0004(.001)	0.4270(.025)	0.1144(.008)	1.3874(1.07)
RE	0.0003(.000)	0.4413(.021)	0.1062(.006)	1.2580(.870)
PC	0.0003(.000)	0.4413(.021)	0.1062(.006)	1.2580(.870)
TE	0.0023(.003)	0.0022(.003)	0.0578(.002)	0.9262(.398)
TE1	0.0111(.015)	0.0109(.014)	0.0754(.020)	0.9638(.542)
TE2	0.0152(.017)	0.0149(.016)	0.0461(.015)	0.7537(.384)
TE3	0.0189(.018)	0.0185(.017)	0.0445(.019)	0.6184(.310)
Reg	0.0865(.045)	0.0870(.045)	0.0365(.019)	0.0367(.019)
BEER	0.0246(.022)	0.1043(.059)	0.0249(.011)	0.2891(.133)

Note: the mean across 10000 simulation iterations is reported, and the standard deviation of the simulation results is reported in parentheses next to the mean.

alternative evaluation statistics, listed in table 2.9.

[10]The simulations are run here with the control variables included in the regression. To test the dependence of the results on the presence of the controls, the simulations were separately rerun without these controls. The results from these simulations were very similar to the ones presented here, and provided identical inferences.

Table 2.9: Absolute Divergence and Coverage Percentage for Each Method Using Each of the Four Data Generating Processes.

	Absolute Deviation				Coverage			
	DGP 1	DGP 2	DGP 3	DGP 4	DGP 1	DGP 2	DGP 3	DGP 4
FE	0.015 (0.01)	0.653 (0.02)	0.338 (0.01)	1.116 (0.38)	95.8%	0.0%	7.0%	12.8%
RE	0.013 (0.01)	0.664 (0.02)	0.326 (0.01)	1.072 (0.33)	86.1%	100.0%	13.7%	49.0%
PCSE	0.013 (0.01)	0.664 (0.02)	0.326 (0.01)	1.072 (0.33)	86.1%	100.0%	13.7%	49.0%
TE	0.038 (0.03)	0.038 (0.03)	0.240 (0.00)	0.941 (0.20)	97.7%	95.8%	20.0%	6.8%
TE1	0.088 (0.06)	0.087 (0.06)	0.272 (0.03)	0.948 (0.26)	95.3%	95.6%	38.9%	14.2%
TE2	0.109 (0.06)	0.108 (0.06)	0.212 (0.03)	0.842 (0.21)	95.3%	95.5%	53.7%	19.4%
TE3	0.125 (0.06)	0.124 (0.06)	0.207 (0.04)	0.763 (0.19)	95.3%	95.5%	60.8%	23.6%
Reg	0.285 (0.07)	0.285 (0.07)	0.185 (0.05)	0.186 (0.05)	68.0%	94.9%	94.8%	67.7%
BEER	0.144 (0.06)	0.311 (0.09)	0.154 (0.03)	0.524 (0.12)	55.8%	98.1%	49.8%	19.1%

Note: the table on the left reports the mean absolute divergence across 10000 simulation iterations, and the standard deviation is reported in parentheses next to the mean. Shaded cells represent the lowest value in the column, and indicate the best quality model. All shaded cells are significantly different from all non-shaded results in the column at the .01 level using a paired two-tailed T-test. The table on the right contains the percentage of the iterations in which the correct parameter was included in 95% confidence interval.

In some ways, the results provide clear support for the expectations described in section 2.7.5, but in other ways the results are ambiguous and lead to more questions. In DGP 1, FE, RE, and PCSE are the most accurate estimators of the between effects. Although RE and PCSE perform significantly better than FE using either the mean squared error or the absolute deviation, the results may be artifacts of the lack of correlation between the residual and the model, and the results do not provide general evidence for selecting between FE, RE, and PCSE. All of these methods, however, appear to be appropriate options under the highly restrictive assumptions made by DGP 1. Under DGP 2, as expected, TE performs significantly better than all alternatives. Likewise, under DGP 3 BEER performs significantly better, and under DGP 4 the individual regressions are the most appropriate method. The inferences regarding the quality of TE, BEER, and Reg under DGPs 2, 3, and 4 respectively are made using both mean squared error

and absolute deviation.

The story becomes somewhat murkier when the coverage percentages are considered. As stated earlier, these measures do not necessarily distinguish well between a high-quality and a low-quality estimator. We might want to require that an estimator fails to reject a null hypothesis that the parameter is equal to the true value, but this requirement is always met when standard errors are very large. An estimate which deviates from the true value by a large amount but also covers the true value in its 95% confidence interval is both biased and inefficient. RE and PCSE in DGP 2 seem to be strong examples of bias and inefficiency.

The most important information to draw from the coverage table is the difference in coverage between BEER and Reg under DGP 3. Although BEER provides point estimates which are closer to the true parameter values than the estimates from Reg, Reg more often includes the true value within a confidence interval. It is troubling that BEER rejects a null of the true value 50.2% of the time. It appears that while the point estimates for BEER are the least biased of all the methods considered here, the standard errors are also biased downwards. In contrast, the standard errors for Reg are likely biased upwards.

2.7.7 Discussion

The results demonstrate conclusively that no one method for estimating the between effects in TSCS data is always the best one. The choice of which method to use should depend on a researcher's theoretical expectations about whether the between and within effects are equal, whether the between effects change over time, and if so, how the between effects change over time. In a simulation, we have the advantage of knowing the true parameter values. In reality we do not have this luxury, so there is no statistical test possible to help a researcher make this decision.

For most questions in political science, however, DGP 3 is much more likely than the other three cases. Few researchers should be comfortable assuming that the between and within effects are equal. The between effect of median state income compares the rich states to the poor states, and the within effect considers the effect of growing and declining wealth. Although both phenomena are important, these are fundamentally different processes and their estimates must require different methods. Likewise, unless the time frame being considered is very short, few researchers should be comfortable with the assumption that the between effects are fixed across time. The differences in the electoral behavior of rich and poor states in 1964 are clearly very different from the differences we observe in 2004. Finally, few researchers should be comfortable assuming that between effects are randomly drawn in each year. The characteristics of the American electorate change, but slowly; and therefore 1996 and 2000 are similar in ways that 1964 and 2000 are not. Only DGP 3 conforms to the assumptions that researchers with TSCS data who are interested in the between effects would be willing to make.

DGP 3 is best modeled by BEER, although the standard errors may be biased downwards. Individual regressions are inefficient because they fail to consider valid information contained in time points which are similar and proximate to the one under consideration. Future work should focus on the question of how much information is appropriate to use to inform estimates in a particular time point given different sample sizes, effect sizes, variable types, and time periods. BEER currently utilizes one mechanism for selecting information, but alternative specifications can and should be made.

2.8 Conclusion

Although the between effects estimation routine and each of the competing methods presented here produce different conclusions about the relationships in the data, it must be emphasized that none of these methods present a complete picture. BEER should

not be used in place of the other methods, but rather it should be used *with* the other methods. An effective statistical investigation of TSCS data should present theory about both the between and within effects, test hypotheses about the within effects with some type of fixed effects model, test the between effects with a method like BEER. One concern might be that running several analyses can give a researcher leeway to report only the most favorable results, and obscure the real relationships. However, if a researcher adopts a design which requires reporting the results from all of these methods, and understands the focus of each method, then the presentation of results will be more nuanced and complete than can be achieved using any single method alone. It is not the case with TSCS data that a relationship simply exists or does not exist; relationships look different over time than they do in a cross-section, and need to be examined from all perspectives.

Chapter 3

Estimation of a Non-linear Logistic Regression With Survey Weights Using Markov Chain Monte Carlo Simulation

3.1 Summary

I develop a method to estimate a non-linear logistic regression with survey weights which uses Markov-Chain Monte Carlo (MCMC) estimation. To demonstrate the utility of the method, I consider a voting model for U.S. presidential general elections which is non-linear, and addresses a long-standing theoretical debate in the study of American elections regarding the moderating role of personal importance on the effect of issue evaluations on the vote. In order to include survey weights, the "Poisson trick" is used which sets the sampling distribution as the mean of a Poisson distribution, and takes advantage of the fact that the distribution is equal to its mean at 0. The convergence of the estimation routine is assessed for each model, and marginal posterior distributions

for each parameter are estimated.

3.2 Introduction

I develop a method to estimate a non-linear logistic regression with survey weights.[1] The method uses Markov-Chain Monte Carlo (MCMC) estimation techniques which are commonly applied to Bayesian models. The likelihood function for the NL-logit model is the joint sampling distribution for the parameters, and uninformed prior distributions are specified. In order to include survey weights, I use the "Poisson trick" which sets the sampling distribution as the mean of a Poisson distribution, and takes advantage of the fact that the distribution is equal to its mean at 0. The convergence of the estimation routine is assessed for each model, and marginal posterior distributions for each parameter are estimated.

The methodology I develop in this paper is general, and can apply to any NL-logit model. To demonstrate the utility of the method, I consider a voting model for U.S. presidential general elections which is non-linear, and addresses a long-standing theoretical debate in the study of American elections. The example is discussed in greater detail in section 3.3: the data are presented in section 3.3.1, and the model is formally described in section 3.3.2. While there are many options for algorithms to estimate the parameters of a non-linear model, most of these options - described in section 3.4 - are designed for continuous dependent variables and fail to converge to results given the complexity of an NL-logit model. One algorithm which does produce estimates for this NL-logit model is idiosyncratic to this particular model, and requires bootstrapping to produce standard errors. The immediate advantage of MCMC is its generalizability: the point estimates and standard errors for any identified NL-logit model can be estimated. The

[1] In regards to probability a logistic regression is already non-linear, but the model presented here involves the extra complexity of a non-linear combination of the covariates.

methodology is discussed in detail in section 3.5. The model is estimated for the 1984, 1996, 2004, and 2008 presidential elections, and the results and convergence diagnostics for the MCMC chain are reported in section 3.6.

3.3 Example: The Role of Personal Importance in Issue Voting in U.S. Presidential Elections

The methodological work in this paper is motivated by the need to estimate a model of voter behavior posited by Gross, Kropko, Krosnick, Macdonald, and Rabinowitz (2010). This methodology, however, can be applied generally to the estimation of any NL-logit model, with or without survey weights.

Gross et al aim to resolve a long-standing debate about the moderating role that personal importance has on the effect of issue agreement on a individual's vote. Previous studies have drawn on formal models of voting behavior to derive a model for vote-choice. In these formulations, an individual's vote is a function of how the individual evaluates each candidate, and these evaluations are penalized if the individual and the candidate disagree on an issue. The penalties are discounted if an individual identifies an issue as being less important. For example, the probability that a voter in 2008 chooses to vote for Barack Obama may depend on the extent the voter agrees with Obama on health insurance reform, but the magnitude of the effect of agreeing should depend on how much the voter cares about the issue of health insurance reform. Some studies have found that issue salience has a significant moderating effect on vote choice, but many others have not found this to be true. A full review of the literature expressing these contradictory findings is presented by Gross et al. Most of these studies have used data from the American National Election Study, which measures the salience of a particular issue by asking respondents "How important is this issue to you personally? Extremely

important, very important, somewhat important, not too important, or not important at all?" (see for example the 2008-2009 ANES Panel Study Questionnaires, January 2009 Advance Release, p. 37 question P3).

Usually, moderation is modeled linearly through the addition of an extra term in the equation which is the product of the two variables being interacted. Although this treatment of moderation is common in regression analysis, the formal voting model which underlies the empirical work on voting implies that importance should be specified as a multiplicative discount on the effect of distance. Niemi and Bartels (1985) and Grynaviski and Corrigan (2006) have previously developed regression models in which the ordinal values of the importance variables are scaled to be between 0 and 1, and are multiplied by the issue distances. Neither analysis, however, determined that importance had a discernible effect on candidate evaluation.

The goal of the Gross et al is to estimate the effect of personally expressed importance as a moderator of issue impact using a less constrained methodology. They develop a model, described in section 3.3.2, which corresponds to those used by Niemi and Bartels and by Grynaviski and Corrigan, but which allows for different underlying voting models, and which estimate the impact of the various importance responses rather than presuming a linear effect. Further, as described in section 3.3.2, they focus on the relative rather than absolute importance values. Of concern here is the case in which the dependent variable is the vote. As a further complication in several of the election studies they analyze there are survey weights. The data, as well as the treatments of importance, and issue voting models, are described in section 3.3.1.

3.3.1 Data

The data are acquired from the American National Election Studies in 1984, 1996, 2004, and 2008. The dependent variable in each year is a vote, among the two-party vote, for

the Democratic candidate: Mondale in 1984, Clinton in 1996, Kerry in 2004, and Obama in 2008. Votes for third party candidates and independents are dropped, leaving sample sizes of 928, 772, 527, and 450 in 1984, 1996, 2004, and 2008 respectively. The proposed model contains three kinds of independent variables. In the discussion below, let C refer to the control variables, I refer to the issue evaluations, and S refer to the importance indicators. In each year, the following demographic controls are included: indicators for race, being from the south, gender, marital status, union membership, self-employement, religion and religiosity, age, income, and education.

Different issues are included in different years. In 1984, views on intervening in Central America and views on the extent to which government should be providing services, jobs, and aid to women are included. In 1996, views on government spending on services, defense, and aid to blacks are included, as well as views on environmental protection and abortion. In 2004, all of the issues from 1996 are included, as well as views on government job provision, gun control, the role of women in the workplace, and diplomacy. Finally, in 2008, views on government spending on services, defense, provision of jobs, and aid to blacks are again included, as well as views on environmental protection, the role of women, abortion, and medical insurance.

For each issue measured in each ANES, the views of the respondent as well as the respondent's perceptions of the candidate's positions are recorded. The average perceived position across all respondents in the survey is used to represent each candidate's position. All 14 model specifications considered by Grynaviski and Corrigan utilize the proximity model to measure issue distance. In the analysis presented by Gross et al, however, the directional model is used instead of the proximity model. For each issue, denote the position of the respondent (voter) as X_v, the average perceived position of the Democratic candidate as X_d, and the average perceived position of the Republican candidate as X_r. Also, let X_o represent the position on the issue scale which represents the center of the

scale (e.g. 4 on a 7 point left-right scale). The agreement between the voter and a candidate $c \in \{d, r\}$ is

$$A_{v,c} = (X_v - X_o)(X_c - X_o). \tag{3.1}$$

Note that larger values represent more agreement between a voter and a candidate, rather than greater distance, and that the calculation of $A_{v,d}$ and $A_{v,r}$ depends on which side of the ideological scale the voter and candidate each fall on. Again, to compare the candidates from the perspective of the voter, the issue index calculates the difference of agreement scores:

$$I_v = A_{v,d} - A_{v,r}, \tag{3.2}$$

where positive values of this difference mean that the voter agrees more with the Democrat than the Republican.

For each issue in the analysis, the salience of the issue to the respondent is recorded. Let $S_{v,j,1}, S_{v,j,2}$, and $S_{v,j,3}$ be dummy variables which indicate that respondent v answers that issue j is "extremely important", "very important", or "somewhat important" respectively. There are few respondents who answer that an issue is "not important at all", so let $S_{v,j,4}$ indicate that respondent v answers that issue j is either "not too important" or "not important at all". Since these four indicators together span every case, "extremely important" is excluded from the model as the reference category.

3.3.2 The Model

The dependent variable y_i is 1 if the respondent i votes for the Democrat, and is 0 if the respondent votes for the Republican. The probability that a respondent votes for the Democrat follows the Bernouli distribution:

$$f(y_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}, \tag{3.3}$$

where

$$\pi_i = \Lambda({y_i}^*) = \frac{1}{1+\exp(-{y_i}^*)}, \tag{3.4}$$

and ${y_i}^*$ is a function of the issue evaluations I with effects denoted by β, the issue saliences S with effects denoted by γ, the demographic controls C with effects denoted by δ, and a constant denoted α. What distinguishes this model from a standard logistic regression is that ${y_i}^*$ is a non-linear function of these variables and parameters. This estimation method can handle any identified specification of ${y_i}^*$. For this application, the functional form of ${y_i}^*$ with J issues and K controls is

$${y_i}^* = \alpha + \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} + \sum_{k=1}^{K} \delta_k C_{i,k}. \tag{3.5}$$

The middle term of equation 3.5 contains the standard linear effect, β_j, for each issue evaluation but this effect is moderated by salience. For example, if the respondent answers that issue 2 is "very important", then $S_{i,2,2}$ is 1, and $S_{i,2,3}$ and $S_{i,2,4}$ are 0, so the term in the brackets reduces to γ_2. These moderating effects γ are fixed across issues, and are multiplicative rather than additive effects. The idea is to model a percent change in the effect of each issue for each salience level relative to the highest level. If γ_2 is estimated to be .8, then the model indicates that the effect of an issue on vote-choice is 20% closer to 0 for an individual who replies that an issue is very important than for an individual who replies that an issue is extremely important.[2]

The model also accounts for a respondent's pattern in selecting which issues are salient. The fraction

$$\frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} \tag{3.6}$$

[2]The term "effect" is used loosely since point estimates do not have a useful direct interpretation in a logistic regression. However, by modeling y^* to satisfy these theoretical requirements, odds ratios and marginal probabilities will also reflect the correct theoretical model specification.

refers to the respondent's pattern in answering all of the salience questions considered in the model. Suppose that there are 5 issues in the model, and two respondents reply that the first issue is extremely important. The first respondent also replies that the other 4 issues are extremely important, and the second respondent replies that the other 4 issues are not important. The response of "extremely important" should be considered to be more meaningful for the second respondent. For the first individual, the effect of the first issue is β_1, and for the second individual, the effect of the first issue is

$$\beta_1 \ \frac{5}{1+4\gamma_3}. \tag{3.7}$$

Theory suggests that $0 < \gamma_3 < 1$, and if this is the case, then the effect of the first issue is more pronounced for the second individual than for the first individual since the effect in equation 3.7 is greater than β_1.

Taking into account the population weight for each respondent, w_i, the log-likelihood function for this model with N respondents is

$$\ell(\beta, \gamma, \delta | y, I, S, C) = \sum_{i=1}^{N} w_i \Big[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \Big]. \tag{3.8}$$

3.4 Methodological Background

There are many options available to researchers who wish to estimate a non-linear model. The most commonly used technique is to specify an NL-model to be "intrinsically linear" (Greene 2000, p. 327). That is, non-linear marginal effects are modeled by including additional linear terms which are transformations of the variables in the model. This class of models includes logged variables, curvilinear, and additive interaction terms. In the model specified here, additive interaction terms do not capture the shape of the moderating effect implied by the theory, and we resort to multiplicative interaction terms

which cannot be modeled linearly.

When a model is sufficiently complex so that it cannot be modeled as intrinsically linear, some researchers use the non-linear least squares estimator (NLS). NLS depends on the same assumptions as the OLS estimator, and can be shown to be consistent and asymptotically normal under these conditions (Davidson and MacKinnon 1993). Computation of point estimates for NLS proceeds in the same way as OLS, but the variance-covariance matrix is computed through a sandwich estimator involving the Hessian of the non-linear model specification (Cameron and Trivedi 2005, p. 151-153). These properties fail, however, when the dependent variable is binary. If OLS is applied to a model with a binary dependent variable, the linear predictions model the probability that that dependent variable equals 1. However, the linear probability model can yield predicted probabilities which are less than 0 or greater than 1 (Greene 2000, p. 813). For the same reasons, NLS fails to be an appropriate estimator of a non-linear model when the dependent variable is binary.

Logistic regressions are estimated through maximum likelihood (ML), and versions of logistic regressions which are non-linear in the variables can theoretically be estimated through ML as well. Logistic regressions are a type of generalized linear model in which the dependent variable is assumed to follow a Bernouli distribution and the covariates form the argument of a logistic link function (Cameron and Trivedi 2005, p. 149). Estimation of point estimates in a generalized linear model requires evaluating the values of the parameters when the first partial derivatives of the likelihood function are 0. The likelihood function is itself a function of $g(x, \theta)$, which represents the specification of the covariates. By the chain rule, each first partial derivative of the likelihood function must be multiplied by the derivative of g. If g is linear or intrinsically linear then the derivatives of g contain single parameters, and the first-order conditions of the likelihood function are straightforward for a computer algorithm to estimate. But if g is non-linear, then an ML

algorithm must either by given or must estimate the gradient of g. The "nlm()" package in R and the "ml" function in Stata allow the user to specify the gradient and Hessian of the log-likelihood function, otherwise numeric first and second derivatives are computed (Venables et al 2010, p. 59-60; Gould et al 2006, p. 21). The use of numeric derivatives, especially the second derivatives, may result in a large increase in computation time, and worse, inaccurate calculations (Cameron and Trivedi 2005, p. 341).

Calculating the analytic first partial derivatives of a log-likelihood function with a non-linear specification of the covariates is non-trival. But if the derivatives can be specified, they form a non-linear system of equations which can be estimated using generalized estimating equations (GEE), which are a special case of the generalized method of moments (GMM) (McGulluch and Nelder 1989). The algorithm will attempt to produce a maximum likelihood solution for the system, but for an NL model, the system of first-order conditions has a more complex non-linear form which often results in flat regions or local optima in the log-likelihood function and may cause hill-climbing algorithms to fail to converge (Cameron and Trivedi 2005, p. 169).

For the particular NL-logit model described in section 3.3.2, George Rabinowitz has developed an estimator that uses numerical first derivatives and a general hill climbing strategy. The method appears to work effectively, but it is idiosyncratic to this particular model, and requires bootstrapping to estimate standard errors. It is not a method that can be applied generally to the class of NL-logit models.

A general technique which does provide a measure of variability for the point estimates is Markov Chain Monte Carlo (MCMC) estimation, operationalized through a Gibbs sampler. As shown by Groenewald and Mokgatlhe (2005), the marginal probabilities required to implement a Gibbs sampler are general to any specification of the

covariates in a logit model.[3] The Gibbs sampler does not suffer from the same convergence problems that are prevalent with ML, GEE, and GMM, and it can be immediately applied to any NL-logit model without calculating analytic derivatives or relying on numeric approximations. The formulation of the MCMC methodology is discussed in section 3.5.

3.5 Methodology

According to Bayes rule, the posterior distribution implied by the model is

$$P(\theta|x,y) \propto P(x,y|\theta) \times P(\theta). \tag{3.9}$$

For the specific example considered in section 3.3, $P(x,y|\theta)$ is equal to the loglikelihood function in equation 3.8 and $P(\theta) = P(\beta,\gamma,\delta)$ is the joint prior distribution for the model parameters. In order to derive results which are theoretically equivalent to maximum likelihood estimates, we want to input as little information into this prior distribution as possible, so that all information comes from the likelihood function. We specify priors for all parameters β, γ, δ which are normal, independent, and very flat. The estimation routine is implemented in WinBUGS version 1.4.3, which uses a Gibbs Sampling technique. WinBUGS has built in routines for several common probability distributions, including the Bernouli distribution that defines the likelihood function considered here. In appendix D, I show that any NL-logit specification can be estimated by MCMC, and I derive the specific marginal posterior probabilities for every parameter in the NL-logit model considered in this paper. A difficulty arises when the population weights are included. In order to customize a probability distribution in WinBUGS, I resort to a trick which takes advantage of a characteristic of the Poisson distribution (Spiegelhalter et al

[3]Albert and Chib (1993) have a similar proof for probit models.

2007). The value of the Poisson distribution at 0 is an exponential function of its mean:

$$f(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!} \implies f(0|\lambda) = e^{-\lambda}. \tag{3.10}$$

If we set

$$\lambda_i = -\log(\ell_i) = -\log\Big[w_i\big[y_i\log(\pi_i) + (1-y_i)\log(1-\pi_i)\big]\Big], \tag{3.11}$$

then

$$f(0|\lambda_i) = w_i\Big[y_i\log(\pi_i) + (1-y_i)\log(1-\pi_i)\Big]. \tag{3.12}$$

The method samples on the customized likelihood function in equation 3.8 by exponentiating the log-likelihood contribution for each element and multiplying them by -1 as in equation 3.11. The method then creates a vector of zeroes with as many elements as cases in the data, and specifies that this vector is distributed by a Poisson distribution with means as defined in equation 3.11. The routine then samples on the sum of the values listed in equation 3.12, which is equivalent to sampling on the likelihood function in equation 3.8.

I run Markov Chain Monte Carlo (MCMC) simulation, implemented as a Gibbs sampler, on the model described here for the ANES data from 1984, 1996, 2004, and 2008. The Gibbs Sampler is run in WinBUGS version 1.4.3, and an example of the code used to run the estimation procedure is listed in appendix E. The chains are not guaranteed to converge to the true posterior distribution in any finite amount of time, so several diagnostics are run post-estimation to assess whether the chains have converged. An excellent general review of available diagnostics was written by Cowles and Carlin (1996). For this application, I use the diagnostics created by Raftery and Lewis (1992), Geweke (1992), and Heidelberger and Welch (1983). All of these diagnostics are implemented in R through the CODA package (Plummer et al 2010).

3.6 Results and Discussion

A Gibbs sampler is run for the model for each year. The chain draws 100,000 values for each parameter in equation 3.5 and computes the log-likelihood value for each iteration's set of drawn parameter values. The 100,000 draws of each parameter are saved and analyzed after they are all drawn. Point estimates and summary statistics for each parameter are listed in table 3.1 for 1984 and 1996, and in table 3.2 for 2004 and 2008. The Gibbs sampler requires many iterations to converge and begin sampling from the actual posterior distribution of the parameters. In addition, consecutive draws may be correlated, which could bias the estimate of the standard deviation. To account for convergence, the first 1000 draws for each parameter are not considered for the calculation of the statistics in tables 3.1 and 3.2, and to account for autocorrelation across the iterations, the chain is thinned so that only every fourth draw is considered. The results are computed using the remaining 24,750 observations.

The first column of results for each year in tables 3.1 and 3.2 contains the values of the parameters for the draw that contained the highest attained log-likelihood. The value of the highest attained log-likelihood is also reported at the bottom of the table. The second column contains the mean and standard deviation across the 24,750 iterations which are used after burning the initial observations and thinning the chain. The last column contains the first and 99th percentiles of the utilized draws for each parameter.

When non-informed prior distributions are used, as is the case the example analyzed here, the joint posterior distribution of the parameters is the same as the log-likelihood function. Note, however, that MCMC is not a hill-climbing algorithm, and there is no reason to expect the maximum attained likelihood in a Gibbs chain to be equal to the solution of an ML algorithm. One advantage of using MCMC over ML is that MCMC does not break down when it encounters flat regions or local maxima. A disadvantage is that MCMC will only return the maximum likelihood by randomly drawing the set of

Table 3.1: MCMC Results for the NL-Logit Model: 1984 and 1996

Results for the 1984 Election Data				**Results for the 1996 Election Data**			
	At Max. LL	Mean (SD)	[1% , 99%]		At Max. LL	Mean (SD)	[1% , 99%]
Constant	-1.02	-0.93 (.37)	-1.80 , -0.06	Constant	0.51	0.60 (.43)	-0.40 , 1.60
Gov. Service	0.20	0.20 (.03)	0.12 , 0.28	Gov. Service	0.37	0.35 (.05)	0.24 , 0.47
Central America	0.25	0.24 (.04)	0.15 , 0.33	Defense Spending	0.58	0.52 (.13)	0.23 , 0.83
Aid Women	0.18	0.19 (.05)	0.08 , 0.29	Aid Black	0.21	0.23 (.05)	0.12 , 0.34
Gov. Jobs	0.19	0.18 (.04)	0.09 , 0.26	Abortion	0.10	0.12 (.02)	0.07 , 0.16
				Environment/Jobs	0.21	0.23 (.08)	0.06 , 0.41
Very Important	0.94	1.25 (.38)	0.63 , 2.43	Very Important	0.49	0.54 (.13)	0.30 , 0.90
Somewhat Important	0.85	1.08 (.37)	0.47 , 2.22	Somewhat Important	0.67	0.68 (.17)	0.35 , 1.17
Not Important	0.86	0.60 (.52)	-0.54 , 2.10	Not Important	0.15	0.29 (.22)	-0.16 , 0.92
Black	1.80	1.93 (.43)	0.99 , 2.98	Black	4.02	4.81 (1.32)	2.23 , 8.46
South	-0.38	-0.33 (.26)	-0.94 , 0.26	South	0.04	0.04 (.25)	-0.54 , 0.62
Female	-0.04	0.03 (.19)	-0.41 , 0.46	Female	0.32	0.40 (.22)	-0.11 , 0.91
Married	0.07	0.06 (.20)	-0.40 , 0.54	Married	0.02	0.05 (.25)	-0.52 , 0.66
Union	1.14	1.12 (.21)	0.64 , 1.62	Union	0.96	0.84 (.28)	0.19 , 1.50
Self Employed	-0.04	-0.14 (.28)	-0.80 , 0.48	Self Employed	-0.08	-0.02 (.30)	-0.73 , 0.67
Protestant	-0.23	-0.24 (.28)	-0.91 , 0.44	Protestant	-0.42	-0.39 (.33)	-1.19 , 0.38
Catholic	0.11	0.06 (.30)	-0.66 , 0.75	Catholic	0.12	0.17 (.36)	-0.67 , 0.98
Jewish	0.96	1.17 (.59)	-0.17 , 2.61	Jewish	1.20	2.35 (1.45)	-0.42 , 6.53
Age: young	-0.32	-0.16 (.23)	-0.69 , 0.38	Age: young	0.20	0.16 (.36)	-0.67 , 1.01
Age: old	0.24	0.08 (.25)	-0.50 , 0.67	Age: old	0.20	-0.06 (.27)	-0.69 , 0.57
Income: low	0.48	0.48 (.28)	-0.17 , 1.14	Income: low	0.64	0.46 (.37)	-0.38 , 1.34
Income: high	-0.53	-0.64 (.20)	-1.11 , -0.17	Income: high	-0.49	-0.53 (.26)	-1.14 , 0.07
Education: low	0.52	0.64 (.32)	-0.08 , 1.39	Education: low	0.75	0.75 (.41)	-0.19 , 1.72
Education: hsplus	0.00	-0.07 (.23)	-0.59 , 0.47	Education: hsplus	-0.41	-0.69 (.28)	-1.35 , -0.04
Education: college	0.71	0.59 (.24)	0.03 , 1.16	Education: college	-0.41	-0.66 (.30)	-1.36 , 0.02
Church frequency	0.08	0.04 (.21)	-0.45 , 0.52	Church frequency	-0.44	-0.26 (.27)	-0.90 , 0.35
N	928			N	772		
Max. Likelihood	-425.20			Max. Likelihood	-296.40		
Iterations	100000			Iterations	100000		
Burn-in	1000			Burn-in	1000		
Thining	4			Thining	4		

Table 3.2: MCMC Results for the NL-Logit Model: 2004 and 2008

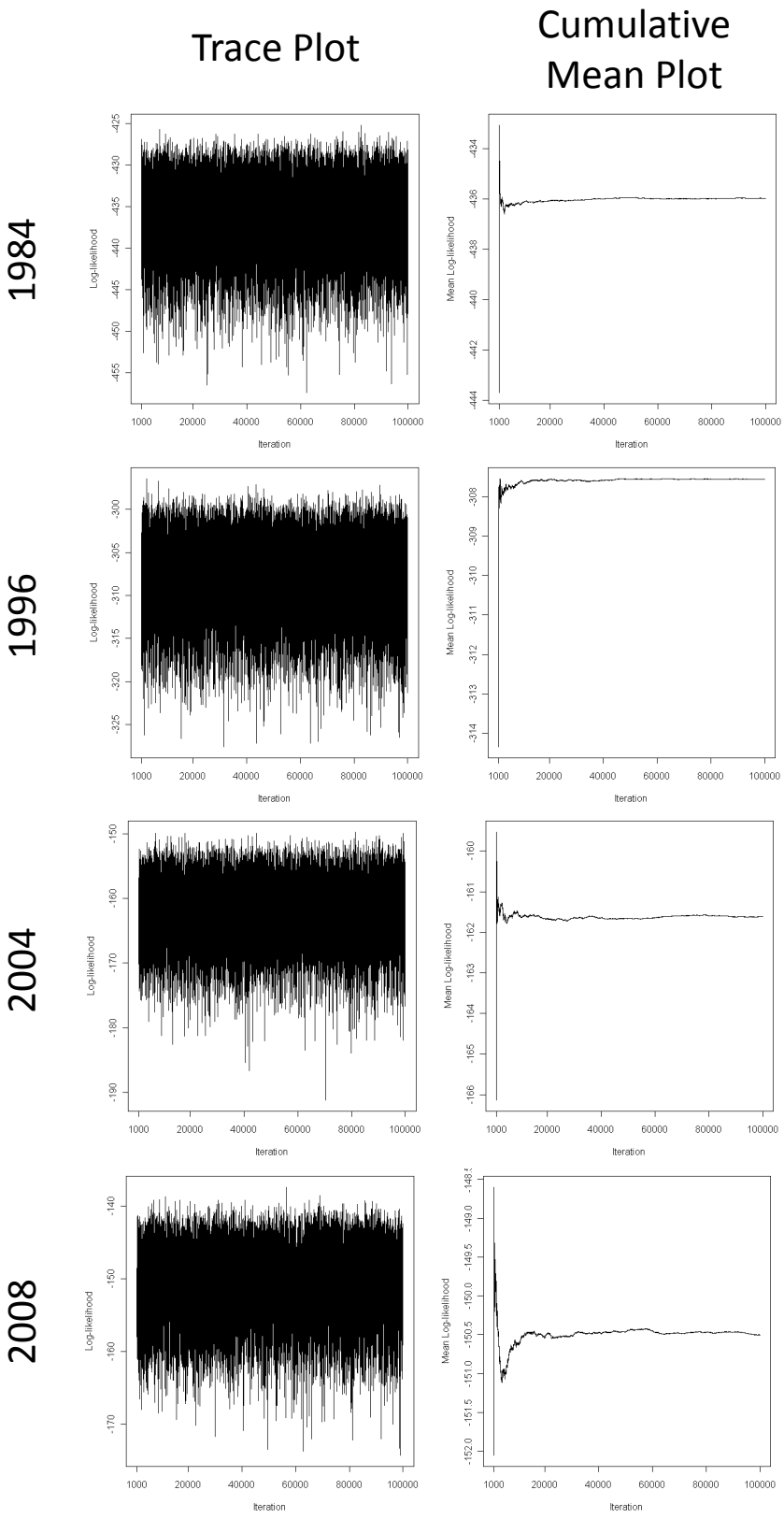
Results for the 2004 Election Data				**Results for the 2008 Election Data**			
	At Max. LL	Mean (SD)	[1% , 99%]		At Max. LL	Mean (SD)	[1% , 99%]
Constant	-1.65	-1.78 (.68)	-3.38 , -0.22	Constant	0.90	0.86 (.63)	-0.57 , 2.37
Diplomacy/Military	0.16	0.16 (.03)	0.09 , 0.24	Gov. Service	0.37	0.42 (.07)	0.26 , 0.59
Gov. Service	0.42	0.40 (.08)	0.22 , 0.60	Defense Spending	0.30	0.28 (.06)	0.14 , 0.44
Defense Spending	0.26	0.25 (.06)	0.11 , 0.41	Medical Insurance	0.01	0.02 (.03)	-0.05 , 0.09
Gov. Jobs	0.05	0.03 (.05)	-0.08 , 0.15	Gov. Jobs	0.06	0.11 (.05)	0.00 , 0.22
Aid Black	0.23	0.27 (.08)	0.09 , 0.47	Aid Black	0.22	0.24 (.07)	0.08 , 0.41
Environment/Jobs	2.40	2.09 (1.09)	-0.41 , 4.67	Environment/Jobs	1.00	0.97 (.22)	0.48 , 1.51
Gun Control	-2.24	-2.34 (.69)	-3.98 , -0.75	Women	0.11	0.14 (.10)	-0.09 , 0.39
Women	-0.01	0.12 (.11)	-0.14 , 0.39	Abortion	0.16	0.18 (.04)	0.09 , 0.29
Abortion	0.17	0.16 (.03)	0.09 , 0.24				
Very Important	1.07	1.00 (.17)	0.65 , 1.48	Very Important	1.11	1.30 (.41)	0.69 , 2.62
Somewhat Important	0.66	0.58 (.18)	0.25 , 1.08	Somewhat Important	0.15	0.23 (.12)	0.03 , 0.58
Not Important	0.01	0.06 (.17)	-0.33 , 0.54	Not Important	-0.04	0.10 (.18)	-0.21 , 0.64
Black	2.81	2.85 (.70)	1.33 , 4.62	Black	7.65	10.45 (5.18)	3.73 , 28.28
South	0.84	0.82 (.43)	-0.18 , 1.83	South	-0.89	-0.85 (.37)	-1.74 , 0.01
Female	-0.26	-0.49 (.35)	-1.32 , 0.31	Female	-0.36	-0.43 (.39)	-1.36 , 0.46
Married	-0.48	-0.56 (.40)	-1.48 , 0.36	Married	-0.04	0.15 (.40)	-0.77 , 1.08
Union	1.12	1.36 (.43)	0.39 , 2.38	Union	0.61	0.79 (.53)	-0.44 , 2.02
Self Employed	-0.91	-0.80 (.46)	-1.89 , 0.23	Self Employed	-1.67	-1.78 (.66)	-3.39 , -0.28
Protestant	0.24	0.61 (.50)	-0.55 , 1.80	Protestant	-0.27	-0.13 (.42)	-1.11 , 0.86
Catholic	1.02	1.10 (.52)	-0.09 , 2.34	Catholic	0.41	0.79 (.48)	-0.32 , 1.91
Jewish	0.75	1.49 (.93)	-0.58 , 3.77	Jewish	1.58	2.00 (1.34)	-1.00 , 5.32
Age: young	0.57	0.59 (.51)	-0.59 , 1.80	Age: young	0.85	0.77 (.68)	-0.77 , 2.44
Age: old	1.47	1.02 (.44)	-0.01 , 2.03	Age: old	0.26	0.24 (.42)	-0.73 , 1.25
Income: low	0.05	0.06 (.55)	-1.23 , 1.33	Income: low	-0.26	-0.62 (.53)	-1.88 , 0.58
Income: high	0.53	0.29 (.41)	-0.66 , 1.23	Income: high	-0.91	-0.98 (.46)	-2.06 , 0.08
Education: low	-0.32	-0.30 (.82)	-2.20 , 1.61	Education: low	0.61	0.20 (.95)	-2.03 , 2.41
Education: hsplus	-0.46	-0.42 (.44)	-1.45 , 0.60	Education: hsplus	-0.92	-1.26 (.50)	-2.46 , -0.13
Education: college	-0.72	-0.88 (.46)	-1.96 , 0.14	Education: college	-1.24	-1.29 (.51)	-2.48 , -0.11
Church frequency	0.51	0.50 (.44)	-0.52 , 1.56	Church frequency	0.13	0.21 (.50)	-0.94 , 1.40
N	527			N	450		
Max. Likelihood	-149.60			Max. Likelihood	-137.40		
Iterations	100000			Iterations	100000		
Burn-in	1000			Burn-in	1000		
Thining	4			Thining	4		

parameters at the modal point and calculating the log-likelihood at that point. Just as the probability of drawing a particular value of a continuous random variable is 0 regardless of the number of draws taken, the probability that MCMC returns the maximum likelihood is 0.

The results in tables 3.1 and 3.2 are only trustworthy if the chain has converged to the actual log-likelihood function. Although there is no way to know with complete certainty whether the chain has converged, there are many useful diagnostics available. First, at a minimum, a plot which traces the drawn values of the log-likelihood function over the iterations must appear to be non-trending. These graphs are called trace plots, and if a trace plot does not resemble a time series of white noise then the chain has not converged. In addition, it is possible to calculate the cumulative mean of the log-likelihood at each iteration. If a chain has converged, then slope of the cumulative mean over the iterations should be flat. The trace plots and cumulative mean plots of the log-likelihood values for each year are listed in figure 3.1.

The trace plots and cumulative mean plots do not indicate that any of the chains have failed to converge, but that does not necessarily mean that the chains have converged. In table 3.3, several tests are run on the chains using the diagnostics mentioned in section 3.5. The Heidleberger and Welch diagnostic contains two tests, called the stationarity test and the halfwidth test. The stationarity test considers whether there is a significant trend in the chain. If a trend is observed, the test drops the first 10% of iterations and reconsiders the trend; the test repeats until no trend is observed or 50% of the observations have been ignored. In this second case, the test determines that the chain has not converged. If the test passes, but only after some iterations are ignored, then the test recommends a particular number of initial iterations to burn-in. For all of the years, the stationarity test passes at iteration 1. The halfwidth test compares the mean of the first half of the sample to the mean of the entire sample, and passes if these means are

Figure 3.1: Trace Plots and Cumulative Mean Plots for the Log-Likelihood Functions



sufficiently similar (Cowles and Carlin 1996). The log-likelihood chains for every year also all pass halfwidth test.

Table 3.3: Convergence Diagnostics for the NL-logit Model of the 1984, 1996, 2004, 2008 ANES.

	1984	1996	2004	2008
Heidelberger and Welch				
Stationarity Test	Passed	Passed	Passed	Passed
Halfwidth Test	Passed	Passed	Passed	Passed
Geweke				
n_A	0.1	0.1	0.1	0.1
n_B	0.5	0.5	0.5	0.5
Z	-0.5	0	-0.93	0.96
Raftery and Lewis				
Burn-in	12	10	18	12
N_{prec}	14574	13562	21933	14700

Note: diagnostics are run to test the convergence only of the log-likelihood function. The Heidelberger and Welch test is run on the entire chain with no initial iterations thrown out. The stationarity tests are run from iteration 1. The Raftery and Lewis diagnostic is run on the chain after burning-in the first 1000 iterations, and the Geweke diagnostic is run on the chain after burning-in the first 10000 iterations.

The Geweke diagnostic is similar to Heidelberger and Welch's halfwidth test. The mean is computed for the first 10% and the final 50% of draws for each parameter, and a comparison of means test is conducted to determine if these means are different (Cowles and Carlin 1996). This procedure reports a Z statistic from this test. To avoid biasing the test, the first 10000 iterations are removed. None of the Z statistics for the chains analyzed here approach a level at which we would suspect the chains have not converged.

Finally, the Raftery and Lewis test is similar to a power analysis in that it considers the number of iterations that would be necessary in order to report a particular statistic within a particular degree of accuracy. The test reports a statistic, N_{prec}, which is the estimated number of iterations necessary to evaluate the 2.5th percentile of the marginal posterior distribution of a parameter with 95% certainty that the estimate is within .005

of the correct value (Cowles and Carlin 1996).[4] In addition, the routine estimates the number of iterations necessary to approach within a specified tolerance of the stationary distribution, and this value reports the number of iterations that should additionally be burnt-in. This diagnostic is run excluding the first 1000 iterations. The recommended iterations are all well below the 100,000 utlilzed in these estimates, and the burn-in values demonstrate that no additional burn-in is necessary.

To summarize the converge diagnostics, the results appear to have converged after 1000 iterations, and specific diagnostic tests support that claim. It is appropriate to interpret the results in tables 3.1 and 3.2.

For the most part, the issue evaluations work in the way we expect: the more that a voter agrees with the Democrat relative to the Republican, the more likely the voter is to vote for the Democrat. When a voter responded that an issue is extremely important, the issues in every year had a significant effect in the correct direction with a few exceptions: in 2004 women's role had no effect, and gun control had a significant effect in the opposite direction, and in 2008 women's role and medical insurance had no effect.

The real theoretical motivation of the model is the evaluation of the moderation effects of the salience levels. If personal importance really changes the effect of the issues in the way expected by the theory, then we should see the greatest issue effects when an individual says an issue is extremely important, a smaller effect when an individual says an issue is very important, an even smaller effect when an individual says that an issue is somewhat important, and the smallest effect when the individual says that an issue is not important. In terms of the parameters in equation 3.5, the theory would suggest that $1 \geq \gamma_2 \geq \gamma_3 \geq \gamma_4$. In the first 6 rows of table 3.4, the percent of iterations which violate this statement are reported, broken down along each type of violation. The analogy to traditional hypothesis testing would suggest that we reject the possibility of

[4]These specifications are defaults, and can be easily changed in R.

Table 3.4: Comparisons of the Moderation Effects of the Levels of Personal Importance

	1984	1996	2004	2008
% of iterations in which				
$\gamma_2 \geq 1$	73.8	0.2	46.1	76.8
$\gamma_3 \geq 1$	52.1	4.6	2.2	0.0
$\gamma_4 \geq 1$	18.5	0.5	0.0	0.0
$\gamma_3 \geq \gamma_2$	26.1	83.5	1.7	0.0
$\gamma_4 \geq \gamma_3$	10.4	12.7	0.1	21.8
$\gamma_4 \geq \gamma_2$	16.2	6.0	0.8	0.0
$1 \geq \gamma_2 \geq \gamma_3 \geq \gamma_4$	12.6	14.3	52.3	18.7

an alternative when it occurs less than 5% of the time. By this standard, we cannot reject any violation in 1984. In 1996 we reject the possibility of the first three violations. In 2004, we reject all violations except $\gamma_2 \geq 1$, and in 2008 we reject the possibility that $\gamma_3 \geq 1$, $\gamma_4 \geq 1$, $\gamma_3 \geq \gamma_2$, and $\gamma_4 \geq \gamma_2$. Considered together, the theory holds in only 12.6%, 14.3% and 18.7% of iterations in 1984, 1996, and 2008 respectively. The theory works best in 2004, where 52.3% of iterations conform to the theory. This joint test is conservative, however, and it is important to observe that the correct pairwise comparisons are generally much more prevalent than incorrect ones. In addition, at least 3 of the 6 possible violations of the theory are rejected in 1996, 2004, and 2008, and any significant result demonstrates a downward trend in the magnitude of the moderation effects as personal importance decreases.

3.7 Conclusion

This paper demonstrates how MCMC can be used to estimate a complicated NL-logit model. With this estimation technique, a researcher does not need to specify analytic derivatives or program an idiosyncratic hill-climbing algorithm. The most difficult challenge is writing the model in a way that can be read by the software. An example of how to write the model described in this paper for the WinBUGS software is provided in appendix E. Gibbs samplers also exist for other generalized linear models, and will be able to handle non-linear specifications. Researchers should be aware of these options and should not compromise complex but theoretically appropriate models.

Appendix A

Regression Results for the Britain Simulation Models

Coefficients for the Britain models are derived from a regression of party affect on 7 policy distance variables and 10 voter characteristics, interacted with dummy variables for the Labour party and the Alliance. Coefficient and variable labels refer to equation 1.11 which describes how latent utility is simulated in these models.

Table A.1: Regression of Party Affect, 1987 Britain Election Study.

Variable	Coef.(S.E.)	Coef. Label	Variable Label
Policy Distances			
Defense	-0.013(.001)***	λ_1	$z_{C,1}, z_{L,1}, z_{A,1}$
Philips Curve	-0.004(.001)***	λ_2	$z_{C,2}, z_{L,2}, z_{A,2}$
Taxes	-0.004(.001)***	λ_3	$z_{C,3}, z_{L,3}, z_{A,3}$
Nationalization	-0.014(.001)***	λ_4	$z_{C,4}, z_{L,4}, z_{A,4}$
Redistribution	-0.010(.001)***	λ_5	$z_{C,5}, z_{L,5}, z_{A,5}$
Crime	0.007(.001)***	λ_6	$z_{C,6}, z_{L,6}, z_{A,6}$
Welfare	-0.011(.001)***	λ_7	$z_{C,7}, z_{L,7}, z_{A,7}$
Voter Characteristics			
Union×Labour	0.173(.047)***	$\beta_{L,1}$	x_1
Gender×Labour	-0.110(.041)***	$\beta_{L,2}$	x_2
Age×Labour	-0.008(.001)***	$\beta_{L,3}$	x_3
Income×Labour	-0.069(.008)***	$\beta_{L,4}$	x_4
South×Labour	-0.598(.112)***	$\beta_{L,5}$	x_5
Midlands×Labour	-0.535(.110)***	$\beta_{L,6}$	x_6
North×Labour	-0.154(.109)	$\beta_{L,7}$	x_7
Wales×Labour	0.107(.132)	$\beta_{L,8}$	x_8
Scotland×Labour	-0.223(.122)*	$\beta_{L,9}$	x_9
Homeowner×Labour	-0.448(.050)***	$\beta_{L,10}$	x_{10}
Labour	0.898(.142)***	$\beta_{L,0}$	
Union×Alliance	0.020(.047)***	$\beta_{A,1}$	x_1
Gender×Alliance	0.148(.041)***	$\beta_{A,2}$	x_2
Age×Alliance	0.003(.001)**	$\beta_{A,3}$	x_3
Income×Alliance	0.000(.008)	$\beta_{A,4}$	x_4
South×Alliance	0.285(.112)**	$\beta_{A,5}$	x_5
Midlands×Alliance	0.207(.110)*	$\beta_{A,6}$	x_6
North×Alliance	0.130(.109)	$\beta_{A,7}$	x_7
Wales×Alliance	0.226(.132)*	$\beta_{A,8}$	x_8
Scotland×Alliance	0.110(.122)	$\beta_{A,9}$	x_9
Homeowner×Alliance	0.066(.050)	$\beta_{A,10}$	x_{10}
Alliance	-1.009(.141)***	$\beta_{A,0}$	
Constant	4.073(.030)***		

$R^2 = .345$, Adjusted $R^2 = .343$.

*** $p < .01$, ** $p < .05$, * $p < .1$

Observations in the regression are uniquely defined by the the individual ID and the party being considered, as in the example dataset in table 1.1. The 1987 British Election Study coding of the variables is as follows: affect is v13a when the choice is Conservative, v13b when the choice is Labour, and the average of v13c and v13d when the choice is Alliance. Labour and Alliance are dummy variables that equal 1 when v8a=2 and 3 respectively. Defense distance through welfare distance are squared differences between

the individual's self placement on the issue (v23a, v28a, v29a, v34a, v35a, v39a, v40a) and the means over all respondents for the party position on each issue (parts b, c, and d of the same question). Union is a dummy that equals 1 if v49c=1 or 2, and 0 if v49c=0. Gender is v58b, age is v58c, and income is v64. Five dummy variables indicate whether the respondent lives in the South, Midlands, North, Wales, or Scotland, and are derived from v48. Homeowner is a dummy that equals 1 if v60ab=02, and 0 otherwise.

Appendix B

Simulation Results for Bayesian MNP in R

The simulations for both data generation models and for all 11 correlation models were also run in R, version 2.6.1. MNP is implemented in the "MNP" library (Imai and van Dyk 2005). MNL is implemented as a conditional logit estimator in the "survival" library (Therneau 2009). MNP is estimated with a Gibbs Sampler, with a default number of draws of 5000. I ran MNL, MNP with the default number of draws, and MNP with 100,000 draws and a burn-in of 50,000 draws. Although any estimation through MCMC needs to be evaluated for convergence, it is evident that the results for the basic simulations converged, while the results for Britain simulations did not converge.

The results for the basic simulations largely resemble the results for the basic simulations in Stata, reported in tables 1.2 and 1.3. Correlations of coefficient point estimates to the known parameters and the percentages of correct signs returned are listed in table B.1. MNP only begins to perform better than MNL for very severe violations of IIA. The Gibbs sampler appears to have converged to the posterior distribution after 5000 draws because estimates are accurate, and do not improve when 100,000 draws are used. In fact, MNP is slightly, though not significantly, less accurate with many more draws. Failure rates for the basic models are reported in table B.2. Note that, as in Stata, MNL rarely fails. Likewise, MNP rarely fails with 5000 draws. However, increasing the number

Table B.1: Basic Model Correlations Between Coefficient Estimates and the True Coefficients, and Percent Correct Signs

	Correlations			*Signs*		
Model	MNL	MNP(5000)	MNP(100,000)	MNL	MNP(5000)	MNP(100,000)
A	0.98(.00)***	0.93(.01)	0.92(.01)	97.0(.43)***	93.1(.67)	93.2(.71)
B	0.97(.00)***	0.94(.01)	0.93(.01)	96.0(.51)***	93.2(.66)	93.5(.70)
C	0.97(.00)***	0.94(.01)	0.93(.01)	95.6(.53)***	93.2(.68)	93.1(.73)
D	0.98(.00)***	0.95(.01)	0.95(.01)	97.1(.43)***	94.7(.60)	94.2(.66)
E	0.98(.00)***	0.97(.00)	0.96(.00)	97.5(.39)***	96.1(.50)	95.7(.55)
F	0.88(.01)	0.96(.00)	0.96(.00)	90.6(.72)	95.5(.57)	95.3(.64)
G	0.93(.01)***	0.89(.01)	0.89(.01)	92.3(.66)**	90.8(.72)	90.4(.80)
H	0.86(.01)	0.94(.01)***	0.93(.01)	91.5(.69)	93.6(.64)	93.9(.68)
I	0.85(.01)	0.93(.01)***	0.91(.01)	89.5(.75)	91.5(.78)	91.6(.86)
J	0.85(.01)	0.92(.01)*	0.92(.01)	88.5(.82)	92.0(.73)	92.1(.78)
K	0.97(.00)***	0.92(.01)	0.92(.01)	94.7(.60)***	91.7(.70)	91.6(.76)

Note: stars indicate that the marked value is significantly greater than the next highest value.
Two-tailed t tests: *** $p < .01$, ** $p < .05$, * $p < .1$

Table B.2: Basic Model Failure Rates.

Model	MNL	MNP(5000)	MNP(100,000)
A	0%	0%	11.0%
B	0%	0%	12.3%
C	0%	0%	11.0%
D	0%	0%	11.0%
E	0.3%	0.3%	12.3%
F	0%	0%	17.0%
G	0%	0%	14.7%
H	4.7%	3.0%	15.3%
I	3.3%	2.7%	16.0%
J	2.0%	1.3%	12.7%
K	0%	0%	14.0%

of draws raises the failure rate of MNP consistently to over 10 percent.

Since the results for the basic simulations in R resemble the Stata results, it is reasonable to expect that the Britain simulations should also resemble the corresponding simulations in Stata. Unfortunately, the Gibbs sampler does not seem to converge to the posterior distribution quickly enough to determine whether or not this is the case. Time is not the issue, rather the algorithm seems to have a non-zero chance of encountering

Table B.3: Britain Model Failure Rates, and Correlations Between Coefficient Estimates and the True Coefficients.

	Fail Rate			*Correlations*		
Model	MNL	MNP(5000)	MNP(100,000)	MNL	MNP(5000)	MNP(100,000)
A	0%	0%	31.3%	0.52(.01)	0.03(.01)	0.05(.01)
B	0%	0%	32.7%	0.54(.01)	0.02(.01)	0.03(.01)
C	0%	0.7%	28.3%	0.53(.01)	0.03(.01)	0.05(.01)
D	2.0%	0.3%	35.7%	0.55(.01)	0.03(.01)	0.05(.01)
E	31.7%	28.7%	52.3%	0.54(.01)	0.00(.02)	0.05(.02)
F	0%	0%	5.7%	0.42(.01)	0.06(.01)	0.07(.01)
G	0%	1.0%	58.0%	0.54(.01)	0.00(.01)	0.03(.02)
H	53.7%	1.7%	21.0%	0.60(.01)	0.04(.02)	0.07(.02)
I	57.3%	0.3%	17.3%	0.57(.02)	0.02(.02)	0.06(.02)
J	51.0%	1.0%	24.7%	0.58(.01)	0.02(.02)	0.05(.02)
K	0%	0%	32.0%	0.53(.01)	0.03(.01)	0.04(.01)

an error and terminating without producing results with each draw. In the three left columns of table B.3, the failure rates for MNL in R are similar to the MNL failure rates reported in table 1.6. MNP with 5000 draws seems to be stable. However, the failure rates for MNP with 100,000 draws approach unacceptable levels. Even model A, the case with completely independent errors, fails nearly a third of the time.

In three right columns of table B.3, the MNP Gibbs sampler clearly has not converged after 5000 draws, yet it also does not converge after 100,000 draws. If more than 100,000 draws are necessary to achieve meaningful results with Bayesian MNP, then it seems likely that the estimator will fail before it can make enough draws. The Britiain simulations are not ridiculous tests of the estimator: only three alternatives are considered with 7 covariates that vary across voters and choices, and 10 covariates that are fixed across choices. Researchers who are considering Bayesian MNP should be aware of the frequency of terminal errors. If the estimator produces results, researchers should seriously consider the question of whether the chain has actually converged

Appendix C

Formulation of BEER

The goal is to estimate the parameters for a linear regression data at one time point, t, by using additional information from other time points. The sufficient statistics for the multivariate normal distribution assumed for the errors in an OLS regression are the coefficients , β, and the residual variance, s^2. The variance-covariance matrix of the coefficients, $V(\hat{\beta}) = \hat{\Sigma}$, is determined by the coefficient estimates and the residual variance. However, in this application, I will treat β_t and Σ_t as unknown. The logic is that information from other time points may change both the coefficient point estimates, and also their standard errors and correlations. The advantage of a Bayesian setup is that information from other time points can be summarized in the prior, which can then be used to update the estimates of β_t and Σ_t. First, in section C.1, the OLS parameters are expressed in a way that conjugate priors exist and can easily be applied. Second, in section C.2, a strategy is discussed to summarize information from different time points within the prior distributions.

C.1 Bayesian Representation of the Regression Parameters

We can express the joint probability of β_t and Σ_t in terms of the conditional probability of β_t given Σ_t and the marginal probability of Σ_t:

$$P(\beta_t, \Sigma_t) = P(\beta_t|\Sigma_t) \cdot P(\Sigma_t). \tag{C.1}$$

Given data X_t and y_t that contain n_t observations and k covariates, the conditional probability of β_t is the likelihood function for β, which follows the multivariate normal distribution,

$$P(\beta_t|\Sigma_t) = L(\beta_t|\Sigma_t, X_t, y_t) = N_k\left(\hat{\beta}_t, \frac{\Sigma_t}{n_t}\right), \tag{C.2}$$

where

$$\hat{\beta}_t = (X_t'X_t)^{-1}X_t'y_t. \tag{C.3}$$

Σ_t follows the inverse-Wishart distribution (Gill 2008, p.570-571):

$$P(\Sigma_t) = L(\Sigma_t|X_t, y_t) = W^{-1}(n_t, \hat{\Sigma}_t), \tag{C.4}$$

where W^{-1} denotes the inverse-Wishart PDF, and

$$\hat{\Sigma}_t = \left[\frac{n_t}{n_t - k}(y_t - X_t\hat{\beta}_t)'(y_t - X_t\hat{\beta}_t)\right](X_t'X_t)^{-1}. \tag{C.5}$$

When expressed with the likelihood functions listed in equations C.2 and C.4, the maximum likelihood estimates for β_t and Σ_t are the same as the OLS estimates for these parameters. This setup for the distributions of β_t and Σ_t may seem overly complicated compared to the standard OLS equations, but this presentation makes it easier to derive the posterior distribution when information from other time points are included as a prior distribution.

Let $L(\beta_t, \Sigma_t|X_t, y_t)$ represent the probability distribution for β_t and Σ_t that depends only on data from time point t. Information from different time points allow us to develop prior beliefs about the parameters, which are contained in the prior distribution

$P(\beta_t, \Sigma_t)$. Bayes' theorem allows us to calculate the posterior distribution:

$$P(\beta_t, \Sigma_t|X, y) \propto L(\beta_t, \Sigma_t|X_t, y_t) \cdot P(\beta_t, \Sigma_t)$$

$$= L(\beta_t|\Sigma_t, X_t, y_t) \cdot L(\Sigma_t|X_t, y_t) \cdot P(\beta|\Sigma) \cdot P(\Sigma). \tag{C.6}$$

Conjugate priors are available for this likelihood function. Following Gill (2008, p.81-84), Gelman et al (2004, p. 87-88), and Robert (2001, p. 189-190), when the following prior distributions are used,

$$\beta_t|\Sigma_t \sim N_k\left(\bar{\beta}, \frac{\Sigma_t}{n_o}\right), \quad \text{and} \quad \Sigma_t \sim W^{-1}(a, B), \tag{C.7}$$

the posterior conditional distribution for β_t is

$$\beta_t|(\Sigma_t, X_t, y_t) \sim N_k\left(\frac{n_o\bar{\beta} + n_t\hat{\beta}}{n_o + n}, \frac{\Sigma_t}{n_o + n_t}\right), \tag{C.8}$$

and the posterior marginal distribution for Σ_t is

$$\Sigma_t|(X_t, y_y) \sim W^{-1}\left(a + n_t, \left[B^{-1} + \hat{\Sigma} + \frac{n_o n_t}{n_o + n_t}(\hat{\beta} - \bar{\beta})(\hat{\beta} - \bar{\beta})'\right]^{-1}\right). \tag{C.9}$$

In order to update the parameters from a regression on data from time point t, we only need to calculate the four posterior arguments using the values of these arguments from the prior distribution and from the regression. The formulas to calculate these posterior arguments are listed in table C.1.

Let n^P, β^P, a^P, and B^P be respectively the sample size, coefficient mean vector, inverse Wishart degrees of freedom, and inverse Wishart scale matrix for the posterior

Table C.1: Formulas to Calculate Posterior Arguments.

Argument	Prior	Data	Posterior
Sample Size	n_o	n_t	$n_o + n_t$
Coefficient means	$\bar{\beta}$	$\hat{\beta}$	$\frac{n_o\bar{\beta} + n_t\hat{\beta}}{n_o + n_t}$
Inverse Wishart Degrees of Freedom	a	n_t	$a + n_t$
Inverse Wishart Scale Matrix	B	$\hat{\Sigma}$	$\left[B^{-1} + \hat{\Sigma} + \frac{n_o n_t}{n_o+n_t}(\hat{\beta} - \bar{\beta})(\hat{\beta} - \bar{\beta})'\right]^{-1}$

distribution. The mean of the inverse Wishart distribution can be used for the variance-covariance matrix of the coefficients (Robert 2001, p. 190):

$$E(\Sigma^P) = \frac{(B^P)^{-1}}{a^P - k - 1}. \tag{C.10}$$

The distribution of the updated coefficients is

$$N_k\left(\beta^P, \frac{\Sigma^p}{n^P}\right), \tag{C.11}$$

therefore β^P contains the coefficient point estimates, and the coefficient standard errors are contained in the square root of the diagonal of $E(\Sigma^P)/n^P$.

C.2 Accumulating Information Over Time Within the Prior Distributions

Information is accumulated through time by using the posterior results for one time point as the prior distribution for the next time point. Two chains are computed which converge on the time point of interest. One chain starts at the first time point and moves forwards through the time point which precedes the one under consideration. The second chain begins at the last time point and updates backwards through the time point under consideration. The posterior results from the first chain are then used as a prior distribution on the posterior results from the second chain.

Suppose that there are T time points in the data which can be labeled times 1 through T. The posterior results for time point $t \in \{1, 2, \ldots, T\}$ can be calculated by computing the forward-moving and backward-moving chains. Unless informed priors are desired, uninformed priors are used for time 1, yielding $\bar{\beta} = \hat{\beta}_1, n = n_1, a = n_1$, and $B = \hat{\Sigma}_1$. These posterior estimates for time 1 are used as the priors for time 2, and β^P, n^P, a^P and B^P are recalculated using the formulas in table C.1. Then these estimates are used as the priors for time 3, and so on, through time $t-1$. Independently, uniformed priors are used for time T, which yields posteriors which are used to compute parameters for time $T-1$, and so on, through time t. Finally, the posterior results for $t-1$ are used as a prior on the results for t.

As noted by Gill, the prior sample size parameter n_o should not necessarily represent the empirically derived prior sample size, rather it is "intended to be a reflection of prior precision relative to the sample size that is tunable by the researcher to reflect prior confidence" (2008, p. 83). When combining information from all of the time points in the data, I reduce the sample size in time point j by multiplying it by a weight parameter $w_{t,j} \in [0, 1]$. This weight, defined in section 2.4, depends on both the similarity and the

proximity of time point j to the time point of interest in the analysis.

The scale matrix of the inverse Wishart matrix, B, must be computed using the algorithm described above. The other three arguments, however, can be updated more efficiently. It can be shown that by combining the forwards-moving and the backwards-moving chains, the posterior or "total updated" sample size is the weighted average of the scaled sample sizes from every time point in the data:

$$n_P = \sum_{j=1}^{T} w_{t,j} \cdot n_j. \tag{C.12}$$

The posterior estimate for the inverse Wishart degrees of freedom is equivalent to posterior sample size. The posterior coefficient means at time point t are also a weighted average of coefficient point estimates:

$$\beta_t^P = \frac{\sum_{j=1}^{T} w_{t,j} \cdot n_j \cdot \hat{\beta}_j}{n_P}. \tag{C.13}$$

Appendix D

Gibbs Sampler for the NL-logit Model

By Bayes' rule, the posterior distribution of parameters in the logistic model are given by

$$P(\theta|X,Y) \propto P(\theta)P(X,Y|\theta), \tag{D.1}$$

where $P(\theta)$ is the prior distribution of the parameters, and $P(X,Y|\theta)$ can be rewritten as the likelihood function for the model:

$$P(X,Y|\theta) = L(\theta|X,Y) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{D.2}$$

where π_i is given by

$$\pi_i = \frac{1}{1+\exp(-y_i^*)}. \tag{D.3}$$

The goal of this discussion is to demonstrate that a general formula for the marginal posterior distributions of the parameters exists for a logistic regression which is general to any specification of $y_i^* = g(x_i, \theta)$ with the restriction that for the J parameters θ, $g : R^J \rightarrow R$. This property of the logistic posterior will justify the use of a Gibbs sampler for the model described in this paper, where y_i^* is defined in equation 3.5. This discussion follows the steps presented by Groenewald and Mokgatlhe (2005, p. 859-860) for a logistic regression which is linear in its variables.

Let each u_i be a random variable distributed Uniform[0,1]. We can rewrite π_i in terms

of u_i:

$$\pi_i = P(y_i = 1) = P\left(u_i < \frac{1}{1 + \exp(-y_i^*)}\right). \tag{D.4}$$

The conditional posterior distribution of u_i is given by

$$u_i | y_i, y_i^* \sim \begin{cases} \text{Uniform}[0, y_i^*] \text{ if } y_i = 1, \\ \text{Uniform}[y_i^*, 1] \text{ if } y_i = 0. \end{cases} \tag{D.5}$$

Given the probability in equation D.4, the following inequalities must be true of y_i^*:

$$y_i^* \geq \log \frac{u_i}{1 - u_i} \text{ if } y_i = 1, \tag{D.6}$$

and

$$y_i^* \leq \log \frac{u_i}{1 - u_i} \text{ if } y_i = 0. \tag{D.7}$$

We use these inequalities to build the marginal distributions of the parameters within y_i^*. Recall that to this point we have not specified the exact specification of y_i^*. Generally, let $y_i^* = g(x_i, \theta)$, and suppose that for a particular set of values for x_i and θ,

$$g(x_i, \theta) = K \tag{D.8}$$

where K is constant. If the set of parameters to be estimated is $\theta = \{\theta_1, \theta_2, \ldots, \theta_J\}$, we can solve equation D.8 for a particular parameter θ_j by specifying functions h and m such that:

$$\theta_j = \frac{K - h(x_i, \theta_{-j})}{m(x_i, \theta_{-j})}, \qquad m(x_i, \theta_{-j}) \neq 0. \tag{D.9}$$

The condition that $m(x_i, \theta_{-j}) \neq 0$ may not be true for all observations i, so the parameter θ_j is only defined when $m(x_i, \theta_{-j}) \neq 0$. If equation D.8 can be solved for parameter θ_j,

which is implied by the restriction that $g : R^J \rightarrow R$, then equations D.6 and D.7 imply that there exists functions h and m such that

$$\theta_j \geq \frac{\log \frac{u_i}{1-u_i} - h(x_i, \theta_{-j})}{m(x_i, \theta_{-j})} \text{ if } y_i = 1, \tag{D.10}$$

and

$$\theta_j \leq \frac{\log \frac{u_i}{1-u_i} - h(x_i, \theta_{-j})}{m(x_i, \theta_{-j})} \text{ if } y_i = 0. \tag{D.11}$$

Given a diffuse prior $P(\theta) \propto 1$, then these inequalities imply that the marginal posterior distribution for a single parameter θ_j is

$$\theta_j | \theta_{-j}, x_i, y_i \sim \text{Uniform}[A_j, B_j], \tag{D.12}$$

where

$$A_j = \min_{i \in \{1,2,\ldots,N\}} \frac{\log \frac{u_i}{1-u_i} - h(x_i, \theta_{-j})}{m(x_i, \theta_{-j})}, \tag{D.13}$$

and

$$B_j = \max_{i \in \{1,2,\ldots,N\}} \frac{\log \frac{u_i}{1-u_i} - h(x_i, \theta_{-j})}{m(x_i, \theta_{-j})}. \tag{D.14}$$

Recall from equation 3.5 that the model considered in this paper is

$$y_i^* = \alpha + \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} + \sum_{k=1}^{K} \delta_k C_{i,k}. \tag{D.15}$$

The parameters for which we need to derive marginal posterior distributions are of four classes: the intercept (α), the effects of demographic controls (δ), the direct effects of the issue distances (β), and the moderating effects of the salience values (γ). The marginal distributions for α, δ, and β are straightforward. For α:

$$\alpha | \beta, \gamma, \delta \sim \text{Uniform}[A, B], \tag{D.16}$$

where

$$A{=}\min_i \log \tfrac{u_i}{1-u_i} - \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} - \sum_{k=1}^{K} \delta_k C_{i,k}, \tag{D.17}$$

and

$$B{=}\max_i \log \tfrac{u_i}{1-u_i} - \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} - \sum_{k=1}^{K} \delta_k C_{i,k}. \tag{D.18}$$

For a particular demographic effect δ_m on variable C_m, the marginal distribution is

$$\delta_m | \alpha, \beta, \gamma, \delta_{-m} \sim \text{Uniform}[A, B], \tag{D.19}$$

where

$$A{=}\min_i \tfrac{1}{C_m} \left[\log \tfrac{u_i}{1-u_i} - \alpha - \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} - \sum_{k \neq m} \delta_k C_{i,k} \right], \tag{D.20}$$

and

$$B{=}\max_i \tfrac{1}{C_m} \left[\log \tfrac{u_i}{1-u_i} - \alpha - \sum_{j=1}^{J} \beta_j I_{i,j} \left[{\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}} \right] \frac{J}{\sum_{j=1}^{J} {\gamma_2}^{S_{i,j,2}} {\gamma_3}^{S_{i,j,3}} {\gamma_4}^{S_{i,j,4}}} - \sum_{k \neq m} \delta_k C_{i,k} \right]. \tag{D.21}$$

The distribution is only defined when $C_m \neq 0$, so the bounds on the uniform distribution are calculated using only the observations for which $C_m \neq 0$.

For a particular issue distance effect β_q on variable I_q, the marginal distribution is

$$\beta_q | \alpha, \beta_{-q}, \gamma, \delta \sim \text{Uniform}[A, B], \tag{D.22}$$

where

$$A{=}\min_i \frac{1}{I_q}\left[\frac{\sum_{j=1}^{J}\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}}}{J}\left(\log\frac{u_i}{1-u_i}-\alpha-\sum_{k=1}^{K}\delta_k C_{i,k}\right)-\sum_{j\neq q}\beta_j I_{i,j}\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}}\right], \tag{D.23}$$

and

$$B{=}\max_i \frac{1}{I_q}\left[\frac{\sum_{j=1}^{J}\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}}}{J}\left(\log\frac{u_i}{1-u_i}-\alpha-\sum_{k=1}^{K}\delta_k C_{i,k}\right)-\sum_{j\neq q}\beta_j I_{i,j}\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}}\right]. \tag{D.24}$$

The distribution is only defined when $I_q \neq 0$, so the bounds on the uniform distribution are calculated using only the observations for which $I_q \neq 0$.

For the moderating effects of salience, a slightly more complex formulation is required. The distribution depends on an individual's particular issue salience response pattern. Suppose that across four issues, a respondent replied that the issues are respectively extremely important, very important, somewhat important, and not important. Then we can rewrite the summation of the salience effects:

$$\sum_{j=1}^{4}\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}} = 1+\gamma_2+\gamma_3+\gamma_4. \tag{D.25}$$

We can also rewrite the summation of the moderated issue effects:

$$\sum_{j=1}^{4}\beta_j I_{i,j}\left[\gamma_2{}^{S_{i,j,2}}\gamma_3{}^{S_{i,j,3}}\gamma_4{}^{S_{i,j,4}}\right] = \beta_1 I_{i,1}+\gamma_2\beta_2 I_{i,2}+\gamma_3\beta_3 I_{i,3}+\gamma_4\beta_4 I_{i,4}. \tag{D.26}$$

For this individual in this example, the candidate bounds for the marginal uniform distribution of γ_2 are given by

$$\frac{(1+\gamma_3+\gamma_4)\left(\log\frac{u_i}{1-u_i}-\alpha-\sum_{k=1}^{K}\delta_k C_{i,k}\right)-J(\beta_1 I_{i,1}+\gamma_3\beta_3 I_{i,3}+\gamma_4\beta_4 I_{i,4})}{\log\frac{u_i}{1-u_i}-\alpha-\sum_{k=1}^{K}\delta_k C_{i,k}-J\beta_2 I_{i,2}}. \tag{D.27}$$

The calculations for γ_3 and γ_4 are symmetric. These candidate bounds are calculated for each individual using their idiosyncratic responses to the salience questions, and the maximum and minimum values across respondents are used for the bounds on the marginal uniform distribution.

Appendix E

Example WinBUGS Code for the NL-Logit Model

The following code implements the Gibbs sampler for the model which uses the 2008 ANES data. In addition to the code listed below, the data need to be specified. I operationalized the data as one large matrix named `X`. Each column of this matrix represents another variable, as defined in the following table:

Code	Variable	Code	Variable	Code	Variable
X[i,1]	Vote (1=Obama)	X[i,18]	Issue: Aid to blacks	X[i,35]	South
X[i,2]	Issue: Gov. service	X[i,19]	Importance: Very	X[i,36]	Gender (1=female)
X[i,3]	Importance: Very	X[i,20]	Importance: Somewhat	X[i,37]	Married
X[i,4]	Importance: Somewhat	X[i,21]	Importance: Not	X[i,38]	Union
X[i,5]	Importance: Not	X[i,22]	Issue: Environment/jobs	X[i,39]	Self employed
X[i,6]	Issue: Defense spending	X[i,23]	Importance: Very	X[i,40]	Protestant
X[i,7]	Importance: Very	X[i,24]	Importance: Somewhat	X[i,41]	Catholoc
X[i,8]	Importance: Somewhat	X[i,25]	Importance: Not	X[i,42]	Jewish
X[i,9]	Importance: Not	X[i,26]	Issue: Womens' role	X[i,43]	Age: young
X[i,10]	Issue: Medical insurance	X[i,27]	Importance: Very	X[i,44]	Age: old
X[i,11]	Importance: Very	X[i,28]	Importance: Somewhat	X[i,45]	Income: low
X[i,12]	Importance: Somewhat	X[i,29]	Importance: Not	X[i,46]	Income: high
X[i,13]	Importance: Not	X[i,30]	Issue: Abortion	X[i,47]	Education: low
X[i,14]	Issue: Gov. jobs	X[i,31]	Importance: Very	X[i,48]	Education: HS plus
X[i,15]	Importance: Very	X[i,32]	Importance: Somewhat	X[i,49]	Education: college
X[i,16]	Importance: Somewhat	X[i,33]	Importance: Not	X[i,50]	Church attendance
X[i,17]	Importance: Not	X[i,34]	Race (1=black)	X[i,51]	Probability weight

```
model{

for (i in 1:N){     ## loop over observations
```

```
g1[i] <- pow(gamma[1],X[i,3])*pow(gamma[2],X[i,4])*pow(gamma[3],X[i,5]);
g2[i] <- pow(gamma[1],X[i,7])*pow(gamma[2],X[i,8])*pow(gamma[3],X[i,9]);
g3[i] <- pow(gamma[1],X[i,11])*pow(gamma[2],X[i,12])*pow(gamma[3],X[i,13]);
g4[i] <- pow(gamma[1],X[i,15])*pow(gamma[2],X[i,16])*pow(gamma[3],X[i,17]);
g5[i] <- pow(gamma[1],X[i,19])*pow(gamma[2],X[i,20])*pow(gamma[3],X[i,21]);
g6[i] <- pow(gamma[1],X[i,23])*pow(gamma[2],X[i,24])*pow(gamma[3],X[i,25]);
g7[i] <- pow(gamma[1],X[i,27])*pow(gamma[2],X[i,28])*pow(gamma[3],X[i,29]);
g8[i] <- pow(gamma[1],X[i,31])*pow(gamma[2],X[i,32])*pow(gamma[3],X[i,33]);

g[i]<-g1[i]+g2[i]+g3[i]+g4[i]+g5[i]+g6[i]+g7[i]+g8[i];

ctrl[i]<-delta[1]*X[i,34]+ delta[2]*X[i,35]+ delta[3]*X[i,36]
+ delta[4]*X[i,37]+ delta[5]*X[i,38]+ delta[6]*X[i,39]
+ delta[7]*X[i,40]+ delta[8]*X[i,41]+ delta[9]*X[i,42]
+ delta[10]*X[i,43]+ delta[11]*X[i,44]+ delta[12]*X[i,45]
+ delta[13]*X[i,46]+ delta[14]*X[i,47]+ delta[15]*X[i,48]
+ delta[16]*X[i,49]+ delta[17]*X[i,50];

ystar[i] <- beta[1] + (8/g[i])*(
 beta[2]*X[i,2]*(g1[i])+ beta[3]*X[i,6]*(g2[i])
+ beta[4]*X[i,10]*(g3[i])+ beta[5]*X[i,14]*(g4[i])
+ beta[6]*X[i,18]*(g5[i])+ beta[7]*X[i,22]*(g6[i])
+ beta[8]*X[i,26]*(g7[i])+ beta[9]*X[i,30]*(g8[i]))
+ctrl[i];
```

```
logit(p[i]) <- ystar[i];  ## logit link

zeros[i]<-0;
zeros[i]~ dpois(phi[i]);
phi[i]<-(-X[i,51]*(X[i,1]*log(p[i]) + (1-X[i,1])*log(1-p[i])))+100;
llh[i]<- X[i,51]*(X[i,1]*log(p[i]) + (1-X[i,1])*log(1-p[i]));
}

sumllh <- sum(llh[]);  # sum of log-likelihood contributions

## priors
for(j in 1:9){
beta[j] ~ dnorm(0 ,.00001) # diffuse Normal priors
}
for (l in 1:3){
gamma[l] ~ dnorm(0,.00001)
}
for (m in 1:17){
delta[m] ~ dnorm(0,.00001)
}

}

Inits
list(beta=c(1,1,1,1,1,1,1,1,1),
delta=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
```

```
gamma=c(1,1,1));
```

References

Albert, James H. and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association.* 88(422): 669-679.

Alvarez, R. Michael and Gabriel Katz. 2009. "Structural Cleavages, Electoral Competition and Partisan Divide: A Bayesian Multinomial Probit Analysis of Chile's 2005 Election." *Electoral Studies.* 28(2): 177-189.

Alvarez, R. Michael and Jonathan Nagler. 1994. "Correlated Disturbances in Discrete Choice Models: A Comparison of Multinomial Probit Models and Logit Models." Working Papers 914, California Institute of Technology, Division of the Humanities and Social Sciences.

—. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science.* 39(3): 714-744.

—. 1998. "When Politics and Models Collide: Estimating Models of Multiparty Elections." *American Journal of Political Science.* 42(1): 55-96.

—. 2000. "A New Approach for Modeling Strategic Voting in Multiparty Elections." *British Journal of Political Science.* 30(1): 57-75.

Alvarez, Michael R., Jonathan Nagler, and Shaun Bowler. 2000. "Issues, Economics, and the Dynamics of Multiparty Elections: The British 1987 General Election." *The American Political Science Review.* 94(1): 131-149.

Bartels, Brandon L. 2009. "Beyond 'Fixed Versus Random Effects': A Framework for Improving Substantive and Statistical Analysis of Panel, Time-Series Cross-Sectional, and Multilevel Data." Under review, *APSR.*

Beck, Nathaniel, and Jonathan Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *American Political Science Review.* 89:(634-647).

Blais, André and Ludovic Rheault. 2010. "Optimists and Skeptics: Why Do People Believe in the Value of Their Single Vote?" *Electoral Studies.* Forthcoming.

Bolduc, Denis. 1999. "A Practical Technique to Estimate Multinomial Probit Models in Transportation." *Transportation Research Part B: Methodological.* 33(1): 63-79.

Brownstone, David and Kenneth Train. 1999. "Forecasting New Product Penetration with Flexible Substitution Patterns." *Journal of Econometrics.* 89(1-2): 109-129.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Campbell, David E. and J. Quin Monson. 2008. "The Religion Card: Gay Marriage and the 2004 Presidential Election." *Public Opinion Quarterly.* 72(3): 399-419.

Cheng, Simon and J. Scott Long. 2007. "Testing for IIA in the Multinomial Logit Model." *Sociological Methods and Research.* 35(4): 583-600.

Chib, Suddhartha. 1998. "Estimation and Comparison of Multiple Change-Point Models." *Journal of Econometrics*. 86:22141.

Clarke, Harold D., Thomas J. Scotto, and Allan Kornberg. 2010. "Valence Politics and Economic Crisis: Electoral Choice in Canada 2008." *Electoral Studies*. Forthcoming.

Cleveland, William S. and Susan J. Devlin. 1988. "Locally Weighted Regression: An Approach to Local Regression by Local Fitting." *Journal of the American Statistics Association*. 83(403): 596-610.

Cowles, Mary Kathryn and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association*. 91(494): 883-904.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

DeBouf, Suzanna and Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science*. 52(1):184-200.

Dow, Jay K. and James W. Endersby. 2004. "Multinomial Probit and Multinomial Logit: A Comparison of Choice Models for Voting Research." *Electoral Studies*. 23(1): 107-122.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth*. 8(2):195-222.

Fisher, Stephen D. and Sara B. Hobolt. 2010. "Coalition Government and Electoral Accountability." *Electoral Studies*. 29(3): 358-369.

Fry, Tim R. L. and Mark N. Harris. 1996. "A Monte Carlo Study of Tests for the Independence of Irrelevant Alternatives Property." *Transportation Research Part B: Methodological*. 30(1): 19-30.

Fullerton, Andrew S., Jeffery C. Dixon and Casey Borch. 2007. "Bringing Registration into Models of Vote Overreporting." *Public Opinion Quarterly*. 71(4): 649-660.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Second Ed. Texts in Statistical Science. New York: Chapman and Hall.

Gelman, Andrew, David Park, Boris Shor, Joseph Bafumi, and Jeronimo Cortina. 2008. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*. Princeton, NJ: Princeton University Press.

Geweke, John. 1991. "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints." *Computer Science and Statistics: Processings of the Twenty-Third Symposium on the Interface*. Alexandria, VA: American Statistical Association. 571-578.

—. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In *Bayesian Statistics 4*. Eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press, 169-193.

Geweke, John, Michael Keane, and David Runkle. 1994 "Alternative Computational Approaches to Inference in the Multinomial Probit Model." *Review of Economics and Statistics*. 76(4): 609-632.

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Statistics in the Social

and Behavioral Sciences Series. New York: Chapman and Hall.

Glasgow, Garrett. 2001. "Mixed Logit Models for Multiparty Elections." *Political Analysis*. 9(1): 116-136.

Gould, William, Jeffrey Pitblado, and William Sribney. 2006. *Maximum Likelihood Estimation with Stata*. College Station, TX: Stata Press.

Greene, William H. 2000. *Econometric Analysis*. Fourth Ed. Upper Saddle River, NJ: Prentice Hall.

—. 2003. *Econometric Analysis*. Fifth Edition. Upper Saddle River, NJ: Pearson Education, Inc.

Groenewald, Pieter C. N. and Lucky Mokgatlhe. 2005. "Bayesian Computation for Logistic Regression." *Computational Statistics and Data Analysis*. 48: 857-868.

Gross, Wendy, Jonathan Kropko, Jon A. Krosnick, Stuart Elaine Macdonald, and George Rabinowitz. 2010. "The Influence of Personal Importance in Issue Voting Models." Presented at the annual meeting of the MPSA Annual National Conference, Palmer House Hotel, Hilton, Chicago, IL, April 2010.

Grynaviski, Jeffrey D. and Bryce E. Corrigan. 2006. "Specification Issues in Proximity Models of Candidate Evaluation (with Issue Importance)." *Political Analysis*. 14(4): 393-420.

Guetzkow, Joshua, Bruce Western, and Jake Rosenfeld. 2007. "State-Level Data on Income Inequality: 1963-2003." Russell Sage Program on the Social Dimensions of Inequality. <http://www.inequalitydata.org/> Data accessed 10 October 2010.

Hajivassiliou, Vassilis A. and Daniel L. McFadden. 1998. 'The Method of Simulated Scores for the Estimation of LDV Models." *Econometrica*. 66(4): 863-896.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Rudd. 1996. "Simulation of Multivariate Normal Rectangle Probabilities and Their Derivatives: Theoretical and Computational Results." *Journal of Econometrics*. 72(1-2): 85-134.

Hausman, Jerry A. and David A. Wise. 1978. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." *Econometrica*. 46(2): 403-426.

Heath, A., R. Jowell, and J.K. Curtice. 1987. *British Election Study: Cross-Section, 1987*. Computer file: ICPSR06452-v1. Colchester, England: ESRC Data Archive/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 1995. doi:10.3886/ICPSR06452.

Heidelberger, Philip and Patrick D. Welch. 1983. "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*. 31(6): 1109-1144.

Hole, Arne Risa. 2007. "Fitting Mixed Logit Models By Using Maximum Simulated Likelihood." *The Stata Journal*. 7(3): 388-401.

Hooghe, Liesbet, Gary Marks and Arjan Schakel. 2008. "Operationalizing Regional Authority: A Coding Scheme for 42 Countries, 1950-2006." *Regional and Federal Studies*. 18(2,3):123-142.

Humphreys, Macartan and Jeremy M. Weinstein. 2008. "Who Fights? The Determinants of Participation in Civil War." *The American Journal of Political Science*. 52(2): 436-455.

Imai, Kosuke and David A. van Dyk. 2005. "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation." *Journal of Econometrics*. 124(2): 311-334.

Ivarsflaten, Elisabeth. 2008. "What Unites Right-Wing Populists in Western Europe?: Re-Examining Grievance Mobilization Models in Seven Successful Cases." *Comparative Political Studies*. 41(1): 3-23.

Kam, Cindy D. 2007. "Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preference." *Political Behavior*. 29(3): 343-367.

Keane, Michael P. 1992. "A Note on Identification in the Multinomial Probit Model." *Journal of Business and Economic Statistics.* 10(2):193-201.

Keane, Michael P. and Kenneth I. Wolpin. 1994. "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence." *Review of Economics and Statistics*. 76(4): 648-672.

Kittle, Bernhard. 1999. "Sense and Sensitivity in Pooled Analysis of Political Data." *European Journal of Political Research*. 35: 225-253.

—. 2006. "A Crazy Methodology? On the Limits of Macro-Quantitative Social Science Research." *International Sociology*. 21(5):647-677.

Leip, David. 2011. "Dave Leip's Atlas of U.S. Presidential Elections." <http://uselectionatlas.org/> Accessed 1 March 2011.

Long, J. Scott and Jeremy Freese. 2005. *Regression Models for Categorical Dependent Variables Using Stata*. Second Edition. College Station, TX: Stata Press.

Macdonald, Stuart Elaine, George Rabinowitz, and Ola Listhaug. 2007. "Simulating Models of Issue Voting." *Political Analysis*. 15(4): 406-427.

Marks, Gary, Liesbet Hooghe and Arjan Schakel. 2008. "Measuring Regional Authority." *Regional and Federal Studies*. 18(2,3):111-121.

McCullagh, P. and John A. Nelder. 1989. *Generalized Linear Models.* Second Ed. London: Chapman and Hall.

McFadden, Daniel and Jerry Hausman. 1984. "Specification Tests for the Multinomial Logit Model." *Econometrica*. 52(5): 1219-1240.

McFadden, Daniel and Kenneth Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics*. 15(5): 447-470.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica*. 46(1): 69-85.

Niemi, Richard G. and Larry M. Bartels. 1985. "New Measures of Issue Salience: An Evaluation." *The Journal of Politics*. 47(4): 1212-1220.

Park, Jong Hee. 2009. "Joint Modeling of Dynamic and Cross-Sectional Heterogeniety: Introducing

Hidden Markov Panel Models." Working paper.

Plummer, Martyn, Micky Best, Kate Cowles, and Karen Vines. 2010. "Package 'coda'". *The Comprehensive R Archive Network*. <http://cran.r-project.org/>

Plumper, Thomas and Vera E. Trogger. 2007. "Efficient Estimation of Time-Invariance and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects." *Political Analysis*. 15:124-139.

Quinn, Kevin M., Andrew D. Martin, and Andrew B. Whitford. 1999. "Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models." *American Journal of Political Science*. 43(4): 1231-1247.

Rabe-Hesketh, Sophia and Anders Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. Second Ed. College Station, TX: Stata Press.

Raftery, Adrian E. and Steven M. Lewis. 1992. "How Many Iterations in the Gibbs Sampler?" In *Bayesian Statistics 4*. Eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press, 763-773.

Revelt, David and Kenneth Train. 1998. "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level." *The Review of Economics and Statistics*. 80(4): 647-657.

Robert, Christian P. 2001. *The Bayesian Choice*. Second Ed. New York: Springer.

Schofield, Normal et al. 1998. "Multiparty Electoral Competition in the Netherlands and Germany: A Model Based on Multinomial Probit." *Public Choice*. 97(2): 257-293.

Spiegelhalter, David, Andrew Thomas, Nicky Best, and Dave Lunn. 2007. "Tricks: Advanced Use of the BUGS Language." *WinBUGS User Manual*. Version 1.4.3.

Spirling, Arthur. 2007. "Bayesian Approaches for Limited Dependent Variable Change Point Problems." *Political Analysis*. 15(4).

Therneau, Terry. 2009. *Survival Analysis, Including Penalised Likelihood*. R-Forge 11384. Available at http://r-forge.r-project.org.

Train, Kenneth E. 2003. *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

United States Bureau of Economic Analysis. 2011. *Gross Domestic Product by State*. <http://bea.gov/regional/gsp/default.cfm> Data accessed 10 October 2010.

Van Groezen, Bas, Hannah Kiiver and Brigette Unger. 2009. "Explain Europeans' Preferences for Pension Provision." *European Journal of Political Economy*. 25(2): 237-246.

Venables, W. N., D. M. Smith, and the R Development Core Team. 2010. "An Introduction to R." Version 2.12.1.

Western, Bruce and Meredith Kleykamp. 2004. "A Bayesian Change Point Model for Historical Time Series Analysis." *Political Analysis*. 12(4):354-374

Whitten, Guy D. and Harvey D. Palmer. 1996. "Heightening Comparativists' Concern for Model Choice: Voting Behavior in Great Britain and the Netherlands." *American Journal of Political Science*. 40(1): 231-260.

Wilson, Carole J. 2008. "Consideration Sets and Political Choices: A Heterogeneous Model of Vote Choice and Sub-national Party Strength." *Political Behavior*. 30(2): 161-183.

Wilson, Sven E. and Daniel M. Butler. 2007. "A Lot More To Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications". *Political Analysis*. 15:101-123.

Zorn, Christopher. 2001. "Estimating Between- and Within-Cluster Covariate Effects, With an Application to Models of international Disputes." *International Interactions*. 27:433-445.