Time Series Models for Discrete Data: solutions to a problem with quantitative studies of international conflict

Simon Jackman[1]

July 21, 1998

[1]Department of Political Science, 455 Serra Mall, Building 160, Stanford University, Stanford, CA 94305-2044, USA. e-mail: jackman@stanford.edu I thank Neal Beck, Brad Carlin, Kate Cowles, Claudia Czado, Richard Juster, Jonathan Katz, and Douglas Rivers for useful comments and discussion, and Adelchi Azzalini, Neal Beck, Jonathan Katz, Richard Tucker, and Joanne Gowa for sharing their data. Errors and omissions remain my responsibility.

Abstract

Discrete dependent variables with a time series structure occupy something of a statistical limbo for even well-trained political scientists, prompting awkward methodological compromises and dubious substantive conclusions. An important example is the use of binary response models in the analysis of longitudinal data on international conflict: researchers understand that the data are not independent, but lack any way to model serial dependence in the data. Here I survey methods for modeling categorical data with a serial structure. I consider a number of simple models that enjoy frequent use outside of political science (originating in biostatistics), as well as a logit model with an autoregressive error structure (the latter model is fit via Bayesian simulation using Markov chain Monte Carlo methods). I illustrate these models in the context of international conflict data. Like other re-analyses of these data addressing the issue of serial dependence, (e.g., Beck, Katz and Tucker 1998), I find economic interdependence does not lessen the chances of international conflict. Other findings include a number of interesting asymmetries in the effects of covariates on transitions from peace to war (and vice versa). Any reasonable model of international conflict should take into account the high levels of persistence in the data; the models I present here suggest a number of methods for doing so.

> *Underlying virtually all the previous discussion lies a quite strong assumption of independence. When, as is quite often the case, this assumption is abandoned, appreciable complications are to expected....*
>
> Cox and Snell, *The Analysis of Binary Data* (2nd edition, p96).

1 Discrete Data, Time Series, and Political Methodology

For most political scientists, training in quantitative methods follows a similar path. In expanding our statistical toolkit beyond linear regression, political scientists traditionally proceed in two directions: (a) models for qualitative dependent variables *in a cross-sectional setting* (e.g., probit / logit for discrete data, Poisson models for counts), and (b) models for continuous dependent variables *in a time series setting* (e.g., Box-Jenkins ARIMA models, error-correction and co-integration, vector auto-regressions).[1]

But data possessing both of these characteristics — that is, qualitative data with a serial structure — pose something of a dilemma even for well-trained social scientists. Most political scientists possess no "off-the-shelf" model for dealing with data of this type. Faced with qualitative time-series data, most political scientists are forced to make a choice, which I sketch in Figure 1. The only way to atend to the dynamic character of the data is to ignore their qualitative structure; on the other hand, the only way to deal with the qualitative structure is to ignore the fact that the data possess serial dependence.

How do political scientists deal with this choice? Almost always the time series characteristics of the data lose out to the qualitative characteristics of the data, and logit / probit or Poisson models are estimated. This occurs for a variety of reasons, some sociological, some statistical. On the former point, models for qualitative dependent variables arguably have more prominence in graduate-level methodological training than time series methods. In addition, in a field like international relations, the methodological "step up" from regression to logit / probit or Poisson models attracts more professional and methodological kudos than, say, running a regression on a categorical dependent variable, perhaps with a Cochrane-Orcutt band-aid for a residual AR(1) process. And frankly, ignoring serial dependencies in the (qualitative) data is probably the lesser statistical evil. For instance, the shortcomings of running a regression on qualitative data (the so-called "linear probability model") are numerous and well-known to political scientists (Aldrich

[1] I note at the outset that models of duration and event histories do not lie comfortably within this dichotomy; and indeed, Beck, Katz and Tucker (1998) show how duration models can be used to capture serial dependence in longitudinal binary data encountered in the democratic peace literature.

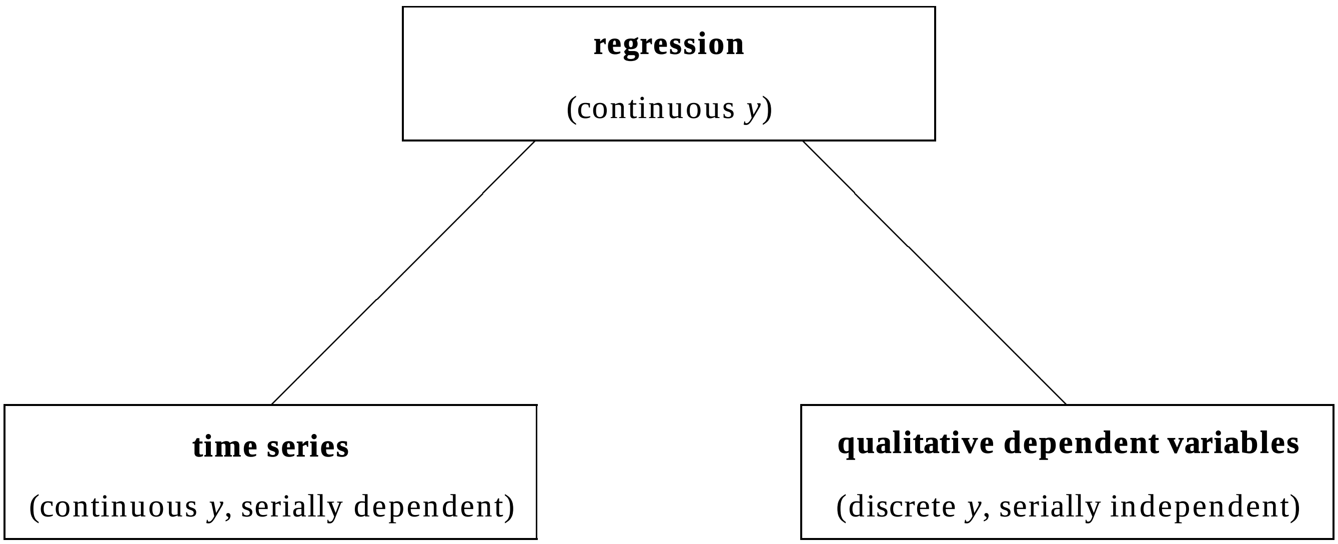

Figure 1: ***A stylized, graphical rendering of quantitative political science.*** *Beyond regression models for continuous, serially independent, dependent variables, political scientists typically proceed in one of the two directions indicated in the graph. When confronted with categorical time series data, political scientists must decide which feature of the data to deal with — qualitative characteristics, or serial structure — since there are no off-the-shelf techniques available for dealing with both.*

and Nelson 1984). On the other hand, it is also well known that estimators of the effects of covariates can remain unbiased and consistent even in the presence of autocorrelation. Indeed, Poirier and Ruud (1988) show that an "ordinary probit" estimated in the presence of autocorrelated disturbances still yields consistent estimates, if less efficient than a "generalized" estimator that dealt with the serial dependence. Nonetheless, the inefficient estimates produced by ignoring any serial dependence raise the possibility of inferential errors, and are a serious problem.

Numerous examples of political scientists encountering this dilemma appear in quantitative studies of international relations. For instance, Oneal and Russett (1997) employ logit models of a binary dependent variable (dispute / no dispute), with data at the level of the dyad-year. They acknowledge that "the greatest danger arises from autocorrelation, but that there are not yet generally accepted means of testing for or correcting this problem in logistic regressions." Working with similar data, Farber and Gowa (1997, 397) note that serial dependence threatens the validity of their analysis of dyad-year data, but "proceed ignoring this lack of independence" since "a better solution is not obvious".

Likewise, consider Gowa's (1998) study of the United States' involvement in "militarized interstate disputes" (MIDs) between 1870 and 1992. The dependent variable is a count of the MIDs the United States is involved in *per year*, data that are simultaneously qualitative and serial.[2] Faced with the choice between treating these data as *either* a time series *or* as qualitative count data, Gowa chose the later, estimating a Poisson model by maximum likelihood. Gowa notes that "[s]ince the values of the dependent variables are discrete and truncated at zero, an OLS regression does not generate efficient parameter estimates" (p18). By the same token, ignoring serial correlation in one's data creates inferential dangers too. But in moving to the discrete (Poisson) data framework, the dangers of serial correlation are ignored, since social scientists typically lack the tools for testing and / or correcting for serial correlation in discrete data.

2 Classes of Models for Discrete Serial Data

In this paper I attempt to fill this methodological void between time series and qualitative dependent variables. A useful starting point is the large literature in biostatistics

[2]These counts come from the Correlated of War (COW) data set (Singer and Small 1994); the operational definition of a MID is as an international event that involves "government-sanctioned" "threats to use military force, displays of military force, ... actual uses of force," and war (Gochman and Maoz 1984), as quoted in (Gowa 1998). Predictor variables include a GNP growth, a series of political indicator variables, and a series of period-specific indicator variables; the hypothesis of interest is the extent to which domestic political or economic conditions increase the involvement of the United States in MIDs.

on models for discrete (and continuous) "time series" data. Biostatisticians encounter an abundance of time series data, both in the form of reasonably short panel studies, and in longer longitudinal designs. Typical examples are monitoring a set of patients or subjects over time, sometimes in an experimental setting (as in pharmaceutical trials), generating "repeated measures" of subjects. Furthermore, bio-statistical outcomes are often discrete in nature. Different types of categorical times series can be distinguished, depending on the nature of the response, and the design of the study. For instance,

- daily rainfall data (1 if measurable precipitation, 0 otherwise) on June days in Madison, Wisconsin for various years (Klotz 1973) are pooled binary time series
- a study in which children are monitored for respiratory disease (Sommer, Katz and Tarwotjo 1984; Ware et al. 1984), or experimental subjects report arthritic pain (Bombardier and Russell 1986) over a sequence of observations generates repeated measures on a binary response;
- the number of asthma presentations recorded per day by an emergency room (Davis, Dunsmuir and Wang 1998), or the number of deaths in the British coalmining industry (Carlin, Gelfand and Smith 1992) is a time series of counts,
- the number of deaths by horse kicks in 14 corps of the Prussian army (Andrews and Herzberg 1985, 18), or the number of patents received by 642 firms between 1975-79 (Hausman, Hall and Grilliches 1984; Chib, Greenberg and Winkelmann 1998) are repeated measures of counts.

Reviews of the biostatistical literature (e.g., Diggle, Liang and Zeger 1994; Ashby et al. 1992) distinguish three classes of models for discrete data with a serial structure. **Marginal models** deliberately divorce the issue of serial dependence from modeling the effects of covariates. Serial dependence is considered a nuisance in this model and captured via association parameters. The model for y is a marginal model in the sense that there is no explicit conditioning on the history of the discrete responses. This model is employed when multiple time series are being modeled (i.e., panels or pooled time-series) and effectively average away any variation reflected in the unit-specific response histories.

Models with **random effects** address the issue of heterogeneity across units by allowing the marginal model to include unit-specific terms, that are drawn from a population distribution. A frequently encountered example in the biostatistical literature is to augment a panel study with unit-specific linear or quadratic time trends, or even simply a unit-specific intercept capturing unobserved heterogeneity across units. However, when the panels are short and there is very little data per unit, estimates of these unit-specific coefficients will tend to be imprecise. The random coefficients approach substitutes the imprecise-but-unbiased unit-specific estimates with estimates drawn from a population, “borrowing strength” across units. Random effects models are far from unknown in political science:[3] applications include King’s (1997) method for estimating unit-specific parameters from aggregate data and Western’s (1998) analysis of variation in the determinants of economic growth among OECD countries. Alvarez and Glasgow (1997) use a simple random effects model to capture individual-level heterogeneity in a panel study of voters over the 1976 and 1980 presidential campaigns (with a continuous dependent variable), while Plutzer (1997) estimated a random effects growth curve tapping change in voter turnout over the three-wave Jennings and Niemi (1981) Student-Parent Socialization Study (with a dichotomous dependent variable). Given the relative familiarity of these models, I will not elaborate further.

Transitional models explicitly incorporate the history of the responses in the model for y_t. In this way each unit-specific history can be used to generate forecasts for that unit, as opposed to the marginal model which makes forecasts solely on the basis of the values of exogenous variables. In addition, Cox (1981) distinguishes two types of transitional models; observation-driven models (where the “transitions” are with respect to observed data) and parameter-driven models (where the “transitions” are with respect to a latent process, and tapped by a transitional parameter). As we shall see, Markov chains are an important component of observation-driven transitional models for discrete data. I also consider a parameter-driven transitional model, where I posit an AR(1) error process on the regression function for the latent variable underlying the discrete responses.

This typology is neither exhaustive nor mutually exclusive. The lengthy review in MacDonald and Zucchini (1997, ch1) reveals no shortage of proposals for characterizing and modeling discrete time series, though only a modest proportion of these proposals have been accompanied by applications. Furthermore, there are also models that are hybrids and there are even differences within categories of these models.

As the research designs shift from short panels to longer time series studies, the models tend to shift in emphasis from attempting to capture “correlation”, “dependence” or “association” in the marginal probabilities, to dealing with the data as time series *per sé*. As the length of the series becomes longer, the serial component of the data is less a “nuisance” and perhaps more a “feature” of the data, and so in these circumstances a transitional model might be preferred to a marginal model. In other instances, the data

[3]Jones and Steenbergen (1997) provides a fairly general summary of the technique for political scientists, from the perspective of multilevel models, also known as hierarchical models.

are explicitly transitional and the past history of an individual subject is important.[4] Then again, despite having a long time series, the nature of the study may be such that interest focuses on a marginal model, in which serial dependence is considered a nuisance. In short, depending on the goal of a particular study and the research design, one modeling strategy may be preferred over others.

3 The Binary Response Model

So as to fix ideas and to introduce some notation, I present a model for independent binary data, $y_t \in \{0, 1\}$, $t = 1, \ldots, T$. Let $\theta_t \equiv \Pr[y_t = 1]$, which in turn depends on covariates via a latent regression function

$$h(\theta_t) \equiv y_t^* = \mathbf{x}_t \boldsymbol{\beta} + u_t,$$

where

- $\mathbf{x}_t$ is a row vector of observations on k independent variables at time t,
- $\boldsymbol{\beta}$ is a column vector of parameters to be estimated,
- $y_t^* \in \mathbb{R}$ is a latent dependent variable, observed only in terms of its sign:

$$y_t = \begin{cases} 0, & \text{if } \; y_t^* \leq 0 \\ 1, & \text{if } \; y_t^* > 0 \end{cases}$$

- u_t is a zero mean stochastic disturbance, identically and independently distributed for all t. For probit, we will assume $f(u_t) = N(0, 1) \equiv \phi(), \forall\ t$, while for logit we assume a logistic distribution, also normalized to have unit variance.[5]

The function $h() : [0, 1] \to \mathbb{R}$ is a *link* function (McCullagh and Nelder 1989) known to the analyst, while the inverse link function $h^{-1}()$ maps the linear predictors into probabilities:

[4] Consider a learning model, in which experimental subjects receive a reward or punishment over a series of trials (e.g., Lindsey 1995, 165ff). In this case the history of a specific individual is of interest, and not just their characteristics as measured with a set of covariates.

[5] The regression parameters $\boldsymbol{\beta}$ are identified only up to the scale factor σ, and so setting $\sigma = 1$ is a convenient normalization with no substantive implications.

i.e.,

$$\begin{aligned}\Pr[y_t = 1] = \Pr[y^* > 0] = \Pr[\mathbf{x}_t\boldsymbol{\beta} + u_t > 0] = \Pr[u_t > -\mathbf{x}_t\boldsymbol{\beta}] &= \int_{-\mathbf{x}_t\boldsymbol{\beta}}^{\infty} f(u_t)du_t \\ &= 1 - \int_{\infty}^{-\mathbf{x}_t\boldsymbol{\beta}} f(u_t)du_t.\end{aligned}$$

For probit and logit $f(u_t)$ is symmetric about 0, so

$$\Pr[y_t = 1] = \int_{-\infty}^{\mathbf{x}_t\boldsymbol{\beta}} f(u_t)du_t \quad \text{and} \quad \Pr[y_t = 0] = \int_{-\infty}^{-\mathbf{x}_t\boldsymbol{\beta}} f(u_t)du_t$$

Substituting the respective functional forms of $f(u)$ yields

$$\theta_t = \Pr[y_t = 1] = h^{-1}(\mathbf{x}_t\boldsymbol{\beta}) = \begin{cases} \Phi(\mathbf{x}_t\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_t\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] dz & \text{(probit)} \\ \Lambda(\mathbf{x}_t\boldsymbol{\beta}) = \frac{\exp \mathbf{x}_t\boldsymbol{\beta}}{1 + \exp(\mathbf{x}_t\boldsymbol{\beta})} & \text{(logit).} \end{cases}$$

Serial independence means that

$$\Pr(y_1 = 1, y_2 = 1, \ldots, y_T = 1) = \Pr(y_1 = 1)\Pr(y_1 = 2)\ldots\Pr(y_T = 1),$$

or, in words, the joint probability equals the product of the marginal probabilities, and so the log-likelihood of the data equals the sum of the individual log-likelihoods. That is,

$$\ln L = \sum_{t=1}^{T} \left[y_t \ln\theta_t + (1 - y_t)\ln(1 - \theta_t)\right]. \tag{1}$$

The log-likelihood function in (1) is easily maximized to yield consistent estimates of $\boldsymbol{\beta}$ whose asymptotic distribution is multivariate normal. Moreover, the matrix of second derivatives of the log-likelihood function (1) with respect to $\boldsymbol{\beta}$ are easily calculated and yield an estimate for the asymptotic covariance matrix of the estimated $\boldsymbol{\beta}$.

4 Observation-Driven Transitional Model

Diggle, Liang and Zeger (1994, 135) define a transitional model as one in which correlation among the discrete responses arises because past responses explicitly influence the present outcome. Past realizations of the discrete response help determine the current response, and so enter the model as additional predictor variables. A two-state Markov chain is a straightforward way to model a binary time series, and forms the basis of the transitional model. A first-order Markov chain for binary data has a transition matrix

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

where $p_{ij} = \Pr(y_t = j | y_{t-1} = i)$. With binary data there are just two unique elements of the 2-by-2 transition matrix. A simple way to relate covariates to the transitional probabilities is via a link function and a regression structure for each of the two transition probabilities. With a logit link we have

$$\text{logit}\left[\Pr(y_t = 1 | y_{t-1} = 0)\right] = \mathbf{x}_t \boldsymbol{\beta}_0$$

and

$$\text{logit}\left[\Pr(y_t = 1 | y_{t-1} = 1)\right] = \mathbf{x}_t \boldsymbol{\beta}_1$$

where the possibility that $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_1$ taps the possibility that the effects of explanatory variables will differ depending on the previous response. Diggle, Liang and Zeger (1994, 195) show that the two equations above can be combined to form the model

$$\text{logit}\left[\Pr(y_t = 1 | y_{t-1})\right] = \mathbf{x}_t \boldsymbol{\beta}_0 + y_{t-1} \mathbf{x}_t \boldsymbol{\alpha} \tag{2}$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\alpha}$, since when $y_{t-1} = 0$ the $\mathbf{x}_t \boldsymbol{\alpha}$ part of the model is zeroed out. This is an extremely simple format with which to test hypotheses about the effects of the covariates on the transition probabilities. Tests of the null hypothesis $\boldsymbol{\alpha} = \mathbf{0}$ tap whether the effects of $\mathbf{x}$ are constant irrespective of the previous state of the binary process. Furthermore, if the first-order Markov specification is correct, then ordinary binary response models can be used to estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}$ and their standard errors. The model is estimated conditional on the first observation in the time series, and care needs to taken interpreting predicted probabilities and model summaries, since the model is now predicting conditional probabilities.

This model enjoys widespread use in the bio-statistics literature, with Cox (1970, 72ff) drawing attention to the link between the transition probabilities for a Markov chain and a logistic regression. Korn and Whittemore (1979) applied the model to panel data on the effects of air pollution; Stern and Cole (1984) used the model to estimate a model of rainfall data. The model extends readily to the case of an ordered or multinomial outcome, and Zeger and Qaqish (1988) extend the model for count data.

5 A Parameter-Driven Transitional Model

I now consider a transitional model but in which the transitions are with respect to the latent dependent variable $\mathbf{y}^*$. In some ways this is a fairly natural way to approach the serial dependence in the binary responses, since it combines familiar time series methods for continuous y variables with the standard binary response model given in section 3. Unfortunately, as we shall see, the offspring of the union of these two familiar models is something of a monster, or at the very least, a problem child.

I introduce serial dependence in the latent variable $\mathbf{y}^*$ by positing a stationary, zero mean, autoregressive process on the disturbances $\{u_t\}$. For simplicity, consider the AR(1) process

$$u_t = \rho\, u_{t-1} + \epsilon_t, \tag{3}$$

where ϵ_t is zero-mean Gaussian white noise. The model for $\mathbf{y}^*$ is now a regression with AR(1) errors, a model with properties well known to political scientists. I briefly re-state a number of these properties, which will help us understand why this seemingly simple model is so complicated to estimate in the context of a discrete response.

We retain the assumption made for the binary response model in section 3 that $\text{var}(u_t) \equiv \sigma_u^2 = 1$, losing the t subscript via the stationarity assumption. Since the ϵ_t are orthogonal with the u_t, the variance of u_t can be decomposed as

$$\text{var}(u_t) = \rho^2\, \text{var}(u_{t-1}) + \text{var}(\epsilon_t).$$

Exploiting the stationarity assumption,

$$\text{var}(u_t) = \rho^2\, \text{var}(u_t) + \text{var}(\epsilon_t)$$

or

$$\sigma_u^2 = \rho^2\, \sigma_u^2 + \sigma_\epsilon^2,$$

and since $\sigma_u = 1$, $\sigma_\epsilon^2 = 1 - \rho^2$; note that stationarity also requires $|\rho| < 1$ and so $0 < \sigma_\epsilon^2 < 1$. Note also that $\text{cov}(u_t, u_s) = \rho^{|s-t|}/(1-\rho^2)$. Thus, the variance-covariance matrix for $\mathbf{u} =$

$(u_1, u_2, \ldots, u_T)'$ is

$$\Sigma_u(\rho) = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix} \neq \mathbf{I}_T, \forall\, \rho \neq 0.$$

The likelihood function for the binary response model with an autoregressive error structure is considerably more complex than the likelihood for the serially independent model. The joint probability $\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T)$ can't be factored into the product of the observation-specific marginal probabilities. Instead,

$$\begin{aligned} L &= \Pr[y_1, y_2, \ldots, y_T] \\ &= \int_{a_1}^{b_1}\int_{a_2}^{b_2}\ldots\int_{a_T}^{b_T} f_T(\mathbf{y}^*|\mathbf{X}\boldsymbol{\beta}, \Sigma_u)\, dy_T^* \ldots dy_2^*\, dy_1^*, \end{aligned} \tag{4}$$

where

$$(a_t, b_t) = \begin{cases} (-\infty, 0) & \text{if} \quad y_t = 0 \\ (0, \infty) & \text{if} \quad y_1 = 1 \end{cases} \tag{5}$$

and $f_T(\mathbf{y}^*|\mathbf{X}\boldsymbol{\beta}, \Sigma_u)$ is the T-dimensional probability density for the latent variable $\mathbf{y}^*$ (Poirier and Ruud 1988, equation 2.8). In the case of probit, this density is the multivariate normal probability density function

$$(2\pi)^{-\frac{T}{2}}|\Sigma_u|^{-\frac{1}{2}} \exp\left[-\frac{\mathbf{u}'\Sigma_u^{-1}\mathbf{u}}{2}\right],$$

with $\mathbf{u} = \mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}$.

The logit model is of limited use given the AR(1) structure of the dependency across observations; the logistic distribution generalizes to a multivariate setting in a fairly cumbersome way. In the time series context, this problem gets more pressing as the length of the series increases, leading to a proliferation of parameters tapping the dependencies across time points which I detail below.[6] Nonetheless, the logit model can still be used for the *marginal* probabilities, even though a multivariate normal distribution is used to cap-

[6]The situation here is directly analogous with the independence of irrelevant alternatives assumption underlying the use of the logit model in a multinomial choice setting. Nesting or grouping alternatives is the way the logit model accommodates interdependence among discrete outcomes; Chamberlain (1980) proposes an analogous logit model for panel data, although the model seems plausible only for short panels and not at all feasible for a single time series.

ture the dependencies among observations (e.g., le Cessie and van Houwelingen 1994).

The likelihood function in (4) poses a ferocious maximization problem, bearing a close resemblance to the intractabilities presented by the multinomial probit (MNP) model for qualitative choice. In MNP, the likelihood function becomes increasingly complex as the number of choices increases; each choice adds another dimension to the integral in the likelihood. Here we have a "multi-period" probit model with the likelihood involving integration of a T-dimensional Normal density. In most time series settings T will be larger than the number of choices in a MNP model, although there are considerably fewer parameters to estimate than in say, a MNP model with four or more outcomes; for instance, in the AR(1) case, $\Sigma_u(\rho)$ is a function of a single parameter, while the corresponding matrix in a MNP setting will usually contain more free parameters than this.

The question of the number of parameters aside, the real difficulty with the binary probit model with AR(1) errors is the T-dimensional integral in the likelihood function. Geweke, Keane and Runkle (1997) report that simulation methods for dealing with the high-dimensional integrals required in multi-period probit models perform poorly as serial dependency increases; the Geweke-Hajivassiliou-Keane (GHK) simulator needs to be run for increasingly longer simulation runs as the magnitude of ρ increases. Indeed, a Markov Chain Monte Carlo (MCMC) approach generally outperforms the GHK simulator for the experimental conditions considered by Geweke, Keane and Runkle.

5.1 Estimation by Bayesian Simulation

The MCMC approach to the time series probit problem has been used by Czado (N.d.). The MCMC approach is attractive in this context because it exploits the fact that the high dimensional integral in the likelihood function in (4) can be well approximated by successively sampling from the series of conditional densities $f(y_t^*|\mathbf{y}_{r<t}^*)$. This sampling algorithm is an example of Gibbs sampling, the workhorse of MCMC. A review of Gibbs sampling need not detain us here; the key idea is that "conditional [densities] determine marginals" (a fact well known to Bayesians), even if the particular details of the relationship in a high dimensional setting can be "obscure" or complicated (Casella and George 1992, 170-1).

In this case we seek the posterior distribution for the unknown parameters and latent data $\pi(\boldsymbol{\beta}, \rho, \mathbf{y}^*|\mathbf{X}, \mathbf{y})$, recalling that $\mathbf{X}$ and $\mathbf{y}$ are the observed data. The MCMC approach breaks this distribution into the conditional distributions $\pi(\boldsymbol{\beta}|\rho, \mathbf{y}^*, \mathbf{X}, \mathbf{y})$, $\pi(\rho|\boldsymbol{\beta}, \mathbf{y}^*, \mathbf{X}, \mathbf{y})$ and $\pi(\mathbf{y}^*|\boldsymbol{\beta}, \rho, \mathbf{X}, \mathbf{y})$. The MCMC algorithm here consists of sampling from each of these distribution, replacing $\boldsymbol{\beta}$, ρ and $\mathbf{y}^*$ when they appear as conditioning arguments with the most recently sampled value for each. At the end a pass m over each of the conditional distributions, the vector of a sampled vectors $(\boldsymbol{\beta}^{(m)}, \rho^{(m)}, \mathbf{y}^{*(m)})'$ consists the state vector of a Markov chain that has the joint posterior as its invariant distribution. When the Markov chain Monte Carlo algorithm has been run for a sufficiently lengthy period, each realization of the state vector is a draw from the joint posterior. These draws from the posterior distribution are saved and summarized for the purposes of statistical inference. Other relevant quantities (e.g., the value of the log-likelihood, percent cases correctly predicted) can also be calculated at each stage of the chain.

In addition, the vector of unknown quantities can be augmented to include missing data; in the application I present below, missing values on the discrete response are treated in this way. This potential for dealing with missing data "on the fly" is an exceptionally useful feature of the MCMC approach.

I turn now to each of the conditional distributions. If the prior distribution for $\boldsymbol{\beta}$ is assumed to $N(\boldsymbol{\beta}_p, \Sigma_p)$, then the conditional distribution for the regression coefficients $\boldsymbol{\beta}$ is multivariate normal with mean

$$(\Sigma_p^{-1} + \mathbf{X}'\Sigma_u^{-1}\mathbf{X})^{-1}(\Sigma^{-1}\boldsymbol{\beta}_p + \mathbf{X}'\Sigma_u^{-1}\mathbf{y}^*)$$

and variance-covariance matrix

$$(\Sigma_p^{-1} + \mathbf{X}'\Sigma_u^{-1}\mathbf{X})^{-1}.$$

The conditional distribution of $\mathbf{y}^*$ is a truncated multivariate normal distribution with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix Σ_u, truncated to $(a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_T, b_T)$, defined in (5). Czado (N.d.) notes that sampling from this truncated multivariate distribution can be accomplished by sequentially sampling from the conditional distributions for each element of y_t^*, where the conditioning is not just on the observed data and the parameters $\boldsymbol{\beta}$ and ρ, but also on the sampled values for $\mathbf{y}^*_{r<t}$. Each conditional distribution is a truncated univariate normal distribution. Given the marginal model for the latent dependent variable

$$y_t^* = \mathbf{x}_t\boldsymbol{\beta} + u_t$$

with the stationary AR(1) error process

$$u_t = \rho u_{t-1} + \epsilon_t,$$

where $|\rho| < 1$ and $\epsilon_t \sim N(0, \sigma_\epsilon^2) \forall\ t$, substitution and re-arranging yields

$$y_t^* | y_{t-1}^* \sim N\left(\mathbf{x}_t \boldsymbol{\beta} + \rho(y_{t-1}^* - \mathbf{x}_{t-1}\boldsymbol{\beta}), \sigma_\epsilon^2\right) I(a_t, b_t) \tag{6}$$

for $t = 2, \ldots, T$, where $\sigma_\epsilon^2 = 1 - \rho^2$ and the function $I(\cdot, \cdot)$ is a binary (0,1) indicator function for the truncation bounds.[7] Note that this sampling algorithm is all conditional on the first observation of the series, which is sampled as

$$y_1^* \sim N(\mathbf{x}_1 \boldsymbol{\beta}, \sigma_u^2) I(a_1, b_1)$$

or use of the Prais-Winsten (1954) transformation might also be used, as is common in GLS fixes for residual autocorrelation in the linear regression setting.

The conditional distribution for ρ is less straightforward. Given the current iteration's estimates of $\mathbf{y}^*$, $\mathbf{X}$ and $\boldsymbol{\beta}$, we obtain $\mathbf{u} = \mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}$. Given our assumption that $\mathbf{u}$ follows an AR(1) process, with $\sigma_u^2 = 1$ and $\epsilon_t \sim N(0, 1 - \rho^2), \forall\ t$, we can obtain a likelihood for ρ, $l(\rho|\mathbf{u})$, equivalent to $f(\mathbf{u}|\rho)$. If all that is known is that ρ is stationary, then the prior for ρ is uniform on the interval [-1, 1], i.e.,

$$\pi(\rho) = \begin{cases} \frac{1}{2} & -1 \leq \rho \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that a posterior distribution is proportional to the likelihood times the prior, or in this case

$$\pi(\rho|\mathbf{u}) \propto f(\mathbf{u}|\rho) \cdot \pi(\rho), \tag{7}$$

and so the posterior will also have zero probability mass outside the stationarity bounds [-1, 1]. Given the normal-based likelihood, this implies that the posterior distribution is proportional to a normal distribution (net of the complications imposed by stationarity), with mean

$$\hat{r} = \frac{\sum_{t=2}^{T} u_{t-1} u_t}{\sum_{t=2} u_{t-1}}$$

and variance $\hat{R} = \sum_{t=2}^{T} u_{t-1} u_t$, but truncated to the [-1,1] interval. In my implementation,

[7] Hajivassiliou (1995, n13) notes it is necessary to restrict the truncation region to a compact region, such that $-\infty < a_t < b_t < +\infty$. This is a technical restriction entailing no loss of generality; in place of $-\infty$ or $+\infty$ we merely substitute an arbitrarily large number (signed appropriately), subject to computational limitations.

a Metropolis method[8] is used to sample from the posterior for ρ. Briefly, the Metropolis method supplies transition probabilities for the Markov chain that "help" it traverse the parameter space and reach its invariant distribution more efficiently. But importantly, the mathematics of the Metropolis method are such that the normalizing constants for the conditional (posterior) distributions are not required, and for this reason Metropolis methods are very attractive in Bayesian analysis, whenever posterior distributions are only defined up to normalizing constants.

In the context considered here, we have a posterior for ρ that is proportional to a normal-based likelihood times a uniform prior. A Metropolis step starts with $\rho^{(q)}$, the current realization of $\{\rho\}$ in the Markov chain. Then,

1. Given $\mathbf{u}^{(q+1)}$, sample ρ^* from $N(\hat{r}^{(q+1)}, \hat{R}^{(q+1)})I(-1, 1)$.

2. With probability

$$\min\left(\frac{f(\mathbf{u}^{(q+1)}|\rho^*)}{f(\mathbf{u}^{(q+1)}|\rho^{(q)})}, 1\right)$$

   accept ρ^* as $\rho^{(q+1)}$, otherwise set $\rho^{(q+1)} = \rho^{(q)}$.

This problem — the absence of a normalizing constant in a posterior distribution — arises frequently in Bayesian analyses of time series, where flat priors over the stationary region for autoregressive parameters give rise to posteriors known only up to a constant factor of proportionality (e.g., Chib and Greenberg 1994; Marriot et al. 1996; West and Harrison 1997). The extension to the discrete time series case introduces no new complications since all the calculations take place with respect to the estimates of the continuous latent quantities (conditional on the observed discrete responses). Previous implementations include work by and has been addressed in the specific context of time series models for discrete data by Geweke, Keane and Runkle (1997) (a multi-period, multinomial probit model) and Czado (N.d.).

6 Marginal Models

Marginal models treat the issue of serial dependence separately from the effect of explanatory variables on the response variable. For instance, consider a binary (0,1) time series $\mathbf{y} = (y_1, \ldots, y_T)$. Let $\theta_t = E(y_t) \equiv \Pr[y_t = 1]$, which in turn depends on covariates

[8] Chib and Greenberg (1995) and Tanner (1996, 176ff) are useful introductions to Metropolis method.

as follows:

$$h(\theta_t) = \mathbf{x}_t \boldsymbol{\beta},$$

where $h()$ is a *link* function known to the analyst and $\mathbf{x}_t$ is the vector of observations on k covariates at time t. Note that although we are modeling serial data, there are no dynamics or conditioning on history in the model for y_t; in this sense the model can be thought of a marginal model, or a model for the marginal expectation of y_t. The marginal variance of y_t is not a critical issue here, and so is assumed to depend on the marginal expectation via a known variance function $v(\theta_t)\upsilon$, where υ is a scale parameter: for instance, when $h()$ is the logit link function, $\text{Var}(Y_t) = \theta_t(1 - \theta_t)$ and υ is set to 1.

Turning to the time series characteristics of the data, consider the correlation between two adjacent observations, y_1 and y_2. By definition,

$$\text{Corr}(y_1, y_2) = \frac{E(y_1, y_2) - E(y_1)E(y_2)}{(\text{Var}(y_1)\text{Var}(y_2))^{1/2}},$$

which in the logit case becomes

$$\text{Corr}(y_1, y_2) = \frac{\Pr(y_1 = 1, y_2 = 1) - \theta_1\theta_2}{(\theta_1(1 - \theta_1)\theta_2(1 - \theta_2))^{1/2}}.$$

However, the marginal quantities θ_1 and θ_2 impose constraints on the correlation across adjacent observations, as the following table helps demonstrate:

	y_2		
y_1	0	1	
0	p_{00}	p_{01}	
1	p_{10}	p_{11}	θ_1
		θ_2	

Our interest is in $\Pr(y_1 = 1, y_2 = 1)$, denoted by p_{11} in the table. Note that this joint probability is bounded by the marginal probabilities θ_1 and θ_2:

$$\max(0, \theta_1 + \theta_2 - 1) < p_{11} < \min(\theta_1, \theta_2) \tag{8}$$

and so the correlation between y_1 and y_2 exhibits an awkward (and potentially implausible) dependency on the marginal probabilities θ_1 and θ_2 (Prentice 1988, 1037). It is more convenient to model the odds ratio of successive observations in the binary time series, effectively collapsing away the marginal terms in equation (8). For the (y_1, y_2) pair considered here, we have

$$\begin{aligned}\psi_{12} &= \frac{\Pr[\text{“no change”}]}{\Pr[\text{“change”}]} \\ &= \frac{\Pr(y_1 = 1, y_2 = 1)\Pr(y_1 = 0, y_2 = 0)}{\Pr(y_1 = 0, y_2 = 1)\Pr(y_1 = 1, y_2 = 0)} \\ &= \frac{p_{11}p_{00}}{p_{01}p_{10}}\end{aligned}$$

The odds ratio takes on values in $(0, \infty)$; if $\psi_{12} = 1$ then "change" is as likely as "no change". Values of ψ_{12} less than 1 indicate that switching or change is more likely than non-switching or stability, while values of ψ_{12} greater than 1 indicate that "no change" is more likely than "change" (Diggle, Liang and Zeger 1994, 150). In this way the odds ratio ψ_{12} taps dependencies between observations 1 and 2 of the time series.

The model parameters are now $\boldsymbol{\beta}$ and ψ, with the log-odds approach ensuring near-orthogonality between the two parameters (Fitzmaurice and Laird 1993). The model is easily extended to handle higher order forms of dependency. For instance, the odds-ratio can be defined in terms of two observations at any arbitrary distance apart in the time series:

$$\psi_{rs} = \frac{\Pr(y_r = 1, y_s = 1)\Pr(y_r = 0, y_s = 0)}{\Pr(y_r = 0, y_s = 1)\Pr(y_r = 1, y_s = 0)} \tag{9}$$

The full distribution of $\mathbf{y}$ can be modeled in terms of the covariates (via $\boldsymbol{\beta}$ and the chosen link function) and the entire $\binom{T}{2}$ set of odds ratios between adjacent observations. Of course, this "saturated" model gives rise to a large number of parameters that can greatly complicate the analysis of lengthy time series. In general, the number of association parameters to estimate increases exponentially with the length of the time series, and net of some simplifying assumptions, the fully unconstrained model is not feasible. Further complications arise should one be dealing with panel or pooled cross-sectional time series data that is unbalanced (the number of observations varies across subjects or units).

Diggle, Liang and Zeger (1994, 151) suggest a number of simplifications for dealing with the proliferation of associational parameters. In a repeated measures or panel setting, one approach might be to set $\psi_{irs} = \psi_i$ (i.e., the degree of longitudinal association is constant for all pairs of observations for unit i). Another approach might be to parameterize ψ_{rs} as a function of temporal distance: e.g.,

$$\ln \psi_{rs} = \alpha_0 + \alpha_1 |r - s|^{-1}$$

such that the degree of longitudinal association is inversely proportional to the time between observations. Finally, Carey, Zeger and Diggle (1993) show how to include covariates in an auxiliary model for the ψ_{rs} terms; they estimate this model using "alternating logistic regressions", where the term "alternating" reflects the fact that the estimation iterates between estimates of $\boldsymbol{\beta}$ in the marginal model for $\mathbf{y}$ and the $\boldsymbol{\alpha}$ coefficients in the auxiliary model for the (log) odds ratios.

Generally, these models are estimates by quasi-likelihood methods, called generalized estimating equations, or GEE (Fitzmaurice, Laird and Rotnitzky 1993). These are "quasi-likelihood" models in the sense that in the discussion thus far, there has been no mention of the distribution of error terms that would typically give rise to a likelihood function for the data. The marginal model is for the the first moment of the y_t, the choice of link function usually implies a simple function for the second moment, and the odds-ratios (with or without regressors) imply some structure on the covariances of the y_t. For Normal-based likelihoods, or for the logit-link binary case, these moments are sufficient to characterize the entire likelihood, and the GEE approach coincides with conventional likelihood based methods. However, the likelihoods for these marginal models are somewhat complicated to directly maximize, and the iterative approach of GEE (successively approximating solutions to the moment equations) has computational advantages, in addition to being free of distributional assumptions.

6.1 A Marginal-Transitional Hybrid

Azzalini (1994) provides an interesting simplification of the marginal model, using a Markov chain to characterize the dynamics in the time series. In this sense Azzalini's model is a hybrid of the marginal and transitional models encountered above. Changes in the binary series $\mathbf{y}$ can be modeled as a Markov chain over the two states (0 and 1), with a transition matrix

$$\begin{pmatrix} 1-p_0 & p_0 \\ 1-p_1 & p_1 \end{pmatrix},$$

where $p_0 = \Pr(y_t = 1 | y_{t-1} = 0)$ and $p_1 = \Pr(y_t = 1 | y_{t-1} = 1)$. These transition probabilities need not be constant over time (and in general will not be constant), and could themselves be modeled with covariates as in the marginal model suggested above. As in the marginal model, it is more convenient to model the odds ratio of the probability of a transition, ψ.

The Markov structure implies

$$\begin{aligned}\Pr(y_t = 1) &= \Pr(y_{t-1} = 1)\Pr(y_t = 1|y_{t-1} = 1) + \Pr(y_{t-1} = 0)\Pr(y_t = 1|y_{t-1} = 0) \\ \theta_t &= \theta_{t-1}p_1 + (1 - \theta_{t-1})p_0.\end{aligned}$$

Thus the transitional parameters p_0 and p_1 appear in the model relating the covariates to the responses (recall that $h(\theta) = \mathbf{x}_t\boldsymbol{\beta}$). Combined with the odds ratio representation of the transitional dynamics

$$\psi = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)} \tag{10}$$

we now have a complete characterization of the process for $\mathbf{y}$, subject to the assumption of a first order Markov process. Because the transitional parameters p_0 and p_1 appear in the expression for θ_t, they will vary over time themselves: MacDonald and Zucchini (1997, 11) use the notation ${}_tp_j$ to reflect this time variation, where $j = 0, 1$.

The odds ratio representation and the Markov representation in equation (10) can be combined and re-arranged to yield the following expression

$${}_tp_j = \left\{ \begin{array}{ll} \theta_t & \text{if } \psi = 1, \\ \frac{\delta-1+(\psi-1)(\theta_t-\theta_{t-1})}{2(\psi-1)(1-\theta_{t-1})} + j\frac{1-\delta+(\psi-1)(\theta_t+\theta_{t-1}-2\theta_t\theta_{t-1})}{2(\psi-1)\theta_{t-1}(1-\theta_{t-1})} & \text{if } \psi \neq 1, \end{array} \right\} \tag{11}$$

for $t = 2, \ldots, T$, and where

$$\delta^2 = 1 + (\psi - 1)\left((\theta_t - \theta_{t-1})^2\psi - (\theta_t + \theta_{t-1})^2 + 2(\theta_t + \theta_{t-1})\right).$$

We can condition on the first observation, setting $\Pr(y_1 = 1) = \theta_1$ to complete the specification of the model. Since $\psi \in (0, \infty)$, it is more convenient to work with $\lambda = \ln(\psi)$, and so the log-likelihood function for this process is

$$\ln L(\beta, \lambda) = \sum_{t=1}^{T} \left[y_t \,\text{logit}({}_tp_{y_{t-1}}) + \ln(1 - {}_tp_{y_{t-1}})\right], \tag{12}$$

This log-likelihood is relatively straightforward to maximize; first and second derivatives of the log-likelihood with respect to to the parameters are provided in Azzalini, or one can also use software specifically designed for this model using `Splus` (Azzalini and Chiogna 1995).

7 An example: dyadic conflict data

To see how these models work in practice, I turn to a large data set gathered for studying the determinants of international conflict. The data consist of 20,990 observations at the level of the dyad-year, where a dyad refers to a pairing of countries, with the binary indicator y_{it} coded 1 if dyad i engaged in a militarized interstate dispute (a MID) in year t, and 0 otherwise.

These data were gathered and analyzed by scholars working on substantive problems international relations (e.g., Russett 1990; Russett 1993; Oneal and Russett 1997). Key issues in this large literature concern the effects of the following covariates:[9]

- **democracy** (the "liberal peace" hypothesis), with the democracy score for each dyad set to the lesser democracy score of the dyad partners, with the scores scaled to run from -1 to 1, using indicators in the Polity III data base (Gurr and Jaggers 1996).

- intra-dyad **trade**, measured as the ratio (percent) of intra-dyad trade to GDP, for each partner, but with the dyad's score set to the smaller of these ratios. This variable is lagged one year, so as not to proxy a current dispute (Oneal and Russett 1997; Beck, Katz and Tucker 1998).

- dyad partners who are **allies** are presumed to be less likely to engage in MIDs. This variable takes the value 1 if the dyad partners were allied, or if both were allied with the United States.

- military **capabilities** are measured as the ratio of the stronger partners' score to the weaker partner's score, on an index based on indicators in the Correlates of War data collection (Singer and Small 1994).

- geographic **contiguity** within a dyad (1 if the partners are geographically contiguous, 0 otherwise), since *ceteris paribus* it is more difficult to engage in a MID with a dyad partner who is distant, or if another country's borders have to be crossed in order to instigate a MID.

- **economic growth** measures the lesser of the rates of economic growth of the partners.

These data have been reanalyzed by a number of scholars attracted to the methodological problems presented by these data. For instance, Beck and Jackman (1998) focus

[9]My source for these data is Beck, Katz and Tucker (1998), who obtained the data from Oneal and Russett (1997). Further details on these data and definitions appear in those articles.

on possible non-linearities and interactions in the latent regression function, ignoring any serial component to the data; Beck, King and Zeng's (1998) reanalysis is in a similar vein, but using neural networks. Beck, Katz and Tucker's (1998) reanalysis of these data focuses the serial properties of these data, and draws attention to some of the shortcomings of previous analyses that omit to control for serial dependence in the data (I mentioned some of the these in section 1). Beck, Katz and Tucker account for serial dependence in these data by augmenting a marginal logit model with non-parametric term (a cubic smoothing spline) for time since last conflict.

Several features of the data should be pointed out. First, MIDs are extremely rare in these data. Of the 20,990 dyad-years in the data I analyze, just 947 or 4.5% are coded 1 for the presence of a MID. The data are also highly unbalanced in their design, in the sense that there is much variation in the length of time each dyad is present in the data (see Figure 2). The maximum length of any dyad-series is 35 yearly observations (1951-1985), but with only 138 (16.7%) of the 827 dyads contributing this much data. This feature of the data creates a formidable barrier for marginal models employing log-odds parameterizations for temporal association; many software implementations of this model require the data to be balanced, and so I do not fit marginal models of this type to these dyadic data.

Coupled with the rarity of MIDs, Figure 3 shows that the overwhelming majority of dyads record no MIDs whatsoever; approximately 75% of the dyad-specific time series on the dependent variable consist of a vector of zeros and only 20 dyads (2.4%) report more years with MIDs than without MIDs. Even this cursory examination of the data suggests tremendous "stability" in the data, and ignoring this serial dependence is likely to lead to faulty inferences.

Finally, 293 (35.4%) of dyads contain temporal breaks. This poses problems for several of the models considered earlier, especially the transitional models, and to a lesser extent the marginal models (by heightening the unbalanced design problem). One solution is to treat temporally-contiguous sequence of observations as a "unit", with possibly multiple "units" per dyad entering the analysis; I adopt this approach when fitting the hybrid marginal-transitional model of Azzalini (1994) I described in section 6.1.

A more attractive solution is to treat the temporal breaks within each dyad as missing data, and make imputations for the data, treating them as missing at random (conditional on the covariates). This is relatively easy to accomplish with the MCMC implementation of the parameter-driven transitional model considered in section 5. I linearly interpolate the missing data on the covariates for the missing observations,[10] but leave the dependent

[10]Most of these temporal gaps are short (1 or 2 yearly observations) and the covariates exhibit modest

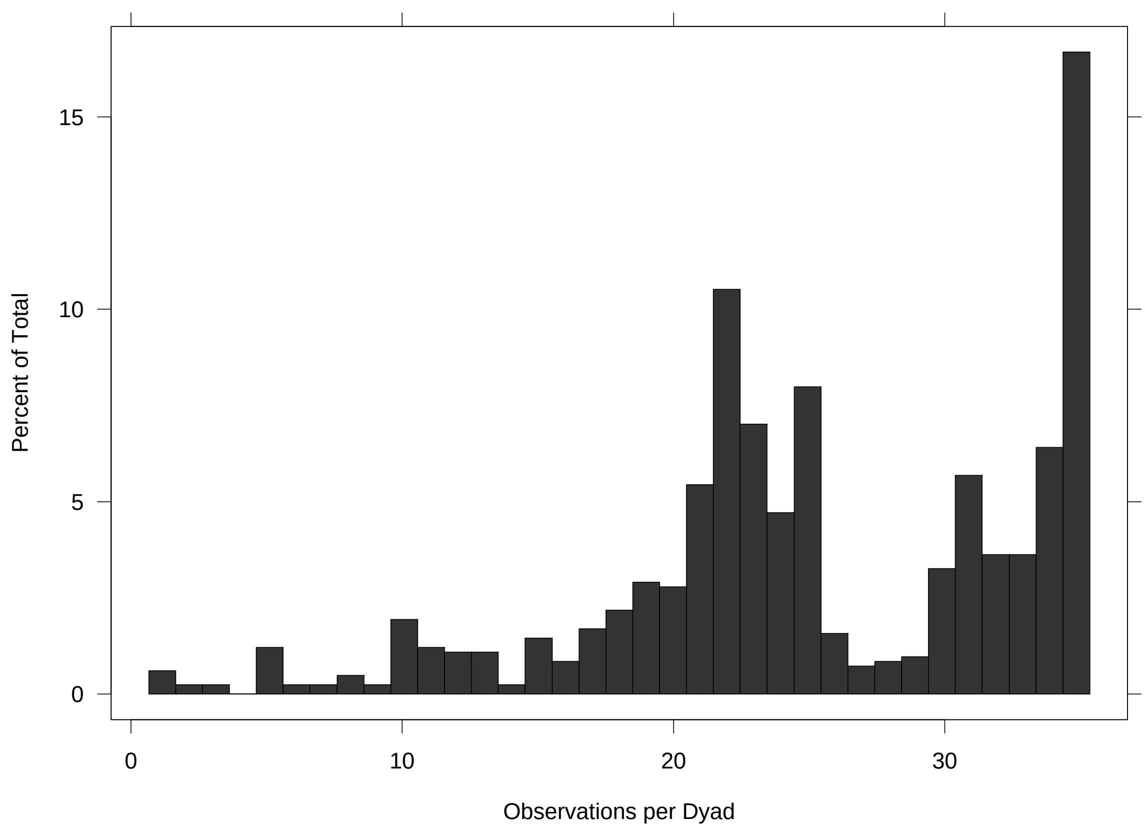

Figure 2: ***Variation in length of time series, data on 827 dyads.***

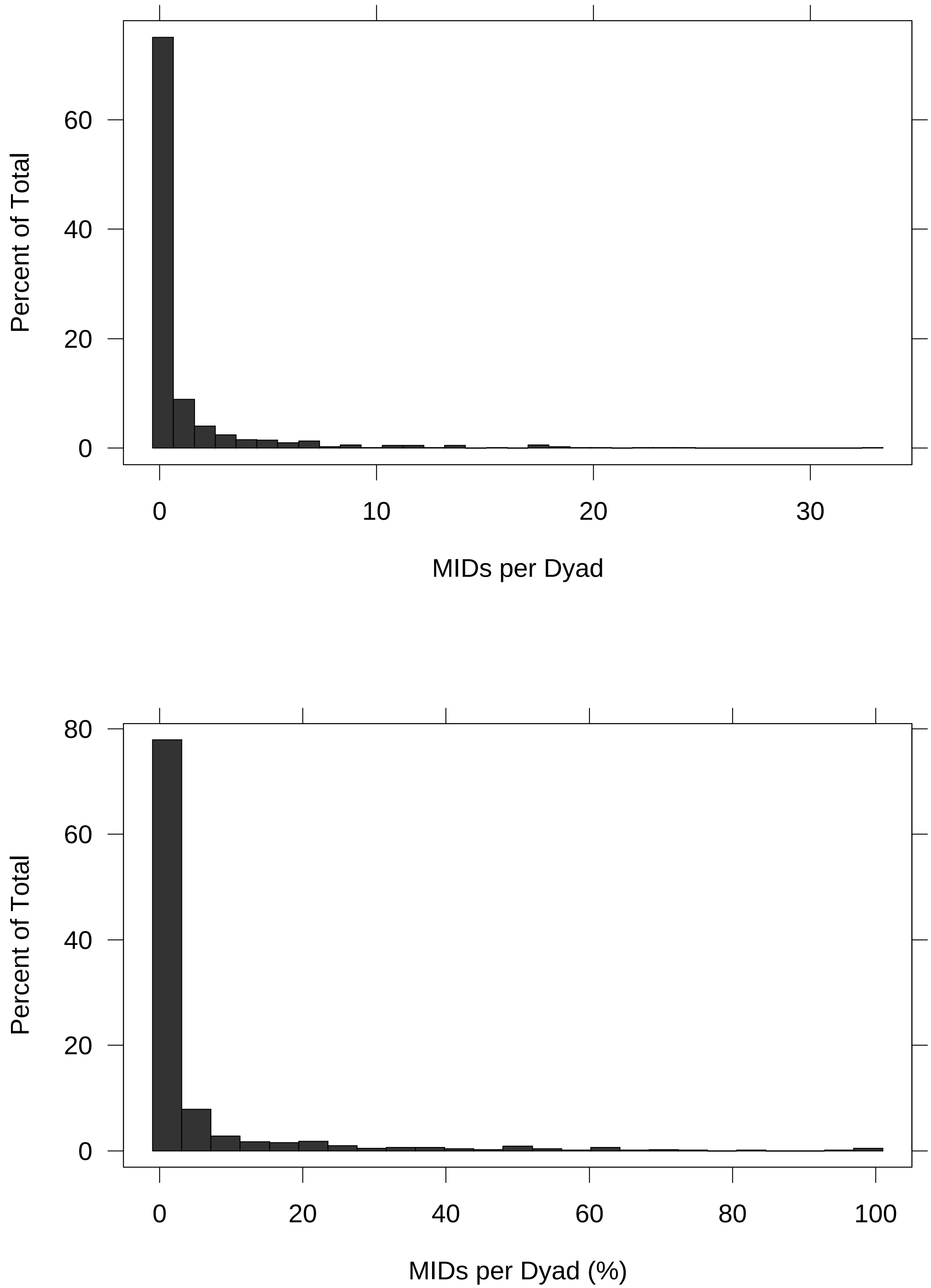

Figure 3: ***The Distribution of MIDs across dyads***. *The modal number of MIDs per dyads is zero, with roughly 75% of the dyad-level time series consisting of a vectors of zeros.*

variable as missing data. The MCMC procedure is augmented to imputations for these missing, discrete y_{it}. When the data are "padded out" so as to remove any temporal discontinuities within dyads, the total number of observations increases from 20,990 to 21,844. This means that there are 854 values of y_{it} to impute, or some 3.9% of the data.

In these senses, these data are a "hard case" with which to assess the utility of the models presented above; these are not at all like the relatively small and well-mannered "canonical" data sets encountered in statistical literature developing these models.

7.1 The Marginal-Transitional Hybrid Model

Results of fitting the marginal-transitional hybrid model of Azzalini (1994) appear in the second column of Table 1, along with the ordinary logit results for comparison.[11] The ordinary logit results suggest that increasing levels of democracy lessens the probability of a MID, as does trade, two variables of particular substantive importance in the international relations literature. The other control variables pick up coefficients large relative to their standard errors.

Controlling for serial dependence via Azzalini's (1994) model alters the results of the unconditional logit analysis. The estimate of the association parameter $\lambda = \ln(\psi)$ is 4.05, which means that the odds of "stability" to "change" in adjacent pairs of observations (within dyads) are over 57:1, an unsurprising result given the crude descriptive statistics reported above. Yet even in the face of this high degree of serial dependence, the coefficient on democracy is essentially unchanged. However, the coefficient on trade shrinks by over a half, and the t ratio shrinks from -4.9 to about -1.9. The coefficients on economic growth, alliances, contiguity and military capability are also all smaller, while their standard errors grow; with the possible exception of economic growth, none of these coefficients fail to attain statistical significance at conventional levels.

Beck, Katz and Tucker (1998) found similar results when they controlled for serial dependence in the data with a variable counting time since last MID. As they point out, the trade coefficient undergoes significant change once serial dependence is introduced to the model because trade tends to drops dramatically once a dyad enters into conflict.[12] Also, both states — conflict and peace — are "sticky", so much so that the best predictor of conflict or peace is what state a dyad was in at $t-1$. These two features of the data

over-time change, so the linear interpolation strategy seems innocuous.

[11]These results were generated using the maximum likelihood procedures in GAUSS Version 3.2.35, with standard errors based on numerical approximations for the Hessians of the respective log-likelihood functions.

[12]The unconditional mean level of the trade variable is .002271, while conditional on two or more consecutive years with MIDs, the average level of trade is .00077, or about a third as small.

	Ordinary Logit	*Marginal-Transitional*
Intercept	−3.3	−3.40
	(.079)	(.12)
Democracy	−.49	−.51
	(.074)	(.11)
Economic Growth	−2.23	−1.91
	(.85)	(1.01)
Alliance	−.82	−.69
	(.080)	(.13)
Contiguity	1.31	1.24
	(.080)	(.13)
Military Capability	−.31	−.22
	(.041)	(.045)
Trade	−66.1	−28.2
	(13.4)	(14.7)
λ		4.05
		(.092)

Table 1: ***Estimates of Ordinary Logit Model, and Marginal-Transitional Hybrid Model.*** *Standard errors appears in parentheses. The parameter* $\lambda = \ln(\psi)$*, where* ψ *is the odds-ratio defined in equation (10). n = 20,990, over 827 dyads of unequal lengths.*

— underlying inertia, and drops in trade once a dyad enters into conflict — serve to undermine the effects of trade once some attempt is made to condition on the history of the dyad, a finding will we see repeated below.

7.2 The Observation-Driven Transitional Model

The Markov-based transitional model reviewed in section 4 permits the exploration of conjectures like those above. In particular, if the effects of covariates are thought to be state-dependent, then the Markov model is a useful way to proceed. Given my assumption of a first-order Markov chain, the effects of the covariates are presumed to be conditional on whether a dyad experienced a MID in the preceding year or not, and the effects of earlier states only effect the present via their indirect effects through $t-1$. As remarked earlier, this model can be estimated simply with software for ordinary binary response models. Estimates of the first-order Markov model appear in Table 2, again, along with the ordinary logit estimates for comparison.

	Ordinary Logit	*Observation-Driven Transitional Model*		
	$\boldsymbol{\beta}$	$\boldsymbol{\beta}_0$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}_1$
Intercept	−3.3	−4.5	5.0	.55
	(.079)	(.12)	(.21)	(.17)
Democracy	−.49	−.33	.075	−.25
	(.074)	(.11)	(.20)	(.17)
Economic Growth	−2.2	−3.5	.13	−3.4
	(.85)	(1.40)	(2.29)	(1.8)
Alliance	−.82	−.39	−.30	−.70
	(.080)	(.12)	(.21)	(.17)
Contiguity	1.31	1.48	−1.41	.070
	(.080)	(.13)	(.21)	(.17)
Military Capability	−.31	−.15	.12	−.028
	(.041)	(.045)	(.08)	(.071)
Trade	−66.1	−31.2	−46.0	−77.2
	(13.4)	(13.5)	(35.45)	(32.8)
n	20,990		19,776	

Table 2: ***Estimates of an Observation-Driven Transitional Model.*** *The estimates for α contrast the effects of the covariates across states; i.e, Pr$[y_t = 1|y_{t-1} = j] = \mathbf{x}\boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\alpha}$. Standard errors appear in parentheses, and the ordinary (unconditional) logit estimates are presented for comparison. The different numbers of observations arise because the transitional-model loses the first observation of each dyad.*

These results are provocative, illustrating that even this modest conditioning on each dyad's history leads to some different interpretations to those we might take away from an ordinary, unconditional analysis. First, note the changes between the ordinary (unconditional) logit analysis and the estimates for β_0 (governing $\Pr[y_t = 1|y_{t-1} = 0]$). The effects of democracy in stopping a transition to a dispute are smaller than the unconditional analysis would suggest, with the coefficient in the transitional analysis about two-thirds of the magnitude of the coefficient obtained from the unconditional analysis (-.33 versus -.49). Other coefficients that appear to change substantially are those for the alliance indicator, military capabilities, and trade; an ordinary logit analysis would appear to overstate the contribution of these variables in preventing MIDs. Note also that the intercept in β_0 for the conditional model is also somewhat larger (more negative) than that for the unconditional analysis, reflecting that "peace" is highly persistent in these data.

The other difference with an ordinary logit analysis arises via the conditioning on the dyad's past, reflected in the estimates of α, and the implied estimates of β_1. The intercept and the contiguity coefficient are distinguishable from zero at conventional levels of statistical significance. The positive intercept in α means that the intercept in β_1 is actually close to zero (about .55), highlighting persistence in MIDs: i.e., putting the predictors to one side, a good predictor of a dyad recording a MID in a given year is whether it recorded a MID in the previous period.

Note also the effects of alliances in the two states. Conditional on a dispute existing between the two countries, alliances help return the dyad to peace, and this contribution of alliances would seem greater than the role of alliances in preventing MIDs arising in the first instance. The coefficient tapping the difference across states is the relevant entry in α, and has a t-statistic of -1.44 ($p \approx$ .075), suggesting that this difference is significant. The magnitude of the difference is also impressive, with the coefficient for alliances conditional on a MID in the previous period being about 170% of the magnitude of the coefficient conditional on peace (-.70 versus -.39). The conventional analysis reports a large effect (-.82), but the transitional model shows that the effect of alliances varies depending on the history of a particular dyad.

A similar story emerges for military capabilities, where a significant negative coefficient conditional on peace (-.15) switches to a zero effect conditional on a dispute (-.028, with a standard error of .071). That is, the balance of military power within a dyad is a determinant of whether the dyad records a MID, but conditional on a dispute existing between the partners, relative military capabilities are not useful predictors of whether the MID continues. Likewise for the effects of contiguity. The significant negative coefficient for the contiguity indicator in α effectively nullifies the significant positive coefficient for contiguity in β_0. This implies that the effects of contiguity are asymmetric, and essentially limited to the decision to instigate a MID: contiguity increases the probability that a dyad will record a MID, but once the dyad is engaged in a militarized dispute, contiguity has no bearing on whether the MID continues.

The trade coefficient is also worth consideration. An ordinary logit analysis assuming the data to be serially independent would appear to overestimate the effects of trade in preventing MIDs by over a factor of 2. But this is only part of the story. The transitional model points to differences in the effect of trade conditional on the state of the dyad, with trade having a much larger role in preventing MIDs from continuing than in preventing a MID from arising in the first instance. Despite the large differences in the magnitudes of the coefficients on trade in the two states (-31.2 and -77.2), the difference between the coefficients struggles to attain statistically significance at conventional levels (t = -1.29, $p \approx .10$, one-tailed). Nonetheless, the difference between the two estimates is impressive, and potentially y more informative than the estimate obtained from a conventional analysis. Note also that this finding mirrors the changes I found in switching from ordinary logit to the marginal-transitional hybrid model in the previous section, and the changes noted by Beck, Katz and Tucker (1998). Here we gain some further insight into this phenomenon, by explicitly conditioning on the history of the process under study. Beck, Katz and Tucker (1998) speculated as to the effects of trade in diminishing the duration of conflicts, while the transitional model I employ allows the conditional or “transitional” effects of trade to be estimated directly.

Finally, consider the effects of democracy and economic growth, two important variables in the liberal peace conjecture. I have already noted that an unconditional model would appear to overestimate the effects of democracy and underestimate the effects of economic growth in maintaining peace. The other conclusion to be drawn from the transitional model is that increasing levels of democracy and economic growth help prevent militarized disputes from starting, and have roughly similar effects in stopping MIDs from continuing.

7.3 Parameter-Driven Transitional Model

I estimated the parameter-driven transitional model discussed in section 5 with MCMC methods. The analysis was run on a subset of the data; only those dyads with data starting in 1951 (the earliest possible start date) and ending in 1985 (the latest possible end date) entered the analysis. Nonetheless, a significant proportion of these dyads have temporal breaks within the 1951-1985 intervals. I “padded out” these temporal discontinuities within dyads (and linearly interpolating on the covariates,as discussed in section 5) such that the full set of 8,540 "observations" analyzed here contain 261 (3%) observations with missing data on the dependent variable. These 8,540 data points thus constitute a balanced design (with missing data) on 244 dyads, over 35 years, versus the full set of 827 dyads and 21,844 observations in the full data set (20,990, plus an extra 854 missing data points). MIDs are slightly more common in these subset of the data than in the full data set: 419 dyad-years (4.9%) of the observations considered here experience a MID, while this percentage is 4.3% in the full data set.[13]

Diffuse normal priors were used for the seven regression parameters, and a uniform prior on [-1,1] for the autoregressive parameter on the latent error process. Multiple runs of the the MCMC algorithm were started with perturbations of the ordinary logit parameter estimates, and with ρ set to a variety of different values over the stationary interval. In each instance the algorithm converges back to the same region of the parameter space, although sometimes quite slowly. The slow performance of the MCMC algorithm is unsurprising, given that the vector $\mathbf{y}^*$ (of length 8,540) is sampled element-by-element, and the high value of ρ (suggesting a high degree of dependence across adjacent y_t^* within dyads). Convergence diagnostics suggested that lengthy runs of the MCMC algorithm are required. The results reported here are based on runs of 5,000 iterations, after a burn-in period of 5,000 iterations, and a trace plot of a typical run is presented in Figure 4. A binning interval of 25 samples is used for calculating the standard deviations of the marginal posterior distributions of the parameters.

There are few differences between the ordinary logit results and those for the transitional model. The intercept for the logit with latent AR(1) errors is substantially greater in magnitude than that for the corresponding ordinary logit model, and the coefficient for democracy is also larger (by about 20%), although this difference is small relative to standard errors of each estimated coefficient. The biggest differences across the models are to do with the trade variable. Augmenting the logit model with an AR(1) process on the disturbances causes the effects of trade to be washed out ($t = 1.26$); recall that trade is one of the few covariates to exhibit a reasonable amount of time variation (and in a way related to the dependent variable), and so might be reasonably expected to be affected by the introduction of a correction for serial dependence in the disturbances.[14] The latent AR(1) model captures the fact that each yearly observations can not be treated *de novo*. But since most of the covariates capture little of the serial structure in the data, the AR1(1)

[13]My excuse for working with this smaller data set is part sloth and part computational. Each MCMC run with even these n=8,540 subset took well over 24 hours.

[14]Greene (1997, 587ff) details the relationship between ρ, serial persistence in the covariates and the inefficiency of OLS in the linear regression setting.

	Ordinary Logit	*Latent Transitional*
Intercept	−3.0	−4.6
	(.10)	(.15)
Democracy	−.79	−1.01
	(.10)	(.15)
Economic Growth	−3.81	−4.25
	(1.70)	(1.60)
Alliance	−1.22	−1.04
	(.12)	(.15)
Contiguity	1.14	1.16
	(.11)	(.19)
Military Capability	.030	−.014
	(.026)	(.063)
Trade	−39.4	−9.3
	(16.6)	(7.4)
ρ	0	.91
	(-)	(.005)
n	8,279	8,540
log-likelihood/n	−.1772	−.1122

Table 3: ***Estimates of Parameter-Driven, Latent, Transitional Model.*** *The ordinary logit estimates (MLEs) are presented for comparison. Estimates of the latent, transitional model are summaries of the marginal posterior distributions for each distribution, based on diffuse priors and the MCMC methods described in section 5.1.*

error process does, and fits the data much better: the estimate of ρ is around .91, and the (normalized) log-likelihood[15] is much higher than that obtained for the ordinary logit analysis.

8 Conclusion

I have surveyed a number of ways of dealing with discrete, serial data. The marginal-transitional hybrid model of Azzalini (1994) and the observation-driven transitional model are easily implemented. The dyadic conflict example shows these models to produce interesting and plausible departures from the ordinary logit results. The latent, parameter-driven, transitional model also produces a number of interesting findings, although is considerably harder to implement, requiring the evaluation of a T-dimensional integral (which I deal with via MCMC methods). In many ways the latent AR() model is more "natural" model for political scientists to work with, combining a familiar discrete choice model with an auto-regressive error process, a well-known and widely understood model for dealing with serial dependence.

On the other hand, the former two models make us confront the fact that these are discrete data, and model the dynamics using Markov chains over the discrete binary responses. In so doing, we probably learn more about the determinants of international conflict from the observation-driven transitional model, and it certainly dominates the other models considered here in terms of substantive mileage per CPU hour, requiring nothing more than some interaction terms to be added to an ordinary binary response model.

In a recent paper analyzing these data, Beck, King and Zeng (1998, 2) claim that a major problem with previous analyses is that "[m]ost scholars use statistical procedures that assume the effects of the causes of war are nearly the same for all dyads". This position is a reasonable characterization, amounting to a criticism of simple marginal models (Beck, King and Zeng go on to advocate and use neural nets; i.e., complicated marginal models). But one extremely easy way to escape this shortcoming is to *not* treat each dyad-year *de novo*, by conditioning estimates of the causes of war on the history of the dyad.[16] Even the most cursory look at the data reveals that the current state of the

[15]For the latent transitional model — estimated by MCMC methods — I take the median of the sampled log-likelihoods produced over the 5,000 iterations of the MCMC algorithm I use for summarizing the parameters.

[16]Beck, King and Zeng go some way towards accomplishing conditioning on the history of the dyad by having time since last dispute enter as a predictor in a neural-net; constrast Beck, Katz and Tucker (1998),who introduce time since last dispute as a predictor in marginal model.

dyad is an important predictor of the future trajectory of a dyad; this is information that would be readily used in an applied setting, and ought to belong in any reasonable model. Moreover, conditioning on the past is an important guiding principle in the analysis of time series data for continuous variables; the models presented here point to ways we can do this in the case of discrete time series data.

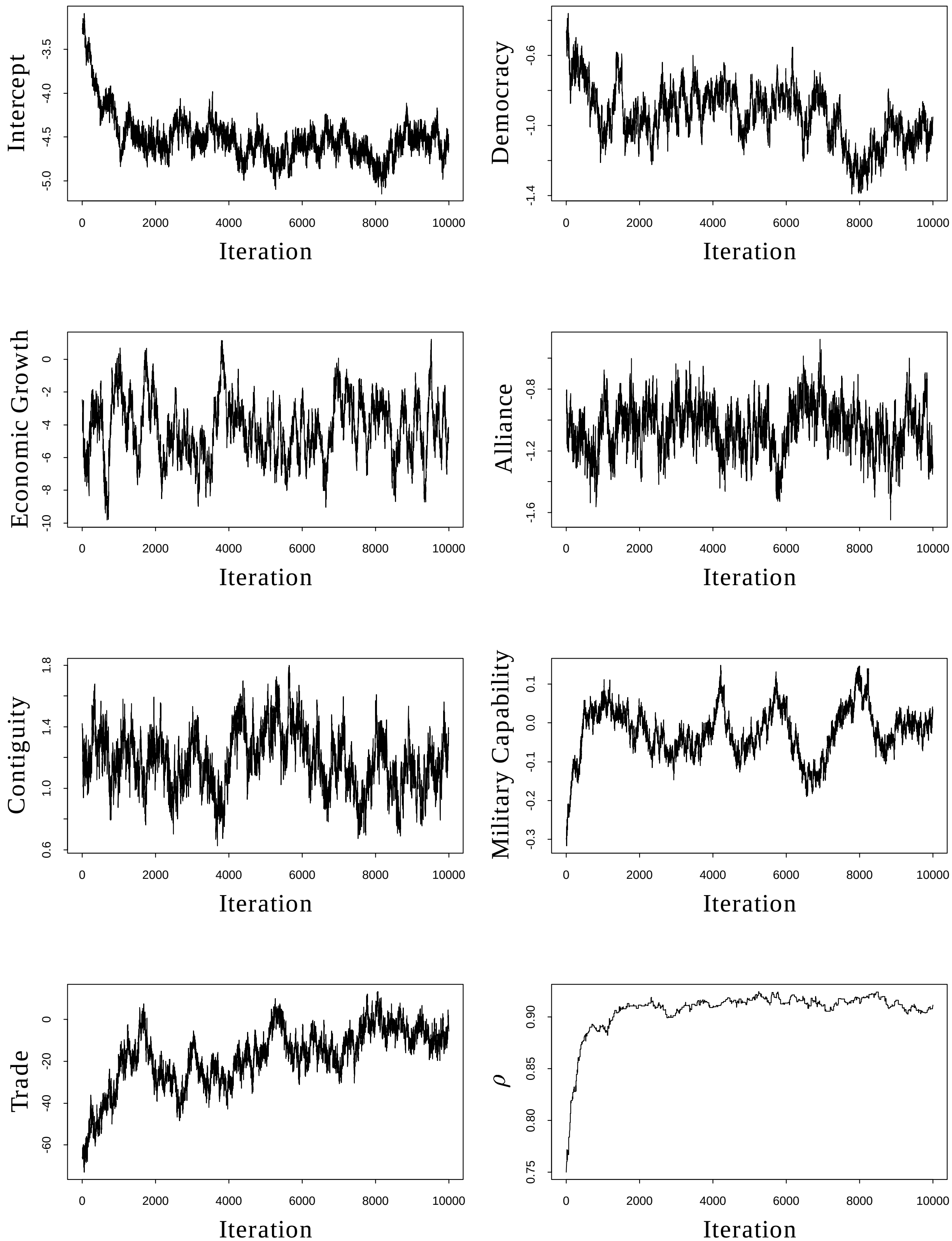

Figure 4: ***Trace Plot of MCMC Algorithm, Parameter-Driven Transitional Model.*** *The last 5,000 simulations are retained for analysis in the results reported in Table 3.*

References

Aldrich, John H. and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Number 07-045 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage.

Alvarez, R. Michael and Garrett Glasgow. 1997. "Do Voters Learn from Presidential Election Campaigns?" `http://wizard.ucr.edu/polmeth/working_papers97/alvar97e.html`.

Andrews, D. F. and A. M. Herzberg. 1985. *Data. A Collection of Problems from Many Fields for the Student and Research Worker*. Berlin: Springer-Verlag.

Ashby, M., J. M. Neuhaus, W. M. Hauck, P. Bacchetti, D. C. Heilborn, N. P. Jewell, M. R. Segal and R. E. Fusaro. 1992. "An Annotated Bibliography of Methods for Analysing Correlated Categorical Data." *Statistics in Medicine* 11:67–99.

Azzalini, A. 1994. "Logistic regression for autocorrelated data with application to repeated measures." *Biometrika* 81:767–75.

Azzalini, A. and M. Chiogna. 1995. `rm.tools`: some `S-plus` tools for the exploratory and parametric analysis of repeated measures data. Technical Report. Dipartimento di Scienze Statistiche, Università di Padova, Italy. `http://lib.stat.cmu.edu/S/rm.tools2`.

Beck, Nathaniel, Gary King and Langche Zeng. 1998. "The Problem with Quantitative Studies of International Conflict." `http://wizard.ucr.edu/polmeth/working_papers98/beck98.html`.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously in Binary Time-Series Cross-Section Models." *American Journal of Political Science*. forthcoming.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42:596–627.

Bombardier, C., J. H. Ware and I. J. Russell. 1986. "Auranofin therapy and quality of in-patients with rheumatoid arthritis." *American Journal of Medicine* 81:565–578.

Carey, Vince C., Scott L. Zeger and Peter J. Diggle. 1993. "Modelling multivariate binary data with alternating logistic regressions." *Biometrika* 80:517–26.

Carlin, B. P., A. E. Gelfand and A. F. M. Smith. 1992. "Hierarchical Bayesian analysis of change point problems." *Applied Statistics* 41:389–405.

Casella, George and Edward I. George. 1992. "Explaining the Gibbs Sampler." *The American Statistician* 46:167–74.

Chamberlain, G. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–238.

Chib, Siddhartha and Edward Greenberg. 1994. "Bayes inference in regression models with ARMA (p, q) errors." *Journal of Econometrics* 64:188–206.

Chib, Siddhartha and Edward Greenberg. 1995. "Explaining the Metropolis-Hastings algorithm." *The American Statistician* 49:327–335.

Chib, Siddhartha, Edward Greenberg and Rainer Winkelmann. 1998. "Posterior simulation and Bayes factors in panel count data models." *Journal of Econometrics* 86:33–54.

Cox, D. 1970. *The Analysis of Binary Data*. London: Chapman and Hall.

Cox, D. R. and E.J. Snell. 1989. *Analysis of Binary Data*. Second ed. London: Chapman and Hall.

Cox, David R. 1981. "Statistical issues in the analysis of time series: some recent developments." *Scandinavian Journal of Statistics* 8:93–115.

Czado, Claudia. N.d. "Multivariate Probit Analysis of Binary Time Series Data with Missing Responses." Department of Mathematics and Statistics, York University. `http://math.yorku.ca/Who/Faculty/Czado/bin.ps`.

Davis, Richard A., W. T. M. Dunsmuir and Y. Wang. 1998. "Modelling time series of count data." In *Asymptotics, Nonparametrics and Time Series*, ed. Subir Ghosh. Marcel Dekker.

Diggle, Peter, Kung-Yee Liang and Scott L. Zeger. 1994. *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Farber, Henry S. and Joanne Gowa. 1997. "Common Interests or Common Polities: Reinterpreting the Democratic Peace." *Journal of Politics* 59(2):393–417.

Fitzmaurice, Garrett M. and Nan M. Laird. 1993. "A likelihood-based method for analysing longitudinal binary responses." *Biometrika* 80:141–51.

Fitzmaurice, Garrett M., Nan M. Laird and Andrea G. Rotnitzky. 1993. “Regression Models for Discrete Longitudinal Responses.” *Statistical Sciences* 8:284–309.

Geweke, John F., Michael P. Keane and David E. Runkle. 1997. “Statistical inference in the multinomial multiperiod probit model.” *Journal of Econometrics* 80:125–165.

Gochman, Charles S. and Zeev Maoz. 1984. “Militarized Interstate Disputes 1816–1976.” *Journal of Conflict Resolution* 28:585–615.

Gowa, Joanne. 1998. “Politics at the Water’s Edge: Parties, Voters, and the Use of Force Abroad.” *International Organization* 52. forthcoming (page numbers refer to typescript version).

Greene, William H. 1997. *Econometric Analysis*. Third ed. New York: Prentice-Hall.

Gurr, Ted Robert and Keith Jaggers. 1996. “Polity III: Regime Change and Political Authority, 1800-1994.” Computer file, Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Hajivassiliou, V. A. 1995. “A Simulation Estimation Analysis of the External Debt Countries of Developing Countries.” In *Econometric Inference Using Simulation Techniques*, ed. Herman K. Van Dijk, Alain Monfort and Bryan W. Brown. New York: Wiley pp. 191–213.

Hausman, J. A., B. H. Hall and Z. Grilliches. 1984. “Econometric models for count data with an application to the patents-R&D relationship.” *Econometrica* 52:909–938.

Jennings, M. Kent and Richard Niemi. 1981. *Generations and Politics*. Princeton, NJ: Princeton University Press.

Jones, Bradford S. and Marco R. Steenbergen. 1997. Modeling Multilevel Data Structures. Technical Report. Political Methodology Working Papers. `http://wizard.ucr.edu/polmeth/working_papers97/jones97.html`.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Klotz, J. 1973. “Statistical inference in Bernoulli trials with dependence.” *Annals of Statistics* 1:373–379.

Korn, E. L. and A. S. Whittemore. 1979. “Methods for analyzing panel studies of acute health effects of air pollution.” *Biometrics* 35:795–802.

le Cessie, S. and J. C. van Houwelingen. 1994. "Logistic Regression for Correlated Binary Data." *Applied Statistics* 43:95–108.

Lindsey, J. K. 1995. *Modelling Frequency and Count Data.* Oxford: Oxford University Press.

MacDonald, Iain L. and Walter Zucchini. 1997. *Hidden Markov Models and Othr Models for Discrete-valued Time Series.* Vol. 70 of *Monographs on Statistics and Applied Probability* London: Chapman and Hall.

Marriot, John, Nalini Ravishanker, Alan Gelfand and Jeffrey Pai. 1996. "Bayesian Analysis of ARMA Processes: Complete Sampling-Based Inference under Exact Likelihoods." In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, ed. Donald A. Berry, Kathryn M. Chaloner and John K. Geweke. New York: Wiley pp. 243–256.

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

Oneal, John R. and Bruce Russett. 1997. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985." *International Studies Quarterly* 41(2):267–94.

Plutzer, Eric. 1997. Voter Turnout and the Life Cycle: A Latent Growth Curve Analysis. Technical Report. Political Methodology Working Papers. `http://wizard.ucr.edu/polmeth/working_papers97/plutz97.html`.

Poirier, Dale J. and Paul A. Ruud. 1988. "Probit with Dependent Observations." *Review of Economic Studies* 55:593–614.

Prais, S.J. and C.B. Winsten. 1954. *Trend estimators and serial correlation*. Number 373 Chicago: Cowles Commission Discussion Paper.

Prentice, Ross L. 1988. "Correlated Binary Regression with Covariates Specific to Each Binary Observation." *Biometrics* 44:1033–48.

Russett, Bruce. 1990. "Economic Decline, Electoral Pressure, and the Initiation of International Conflict." In *The Prisoners of War*, ed. Charles Gochman and Ned Allan Sabrosky. Lexington, MA: Lexington Books pp. 123–140.

Russett, Bruce. 1993. *Grasping the Democratic Peace.* Princeton, New Jersey: Princeton University Press.

Singer, J. David and Melvin Small. 1994. *Correlates of War Project: International and Civil War Data, 1816-1992*. Ann Arbor, Michigan: Inter-university Consortium for Political and Social Research. ICPSR Study Number 9905.

Sommer, A., J. Katz and I. Tarwotjo. 1984. "Increased risk of respiratory infection and diarrhea in children with pre-existing mild vitamin A deficiency." *American Journal of Clinical Nutrition* 40:1090–95.

Stern, R. D. and R. Cole. 1984. "A model fitting analysis of daily rainfall data (with discussion)." *Journal of the Royal Statistical Society, Series A* 147:1–34.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Postrior Distributions and Likelihood Functions*. Third ed. New York: Springer-Verlag.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer and B. G. Ferris Jr. 1984. "Passive smoking, gas cooking, and respiratory health of children living in six cities." *American Review of Respiratory Disease* 129:366–374.

West, Mike and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42. in press.

Zeger, S. L. and B. Qaqish. 1988. "Markov regression models for time sries." *Biometrics* 44:1019–31.