Autoregressive Count Models for the Applied Social Scientist

Garrett N. Vande Kamp*

Soren Jordan †

1 Introduction

Time series data in the social sciences are routinely characterized by serial autocorrelation. Such data are typically modelled using autoregressive distributed lag models, which incorporate lags of the dependent and independent variables in order to capture the dynamic processes in the data (Webb and Linn 2020). The autoregressive features of these models - that is, the inclusion of lagged dependent variables as regressors in these models - are particularly important, both for theoretical and methodological reasons.

But while the use of lagged dependent variables is trivial when applied to the linear model for continuous data, their use is much more difficult when applied to the generalized linear model for bounded, discrete data. Such is the case for dynamic count data, where the inclusion of lagged dependent variables as regressors results in econometric difficulty (Brandt et al. 2000). For this reason, dozens of stationary, autoregressive count models have been

*University of Georgia; garrettvandekamp@uga.edu
†Auburn University; scj0014@auburn.edu

proposed, including some that are original to political science (Brandt and Williams 2001). As a result, multiple scholars have made an effort to categorize these models in literature reviews published in the last ten years (Weib 2021, Davis et al. 2021, Cameron and Trivedi 2014).

Unfortunately, the wide variety of autoregressive count models that are available "on the market" actually imply vastly different data-generating processes. Unlike the standard ADL model applied to continuous data, autoregressive count models are not a one-size-fits-all solution, whereby a single model is developed and consistently applied each time a scholar observes a count time series that appears to be autoregressive. And while existing literature reviews do well in explaining the statistical properties of existing models, they do not focus on the questions that applied social scientists have when using these models. There is largely no discussion of what types of social phenomena may be most appropriate for a particular autoregressive count model. Even assuming that the appropriate autoregressive count model is estimated (i.e. the model that most closely matches the implied DGP), there is largely no discussion of relevant quantities of interest, like the long-run effect. And perhaps most importantly, there is little evidence showing the consequences of model misspecification. As such, scholars end up either using the model that is easiest for them to implement or, rather unfortunately, end up using no autoregressive model at all.

This study seeks to resolve these problems by examing autoregressive count models from the perspective of an applied social scientists and formally testing their statistical properties. Our goal is three-fold: identify models relevant to political science, provide guidance on choosing the best model in any given circumstance, and create formula for useful quantities of interest to facilitate interpretation. We start by highlighting the important of autoregression in dynamic linear models. We next review the most popular types of autoregressive count models, paying close attention to substantive theory, quantities of interest, and implementation concerns. We then demonstrate the consequences of model misspecification via

Monte Carlo simulation, a demonstration that to our understanding is the first of its kind. Finally, we conclude with a demonstration on how to apply these tools in an applied setting; specifically, in a model of detentions of democratic dissidents in China (Truex 2019).

2 The Theory of Autoregression

A time series featuring autocorrelation offers both challenges and opportunities to scientists using regression models to make causal inferences on observational data. The dependence caused by autocorrelation violates the assumptions of most regression-based modelling techniques, including ordinary least squares and maximum likelihood. At minimum, estimates may be inefficient; if the dependence represents a misspecification of the regression model, such estimates may also be inconsistent (Keele and Kelly 2006; Wilkins 2016). But autocorrelation also represents an opportunity for scholars to test dynamic theories of causality (Beck 1985). A change in the contemporaneous value of the independent variable might have a causal effect on the dependent variable in the contemporaneous time period, but it may also have a causal effect on the dependent variable in future time periods as well.

In the early days of time series analysis, scholars tested dynamic theories of causation using finite distributed lag (FDL) models that only included lags of the independent variables. While these models could be easily estimated, scholars feared that the number of lags necessary to capture the dynamics in the data would lead to substantial inefficiencies in estimation. Koyck (1954) first demonstrated that a model including a first-order lagged dependent variable as a regressor in a dynamic linear regression model was equivalent to including an infinite number of lags of the independent variables and error term, subject to a geometric functional form. If one assumes that geometric decay reflects the underlying data-generating process of an FDL model, then using a lagged dependent variable allows for more efficient estimation on an FDL model:

$$y_t = \alpha_0 + \alpha_0\alpha_1 + \alpha_0\alpha_1^2 + ...\beta_0 x_t + \beta_0\alpha_1 x_{t-1} + \beta_0\alpha_1^2 x_{t-2} + ... + \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + ...$$

$$\Leftrightarrow y_t = \frac{\alpha_0}{1-\alpha_1 L} + \frac{\beta_0}{1-\alpha_1 L}x_t + \frac{1}{1-\alpha_1 L}\varepsilon_t$$

$$\Leftrightarrow y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \varepsilon_t \tag{1}$$

where L is a lag operator and ε_t is a white noise error term.

Aside from its methodological convenience, however, there are theoretical reasons that many political phenomena are best understood as being autoregressive, with current observations of a series being directly dependent on previous previous realizations of that same phenomena. For example, a legislator may use a government agency's budget in the previous fiscal year as a blueprint for creating a new budget for the coming fiscal year, making changes to it based on contemporaneous factors. This would imply an AR(1) model like equation 1.

The autoregressive-distributed lag model, or ADL(p,q) model, is a general dynamic regression model that currently dominates the study of a real-valued time series by political scientists.[1] The model features p lags of the dependent variable and q lags of the independent variable:

$$y_t = \alpha_0 + \Sigma_{i=1}^{p}\alpha_i y_{t-i} + \Sigma_{i=0}^{q}\beta_i x_{t-i} + \varepsilon_t \tag{2}$$

The inclusion of lags of both the dependent and independent variables allows for easy calculations of the causal effect an independent variable has on the dependent variable, both in the current time period and at periods into the future. While there are many dynamic quantities of interest that can be calculated, two stand out as the most commonly used by

[1]Scholars can easily extend this model to account for autocorrelation in the dependent variable due to a moving average: see Vande Kamp and Jordan (Forthcoming).

political scientists (Deboef and Keele 2008). The short-run discrete effect is the effect that a one-unit increase in the independent variable in the contemporaneous period has on the conditional mean of the dependent variable in that same time period.[2] Due to the nature of the linear model, this discrete effect is also equal to the marginal effect, a name which may be more familiar to scholars (Long 1997). This effect is simple to calculate, as it is simply the coefficient of the independent variable, β_0:

$$\frac{\Delta y_t}{\Delta x_t} = \frac{\partial y_t}{\partial x_t} = \beta_0 \tag{3}$$

The long-run discrete effect is the effect of a permanent, one-unit increase in the independent variable on the conditional mean of the dependent variable in each time period after the series reaches a new equilibrium in the long-run (Blackwell and Glynn 2018).[3] Alternatively called the long-run marginal effect, it can also be interpreted as the cumulative effect a one-time, one-unit increase in the independent variable has on the dependent variable, starting at the time point in which the one-time increase occurred and finishing once the effect of that shock has dissipated over the long-term. Either way you interpret it, the quantity has a simple, closed-form solution:

$$\lim_{h\to\infty} \frac{\Delta y_{t+h}}{\Delta x} = \lim_{h\to\infty} \frac{\partial y_{t+h}}{\partial x} = \sum_{h=0}^{\infty} \frac{\Delta y_{t+h}}{\Delta x_t} = \sum_{h=0}^{\infty} \frac{\partial y_{t+h}}{\partial x_t} = \frac{\sum_{i=0}^{q} \beta_i}{1 - \sum_{j=1}^{p} \alpha_j} \tag{4}$$

At this point, those who are not familiar with time series may question the necessity of dynamic regression models. If β_0 is an important quantity of interest, why not simply estimate a static regression model with x_t as a covariate without all the lags included? While an appealing on its face, such an approach ultimately results in invalid inference. Excluding

[2]This has elsewhere been called the contemporaneous effect or the instantaneous effect.

[3]Some have taken to calling the long-run effect the long-run multiplier (Deboef and Keele 2008). This is a slight misnomer. The long-run multiplier is the quantity $(1 - \sum_{j=1}^{p} \alpha_j)^{-1}$ that multiplier the short-term effect of an independent variable, β_0. Assuming no other lags of the independent variable, the product of these two elements is the long-run effect.

a lagged dependent variable from a regression model when it is needed will result in biased estimates of β_0 if the independent variable is itself autocorrelated (Keele and Kelly 2006; Wilkins 2016).[4] Given that most time series are autocorrelated, including the series that are used as independent variables, ADL models are necessary when using regression for causal inference with time series data.[5]

The ADL(p,q) model is the dominant approach to modelling time series data in political science. But it is far from the only one. Some models focus on introducing dynamic processes into the error term. The ARMAX model specifies that the error term of the dependent variable is an ARMA(p,q) process. Similarly, state-space models specify that the observed dependent variable is a function of a latent state variable, and the error term of that state variable follows an ARMA(p,q) process. While such models are useful, they do not enable the independent variables in the model to have dynamic causal effects. As such, we decline to analyze these models in this paper.[6]

Finally, some dynamic models focus on structural breaks, in which the relationship between the dependent variable and the independent variable(s) fundamentally changes as time passes. If the time period of a potential structural break is known, it can be tested using the Chow test (1960). But if it is not known, it can be estimated using a variety of models, such as a Bayesian changepoint model. Structural breaks are doubtlessly important to time series analysis and complement well the ADL(p,q) model. But one paper cannot cover all topics, so in the interest of parsimony, they will not be discussed in this paper.[7]

[4]Excluding a lagged independent variable when the independent variable is autocorrelated will also result in bias.

[5]We demonstrate this point in the upcoming Monte Carlos.

[6]We note here that generalizations of ARMA error models do exist for count data, including the generalized linear autoregressive moving average model (GLARMA, Dunsmuir and Scott 2018) and the Poisson exponentially weighted moving average model (PEWMA, Brandt et al 2000). We refer interested readers to these papers, as well as the previously cited literature reviews.

[7]We note generalizations of such models have been created for count data, such as the Poisson changepoint model of Park (2010) and its generalization by Blackwell (2018).

3 Extending Autoregression to Count Data

While the above linear model can generate time series data that are continuous on the real number line, it is inappropriate for a variety of situations involving limited dependent variables, including count data. Whenever scholars analyze count data, they typically move away from the linear model and its estimation via OLS. Instead, they use a generalized linear model that features a count distribution with a conditional mean term that, when combined with a link function, is equal to an underlying linear predictor (Long 1997). For time series data, a FDL count model can be written as:

$$y_t = \mathcal{C}(\mu_t, \theta) \tag{5}$$

$$f(\mu_t) = \alpha_0 + \Sigma_{i=0}^{q} \beta_i x_{t-i} \tag{6}$$

where $\mathcal{C}$ is a count distribution such as Poisson or negative binomial, μ_t is the conditional mean of y_t, and $f()$ is a link function. The canonical link function for the Poisson distribution is the natural logarithm, and it is the most commonly assumed link function for count variables. Calculation of discrete effects or marginal effects using a log-link function results in estimates that are different for every observation of the data; as a result, they are not preferred interpretational tools. Instead, percentage effects can be calculated that are universal to every observation; this is the percentage change in the conditional mean of the dependent variable based on a one-unit increase in the independent variable. The short-run percentage effect for a FDL model with a log link functions uses the standard formula:

$$\frac{\%y_t}{\Delta x_t} = (\exp(\beta_0) - 1) * 100 \tag{7}$$

There are those who advocate the identity link function for counts, however, which allows

a scholar to estimate a count model with either OLS or with MLE and an identity link (WHO?). But such a model is not feasible unless the linear predictor is always positive. Furthermore, the only way to prevent the calculation of nonsensical, negative conditional means is to transform the independent variables such that they are non-negative and have a non-negative effect on the dependent variable.

The generalized linear model may seem trivial to political scientists at this point, as such models for count data have long been commonplace (King 1989). But the theory of autoregression makes count models complicated. Simply including a lag of the dependent variable in a generalized linear model with a log link function results in an explosive series if autoregression is positive (Brandt et al. 2000). Statisticians have developed dozens of stationary count models incorporating autoregression in some way at this point. Each model has its own assumptions about its statistical moments, the types of relationships it allows, and the structure of causal effects. Making it even more complicated is that such models have very similar sounding names and acronyms: Integer ARMA, Integer GARCH, linear Poisson autoregression, log-linear Poisson autoregression, Poisson AR, etc. The topic has reached such complexity that multiple literature reviews have been published in recent years covering the most popular of these models (Weib 2021, Davis et al. 2021, Cameron and Trivedi 2014).

While the existing literature reviews are great in many regards, they lack important features to make them relevant to social scientists. First, these literature reviews typically lack a theory-driven explanation of models that give empirical examples that are appropriate for each model. Second, these literature reviews often do not cover estimators for such models in statistical programs. Third, they typically do not list formulas for important quantities of interest, like the short-run effect and the long-run effect. And finally, these reviews do not describe tools that can be used to discriminate between these models if the

scholar is uncertain about which to choose.[8] We therefore seek to review the literature as a social scientist would, paying attention to each of these questions for the most popular autoregressive count models.

4 Choices for Autoregressive Count Models

In this section and its subcomponents, we review the most popular autoregressive count model choices from a theoretical perspective.

4.1 Autoregressive Count Models Using Binomial Thinning

Some of the most appealing autoregressive models for count data subject the lagged dependent variable to a binomial thinning operator (McKenzie 1985).[9] These models treat a count variable as a sum of individual units, each of which might persist into a future period or expire according to a Bernoulli distribution.[10] Mathematically, these models are the sum of two distinct components: a standard count distribution, similar to those used for FDL count models, and a lag of the observed count subject to a thinning operation. The most common thinning operation is the binomial thinning operation; a first order autoregressive model using this operation can be written as:

$$y_t = \alpha_1 \circ y_{t-1} + \mathcal{C}(\mu_t, \theta) \tag{8}$$

$$\alpha_1 \circ y_{t-1} = \Sigma^{y_{t-1}} \mathcal{B}(\alpha_1) = \mathcal{BIN}(\alpha_1, y_{t-1}) \tag{9}$$

where $\alpha_1 \circ y_{t-1}$ is equal to the realization of a y_{t-1} series of independent Bernoulli distributions sharing the success parameter α_1, which can be equivalently expressed as a single

[8] Brandt and Williams (2001) provide a notable exception.

[9] These models have been described as Integer ARMA models, or INARMA models, in the literature.

[10] Other thinning operations exist; see cited literature reviews for references.

Binomial trial with the success parameter α_1 and the number of trials equalling y_{t-1} (Du and Li 1991). As before, $\mathcal{C}$ is a count distribution and μ_t is subject to a link function to make it equivalent to a linear predictor as in equation 6.

In a first-order binomial thinning model, the two parts can be viewed as a birth term, in which new counts are created, and a death term, in which existing counts either persist to become part of the new count or perish and are removed from the model. As such, it is particularly compelling when a count variable consists of individual entities that are capable of persisting.[11] For example, a count of the number of bills a legislator proposes in a legislative session may be a thinning process because the legislator can propose unsuccessful bills from previous legislative sessions (Cox and Terry 2008). Similarly, counts of prisoners in a given year may be a thinning process because existing prisoners can either be released from prison or held for another year (Horz and Marbach 2022). But the same feature that makes it appealing also makes it limiting; the model cannot typically feature negative autoregression.

The model has a number of desirable statistical properties. The model is stationary, so long as $\sum_{i=1}^{p} \alpha_i < 1$, which mirrors the stationarity conditions in the linear ADL model (Scotto, Weiß, and Gouveia 2015). Additionally, the autoregressive parameters of the model can be estimated using the Yule-Walker equations, providing a non-parametric method that helps with both model specification and estimation (Weiß 2021). In addition, the first-order model has the further property that the distribution of the dependent variable, y_t, is identical to the distribution for the birth term, $\mathcal{C}$.

The estimation of a thinning model depends on the choice of link for the linear predictor. Binomial thinning results in a conditional mean that is subject to geometric decay. As such, the choice of an identity link function results in a linear model for the conditional mean. While a maximum likelihood estimator is not currently available in software programs, the

[11] Unfortunately, thinning models lose this compelling interpretation whenever there is autoregression higher than one. Then it has an interpretation as a branching process, which seems to have little utility in a social science setting.

model can be consistently estimated using OLS since it is a linear model. Choosing an identity link function typically requires non-negative independent variables and non-negative relationships. In addition, the very nature of binomial thinning means that only positive autoregression can be represented.

If one chooses the log-link function, the `coconots` package in `R` estimates a thinning model with a generalized thinning operator that is inclusive of binomial thinning when using a Poisson distribution (Jung and Tremayne 2011). This estimator is also restricted to positive autoregression, but can handle any type of covariate or coefficient in its link function. The package allows users to estimate either first-order or second-order autoregression and use either a Poisson or generalized Poisson distribution. Unfortunately, it does not allow for techniques that handle excess zeroes. Semiparametric and Bayesian techniques also exist in other `R` packages, though there models do not allow for covariates and are therefore of limited utility to social scientists.[12]

Calculation of quantities of interest are relatively straightforward. When using the identity link function, one simply uses the formulas in equation 3 and equation 4. When using the log link function, the short-run percentage effect uses the formula in equation 7. Unfortunately, there is no simple way to calculate a long-run percentage effect, and the actual formula appears difficult to use in practice. In its place, we follow Brandt and Williams (2001) practice and instead choose to report the short-run and long-run marginal effects:

$$\frac{\partial y_t}{\partial x_t} = \exp(\alpha_0 + \Sigma_{i=0}^{q} \beta_i x_{t-i})\beta_0 \tag{10}$$

$$\lim_{h\to\infty} \frac{\partial y_{t+h}}{\partial x} = \sum_{h=0}^{\infty} \frac{\partial y_{t+h}}{\partial x_t} = \frac{\exp(\alpha_0 + \Sigma_{i=0}^{q} \beta_i x_{t-i})\beta_0}{1 - \sum_{j=1}^{p} \alpha_j} \tag{11}$$

While the marginal effects are easy to calculate, the conditional mean from the birth

[12]We refer to the ZINARp and spINAR packages.

term results in a marginal effect that is different for each observation. While unfortunate, scholars have devised multiple solutions to present summaries of these effects. Hanmer and Kalkan (2013) advise reporting the sample average of effect estimates, which could easily be computed and given.

4.2 Autoregressive Count Models using Transformations of the Lagged Dependent Variable

A second class of autoregressive models feature lags of the dependent variable in the linear predictor (Zeger and Qaqish 1988).[13] These models treat the count variable much like a FDL model; the dependent variable is distributed as it is in equation 5. The difference is that the linear predictor includes a lagged dependent variable transformed by the link function assumed by the scholar for the conditional mean. The conditional mean of a first-order autoregressive model using a transformation of the lagged dependent variable can be written as:

$$f(\mu_t) = \alpha_0 + f(y_{t-1}) + \Sigma_{i=0}^{q} \beta_i x_{t-i} \tag{12}$$

When an identity link is assumed, the conditional mean is linear and functions as expected for all values of y_t. But when the log-link is assumed, zero becomes an absorbing state: any observed value of zero will cause the rest of the series to permanently become zero. This is due to the fact that the log of zero is negative infinity. While such a model is still mathematically sound, it does not likely represent many real world data-generating processes. When using the log link function, Zeger and Qaqish (1988) propose the use of a modified lagged dependent variable that removes zeroes via a mathematical operation. This results in a slightly different conditional mean function:

[13] These models have been described as Integer ARCH models, or INARCH models, in the literature.

$$\log(\mu_t) = \alpha_0 + \log(y^*_{t-1}) + \Sigma^q_{i=0}\beta_i x_{t-i} \tag{13}$$

Zeger and Qaqish propose two choices for y^*. The first is to set a minimum value for the observed series such that the observed zero values are replaced with that minimum value: $y^*_{t-s} = \max(y_{t-s}, c)$, $0 \leq c \leq 1$.[14] The second is to add a constant to every observation in the time series so that no value equals zero: $y^+_{t-s} = y_{t-s} + c$, $0 \leq c \leq 1$.

By putting the lagged dependent variable in the conditional mean, individual realizations of a count are no longer capable of persisting into future time periods. Instead, the process that caused those counts is allowed to have a dynamic effect within the link function, subject to geometric decay. These models are therefore ideal for those phenomena in which individual instances of a count are not capable of persistence but, nonetheless, observed counts are autocorrelated. Counts of terror attacks are an ideal example (Brandt and Sandler 2010). A terror attack does a unique amount of damage each time, injuring people and damaging property in such a way that cannot be repeated. Even so, counts of terror attacks are autocorrelated, indicating that the factors that cause one terror attack are likely to persist into future attacks. The log link function has the additional feature of being able to handle both positive and negative autoregression without issue. The identity link function typically cannot handle such a relationship.

These models have desirable statistical properties. The models are identified and stationary, so long as $|\sum^p_{i=1}\alpha_i| < 1$, which mirrors the stationarity conditions in the linear ADL model (Doukhana, Fokianos, and Tjøstheim 2012). Additionally, the autoregressive parameters of the identity link model can be estimated using the Yule-Walker equations, while the parameters of the log link model can be estimated using a variant of the Yule Walker equations (Weiß 2021). Because the lagged dependent variable is part of the conditional

[14]It is feasible to imagine a c larger than 1, but this is rarely done because 0 is the only problematic count in the log-linear autoregressive model. In keeping with the literature, we assume $0 \leq c \leq 1$ from now on.

mean, the distribution of the observed variable, y_t, is identical to $\mathcal{C}$.

Estimation of models with lagged dependent variables is trivial and can be done with standard estimators for count models. If a scholar chooses the identity link function, a linear model results and estimation can proceed as it was done for thinning models. If a scholar chooses the log link, estimation can still be done using existing estimators for generalized linear models. However, the scholar must now choose which modification of the lagged dependent variable to use. If the scholar chooses to use the lag of the dependent variable plus a constant, then the scholar must also choose the value of the constant prior to model estimation. Failure to pick the correct value will result in inconsistent estimates; as such, the model does not seem practical for scholars attempting to estimate an unknown data-generating process.

In contrast, the scholar may choose to use the lag of the maximum between the observed dependent variable and a constant. Cameron and Trivedi (2014) state that for this model, there is an estimation procedure that allows scholars to avoid choosing the constant and, as such, allows for consistent estimation. They start by creating a lag of the dependent variable as normal, assuming that the constant is equal to 1 ($y^{*1}_{t-s} = \max(y_{t-s}, 1)$). They then include a companion dummy variable, d_{t-s} in which a 1 indicates that the observed lag dependent variable, y_{t-s}, is equal to zero and 0 otherwise. For an AR(1) process, the conditional mean scholars will estimate is:

$$\log(\mu_t) = \alpha_0 + \alpha_1 \log(y^{*1}_{t-1}) + \theta_1 d_{t-1} + \Sigma^q_{i=0}\beta_i x_{t-i} \tag{14}$$

The intuition behind this approach is that an arbitrary value for c is chosen, inducing measurement error that is then captured by the lagged dummy. When estimating this model, one can recover an estimate of c using the calculation $c = \exp(\frac{\theta}{\alpha_1})$ (Cameron and Trivedi 2013). To demonstrate this formula, and the equivalence of equation 13 and equation 14 in

general, start by recognizing that $y_t^* = y_t^{*1} c^{d_t}$. Then, simply substitute this equation into equation 13 and apply basic logarithm rules until you get equation 14.

Calculation of quantities of interest are relatively straightforward. When using the identity link function, one simply uses the formulas in equation 3 and equation 4. When using the log link function, the short-run effect uses the formula in equation 7. The long-run percentage effect differs depending on which implementation you use. Because only the model featuring the maximum between the observed and a constant can be consistently estimated, we only provide its long-run percentage effect:

$$\lim_{h\to\infty} \frac{\%y_{t+h}}{\Delta x} = \sum_{h=0}^{\infty} \frac{\%y_{t+h}}{\Delta x_t} = (\exp(\frac{\sum_{i=0}^{q} \beta_i}{1 - \sum_{j=1}^{p} \alpha_j}) - 1) * 100 \tag{15}$$

4.3 Other Notable Autoregressive Models

There are a few additional autoregressive models political scientists have used that must be mentioned. The first class of models feature a transformation of the lagged conditional mean as an autoregressive term in the conditional mean (Ferland, Latour, and Oraichi 2006; Fokianos and Tjøstheim 2011). Such models can be viewed as the count data counterpart to an autoregressive model for binary data developed by political scientists for binary data (Wucherpfenig et al. 2021). As with models using a transformation of the lagged dependent variable, the dependent variable is distributed as it is in equation 5. The conditional mean of a first-order autoregressive model using a transformation of the lagged conditional mean can be written as:

$$f(\mu_t) = \alpha_0 + f(\mu_{t-1}) + \Sigma_{i=0}^{q} \beta_i x_{t-i} \tag{16}$$

This models shares many attributes with models using transformations of the lagged dependent variable. It is appealing for autocorrelated count data where individual counts

are not capable of persisting, and the log link can handle negative autoregression. The identification and stationarity properties are the same. The calculation of dynamic effects are also identical.

But there are some differences as well. First, the use of the lagged conditional mean prevent the error term from having a geometric decay. This is notably different the previously articulated autoregressive count models, as well as the autoregressive model in the continuous case. This is not necessarily a problem, as the model may be a closer approximation of real-world data-generating processes than their counterparts that use the lagged dependent variable. But it is something that must be mentioned because it is a departure from standard assumptions. Second, the use of the unobserved conditional mean in the linear predictor requires a new likelihood function and, therefore, new estimation tools. The `tscount` package in `R` develops an estimator using either the Poisson or negative binomial distributions. There do not appear to be any implementations that take into account excess zeroes, however.

The final model that we want to cover is one original to political science. Brandt and Williams (2001) create a unique autoregressive count model within the state-space framework, an approach that replaces the deterministic conditional mean used in other autoregressive models with a latent state variable that is of theoretical interest to scholars. While the state-space framework is attractive for particular types of time series applications, it is not universally required to estimate an autoregressive model. As such, we will not be going into the state-space aspects of Brandt and Williams' model. Nonetheless, the state variable from Brandt and Williams model can be replaced with a deterministic conditional mean, resulting in a unique autoregressive model. Brandt and Williams assume the data-generating process of equation 5, specifically with a Poisson distribution. The conditional mean for a first-order autoregressive model is written as:

$$\mu_t = \alpha_1 y_{t-1} + (1 - \alpha_1) \exp(\alpha_0 + \Sigma_{i=0}^{q} \beta_i x_{t-i}) \tag{17}$$

The lag of the dependent variable is included as part of the conditional mean, but it is not inside the inverse link function that contains the linear predictor. As such, it draws elements from both the thinning model with a log link and the transformed dependent variable model with a log link. Additionally, the conditional mean subjects the "birth" term to the tradeoff term $(1 - \alpha_1)$. In other words, the dynamics in the model are zero-sum: an increase in influence of autoregression in the realization of counts comes with a decrease in the influence of contemporaneous independent variables in the realization of counts. This results in a strictly smaller conditional mean than a thinning model with a log link function. The presence of the tradeoff term makes the conditional mean similar to finite mixture models, in which the two populations the data are drawn from are the current values of the linear predictor and lagged values of the linear predictor, subject to a geometric decay.

A state-space estimator of this model in R has been provided by Brandt on his website.[15] There has yet to be an implementation of the model for more general count distributions, nor has a model been developed for excess zeroes. There has also not been a non state-space estimator of this model.

Calculating quantities of interest is strikingly similar to that of thinning models. As reported by Brandt and Williams, the short-run marginal effect is calculated using equation 11, which is the same formula as the long-run marginal effect for a thinning model with the log link function. Similarly, the formula for the long-run marginal effect is equation 10.

Having now reviewed the most popular and useful count approaches, we move to Monte Carlo simulations. As a reminder, our goal is to identify the costs and benefits of applying different count models across multiple data-generating processes. Sometimes these choices will be appropriate: for instance applying a model that suggests the persistence of a lagged conditional mean to a DGP that contains the same. The benefit in these "ideal cases" should be clear and obvious. Often, though, our choices are ill-informed or uninformed, and we use

[15]https://personal.utdallas.edu/ pxb054000/code–software.html

models simply out of convenience or previous experience. These costs are unknown, so we estimate them.

5 Simulating the Consequences of Misspecifying an Autoregressive Count Model

Because a scholar never truly knows the data-generating process of a real world phenomena, it is impossible to say ex ante what dynamic model is most appropriate for a particular count time series. As such, it is critical to know the consequences of model misspecification and find techniques for avoiding misspecification. Some existing studies have applied multiple autoregressive count models to the same dataset in an effort to illustrate how scholars might choose among these models. Cameron and Trivedi (2013) carry out an analysis of the monthly count of strikes in the U.S. manufacturing sector. Their analysis demonstrates that both the thinning model and the transformation of the lagged dependent variable model using log links fit the data much better than a FDL model, though the transformed dependent variable model appears to have slightly more explanatory power. They do a second analysis on the number of stock trades in five-minute intervals during a day at the New York Stock Exchange. The autoregressive models again perform substantially better than FDL models, though this time the thinning models edge out the transformation of the lagged dependent variable model. Weib (2021) does an analysis of weekly sales of a soap product in U.S. supermarkets. He finds that model fit substantially improves for both the thinning model and the transformed dependent variable model when the negative binomial distribution is used instead of a Poisson distribution, with the transformed dependent variable model having slightly better explanatory power.

While these demonstrations are certainly informative, they fail to detail the consequences of choosing an incorrect autoregressive count model but is otherwise correctly specified.

Therefore, we seek to fill this gap with a Monte Carlo experiment. The construction of a proper test of these models is difficult given that each model has various restrictions on what types of relationships it can accommodate. Models using an identity link function cannot accommodate a non-negative independent variable or a non-negative coefficient. Similarly, thinning models or transformed models with an identity link cannot accommodate negative autoregression. Furthermore, estimators of each type of model are limited in development and are inconsistent in what types of distributions they provide and whether they can accommodate excess zeroes.

In order to best demonstrate the consequences of choosing an incorrect autoregressive count model, an overarching meta-model is developed. Every data-generating process in the simulation is a first order autoregressive count model with a Poisson distribution. The exogenous covariates in each model are a single independent variable and an intercept, ($x_t\beta + \alpha_0$), in which $\beta = 1$ and $\alpha_0 = 0.1$. The autoregressive term is experimentally varied, $\alpha_1 \in (0, 0.5)$. The independent variable is a binary variable that is either white noise or positively autocorrelated. It is generated as:

$$x_t^* = \rho x_{t-1}^* + u_t \tag{18}$$

$$x_t = \begin{cases} 1 & \text{if } x_t^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where u_t is normally distributed white noise and $\rho \in \{0, 0.5\}$

Using this overarching model, there are seven different autoregressive count models the data can be generated from. Two are the binomial thinning models of equation 8, with the identity and log link function. Two are transformed lags of the dependent variable models of equation 12, with the identity or log link function.[16] Two are transformed lags of the

[16]The log-link function uses $\log(y_{t-1}^*) = \log(max(y_{t-1}, 0.5))$

conditional mean mode of equation 16, with the identity or log link function. And the last is the autoregressive model developed by Brandt and Williams (2001) in equation 17. For each data-generating process, 500 data points are generated for the independent and dependent variable.[17]

Once the data are generated, we estimate seven different models to the data: one correct specification and six misspecifications. The `lm` function is used as an estimator for the conditional mean of the of the binomial thinning model with identity link. The `cocoReg` function in the `coconots` package is used to estimate the binomial thinning model with the log link. The `glm` function is used to implement the transformed lags of the dependent variable models, implemented in the form recommended by Cameron and Trivedi (2013). The `tsglm` function in the `tscount` package is used to estimate the models with the transformed lags of the conditional mean. Finally, the code available of Brandt's website is used to estimate the Brandt and Williams model, as mentioned in an earlier footnote. In addition, two models without autoregressive parameters are estimated using the glm function.

Once each model was estimated, three different measures of model fit were calculated and stored. The first two should be familiar to scholars: Akaike's Information Criteria (AIC) and Root Mean Squared Error (RMSE). The third, the log-score rule, is a general assessment tool for probabilistic models (Czado, Gneiting, and Held 2009). It requires the calculation of a model's predicted probability of each observed value of the dependent variable in the sample. These probabilities are then transformed into a log-score, $-\log(\hat{Pr}(y_t))$, and the average is calculated as a measure of model assessment. Log-scores work similar to information criteria tools like AIC, as lower number indicates better model fit. But their use of predicted probabilities is reminiscent of expected percentage of correct predictions measure that Herron (1999) proposed for binary models.

After each model is estimated, coefficients and standard errors are stored. With seven

[17]We specify $y_0 = \mu_0 = 0$.

data-generating processes and four parameter combinations, this results in 28 unique data-generating processes. With nine estimators used per dataset, that is 252 unique data-model pairs simulated. The process repeats 1000 times, resulting in 28,000 models simulated and 252,000 models estimated. Simulation took place on a Precision 5820 Tower Workstation with 16 cores and 128 mb of RAM. The simulation took approximately two hours.

6 Simulation Results

We present three sets of results. First, we demonstrate the consequences of ignoring autoregression and instead simply estimating a static model with the correct link function. The results are contain in Figure **??**. In the left-hand column, the data-generating process has no autoregression because $\alpha_1 = 0$. As such, static models can recover the correct parameter estimate of $\beta = 1$. In the right-hand column, however, the static models are misspecified.

As seen by comparing the lower right pane to the other panes in Figure **??**, the estimates of β for most models are consistent unless both the dependent variable and independent variable are autoregressive. If the independent variable is not autoregressive, as shown in the top right pane of Figure **??**, the models featuring transformed lags and the thinning model with the identity link are still consistent. In contrast, bias in β will always occur when a static model is estimated for either a thinning data-generating process with a log link or the Brandt and Williams data-generating process. In both instances, the data-generating process features a log link for the linear predictor but is a linear function of the lagged dependent variable. This unusual mistmatch between the functional forms of the linear predictor and the lagged dependent variable seem to drive the result that omitting autoregressive terms for either data-generating process will result in bias, regardless of whether the independent variable is autoregressive.

We have two general interests: which model fits the data best, and which model best

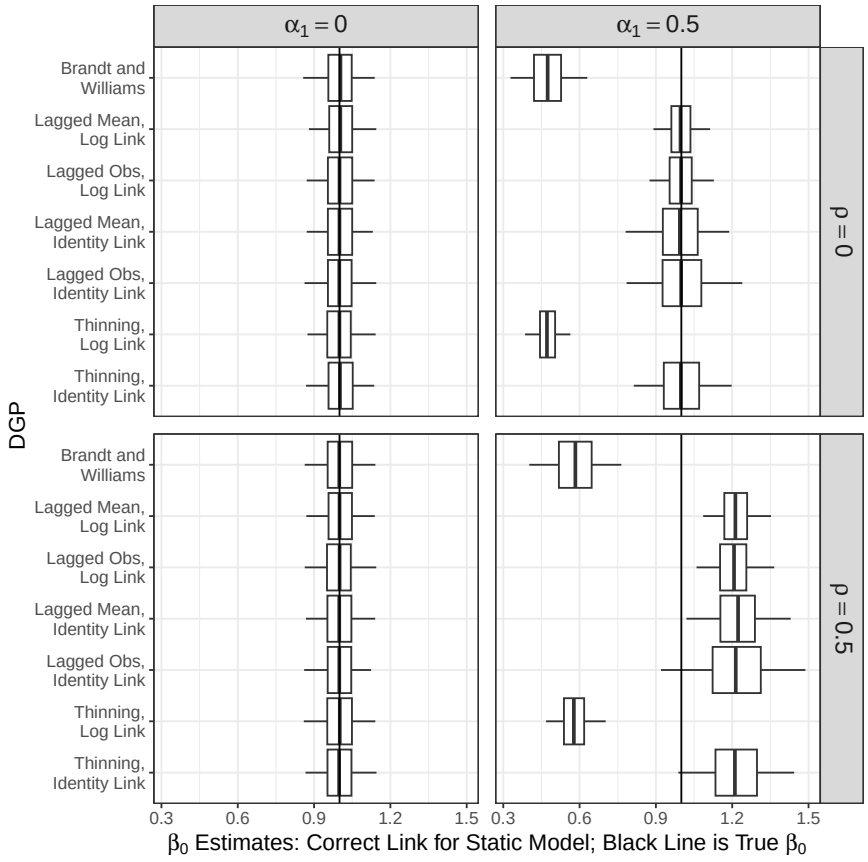

Figure 1: Estimates of β from a dynamic data-generating process using a static model with the correct link function

uncovers the true parameter estimate. For the purpose of MPSA 2024, we have not yet developed a generalized way of displaying our results—on this point, we would like feedback. The challenge is the number of possible combinations: for each of the seven DGPs, we estimate all seven possible models, varying autoregression in x and y. At this point, we provide a general plot (one DGP) for fit statistic and parameter estimate. We discuss the pattern of our results in the presentation.

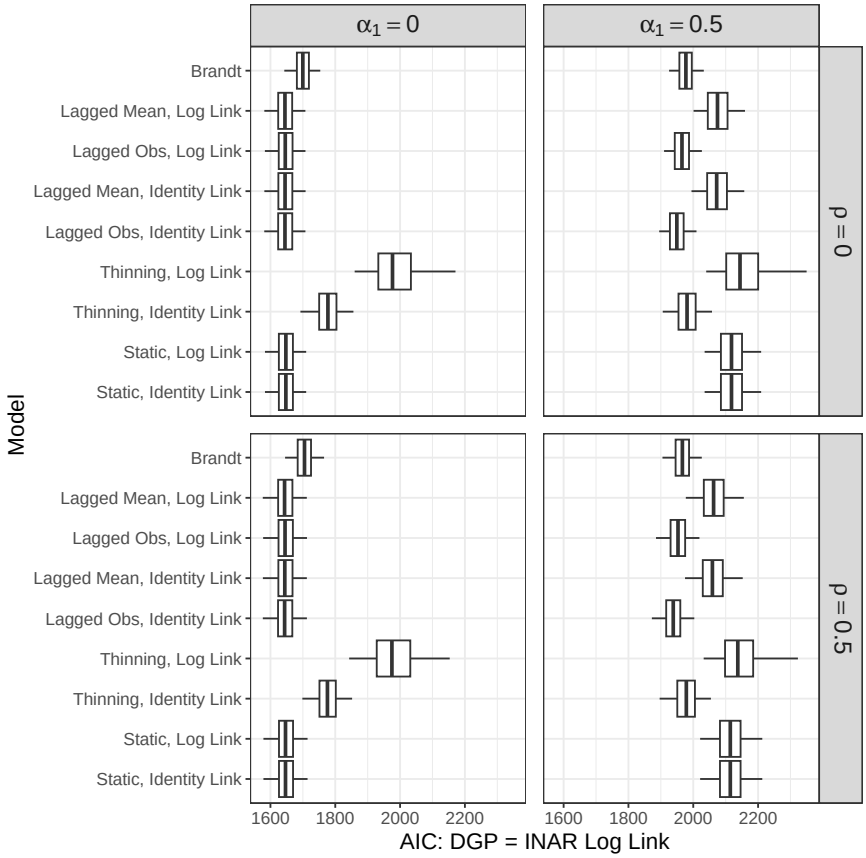

Figure 2: AIC

Figure **??** shows the fit statistics for the model estimate on the y-axis, given that the true DGP follows an INAR process with a log link. Also note that the Brandt and Williams estimator is excluded, as its AIC was universally dominated and exploded the scale of the figure. The correct model to choose would be the Thinning Model with a Log Link. The AIC

suggests that an incorrect choice in this context is relatively costless when there is no autoregression in the dependent variable (the left column). The AICs for multiple models—with the odd exception of the correct choice—are relative equitable. When there is autoregression in y (right column), the INARCH model with the lagged conditional mean performs best, as long as an identity link function is specified.

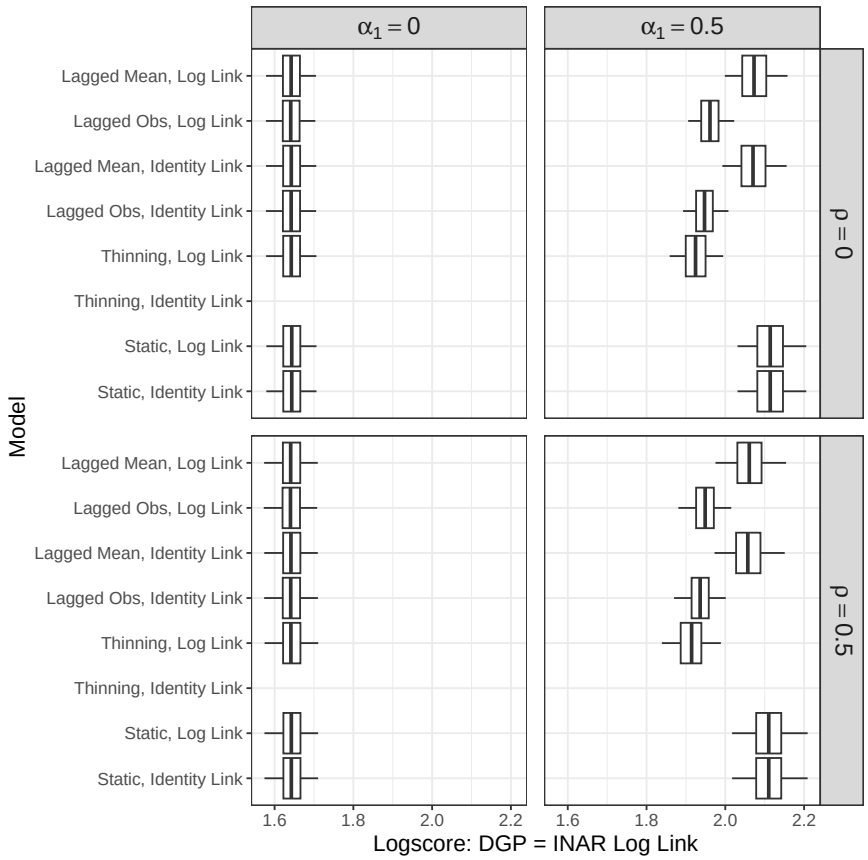

Figure 3: Log Score

Figure **??** demonstrates the same comparisons, this time measuring the log score. The left column largely echoes the same conclusions: the choice of model results in a negligible change in the fit statistic, so the choice of modeling strategy seems arbitrary. The story is quite different in the right column, where the correct modeling function (the Thinning Model, Log Link) outperforms all others. Given that we know the truth in this circumstance—the true

DGP—we get some evidence that there is also a preferred fit statistic in this circumstance, in addition to a preferred model.

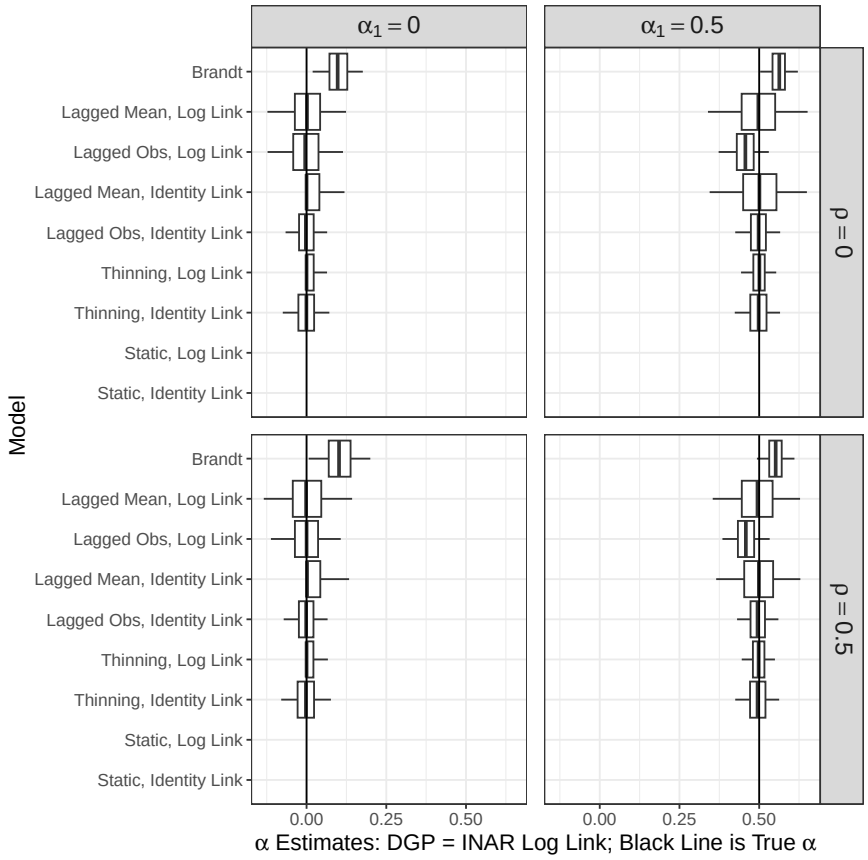

Figure 4: Estimate: y lag.

Figure **??** shows the estimates recovered for the lag of the dependent variable for the same data generating process. The static models are omitted, as they estimate no such coefficient. Focusing again on the scenario in which there is autoregression in y (the right column), we see that two models are biased—the Brandt and Williams model as well as the INARCH model with an identity link. Estimating these models instead of the model that matches the true DGP results in biased inferences (in different directions). The other five models are generally unbiased, though their efficiency varies greatly. The INARCH model with a lagged conditional mean (with either a log or identity link function) uncovers the correct estimate

on average, but there is considerable dispersion around these estimates. Especially in series with limited observations, as is often the case in time series, these inefficiencies could lead to changes in inferences.

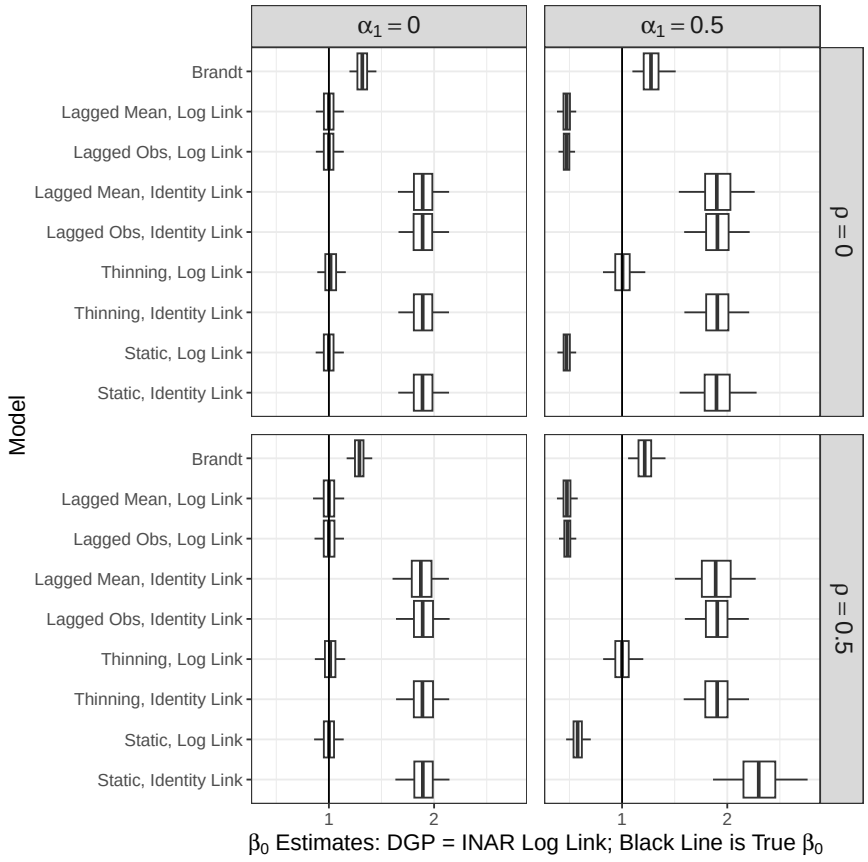

Figure 5: Estimate: x lag.

Figure **??** finally shows the resutls for recovered estimates on the x variable. These estimates are untransformed. Here, we see dramatic variation in the quality of the estimates. When there is no autoregression, any model that uses a log link will recover similarly quality inferences (as long as those estimates are not transformed before interpretation). However, when there is autoregression in y (right column), *only* the true model that represents the true DGP uncovers the correct parameter estimate. Bias—sometimes dramatic bias—results when the wrong model is applied.

7 Application: Detentions in Autocratic States (Truex 2019)

To demonstrate how a scholar would employ autoregressive count models, we focus on a study that has a count time series as a dependent variable but does not use an autoregressive model. Truex (2019) seeks to explain variation in the Chinese government's detention of democratic dissidents. Truex theorisizes that autocratic governments like China will preemptively engage in repression before collective demonstrations occur. In particular, he focuses on the "dissident calendar," or "the set of events known in advance that serve as natural focal points for coordination and collective action." Truex hypothesizes that as a focal event on the dissident calendar approaches, the Chinese government will detain more democratic dissidents as a preemptive response to collective action.

Truex tests this hypothesis on a time series of the monthly number of detentions of democratic activists from 1998-2014. His main independent variable is a coding of focal events, a count variable that indicates how many focal events occur during a given month. Focal events include events like "the major anniversaries of the Tiananmen Square Massacre, the selection of key leaders at the five-year Party Congress, and even the Beijing Olympic Games." He controls for events that are unforeseen by autocratic governments but nevertheless may influence detentions, such as governance scandals, leadership purges, and foreign revolutions. He also includes a cubic trend term.

While Truex does not estimate an autoregressive count model in his analysis, there is strong reason to believe that autoregression exists in the data. In the supplemental materials, he plots the estimates of the autocorrelation function for the data. The plots indicate a positive, statistically significant first-order autocorrelation. This is preliminary evidence that the model is first-order autoregressive.

Truex also makes an interesting choice not to specify a distributed lag model; there are no lags of any of the variables created in the data. Instead, Truex chooses to code each focal event that occurs as happening during the month it actually occurs as well as the two months preceding the event; a table of this coding scheme can be seen in Table **??**. This coding strategy is done to capture "the preemptive logic of repression"; because autocratic governments know when these focal events will occur, they start detaining individuals in the two months leading up to the focal event. Mathematically, this choice is equivalent to specifying a model using the standard focal event coding and its first two temporal leads, but assuming the set of variables to all have the same coefficient. This coding scheme is not unique to the focal event variable, either: all of the control events, which involve unforeseen events, are coded as occurring the month they occur and two months following that event.[18]

Table 1: Truex's Coding of Focal Event Variable, with Leads as Comparison

Time point	Event	1st Lead	2nd Lead	Truex Specification
t - 2	0	0	1	1
t - 1	0	1	0	1
t	1	0	0	1
t + 1	0	0	0	0
t + 2	0	0	0	0

In order to improve model fit, we first start with the task of identifying a set of autoregressive models that could plausibly describe the data. First, the data are positively autocorrelated. This means that models that assume positive autoregression - including thinning models and the Brandt and Williams (2001) model - are plausible.[19] Second, the main independent variable and all of the controls are strictly non-negative. This means that models using an identity link function are also plausible. Of the seven types of autoregressive models described, all seven could be responsible for generating the data.

[18]With the exception of the leadership transition variable, which codes the twelve months following the event.

[19]The thinning models are additionally supported by the idea that detentions are repeatable events: a dissident can be detained and released in the same month, then detained again in the next month.

Second, we estimate the seven first-order autoregressive versions of the model Truex estimates in Table 1, Column 3 in their original paper. In order to ensure that the autoregressive models are all comparable, they are all estimated assuming the Poisson distribution. While the choice may appear to induce bias due to dispersion issues in the data, it actually does not. The Poisson distribution is one of the few generalized linear models in the exponential family that are consistent so long as the conditional mean is correctly specified; it shares this property with the normal distribution, generalizing one of the desirable features of the ordinary least squares estimator (Cameron and Trivedi 2014; Gourieroux, Monfort, and Trognon 1984). Goodness of fit statistics for these estimates are presented in Table 2. The lagged dependent variable model with the log link, as shown in equation 14, is the best fitting model. As such, it is adopted as the autoregressive model of choice.

Table 2: Fit Statistics for Autoregressive Models of Detentions of Democratic Dissidents

	Thinning Id Link	Thinning Log Link	Transformed DV, Id Link	Transformed Mean, Id Link	Transformed Mean, Log Link	Transformed DV, Log Link	E Log
AIC	1011.83	1469.05	656.09	830.01	636.90	578.34	5(
RMSE	2.77	2.62	2.85	3.01	2.57	2.41	
ln(score)	N/A	1.46	1.57	2.00	1.52	1.37	

Finally, a series of dynamic count models are estimated and reported in Table 3. The first model is Truex's original model. The second is a finite-distributed lag model using focal event and its leads as independent variables rather than Truex's unusual specificaiton. The third is the lagged dependent variable model using Truex's original independent variable. The last model incorporates both innovations: the focal event, its leads, and a lagged dependent variable.

The replacement of Truex's coding of the focal events variable with a more standard set of variables leads to a consistent improvement in model fit. AIC, RMSE, and Log Score all decline when these changes are made, regardless of whether the model includes autoregressive terms or not. The inclusion of the lagged dependent variable and its associated dummy term,

Table 3: ???

	Original	FDL	PA	ADL
(Intercept)	−0.26	−0.29	−0.04	−0.17
	(0.50)	(0.50)	(0.53)	(0.53)
Focal Event$_{t+2:t}$	0.91***		0.79**	
	(0.26)		(0.24)	
Focal Event		0.92*		0.87*
		(0.41)		(0.36)
Focal Event$_{t+1}$		0.98*		0.93**
		(0.41)		(0.36)
Focal Event$_{t+2}$		0.80+		0.53
		(0.42)		(0.39)
Log(Detentions$^{*}_{t-1}$)			0.47*	0.59**
			(0.18)	(0.19)
Dummy (Detentions$_{t-1}$=0)			−0.82**	−0.76**
			(0.27)	(0.27)
Num.Obs.	204	202	203	201
AIC	505.4	501.3	484.8	477.7
RMSE	2.67	2.59	3.13	3.04
Log.Score	1.55	1.52	1.45	1.38
SRPE - Focal Event	149.25	150.19	279.92	395.11
LRPE - Focal Event	1448.48	1372.54	8576.91	29 079.32

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

however, causes a split in the fit statistics. AIC and the log score are both lower when the autoregressive terms are included, but RMSE is lower when it is not. For their part, the terms themselves are statistically significant. On the whole, the evidence suggests that the full ADL model is the best fit for the data.

The inclusion of dynamic terms in Truex's model changes the effect estimates in the data, as can be seen by the short-run percentage effect and long-run percentage effect in Table 3. Under Truex's model, the conditional mean of detentions is 150% higher during focal events than when a focal event does not occur. In the ADL model, the conditional mean is 400% higher; this is more than double the previous effect estimate. The long-run percentage effect is even more dramatic. The long-run total percentage point increase in the conditional mean is about 1500% in the original model. When the ADL is estimated, the long-run percentage point increase is 29000% percent; this is roughly a 20x increase in the long-run effect.

8 Conclusion

The authors didn't write a conclusion because they hate science :(

9 References

Beck, Nathaniel, Kristian Skrede Gleditsch, and Kyle Beardsley. "Space is more than geography: Using spatial econometrics in the study of political economy." International studies quarterly 50.1 (2006): 27-44.

Benjamin, Michael A., Robert A. Rigby, and D. Mikis Stasinopoulos. "Generalized autoregressive moving average models." Journal of the American Statistical association 98.461 (2003): 214-223.

Brandt, Patrick T., and John T. Williams. "A linear Poisson autoregressive model: The

Poisson AR (p) model." Political Analysis 9.2 (2001): 164-184.

Brandt, Patrick T., et al. "Dynamic modeling for persistent event-count time series." American Journal of Political Science (2000): 823-843.

Cameron, A. Colin, and Pravin K. Trivedi. Regression analysis of count data. Vol. 53. Cambridge university press, 2013.

Franzese, Robert J., Jude C. Hays, and Scott J. Cook. "Spatial-and spatiotemporal-autoregressive probit models of interdependent binary outcomes." Political Science Research and Methods 4.1 (2016): 151-173.

Davis, Richard A., et al. "Count time series: A methodological review." Journal of the American Statistical Association (2021): 1-15.

Douc, Randal, Paul Doukhan, and Eric Moulines. "Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator." Stochastic Processes and their Applications 123.7 (2013): 2620-2647.

Fokianos, Konstantinos, and Dag Tjøstheim. "Log-linear Poisson autoregression." Journal of Multivariate Analysis 102.3 (2011): 563-578.

Hays, Jude C. and Robert J. Franzese. "A Comparison of the Small-Sample Properties of Several Estimators for Spatial-Lag Count Models" in Franzese, ed., Advances in Political Methodology, Elgar Research Collections (2017): 180-207.

Jordan, Soren, and Andrew Q. Philips. "Exploring meaningful visual effects and quantities of interest from dynamic models through dynamac." Journal of Open Source Software 5.54 (2020): 2528.

King, Gary. "Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model." American Journal of Political Science (1988): 838-863.

King, Gary. "Variance specification in event count models: From restrictive assumptions to a generalized estimator." American Journal of Political Science (1989): 762-784.

Sim, Tepmony, Randal Douc, and François Roueff. "General-order observation-driven models: ergodicity and consistency of the maximum likelihood estimator." arXiv preprint arXiv:2106.05201 (2021).

Whitten, Guy D., Laron K. Williams, and Cameron Wimpy. "Interpretation: the final spatial frontier." Political Science Research and Methods 9.1 (2021): 140-156.

Wucherpfennig, Julian, et al. "A Fast Estimator for Binary Choice Models with Spatial, Temporal, and Spatio-Temporal Interdependence." Political Analysis (2021): 1-7.

Zeger, Scott L., and Bahjat Qaqish. "Markov regression models for time series: a quasi-likelihood approach." Biometrics (1988): 1019-1031.

Liboschik, Tobias, Fokianos, Konstantinos and Fried, Roland. (2017). tscount: An R package for analysis of count time series following generalized linear models. Journal of Statistical Software 82(5), 1–51