Discrete Choice Data with Unobserved Heterogeneity: A Conditional Binary Quantile Model*

Xiao Lu[†]

[†]Department of Political Science, University of Mannheim, A5, 6, 68159, Mannheim, Germany. Email: xiao.lu@gess.uni-mannheim.de

Abstract

In political science, data with heterogeneous units are used in many studies, such as those involving legislative proposals in different policy areas, electoral choices by different types of voters, and government formation in varying party systems. To disentangle decision-making mechanisms by units, conditional mean-based discrete choice models focus exclusively on the conditional mean and ignore the heterogeneous effects within a population. This paper proposes a conditional binary quantile model that goes beyond this limitation to analyze discrete response data with varying alternative-specific features. This model offers an in-depth understanding of the relationship between the explanatory and response variables. Compared to conditional mean-based models, the conditional binary quantile model relies on weak distributional assumptions and is more robust to distributional misspecification. The model also relaxes the assumption of the independence of irrelevant alternatives, which is often violated in practice. The method is applied to a range of political studies to show the heterogeneous effects of explanatory variables across the conditional distribution. Substantive interpretations from counterfactual scenarios are used to illustrate how the conditional binary quantile model captures unobserved heterogeneity, which extant models fail to do. The results point to the risk of averaging out the heterogeneous effects across units by conditional mean-based models.

Keywords: discrete choice, unobserved heterogeneity, conditional binary quantile regression, Bayesian inference

1 Introduction

In this paper, I propose an alternative approach to conditional mean-based models for analyzing discrete choice data with alternative-specific features. While there has been tremendous progress in modeling discrete choice data in political science (see, e.g., Alvarez and Glasgow, 1999; Poole, 2001; Sartori, 2003; Carter and Signorino, 2010; Glasgow, 2011; Traunmüller et al., 2014; McGrath, 2015; Rainey, 2016), little attention has been paid to the incompleteness of existing estimators that produce only a single estimate for each covariate across units. When heterogeneity exists among individuals or subgroups of a population, which is often the case in research on topics such as party politics, legislative decision making, election studies, trade policies, wars and conflicts, etc., the existing estimators risk averaging out heterogeneous effects across units and thus providing biased estimates. Based on the previous progress, this paper develops a conditional binary quantile (CBQ) model that is capable of describing the full conditional distribution of the response variables and incorporating alternative-specific features into the estimation. The strength of the CBQ model is that it allows the estimation of unobserved effects across units without imposing restrictive distributional forms. In contrast to the conditional mean-based discrete choice model that is inconsistent when the latent error distribution differs from the assumption, the CBQ model is more robust to distributional misspecification. Compared to the conditional logit (CL) model, the CBQ model inherits no assumption of the independence of irrelevant alternatives (IIA), which is often violated in practice.

Conditional mean-based discrete choice models, such as logit and probit models,

have been used extensively to model data with discrete outcomes. However, those conventional models only estimate the conditional mean and ignore heterogeneity among individuals or subgroups within a population. In many situations, we are interested in the full behavior of the explanatory variables beyond the conditional mean. For example, we may be interested in how bichamber conflict affects the adoption of bills with high and low success rates and how ideological factors affect coalition choices among coalitions with high and low likelihoods of formation.

Quantile regression has been developed to remedy the shortcomings of the conditional mean-based model that fails to assess the full conditional properties of continuous response variables (a thorough overview of the development of quantile regression can be found in Koenker (2005)). Similar to the continuous case, in a discrete choice setting, we are also interested in disentangling heterogeneous effects among units. However, due to the difficulty of estimation, quantile regression is seldom used to model discrete responses (for an exception, see Benoit and Poel (2012)). There is no closed-form solution for quantile regression even in the simplest binary setting, and the traditional optimization-based estimators, such as the maximum likelihood estimator, fail in such cases. The exceptional work by Benoit and Poel (2012) develops a simulation-based estimator for binary quantile models that draws posterior samples through the Markov chain Monte Carlo (MCMC) process. However, until now, it has been applied only to the simplest binary setting, where the features of the choice alternatives within each choice set remain the same.

Because quantile values are equivalent under monotonic transformations (Koenker and Machado, 1999), quantile regression is particularly suitable for discrete choice

models that require a certain form of nonlinear transformation from continuous latent utilities to discrete outcomes (Horowitz, 1992). Compared to the conditional mean-based regression, quantile regression is robust to location-scale shifts of the conditional distribution of the response variable (Benoit and Poel, 2012). It is less sensitive to distributional misspecification and provides a richer view of the influence of the explanatory variables on the response. However, compared to the models with continuous dependent variables, quantile regression with a discrete dependent variable has been underappreciated. Only recently have researchers in both the theoretical and applied fields started to recognize the potential benefits of quantile regression to model discrete outcome data (Koenker and Hallock, 2001; Kordas, 2006; Benoit and Poel, 2012; Benoit et al., 2013; Oh et al., 2016).

To deal with the varying features of alternatives within each choice set, the conditional binary model extends the simplest binary model to a multinomial setting where choice-specific features are also analyzed. The CL model, one of the most popular conditional binary models, inherits from the binomial logit model the often-violated IIA assumption (McFadden, 1986). To partly relax the IIA assumption, the mixed logit (MXL) model introduces random coefficients that are assumed to take a certain distributional form (Hensher and Greene, 2003; Hensher et al., 2015). As a result, the MXL model produces biased estimates when the strong distributional assumption of random coefficients is violated. It also suffers an identification problem when the number of covariates is large. In addition, the choice regarding which parameters are random coefficients is often arbitrary. The conditional probit model overcomes the IIA assumption by adding a correlation structure to the covariates.

However, additional constraints on the covariance matrix are needed to identify the model. When the error is not normally distributed, the estimates from the conditional probit model are biased. The fact that many probit models are unidentified also indicates the complexity of the conditional probit model. When decision makers hold heterogeneous preferences, the above models provide only an incomplete account of the underlying mechanism. Aware of the pitfalls of the existing methods, I present a CBQ model to provide a novel approach to remedy the above shortcomings.

This paper contributes in several respects. First, beyond the limitations of the conditional mean-based models, this paper devises a CBQ model that is capable of representing the full conditional distribution through a range of quantiles to model unobserved heterogeneity across units. Second, it relaxes the often-violated distributional assumption in traditional discrete choice models, such as probit and logit. Compared to the MXL model, which partly relaxes the IIA assumption by assuming a certain distributional form of random coefficients, the CBQ model has been proven to be robust to a wide range of error distributions. Third, this paper wraps the model into a Bayesian framework and enables efficient and exact inferences of parameter values through the MCMC process. Consequently, in this paper, quantities of interest are drawn to facilitate substantive interpretations. Finally, by applying the model to a range of political studies, I compare the performance of the CBQ model with that of a range of popular discrete choice models. To illustrate how the CBQ model is able to capture unobserved heterogeneity, which extant models fail to do, I devote great effort to substantive interpretations from counterfactual

scenarios.

The rest of the paper proceeds as follows. First, the CBQ model is developed, and its estimation strategy is described. Second, using both a homogeneous and a heterogeneous dataset with two choice alternatives, the simulation studies demonstrate how well the CBQ model approximates the true conditional distribution and how it outperforms the often-used ordinary logit model. The third part focuses on real-world applications in a range of political studies, namely, the EU legislature, the US presidential election and government formation in parliamentary democracies. The ordinary logit, multinomial probit, CL and MXL models are originally used in the respective applications. The complexity of the data structure in those applications is in increasing order; whereas the 1992 US presidential election had three candidates, the EU legislature has only two choice alternatives. The government formation dataset has the most complicated structure, with different numbers of potential governments across countries and elections. This arrangement helps illustrate the applicability of the CBQ model to a wide range of discrete choice data. Substantive interpretations based on counterfactual scenarios are used to show how the CBQ model accounts for unobserved heterogeneity and, in comparison, how the conditional mean-based models fail to do so. The final section concludes by pointing to areas of future research. For interested readers, I provide a brief discussion on the limitations of the conditional mean-based methods in the Appendix.

2   The conditional binary quantile model

Since the seminal paper by Koenker and Bassett (1978), quantile regression has been extensively used to model across quantiles the varying effects of the explanatory variables on the continuous response variable. Instead of modeling the conditional means of the response variable, quantile regression estimates the conditional quantiles that provide extension to the linear regression that fails to assess heterogeneity among units or subgroups of a population. Similarly, in the face of a discrete response variable, quantile regression enables a full analysis of its conditional properties. Manski (1975) pioneers the research by introducing a semiparametric estimation of a binary quantile regression. The main idea is to relax the distributional assumption of the error term and allow the data to determine the value of the parameters. Such an estimation is originally conducted using a maximum score estimator (MSE). However, the implementation of the MSE within the framework of mixed integer programming is computationally difficult and has a very slow convergence rate (Kim and Pollard, 1990). An alternative subsampling approach that limits its focus on the conditional median is also computationally intensive (Delgado et al., 2001). Nonparametric smoothed density estimation presents an approximating approach to the estimation (Horowitz, 1992), which is later extended to the conditional quantiles (Kordas, 2006). However, the smoothing of the error distribution is restrictive and performs badly at the tails of the distribution (Kotlyarova, 2005). To overcome this problem, scholars have recently introduced a Bayesian analogue to Manski's approach (Kordas, 2006; Florios and Skouras, 2008; Benoit and Poel, 2012; Benoit et al., 2013). The main idea of the Bayesian approach is to incorporate hyperpa-

rameters with prior distributions in forming the joint posterior distribution with the likelihood function, which can then be sampled using MCMC methods. Therefore, rather than approximating the conditional distributions of the quantiles by using smoothing methods, the Bayesian approach allows the exact inference from the joint posterior distribution. However, until now, the method has only been sparsely applied to the simplest binary setting without alternative-specific features. It cannot handle a situation in which there are multiple alternatives with varying features in each choice set, which is often the case in political science research.

To complement the existing methods in modeling data with the discrete response variable, this section develops a CBQ model where the variables of a dataset vary not only between units but also between the choice alternatives. The CBQ model provides reliable estimates by calculating the conditional probabilities of alternatives within each choice set, and it is robust to distributional misspecification without relying on the IIA assumption.[1] The structure of this section is as follows. First, the binary quantile model in the simplest setting is introduced. Then, the full-fledged CBQ model is formed by imposing a conditional structure over the simple binary quantile model. Finally, in this section, an estimation strategy to infer parameter values from the CBQ model is developed.

2.1 The conditional quantiles

Quantiles are a set of values that divide a distribution or a population into equal groups. Instead of assuming a data generating process (DGP) conditioning on the

---

[1]For formal proof, I refer to Manski (1985). For a brief discussion of conditional mean-based models, see the Appendix.

mean, it is often reasonable to assume a quantile DGP (Horowitz, 1992; Kordas, 2006; Florios and Skouras, 2008; Benoit and Poel, 2012; Alhamzawi and Yu, 2013; Benoit et al., 2013; Alhusseini, 2017). The earlier version of quantile regression focuses on the conditional median as a robust alternative to the conditional mean. However, when the numbers of individuals within each subgroup are different, the conditional median poorly represents the whole sample (Kordas, 2006). The same is true for arbitrarily set quantiles when between-group imbalance exists. Therefore, unless theoretically driven, it is desirable to model data with a rich set of quantiles instead of using only one of them.

Quantile regression models the data with a τth conditional quantile as

$$Q_\tau(y|\boldsymbol{x}) = F_\tau^{-1}(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1}$$

where $F(.)$ is a cumulative density function, $\boldsymbol{x}$ corresponds to varying features of the data, and $\boldsymbol{\beta}$ is a vector of coefficients associated with the features. By inserting a monotone indicator function into Equation (1), the quantile representation of the binary response variable conditioning on a set of covariates has the following form:

$$Q_\tau(y = 1|\boldsymbol{x}) = I(F_\tau^{-1}(\boldsymbol{x}'\boldsymbol{\beta}) > 0), \tag{2}$$

where $I(.)$ is an indicator function that takes the value of 1 when the condition within the bracket is satisfied. Consequently, the τth quantile of a random variable Y, which is $q_Y(\tau)$, is defined as the inverse of its cumulative density function or the inferior of all supported values in its cumulative density distribution greater than

8

the quantile:

$$q_Y(\tau) := F_Y(y)^{-1} = \inf\{y : F_Y(y) \geq \tau\}. \tag{3}$$

In contrast to the conditional mean-based approach that minimizes a squared loss function, the τth quantile is obtained by minimizing an asymmetric loss function $\mathcal{L}$, whose components are weighted by τ (Yu and Moyeed, 2001):

$$\mathcal{L} = \tau \int_{y>q} |y - q| dF_Y(y) + (1 - \tau) \int_{y<q} |y - q| dF_Y(y), \tag{4}$$

where q represents the value of the τth quantile as calculated by Equation (3). From Equation (4), it is obvious that instead of splitting the sample, the quantile loss function uses all the sample information to derive the quantile values. Because the above linear loss function depends not on certain covariate values but on covariate orderings, the quantile estimators are more robust than the least square estimators in the presence of extreme values. By solving the first order condition, we have the τth sample quantile of y as

$$\frac{1}{N} \sum_{i=1}^{N} \rho_\tau(y_i - q), \tag{5}$$

where $\rho_\tau(x) = x(\tau - I(x < 0))$ is known as the check function (Koenker and Bassett, 1978). Given a linear model $y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i$, the τth quantile regression estimator minimizes the above equation by replacing q with $\boldsymbol{x_i'}\boldsymbol{\beta}$. In such models, we often assume $Q(\varepsilon_i|\boldsymbol{x_i}, \tau) = 0$. If $\tau = 0.5$, the model is degenerated to a median regression.

The goal of the estimation is to discover the values of $\boldsymbol{\beta}$ by solving

$$\arg \min_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - \boldsymbol{x_i'}\boldsymbol{\beta}|. \tag{6}$$

Or, equivalently,

$$\boldsymbol{\beta_\tau} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \rho_\tau \left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta} \right). \tag{7}$$

2.2   Incorporating alternative-specific features

In the simplest two-alternative (binary) setting, the conditional probability that a particular alternative $j = 1$ from a choice set $J = \{1, 2\}$ is chosen is simply the probability that the utility of that alternative y_{i1}^* is greater than that of the other alternative, which is y_{i2}^*. In formula, note that $Pr(y_{i1} = 1 | \boldsymbol{x_i}) = Pr(y_{i1}^* > y_{i2}^*)$, where $\boldsymbol{x_i}$ is a vector of covariates. To identify the model, one of the alternative utilities, y_{i2}^*, is always normalized to zero, which results in $Pr(y_{i1} = 1 | \boldsymbol{x_i}) = Pr(y_{i1}^* > 0)$. In such a simple binary setting, the above probability can be represented as a cumulative density function of the latent variable y_{i1}^* (see e.g., Benoit and Poel, 2012).

When the number of choice alternatives increases, the probability calculation of specific outcomes becomes more complicated. Consider the following example with three alternatives in a choice set $J = \{1, 2, 3\}$. The probability of choosing $j = 1$ conditioning on a set of covariates $\boldsymbol{x_i}$ is simply the probability that the utility from alternative 1 is greater than that of alternatives 2 and 3: $Pr(y_{i1} = 1 | \boldsymbol{x_i}) = Pr(y_{i1}^* > y_{i2}^* \quad \text{and} \quad y_{i1}^* > y_{i3}^* | \boldsymbol{x_i})$, where $\boldsymbol{x_i} = (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \boldsymbol{x_{i3}})'$ are choice-specific features. Since in a latent linear specification $y_{ij}^* = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \varepsilon_{ij}$, the latent utility consists

10

of a deterministic component and a random component, after reordering the random and deterministic components in the equation, we have the above conditional probability equal to the double integrals of the corresponding random components:

$$Pr(y_{i1} = 1|\boldsymbol{x_i}) = \int_{-\infty}^{(\boldsymbol{x_{i1}}-\boldsymbol{x_{i2}})'\boldsymbol{\beta}} \int_{-\infty}^{(\boldsymbol{x_{i1}}-\boldsymbol{x_{i3}})'\boldsymbol{\beta}} f(\varepsilon_{i3} - \varepsilon_{i1}, \varepsilon_{i2} - \varepsilon_{i1})d(\varepsilon_{i3} - \varepsilon_{i1})d(\varepsilon_{i2} - \varepsilon_{i1}),$$

(8)

where $f(.)$ denotes the joint probability density function of $\varepsilon_{i3} - \varepsilon_{i1}$ and $\varepsilon_{i2} - \varepsilon_{i1}$. We can simplify the above equation by regarding the difference between two random components as a single random term: $\eta_i^{jk} = \varepsilon_{ij} - \varepsilon_{ik}$. More generally, let u_i^{jk} denote $(\boldsymbol{x_{ij}} - \boldsymbol{x_{ik}})'\boldsymbol{\beta}$; the probability of choosing alternative $j \in J$ in the choice set with the number of alternatives K_i is

$$Pr(y_{ij} = 1|\boldsymbol{u_i}) = \int_{-\infty}^{u_i^{j1}} \cdots \int_{-\infty}^{u_i^{j(j-1)}} \int_{-\infty}^{u_i^{j(j+1)}} \cdots \int_{-\infty}^{u_i^{jK_i}}$$

(9)

$$f(\eta_i^{j1}, \ldots, \eta_i^{j(j-1)}, \eta_i^{j(j+1)}, \ldots, \eta_i^{jK_i})d\eta_i^{j1} \ldots d\eta_i^{jK_i}.$$

If the number of choice alternatives K_i $\forall i \in \{1, \ldots, N\}$ is the same across all choice sets, we can simply multiply the probabilities of each choice alternative to calculate the joint likelihood of the whole dataset. However, in practice, we often encounter data with different numbers of choice alternatives. This leads to a biased likelihood towards choice sets with high numbers of alternatives by equally multiplying the probabilities of each choice alternative. To have an unbiased likelihood, a multinomial structure must be embedded in the calculation of the joint probability (McFadden, 1974). Therefore, before the joint probabilities of all choice sets are calculated, the probability for each alternative is weighted by the number of

choice alternatives in the corresponding choice set. Each weight is formulated as $\frac{K_i!}{\prod_{l \in \{1...S_i\}} s_{il}!}$, where K_i is the number of alternatives in choice set i, and s_{il} is the number of alternatives l in choice set i with the total number of alternatives S_i.

In summary, we have the following conditional likelihood function given N choice sets:

$$\mathcal{L}_{\text{condition}} = \prod_{i=1}^{N} \frac{K_i!}{\prod_{l \in \{1...S_i\}} s_{il}!} \prod_{j=1}^{K_i} Pr(y_{ij} = 1|\boldsymbol{x_i})^{I(y_{ij}=1)} Pr(y_{ij} = 0|\boldsymbol{x_i})^{I(y_{ij}=0)}. \quad (10)$$

Notice from the above equation that for alternatives not chosen, we must calculate the inverse probabilities: $Pr(y_{ij} = 0|\boldsymbol{x_i}) = 1 - Pr(y_{ij} = 1|\boldsymbol{x_i})$.

Equation (10) is a very general form of the conditional binary choice model. The variation includes the CL model (McFadden, 1974), the conditional probit model (Hausman and Wise, 1978), and the MXL model (Hensher and Greene, 2003). The CL model specifies the probability of choosing j from K alternatives in each choice set i as

$$Pr(y_{ij} = 1|\boldsymbol{x}) = \frac{exp(\boldsymbol{x'_{ij}\beta})}{\sum_{k=1}^{K} exp(\boldsymbol{x'_{ik}\beta})}. \quad (11)$$

In contrast to the CL model, the conditional probit model assumes a normal distribution of the error (random) terms and incorporates the correlation of the errors among the units (for details on the conditional probit model, I refer to Hausman and Wise (1978)). The conditional probit model relaxes the assumption of IIA by imposing correlation among the units. However, the specification of the covariance matrix and the estimation of the model are very complicated. Compared to the conditional probit model, the MXL model adopts a random coefficient approach to

relax the assumption of IIA. It is specified as follows:

$$Pr(y_{ij} = 1|\boldsymbol{x}) = \frac{exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\eta_i})}{\sum_{k=1}^{K} exp(\boldsymbol{x}'_{ik}\boldsymbol{\beta} + \boldsymbol{z}'_{ik}\boldsymbol{\eta_i})}, \tag{12}$$

where $\boldsymbol{z_{ij}}$ is a vector of features varying over choice sets and $\boldsymbol{\eta_i}$ is a vector of the corresponding random coefficients. However, since the true DGP is unobserved, the distributional assumptions of the error term in the CL model and the random coefficients in the MXL model are likely to be violated. In addition, the CL model is also sensitive to the violation of the IIA assumption.

In contrast to the above-mentioned models, the CBQ model, whose estimation is deliberated in the following subsection, relies on less restrictive distributional assumptions and depends on no IIA assumption. Compared to the conditional probit model, the CBQ model has a much simpler structure.

2.3 The estimation strategy

In general, the binary quantile regression has no closed-form solution. To solve this problem, Manski (1975) relies on mixed linear programming that is computationally intensive and unfeasible for a large dataset. Other scholars rely on continuous approximation using kernel densities but obtain poor performance around the tails of the distribution (Kotlyarova, 2005). Koenker and Machado (1999) find that the minimizing problem in Equation (7) is closely related to the asymmetric Laplace distribution (ALD). Therefore, without using approximation methods, we can estimate the quantile regression within the framework of a special distribution. I adopt
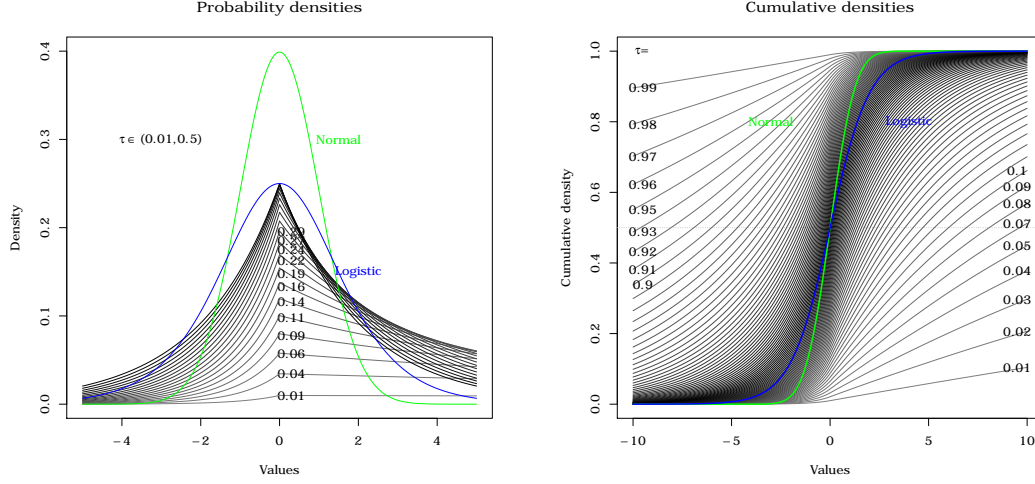
Figure 1: Comparison between the standard normal distribution, the standard logistic distribution, and the ALDs, with different quantile specifications τ (scale fixed to one)

this approach in my model.

The ALD density of a random variable y conditioning on three parameters, μ, σ and τ, can be written as

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\{-\rho_\tau(\frac{y-\mu}{\sigma})\}, \tag{13}$$

where $\rho_\tau(x) = x(\tau - I(x < 0))$ and $\mu = \boldsymbol{x}'\boldsymbol{\beta}$. In a discrete choice setting, σ is normalized to 1 for the purpose of identification. Figure 1 compares the probability densities and cumulative densities between the normal distribution, the logistic distribution, and the ALD with different quantiles.

For a binary response variable y_i, the probability of $y_i = 1$ conditioning on the

14

data and the quantile is as follows:

$$
\begin{aligned}
Pr(y_i = 1|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau) \ &= Pr(y_i^* > 0|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau) \\
&= Pr(\boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i > 0|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau) \\
&= Pr(\varepsilon_i > -\boldsymbol{x_i'}\boldsymbol{\beta}|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau) \\
&= 1 - Pr(\varepsilon_i < -\boldsymbol{x_i'}\boldsymbol{\beta}|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau) \\
&= 1 - F_{\text{ALD}}(-\boldsymbol{x_i'}\boldsymbol{\beta}|\boldsymbol{x_i}, \boldsymbol{\beta}, \tau).
\end{aligned}
\tag{14}
$$

Careful readers may notice that the ALD replaces the logistic or normal distribution as another error distribution and may doubt whether the ALD suffers a similar problem of distributional misspecification. Fortunately, it has been formally shown that in fairly general conditions, the posterior consistency holds under the ALD, even if the true error distribution is misspecified (Sriram et al., 2013). Empirical studies also support this finding (Yu and Moyeed, 2001).

Instead of using Metropolis-Hasting within the Gibbs algorithm to sample the joint posterior distribution, the ALD can be represented as a location-scale mixture of normal distributions (Kotz et al., 2001; Yu and Moyeed, 2001; Kotz et al., 2003; Reed and Yu, 2009; Kozumi and Kobayashi, 2011):

$$
\varepsilon = \theta z + \tau\sqrt{z}u,
\tag{15}
$$

where

$$
\theta = \frac{1 - 2\tau}{\tau(1 - \tau)}, \text{and } \tau = \sqrt{\frac{2}{\tau(1 - \tau)}}.
\tag{16}
$$

As a result, latent variable y_i^* can be rewritten as

$$y_i^* = \boldsymbol{x_i'\beta} + \theta z_i + \tau\sqrt{z}u, \tag{17}$$

where $z_i \sim Exponential(1)$ and $u_i \sim Normal(0,1)$ are auxiliary variables. Therefore, we have the joint distribution as

$$f(\boldsymbol{y^*}|\boldsymbol{\beta}, z) \propto \left(\prod_{i=1}^{N} z_i^{-1/2}\right) \exp\left\{-\sum_{i=1}^{N} \frac{(y_i^* - \boldsymbol{x_i'\beta} - \theta z_i)^2}{2\tau^2 z_i}\right\}. \tag{18}$$

In the simple binary choice model, it is possible to augment the data in the sampling procedure by adding a latent variable y_i^*, which has continuous support over the whole real line:

$$\pi(y_i^*|y_i, \boldsymbol{\beta}, \tau) \propto \{I(y_i^* > 0)I(y_i = 1) + I(y_i^* < 0)I(y_i = 1)\}f_{\mathrm{ALD}}(y_i^*|\boldsymbol{x_i'\beta}, \tau). \tag{19}$$

As a result, the inference of $\boldsymbol{\beta}$ is simply the integral of latent variable y_i^*:

$$Pr(\boldsymbol{\beta}|y_i) = \int_{y_i^*} p(\boldsymbol{\beta}|y_i, y_i^*)f(y_i^*|y_i)dy_i^*. \tag{20}$$

However, in the CBQ model, the technique of data argumentation cannot be used directly. Because each choice set contains multiple alternatives, there is no clearcut threshold for the latent utility function. Instead, the inference of the parameters relies on the joint posterior density of all random components in the model. Based on the above procedure and Equation (10), the joint posterior distribution of the

CBQ model is given by

$$
Pr\left(\boldsymbol{\beta}, \boldsymbol{y}^* | \boldsymbol{y}, \tau\right) \propto
$$

$$
Pr\left(\boldsymbol{\beta}\right) \prod_{i=1}^{N} \frac{K_i!}{\prod_{l \in \{1...S_i\}} s_{il}!} \prod_{j=1}^{K_i} \left(\prod_{k \in \{1...K_i\}/j} F(y_{ij}^*|(\boldsymbol{x_{ij}} - \boldsymbol{x_{ik}})'\boldsymbol{\beta}, \tau) \right)^{I(y_{ij}=1)} \times
$$

$$
\left(1 - \prod_{k \in \{1...K_i\}/j} F(y_{ij}^*|(\boldsymbol{x_{ij}} - \boldsymbol{x_{ik}})'\boldsymbol{\beta}, \tau) \right)^{I(y_{ij}=0)} ,
$$

(21)

where $F(.)$ is the cumulative density function of the ALD, N is the number of choice sets, K_i is the number of alternatives in the choice set i, and s_{il} is the number of alternatives l in the choice set i, with the total number of alternatives S_i. The posterior distribution is assessed using a gradient-based MCMC algorithm (Hamiltonian Monte Carlo) that provides an efficient sampling routine (Hoffman and Gelman, 2014). The sampler is programmed in an open-source software called stan (Carpenter et al., 2017).

3 Simulation studies

Applying a CBQ DGP will doubtlessly lead to better performance of the CBQ model compared to the performance of other models. Nevertheless, in reality, we are always left with some DGP that is unknown to researchers *a priori*. Therefore, for practical purposes and to illustrate the usefulness of the CBQ model across a wide range of situations, it is better to simulate datasets using a DGP other than CBQ. For the simplest binary choice data, researchers always impose a logit or probit form whose estimates are rather similar. In this simulation study, I want to show that even with

a DGP other than CBQ, the CBQ model can well approximate the true DGP. With a heterogeneous dataset, the CBQ model not only captures inherent heterogeneity but also outperforms the traditional conditional mean-based models in terms of prediction accuracy, at least over certain quantiles.

This simulation study generates two sets of data: one homogeneous and the other heterogeneous.[2] The homogeneous dataset is intended to illustrate how well CBQ can approximate the true DGP. I use the heterogeneous dataset to show why the conditional mean-based models fail to capture heterogeneity and how the CBQ models are able to do so. To simulate homogeneous datasets, I assume an ordinary logit DGP in which the deterministic component consists of a constant and a single covariate. The value of the constant is set to zero ($\alpha = 0$), and the coefficient of the covariate is set to five ($\beta = 5$). Thus, $Pr(y_i = 1) = \text{logit}^{-1}(5 \times x_i) \quad \forall i \in I$, where $x_i \sim N(0, 1)$ is predetermined.

The homogeneous dataset assumes the same coefficient value across observations. In contrast, in the heterogeneous dataset, the coefficients vary across different covariate values. Intuitively, this variance means that the preference of individuals differs according to their own characteristics. For example, voters' intention to vote for a certain party depends on their distance from that party. Compared to those with a small distance, voters with a large distance from a party may be less responsive to other characteristics of the party in their vote choice. To model this kind of heterogeneity, I assume coefficients equate to the second power of the values of the covariate. Therefore, $\beta_i = x_i^2, \quad \forall i \in I$. The rest of the specifications remain the

---

[2]Codes for replicating the analyses are available (Lu, 2019).

same as the homogeneous DGP. For each dataset, I simulate 200 observations based on a binomial realization (0 or 1) of the calculated probabilities from the ordinary logit model. To map this DGP to real-world examples, I also replicate the analysis of the EU legislature in the application section. More complicated real world examples with multiple choices (US presidential voting) and an inconstant number of choice alternatives across choice sets (government formation) are also illustrated in the application section.

Figure 2 shows the predicted probabilities based on the estimates from both the ordinary logit and the CBQ model. In the left panel, where a homogeneous logit DGP is assumed, the predicted probabilities of the ordinary logit (black dots) and the CBQ model (gray lines) show rather similar shapes. The CBQ model approximates the true ordinary logit DGP very well, as the gray lines almost cover the true predicted probabilities with very little variation. The variation comes from the stochastic component that maps the continuous probability measure into binary values of the response variable. At the tails, the predicted probabilities of the two types of estimates almost overlap with each other. In comparison, the right panel shows a different picture when the dataset is heterogeneous. Both at the tails and in the middle of the predicted probabilities, we observe a relatively large variation over different quantile estimates, indicating a clear sign of unobserved heterogeneity. Because the ordinary logit model only relies on the conditional mean for estimation, it can only produce the conditional mean estimates and the predicted probabilities based on the condition mean, which clearly fails to capture unobserved heterogeneity from the true DGP. Corresponding to the true DGP, the predicted probabilities from

the CBQ model over different quantiles show that individuals indeed have varying preferences over X in their decision making. If computational power allows it, people can always increase the number of quantiles to obtain a better approximation of the whole conditional distribution of the true DGP. In this case and the following cases, nine quantiles are already enough to show a rich picture of unobserved heterogeneity among observations.
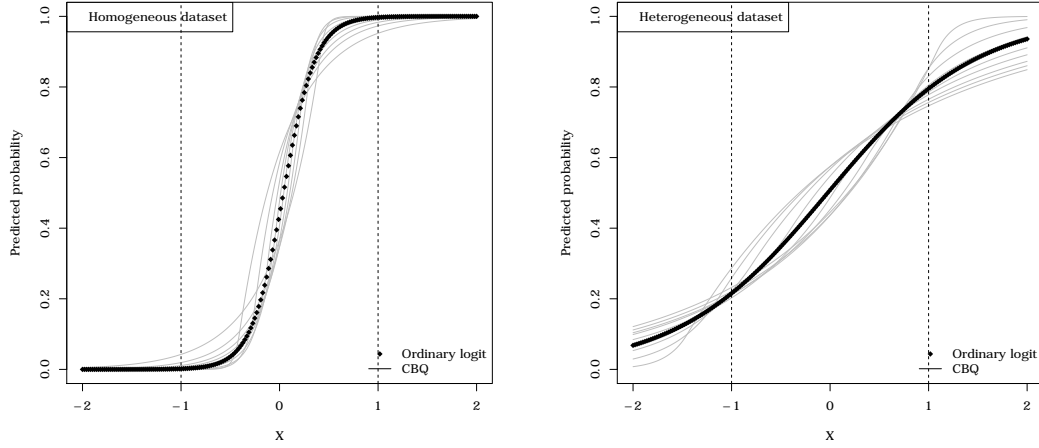

Figure 2: Comparison between the CBQ model (from 0.1 to 0.9 quantiles) and the ordinary logit model

To benchmark the performance of the CBQ model against the ordinary logit model in the above simulation studies, Table 1 compares the prediction accuracies between the two types of models.[3] When the true DGP is a homogeneous logit, the prediction accuracies of the CBQ model across quantiles is very similar to that of the

---

[3]The predicted response value equals 1 when the predicted probability is larger than 0.5, and 0, otherwise.

ordinary logit model; five out of nine CBQ estimators perform equal to or slightly better than the ordinary logit. However, in regard to the heterogeneous dataset, the prediction accuracies of most CBQ estimators are equal to or outperform that of the ordinary logit, except for the median estimator (CBQ-Q5). In this particular example, this outcome indicates that the estimators based solely at the middle (mean or median) of the conditional distribution are inferior to the estimators with a rich set of quantiles. Any single quantile provides only a piece of the information that is necessary to construct a complete picture of the underlying mechanism. There are always quantiles producing better predictions than others, but that does not mean that other quantiles are less important. The varying performance across quantiles may result from the fact that the number of observations surrounding a particular quantile is smaller than that of other quantiles. This variation can also result from stochastic components in the DGP.

Table 1: Benchmarking performance of the ordinary logit and the CBQ model

	Prediction accuracy	
	Homogeneous data	Heterogeneous data
Ordinary logit	43.0%	40.5%
CBQ-Q1	42.0%	46.0%
CBQ-Q2	42.5%	45.5%
CBQ-Q3	42.5%	44.0%
CBQ-Q4	42.5%	42.5%
CBQ-Q5	43.0%	40.0%
CBQ-Q6	43.0%	40.5%
CBQ-Q7	43.5%	40.5%
CBQ-Q8	44.0%	41.0%
CBQ-Q9	44.5%	40.5%
Proportion of equal or better predictions	5/9	8/9

Overall, this simple simulation study shows that even with a DGP other than the CBQ, with very similar performance, the CBQ model can approximate well the true distribution. Most importantly, the CBQ model outperforms the conditional

mean-based models, that is, the ordinary logit in this example, when unobserved heterogeneity is an inherent nature of the dataset. For real-world examples and a comparison with additional models, the following application section replicates three studies on the EU legislature, the 1992 US presidential election and government formation.

4   Applications

The above simulation study shows how the CBQ model can approximate the true DGP and how the model accounts for unobserved heterogeneity. To lend more intuition and incorporate substantive interpretations, this section applies the model to real-world examples from a range of political studies. The first application focuses on the EU codecision legislative procedure in the period between 1999 and 2004: for this procedure, the author tries to explain the drastic rising of the first reading agreements after the introduction of the Amsterdam treaty. The second leads us to the US presidential election in 1992, and the authors endeavor to understand the final win of Bill Clinton in competition with George H. W. Bush and Ross Perot. The final application focuses on government formation with respect to over 220 formation opportunities in 14 countries from 1945 to 1987. The order of the applications follows the increasing complexity of the data structure: in the EU legislature case, only two alternative choices are available, namely, the first reading agreement or not. In the 1992 US presidential election, the voters faced three choices among Clinton, Bush and Perot. Finally, the dataset on government formation consists of a different number of potential governments across countries and elections. As a result, these

studies provide the opportunity to compare the CBQ model with different types of discrete choice models, namely, the ordinary logit model of the EU legislature, the multinomial probit model of the US presidential voting, and the CL and MXL models of government formation. The CBQ model identifies among heterogeneous observations which subgroups the conditional mean-based models fail to account for. In addition to overall better predictive performance in those applications, the CBQ model shows supremacy over different types of conditional mean-based models in explaining unobserved heterogeneity.

Refraining from lengthy discussions of all aspects of replication studies, I delegate the full estimates to the Appendix and focus on substantive interpretations of some key variables in the main text. In these replication studies, I address unobserved heterogeneity, which conditional mean-based models fail to handle. To briefly summarize the results, compared to the average bills, EU colegislators are less concerned about delegation costs when a bill is very likely or unlikely to be concluded in the first reading stage. In the 1992 US presidential election, the levels of heterogeneity vary among voters with different characteristics when voting for a particular candidate. In particular, the preference for heated topics and party-voter alignment influence voter heterogeneity in vote choice. While the main results of the original study remain, the replication by the CBQ model indicates that voter heterogeneity can also induce heterogeneity in the predicted vote shares. In regard to government formation, substantive interpretations from counterfactual scenarios show that the potentially promising coalitions and misery coalitions respond differently to the status change of the largest party.

To sufficiently capture unobserved heterogeneity, all of the following applications use a nine-quantile specification of the CBQ model. Abbreviations Q1 to Q9 denote quantiles from 0.1 to 0.9. All CBQ estimations are performed with five parallel chains, except for cases that reach the maximum tree depth of the sampler. In those cases, the number of chains is reduced to four or, in the most extreme cases, two to avoid infinite computational time. The computational facility used is a high-performance computing cluster, with each node equipped with two Intel Xeon E5-2640v3 CPUs, 128 GB working memory and a 128 GB local disk (solid-state drive). Visual and numerical examinations all indicate that the chains are properly converged.

4.1   EU legislature

The Maastricht Treaty in 1992 established a codecision legislative procedure in which the European Council had to negotiate with the European Parliament (EP) in order to pass a bill. While the introduction of the codecision procedure empowered the EP to enhance the democratic process, the Amsterdam treaty in 1999 amended the procedure to allow legislation to be concluded as early as in the first reading stage. This reform raised trade-offs between democracy and efficiency and provided colegislators alternative choices between concluding a bill earlier or later. Against this context, Rasmussen (2010) studies the conditions under which colegislators decide to have an early conclusion. She proposes three sets of hypotheses on early conclusion, namely, the delegation costs of the rapporteurs from the EP, bargaining certainty and bargaining impatience. While the first factor is expected to reduce the

likelihood of a first reading agreement, the latter two factors are expected to increase this likelihood. Using 487 codecision bills between 1999 and 2004, the empirical results from the ordinary logit regression offer support for all three hypotheses.

In this application, the ordinary logit and ordinary probit models provide almost identical results, except for the differences in the scale of estimates. The IIA assumption is also not relevant here because only two alternatives are available. The key issues lie in the identification of unobserved heterogeneity. Although the author controls for a group of variables relating to the characteristics of the bills, this provides no guarantee that unobserved heterogeneity among bills has been well counted. In particular, those control variables provide only very rough indicators for the features of the bills. To identify unobserved heterogeneity, I replicate the full model (Model III in Table 2 and Figure 4 of the original paper) using the CBQ model, with quantiles ranging from 0.1 to 0.9. Compared to the 76.6% prediction accuracy of the ordinary logit model, the prediction accuracies of the CBQ model in nine quantiles range from 75.1% to 77.7%. Five out of nine quantile predictions perform equal to or better than the ordinary logit model. In essence, if unobserved heterogeneity has been well controlled for, we should expect no large variation among CBQ estimates across quantiles. Here, I illustrate the effects of the delegation costs of the rapporteurs from the EP on the probability of a first reading agreement. The full estimates are delegated to the Appendix.

As is clear from the left panel of Figure 3, we observe considerable variation among the effects of the interaction between the big party group and the distance between the rapporteur and the EP median.[4] In particular, CBQ estimates at

_____

[4]The variable "policy distance rapporteur - EP median" alone shows no significance over all

quantiles 0.1 and 0.9 show no significance. Since the CBQ and the ordinary logit estimates are incomparable in scale, I plot the predicted probabilities of different models against each other in the right panel of Figure 3. According to the plot, we clearly observe the difference between the estimates from the lowest (CBQ-Q1) and highest quantiles (CBQ-Q9) and those from the ordinary logit and the rest of the CBQ model.
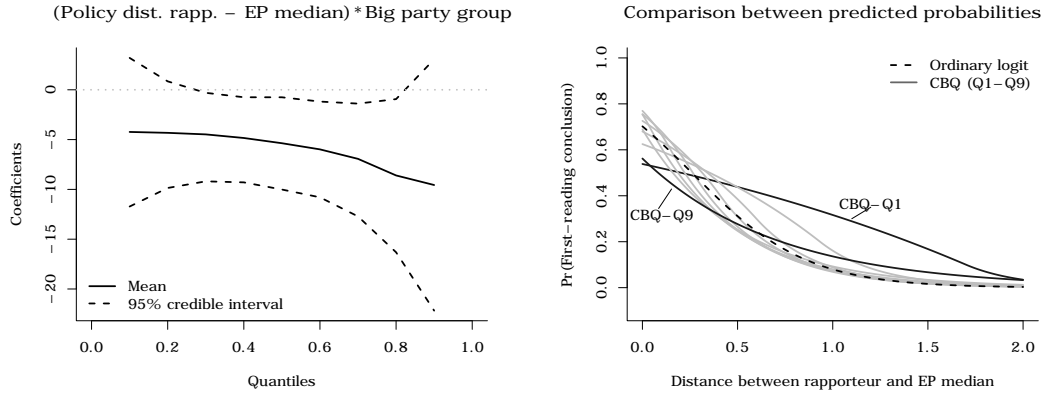

Figure 3: Coefficients and predicted probability of distance between the rapporteur and the EP median

The predicted probabilities at quantiles 0.1 and 0.9 have the smallest reductions when delegation costs are measured by the distance, which can increase from 0 to 2, between the rapporteur and the EP median; this indicates that EU colegislators are less concerned about delegation costs when the bills are very likely or unlikely to be passed early on in nature. Such bills, for example, include those that are of an administrative nature and thus need no further negotiation or those on which

quantiles and therefore is not plotted here.

26

it is difficult for member states to reach compromise, and they consequently reach no agreement at all. Such bill characteristics are not captured by the specified variables and induce unobserved heterogeneity. In summary, even in the simplest two-alternative case, the presence of unobserved heterogeneity in the dataset biases the estimates by the conditional mean-based models. In contrast, the CBQ model successfully captures this unobserved heterogeneity.

4.2 US presidential election in 1992

Alvarez and Nagler (1995) study the 1992 US presidential election with three candidates: Clinton, Bush and Perot. Using a subsample of 909 respondents from the 1992 American National Election Study (Miller et al., 1993), they offer thorough coverage of the mainstream explanations for Clinton's success, namely, economic voting, spatial voting and so-called "angry voting."[5] Their results from a multinomial probit model show that economic perception is the dominant factor determining the election outcome. While issue and ideological distances between the electorate and the candidates have some explanatory power, they cannot alter the result. Finally, the "angry voting" hypothesis finds no empirical support from their analysis.

Compared to the ordinary logit model, the multinomial probit model used in the paper allows the simultaneous analysis of multiple choices. More importantly, it offers a correlation structure between random components of the utility function and therefore eliminates the restrictive IIA assumption (Hausman and Wise, 1978). The IIA assumption is relevant in this example because the availability of the third choice

---

[5]In this subsample used for estimation, Clinton received 45.8% share of votes, Bush 34.1% and Perot 20.0%, which is close to real outcomes.

alternative may alter the voters' utility, which will not be captured if IIA holds. Nevertheless, the correlation structure embedded in the multinomial probit model only differentiates preferences among choice alternatives and does not differentiate the heterogeneous preferences of individuals for each choice alternative.

I replicate their analysis using the CBQ model and compare the result with that from the multinomial probit model.[6] The prediction accuracies of the CBQ model at different quantiles range from 74.4% to 75.6%, which are all better than the 74.1% of the multinomial probit. The overall results of the CBQ model confirm the sign and significance of the original analysis (see the coefficient plots in the Appendix).

However, the CBQ estimates show considerable variation across quantiles, which is a clear indicator of the existence of unobserved heterogeneity in the sample. To quantify unobserved heterogeneity among voters for each choice alternative, I calculate the standard deviation of nine quantile estimates for each covariate that is related to the voters' characteristics.[7] As shown in Table 2, the largest differences between unobserved heterogeneity of the two choice alternatives appear in the variables *Abortion*, *Democrat*, *Republican* and *Respondents' education*. Everything else being equal, preference heterogeneity is low for voters who identify themselves as Democrat when voting for Clinton. The same is true for voters who identify themselves as Republican when voting for Bush. This finding is in accordance with the

---

[6]In the original paper, Alvarez and Nagler (1995) mistakenly code the ideological placement of the three candidates. I use the corrected version that they provided, which slightly increased the prediction accuracy of the multinomial model from 74.0% to 74.1%. The original analysis was run in GAUSS. I replicated their analysis using an R package called "mlogit", which provides multinomial probit functionality and produces very similar estimates (Croissant et al., 2012).

[7]The choice alternative Perot is set as the reference category, as in the original analysis, and therefore has no estimates.

literature on party alignment (see, e.g., Miller, 1991). However, when deciding on candidates from other parties, there is considerably high preference heterogeneity among voters. In regard to abortion, a heated topic during the election, pro-abortion voters have more coherent preferences in voting for Clinton rather than for Bush. This makes sense given the unequivocal opposition of pro-abortion voters against Bush. Finally, even though educated voters tend to vote for Bush, they show higher heterogeneity in voting for Bush than in voting for Clinton.

Table 2: Voter heterogeneity (calculated as the standard deviation of the nine quantile estimates for each covariate)

Covariates	Alternatives	
	Bush	Clinton
Felt personal finances were worse	0.025	0.024
Felt national economy was worse	0.085	0.138
Oppose government jobs	0.074	0.020
Oppose government health care	0.084	0.027
Oppose government minority assistance	0.044	0.139
Abortion	0.280	0.032
Region (East)	0.242	0.353
Region (South)	0.333	0.397
Region (West)	0.177	0.221
New or returning voter	0.247	0.238
Term limits	0.155	0.102
Felt deficit was a major problem	0.554	0.271
Democrat	2.657	0.512
Republican	0.779	2.723
Gender (Female)	0.354	0.132
Respondents' education	0.096	0.019
Age: 18-29	0.667	0.412
Age: 30-44	0.491	0.381
Age: 45-59	0.464	0.144

In addition to voter characteristics, unobserved heterogeneity among voters can also impact the effect of the ideological placement of the candidates on vote shares. To conduct a substantive interpretation, I focus on the effects of Bush's ideological

placement on the candidates' vote shares. Corresponding to Figure 1A in the original paper by Alvarez and Nagler (1995), Figure 4 compares the predicted vote shares of the three candidates by varying Bush's position while keeping the other variables constant. The red lines are the predictions by the multinomial probit model, and the black lines are those by the CBQ model. The red dots in the plot represent the real vote share in the sample of the study. Shown from top to bottom are Clinton, Bush and Perot. While the overall predictions are similar, the plot shows that due to voter heterogeneity, the effects of the candidates' ideological movement also vary. The largest variation occurs when voters decide to vote for or against Perot. By comparing the predicted vote shares and the real vote shares, it is clear that the multinomial probit estimator fails to account for unobserved heterogeneity among voters. The multinomial probit model is among the worst predictors when Bush's position is set at the actual value: 5.32. The best predictive quantile is 0.9. This means that loyal voters, who have a high likelihood of voting for the corresponding candidate, have a larger influence over the final result than do the rest of the voters. The underestimation of the multinomial model occurs because loyal voters have different preferences than the rest of the voters, and the multinomial model estimates only the average preferences.
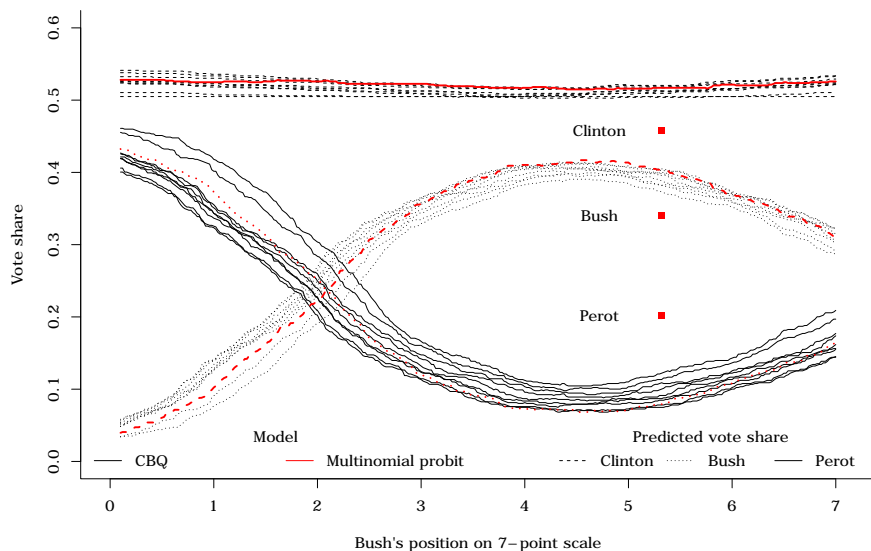
Figure 4: Predicted vote shares of the three candidates by varying Bush's position (corresponding to Figure 1A in Alvarez and Nagler (1995)).

4.3 Government formation

In parliamentary democracies, government formation bridges the gap between the electoral outcome and government practice. Given electoral outcomes, competing parties must negotiate among themselves about who will constitute the elected government. Martin and Stevenson (2001) (hereafter MS) generated one of the most comprehensive government formation datasets, which contains in total 220 formation opportunities and 33,256 potential governments in 14 countries between 1945 and 1987. This dataset provides a valuable opportunity to check the existing explanations of government formation across countries.

Compared to previous applications, the data structure of government formation is the most complicated. In the analysis of government formation, the numbers of

31

coalition alternatives differ across countries and elections, and the characteristics of potential governments within each formation opportunity can be intercorrelated. For example, the status of the minority coalition and the minimal winning coalition may depend on the inclusion or exclusion of the largest party in the coalition; the coalition with the previous prime minister can be correlated with the status of the incumbent coalition, and so on. Therefore, when estimated by the traditional discrete choice model with equal weights over all alternatives, a formation opportunity with a large number of coalition alternatives will bias the likelihood of others.

In terms of the IIA assumption, as the dataset of MS contains so many government alternatives, it is difficult to comprehend the violation of the IIA assumption. Compared to vote choice, the violation of IIA in the case of government formation does not come from any concrete utility functions of particular actors. Instead, it comes from the ratio of probabilities of government alternatives that is operationalized by a logit map from a set of government characteristics to a probability space. The intuition has already been stated elsewhere (see, e.g., Martin and Stevenson, 2001; Glasgow et al., 2012), and it is worth some additional notes for clarity. According to the logit formula, we know that the probability ratio between two governments depends only on the difference between the two governments' characteristics and remains independent of the characteristics of other government alternatives. Therefore, the IIA assumption is violated when unobserved government characteristics are correlated. This is the case when, for example, some government alternatives are considered substitutes or complements to each other for reasons that are unobserved by researchers. With a growing number of government alternatives, the

likelihood that the unobserved characteristics of government alternatives are correlated also increases. In summary, while vote choice is theorized to be made out of some latent utility functions, government formation can also be theorized to result from some latent characteristic functions of government. While voters choose candidates/parties that give them the highest utilities, governments are formed when they have the "best characteristics" compared to those of other alternatives. Therefore, when unobserved government characteristics are correlated, the IIA assumption is violated.

Attempting to overcome the problem of unbalanced numbers of choice alternatives, MS use the CL model, whose estimates depend on each formation opportunity (choice set). However, the CL model relies on the strong IIA assumption, which is often violated in practice. In another replication study, Glasgow et al. (2012) (hereafter GGG) apply an MXL model to relax the restrictive IIA assumption. They notice some differences between their results and those of MS. However, both models assume a logistic form of the underlying error distribution and are unable to assess the full conditional distribution of the dependent variable. When unobserved heterogeneity exists in the dataset, both estimators produce biased estimates.

I replicate the study by MS using the CBQ model, which is intended to handle both unbalanced numbers of choice alternatives and the restrictive IIA assumption. In addition, the CBQ model accounts for unobserved heterogeneity by specifying nine quantile estimators. By delegating the resulting estimates to the Appendix, I focus on an in-depth interpretation of a selected number of factors in the main text.

Since most potential governments are not formed, prediction accuracy is inflated

when all potential governments are included. With the increasing number of choice alternatives, the prediction accuracy will automatically increase, although nothing else has been changed. Indeed, the prediction accuracies of all models are close to 99.2% when all potential governments are considered. To overcome this problem, I compare the predicted probabilities and the prediction accuracies for those governments that were actually formed. For 220 actually formed governments, the prediction accuracies of the CL and MXL models are 40.5% and 40%, respectively. The prediction accuracies of the CBQ model range from 34.5% to 42.3%. The best predictive quantiles are 0.1 and 0.9, with prediction accuracies of 42.3% and 40.9%, respectively, and the worst predictive quantiles are actually in the middle. This indicates that the estimates close to the middle of the conditional distribution are unrepresentative of the whole sample. Therefore, the conditional mean-based models such as CL and MXL fail to capture unobserved heterogeneity of potentially promising (high quantiles) and very unpromising governments (low quantiles).

Due to the complicated data structure, the traditional counterfactual analysis, as previously illustrated, cannot be easily performed under the context of government formation. As pointed out by GGG, performing a counterfactual analysis by simply varying a single variable while keeping others constant is likely to produce illogical results. Therefore, following their practice, I produce logically coherent counterfactual scenarios by switching the largest party to the second largest party, which consequently leads to changes in the features of other potential governments. The predicted probabilities of each model are based on the new counterfactual dataset, which is logically coherent. Figure 5 plots the substitution patterns and compares

34

changes in the predicted probabilities between CBQ and CL and between CBQ and MXL. In the plot, the best predictive quantiles—Q1 (0.1 quantile) and Q9 (0.9 quantile)—are used in comparison. The upper panels show the comparison between CBQ and CL, and the lower panels show the comparison between CBQ and MXL. The hollow points represent pairs of changes in the predicted probabilities when the largest party becomes the second largest party. Black points are pairs whose differences are statistically significant at the 95% confidence level, and gray points are those that are insignificant. The dashed lines are 45-degree lines on which the changes in probabilities are the same. We can observe that the patterns between the upper and lower panels are very similar and that the main difference comes from comparison with CBQ-Q1 and CBQ-Q9. For those significantly different pairs, CBQ-Q1 predicts, on average, a negative change in probabilities, while CL and MXL predict almost no change. In contrast, CBQ-Q9 predicts no change in probabilities, while CL and MXL predict, on average, a negative change. The differing substitution patterns between Q1 and Q9 indicate that potentially promising coalitions and misery coalitions respond differently than average governments do when the largest party becomes the second largest party. In particular, the promising coalitions suffer little from losing the largest party, while the misery coalitions' chances to be in the government become less likely. This heterogeneity among different types of coalitions is not captured by the conditional mean-based CL or MXL models.
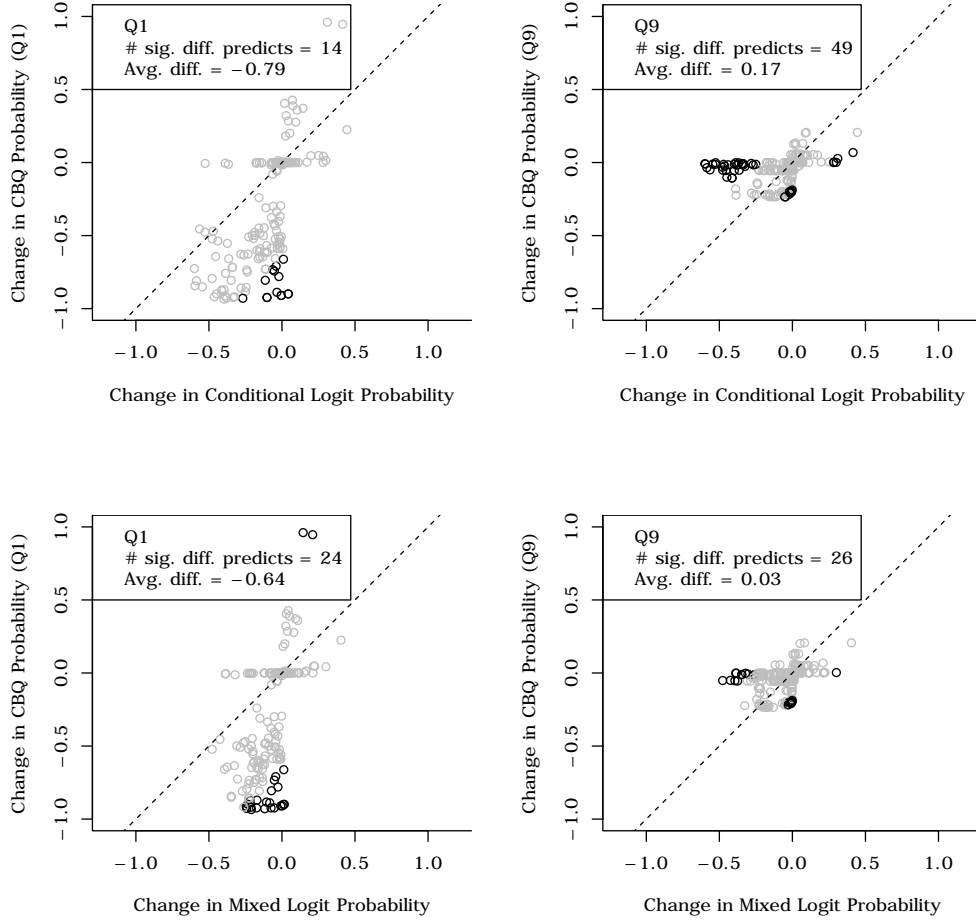
Figure 5: Comparing the CBQ, CL and MXL substitution patterns. The dashed lines are the 45-degree equal-division lines. Black points represent the pairs that are significantly different at a 95% confidence level while the gray points are insignificant. Q1 and Q9 indicate 0.1 and 0.9 quantile estimators, respectively. The average difference is calculated based on significantly different pairs.

36

Figure 6 shows the kernel density of change in the predicted probabilities when the largest party becomes the second largest. From the top to the bottom are the CL, MXL, CBQ-Q1 and CBQ-Q9 predictions. While the CL and MXL predictions of the counterfactual scenarios are similar, they show a clear difference from the CBQ-Q1 and CBQ-Q9 predictions. CBQ-Q1 has a much wider range of change in the predicted probability, while CBQ-Q9 is more concentrated at the center. In accordance with the substitution patterns in Figure 5, the change in the predicted probability from CBQ-Q1 has a fat tail at the negative values.
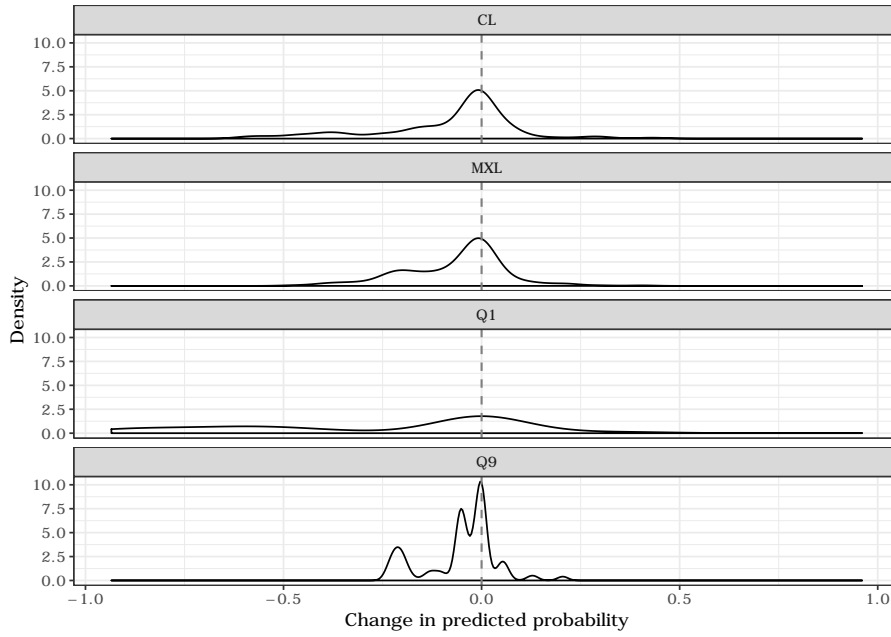

Figure 6: Change in predicted probabilities when the largest party becomes the second largest

Overall, the results from the CBQ model suggest that the previous analyses fail to account for unobserved heterogeneity among potential governments. Most variables show considerable variation across the nine quantile estimates. According

to the counterfactual scenarios, unobserved heterogeneity among coalitions leads to substantively different predictions between different kinds of coalitions. In particular, I find that compared to the average coalitions, the promising coalitions and the misery coalitions have a different response to the status change of the largest party.

5   Conclusion

Despite its sparse usage in political science, quantile regression has demonstrated great potential in producing complete and robust analyses in studies on, for example, party politics, legislative decision making, voting behavior, international organizations, and trade policies, among others, and continues to gain popularity in various fields of political science (see, e.g., Rosenberg et al., 2017; Ratkovic and Tingley, 2017; Cantú, 2014; Betz, 2017). However, most applications focus on the continuous response variable and ignore a large and important branch of data containing discrete response variables. Discrete response variables, such as the occurrences of wars, the voting behavior of electorates, a failure or passage of a legislative bill, agreements of international organizations, and the implementation of trade policies, are often core research interests in political science.

As far as I am aware, this paper is the first to provide an approach to estimate quantile regression in a discrete choice setting with alternative-specific features. The conditional mean-based methods fail to handle unobserved heterogeneity across units. Additionally, the existing binary quantile models are applicable to only the simplest binary outcome settings and fail to account for choice-specific features in each choice set. This paper attempts to bridge the gap by introducing the CBQ

regression that describes the full conditional properties of the response.

Since the behavior of any random variable is governed by its distribution, and the fitness of any model to the data revealed from random variables is therefore affected by the full characteristics of the underlying distribution, quantiles provide a much more thorough picture of the data than the conditional mean does. To cite Mosteller and Tukey (1977): "What the regression curve does is (to) give a grand summary for the averages of the distributions corresponding to the set of xs. We could go further and compute several different regression curves corresponding to the various percentage points of the distribution and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture." Instead of having an *ad hoc* assumption of a certain error distribution, which can rarely be justified by the theoretical arguments, the CBQ approach models the error term more flexibly and adjusts the quantiles to fit the data. When the error distribution is unobserved or unknown to the researcher, the CBQ model is more resistant to the misspecification of error distributions. In summary, compared to the conditional mean-based models, the CBQ model provides a much more robust and comprehensive estimator, even when the features of alternatives within each choice set vary.

By applying the CBQ model to a range of political studies, the results show interesting variations with the previous analyses. Accordingly, the conditional mean-based discrete choice models fail to account for unobserved heterogeneity, which is well captured by the CBQ model. From the analyses, we observe that particular subgroups of a population may have different preferences and behavior compared to

that of the average.

I admit that the adoption of the CBQ model comes with a mild efficiency loss when the error distribution of the latent utility is known to the researchers. Under such situations and when the researcher is solely interested in the conditional mean of the response variable, it is of course better to use the model with the known error distribution without relying on the CBQ model. However, it is often the case in practice that we do not know from empirics or theories the true form of the error distribution in a discrete choice setting, and the choice of modeling the conditional mean is mostly *ad hoc*. In this case, it is always better to use the CBQ model at least as a robustness check than to simply apply conditional mean-based models with an *ad hoc* specified distribution. It is also important to note that the ultimate goal of the CBQ model is to reveal unobserved heterogeneity rather than to obtain superior predictive performance. While the prediction accuracies of each quantile differ due to differing proportions of heterogeneous subgroups, the CBQs as a whole provide a rich picture of unobserved heterogeneity within a population.

The model can be further extended to model dynamic choices in longitudinal data and strategic interaction among actors. When the number of covariates is large or when researchers are uncertain about the composition of the utility function, penalty terms can also be added to the model to automate variable selection. I leave these as agendas for future research.

References

Alhamzawi, R. and K. Yu (2013). Conjugate Priors and Variable Selection for Bayesian Quantile Regression. *Computational Statistics and Data Analysis 64*, 209–219.

Alhusseini, F. H. H. (2017). Bayesian Quantile Regression with Scale Mixture of Uniform Prior Distributions. *International Journal of Pure and Applied Mathematics 115*(1), 77–91.

Alvarez, R. M. and G. Glasgow (1999). Two-stage Estimation of Nonrecursive Choice Models. *Political Analysis 8*(2), 147–165.

Alvarez, R. M. and J. Nagler (1995). Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election. *American Journal of Political Science 39*(3), 714–744.

Benoit, D. F., R. Alhamzawi, and K. Yu (2013). Bayesian Lasso Binary Quantile Regression. *Computational Statistics 28*, 2861–2873.

Benoit, D. F. and D. V. Poel (2012). Binary Quantile Regression: A Bayesian Approach based on the Asymmetric Laplace Distribution. *Journal of Applied Economics 27*, 1174–1188.

Betz, T. (2017). Trading Interests: Domestic Institutions, International Negotiations, and the Politics of Trade. *The Journal of Politics 79*(4), 1237–1252.

Cantú, F. (2014). Identifying Irregularities in Mexican Local Elections. *American Journal of Political Science 58*(4), 936–951.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software 76*(1), 1–32.

Carter, D. B. and C. S. Signorino (2010). Back to the Future: Modeling Time Dependence in Binary Data. *Political Analysis 18*(3), 271–292.

Croissant, Y. et al. (2012). Estimation of Multinomial Logit Models in R: The mlogit Packages. *R package version 0.2-2. URL: http://cran. r-project. org/web/packages/mlogit/vignettes/mlogit. pdf*.

Delgado, M. A., J. M. Rodrıguez-Poo, and M. Wolf (2001). Subsampling Inference in Cube Root Asymptotics with An Application to Manski's Maximum Score Estimator. *Economics Letters 73*(2), 241–250.

Florios, K. and S. Skouras (2008). Exact Computation of Max Weighted Score Estimators. *Journal of Econometrics 146*, 86–91.

Glasgow, G. (2011). Introduction to the Virtual Issue: Recent Advances in Discrete Choice Methods in Political Science. *Political Analysis 19*(1), 1–3.

Glasgow, G., M. Golder, and S. N. Golder (2012). New Empirical Strategies for the Study of Parliamentary Government Formation. *Political Analysis 20*, 248–270.

Hausman, J. A. and D. A. Wise (1978). A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica 46*(2), 403.

Hensher, D. A. and W. H. Greene (2003). The Mixed Logit Model: the State of Practice. *Transportation 30*(2), 133–176.

Hensher, D. A., J. M. Rose, and W. H. Greene (2015). *Applied Choice Analysis (Second Edition)*. Cambridge: Cambridge University Press.

Hoffman, M. D. and A. Gelman (2014). The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Horowitz, J. L. (1992). A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica 60*(3), 505–531.

Kim, J. and D. Pollard (1990). Cube Root Asymptotics. *The Annals of Statistics 18*(1), 191–219.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression Quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and K. F. Hallock (2001). Quantile Regression. *Journal of Economic Perspectives 15*(4), 143–156.

Koenker, R. and J. A. F. Machado (1999). Goodness of Fit and Related Process for Quantile Regression. *Journal of American Statistical Association 94*(448), 1296–1310.

Kordas, G. (2006). Smoothed Binary Regression Quantiles. *Journal of Applied Economics 21*(3), 387–407.

Kotlyarova, Y. (2005). *Kernel Estimators: Testing and Bandwidth Selection in Models of Unknown Smoothness.* Ph. D. thesis, McGill University.

Kotz, S., T. J. Kozubowski, and K. Podgorski (2001). *The Laplace Distribution and Generalizations: A Revisit with New Applications.* Springer Science+Business Media, LLC.

Kotz, S., T. J. Kozubowski, and K. Podgorski (2003). An Asymmetric Multivariate Laplace Distribution. *Working Paper*, 1–26.

Kozumi, H. and G. Kobayashi (2011). Gibbs Sampling Methods for Bayesian Quantile Regression. *Journal of Statistical Computation and Simulation 81*(11), 1565–1578.

Lu, X. (2019). Replication data for "Discrete Choice Data with Unobserved Heterogeneity: A Conditional Binary Quantile Model", https://doi.org/10.7910/DVN/1WZCEA, Harvard Dataverse, DRAFT VERSION, UNF:6:LLGCgS7NqQh4oXw4N5b52A==[fileUNF].

Manski, C. F. (1975). Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics 3*, 205–228.

Manski, C. F. (1985). Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics 27*(3), 313–333.

Martin, L. W. and R. T. Stevenson (2001). Government Formation in Parliamentary Democracies. *American Journal of Political Science 45*(1), 33–50.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

McFadden, D. (1986). The Choice Theory Approach to Market Research. *Marketing Science 5*(4), 275–297.

McGrath, L. F. (2015). Estimating Onsets of Binary Events in Panel Data. *Political Analysis 23*(4), 534–549.

Miller, W. E. (1991). Party Identification, Realignment, and Party Voting: Back to the Basics. *American Political Science Review 85*(2), 557–568.

Miller, W. E., D. R. Kinder, and S. J. Rosenstone (1993). The National Election Studies. 1993. American National Election Study, 1992: Pre-and Post-election Survey CPS Early Release Version, computer file. *Ann Arbor: University of Michigan, Center for Political Studies, and Inter-University Consortium for Political and Social Research.*

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods. Pearson Education.

Oh, M.-s., E. S. Park, and B.-S. So (2016). Bayesian Variable Selection in Binary Quantile Regression. *Statistics and Probability Letters 118*, 177–181.

Poole, K. T. (2001). Nonparametric Unfolding of Binary Choice Data. *Political Analysis 8*(3), 211–237.

Rainey, C. (2016). Dealing with Separation in Logistic Regression Models. *Political Analysis 24*(3), 339–355.

Rasmussen, A. (2010). Early Conclusion in Bicameral Bargaining: Evidence from the Co-decision Legislative Procedure of the European Union. *European Union Politics 12*(1), 41–64.

Ratkovic, M. and D. Tingley (2017). Sparse Estimation and Uncertainty with Application to Subgroup Analysis. *Political Analysis 25*(1), 1–40.

Reed, C. and K. Yu (2009). A Partially Collapsed Gibbs Sampler for Bayesian Quantile Regression. *Technical report, Department of Mathematical Sciences, Brunel University London.*

Rosenberg, A. S., A. J. Knuppe, and B. F. Braumoeller (2017). Unifying the Study of Asymmetric Hypotheses. *Political Analysis 25*(3), 381–401.

Sartori, A. E. (2003). An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions. *Political Analysis 11*(2), 111–138.

Sriram, K., R. Ramamoorthi, P. Ghosh, et al. (2013). Posterior Consistency of Bayesian Quantile Regression based on the Misspecified Asymmetric Laplace Density. *Bayesian Analysis 8*(2), 479–504.

Traunmüller, R., A. Murr, and J. Gill (2014). Modeling Latent Information in Voting Data with Dirichlet Process Priors. *Political Analysis 23*(1), 1–20.

Yu, K. and R. A. Moyeed (2001). Bayesian Quantile Regression. *Statistics & Probability Letters 54*, 437–447.